Australian Government
Geoscience Australia

HiperSeis: Supercharging Seismic Workflows on High Performance Computing Platforms

Rakib Hassan, Babak Hejrani, Alexei Gorbatov and Fei Zhang

Introduction

Geoscience Australia (GA) maintains a collection of permanent seismic stations scattered around continental Australia. GA also deploys temporal arrays of seismic stations, progressively spanning the entire continent (Fig. 1). AusArray unites data collected from the Australian National Seismological Network (ANSN), multiple academic transportable arrays (supported by AuScope and individual grants) as well as the seismometers in schools program. A recent addition has been Geoscience Australia's Exploring for the Future program (EFTF) which has doubled the national rate of such data collection.
These data are stored on traditional file systems in legacy formats and are not amenable to data- and compute-intensive seismic workflows.

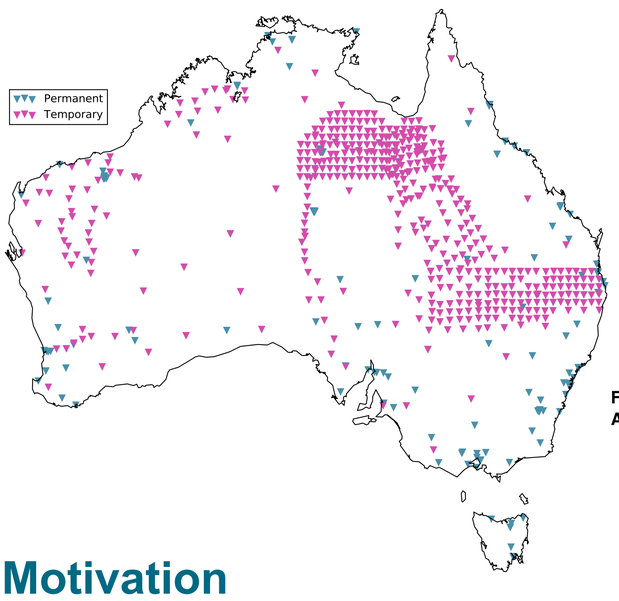

Figure 1. Permanent and temporary stations around Australia

Motivation

High-resolution P and S body wave travel-time tomography of continental Australia are key deliverables for the EFTF program at GA. The primary input for generating a travel-time tomography is travel-time residuals – the difference between observed travel-times for earthquakes recorded at seismic stations and their corresponding theoretically predicted values (Fig. 2) – for a collection of stations andearthquakes such that their corresponding ray-paths intersect the region of interest.
GA maintains a collection of over 20 TB of waveform data, from ~900 stations, spanning over more than two decades. Automatically detecting body wave arrivals and deriving travel-time residuals for selected earthquakes (~25000) recorded over this period (Fig. 3) is an extremely IO- and compute-intensive exercise – thus necessitating highly scalable parallel filesystems and high performance computing environments, e.g. at the National Computational Infrastructure (NCI).

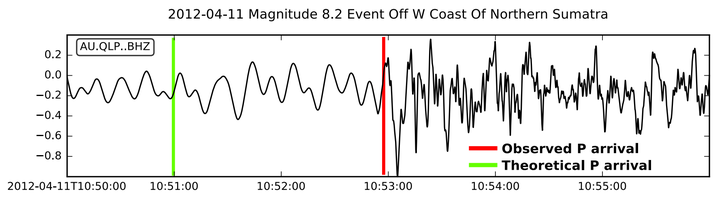

Figure 2. Z-component waveform data from station QLP, centred on the P wave arrival for a magnitude 8.2 earthquake off the West coast of Northern Sumatra.

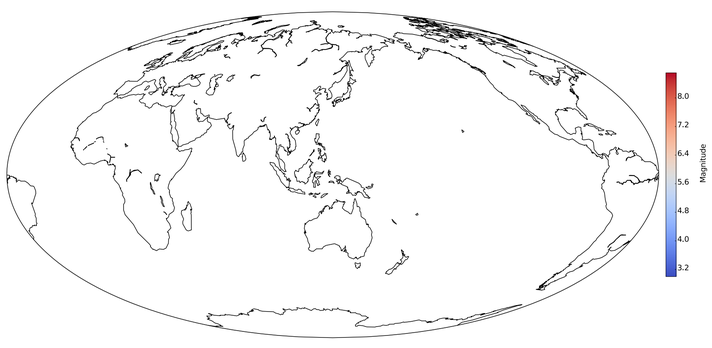

Figure 3. Earthquake events used in this study, coloured and sized according to magnitude. Events have been adaptively selected to provide good azimuthal coverage of the Australian continent.

