The University of Auckland

Centre for eResearch

Research Case Studies 2013

THE UNIVERSITY OF AUCKLAND

NEW ZEALAND

Te Whare Wānanga o Tāmaki Makaurau

Contents

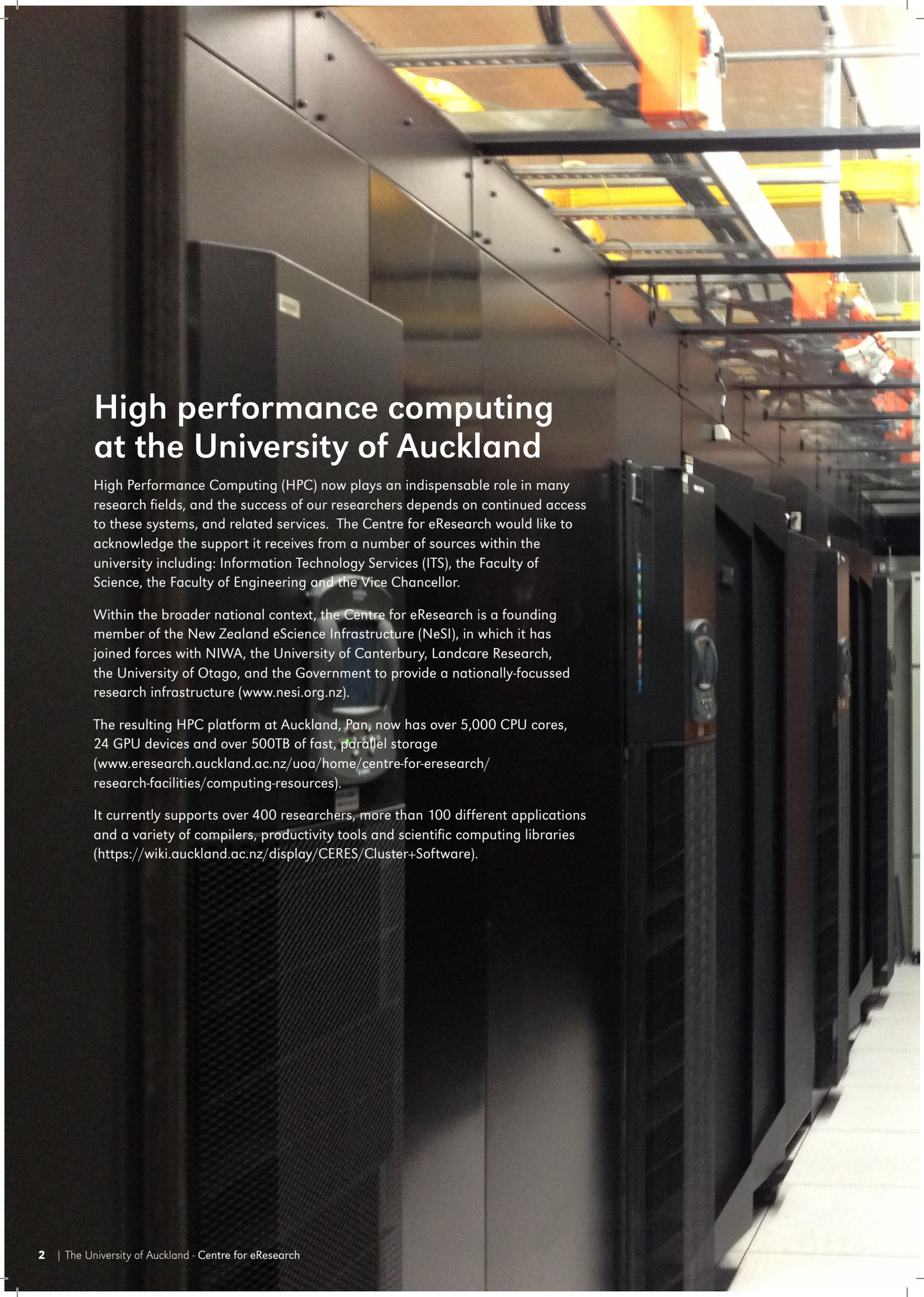
Our Vision

The computational demands of today's research questions often require specialised computing and storage infrastructure, along with advanced software tools and services. The Centre for eResearch pioneers the development of such infrastructure and services on behalf of the University's research communities. We work hand-in-hand with researchers to help them leverage computational advances in pursuit of their own research questions. To this end we provide a sophisticated high performance cluster computing environment, advanced storage solutions tuned for managing large research data collections and a variety of software services.

Our vision at the Centre for eResearch is to be a nationally and internationally influential community of eResearch practitioners, developers and thought-leaders. We aim to excel in facilitating the research of others, as well as contributing to the evolving research themes within the field of eResearch itself.

This publication showcases a small fraction of the research projects from 2013 that the Centre for eResearch has helped to facilitate.

Professor Mark Gahegan
Director, Centre for eResearch



High performance computing at the University of Auckland

High Performance Computing (HPC) now plays an indispensable role in many research fields, and the success of our researchers depends on continued access to these systems, and related services. The Centre for eResearch would like to acknowledge the support it receives from a number of sources within the university including: Information Technology Services (ITS), the Faculty of Science, the Faculty of Engineering and the Vice Chancellor.

Within the broader national context, the Centre for eResearch is a founding member of the New Zealand eScience Infrastructure (NeSI), in which it has joined forces with NIWA, the University of Canterbury, Landcare Research, the University of Otago, and the Government to provide a nationally-focussed research infrastructure (www.nesi.org.nz).

The resulting HPC platform at Auckland, Pan, now has over 5,000 CPU cores, 24 GPU devices and over 500TB of fast, parallel storage (www.eresearch.auckland.ac.nz/uoa/home/centre-for-eresearch/research-facilities/computing-resources).

It currently supports over 400 researchers, more than 100 different applications and a variety of compilers, productivity tools and scientific computing libraries (https://wiki.auckland.ac.nz/display/CERES/Cluster+Software).

CASE STUDY 1

Simulating quantum mechanics on high performance computing cluster

Arne L. Grimsmo, Department of Physics

Emerging advances in laboratory methods allowing fundamental laws of nature to be studies more closely

The laws of quantum mechanics are often considered mysterious and counter-intuitive where many elusive phenomena predicted by quantum physics have not been directly observable. Fortunately in the last couple of decades impressive progress in the control of single atoms and quanta of light has opened up the fundamental laws of nature to direct experimental study. This offers a promising future where we can employ "quantum technology" to perform tasks not achievable by other means. Examples include quantum cryptography for secure communication or quantum computing for breaking cryptography schemes. The interest in the experimental progress in this field was brought to light when the Nobel prize in physics in 2012 was given to Serge Haroche and David J. Wineland, for their groundbreaking experiments in the manipulation of single atoms and single light particles.

The interaction of light with matter

My research interest within quantum mechanics involves studying the interaction of light with matter on the most fundamental level. Progress in high performance computing (HPC) has helped me simulate the laws of quantum mechanics. This is a very numerically demanding task due to how the size of a quantum system "scales" with the number of particles it consists of. It is therefore necessary to use high performance computer clusters to simulate quantum systems of even moderate size. Such simulations are of great value as outcomes can be used to guide the experimental efforts in the lab and also because simulations offer access to information that is hard or impossible to retrieve by measurement. In this way one can study many elusive phenomenon, such as quantum entanglement, in detail.

At the University of Auckland my co-workers and I have recently made an effort in this direction by simulating a large number of atoms interacting with the electromagnetic field on the NeSI Pan high performance computing cluster. We assume that the atoms are trapped inside a cavity, consisting of highly reflective mirrors, which increases the strength of the interaction between the electromagnetic field and the atoms. This system has a fascinating property, first predicted theoretically in the 1950s, but not witnessed in experiment until 2010; namely, if the interaction strength between the atoms and electromagnetic field becomes sufficiently strong, the system enters a so-called "super-radiant phase" emitting light at high intensity higher than what one would expect just by adding up the rate of emission from each atom. In the presence of this "collective effect", where the atoms cooperate to generate light, the atoms and the photons present inside the cavity can no longer be viewed as individual constituents, but are entangled in such a way that they create a fundamentally new type of matter.

By simulating this transition from "normal phase" where the atoms and photons act on an individual basis to the "super-radiant phase" where they collectively cooperate on a computer cluster, we can gain insight into this entangling of light and matter. Our goal is to capture a genuinely quantum phenomenon of entanglement between atoms and an electromagnetic field.

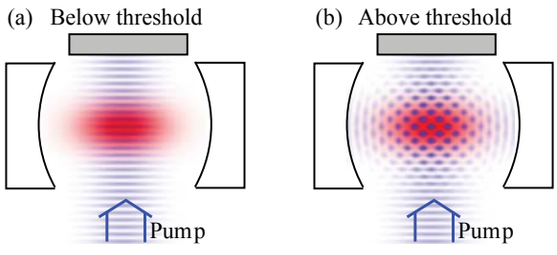


Figure 1

***Figure 1 (a) and (b)** show the cavity mirrors in white blocks and the cloud of atoms in red. The strength of the interaction between the atoms and the cavity is controlled by the strength of a laser pump shown in blue. (a) Below threshold: Here the coupling strength is weak and the atoms act independently of each other. (b) Above threshold: Here the coupling strength is strong, and the atoms self-organised in a lattice, and cooperate to generate light at high intensity; the so-called super-radiant phase. The atoms and the light are now entangled to make up a new type of matter that is more than just the sum of its parts.*

Simulation design

Simulations for a quantum state was represented by 720×720 complex numbers. We applied a matrix to the quantum state viewed as a vector which involved a matrix size of $(720 \times 720) \times (720 \times 720)$. To do one run approximately 90 GB of RAM was needed which is largely available on the NeSI Pan cluster.

One hundred simulations were run in parallel which quickly mapped out the behaviour of the system over a large range of model parameters. By looking at the numerical results for the quantum state, one can witness the system going through several transitional regions of different behaviour as we vary the model parameters. Qualitatively, these different behaviours of the quantum state are different phases of the "quantum matter" inside the cavity. The Figures below show visualisation of the quantum state of the electromagnetic field in the cavity, in two different phases for the quantum matter, inferred from our numerical simulation.

International collaboration

Parallel to our numerical simulations, we are now in collaboration with a group of researchers at the National University of Singapore (NUS) who are implementing laboratory experiments to realise the same model. In drawing upon our results, these simulations can provide a useful guide to the researchers at NUS and help them understand the dynamics. We also believe that through the collaboration, we can anticipate novel experimental results and probe deeper into the quantum nature of the interaction of light and matter.

My research efforts in Auckland, and the collaboration my co-workers and I now have with Singapore, would not have been feasible without access to the NeSI computing cluster and the help received from the technical staff at the Centre of eResearch in setting up our simulations. The group at the Centre of eResearch stand out for how they understand the needs of the scientists, and the effort they make to streamline the use of the cluster for us on an individual basis.

Figures 2-3 *show a visualisation of the quantum state of the electromagnetic field inside the cavity. Figure 2: This is what the EM field looks like below threshold. It is in the "vacuum" state (zero photons) which is the same state it would be in if there were no atoms present. Interestingly, due to quantum mechanical fluctuations, even the vacuum state has an interesting structure. Figure 3: The EM field above threshold; the superradiant phase. The two-peak structure with concentrations away from the origin (0,0) is a quantum state with a large number of photons.*

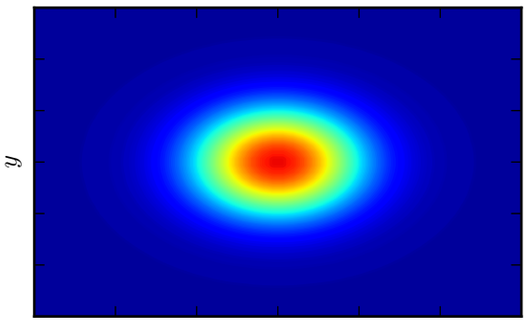


Figure 2

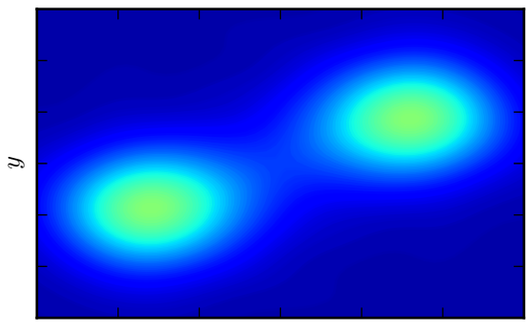


Figure 3

CASE STUDY 2

Planet hunting

Philip Yock, Department of Physics

Using microlensing to detect planets

Astronomers and philosophers have long thought about planets like ours, possibly inhabited, orbiting the stars of the night sky. Assuming the solar system is typical, a rough estimate of planets in the Milky Way would be around one trillion. Confirming this prediction has, however, turned out to be far from trivial.

The first planet to be found orbiting a star like the Sun was not found until 1995. The reason is straightforward. If we attempt to see a planet like Earth orbiting a star like the Sun, we need to search very close to the star, because even the nearest stars are at vast distances. The proximity of the planet to its host star implies it is lost in the glare of the star. For this reason, indirect detection methods are required.

New Zealand astronomers have been largely responsible for pioneering a planet detection technique which utilises the bending of light by the gravitational field, an effect that was predicted by Einstein early last century.

The method is illustrated as Figure 1. If two stars are aligned, as seen from Earth, then the nearer one acts as a gravitational "lens" of the more distant one. If planets orbit the lens star, their gravitational fields can perturb the lensing action to a detectable degree.

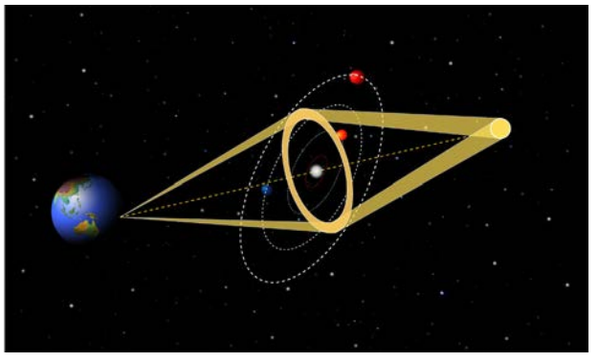


Figure 1: *The bending of light in the gravitational field of a star leads to the formation of an Einstein ring. The size of the Einstein ring is comparable to that of typical planetary systems. It is therefore possible for planets orbiting the "lens" star to betray their presence by perturbing the ring.*

NZ astronomers using the microlensing technique are members of international collaborations known variously as MOA, MicroFUN and PLANET. Of these groups, MOA has the largest presence in New Zealand with members from Auckland, Massey and Canterbury universities.

Simulating light rays

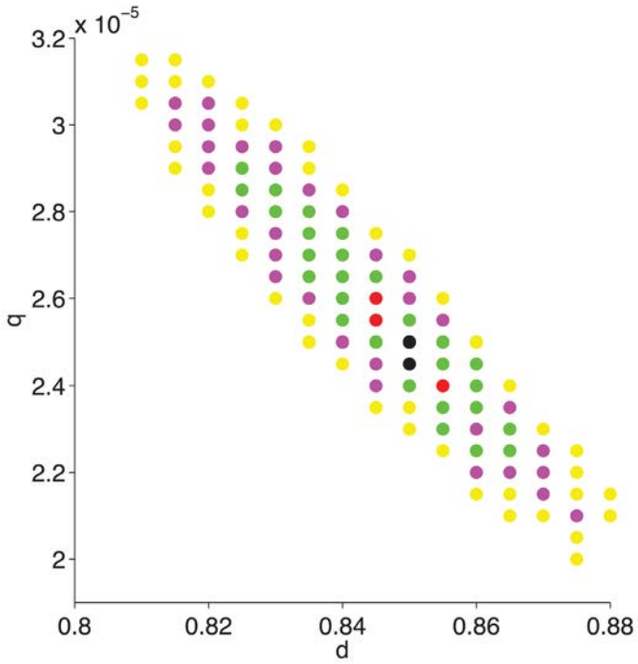
All stars are in motion. The formation of an Einstein ring is therefore a transient phenomenon only, visible for a matter of days. As two stars move into and out of alignment, the effect grows and subsequently diminishes. The effect of planets on the lensing process is most easily seen when the alignment is highest. The NZ groups have made good use of this seemingly self-evident effect. Analysing data on these "microlensing events" is not straightforward. It is very much a multi-parameter problem with at least five, ten or fifteen parameters that need to be varied to find the best match to the data.

