THE UNIVERSITY
OF ADELAIDE
AUSTRALIA

Investigation into High Performance Computing Technologies for Geophysics

Tristan M. Wurst,

B.Sc

Geology and Geophysics
School of Earth and Environmental
Sciences The University of Adelaide

October 2010

Contents

List of Figures

Abstract

The processing of magnetotelluric (MT) data is typically carried out on a desktop computer and as a result suffers from a number of drawbacks. The time taken to process the data on the desktop computer is unacceptably long and can take approximately a month. The limited amount of random-access memory (RAM) in the desktop computer limits the length of the time series that can be used in the bounded influence remote referencing processing (BIRRP) program. Cloud computing is a new high performance computing (HPC) technology that can be accessed over the internet and has the potential to address the drawbacks presented by the desktop computer. Cloud computing reduces the cost of HPC by pooling computing resources on a large scale. Cloud computing offers on-demand resources allowing the user to use only what they need and to change the type of resources they require to suit an evolving need. To utilise the HPC capabilities of the cloud, a problem must exhibit a high degree of parallelisation. MT processing is particularly well suited to cloud computing because of its inherent ability to parallelise by the number of stations. To enable automatic utilisation of the cloud resources, workflow technology can be used in conjunction with the existing MT processing codes. This new approach to MT processing presents the opportunity to addresses other inefficiencies in the processing. As the cloud is accessible over the internet, this presents the opportunity to perform some processing in the field. The ability to process data in the field is advantageous because it allows for near instant feedback about the quality of the obtained data. This feedback can then be used by the survey team to change the survey to optimise the quality of the obtained data if required. However, to achieve this, a number of new processing techniques need to be introduced into the workflow.

Chapter 1: Introduction

1.1 Overview

Geophysics is one of the most computationally intensive scientific domains, utilising both complicated theoretical models and large datasets. Some of the computationally demanding applications of geophysics include seismic processing (Tulchinsky & Tulchinsky 2009), 3D geothermal simulations (Wolf *et al.* 2007), 3D simulations of the dispersal of volcanic particles (Ongaro *et al.* 2007), inversion of 3D MT data (Newman & Alumbaugh, 2000; Zyserman & Santos, 2000; Tang-pei & Qun, 2008; Lin *et al.* 2009; Siripunvaraporn & Egbert 2009), geodynamo (Glatzmaier and Roberts, 1995), seismic wave propagation (Komatitsch & Tromp, 2002a, 2000b), mantle convection (Kameyama & Yuen, 2006; Matyska & Yuen, 2005), and lithospheric dynamics (Surussavadee & Staelin, 2006). Although the above examples vary in the degrees to which they are data- and compute-intensive problems, what is common to all of them is the interest in parallel computation to reduce the time taken to compute the problem.

To overcome the limitations of serial processing and to address the problems of data-intensive computing, a number of different computing architectures have been used to implement applications that can be decomposed and performed in parallel. These include the dedicated supercomputer, grid computing, clusters of computers and, more recently, cloud computing. Considerable progress in parallel computing with these architectures has already been made but many people are still using desktop computers to do their modelling. The desktop computer is fast reaching its limitations for even mildly data-intensive applications (Szalay & Gray, 2006). This is driving the need to make the transition from the desktop computer to a parallel architecture capable of large-scale parallelisation. The newest and most exciting of these architectures is cloud computing. Cloud computing is exciting for a number of reasons, the most notable being its ability to provide on-demand computing resources. This new paradigm of high performance computing gives the user the ability to request and use resources to suit their demand. Further, the elastic nature of this service allows the user to change the resource type such as instance type, and number, with relative ease if they find that their current configuration does not suit their purpose. To allow the effective utilisation of these resources, workflow technology can be used to execute parallel computations automatically across the cloud infrastructure.

The purpose of this investigation is to assess the suitability of cloud computing, workflow technology and parallelisation to overcome the current limitations imposed by the desktop computer in the processing of MT data. Currently, MT data is processed in a sequential manner in which each station in a survey is processed in sequence. The fact that the processing of these stations occurs independently means that this problem is very amenable to parallel execution, in which all stations can be processed concurrently. Problems in the current processing methodology are threefold. Firstly, the time taken to process the data can be as long as a month and needs to be substantially reduced by using parallelisation. Secondly, the limitations of the available RAM in the desktop computer means that only approximately a third of the obtained time series can be used in obtaining the MT responses. It is hoped that by using on-demand resources that an appropriate amount of RAM can be requested to address this problem. Thirdly, the processing itself is quite laborious and fragmented consisting of many folders of data that utilises different codes at different stages, resulting in a processing methodology that takes a considerable time to share, teach and master. Workflow technology has the potential to address this problem by providing an intuitive platform for the parallel execution of workflows. Additional benefits of a workflow system are that it allows for easy sharing of the processing methodology. The major barrier to progress by utilising these technologies is that they are still in their infancy and are still evolving to suit the demands required of them. For this reason, very little work utilising these technologies has been demonstrated in the domain of geophysics. It is hoped that this pioneering work will reveal the current potential and applicability of these technologies for geophysical applications.

Chapter 2: Background

2.1 A New Paradigm

It has long been accepted that theoretical and experimental sciences have been the basic research paradigms for exploring nature (Bell *et al.* 2009). In recent decades, a third paradigm has emerged as specialisation of experimentation, and focuses on the unique opportunities provided by numerical techniques offered by computers. The relationship between the different paradigms is one of data volume. From theory, to experimentation, to computation, each level requires an increase in the amount of data consumed and produced (Nelson 2009). Science is currently facing a new crisis that is again driven by data volume. The use of sensor network, satellites, high throughput instruments and supercomputers has led to an exponential increase in data volumes as compared to only a decade ago (Bell *et al.* 2009). The desktop computer and data analysis programs such as Matlab and Excel are not capable of processing millions of data records and are primitive by most standards. This increase in data volume requires scalable processing methods to extract knowledge from the data (Szalay & Gray 2006). For data-intensive applications to be effective, they must be able to manage and process these large data volumes in an acceptable amount of time (Gorton *et al.* 2008). This data-intensive approach to discovering new knowledge has been described as a paradigm in its own right and has come to be known as the fourth paradigm.

2.2 Parallelisation

Traditionally, software has been written for serial execution in which a single central processing unit (CPU) is used to process the problem. This is achieved by decomposing the problem into a series of discrete instructions that are then executed one after another (Barney 2010). In contrast, parallel programming seeks to use multiple computer resources simultaneously, such as a computer with multiple processors or a number of computers connected by a network, or both, to solve problems. This is achieved by breaking a problem down into a number of discrete parts (Barney 2010). The discrete parts are then broken down into a series of instructions that can be solved concurrently. Lower levels of parallelisation include a shared nothing approach in which totally independent processes can be executed concurrently instead of sequentially (see Figures 2.1 and 2.2). The adoption of a parallel approach depends on the ability of the problem to be decomposed into smaller independent pieces and the ability of the parallel problem to be solved in less time than the serial version (Barney 2010). Barriers to the adoption of parallel computing by geoscientists include developing the parallel algorithms and the associated software, which is time consuming and often requires multidisciplinary expertise (Zhang *et al.* 2007).

The current interest in parallel computation stems in part from the inability of desktop CPUs to process data in a timely manner and limitations in the amount of available RAM. In the past, the usual response to this problem was to produce increasingly faster CPUs. This increase in CPU speed has grown at almost a constant rate for the last two decades; a phenomenon known as Moore’s law (Sava 2010). Moore’s law states that the number of transistors that can be placed inexpensively on an integrated circuit has doubled approximately every 18 months, which can be seen in Figure 2.3 (Ongaro *et al.*

2007). The performance of many electronic devices such as processing speed and memory capacity are strongly dependent on Moore’s law and follow this exponential rate of improvement. However, Moore’s law cannot be sustained indefinitely.

Ultimately, the physical limits of packing more and more transistors together on silicon chips will be reached as the miniaturisation of the transistors reaches its limits. However, a more immediate problem to the packing of additional transistors into one processor is one of heat dissipation. The smaller the transistors are, the faster they switch. This switching uses energy, producing power that is dissipated as heat. The transistors are located on a particular piece of dye and having more transistors on the dye results in the generation of increasing amounts of power. The amount of power that can be generated on any one processor before its performance is compromised is known as the power ceiling and can be seen in Figure 2.4. To overcome this problem, multicore architectures have been introduced that utilise multiple processors that each generates power under the 130W power ceiling, Figure 2.4. However, the multicore architecture brings with it the complication of a different programming methodology needed to allow for communication between the different processors or cores (Mudge, C. personal communication, 27 September 2010).

2.3 Cloud Computing

Cloud Computing is the newest iteration in the HPC market and was born from the culmination of years of experience with grid computing, high performance clusters and service oriented architecture fields (Simmhan *et al.* 2009). This technology is still in its infancy and as a result, its true potential for scientific research has yet to be realised. The word cloud itself is an ambiguous term and can lead to some confusion when talking about the technology. ‘The cloud’ is simply a reference to the internet that is often reduced to a depiction of a cloud as an abstraction of its role in how resources are accessed over it (Christina *et al.* 2008). The cloud computing model has large clusters of computers typically contained in data centres, the construction of which can be seen in Figure 2.5. These clusters are remotely accessed over the internet.

Further ambiguity about the nature of cloud computing is introduced because no formal definition about what it actually is has been decided on. While some informally use the term to denote any service that is accessed over the internet more formal definitions are beginning to come about, most of which agree that the basic features that define cloud computing are, a pay-per-use utility model, the potential for massive scalability, and virtualisation (Vaquero *et al.* 2009). Based on the economy of scale, cloud computing exploits the resources of massive data centres that are powered by many computer clusters composed of relatively inexpensive commodity components. The pooling of resources on this scale combined with load balancing lead to a resource optimisation, resulting in a very economically attractive model (Vaquero *et al.* 2009). The cloud computing paradigm has the potential to offer a number of services that include infrastructure as a service, which allows the user to interact with virtual servers and storage; software as a service, which allows the user to remotely access and operate applications; and platform as a service, in which the user can create and develop applications on the provider’s platform (Vaquero *et al.* 2009). Cloud computing is attractive to the scientific domain because it puts a layer between the scientist and the underlying infrastructure so they do not need to worry about managing the infrastructure or periodically upgrading a local computing infrastructure. This results in an accessible HPC technology allowing scientists to experiment and compete based on ideas and not budgets. Although cloud computing reduces some overheads, it also introduces some others. Before the resource can be utilised for computation, a network must be set up and an image must be created or discovered before any computation can take place (Christina *et al.* 2008). An image is a virtual appliance that is used to create a virtual machine. The main components of an image are a read-only file system image, which includes an operating system such as windows, and any additional software required to deliver a service. By using virtual machines, the working environment is configured independently from the underlying resource allowing for multiple environments to be deployed on the resource at the same time (Christina *et al.* 2008).

2.4 Workflow for E-science

While much effort and progress had been made in developing high performance computing technologies that exploit parallelisation what scientists really need are tools that bring the power of these resources to their desktops. Traditionally, batch files, shell scripts, and general-purpose scripting languages such as Python have been used for tool integration. Scripting languages are used to complete tasks such as specifying the data and software to be used and coordinating the assignment and movement of data across locations (McPhillips *et al.* 2009). Although scripted applications have the ability to manage and control computations, they also suffer from some major drawbacks. To update an existing script by adding a new code or by updating the version of an existing code can prove to be error prone and costly. This is because to achieve this you have to manually scan through the scripts to make changes (Gil 2009). More importantly, scripting languages are programming languages and fall outside the expertise of many domain scientists who need to concentrate on their research and not computation (Gil 2009).

The concept of the workflow has emerged in recent years as a challenger to traditional approaches to automating computational tasks. A scientific workflow acts like a specialised script and orchestrates the execution of a multi-step process (Deelman *et al.* 2009). A typical workflow will be composed of a number of different processes that are linked by their dependencies to one another. For example, a workflow may be a data analysis protocol that consists of a sequence of calling data from a database, pre-processing, submitting a job to a cloud computer, and post-processing steps (Goble & Roure 2009). Workflow systems vary in many respects such as what resources they use, how control flow is handled, how interactive they are and how tasks are allocated to resources (De Roure *et al.* 2009). However, the common goals and characteristics that differentiate them with respect to traditional tool integration approaches based on scripting languages are many. Firstly, scientific workflow systems are based on dataflow languages in which workflows are represented in directed graphs. The nodes in the graph represent computational stages and the pipeline between them represents data flow and dependencies between the nodes as can be seen in Figure 2.6 (McPhillips *et al.* 2009). Secondly, many workflow systems use a graphical interface to allow visual authoring of a workflow. This is especially advantageous because it allows scientists, who may only have a basic understanding of programming concepts, to compose workflows (McPhillips *et al.* 2009). Thirdly, the dataflow programming language combined with the visual authoring mechanic allows data that is produced at one node to be easily routed downstream to many nodes. This results in workflows being more declarative about the interactions between the nodes, something that scripting languages have difficulty in achieving due to the flow between components being hard to visualise in the text of a complex code (McPhillips *et al.* 2009). Other advantages workflows have over traditional scripting processes are that workflows can automatically record and process provenance and allow for easy concurrent execution of workflow tasks (McPhillips *et al.* 2009).

Data provenance involves recording the creation history of a data object created by a workflow. The importance of provenance capture is something that should not be underestimated. Provenance capture allows for the possibility of reproducibility, which is at the heart of the scientific method (Deelman *et al.* 2009). The immediate benefits of provenance are obvious to the scientist who may want to collect enough information about the workflow to be published in research papers. However, it also allows the readers of the research to run the experiment to interrogate the results (Araujo *et al.* 2009). Further, provenance allows the scientist to go back to dated research and re-run experiments performed long ago with new parameters, inputs or configurations to observe if there are any changes in the results (Araujo *et al.* 2009). A higher degree of reproducibility on this level is advantageous because it gives more people the ability to run the experiment ensuring that errors are found and addressed faster (Araujo *et al.* 2009).

