A grid-based facility for large-scale cross-correlation of continuous seismic data

John Townend[1], Yannik Behr[1], Kevin Buckley[2], Martha Savage[1], and John Hine[2]
[1]School of Geography, Environment, and Earth Sciences, Victoria University of Wellington
[2]School of Engineering and Computer Science, Victoria University of Wellington

Report completed for the Earthquake Commission
(Project EQC08/556)

30 June 2009

Executive summary

The vast volumes (more than 3.5 gigabytes per day) of seismic data being recorded by the *GeoNet* network of geophysical instruments provide exciting opportunities for studying the New Zealand plate boundary and better characterising earthquake and volcanic hazard sources. To take full advantage of these data and recent developments in seismological data analysis requires efficient computer algorithms capable of automatically extracting and processing long streams of data.

Geophysical research, as is the case in many other data-intensive fields of science, often involves a large number of incremental processing steps, during each of which various decisions must be made regarding the particular choice of parameters or even algorithms to use. One way of describing a sequence of processing steps is to use a "computational workflow", a documented chain of modular components representing different stages in the overall analysis. Such a workflow enables researchers to examine the effects of different processing sequences by re-running an analysis using modified input parameters or by replacing portions of the workflow with alternative analytical components.

In this project, we have developed a computational workflow describing the analysis of seismic records obtained from the *GeoNet* seismic data archive or other repositories. We focus on a process known as "cross-correlation", in which two signals — such as seismograms — are compared mathematically to determine their similarity and make accurate timing measurements. Cross-correlation underpins much of modern seismology, including earthquake detection and location, and analysis of the seismic noise field, which our workflow is designed to address.

More than 95% of the time, seismometers designed to record earthquakes are actually recording continuous, low-pitched noise—the incoherent background hum of the Earth. By comparing long records of seismic noise recorded at two different locations using cross-correlation techniques, a small amount of coherent seismic energy propagating directly between the two sites can be detected. This energy propagates as a seismic wave at a speed governed by the physical properties of the rocks it passes through. By measuring this speed, we can map the Earth's deep structure in much the same way as ultrasound is used to look inside human bodies.

The computational workflow we have developed has been configured to run on a grid-computing system consisting of 230 networked desktop computers that can be harnessed for computationally intensive computing when otherwise idle. Distributing a computation among many separate processors within the grid reduces the time required to analyse large volumes of seismic data for noise imaging purposes.

The key outcomes of this project are as follows:

1. A two-stage computational workflow enabling continuous seismic waveform data extracted from the *GeoNet* archive, or elsewhere, to be systematically pre-processed and analysed;
2. Deployment of this workflow on a grid computer in a way that makes use of the maximum amount of the computational power available at any one time;
3. Development of a web interface that facilitates running the workflow;
4. Successful application of the workflow to datasets of various scales.

Technical abstract

As in many other data-intensive fields of science, geophysical research often involves a large number of incremental processing steps, during each of which decisions must be made regarding the particular choice of parameters or even algorithms to use. The advent of the world-class *GeoNet* network and data archive and the high-speed Kiwi Advanced Research and Education Network (KAREN) linking New Zealand universities and research institutions provides a key opportunity to analyse an increasingly routine geophysical task, the cross-correlation of continuous seismic waveforms, from the viewpoint of computational efficiency.

In this project, we have developed a computational workflow describing the cross-correlation of seismic waveform data using grid-computing resources. Such a workflow enables researchers to test different processing choices, or to replace one or more steps in the overall processing stream with an alternative algorithm. Cross-correlation underpins much of contemporary seismology, particularly in the areas of differential earthquake location, earthquake and tremor detection, and ambient noise tomography. We focus here on ambient noise correlation, a field of active research in New Zealand and overseas, which typically involves the analysis of many months of data recorded at dozens of seismographs and hence represents a large and parallelisable computational task.

The workflow has been implemented on a 230-node grid controlled by the Sun Grid Engine, and is invoked via the command line or a webpage interface. Data are retrieved from the *GeoNet* archive using web service clients developed using the gSOAP package, providing high-speed, autonomous data transfer.

Experimentation using two workflow implementation packages, Kepler and Taverna, revealed unanticipated difficulties in incorporating in the workflow seismic data acquisition web services developed in conjunction with this project. Both packages appear to hold promise for describing geophysical workflows of this sort, but for the purposes of this study we have constructed the cross-correlation workflow from existing codes called in sequence by a shell script and invoked via the command line or a webpage interface.

The key outcomes of this project are: (1) a two-stage workflow enabling continuous seismic waveform data extracted from the *GeoNet* archive to be pre-processed and correlated in a systematic manner; (2) execution of this workflow within a grid processing environment, dynamically using the maximum processing power made available via the cycle-stealing infrastructure; (3) a web interface enabling the workflow to be invoked without the need for command line interaction; and (4) successful application of the workflow to datasets of various scales.

The collaboration throughout this project between geophysicists and computer scientists has highlighted the challenges of adapting specialist software developed for single-user, single-processor, serial use to a distributed or parallel environment. Our reliance on a cycle-stealing computational resource has also underscored the importance of designing workflow components in as modular a fashion as possible.

A grid-based facility for large-scale cross-correlation of continuous seismic data

John Townend[1], Yannik Behr[1], Kevin Buckley[2], Martha Savage[1], and John Hine[2]

[1]School of Geography, Environment, and Earth Sciences, Victoria University of Wellington

[2]School of Engineering and Computer Science, Victoria University of Wellington

Executive summary i

Technical abstract ii

List of figures iv

List of tables vi

Acknowledgments vi

Introduction 1

- Computational workflows 1
- Seismic cross-correlation 2
- Ambient noise tomography 3
- Links to other research projects 6

Technical requirements and software design 10

- Workflow description 10
- Grid resources 12
- Data acquisition 13
- Pre-processing (Stage 1) 13
- Cross-correlation (Stage 2) 14
- Post-processing 14

Deployment and evaluation 14

Recommendations for further research 16

- Incorporating other pre- or post-processing modules 16
- Future applications 17
- Improvements to the job allocation 18

Summary 19

Glossary 20

References 21

Appendix A: Documentation 24

Appendix B: Conference outputs 26

Appendix C: eResearch Australasia Abstract 27

List of figures

Figure 1 An example of the cross-correlation of two seismic waveforms. (Top) Two seismograms plotted as functions of time. The red and blue seismograms are near-exact copies of each other, although the red one has been degraded by the addition of a random noise signal, and delayed by several seconds. (Middle) Cross-correlation results for the red and blue signals showing that very high correlation is obtained for a lag of 5 s. (Bottom) Similar plot to the top one but with the blue curve shifted by the 5 s lag determined by cross-correlation.....2

Figure 2 An example of the tomographic imaging possible using cross-correlation of ambient seismic noise. (Left) Rayleigh wave group velocity map for a period of 13 s, computed from one year of continuous seismic noise recorded by 42 stations in the *GeoNet* network. This image corresponds to an approximate depth of 15 km in the crust (Lin et al., 2007a). (Right) An interpretation of the tomographic results. Clearly visible at this period (depth) are the low-velocity Taranaki and Canterbury sedimentary basins (TB and CB, respectively) and Hikurangi subduction accretionary prism (east of the North Island), high seismic-velocity material extending the length of the South Island as into central North Island, and the low-velocity Taupo Volcanic Zone near Rotorua.....4

Figure 3 Schematic representation of the data processing scheme used for ambient noise tomography (Bensen et al., 2007a). This project focuses on the first and second phases, which we refer to in this document as "pre-processing" and "cross-correlation". 5

Figure 4 Schematic illustration of the ambient noise correlation tomography procedure in the case of five seismographs and a single day's data. The calculations performed in phases 2–4 can be performed independently of each other.....6

Figure 5 Map showing broadband seismographs in the *GeoNet* national network (as of 2008) and temporary deployments providing data for ambient noise cross-correlation. SAPSE — Southern Alps Passive Seismic Array; (W)CNIPSE — (Western) Central North Island Passive Seismic Experiment; NORD — Northland Deployment; TVZ — Taupo Volcanic Zone. 7

Figure 6 Schematic representations of the processes by which continuous waveform data stored by channel (i.e. seismograph and sensor) and day can be extracted from the *GeoNet* archive using web services operated via KAREN. The web services pictured here were developed as part of the GNS Science/VUW Seismographic Information Services project. 8

Figure 7 An example of computational workflow construction in Kepler. The main panel shows different components ("actors") of the workflow, whose inputs and outputs are linked together to represent a sequence of processes. This example was constructed to represent acquiring data for a particular seismic network, station (seismograph), channel (sensor) and date using the SIS web services. 11

Figure 8 Screen-shot showing the webpage interface used to set up and submit a Stage 1 computation. A similar webpage has been created to submit Stage 2 jobs. See Appendix A for details of the different input parameters. 12

Figure 9 An example of the results of the cross-correlation analysis for 10 stations (90 days, 1 Hz sampling, vertical component, 3–40 s filtering). Each trace

represents the cross-correlation of data from a pair of seismographs, and is plotted on the vertical scale at a height corresponding to the great-circle distance between the two sites. This reveals pulses of surface wave energy propagating at speeds of 2–3 km s^{-1} across the network. Both positive ("causal") and negative ("acausal") lags are shown: signals of opposite lag represent waves propagating in opposite directions between the two stations in a cross-correlation pair. 15

Figure 10 Histogram of times taken for each of 3600 Stage 1 pre-processing tasks distributed amongst 180 processors on the 230-processor ECS grid. Each processor completed between 9 and 25 separate tasks.....16

List of tables

Table 1 Summary of the time taken using a single desktop computer and the ECS grid to complete the same task. 15

Table 2 Summary of the scales of cross-correlation encountered in different seismological applications. 18

Acknowledgments

We gratefully acknowledge the EQC Research Foundation's support of this project, the *GeoNet* project, and allied research. Fan-Chi Lin and Michael Ritzwoller (University of Colorado at Boulder) kindly provided the original ambient noise processing codes, which Stephen Bannister (GNS Science) has helped refine for New Zealand conditions. We also thank Paul Grimwood (GNS Science) for ongoing advice related to seismic data transfer and the *GeoNet* archive, and Christo Muller and Mark Davies (School of Engineering and Computer Science) for technical assistance.

Introduction

The advent of the world-class *GeoNet* network (http://www.geonet.org.nz/) and data archive and the Kiwi Advanced Research and Education Network (KAREN; http://www.karen.net.nz/) linking New Zealand universities and research institutions provides a key opportunity to analyse an increasingly routine geophysical task, namely the cross-correlation of continuous seismic waveforms, from the viewpoint of computational efficiency.

This project has two complementary goals. The first is to develop a computational workflow — a documented sequence of analytical steps — allowing automated or interactive analysis of continuous raw seismic data using grid-computing resources. The computational grid used in this study comprises approximately 230 desktop computers within the School of Engineering and Computer Science (ECS) at Victoria University of Wellington (VUW), each of which can be employed as a remote processor whenever it is otherwise idle. We focus our attention on the problem of analysing very long streams of ambient seismic noise data in order to obtain information about lithospheric velocity structure. The second goal of this project is to develop an interface to the computational workflow that facilitates its use in an efficient and effective manner by researchers in the broader geophysical community.

Computational workflows

As in many other data-intensive fields of science, geophysical research often involves a large number of incremental processing steps, during each of which decisions must be made regarding the particular choice of parameters or even algorithms to use. The research undertaken in this project is intended to combine an increasingly routine geophysical task, cross-correlating large data sets, with modern e-research approaches to systematising and documenting the research process itself.

The idea of encapsulating within a particular research output all of the processing parameters used in obtaining that output (a figure, table, or parameter value) from the original input (raw data) is not new. Examples within the geophysical realm include the Complete PostScript System (Wessel, 2003), an archival and exchange format that incorporates details of the processing parameters and algorithms used to generate it within an output image; the SINEX format (Solution-INdependent Exchange Format; Blewitt et al., 1994) used to exchange geodetic locations and the modelling parameters they depend on; and the Stanford Exploration Project reproducible electronic document protocol (Schwab et al., 1997), which consists of rules used to reproduce entire books from the original source code and data. What each of these examples — which are often referred to as “reproducible research” — provides, to varying degrees, is the potential for researchers to ask questions of their own or others’ research such as:

- Could we plot that data a different way to see things more clearly?
- What effect on the results would changing that parameter have?
- Would she have got the same result if she’d used his algorithm at that step instead of mine?

