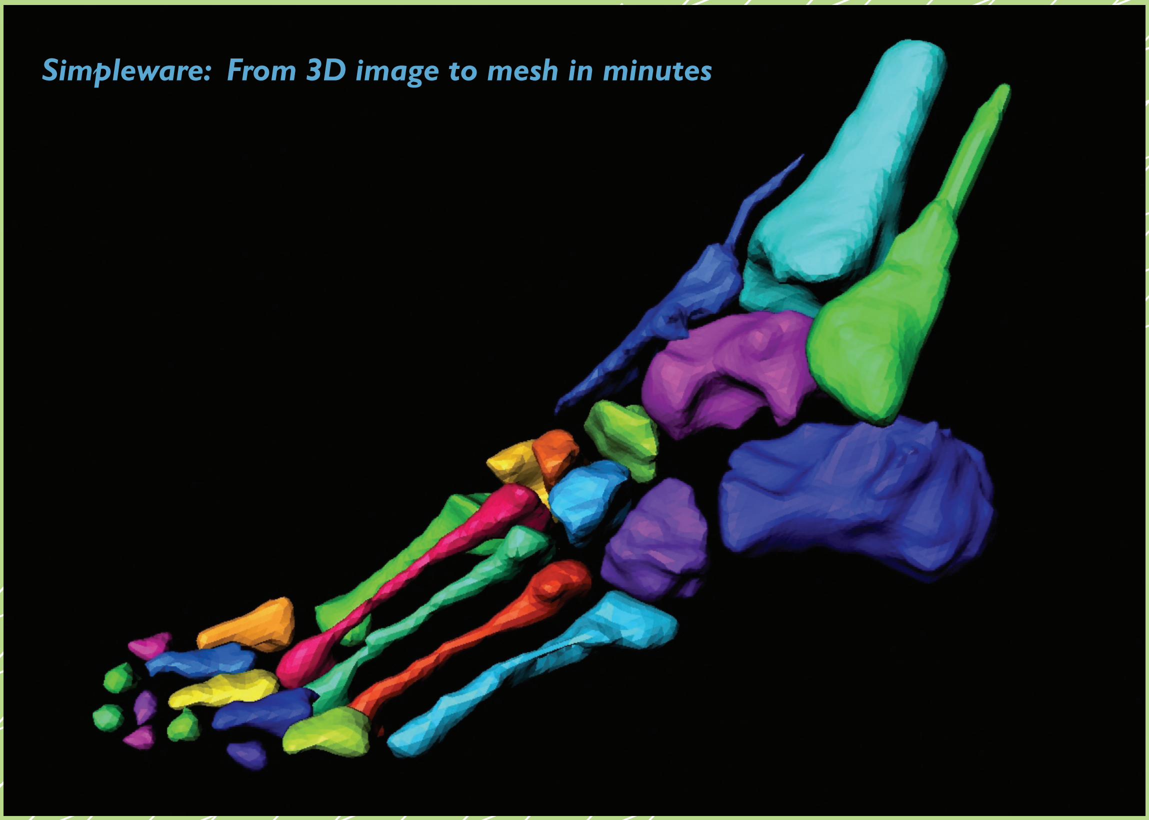
# CSAR FOCUS

Edition 14
Autumn - Winter 2005

high performance computing at Manchester ~ taking research into the future

www.csar.cfs.ac.uk



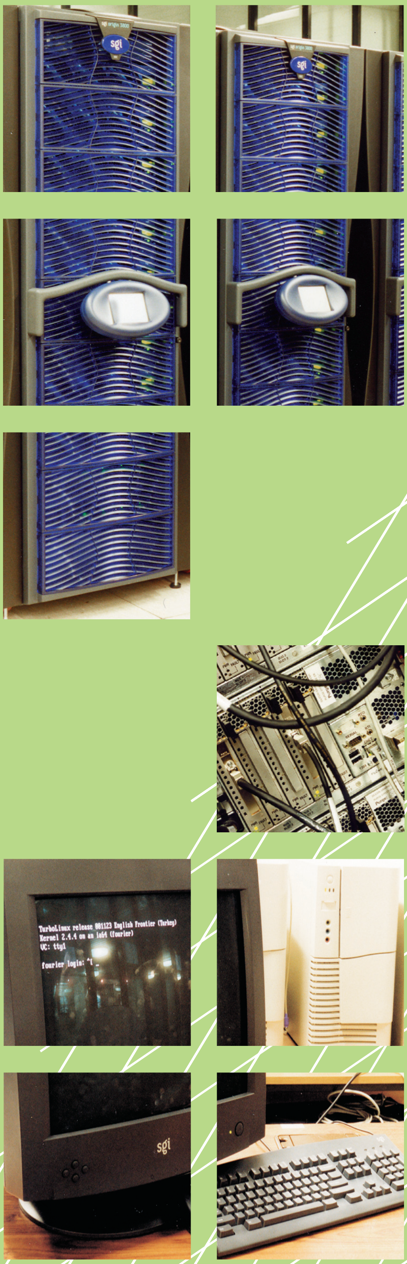
Simpleware: From 3D image to mesh in minutes

MANCHESTER
1824
The University of Manchester

CSC

sgi

# Contacts



## CSAR Helpdesk

Telephone: 0161 275 5997 / 275 6824
Fax: 0161 275 6800
Email: csar-advice@cfs.ac.uk

## Terry Hewitt

*Deputy Director, Manchester Computing, University of Manchester*

Telephone: 0161 275 6095
Fax: 0161 275 6800
Email: w.t.hewitt@manchester.ac.uk

## Dr Mike Pettipher

*Head of Scientific Computing Services, Manchester Computing, University of Manchester*

Telephone: 0161 275 6063
Fax: 0161 275 6800
Email: m.pettipher@manchester.ac.uk

## Claire Green

*Editor of CSAR Focus, Research Support Services, Manchester Computing, University of Manchester*

Telephone: 0161 275 5997
Fax: 0161 275 6800
Email: claire.l.green@manchester.ac.uk

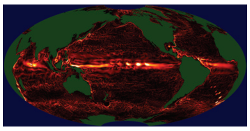
## Copyright

*All pictures are reproduced by permission. Any further reproduction will require the owner's permission.*

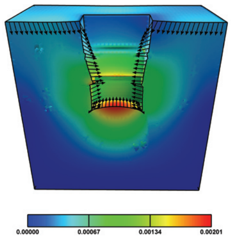
## Disclaimer

*Articles are the responsibility of the individual author. The Editor, CSAR and the University of Manchester assume no responsibility or liability for articles or information appearing in this publication. Opinions expressed in some articles may not reflect those of CSAR.*

# Contents



Page 10



Page 27

# Editorial

Welcome to the penultimate edition of *CSAR Focus*. We are delighted to include articles from the CSAR user community on topics such as 3D imaging for use in Finite Element Analysis and using the CSAR systems to speed the training of neural networks modelling linguistic processes.

Walter F. Brooks, Chief of the NASA Advanced Supercomputing Division at Ames Research Center, details the successful deployment of *Columbia* - the largest shared memory system in the world - on pages 10-12. The system is currently being used for pioneering work and testing, specifically looking at large-scale processing and scaling on 2,000 processors and beyond and is already enabling groundbreaking computations including two promising climate models and a global-ocean circulation and sea-ice model, ECCO (Estimating the Circulation and Climate of the Ocean). Improved resolution and accuracy of global-ocean circulation and sea-ice estimates made possible by *Columbia* are yielding an increasingly accurate picture of the effect of the oceans on the Earth's climate.

On pages 22-24 we take a look at the Swiss national supercomputing centre, CSCS. Thanks go to Marie-Christine Sawley, the Chief Executive Officer for providing us with this insight and to all other authors who contributed to this edition. If you would like to contribute to the final edition of *CSAR Focus*, due out early in 2006, please contact me via the CSAR Helpdesk (csar-advice@cfs.ac.uk).

Claire Green
Editor, *CSAR Focus*

# New consortium aims to give the UK a world lead in research using High Performance Computing

***Press Release***
***Issued 20 September 2005***

The largest ever consortium to support UK academic research using high performance computers (HPC) is being established by the University of Edinburgh, the University of Manchester and the Council for the Central Laboratories of the Research Councils' (CCLRC) Daresbury Laboratory.

Professor Timothy O'Shea, Principal and Vice-Chancellor of the University of Edinburgh, Professor Alan Gilbert, President and Vice-Chancellor of the University of Manchester and Professor John Wood, Chief Executive of CCLRC today approved this joint venture, called HPC-UK, to couple their computational science and engineering support teams. These have supported all UK national academic HPC facilities for 20 years. From now on they will work together, pooling their complementary expertise, to provide UK researchers with unprecedented breadth and depth of support in the application of HPC technology to the most challenging scientific and engineering problems.

Professor O'Shea, said: "Edinburgh has pioneered novel computing solutions to challenging science problems for 25 years. HPC-UK is a natural next step that will enable us to continue to innovate while tackling the most ambitious projects." These range from climate change to personalised medicine, from understanding the fundamental building blocks of nature to the evolution of the whole universe, from modelling fusion reactors to financial forecasting. They affect all our lives and HPC is the key.

Professor Wood, said: "CCLRC has led HPC applications enabling predictive calculations of enzymes and catalysts, simulations of cell membranes, accurate models of combustion and aerodynamics, and detailed nutrient and fluid flow models of off-shore water quality. Already working with the key UK projects across the community, HPC-UK plans to engage with, and encourage, new and emerging disciplines." These demand a multidisciplinary approach: computer design, software engineering, numerical algorithm and visualisation expertise combine, often with several fields of science, to tackle the most complex problems. The consortium will grow to include new areas of activity, so that its expertise continues to match the requirements of UK research.

Professor Gilbert, said: "As the problems get more complex, and computers get bigger and more expensive, simulation is turning into 'big science' - becoming a multinational and uniquely multidisciplinary enterprise. HPC-UK is our response to keep the UK at the forefront." HPC-UK will ensure that UK researchers are best placed to exploit emerging international facilities and it will provide the UK with the capability to host them. It aims to place the UK at the top of nations exploiting this strategic technology.

**www.hpc-uk.ac.uk**

# Heterogeneous Catalysis at the Molecular Level

*Rudy Coquet, Edward L. Jeffery and David J.Willock*
*School of Chemistry, Cardiff University*

Heterogeneous catalysts provide efficient processes to produce many of the chemicals we take for granted in our everyday lives. By reducing the energy barrier to difficult chemical transformations they both reduce the time required for reactions and the size of chemical plants required. The reduced energy requirements for a catalysed reaction also allow us to obtain the desired product with less environmental impact than would be incurred using only non-catalysed reactions. This latter point is becoming ever more important with new catalysts which can improve product selectivity and play a role in green technologies continually in demand.

To obtain new catalysts we need as much information as possible on the operation of existing materials and an ability to find out what will happen if the catalyst structure or composition is altered. Even for a comparatively simple reaction such as the hydrogenation of acetone to propan-2-ol, the complexity of the surface reaction can be difficult to understand without the aid of computer simulation at the atomic level. The basic mechanism can be adapted from the Horiuti-Polanyi scheme for ethene hydrogenation over Pt group metals shown in Figure 1. Hydrogen adsorbs and dissociates on Pt surfaces to give atomically adsorbed hydrogen. From surface science temperature programmed desorption (TPD) measurements and vibrational spectroscopy acetone is thought to adsorb molecularly[1]. Surface hydrogen can then add across the C=O double bond to produce the secondary alcohol, propan-2-ol. The alcohol has a much lower affinity for the surface than the ketone and so readily desorbs and is detected as product.

The first thing to note about this type of scheme is that the detailed structure of the surface is rather limited; it is a place where the reactants congregate, are activated and react. However, this reaction will not occur in the gas phase at any appreciable rate without the presence of the catalyst and so the surface itself must be more actively involved than this picture suggests. Computer simulation at the atomic scale is now able to provide this more detailed picture and allow us to review experimental data in the light of calculations.

To tackle these sort of processes we need to be able to describe the bond forming and breaking processes taking place and so some level of quantum mechanical treatment of the electronic states of the system is required. We have chosen periodic DFT with a planewave basis set and gradient corrected functionals (PW91) as implemented in the VASP program[2]. In earlier calculations it has been shown that this approach is able to give a good description of the $H_2$ dissociation process on Pt group metals[3] and with the additional computer power available from Newton and HPCx we have now been able to study the ketone adsorption[4].

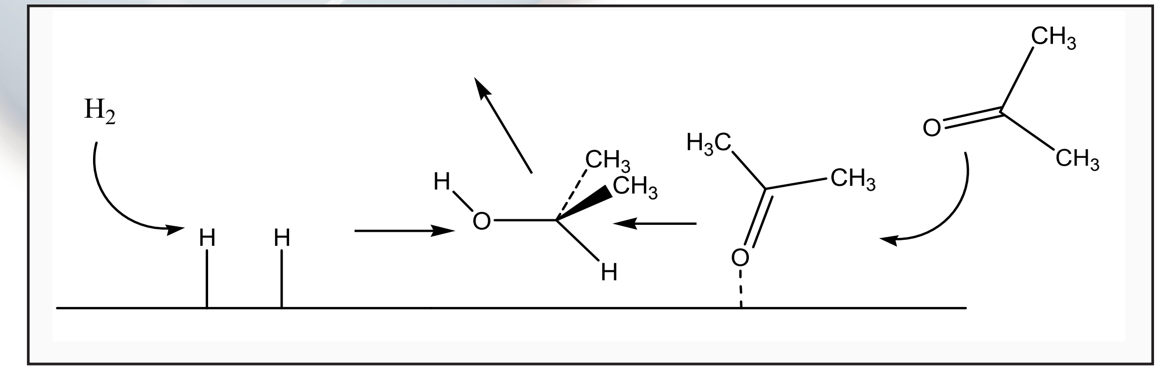


Figure 1: The Horiuti-Polanyi scheme adapted for ketone hydrogenation.

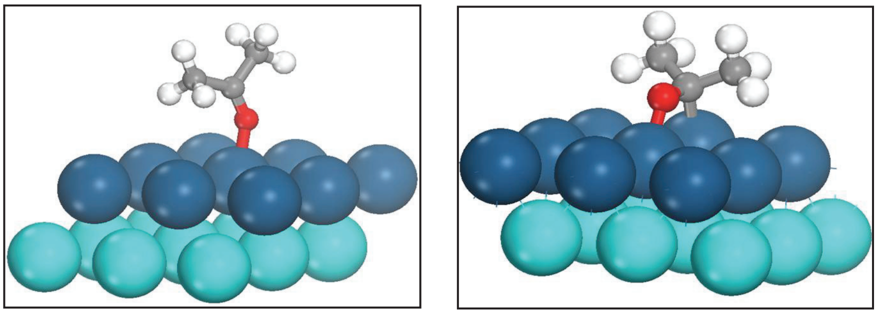

(a) (b)

Figure 2: Optimised structures for acetone keto-isomer adsorbed on Pt(111). (a) end-on, and (b) side-on. Only two layers of the three layer simulation slab are shown.