At the University of Auckland a 'ray shooting' method has been devised. In this, simulated light rays are shot through an assumed lens system configuration, rather like optical lens designers do, and the computed image is compared with observation. Very large numbers of trials are made especially for a lens systems with multiple planets.

Trials are made one at a time in a procedure that is parallelisable and ideally run on clusters of personal computers (PCs). The members of the MOA group at the University of Auckland have been very fortunate to receive much assistance with this.

Dr Nicholas Rattenbury pioneered the use of a cluster for microlensing, and his work was extended by Drs Christine Botzler and Yvette Perrott. They used teaching PCs that were kindly made available on evenings, weekends and holidays thanks to James Harper and Sean Davidson of the University.

More recently, the Pan cluster became available which was used successfully by MSc student Matthew Freeman and summer scholar Charlotte Airey with assistance provided by Sina Masoud-Ansari. Some of their results appear below.



Series of over one hundred trials showing colour coded chi square values that was carried out in one night's computing on the Pan. The results would have taken some months to obtain on a single PC. They were published in a paper on the detectability of Earth-like planets that appeared in the Monthly Notices of the Royal Society, Vol 431, pp 2975-2985 (2013).

The telescopes used by the MOA group at the Mt John University Observatory in Canterbury. MOA is a Japan/NZ collaboration, and the nearer of the two telescopes shown above was supplied by Japan. With a diameter of 1.8m, it is the world's largest telescope dedicated to microlensing.

Future plans

The international microlensing community currently operate a growing number of telescopes, currently in excess of twenty, to monitor microlensing events. The quality of data being obtained with these telescopes is growing continuously and also the sophistication of the data analysis. Already several exotic discoveries have been made including the discovery of a large population of "free-floating" planets roaming interstellar regions in the Milky Way.

Further information

MOA: www.phys.canterbury.ac.nz/moa

MicroFUN: www.astronomy.ohio-state.edu/~microfun

PLANET: www.planet-legacy.org

CASE STUDY 3

Studying the shape and the size of the universe

Grigor Aslanyan, Department of Physics

Data from spacecraft helps refine theoretical models of the universe

The study of the universe as a whole goes back to as early as 1915 when Einstein introduced General Relativity. Only two years later, Willem de Sitter realized that General Relativity did not predict or constrain the global shape of the universe. There is no theoretical reason to believe that the universe is infinite. In fact, certain theoretical arguments about the birth of the universe give preference to finite non-trivial shapes such as a 3-torus (A. Linde, 2004). This subject remained purely speculative until very recent accurate measurements of the cosmic microwave background (CMB) anisotropies were made by the Cosmic Background Explorer (COBE) satellite (launched in 1989) and the Wilkinson Microwave Anisotropy Probe (WMAP) spacecraft (launched in 2001).

The CMB was emitted very shortly after the big bang so it gives a snapshot of the very early universe. A non-trivial shape of the universe changes the pattern of the CMB in certain ways which makes CMB an excellent tool to study the global properties of the universe. Several groups have addressed the problem before some of which found a slight preference for a finite rather than infinite shape of the universe. However, studying the statistical significance of the results (i.e., what the probability is that the universe has a non-trivial shape) is computationally challenging and certain approximations need to be used.

My research studying the shape and size of the universe involves doing a precise statistical analysis of three different shapes of the universe with the most recent data from WMAP (Figure 1) and testing the approximations used in previous studies.

Moreover, the Planck satellite which was launched in 2009 is expected in 2014 to deliver precise measurements of the polarisation of CMB. This essentially will double the amount of data currently available and can greatly improve our understanding of the shape of the universe. For the first time my colleagues and I developed the formalism to include the polarization data in the analysis. We then used simulations to predict how far the polarisation data from the Planck satellite can push the current limits on the shape and the size of the universe.

The shapes we studied are that of the torus in 1, 2, and 3 dimensions. These are among the most popular theoretical candidates. A torus in 1 dimension is simply a circle (the other two dimensions are infinite). If one travels in that direction she will eventually come back to the starting point. A 2-dimensional torus has the shape of a donut (the third dimension is infinite) (Figure 2). It is impossible to draw a torus in 3-dimensions, but it is a simple generalisation of the previous cases.

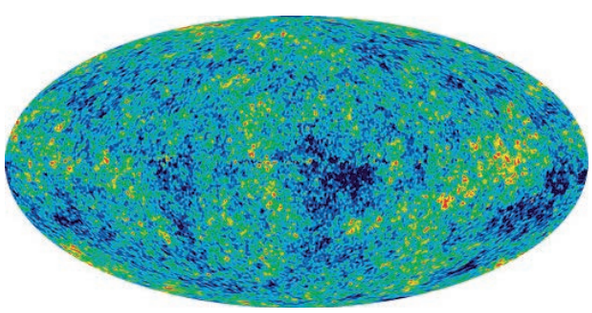


Figure 1: *The temperature anisotropies of the cosmic microwave background from the final release of WMAP data.*

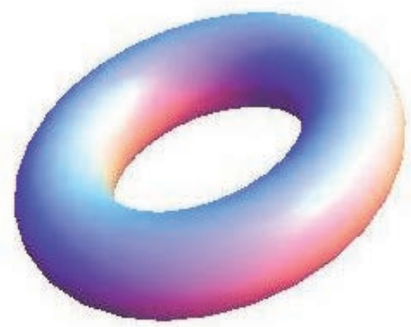


Figure 2: *A torus in two dimensions*

One can imagine it as a cube with opposite edges identified; i.e. if one hits the edge of the cube she simply enters the cube from the opposite side.

Our results were recently published in JCAP 08 (2013) 009, also available on arXiv, 1304.1811. They indicate that certain approximations used in previous studies tend to overestimate the statistical significance of a detection of non-trivial shapes. Simulations are crucial to accurately estimate the implications of the data about the shape and the size of the universe.

We found that the current data does not give a statistically significant preference to a non-trivial shape of the universe. Our results have also indicated that the polarisation data from the Planck satellite can push the lower limits on the size of the universe beyond what has been obtained so far with any methods.

It can also predict with high significance a torus topology with a size even bigger than the observable part of the universe. The techniques we have developed can be used with the actual data once it is released. The results could provide a very important piece of information about the connection between quantum physics and General Relativity and how the universe emerged from the quantum foam of spacetime.

Simulating the size and shape of the universe

Many simulations of the shape and size of the universe were needed to accurately estimate the statistical significance of the possible detection from the real experimental data. If the universe has a non-trivial shape then it must have a certain size and orientation. One needs to simulate data with all possible sizes and orientations with a sufficiently small step size. In order to obtain accurate results, on the Pan cluster we simulated 500 skies for 25 different sizes of the three topologies analysed – about 40,000 simulations in total. For each simulation we had to do a detailed fit with all possible orientations. As our results confirmed, doing this analysis was very important in order to find accurate results. This task would not be possible to accomplish without the usage of the Pan cluster. Each simulation could be analysed independently so it was very easy to use parallel computation. We were able to run many simulations on different cores of the cluster at the same time which significantly increased the speed of the analysis.

The most important next step will be to analyse the actual polarisation data from the Planck satellite when it becomes available in 2014. Another important future study would be to analyse a torus topology allowing all 3 sides to have different lengths. Unfortunately, this is computationally much more demanding than our current study, and even the usage of the Pan cluster would not help accomplish the task with the same techniques. One needs to come up with a different theoretical approach for that analysis, and this is one interesting and important study for the future.

CASE STUDY 4

Modelling dispersal and ecological competition in a statistical phylogeographic framework

Louis Ranjard, David Welch, Marie Paturel and Stéphane Guindon, Department of Statistics

Developing a new model for reconstructing colonisation history of habitats

Competition between organisms influences the processes governing the colonisation of new habitats. As a consequence, species or populations arriving first at a suitable location may prevent secondary colonisation. While adaptation to environmental variables (e.g., temperature, altitude, etc.) is essential, the presence or absence of certain species at a particular location often depends on whether or not competing species co-occur. Despite the potential of competition to maintain populations in isolation, past quantitative analyses of evolution have largely ignored that factor because of technical difficulties.

We have developed a model that integrates competition along with dispersal into a Bayesian phylogeographic framework. The proposed model takes genetic sequences and their spatial coordinates as input in order to date dispersal events and estimate dispersal and competition parameters.

Simulating landscape colonisation

Our statistical modelling on Pan uses the R software and high throughput is obtained by using parallel processing. In the first step to get an accurate understanding of how our approach performs, an extensive test on the NeSI Pan cluster was done by (i) performing simulations of organisms colonising space and competing while evolving and (ii) estimating the competitive and dispersal parameters from the genetic diversity generated by these simulations (Figure 1).

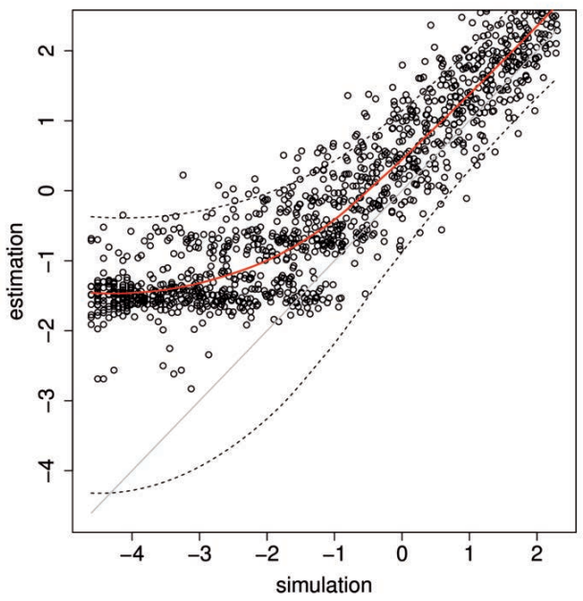


Figure 1: *Results from simulations shows that our approach can successfully detect the influence of competition on the history of colonisation of space inferred from molecular sequences. Each dot on the graph corresponds to a different simulation instance performed on the Pan cluster.*

In the second step, Pan was used to analyse a dataset of Hawaiian endemic species of the Banza genus called Hawaiian katydids –grasshopper-like insects. The genetic relationships among these organisms show evidence for island radiation with different species inhabiting the Hawaiian archipelago.

The phylogenetic tree built from mitochondrial sequences exhibits a striking "progression rule" pattern whereby each island appears to have been colonised only once. No obvious environmental feature varying across islands explains the observed geographical distribution of these organisms. Therefore, adaptation does not seem to be the factor causing the absence of secondary colonisaton.

We fitted our model to a rooted phylogenetic tree with node heights expressed in calendar time units (Figure 2). Using Pan we could demonstrate that competition can be detected with high sensitivity and specificity from the analysis of genetic variations in space.

In the next step, we are planning on extending our model to integrate extra environmental parameters that may have affected the colonisation of space. The NeSI Pan cluster will be used similarly to investigate the characteristics of the new model and how it can be applied to any biogeographical datasets for which genetic data is available.

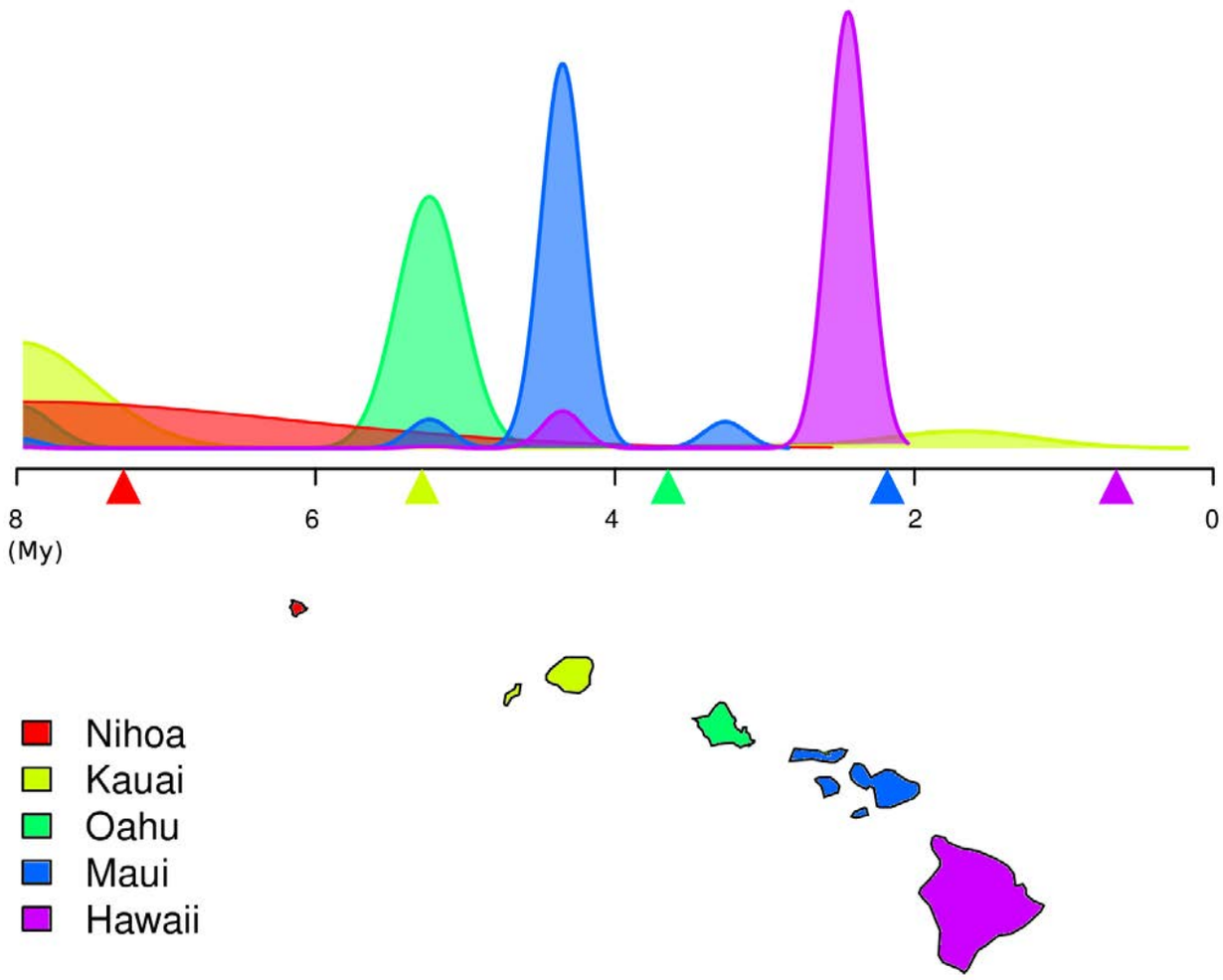


Figure 2: *Colonisation date estimates of each Hawaiian Island from molecular sequences for the Hawaiian katydids (in million years, My). Geological estimates of islands formation are indicated by coloured triangles under the time axis. While the order of colonisation and geological formation agree, the time shift between the two dating approaches can probably be explained by an underestimate of the genetic mutation rate in these organisms.*

CASE STUDY 5

Fully coupled thermo-hydro-mechanical modelling of permeability enhancement by the finite element method

Justin Pogacnik, Department of Engineering Science

Permeability controls the percolation of fluid in porous and fractured rock and is one of the most crucial hydrologic parameters.

In geothermal systems, permeability is typically controlled by fractures and fracture connectivity within the rock mass. Because of its complexity, permeability is often a very difficult parameter to evaluate and apply in any meaningful way. This is especially true in the area of permeability enhancement. For many problems, permeability should be regarded as a time-dependent parameter, that can be enhanced or inhibited over time by various processes such as chemical species dissolution and precipitation, changes in stress or pore pressure (effective stress), and by thermo-elasto-plastic effects (thermal cracking).

Permeability often determines the feasibility of some important geologic processes and their economic potential. This is especially true in geothermal energy production. If permeability is too low, then a geothermal well is often stimulated by over pressurisation (hydrofracking or hydroshearing) or injection of cold water to induce fracturing and enhance permeability. It is the understanding of the permeability enhancement process that is the subject of my research.

We have constructed a medium description that is based on the true spatial fluctuation of material properties taken from well log and well core data, a 'poroperm' medium. The poroperm medium is characterised by fracture density distribution that follows a pink noise spatial fluctuation and a related long-tailed permeability distribution that describes the fracture connectivity in the medium. Small changes in fracture density result in large changes in permeability. Specifically, we are simulating fluid percolation flow and heat flow through a deformable poroperm medium using the finite element method. Due to spatial fluctuations in porosity, and thus permeability, fluids tend to percolate through native permeability pathways in the medium. The key to sustainable and productive heat extraction lies in the enhancement of those native percolation pathways to allow sufficient flow rates without allowing cold water to pass to the receiver wells. We are seeking to understand the physics of injection induced strain damage that result in permeability enhancement of the existing native permeability pathways.

