SAND2011-4691C

PEDAL: A NEW EVENT DETECTION AND SIGNAL ASSOCIATION ALGORITHM

Sanford Ballard, Timothy J. Draelos, Mark A. Gonzales, and Christopher J. Young

Sandia National Laboratories

Sponsored by the National Nuclear Security Administration

Contract No. DE-AC04-94AL85000/SL08-IRP-NDD02

ABSTRACT

The Probabilistic Event Detection, Association and Location algorithm (PEDAL) is a new approach to the problem of associating isolated seismic observations from a network of stations into a list of hypothesized seismic events consistent with those observations. In our method, the Earth is discretized into a dense 3D grid of nodes that spans the globe from the surface down to the maximum depth at which earthquakes are likely to occur. The grid is extended to 4D by the addition of a time dimension. Given a set of seismic observations within a 23-minute time window, a network 'fitness' value is calculated at each grid node by summing the station-specific conditional fitnesses, which are proportional to the conditional probabilities that each observation was generated by a seismic event at the grid node and assuming that each observation was generated by a refracted P wave. The node with the highest fitness value is accepted as a hypothetical seismic event location, subject to some minimal fitness value, and all seismic arrivals within a 40-minute time window that are consistent with that event are associated with it. During this association step, the assumption that the arrival was a direct P arrival is relaxed and many different phases are considered. Once an arrival is associated with an event, it is removed from further consideration. While there are still unassociated arrivals, the search is repeated to find other hypothetical seismic events until no more seismic events are identified that satisfy the minimum fitness criteria.

Because the exhaustive search approach is computationally expensive, we have implemented the algorithm on Graphics Processing Units (GPUs), thereby achieving performance levels needed to meet the requirements of real time monitoring systems like the CTBTO's IDC, while still running on a single machine. We evaluate performance relative to current association algorithms by processing an interval of IMS data, and comparing our results to both the SEL3 and LEB bulletins.

Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

OBJECTIVES

The objective of this work is the development of an improved system to address the identification of seismic events from a stream of monitoring station observations (detections and corresponding feature measurements). Generalized Association, the current seismic event association technology in use by the Comprehensive Test Ban Treaty Organization (CTBTO) for their International Data Center (IDC), is almost 15 years old (Le Bras*, et al*. 1994), predating some important innovations in computer software and hardware. While this easily parallelizable technique was an improvement over the previous rule-based expert system approach that scaled poorly with large numbers of station signal detections, it has proven difficult to tune, and the quality of the IDC automatically produced bulletin (SEL3) has changed little over the past decade, suggesting that a new approach is warranted. We believe that the association problem could benefit greatly from new algorithms that exploit modern data and computational resources. The availability of many years of historical data (observables) for known sources and the steady improvement in advanced algorithms, such as machine learning techniques, as well as the considerable advances in computational power provide an opportunity to address the automated association problem in a new manner. A set of observables readily available from IDC database tables, such as arrival time and amplitude, signal-to-noise ratio, slowness, and azimuth, from all monitoring stations over a window of time (e.g., 1 hour) can be combined in elemental and differential form to create a global feature vector representing all source events occurring in the specified time window. Key to accurate event identification is comparing the differences in time and other features of seismic phases received at various monitoring stations in proximity of the actual source event: the relative patterns seen between stations are characteristic of signals coming from a particular source location. Information from historical analyst-reviewed event bulletins can be used as ground-truth target information to develop and test an algorithm that maps these arrival features to event probabilities for a given location. This is of great importance because accurately locating a seismic event in 3-dimensional space (latitude, longitude, depth) is a key step in identifying nuclear explosions.

RESEARCH ACCOMPLISHED

Introduction

When a nuclear explosion occurs below the surface of the Earth, signals created by the source propagate through the Earth and are recorded by a network of sensors measuring ground motion. To identify such an event, data must be carefully processed through a standard series of steps. First, data from each sensor are processed separately to find signals of interest, then the set of signals from the full network are associated to generate hypothetical events (times and locations) that can account for the signals. Next, magnitudes (sizes or yields) of the events are estimated, along with likely source type (e.g. earthquake or explosion). All of this is done automatically, but due to the complexity of the problem and poor station coverage for many areas of the world, particularly for smaller events, the automatic results must be reviewed carefully by an analyst and any errors/omissions must be corrected before further action will be taken on any events of interest for nuclear explosion monitoring. Nuclear explosions are rare, but other types of seismic events (e.g. earthquakes) are not and identifying event type with seismic data is difficult, so the number of events to be examined each day is large (typically more than 100).

All of the processing steps are important, but association of the signals to form the events is perhaps the most challenging and critical. The associator produces the list of event hypotheses which must be reviewed by the analysts to determine if any nuclear explosions have occurred, so the amount of work that the analysts must do is directly controlled by the associator. Further, the association step typically controls the level at which the station signal detection thresholds are set, because the number of possible event hypotheses scales exponentially with the number of detections. Thus, setting very low station thresholds (so as not to miss an explosion), can overwhelm the associator, slowing or even stopping the data processing flow. An ideal associator would be able to efficiently process very large numbers of detections, hence allowing station detection thresholds to be set as low as possible, but would also produce only very high quality events, requiring minimal analyst review. In reality, a compromise must be chosen between missing legitimate events and creating false events that will be rejected by analysts.