The ability and importance of sharing workflows goes well beyond attaching workflows to research papers for critiquing. The myExperiment website has emerged as a social website that is based on the web 2.0 approach of social networking sites such as MySpace and Facebook. However, the big difference between these sites and myExperiment is that it is a social website for scientists, specifically designed around the sharing of workflows (De Roure *et al.* 2009). Sharing workflows in this way has many advantages. The most obvious advantage is for the ability of reuse. Workflows capture pieces of scientific processes and know-how that is often tacit and otherwise hard to share. The myExperiment website makes it possible to share this knowledge with other scientists resulting in an acceleration in the time taken for a new scientist to perform experiments. Further, workflows can be branched, and workflow patterns and fragments can be reused to quickly adapt to new applications within and even outside their original purpose and domain (De Roure *et al.* 2009).

2.5 The Trident Workbench

Very recently, Microsoft weighed into the workflow paradigm with the release of Project Trident. Project Trident gets its name from the Ocean Observatories Institute Project, which was formerly known as NEPTUNE. The overriding goal of this project is to turn oceanography from a data/knowledge poor science to a data/knowledge rich science by creating the first plate scale observatory, which can be seen in Figure 2.7. To do this, the project will deploy approximately 2,000 km of fibre optic cable along the seafloor to which a range of chemical, geological and biological sensors will be attached. It is hoped that gathering this data will allow for a better understanding of issues such as the ability of the ocean to absorb greenhouse gases and how stresses on the seafloor cause earthquakes and tsunamis (Barga *et al.* 2008b; Knies 2009; Simmhan *et al.* 2009). However, the collection of data on this scale brings with it the problem of how to manage and process the data to gain knowledge from it. From this, Project Trident was conceived and would serve as a tool for scientists to understand oceanography.

The potential for applications of the workbench outside of oceanography was quickly recognised. A collaboration with Johns Hopkins University was undertaken to develop an astronomer’s workbench for the Panoramic Survey Telescope and Rapid Response System (Knies 2009). The goal of this project was to identify objects that were approaching Earth such as asteroids and comets (Barga *et al.* 2008a; Knies 2009). This project helped in developing the versatility of the workbench, which is best described by Roger Barga, who led the Advanced Research Services and Tools team that was responsible for developing the Trident Scientific Workflow Workbench:

> If you design a system with two or three different customers in mind, you generalize very well. You come up with a very general architecture. One of the

> challenges we had to overcome was to not specialize on just one domain, or it would be too specialized a solution. Pick two or three, and balance the requirements so you build a general, extensible framework. We think we've done that (Knies 2009).

2.6 MT Concepts

The physical property that MT measures is resistivity. Resistivity is a measure of how well rocks conduct electricity and varies over seven orders of magnitude in the Earth. This property can be used to separate the Earth into zones based on conductivity differences in the rocks. For example, ground water, minerals, and molten rock conduct electricity very well and have low resistivities. Water in sedimentary rocks is contained between the grains and has a larger range of conductivities. Drier crystalline rocks have high resistivities. Therefore, this property can be used to separate the Earth into zones based on conductivity differences in the rocks (Heinson, G. Personal communication, 12 July 2010).

The method itself is an electromagnetic method that uses time varying electric and magnetic fields that propagate into the earth to determine the resistivity structure of the earth. The bandwidth of MT ranges from 10,000Hz to ten thousandths of a hertz. High frequencies are used to investigate relatively shallow anomalies in the upper crust and can be used to explore for minerals, ground water, geothermal anomalies, and oil. In contrast, lower frequencies are used to investigate deeper into the Earth to the middle crust and down into the upper mantle approximately 50 km beneath our feet. MT can also be used to investigate the mantle transition zone, at about 400km, and even right down into the core (see Figure 2.8) (Heinson, G. Personal communication, 12 July 2010).

Unlike other electromagnetic methods such as GPR, which uses a human-supplied signal source known as a transmitter to provide the signal, MT is a passive technique and uses naturally occurring phenomenon as a signal source. For the bandwidth from 1,000Hz to a few hertz, the signal is due to lightning strikes occurring around the equatorial band. At lower frequencies, magnetic storms cause a squashing of the magnetic field lines resulting in a change in the magnetic field strength. At the lowest frequencies, the movement of the conductive ocean water through the Earth's core magnetic field can produce a signal (see Figure 2.8) (Heinson, G. Personal communication, 12 July 2010).

In the field, electrodes are put out that are typically pot electrodes containing an electrolytic solution. These electrodes are typically connected by wires to a voltmeter. The electrodes measure the natural voltage difference in the ground between two points. Two electrodes are deployed at each site and they are located orthogonal to each other, with one measuring a voltage gradient between two pot electrodes in the north-south orientation, and the other measures the voltage gradient in the east-west orientation. These measurements are measured with time because there is a time dependent change in all of these parameters. The electric field is not typically measured in the vertical direction because the air is a resistor, resulting in all electric current flowing parallel to the surface. A magnetic sensor is used to measure the magnetic field in three orthogonal directions typically one in the east, one to the north, and one vertically. The type of magnetic sensor used will depend on the purpose of the survey. The broadband method, which measures relatively shallow responses, uses induction coils. For the long period method, which measures relatively deep responses, a fluxgate magnetometer is used (Heinson, G. Personal communication, 12 July 2010).

The magnetic field is considered the input signal into the earth and causes the electric currents to flow in the earth by induction. The electric components Ex and Ey are considered the responses or outputs of the time varying magnetic field interacting with the earth. Linking these two is a two by two tensor, which represents the filtering effect that the earth applies to the signal. If the earth were a perfect resistor, we would have no current flowing so this effect would be small. If it is a good conductor then a great deal of current flows and the effect of the earth on the signal is large. By solving for how the earth acts as a filter on the signal the resistivity structure of the earth can be deduced, because the filtering effect is directly related to resistivity (Heinson, G. Personal communication, 12 July 2010).

2.7 MT Processing

To obtain a resistivity model of the earth, a number of different manipulations are applied to the raw field data. Figure 2.9 shows a general flowchart that outlines these manipulations applied to one station. This a general flowchart only because the order that these manipulations may be carried out can vary from processor to processor, who may use different programs to perform the manipulations. Note this process is repeated for every station in the survey.

1. The first step in the processing stream is to transfer the field data from the logger to a computer. The data is contained in a folder that is named as a day number of the year. This folder contains the time series of the various components of the

magnetic and electric fields. Usually, at least two components are always measured for the electric field while at least two and up to three may be recorded for the magnetic field. Other files that are contained within the folder include a GPS file and an ambient temperature file. All files in this folder are generically named as station name, year, month, day, hour, minutes and seconds. This folder will typically contain many files because the loggers are only capable of recording the files at ten-minute intervals. Therefore, every ten minutes, a new group of ten-minute long files are recorded. Since the loggers can be recording for days at a time, this will result in the folder containing many ten-minute files. Although the amount of data collected will vary from survey to survey, the station folder typically contains around 1.5GB of data (Thiel, S. Personal communication, 19 July 2010).

2. The second step in the processing stream is to erase spurious files from the day folder. A typical ten-minute time series file with a sampling rate of 500Hz will contain approximately 300000 values. Files that have large discrepancies from this value will need to be deleted. Spurious files typically occur at the start and the end of the recording and created when the logger is recording during the field setup (Thiel, S. Personal communication, 19 July 2010).
3. In the third step, the discontinuous ten-minute time series files are merged into one coherent file resulting in one Ex, Ey, Bx and By file (Thiel, S. Personal communication, 19 July 2010).
4. After the time series is merged, it is plotted and viewed to determine if there is any obvious noise present (Thiel, S. Personal communication, 19 July 2010).
5. This step involves doing a unit conversion but depends on the type of survey being done. If a broadband survey is being done, then the unit conversion can only be done on the electric field. This is because in a broadband survey the

response of the induction coils is only known in the frequency domain so the magnetic field conversion must be performed in the frequency domain. For a long period survey, this is not an issue and so both the electric and magnetic field conversion are done here (Thiel, S. Personal communication, 19 July 2010).

6. The time series of the electric and magnetic fields is converted to the frequency domain using the BIRRP program. The third and fourth Fourier coefficients are calculated for 12 decimations of the time series producing 24 Fourier coefficients that are essentially the MT responses (Thiel, S. Personal communication, 19 July 2010).
7. The BIRRP output files are reformatted into the .edi, .dat, .coh, .imp formats (Thiel, S. Personal communication, 19 July 2010).
8. In this step, all of the MT responses from every station are used. The phase tensor analysis provides information about the dimensionality and strike of the site and is needed in the inversion package to rotate the data to the right strike and to determine if the site is suitable for an inversion (Thiel, S. Personal communication, 19 July 2010).
9. The inversion takes the MT responses and produces a resistivity model of the earth. This is an iterative process and can take 10–20 iterations for a 3D inversion or hundreds of iterations for a 2D inversion. However, a single 3D iteration takes much longer than a single 2D iteration resulting in the 3D inversion process taking much more time overall to execute (Thiel, S. Personal communication, 19 July 2010).

Chapter 3: Approach

3.1 Approach

The approach to using a HPC technology to process MT data involved a number of discrete steps. The aim of the investigation is to explore the feasibility of taking an existing processing code to see if it lends itself to concurrent execution. This could be achieved by simultaneously processing a number of stations in parallel that would otherwise occur sequentially. Using existing processing codes would be advantageous because it will allow for a large speed up with a relatively small energy investment. To achieve this, all of the available processing methodologies were catalogued and assessed for their suitability to be used in a workflow and cloud environment. Secondly, familiarity with the Trident workflow system was developed to determine how transfer and execute programs in this environment. Additionally, familiarity with how to construct and execute workflows was developed. Thirdly, a cloud infrastructure was chosen based on ease of use and economics.

3.2 Applications of Parallel Processing to MT Data

The problems with processing MT data are primarily caused by data overflow. In the field, a variable number of stations are deployed. The exact number of stations deployed will depend on a number of factors such as the purpose of the survey, the nature of the terrain, and the size of the field team. A typical survey may consist of 50 stations. Each station records orthogonal components of both the electric field and magnetic field as a time series. The time series is composed of a number of samples of the magnetic and electric fields. The number of samples recorded as a time series will depend on the amount of time the instruments are left in the field and the frequency at which they are sampling. A typical broadband survey will sample the fields at 500Hz and will collect data for approximately 24hrs. This will result in around 50 million samples being collected for one component, of one field, at one station. It is the analysis of this large number of samples that is primarily responsible for the inconveniently large amount of time required to process the data.

The time series data need to have a number of manipulations performed on them as outlined in section 2.7, steps 1–7. The primary purpose of these manipulations is to obtain the MT responses in an edi format. The edi format is a format specified by the MT community and serves as a standardisation of results it is also the format of data required to perform the inversion. The diagram in Figure 3.1 shows how the data from each site is typically processed in a sequential manner using one desktop computer. The time taken to process each site using the desktop computer typically takes four hours for a broadband survey resulting in the processing from the raw time series from the logger to the EDI file taking $4\text{hrs} \times \text{N sites}$. For 50 sites, this equates to $4\text{hrs} \times 50\text{ sites} = 200\text{hrs}$. Assuming the processor is working an eight-hour day this easily equates to a month of work.

From Figure 3.1, it can be seen that the time series from each site is processed independently from each other, with the time series data from one site producing one EDI file for each site. As there is no interdependence between the sites, this immediately lends itself to a simple shared nothing type of parallelisation. Figure 3.2 shows the conceptual framework of a shared nothing parallelisation approach to the processing of MT data on different nodes. In this framework, instead of the time series being processed sequentially, one after the other, they are processed at the same time. In this scheme, the nodes are analogous to different desktop computers all running at the same time. The immediate benefit of this is that instead of the execution time being the number of sites multiplied by the time taken to process one site, it becomes the time taken to process on site because they are all being processed concurrently.

To assess the suitability of MT processing for concurrent execution, the current processing methodologies were catalogued. The purpose of this was to try to find a processing methodology that would best suit the ultimate goal of automated concurrent execution. The advantage of using an existing processors code base to achieve concurrent execution is that it allows for a relatively large speed-up, with a relatively small energy investment because the processing algorithms do not have to be written from scratch. Figure 3.3 shows an overview of the MT processing performed by the different processors. Essentially, they all manipulate the data in the same way as outlined in section 2.7, steps 1–7. Noticeable differences between them include the degree to which they are automated and the different code bases that they utilise. Immediately, it appears that processor 1 has developed the most attractive methodology for concurrent execution. The reasons for this is that it is the most automated and is also capable of automatically detecting and processing both broadband and long period data. This is in contrast to the other MT processing methodologies that only process one or the other, which can be seen in Figure 3.3.

A more detailed and accurate diagram of processor one's processing methodology can be seen in Figure 3.4. From Figure 3.4, it can be seen that the processing methodology consists of two discrete Python programs. The first program is called DataPrep and the second program is called BIRRPInterface. The automation of the second program is driven by references to a spreadsheet, containing field data, which enables the automatic creation of header and script files. However, the process is not fully automatic because to obtain the length of the time series to be analysed the time series must first be plotted. This is necessary because perturbations that occur in the time series must be accounted for before further processing and to date the most effective way to determine this is by visual inspection of the time series.