Simulating finite elements

Finite element simulations were performed on the NeSI Pan cluster using my own C++ software and using LAPack++ and Magma linear-solver software. Pan allowed me to run these simulations that would otherwise not be possible with just the resources in our lab. In the finite element method, a finer mesh solves the governing partial differential equations more accurately. CPU memory is often the limiting factor for the spatial resolution of the mesh. Finer meshes create many more degrees of freedom, which consume copious amounts of random access memory (RAM). The simulations performed in my research are fully coupled multiphysics simulations. Fully coupling different partial differential equations creates an additional memory burden by increasing the size of the linear system that needs to be solved. Adding solid mechanics to a fluid and heat transport system further increases the size of the linear system by a factor of three.

The simulations shown in figures (1,2) were performed on a single node of the NeSI cluster that included 12 cores and 96GB of RAM. Compared to my desktop computer which has 8GB of RAM, I have been able to run significantly finer grids in less time. I could not perform a finite element simulation with even half the mesh resolution of the images seen on my desktop computer.

While the initial thought of parallelising a serial code can be daunting, I found the process to be very simple using OpenMP and shared-memory parallelisation (SMP). The commands are easy to use and many linear solver packages have SMP versions that can be downloaded, compiled, and installed fairly easily. I also found the eResearch staff to be very helpful when installing necessary packages for my research. I'm able to run simulations with much finer mesh resolution that are more accurate as a result of the NeSI Pan cluster.

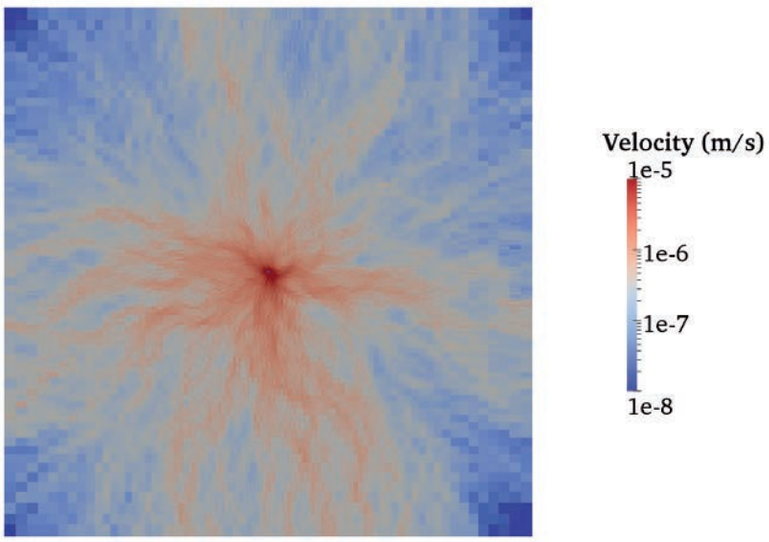


Figure 1: *Vertical 2D section that shows normalised fluid velocities around an injection well. Fluids tend to percolate in native permeability pathways as a result of fracture connectivity in the rock. The goal of native permeability enhancement is to stimulate these native pathways through shear deformation and damage to increase flow rates without compromising the heat of the reservoir.*

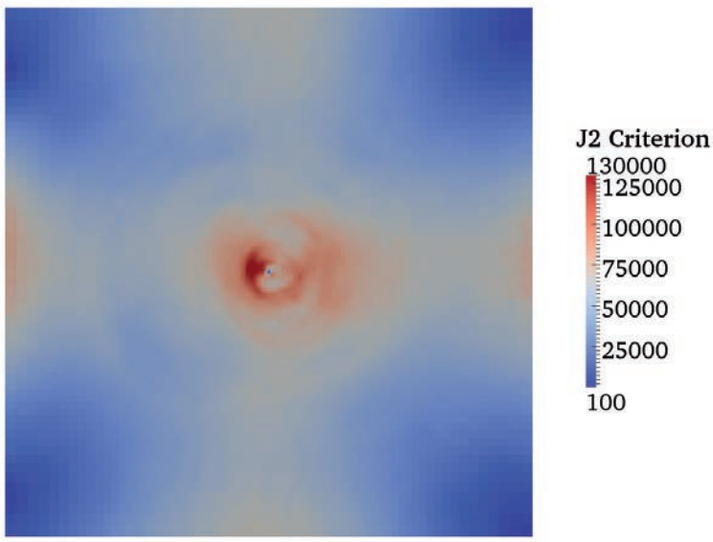


Figure 2: *Von Mises J2 yield stress criterion around the injection well from Figure 1. The J2 criterion is the second principal stress invariant or the difference between the maximum and minimum principal stresses. Areas of dark red indicate the areas where the most damage and yielding are likely to occur due to stress concentrations that result from native fractures in the rock. The damage in those areas results in an increase in fracture connectivity and thus fluid flow in the percolation pathways.*

CASE STUDY 6

Bayesian additive regression trees vs logistic regression – estimation of propensity scores

Samuel Passmore, Department of Statistics

Accurate comparison of hospital performance is crucial to the allocation of funding in New Zealand hospitals.

A problem for such comparisons is that patients and conditions are not randomly spread across hospitals. A range of factors, such as the economic environment surrounding the hospital, or a specialty unit within the hospital, mean that there may be disproportionate groups of demographics within any one hospital. When comparing hospitals' performance, the non-random spread of patients can be accounted for through the use of propensity scores. Propensity scores weigh the importance of patients so that the demographics are balanced between hospitals. A number of methods can be used to estimate propensity scores.

My Honours project research compared the performance of Bayesian additive regression trees (BART) to a logistic regression. BART is a sum of trees model where the growth of a tree is constrained by its priors, then using an iterative Markov-chain Monte-Carlo algorithm, back fits the model for optimal fit. This method is computationally expensive.

Analysing propensity scores

The NeSI Pan cluster allowed me to perform a parameter sweep to optimise the performance of BART and compare the results to a logistic regression. Results showed that a tailored BART performed marginally better than logistic regression in the estimation of propensity scores across the 9 hospitals. Figure 1 shows the difference in the proportion of patients in each age category of Hospital 17 when compared to all other hospitals. This result was fairly typical across the hospital comparisons.

Pan allowed me to run my analyses across 15 hospitals, with a sweep of parameter settings, in parallel. Considering a single run of a BART model would take between 10 – 16 hours, being able to run this in parallel saved me weeks in time and allowed me to focus on the results of the analyses rather than waiting for the calculation to complete.

For a project that only had a year time-frame this was very helpful. The staff at the Centre for eResearch were very helpful in helping me set up the analyses to run in the most efficient way which saved a lot of time and effort.

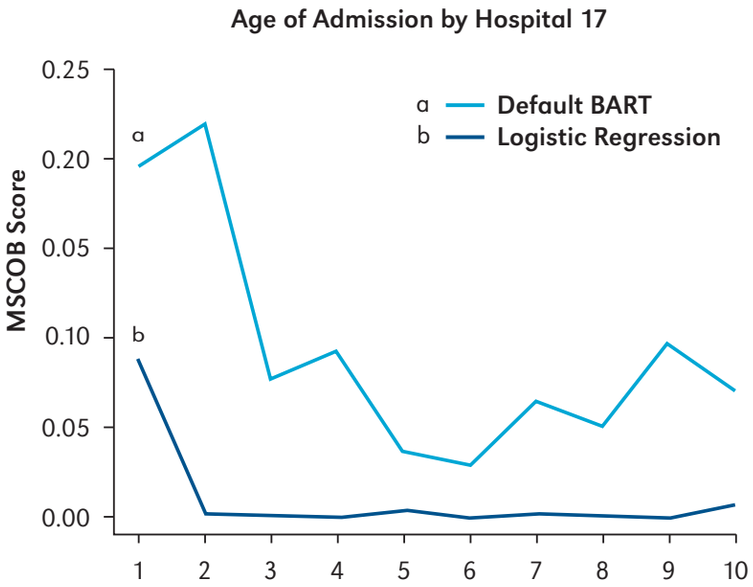


Figure 1: *shows for hospital 17 the Mean Squared Covariate Balance score (MSCOB) for both the Default BART and Logistic regression models across the 10 propensity score bins. Propensity scores are continuous values from 0 – 1; however, to compare between hospitals we computed the mean-squared difference for each decile calling each a bin.*

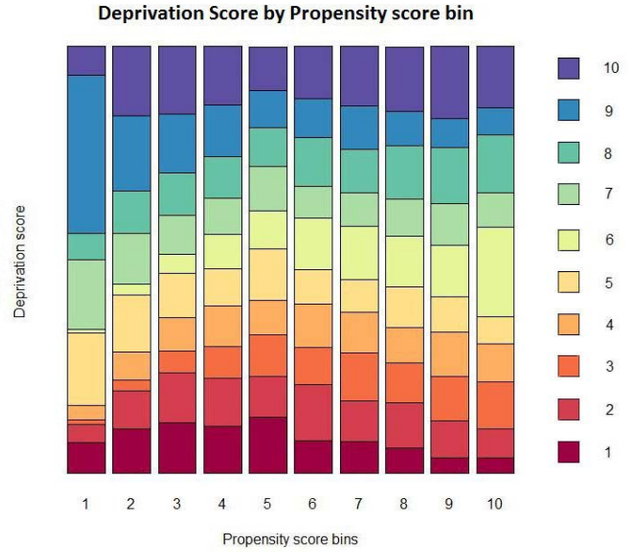


Figure 2: *shows the MSCOB for deprivation score across the deciled bins. Deprivation score is a poverty scale ranging from 1, most impoverished, to 10, least impoverished. A perfect balance would show a matrix of squares. We can see from this graph that the balance is not perfect particularly in the first bin with a deprivation score of 9. This indicates there are a disproportionate number of wealthy people who have a 0 - 10% chance of attending this hospital.*

CASE STUDY 7

Why are some molecules drugs?

Jóhannes Reynisson, School of Chemical Sciences

The fundamental question of why a few molecules are beneficial drugs whereas most molecules are just molecules is still unanswered.

By calculating the properties of known drugs using quantum chemical methods a region of properties can be established and used as a metric for drug design.

For example, by analysing drugs with different oral activity (how well can the drug be administered as a pill) and calculating their polarisabilities and dipole moments it is clear that a preferred region is favoured as shown in Figure 1.

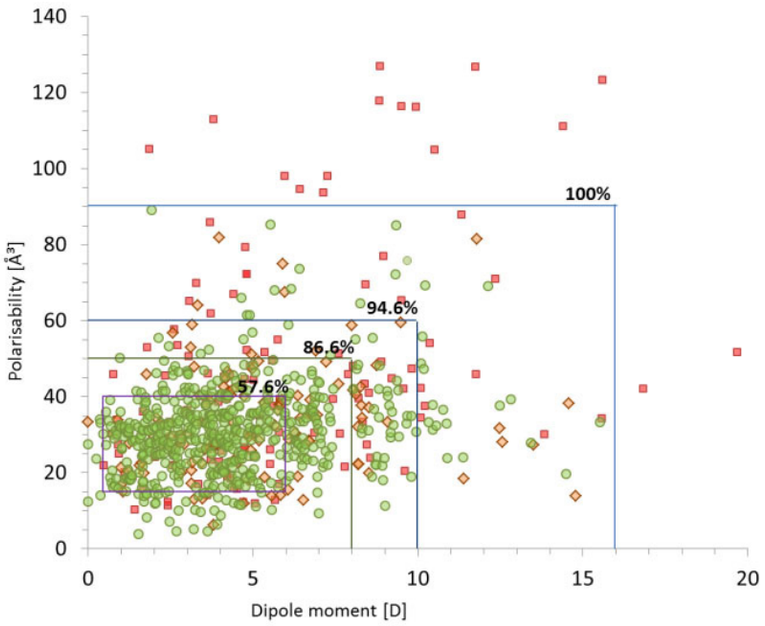


Figure 1: *Dipole moment versus polarisability presenting the distribution of bioavailability groups with boundaries determined for Group. Group 1– low (red squares), group 2 - moderate (orange diamonds), group 3– high oral activity (green circles). In the range of 15-40 Å³ (polarisabilities) and 0.5 – 6 D (dipole moments) ~58% of high oral activity drugs are found meaning that this is a target area for designing new drugs with the ability to cross cell membranes.*

Calculating the physiochemical properties of drug compounds

Having a computational resource such as the Pan cluster allows us to calculate the physicochemical properties of the drug compounds under study on a high level of theory. Gaussian software is used. A large host of drugs/compounds need to be processed and the calculations required are computationally demanding.

This is instrumental in defining the boundaries of Known Drug Space and without the Pan cluster this research is not possible.

The next challenge in our research is to use the polarised continuum model (PCM) to predict the water solubility of drug compounds and establish its boundaries. The PCM method is a sophisticated theoretical approach that requires substantial computational resource to process all known drugs.

Papers published with data generated with NeSI reseources

1. P.A. Hume, M.A. Brimble, J. Reynisson, Aust. J. Chem., 65 (2012) 402-408.
2. P.A. Hume, M.A. Brimble, J. Reynisson, Comp. Theor. Chem., 1005 (2013) 9-15.
3. K.L.M. Drew, J. Reynisson, Eur. J. Med. Chem., 56 (2012) 48-55.
4.B. Yu, J. Reynisson, Eur. J. Med. Chem., 46 (2011) 5833-5837.

CASE STUDY 8

Putting turbulence to work

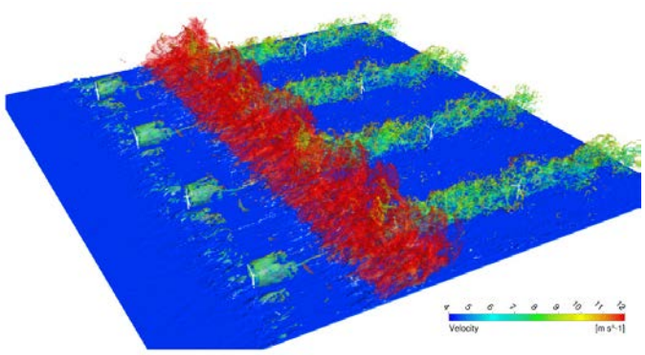
John Cater, Department of Engineering Science

Optimising the new generation of wind turbines.

As the energy sector seeks to provide more of our power consumption needs from renewable resources, there is a constant pressure to extract more generation from existing infrastructure, such as existing wind farms, or to produce more from a particular new site.

Research in the Department of Engineering Science led by Dr. John Cater, working along with the Department of Mechanical Engineering, is focussed on optimising the power generated from the new generation of wind turbines currently being installed around the world. High performance computing is used to solve the non-linear equations of motion for a wide number of wind flows and weather scenarios, and the output data is then used to change the way turbine controllers behave in extreme wind events to decrease turbine damage and to increase power output.

The figure below shows a gust, a region of high speed wind flow (shown in red), encountering an off-shore wind farm. In this simulation, each turbine operates individually and responds to the increased speed by changing its blade rotation speed and pitch angle. Although the gust starts out a large coherent structure (at the left-hand side of the picture), it is broken up by the presence of the front row of turbines and turbulent wakes are generated, which produce a very 'messy' flow at subsequent turbines. These messy wakes are responsible for high fluctuating loads on the turbine shafts. This work is world-leading, and has led to a number of recent high-profile publications. Other projects using the computing facilities of the NeSI Pan cluster include simulating the flow through the human upper airway, where turbulence is important in the mixing of gases, such as CO_2 during the breathing cycle, and in the generation of some parts of speech (called fricatives).



Simulating turbulent flows

The Pan cluster at the University of Auckland is used for a wide range of turbulent flow problems to generate statistically significant data sets to better model momentum and energy transport. These simulations produce enormous file sizes that are too large to be processed on desktop computers (more than 30GB per timestep, and hundreds of timesteps are needed). The large memory and storage capacity of Pan made these simulations possible. The turbulent flow data generated to date has been used to improve the efficiency of offshore wind farms, to create more sophisticated breathing support devices and to improve our understanding of sound production. The computational fluid dynamics (CFD) software called ANSYS with CFX is used for the simulations.

The next projects that will use Pan will be focussed on using turbulent flows to improve the productivity of the primary industries sector in NZ. Dr Cater's group will use the parallel computing facilities on the Pan cluster to model the complex mechanics of the bovine rumen and to study the discharge of effluent in marine aquaculture such as that from mussel farms and fish cages.

CASE STUDY 9

Hemodynamics in the microcirculation

Tet Chuan Lee, David Long, Richard Clarke,
Department of Engineering Science

The microcirculation is the network of arterioles, capillaries and venules that deliver blood to the body's tissues.

