SAND2017-8216R

Trinity Phase 2 Open Science: CTH

Kevin Ruggirello and Tracy Vogler
Sandia National Laboratories[1]

Abstract:

CTH is an Eulerian hydrocode developed by Sandia National Laboratories (SNL) to solve a wide range of shock wave propagation and material deformation problems. Adaptive mesh refinement is also used to improve efficiency for problems with a wide range of spatial scales. The code has a history of running on a variety of computing platforms ranging from desktops to massively parallel distributed-data systems. For the Trinity Phase 2 Open Science campaign, CTH was used to study mesoscale simulations of the hypervelocity penetration of granular SiC powders. The simulations were compared to experimental data. A scaling study of CTH up to 8192 KNL nodes was also performed, and several improvements were made to the code to improve the scalability.

Background and Research Objectives:

Anderson et al. [2008] performed experiments on the penetration of gold rods into targets of pressed SiC powder (~72% TMD) at velocities of 1-3 km/s. Mesoscale simulations in which each individual grain is resolved in CTH have been used previously to study compaction [Borg & Vogler, 2013] and penetration [Borg & Vogler, 2008] processes, but the Anderson et al. experiments represented a new regime to study the applicability of these tools. One goal of the simulations was to explore differences between 2-D and 3-D mesoscale calculations since 2-D calculations are less computationally expensive.

A strong and weak scaling study was also performed on a multi-material shock wave problem out to 8192 nodes (524,288 MPI tasks) to examine the scalability of the code on the new platform. Areas of improvement were identified and features were added to the code to help address them.

Importance to Sandia National Labs and Mission:

Given the wide use of CTH for various problems at SNL, the scalability and performance of the code is important for analysts. The usability of the code on the new platform was evaluated during the Open Science period to make sure analysts would be productive on the Trinity Knights Landing partition. The Open Science period was also used to perform mesoscale simulations of the impact of pressed SiC powders which are important for several applications.

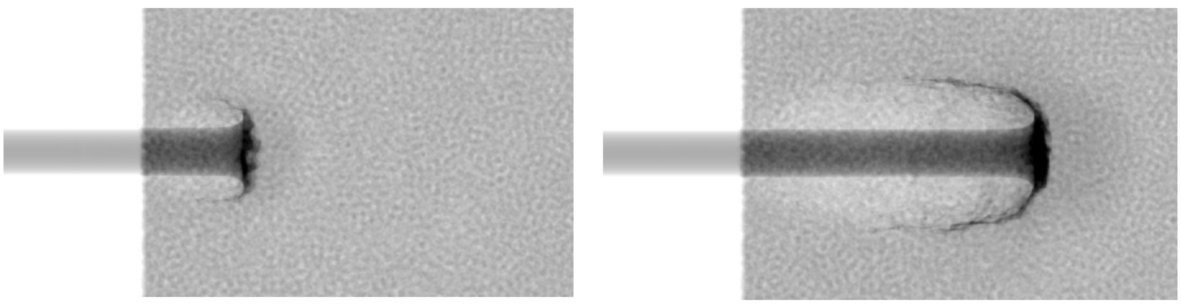
---

[1] Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

Scientific Approach and Accomplishments:

Mesoscale Simulations of Hypervelocity Penetration of Granular SiC:

The mesoscale simulations were modeled after the Anderson et al. experiments, though certain aspects proved impossible to duplicate due to limitations in generating mesoscale volume elements. In particular, relatively large particles were used in the simulations compared to those used in the experiments. Also, the volume fraction used in the calculations was 55% rather than the 72% used in the simulation. Despite these limitations, the simulations proved interesting. Synthetic radiographs from a simulation at two points in time can be seen in Fig. 1; they are qualitatively similar to those reported by Holmquist & Johnson [2008] for these experiments.



(a) (b)

Fig. 1 Synthetic radiograph generated from mesoscale simulations of porous SiC for a velocity of 2.5 km/s at 1.5 and 4.5 us.

From the images shown in Fig. 1, it is possible to extract the penetration velocity of the gold rod for different impact velocities. These penetration velocities are shown in Fig. 2 along with the experimental results from Anderson et al. For comparison, predictions using the Bernoulli equation for hydrodynamic penetration (no strength in projectile or target) are shown for the densities used in the experiments (2.35 g/cc) and the simulations (1.78 g/cc). Based on the effect of initial target density, the simulations are expected to lie above the experiments as is seen to occur. Some differences are apparent when "slide" interfaces are used for the particles.

To examine the difference between 2D and 3D simulations, 2.5-D calculations were conducted using a slab (extending across the domain in one direction) penetrator rather than a cylinder. Although analysis of these results is not complete, comparison of the 2-D, 2.5-D, and 3-D calculations will provide insight into the value of 2-D simulations.

