ENHANCING SEISMIC CALIBRATION RESEARCH THROUGH SOFTWARE AUTOMATION AND SCIENTIFIC INFORMATION MANAGEMENT

Stanley D. Ruppert, Douglas A. Dodge, Annie B. Elliott, Michael D. Ganzberger, Teresa F. Hauk, and Eric M. Matzel

Lawrence Livermore National Laboratory

Sponsored by National Nuclear Security Administration
Office of Nonproliferation Research and Engineering
Office of Defense Nuclear Nonproliferation

Contract No. W-7405-ENG-48

ABSTRACT

The National Nuclear Security Administration (NNSA) Ground-Based Nuclear Explosion Monitoring Research and Engineering (GNEM R&E) program has made significant progress enhancing the process of deriving seismic calibrations and performing scientific integration with automation tools. We present an overview of our software automation and scientific data management efforts and discuss frameworks to address the problematic issues of very large datasets and varied formats utilized during seismic calibration research. The software and scientific automation initiatives directly support the rapid collection of raw and contextual seismic data used in research, provide efficient interfaces for researchers to measure and analyze data, and provide a framework for research dataset integration. The automation also improves the researchers ability to assemble quality controlled research products for delivery into the NNSA Knowledge Base (KB). The software and scientific automation tasks provide the robust foundation upon which synergistic and efficient development of GNEM R&E program seismic calibration research may be built.

The task of constructing many seismic calibration products is labor intensive and complex, hence expensive. However, aspects of calibration product construction are susceptible to automation and future economies. We are applying software and scientific automation to problems within two distinct phases or "tiers" of the seismic calibration process. The first tier involves initial collection of waveform and parameter (bulletin) data that comprise the "raw materials" from which signal travel-time and amplitude correction surfaces are derived and is highly suited for software automation. The second tier in seismic research content development activities includes development of correction surfaces and other calibrations. This second tier is less susceptible to complete automation, as these activities require the judgment of scientists skilled in the interpretation of often highly unpredictable event observations. Even partial automation of this second tier, through development of prototype tools to extract observations and make many thousands of scientific measurements, has significantly increased the efficiency of the scientists who construct and validate integrated calibration surfaces. This achieved gain in efficiency and quality control is likely to continue and even accelerate through continued application of information science and scientific automation.

Data volume and calibration research requirements have increased by several orders of magnitude over the past decade. Whereas it was possible for individual researchers to download individual waveforms and make time-consuming measurements event by event in the past, with the terabytes of data available today, a software automation framework must exist to efficiently populate and deliver quality data to the researcher. This framework must also simultaneously provide the researcher with robust measurement and analysis tools that can handle and extract groups of events effectively and isolate the researcher from the now onerous task of database management and metadata collection necessary for validation and error analysis. Lack of information-management robustness or loss of metadata can lead to incorrect calibration results in addition to increasing the data-management burden. To address these issues, we have succeeded in automating several aspects of collection, parsing, reconciliation, and extraction tasks, individually. Several software automation prototypes have been produced and have resulted in demonstrated gains in efficiency of producing scientific data products. Future software automation tasks will continue to leverage database and information-management technologies in addressing additional scientific calibration research tasks.

OBJECTIVES

The National Nuclear Security Administration (NNSA) Ground-Based Nuclear Explosion Monitoring Research and Engineering (GNEM R&E) program has made significant progress enhancing the process of deriving seismic calibrations and performing scientific integration with automation tools. We present an overview of our software automation efforts and framework to address the problematic issues of very large datasets and varied formats utilized during seismic calibration research and the attributes required to construct next-generation data acquisition. The scientific automation engineering and research will need to provide the robust hardware, software, and data infrastructure foundation for synergistic GNEM R&E program calibration efforts. The current task of constructing many seismic calibration products is labor intensive and complex, hence expensive. However, aspects of calibration-product construction are susceptible to automation and future economies. Data volume and calibration research requirements have increased by several orders of magnitude over the past decade. We have succeeded in automating many of the collection, parsing, reconciliation, and extraction tasks, individually. Several software automation prototypes have been produced and have resulted in demonstrated gains in efficiency of producing scientific data products. In order to fully exploit voluminous real-time data sources and support new requirements for time critical modeling, simulation, and analysis, a more scalable and extensible computational framework will be required.

