Annual Report for Blue Waters Professor Allocation

- **Project Information**
  - Title: Algorithms for extreme scale systems
  - PI: William Gropp, University of Illinois Urbana-Champaign
  - Collaborators: Luke Olson, University of Illinois Urbana-Champaign
  - Contact: wgropp@illinois.edu

- **Executive summary (150 words)**

Continued increases in the performance of large-scale systems will come from greater parallelism at all levels. At the node level, we see this both in the increasing number of cores per processor and the use of large numbers of simpler computing elements in GPGPUs. The largest systems must network tens of thousands of nodes together to achieve the performance required for the most challenging computations. Successfully using these systems requires new algorithms and new programming systems. My research looks at the effective use of extreme scale systems. Over the last year, we have shown the benefit of lightweight intranode balancing on scalability and performance. We continue to explore alternative formulations of conjugate gradient that eliminate some of the strict barrier synchronization as well as better use the memory hierarchy, as well as looking at ways to reduce the impact of communication on the scalability of algebraic multigrid as well as algorithmic approaches to resilience that exploit the multilevel representation in multigrid.

- **Description of research activities and results**
  - *Key Challenges:* At extreme scale, even small inefficiencies can cascade to limit the overall efficiency of an application. New algorithms and programming approaches are needed to address barriers to performance.
  - *Why it Matters*: This work directly targets current barriers to effective use of extreme scale systems by applications. For example, Krylov methods such as Conjugate Gradient are used in many applications currently being run on Blue Waters (MILC is one well-known example). Developing and demonstrating a more scalable version of this algorithm would immediately benefit those applications. In the longer term, the techniques that are developed will provide guidance for the development of highly scalable applications.
  - *Why Blue Waters:* Scalability research relies on the ability to run experiments at large scale, requiring tens of thousands of nodes and hundreds of thousands of processes and cores. Blue Waters provides one of the few available environments where such large-scale experiments can be run. In addition, only Blue Waters provides a highly capable I/O system, which we plan to use in developing improved approaches to extreme-scale I/O.
  - *Accomplishments:* Early results with alternative Krylov formulations have revealed several performance effects that can provide a factor of 2 or more improvement in performance at scale. Current work has been limited by the fact that the nonblocking MPI_Allreduce on Blue Waters is functional but does not provide the expected (or perhaps hoped for) performance,

particularly in terms of the ability to overlap the Allreduce operation with other communication and computation.

- **List of publications and presentations associated with this work**

Publications

Low-overhead scheduling for improving performance of scientific applications, Ph.D. Thesis, Vivek Kale. Spring 2015. https://www.ideals.illinois.edu/handle/2142/78642

Exploiting nonblocking collective operations in Conjugate Gradient, abstract for 2015 Copper Mountain Meeting, P. Eller and W. Gropp

Analyzing the Performance of a Sparse Matrix Vector Multiply for Extreme Scale Computers, Amanda Bienz, Jon Calhoun, Luke Olson, Marc Snir, William D. Gropp. Poster at SC15, Austin, TX, Nov. 2015. http://sc15.supercomputing.org/sites/all/themes/SC15images/tech_poster/tech_poster_pages/post327.html

Non-Blocking Preconditioned Conjugate Gradient Methods for Extreme-Scale Computing, P. Eller, ACM Student Research Competition Poster at SC15, Austin, TX, Nov. 2015. http://sc15.supercomputing.org/schedule/event_detail?evid=spost116

Publications in preparation

Scalability of non-Galerkin Parallel Algebraic Multigrid, A. Bienz, R. Falgout, W. Gropp, L. Olson, and J. Schroder, in preparation.

Relevant related work includes

A Hybrid Format for Better Performance of Sparse Matrix-Vector Multiplication on a GPU, Dahai Guo, William Gropp and Luke Olson, July 2015, International Journal of High Performance Computing Applications.

A multiplatform study of I/O behavior on petascale supercomputers. Huong Luu, Marianne Winslett, William Gropp, Robert B. Ross, Philip H. Carns, Kevin Harms, Prabhat, Surendra Byna, and Yushu Yao. In Thilo Kielmann, Dean Hildebrand, and Michela Taufer, editors, HPDC, pages 33–44. ACM, 2015.

Towards a more fault resilient multigrid solver. Jon Calhoun, Luke Olson, Marc Snir, and William D. Gropp. In Proceedings of the High Performance Computing Symposium, HPC '15, San Diego, CA, USA, 2015. Society for Computer Simulation International.

Presentations

These are some of the presentations that included reference to Blue Waters, including work performed using the Blue Waters professor allocation.

- Do You Know What Your I/O is Doing?, at HPC China 2015, Wuxi, China, Nov 9-12, 2015.
- Do You Know What Your I/O is Doing?, at Russian Supercomputing Days, Moscow, Russia, Sep. 28-29, 2015.
- Do You Know What Your I/O is Doing?, at 2015 Smokey Mountains Computational Sciences and Engineering Conference, Gatlinburg, Tennessee, Aug. 31-Sep. 2, 2015.
- Engineering for Performance in High Performance Computing, at ISC'15, invited workshop keynote, Frankfurt, Germany, July 13-16, 2015.
- Is MPI+X Enough for Exascale, Keynote for International High Performance Computing Forum, Tianjin, China, May 2015.
- The Future of the Message-Passing Interface, invited keynote at ISUM - 6th International Supercomputing Conference in Mexico, Mexico City, Mexico, March 9-13, 2015.
- Using MPI I/O for Big Data, invited presentation that the International Winter School on Big Data, Tarragona, Spain, January 26-30, 2015.

Plan for next year

These projects have made good progress over the last year and are expected to expand their need for scalability studies. In addition, based on the experience with poor I/O performance, we expect to begin looking at parallel I/O approaches, beyond what can be supported in our PAID IO project. The research efforts for the next year include

1. Parallel I/O autotuning and adaptivity
2. Communication optimized Krylov methods.
3. Resilient algorithms for Multigrid and multigrid preconditioned Krylov methods.

Most of these experiments study behavior at scale and typically need only a short run time but with 10,000-20,000 nodes. In order to produce timings at scale that are consistent, reproducible, and accurate, a typical run may require anywhere from a few minutes to 30 minutes per test. Thus, tests at scale may require 1,000-10,000 node-hours each. Because these tests are being used to evaluate different algorithms, most of which are developed as a result of evaluating the results of experiments on Blue Waters and at scale, it is difficult to determine a priori the amount of time that will be needed. Over the past year, we were careful to limit the scale for tests in order to limit the amount of resources consumed; as a result, we used only about 40,000 node hours. In the upcoming year, I expect several projects to run at full scale by the end of the year. If each of the 3 projects requires 2 tests at full scale (30 minutes at 20,000 nodes), along with a sequence of scaling tests (another 50%) and some development time, 100,000 node hours would be needed. Depending on the progress of the algorithm development efforts, more time (as much as the 245,000 node-hour allocation) or less may be required. An exact estimate simply is not possible for this type of basic computer science research. For a specific request, 100,000 node hours should be sufficient; however, the option for more time, up to the original allocation, is highly desirable.

Few other resources will be needed. While some IO scalability studies may require files in the multi-Terabyte range, these will be temporary files. Similarly, little networking is expected.

Estimated distribution of time: Q1: 20%, Q2: 20%; Q3: 30%; Q4: 30%

Rationale for the distribution: In the first quarter, the research projects will continue. Over the course of the year, greater scalability will be investigated, requiring larger runs. In addition I expect to add work on graph algorithms, again running some benchmarks at scale.