Experimental TPD data and vibrational spectroscopy[1] suggested that two species are present on the Pt(111) surface; the end-on ketone (Figure 2*a*) which desorbs at 184 K and shows vibrational modes only weakly perturbed by interaction with the metal and a second species which desorbs at the higher temperature of 199 K and gives a lower vibrational frequency for the C=O stretch vibration between 1511 and 1610 $cm^{-1}$. This second species was assigned to acetone in a side-on adsorption mode (Figure 2*b*).

The structures in Figure 2 are actually from our optimisation calculations which showed that the side-on adsorption has a lower binding energy to the surface than the end-on mode (21 *cf* 40 kJ $mol^{-1}$). This can be understood in terms of the steric interaction of the acetone methyl groups with the surface which is greater in the side-on than in the end-on case. However this contradicts the experimental result since the lower binding energy should correspond to a lower desorption temperature.

Vibrational spectroscopy is one area in which simulation can make direct contact with experiment and so we also calculated the adsorbate vibrational modes for comparison with the experimental spectra. We were able to confirm the assignment of the low temperature peak as the end-on adsorbate with a calculated CO stretch frequency of 1600 $cm^{-1}$ which compared well with the experimentally reported value of 1640 $cm^{-1}$. However a calculation of the vibrational frequencies for the side-on adsorbate gave a C=O stretch of 1161 $cm^{-1}$. The lowering in frequency compared to the end-on form can be attributed to interaction with the surface, both O and C form bonds to Pt atoms and so the C=O bond itself is weakened. This also leads to the loss of planarity of the adsorbate as can clearly be seen in Figure 2*b*. So the side-on adsorption mode shows both the wrong adsorption energy and vibrational data to be reconciled with experiment. This leaves the question of what surface species actually gives rise to the second signal observed in the experimental vibrational spectra?

To answer this we looked at isomerisation of the adsorbate. A molecule with C-H bonds adjacent to the C=O ketone group can undergo keto-enol isomerisation (Figure 3). In the gas phase the enol form of the molecule is only a tiny fraction of any given sample. Indeed we calculate an energy difference between the gas phase isomers of 50 kJ $mol^{-1}$. However the adsorption of the molecule to the surface lowers the energy of the system. This means that if the adsorption of the enol form to the surface is more favourable than the ketone by 50 kJ $mol^{-1}$ or more the balance could tip in its favour. Accordingly we studied the adsorption of enol and the de-protonated enolate forms of acetone on the surface. The most stable forms with energies quoted relative to gas phase ketone are shown in Figure 4. Both the enol and enolate are more favourable on the Pt surface than the end-on ketone form. In addition the enolate gives a C=O stretching frequency of 1574 $cm^{-1}$, within the experimental range for the mystery "species 2".

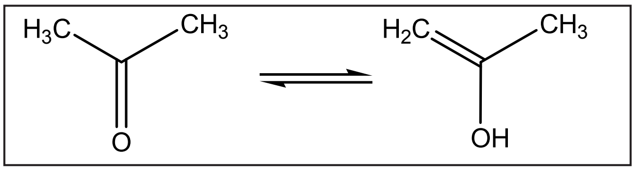
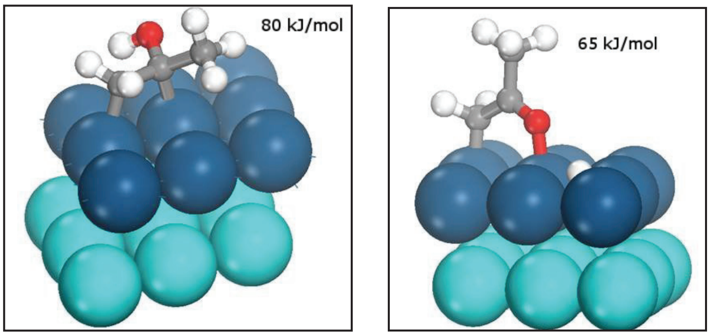


Figure 3: Isomers of acetone the ketone form (left) can undergo a proton transfer to give the enol (right).



(a) (b)

Figure 4: Optimised structures for (a) enol and (b) enolate with H adsorbed on Pt(111). Energies are quoted relative to the gas phase ketone with a positive value indicating favourable adsorption.

From these calculations we conclude that acetone adsorbing on a Pt surface will be present both as end-on and enol/enolate forms, *i.e.* the molecule will isomerise on the surface. This adds additional steps to the Horiuti-Polanyi scheme and suggests that hydrogenation may be across a C=C rather than a C=O double bond under conditions in which the isomerisation is faster than the hydrogenation of the ketone.

Most catalysts employ metal particles supported on oxides and so reactions may occur not only on the simple surfaces of the active metals but also at the interface between the oxide and metal particle. A case in point is the relatively new area of nano-particulate Au supported on various oxides. The particle size in this case is critical to the catalytic performance, bulk Au is inactive but particles on the nanometre scale have shown room temperature activity for the oxidation of CO to $CO_2$[5]. To study this chemistry the inclusion of the Au/oxide interface requires large unit cells to be constructed so that the periphery of the Au cluster can be included without interference between periodic images. To achieve this we are using the SIESTA[6] code to simulate a $Au_{10}$ particle supported on a $Mg_{96}O_{96}$ supercell. One difference between this and periodic metal simulations of larger particles is that charge transfer between Au and the support oxide can significantly influence the electronic character of the metal particles. This is particularly important if the oxide surface is defective. Figure 5 shows the charge transfer from a surface oxygen vacancy to the $Au_{10}$ cluster. This gives a cluster with an overall negative charge and we are currently investigating the effect this has on the adsorption and vibration of reactant molecules.

The calculations covered in this report require the consideration of many alternative positions on the surface for each isomer of a reactant. Each individual calculation of optimal adsorption geometry and vibrational frequencies are computationally intensive. The electron density for the metal surfaces is typically represented by a Fourier transform in the simulation cell requiring around 800,000 plane wave coefficients and the system sizes for the supported metal particles would have been impractical a few years ago. The large shared memory resources available on HPCx and Newton make such calculations possible on the required timescales. The accuracy of the calculations give us confidence to use the simulation results to help interpret experimental data from surface science experiments and are leading to an improved understanding of catalytic reaction mechanisms.

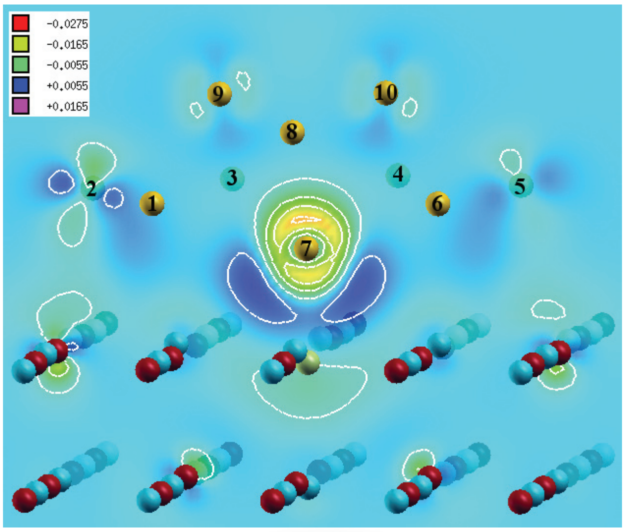


Figure 5: A $Au_{10}$ cluster adsorbed at a surface oxygen defect on MgO(001) the contour plot shows the transfer of charge from the defect to the central Au atom of the cluster.

## Acknowledgements

This project was funded, in part, from an EPSRC IMI project involving Johnson Matthey, BP-Amoco, Robinson Brothers, Grace Davidson and Accelrys. Time on the UK HPCx facility was obtained through the Materials Consortium.

## References

1. N.R.Avery, *Surf.Sci.*, **125** (1983) 771.; M.A.Vannice, W. Erley and H. Ibach, *Surf. Sci.*, **254** (1991) 1.

2. G. Kresse and J. Hafner, *Phys. Rev. B*, 1993, **47**, 558; G. Kresse and J. Hafner, *Phys. Rev. B*, 1994, **49**, 14251.

3. G.W. Watson, R.P.K.Wells, D. J. Willock and G. J. Hutchings, *Chem. Commun.* 705, (2000); G.W. Watson, R.P.K.Wells, D.J. Willock and G.J. Hutchings, *J.Phys.Chem.B*, **105**, 4889, (2001).

4. E.L. Jeffery, R.K. Mann, G.J. Hutchings, S.H. Taylor and D.J. Willock, *Catalysis Today*, **105**, 85, (2005).

5. M. Haruta, *Catalysis Today*, **36** (1997) 153.; G.J. Hutchings, *Gold Bull.*, **29**, (1996), 123.

6. J.M. Soler, E. Artacho, J.D. Gale, A. García, J. Junquera, P. Ordejón, and D. Sánchez-Portal, *J. Phys.: Condens. Matter*, 2002, **14**, 2745.

## Contact Details

Rudy Coquet, Edward L. Jeffery and David J.Willock
School of Chemistry
Main Building,
Cardiff University
Cardiff CF10 3AT.

email: willockdj@cf.ac.uk

# Teaching Newton to speak: Using CSAR services to speed the training of a neural network that models human language processing

*Stephen Welbourne and Matthew Lambon Ralph*
*University of Manchester*

## *Introduction*

Neural networks are now well established tools in the study of language processes (Dell, Schwartz, Martin, Saffran, & Gagnon, 1997; Harm & Seidenberg, 2004; Plaut, 1996; Plaut, McClelland, Seidenberg, & Patterson, 1996; Seidenberg & McClelland, 1989; Welbourne & Lambon Ralph, 2005). These networks are attractive as models because they process information in a similar way to the human brain; using a large number of simple information processing units, in parallel, to map representations across domains. In the case of the brain these processing units are neurons whereas in the models they are artificial neurone-like constructs, built from code that runs on a serial processor (usually a PC). In both cases learning occurs as a result of modification to the weights (synaptic strengths) of the connections between units, with these modifications occurring slowly over a large number of repeated trials.

The main limiting factor in the use of these models is the computational resource that is required to train them. Unlike the brain, these models cannot truly process information in a parallel manner, but have to simulate parallelism by cycling through the units serially. In addition, the kind of language tasks that are interesting to model tend to require training on large corpuses of words, typically thousands of items. As a result it is not unusual for these models to require weeks or even months of processing.

The purpose of this Class 3 project was to test the feasibility of using parallel supercomputers to significantly reduce the processing time required for this kind of model. Ultimately we would like to model speech, verbal comprehension and reading behaviours simultaneously, within the same generalised language model. However, for the purposes of this simulation, we elected to concentrate solely on the mapping from meaning to phonology (speech).

## *Simulation Details*

The training corpus consisted of 2998 monosyllabic words with phonological representations taken from Plaut et al. (1996). The semantic representations were constructed by generating unique random binary vectors of length 100 with an average of 20 units set to 1 and 80 set to 0. This ensured that we preserved two important features of human semantics: firstly, that semantic representations are relatively sparse, and secondly, that the mapping between semantics and phonology is not in any way systematic.

Figure 1 shows the architecture of the recurrent network that was used for these simulations with semantic and phonological layers connected by hidden layers consisting of 1500 units. Where layers of units are shown as connected it was always the case that every unit in the sending layer was connected to every unit in the receiving layer. Activation functions for the units were logistic with time integrated inputs. The network was trained using standard backpropogation through time with a learning rate of 0.05 and momentum of 0.9, applied only when the gradient of the error slope was less than 1.

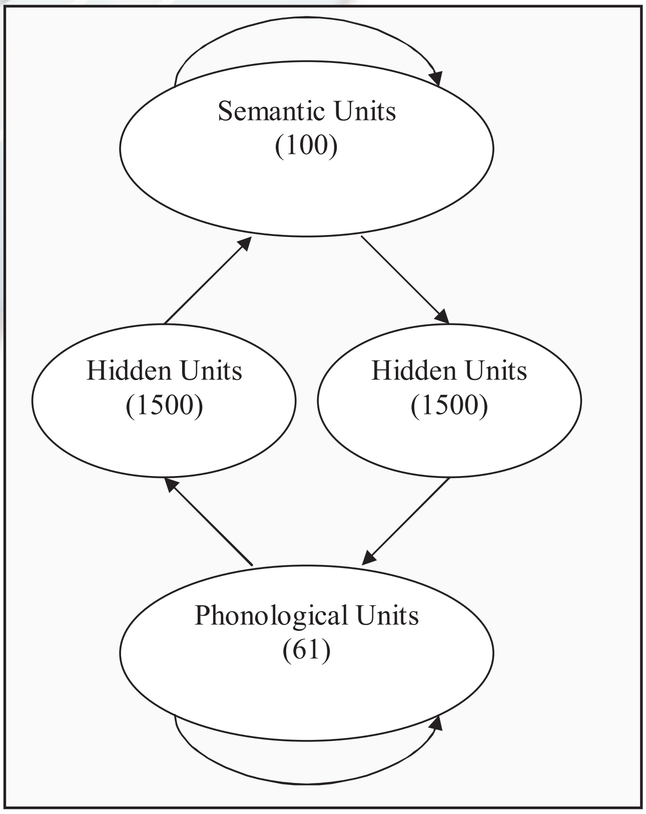


Figure 1: Network Architecture.

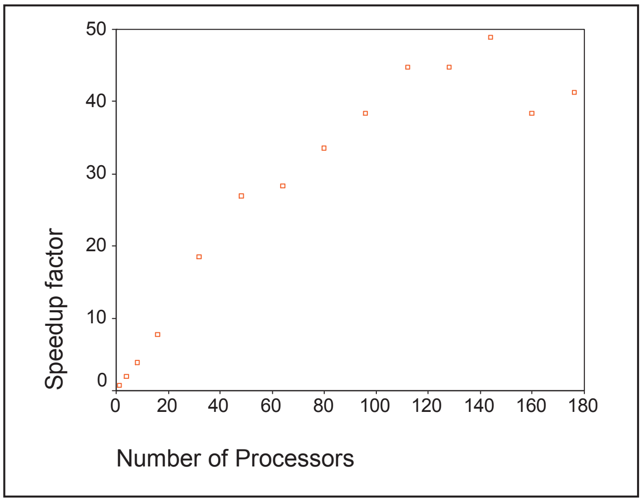


Figure 2: Showing how speed of training scales with number of processors compared to a desktop PC.

## Parallelisation Method

In this task there are two obvious approaches to parallelisation. Either one can parcel out the units between processors, or one can replicate the entire network on each processor and divide the training batch by processor. We elected to adopt the latter approach and, with the extensive help of the CSAR support team, we adapted our existing network simulator code to run on Newton.

## Results

For the purposes of this feasibility study, we were not interested in how well the network could perform the task, but merely in how long it took to accumulate the weight updates for one pass of the entire training corpus (1 epoch of training). In, particular we were interested in how the speed of training would scale with the number of processors. Accordingly, we ran trials of 2 epochs of training over increasing numbers of processors up to a maximum of 176. Figure 2 shows the results of this investigation. For convenience the y scale is expressed in multiples of single processor speeds (measured on a standalone Pentium 4 3.2Ghz PC running windows XP). When using only one processor Newton actually runs the code more slowly than on a standalone PC (speedup=0.76). However, the speed of processing scales reasonably linearly all the way up to 100 processors (speedup≈40). After this processing speed continues to improve slightly up to about 140 processors. Beyond that adding extra processors actually reduces processing speed.