PEDAL Earth Grid

Dividing the Earth into a 3D grid of location nodes, each node will have a unique feature signature to estimate the probability that a seismic event occurred at that location and generated a subset of the observations in the feature vector. We use a uniform grid spacing of 0.5 degrees at the surface of the Earth (Figure 1). In addition, we add additional depth nodes in regions where deep seismicity has occurred in the past (Figure 2). Altogether, this gives us

a total of 383,753 nodes to cover the Earth, which is a significant computational challenge. Fortunately, since each node has a unique set of predicted observables, both algorithm development (i.e., training) and operational use are fully parallelizable. A further complication that must be addressed specifically for the nuclear monitoring application is that many of the node locations will not have historical data available because seismicity has a limited distribution. Training data for these nodes must be extrapolated from other nodes or generated using modeling. Determining which approach is most appropriate is an important goal of our research.

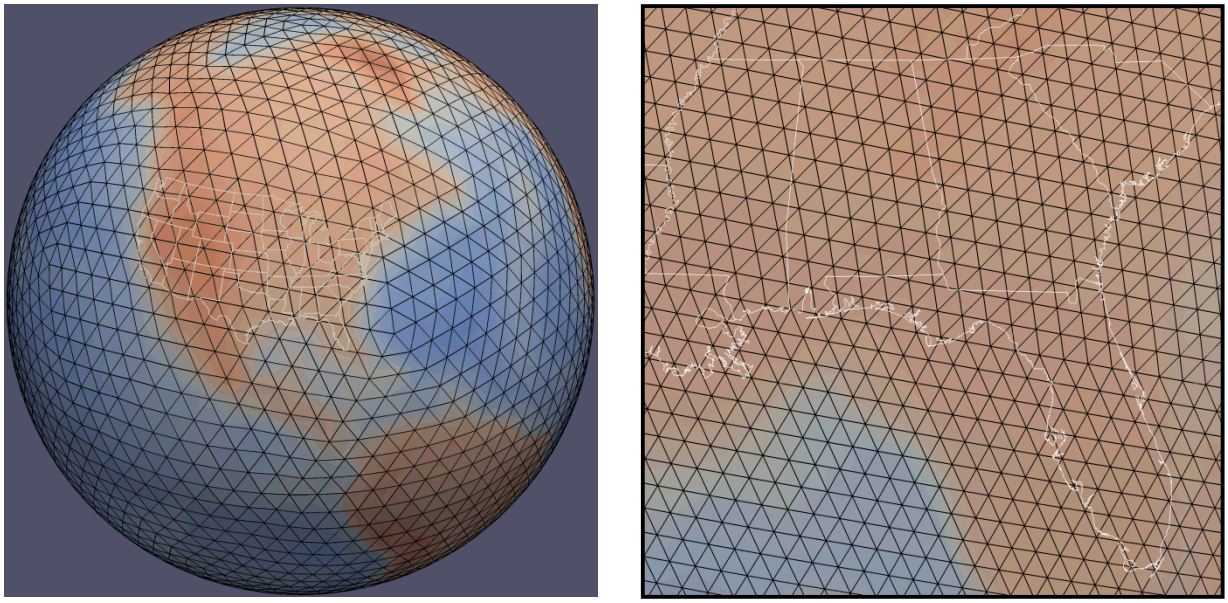

Figure 1. Left: **Coarse event location grid (4°).** Right: **Actual ½° grid spacing (over southeastern United States).**

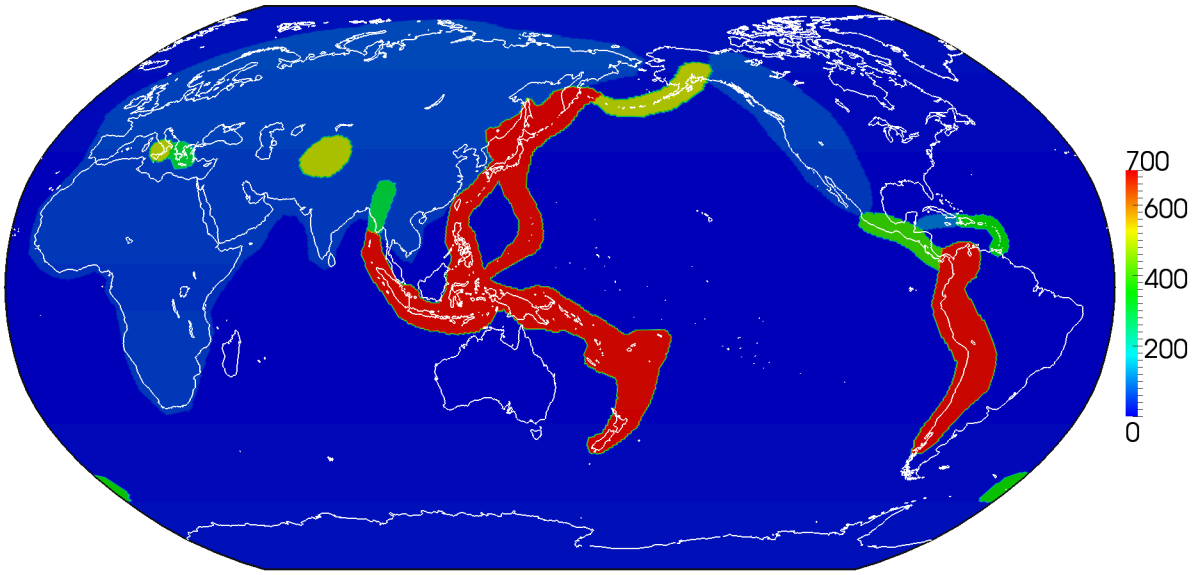

Figure 2. Depth grid based on seismicity.