RESEARCH ACCOMPLISHED

The primary objective of the Scientific Automation Software Framework (SASF) efforts is to facilitate development of information products for the Ground-Based Nuclear Explosion Monitoring Research and Engineering (GNEM R&E) regionalization program. The SASF provides efficient access to, and organization of, large volumes of raw and derived parameters, while also providing the framework to store, organize, integrate and disseminate information products for delivery into the National Nuclear Security Administration Knowledge Base (NNSA KB). The current framework supports integration, synthesis, and validation of the various different information types and formats required by each of the seismic calibration technologies (Figure 1). For example, the seismic location technology requires parameter data (site locations, bulletins), and time-series data (waveforms) and produces parameter measurements in the form of arrivals, gridded geospatially registered corrections surfaces and uncertainty surfaces through the use of various tools and information-processing frameworks (relational databases (RDBs), Geographical Information Systems (GISs), and associated product and data visualization and data management tools (e.g., RBAP, KBALAP, KBCIT, DM).

These information-management and scientific automation tools are used together within specific seismic calibration processes to support production of tuning parameters for the United States Atomic Energy Detection System operated by the Air Force (Figure 2). The calibration processes themselves appear linear (Figure 2), beginning with data acquisition and extending from reconciliation, integration, and measurement and simulation to construction of calibration / run-time parameter products. Efficient production of calibration products, however, requires extensive synergy and synthesis not only between data-types (Figure 1), but also between measurements and results derived from the different calibration technologies (e.g., Location, Identification, Detection) (Figures 1 and 2). Even with successful implementation of automation within many of the individual steps, the current infrastructure will not scale to handle order-of-magnitude additional data or extend to handle time-critical data acquisition or analysis. This lack of scalability and flexibility limits efficient production and delivery of run-time calibrations to the operational seismic monitoring pipeline (Figure 2, bottom) as a large manual effort is still required to acquire and integrate streaming (10–20 GB/day) signals with associated metadata. This synergy and synthesis between complex tools and very large datasets is critically dependent on having a scalable and extensible unifying framework. These requirements of handling large datasets in diverse formats and facilitating interaction and data exchange between tools supporting different calibration technologies led to an extensive scientific automation software engineering effort to develop an object-oriented database-centric framework (Figure 3) as a unifying foundation.

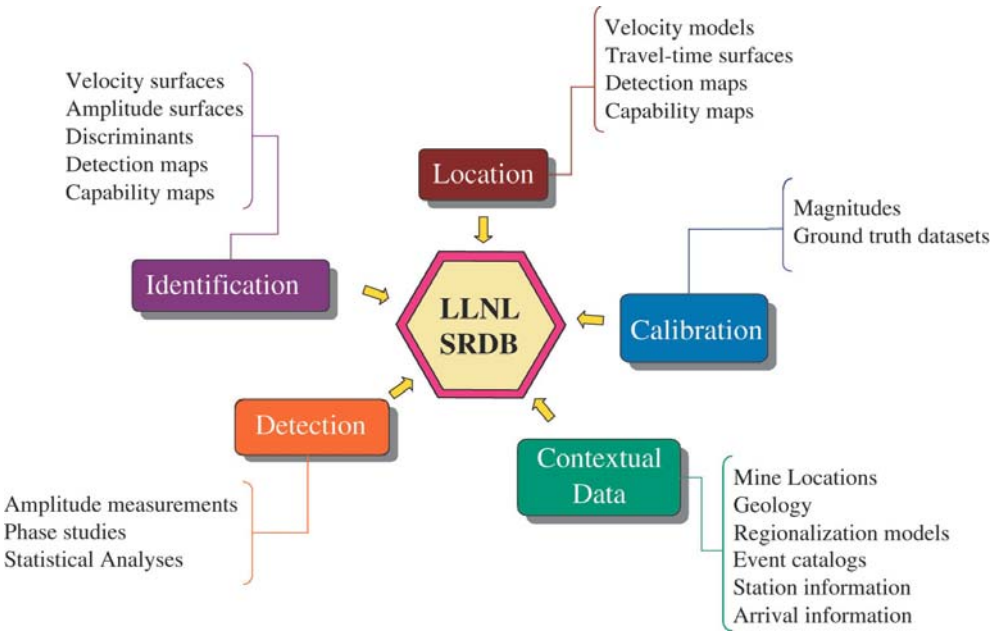

Figure 1. The Scientific Automation Software Framework provides a unifying framework for contextual/reference data and information products.