Lining the walls of these microvessels is a layer known as the Endothelial Glycocalyx Layer (EGL). The diameter of the vessels we consider (e.g. post-capillary venules) typically lie in the range of about ten microns and the EGL can extend half a micron or more into the blood flow. It plays an important role in the microcirculation as it modifies the velocity profile of blood flow and is believed to alleviate the fluid shear stresses exerted upon the vessel wall. It is also theorised to play a role in regulating the permeability of the vessel as well as being involved in mechanical signalling and inflammatory cell trafficking. Moreover, it is also believed that it plays an important function in disease states such as ischemia-reperfusion injury, diabetes and atherosclerosis.

However, the EGL does not behave the same in the laboratory as in the body. This issue is compounded by the fact that in-vivo measurements of the EGL are extremely challenging. For these reasons, it is hoped that models can help us to better understand the EGL and confirm some of the above-hypothesised behaviours.

Current models of the EGL, however, typically assume an idealised vessel geometry which may not necessarily represent biological reality. In order to model a more realistic microvessel, a Boundary Element computational model was developed that can simulate blood flow through a physiologically realistic microvessel based upon real biological data obtained using confocal microscopy.

Figure 1 shows a computational mesh for a post-capillary venule obtained from the Centre for Microvascular Research at Queen Mary University of London. Triangular elements were used to approximate the geometry with a barycentric coordinate system used on each element. Examples of the fluid shear stresses predicted by simulations for a particular choice of EGL and vessel geometry are shown in Figure 2.

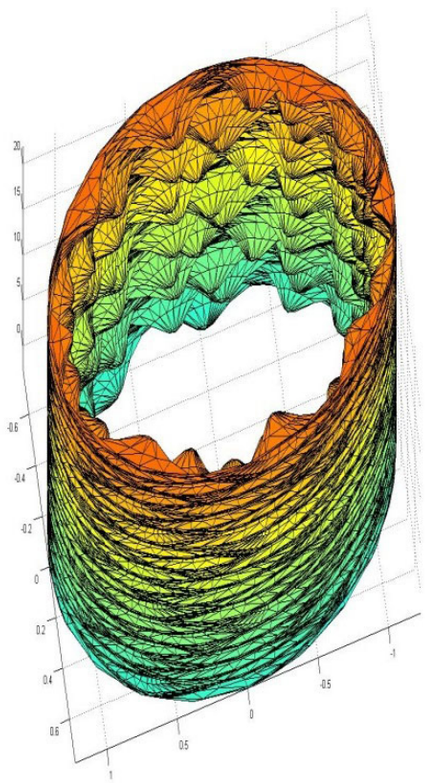


Figure 1: *Computational mesh of a post-capillary venule based on real biological data. The endothelial cells (exaggerated here) can be seen to protrude into the vessel.*

Constructing matrices

The Boundary Element Method, which was written in the C programming language, readily lends itself to parallelisation. Construction of the full matrices required by the Boundary Element Method was multithreaded using OpenMP, which enabled the multi-core environment on the University of Auckland's high performance computing to be exploited. For the calculations shown in Figure 2, the simulations were run on a single node across sixteen cores, although larger computations (i.e. relating to a greater number of partitions, necessary for finer meshes) could easily be run across multiple nodes using MPI. Once built, the matrices were assembled into a linear system which was then solved iteratively using Generalised Minimal Residual (GMRES) again multithreaded across multiple cores using OpenMP. Figure 3 provides details of computational time as a function of the number of cores and memory usage as a function of mesh size.

There are numerous physiological effects which are still to be incorporated, including the elastic behaviour of the EGL, as well as possible charge and osmotic effects, all of which will increase the size of the numerical scheme and the computational burden.

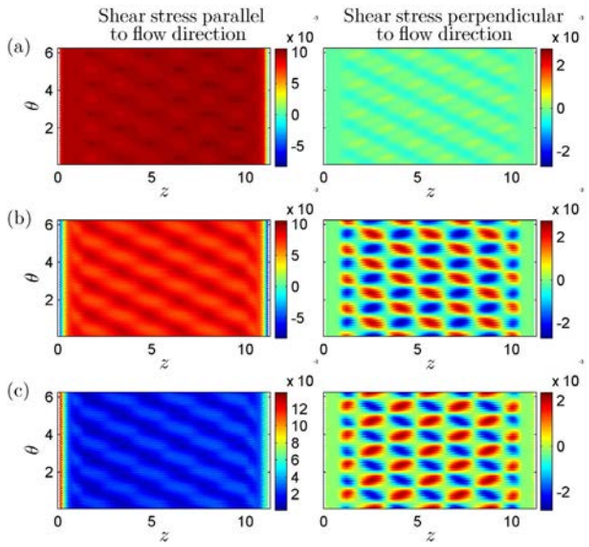


Figure 2: *Examples of Fluid Shear Stresses exerted upon the vessel wall (taken from Lee, T-C. Masters Thesis 2013, University of Auckland)*

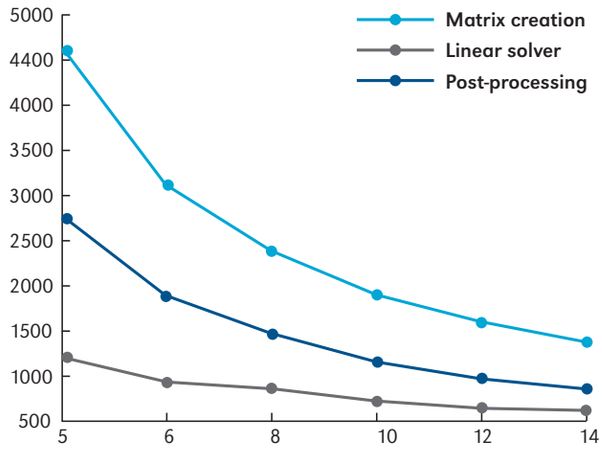


Figure 3: *(Top) Computational Speed as a function of cores (Bottom) Memory usage as a function of mesh size (Taken from Lee, T-C. Masters Thesis 2013, University of Auckland)*

CASE STUDY 10

3D Electromagnetic modeling and simulation using heterogeneous computing

John Rugis, Institute of Earth Science and Engineering

In the earth sciences, a number of different electromagnetic techniques are used to non-invasively acquire information about underground material properties and objects.

One of these techniques, magnetotellurics, takes advantage of the naturally occurring time varying electromagnetic field that surrounds the earth. The lowest frequency components of this field penetrate through the surface of the earth to a depth of up to tens of kilometres. Differing subsurface materials reflect this field and careful measurements of these reflections can be used to create 3D images of the subsurface through a computational process known as inversion.

The goal in this project was to explore the possibility of speeding-up the most time consuming step of the inversion process - the forward modelling. To achieve this goal, we selected one of the highest performance options available in computing today: Intel CPU's and NVIDIA GPU's across multiple nodes.

Simultaneous utilisation of multiple compute nodes requires careful problem decomposition and parallel programming techniques. However, before taking on this complexity, it is generally best to start with a reference model. Results produced by subsequent parallelisation can then be verified against the reference. Figure 1 shows a reference visualisation of simulated electric field in a $100 \times 100 \times 100$ model space after 120 time steps. The results after parallelisation and scaling-up were a perfect match!

The final model size in this project was increased to $848 \times 848 \times 848$. A slice of the electric field after 550 steps is shown in Figure 2 where we see the concentric rings which is exactly what electromagnetic theory says should happen.

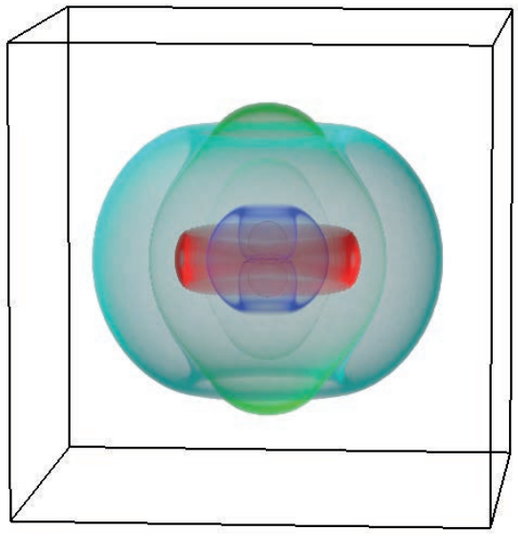


Figure 1: *Electric field in reference model*

Using GPU nodes

Using eight GPU nodes on the Pan cluster at the University of Auckland, we were able to achieve an 80-fold overall speedup in forward modelling run time. With this success, we are on-track to reduce typical 3D magnetotelluric inversion run times from over a week to less than a day. The NVIDIA CUDA software development tools on the NeSI Pan were used extensively for coding, debugging and profiling. OpenMPI libraries were used for inter-node communication.

We could not have even considered this work without access to specific high performance computing facilities such as those available at the University of Auckland. The support and technical assistance from the staff at the Centre for eResearch of the University of Auckland was instrumental in our success.

Parallel computing is central in our research leading to more robust and efficient methods. Our parallel computing success has helped us to meet and exceed client expectations in our commercial work.

We plan to further optimise our techniques and to continue scaling up to use additional GPU nodes as they become available on the NeSI Pan cluster.

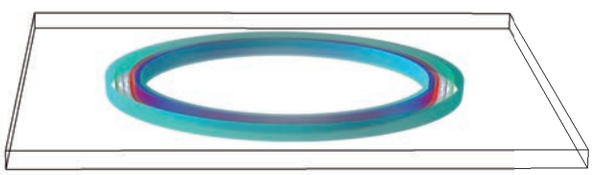


Figure 2: *Electric field slice in final model*

CASE STUDY 11

Revealing key processes in enzyme efficiency through high performance computing

Davide Mercadante, School of Chemical Science

The fascinating ability of enzymes to act processively

Processivity is a word used to define the ability that some enzymes have to catalyze several reactions in a single binding cycle to their substrate. In other words, an enzyme binds its substrate and does not dissociate from it until more than one modification to it has been performed. Usually, processivity is exerted on substrates that have a linear conformation such as polymeric chains of DNA or polysaccharides, which need to be chemically modified in order to prompt a function.

In the plant world, there are many examples of substrates as long polymeric chains: for example the polysaccharides that are embedded in the plant cell wall, which need processive enzymatic activity in order to be dynamically re-arranged. Interestingly, plant pathogen bacteria also express enzymes that potentially have a processive activity on plant cell wall polysaccharides. One of the most interesting class of these enzymes are the bacterial Pectin methylesterases (PME), which catalyze the conversion of pectin linear regions from highly methylesterified to lowly methylesterified chains. By doing this, the chemical and physical properties of the plant cell wall are modified in a way that favors the breach of the plant cell by the pathogen. Through means of Molecular Dynamics simulations carried out using the NeSI Pan high performance computing (HPC) cluster, I investigated the ability of a bacterial PME, expressed by the plant pathogen *Erwinia chrysanthemi*, to slide along pectin chains after each reaction cycle. Molecular Dynamics simulations represent a gold standard in the study of molecular motions through the application of principles that describe molecular mechanics.

The study revealed for the first time the mechanism, at an atomic level, by which the protein accomplishes a processive run along the polysaccharide. It was possible to solve the puzzling problem posed by the conformation of each monosaccharide along the chain. Indeed, the functional groups that are processed by the enzyme are alternately-facing conformations suggesting that a simple sliding of the enzyme along the chain would not be sufficient to assure any processive catalysis (Figure 1).

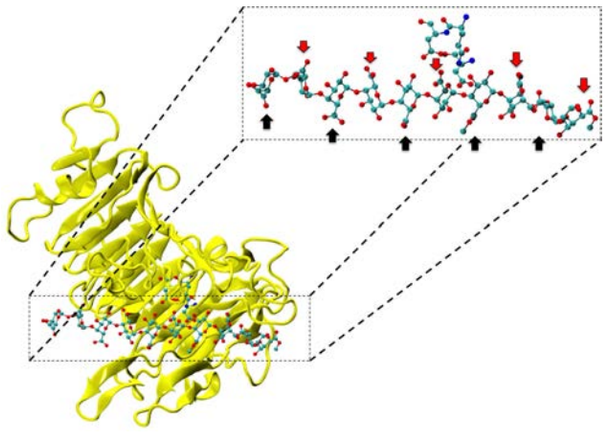


Figure 1: *Complex between Erwinia chrysanthemi pectin methylesterase (shown in yellow) and pectin decasaccharide (shown in ball and stickes). The inset shows the opposite orientations of contiguous monosaccharides along the decamer (arrow).*

Simulations revealed that after the first catalytic event the conformational freedom of the substrate is locally enhanced by the electrostatic repulsion between the enzyme and the freshly generated product of reaction. This repulsion promotes a rotation of the functional groups along the polysaccharides so that a new group, susceptible to catalysis, can be approached by the enzyme (Figure 2). The above described study was useful to sample both the phenomenon of monosaccharide rotations around the glycosidic bond and the subsequent sliding of the enzyme along the substrate.

These findings are remarkable with respect to the mechanism adopted by these enzymes and give insights into a strategy for processive catalysis that does not need the hydrolysis of any externally available high-energy co-factor. Instead it uses the potential already embedded in the product of reaction to yield the energy useful to

catalyze more reactions along the polymer. Moreover, on a point of view more closely related to the investigated system, the discovery of this mechanism provides new evidence of how the plant cell wall is dynamically modified to favour physiological process in plants and how the infection of plant hosts by bacteria is efficiently carried out.

The sampling of substrate conformational variations crucial to bacterial infection of plant hosts will be of great help to develop new compounds able to modify pathogen defense responses in plant cells.

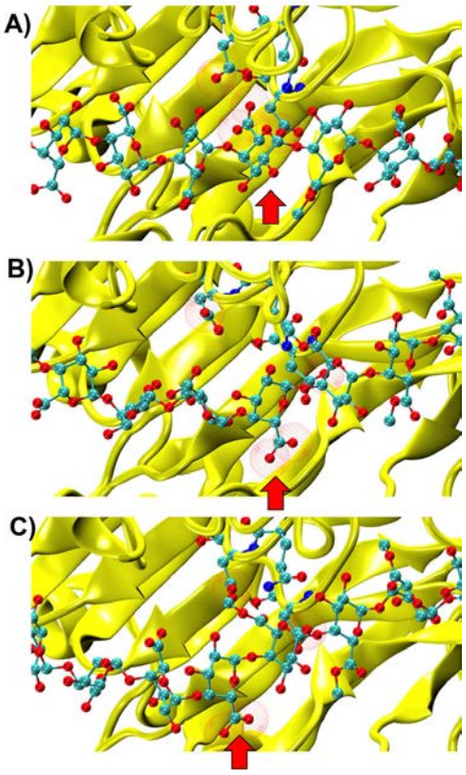


Figure 2: *Conformational variations of the monosaccharide in proximity of the catalytic site of Erwinia chrysanthemi PME. The red spheres surround the atoms involved in the electrostatic repulsion between the monosaccharides and the active site (A). Such repulsion displaces the catalyzed monosaccharide out of the binding site and favours a rotation along the glycosidic bond (B). Thermal fluctuations will accomplish the sliding useful for the docking of another monosaccharide into the active site (C).*

Why using high performance computing?

Molecular Dynamics simulations compute the motions of particles (which can be part of a biological or a non-biological system) by integrating, at each time-step, Newton's equation of motion. In this equation the force acting between two particles is described as proportional to the mass of the particles and their acceleration at each time-step.

$$\vec{F} = m\vec{A}$$

It is interesting to see that this is a very simple equation that, given all the parameters, can be easily solved by any simple calculator. Nevertheless, proteins and whatever surrounds them (water, ions, membranes or some other biological structures) are constituted by hundreds of thousands of particles (sometimes millions) that exert forces over both short (these are Lennard-Jones, hydrogen bonding and hydrophobic forces) and long-range (electrostatic forces). Computing all the forces at each time-step for hundreds of thousands or sometimes millions of particles is not an easy task although the equations involved are fairly simple. Scaling the calculations for some defined subsets of particles across different CPUs of multiple-core computers can considerably speed up the Molecular Dynamics simulations. The simulations were carried out on the Pan cluster using GROMACS 4.5.5 software as the Molecular Dynamics engine.