At its simplest, the implementation of a workflow is a computer program. However, when there are large amounts of data and possibly different computational resources involved the problem of tracking and documenting input parameters and processing

requirements becomes more complex. Several workflow modelling environments have been developed to enable researchers in various disciplines to design processing sequences by linking modular components. As discussed further below, we have experimented with using two such workflow modelling environments, Kepler and Taverna, for the cross-correlation processing, although in this case we have opted to use a simpler text-based system for tracking workflows.

Seismic cross-correlation

Cross-correlation underpins much of contemporary seismology, particularly in the areas of differential earthquake location (e.g. Du et al., 2004; Schaff et al., 2004; Bannister et al., 2006; Shelly et al., 2006; Shelly et al., 2007; Clarke et al., 2009) and ambient noise tomography (Shapiro et al., 2005b; Bensen et al., 2007a; Lin et al., 2007a). The cross-correlation operation computes the similarity between two signals and the optimal time-shift required to align one with the other: in broad terms, accurate earthquake location is based on measurements of this time-shift, and ambient noise tomography (Figure 1) on the degree of similarity.

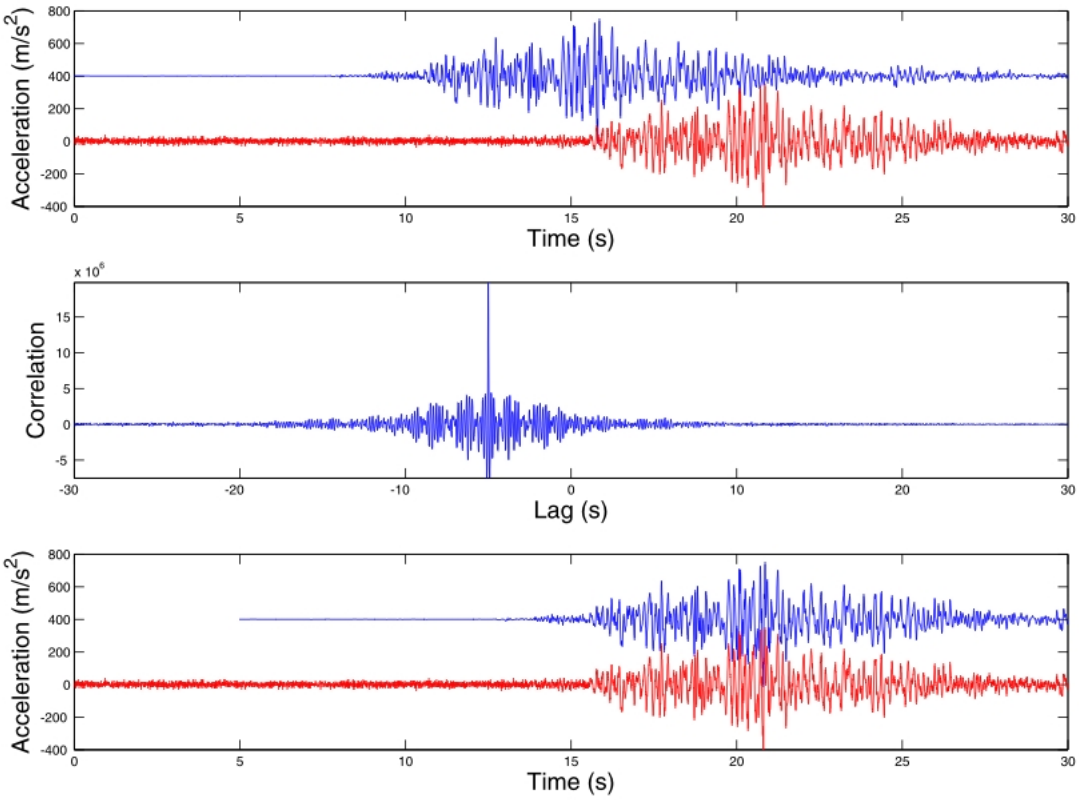

Figure 1 An example of the cross-correlation of two seismic waveforms. (Top) Two seismograms plotted as functions of time. The red and blue seismograms are near-exact copies of each other, although the red one has been degraded by the addition of a random noise signal, and delayed by several seconds. (Middle) Cross-correlation results for the red and blue signals showing that very high correlation is obtained for a lag of 5 s. (Bottom) Similar plot to the top one but with the blue curve shifted by the 5 s lag determined by cross-correlation.

In the last five years, it has become feasible to cross-correlate large data sets containing several tens of thousands of earthquakes or spanning many months of continuously recorded seismic noise. The computational demands of these calculations, however, preclude much in the way of experimentation with respect to parameter choices or alternative algorithms. Moreover, transferring data files from

an archive to the computer system used for the analysis is a non-trivial, time-consuming process in itself.

Ambient noise tomography

We focus the cross-correlation processes underpinning this project on ambient noise tomography, a novel field of research in which long streams of apparently random noise recorded at different positions are analysed to yield coherent information about seismic wave propagation (e.g. Shapiro and Campillo, 2004a; Shapiro et al., 2005b; Lin et al., 2006; Lin et al., 2007b). The speeds at which seismic waves propagate are governed by the physical properties of the rocks through which they pass. By measuring these speeds for waves of different wavelengths, we can map the Earth's deep structure in much the same way as ultrasound is used to look inside human bodies.

Just as radiographers use X-rays and ultrasound to image the internal structure of a human body, so seismologists use seismic waves generated by earthquakes or artificial sources to study the earth's interior. Earthquakes produce abundant seismic energy, but occur sporadically in unpredictable locations, whereas artificial sources can be tailored to specific targets but are less energetic than earthquakes and costly to deploy. Modern seismological networks, such as *GeoNet*, are designed to record seismic waves generated by earthquakes, but more than 95% of the "signal" recorded by such networks is ostensibly incoherent noise generated by the interaction of ocean waves with the seabed and coastline (Bromirski and Duennebier, 2002; Schulte-Pelkum et al., 2004; Barruol et al., 2006).

Cross-correlation plays a central role in ambient noise analysis. It can be demonstrated that cross-correlating long noise records from pairs of seismometers reveals a coherent signal corresponding to the propagation of a seismic wave from one instrument to the other (Shapiro and Campillo, 2004b; Sabra et al., 2005a; Shapiro et al., 2005a; Weaver, 2005). Over distances greater than ~10 km, this signal corresponds to a wave propagating just below the earth's surface (Roux et al., 2005) at speeds determined by the earth's elastic properties. Lower-frequency energy samples greater depths: measurements at periods of 5–25 s provide images of geological structure to depths of ~30 km (Sabra et al., 2005b; Shapiro et al., 2005a). Using waves of higher frequencies, researchers have also used ambient noise correlation to determine the shear-velocity structures of shallow sedimentary basins and building foundations for geotechnical purposes (Gouédard et al., 2008; Bonnefoy-Claudet et al., 2009).

The appeal of ambient noise imaging lies in using pervasive, high-amplitude seismic energy to map the earth's structure over large areas (Sabra et al., 2005b; Shapiro et al., 2005a; Bensen et al., 2007b; Yang et al., 2007), and this is now the focus of vigorous international research (Gerstoft et al., 2006a; Gerstoft et al., 2006b; Sens-Schonfelder and Wegler, 2006; Stehly et al., 2006; Bensen et al., 2007b; Brenguier et al., 2007). In 2006, we successfully demonstrated the potential for noise imaging in New Zealand. Using data from 42 *GeoNet* instruments, we undertook the first New Zealand-wide study of lithospheric Rayleigh wave speeds(Lin et al., 2007a) and for the first time characterised crustal structure over the entire country using a single technique (Figure 2).

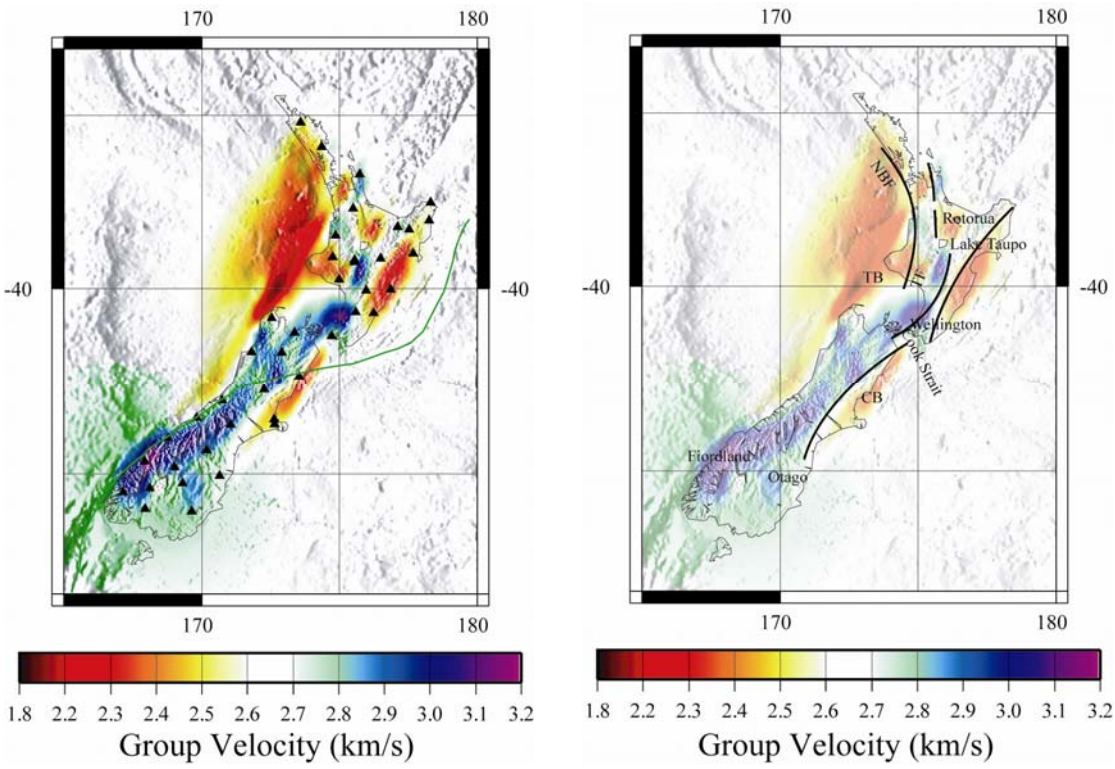

Figure 2 An example of the tomographic imaging possible using cross-correlation of ambient seismic noise. (Left) Rayleigh wave group velocity map for a period of 13 s, computed from one year of continuous seismic noise recorded by 42 stations in the *GeoNet* network. This image corresponds to an approximate depth of 15 km in the crust (Lin et al., 2007a). (Right) An interpretation of the tomographic results. Clearly visible at this period (depth) are the low-velocity Taranaki and Canterbury sedimentary basins (TB and CB, respectively) and Hikurangi subduction accretionary prism (east of the North Island), high seismic-velocity material extending the length of the South Island as into central North Island, and the low-velocity Taupo Volcanic Zone near Rotorua.

A key advantage of ambient noise over either natural seismicity or artificial shots for subsurface imaging is that it provides a fixed, repeatable energy source. This has two important implications: (1) robust wave speed measurement uncertainties can be estimated by analysing data from different time periods (e.g. Bensen et al., 2007b; Yang et al., 2007); and (2), one set of measurements can be used as a benchmark for detecting subsequent changes (Brenguier et al., 2007).

Ambient noise analysis for tomographic purposes typically consists of four distinct phases (e.g. Bensen et al., 2007a, and references therein), summarized in the following paragraphs: (1) station-by-station data preparation; (2) cross-correlation and stacking; (3) dispersion analysis; and (4) quality control (Figure 3). This project focuses on the first two phases, the first involving data from separate stations individually and the second applied to data from pairs of stations — for the sake of completeness, we also outline the third and fourth phases.