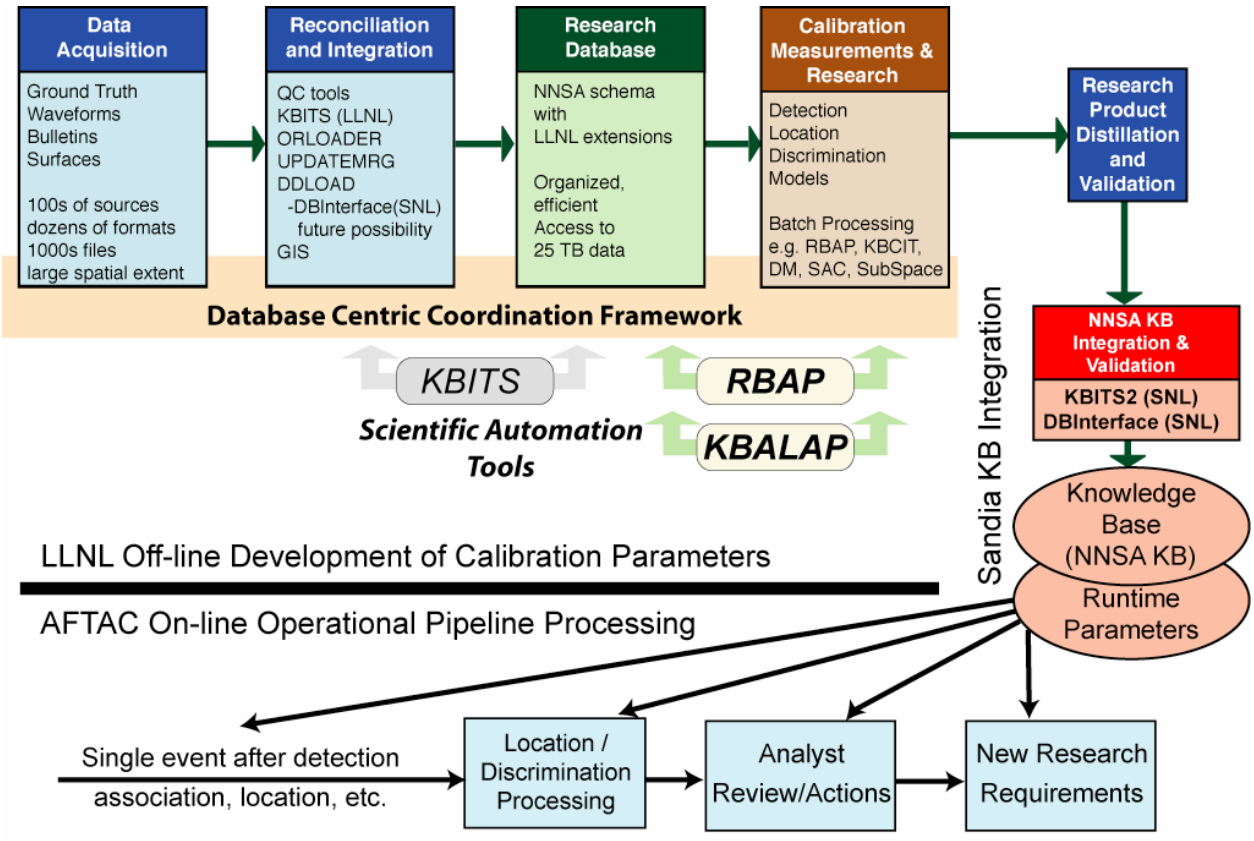

Figure 2: Summary of the processes of data collection, research, and integration within the LLNL calibration process that result in contributions to the NNSA KB. The relationships of the current LLNL calibration tools, scientific automation tools, and database coordination framework to those involved in the assembly of the NNSA KB are delineated.