Event Detection and Location using Spatial Fitness

Hypothetical events are located by identifying the largest "fitness" value calculated for each of the nodes in the PEDAL Earth grid. Consider an arbitrary grid node ω and event origin time T_o and a single arrival A_i, from station

S_i, that includes observations (attributes) of arrival time (T_i.), azimuth (az_i), and horizontal slowness (sh_i). The conditional fitness, proportional to the probability that A_i was generated by a seismic event, $E_{\omega,T0}$, which occurred at (ω, T_o), is given by

$$F(A_i \mid E_\omega) = \exp\left[-\left(\frac{T_i - T_0 - p_{t,i,\omega}}{\varepsilon_{t,i,\omega}}\right)^2\right] \bullet \exp\left[-\left(\frac{az_i - p_{az,i,\omega}}{\varepsilon_{az,i,\omega}}\right)^2\right] \bullet \exp\left[-\left(\frac{sh_i - p_{sh,i,\omega}}{\varepsilon_{sh,i,\omega}}\right)^2\right] \quad (1)$$

where $p_{t,i,\omega}$, $p_{az,i,\omega}$ and $p_{sh,i,\omega}$ are the expected travel time, azimuth and horizontal slowness for an event at ω, and $\varepsilon_{t,i,\omega}$, $\varepsilon_{az,i,\omega}$, and $\varepsilon_{sh,i,\omega}$, are tolerance values. Figure 3 shows a plot of the implied Gaussian distribution associated with each attribute. Note that the product of the attribution distributions will at most be equal to the maximum individual attribute value. Therefore, if any attribute value is small, then the contribution to spatial fitness will be small.

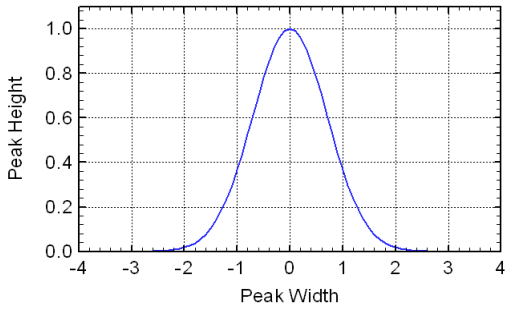

Figure 3. General attribute distribution: $P(a_i, E_\omega)$ = probability of arrival attribute a_i from station i caused by an event at location ω.

Spatial Fitness using Pairs of Arrivals

For computational efficiency, PEDAL uses pairs of arrivals to compute the spatial fitness values at each grid node. For all events E_ω, ($T_i - T_j$) observed at two stations S_i and S_j is independent of event origin time. So the joint fitness of any two arrivals is

$$F(A_i \cap A_j \mid E_\omega) = \exp\left[\begin{array}{l} -\dfrac{\left((T_i - T_j) - (p_{t,i,\omega} - p_{t,j,\omega})\right)^2}{\varepsilon_{t,i,\omega}^2 + \varepsilon_{t,j,\omega}^2} \\ -\left(\dfrac{az_i - p_{az,i,\omega}}{\varepsilon_{az,i,\omega}}\right)^2 - \left(\dfrac{az_j - p_{az,j,\omega}}{\varepsilon_{az,j,\omega}}\right)^2 \\ -\left(\dfrac{sh_i - p_{sh,i,\omega}}{\varepsilon_{sh,i,\omega}}\right)^2 - \left(\dfrac{sh_j - p_{sh,j,\omega}}{\varepsilon_{sh,j,\omega}}\right)^2 \end{array}\right] \quad (2)$$

where subscripts t refer to travel time and T to arrival time. Now we sum the conditional fitnesses of all pairs of arrivals in a specified time interval (23 minutes)

$$F_\omega = \sum_{i=1}^{NA-1} \sum_{j=i+1}^{NA} F(A_i \cap A_j \mid E_\omega) \quad (3)$$

and search 3D grid space for the point with the highest F_{ω}. Note that if there are arrivals from more than one event in the time interval (not uncommon for a global monitoring system), then there will be multiple peaks in the 3D grid space. Our goal at this stage is to find the overall highest peak, i.e. the event with the most arrivals.

Once the overall peak has been found, the next step in PEDAL is to search the time axis for an origin time, T_o, with maximum temporal fitness.

$$F_{\omega,T_0} = \sum_{i=1}^{NA} \exp\left[-\left(\frac{(T_i - T_0) - p_{t,i,\omega}}{\varepsilon_{t,i,\omega}} \right)^2 - \left(\frac{az_i - p_{az,i,\omega}}{\varepsilon_{az,i,\omega}} \right)^2 - \left(\frac{sh_i - p_{sh,i,\omega}}{\varepsilon_{sh,i,\omega}} \right)^2 \right] \tag{4}$$

Thus, we have identified both the spatial and temporal position of an event in our time sequence.

Association of Arrivals with Events

After an event, $O_{\omega,To}$, is detected, located, and its origin time established, the arrivals that this event generated are identified and associated with the event. Given a detected event at location ω and origin time T_o, we associate arrivals with $O_{\omega,To}$ using phase-specific predictions and tolerance values.

$$F_{ph,\omega,T_O} = \sum_{i=1}^{NA} \exp\left[\begin{array}{l} -\dfrac{\left((T_i - T_O) - p_{t,i,\omega,ph}\right)^2}{\varepsilon^2_{t,i,\omega,ph}} \\ -\left(\dfrac{az_i - p_{az,i,\omega,ph}}{\varepsilon_{az,i,\omega,ph}} \right)^2 \\ -\left(\dfrac{sh_i - p_{sh,i,\omega,ph}}{\varepsilon_{sh,i,\omega,ph}} \right)^2 \end{array} \right] \tag{5}$$

For each unassociated arrival in current time window, we find the phase, *ph*, for which $F_{ph,\omega,To}$ is greatest. If $F_{ph,\omega,To}$ is greater than an established threshold, then we associate the arrival with $O_{\omega,To}$ and remove arrival from the window.