During the first phase, data recorded at each of N stations on a single day are prepared for cross-correlation by removing the instrument response, removing the mean signal and any linear trend, and if necessary amalgamating several short files into day-long records. To ameliorate the unwanted effects of any earthquakes or spikes in the data set, each chunk of data is normalized using one of several possible methods and spectrally whitened in the Fourier domain. This stage of analysis involves N multi-step operations, each of which currently takes approximately 60 s on a 2.8 GHz

desktop PC for one day's data recorded at 50 Hz. If *M* days' data are being analysed, the number of operations is *MN*.

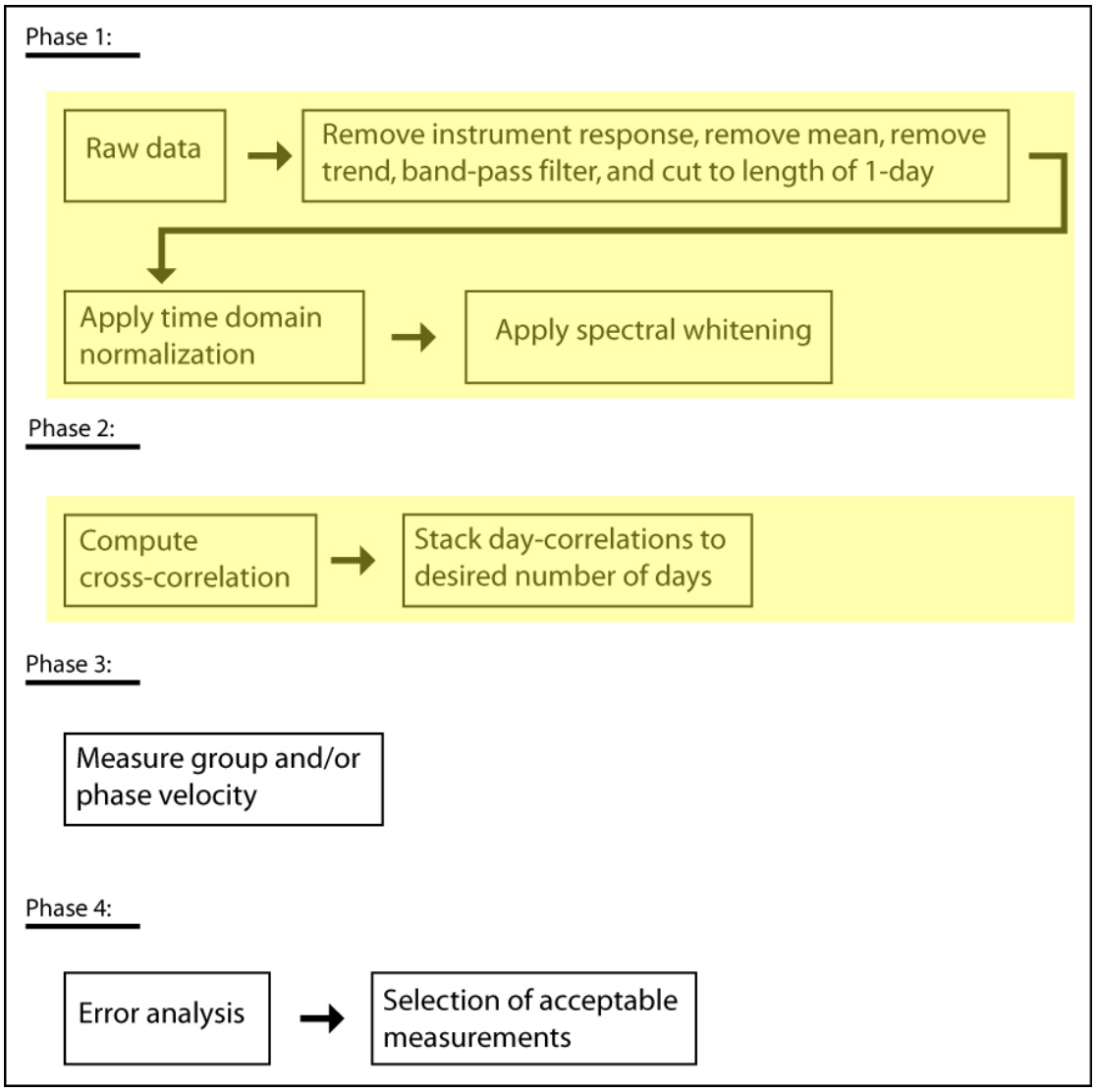

Figure 3 Schematic representation of the data processing scheme used for ambient noise tomography (Bensen et al., 2007a). This project focuses on the first and second phases, which we refer to in this document as "pre-processing" and "cross-correlation".

The second phase, the cross-correlation itself, involves a pair-wise multiplication of the outputs of phase one. For each day-long chunk of data, this requires a total of $1+2+...+N = N(N–1)/2$ processes, each of which currently takes approximately 1 s and can be completed independently of the other pairs' processes. The cross-correlation results are transformed back to the time-domain and stacked to the corresponding results for any other days being analysed. If *M* days' data are analysed, a total of $MN(N–1)/2$ processes are involved.

The dispersion analysis carried out in the third phase of analysis is performed on each cross-correlation stack separately, meaning that $N(N–1)/2$ independent processes are required. Most groups doing ambient noise correlation research (including our group; Lin et al., 2007b; Behr et al., 2009) use automated frequency–time analysis incorporating a phase-matched filter determined by preliminary measurements of the group speed dispersion curve (Levshin and Ritzwoller, 2001). In essence, this involves measuring group speed twice over a range of frequencies: the first suite of

measurements defines a filter that is used to emphasise the signal of interest for the second, refined suite of measurements.

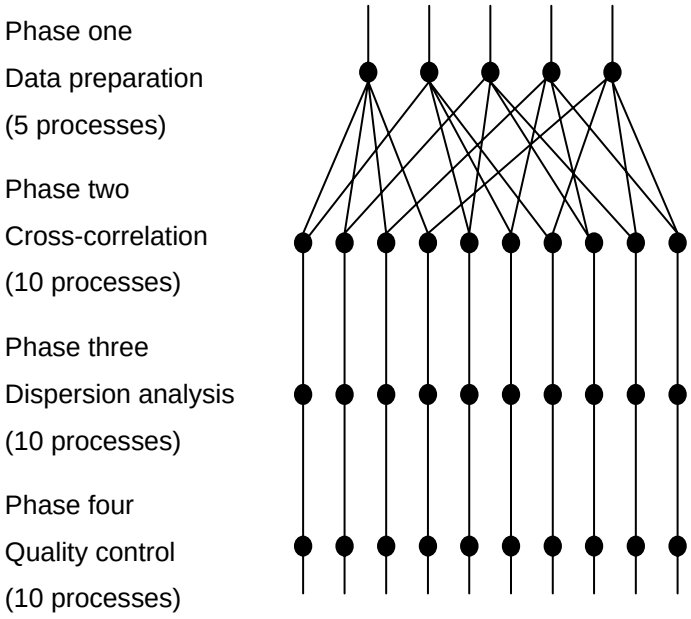

Figure 4 Schematic illustration of the ambient noise correlation tomography procedure in the case of five seismographs and a single day's data. The calculations performed in phases 2–4 can be performed independently of each other.

Finally, the dispersion analysis results undergo quality control based on a number of factors including the measured signal-to-noise ratio, temporal repeatability (with respect to measurements made for other time periods), and compatibility with measurements along similar paths between other pairs of stations. This again requires $N(N-1)/2$ processes.

Overall, therefore, the complete analysis can be represented as requiring a total of $MN+(M+2)(N-1)/2$ processes, each of which involves a number of smaller operations (cf. Figure 4). For the 12 months of data from the then-42-station national component of *GeoNet* we worked with in our earlier study (Lin et al., 2007a), this corresponded to a total of ~330,000 separate processes; with *GeoNet*'s ongoing expansion and the incorporation of data from temporary deployments in the noise correlation work (Figure 5), the scale of the computation becomes ever-larger. This analysis represents the case in which data from a single seismograph sensor (typically the vertical component) are being analysed: working with all three seismograph components, increases the computational requirements by a factor of three.

Links to other research projects

The cross-correlation workflow modelling has been conducted in close conjunction with several related VUW and GNS Science projects, pertinent details of which we summarise briefly here.

Seismographic Information Service (SIS)

GeoNet collates ~3.5 Gb of continuous seismic data each day and the current volume of archived data is >6 Tb. Until recently, external users accessed the data using an email-based data request procedure ("autoDRM", for waveform data) or a searchable web-interface ("QuakeSearch", to access a subset of earthquake catalogue parameters

such as origin time, hypocenter, and magnitude). Neither tool was well-suited to automated transfer, nor was it possible to execute large requests such as “give me a year’s-worth of continuous waveform data, please”.

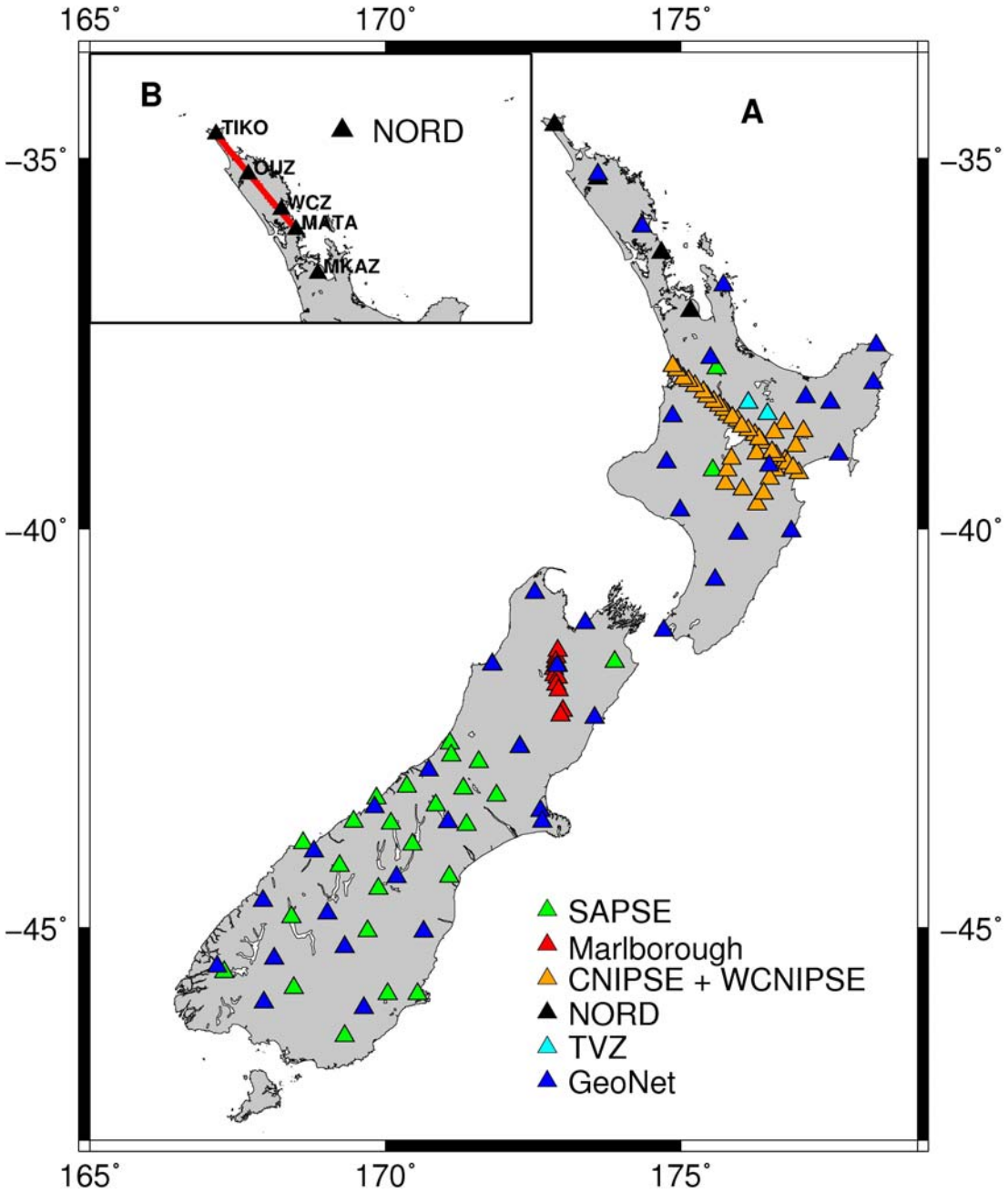

Figure 5 Map showing broadband seismographs in the *GeoNet* national network (as of 2008) and temporary deployments providing data for ambient noise cross-correlation. SAPSE — Southern Alps Passive Seismic Array; (W)CNIPSE — (Western) Central North Island Passive Seismic Experiment; NORD — Northland Deployment; TVZ — Taupo Volcanic Zone.