The high scaling performances of the hardware, in association to an appropriate compilation of the software relative to the cluster architecture, were fundamental in achieving the results as the conformational variations described above do not only belong to large molecules but occur on timescales for which Newton's equation needs to be integrated thousands of times. Thanks to the high computing power provided by the New Zealand e-Science infrastructure (NeSI) simulations were highly parallelised. In addition, with the available general parallelised file-system (GPFS), the collection of data after computation on the required CPU cores was remarkably sped up. These advances have lowered the barrier between the integration of the equations to be solved and the writing of the calculated results on the file-system – previously a bottleneck for large systems.

The results of the study have been published and featured on the cover of Biophysical Journal[1] – a high-impact factor journal in the field of biophysics and computational or experimental structural biology. Such results are also part of my successfully completed PhD in Chemistry at the University of Auckland. Moreover, part of the results obtained from the collected trajectories calculated using the Pan cluster has been submitted and is under revision in the journal PLoS Computational biology[2].

Further simulations will be carried out partially on the Pan cluster for a higher scalability of the software and partially on the high performance supercomputing center located at JULICH in Germany. At the Heidelberg Institute of Theoretical Studies (HITS) in Heidelberg, I am continuing the research that I started during my PhD at the University of Auckland.

[1] *Mercadante, D., Melton, L. D., Jameson, G. B., Williams, M. A. & De Simone, A. Substrate dynamics in enzyme action: rotations of monosaccharide subunits in the binding groove are essential for pectin methylesterase processivity. Biophysical journal 104, 1731-1739, doi:10.1016/j.bpj.2013.02.049 (2013).*

[2] *Mercadante, D., Melton, L. D., Jameson, G. B. & Williams, M. A. K. Processive pectin methylesterases: the role of electrostatic potential, breathing motions and bond cleavage in the rectification of Brownian motions. PLoS Computational Biology Submitted (2013).*

CASE STUDY 12

Finding genetic variants responsible for human disease hiding in the universe of benign variants

Klaus Lehnert and Russell Snell,
School of Biological Sciences

Our programs aim to unravel the genetic basis of human diseases using new approaches enabled by recent step-changes in genetic sequencing technologies (aka the "$1000 genome").

The human genome comprises 3 billion loci and individuals typically differ from this 'reference' at millions of sites. These differences are the result of a complex interplay between ancient mutations, selection for survival fitness, mating between populations, events of near-extinction, and a very strong population expansion in the last 100 years.

A constant supply of new mutations creates new variants that are extremely rare. Some of these variants are directly responsible for disease and others cause genetic diseases in unknown combinations. We combine classical genetics approaches with genome sequencing to identify potentially disease-causing variants for experimental validation. One of our focus areas are neurological diseases, and we have just started a large project to understand the genetic nature of autism-spectrum disorders (www.mindsforminds.org.nz) – a debilitating neurodevelopmental condition with increasing prevalence in all human populations. We expect that identification of genetic mutations will help us to better understand the disease process and identify new targets for therapeutic intervention. This is 'big data' research – we typically process 100 billion 'data points' for each family, and the data analysis and storage requirements have posed significant challenges to traditionally data-poor biomedical analysis.

Analysing human genomes

Using the parallel processing options available on the Pan cluster we were able to derive optimal combinations for multiple interdependent parameters to align 100's of millions of sequence reads to the human genome. These read sequences are strings of 100 nucleotide 'characters' (one of the four DNA 'bases' plus 'not known'), including a confidence score for each base call, and contain a small number of differences to the reference string – some of these are the patient's individual variants, and some are technical errors.

The non-uniform nature of the human genome reference further complicates the alignment problem. The goal is to assign a unique position for each read in the genome using mixed algorithms employing string matching to 'seed' the alignment and a combination of string matching and similarity scoring to extend the alignment through gaps and differences taking into account the confidence score for each base call. Through parameter optimisation and multi-threading we successfully reduced run times from several hours to seven minutes per patient.

The second step of each patient's genome analysis aims to derive 'genotypes' for each of the millions of loci that differ from the reference in each patient. A genotype is a 'best call' for the two characters that can be observed at a single position (humans have two non-identical copies of each gene, one inherited from each parent). Genome sequencing creates 20-500 individual 'observations' of each position and the observations may be inconsistent, and/or may be different to the reference. We obtain a 'consensus call' for each position through a process that first proposes a *de novo* solution for the variant locus (i.e., not influenced by the reference), and then applies complex Bayesian framework to compute the most probable genotype at each locus.

This process requires approximately 2000 hours of computation for a single family. However, on the Pan cluster we can apply a classic scatter-gather approach: we split the genome into dozens of segments, compute genotypes and probabilities for all segments in parallel, then combine the results from the individual computations to generate a list of all variants in each

individual – with confidence scores. Total compute time remains unchanged, but the process completes overnight!

We use standard bioinformatics software programs such as Burrows-Wheeler aligner, the Genome Analysis Toolkit (GATK), samtools, and others; these work very well but require multi-dimensional parameter optimisation. Figure 1 is an example of the alignments we work with.

It shows a software visualisation of the alignment of 360 million 100-nucleotide reads for a parent-child trio against an unrelated genome reference. Our analysis on Pan has clearly identified a new mutation in the child (orange squares, bottom panel), and it appears to affect only one of the child's two gene copies.

Being able to obtain results in hours instead of months will not only permit us to analyse more patients than otherwise possible, it also makes it possible to evaluate the effects of reduced coverage on genotyping confidence. If successful, this will reduce our cost of data generation, which in turn will increase the number of cases we can analyse.

What next?

While our program is under way, we are now starting collaboration with other researchers investigating other genetic diseases around New Zealand. The Pan cluster is the perfect home for this collaboration, and we hope to enrich further optimisation of our approaches. And of course there are disease-causing variants to be discovered!

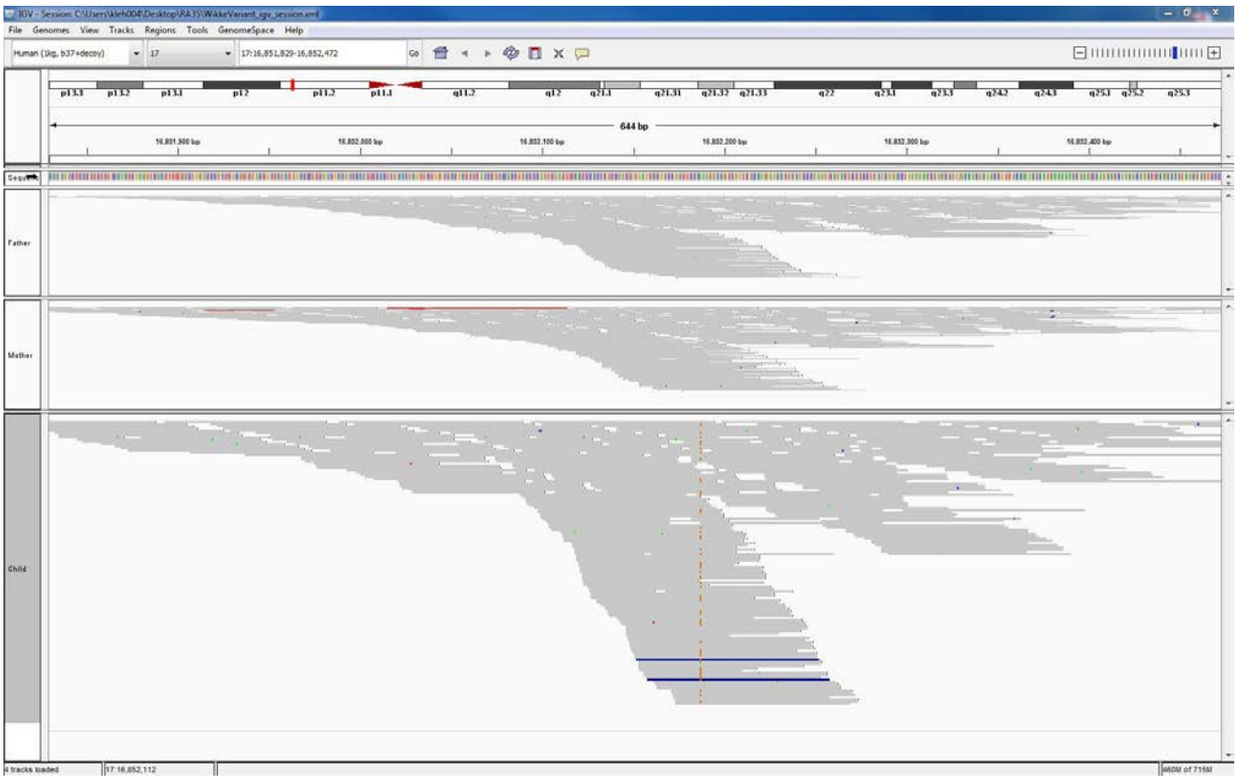


Figure 1: *A visualisation of the alignment of 360 million 100-nucleotide reads for a parent-child trio against an unrelated genome reference. The window is zoomed on 644 of the 3 billion bases in the human genome (horizontal axis), the vertical axis shows features in the reference genome (top), the reference sequence (colour-coded), followed by three panels displaying several hundred reads (grey lines) obtained from the genomic DNA from father, mother, and child. Differences to the reference sequence are indicated in colour, positions identical to the reference are in grey (majority).*

CASE STUDY 13

Improving the short term precipitation forecasts for New Zealand

Sijin Zhang, Geoff Austin, Atmospheric Physics Group

Precipitation affecting New Zealand is mainly initialised in the Tasman Sea and then developed and moved east-wardly with north-westerly winds.

The National Radar Network operated by the New Zealand Meteorological Service Ltd. (MetService) is capable of providing the observations with very high spatial and temporal resolution over the area within approximately 200 kilometres of the coast, which covers most regions of the country except some isolated islands. The use of such high resolution data shows the potential to correct the initial moisture related backgrounds of mesoscale numerical precipitation forecast (NPF) using the Weather Research and Forecasting (WRF) model, which is implemented by MetService operationally, and thus reduces the impact of the "spin-up" problem. While in the region where is not covered by radar, only geostationary satellite (e.g., GOES) is capable of providing relative high resolution data and the availability of such data, which usually can be obtained around 30 min once, is expected to be useful to delineate precipitation information out of the radar range.

In this project, the assimilation of radar and satellite data using WRF are being investigated. At the current stage, a total of 13 heavy rainfall events occurred during the summer of New Zealand (November 2011 to January 2012) were selected to evaluate the assimilation results. Two data assimilation schemes were used: one is the so called Water Vapour Correction (WVC) method, which is set up based on the empirical Z-q_r relationship, and another one is the Reverse Kessler method (RK-nudging) which is set up based on the reverse Kessler warm rain processes and the associated saturation adjustment. Satellite precipitation was retrieved using the RainSat technique. Both methods were used to adjust the model water vapour mixing ratio profile according to radar and satellite observed/retrieved rain rates and the results were analysed and compared. Furthermore, precipitation cloud, which is derived from the infrared difference from two channels of satellite, is also assimilated into model with a method similar to the WVC method. The results show that, the assimilation of radar reflectivity data could significantly improve the precipitation forecasts up to around 9 hours. The addition of satellite retrieved precipitation in the assimilation system could enhance the forecast ability further, especially after the lead time of around 3-4 hours. The RK-nudging method outperforms the WVC method on average. The assimilation of precipitation cloud might improve the precipitation formation and development in some cases, but the improvement resulting from the precipitation cloud assimilation was not as significant as the rain rate assimilation on average, especially for the first several hours. Figure 1 shows a precipitation forecast run using WRF software at the spatial resolution of 3km over New Zealand.

Forecasting weather

Our group started using the NeSI's Pan cluster in 2012. Before that, our research was limited to a local computing cluster run by the physics department which has only 15 nodes and very limited disk space. It was quite difficult to provide accurate weather forecasts since the available computational resources did not allow us to run our model at relative high resolution. Generally, our model runs at a nested domain configuration with the resolutions of 30km (outer domain) and 10km (inner domain), and apparently, such resolution has no capability of capturing most small scale features like tornado. Moreover, it was painful to investigate the effects of either radar or satellite high resolution data assimilation since that even for a single case, the local cluster usually took more than one week to finish. At present, we are able to improve our highest model resolution to 3km, although it is still not perfect, but it already makes our modelling system matching to most operational weather forecast centres around the world. Our data assimilation studies usually can be done within 24 hours for a single case. Furthermore, we can run a number of cases at the same time which really saves us a lot of time compared to run them one by one as we used to do with our old cluster.

Overall, weather forecasting is usually one of the major applications of parallel computer. In New Zealand, we are so glad that we can have the powerful computational resources supported by NeSI. By employing the high performance computing facilities, we can develop and investigate the advanced techniques used in weather forecasts and thus it gives us the opportunity to make more accurate forecast for New Zealand at the next step.

We are planning to start four dimensional variational (4DVar) data assimilation experiment at the next step; similar studies began recently in Japan, UK and American meteorological agencies. The 4DVar technique usually requires massive computational resources so not so many countries can afford that so far. However, we believe that the NeSI's facilities are still capable of supporting us to do some preliminary studies and the results are expected to provide large benefits to the New Zealand's weather forecasts and the associated risks management in the future.

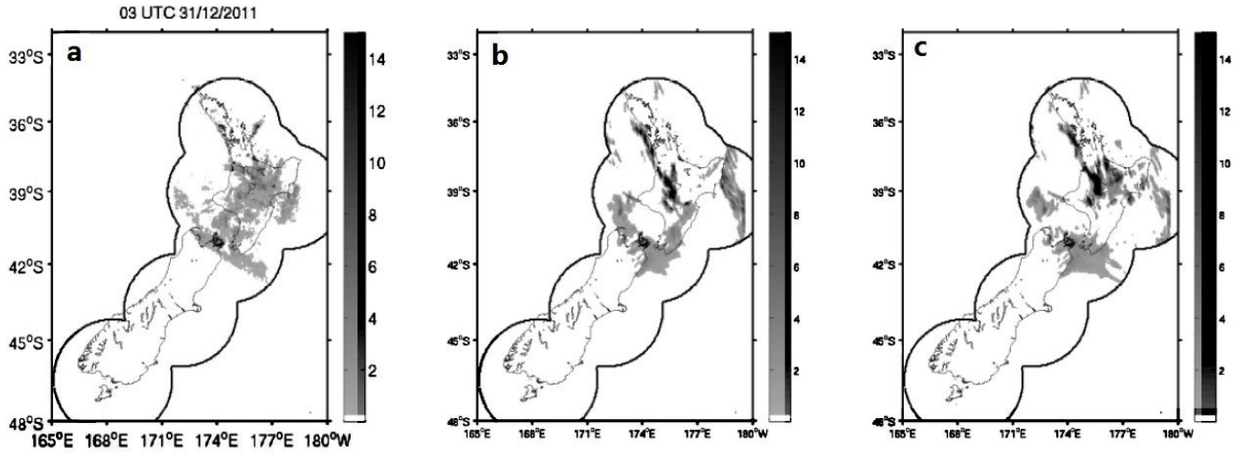


Figure 1: *The precipitation forecasts run by WRF at the spatial resolution of 3 km over New Zealand. From left to right: (a) radar observations, (b)Control experiment (no data assimilation) and (c) radar data assimilation experiment*

CASE STUDY 14

Accelerating the discovery of natural products made by orphan megasynthases

Verne Lee, School of Biological Science

Natural products and orphan megasynthases

Natural products are bioactive molecules produced by living organisms. These molecules play important biological roles in signalling, nutrient acquisition and defence. They are also very important compounds in human and animal health as natural (often foodborne) toxins and particularly as sources of novel pharmaceuticals. Some of the most chemically diverse and useful natural products are made by microbes using a specialised class of enzymes known as "megasynthase" enzymes (see Figure 1).

Large-scale sequencing of microbial genomes has revealed a large number of genes that code for megasynthase enzymes. However, for a big subset of these megasynthases nothing is known about the natural products that they produce. These megasynthase are often referred to as "orphan" megasynthases. The production of natural products by microbes is typically under very tight control and they are only produced in specific conditions. This makes identifying and studying the natural products made by orphan megasynthases very challenging.

Megasynthase enzymes function in a manner analogous to an assembly line, joining together multiple chemical building blocks one after another to produce the finished natural product. Each unique natural product is made by a unique megasynthase enzyme. The chemical building blocks that will be used to assemble the natural product are selected by specific parts of the megasynthase enzymes. These parts of the megasynthases can be referred to as "binding domains" as they select the specific building blocks from the pool available in the cell by specifically binding them in a binding pocket. Figure 2 shows the binding pocket of a binding domain with a building block bound into it.