The Amazon Elastic Compute Cloud was chosen as the cloud infrastructure to develop with. It was chosen because it provides the most comprehensive service, and is easy to use and access. The Amazon EC2 web service offers a number of services that are too numerous to list here. Table 3.1 shows some of the on-demand instances that are available. To process the data for each site, each site will require an instance type. Table 3.1 shows some of the different instance types and their cost. Ultimately using concurrent execution the processing will take approximately 4 hours. For a survey consisting of 50 sites, this could equate to a cost of $\$0.12 \times 4\text{hrs} \times 50\text{sites} = \$24$ using the small instance type. However, one of the problems with the current processing methodology is the inability of the desktop computer to process all of the time series due to the limited amount of available RAM. Clearly as the desktop computer possesses 2GB of RAM, then the small instance type is not going to be an efficient final solution. However, the small instance type may be economically attractive in performing initial test experiments. To process the entire time series approximately 48 GB would be required. From Table 3.1 the only instance type that can meet this stipulation is the quadruple extra large instance type. Using this instance type, the cost of processing a 50-site survey would become $\$2.48 \times 4\text{hrs} \times 50\text{sites} = \$496$. This represents a significant increase of a factor, of approximately 20, as compared to using the small instance type. However, this cost is still only small fraction of the cost of the entire survey. Possible ways of reducing cost include exploring other instance types such as the double extra large or extra large types that may produce data of comparable quality to the quadruple extra large instance type, but at a fraction of the cost.

To enable concurrent execution of this processing, methodology on a cloud infrastructure it is proposed that a workflow system is used. The advantages of using a workflow system are that it allows for an easy parallel execution environment, further enables automation and results in a methodology that is easily shared. Figure 3.5 demonstrates how this processing methodology could be executed using the Trident workbench in a sequential manner. This could be achieved by wrapping the Python codes and using Visual Studio 2010 to generate the programs as activities. This method is advantageous because allows the current codes to be used with little rewrite. However, some code around the activity needs to be written so that the station information required by the program to run automatically can be inputted. Once this has been achieved, a parallel activity can then be used to execute the processing of any number of sites at the same time. Figure 3.6 shows the potential for concurrent execution of two sites at the same time, which can easily be extrapolated to any number of sites. The major disadvantage with this approach is that because it relies on calling the BIRRP program and initiation files from a folder, then such a folder would need to be set up on every instance required to process a station. This approach could lead to potential inconveniences and inefficiencies if for any reason parameters within the initiation files or BIRRP program need to be altered. Due to the flexibility of the workflow technology, there are a number of different ways that the same goal can be obtained. Instead of using the approach above, a more streamlined approach may involve making actors out off each program. For example, the BIRRP would be an actor in the workflow therefore eliminating the need to call it from a folder. Further advantages of breaking up the processing methodology into as many discrete actors as possible include the ability to slot new actors that may be required in the future such as a filter, between existing actors. This may not be possible using the BIRRP interface program as an actor because it does many discrete processes. Another advantage of this method is that by having more discrete actors with a specific purpose the code behind them becomes easier to understand and change. Disadvantages of this approach include a larger energy investment in creating more discrete actors, integrating automation in the way of automatic generation of script and header files and long linear segments of the workflow.

Chapter 4: Discussion

Although this provides a good mechanism for concurrent execution of the MT processing, it does not address all of the problems. This is because the MT processing methodology is still in a constant state of improvement driven largely by new tools and necessities. The current necessity driving the evolution of the MT processing is the need to assess data quality in the field. The need to check data quality in the field is something that should not be underestimated. Conducting surveys is expensive and time consuming and nothing could be more inconvenient than coming back from a survey only to find that the data is not of the required quality. This could happen for a number of reasons, such as a nearby generator contaminating the electrical signal with 50Hz noise. The ability to check the quality of data in the field means that these problems can be identified and if possible, the survey can be redesigned in near real time to address the problem. Further, checking of the quality of the data in the field means that the survey can be redesigned to optimally accommodate new information obtained about the sub-surface geology. Again, the new tool that has the potential to address this necessity is cloud computing. By accessing a satellite from the field, the site data can be uploaded to a data centre for processing much like that outlined in Figure 3.6. However, to concurrently execute a workflow on the cloud in the field, an additional number of caveats need to be introduced to make such a system viable. First and most pressingly, the bandwidth limitations mean that the site data needs to be reduced to allow for uploading the data in a time efficient manner. Currently, the time series data is stored as ASCII. This is a very inefficient way to store the data. To overcome this limitation, an algorithm to compress the time series would need to be developed to reduce the volume of data before it is uploaded. In addition, another algorithm would need to be developed to uncompress the data before further processing in the cloud.

Due to the limited time available for processing in the field, further automation needs to be integrated into the workflow to detect automatically the length of the usable time series and account for possible noise in the data. Figure 4.1 shows a sample of the time series collected for orthogonal components of the magnetic and electric fields for 300 seconds. The sharp spikes represent lightning strikes from around the world, with the larger spikes representing lightning strikes that are closer to the site while the smaller spikes represent more distant strikes. The longer periods in the data are due to magnetospheric activity. This is an example of good data because it has a continuity that is expected from MT data. Figures 4.2 and 4.3 show the power spectra of the time series for the By and Ex components of the time series shown in Figure 4.1. The power spectra is the Fourier transform of the time series and in the diagrams shown is for ten minutes of data. The power spectrum goes from zero frequency to the nyquist frequency of 250Hz. The power spectrum shows the variation in signal strength with change in frequency. Figures 4.2 and 4.3 show a smooth, continuous spectra, which means that the energy is fairly evenly distributed across the frequencie range. At 200Hz the roll off starts to occur and is due to the sensitivity of the induction coil. The small peaks that begin at around 8Hz are due to the Schumann resonance. The Schumann resonance is a cavity resonance and is related to the thickness of the ionosphere. The source of the resonance is a lightning strike at the equator that then causes the signal to propagate around the earth as a standing wave. The following peaks at around 16Hz, 24Hz and 32Hz are harmonics. Figures 4.4 and 4.5 show the power spectrum map of By and Ex. The power spectrum map shows the intensity of the power spectrum in colours. This plot differs from Figures 4.2 and 4.3 by showing the power spectrum at each time, which is denoted on the y-axis. Again, the Schumann resonances and harmonics are clearly visible. The roll off is also clearly visible as the graduation from yellow to blue.

From Figures 4.4 and 4.5, it is clear by the uniformity of the image with time that there is little to no change in the power spectrum with time. This is an example of good MT data.

An example of a time series that has been affected by noise is shown in Figure 4.6 and displays the By and Ex channels again. The window length here is much shorter than that of Figure 4.1 and is only 0.3 seconds long. In Figure 4.7, the time series is dominated by a 50Hz signal from powerlines. Figures 4.7 and 4.8 show the power map of the 50Hz affected time series. From these power maps, it is obvious that the spectrum is dominated by spectral lines at 50Hz and the 3rd 5th and 7th harmonics. This noise is described as stationary noise because the noise remains at the same frequency with time, which can be seen in Figures 4.7 and 4.8.

In Figure 4.9, another time series is shown. From 0 to 225 seconds, the signal has lost the typical continuity expected from good MT data. The reason for this is that for this period a transmitter was active. From 225 seconds onwards the transmitter was turned off and the signal returns to that expected for MT data. For the time from 0 to 225 seconds, the noise is so bad that this data should be excluded from further processing. Although the longer periods are still faintly visible in this window, to get them out would require a great effort. Figures 4.10 and 4.11 show the power map of this time series, which contains noise past the time interval, expected from that expected in the time series. From this, we may conclude that the data is also influenced by cultural noise. Figure 4.10 shows the By component and as can be seen from the power map contains both stationary and non-stationary noise. Figure 4.11 shows the Ex component, which only shows stationary noise.

To enable automatic detection of the useable length of the time series, an algorithm could be developed that takes into account large variations from that expected for good MT responses. For example, Figure 4.9 shows noise for approximately 0–225 seconds that has a much larger variation and magnitude than expected for good MT data. A further stipulation that should be built into the algorithm would be to take into account if there is little or no variation between a certain number of data points. This can happen in the field if one of the channels is disconnected from the interface box. A final stipulation that might be included in such an algorithm would be to detect if there is enough time series present from which to obtain good MT responses.

Pre-BIRRP quality control should include an inspection of the power spectrum as explained above. If the power spectrum is clean, as shown in Figures 4.3, 4.4 and 4.5, then the data can move straight on to being processed by BIRRP. However, if the data is affected by electrical noise, such as the data in Figures 4.6, 4.7 and 4.8, then a filter maybe applied to it to eliminate the noise before passing the time series to the BIRRP program. Further, this process could be made automatic. Electrical noise occurs at 50Hz and has predictable harmonics. Figures 4.2, 4.3, 4.4 and 4.5 show the power spectrum for good MT data, which has a good consistency. Deviations from this consistency at frequencies where electrical noise is known to occur can then be attributed to electrical noise. Using this information, an algorithm could be developed that compares the intensity of the power spectrum at the frequencies responsible for electrical noise with all other frequencies to check for consistency. If large deviations from the expected consistency are found at 50Hz, then this data should be passed to a 50Hz filter, before being passed to BIRRP.

Another scenario is that the data has non-stationary noise such as that shown in Figure 4.10. To deal with this scenario, a further stipulation would need to be built into the algorithm discussed to detect electrical noise. The stipulation would simply detect discontinuities in the intensity of the power spectrum with time and frequencies. If the discontinuity is present in many frequencies, then it can be flagged as non-stationary noise. If non-stationary noise is detected in the workflow, then this information must be returned to the field team. The field team may then want to redo the site or change its location to try to avoid the noise. This measure is necessary because non-stationary noise cannot be easily filtered.

Post-BIRRP quality control measures may include analysis of plots of coherence. Plots of coherence show how well the electric channel is predicted by the magnetic channel. This is possible because the electric fields are caused by induction. Plots showing relatively high coherence, coherences of above 0.8, are considered to result from good MT data. Figure 4.12 shows mostly high coherence with different periods and represents good MT data. This contrasts to Figure 4.13, which shows some good coherence for the high frequencies but poor coherences for lower frequencies. As the coherence is a measure of how good the data is, this may also be used as a source of quality control. In addition, plots of coherence may be used to alter a survey to improve data quality. In Figure 4.12, the data displays high coherences for the short and intermediate wavelengths but at a periodicity of approximately 50 seconds, the coherences start to decrease dramatically. If analysis of the longer periods is essential in the survey, then the field crew may decide to leave the instruments out for longer to increase the coherence values at longer periods.

Another area in which near instant feedback in the field would be beneficial is dimensionality analysis. Dimensionality analysis of a site tells the geophysicist about the dimensionality of the sub-surface geology. Having access to this information in the field is advantageous because it would allow the geophysicist to redesign the survey to highlight areas of interest or deploy more sites over an area of interest if need be. A recent Fortran code that was developed to determine the dimensionality of MT data is called WALDIM. Figure 4.14 shows an overview for how the WALDIM application works. Integration of such an application into a workflow would not be difficult to achieve and would be a valuable tool to have in the field.

Chapter 5: Conclusion

Despite increases in the performance of desktop computers, in accordance with Moore's law, the growing amount of data needing to be processed in domains such as geophysics is facilitating the transition to HPC technologies. It has been shown that the processing of MT data lends itself particularly well to parallelisation and consequently is ideally suited to the cloud computing paradigm. Further, the level of parallelisation addressed by the MT processing is well suited to workflow technology, which allows for the easy execution of the processing methodology across the cloud. However, to fully utilise this approach, a number of changes and additions to the MT processing methodology need to be made so the full potential of the technology can be realised. From this research, it is not difficult to conceive that this approach would be widely applicable to other geophysical applications that have an inherent ability to be parallelised. For example, large-scale gravity and magnetic maps are composed of a composite of smaller maps. It is not difficult to imagine that the creation of the smaller maps could be concurrently executed in the cloud analogous to the way the MT sites were. In general, the cloud computing paradigm is attractive for a number of reasons. Firstly, the pooling of resources on this scale result in a HPC technology that is affordable for everyone. Secondly, the on-demand nature of these resources means that the HPC resources can be accessed, terminated or changed to suit the changing needs of the user. Finally, the cloud computing paradigm furthers the current trend of accessing services over the internet, resulting in a service that is easily accessed and used. From the geophysicist's perspective, this is especially advantageous because it means that for the first time, HPC technology can be accessed in the field. This is especially important because it has the potential for field surveys to be redesigned in almost real time to accommodate the incoming data. Disadvantages of the technology include its newness and the small amount of similar work that has been conducted with cloud computing. However, these disadvantages are true of all new technologies that require pioneers to reveal their potential. The biggest disadvantage of the use of this technology lies within bandwidth considerations. Unfortunately, by using this technology to address data-intensive applications, the limiting factor becomes how fast you can transfer information across the internet. For applications such as the MT processing, it is not difficult to conceive ways of compressing the data so that this is achievable. However, the applicability of this approach to other geophysical applications remains uncertain, and subject to further research.

References

BARGA R., JACKSON J., ARAUJO N., GUO D., NITIN, GAUTAM & SIMMHAN Y. 2008a. The Trident scientific workflow workbench. IEEE International Conference on eScience (unpubl.).

BARGA R. S., JACKSON J., ARAUJO N., GUO D., GAUTAM N., GROCHOW K. & LAZOWSKA E. 2008b. Trident: Scientific workflow workbench for oceanography. *IEEE Congress on Services*, pp. 465–466.

BARNEY B. 2010. Introduction to parallel computing <https://computing.llnl.gov/tutorials/parallel_comp/>. (retrieved 14/08/2010).

BELL G., HEY T. & SZALAY A. 2009. Beyond the data deluge. *Science* **323**, 1297–1298.

