SPARSE MATRIX FACTORIZATION IN SOLID MECHANICS AND GEOPHYSICS ON CPUS AND GPUS

Allocation: Industry/60 Knh
PI: Seid Koric[1,2]
Co-PI: Anshul Gupta[3]
Collaborators: Natalia Gimelshein[4], Steve Rennich[4], Vladimir Puzyrev[5], Scott Wilkin[1,2]

[1]National Center for Supercomputing Applications
[2]University of Illinois at Urbana–Champaign
[3]IBM T. J. Watson Research Center
[4]NVIDIA Corp.
[5]Barcelona Supercomputing Center

EXECUTIVE SUMMARY:

Sparse matrix factorization is a critical algorithm in many science, engineering, and optimization applications. The performance of the massively parallel direct multi-frontal solver Watson Sparse Matrix Package (WSMP) [1] for solving large sparse systems of linear equations arising in an implicit finite element method in solid mechanics and an inversion problem in geophysics was evaluated on Blue Waters and achieved new records in sparse matrix factorizations both with CPU and GPU solvers. We performed full-scale benchmarking tests up to 65,536 cores (100 Tflop/s) on XE6 nodes using assembled global stiffness matrices and load vectors with 5 million to 40 million unknowns extracted from "real-world" commercial implicit finite element analysis (FEA) and academic geophysics codes. We also present a minimally invasive approach to the GPU acceleration of WSMP that can more than double the CPU-only numerical factorization performance and scale beneficially up to as many as 512 Cray XK7 nodes.

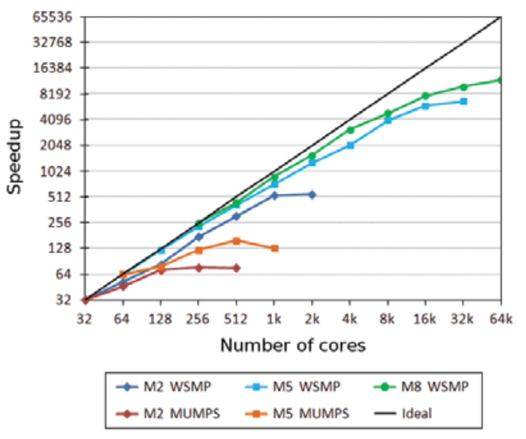

FIGURE 1: WSMP parallel speedup for geophysical tests on XE6 nodes of Blue Waters.

INTRODUCTION

Across a range of engineering fields, the use of simulation and computational models is pervasive for designing engineered systems. HPC systems play an essential role in simulations and modeling. Researchers and manufacturing teams depend on HPC to create safe cars, energy-efficient aircraft, effective communication systems, and efficient supply chain models. HPC is also necessary for large-scale geophysical simulations; in order to obtain a realistic image of a geologically complex area, industrial surveys collect vast amounts of data, making the computational cost extremely high for the subsequent simulations.

A major computational bottleneck of implicit finite modeling in solid mechanics and the inversion algorithms in geophysics is the solution of large sparse systems of linear ill-conditioned equations $Ax = b$ in real or complex domains. There has been considerable interest in the development of numerical algorithms for solving large sparse linear systems of equations and their efficient parallel implementation on HPC systems for more than three decades. The algorithms may be grouped into two broad categories: direct methods and iterative methods. Finding and computing a good preconditioner for use with an iterative method can be computationally more expensive or often impossible, thus making implicit numerical methods with direct solvers often the only feasible methods with ill-conditioned systems. The limitation on CPU and memory requirements made the use of direct solvers uneconomical in the past, resulting in broad use of iterative solvers. The recent rise of petascale computational resources, however, has greatly increased the efficiency and practicality of using direct solvers for large sparse systems.

METHODS & RESULTS

For this study, we have chosen a few test systems resulting from large-scale geophysics electro-magnetic (complex domain) and solid mechanics (real domain) problems.