## Discussion

This project set out to test the feasibility of using supercomputing services to speed the training of neural networks modelling linguistic processes. Using a typical network setup, modelling the mapping from semantics to phonology, we have demonstrated that speedup factors in excess of 40 are achievable. Time constraints prevented us from conducting further empirical investigations; it would be interesting to know how the network parameters (number of units and batch size) would affect the scaling performance. Nevertheless, we have clearly shown that this approach has considerable potential to reduce the time required to train these kinds of networks.

## References

Dell, G. S., Schwartz, M. F., Martin, N., Saffran, E. M., & Gagnon, D. A. (1997). Lexical access in aphasic and nonaphasic speakers. Psychological Review, 104(4), 801-838.

Harm, M.W., & Seidenberg, M. S. (2004). Computing the Meanings of Words in Reading: Cooperative Division of Labor Between Visual and Phonological Processes. Psychological Review, 111(3), 662-720.

Plaut, D.C. (1996). Relearning after damage in connectionist networks: toward a theory of rehabilitation. Brain And Language, 52(1), 25-82.

Plaut, D.C., McClelland, J. L., Seidenberg, M. S., & Patterson, K. (1996). Understanding Normal and Impaired Word Reading: Computational Principles in Quasi-Regular Domains. Psychological Review, 103(1), 56-115.

Seidenberg, M. S., & McClelland, J. L. (1989). A Distributed, Developmental Model of Word Recognition and Naming. Psychological Review, 96(4), 523-568.

Welbourne, S. R., & Lambon Ralph, M. A. (2005). Exploring the impact of plasticity-related recovery after brain damage in a connectionist model of single-word reading. Cognitive, Affective & Behavioral Neuroscience, 5(1), 77-92.

# World's Largest Altix Deployment Tackles Real-World Challenges

*Walter F. Brooks, Chief, NASA Advanced Supercomputing Division, Ames Research Center*

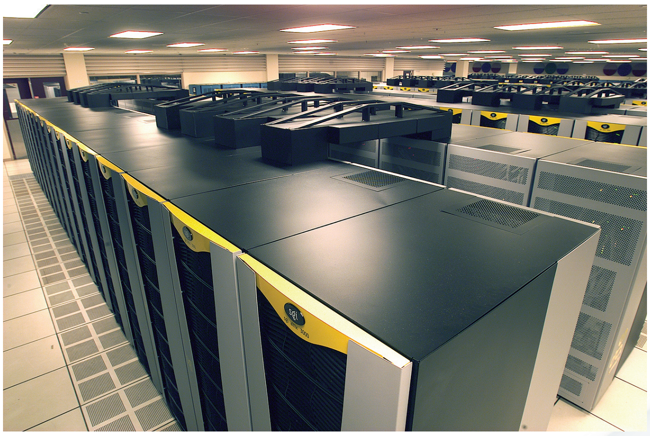


Figure 1: NASA's 10,240-processor SGI Altix supercluster, "Columbia." *Photo courtesy of Tom Trower, NASA Ames Research Center.*

NASA's Columbia supercomputer is a highly integrated "constellation" of twenty 512-processor Silicon Graphics, Inc. (SGI) Altix 3700 nodes using Intel Itanium 2 processors, interconnected via the high-speed InfiniBand fabric and running the Linux operating system. Named to honour the late crew of the U.S. Space Shuttle Columbia, the 10,240-processor system is housed at NASA's Ames Research Center (ARC) in California's Silicon Valley. This system, which has helped put the U.S. back on the high-end computing technology leadership track, is ranked as the third fastest supercomputer in the world on the June 2005 Top500 supercomputing list with a LINPACK rating of 51.9 teraflops.

Columbia is the product of a strong collaboration between NASA; major partners SGI and Intel Corp.; and supporting partners Voltaire, Inc., Mellanox Technologies, Inc., Cisco Systems, Inc., and Altair Engineering. NASA's expertise in high-end computing, the space agency's experience with exploiting SGI's single-system image (SSI) architecture, and the unprecedented dedication of and cooperation among all partners have been key to Columbia's successful deployment.

NASA's quest for building increasingly larger SSI systems is driven by the goal of providing its scientific and engineering users with the simplest and most efficient platform for high-performance computing. As the main computational resource for the entire agency, and as a national leadership-class system, Columbia is designed to use off-the-shelf components, including hardware and software compliant with industry and de-facto standards. With the open source Linux operating system, Columbia is accessible to a wider community of scientists and engineers.

### *Why Altix?*

As part of an ongoing effort to push the limits of high-performance computing, NASA installed the world's first 512-processor SGI Altix 3000 SSI supercomputer in 2003, named Kalpana to honor the memory of astronaut Kalpana Chawla, one of the seven crew members aboard the Columbia shuttle.

Although Kalpana increased NASA's ability to make science and engineering contributions to the Columbia accident investigation and Return to Flight activities at Ames, and enabled significant strides in climate and ocean modeling, it was clear that an even larger, more powerful system was needed to meet the agency's steadily growing demands for high-fidelity modeling and simulation, and to help advance all four of NASA's missions simultaneously. A year after Kalpana was installed, NASA began exploring the possibility of building a larger system - an endeavour sparked by their participation in the High-End Computing Revitalization Task Force (HECRTF), which was established under the U.S. National Science and Technology Council to guide future Federal high-end computing investments.

NASA formed a team of experts in high-end computing and examined possible architectures. As the benefits and potential risks of a small-node cluster emerged, and considering the success of Kalpana, the NASA team presented SGI and Intel with an alternative solution: a fat-node cluster based on the well-established Altix architecture, which would link twenty 512-processor nodes, connected with an input/output subsystem architecture delivering 10-gigabit-per-second performance.

### *Shared Memory Across 2,048 Processors*

Once all 20 of Columbia's nodes were installed, the team began another project, linking four of these systems together to achieve a shared-memory environment across 2,048 processors, making it the largest (shared-memory system) in the world. All 512 processors within each of the Bx2 nodes used for the 2048 are interconnected via NUMAlink4 using a fat-tree topology. In turn, each of these boxes is connected through six networks: NUMAlink4, InfiniBand, 10-gigabit Ethernet, and three gigabit Ethernet networks. To enable a NUMAlink connection between each of the four boxes, SGI designed a wall to hold the NUMAlink cables.

The main purpose for creating this system was to find ways to handle large, tightly coupled jobs, and to experiment with scaling. The 2048 system is currently being used for pioneering work and testing - specifically looking at large-scale processing and scaling - 2,000 processors and beyond. The 2048 system is already enabling groundbreaking computations including two promising climate models and a global-ocean circulation and sea-ice model, ECCO (Estimating the Circulation and Climate of the Ocean).

The ECCO work has utilized over 2,000 processors, achieving 1/16th of a degree resolution, which has enabled discovering phenomena in the ocean not previously possible due to the limited resolution attainable before the shared-memory environment of the 2048 was available. Such results are yielding an increasingly accurate picture of the effect of the oceans on the Earth's climate.

## Columbia System in Action

Columbia's capabilities have been applied to some key NASA engineering challenges in returning the space shuttle to flight, as well as scientific breakthroughs in Earth science, and space exploration. For example, a single simulation of a damaged area (such as a tile) on the space shuttle orbiter took three weeks running 24x7 on ARC's previous systems. In comparison, 100 of these damage site simulations were run in just 12 hours using eight of Columbia's nodes. This particular application simulated temperature and heating rates of damaged tiles on an orbiter as it re-enters Earth's atmosphere, travelling at speeds between Mach 16 and 26. (Figure 3). In some cases, the simulations showed the temperature of the damaged tiles rising far beyond 2,800 degrees Fahrenheit—the point at which they begin to fail and would not allow a shuttle crew to return safely to Earth.

During the recent STS-114 Discovery Mission, NASA utilized the same near-real-time simulation and modeling analysis capability to analyze the foam incident, tile filler gap material and the Advanced Flexible Reusable Surface Insulation (AFRSI) thermal blanket damage. Having this modeling capability was extremely valuable when assessing risks and making decisions during the mission.

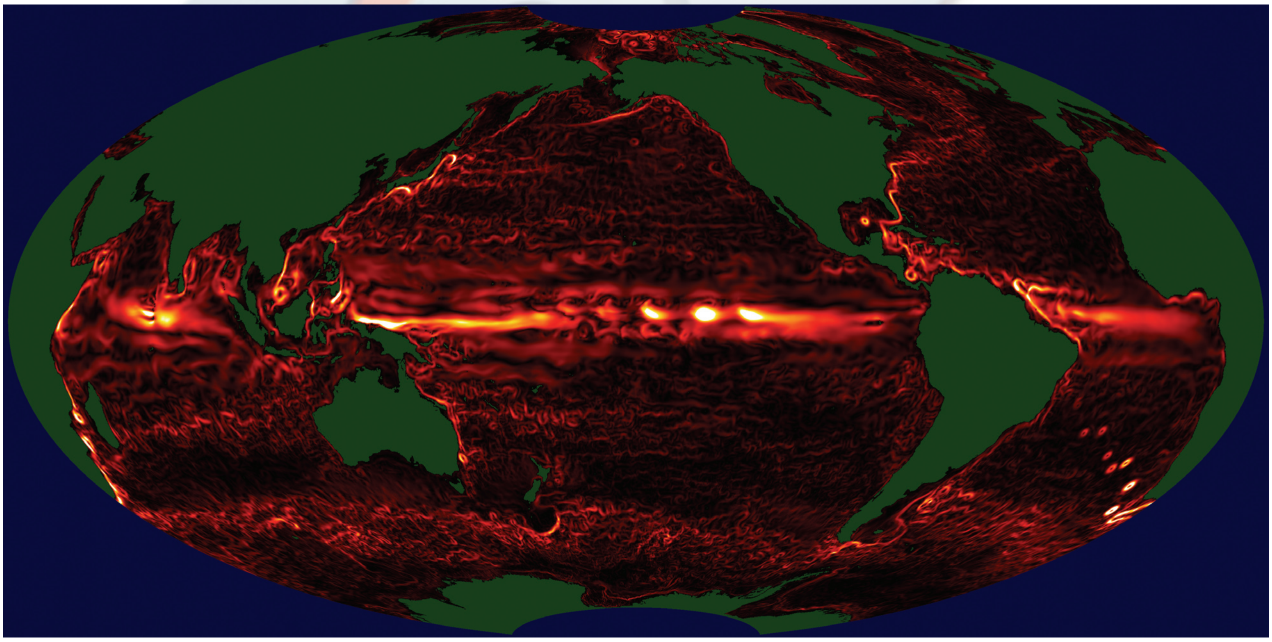


Figure 2: Improved resolution and accuracy of global-ocean circulation and sea-ice estimates made possible by Columbia are yielding an increasingly accurate picture of the role of the oceans in the Earth's climate as shown here. This work is a contribution to the consortium for Estimating the Circulation and Climate of the Ocean (ECCO).

*Image courtesy of Christopher Henze, NASA Ames Research Center.*

Columbia's architecture also lends itself well to large-scale molecular dynamics simulations involving millions to billions of atoms. These simulations can be important for designing new materials for thermal protection systems or for building future spacecraft. Preliminary tests on 1,920 processors of the 2048 system have achieved incredible results including a 1.4 million- and 18.9 billion-atom simulation using a space-time multi-resolution molecular dynamics algorithm.

### 63 Teraflops and Beyond

To remain a viable resource capable of meeting the agency's and the country's growing high-end computing requirements for advanced modeling and simulation, NASA is planning to work with other vendors to explore emerging architectures and cluster solutions.

While Columbia has already increased NASA's compute capability ten-fold and has attained science results not previously possible, the agency's demand for high-end computing continues to grow rapidly. To keep up with this demand for high-end computing capabilities, NASA will continue working with the National Coordination Office and sister agency programs like the U.S. Defense Advanced Research Projects Agency's (DARPA) High Productivity Computing Systems (HPCS) program to bring sustained petaflop-scale computing to the U.S. by the end of the decade.

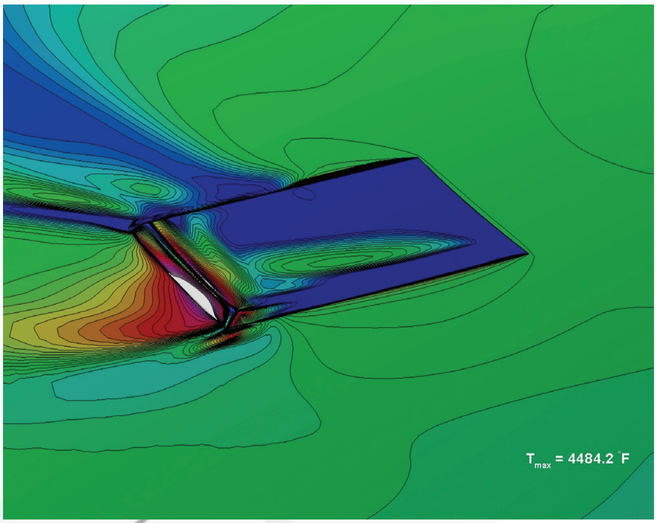


Figure 3: A new, rapid aerothermal computational fluid dynamics (CFD) analysis capability using Columbia permits near-real-time analysis of observed Orbiter damage during flight. Shown here: temperature contours of a shuttle tile cavity case traveling at Mach 22.9. *Image courtesy of the NASA Debris Transport Team.*

# Newton as a Single System

***Kevin Roy***

***Senior HPC Consultant, Manchester Computing, University of Manchester***

CfS (the consortium that provides the CSAR service) has continually pushed Newton to expand its potential and make it a more productive, easier to use system for those who use the system. As a result we have implemented our final planned hardware alteration; coupling both halves of the system into a single system with full NUMAlink interconnect and running a single operating system.

Before Newton became a 512 processor machine it had really only operated as two 256 processor machines with batch jobs running on one of the two hosts and never across both of them. The effect of this was that the maximum job size was restricted to the size that would fit on each machine.

The reason for this limitation was that although each of the two 256 processor machines were connected using SGI's Numalink interconnect, which gives 3.2Gb/s between pairs of processors, the machines were connected to each other with a single gigabit ethernet connection, which needed to be shared amongst all the processors.

The two 256 processor machines have now been reconfigured architecturally as a single 512 processor machine allowing batch jobs to span the whole system; this allows jobs of up to 496 processors to run. Newton is a 512 processor machine but some processors are reserved for interactive (4) and system processes (12), leaving the remaining 496 available for batch. This has also enabled the scheduling to improve, providing better turnaround for medium and small jobs.

We now have a single point of entry into the machine for all processors, allowing networked applications (such as grid projects) to connect directly to the internet.

The simplicity of a single system is evident in the file systems too. Operating and system software is now identical across the whole machine and all local temporary disk can now be viewed from the interactive area of the machine.

# Simpleware: From 3D image to mesh in minutes

*Emma AC Johnson, Communications Officer and Dr Philippe G. Young, Managing Director, Simpleware Ltd.*

In the past, converting 3D images into meshes for use in Finite Element (FE) analysis often necessitated time consuming processes and a gross simplification of the model geometry. UK-based, imaging specialist Simpleware Ltd. currently offers two new software products, *ScanIP* and *+ScanFE*, as an advanced solution to this problem, enabling users to quickly and accurately convert any 3D dataset, such as a magnetic resonance imaging (MRI) scan, into high-quality meshes in minutes.