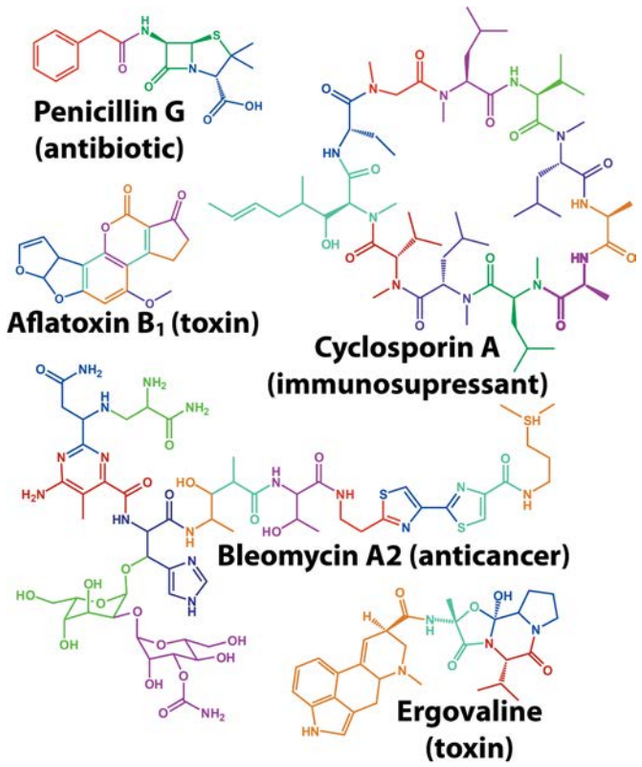


Figure 1: *Megasynthase enzymes are capable of producing a diverse range of natural products. A few examples of natural products produced by megasynthase enzymes are shown. Megasynthases assemble the natural products from multiple chemical building blocks, depicted here in different colours.*

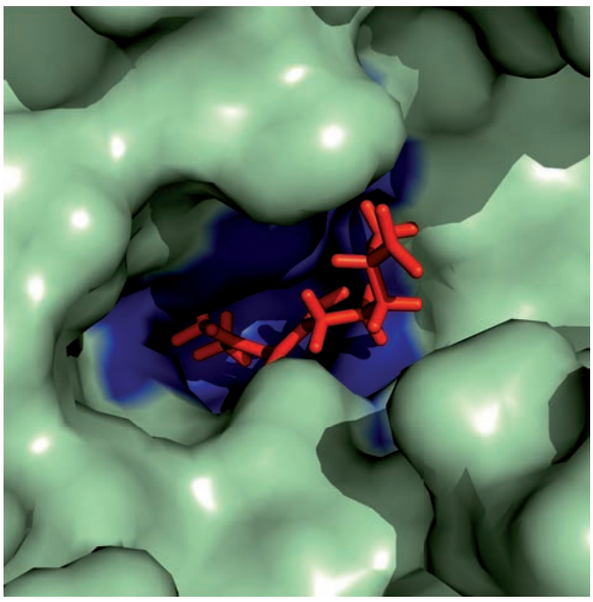


Figure 2 (above): *The chemical building blocks used to assemble the natural products are selected by binding in the binding pockets of the megasynthase binding domains. A chemical building pocket (red) is depicted bound into the binding pocket (blue) of a binding domain (green).*

Predicting the products of orphan megasynthases

If we were able to predict the characteristics of the natural product that a megasynthase produces based on the gene sequence of the orphan megasynthase, this would go a long way towards identifying and investigating these novel natural products. A key part of this effort is predicting which chemical building blocks are used by a megasynthases to assemble its natural product. As the building blocks are selected by the specific binding pocket in the binding domain, this information is available in the gene sequence. However, the predictions cannot be made directly from sequences. Instead, a two-step process is required. The shape and nature of the binding domain and respective binding pocket must be modelled computationally. Then, the process of binding between potential building blocks and the modelled binding pocket must then be simulated to predict which building block is bound tightest by the binding domain.

Simulating binding

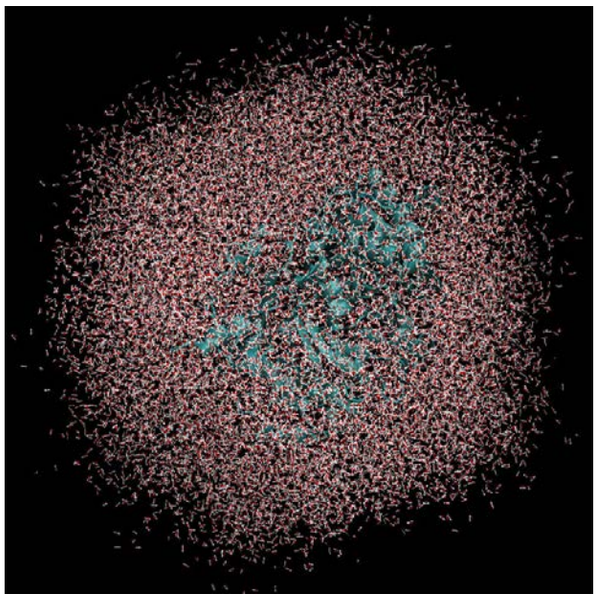
Simulating the binding of the potential building blocks to the modelled binding domains is particularly challenging.

Figure 3 (right): *In this snapshot from a binding simulation, a binding domain (cyan) can be seen surrounded by water molecules (white and red). The chemical building block being tested for binding is present in the simulation but is not visible in this figure.*

Our technique involves simulating the forces involved in the interactions between the atoms of the binding domain, the potential building block molecules and surrounding water molecules using the Amber molecular dynamics package on the Pan cluster. The system simulated can include over 50,000 atoms that are in constant motion (Figure 3). Furthermore, these simulations must be done accurately for a fairly large number of potential building blocks (we have been conducting trials with 160 potential building blocks) within a reasonable time frame. With the binding simulations on each building block able to be run independently, parallel computation on the Pan cluster was ideal for our needs. Together with the parallelisation capabilities of the Amber package, which allowed each simulation to be spread over multiple cores, the run time for a full set of 160 building blocks is on the order of 3-4 days. Without the Pan cluster, our simulations are simply not feasible.

International collaboration – putting it together

We are conducting our research in collaboration with researchers at Aarhus University, Denmark, who are developing computation techniques to predict how a megasynthase assembles the building blocks into the final natural product. The aim is to combine our predictions of the chemical building blocks with their predictions of natural product assembly to result in complete prediction of the natural products made by the orphan megasynthases. This research and collaboration is only possible with the computing resources provided by the NeSI and the excellent support from the staff at the Centre of eResearch.



CASE STUDY 15

The landscape costs of brushtail possum dispersal

Thomas Etherington, School of Environment

The brushtail possum is a notorious invasive species in New Zealand.

Dispersal is a key issue for possum management as it results in the spread of bovine tuberculosis and recolonisation of conservation areas. Landscape features will act to limit dispersal by exerting costs through energetic expenditure, behavioural aversion, and mortality risk. An understanding of how these landscape costs influence dispersal is important for large-scale pest suppression programs as they must take into account the potential for possums to disperse back into an area under management.

The brushtail possum (© Nga Manu images)

Landscape genetics data is being used to identify the costs imposed on possum dispersal by landscape features. The assumption used is that possums that are more closely related are likely to be separated by lower landscape costs as possums must be successfully dispersing between those locations in order to transfer their genetic information. The costs associated with different landscape features is being determined by analysing the landscape genetics data using a geographic information system (GIS) approach called least-cost modelling. Least-cost modelling can find least-cost pathways between pairs of locations that represent the most efficient route that balances the distance travelled and the landscape costs traversed. An appropriate set of landscape costs would produce least-cost paths that would explain a large amount of the variation in genetic distances.

Analysing landscape scenarios

In order to apply this approach a large number of landscape scenarios consisting of different combinations of landscape features and different sets of costs values needs to be analysed. However, to calculate least-cost paths between possum sampling locations for a single landscape scenario takes on average nearly an hour to process on a desktop computer. Therefore, using a desktop computing approach would seriously limit the number of landscape scenarios that could be analysed. With help from staff at the NeSI, the Python code developed to analyse a landscape scenario was modified such that a single shell script could be used to automatically create and submit a large number of landscape scenarios to the Pan cluster for least-cost path analysis. NetworkX, GDAL, SciPy and NumPy software extensions formed a critical part of the workflow. The result of this was that 47,000 different landscape scenarios could usually be processed within 1–2 days depending on the level of activity on the cluster rather than 5 years on a desktop! Essentially, this research could not be conducted to an international standard without the resources and support provided by the NeSI.

Preliminary results suggest that large improvements in understanding the dispersal of possums across landscapes can be achieved by accounting for the costs associated with landscape features (Figures 2a and 2b). The analyses indicate that the main factors controlling possum dispersal are the size of major rivers followed by the absence of tree and scrub landcover.

Once the analyses are completed, GIS maps that represent the landscape in terms of costs to possum dispersal will be produced. These cost maps will then be used as inputs to further GIS analyses that will enable large- scale pest suppression programs to tailor management based on whether parts of a landscape are more or less isolated, or are more or less likely to act as dispersal pathways.

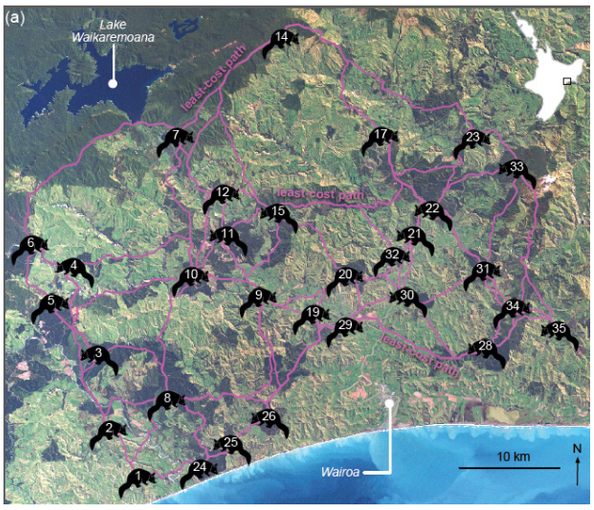


Figure 2a: *Preliminary results from the least-cost modelling have identified pathways that represent the most efficient route across the landscape between neighbouring possum sampling locations.*

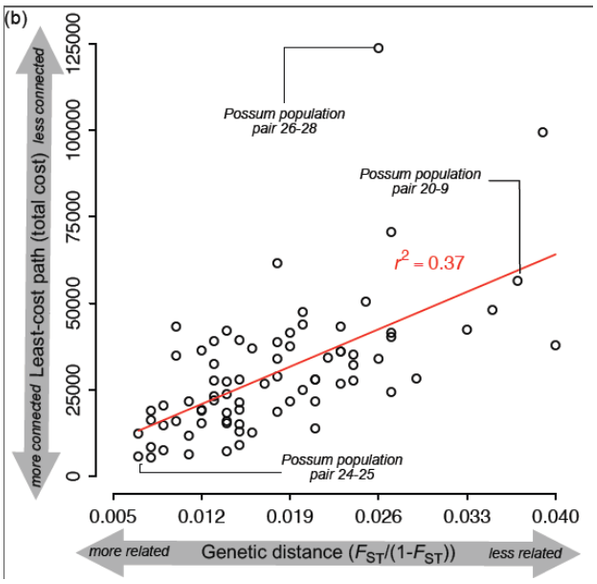


Figure 2b: *A correlation between genetic distance and least-cost path total cost, where each point represents a neighbouring possum population pair. If the landscape does affect dispersal then possum sampling locations that are more connected should be more related, and possum sampling locations that are less connected should be less related. The amount of variation in the genetic distances that can be explained by the least-cost paths suggest that landscape features do affect possum dispersal.*

CASE STUDY 16

Using data mining for digital ink recognition

Rachel Blagojevic and Beryl Plimmer,
Department of Computer Science

Computer-based diagramming is often cumbersome to achieve with typical mouse and keyboard input.

With recent advances in hardware, such as touch and stylus detection, computer-based sketch tools can offer a similar interaction experience to pen and paper. By imitating the pen and paper environment, sketch tools permit the quick construction of diagrams.

Automatic recognition of sketches enables benefits, such as the translation and execution of sketched models, and intelligent editing (see Figure 1). By developing more accurate recognisers, greater functionality can be supplied by computer-based sketch tools.

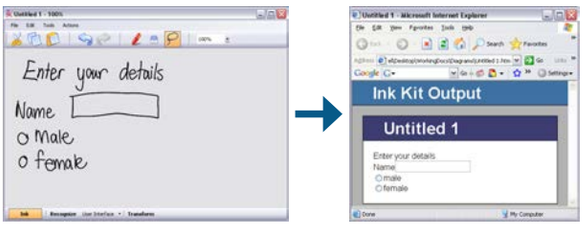
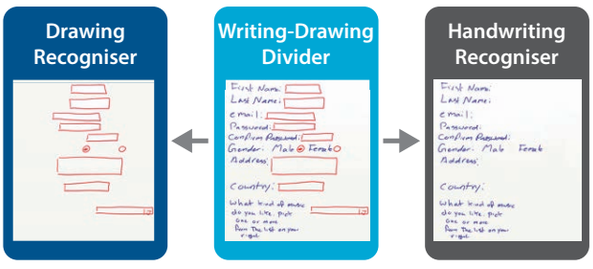


Figure 1: *Benefits of automatic sketch recognition.*

Automatic digital-ink diagram recognition technology is still inadequate for general use. Our research uses data mining techniques to improve the accuracy of recognition. As a challenging example that benefits from this approach, we have focused on the automatic separation of the writing and drawing components of diagrams. People are able to comprehend writing and drawing seamlessly, yet there is a clear semantic divide that suggests, from a computational perspective, that it is sensible to deal with them separately. Feature based recognition is a common approach to writing/drawing division. The choice of distinguishing features and algorithms is critical to its success. We have used data mining techniques to build more accurate writing /drawing dividers.



The Pan cluster was used to perform our analysis of features and algorithms. The computational requirements of this analysis proved to be demanding, due to the complexity of the algorithms and the large number of features and instances included in the training data set. We had previously tried several hardware solutions, including standard desktop machines and several servers. However, when using these resources, many simple experiments took several days, or even weeks, to complete. In order to use the Pan cluster to its full capability, our experiments were distributed across several nodes. For example, for a ten-fold cross-validation trial, each fold was run in parallel by a separate cluster node. Parallelising the folds of each experiment greatly decreased the time required for the analysis, thereby allowing us to run more complex algorithms than before. Overall, excellent results were obtained from the cluster, which could not have been achieved otherwise. It enabled us to use large and varied datasets, ensuring that the levels of accuracy found were reliable and the algorithms were well trained. Our resulting dividers are significantly more accurate than three existing dividers.

Outcome

We have demonstrated that a systematic analysis of data mining techniques can significantly improve the accuracy of writing/drawing dividers. In the future, we plan to investigate how this methodology could be used to improve other parts of the recognition process for digital ink diagrams.

Figure 2 (left): *Sketch recognition process with a writing-drawing divider.*

CASE STUDY 17

The complex unsteady flow within a fluid-filled annulus and its transition to turbulence

Sophie Calabretto, Department of Engineering Science

The spin-up from rest of a fluid enclosed by a container can yield some interesting and beautiful phenomena.

Initially, the container and the fluid are both stationary. A sufficiently long time after the container is set spinning, the fluid rotates with the container as if it were a solid body.

Between these two simple, well-understood, equilibrium states a complicated process occurs involving the transient formation of boundary layers (thin layers of moving fluid) and turbulence (an unsteady, disorganised, three-dimensional state of flow). The aim of this project is to investigate the unsteady dynamics resulting from these rotating flows. The transition between slightly disturbed laminar flow and fully turbulent flow is of particular interest. Turbulent transient flows appear in a broad range of fields, including the study and optimisation of pipe networks, and the analysis of abnormalities in the cardiovascular system.

The flow of fluid in a rotating annulus is used as a paradigm for studying the complex dynamics of these transient flows. As with all real fluid motion, the flow in this problem is governed by the full, three-dimensional, Navier-Stokes equations. Analytic solutions are known only for a small number of particular cases, and so computational schemes are used to calculate approximations to the pressure and velocity fields. In this project, the rotationally symmetric Navier-Stokes equations are numerically solved by exploiting the capabilities of semtex (http://users.monash.edu.au/~bbun/semtex.html) a quadrilateral spectral element direct numerical simulation (DNS) code that is ideal for solving problems in cylindrical coordinate systems.