Parallel speedup of the factorization of the three largest geophysics test matrices is shown in fig. 1. In these experiments, WSMP reached a performance of 97 Tflop/s on 65,536 cores of Blue Waters in the solution of the linear system with 8 million unknowns in the complex domain, or 16 million unknowns in real domain. Our scalability study credibly showed that the direct solver WSMP is prepared for even larger HPC simulations with systems having upwards of 10 million complex unknowns, which will provide high-fidelity results for the large inversion problem (example in fig. 2).

Besides running the standard CPU-only version of WSMP, we have shown that, with remarkably minor modifications to the original CPU code, the numerical factorization performance of WSMP can be accelerated on a reasonably large number of GPU-enabled XK7 nodes on Blue Waters. Modifications to the original code were mostly limited to intercepting high-level BLAS operations and handling them using a new GPU library.

The factorization performance of GPU-accelerated WSMP, so called ACCEL_WSMP, is given in fig. 3. Benchmarking with industrially derived finite-element analysis solid mechanics matrices showed that on 128 nodes, the GPU-accelerated code was two times faster than the CPU-only code using the same number of GPU or CPU nodes. A fairly high degree of acceleration was observed for lower node counts. The WSMP numerical factorization showed GPU acceleration by a factor of 2.5 to 3.5 on up to 32 nodes. Due to issues with load balancing and difficulty maintaining sufficient computational intensity to hide CPU/GPU transfers, the observed speedup was reduced at larger node counts. Work is underway to further improve the performance of the GPU code.

WHY BLUE WATERS?

Blue Waters is the only place where massively parallel sparse solver technology such as WSMP can be tested by taking full advantage of large amounts of distributed memory, tens of thousands of modern multicore processors and GPUs, and low latencies and increased bandwidth of leading interconnect network technologies. This exciting technology advancement will lead to a massive leap in terms of advances in design and manufacturing, and understanding the properties of the Earth subsurface to allow a major breakthrough in oil exploration, to name two of many applications in science and industry that can benefit from this work on current and future petascale architectures hosted at NCSA.

PUBLICATIONS

Gimelshein, N. E., A. Gupta, S. C. Rennich, and S. Koric, GPU acceleration of WSMP. GPU Technology Conference 2015, San Jose, Calif. , March 17–20, 2015.

Koric, S., Q. Lu, and E. Guleryuz, Evaluation of massively parallel linear sparse solvers on unstructured finite element meshes. *Computers and Structures*, 141 (2014), pp. 19–25, doi:10.1016/j.compstruc.2014.05.009.

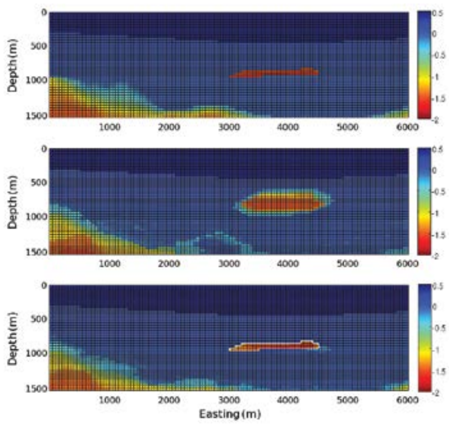

FIGURE 2: Real model (top) and the results of the unconstrained stand-alone electromagnetic inversion (middle) and constrained by the seismic results (bottom). Color represents electric conductivity on a logarithmic scale.

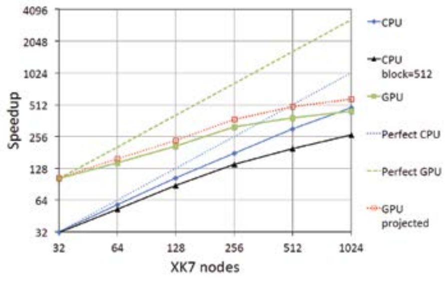

FIGURE 3: Numerical factorization speedup vs. number of XK7 nodes used for the M20 test system originating from solid mechanics finite element analysis.