Simpleware Ltd. has established itself as the world leader in the provision of software and services for the conversion of 3D imaging data into numerical models. *ScanIP* provides image visualisation and processing and allows files to be exported to other manufacturing software programs. *+ScanFE* is a mesh-generation module that creates volumetric meshes which can be exported directly to leading FE (and CFD) commercial solvers.

Simpleware's flagship software product *ScanIP* provides an extensive range of image processing and meshing tools to generate highly accurate subject specific models based on data from 3D imaging modalities such as MRI, Ultrasound and Computed Tomography (CT). Features of particular interest include a metal artifact removal filter (for artifacts in CT scans), improved topology and volume preserving smoothing algorithms and a broader range of visualisation modalities.

Novel proprietary algorithms and techniques developed by Simpleware, also permit fully automated, robust generation of FE models based on 3D imaging data and these have been implemented into a commercial code, *+ScanFE*. Mesh generation based on imaging data is an area of great interest in the FE community but the majority of approaches to date have involved generating a surface model from the scan data which is then exported to a commercial mesher – a process which is time consuming, not very robust and virtually intractable for the complex topologies typical of actual composites. A more direct approach is to combine the geometric detection and mesh creation stages in one process. The process incorporates an adaptive meshing scheme, which is fully automated and robust, creating smooth meshes with low element distortions regardless of the complexity of the segmented data.

A sophisticated assignment of material properties is based on signal strength allowing a general mapping function between greyscale and density or Young's Modulus to be defined (several different functions can be assigned to each part). These are all in addition to the proprietary technology which ensures high quality multi-part meshes which conform perfectly at part interfaces (both STL and volume meshes).

In addition to simplifying the meshing process dramatically, the mesh generation from scan data has several important advantages:

a) accuracy of meshed topology is only contingent on image quality. The geometry of the structure is reproduced in the finite element mesh at sub-voxel accuracy.
b) structures consisting of several different materials can be meshed automatically.
c) interfacial contacts can be modelled.

As well as performing convergence studies of field parameters of interest by increasing mesh density, convergence of models to morphology with increased image resolutions can be carried out. Where the properties vary in a continuous fashion throughout the structure, the approach can be used to derive a relationship or mapping function between the signal strength and the material properties which can be extremely useful for studying a wide range of problems including open celled foams, soil samples, and bone.

## Case Study

A case study was carried out to explore the feasibility of using clinical data for post-clinical structural evaluation of implant performance. An *in vivo* clinical scan of a patient fitted with a total hip replacement (THR) system was used to explore the influence of mesh density on the predicted response as well as the influence of the assumed contact model at the cup–implant interface.

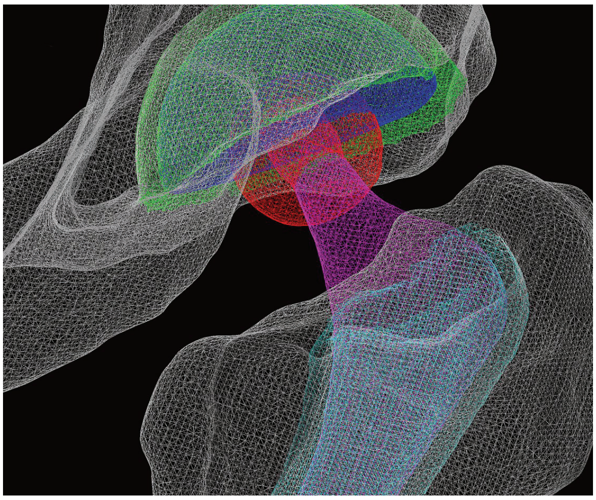
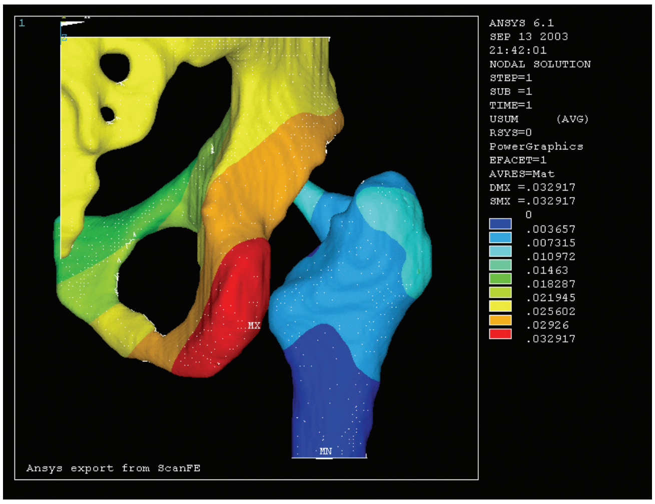


Figure 1: Hip, Femur and Implant in $^{+}$*ScanFE*.



```
ANSYS 6.1
SEP 13 2003
21:42:01
NODAL SOLUTION
STEP=1
SUB =1
TIME=1
USUM      (AVG)
RSYS=0
PowerGraphics
EFACET=1
AVRES=Mat
DMX =.032917
SMX =.032917
     0
     .003657
     .007315
     .010972
     .01463
     .018287
     .021945
     .025602
     .02926
     .032917
```

Figure 2: FEA Analysis Results.

## Methodology

A CT scan of in-plane resolution, 0.77 mm and slice-to-slice separation, 1 mm was re-sampled and a Metal Artefact Removal (MAR) filter applied. Six masks were created using ScanIP: (1) Pelvis, (2) Cement, (3) Cup, (4) Stem, (5) Cement mantle, (6) Proximal Femur. Based on the six segmented structures, two smooth models of different mesh densities were generated using $^{+}$*ScanFE* taking less than 3 minutes each. Additionally, a rapid prototyped model replica with the exact geometry as the FE mesh topology, was generated. Using commercially available FEA software, material properties, boundary conditions and loads – including muscle forces – were applied. All materials were assumed to be homogenous, isotropic and to behave linear elastically. Nodes at the top of the pelvis and distal part of femur were defined in $^{+}$*ScanFE*. The response of the system was analysed under static loading conditions with a sliding interface at cup-implant interface. The total solution time (on an Intel 2.8 GHz) for the low density mesh was a little over 2 hours and 6 hours for the high density model.

## Results

The study demonstrated the potential of the proposed approach for the generation of patient specific FE models based on *in vivo* clinical scans. In spite of their complexity and sophistication, full FE simulations can be carried out on an inexpensive and commonly available hardware platform.

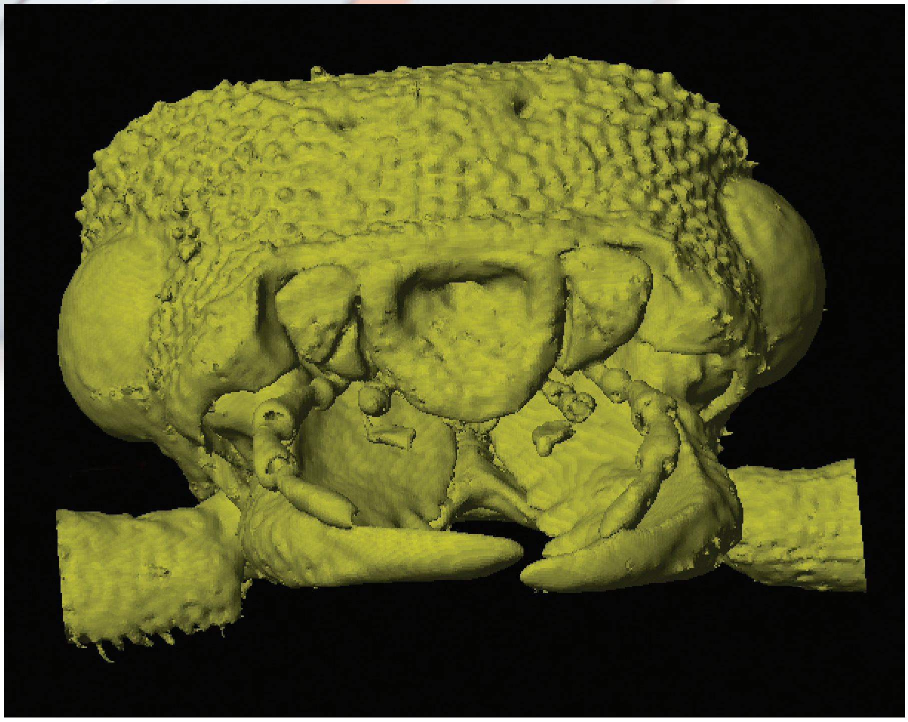


Figure 3: Beetle Mandible.

## Natural Sciences

As a joint venture between the University of Exeter and Exeter Advanced Technologies (X-AT), Simpleware was asked to contribute to a set of exhibits on a theme of Biomimicry for the Eden Project's new Education Centre which opened recently. Biomimicry is the imitation of nature to develop scientific and engineering processes and classic examples include Velcro®, inspired by burdock seeds; parachutes inspired by dandelion seeds and in particular relation to the Eden Project, the structure of a dragonfly's eye as the inspiration behind the design of the Biomes. Rapid prototyped models and computer simulations have been developed from CT scan images of plant and insect specimens. The resultant models form elements of visual displays from which children and adults of all ages can learn about how scientists and engineers use nature as an invaluable source of inspiration.

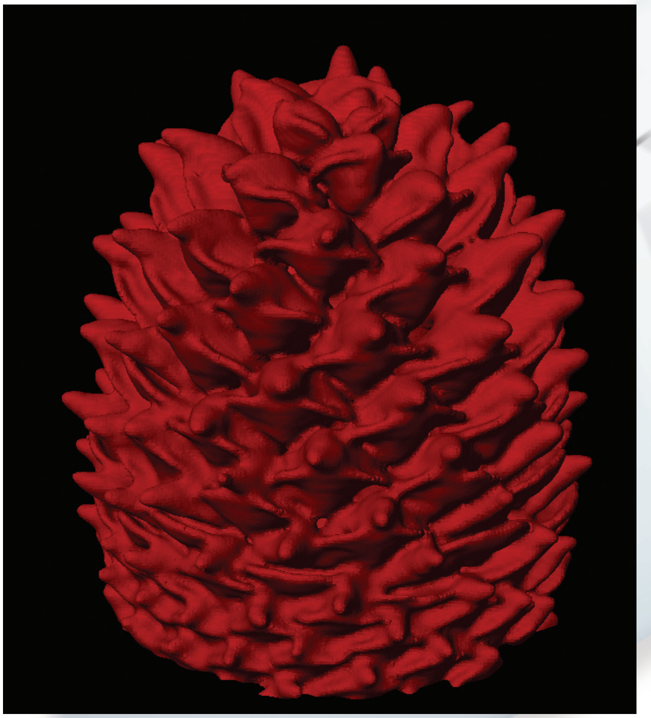


Figure 4: Pine Cone.

Simpleware is also currently involved in a range of natural science projects in collaboration with the Research Support Services (RSS) group at the University of Manchester. Using the high quality mesh generation techniques provided by Simpleware, FEA techniques will be used to study the structure and behaviour of a range of natural organisms. Once funding is secured, Simpleware and its co-investigators will explore the strength and functionality of a *Baryonyx* dinosaur skull (Natural History Museum) and the stress distributions within a beetle mandible (University of Gottingen) to identify the feeding habits of each creature. CFD analysis will also be used to study the hydrodynamics of graptolite locomotion to ascertain its marine environment (University of Edinburgh) and the wind driven pollination mechanisms of pine cones to determine their efficacy at pollen capture (University of Exeter).

The ease and accuracy with which models can be generated have opened up a wide range of previously difficult or intractable problems to numerical analysis, including blood flow, material characterisation of nanostructural composites and patient specific implant design. If the system is coupled with rapid prototyping hardware, it is also possible to produce a solid polymer or metal facsimile of the object in question - the process can then be effectively conceived as a 3D photocopier.

Simpleware has developed a new module, $^{+}$*ScanCAD*, which allows the import and interactive positioning of CAD models within the image masks. This can be used to bring in reaming tools, implants etc. and integrate them into the image. STL or FE models can then straightforwardly be generated. The new release will also include level set methods which are very powerful techniques for segmenting images.

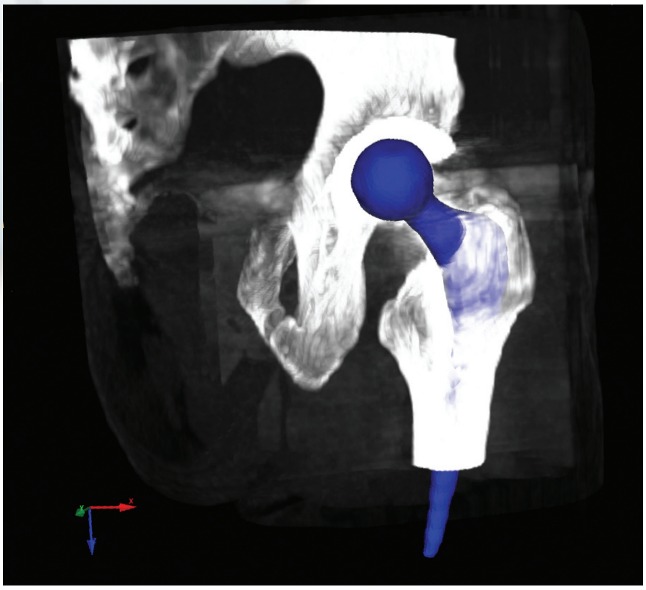


Figure 5: Hip with positioned CAD implant (Stryker).

# Grid Enabled Interactive Molecular Dynamics Simulations

*Peter V. Coveney and Shantenu Jha,*
*Centre for Computational Science, University College London*
*Stephen Pickles,*
*Manchester Computing, The University of Manchester*

The importance of computational approaches in providing quantitative information as well as qualitative insight has been widely acknowledged. For biomolecular systems, classical molecular dynamics (MD) simulations have the greatest ability to provide insight into specific aspects of a system at a level of detail not possible for other simulation techniques and often not even accessible experimentally [1]. However, the ability to provide such specific information comes at the price of making such simulations extremely intensive computationally. Although so far high-performance computational approaches have addressed this requirement to a great extent and in the process have played a critical role in enhancing our understanding of biomolecular systems, it has been shown [2] that if larger and more "real systems" are to be simulated over meaningful timescales, then advances in both the algorithms and the computing approach used are imperative. In this article we will focus on new computing approaches that facilitate *interactive* simulations of large-scale systems and will discuss some of the issues that require attention when attempting to use such an approach effectively and in a routine manner. By interactive simulations, we mean simulations with which the end-user can interact though a visualization component (visualizer) and/or computational steering client in near real-time.

## *Using Interactive Simulations to Compute Free Energy Profiles: An Exemplar*

The transport of biomolecules like DNA, RNA and poly-peptides across protein membrane channels is of primary significance in a variety of areas. Although there has been a flurry of recent activity, both theoretical and experimental [3, 4] aimed at understanding this crucial process, many aspects remain unclear.