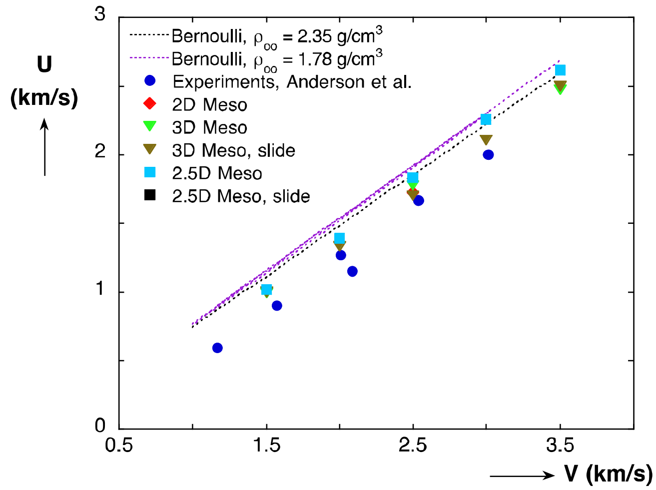


Fig. 2 Penetration velocities from the mesoscale calculations compared to experimental results and hydrodynamic predictions.

Scaling Study:

Phase 2 of the Trinity High Performance Computing platform consists of more than 9500 nodes based on the Intel Knights Landing (KNL) chip. Each KNL node consists of 68 cores with up to 4 threads per core, the AVX512 instruction set for improved vector operations, and 384 GB of DDR4 memory. Each node also contains 16 GB of stacked High Bandwidth Memory (HBM) which is designed to accelerate application performance by increasing memory bandwidth.

The significant increase in parallelism and the addition of HBM represents several new challenges to applications. As part of the Open Science campaign, various HBM modes were evaluated for CTH and it was found that utilizing the HBM in cache mode consistently gave the best performance.

Currently CTH is using MPI across all the cores on a node with only 1 thread per core. Figure 3 shows the strong and weak scaling up to 8192 nodes (557,056 MPI tasks) for a flat mesh problem. For strong scaling the total computational work is kept constant and the number of processors are varied while weak scaling scales the computational work with the number of processors. Ideally the strong scaling time should decrease linearly with the number of nodes while the weak scaling time remains constant. While there is deviation from ideal scaling for smaller problems at large node counts (approximately 1024 cells or less per task), this is expected due to increased communication costs and is below the work per task that problems are typically run at.

