

**CUDA Center of Excellence at Illinois
CUDA Achievement Award Submission**

“GPU Supercomputing in Blue Waters”

Wen-mei Hwu, UIUC and NCSA

The Blue Waters supercomputer at the University of Illinois contains 3072 Kepler GPUs, totaling about 4 Peta FLOPS of peak compute throughput. There are a total of 25,712 nodes in Blue Waters. 22,480 of them are Cray XE6 nodes that contain two AMD Interlagos CPUs each, totaling 7.1 Peta FLOPS. 3,072 of these nodes are Cray XK7 nodes that consist of one AMD Interlagos CPU and one NVIDIA Kepler GPU, totaling 4 Peta FLOPS.

The UIUC CUDA Center of Excellence PI has been a Co-PI of the Blue Waters project and proposed the inclusion of Kepler GPUs in 2011. At the time, the NSF review panel and the Blue Waters leadership team were concerned about the lack of whole-application, production-use evidence of GPUs at scale. The UIUC CCoE PI worked closely with Marc Adams and an NVIDIA team to assemble such data. Unfortunately, at the time of the proposal, the data collected from the application teams fell short due to one or more of the following issues (1) experiments were done with paths/models in applications that are not typically exercised in production use, (2) the time comparisons were done for computation intensive phases of the applications rather than whole-application execution time including I/O and start-up time, (3) the experiments were done at a scale that is too small for Blue Waters. A survey showed that about 1/3 of the Blue Waters allocation teams have significant GPU activity. Going all the way to using XK7 nodes for the entire machine was viewed to be too risky for the NSF research community. Ultimately, a little over 1/3 of the peak performance of Blue Waters was assigned to GPUs, which translates into 3,072 Kepler GPUs.

As part of the agreement, the UIUC CCoE PI agreed to assemble and lead a taskforce to prove that the GPUs in XK7 nodes provides at least 3X full-application performance advantage over using CPUs alone in these nodes when running NSF-designated Blue Waters applications (NAMD, QMCPACK, Chroma, and GAMESS), at a scale no less than 600 nodes. This was part of the Blue Waters acceptance criteria. Additionally, the taskforce is strongly encouraged to demonstrate that X7 nodes significant performance advantage over XE6 nodes. This was not set as an acceptance criteria because NSF was concerned about increasing the risk of the acceptance process. Furthermore, the UIUC CCoE PI was asked to offer a VSCSE Summer School on Proven Algorithm Techniques for Many-core Processors to enhance the GPU use capability of the Blue Waters application teams. The taskforce consists of UIUC CCoE researchers, NVIDIA engineers, Cray engineers, and NCSA staff, listed alphabetically:

Marc Adams (NVIDIA), Sarah Anderson (Cray), Greg Bauer (NCSA), Eric Bohm (UIUC), Mark Berger (NVIDIA), Bret Bode (NCSA), Robert Brunner (NCSA), Jerry Chen (Mike Clark (NVIDIA), Massimiliano Fatica (NVIDIA), Joe Glenski (Cray), Wen-mei Hwu (UIUC, Lead), Sanjay Kale (UIUC), Hee-Seok Kim (UIUC), Jeonnim Kim (NCSA/ORNL), David Kirk (NVIDIA), Ryan Mokos (NCSA), Ryan Olson (Cray), Jim Phillips (UIUC), Carl Ponder (NVIDIA), Dale Southard (NVIDIA), Sarah Tariq (NVIDIA), Kevin Thomas (Cray), Przemyslaw Tredak (NVIDIA), Cliff Woolley (NVIDIA), Cyril Zeller (NVIDIA)

The task force is a great example of the type of large-scale, cross-institution collaborative initiative that can only be successfully executed with the support of a CCoE.

Performance Results

The first set of results compare the performance of running applications on both the CPU and GPU in X7 nodes versus using only the CPU in these X7 nodes. This measurement quantifies the multiplicative factor of performance by adding a Kepler into each XK7 node. The execution time is measure from the launch of the application to the end of the application, including all I/O and data transfer activities. The results are summarized as follows in Table 1. Adding a Kepler GPU multiplies the whole application performance by 3.1 to 4.9, exceeding the expectations of NSF.