A back-of-the-envelope estimate of the computational resources required helps us to appreciate why there has not been any significant computational contribution to understanding the dynamical aspects of the translocation problem. The physical time scales for translocation of large bio-molecules through a trans-membrane pore is typically of the order of tens of microseconds. It currently takes approximately 24 hours on 128 processors to simulate one nanosecond of physical time for a system of approximately 300,000 atoms. Thus it takes about 3000 CPU-hours on a tightly coupled machine to simulate 1ns. Therefore a straightforward "simple" MD simulation will take $3 \times 10^7$ CPU-hours to simulate 10 microseconds – a prohibitively expensive amount. Thus approaches that are "smarter" than vanilla classical equilibrium MD simulations are required.

As part of the SPICE project[1] we are implementing a method (here referred to as SMD-JE, see Refs. [2, 6] for more details) to compute the free energy profile (FEP) along the vertical axis of the protein pore. By adopting the SMD-JE approach, the net computational requirement for the problem of interest can be reduced by approximately a factor of 50-100. The methodology however, requires the introduction of two new parameters, with a corresponding uncertainty in the choice of the values of these parameters. The situation can be addressed by performing a set of "preprocessing simulations", which, along with a series of interactive simulations, help determine an appropriate choice of the parameters. To benefit from the potential advantages of the SMD-JE approach and to facilitate its implementation at all levels — interactive simulations for such large systems, the pre-processing simulations and finally the production simulation set — we use techniques developed in the RealityGrid project[2] to make optimal use of the computational resources of a federated trans-Atlantic grid.

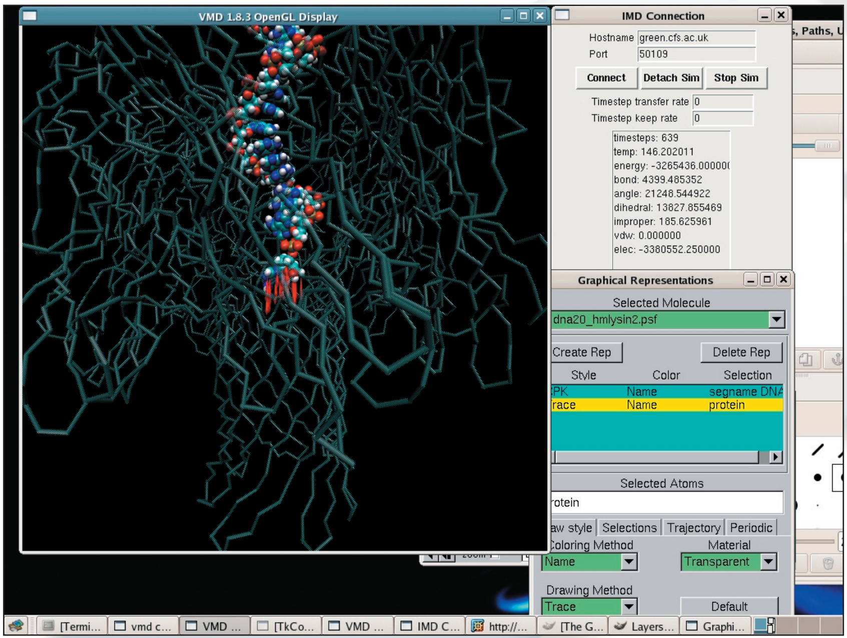
## *Making Interactive Simulations Possible: Some Technical Issues*

As the model being studied is very large and complex, a large number of processors are needed to provide sufficient compute-power for the simulation to be interactive. Thus, it is rather unlikely that all the computational and visualization resources required for interactive simulations will be available in one place, and even less likely that these will be co-located with the user.

[1] SPICE is the UK part of a joint UK-US high-end computing project funded by the EPSRC in the UK and the NSF in the US that shows the advantages of grid-enabled interactive simulations. SPICE will be exhibiting at SC05 on the University of Manchester and UK e-Science booths, amongst others.

[2] See http://www.realitygrid.org.

As a consequence, we have to deal with a geographically distributed set of resources, and the performance of high-end interactive simulations becomes dependent on the performance of the network. Furthermore, since these interactive simulations involve bi-directional communication - there is a steady flow of data from the simulation to the visualizer, as well as flow in the reverse direction when, for example, the user applies a force to a subset of atoms through the visualizer — unreliable communication can lead not only to a possible loss of interactivity, but also to a slowdown of the simulation if it has to wait for data from the visualization [6]. Thus interactive MD simulations place rather unique and demanding quality of service (QoS) requirements (low latency, jitter and packet loss) on the network linking the simulation and visualization components. The recently deployed UKLight [7] network and the optically networked Global Lambda Infrastructure Facility [8] provide us with the necessary QoS. The CSAR systems – Newton and Green – are already connected to UKLight and the connection of HPCx should be in place by Supercomputing 2005.



Hostname green.cfs.ac.uk
Port 50109
Timestep transfer rate 0
Timestep keep rate 0

```
timesteps: 639
temp: 146.202011
energy: -3265436.00000(
bond: 4399.485352
angle: 21248.544922
dihedral: 13827.855469
improper: 185.625961
vdw: 0.000000
elec: -3380552.250000
```

Selected Molecule dna20_hmlysin2.psf

| Style | Color | Selection |
|---|---|---|
| CPK | Name | segname DNA |
| Trace | Name | protein |

Selected Atoms protein
Coloring Method Name
Material Transparent
Drawing Method Trace

Figure 1: Snapshot of an interactive session in progress. A single stranded DNA polymer is beginning its translocation through the alpha-hemolysin protein pore. The red arrow-headed lines represent the forces that are applied to the end residues of the DNA to guide and speed up its translocation through the pore. When grid infrastructure is not used the user is responsible for providing host and port information. The use of RealityGrid tools simplifies this situation and the user only need select the simulation of interest. The steering infrastructure [5] then handles establishing the details of communication and exchanging this information in a manner that is essentially invisible to the user.

This style of interactive computing requires co-allocation of computation and visualization resources as well as network circuits of sufficient bandwidth and QoS between the simulation and visualization components so that the simulation does not stall waiting for the visualization and interactivity can be maintained. The need for community-wide agreements and protocols for co-allocation is obvious, but today, *ad hoc*, pair-wise agreements are the norm. For example, co-allocation [9, 10] is currently handled by manually reserving the necessary resources (we are grateful for the co-operation of HPCx, CSAR and TeraGrid staff in this matter). The current mode of operation is cumbersome and error-prone (one of the authors had to exchange about a dozen emails correcting three distinct errors introduced by two different administrators for one reservation request!) and is not in general a scalable solution. The situation gets even more complicated when attempting to schedule non-compute resources like network light-paths simultaneously. At present, we find ourselves in the fortunate situation where we have a quasi-persistent connection between our visualization machine (at UCL) and most of our compute resources (at CSAR and on the TeraGrid) and thus have to worry only about co-allocating compute resources. But sooner or later, demand for switched light-paths will increase and we will be faced with the more general situation of having to worry about coordinating and co-scheduling network circuits with compute resources, for it is of no use if the circuit between two machines is provisioned only after the user's reservation on the two machines has expired! There have been encouraging attempts and progress has been in this direction [11, 12, 13], but much remains to be done. This is a challenging problem and will need to be addressed by all the communities involved, for any bespoke or partial solution will present problems when used in a true grid context.

By facilitating the coupling of the many different components involved — geographically distributed visualization, computational and software resources in distinct administrative domains — grid infrastructure enables us to circumvent traditional constraints of batch computing style HPC by allowing distributed large-scale interactive simulations. The computation of FEP is an example of a large-scale problem that benefits tremendously from using HPC grids. In particular it is a good example of the advantages — both quantitative and qualitative — that steering simulations of large biomolecular systems can provide. The use of interactive simulations to explore further interesting and important problems, as well as its uptake as a generally useful computational approach, will require a stable and easy to use infrastructure that addresses satisfactorily the issues discussed and thus makes performing large-scale steered simulations more convenient and thereby routine.

## Acknowledgements

This work has been funded as part of the RealityGrid, RealityGrid-Platform, ESLEA and SPICE projects and has been possible thanks to the assistance of many people. In particular we'd like to mention Andrew Porter, Robin Pinning and Rob Haines at the University of Manchester for their work on the software infrastructure. We would like to thank Nicola Pezzi, Peter Clarke and John Graham for facilitating the use of UKLight. Finally we'd like to acknowledge Matt Harvey from UCL for assistance in the installation and deployment of the required infrastructure at many levels.

## Contact Details

Peter V. Coveney and Shantenu Jha
Centre for Computational Science,
University College London
20 Gordon Street
London
WC1H 0AJ

Stephen Pickles
Manchester Computing,
The University of Manchester
Oxford Road
Manchester
M13 9PL

## Further Reading

Volume 363, Number 1833 / August 15, 2005 of *Philosophical Transactions: Mathematical, Physical and Engineering Sciences*: http://www.journals.royalsoc.ac.uk/link.asp?id=h88056h75554

Volume 7, No. 5 and 6 of *Computing in Science and Engineering*, Guest Editors: Bruce Boghosian and Peter Coveney: http://www.computer.org/portal/site/cise/

## References

[1] M. Karplus and J. McCammon. *Nature Structural Biology*, 9(9):646–652, 2002.

[2] S. Jha, P. V. Coveney, M. Harvey, and R. Pinning. *Proceedings of ACM/IEEE Supercomputing Conference*, 2005, in press. http://sc05.supercomputing.org/schedule/pdf/anal109.pdf

[3] D. K. Lubensky and D. R. Nelson. *Phys. Rev E*, 31917 (65), 1999; Ralf Metzler and Joseph Klafter. *Biophysical Journal*, 2776 (85), 2003; Stefan Howorka and Hagan Bayley, *Biophysical Journal*, 3202 (83), 2002.

[4] A. Meller et al, Phys. Rev. Lett., 3435 (86) 2003; A. F. Sauer-Budge et al. *Phys. Rev. Lett.* 90(23), 238101, 2003.

[5] S. M. Pickles, R. Haines, R. L. Pinning and A. R. Porter, A Practical Toolkit for Computational Steering, *Phil Trans. R. Soc.*, 363, 1843-1853, 2005.

[6] S. Jha, M. J. Harvey, P. V. Coveney, N. Pezzi, S. Pickles, R. L. Pinning, and P. Clarke. *Proceedings of the UK e-Science All Hands Meeting*, 2005. http://www.allhands.org.uk/2005/proceedings/papers/455.pdf.

[7] http://www.uklight.ac.uk.

[8] http://www.glif.is.

[9] B. Boghosian and P. V. Coveney. Guest Editors' Introduction: Scientific Applications of Grid Computing – part ii. *Computing in Science and Engineering,*. DOI Bookmark: http://doi.ieeecomputersociety.org/10.1109/MCSE.2005.123.

[10] Karl Czajkowski, Stephen Pickles, Jim Pruyne, Volker Sander, Usage Scenarios for a Grid Resource Allocation Agreement Protocol, GRAAP-WG, Global Grid Forum, 2003.

[11] G-lambda: Coordination of a grid scheduler and lambda path service over GMPLS. http://www.gtrc.aist.go.jp/g-lambda/.

[12] http://www.cct.lsu.edu/personal/maclaren/CoSched/.

[13] https://forge.gridforum.org/projects/graap-wg.

# Integrated Parallel Rendering for AVS/Express

*Louise M. Lever and James S. Perrin*
*Manchester Computing, University of Manchester*

*Enabling integrated multi-processor computation and rendering for High-Performance Visualization*

## Introduction

AVS/Express is a leading visualization application development package, which was described in CSAR Focus Issue 10. To summarize, the AVS/Express visualization package was extended by the Manchester Visualization Centre, in collaboration with AVS, KGT and SGI. The first goal was to provide a multi-pipe parallel renderer, enabling AVS/Express to be used in immersive environments such as a CAVE or RealityCenter, while increasing rendering performance by parallelizing the rendering as much as possible. The shortcoming of this project was that many users were left with a single CPU for the computation of their visualization application. Our second goal was therefore to develop a toolkit to enable parallel computation within AVS/Express. The two projects are respectively known as Multi-Pipe Express (MPE) and the Parallel Support Toolkit (PST).

This article continues with the progress of the integration of the MPE and PST projects, to produce a powerful visualization architecture for high-performance computation and visualization.

## Overview

AVS/Express MPE and PST provide a strong visualization application development system. PST allows users to use parallelized modules within their applications to perform visualization tasks across many processors, in both shared memory systems and clusters. The end result of PST computation is typically a mesh dataset e.g., the sets of triangles produced in an isosurface calculation. For desktop users and existing MPE users, this data would normally be returned to the main application, pre-processed for rendering i.e., conversion to triangle strips and mapping of datamaps to RGBA colours, and then passed to the rendering system. For MPE users wishing to visualize large datasets this is a serious drawback, as all renderable geometry is channelled through the single threaded main application for pre-processing and rendering distribution. Even shared memory systems are impaired by the serialized pre-processing stage. Figure 1 shows the standard rendering architecture.

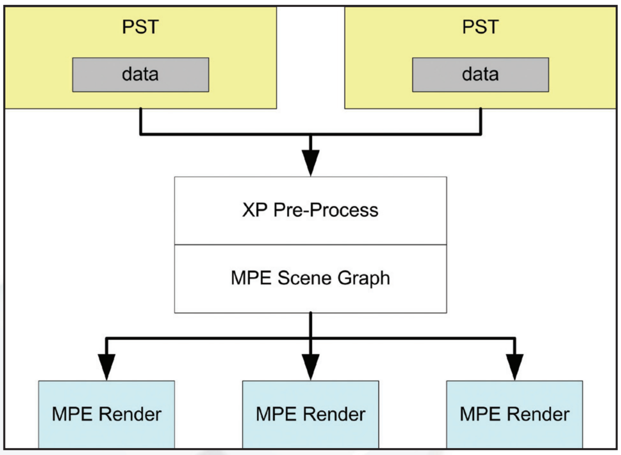


Figure 1: Standard Rendering Architecture, where PST output data is sent back to the main application for pre-processing before being distributed for parallel rendering.

The first step towards the development of an integrated system was to parallelize the pre-processing stage. In this mode of operation the PST nodes are responsible for pre-processing of their generated output. For desktop or single-pipe users this parallelization still provides a considerable performance gain. Figure 2 shows the Level 1 Parallel Rendering Architecture, where pre-processing is performed in parallel by PST before being sent back to the main application for rendering distribution. The second step is to enable tighter integration with the MPE rendering system. Typically, renderable data is encapsulated by the MPE renderer and distributed to the rendering nodes. Recent changes to MPE now effectively enables "empty" objects to be placed into the scene-graph for rendering. The main application is now only responsible for controlling the position and orientation of these objects with the context of the scene. In this mode, PST is now also responsible for encapsulation of its renderable data, ready for rendering by MPE. On both shared memory systems and cluster solutions one or more Distributed Object Handlers (DOH) listen for incoming objects from the PST nodes, the contents of which are placed into the "empty" objects. Figure 3 shows the Level 2 Parallel Rendering Architecture, where both pre-processing and rendering encapsulation are performed by PST in parallel, with the resultant data sent directly to the rendering nodes via the Distributed Object Handlers.

