ORNL/TM-2019/1140
CRADA/NFE-14-05227

CRADA Final Report: CRADA Number NFE-14-05227 with Total E&P Research and Technology USA LLC

David E. Bernholdt

2019-03-10

CRADA final report for
CRADA number NFE-14-05227

OAK RIDGE NATIONAL LABORATORY
MANAGED BY UT-BATTELLE FOR THE US DEPARTMENT OF ENERGY

ORNL/TM-2019/1140
CRADA/NFE-14-05227

Computer Science and Mathematics Division

CRADA Final Report: CRADA Number NFE-14-05227 with Total E&P Research and Technology USA LLC

David E. Bernholdt

Date Published: 2019-03-10

Prepared by
OAK RIDGE NATIONAL LABORATORY
Oak Ridge, Tennessee 37831-6283
managed by
UT-BATTELLE, LLC
for the
US DEPARTMENT OF ENERGY
under contract DE-AC05-00OR22725

Approved for Public Release

1. ABSTRACT

Cooperative Research and Development Agreement (CRADA) NFE-14-05227 between Oak Ridge National Laboratory (ORNL) and Total E&P Research and Development USA LLC (Total) focused on exploration of programming models and optimization opportunities for two families of seismic imaging algorithms frequently used within the oil and gas sector. Studies involved the use of OpenACC and OpenMP directive-based programming models for GPU accelerators on Reverse Time Migration (RTM), and Time Reverse Time Migration (tRTM) model applications. The work provided insights into both the directive language capabilities and gaps with respect to the needs of these algorithms and how the algorithms can be ported to GPU accelerators using directive languages.

2. STATEMENT OF OBJECTIVES

The following are the technical objectives set out in the original CRADA and the two subsequent amendments.

- **Year 1**
  - **Task 1: Identification and understanding of performance-critical sections of RTM code suite, implement OpenACC directives and explore performance and optimization issues.** The team will identify the computational kernels of the original RTM code suite which limit performance and understand, through code inspection and profiling, their resource utilization characteristics. This information will also be used in the subsequent explorations of other programming models. OpenACC versions of selected kernels, possibly including multiple variants, will be developed and evaluated for both performance and expressiveness. Important kernels which cannot be expressed using OpenACC, or are awkward to express or do not perform as expected are of interest to help drive work outside of this agreement to improve the OpenACC standard and implementations. To this end, another desirable product of this work will be the development of compact code examples that can used to drive such improvements. The preference and goal will be to abstract the challenging code to a degree that it can be shared publicly as non-proprietary code. Where this is not possible, the code will remain proprietary and will not be shared outside of the Contractor and Participant organizations.
  - **Task 2: Explore programming model and optimization issues for partitioned global address space (PGAS) models.** Similar to Task 1, the team will identify kernels that are good candidates for implementation using CoArray Fortran (CAF), as an exemplar of the partitioned global address space (PGAS) programming model. As appropriate, we will consider the version of CoArray Fortran embodied in the Fortran 2008 standard, as well as the Rice University "CoArrary Fortran 2.0" variant of CAF. As in Task 1, we also seek to produce code examples that can be used to drive improve the CAF standard and implementation.
  - **Task 3: Explore programming model and optimization issues for heterogeneous systems using OpenMP 4.** Similar to Task 1, the team will identify kernels that are good candidates for OpenMP version 4, focusing particularly on the use of accelerators (`target` directive). As in Task 1, we also seek to produce code examples that can be used to drive improve the OpenMP standard and implementation.
  - **Task 4: Explore I/O performance and optimization issues.** The team will develop a seismic depth imaging I/O benchmark kernel, based on the RTM suite, port it to use the ADIOS I/O system, and compare the performance of the original and ADIOS approaches.

- **Task 5: Evaluate programming environments for future systems.** The team will utilize the results of Tasks 1-4 to carry out both performance modeling studies and actual benchmarking to evaluate and compare the performance and scalability of the different approaches. The primary target system will be the OLCF Cray XK7, Titan, though other systems will also be used, as appropriate. The various approaches will also be evaluated for performance portability (e.g., different implementations of the programming model, as well as different computer systems), and expressiveness and ease of use.