The 12-month “Seismographic Information Service” (SIS) KAREN Capability Build Project was led by Paul Grimwood of GNS Science and addressed methods of accessing and distributing broadband seismographic data via the KAREN network.

Under the auspices of the SIS project, data access tools better suited to current and projected data volumes and research demands were developed for different kinds of seismic data:

- Event metadata ("earthquake catalogue") — the *GeoNet* catalogue of earthquake locations, arrival time measurements ("picks"), magnitude information etc. has been mapped to a standardized XML schema and controlled vocabulary known as "QuakeML";
- Continuous waveform data (hour–year durations) — data stored in a channel-day miniSEED archive can be requested using client-side services that interact with the server-side providers via a Web Services Description Language (WSDL) template;
- Event waveform data (1–100 s durations) — waveform data for a particular event can be obtained by first parsing the event's QuakeML metadata and then requesting continuous data spanning the time period and stations of interest.

As discussed further below, the second of these web services is the one used in the present project to acquire data from the *GeoNet* archive. Using a standard desktop computer, it is currently possible to download continuous waveform data at rates of several megabytes per second.

Ambient noise field characterisation and tomography

The appeal of ambient noise imaging lies in using pervasive, high-amplitude seismic energy to map the earth's structure over large areas, and as noted in the Introduction this is now the focus of vigorous international research. Since demonstrating the potential for New Zealand-wide ambient noise correlation tomography (Lin et al., 2007b), the VUW group has begun focussing on applying ambient noise correlation to specific tectonic targets.

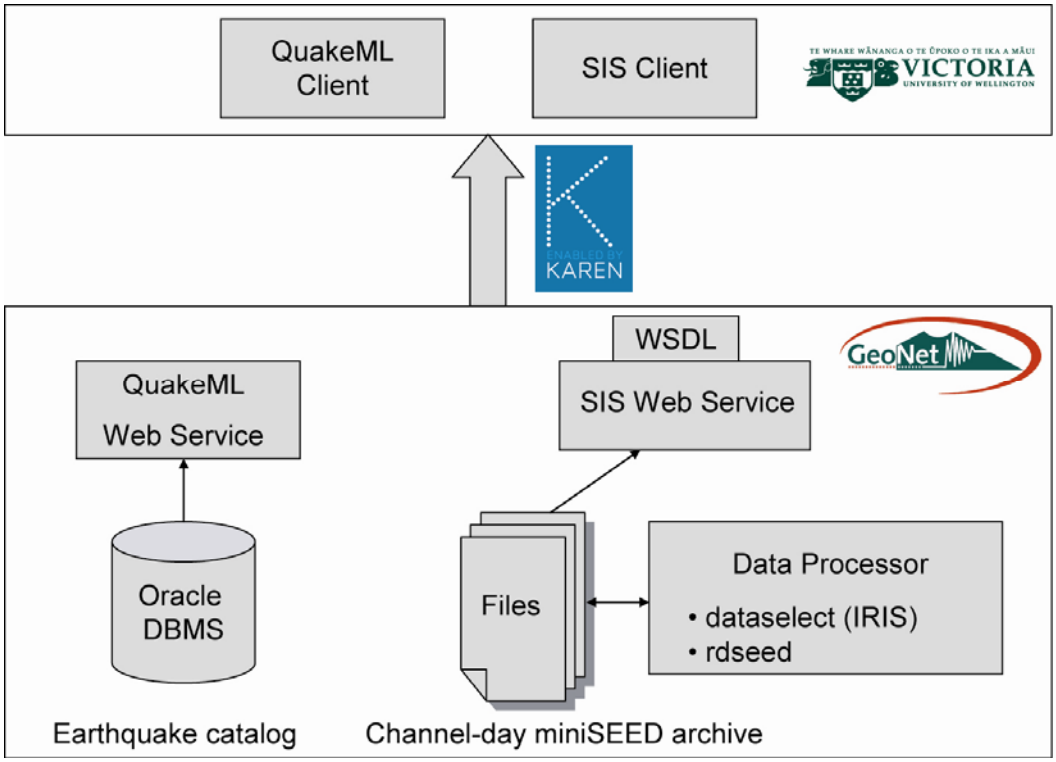

Figure 6 Schematic representations of the processes by which continuous waveform data stored by channel (i.e. seismograph and sensor) and day can be extracted from the *GeoNet* archive using web services operated via KAREN. The web services pictured here were developed as part of the GNS Science/VUW Seismographic Information Services project.

With the support of the Royal Society of New Zealand's Marsden Fund, we are now analysing on a routine basis multi-year noise records from more than 100 seismometers in *GeoNet*'s national and regional networks and recent temporary deployments. In conjunction with ocean wave state data, we are also working to identify distinct noise sources and characterise the noise field itself.

Two of the Marsden project's three main objectives are closely linked to this cross-correlation research and the SIS data acquisition methods that underpin it:

1. Systematic correlation of *GeoNet* and temporary network noise records and computation of seismic velocities at periods of <2–25 s for surface (Rayleigh, Love) and local body (P) waves (Behr et al., 2009);
2. Quantification of spatial and temporal (seasonal) variations in the seismic noise field in conjunction with an oceanic wave state model (preliminary results now in prep., Brooks et al., 2009);

Seismic tremor detection

Slow earthquakes are episodes of fault slip occurring over days or weeks, and were first identified using global positioning system instruments. They have been detected in several subduction zones, where one tectonic plate is thrust beneath another, most notably in the western United States and Canada, Japan, Mexico, and the Hikurangi subduction zone east of the North Island of New Zealand (Douglas et al., 2005). Geodetically-detected episodes of this slow slip appear in several subduction zones to be accompanied by bursts of low-frequency coherent noise known as seismic tremor, but whether a single physical process governs this association or even whether slow slip is invariably accompanied by tremor remains unresolved.

In a recent completed study, VUW and GNS Science researchers conducted a detailed analysis of broadband seismic data spanning a slow slip episode near Gisborne (Delahaye et al., 2009). This analysis revealed that slow slip was accompanied by distinct reverse-faulting microearthquakes, rather than tremor. The timing, location, and faulting style of these earthquakes are consistent with stress triggering down-dip of the slow slip patch, either on the subduction interface or just below it.

Delahaye et al. (2009) observed that the Gisborne slow earthquake occurred in a tectonically different environment from those in which tremor has been reported to accompany slow slip, and suggested that temperature or pressure conditions govern the seismic response to slow slip. With funding provided by the VUW Faculty of Science and the EQC Research Foundation, we have begun developing methods of monitoring continuous seismic data recorded during episodes of slow slip for seismic tremor. We have focussed to date on a slow earthquake that began in December 2007 between the Kapiti Coast and the Marlborough Sounds, which has very different depth, duration, and size characteristics from the previously studied event near Gisborne and may enable Delahaye et al.'s hypothesis about the roles played by temperature and pressure to be tested.

Work is ongoing to implement two tremor-detection algorithms (Kao and Shan, 2004; Wech and Creager, 2008) on VUW computers for near-real-time analysis of continuous waveform data provided by *GeoNet*. The tremor detection process is well-suited to a grid-computing environment, and to the tools being developed here, but further research on that topic has yet to commence.

Technical requirements and software design

The cross-correlation workflow described below incorporates already developed and operational processing codes (originally provided by Fan-Chi Lin and Michael Ritzwoller, University of Colorado at Boulder, and since modified at VUW in collaboration with Stephen Bannister, GNS Science). Initial discussions between the earth science (JT, YB, MKS) and computer science teams (KB, JH) regarding the computational demands posed by the cross-correlation problem and the grid environment in which these would be addressed suggested that a two-stage process be adopted, each independent of the other.

In the first stage, which corresponds to phase 1 of the ambient noise processing scheme illustrated in Figure 3, individual traces are passed along the data manipulation pipeline, whilst in the second stage (phase 2 in Figure 3) pairs of pre-processed files are operated on together.

A benefit of the two-stage approach adopted here is that pre-processed data can, where storage is available, be stored offline. This removes the need to return to a remote provider of the initial data — the *GeoNet* archive in this case — each time a new analysis of already processed data is undertaken, during testing of new activities against some data, or when new users are being instructed in how to use the process. Such "connectivity vs. storage vs. processing" trade-offs may be of more concern to researchers in the future as larger computational problems require computing to be done on remote resources that incur connectivity, storage, or processing costs.

Workflow description

Different researchers may wish to use different algorithms at each stage of the analysis, or experiment with altering intermediate parameters to see how sensitive the results are to various decisions. To completely describe the computation as a whole, a cross-correlation workflow needs to record:

- The initial input data sets;
- The programs/algorithms used at each stage of the analysis;
- The input parameters required by each of those programs;
- The results, in the form of numerical data and/or graphical output.

At the outset of this project, we envisaged creating a catalogue of cross-correlation tasks comprising two components:

- A structured file system holding pointers to the input data sets and the results;
- A database holding the meta-information describing the execution of each workflow and the parameters used for the run.

We experimented with using two workflow implementation packages, Kepler and Taverna, to describe and invoke the cross-correlation processing. Each of these packages enables a sequence of computational steps to be assembled in a modular fashion. Our experience with Kepler (https://kepler-project.org/) was that while the graphical interface (Figure 7) made workflow description straightforward, the available modules for obtaining data remotely — specifically, the "Web Service Actor" — did not interface well with the SIS web services. Output from the Web Service Actor could not be piped into other modules we wished to use, and we were unable to make the workflow interactive without the use of a separate "interaction

server", which in Kepler's default configuration is based at the San Diego Supercomputer Center. Taverna (http://www.mygrid.org.uk/tools/taverna/), which has been largely developed within the computational biology community, provides a more involved (i.e. less graphical) method of assembling workflows. Our experience with Taverna was, similarly, that remote data acquisition — using the "WSDL scavenger" module — was difficult to implement and ultimately failed when the *GeoNet* WSDL was updated, in spite of a series of discussions with the Taverna development team.

Kepler and Taverna do appear to hold promise for describing workflows of the sort involved with the cross-correlation process, but establishing how to use them for seismic data acquisition and grid-processing has proven to be beyond the scope of this project. A more extensive evaluation of how to use these or similar workflow implementation packages for seismological purposes should be carried out, perhaps in the form of a student research project.

In light of our experience with Kepler and Taverna, we have opted to use a much simpler representation of the workflow that is more immediately compatible with the software modules required to acquire *GeoNet* data and perform the different processing steps. Our workflow consists of several existing or customised modules developed as part of the larger ambient noise tomography project, which are called in sequence by a master program written as a shell script. The master program is run directly from the command line (using standard UNIX syntax; see Appendix A) or via a web interface (Figure 8).