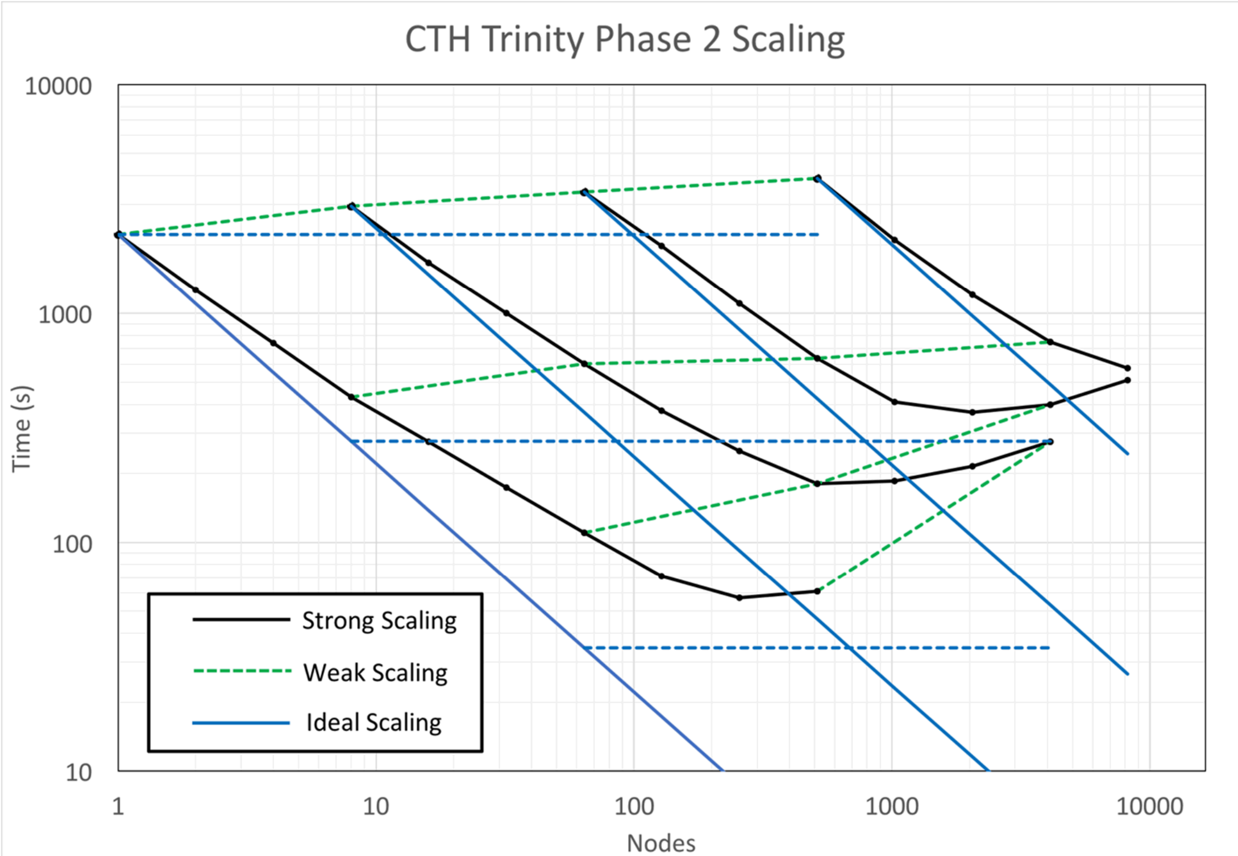


Fig. 3: Strong and weak scaling behavior on the Knights Landing nodes of Trinity up to 8096 nodes. The strong scaling falls off at higher node counts due to communication costs.

The performance of the Haswell versus the KNL nodes is shown in Fig. 4. MPI was used across all the nodes. For a single node, the KNL was approximately 17% faster than the Haswell node for an equivalent problem. However, as the number of nodes is increased the Haswell nodes outperform the KNL nodes beyond 8 nodes. This is because twice the number of MPI tasks were used on the KNL's versus the Haswell's so each core has half the work to do and was more sensitive to load imbalances. This was especially apparent when looking at AMR cases.

To help mitigate the sensitivity to load imbalances, a new timer based load balancing was added to version 12.0 of CTH. In previous versions of CTH, every AMR block was assumed to have the same computational cost but due to more advanced material models and other algorithms certain blocks could consume more computational resources than others. With the cost based load balancing, a high-resolution timer is used to measure the computational time of each block and the load is balanced based on this time. In test cases performance gains from 5-25% were observed depending on the problem and scale it was run at. In the future, the cost based load will be modified to include the cost of balancing it (communication cost), and other improvements.

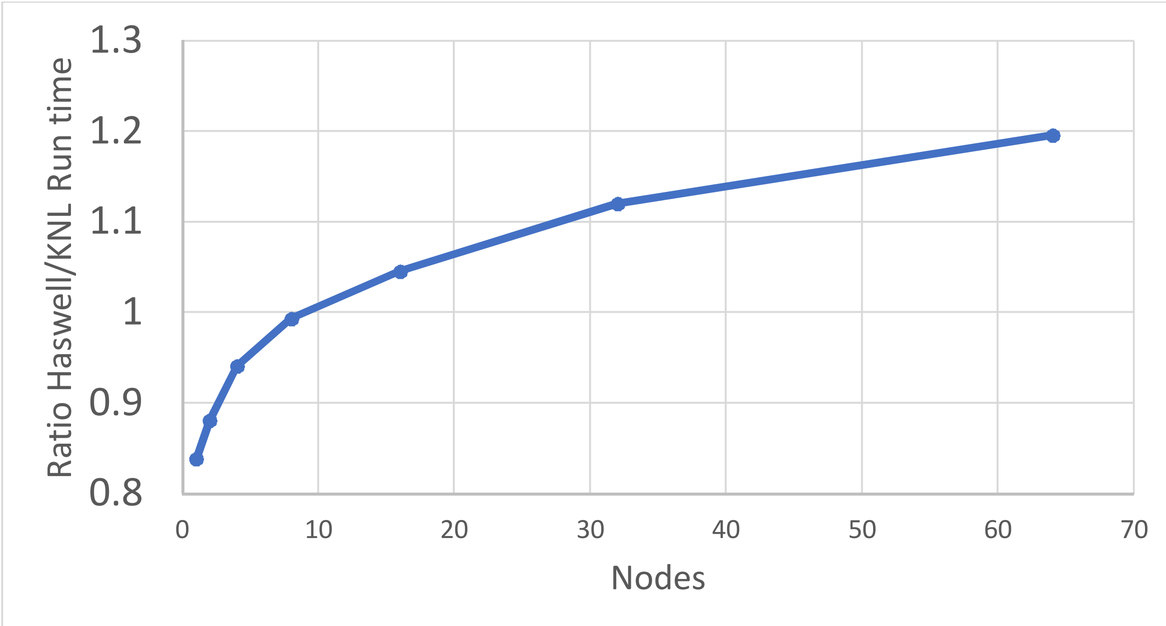


Fig. 4: Ratio of Haswell to KNL scaling performance.