Methodology

• We convert waveform data from legacy miniseed format to ASDF (Krischer et al. 2016), which is based on HDF5 containers. This allows us to take advantage of the parallel Lustre filesystem at NCI.

• Automatically detecting body wave arrivals is an embarrassingly parallel problem. Each processor reads time-contiguous chunks of waveform data, in parallel, and runs computationally intensive algorithms for detecting body wave arrivals for earthquakes that fall within the time-range.

• We use the computationally expensive, but robust AIC arrival detection algorithm implemented in PhasePAPY (Chen et al. 2016), which we converted to Cython to speed up the original pure-python implementation.

• While processing each waveform, we iteratively apply six different – progressively less stringent – parameterizations of the AIC algorithm until an arrival is detected or all parameterizations are exhausted. Although it increases computational cost significantly, this approach allows us to categorize arrivals by quality – the higher quality arrivals being detected by more stringent parameterizations.

• Although AIC is widely used for arrival detection, it is not immune to misidentifications. We compute additional quality measures based on sinuosity and wavelet analysis of a given waveform around a putative arrival (Fig. 4).

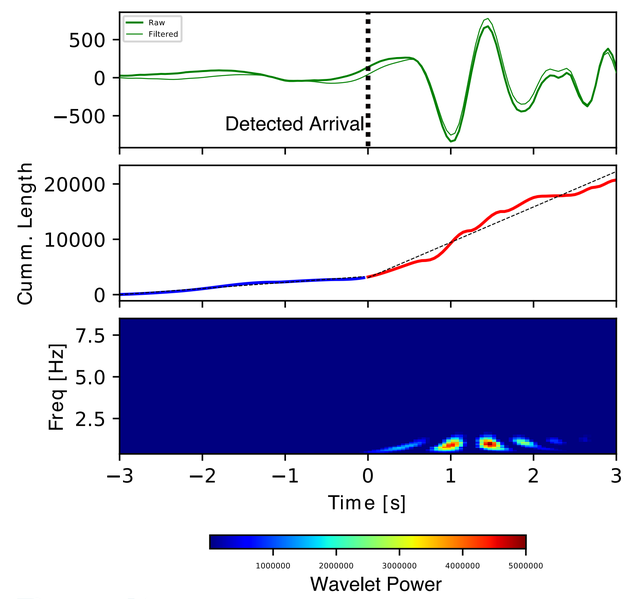

Figure 4. The top panel shows raw and filtered waveforms centred on the detected arrival time. The middle panel shows the cumulative integral of the absolute raw waveform, before and after the detected arrival. A marked change in gradient indicates a robust arrival. The bottom panel shows the continuous wavelet transform of the raw waveform. A marked change in wavelet power after the detected arrival indicates a robust arrival.

Results

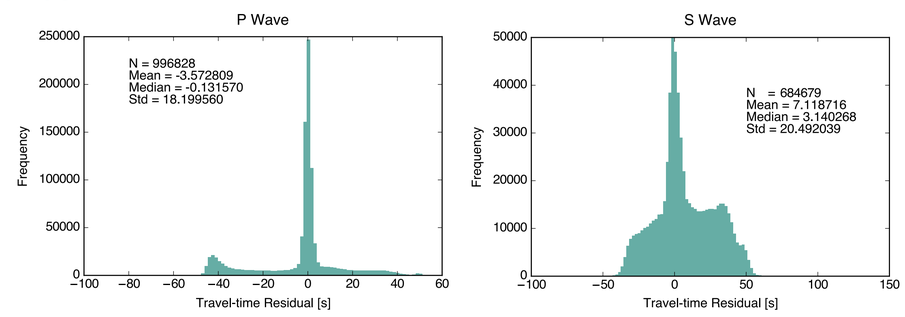

Figure 5. (a) Distribution of travel-time residuals for detected P arrivals are shown split into 100 bins. (b) Same as in (a), but for detected S arrivals.

