SAND2020-0478C

Exascale Method of Moments for Linear Electromagnetics with Gemma

S. A. Blake, B. Zinser, R. Pfeiffer, J. Kotulski, V. Dang, S. Campione, and W. Langston
Dept. of EM Theory & Simulation
Sandia National Laboratories
Albuquerque, USA
sablake@sandia.gov, bzinser@sandia.gov, rapfeif@sandia.gov, jdkotul@sandia.gov, vqdang@sandia.gov, sncampi@sandia.gov, wllangs@sandia.gov

***Abstract*—Gemma is a linear electromagnetic code developed at Sandia National Laboratories based on the method of moments. It is part of the Advanced Simulation and Computing program that has the objective of supporting next generation high performance computing (HPC) platforms. Next generation HPC platforms are moving from homogeneous machines utilizing central processing units (CPUs) to heterogeneous machines that utilize CPUs alongside hardware based accelerators like graphical processing units (GPUs) and many integrated core architectures (MICs). This report describes the progress Gemma has made in utilizing large scale heterogeneous computing resources and reports current strong scaling results leveraging Sandia's Kokkos library.**

***Index Terms*—Gemma, GPU, Method of Moments.**

I. Introduction

The Electromagnetic (EM) Theory & Simulation department at Sandia National Laboratories supports design and qualification of components sensitive to EM radiation. The complex nature of this problem has led to the development of several specialized high performance computing (HPC) software packages like EIGER [1]–[3]. Next generation exascale HPC architectures are continuously evolving to allow for larger, more computationally intensive problems to be solved. At the same time, the need for designers and analysts to rapidly iterate on new designs while still maintaining precise results for qualification efforts has become increasingly important. Gemma, EIGER's successor, seeks to optimize the algorithms and methods used in established computational electromagnetic codes for newer HPC architectures while introducing new algorithms suitable to support rapidly iterating design processes.

Gemma is based upon the frequency domain method of moments (MOM) [4], [5], which is a useful technique for solving the electric/magnetic field integral equations after surfaces are discretized using planar or curvilinear mesh elements. This technique can be applied to nanophotonics, radar cross-section, and coupling through apertures in shields [1], [6]. This formulation results in a dense, complex valued, typically ill-conditioned system matrix. For N unknowns, the amount of memory required to store the system matrix scales as $O(N^2)$ and the number of computational operations as $O(N^3)$.

This scaling presents unique challenges in terms of runtime and size. Recent advances in methods like multi-level fast multipole method and preconditioners such as the sparse approximate inverse preconditioner have shown promise for accelerating a subset of this type of workload [7], [8]. Additionally, Gemma is extending proven HPC capabilities to allow for high precision qualification studies [1].

Scaling to heterogeneous HPC systems requires dealing with both multi-node communication and intra-node threading with hierarchical memory spaces. Gemma uses Kokkos [9] as an abstraction layer to support multiple architectures from a single code base. Using an abstraction layer between Gemma and accelerator specific programming interfaces allows Gemma to support current accelerators (GPUs, CPUs, and MICs) as well as providing Gemma a path to support other architectures developed for upcoming next generation platforms.

Gemma's primary motivation for utilizing HPC systems is to solve larger, more computationally complex problems. The amount of memory and processing resources a problem requires is determined by the electrical size of the problem's surface and how many small geometric features need to be captured. As the problem becomes electrically large or more small features are added, both computational complexity and memory required for solving the system matrix increases as well. Individual nodes typically have larger allocations of RAM, leading to CPUs being computationally bound and accelerators being memory bound. Prior results from Gemma have focused on multi-threaded single node performance to exploit accelerators. Gemma has recently implemented an additional MPI layer to fully utilize entire HPC systems. Here initial scaling results across different platforms are reported.

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 Energys National Nuclear Security Administration under contract DE-NA-0003525. This paper describes objective technical results and analysis. Any subjective views or opinions that might be expressed in the paper do not necessarily represent the views of the U.S. Department of Energy or the United States Government.

978-1-7335096-0-2©2020 ACES

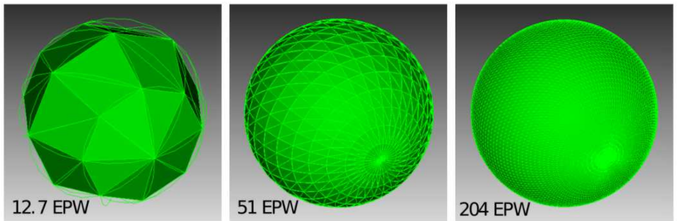

Fig. 1. Example of mesh refinement using a simple sphere for different EPWs.

