Award: DE-SC0005050
Title: Enabling Exascale Hardware and Software Design through Scalable System Virtualization

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

Date of Final Report: February 1, 2015

1 Summary of Results

In this grant, we enhanced the Palacios virtual machine monitor to increase its scalability and suitability for addressing exascale system software design issues. This included a wide range of research on core Palacios features, large-scale system emulation, fault injection, perfomrance monitoring, and VMM extensibility. This research resulted in large number of high-impact publications in well-known venues, the support of a number of students, and the graduation of two Ph.D. students and one M.S. student. In addition, our enhanced version of the Palacios virtual machine monitor has been adopted as a core element of the Hobbes operating system under active DOE-funded research and development.

2 Research Conducted

Research conducted at UNM on this project covered a wide range of areas in collaboration with research at Northwestern University, the University of Pittsburgh, Oak Ridge National Lab, and Sandia National Labs. UNM research focused primarily in the following areas: (1) Basic VMM Enhancements; (2) Virtual network integration and optimization; (3) Fault injection for resilience studies; (4) Architecural simulation; and (5) VMM extensibility.

2.1 Basic VMM Enhancements

Our first work on the project was with our collaborators at Northwestern and Sandia on basic enhancements to Palacios for scalability and supporting modern hardware and operating systems. A general description of Palacios and its scalability, partially as a result of this work, was published in the 2011 ACM SIGPLAN/SIGOPS International Conference on Virtual Execution Environments in Newport Beach, CA .

In terms of defined milestones, this work covered multi-core support to Palacios for both Intel and AMD processors **(Year 1 Milestone 1)** and pass-through support for Infiniband, Intel 10GigE, and Cray SeaStar network interfaces **(Year 1, Milestone 2)**. In addition, it included initial work o integrating AMD I/O MMU support done in a prior version of Palacios **(Year 1, Milestone 3)** which was later integrated into Palacios. We also worked to ported Palacios to the Linux host operating systems and integrated into into the Cray production operating systems **(Year 2, Milestone 2)**. Finally, we added support for the virtualizing hardware machine check architecture **(Year 2, Milestone 7)** and for virtualizing system performance counters **(Year 2, Milestone 4; Year 3, Milestone 4)**.

2.2 Overlay Network Integration and Optimization

UNM has collaborated with Northwestern on successfully integrating the VNET overlay network capabilities in Palacios **(Year 1, Milestone 4)**. UNM's focus in this collaborative research with Northwestern has been on supporting high-speed VM-level networking using Infiniband and 10 Gigabit Ethernet adapters. This has included support for a variety of network offload techniques, and provides support for a wide range of networking technologies in the virtual machine monitor. This research resulted in high impact publications [12, 8, 7]. UNM student Zheng Cui who led research in this direction at UNM received her Ph.D. in 2013 [6] and now works at VMWare, Inc. in Palo Alto, California.

2.3 Large-scale Emulation

We also demonstrated the ability of Palacios to aid in the simulation and emulation of large-scale systems with Palacios. In particular, we implemented full time-dilation support in Palacios, allowing us to emulate large systems using the VMM than the actual hardware on which Palacios runs. As part of this work, we also rewrote large portions of Palacios' time management system and extended it to support emulating performance-heterogeneous cores in Palacios. We are also worked in collaboration with Sandia National Labs and Northwestern to integrating this work with the SST toolkit and the GEM5 architectural simulator **(Year 2, Milestone 3)**. A workshop paper and two journal publication describing our work in this direction were also published [3, 1].

2.4 Fault Injection

We also conducted substantial work increasing Palacios's support for fault injection. Specifically, we added support for injecting DRAM memory failures and disk failures into Palacios, two failure modes prevalanet in large-scale computational and storage systems **(Year 2, Milestone 8).** As part of this work, we also examined the potential for application level recovery from failures to drive our work in what faults to inject and what kind of interfaces to present to the OS and application in terms of failures. This work resulted in several different publications [2, 10].

2.5 VMM Extensibility

To support VMM extensibility and configurability, we added a simple extension interface to Palacios to support extending the virtual machine monitor in simple ways; the initial interface eas simple, allowing extensions to hook into existing VM hooks on individual processors and supporting simple lock-based/single-processor extensions **(Year 2, Milestone 6)**. We extended this interface in two ways: first, we added support for hooking additional features such as individual exits on particular cores to this interface to enable a wider range of more powerful extensions and configurations. Second, we examined the use of more complex event-based lock-free in Palacios **(Year 3, Milestone 5)**. While this work resulted in publications [4] and partially supported a graduate student who received his Ph.D. [11] and now works at Two Sigma Investments in New York, NY, in the end it proved unnecessarily complicated for inclusion in Palacios.