- **Year 2**
  - **Task 1: Identification and understanding of performance-critical sections of Time Marching Based Method (TMBM) code suite, explore performance and optimization issues of the actual OpenACC implementation.** The Parties will start from the actual OpenACC implementation done by the Participant in 2015 and will understand, through code inspection and profiling, their resource utilization characteristics. TMBM is based on 3 different wave equation finite difference kernel propagators (acoustic, TTI-acoustic and elastic). The performance analysis will be performed on all of them. Based on the optimization analysis and improvement, TMBM will be tested at large scale on the elastic kernel. Starting from the elastic SEAM model, elastic-RTM performance of TMBM will be evaluated. In addition, the performance of the ADIOS library, already implemented in the actual TMBM version for checkpointing, will be evaluated and improved where possible.
  - **Task 2: Explore programming model and optimization issues for heterogeneous systems using OpenMP 4.** Similar to Task 1, the team will identify kernels that are good candidates for OpenMP version 4, focusing particularly on the use of accelerators (target directive). As in Task 1, we also seek to produce code examples that can be used to drive & improve the OpenMP standard and implementation. Explore OpenMP 4.0/4.1 constructs for GPU parallelization.
- **Year 3**
  - **Task 1: Identification and understanding of performance-critical sections of Time Reverse Time Migration (tRTM) code suite, explore performance and optimization issues of the actual OpenACC implementation.** The Parties will start from the actual OpenACC implementation done by the Participant in 2015 and will understand, through code inspection and profiling, their resource utilization characteristics. tRTM is based on 3 different wave equation finite difference kernel propagators (acoustic, acoustic TTI , elastic and elastic TTI). The performance analysis will be performed on all of them. Based on the optimization analysis and improvement, tRTM will be tested at large scale on the elastic kernel. Starting from the elastic SEAM model, elastic-RTM performance of tRTM will be evaluated. In addition, the performance of the ADIOS library, already implemented in the actual tRTM version for checkpointing, will be evaluated and improved where possible.
  - **Task 2: Explore programming model and optimization issues for heterogeneous systems using OpenMP 4.5.** Similar to Task 1, the team will identify kernels that are good candidates for OpenMP version 4.5, focusing particularly on the use of accelerators (target directive). As in Task 1, we also seek to produce code examples that can be used to drive & improve the OpenMP standard and implementation. Explore OpenMP 4.0/4.5 constructs for GPU parallelization and for performance portability.
  - **Task 3: Explore multi-level storage hierarchy for future systems by extending ADIOS capabilities.** Future extreme computing system will have different levels of storage a complex to deliver high bandwidth IO and reduce data movement by bringing data closer to the compute element. Those multiple levels of storage will have different latencies and speed. The management of those different levels by the users may be

complicated and needs to be simplified and transparent. ADIOS is good to candidate to be extended to handle those different storage levels.

3. BENEFITS TO THE FUNDING DOE OFFICE'S MISSION

Oak Ridge National Laboratory hosts the Oak Ridge Leadership Computing Facility (OLCF), which, fields some of the most powerful supercomputers in the world and carries out a broad range of research in computational science and computer science. ORNL researchers have developed extensive expertise in porting and optimizing applications for supercomputers, such as the OLCF's Titan and Summit systems, as well as insight into future computer architectures and how they will influence applications. ORNL's computational and computer science research programs benefit from deep interactions with scientific applications from a variety of domains, which allow researchers to better understand the applicability and limitations of their tools and techniques.

4. TECHNICAL DISCUSSION OF WORK PERFORMED BY ALL PARTIES

Many of the technical outcomes of this work have been summarized in a peer-reviewed paper: Kshitij Mehta, Maxime Hugues, Henri Calandra, Oscar Hernandez, and David E. Bernholdt, One-Way Wave Equation Migration at Scale on GPUs using Directive Based Programming in 2017 IEEE Parallel and Distributed Processing Symposium (IPDPS), Orlando, Florida, May, 2017, pp. 224-233. DOI: 10.1109/IPDPS.2017.82.

Work started with the one-way wave equation migration (OWEM), a simplified form of the Reverse Time Migration algorithm. The OWEM application was ported to OpenACC, resulting in a speedup of 3x on a single NVIDIA K20X GPU (on the Titan supercomputer) compared to the performance obtained on an 8-core Intel Sandy Bridge CPU. The port made use of the cuFFT library and customized sparse solvers which used CUDA kernels for certain operations to obtain best performance. The computational kernels ported to OpenACC were also identified as good candidates for OpenMP offloading. This work represents Tasks 1 and 3 in the Year 1 statement of work. Task 2 (PGAS programming models) was not completed due to time constraints. Experience implementing the OWEM application in OpenACC also served as an important driver for the design of the "deep copy" extension to OpenACC, which was included in version 2.6 of the language standard.