Scientific Automation Software Tools

Information products created using the Lawrence Livermore National Laboratory (LLNL) Seismic Research Database (SRDB) may be grouped under two major categories or tiers: Tier 1 - primary data products, and Tier 2, derived products. In order to calibrate seismic monitoring stations, the LLNL SRDB must incorporate and organize the following categories of primary and derived measurements, data, and metadata:

Tier 1: Contextual and Raw Data
Station Parameters and Instrument Responses
Global and Regional Earthquake Catalogs
Selected Calibration Events
Event Waveform Data
Geologic/Geophysical Datasets
Geophysical Background Model

Tier 2: Measurements and Research Results
Phase Picks
Travel-time and Velocity Models
Rayleigh and Love Surface Wave Group Velocity Measurements
Phase Amplitude Measurements and Magnitude Calibrations
Detection and Discrimination Parameters

Automating Tier 1

Corrections and parameters distilled from the calibration database provide needed contributions to the NNSA KB for the Middle East/North Africa/Western Europe (ME/NA/WE) region and will improve capabilities for underground nuclear explosion monitoring. The contributions support critical functions in detection, location, feature extraction, discrimination, and analyst review. Figure 2 outlines the processes of data collection, research, and integration within the LLNL calibration process that result in contributions to the NNSA KB and the relationship of the LLNL calibration tools to those involved in the assembly of the NNSA KB or within the AFTAC operational pipeline. Within the major process steps (data acquisition, reconciliation/integration, calibration research, product distillation) are many labor intensive and complex steps. The previous bottleneck in the calibration process was in the reconciliation/integration step (Figure 2). This bottleneck became acute in 1998, and the KBITS suite of automated parsing, reconciliation, and integration tools for both waveforms and bulletins (ORLOADER, DDLOAD, UpdateMrg) was developed. The KBITS suite provided the additional capability required to integrate data from many data sources and external collaborations. Data volumes grew from the 11,400 events / 1 million waveforms in 1998 to the 6 million events / 70 million segmented waveforms and terabytes of continuous data today (e.g., Ruppert et al., 1999, Ruppert et al., 2004). This rapid increase in stored parameters soon led to new bottlenecks hindering rapid development and delivery of calibration research.

Automating Tier 2

As the number of data sources required for calibration increased in number and source location, it became clear that the manual and labor-intensive process of humans transferring thousands of files and unmanageable metadata could not keep the KBITS software fed with data to integrate, nor could the seismic research efficiently find, retrieve, validate, or analyze the raw parameters necessary to effectively produce seismic calibrations in an efficient manner. Significant software engineering and development efforts were applied to address this critical need to produce software aids for the seismic researcher. Two scientific automation tool prototypes (RBAP, KBALAP) (Figure 2) are under development for seismic location and seismic identification calibration tasks.

Both of these prototypes include methods and aids for efficiently extracting groups of events and waveforms from the millions contained in the SRDB and making large numbers of measurements with metadata in a batch mode. The concept of event sets (groups of related seismic events or parameters that can be processed together, e.g., either station centric or event centric) was introduced as previous SAC scripts and macros could not scale to the task.

The KBALAP Program

The Knowledge Base Automated Location Assessment and Prioritization (KBALAP) program is a set of database services and a client application that combine to efficiently produce location ground truth (GT) data that can be used in the production of travel time correction surfaces and as part of the preferred event parameters used by other tools in our processing framework.

The part of KBALAP that runs as a database service is responsible for evaluating bulletin and pick information as it enters the system to identify origin solutions that meet predefined GT criteria with no further processing, and to identify events that would likely meet a predefined GT level if a new origin solution were produced using available arrivals. The database service is also responsible for identifying events that should have a high priority for picking based on their existing arrival distribution and the availability of waveform data for stations at critical azimuths and distances.

The interactive portion of KBALAP has three principal functions. These are

- interactive production of GT origins through prioritized picking and location,
- interactive specification of GT-levels for epicenter, depth, origin time, etype, and
- batch-mode location of externally-produced GT information.