CHINTHAKA E., BARGA R., PLALE B. & ARAUJO N. 2009. Workflow evolution: Tracing workflows through time.
<http://research.microsoft.com/apps/pubs/default.aspx?id=119074> (retrieved 25/08/2010)

CHRISTINA, MEHTA G., FREEMAN T., DEELMAN E., KEAHEY K., BERRIMAN B. & GOOD J. 2008. On the use of cloud computing for scientific workflows. *ESCIENCE '08: Proceedings of the 2008 Fourth IEEE International Conference on eScience,* Washington, DC, USA, pp. 640–645. IEEE Computer Society.

DE ROURE D., GOBLE C. & STEVENS R. 2009. The design and realisation of the (my)Experiment Virtual Research Environment for social sharing of workflows. *Future Generation Computer Systems—the International Journal of Grid Computing—Theory Methods and Applications* **25**, 561–567.

DEELMAN E., GANNON D., SHIELDS M. & TAYLOR I. 2009. Workflows and e-Science: An overview of workflow system features and capabilities. *Future Generation Computer Systems—the International Journal of Grid Computing—Theory Methods and Applications* **25**, 528–540.

GIL Y. 2009. From data to knowledge to discoveries: Artificial intelligence and scientific workflows. *Scientific Programming* **17**, 231–246.

GLATZMAIER G. A. & ROBERTS P. H. 1995. A 3-dimensional convective dynamo solution with rotating and finitely conducting inner-core and mantle. *Physics of the Earth and Planetary Interiors* **91**, 63–75.

GOBLE C. & ROURE D. D. 2009. *The impact of workflow tools on data-centric research* (The Fourth Paradigm Data-Intensive Scientific Discovery). Microsoft Research, Redmond, Washington.

GORTON I., GREENFIELD P., SZALAY A. & WILLIAMS R. 2008. Data-intensive computing in the 21st Century. *IEEE Computer* **41**, 30–32.

KAMEYAMA M. & YUEN D. A. 2006. 3-D convection studies on the thermal state in the lower mantle with post-perovskite phase transition. *Geophysical Research Letters* **33**, 12.

KNIES R. 2009. Project Trident: Navigating a sea of data <http://research.microsoft.com/en-us/news/features/projecttrident-070909.aspx#rmcTop>. (retrieved 25/08/2010).

KOMATITSCH D. & TROMP J. 2002a. Spectral-element simulations of global seismic wave propagation—I. Validation. *Geophysical Journal International* **149**, 390–412.

KOMATITSCH D. & TROMP J. 2002b. Spectral-element simulations of global seismic wave propagation—II. Three-dimensional models, oceans, rotation and self-gravitation. *Geophysical Journal International* **150**, 303–318.

LIN C. H., TAN H. D. & TONG T. 2009. Parallel rapid relaxation inversion of 3D magnetotelluric data. *Applied Geophysics* **6**, 77–83.

MARTI A., QUERALT P. & LEDO J. 2009. WALDIM: A code for the dimensionality analysis of magnetotelluric data using the rotational invariants of the magnetotelluric tensor. *Computers & Geosciences* **35**, 2295–2303.

MATYSKA C. & YUEN D. A. 2005. The importance of radiative heat transfer on superplumes in the lower mantle with the new postperovskite phase change. *Earth and Planetary Science Letters* **234**, 71–81.

McPhillips T., Bowers S., Zinn D. & Ludascher B. 2009. Scientific workflow design for mere mortals. *Future Generation Computer Systems—the International Journal of Grid Computing—Theory Methods and Applications* **25**, 541–551.

Nelson M. L. 2009. Data-driven science: A new paradigm? *EDUCAUSE* **44**, 6–7.

Newman G. A. & Alumbaugh D. L. 2000. Three-dimensional magnetotelluric inversion using non-linear conjugate gradients. *Geophysical Journal International* **140**, 410–424.

Ongaro T. E., Cavazzoni C., Erbacci G., Neri A. & Salvetti M. V. 2007. A parallel multiphase flow code for the 3D simulation of explosive volcanic eruptions. *Parallel Computing* **33**, 541–560.

Sava P. 2010. Introduction to this special section: High-performance computing. *The Leading Edge* **29**, 42–43.

Simmhan Y., Barga R., Ingen C. V., Lazowska E. & Szalay A. 2009. Building the Trident scientific workflow workbench for data management in the cloud. Third International Conference on Advanced Engineering Computing and Applications in Sciences (unpubl.).

Siripunvaraporn W. & Egbert G. 2009. WSINV3DMT: Vertical magnetic field transfer function inversion and parallel implementation. *Physics of the Earth and Planetary Interiors* **173**, 317–329.

SURUSSAVADEE C. & STAELIN D. H. 2006. Comparison of AMSU millimeter-wave. satellite observations, MM5/TBSCAT predicted radiances, and electromagnetic models for hydrometeors. *IEEE Transactions on Geoscience & Remote Sensing* **44**, 2667–2678.

SZALAY A. & GRAY J. 2006. Science in an exponential world. *Nature* **440**, 413–414.

TANG-PEI C. & QUN W. 2008. A distributed parallel algorithm for magnetotelluric forward modeling. *2008 International Conference on Computer Science and Software Engineering* **3**, 295–298.

TULCHINSKY V. G. & TULCHINSKY P. G. 2009. Application of SCIT supercomputers to develop and execute parallel geophysical programs. *Cybernetics and Systems Analysis* **45**, 902–914.

VAQUERO L. M., RODERO-MERINO L., CACERES J. & LINDNER M. 2009. A break in the clouds: Towards a cloud definition. *Computer Communication Review* **39**, 50–55.

WOLF A., RATH V. & BUCKER M. H. 2007. Parallelisation of a geothermal simulation package: A Case study on four multicore architectures. *Parallel Computing: Architectures, Algorithms and Applications* **38**, 451–458.

ZHANG H., LIU M., SHI Y. L., YUEN D. A., YANA Z. Z. & LIANG G. P. 2007. Toward an automated parallel computing environment for geosciences. *Physics of the Earth and Planetary Interiors* **163**, 2–22.

ZYSERMAN F. I. & SANTOS J. E. 2000. Parallel finite element algorithm with domain decomposition for three-dimensional magnetotelluric modelling. *Journal of Applied Geophysics* **44**, 337–351.

Tables and Figures

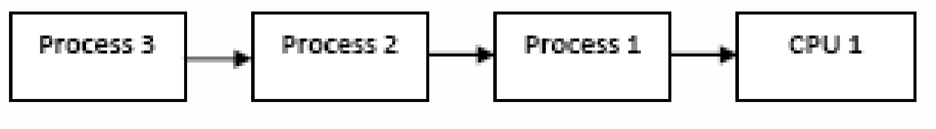

Figure 2.1: Sequential Execution of Independent Processes at Different Times

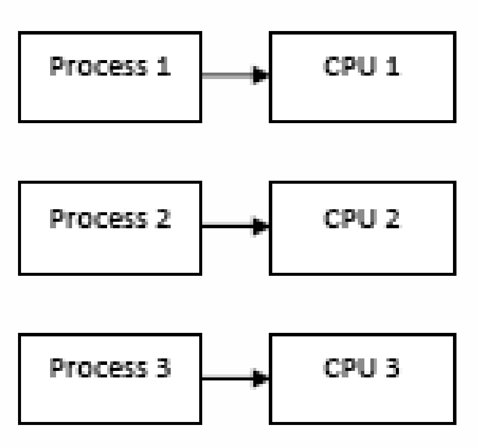

Figure 2.2: Parallel Execution of Different Processes at the Same Time

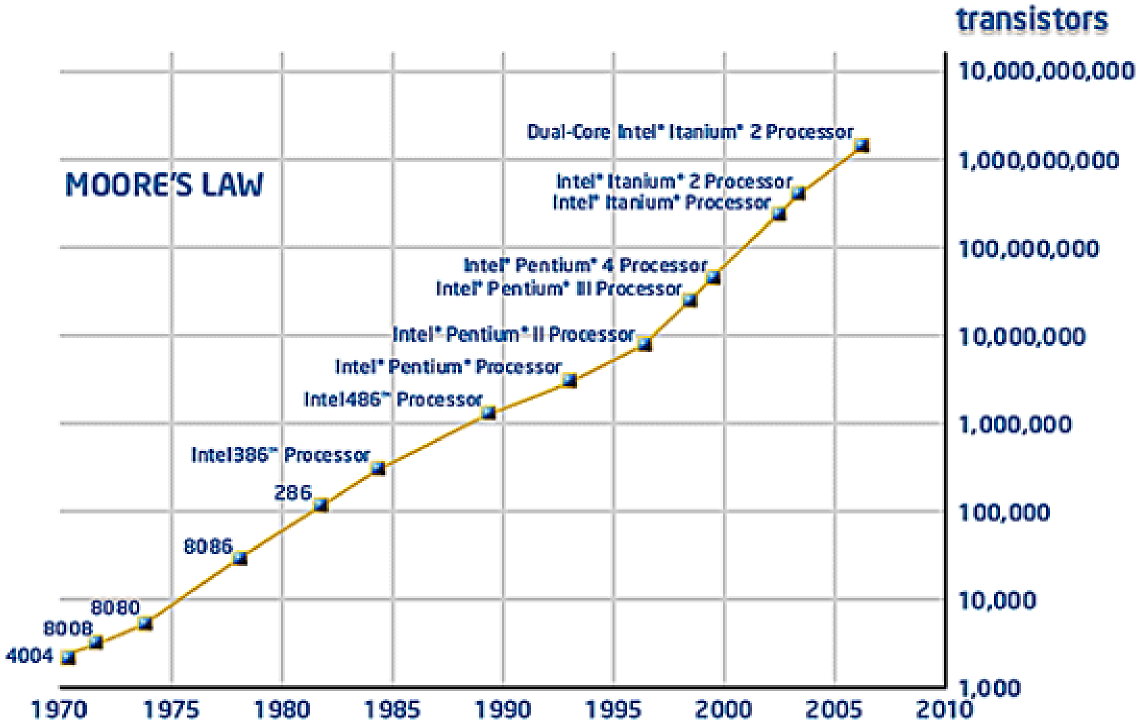

Figure 2.3: Graph Showing the Increase in the Number of Transistors on Processors with Time

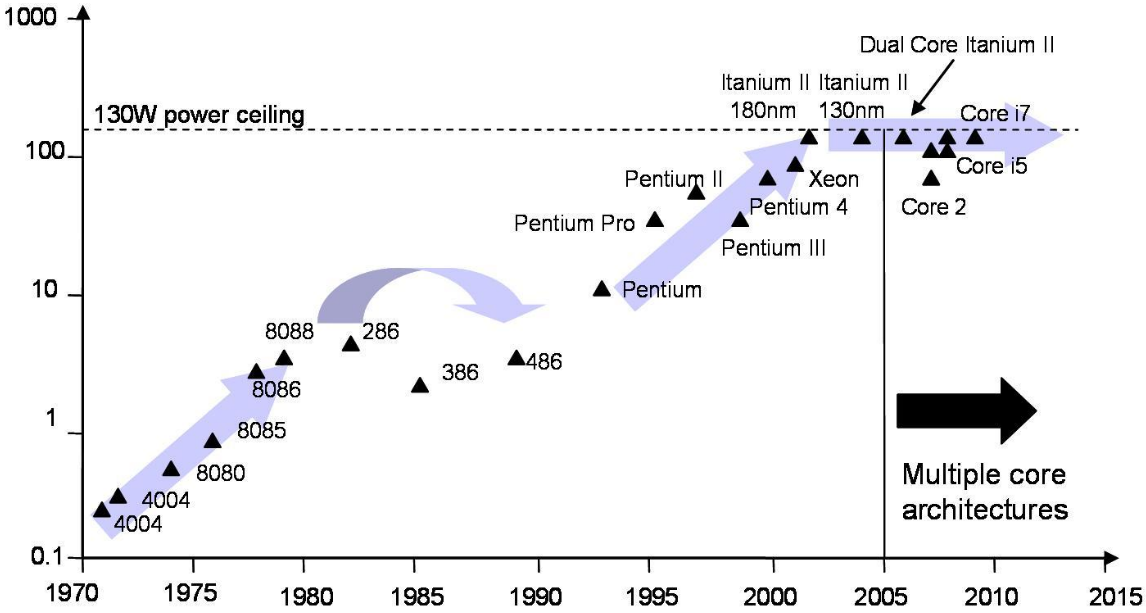

Figure 2.4: Relationship Showing the Power Ceiling Wall with Increasing Number of Transistors on Processors (Mudge, C. personal communication, 27 September 2010)

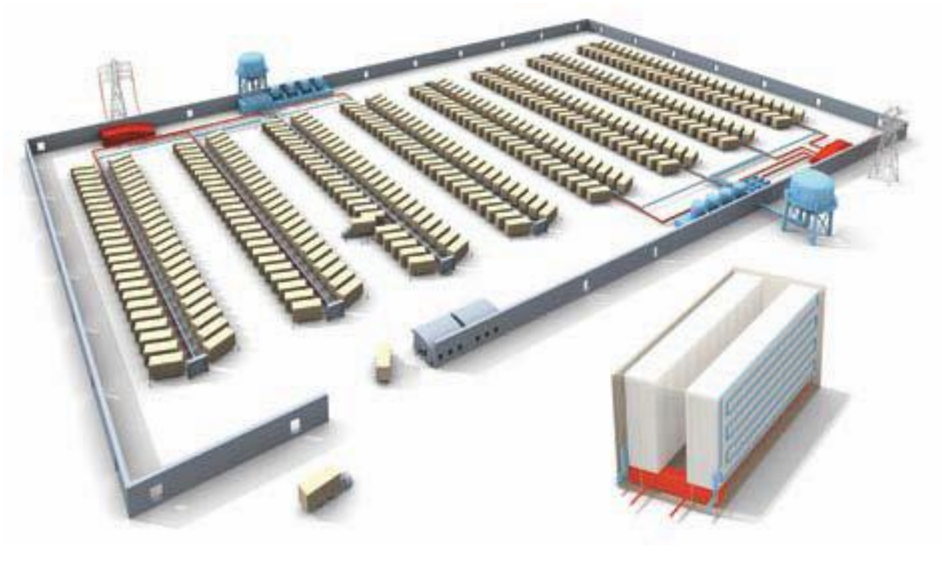

Figure 2.5: Depiction of a Data Centre in which the Servers are Located in Shipping Containers (Mudge, C. personal communication, 27 September 2010)