Once all the appropriate arrivals have been associated with the current event, we recalculate the fitness for the 3D grid with the remaining arrivals and look for another peak. The process of identifying events and associating arrivals based on travel time, azimuth, and slowness continues until there are less than two arrivals in the window or the peak spatial fitness is less than an established threshold.

At this point, one final association step is performed, based solely on travel time, in an attempt to sweep up any remaining arrivals that can be associated with our events. The fitness value is computed for each detected event and, assuming the highest fitness value is greater than an established threshold, the arrival is associated with the event having the highest fitness.

$$F_{ph,\omega,T_O} = \sum_{i=1}^{NA} \exp\left[-\frac{\left((T_i - T_O) - p_{t,i,\omega,ph}\right)^2}{\varepsilon^2_{t,i,\omega,ph}} \right] \tag{6}$$

PEDAL Process

For the following discussion, refer to

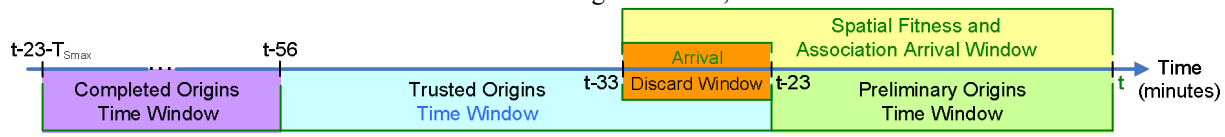

Figure 4 and Figure 5. The idea is to process a time interval of detections, and from these find events to add to the event bulletin. However, we do not want to add an event to the bulletin unless we are certain that there are no additional arrivals that might be in the next interval to process, i.e. that the event is truly "final" or complete. We, therefore, establish three consecutive time windows to describe origins at varying degrees of processing.

1. *Preliminary Origins* – This time window includes origins that have been detected by virtue of peak spatial fitness, but are preliminary because their origin times are greater than t-23 minutes or their sets of associated arrivals do not pass the trusted origin criteria given below. Preliminary origins will be labeled as such, but their associated arrivals will be returned to the arrival list so that this origin can be detected again in the next time interval, potentially with a richer set of arrivals contributing to its detection.

2. *Trusted Origins* – This time window includes origins that are considered trusted because its set of arrivals passes the trusted origin criteria, but their origin times are such that additional arrivals of secondary phases associated with them could exist in the future (next time interval). Note that trusted origins cannot have any more first-P arrivals associated with them.

3. *Completed Origins* – This time window includes origins that are considered complete and final because all possible arrivals that could be associated with these origins have been seen by virtue of their origin times and the travel time of the latest arriving expected secondary phases (T_{Smax}).

For our testing, PEDAL operates on IDCX arrivals in the 33-minute Spatial Fitness Arrival Window, A', from 151 primary and secondary IMS seismic stations (

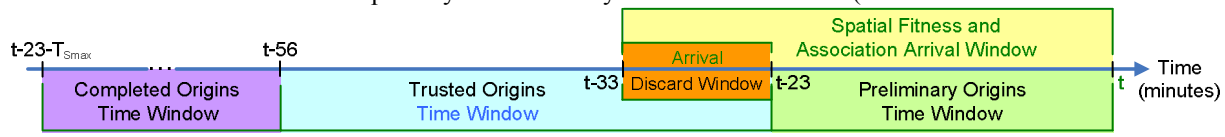

Figure 4). For the spatial fitness calculation, first-P predictions are used and first-P arrivals take a maximum of 23 minutes to travel from any grid node to a station on the opposite side of the Earth, hence all arrivals within a 23-minute window must be processed. However, because we will be trying to find events within a 10-minute span (the Discard Window), the total arrival window to process is 23+10 = 33-minutes. Following identification of the location of a hypothetical event, the origin time is computed using first-P predictions and arrivals in the Spatial Fitness Arrival Window (see the Origin Time Window in

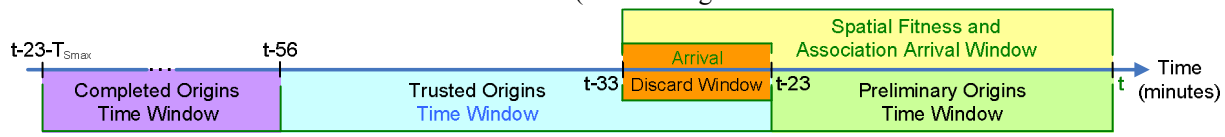

Figure 4). At this point, association of all arrivals is performed on the hypothetical event. All associated arrivals are removed from the arrival list and a search for new hypothetical events is repeated until no more seismic events are identified that satisfy the minimum fitness criteria.

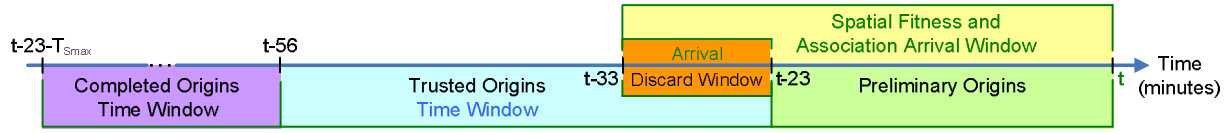

Figure 4. PEDAL arrival and time windows.