The first of these capabilities allows the user to view epicenters and GT information on a map based on selection criteria input by the user. The user can select any GT or potential GT event and observe the distribution of stations with picks and stations with available waveforms. The user can select any station with available waveforms and open a picker with any current picks displayed. There the user can adjust existing picks, add new picks, mark bulletin picks as unusable, and relocate the event. A new GT level is calculated, and the user can choose to accept that origin solution and GT level or continue working with other stations.

The interactive GT entry mode of KBALAP allows the user to retrieve information about a specific event and add or update that event's GT parameters. The program can also create a new event with a GT level for cases where epicenter, time, depth and magnitude GT data are available. Similarly, the batch mode part of the program allows specification of flat files containing GT data for events already in the database.

The RBAP Program

The Regional Body-wave Amplitude Processor (RBAP) is a software tool to help automate the process of making amplitude measurements of regional seismic phases for the purpose of calibrating seismic discriminants at each station. RBAP generates station-centric raw and Magnitude Distance Amplitude Correction (MDAC) corrected Pn, Pg, Sn and Lg amplitudes along with their associated calibration parameters (e.g., phase windows, MDAC values, reference events, etc.) in database tables. It strictly follows the Working Group (WG) 2 standardized processing described in the MDAC White Paper (Walter et al., 2003) and it replaces the original collection of scripts described by Rodgers (2003). RBAP has a number of advantages over the previous scripts. It is much faster, significantly easier to use, scales more easily to a larger number of events and permits efficient project revision and updating through the database.

RBAP integrates the functions of the modules in the previous LLNL scripts into a single program that is designed to perform the amplitude measurement task efficiently and to require a minimum effort from the users for managing their data and measurements. For well-located events with pre-existing analyst phase picks, the user reviews for quality control and then generates all the amplitudes with just a few mouse clicks. For events needing more attention, the user has complete control over the process (e.g., window control, ability to mark bad data, define regions, define MDAC parameters, and define the events to be used in the overall calibration process). RBAP shortens the time needed by the researcher to calibrate each station while simultaneously allowing an increase the number of events that can be efficiently included. RBAP is fully integrated with the LLNL research database. Data are always read directly from the appropriate tables in the research database rather than from a snapshot, as was done in the previous system. All RBAP result tables have integrity constraints on the columns with dependencies on data in the LLNL research database. This design makes it very difficult for results produced by RBAP to be stale and also ensures that as the research database expands, RBAP automatically becomes aware of new data that should be processed. RBAP initial users will be LLNL WG 2 members working on Integrated Research Products for FY04.

Some RBAP Key Features

• *Based on WG 2 Standardized Algorithm*
-RBAP is built on the WG 2 standardized body-wave amplitude measurement algorithms documented in the "MDAC White Paper" (Walter et al., 2003). Its results are completely consistent with the last version of the LLNL scripts (Rodgers, 2003) that were vetting in the February 2003 WG 2 exercise between LLNL, LANL, and AFTAC.

• *Fast and Efficient Calibration*
RBAP is self-contained and optimized for station-centric body-wave processing. "Good" events can be handled with just a few mouse clicks. The researcher has direct control over key calibration parameters within the tool such as phase amplitude windows and migration, marking bad segments, defining distinct geophysical regions, event types to process, etc. We expect RBAP to provide roughly a factor of 5 increase in calibration speed compared with the original scripts, enabling us to calibrate more stations, with more events per station.

• *Project Management*
RBAP is designed so that a calibration project can be put down for a day, month, or year, and easily picked up, by the same researcher or a new one. All processing metadata are saved and events are easily tracked as processed, unprocessed, or outside the current project definitions. This allows a researcher to efficiently work through a huge data list without repetition and to easily identify and incorporate new events as they become available in the database.

• *Utilizes Database for Up-to-Date results*
RBAP can draw on the latest calibration parameters being generated by other working groups, such as the most recent phase picks, relocations, magnitudes, instrument response information, or event type ground truth.

• *Batch Processing*
RBAP is designed to allow simple batch updating of the amplitude results, whether the change is small (e.g., one-event is relocated) or large (instrument response is changed, affecting all events).