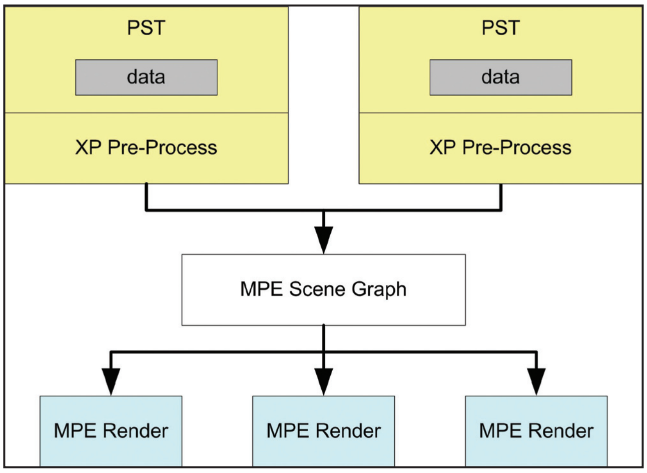


Figure 2: Level 1 Parallel Rendering Architecture, where PST output is pre-processed in parallel before it is returned to the serial application for rendering distribution.

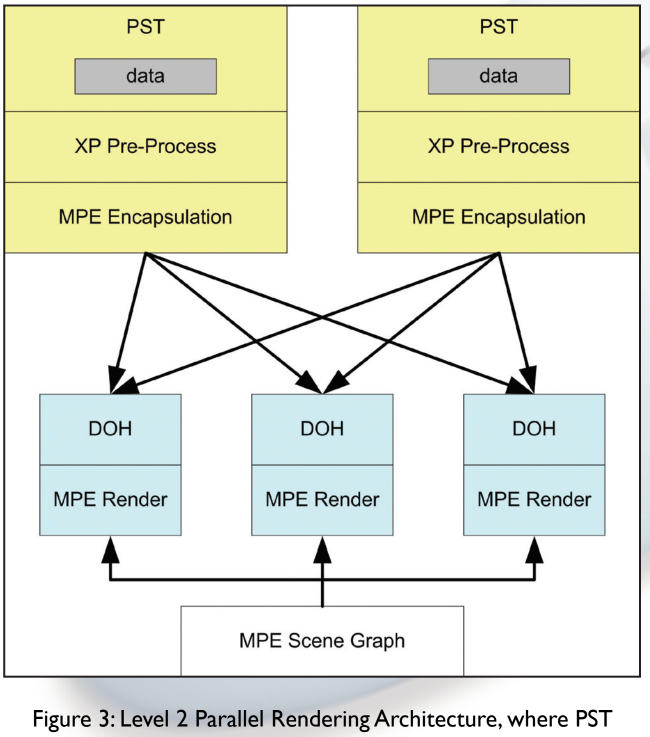


Figure 3: Level 2 Parallel Rendering Architecture, where PST output is both pre-processed and encapsulated ready for rendering and is sent directly to the Distributed Object Handlers as needed. Only empty objects are contained within the main scene graph.

The previous bottleneck is now avoided and reference objects are rendered within the application scene-graph. By providing a set of DOH threads, further improvements have been made, adding modes of operation allowing the reference objects to be rendered synchronously or asynchronously. Most visualization tasks within AVS/Express (and PST) are carried out in chunks, so for example one isosurface is comprised of many chunks of a chosen number of triangles. Such chunking of data improves handling of memory and rendering speed. Under asynchronous behaviour, chunks of renderable data are rendered and hence cached by the graphics hardware immediately. This approach has two benefits, a) users can see the visualization progressively appear as it is produced, which is particularly useful for large datasets and b) the graphics pipes can cache more easily when geometry is streamed into them.

Synchronous behaviour is used when progressive updates are not desired and the user wishes to see the new output when it is all available and ready to display.

With appropriate multipipe configurations it is possible to achieve database decomposition techniques for both shared memory and distributed cluster architectures. Database decomposition allows many direct PST to MPE connections, where each MPE node renders only the partial scene-graph generated by a local PST node. Only a final compositing stage is required to produce the final image. Figure 4 shows a distributed architecture enabling Database Decomposition, where each rendering node only renders 1/3 of the isosurface data and frame-buffer compositing is used to get the final image. Figure 5 shows an example of the Lattice Boltzmann dataset as rendered using database decomposition.

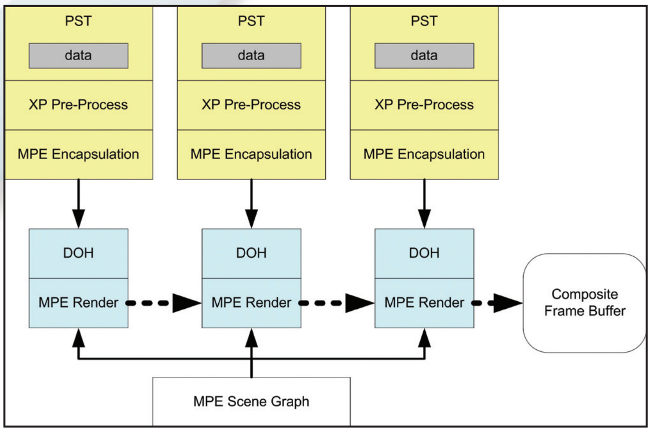


Figure 4: Example of a configuration for enabling Database Decomposition, where each render node only renders a partial scene-graph and frame-buffer compositing is used to get the final image.

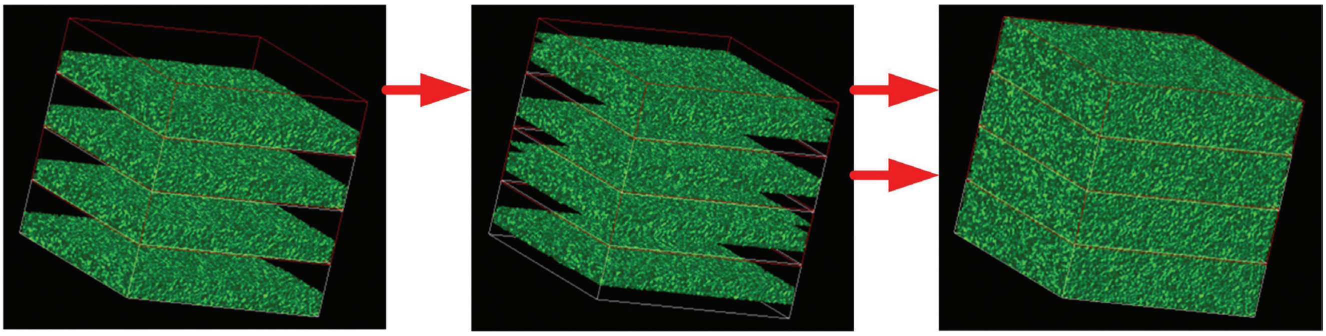


Figure 5: The Lattice-Boltzmann dataset as rendered with Database Decomposition and Frame-Buffer Compositing.

## *The Next Hurdle*

As high-performance computation and visualization increase in complexity and sheer problem size, the availability of graphics resources becomes the next major bottleneck facing users. While many customers do possess systems with thousands of CPUs, most current visualization solutions are limited to a few graphics pipes, typically in the region of 4 to 16.

Given the size of some problems and the utilization of thousands of CPUs using PST, the relatively small number of pipes becomes a bottleneck in itself. Hundreds of CPUs producing renderable data will now swamp the pipes, leaving users unable to visualize satisfactorily.

Our next major task therefore is to develop a Massively Parallel Renderer (MPR) which will use the available CPUs as software based render nodes, bypassing the need for hardware graphics pipes. The AVS/Express MPE Renderer will be extended to scale up to possibly thousands of processors and enable a software based solution. To achieve this, the MPR will build upon the existing OpenGL solution by employing the MesaGL library. In this manner the existing framework within MPE and PST will be able to migrate without the need for a whole new render system to be developed. However, there are still remaining issues to be considered, which will require investigation if the MPR is to be successful. The main obstacles to success are a) scalability of PST computation, b) scalability of MPE render nodes and c) efficient management of large-scale frame-buffer composition. The latter problem requires careful consideration of distribution and composition techniques, if the MPR is to handle thousands of generated frame-buffers.

## *Contacts*

Louise M. Lever:
louise.lever@manchester.ac.uk

James S. Perrin:
james.perrin@manchester.ac.uk

George W. Leaver:
george.leaver@manchester.ac.uk

http://www.sve.man.ac.uk/mvc/Research/mpe/
http://www.sve.man.ac.uk/mvc/Research/PST/

# CSCS, the Swiss national supercomputing centre

*Marie-Christine Sawley*
*CEO, CSCS*

Switzerland is a country known for its mountains and lakes, three national languages, ski slopes, cheese and chocolate: that is for the postcard description of it. It is also known for the quality of its education and world class research institutions which attract students, post docs and researchers from different horizons. Its industrial scene offers a wealth of players active in high value added goods and services, such as engineering, pharmacy, food and nutrition, financial institutions and biotechnology, a reflection of a country that has virtually no natural resources other than its skilled workforce and grassy landscapes.

The southern slope of the Swiss Alps is the base for the CSCS, the Swiss national supercomputing centre: this region which offers some beautiful scenery and hosts a small thriving university, is a natural bridge towards the vast Lombardy. The centre opened in 1991 and has since then offered resources and services to the Swiss research community. Since 2004, the CSCS is developing its activities along a new strategic quadriennial plan, comprised of the extension and the replacement of the computing infrastructure, complex and massive data analysis, benchmarking, and national hub for the Grid. It participates in European collaborations like EGEE and is actively involved in regular exchanges with leading HPC centres in Europe and beyond, like Manchester Computing.

The centre offers resources to scientists working in varied domains such as physics, chemistry and biochemistry, engineering science, CFD, climate modeling and environmental science. Presently the computing time is being allocated under scientific review to a number of institutions:

- Swiss institutes of Technology in Zurich, Lausanne, state universities (Zurich, Bern, Geneva, Basel, Fribourg, Neuchatel)
- MétéoSuisse for its daily predictions
- Industries, especially through funding agencies for technical transfer
- International research partners, such as EGEE.

In addition, the centre hosts on one of its supercomputers, the Nec SX5, the suite for National weather predictions for MétéoSuisse.

Figure 1: Marie-Christine Sawley, CEO, CSCS.

## *Extension and renewal of the computing infrastructure*

The development plan of CSCS for the years 2004-2007 proposed and secured the funding for a number of developing activities and projects. The first chapter concerned an extension and a replacement of the computing infrastructure. The extension, called Horizon, unfolded during 2004 and was conceived for the installation of a MPP system at CSCS for highly scalable nodes, to be put at the disposal of scientists as a tool for enabling new frontiers in science to be addressed. CSCS conducted this project in collaboration with the Paul Scherer Institute, which is the largest research institute in Switzerland, conducting multi-disciplinary research in areas such as beam and reactor physics, material, environmental and biomedical science. CSCS and PSI wanted to join efforts to buy the first 1000 processors and plus MPP machine in Switzerland, and to be able to deploy this as a first class instrument for some key areas in science. The procurement that we ran had very strict electricity and power requirements, was very demanding on the performance side — benchmarks were based on a mix of the new HPCC suite and user codes — and the total cost of ownership for a period of 3 years played an important role in the evaluation criteria.

CSCS ran a call for tender in 2004, and after the analysis of seven different offers, decided on the purchase of a Cray XT3, comprised of 1100 AMD processors connected via a very fast interconnect toroidal communication network based on the Seastar® chip. The system is currently being installed, hosts some early users and we plan to put the system in production during the first quarter of 2006.

The second project, called Zenith, will target the replacement of the existing computing infrastructure. Very complementary to Horizon, the system must offer capability computing resources for codes that do not scale very well, which need shared memory capacity, but will also exhibit some very good bandwidth between memory and processors.

### *The user community*

The CSCS application portfolio reflects both the history of the centre and the pattern of strong research areas in Switzerland, be it in public or private sector: fundamental sciences (physics and chemistry), as well as material and biomaterial, nanotechnology, environmental and engineering sciences, use the major share. As has already been pointed out, the CSCS hosts the daily suite for National weather services in Switzerland and this long time collaboration has brought a share of mutually beneficial activities, and has ignited natural synergies with some of the world class research community in Switzerland, like the groups working in Climate modeling which make use of the services.

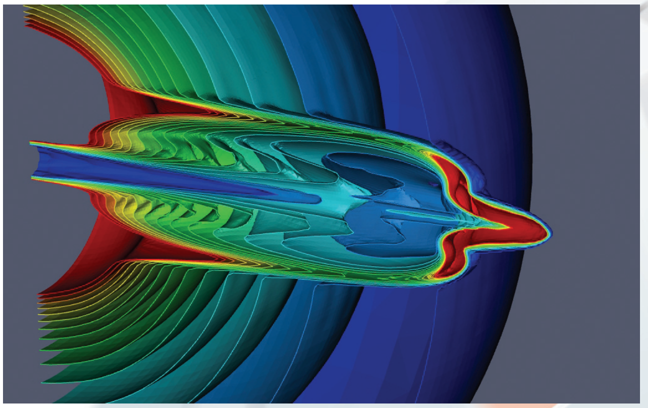


Figure 2: Magnetic jet from a young stellar object.

There exist a wealth of new emerging applications that bring to CSCS new development opportunities. Some of them are compute intensive, like in biomaterial sciences, or data intensive like particle physics or bioinformatics. Today some require the Grid infrastructure, and this was an opportunity for CSCS to enter into close contact with some new partners who require the weaving of a strong SW fabric around services, data and compute. This is the case of the Swiss Biogrid project, a joint initiative between the Biozentrum and the Novartis research institute, the Swiss Institute for Bioinformatics, the Functional Genomics Centre in Zurich and CSCS. It is also the case for the LCG computing since CSCS is hosting the Tier 2 level of national infrastructure used by the nine labs that are conducting research in particle physics. In the long term, we are also working in establishing contacts with the financial community who are making use of more and more complex mathematical modeling for portfolio management.

Figure 3: View of the Horizon system newly installed at CSCS.

### *The real challenges*

For a centre like CSCS, the challenges facing us in the near and medium term future are in two areas:

- capacity to continue deploying capability computing and complementary architectures, which is certainly in synchronism with funding but also in the capacity and the willingness to continue to be a pioneer;
- software, tools, advanced scientific visualization, performance monitoring, mathematical modeling, which gives the unique flavour to the atmosphere our users and partners experience by working with us and in addition the value of the computing cycles we make available.

Both relate to human competencies and skills, they make our work exciting and are the best way for us to face the rapid speed of changes we see in technology and in the way our users, mostly researchers, conduct their research and make their choices. At CSCS, there are 33 of us today, from 10 different nationalities and pluri-lingual (very few up to 6: not the majority!)

We strongly feel that organizing the access to today's Teraflops computers with the right combination of good quality software — scaling to hundreds of processors — , value added tools such as analytic benchmarking, performance tuning and evaluation and advanced visualization is the best way to build a sound road to Petaflop computing.

Marie-Christine Sawley, CEO, CSCS.
sawley@cscs.ch
www.cscs.ch

# What could GridSolve do for you?

*Jon Gibson*
*HPC Consultant, Manchester Computing, University of Manchester*

The NetSolve/GridSolve project is an attempt by the Innovative Computing Laboratory at the University of Tennessee to enable a scientific/engineering community, using its standard programming interfaces and desktop computing environments, to easily and efficiently access distributed hardware and software resources. The GridSolve software, a development from NetSolve, is the middleware necessary to provide this bridge. Built using standard internet protocols, it is available for most variants of UNIX and parts of the system are available for Windows 2000 and XP. It supports a fully heterogeneous environment. Operating a client-agent-server system, it searches for computational resources on a network, chooses the best available, then using retry for fault-tolerance, solves the requested problem and returns the answer to the user. Available resources are load-balanced to ensure their efficient use. The system is versatile enough to allow the integration of any arbitrary piece of software as a resource. For user accessibility, interfaces are currently available in Fortran, C, Matlab, Mathematica and Octave.