The current threading model in CTH was found to not give any speed up. Figure 5 shows the ratio of run times for multi-threaded cases to a single thread case. For all cases the number of threads times PE's was kept constant at 64 on a single node. The current threading model in CTH uses openMP compiler directives to parallelize the J index loops in the code. For many of the loops there is not enough parallelism to overcome the cost of starting/joining threads and load imbalances of some of the loops.

Threading improvements are being investigated to increase performance and allow less MPI tasks per node to be used to decrease communication cost and load imbalances. Parallel threading over AMR blocks is one possible change that could increase the threaded performance, but substantial code modifications are required to investigate this.

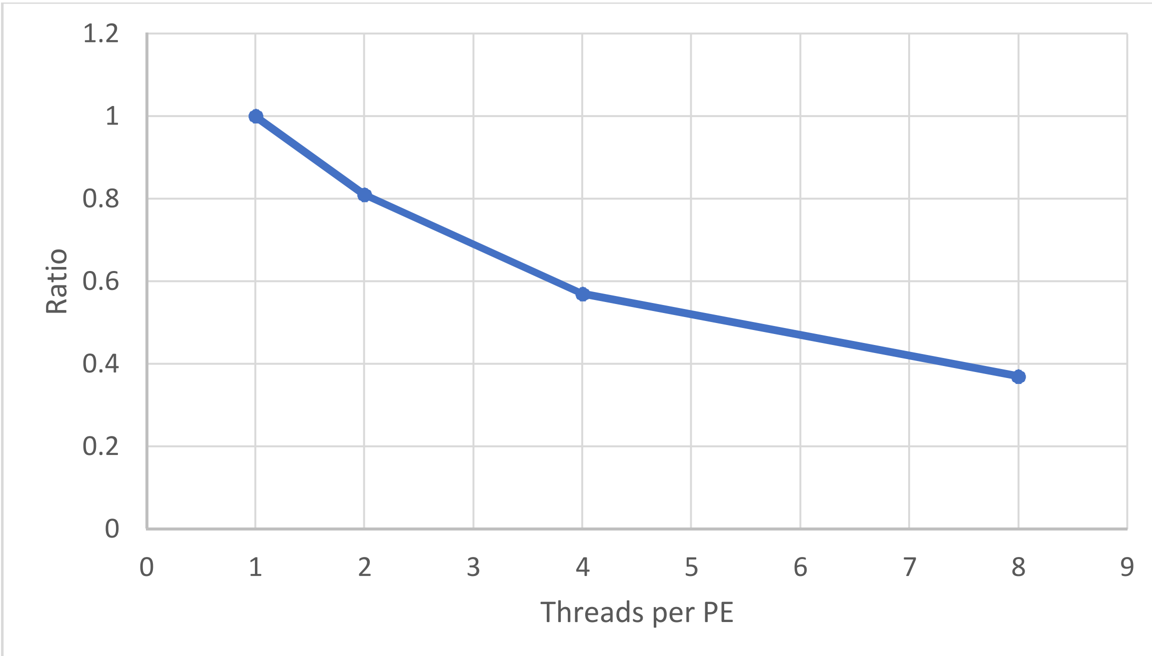


Fig. 5: Ratio of threaded run time to single thread case. The number of threads x PE's was kept constant.

Planned Publications:

Several conference and journal publications are planned in relevant venues.

Unresolved Problems:

Due to the lustre file system available during the Open Science campaign being undersized for the system, IO performance was not examined extensively. Using the post processing and on the fly visualization capabilities of Spymaster, data storage was kept to a minimum for all runs. Due to the KNL chip design it is anticipated that a staggered IO approach might improve the performance by limiting the number of concurrent IO tasks per node and will be investigated in the future. The usage of burst buffers on the machine was also not explored due to difficulty in properly utilizing them.

References:

Anderson, C.E., T. Behner, D.L. Orphal, A.E. Nicholls, and D.W. Templeton. 2008. 'Time-resolved penetration into pre-damaged hot-pressed silicon carbide', *International Journal of Impact Engineering*, 35: 661-73.

Borg, J.P., and T.J. Vogler. 2008. 'Mesoscale simulations of a dart penetrating sand', *International Journal of Impact Engineering*, 35: 1435-40.

Borg, J.P., and T.J. Vogler. 2013. 'Rapid compaction of granular materials: characterizing two and three-dimensional mesoscale simulations', *Shock Waves*, 23: 153-76.

Holmquist, T.J., and G.R. Johnson. 2008. 'The failed strength of ceramics subjected to high-velocity impact', *Journal of Applied Physics*, 104: 013533.