Award: DE-FG02-05ER25662
Title: A Framework for Adaptable Operating and Runtime Systems

PI: Prof. Patrick G. Bridges
Institution: University of New Mexico

Date of Final Report: February 1, 2012

1 Summary of Results

In this grant, we examined a wide range of techniques for constructing high-performance configurable system software for HPC systems and its application to DOE-relevant problems. There were three major directions to the research which resulted in a number of publications presented at high impact venues. In addition, this research spawned follow-on research projects with significant DOE research impact, and supported the education of several students who now work for DOE labs after receiving their Ph.D. degrees. In the remainder of this report, we describe these results in detail.

2 Research Conducted

Overall, research and development on this project focused in three specific areas: (1) software frameworks for constructing and deploying configurable system software, (2) applcation of these frameworks to HPC-oriented adaptable networking software, (3) performance analysis of HPC system software to understand opportunities for performance optimization.

2.1 Frameworks for HPC System Software

We examined two specific configurable system software frameworks in this project, specifically the THINK and Cactus software frameworks. THINK provided a framework for fine-grained composition of operating systems, while Cactus was focused more on networking software.

Our work with THINK was partially successful, as we ported it to X86 systems and demonstrated its ability to be configured to provide different system software functionality. However, porting THINK to real HPC systems, applications, and runtimes proved daunting—the specialized functionality required by these systems made such work time consuming and difficult. As a result, we eventually moved away from allowing every component in the operating system to configured, and focused instead of configurability in key subsystems. A description of some of our work on using THINK for HPC systems was published in the journal *Operating Systems Review* [8].

In contrast to the work in THINK, Cactus proved much more appropriate to HPC system software, particularly networking software. Because it was specifically crafted to address networking configurability challenges, it was comparatively easy to modify it to address HPC-oriented runtime issues. Our initial work in this direction, MPI/CTP, resulting in a highly configurable MPI implementation built in the Cactus framework. Results of this work were published in the well-known EuroMPI conference [10].

2.2 Adaptable Networking Software

Building on the work on the Cactus framework and MPI/CTP, we also conducted research on the potential for application-oriented system software adaptation in the networking stack. Our specific focus in this work was communication support for dynamic HPC applications. Previous MPI implementations did not change transport protocols or allocate resources based on the application characteristics, resulting in degraded application performance. To address this, we built the Protocol Reconfiguration and Optimization system for MPI (PROMPI) to meet the needs of dynamic modern HPC applications. PROMPI used profiles of past application communication characteristics to dynamically reconfigure MPI protocol choices.

In this work, we demonstrated that dynamic reconfiguration can improve the performance of important MPI applications when exact communication profiles are known. We also demonstrated that profiles from past application runs with different (but related) inputs can be used to optimize the performance of later application runs. The results of this work were presented in the 2009 Workshop on Communication Architectures for Clusters (CAC 2009) [11].

2.3 OS Noise Analysis

In addition to this research, we also conducted important performance analyses in existing HPC operating systems to understang their key characteristics. Because of its importance to HPC systems, much of this work focused on understanding the impact of OS noise on HPC application performace. This performance analysis was essential to understand potential performance tradeoffs in different HPC system software configurations.

To quantify the impact of OS noise on DOE application performance, we modified the Sandia catamount operating system so that we could inject different levels of noise into the operating system. We then ran a number of DOE-relevant applications (CTH, SAGE, and Pop) on the Sandia ASC Red Storm system with different amounts and kinds of injected noise. Our results showed that key DOE applications are highly sensitive to noise, with high-frequency/low-duration noise having less impact on application performance than low-frequency/high-duration noise. Results of this work were published in a number of places, including the 2008 ACM/IEEE Supercomputer Conference (SC'08), where it was nominated for Best Paper and Best Student Paper awards [4, 5].

3 Related Follow-on Projects and Publications

In addition to the direct research conducted on this project, the frameworks and infrastructure developed as part of this project resulted in several related projects. First among these is our current DOE grant on scalable virtualization for HPC systems (DE-SC0005050) which has resulted in many high-impact results. In addition, work on replication-parallel networking for HPC systems leveraged the work on this project on the Cactus framework and resulted in a high-impact publication in the well-known HotOS Symposium [2].