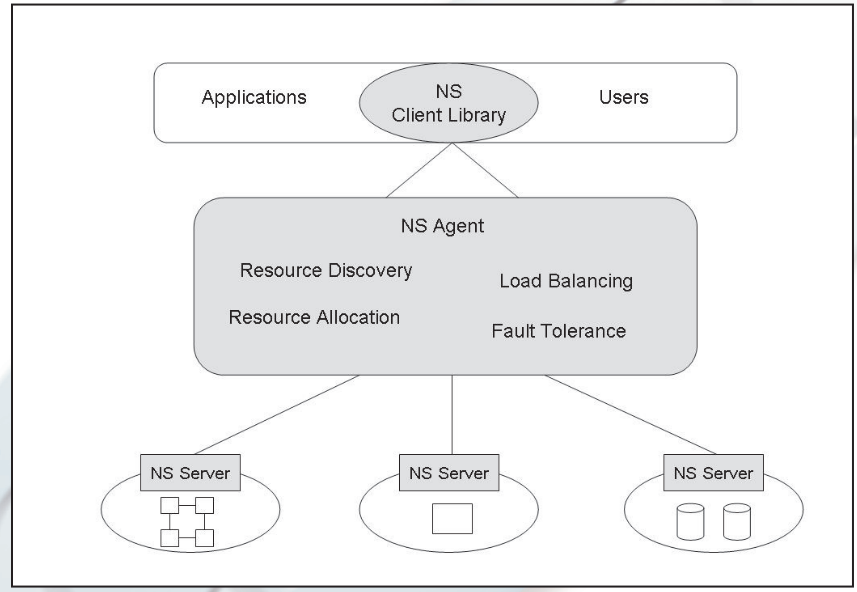


Figure 1: The GridSolve System.

Figure 1 illustrates the architecture of a GridSolve system (with the shaded parts being part of GridSolve). To access GridSolve, the user must link their application with the GridSolve library. This application will make calls to GridSolve's API requesting specific services, so gaining access to remote computing resources without the user having to know what resources or requiring specialist knowledge. The GridSolve agent maintains a database of servers, along with their capabilities (hardware performance and allocated software) and dynamic usage statistics, from which it allocates server resources to client requests, looking for the quickest solution time whilst load balancing the system. The agent keeps track of failed servers. The GridSolve server is a daemon process that awaits client requests. It could potentially be run on anything from a PC to a cluster to a large HPC machine, such as CSAR's Altix or Origins. An important part of the server is a source code generator, which parses a problem description file (PDF), essentially a standard wrapper around the functions being called. Use of these PDF's allows new functionality to be incorporated into the system. Hidden from the user, a given GridSolve request proceeds as follows: the client contacts the agent for a list of capable servers; the client contacts the designated server, sending the input parameters of the request; the server runs the requested routine and returns either the output parameters or an error status to the client.

Obviously, given the limited space, I have only really been able to provide a short introduction; more information about the GridSolve project can be found on their website: http://icl.cs.utk.edu/netsolve. It is likely that future developments in the GridSolve software, such as an interface to the LSF batch queuing system and GSI authentication, will potentially allow GridSolve to be run as a service on CSAR machines. Monitoring usage of resources will, of course, be integral to such as service. If anyone feels they would be particularly interested in using it or has any comments or questions, then please e-mail me at jon.gibson@manchester.ac.uk.

# The SGI® Roadmap – Taking HPC to the Next Level

*Michael Woodacre, Chief Engineer,*
*Server Platform Division, SGI*

## Background

A number of technology trends over the past decade have shaped the direction of high performance computing. Commodity clusters, which now dominate HPC, have benefited from the continued performance increases seen with x86 processor cores. Consistent with the rise in cluster use, the MPI programming paradigm has become extremely popular over the last decade as it helped programmers build their codes in a highly portable fashion. These trends have allowed applications to ride the rapidly improving performance curves of mass market processors instead of being captive to a particular architecture or vendor.

But while Moore's Law growth of transistors promises at least a few more process generations ahead, the ability to translate this directly into processor core performance has become a major issue. Similarly, the expansion of cluster systems to larger scale has not yielded the performance gains hoped for by high end computing users. SGI has adapted its HPC strategy to these trends, taking a commodity-based approach, while developing important capabilities that are critical to moving beyond the limitations of Moore's Law and cluster-based parallelism

## The SGI® Altix® System Development

SGI has been designing scalable shared memory systems for many years. By the late 1990's, proprietary components were increasingly hard to justify given the rising performance of lower cost, mass market parts. SGI chose to focus on providing value to its customers through its scalable system architecture.

The first system to be delivered to the market using this approach was the SGI Altix server. First introduced in 2003, the Altix moved away from the proprietary based designs of the earlier Challenge and Origin systems, to an approach that leverages more COTS (common off the shelf) components.

1. Intel® Itanium®2 processor (replacing MIPS).
2. ATI Graphics processor (replacing proprietary SGI designs).
3. Standard Linux® OS (replacing proprietary IRIX).
4. Commodity SDRAM DIMM (replacing proprietary DIMM designs).

This strategic direction has allowed SGI to continue delivering state of the art system performance to its customers, while driving down system costs to meet ever increasing market expectations. This performance is delivered through the powerful SGI NUMAlink™ interconnect and NUMAflex™ scalable shared memory architecture.

## SGI NUMAlink Interconnect and NUMAflex Architecture

The NUMAlink interconnect is the fabric that binds all the elements together in SGI's NUMAflex system architecture. Multiple generations of NUMAlink have been designed sharing a number of common elements. Firstly, high bandwidth per pin. This is key to building scalable systems where physical packaging constraints come into play. For example, it helps the design of router ASICs with strong bandwidth and radix characteristics. Low latency interconnect is also key to building scalable systems. Improving the latency of critical operations within the interconnect provides the ability to increase the level of parallelism that applications will be able to effectively exploit. NUMAlink was designed to have an encoding of messages on the physical channel that is optimized for the transfer of small messages, while maintaining highly reliable data transport.

System packaging of scalable systems also plays significant importance as designers look to squeeze latency from interconnect paths. NUMAlink has been designed to deliver high performance over printed circuit boards as well as cables. This allows a uniform interconnect approach for scalable systems, both within a packaging enclosure, and for cabling packaging enclosures together. It also removes the need to add additional buffers etc. to drive signals between different components.

A fundamental aspect of the NUMAflex system architecture is the provision of globally addressable memory. This allows the processing elements to directly operate on data through load/store semantics at the hardware level, without the need to go through a software layer as a cluster based approach requires. It's also important to note that a system with shared memory supported at the hardware level does not restrict you to only running applications designed to take advantage of shared memory. In fact, the SGI Altix system delivers industry leading MPI performance, even though MPI was designed to suite distributed memory or cluster style systems. In addition, globally addressable memory also makes operations such as dynamic load balancing more effective.

Scalable shared memory provides the foundation to build an operating system that can handle large processor counts. Altix currently supports a single operating system image size up to 512 processors. This can significantly reduce the administration costs of large HPC systems. Imagine the ease of managing just one operating system for 512 processors, as compared with the effort involved with 256 copies of the operating system in a cluster of 2P nodes. Beyond this, a large operating system image can also provide users with a more productive environment to work within. Users can easily manage their processes and datasets without a need to co-ordinate between systems. Users are also able to take advantage of a broader set of programming paradigms, such as OpenMP programming, up to much larger processor counts.

## Multi-Paradigm Computing

With the barriers encountered to the advancement of processor core performance, general purpose processors have now turned to integrating multi-cores on a single chip to make use of the increasing number of transistors. This approach is attractive to the chip manufacturers as it is a relatively easy way to make use of the silicon area. However, not all applications will benefit, especially HPC applications that tend to require high bandwidth access to off chip memory. Further, the power and cooling problems associated with these designs present additional barriers to progress in this mode.

Given this scenario, industry observers and users alike have begun to recognize the need for a more effective approach to performance. SGI recognized this some time ago and therefore embarked on a development path to advance its shared memory architecture by supporting a variety of tightly coupled alternative processing elements. The first example of this approach can be seen with the visualization subsystem of SGI servers, where GPUs (graphics processing units) are deployed to accelerate graphics processing. These were initially connected into the system through standard IO interfaces (PCI-X/AGP).

More recently, SGI introduced the capability to *directly* couple novel processing elements into the NUMAflex system architecture. SGI's TIO ASIC provides a standard IO interface and in addition, it has a scalable system port (SSP), which opens up the NUMAlink interconnect to novel devices. Not only can these devices have full access to the high bandwidth NUMAlink interconnect, but the NUMAlink communication protocol is also opened up so these devices can interact directly with all shared memory in the system, as well as other mechanisms such as atomic memory operations.

The first product to exploit the SSP port couples FPGA (field programmable gate array) technology directly into NUMAlink. FPGAs, like general purpose processors have gained greatly from Moore's law. However, unlike scalar processors which have hit the wall with regard to increasing single thread performance, FPGAs can directly take advantage of the continued growth of available transistors. FPGA programming techniques directly expose the parallelism in an algorithm to the hardware so that it is possible to gain orders of magnitude performance improvements for some algorithms/ applications. SGI has introduced this technology to the marketplace as RASC™ – Reconfigurable Application Specific Computing.

Now that SGI has opened up the capability to attach novel devices to its scalable system architecture, it is working with a number of partners to exploit this ability and deliver exceptional performance gain for a variety of HPC applications. One example is the Clearspeed™ floating-point accelerator. Systems can be configured to deliver the best-performing processing elements for particular applications.

The move to multi-core processors also raises interesting questions about the direction of programming models. MPI programs have typically been written to run well on cluster based systems with multiple micro-second latencies. With multi-core and multi-threaded processors, it would be of great benefit if the programming model could take advantage of the evolving hierarchy of communications latency. Here is where the global addressable memory approach of the NUMAflex architecture allows seamless scaling of programming models. Programmers can and already do take advantage of the growing power of multiple processors and compute nodes, by reducing the need to re-code or re-optimize as operations that were once separated by node or chip boundaries move across those boundaries with ease.

### Looking Ahead: SGI's Project Ultraviolet

Project Ultraviolet will build on the multi-paradigm capabilities delivered with Altix. The system architecture will be able to scale from a single node, all the way up to Petascale systems. Key advances will include:

- A new generation of interconnect, that will increase the global addressing reach, and implement communications protocols to increase the efficiency of packet and message level data transport.
- The incorporation of novel processing elements in addition to robust support for the next generation of Intel processors as the general purpose processing elements
- A second generation of the SSP to provide even greater control to these devices to increase data transport efficiency within the system architecture.
- A new data transport capability to deal with algorithms that traditionally mapped onto a vector paradigm. This will be used to supplement the microprocessors when dealing with data items that don't fit the cache-line orientated designs that mass market processors use.

Ultraviolet comprises a truly elegant combination of both flexibility and performance; one that can support everyday workload demands with a new level of productivity, while scaling up to power the next grand challenge problems which can't afford the limitations of today's clustered processor approach.

### Summary

The HPC industry is facing new challenges as IC technology continues to deliver on Moore's law growth of transistors, but cores used at the heart of many cluster based systems stall in the delivery of better single thread performance. Novel computing elements, such as FPGAs and highly parallel floating point accelerators, are offering new potential to drive application performance forward. As we move towards Peta-scale computing, what will the programming models be to make effective use of such systems? SGI is taking a multi-paradigm approach, with its globally addressable memory architecture as the foundation, to build cost effective, scalable and versatile systems. SGI is already delivering on this vision, with the technology to build innovative solutions that directly address the problems of building high performance, high productivity systems.

# HDF2AVS - A Simple to Use Parallel IO Library for Writing Data for AVS

Craig Lucas and Joanna Leng
Manchester Computing, University of Manchester

### Motivation and Background

HDF2AVS is a library of routines to write out data, in parallel, in the HDF5 (Hierarchical Data Format) [3,4] data format, for input into the visualization system AVS/Express (Advanced Visual Systems) [1,2]. It is available as a Fortran 90 module and consists of various routines for writing out different types of array or coordinate data.

HDF uses a tree structure of groups and datasets. A group consists of zero or more datasets and other groups, together with supporting metadata. Datasets contain multidimensional arrays of data elements. The library is still in development at NCSA (National Center for Supercomputing Applications) [6]. We chose HDF for many reasons:

- it is a user defined format like XML
- it is a binary format that allows compression so drastically reducing the size of data files
- it is a format with longevity (NetCDF4 [7] is to be implemented on top of it)
- there is a dump facility that allows users to investigate the contents of binary files easily
- there is a reader for HDF already within AVS/Express
- parallel IO is supported.

There are many advantages to writing out data in parallel. On parallel machines there are two traditional approaches to writing data. One is to collect all data to one processor and then write this to disk. This obviously creates a communication overhead and can be slowed down further if there is not enough local memory on this master processor to hold the entire data set. These problems can be avoided by the other standard approach

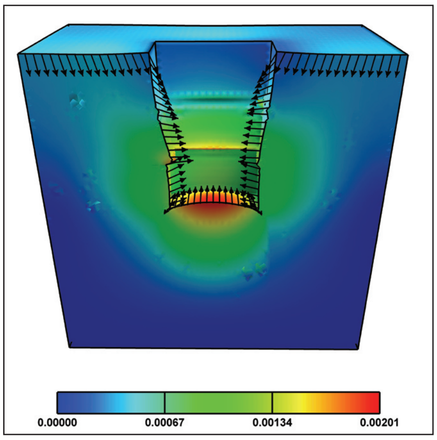


Figure 1: The results of an FEA soil excavation and subsidence experiment using an inhomogeneous soil model, generated by a random field generator.

of having each processor write to its own file. This introduces a post processing overhead, and also limits you to read the files back in on the same number of processors. Parallel HDF is built on top of MPI-IO [5] which allows simultaneous access to the same file, thus removing all of the problems above.

The HDF files are in binary. In fact, MPI-IO itself only supports binary data. Binary (and compressed binary) files are much smaller than ASCII files, easily by a factor of 3 or 4. Our experience shows that many computational scientists write ASCII data, as they are human readable, and binary files, of course, are not. However, HDF supply the excellent utility, h5dump, to dump to the screen all or part of a binary HDF file, thus removing any disadvantages of binary data. Figure 1 shows the visualization of some FEA data [7]. Stored as ASCII the total file size for this data is 57MB. Converted to HDF format the size is 20MB and a further reduction can be obtained by converting to a compressed HDF format giving a file size of 13MB. Although this is a small example, the potential in reducing file sizes is obvious for very large datasets.

We chose to write the HDF2AVS wrapper library for many reasons. Firstly, the AVS reader requires much metadata; for example the dataset requires metadata for the data type, its dimensions and the grid used. In fact, to write out a 2 dimensional array, in a form that the AVSreader understands, requires 87 HDF subroutine calls! However, only one call to HDF2AVS.