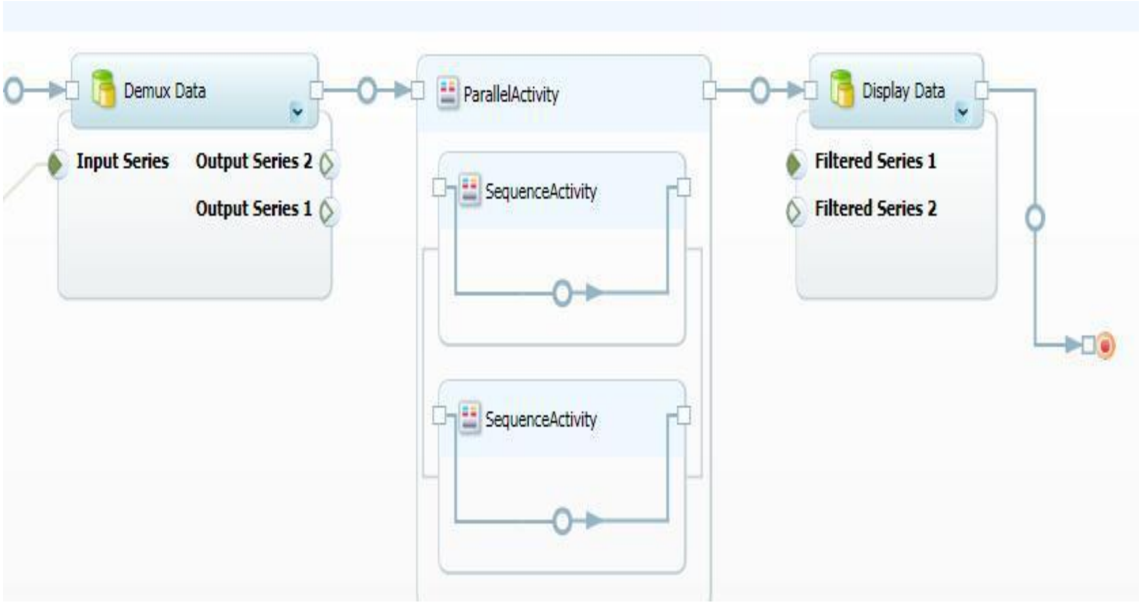

Figure 2.6: A Workflow Utilising the Trident Workflow Workbench Showing the Nodes Called Actors and the Flow of Data between Them

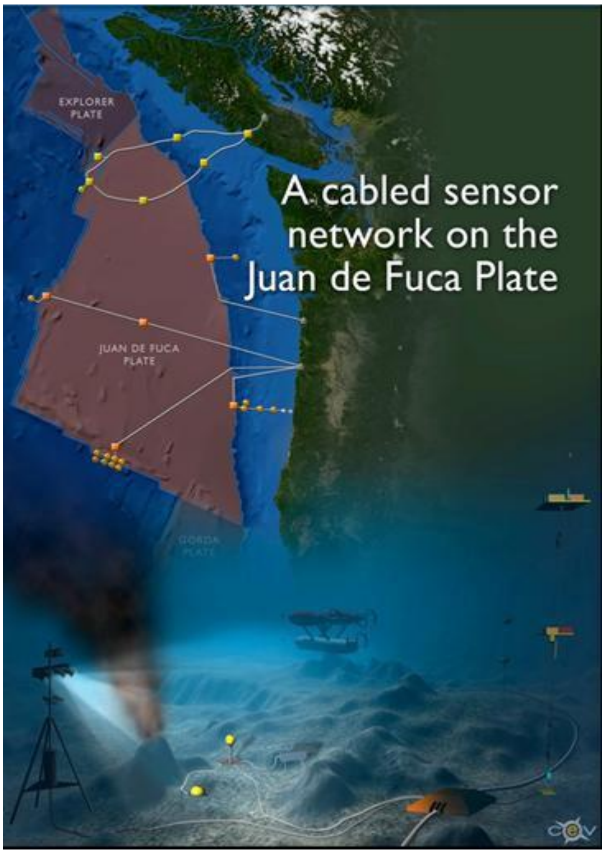

Figure 2.7: The Proposed Plate Scale Observatory on the Juan De Fuca Plate (Barga et al. 2008b)

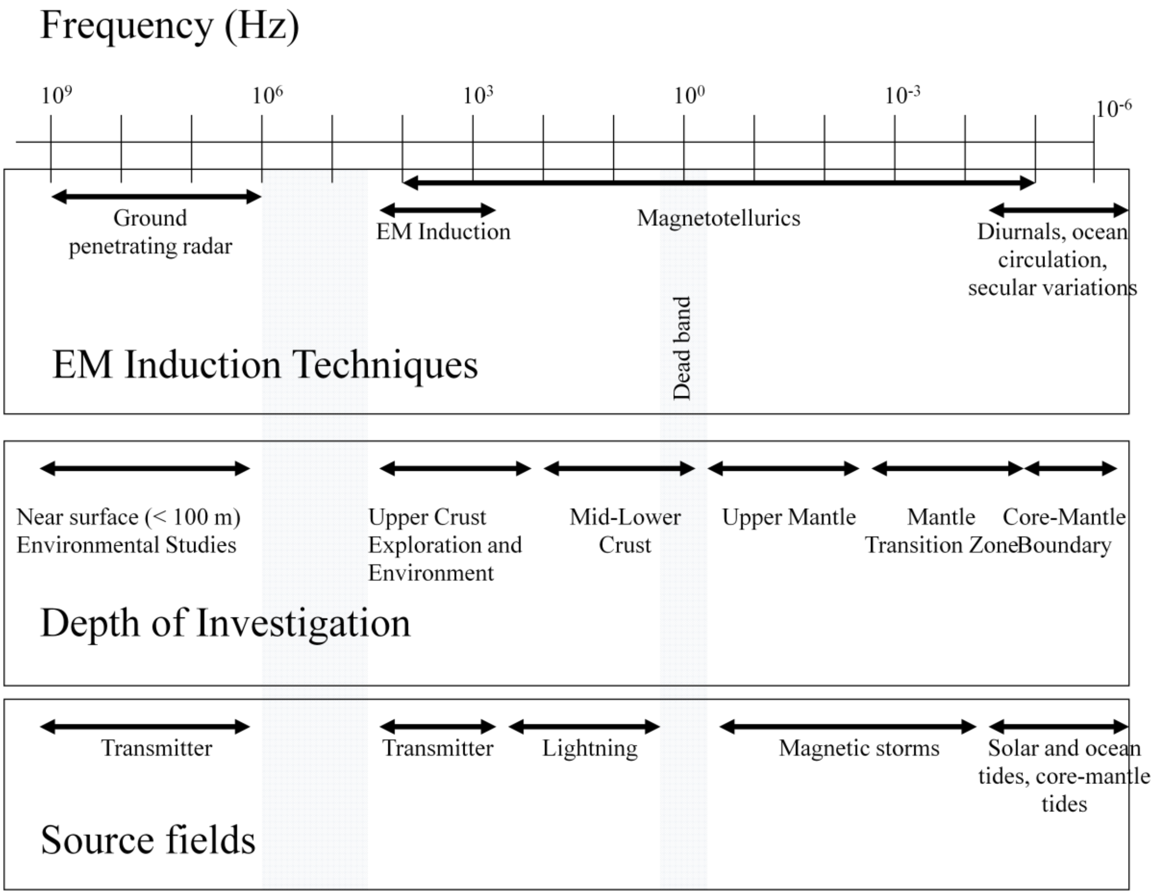

Figure 2.8: Comparisons between Frequency and the Properties of different EM Induction Techniques (Heinson, G. Personal communication, July 12, 2010)

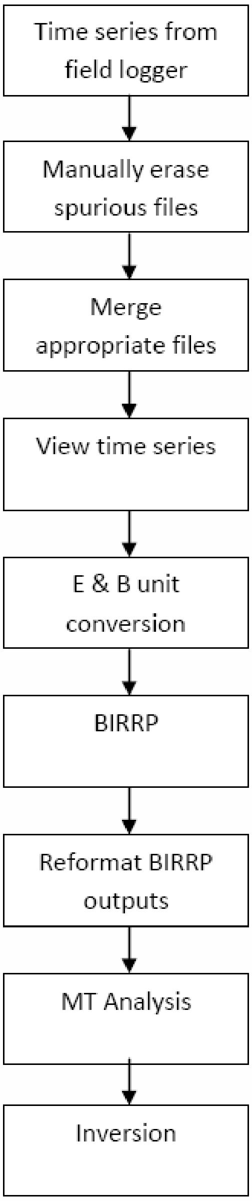

Figure 2.9: General Flowchart of the Processing Methodology

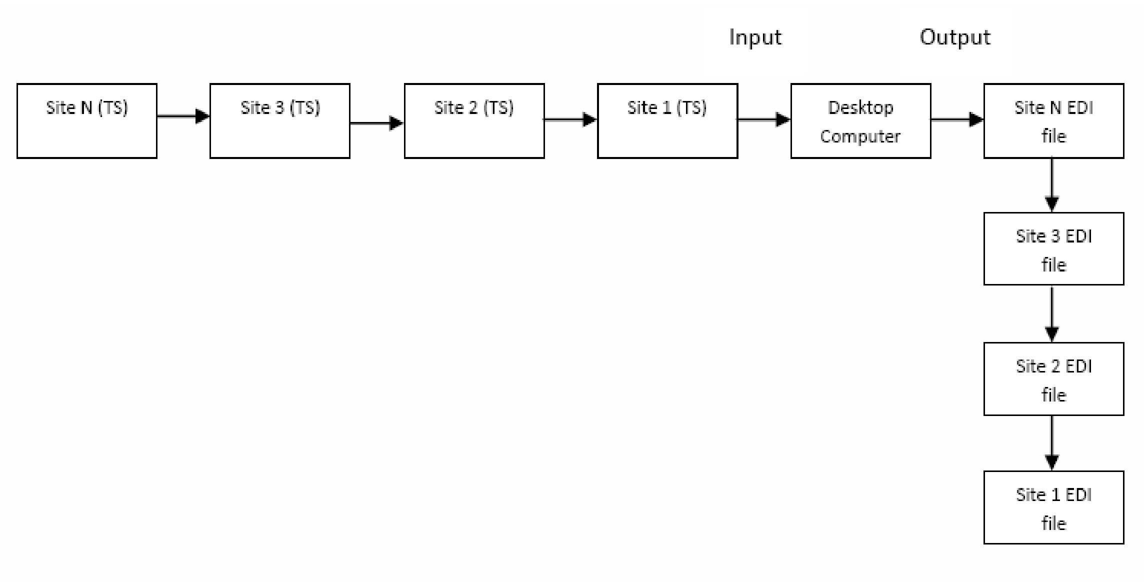

Figure 3.1: Depiction of the Serial Processing of the Time Series Data from the Field to the EDI Target Format

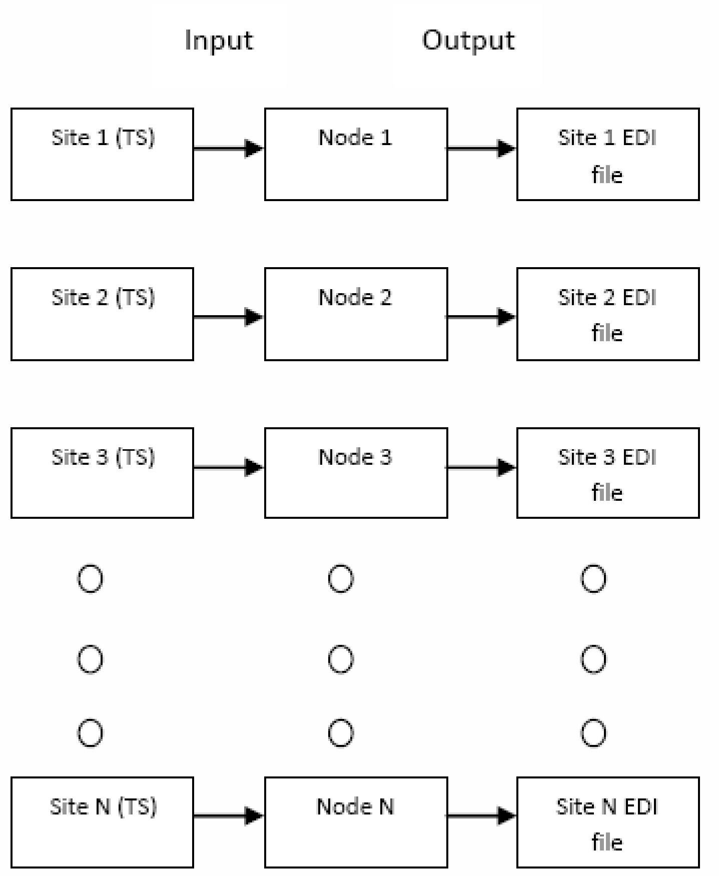

Figure 3.2: Depiction of a Shared Nothing Parallel Processing Framework of the Sites.