As we discussed above, once all hypothetical events have been indentified for the current time window, a final association step is performed based on arrival time fitness alone. Each remaining unassociated arrival is fitted with

each hypothetical origin and the arrival is associated with the origin that results in the highest fitness, assuming a minimal fitness threshold. At this point, we are ready to move on to the next time interval. First, however, we must add to the final event bulletin events that satisfy the following criteria.

Trusted Origin Criteria: Test each hypothetical origin against the following criteria, in order:

- At least one of its associated arrivals is in the Discard Window. This indicates that the event cannot have additional P arrivals beyond A'. If there is not at least one associated arrival in the Discard Window, then return all arrivals to A'. This event will likely get formed again in the next processing interval.
- Has at least two associated arrivals (as a result of the previous step, at least one of them will be in the Discard Window). This test insures that all events must have more than one associated arrival. If so, then add origin to the permanent set of origins. If there are not two or more associated arrivals, then return all arrivals to A'. This event will likely get re-built in the next processing interval.

Arrivals associated with origins that don't satisfy the criteria are placed back into the arrival list, A', to be processed in the next time interval or permanently stored as unassociated arrivals, depending on the arrival times. The entire iterative PEDAL process is depicted in a flowchart in Figure 5.

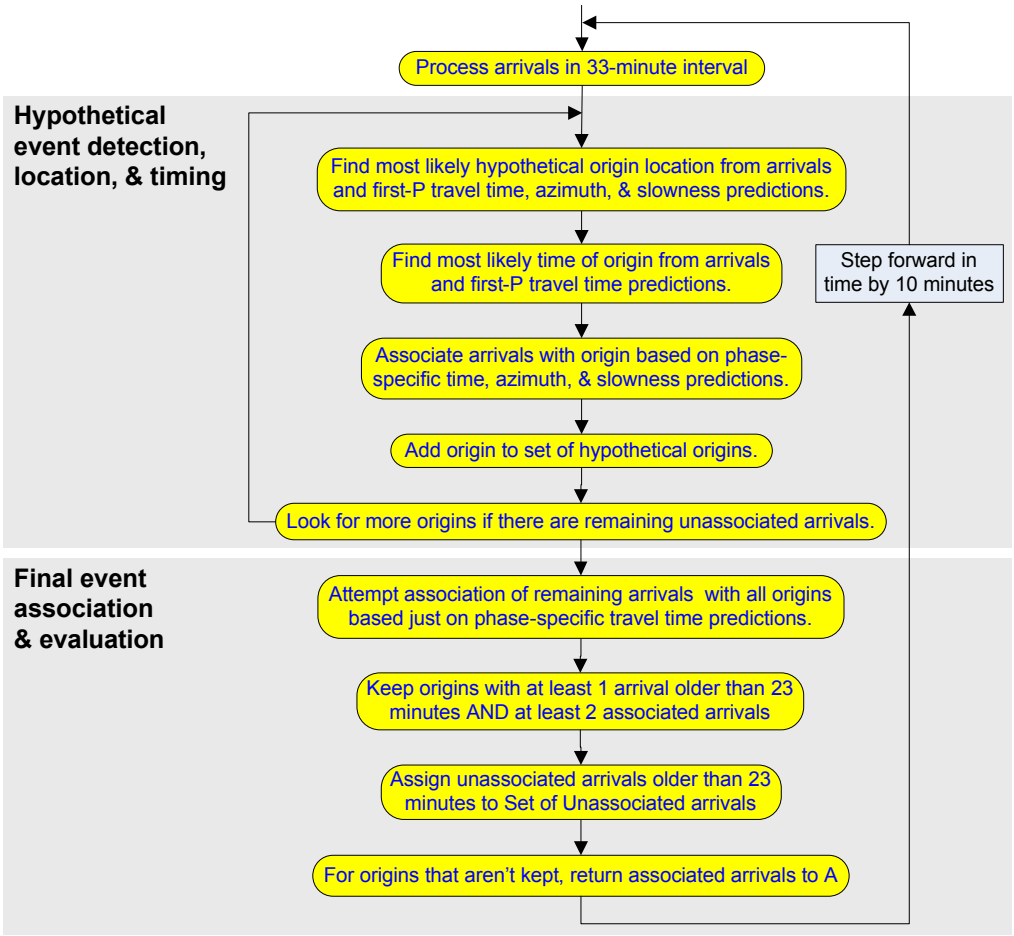

Figure 5. PEDAL flowchart.

Computational Issues and GPUs

Although the calculation of the spatial fitness values must be done for each of the 383,753 grid nodes, each grid node's calculation is independent of all the others, so the problem is fully parallelizable. Each node has a unique set of prediction values (expected value and tolerance) for each attribute (travel time, azimuth, and slowness) and for

each of 151 stations. This results in a memory requirement to store all predictions of ~1.3 gigabytes. During PEDAL operation, each node will use the same set of arrivals in its computation as every other node.

The spatial fitness calculation requires prediction values (expected values and tolerances) of first-P travel time, azimuth, and slowness at every node for each station. The spatial fitness calculation is $O(n^2)$, where n is the total number of arrivals. For $n = 100$, the calculation involves evaluation of the exponential function approximately 2 billion times. For $n = 500$, it must be evaluated approximately 47 billion times. Recall that PEDAL performs the spatial fitness calculation in three dimensions followed be a temporal fitness calculation in one dimension. Calculation of fitness in four dimensions over a 70 minute window with 0.1 second time steps would require the evaluation of the exponential function approximately 2 quadrillion times, hence our preference for considering pairs of arrivals.