Database Centric Coordination Framework

As part of our effort to improve our efficiency, we have allowed researchers to easily share their results with one another. For example, as the location group produces GT information, that information should become available for other researchers to use. Similarly, phase arrival picks made by any qualified user should also become immediately available for others to use. This concept extends to the sharing of information about data quality. It should not be necessary for multiple researchers to have to repeatedly reject the same bad data. Rather, once data are rejected because of quality reasons, they should automatically be excluded from processing by all tools. We are implementing this system behavior using database tables, triggers, stored procedures, and application logic. Although we are at the beginning of this implementation, we have made significant progress over the last year with several kinds of information sharing using the new Database Centric Coordination Framework. These are discussed below.

Significant software engineering and development efforts have been applied successfully to construct an object-oriented database framework that provides database-centric coordination between scientific tools, users, and data (Figure 3). A core capability this new framework provides is information exchange and management between different specific calibration technologies and their associated automation tools such as Seismic Location (e.g. KBALAP), seismic identification (e.g., RBAP), and data acquisition / validation (e.g., KBITS). A relational database (ORACLE) provides the current framework for organizing parameters key to the calibration process from both Tier 1 (raw parameters such as waveforms, station metadata, bulletins, etc.) and Tier 2 (derived measurements such as ground-truth, amplitude measurements, calibration and uncertainty surfaces etc). Efforts are underway to augment the current relational database structure with semantic graph theory structured queries for handling complex queries.
Seismic calibration technologies (location, identification, etc.) are connected to parameters stored in the relational database by an extensive object-oriented multi-technology software framework (Figure 3, middle) that include elements of schema design, stored procedures, real-time transactional database triggers, and constraints, as well as

coupled Java and C++ software libraries to handle the information interchange and validation requirements. This software framework provides the foundation upon which current and future seismic calibration tools may be based.

Sharing of Derived Event Parameters

We have long recognized the inadequacies of the CSS3.0 origin table to serve as a source of information about the "best" parameters for an event. One origin solution may have the best epicenter but poor information on other parameters.

