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; Jack Dongarra, University of Tennessee
  - 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 been exploring alternative formulations of conjugate gradient that eliminate some of the strict barrier synchronization as well as better use the memory hierarchy. Other exploratory studies have begun looking at the scalability and fault tolerance of algebraic multigrid, scaling for large graph problems, and the benefit of lightweight intranode balancing on scalability and performance.

- **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 with other communication and computation. Work on sparse matrix formats for GPUs, while it did not use any of my Blue Waters allocation in this year, was driven by the need for this capability for several of the other areas of interest; that work, which has a journal paper under review, will be used with the year two research.

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

No publications on this work so far. While I've given many talks that mention Blue Waters, and a few that have covered some of the general ideas behind this research program, the work to date has been too preliminary to feature in a presentation.

Work in preparation includes:
Low-overhead scheduling for improving performance of bulk-synchronous program on next-generation clusters of SMPs, Ph.D. Thesis, Vivek Kale. Expected in Spring 2015.

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

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, revision submitted to IJHPCA.

- **Plan for next year**

Several projects are now reaching the point where they are ready to perform scalability studies. These include

1. MPI-aware intranode load balancing.
2. Communication optimized Krylov methods.
3. Resilient algorithms for Multigrid and multigrid preconditioned Krylov methods.

In addition, new work in the following areas is beginning as part of collaborations with Jack Dongarra's group at the University of Tennessee (we may also begin a collaboration with Jim Demmel's group on some related projects).

1. Communication limited algorithms, including ones for graph problems.
2. Resilient algorithms for linear algebra.

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 20,000 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 5 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, 170,000 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, 170,000 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, we will be gathering data at scale for fine-grain load balancing approaches applied to several benchmarks and one application. Startup work in resilience and continuing work on Krylov methods will also take place. In the second quarter, collaborations with Dongarra's group will gain steam. In the third and fourth quarters, those projects will continue, running at greater scale, and in addition I expect to add work on graph algorithms, again running some benchmarks at scale.

The report should be submitted as a PDF file at: https://bluewaters.ncsa.illinois.edu/blue-waters-professors-reports. There are no specific formatting instructions. It is recommended to include visualizations, charts and other images in the document to help illustrate the results.