Table 1 Performance advantage of using CPU+GPU in each X7 node versus using only the CPU in each X7 node

	NAMD	QMCPACK	Chroma	GAMESS
# of nodes used	768	700	768	1536
CPU-only time (sec)	11833.7	4477.0	1244.5	14637.5
CPU+GPU time (sec)	3484.5	908.3	320.2	4682.7
Ratio	3.4	4.9	3.9	3.1
Test case used (Each is selected to reflect the intended production runs on Blue Waters).	The 100M atom "chromataphore" benchmark was run for 60000 time steps. A 2-femtosecond time-step was used, with output of atom positions to DCD files with parallel writers.	Graphite 4x4x1 (256 electrons), VMC followed by DMC with 179,200 DMC Walkers. The scientific result of each run is an energy value with a computed error bar.	Solution for all 12 spin-color components of the quark propagator. On the GPU two solvers were used for the Dirac equation. One was BiCGStab with algorithmic improvement to allow mixed precision. The other was a GCR algorithm with a domain-decomposed Additive-Schwarz solver.	A many-body expansion to estimate the full CCSD(T) correlation energy for a system of 32 water molecules by calculating 1-, 2-, and 3-body terms. The monomer, dimer and trimer CCSD(T) calculations are performed in parallel with 384 concurrent calculations taking place over 1536 XK7 nodes.

The second set of results compare the performance of running applications on CPU and GPU in X7 nodes versus using both CPUs in XE6 nodes. The experiments are done with the same test cases as the previous experiment. The execution time is measure from the launch of the application to the end of the application, including all I/O and data transfer activities. This measurement quantifies the performance advantage of X7 nodes over XE6 nodes. Replacing XE6 nodes with XK7 nodes multiplies the whole-application performance by 1.8 to 2.7 times. These results indicate that including GPUs in Blue Waters was definitely the right decision for these applications. Keep in mind that the XE6 and XK7 nodes have comparable price and power. The improvements in performance per watt and performance per watt both exceeded the expectation of the original NSF review panel.

Table 2 Performance Advantage of using XE6 nodes (2 CPUs per node) vs. using XK7 nodes (CPU+GPU per node)

	NAMD	QMCPACK	Chroma	GAMESS
# of nodes used	768	700	768	1536
XE6 time (sec)	6620.6	2452.4	1244.5	To be confirmed
XK7 time (sec)	3484.5	908.3	320.2	4682.7
Ratio	1.8	2.7	2.4	To be confirmed

The Blue Waters GPU Computing experience has validated that GPUs are ready for prime time. The GPUs improve the performance per watt of Blue Waters for these applications by 1.8-2.7X for these NSF selected applications. These results are being submitted to a major ACM publication venue.

GPU Programming Libraries, Tools and Education

The current concern is to enable many more Blue Waters application teams to take advantage of GPUs. The UIUC CCoE is working to provide more productivity tools and libraries to the Blue Waters application teams. The two that are most mature are the data movement library GAMC [Asplos'10] and data layout transformation library DL [InPar'12-b]. A tri-diagonal solver that matches the speed of the current CUSPARSE solver and the numerical stability of the Intel MKL has become available to the application teams and being incorporated into CUSPARSE [SC'12]. A Linux prototype implementation of C++AMP based on LLVM, GMAC and DL has also been developed. These tools and libraries complement those provided by NVIDIA and Cray.

The UIUC CCoE researchers have published their work on algorithm and data optimizations at InPar'12 [InPar'12-a] and the widely circulated IEEE Computer Magazine [Computer'12]. These techniques have been taught in VSCSE 2012 Summer School, BSC 2012 PUMPS Summer School, and a 2013 UIUC/Minnesota/Ohio State shared graduate course. The UIUC CCoE has also offered a Coursera Heterogeneous Parallel Programming MOCC course with more than 10,000 active students and 25,000 students who registered for a future offering.

[Asplos'10] Gelado, J.E. Stone, J. Cabezas, S. Patel, N. Navarro and W.W. Hwu, "An Asymmetric Distributed Shared Memory Model for Heterogeneous Parallel Systems," The ACM/IEEE 15th International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS'10). March 2010. Pittsburgh, PA.

[InPar'12-a] J. A. Stratton, N. Anssari, C. I. Rodrigues, I. Sung, N. Obeid, L. Chang, G. Liu, and W. Hwu, Optimization and Architecture Effects on GPU Computing Workload Performance, IEEE Innovative Parallel Computing, San Jose, May, 2012.

[InPar'12-b] I. Sung, G. Liu, W. Hwu, DL – A Data Layout Transformation System for Heterogeneous Parallel Computing, IEEE Innovative Parallel Computing, San Jose, May, 2012.

[Computer'12] Algorithm and Data Optimization Techniques for Scaling to Massively Threaded Systems, J. A. Stratton, N. Anssari, I. Sung, N. Obeid, L. Chang, G. Liu, and W. Hwu, IEEE Computer Special Issue on Massively Threaded Systems, 2012.

[SC'12] L. Chang, J.A. Stratton, H. Kim, and W.W. Hwu, "A Scalable, Numerically Stable Tridiagonal Solver Using GPUs," The International Conference for High-Performance Computing Networking, Storage, and Analysis (SC'12), Salt Lake City, 2012.