Because the exhaustive search approach is computationally expensive using sequential processing, we use Graphics Processing Units (GPUs) to perform this task. We currently use an NVidia Tesla C1060 Computing Processor Board, which has 240 processor cores and 4 gigabytes of memory. Figure 6 shows the extreme difference in computational cost between sequential and parallel processing of the spatial fitness function. Figure 7 shows the impact of using multiple GPU boards for the spatial fitness calculation.

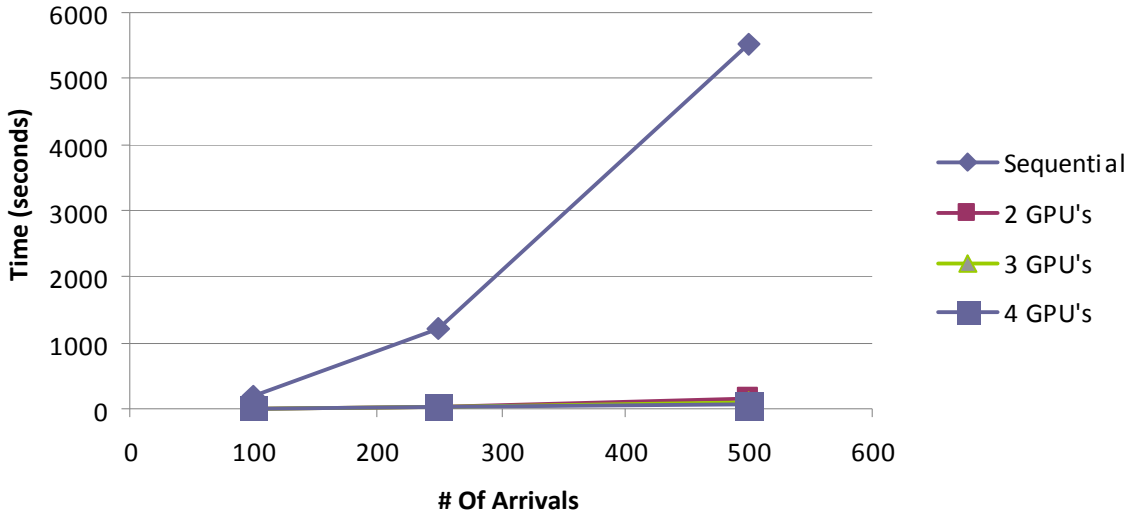

Figure 6. Sequential processing vs. GPU processing.

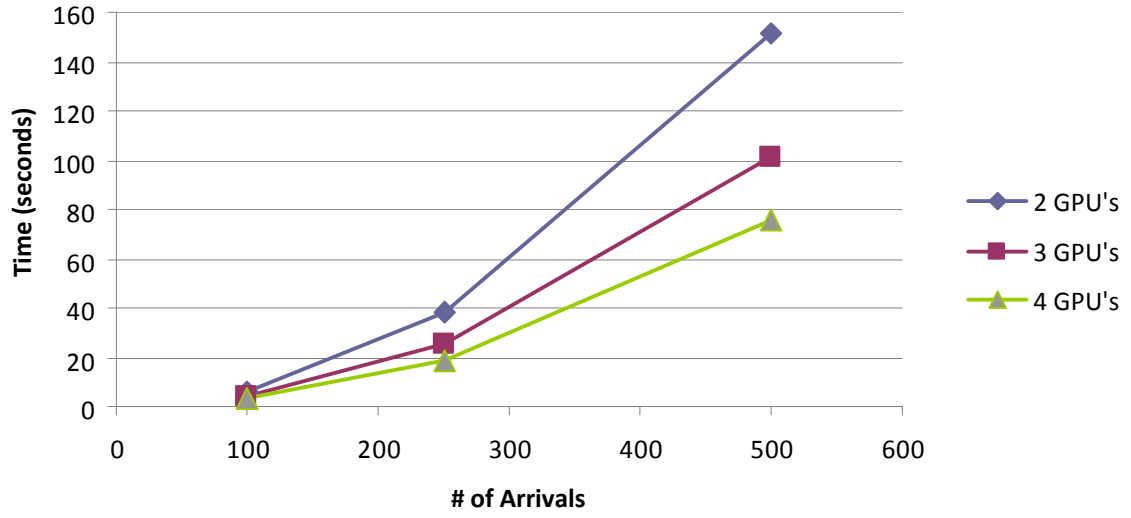

Figure 7. Processing with different numbers of GPUs.

Performance Evaluation and Experimental Results

PEDAL was evaluated on a 60-minute window of IDCX arrivals from December 18, 2008, where 5 origins of a variety of magnitudes exist in the LEB.