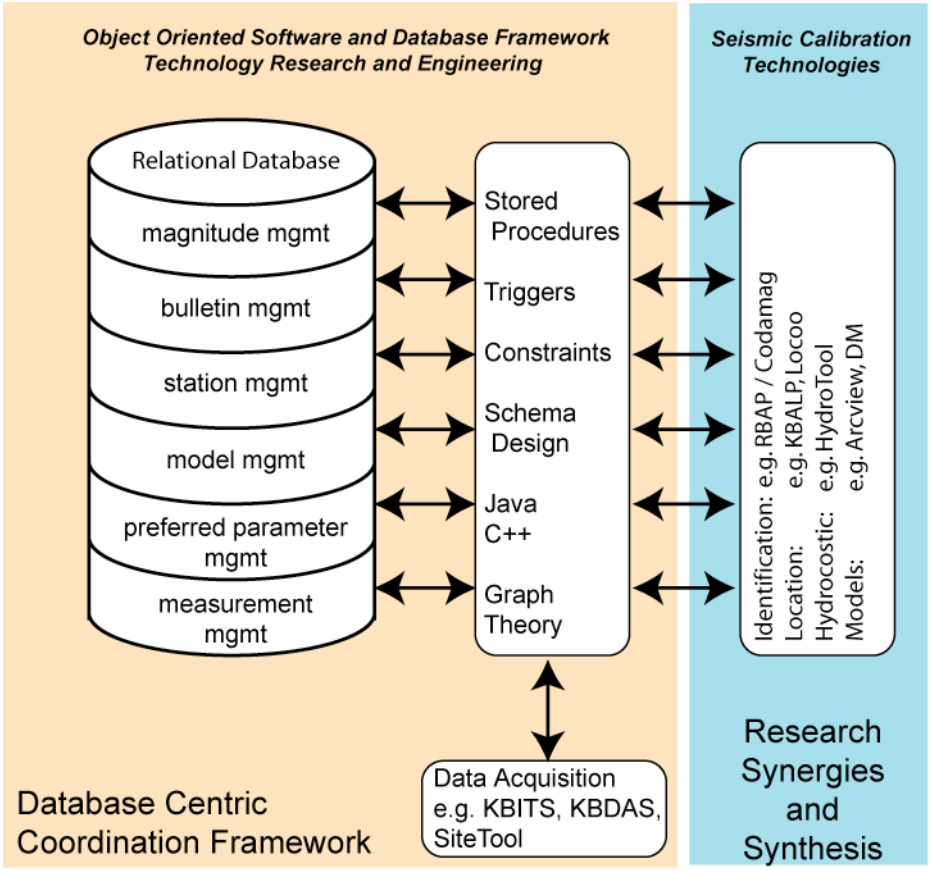

Figure 3. Overview of the Database Centric Coordination Framework that provides the enabling information technology to allow synergy and synthesis of data and calibration technologies for the efficient production of calibration deliverables.

Another may have the correct event type but be poor in other respects, and so on. We had discussed producing origin table entries, with our organization as the author, but that approach has difficulties. Different groups would have responsibility for different fields in the origin. Because their information would not be produced in synchronization, we would always have to be either updating the preferred origin or producing new preferred origins. Also, there would be difficulties in tracking the metadata associated with each field of the preferred origin. Our solution was to create a set of new tables and associated stored procedures and triggers that collectively maintain the "best" information about events.

Enhancements to Efficiency through Cluster-Based Computing

We have begun to leverage scalable and reconfigurable cluster computing resources to improve the efficiency of our computational infrastructure. Just as the database-centric approach to information management provided important gains in efficiency, we needed to move to a different computational paradigm to provide the computational power

necessary during calibration production and research. We have begun developing a set of flexible and extensible tools that are platform independent and parallelizable. These research tools will provide an efficient data processing environment for all stages of the calibration work flow, from data acquisition through making measurements to calibration surface preparation. This scalable and extensible approach (Figure 4) will result in more coupled and dynamic work flow in contrast to the linear work flow of the past, and allow more interaction between data, model creation, and validation processes.