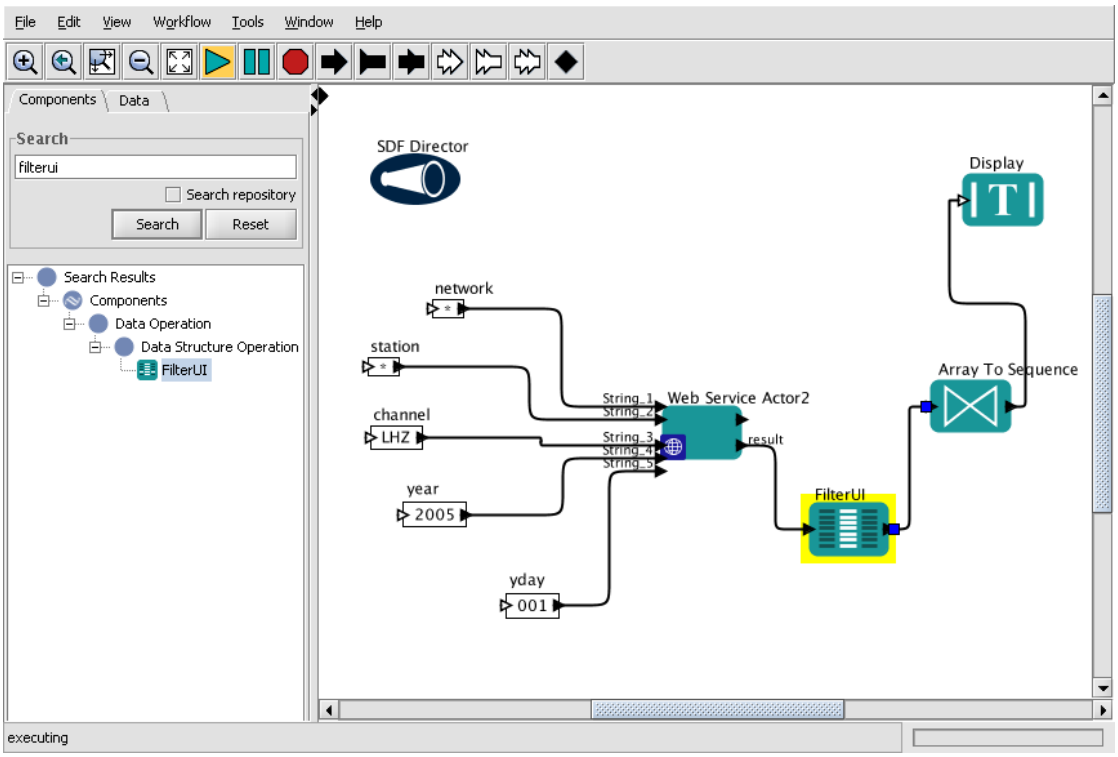

Figure 7 An example of computational workflow construction in Kepler. The main panel shows different components ("actors") of the workflow, whose inputs and outputs are linked together to represent a sequence of processes. This example was constructed to represent acquiring data for a particular seismic network, station (seismograph), channel (sensor) and date using the SIS web services.

Grid resources

VUW currently operates two distributed computing facilities ("grids"), each of which comprises a large number of desktop computers ("hardware") running common software within a centralised job allocation framework ("middleware").

The Engineering and Computer Science (ECS) Grid consists of approximately 230 desktop computers running the netBSD operating system under the control of a SUN Grid Engine (SGE). When a particular desktop computer is idle (such as overnight or whenever its owner is not logged on), its processing power can be utilised by the grid, in a process known as cycle-stealing or cycle-scavenging. This processor is available until the owner logs on. The Student Computing Services (SCS) grid operates within a similar cycle-stealing framework, but consists of a larger number (~1000) of desktop computers running the Windows XP operating system under the control of Condor.

We have chosen to use the ECS grid for this project, as many of the component codes cannot easily be deployed on the Windows XP-based SCS system, including the web-services required to obtain data from *GeoNet*, Python, and the SAC suite of seismological utilities.

An obvious limitation of the cycle-stealing Sun Grid Engine is that one or more desktop computers may be removed from the pool by its owner logging on. In that case, any tasks allocated to that processor are suspended until the machine becomes available once again. We discuss ways in which this issue might be addressed in future work in the "Recommendations for future research" section below.

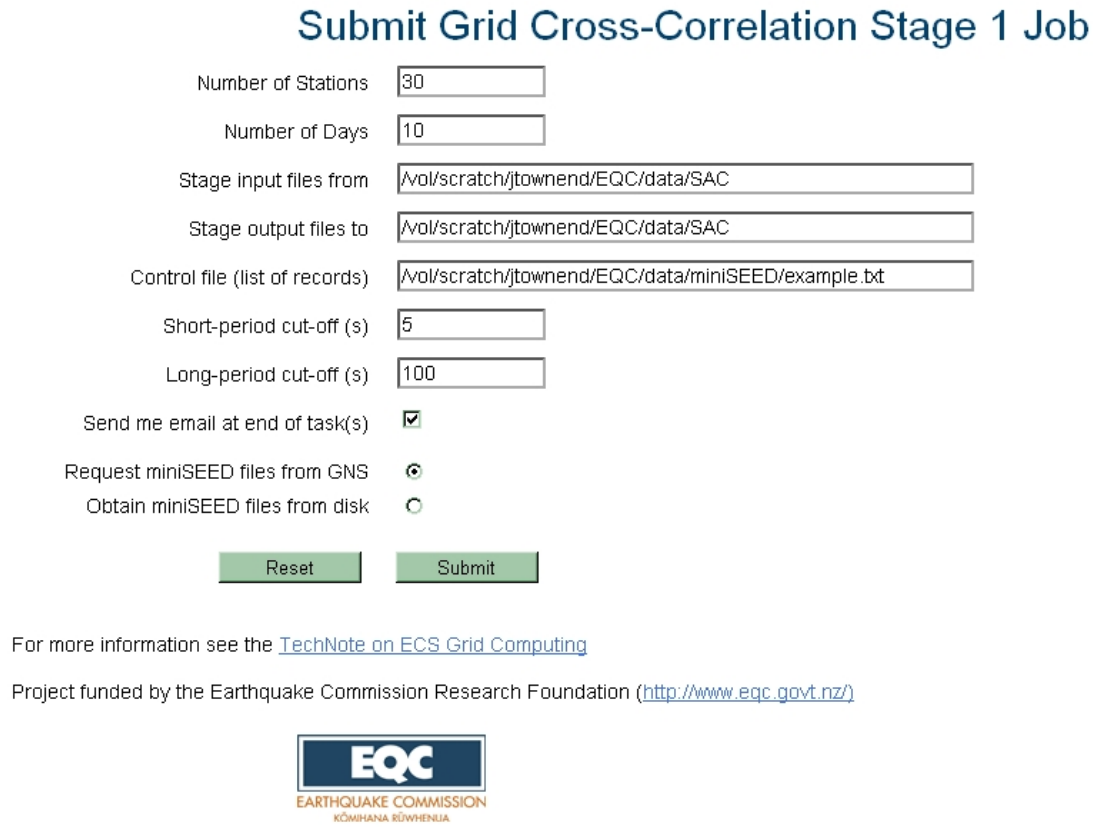

Figure 8 Screen-shot showing the webpage interface used to set up and submit a Stage 1 computation. A similar webpage has been created to submit Stage 2 jobs. See Appendix A for details of the different input parameters.

Data acquisition

Access to the *GeoNet* continuous seismic waveform dataset is provided via a web-service defined by a Web Services Description Language (WSDL) file hosted on the *GeoNet* website (Figure 6). We have used the gSOAP package (http://gsoap2.sourceforge.net) to create core client applications in C from the specifications provided in this WSDL file. This provides high-speed access to data with minimal user intervention: convenience functions have also be written around this code skeleton to provide a more user-friendly interface, which has since been used by colleagues working on allied projects.

We have also successfully constructed a second client-interface to the web-service using Java and the Netbeans IDE, which highlights the versatility of the WSDL data transfer method.

The workflow has been designed to either download the data files specified in the Stage 1 parameter file from the *GeoNet* archive, or to use previously downloaded files already stored within the grid.

Pre-processing (Stage 1)

Each of the original data files can be operated on in Stage 1 independently of all other data files. Distributing the computation between the available processors is achieved by submitting the total number of tasks to the Sun Grid Engine (see Appendix A), which allocates them to the currently available processors. The files themselves are stored contubernially on a central disk visible to all processors, facilitating file access across the distributed grid.

Reformatting

The data extracted from the *GeoNet* archive are in miniSEED format (a standard for seismic data exchange) and require reformatting prior to any further processing. We first convert each miniSEED file into Seismic Analysis Code (SAC; http://www.iris.washington.edu/software/sac/) format, which is more suitable for the subsequent analysis, using a standard conversion utility known as rdseed (http://www.iris.washington.edu/manuals/rdseed.htm).

Removal of the instrument response (deconvolution)

In order to extract a true record of ground motion from a seismogram, we need to account for and remove the effects of the instrument that recorded the ground motion. This process is referred to as “deconvolution”, and represents the single-most time-consuming stage in the overall process.

The necessary instrument response files for all the seismographs of interest are downloaded from *GeoNet* in a standard format (dataless SEED volume), from which a single file corresponding to each station is obtained using rdseed. Each of these files contains a complete mathematical representation of the recording process, including the sensor, digitizer, and any filtering. The deconvolution itself is performed using SAC routines.

Signal cleaning and filtering

The mean and trend of each waveform are removed using SAC and standard Python modules, and the record filtered to the period band of interest (typically 5–100 s periods for the ambient noise correlation tomography work currently being

undertaken) using Python, Fortran and the fftw subroutine library for computing discrete Fourier transforms (http://www.fftw.org/).

Temporal normalisation and spectral whitening

To avoid any earthquakes or instrumental glitches in the data artificially dominating the cross-correlations, we create signal envelope functions and use these to down-weight high-amplitude signals in the filtered data.

The output of the temporal normalisation and spectral whitening process is a SAC-formatted file corresponding to a particular station and day. This represents the final output from the pre-processing stage (Stage 1), and the input to the cross-correlation process (Stage 2).

Cross-correlation (Stage 2)

Once Stage 1 has been completed, we can invoke the Stage 2 cross-correlation from the command line or using the webpage interface. The inputs to Stage 2 are the same as for Stage 1, with the addition of a parameter specifying the total number of days for which to perform the cross-correlation. As in Stage 1, the necessary input files are stored in a common directory visible by all processors in the grid. We have implemented a job allocation scheme, tailored to the specific nature of the ECS grid and the processing power available at any one time, in which the total number of cross-correlations to be performed is divided equally amongst the available processors. This approach should facilitate the transition to other grid-infrastructure in the future.

Post-processing

The Stage 2 output files are returned to a directory specified in the parameter file or via the web interface, and can then be transferred to another site for further processing. To date, we have not included any graphical output in the workflow, since individual requirements differ at this point in the analysis, but codes have been developed to enable some commonly used diagrams to be generated with minimal user input. One such diagram is illustrated in Figure 9.

Deployment and evaluation

We have successfully tested the workflow using datasets of various sizes: sampling rates of 1 Hz and 100 Hz; data set durations of up to 365 days; and 2–10 stations.

Since the number of grid processors available at any particular time is variable, given that the ECS grid operates via cycle-stealing, the gain in processing speed we gain by using the grid also varies. As noted above, the workflow dynamically utilises the maximum number of processors available at any one time, and to date we have made concurrent use of as many as 180 otherwise idle processors (78%).

Table 1 illustrates the times taken to perform the same job on a single-processor desktop computer and the ECS grid. In this particular case (5 stations, 99 days, data sampled at 1 Hz), using the grid resources available at the time resulted in an overall reduction in run-time of approximately 87%, from 39.4 minutes to 5.1 minutes. It is important to highlight that most of this gain took place in Stage 1, which ran approximately 12 times faster on 58 processors than on a single processor. As noted above, this is because removal of the instrument response represents the single most time-consuming step in the pre-processing.