II. Heterogeneous Platform Scaling

To test the scaling properties of Gemma, the scattering from a 1m radius sphere at 30 MHz was investigated. The sphere was meshed to give a fixed number of elements per wavelength (EPW) as shown in Fig. 1. Table I lists N for various levels of mesh refinement. These were then run on two platforms (referenced jointly under the name Mutrino). One platform consisted of Intel Haswell (HSW) processors, each node containing 32 CPU cores and 128 GB of RAM. The other consisted of Intel Knights Landing (KNL) processors, each node containing 64 CPU cores with 96 GB of RAM. Strong scaling results are shown in Fig. 2 for $N = 55296$ and $N = 221184$. Note that $ReferenceNodes$ in Fig. 2 was 8 for $N = 55296$ and 24 for $N = 221184$. In both cases, the maximum number of nodes was restricted to 44 nodes due to hardware availability. With this preliminary result we are not able to identify the upper limit of scaling. Results for GPUs have been generated, but are currently undergoing further review.

TABLE I
Unknowns per Elements for Refined Sphere.

EPW	Vertices	Elements	Unknowns (N)
12.7	20	36	54
51	290	576	864
204	4610	9216	13824
408	18434	36864	55296
816	73730	147456	221184

Ideal strong scaling would correspond to a line with a slope of 1 (plotted for reference). For the $N = 221184$ problem, scaling was almost linear on both HSW and KNL with slopes of 0.93 ± 0.03 and 0.86 ± 0.01 respectively. Scaling drops slightly in the $N = 55296$ case with slopes of 0.78 ± 0.01 and 0.77 ± 0.01 (HSW/KNL). This reduction is mainly due to the problem becoming more communication bound as the amount of computational work per node decreases. It is worth noting that the increased computational power of the KNL nodes did not result in better overall performance in this case. Investigation showed that this was largely due to the smaller amount of memory available on these nodes. The solvers quickly become memory bound due to the $O(N^2)$ requirement for storing the system matrix. Rearranging memory to better utilize increased compute capacity of accelerators is an active area of research.

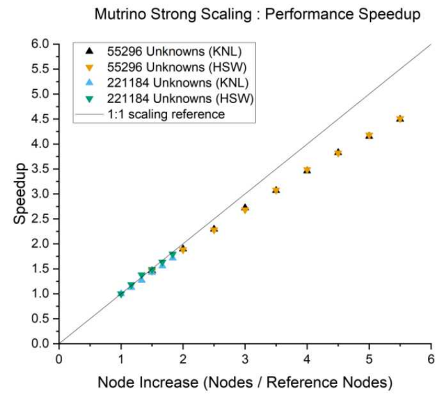

Fig. 2. Strong scaling results on HSW and KNL platforms. The black line depicts ideal strong scaling.

III. Conclusion

Gemma has demonstrated the ability to run on multiple architecture types. In addition, it has shown strong scaling results for HSW and KNL processors that indicate no loss of scaleability when running on different architectures. These results indicate that Gemma is currently memory bound, particularly at the register level. Current results also indicate that Gemma is not compute or communications bound. Gemma's next efforts are in demonstrating similar scaling when running on GPUs as well as improving overall utilization of compute capability when running on CPUs, MICs, and GPUs.

References

[1] W. L. Langston, J. Kotulski, R. Coats, R. Jorgenson, S. A. Blake, S. Campione, A. Pung, and B. Zinser, "Massively parallel frequency domain electromagnetic simulation codes," in 2018 International Applied Computational Electromagnetics Society Symposium (ACES), pp. 1–2, Mar. 2018.

[2] R. M. Sharpe, J. B. Grant, N. J. Champagne, W. A. Johnson, R. E. Jorgenson, D. R. Wilton, W. J. Brown, and J. W. Rockway, "Eiger: Electromagnetic Interactions GEneRalized," tech. rep., Sandia National Laboratories Report, 1997.

[3] D. R. Wilton, W. A. Johnson, R. E. Jorgenson, R. M. Sharpe, and J. B. Grant, "Eiger: A new generation of computational electromagnetics tools," tech. rep., Sandia National Laboratories, 1996.

[4] W. C. Gibson, The Method of Moments In Electromagnetics. CRC Press, 2015.

[5] R. F. Harrington, Time-Harmonic Electromagnetic Fields. McGraw-Hill Book Company, 1961.

[6] L. I. Basilio, S. Campione, L. K. Warne, and W. L. Langston, "Electromagnetic theory at Sandia National Laboratories," in 2017 International Conference on Electromagnetics in Advanced Applications (ICEAA), pp. 1879–1882, Sep. 2017.

[7] J.-G. Wei, Z. Peng, and J.-F. Lee, "Multiscale electromagnetic computations using a hierarchical multilevel fast multipole algorithm," Radio Science, vol. 49, no. 11, pp. 1022–1040, 2014.

[8] M. T. Bettencourt, B. Zinser, R. E. Jorgenson, and J. D. Kotulski, "Performance portable sparse approximate inverse preconditioner for efie equations," in 2017 International Conference on Electromagnetics in Advanced Applications (ICEAA), pp. 1469–1472, Sep. 2017.

[9] H. C. Edwards, C. R. Trott, and D. Sunderland, "Kokkos: Enabling manycore performance portability through polymorphic memory access patterns," Journal of Parallel and Distributed Computing, vol. 74, no. 12, pp. 3202–3216, 2014. Domain-Specific Languages and High-Level Frameworks for High-Performance Computing.