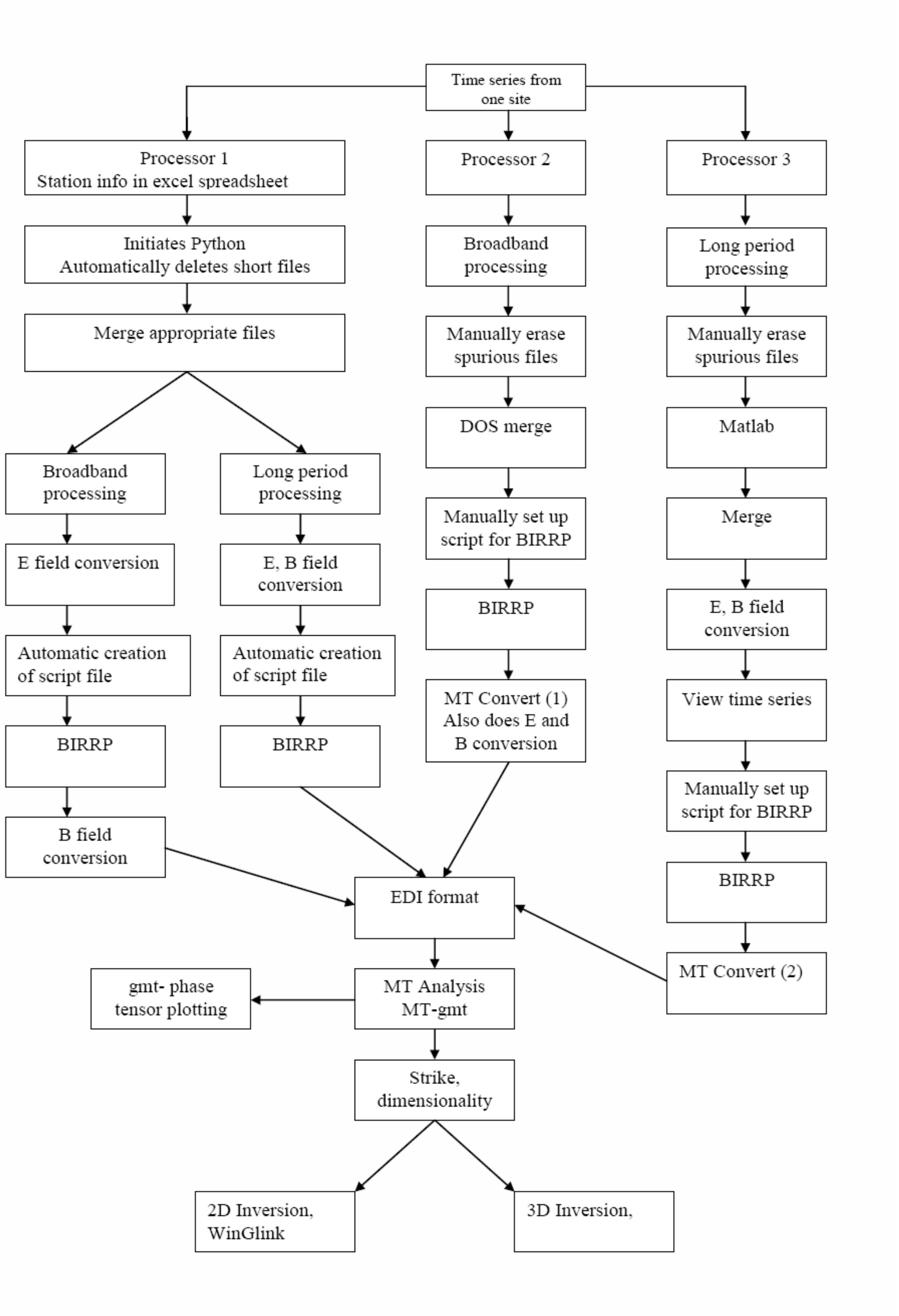
Time series from one site
Processor 1 Station info in excel spreadsheet
Processor 2
Processor 3
Initiates Python Automatically deletes short files
Broadband processing
Long period processing
Merge appropriate files
Manually erase spurious files
Manually erase spurious files
DOS merge
Matlab
Broadband processing
Long period processing
Manually set up script for BIRRP
Merge
E field conversion
E, B field conversion
BIRRP
E, B field conversion
Automatic creation of script file
Automatic creation of script file
MT Convert (1) Also does E and B conversion
View time series
BIRRP
BIRRP
Manually set up script for BIRRP
B field conversion
EDI format
BIRRP
gmt- phase tensor plotting
MT Analysis MT-gmt
MT Convert (2)
Strike, dimensionality
2D Inversion, WinGlink
3D Inversion,

Figure 3.3: Overview of the Different MT Processing Methodologies of the Three Different Processors

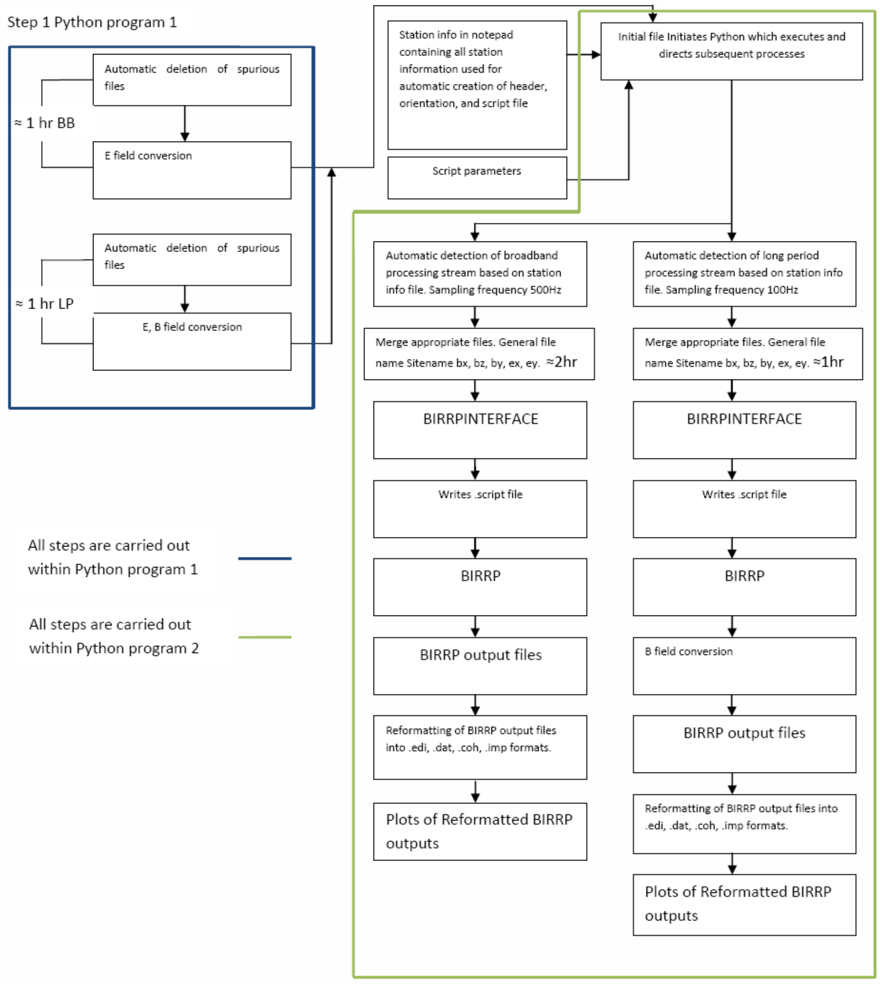
Step 1 Python program 1
Automatic deletion of spurious files
≈ 1 hr BB
E field conversion
Automatic deletion of spurious files
≈ 1 hr LP
E, B field conversion
Station info in notepad containing all station information used for automatic creation of header, orientation, and script file
Script parameters
Initial file Initiates Python which executes and directs subsequent processes
Automatic detection of broadband processing stream based on station info file. Sampling frequency 500Hz
Merge appropriate files. General file name Sitename bx, bz, by, ex, ey. ≈2hr
BIRRPINTERFACE
Writes .script file
BIRRP
BIRRP output files
Reformatting of BIRRP output files into .edi, .dat, .coh, .imp formats.
Plots of Reformatted BIRRP outputs
Automatic detection of long period processing stream based on station info file. Sampling frequency 100Hz
Merge appropriate files. General file name Sitename bx, bz, by, ex, ey. ≈1hr
BIRRPINTERFACE
Writes .script file
BIRRP
B field conversion
BIRRP output files
Reformatting of BIRRP output files into .edi, .dat, .coh, .imp formats.
Plots of Reformatted BIRRP outputs
All steps are carried out within Python program 1
All steps are carried out within Python program 2

Figure 3.4: Detailed Overview of Processor One's MT Processing Methodology

Table 3.1 Different Instance Types and Prices Offered by Amazon EC2

Standard On-Demand Instances	**Windows Usage**	**Memory**
Small (Default)	$0.12 per hour	1.7 GB
Large	$0.48 per hour	7.5 GB
Extra Large	$0.96 per hour	15 GB
Micro On-Demand Instances	**Windows Usage**	
Micro	$0.035 per hour	613 MB
High-Memory On-Demand Instances	**Windows Usage**	
Extra Large	$0.62 per hour	17.1 GB
Double Extra Large	$1.24 per hour	34.2 GB
Quadruple Extra Large	$2.48 per hour	68.4 GB

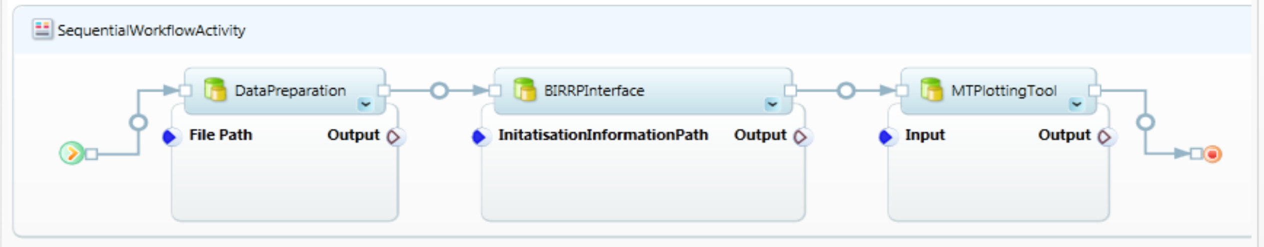

Figure 3.5: Sequential Execution of the MT Processing Using Existing Processing Codes

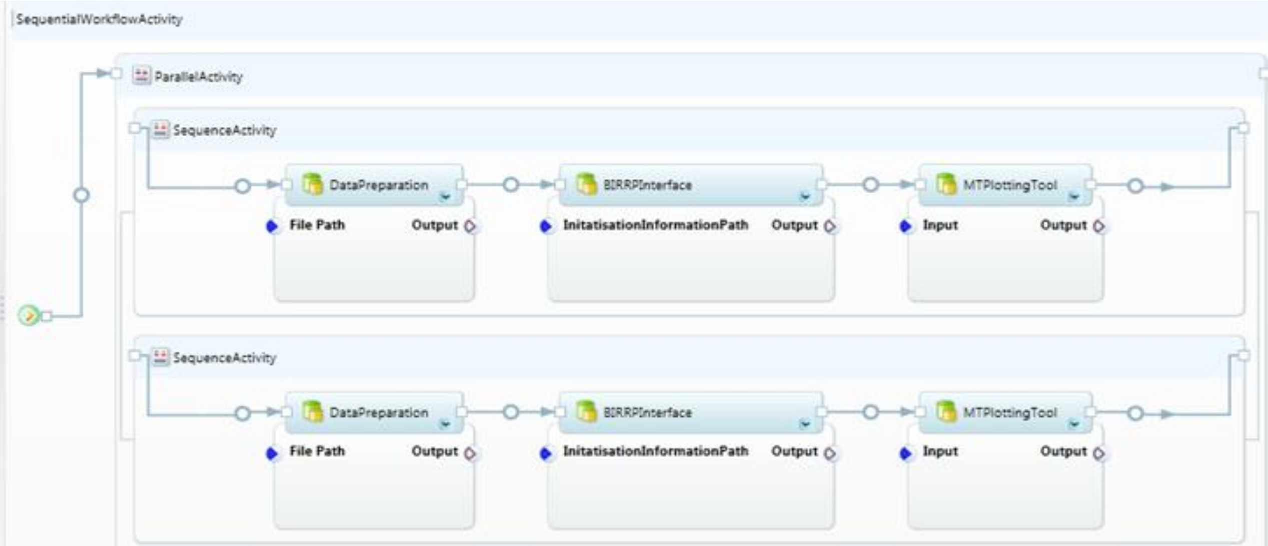

Figure 3.6 Parallel Execution of the MT Processing Using Existing Processing Codes

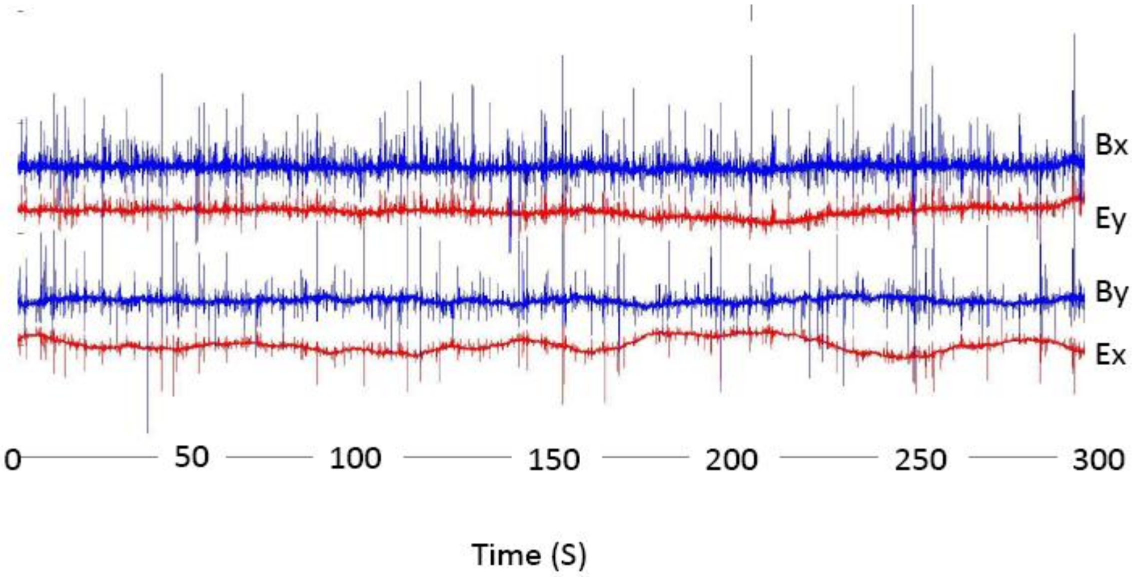

Figure 4.1: An Example of Good MT Data for Orthogonal Components of the Electric and Magnetic Fields Sampled at 500Hz