Once we had a working OpenACC implementation of the OWEM application, we carried out an at-scale seismic imaging workload using 2793 shots from the SEAM Phase I Isotropic dataset. Each shot was processed with an instance of the OWEM application running on 28 nodes of Titan (28 GPUs), utilizing a total of 18,424 out of the total 18,688 nodes on Titan. At the time, this represented the largest scale OpenACC application even run. The experiment provided a number of insights about the OpenACC implementation on Titan, and the execution of very large-scale seismic imaging workloads on supercomputers. The work uncovered a bug in the OpenACC runtime, which we worked around by making explicit calls to clean up memory usage on the GPU after the completion of a task. The work also illustrated resilience issues in large workloads, which can be addressed by the inclusion of checkpoint/restart capabilities into the workload. Finally, the experiment highlighted I/O-related performance issues with the workload to be addressed in the future (Year 1/Task 4).

The benchmarking activities in the work described above, both individual kernels within the OWEM application and for the overall SEAM workload constituted Year 1/Task 5. These results were used internally to guide our work, and some were reported in the IPDPS paper.

Work during Year 2 extended the model application to the more sophisticated Time Marching Based Method (TMBM), which includes a modeling component and an imaging component. For the imaging, TMBM uses Reverse Time Migration (RTM), which is more computationally demanding (~2x) and complex than the OWEM application used previously. As part of the effort to optimize RTM for running on Titan, many small-scale experiments were performed and its OpenACC kernels were fine-tuned to optimize performance per node (Year 2/Task 1). Additionally, techniques such as GPUDIRECT were utilized to reduce communication times between the host and device. We also looked at a preliminary OpenMP 4.5 offload implementation for key RTM kernels (Year 2/Task 2). Translation is mostly straightforward, though OpenMP 4.5 is prescriptive, whereas OpenACC is more descriptive. As a consequence, where one can rely on the OpenACC compiler to provide appropriate optimizations of GPU kernels, the OpenMP version must be more explicitly defined, which reduces the performance portability. OpenMP 5.0 (released in November 2017) provides more flexibility.

At the workload level, we also built upon the previous SEAM experience, primarily focusing on improvements to the I/O aspect of the application using the ADIOS library (Year 2/Task 1). A large scale run of the 2793 shots of the SEAM dataset using 18,000 nodes on Titan was estimated to consume over 10 hours on Titan. Each shot generated under 10 TB of wavefield data. Of the 2793 shots, a maximum of 1000 shots can be run at a time (18 GPUs per shot), which leads to storing a maximum of 10 PB of wavefield data on the file system at any time. Thus, the full dataset would have generated an aggregate of 25 PB of wavefield data during the 10-hour run. Multiple optimizations were implemented at the file system level with the help of the Technology Integration team to optimize file I/O in preparation for the large run. This included pre-creating output files on the metadata server, unlinking as opposed to removing wavefield data at the end of a shot, and writing data separately to the atlas1 and atlas2 filesystems for improved bandwidth. Smaller scale runs using 2000, 5000, and 9000 nodes on Titan showed that file I/O took at least 25% of the total application runtime. Profiling using Darshan showed that a majority of writes from all processes were small-sized writes in the 1KB – 10 KB range. As the application used POSIX I/O to read and write data, POSIX I/O was replaced with the ADIOS I/O library for optimizing I/O performance. Small-scale tests showed that the write sizes were now in the 10 MB – 100 MB range. A run with ADIOS on Titan experienced failures due to insufficient disk space. No more tests were run as the project had exhausted its allocation time on Titan (including additional time awarded).

None of the planned Year 3 activities were completed. After the extraordinary delays in obtaining approval for the final amendment, both parties had moved on to other interests.

5. SUBJECT INVENTIONS (AS DEFINED IN THE CRADA)

None

6. COMMERCIALIZATION POSSIBILITIES

There is no specific technology to commercialize. However it is expected that Total will use the experience gained in this CRADA to adapt other of their in-house applications to GPU accelerators.

7. PLANS FOR FUTURE COLLABORATION

There are no plans at this time for further collaboration in relation to programming of GPU accelerators.

8. CONCLUSIONS

ORNL and Total cooperated to study several seismic imaging applications on GPU accelerators. The work included porting of both one-way wave equation migration and reverse time migration applications to OpenACC as well as explorations with OpenMP 4.5. The applications were also used to explore at-scale workloads based on the industry SEAM database, run at nearly the full scale of Titan (18,424 nodes). Total gained experience and better understanding of key seismic imaging algorithms and workloads on GPU accelerators and on large HPC systems. ORNL gained experience with GPU acceleration in another application domain, which motivated and drove improvements in both the OpenACC and OpenMP standards as well as the OpenACC implementation available at the Oak Ridge Leadership Computing Facility (and elsewhere).