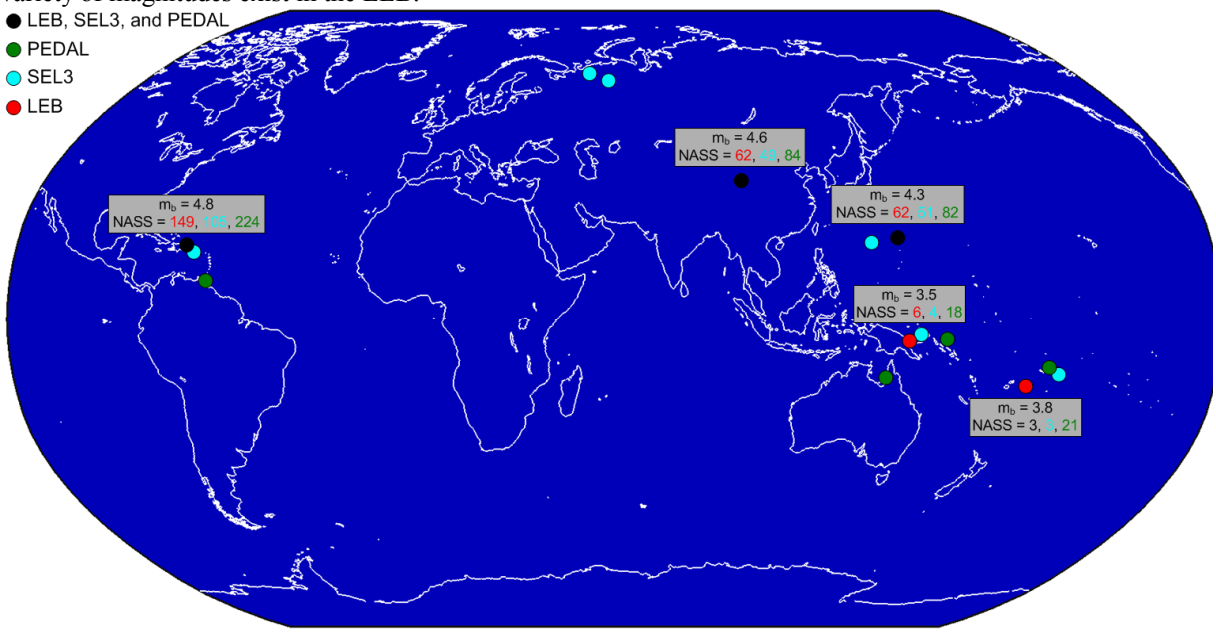

Figure 8 shows a map of these events along with their magnitudes and number of associated arrivals, which is a more robust measure of the quality of an event. We show LEB, SEL3, and PEDAL results. Where the bulletin events are in nearly the same location, a single black circle is shown. Where they deviate, separate symbols are shown for LEB (red), SEL3 (cyan), and PEDAL (green). In the accompanying text boxes, the number of associated arrivals is shown for each bulletin. Classifying these events by the LEB number of associated arrivals, we can see that both PEDAL and GA (SEL3) clearly find the NASS=149, 62, and 62 events. We note that the number of associated phases for PEDAL is much higher than either LEB or GA, probably reflecting some incorrect secondary phase associations. Event detection using first-P phase predictions has been the focus of development for PEDAL thus far. Our attention will turn next to high quality association of secondary phases. PEDAL and GA also seem to find the NASS=3 event in the SW Pacific, although with imprecise location, but PEDAL does struggle with the NASS=6 event in Indonesia. It identifies two events, the "average" of which may result in a good detection. Post processing of event detections and phase associations is another important area of future work. Finally, GA detects two extraneous events in northern Europe.

For more insight, Figure 9 shows the results of four iterations of the PEDAL process. At each iteration of the PEDAL event detection, location, and association process, one can see the location of the largest spatial fitness, indicating the largest event resulting from the current set of arrivals and other smaller peaks that are hints of other events of lesser magnitude. At iteration 2, arrivals associated with the previously identified event in Iteration 1 are no longer available for the current spatial fitness calculation. By iteration 4, it is difficult to identify peaks (potential real events) above the background, though we can still see the peak corresponding to the event at the northern edge of South America. After removal of this event, identification of the next highest peak is very ambiguous. Therefore, a spatial fitness threshold is used as a criteria to indicate when to advance the PEDAL time window into the future.

Evaluating the performance of an associator is a very difficult task. Establishing the true set of detectable events is generally not clear. Even deciding that events from two different bulletins are the same is not always straightforward. Evaluating proper associations is challenging as well. PEDAL processed arrivals in approximately 20 minutes using a single GPU in a desktop computer and generated 7 events, matching all 5 of the LEB events to some degree, but also clearly generating some false events. Both SEL3 and PEDAL are tuned to not miss legitimate

events. As a result, it is not uncommon for the automated systems (SEL3 and PEDAL) to find multiple events near an event listed in LEB after review by analysts. Therefore, one next step is to develop an algorithm for combining events that are close together in both time and space. Utilizing historical data in establishing the probability that an event originate from a particular location on the Earth is another capability that will greatly improve PEDAL's ability to eliminate bogus events while retaining legitimate ones.