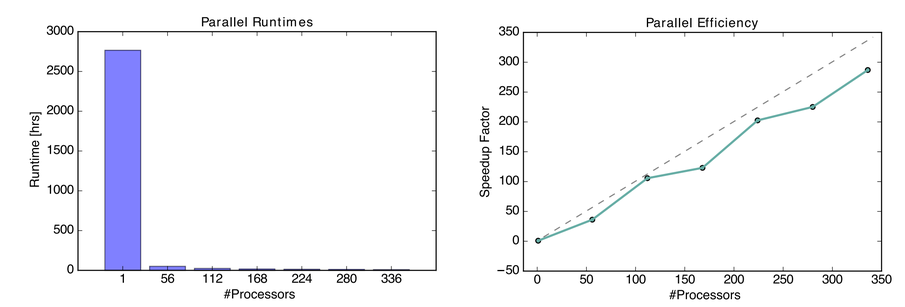

Figure 6. (a) Parallel runtimes are shown in comparison to a serial case – see text below. (b) The teal curve shows speedup factor as a function of the number of processors. The dashed grey line shows linear speedup.

• Waveform data from ~900 stations were analyzed for ~25000 earthquake events, leading to ~22500000 waveforms being processedfor detecting P and S wave arrivals, followed by quality assessment.

• Figures 5a and 5b show the number of P and S wave arrivals detected, respectively. The majority of these arrivals agree well with the theoretical arrival time, as shown by their clustering around 0.

• Figure 6a shows parallel runtimes, which were computed based on 10% of the data and then projected to the actual problem size.

• Parallel efficiency shown in Fig. 6b suggests a near-linear speedup with an increasing number of processors, which is generally expected for an embarrassingly parallel problem. We expect this near-linear trend in speedup to continue up to the peak IO throughput of the Lustre filesystem, at which point the problem becomes IO-bound and increasing the number of processors further may degrade performance instead.

Discussion and Conclusion

All software programs used in this work are available in the open source Github repository HiperSeis (Zhang et al 2019), which is developed at GA. In particular, it comprises scripts for converting seismic waveform data into ASDF format, amenable to highly scalable parallel file-systems. It also contains parallelized modules for detecting earthquake arrivals and computing cross-correlation of waveform data.

Current results from parallel earthquake arrival detection, run on 336 cores, over more than 20 TB of combined waveform data, suggest a speed up by a factor of ~300. An exercise that would have otherwise taken in the order of three months can now be completed overnight. We expect similar, potentially improved, speedups for more computationally intensive workflows, e.g. cross-correlation of waveform data. Short turnaround times of these workflows facilitate experimentation with enhanced algorithms for seismic data analysis.

Acknowledgement

Geoscience Australia acknowledges the traditional custodians of the country where this work was undertaken. We also acknowledge the support provided by individuals and communities to access the country, especially in remote and rural Australia.

References

Lion Krischer, James Smith, Wenjie Lei, Matthieu Lefebvre, Youyi Ruan, Elliott Sales de Andrade, Norbert Podhorszki, Ebru Bozdağ, Jeroen Tromp, An Adaptable Seismic Data Format, Geophysical Journal International, Volume 207, Issue 2, November, 2016, Pages 1003–1011

Chen Chen, Austin A. Holland; PhasePApy: A Robust Pure Python Package for Automatic Identification of Seismic Phases. Seismological Research Letters ; 87 (6): 1384–1396

Zhang, F., Hassan, R., Medlin, A., Gorbatov, A., & Hejrani, B. (2019, April 17). GeoscienceAustralia/hiperseis. Retrieved April 30, 2019, from https://github.com/GeoscienceAustralia/hiperseis

GA 15-xxxx | GeoCat xxxx)

GEOSCIENCE AUSTRALIA
APPLYING GEOSCIENCE TO AUSTRALIA'S MOST IMPORTANT CHALLENGES

For Further Information:
Email: rakib.hassan@ga.gov.au
Ph: +61 2 6249 9416 **Web:** www.ga.gov.au