Also, some of the concepts in HDF are not straightforward, with data and file spaces, property lists and unique data types. But to use our library the user is not required to understand any of these.

### Using HDF2AVS

To write out an array in HDF2AVS you just need to pass the dimensions of your global data, and the part of that global data owned by the calling process. For example consider writing a 2 dimensional array.

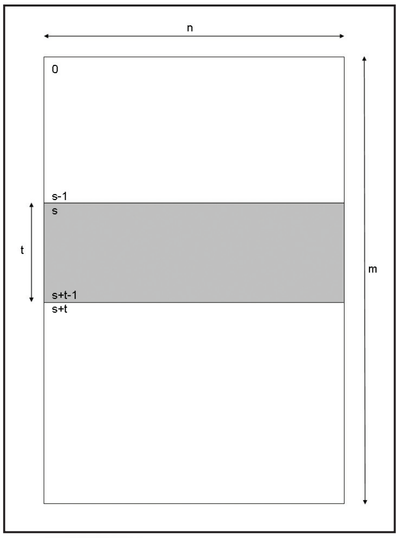


Figure 2: An example data distribution for HDF2AVS. The global array has *m* rows and *n* columns of data. The local array has *t* rows of data, and is shown shaded.

Figure 2 shows a possible data distribution for the local arrays distributed over blocks of rows of the global array. The dimension of the global array is *m* x *n*, and the process here owns *t* rows of data, the first of which is in row *s* of the global array. Note HDF numbers its coordinates from zero.

There is no constraint in where or how big the local array is. For instance, local arrays can overlap, and need not span the whole width and not all of the global array needs to be written to. We do, however, check that you don't try to write outside of the global array, which is not permitted in HDF.

```
PROGRAM using_hdf2avs
USE MPI
USE HDF2AVS
.
! Could be REAL etc
INTEGER, ALLOCATABLE, DIMENSION(:,:) :: my_data .
INTEGER(KIND = 8), DIMENSION(2) :: global_dims
INTEGER(KIND = 8), DIMENSION(2) :: local_dims
INTEGER(KIND = 8), DIMENSION(2) :: global_origin
.
global_dims = (/m,n/)
! Data split in rows
local_dims = (/t,n/)
global_origin = (/s,0/)
ALLOCATE( my_data(t,n) )
.
CALL MPI_INIT(error)
.
CALL HDF2AVS_WRITE_2D_ARRAY(my_data, global_dims,
&
           local_dims, global_origin, ‘filename.
h5’ ) .
CALL MPI_FINALIZE(error)
END PROGRAM using_hdf2avs
```

Figure 3: A code fragment, in Fortran, calling HDF2AVS to write out a 2D array.

In Figure 3 we give a code fragment, in Fortran, that calls the HDF routine HDF2AVS_WRITE_2D_ARRAY. Note the generic routine name for our supported data types: `integer, real` and `real*8.`

The routine takes just five arguments:

- `my_data` is the local data on the process
- `global_dims` is the dimension of the global dataset distributed over all processes
- `local_dims` are the dimensions of `my_data`
- `global_origin` gives the position of the local data in the global array
- the file name is also required.

This is a collective call made by all processes.

## Performance

In writing the library we were not driven by performance issues. We have concentrated on functionality rather than performance. We wanted to help users to write their data out easily to a useful visualization format. It is not intended to replace your bespoke IO routines.

Consider the example where you have data on each processor and you wish for this data to be written to a single file. We have 1GB of data on each process, so the more processors used, the larger the file. We can write out the data using HDF2AVS with a single collective call. Alternatively, a master process receives the arrays from each of the other processors and writes them individually to the file.

Figure 4 shows the time to write the HDF file and also the time to receive and write the individual files in a binary format. Note that the time taken to write ASCII data is very much increased, and it not shown. You can see that the times vary enormously for different runs, up to a factor of 10. There are many issues that affect the performance at a particular time: how busy the machine is, where the processors are on the machine, whether there is enough memory available to buffer the data before writing and what other codes running are demanding OS services etc. We used the CSAR machine Newton for these timings, which was heavily loaded.

## Further Work and Information

The library is still under development, planned improvements include:

- Support for cell data for FEA applications.
- Optional arguments for altering the metadata.
- MPI-IO optimizations.
- Routines for reading files.
- Other data distributions.
- Documentation.

The aim of this article is to gauge interest. The development can be driven by user requirements.

If you think this library could be useful to your applications please contact either Craig Lucas or Joanna Leng. We can advise on how the library could be used or adapted for your application and dataset.

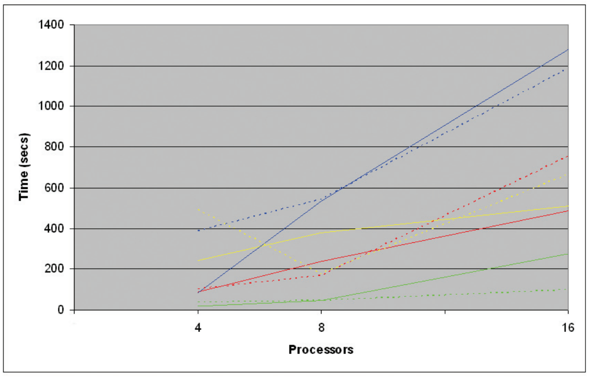


Figure 4: Comparison of HDV2AVS and gathering data on one process to write to a single file, for a 2D array. Each colour relates to the same 16 processors for both methods run concurrently. HDF2AVS is shown solid and gathering dashed.

## References

[1] Advanced Visual Systems Inc. *AVS/Express Developer's Reference*, Release 3.1.

[2] Advanced Visual Systems Inc. Homepage: http://www.avs.com/

[3] HDF Homepage: http://hdf.ncsa.uiuc.edu/index.html

[4] HDF5 API *Specification Reference Manual*

[5] MPI Forum Documentation: http://www.mpi-forum.org/docs/docs.html

[6] NCSA Homepage: http://www.ncsa.uiuc.edu/

[7] NetCDF Homepage: http://www.unidata.ucar.edu/software/netcdf/

[8] Smith I.M., Leng J. and Margetts L., *Parallel Three Dimensional Finite Element Analysis of Excavation*, ACME, Sheffield, March 2005.

# SIGGRAPH 2005

*Javier Gomez Alonso, Michael Daw and Martin Turner*
*Manchester Computing, University of Manchester*

SIGGRAPH 2005, the 32nd International Conference on Computer Graphics and Interactive Techniques, took place in Los Angeles between 31 July and 4 August 2005. There were more than 29,000 attendees and 250 exhibitors at this prestigious event . An important part of the conference included the exhibition that ran over the last three days of the conference.: James Perrin from Manchester Computing was invited by SGI Inc. to demonstrate the recently developed fully parallel version of the scientific visualization software AVS/Express using the next generation of SGI graphics hardware.

The Emerging Technologies programme, a part of SIGGRAPH 2005 aimed at 'interacting with digital experiences that move beyond digital tradition, blurring the boundaries between art and science, and transforming social assumptions', saw part of the conference broadcasted over the network in an attempt to reach a broader audience. Broadcast options included Multicast MPEG-2 streams and the Access Grid (next generation of video conferencing). Eleven carefully selected remote locations participated in what happened to be the debut of Access Grid in SIGGRAPH's history. Because of its excellent reputation, Manchester Computing was invited to be one of these selected locations.

A combined Access Grid and stereoscopic visualization facility available at The University of Manchester's e-Science North West Centre provided the key UK venue. Food and drink were provided for local attendees, whose backgrounds ranged from the arts to sciences. The sessions covered a variety of topics from distributed live performances, to art and media panels, to the highlight of the Emerging Technologies programme: the "AGJuggler" demonstration. Manchester Computing's Manchester Visualization Centre took an active part in this demonstration, in collaboration with Purdue University and presenters at the Los Angeles Convention Center.

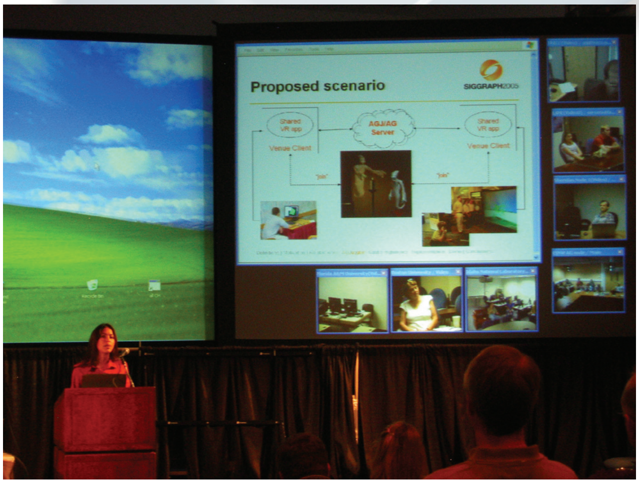


Figure 1: AGJuggler presenter at the LA Convention Center. Background screen showing AG videos of the remote locations.

AGJuggler is a toolkit for collaborative Virtual Reality (VR). It consists of a set of libraries that provide routines and functions that can be added to existing VR applications in order to have them running in geographically distant Access Grid nodes. The routines offered by the libraries enable them to be Grid-aware, i.e. they can be run collaboratively on any Access Grid node. A typical AGJuggler session uses the Access Grid toolkit to share data and communications between all remote participants. By using the toolkit, each client maintains local information about participants, their status and device data. For communication among participants, AGJuggler does not transmit video frames or geometrical models, but only the data relevant to the users' state, thus using a very low network bandwidth and providing a higher frame rate than alternative approaches. Supported hardware setups range from fully immersive CAVE-like systems with user motion tracking, to PCs equipped with active stereographics capability, or even personal laptops. The Manchester Visualization Centre Access Grid-enabled their passive stereoscopic visualization system in order to join the demonstration, which consisted of a 3D tour through the virtual haunted house of 'Castle Highmoore'.

Large projection systems are often expensive and fixed to a specific location. A portable (luggable) passive stereoscopic visualization facility has now been constructed based on the Geowall design (www.geowall.org) that incorporates the Access Grid. This is available for local or remote installation and hire, and after being demonstrated at this event, was available for viewing at certain conferences and shows during

September. This construction was funded by the JISC project SAGE, Stereoscopic Access Grid Environment.

The next similar major international events to be made available over the Access Grid are the Fifth Virtual Conference on Genomics and Bioinformatics and SC Global 2005, for which cheese and wine will be provided to encourage attendance at the late sessions. We hope to see you all there.

For more information on these events, please visit http://www.virtualgenomics.org/vcgb/conference_2005.htm and http://sc05.supercomputing.org/programs/sc_global.php.

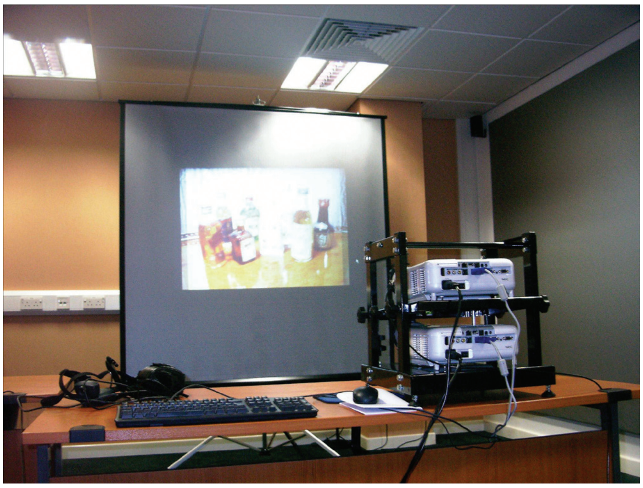


Figure 2: Manchester Computing's portable passive stereoscopic unit, showing 'intoxication' a 3D art piece by Karen Grainger.

# ISC 2005

*Carl Ward*
*Manchester Computing, University of Manchester*

This year ISC2005 in Heidelberg celebrated its 20th anniversary. The conference began its life as the Mannheim Supercomputer Seminar in 1986 when it was attended by 80 participants. This year's annual event saw attendance swell to over 650 participants from 30 different countries.

The University of Manchester has held a booth at ISC since 2001. This time I hosted the booth along with Mike Pettipher and James Perrin. It was my first visit to the conference and apparently I was lucky enough to help host a particularly busy year. For the first time in the event's history the organisers had decided to release over 80 exhibition day passes which increased attendance. Over the three days I was able to speak to people from a variety of institutions and organisations including Warsaw University and the Joint Supercomputer Centre in Moscow.

Mike and James attended a number of the conference presentations, including the opening day's keynote presentation "Progress in Supercomputing: the Top Three Breakthroughs of the Last 20 and the Top Three Challenges for the Next 20 Years" given by Horst Simon. James also worked on the booth to demonstrate Reality Grid's computational steering, on-line visualization and check-pointing tools. The evening's Get Together Party included the biggest bowl of ice cream I've ever seen in my life, all flavours heaped together to form an impressive mountain.

The following day saw the Steve Louis keynote presentation "Peta-scale Computing During Disruptive Times". Visitors continued to peruse the work of Manchester Computing and this day, being the most hectic, passed very quickly. In the evening it was time for the ISC 20th Anniversary Gala Event which took place at the idyllic Castle Neckarbischofsheim, a former bishop's residence. Excellent food and drink was served, including spaetzle, something I'd never had until this my first visit to Germany, it's an alternative to the potato (tiny noodles or dumplings made with flour, eggs, water or milk, salt and sometimes nutmeg) - very nice but think I prefer the trusty spud. The Gala Event was a great opportunity to meet fellow participants in a pleasant location away from the confines of the conference hall.

Wolfgang Gentzsch's keynote speech on the final day entitled "Grid Computing in Research and Business around the World", brought ISC2005 to a close and ended the event's five year association with Heidelberg. Sadly this was the last time the conference will be held in this beautiful part of Germany. Next year it moves to a new home in Dresden and promises even further improvements to what is an already well attended and respected event.

CSAR,
Manchester Computing, Kilburn Building,
University of Manchester, Oxford Road,
Manchester. M13 9PL
Tel: +44 161 275 6824/5997
Fax: +44 161 275 6040
E-mail: csar-advice@cfs.ac.uk

## CSAR Information

The CSAR Website - http://www.csar.cfs.ac.uk - contains help and information on all aspects of the service, including sections on Software, Training, Registration and Project Management and Accounting.

Additional information, particularly with regards to service developments and other events associated with the service, is also provided via a monthly bulletin issued by email to all users. An archive of these bulletins is available at http://www.csar.cfs.ac.uk/about/bulletin

CSAR Focus is published twice a year and is also available to view on the Web - http://www.csar.cfs.ac.uk/about/csarfocus. To change your mailing address please email your new details to the CSAR Helpdesk.

## Getting Help

If you require help on any aspect of using the CSAR service you can contact the CSAR Helpdesk team who will deal with your query promptly and efficiently.

Telephone: 0161 275 5997 / 0161 275 6824

Email: csar-advice@cfs.ac.uk

The CSAR Helpdesk is open from 08:30 - 18:00 Monday to Friday, except on Public Holidays.

**ISBN 1470-5893**

CSAR FOCUS ~ Computer Services *for* Academic Research