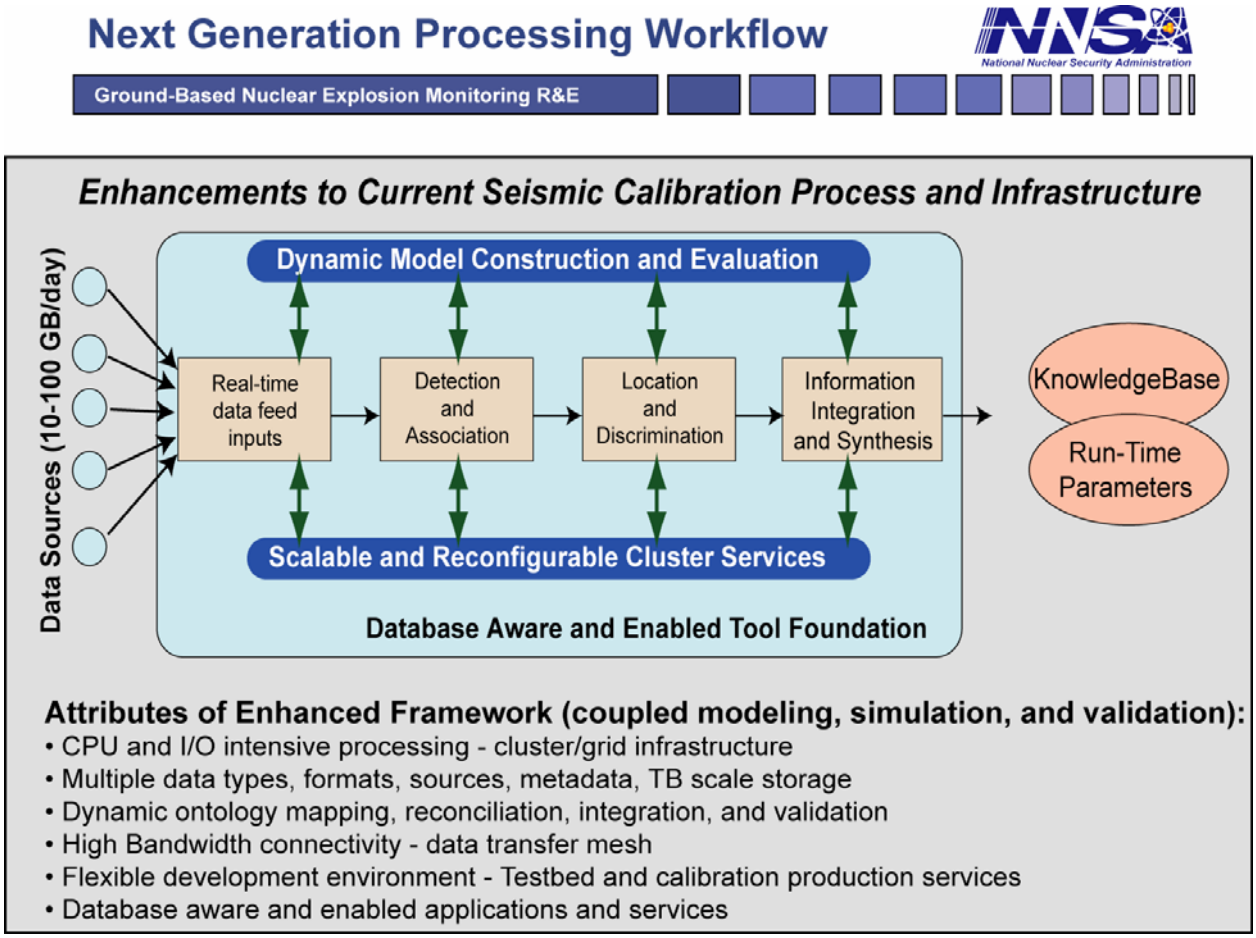

Figure 4. Overview of the cluster based calibration workflow under design and development.

Initial development and modification of existing codes and algorithms of the cluster based computing environment, has yielded significant efficiency improvements in RBAP and other measurement tools. Modification of RBAP to incorporate threads to isolate computationally intensive operations has provided a more interactive and responsive environment for the researcher and has laid the ground work for moving the threads to cluster-based computing resources. Other areas under investigation for taking advantage of cluster resources are for waveform correlation and subspace detector work, in addition to providing the ability to efficiently perform large-scale event relocations to evaluate ground truth and model calibrations.

CONCLUSIONS AND RECOMMENDATIONS

We present an overview of our software automation efforts and framework to address the problematic issues of very large datasets and varied formats utilized during seismic calibration research and the attributes required to construct next-generation data acquisition. By combining both a database centric information management system coupled with scalable and extensible cluster-based computing, we have begun to leverage a high-performance computational framework to provide increased calibration capability. These new software and scientific automation initiatives

could directly support our current mission, including rapid collection of raw and contextual seismic data used in research, providing efficient interfaces for researchers to measure/analyze data, and providing a framework for research dataset integration. The initiatives would improve time-critical data assimilation and coupled modeling/simulation capabilities necessary to efficiently complete seismic calibration tasks. The scientific automation engineering and research will need to provide the robust hardware, software, and data infrastructure foundation for synergistic GNEM R&E program calibration efforts.

ACKNOWLEDGEMENTS

We acknowledge the assistance of the LLNL computer support unit in implementing and managing our computational infrastructure. We thank Jennifer Aquilino and Laura Long for their assistance in configuration and installation of our Linux cluster.

REFERENCES

Ruppert, S., T. Hauk, J. O'Boyle, D. Dodge, and M. Moore (1999), Lawrence Livermore National Laboratory's Middle East and North Africa Research Database, in *Proceedings of the 21st Seismic Research Symposium: Technologies for Monitoring The Comprehensive Nuclear-Test-Ban Treaty*, LA-UR-99-4700, Vol. 1, pp. 234-242, Las Vegas, NV.

Ruppert, D. Dodge, A. Elliott, M. Ganzberger, T. Hauk, E. Matzel, and F. Ryall (2004), Enhancing seismic calibration research through software automation, in *Proceedings of the 26th Seismic Research Review: Trends in Nuclear Explosion Monitoring*, LA-UR-04-5801,Vol. 2, pp. 780-789, Orlando, FL.