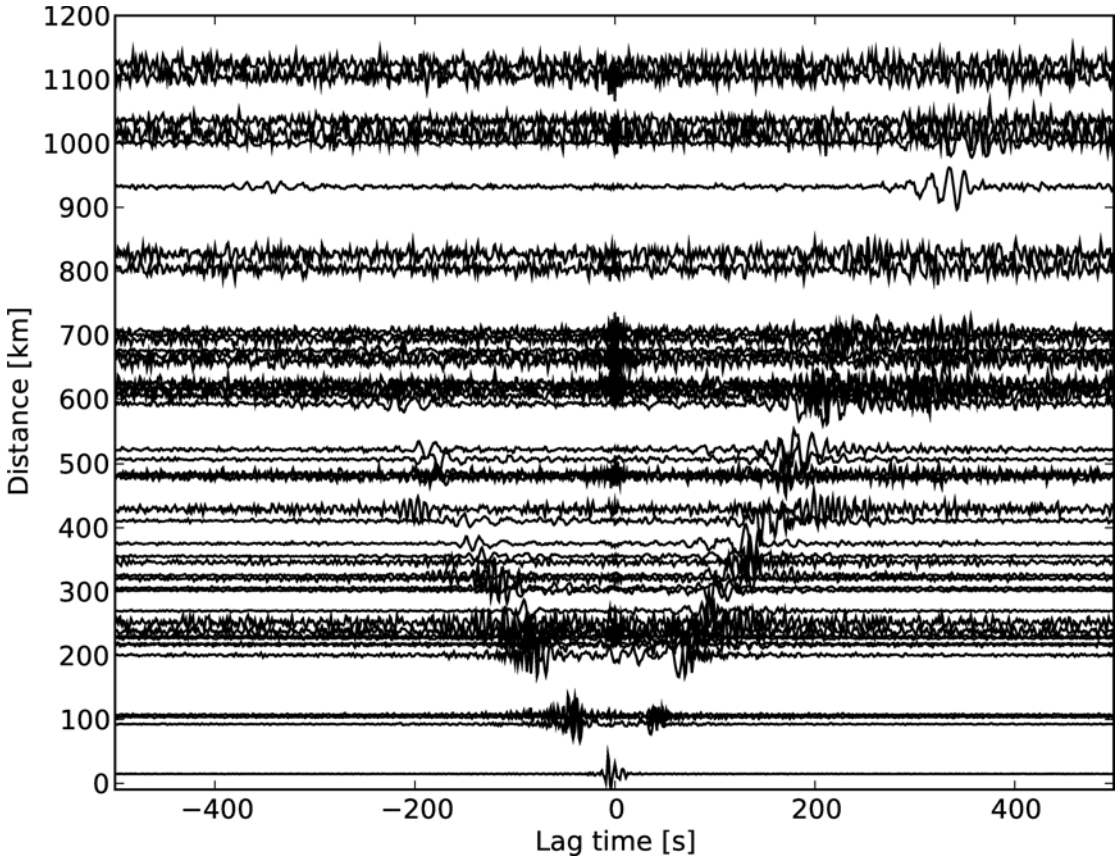

Figure 9 An example of the results of the cross-correlation analysis for 10 stations (90 days, 1 Hz sampling, vertical component, 3–40 s filtering). Each trace represents the cross-correlation of data from a pair of seismographs, and is plotted on the vertical scale at a height corresponding to the great-circle distance between the two sites. This reveals pulses of surface wave energy propagating at speeds of 2–3 km s^{-1} across the network. Both positive ("causal") and negative ("acausal") lags are shown: signals of opposite lag represent waves propagating in opposite directions between the two stations in a cross-correlation pair.

	Desktop computer	ECS grid
Number of stations	5	5
Number of days	99	99
Sampling frequency (Hz)	1	1
Number of processors	1	58 (Stage 1), 117 (Stage 2)
Time for Stage 1 (min)	34.2	2.9
Time for Stage 2 (min)	5.2	2.2
Total time (min)	**39.4**	**5.1**

Table 1 Summary of the time taken using a single desktop computer and the ECS grid to complete the same task.

The largest workflow executed on the ECS grid to date involved the Stage 1 processing of 360 days' data recorded at 100 Hz sampling at 10 stations. This represents a total of 3600 tasks, which were completed on 180 processors in a total time of three hours. By way of comparison, the time required to perform one of these tasks (1 station, 1 day) on a single processor was found to be approximately two

minutes: to process all 3600 records using a single processor would require approximately 120 hours. And using the grid resulted in an approximately 40-fold reduction in run-time.

This job was executed at 2 am on a weekday morning, when a large number of grid processors were available. Inspection of the job-completion reports reveals that 180 processors performed between 9 and 25 tasks each (Figure 10), and that the entire Stage 1 calculation used a total of ~114 Gb of disk space. The difference between the number of processors used (180) and the overall improvement in calculation speed (40-fold) represents a combination of the unequal workloads performed by each processor, the different speeds of the processors, and the computational overhead associated with deploying the workflow on the grid.

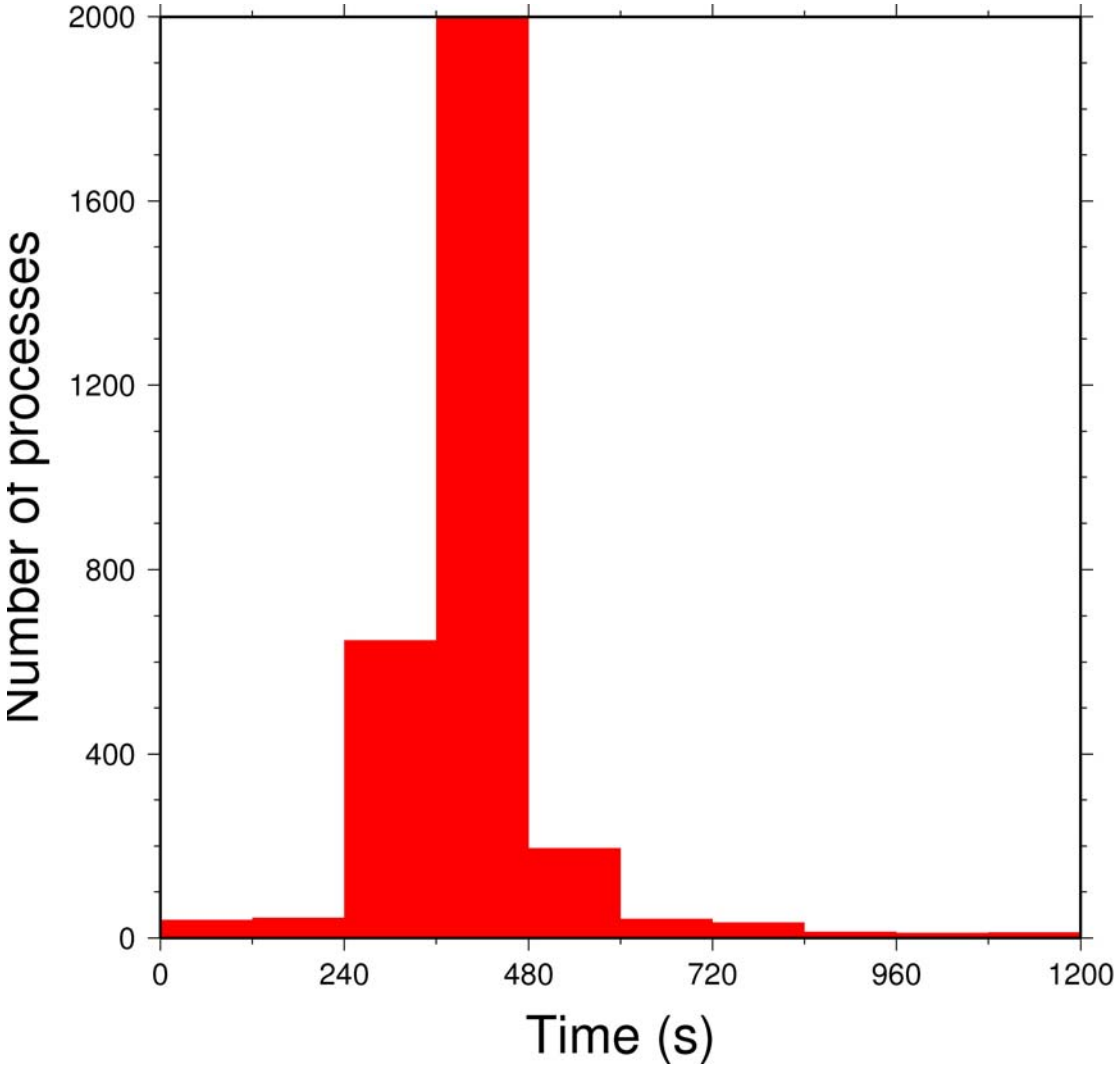

Figure 10 Histogram of times taken for each of 3600 Stage 1 pre-processing tasks distributed amongst 180 processors on the 230-processor ECS grid. Each processor completed between 9 and 25 separate tasks.

Recommendations for further research

Incorporating other pre- or post-processing modules

We have focussed in this project on a workflow addressing the immediate needs of several staff and students at VUW. Several facets of this workflow (notably the

deconvolution, temporal normalisation, and spectral whitening) are required for ambient noise studies, but would not necessarily be used by researchers interested in, for instance earthquake relocation or tremor detection. Similarly, the focus here has been on continuous records rather than, say, earthquake seismograms.

Constructing the workflow from pre-existing modular components, most of which have been developed independently by different researchers for specific applications, imposes some constraints on the file structure that must be used. In particular, modules designed to operate serially on files stored in a particular static directory structure on a single computer cannot be straightforwardly linked to one another in a distributed, parallel-processing realm unless the same directory structure is used or the modules modified accordingly.

What this has meant in the case of the cross-correlation workflow developed here is that file structures created at intermediate steps in Stage 1 have had to be made explicit during Stage 2 even though the Stage 2 calculations depend solely on the final Stage 1 outputs. In other words, by adopting processing codes developed originally for sequential use by a single user with a single computer, we have not been dealing with a situation in which each component reads input and generates output without any reliance on an over-arching file structure.

This situation does not make it straightforward for new users of the workflow to modify it by replacing, adding, or removing specific modules without having to retain the existing file-naming conventions and implicit directory structures. In order to enable future users of the workflow to customise it for specific requirements, it may prove helpful to progressively replace the modules now being used with versions that do not carry inherited file-structure overheads. Two options to consider as initial steps in this process would be to redesign the outputs of Stages 1 and 2 as single files independent of directory structure, or to make the creation the necessary directory structure a workflow module itself.

Future applications

The ambient noise cross-correlation addressed here typically involves cross-correlation of hundreds or thousands of long data files. We anticipate future research at VUW in seismological data processing to also involve either asymmetric cross-correlation (in which the two files in a pair are of different lengths, such as is the case for seismic tremor detection using matched filters), or much greater numbers of smaller files (such as with differential earthquake location). Table 2 summarises the different scales of cross-correlation encountered when performing cross-correlation for different seismological purposes. It is likely that new processing modules will need to be developed to accommodate the different scales of cross-correlation.

We noted above that detecting tremor associated with slow slip in the Hikurangi subduction zone is an ongoing research effort at VUW. The cross-correlation workflow could also be straightforwardly adapted for low-magnitude earthquake detection and applied, for example, to the problem of detecting earthquakes similar to those accompanying the 2004 Gisborne slow slip event (Delahaye et al., 2009) in spanning longer durations.

We have obtained a cross-correlation data set suitable for analysis using the existing cross-correlation workflow, and will begin this analysis in July 2009. Using data recorded at *GeoNet* seismographs in western New Zealand and sites in eastern

Australia and New Caledonia, we will conduct an ambient noise correlation study of the Tasman Basin based on results obtained in a preliminary study in 2008 by VUW undergraduate Zara Rawlinson (Rawlinson et al., 2008). We intend to publish this work on computational workflows and grid-based cross-correlation using the trans-Tasman data as a case study.

	Record length	**Template length**	**Number of records**
Tremor/earthquake detection	Long (10^4 s)	Short (10^{-1}–10^3 s)	10^3–10^4
Ambient noise correlation	Long (10^4 s)	Long (10^4 s)	10^3–10^4
Relative relocation	Short (10^{-1}–10^1 s)	Short (10^{-1}–10^1 s)	10^5–10^6

Table 2 Summary of the scales of cross-correlation encountered in different seismological applications.

Improvements to the job allocation

To encourage desktop computer users to contribute as many surplus central processing unit cycles to the ECS grid as possible, the Sun Grid Engine has been configured so that any console use will either suspend a grid task that is already running on that processor, or prevent any further tasks from beginning until the console user has logged out. This protects the desktop user from possible performance degradation caused by the grid job. It also means, however, that the grid user is not guaranteed a specific number of processors for the duration of a large job, and can lead to some jobs being interrupted part-way through. Among other things, running grid computations in an environment in which the pool of available resources may vary without warning constrains the use of message-passing methodologies between parallel tasks.

Such an operating environment lies some way from the idealised “cloud” or “cluster” metaphors of distributed computing, in which a grid user is assured that certain distributed computational resources are theirs and only theirs for a specific length of time once access has been obtained. Indeed, in many such environments a grid user will know the waiting time for resources to become available as well.

Knowledge of the expected wait and execution times can serve to inform the way in which a grid computation is partitioned within a distributed environment to maximise the use of the resources. In the absence of that knowledge, as has been the case in this project, it is necessary to partition the overall computation into a large number of small tasks so that no one task’s failure to complete on schedule jeopardises the whole computation.