Solving the rotationally symmetric Navier-Stokes equations

Semtex is a family of spectral element simulation codes, written in C++,developed by Prof. Hugh Blackburn of Monash University, Australia. Semtex uses parametrically mapped quadrilateral elements, the classic Gauss-Lobatto-Legendre nodal shape function basis, and continuous Galerkin projection to solve the conservation of momentum (Navier-Stokes) equations and the conservation of mass (continuity) equation that govern the flow of an incompressible fluid. In order to obtain a solution that is converged both spatially and temporally, a very fine computational mesh and a small time step are used when unning the DNS code. Using multiple threads on a single node, may require 20 GB of RAM and CPU times of up to a week, while producing datasets of up to 80 GB. The next step will be to use the Pan cluster to run fully three-dimensional simulations of the flow (rather than looking at the rotationally symmetric cross-section). Pan will be crucial for this task, as the code will be run in parallel (using MPI to distribute jobs over a number of processors). Distributing this huge problem over many nodes makes the simulation feasible – drastically reducing the computation time required, while still allowing highly-resolved, converged solutions. Without access to the NeSI Pan cluster, these three-dimensional simulations would not be practicable.

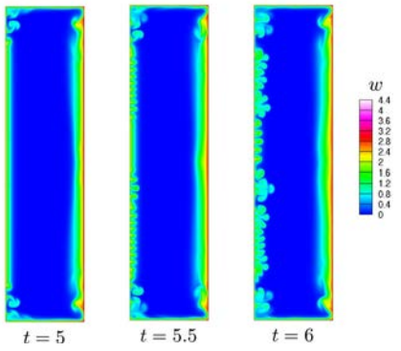


Figure 1 (above): *Snapshots of the azimuthal velocity component (velocity in the direction of rotation), showing the development of an unsteady boundary layer and secondary phenomena.*

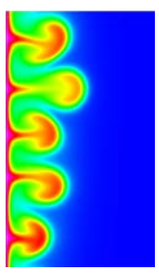


Figure 2 (left): *Close-up of the secondary non-linear Görtler instabilities seen on the inner wall of an annulus with a rectangular aspect ratio.*

CASE STUDY 18

Engine knock in a spark-ignition engine with hydrogen supplementation

Yu Chen, Department of Mechanical Engineering

Engine knock

Spark-ignition (SI) engines play an important role in our lives, particularly in the transportation sector. Although these engines have been studied widely over the past decades, there are still some crucial challenges that we have to face.

The most important challenges are:

- Reserves of fossil fuels are decreasing year by year thus, substitutes for gasoline need to be found.
- Emissions, such as CO_2, are of concern. They have become a significant environmental issue in recent years, and need to be reduced.
- Knock – an abnormal combustion that is caused by autoignition of the unburnt air-fuel mixture. The efficiency of an SI engine is limited by this phenomenon.

H_2 is renewable and carbon free, and has been considered as an alternative fuel for SI engines. The unique properties of H_2 give it a higher knock resistance, and technically, the efficiency of H_2 fuelled SI engines should be improved. However, the low energy density of H_2 suggests that it is more feasible to use it as a supplement, rather than as the sole fuel in an SI engine.

My research is focussed on the knock behaviour of the SI engine with H_2 supplementation. The quantification of knock can be categorised into two aspects: knock tendency and knock intensity. Knock tendency is normally related to the possibility of occurrence of knock under given conditions. Knock intensity is normally considered to deal with the pressure oscillation and its energy content during knocking. While these two aspects can be studied experimentally, the chemical kinetics of gasoline, air, and H_2 reactions under engine operating conditions can be investigated by computer simulation.

Multi-processing on the Pan cluster

The comprehensive model used in this study was developed by a group at Lawrence Livermore National Laboratory. This chemical kinetic model includes n-heptane, iso-octane, and hydrogen mechanisms, with approximately 1500 species and 6000 elementary reactions. The open-source package Cantera is used to solve the kinetic model. One of my goals is to investigate how H_2 affects the knock behaviour of gasoline under engine operating conditions. This is done by analysing the relative sensitivities of ignition delay time (a key factor that influences engine knock) with respect to the chemical rate constants. An example is shown in Figure 1.

Note that a sensitivity of x percent with respect to a particular rate constant indicates that, in response to a 100 percent change of the rate constant, the ignition delay time is estimated to change by x percent. As there are approximately 6000 elementary reactions in the mechanism, the forward and backward reactions of many of these elementary reactions are separated. This means that approximately 9000 simulations have to be done for one sensitivity analysis at a particular engine operating condition (temperature and pressure). Based on a computer with Intel Core i5 3.33 GHz processor, the time consumption for an individual simulation of the sensitivity analysis ranges from a few hours to a few days, depending on the engine operating conditions. This means that completing one sensitivity analysis on one personal computer would require a few years, which is infeasible. However, by utilising the Pan cluster facility at the University of Auckland, hundreds of simulations can be performed at the same time. Now, the time consumption of a sensitivity analysis only requires from a few days to a fortnight, which makes the work feasible.

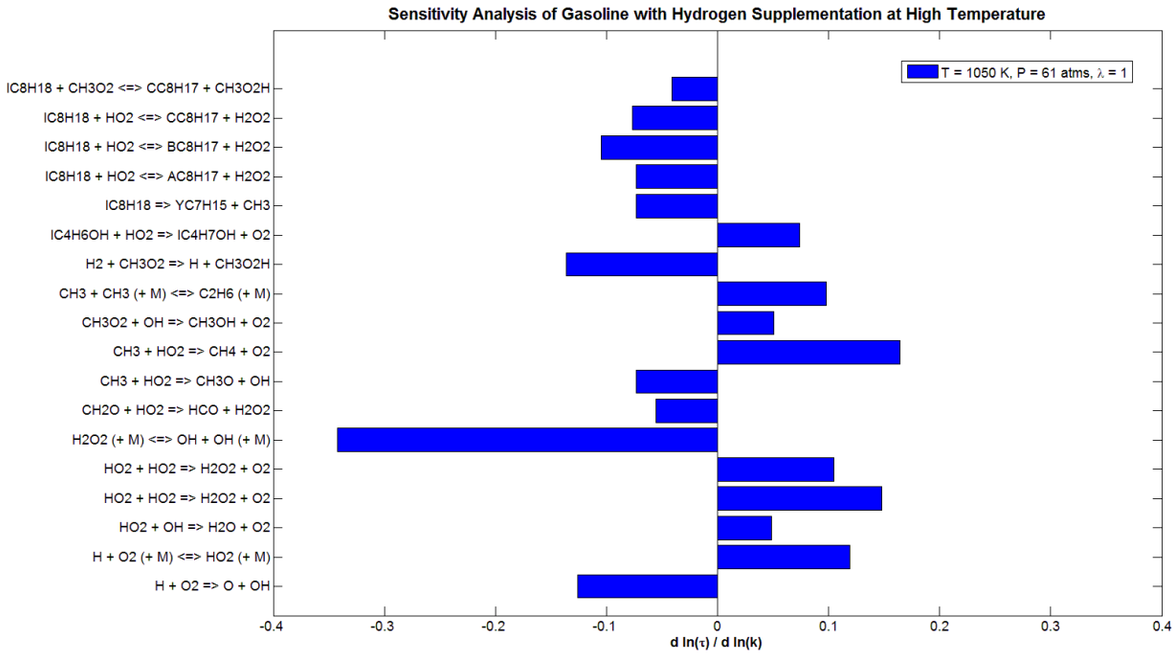


Figure 1: *Simulated sensitivity coefficients of the ignition delay time (T) with respect to the rate constants (k) at the stoichiometric condition. Sensitivity coefficients are shown (as percentages) on the x-axis, and elementary reactions are shown on the y-axis. Only sensitivities greater than 3% are shown. Fuel: gasoline (91 octane rating) with 12.5% of the available energy being due to the H2 supplement. Engine operating conditions are temperature = 1050 K, pressure = 61 atm.*

The next task will be to compute temperature profiles of the unburnt air-fuel mixture in the compression and combustion processes. This will be based on in-cylinder pressure data from my experiments. From there, I will apply the chemical kinetic model under the temperature profile, to determine the mole fraction of species that play a significant role in engine knock, thereby finding kinetically how H_2 supplementation affects the knock behaviour of an SI engine.

CASE STUDY 19

3D Cryo-EM reconstructions of macromolecular complexes

Hariprasad Venugopal,
School of Biological Sciences

To understand biology, it is key to understand the structure of bio-molecules.

Late 20th century has yielded great advancements in techniques like X-ray crystallography and Nuclear Magnetic Resonance (NMR) to delve into the structures of biomolecules to atomic resolution as well as to understand its internal dynamics. Cryo–electron-microscopy (Cryo-EM) is a recent and fast developing tool to directly visualize these structures. The study of large biomolecular assemblies, which are generally intractable by X-ray and NMR has been opened up to near-atomic resolution to atomic resolution with the recent advances in Cryo-EM imaging and image processing. With the procurement of the TF20 transmission electron microscope enabled with energy filtering and tomographic capabilities at the University of Auckland, this kind of work has become feasible for the first time in New Zealand Pan cluster provides the necessary computational platform for processing high resolution images which are computationally very expensive.

Armed to deliver – Insight into the action of a microinjection nanodevice

This is a Marsden funded project for elucidating the functional mechanism of a microinjection device, AFP (anti-feeding Prophage), a complex of 18 different protein, evolved by bacteria *Serratia entomopihila* that delivers a toxin to kill *Cosselyira zealandica*, commonly known as grass grub, a widespread pest of pasture that causes huge economic loss to New Zealand. This work is in collaboration with Dr. Mark Hurst of AgResearch. In a recent paper titled *Three-dimensional Structure of the Toxin-delivery Particle Antifeeding Prophage of Serratia entomophila* (Heymann, J. B., J. D. Bartho, et al. (2013). JBC 288(35): 25276-25284), we have elucidated the structure of the protein complex in resting state to a resolution of ~20Å, which shows the domain level architecture of the nano-device as well as its toxin cargo encased in the central sheath. With high resolution data from the new microscope we will be pushing towards capturing the particle in its various conformational snapshots leading all the way to the contracted state at atomic resolution. This detailed understanding of such a mobile protein delivery system and its unusual eukaryotic target specificity could guide rational designing of nanoscale devices and hence opening cutting edge therapeutic avenues such as delivering antigenic proteins in anti-tumour immunotherapy.

Why Pan cluster?

3D reconstructions from 2D EM images of biomolecules are performed from thousands of images of the molecule. This kind of image processing is computationally very expensive. Modern Cryo-EM image processing software like BSOFT and EMAN (1, 2.1) acknowledges this issue and is written to incorporate scalability in terms of multi thread processing abilities. Pan cluster offers an ideal platform to make use of the parallelisation in these computational routines to fasten the calculation. A single particle analysis routine consists of aligning the particles picked from the micrographs to a reference model. This reference is refined in an iterative way to reach final 3D reconstruction. For example in BSOFT, such an alignment of Afp particle defined in a 400*400 pixel image at 42k magnification scanned at 3.02 Å/pixel raster-step, takes 180 seconds per particle on our workstation (Intel® Core™ i7 CPU 960 @ 3.20GHz processor with 8 threads and 8GB RAM). This single particle orientation and alignment searches are linearly scaled, i.e. 1 processor per particle. When the datasets consists of 10^5-10^6 particles (critical for atomic resolution studies) the computational time required on a standard PC will be in the order of months.

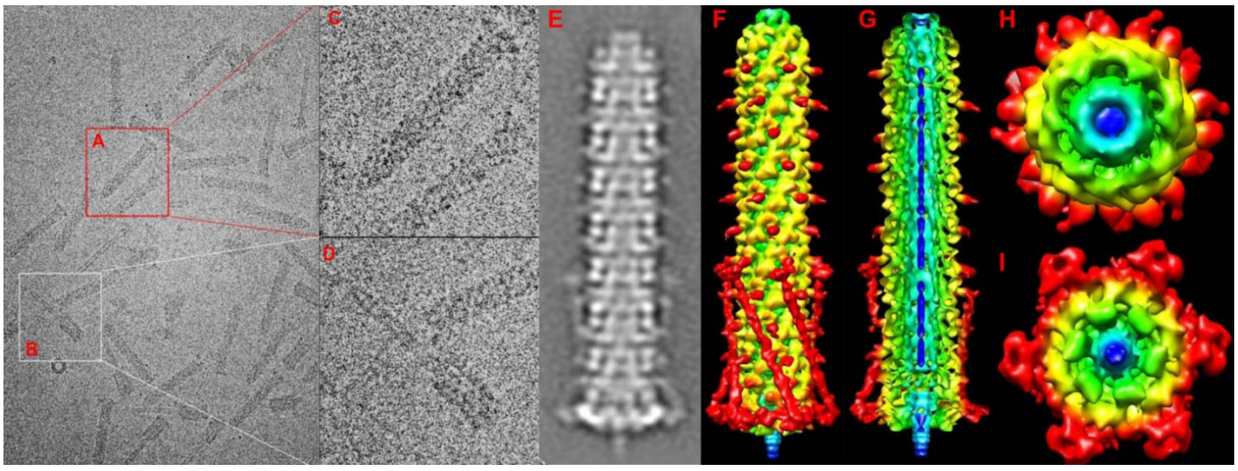


Figure 1: *A: 400*400 pixel boxed AFP particle B: Boxed AFP contracted form. C: Enlarged view of Afp Particle D: enlarged view of the contracted particle. E: Stacked image of Afp particle of similar view. F: Radially coloured 3D reconstruction of Afp particle from blue (center) to red (periphery). G: Vertical cross section of the Afp particle showing the central cyan tube containing the putative toxin colured blue. H: Shows the top view of the Afp particle with sheath proten exhibiting (orange coloured) helical rotation. I: Bottom view of the base plate with the blue puncturing tip.*

Pan offers better hardware and large number of multi-core-processors which are set up for parallel computation. Present calculations use 10 processors on a single node but for higher number of particles as mentioned earlier this can be scaled to use multiple nodes. Such an infrastructure enables us to run multiple jobs with large number of processors in order to perform various computational tests in a timely manner hence making this endeavour feasible.

CASE STUDY 20

Mathematically modelling gastrointestinal electrical activity

Shameer Sathar, Jerry Gao, Mark Trew, Leo Cheng, Auckland Bioengineering Institute

The gastrointestinal system

The gastrointestinal (GI) system comprises a series of discrete organs, from mouth to anus, with distinct functions - digestion, absorption of nutrients, excretion, and protection from digestive agents and pathogens. GI motility, the regulated contraction of muscles along the GI tract, is coordinated by the underlying omnipresent electrical activity termed as slow waves. These are initiated and propagated through networks of the interstitial cells of Cajal (ICC) to the smooth muscle cells (SMC).

Gastric dysrhythmias, or abnormal electrical activity in the stomach, have been implicated in the pathophysiology of several motility disorders, such as gastroparesis. A validated mathematical model offers a virtual medium in which a variety of hypotheses can be comprehensively investigated. The effects of different treatment strategies can also be predicted, potentially more cost effectively than by other experimental approaches. An exciting application, is using models to relate body surface electric field, or magnetic field, potentials to the underlying gastric slow wave activity, as shown in Figure 1.

Modelling on the Pan cluster

Multiscale models of gastric electrophysiology (encompassing the subcellular, cell, tissue, organ, and whole-body scales) can be mathematically described by continuum modelling frameworks such as the monodomain or bidomain equations. Spatial discretisation of the 3D geometrical domain, using a finite element method, leads to a system of differential-algebraic equations.

A semi-implicit method is used for temporal discretisation, where the linear term is treated implicitly and the nonlinear terms explicitly. The numerical solution of these equations on tissue-specific geometric models is typically computationally expensive. The CHASTE computational package, originally developed by the Computational Biology Group at the University of Oxford for solving tissue electrophysiology problems with cardiac applications, is being used for this research. It is built upon an MPI-based library, enabling it to efficiently split the high computational demands of simulation over many parallel processes. The CHASTE code scales linearly across multiple cores of the NeSI Pan cluster.

Recently, we developed a mathematical model for quantifying ICC network-function relationships by embedding biophysically-based ICC cell models into tissue-specific ICC network imaging data. Figure 2 shows the activation-time map of slow waves over an ICC network. This particular problem was relatively small, with about 300,000 solution points and a simulation time of 1000 ms. However, multiple sets of ICC network samples were analysed in parallel, as multiple independent jobs, each using up to 32 cores. The entire set of ICC networks were fully analysed in less than 48 hours using the cluster. If the NeSI Pan cluster had not been available, this data may have required weeks for analysis.

In the future, we plan to simulate gastric electrophysiology on a 3D computational domain (approximately 3 million computational nodes) over temporal durations of up to several minutes. This model will incorporate recently developed, biophysically based, ICC and SMC cell models, along with analysed ICC network function data. We aim to simultaneously utilise up to 1024 cores of the NeSI Pan cluster and concurrently employ the cluster GPU resources. Figure 3 shows preliminary results for slow wave simulations in the stomach.