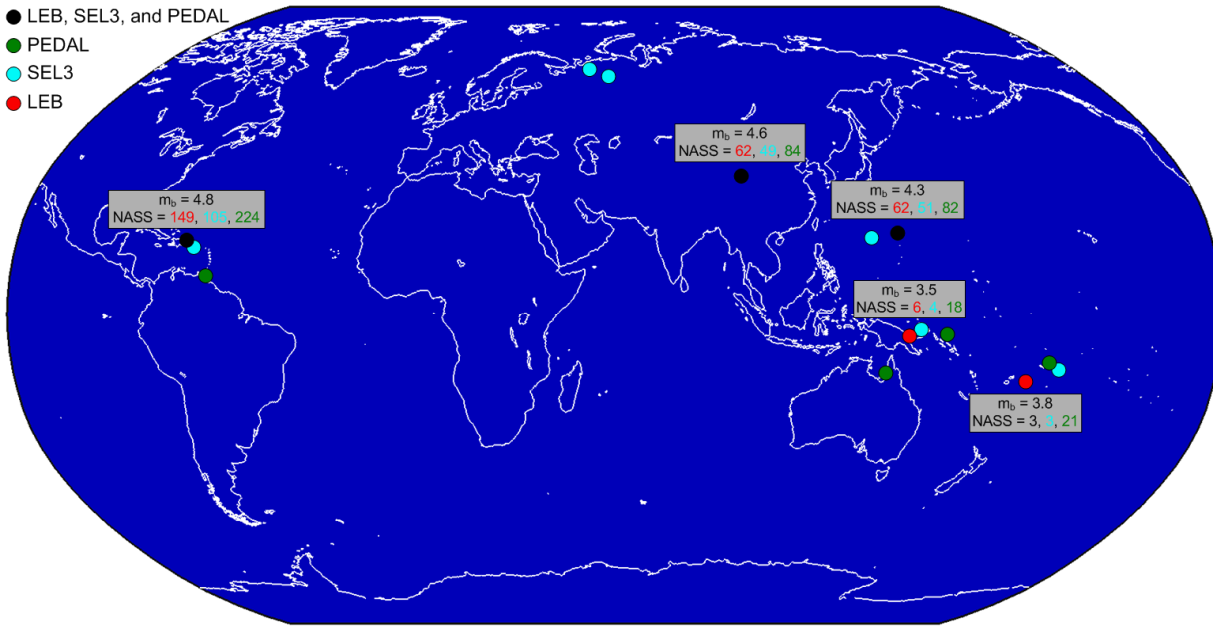

Figure 8. Associator results from a 60-minute time window on 12/18/2008 with 5 LEB events in red and black. Black points indicate that PEDAL and SEL3 detected events collocated with the LEB event. Green (PEDAL) or cyan (SEL3) dots close to a black dot indicate an additional "split" event by the associator.

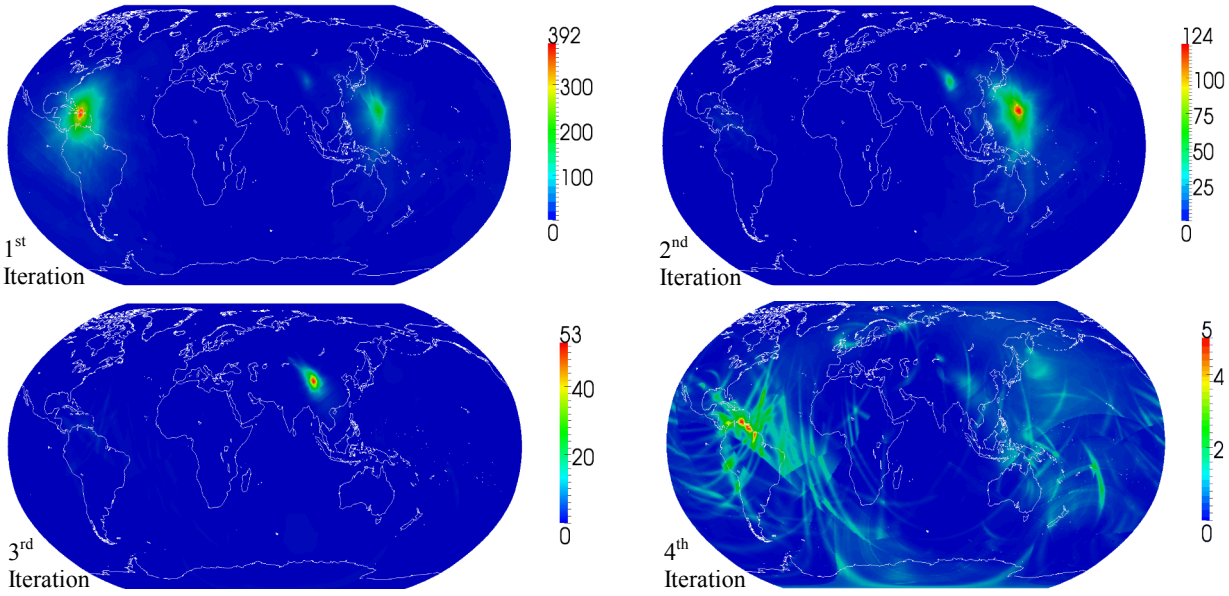

Figure 9. Spatial fitness maps from first four iterations of PEDAL.

CONCLUSIONS AND RECOMMENDATIONS

We are developing a new signal association algorithm using exhaustive search of 4-dimensional parameter space (3 spatial dimensions and time). We identify event locations from the maximum 'fitness' assuming all arrivals are P phases. Associations of arrivals to located events are based on the maximum fitness over all possible phases. Where historical data exists, empirical expected values and tolerances are established for use in the fitness calculations. The use of GPUs allows PEDAL to run in real-time on a desktop computer.

Performance on small time windows of IDCX arrivals is promising. However, testing over large periods of time that includes large events and aftershock sequences is necessary. Various parameters used in the PEDAL process must be optimized for peak performance. An important task that we must address for measuring associator performance is the establishment of a good metric for evaluating the quality of PEDAL processing, which must consider both missed events and false events. Also, we will develop a method to allow us to calculate the probability that an event actually occurred for each of the events that PEDAL produced. While we know that fitness is proportional to probability, the proportionality relationship is not simple. Actual probabilities will be much easier to interpret than fitness values in deciding what to do with an event produced by PEDAL. Further, a Bayesian probabilistic calculation should allow us to incorporate important prior information on the likelihood of events occurring in specific regions, which is definitely not the same for all locations.

REFERENCES

Le Bras, R., H. Swanger, T. Sereno, G. Beall, R. Jenkins, W. Nagy, and A. Henson (1994). Global Association Final Report, SAIC Technical Report SAIC-94/1155.