The experience we have gained in deploying geophysical tools on the ECS grid has highlighted the opportunities to further investigate optimum job allocation strategies. The workflow developed to date invokes the separate tasks, in either Stages 1 or 2, as a single job that is then distributed by the Sun Grid Engine to whatever processors are currently available. We have not yet explored the options and benefits of reserving a certain number of processors (K) for the expected duration of the entire job and allocating each processor $1/K$ of the total computation.

We intend to investigate further different mechanisms for optimising task allocation within the constraints of the ECS grid's cycle-stealing operation. Preliminary discussions have been held with staff in ECS about involving senior undergraduate students in this work in the second half of 2009.

Summary

This project has focussed on implementing an increasingly routine geophysical task, the cross-correlation of long streams of seismic data, as a computational workflow in a grid-computing environment. The key outcomes of this project are as follows:

1. A two-stage workflow enabling continuous seismic waveform data extracted from the *GeoNet* archive to be pre-processed and correlated in a systematic manner;
2. Execution of this workflow within a grid processing environment, dynamically using the maximum available computing power made available by the cycle-stealing infrastructure and using, to date, up to 180 otherwise idle processors (78%) within the 230-processor ECS grid;
3. A web interface enabling the workflow to be invoked without the need for command line interaction;
4. Successful application of the workflow to datasets of various scales.

Experimentation using the Kepler and Taverna workflow implementation packages revealed difficulties in incorporating in the workflow seismic data acquisition web services developed in conjunction with this project. Both packages appear to hold significant promise for describing geophysical workflows of this sort, but further research is required to determine how to efficiently integrate web services providing high-speed access to *GeoNet* data archives in a Kepler or Taverna workflow, and execute such workflows using grid-computing resources.

Melding geophysical demands with computing infrastructure and protocols has emphasised the need for and benefits of close multidisciplinary collaboration in order to take maximum advantage of technological and scientific developments. This work has highlighted the programming challenges involved in assembling sequential workflows from programs originally developed for use in a single-user, serial processing environment. Our use of existing seismological processing codes that tacitly assume certain file structures does not facilitate a fully modular workflow, but we have identified ways in which this approach to seismic data analysis might be addressed in future research.

Glossary

Actor Terminology used in the Kepler project to refer to a single processing module

Client A networked computer that requests a service (such as data or an application) from another computer (the server)

Condor A software framework for managing computationally intensive tasks, commonly in a cycle-scavenging distributed network such as VUW's SCS grid

GPL (GNU General Public License) A software licence granting rights and obligations to users of free software

gSOAP (GPL Simple Object Access Protocol) A standard syntax for applications to call each other's methods and exchange data; also, a C and C++ software development toolkit for building interfaces to SOAP web services

KAREN (Kiwi Advanced Research and Education Network) A high-speed network connecting New Zealand universities and research institutions and providing data transfer rates of up to 10 gigabits per second

Kepler A software system for developing and operating scientific workflows

miniSEED A data-only version of SEED containing no metadata

Netbeans IDE (Integrated Development Environment) A software development framework for developing Java applications

netBSD (Berkeley Software Distribution) An open-source version of the BSD computer operating system

QuakeML (Markup Language) A modular representation of seismological data, particularly earthquake observations and location parameters

REANNZ (Research and Education Advanced Network New Zealand) The Crown-owned company set up to establish, own and operate a high-speed telecommunications network in New Zealand for the research and education sectors

SEED (Standard for the Exchange of Earthquake Data) An international standard format for the exchange of digital seismological data including waveform data and metadata such as instrument response files

Server A program or computer managing shared access to a centralized resource or service

SGE (Sun Grid Engine) An open-source batch-queuing system used to manage tasks within a distributed computing environment such as VUW's ECS grid

Taverna A software tool for designing and executing workflows, particularly in the field of computational biology/bioinformatics

Web services Network services provided via a generic interface that enables users to build there own client applications

Workflow A representation of a sequence of operations

WSDL (Web Service Description Language) An XML format for describing web-services

References

Bannister, S., Thurber, C., and Louie, J., 2006, Detailed fault structure highlighted by finely relocated aftershocks, Arthur's Pass, New Zealand: *Geophysical Research Letters*, 33

Barruol, G., Reymond, D., Fontaine, F.R., Hyvernaud, O., Maurer, V., and Maamaatuaiahutapu, K., 2006, Characterizing swells in the southern Pacific from seismic and infrasonic noise analyses: *Geophysical Journal International*, 164: 516-542

Behr, Y., Townend, J., Bannister, S., and Savage, M.K., 2009, Shear-velocity structure of the Northland Peninsula, New Zealand, inferred from ambient noise correlation: *Journal of Geophysical Research*, In preparation.

Bensen, G.D., Ritzwoller, M.H., Barmin, M.P., Levshin, A.L., Lin, F., Moschetti, M.P., Shapiro, N.M., and Yang, Y., 2007a, Processing seismic ambient noise data to obtain reliable broad-band surface wave dispersion measurements: *Geophysical Journal International*, 169: 1239-1260

—, 2007b, Processing seismic ambient noise data to obtain reliable broad-band surface wave dispersion measurements: *Geophysical Journal International* doi:10.1111/j.1365-246X.2007.03374.x.

Blewitt, G., Bock, Y., and Kouba, J., 1994, Constraining the IGS polyhedron by distributed processing, *in* Bureau, I.C., ed., IGS Analysis Workshop Proceedings: Densification of ITRF through regional GPS networks: Pasadena, Jet Propulsion Laboratory, p. 21-27.

Bonnefoy-Claudet, S., Baize, S., Bonilla, L.F., Berge-Thierry, C., Pasten, C., Campos, J., Volant, P., and Verdugo, R., 2009, Site effect evaluation in the basin of Santiago de Chile using ambient noise measurements: *Geophysical Journal International*, 176: 925-937

Brenguier, F., Shapiro, N.M., Campillo, M., Nercessian, A., and Ferrazzini, V., 2007, 3D surface wave tomography of the Piton de la Fournaise volcano using seismic noise correlations: *Geophysical Research Letters*, 34: L02305 doi:10.1029/2006GL028586.

Bromirski, P.D., and Duennebier, F.K., 2002, The near-coastal microseism spectrum: spatial and temporal wave climate relationships: *Journal of Geophysical Research B: Solid Earth*, 107: ESE 5–1 – 5–20

Brooks, L.A., Townend, J., Gerstoft, P., and Bannister, S., 2009, Analysis of ambient seismic noise fundamental and first order Rayleigh waves in the Taranaki region of New Zealand: *Geophysical Research Letters*: In prep.

Clarke, D., Townend, J., Savage, M.K., and Bannister, S., 2009, Seismicity in the Rotorua and Kawerau geothermal systems, Taupo Volcanic Zone, New Zealand, based on improved velocity models and cross-correlation measurements: *Journal of Volcanology and Geothermal Research*, 180: 50-66

Delahaye, E.J., Townend, J., Reyners, M.E., and Rogers, G., 2009, Microseismicity but no tremor accompanying slow slip in the Hikurangi subduction zone, New Zealand: *Earth and Planetary Science Letters*, 277: 21-28

Douglas, A., Beavan, J., Wallace, L., and Townend, J., 2005, Slow slip on the northern Hikurangi subduction interface, New Zealand: *Geophysical Research Letters*, 32: 1-4

Du, W.X., Thurber, C.H., and Eberhart-Phillips, D., 2004, Earthquake relocation using cross-correlation time delay estimates verified with the bispectrum method: *Bulletin of the Seismological Society of America*, 94: 856-866

Gerstoft, P., Fehler, M.C., and Sabra, K.G., 2006a, When Katrina hit California: *Geophysical Research Letters*, 33

Gerstoft, P., Sabra, K.G., Roux, P., Kuperman, W.A., and Fehler, M.C., 2006b, Green's functions extraction and surface-wave tomography from microseisms in southern California: *Geophysics*, 71: SI23-SI31

Gouédard, P., Cornou, C., and Roux, P., 2008, Phase-velocity dispersion curves and small-scale geophysics using noise correlation slantstack technique: *Geophysical Journal International*, 172: 971-981

Kao, H., and Shan, S.J., 2004, The Source-Scanning Algorithm: Mapping the distribution of seismic sources in time and space: *Geophysical Journal International*, 157: 589-594

Levshin, A.L., and Ritzwoller, M.H., 2001, Automated detection, extraction, and measurement of regional surface waves: *Pure and Applied Geophysics*, 158: 1531-1545

Lin, F.-C., Ritzwoller, M.H., Townend, J., Bannister, S., and Savage, M.K., 2007a, Ambient noise Rayleigh wave tomography of New Zealand: *Geophysical Journal International*: doi: 10.1111/j.1365–246X.2007.03414.x

Lin, F.C., Ritzwoller, M.H., and Shapiro, N.M., 2006, Is ambient noise tomography across ocean basins possible?: *Geophysical Research Letters*, 33

Lin, F.C., Ritzwoller, M.H., Townend, J., Bannister, S., and Savage, M.K., 2007b, Ambient noise Rayleigh wave tomography of New Zealand: *Geophysical Journal International*, 170: 649-666

Rawlinson, Z., Behr, Y., Bannister, S.C., and Townend, J., 2008, Ambient noise correlation across the Tasman Basin, Geosciences '08: Wellington.

Roux, P., Sabra, K.G., Gerstoft, P., Kuperman, W.A., and Fehler, M.C., 2005, P-waves from cross-correlation of seismic noise: *Geophysical Research Letters*, 32: 1–4

Sabra, K.G., Gerstoft, P., Roux, P., Kuperman, W.A., and Fehler, M.C., 2005a, Extracting time-domain Green's function estimates from ambient seismic noise: *Geophysical Research Letters*, 32: 1–5

—, 2005b, Surface wave tomography from microseisms in Southern California: *Geophysical Research Letters*, 32: 1–4

Schaff, D.P., Bokelmann, G.H.R., Ellsworth, W.L., Zanzerkia, E., Waldhauser, F., and Beroza, G.C., 2004, Optimizing correlation techniques for improved earthquake location: *Bulletin of the Seismological Society of America*, 94: 705-721

Schulte-Pelkum, V., Earle, P.S., and Vernon, F.L., 2004, Strong directivity of ocean-generated seismic noise: *Geochemistry, Geophysics, Geosystems*, 5: doi:10.1029/2003GC000520 doi:10.1029/2003GC000520.

Schwab, M., Karrenbach, M., and Claerbout, J., 1997, Making Scientific Computations Reproducible, Computing in Science and Engineering, Volume 2, p. 61-67.

Sens-Schonfelder, C., and Wegler, U., 2006, Passive image interferometry and seasonal variations of seismic velocities at Merapi Volcano, Indonesia: *Geophysical Research Letters*, 33

Shapiro, N.M., and Campillo, M., 2004a, Emergence of broadband Rayleigh waves from correlations of the ambient seismic noise: *Geophysical Research Letters*, 31: L07614 1-4

—, 2004b, Emergence of broadband Rayleigh waves from correlations of the ambient seismic noise: *Geophysical Research Letters*, 31: L07614 1–4

Shapiro, N.M., Campillo, M., Stehly, L., and Ritzwoller, M.H., 2005a, High-resolution surface-wave tomography from ambient seismic noise: *Science*, 307: 1615–1618

—, 2005b, High-resolution surface-wave tomography from ambient seismic noise: *Science*, 307: 1615-1618

Shelly, D.R., Beroza, G.C., and Ide, S., 2007, Non-volcanic tremor and low-frequency earthquake swarms: *Nature*, 446: 305-307

Shelly, D.R., Beroza, G.C., Ide, S., and Nakamula, S., 2006, Low-frequency earthquakes in Shikoku, Japan, and their relationship to episodic tremor and slip: *Nature*, 442: 188-191

Stehly, L., Campillo, M., and Shapiro, N.M., 2006, A study of the seismic noise from its long-range correlation properties: *Journal of Geophysical Research*, 111

Weaver, R.L., 2005, Information from seismic noise: *Science*, 307: 1568–1569

Wech, A.G., and Creager, K.C., 2008, Automated detection and location of Cascadia tremor: *Geophysical Research Letters*, 35

Wessel, P., 2003, Complete PostScript: an archival and exchange format for the sciences?: *EOS*, 84: 351

Yang, Y., Ritzwoller, M.H., Levshin, A.L., and Shapiro, N.M., 2007, Ambient noise Rayleigh wave tomography across Europe: *Geophysical Journal International*, 168: 259–274

Appendix A: Documentation

Invoking a cross-correlation job on the ECS grid is done by entering control parameters into a webpage interface or by specifying those control parameters directly via the command line. In each case, the job is submitted to the Sun Grid Engine (SGE) via the 'qsub' command; further details regarding this process are available online at http://ecs.victoria.ac.nz/Support/TechNoteEcsGrid.