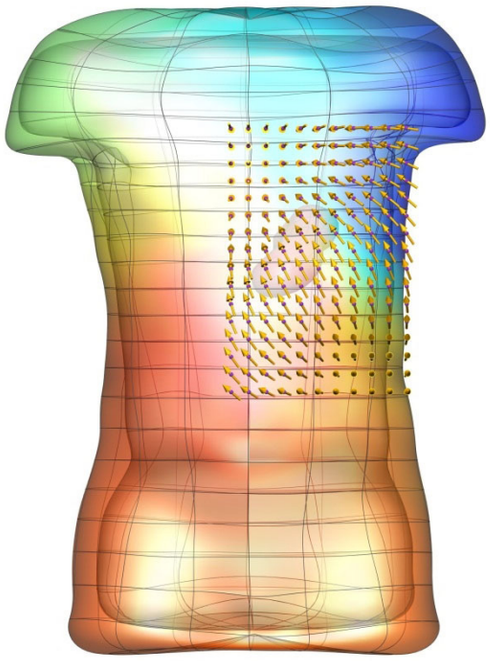


Figure 1: *Torso model with electric potential fields (coloured surface) and magnetic fields (gold arrows) resulting from simulated gastric slow wave activity. Work conducted in conjunction with Prof. Alan Bradshaw (Vanderbilt University, Nashville, TN).*

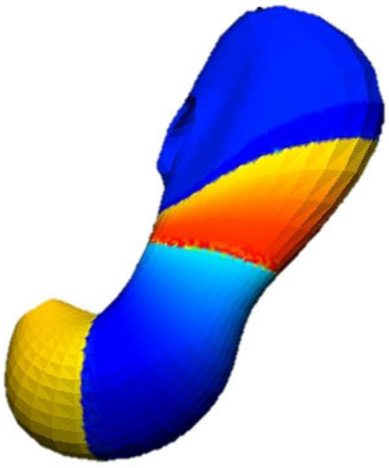


Figure 3: *Simulated slow wave activity on an anatomically realistic stomach geometry.*

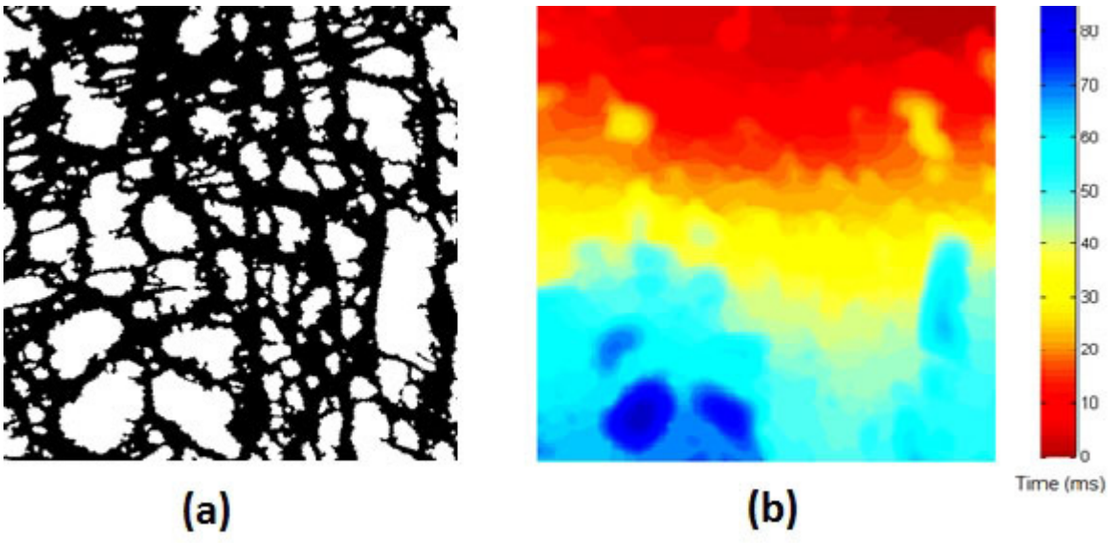


Figure 2: *Slow wave simulation over an ICC network. (a) Tissue-specific ICC network imaging data. (b) Activation time map of slow waves over the network.*

CASE STUDY 21

1-D numerical models of post-glacial river evolution

Jon Tunnicliffe, School of Environment

Simulating river evolution

Since the close of the last glacial maximum, many mountain valley river systems with large glacial fills have undergone complex rearrangement of channel morphology, sedimentology, and long profile due to changing sediment supply conditions and fall in base-level. Our research has focused on the question of whether we can apply model representations of sediment transport processes - derived at the annual to decadal scale – to large valley systems at the millennial scale. It would appear that many aspects of the process of erosion of these valleys fills, and the enduring influence of this glacial legacy on the modern river, can be captured quite well (see Figure 1).

In this research, we have been using cross-section-based ('1-D') representations of the channel-floodplain system and substrate composition (grain-size distribution) coupled with an implicit finite-difference-based hydraulics code to simulate the evolution of a river's long profile over time. The chronology of river down-cutting has been constrained using evidence from radiocarbon and infrared-stimulated luminescence dating. As the valley fill is evacuated, changes in valley slope, and an evolving lag of coarse, bouldery material, result in a river of much different character. By systematically changing the governing parameters of this river model, we can determine which factors most strongly influence the long-term evolution of the fluvial system (see Figure 2).

The need for Pan

Establishing model uncertainties, and the relative importance of each transport model parameter, is a highly complex and non-linear problem. By running a considerable number of parallel instances of the model, we are better able to assess the many contingencies and interactions amongst all of the model variables. For instance, a heightened proportion of boulders in the original valley fill results in a much deeper concavity of the river profile (thus an improved match with the modern river), while higher assumed rates of gravel abrasion will act to suppress concavity but promote more prominent downstream diminution in median grain size. We have also gained insights into how multiple tests of fitness for model results (e.g., proportion of sand in the river bed or the rate of grain size diminution downstream) cannot all be satisfied at once, highlighting some important areas for future model refinement. Typical model run times are more than one day; having access to a large number of cores makes it feasible to achieve what is an otherwise impossible task. We are able to comprehensively assess the dynamic behaviour of mophodynamic models in ways that we previously could not (see Figure 3).

Future work

We are in the process of completing our model runs; our intention is to further explore how climate and vegetation changes throughout the Holocene may have altered the pace of morphologic development of the river system. Results from parameter sensitivity work will feed back into further model refinement. The power of the grid could also be used to explore a number of random forcings on the valley-floor system, such as mass movements and debris flows, which have influenced the river system over time. More powerful 2-dimensional models could also be put to the task of exploring the long-term evolution of post-glacial mountain river systems.

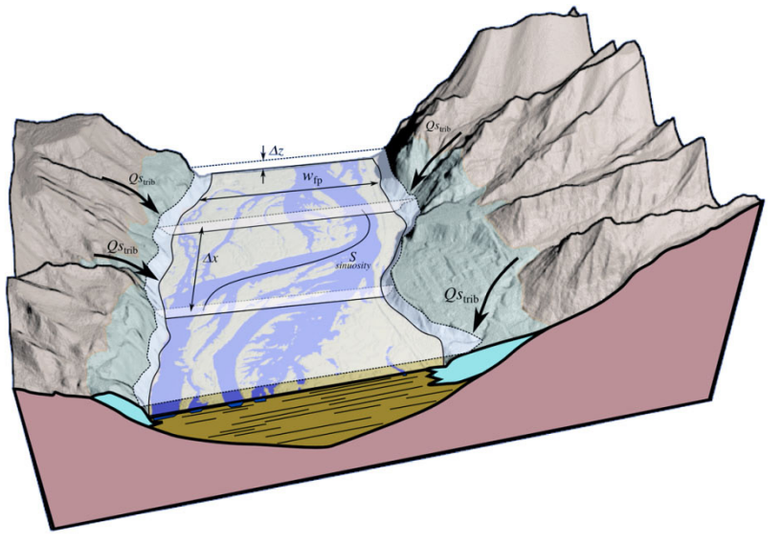


Figure 1: *Model discretization of the post-glacial valley floor. We employ a 1-dimensional finite-volume framework that encompasses the river-floodplain system and contributions from tributary sources, to simulate down-cutting and evacuation of Pleistocene glacial sediments.*

Figure 2: *A river inset within deep (>100m) relict glacial fill. As the river eroded through these layers, it recruited material from the substrate, progressively concentrating the proportion of coarser, bouldery material.*

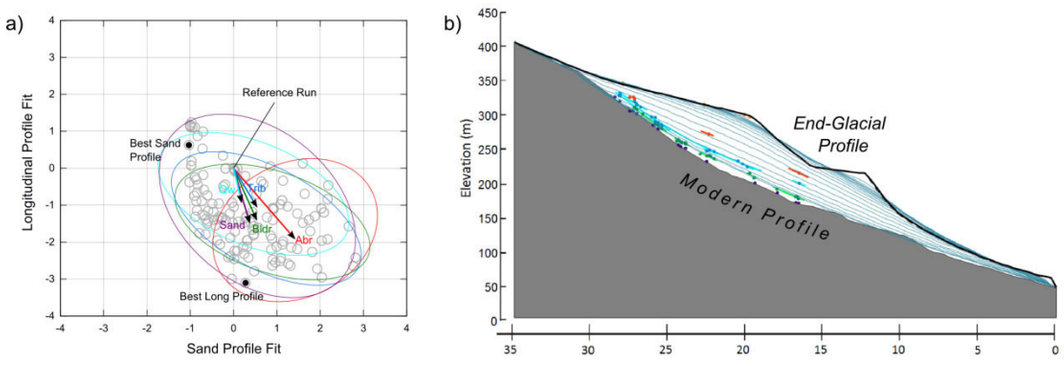


Figure 3: *(a) Model results: multiple model runs provide an indication of the relative influence of governing variables. (b) The trajectory of model river downcutting is compared with measurements and dating of various terrace levels in the modern system.*

About the authors

Arne L. Grimsmo, PhD candicate
Department of Physics
Now at NTNU, Norway
Email: arne.grimsmo@ntnu.no

Assoc Prof Philip Yock
Department of Physics
Email: p.yock@auckland.ac.nz

Dr Grigor Aslanyan, Research Fellow
Department of Physics
Email: g.aslanyan@auckland.ac.nz

Louis Ranjard, PhD candidate
Bioinformatics Institute
Email: l.ranjard@auckland.ac.nz

Dr. David Welch, Senior Lecturer
Computer Science
Email: David.welch@auckland.ac.nz

Marie Paturel, MSc Internship France
Department of Statistics
Email: paturel.marie@gmail.com

Dr Stéphane Guindon, Senior Lecturer
Department of Statistics
Email: s.guindon@aucklandl.ac.nz

Justin Pogacnik, Research Fellow
Department of Engineering Science
Email: j.pogacnik@auckland.ac.nz

Sam Passmore, Masters student
Department of Statistics
Email: sam.passmore@auckland.ac.nz

Dr Jóhannes Reynisson, Senior Lecturer
School of Chemical Sciences
Email: j.reynisson@auckland.ac.nz

Dr John Cater
Department of Engineering Science
Email: j.cater@auckland.ac.nz

Dr. David Long, Principal Investigator and Lecturer
Department of Engineering Science
Email: d.long@auckland.ac.nz

Tet Chuan Lee, Masters student
Auckland Bioengineering Insititute
Email: tlee114@aucklanduni.ac.nz

Dr Richard Clarke, Senior Lecturer
Department of Engineering Science
Email: rj.clarke@auckland.ac.nz

John Rugis
Institute of Earth Science and Engineering
Auckland UniServices Ltd.
Email: j.rugis@auckland.ac.nz

Dr Davide Mercadante, Postdoctoral Research Fellow
School of Chemical Sciences
Now at HITS, Germany
Email: davide.mercadante@h-its.org

Assoc Prof Klaus Lehnert
School of Biological Sciences
Email: k.lehnert@auckland.ac.nz

Prof Russell Snell
School of Biological Sciences
Email: r.snell@auckland.ac.nz

Sijin Zhang, PhD candidate
Atmospheric Physics Group
Email: sijin.zhang@auckland.ac.nz

Prof Geoff Austin
Department of Physics
Email: g.austin@auckland.ac.nz

Dr Verne Lee, Research Fellow
Biological Science
Email: t.lee@auckland.ac.nz

Thomas Etherington, Research Fellow
School of the Environment
Email: thomas.etherington@auckland.ac.nz

Dr Rachel Blagojevic
Department of Computer Science
Email: rachel.blagojevic@auckland.ac.nz

Assoc Prof Beryl Plimmer
Department of Computer Science
Email: b.plimmer@auckland.ac.nz

Sophie Calabretto, PhD candidate
Department of Engineering Science
Email: s.calabretto@auckland.ac.nz

Yu Chen, PhD candidate
Mechanical Engineering
Email: yche441@aucklanduni.ac.nz

Hariprasad Venugopal
School of Biological Science
Email: h.venugopal@auckland.ac.nz

Shameer Sathar, PhD candidate
Auckland Bioengineering Institute
Email: ssat335@aucklanduni.ac.nz

Jerry Gao, PhD candidate
Auckland Bioengineering Institute
Email: jgao032@aucklanduni.ac.nz

Dr Mark L. Trew, Senior Research Fellow
Auckland Bioengineering Institute
Email: m.trew@auckland.ac.nz

Dr Leo Cheng, Senior Research Fellow
Auckland Bioengineering Institute
Email: l.cheng@auckland.ac.nz

Dr Jon Tunnicliffe, Lecturer
School of Environment
Email: j.tunnicliffe@auckland.ac.nz

Computational and technical support team, Centre for eResearch and the NeSI Pan cluster

Aaron Hicks
Software and Systems Engineer, NeSI
Email: HicksA@landcareresearch.co.nz

Bart Verleye
eResearch Support, Faculty of Engineering
Email: b.verleye@auckland.ac.nz

Ben Roberts
HPC Application Scientist, NeSI
Email: b.roberts@auckland.ac.nz

Daniela Dunn
Tuakiri Representative, NeSI
Email: d.dunn@auckland.ac.nz

Eva Choi
Web Developer, NeSI
Email: eva.choi@auckland.ac.nz

Gene Soudlenkov
Technical Lead, HPC Services, NeSI
Email: g.soudlenkov@auckland.zc.nz

Grant Wooding
Technical Writer, NeSI
Email:grant.wooding@auckland.ac.nz

Greg Hall
System Engineer, NeSI
Email: g.hall@auckland.ac.nz

Jaison Mulerikkal
HPC Application Scientist, NeSI
Email: j.mulerikkal@auckland.ac.nz

John Rugis
HPC Application Scientist, NeSI
Email: j.rugis@auckland.ac.nz

Jordi Blasco
HPC Application Scientist, NeSI
Email: j.blasco@auckland.ac.nz

Markus Binsteiner
Software Developer, NeSI
Email: m.binsteiner@auckland.ac.nz

Martin Feller
eResearch Support, Faculty of Science
Email: m.feller@auckland.ac.nz

Matthew Healey
Project Coordinator, NeSI
Email: matthew.healey@otago.ac.nz

Mike Lee
eResearch Support, Faculty of Science
Email: mike.lee@auckland.ac.nz

Nick Young
Web Developer, NeSI
Email: nick.young@auckland.ac.nz

Peter Higbee
Software and Systems Engineer, NeSI
Email: peter.higbee@otago.ac.nz

Peter Maxwell
HPC Application Scientist, NeSI
Email: peter.maxwell@nesi.org.nz

Robert Burrowes
Technical Lead, Research Data Services
Faculty of Science
Email: r.burrowes@auckland.ac.nz

Sina Masoud-Ansari
eResearch Support, Faculty of Science
Email: s.ansari@auckland.ac.nz

Yuriy Halytskyy
Software and Systems Engineer, NeSI
Email: y.halytskyy@auckland.ac.nz

THE UNIVERSITY OF AUCKLAND
NEW ZEALAND
Te Whare Wānanga o Tāmaki Makaurau

Centre for eResearch

Postal address

Centre for eResearch

Computer Science

University of Auckland

Private Bag 92019

Auckland 1142, New Zealand

Physical address

Centre for eResearch

Room G021, LG

Building 409

24 Symonds Street

Auckland 1010, New Zealand

Mark Gahegan, Director

Phone: +64 9 373 7599 ext 81891 | DDI: +64 9 923 1891

Email: m.gahegan@auckland.ac.nz

www.eresearch.auckland.ac.nz