4 Funded Students Degrees Conferred

Manjunath Gorentla Venkata: Conducted research on the use of the Cactus framework for configurability and adaptaion in MPI, and the use of these adaptation capabilities to improve HPC application performance.

Degree Granted: Ph.D., UNM, 2010

Current Position: R&D Associate, Oak Ridge National Laboratory

Kurt B. Ferreira: Conducted research on the impact of HPC OS noise on HPC application performance

Degrees Granted: M.S., UNM, 2008; Ph.D., UNM, 2011;

Current Position: Member of Technical Staff, Sandia National Laboratories

Publications

[1] Patrick G. Bridges, Arthur B. Maccabe, and Orran Krieger. System software for high-end computing. *Operating Systems Review: Special Issue on System Software for High-End Computing Systems*, 40(2), April 2006.

[2] Patrick G. Bridges, Donour Sizemore, and Scott Levy. Exploiting MISD performance opportunities in multi-core systems. In *Proceedings of the 13th Workshop on Hot Topics in Operating Systems (HotOS XIII)*, Napa, CA, May 2011.

[3] Kurt B. Ferreira. *Characterizing HPC Application Sensitivity to OS Interference Using a Kernel-level Noise Injection Framework*. PhD thesis, The University of New Mexico, Computer Science Department, Albuquerque, NM, 87131, 2008.

[4] Kurt B. Ferreira, Ron Brightwell, and Patrick G. Bridges. An infrastructure for characterizing the sensitivity of parallel applications to OS noise. In *Proceedings of the 2006 Symposium on Operating System Design and Implementation*, 2006. Work-In-Progress Session.

[5] Kurt B. Ferreira, Ron Brightwell, and Patrick G. Bridges. Characterizing application sensitivity to OS interference using kernel-level noise injection. In *Proceedings of the 2008 ACM/IEEE Conference on Supercomputing (SC08)*, November 2008.

[6] Arthur B. Maccabe, Patrick G. Bridges, Ron Brightwell, and Rolf Riesen. Recent trends in operating systems and their applicability to HPC. In *Proceedings of the Cray User Group 2006 Technical Meeting*, Lugano, Switzerland, May 2006.

[7] Rolf Riesen, Ron Brightwell, Patrick G. Bridges, Trammell Hudson, Arthur B. Maccabe, Patrick M. Widener, and Kurt B. Ferreira. Designing and implementing lightweight kernels for capability computing. *Concurrency and Computation: Practice and Experience*, 21(6):791–817, April 2009.

[8] Jean-Charles Tournier, Patrick G. Bridges, Arthur B. Maccabe, Patrick M. Widener, Zaid Abudayyeh, Ron Brightwell, Rolf Riesen, and Trammell Hudson. Towards a framework for dedicated operating systems development in high-end computing. *Operating Systems Review: Special Issue on System Software for High-End Computing Systems*, 40(2):16–21, April 2006.

[9] Manjunath Gorentla Venkata. *A Protocol Reconfiguration and Optimization System for MPI*. PhD thesis, The University of New Mexico, Computer Science Department, Albuquerque, NM 87131, 2010.

[10] Manjunath Gorentla Venkata and Patrick G. Bridges. MPI/CTP: A reconfigurable MPI for HPC applications. In Dieter Kranzlmüller, Peter Kacsuk, Jack Dongarra, and Jens Volkert, editors, *Recent Advances in Parallel Virtual Machine and Message Passing Interface: 13th European PVM/MPI Users' Group Meeting*, volume 4192 of *Lecture Notes in Computer Science*. Springer-Verlag, 2006.

[11] Manjunath Gorentla Venkata and Patrick G. Bridges. Using application communication characteristics to drive dynamic MPI reconfiguration. In *Proceedings of the 2009 Workshop on Communication Architectures for Clusters (CAC 2009)*, May 2009.