General syntax

qsub	-t *range_of_tasks_to_be_completed*	(SGE parameter; number of required tasks and process placement information, in this case known in advance from the number of data transformations or pair-wise correlations required)
	-q *SGE_queue_instance*	(SGE parameter; job queuing)
	-wd *working_directory*	(SGE parameter; working directory)
	-M *username@domain*	(optional SGE parameter; email address for notifications related to each task)
	-m *notification*	(optional SGE parameter; explicit specification notification required)
	stage1.sh or *stage2.sh*	(SGE parameter; process to be executed)
	base_directory	(program parameter)
	input_stage_directory	(program parameter; grid-visible location from which to stage initial data)
	control_file	(program parameter; list of stations/days to be operated on)
	output_stage_directory	(program parameter; grid-visible location to which to stage output data)
	low-period_limit	(program parameter; short-period filtering limit)
	high-period_limit	(program parameter; long-period filtering limit)

The Stage 2 (pair-wise correlation) program takes an extra parameter making explicit the number of days being considered without needing to inspect the control file; this is used to determine the pair of files that an individual processor will operate on:

number_of_days_to_process	(program parameter; number of days to process)

For example, consider user 'jtownend' processing the data listed in '/control_path/example.txt' from 30 stations recorded over 90 days (implying $30 \times 90 = 2700$ data files and $30 \times 29/2 \times 90 = 39150$ separate correlations), and filtering the data between upper and lower period bounds of 5 and 100 s. If the input data are to

be staged from /input_path/data, the output data to /output_path/stage1 and /output_path/stage2 in Stages 1 and 2, respectively, whilst using a working directory /local/tmp/jtownend on each processor accessible to the user running the job, then the command line invocations for Stages 1 and 2 would be:

Stage 1

qsub -t 1-2700 -q all.q -wd /vol/grid/sgeusers/jtownend -M john.townend@vuw.ac.nz -m e stage1.sh /local/tmp/jtownend /input_path/data /control_path/example.txt /output_path/stage1 5 100

Stage 2

qsub -t 1-2700 -q all.q -wd /vol/grid/sgeusers/jtownend -M john.townend@vuw.ac.nz -m e stage2.sh /local/tmp/jtownend /output_path/stage1 /control_path/example.txt /output_path/stage2 5 100 90

Note that the *output_stage_directory* for Stage 1 here corresponds to the *input_stage_directory* for Stage 2, and the *control_file* ('example.txt') would contain a list of files similar to the following:

example.txt

2008.032.LTZ.10-LHZ.NZ.D

2008.032.MLZ.10-LHZ.NZ.D

.

.

.

2008.122.MLZ.10-LHZ.NZ.D

Note that for the purposes of illustration, the 'qsub' has been instructed to email 'jtownend' at the end of each completed task ('-m e') —this would normally not be the case for large numbers of tasks, to avoid the consequent plethora of emails.

Appendix B: Conference outputs

Behr, Y., Townend, J., Savage, M.K., and Bannister, S., 2008. New Zealand surface wave velocity maps and S-velocity profiles from ambient seismic noise correlation. Geosciences ’08, Wellington.

Buckley, K., 2009. Geoscience eResearch at Victoria University of Wellington. BeSTGRID GeoScience Strategy Group Meeting, Auckland.

Grimwood, P., Behr, Y., Townend, J., Hine, J., and Savage, M.K., 2008. New Zealand Seismographic Information Service — enhanced access to seismographic data for research and teaching. American Geophysical Union Fall Meeting, San Francisco.

Townend, J., 2008. Expanding geophysical data sets and computing requirements. Victoria eResearch Symposium, Wellington.

Townend, J., Behr, Y., Buckley, K., Savage, M.K., and Hine, J., 2009. A grid-based facility for large-scale cross-correlation of continuous seismic data. eResearch Australia, http://www.eresearch.edu.au/, 9–13 November 2009, Sydney (abstract submitted 29 June 2009; Appendix C).

Appendix C: eResearch Australasia Abstract

A Grid-Based Facility for Large-Scale Cross-Correlation of Continuous Seismic Data

John Townend[1], Yannik Behr[2], Kevin Buckley[2], Martha Savage[1], John Hine[2]

[1]School of Geography, Environment, and Earth Sciences, Victoria University of Wellington, Wellington, New Zealand, john.townend@vuw.ac.nz

[2]School of Engineering and Computer Science, Victoria University of Wellington, Wellington, New Zealand

INTRODUCTION

Just as radiographers use X-rays and ultrasound to image the internal structure of a human body, so seismologists use seismic waves generated by earthquakes or artificial sources to study the earth's interior. More than 95% of the time, however, seismometers designed to record earthquakes are actually recording continuous, low-pitched noise—the incoherent background hum of the earth. Much of this noise is produced by ocean waves hitting the coastline, and New Zealand's geographic isolation exposes it to a particularly energetic ocean. Recent studies reveal that this noise is not entirely random. By comparing long records of seismic noise recorded at two different locations, a small amount of coherent seismic energy propagating directly between them can be detected. This energy propagates as a seismic wave at a speed governed by the physical properties of the rocks it passes through. By measuring this speed, geophysicists can map Earth's deep structure in much the same way as ultrasound is used to look inside human bodies (Figure 1).

This work described here addressed two complementary goals. The first was to develop a computational workflow — a documented sequence of analytical steps — allowing automated or interactive analysis of continuous raw seismic data using grid-computing resources. The second goal was to develop an interface to the computational workflow that facilitates its use in an efficient and effective manner by researchers in the broader geophysical community.

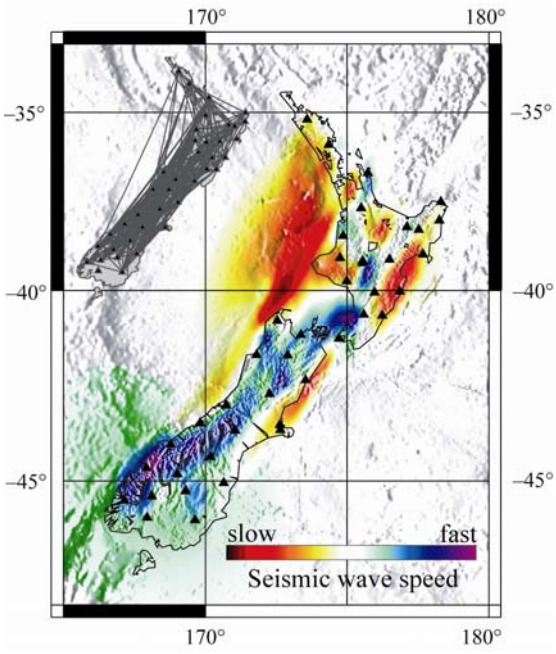

Figure 1: Map of seismic wave speeds at ~15 km depths obtained using one year of continuous seismic noise recorded by the *GeoNet* network (triangles; http://www.GeoNet.org.nz/). Red colours indicate slow speeds (e.g. Taranaki and Hikurangi basins west and east of the North Island), and blue colours indicate high speeds. The inset map shows the propagation paths analysed [1].

AMBIENT SEISMIC NOISE TOMOGRAPHY

Modern seismological networks, such as New Zealand's *GeoNet* system (Figure 1), are designed to record seismic waves generated by earthquakes, but more than 95% of the "signal" recorded by such networks is ostensibly incoherent noise

generated by the interaction of ocean waves with the seabed and coastline. It has been shown recently that cross-correlating long noise records from pairs of seismometers reveals a coherent signal corresponding to the propagation of a seismic wave from one instrument to the other [2-5]. Over distances greater than ~10 km, this signal corresponds to a wave propagating just below the earth's surface [6] at speeds determined by the earth's elastic properties. Lower-pitched energy samples greater depths: measurements at periods of 5–25 s provide images of geological structure to depths of ~30 km [4, 7]. By analysing cross-correlation measurements at different frequencies, we can therefore determine the seismic velocity structure of the earth's crust at various depths [1].

COMPUTATIONAL WORKFLOW MODELLING

As in many other data-intensive fields of science, geophysical research often involves a large number of incremental processing steps, during each of which decisions must be made regarding the particular choice of parameters or even algorithms to use. We have been working on methods of combining the increasingly routine geophysical task of cross-correlating large data sets to image the earth, with modern e-research approaches to systematising and documenting the research process itself.

The idea of encapsulating within a particular research output all of the processing parameters used in obtaining that output (a figure, table, or parameter value) from the original input (raw data) is not new. Examples within the geophysical realm include the Complete PostScript System [8], an archival and exchange format that incorporates details of the processing parameters and algorithms used to generate it within an output image; the SINEX format [Solution-INdependent Exchange Format; 9] used to exchange geodetic locations and the modelling parameters they depend on; and the Stanford Exploration Project reproducible electronic document protocol [10], which consists of rules used to reproduce entire books from the original source code and data. What each of these examples — which are often referred to as "reproducible research" — provides, to varying degrees, is the potential for researchers to ask questions of their own or others' research such as:

- Could we plot that data a different way to see things more clearly?
- What effect on the results would changing that parameter have?
- Would she have got the same result if she'd used his algorithm at that step instead of mine?

At its simplest, the implementation of a workflow is a computer program. However, when there are large amounts of data and possibly different computational resources involved the problem of tracking and documenting input parameters and processing requirements becomes more complex.

CROSS-CORRELATION OF CONTINUOUS SEISMIC DATA IN A GRID-COMPUTING ENVIRONMENT

We have developed methods of extracting continuous seismic waveform data from the *GeoNet* archives via the Kiwi Advanced Research and Education Network (KAREN) using web services, and of distributing the data pre-processing and cross-correlation tasks amongst processors operating within a Sun Grid Engine environment.

To date, our research has focussed on tailoring geophysical demands to the available computational resources and protocols. The collaborative nature of this project has highlighted how important close interaction between end-users (geophysicists) and computing specialists is in ensuring that complex problems are described and represented in the most computationally efficient way possible. We have successfully demonstrated the suitability of a two-stage processing sequence for cross-correlation jobs involving data sets of a range of sizes. Future work will address questions such as how to most efficiently allocate tasks between the processors available at any one time, and how to monitor resources and job progress.

REFERENCES

1. Lin, F.C., et al., *Ambient noise Rayleigh wave tomography of New Zealand.* Geophysical Journal International, 2007. **170**(2): p. 649-666.
2. Shapiro, N.M. and M. Campillo, *Emergence of broadband Rayleigh waves from correlations of the ambient seismic noise.* Geophysical Research Letters, 2004. **31**(7): p. L07614 1–4.
3. Sabra, K.G., et al., *Extracting time-domain Green's function estimates from ambient seismic noise.* Geophysical Research Letters, 2005. **32**(3): p. 1–5.
4. Shapiro, N.M., et al., *High-resolution surface-wave tomography from ambient seismic noise.* Science, 2005. **307**(5715): p. 1615–1618.
5. Weaver, R.L., *Information from seismic noise.* Science, 2005. **307**(5715): p. 1568–1569.
6. Roux, P., et al., *P-waves from cross-correlation of seismic noise.* Geophysical Research Letters, 2005. **32**(19): p. 1–4.
7. Sabra, K.G., et al., *Surface wave tomography from microseisms in Southern California.* Geophysical Research Letters, 2005. **32**: p. 1–4.
8. Wessel, P., *Complete PostScript: an archival and exchange format for the sciences?* EOS, 2003. **84**: p. 351.

9. Blewitt, G., Y. Bock, and J. Kouba, *Constraining the IGS polyhedron by distributed processing*, in *IGS Analysis Workshop Proceedings: Densification of ITRF through regional GPS networks*, I.C. Bureau, Editor. 1994, Jet Propulsion Laboratory: Pasadena. p. 21-27.

10. Schwab, M., M. Karrenbach, and J. Claerbout, *Making Scientific Computations Reproducible*, in *Computing in Science and Engineering*. 1997. p. 61-67.