3 Related Follow-on Projects and Publications

In addition to the direct research conducted on this project, the frameworks infrastructure developed as part of this project resulted in several related projects. Most visible of these is the recent DOE-funded Hobbes Exascale Operating System Project led by Oak Ridge and Sandia National Labs. This project is using virtualization to research and develop techniques for next-generation operating systems, and the Palacios virtual machine monitor enhancements developed as part of this project are a key element of the Hobbes operating system [5].

4 Funded Students and Degrees Conferred

Donour Sizemore: Conducted research on parallelization of high-speed networking systems, particularly in component-based software

Degree Granted: Ph.D., UNM, 2011

Current Position: Two Sigma Investments, New York, NY

Philip Soltero: Conducted research on failure injection with Palacios, particularly the injection of memory failures into virtual machines.

Degree Granted: M.S., UNM, 2012

Current Position: Engineer Analyst, Tau Technologies

Cui Zheng: Conducted research on implementing and optimizing virtual networking for HPC virtual machine monitors.

Degrees Granted: Ph.D., UNM, 2013

Current Position: Member of Technical Staff III, VMWare

Publications

[1] Patrick G. Bridges, Dorian Arnold, Kevin T. Pedretti, Madhav Suresh, Feng Lu, Peter Dinda, Russ Joseph, and Jack Lange. Virtual machine-based emulation of future generation high-performance computing systems. *International Journal of High Performance Computing Applications*, 26(2):125–135, May 2012.

[2] Patrick G. Bridges, Mark Hoemmen, Kurt B. Ferreira, Michael A. Heroux, and Ron Brightwell. Cooperative application/OS DRAM fault recovery. In *Fourth Workshop on Resiliency in High Performance Computing in Clusters, Clouds, and Grids (Resilience'11)*, Bordeaux, France, August 2011.

[3] Patrick G. Bridges, Kevin Pedretti, and Dorian Arnold. VM-based slack emulation of large-scale systems. In *Proceedings of the 2011 International Workshop on Runtime and Operating Systems for Supercomputers (ROSS 2011)*, Tucson, AZ, May 2011.

[4] 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.

[5] Ron Brightwell, Ron Oldfield, Arthur B. Maccabe, and David E. Bernholdt. Hobbes: Composition and virtualization as the foundations of an extreme-scale os/r. In *Proceedings of the 3rd International Workshop on Runtime and Operating Systems for Supercomputers*, ROSS '13, pages 2:1–2:8, New York, NY, USA, 2013. ACM.

[6] Zheng Cui. *Enhancing HPC on Virtual Systems in Clouds through Optimizing Virtual Overlay Networks.* PhD thesis, University of New Mexico, Albuqueruque, New Mexico, 2013.

[7] Zheng Cui, Patrick G. Bridges, John R. Lange, and Peter A. Dinda. Virtual TCP offload: optimizing Ethernet overlay performance on advanced interconnects. In *Proceedings of the 22nd International Symposium on High-performance Parallel and Distributed Computing (HPDC'13)*, pages 49–60. ACM, 2013.

[8] Zheng Cui, Lei Xia, Patrick G. Bridges, Peter A. Dinda, and Jack R. Lange. Optimizing overlay-based virtual networking through optimistic interrupts and cut-through forwarding. In *Proceedings of the 2012 ACM/IEEE Conference on Supercomputing (SC'12)*, Salt Lake City, UT, November 2012.

[9] Jack Lange, Kevin Pedretti, Peter Dinda, Patrick G. Bridges, Chang Bae, Philip Soltero, and Alexander Merritt. Minimal-overhead virtualization of a large scale supercomputer. In *Proceedings of the 2011 ACM SIGPLAN/SIGOPS International Conference on Virtual Execution Environments (VEE 2011)*, Newport Beach, CA, March 2011.

[10] Scott Levy, Matthew G.F. Dosanjh, Patrick G. Bridges, and Kurt B. Ferreira. Using unreliable virtual hardware to inject errors in extreme-scale systems. In *Proceedings of the 3rd Workshop on Fault-tolerance for HPC at extreme scale*, pages 21–26. ACM, 2013.

[11] Donour Sizemore. *Parallel Network Protocol Stacks Using Replication.* PhD thesis, The University of New Mexico, Computer Science Department, Albuquerque, NM, 87131, 2011.

[12] Lei Xia, Zheng Cui, John Lange, Yuabn Tang, Peter Dinda, and Patrick G. Bridges. VNET/P: Bridging the cloud and high performance computing through fast overlay networking. In *Proceedings of the 21st International ACM Symposium on High-Performance Parallel and Distributed Computing (HPDC'12)*, June 2012.