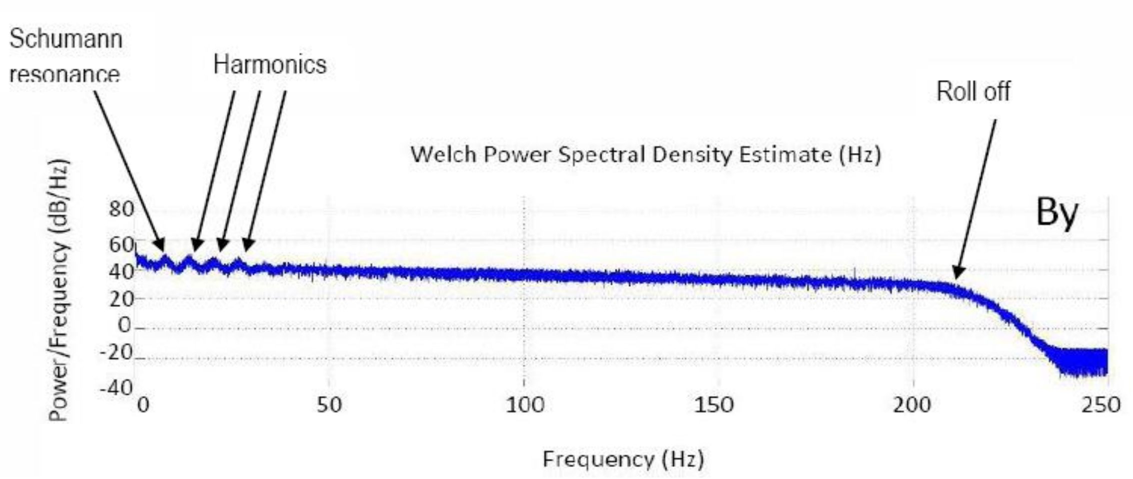

Figure 4.2: Power Spectra of the Time Series for the By Component Shown in Figure 4.1

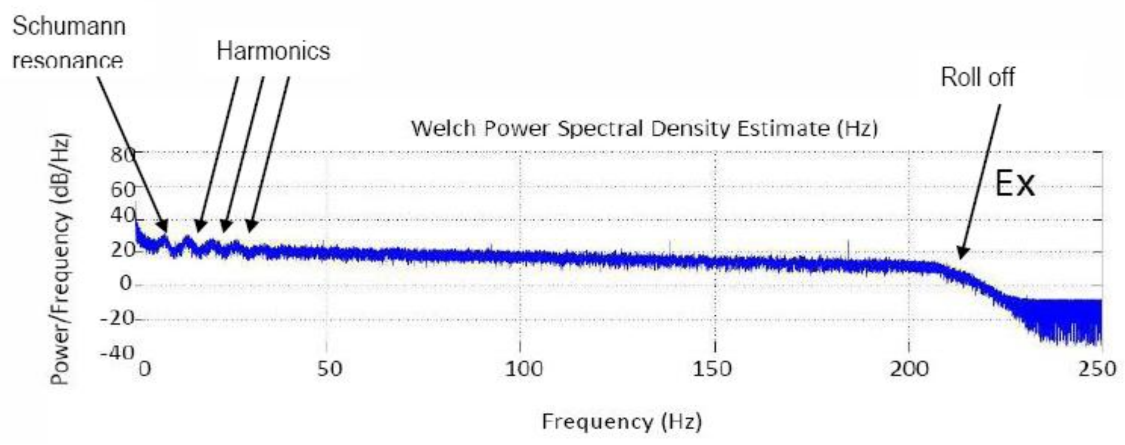

Figure 4.3: Power Spectra of the Time Series for the Ex Component Shown in Figure 4.1

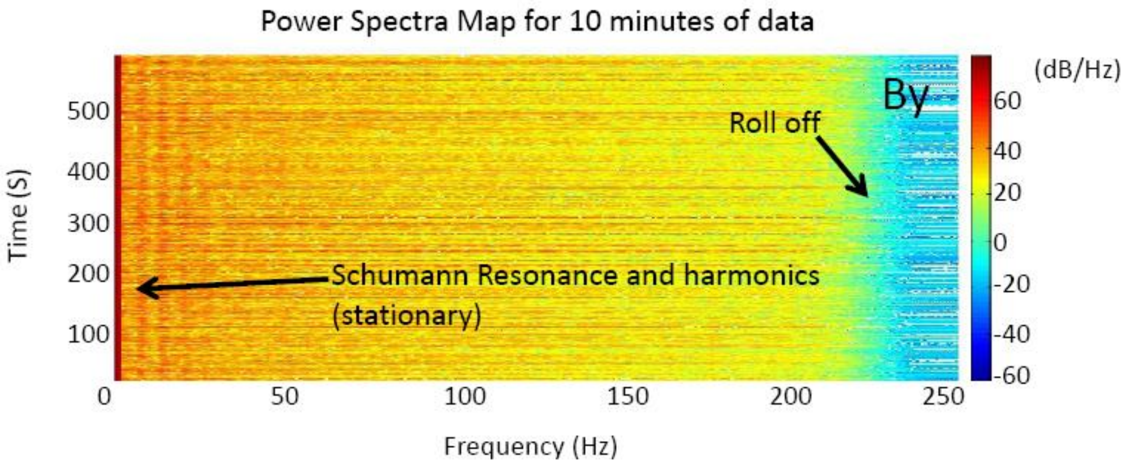

Figure 4.4: Power Spectra Map of the Time by Component Showing the Change of the Power Spectrum with Time

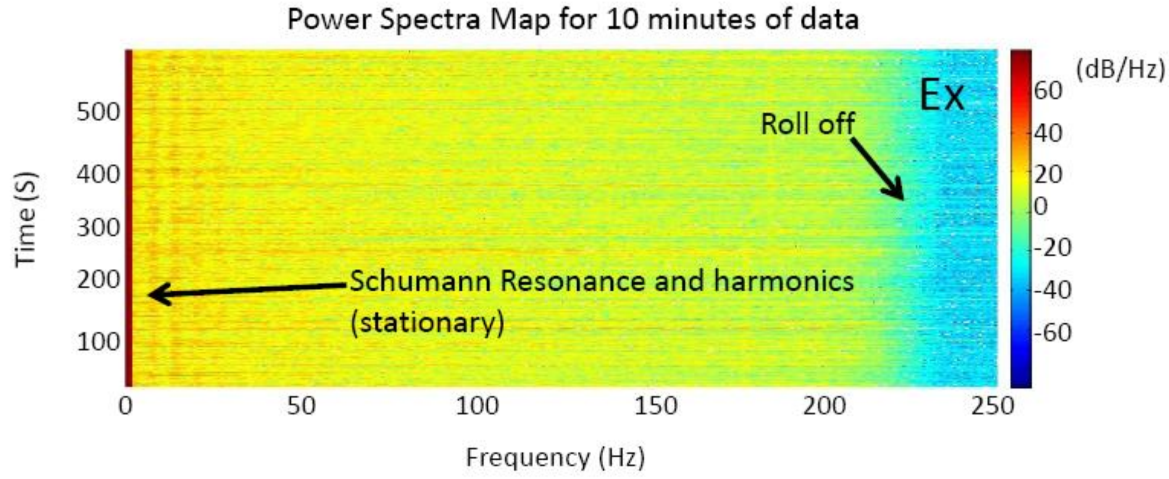

Figure 4.5: Power Spectra Map of the Ex Component Showing the Change of the Power Spectrum with Time

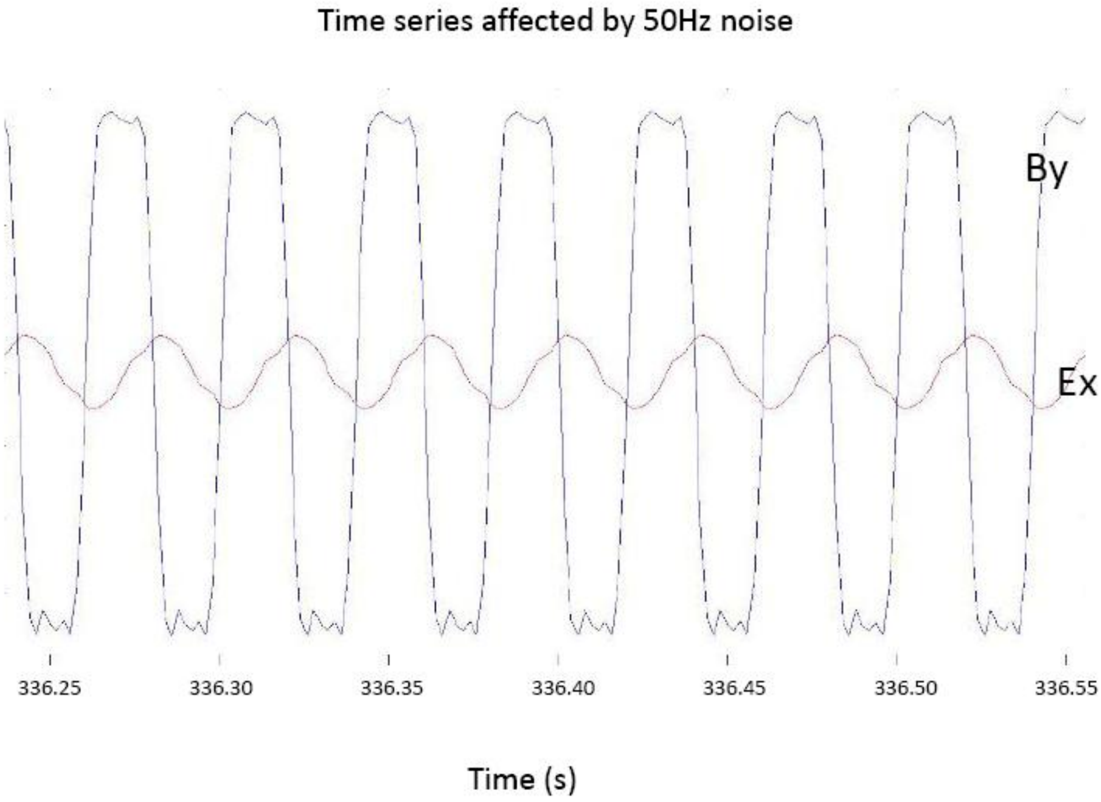

Figure 4.6: An Example of Time Series Data Affected by 50Hz Electrical Noise

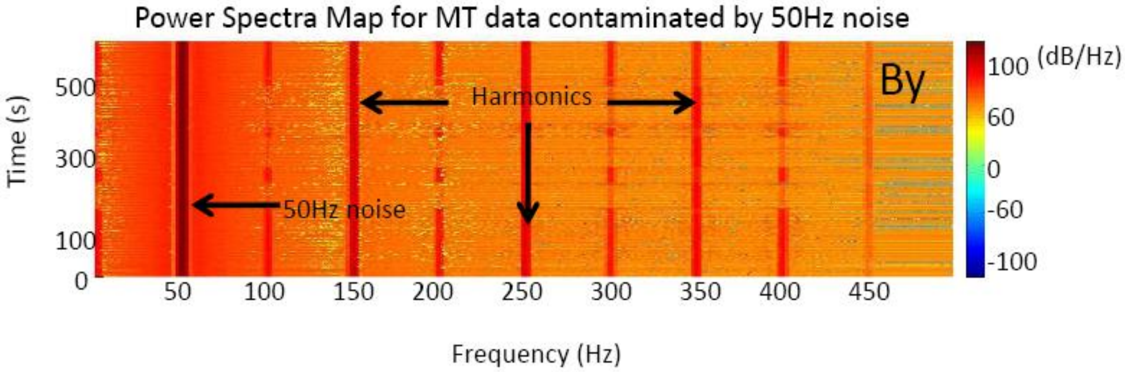

Figure 4.7: Power Spectra Map of the By Component Showing the Change of the Power Spectrum with Time

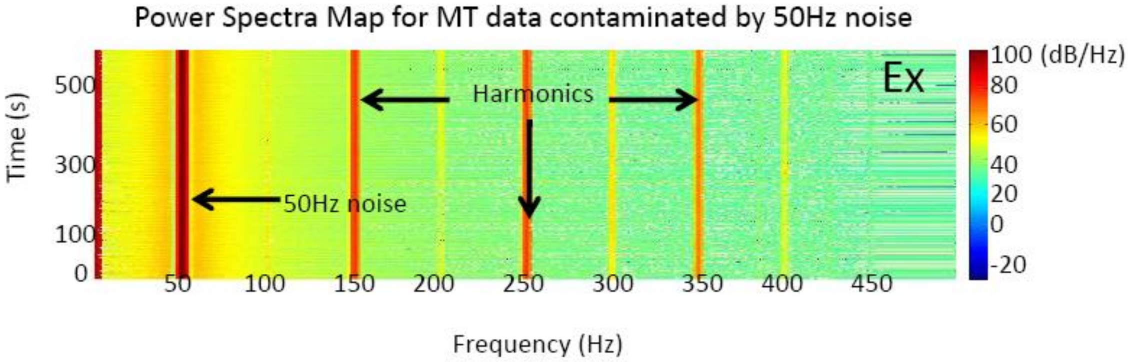

Figure 4.8: Power Spectra Map of the Ex Component Showing the Change of the Power Spectrum with Time

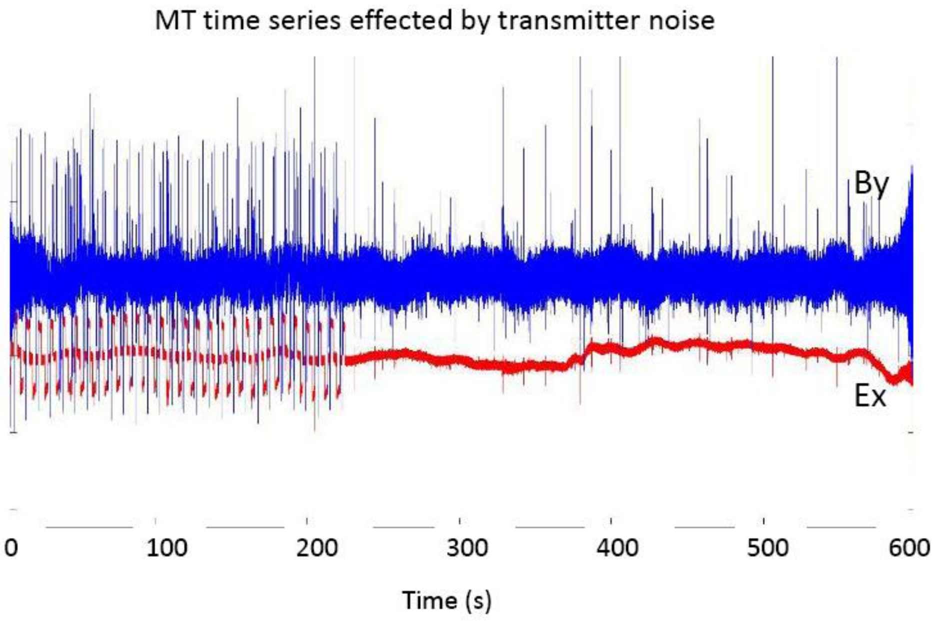

Figure 4.9: An Example of a Time Series Affected by Intermittent Transmitter Noise

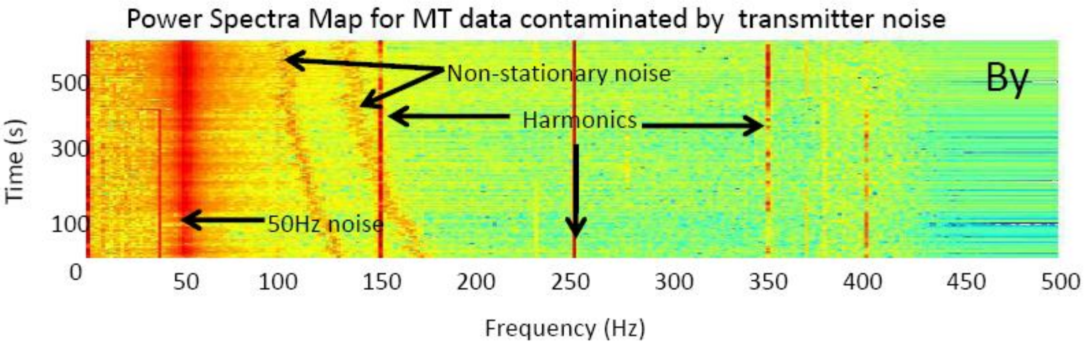

Figure 4.10: Power Spectra Map of the By Component Affected by the Transmitter Showing Both Stationary and Nonstationary Noise

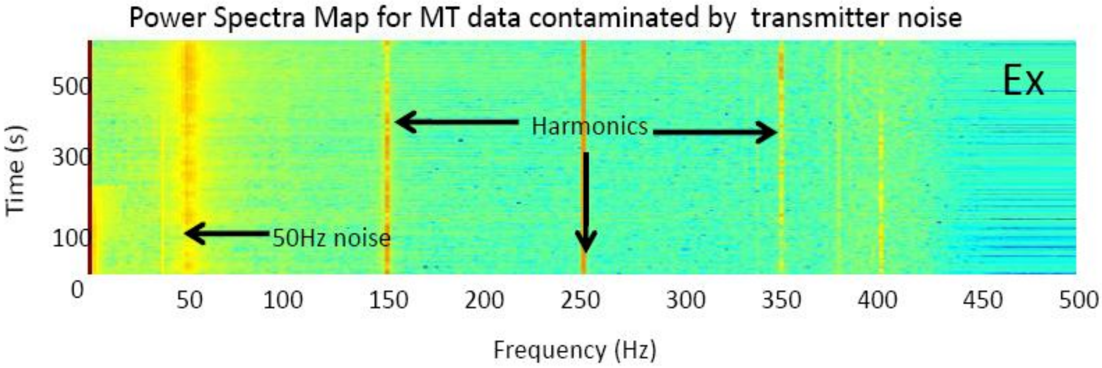

Figure 4.11: Power Spectra Map of the Ex Component Affected by the Transmitter Showing Stationary Noise

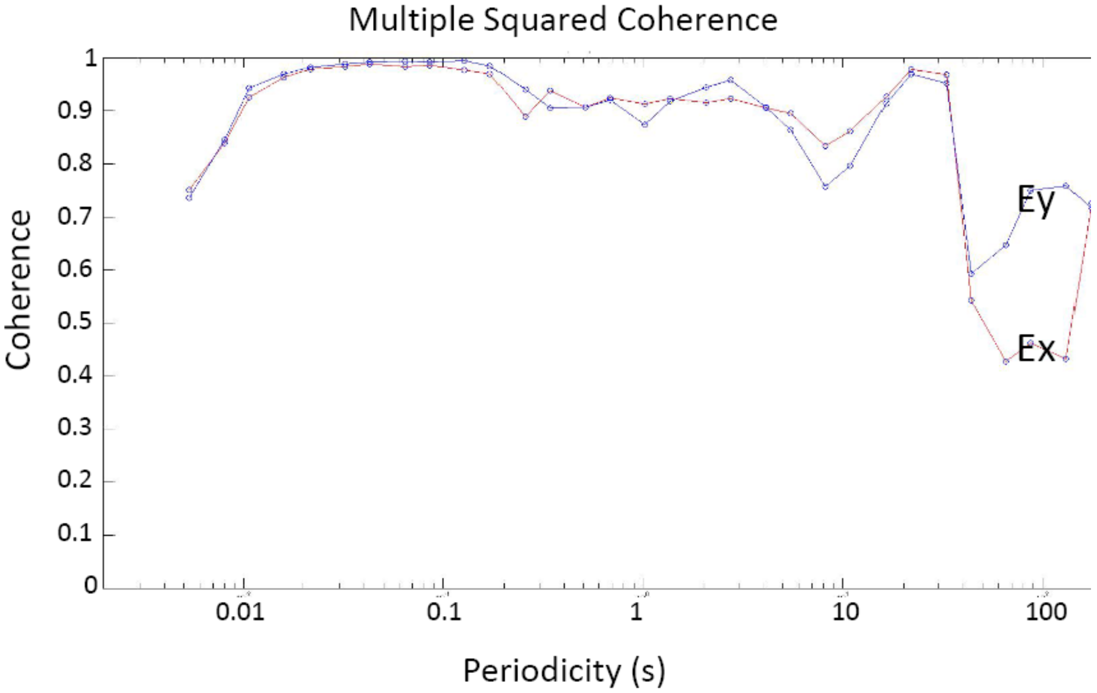

Figure 4.12: A Coherence Plot Characteristic of Good MT Data

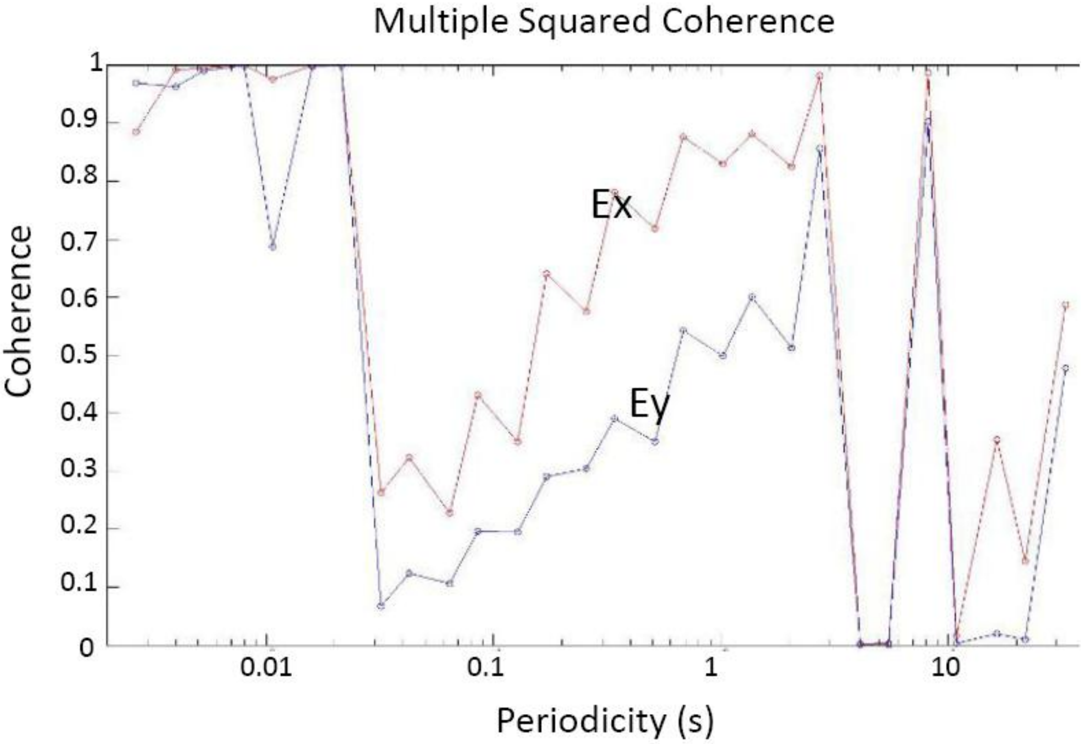

Figure 4.13: A Coherence Plot Characteristic of Poor MT Data

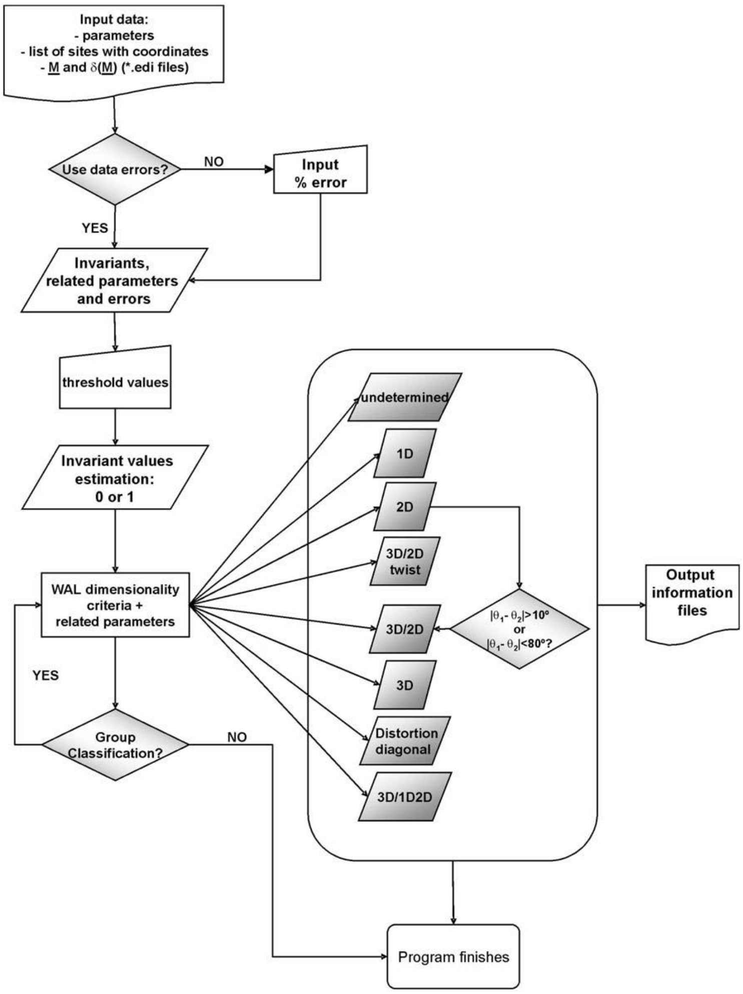
Input data:
- parameters
- list of sites with coordinates
- M and δ(M) (*.edi files)
Use data errors?
NO
Input % error
YES
Invariants, related parameters and errors
threshold values
Invariant values estimation: 0 or 1
WAL dimensionality criteria + related parameters
YES
Group Classification?
NO
undetermined
1D
2D
3D/2D twist
3D/2D
3D
Distortion diagonal
3D/1D2D
|θ1- θ2|>10° or |θ1- θ2|<80°?
Output information files
Program finishes

Figure 4.14: An Overview Flowchart of how the WALDIM Dimensionality Analysis Program Works (Marti et al. 2009).