Final Scientific/Technical Report for "Enabling Exascale Hardware and Software Design Through Scalable System Virtualization"

DOE Award: DE-SC0005343

Institution: Northwestern University

Project Title: Enabling Exascale Hardware and Software Design Through Scalable System Virtualization

PI: Peter A. Dinda

Consortium: Northwestern University (Peter A. Dinda, PI, Russ Joseph, Fabian Bustamante),
University of Pittsburgh (Jack Lange, PI, via subcontract from Northwestern),
University of New Mexico (Patrick Bridges, PI and coordinating PI, Dorian Arnold)
Sandia National Labs (Kevin Pedretti, PI)
Oak Ridge National Lab (Stephen Scott, PI)

Date: 03/17/2015

Executive Summary

The purpose of this project has been to extend the state of the art of systems software for high-end computing (HEC) platforms, and to use systems software to better enable the evaluation of potential future HEC platforms, for example exascale platforms. Such platforms, and their systems software, have the goal of providing scientific computation at new scales, thus enabling new research in the physical sciences and engineering. Over time, the innovations in systems software for such platforms also become applicable to more widely used computing clusters, data centers, and clouds.

This was a five-institution project, centered on the Palacios virtual machine monitor (VMM) systems software, a project begun at Northwestern, and originally developed in a previous collaboration between Northwestern University and the University of New Mexico. In this project, Northwestern (including via our subcontract to the University of Pittsburgh) contributed to the continued development of Palacios, along with other team members. We took the leadership role in (1) continued extension of support for emerging Intel and AMD hardware, (2) integration and performance enhancement of overlay networking, (3) connectivity with architectural simulation, (4) binary translation, and (5) support for modern Non-Uniform Memory Access (NUMA) hosts and guests. We also took a supporting role in support for specialized hardware for I/O virtualization, profiling, configurability, and integration with configuration tools.

The efforts we led (1-5) were largely successful and executed as expected, with code and papers resulting from them. The project demonstrated the feasibility of a virtualization layer for HEC computing, similar to such layers for cloud or datacenter computing. For effort (3), although a prototype connecting Palacios with the GEM5 architectural simulator was demonstrated, our conclusion was that such a platform was less useful for design space exploration than anticipated due to inherent complexity of the connection between the instruction set architecture level and the microarchitectural level. For effort (4), we found that a code injection approach proved to be more fruitful.

The results of our efforts are publicly available in the open source Palacios codebase and published papers, all of which are available from the project web site, v3vee.org. Palacios is currently one of the two codebases (the other being Sandia's Kitten lightweight kernel) that underlies the node operating system for the DOE Hobbes Project, one of two projects tasked with building a systems software prototype for the national exascale computing effort.

Comparison of Accomplishments and Goals/Objectives

The following analysis uses Sections 6.1-6.10 of our original proposal (the proposed work sections) and Table 1 (Milestones). We focus on the specific items in which Northwestern (including our subcontract with the University of Pittsburgh) took a leadership or participatory role.

- (Participant) Section 6.1, "Continued expansion of architectural support", specifically I/O MMU support. I/O MMU support was developed as anticipated. Two frameworks for guest access to hardware PCI devices (one that is "always on", and one that is selectively enabled only for trusted components of the guest) were developed to make can make use of it.

- (Lead) Section 6.2 "Multicore and multiprocessor support", specifically support for Intel/AMD multicore, and NUMA support. Both functionalities were developed as anticipated. At the beginning of the project, Palacios supported single core guests on SMP hosts. At the end of the project, it supported the creation of SMP and NUMA guests on SMP and NUMA hosts. Additionally, we developed mechanisms to support both the static and dynamic mapping of the guest NUMA configuration to the host NUMA configuration. We evaluated NUMA support both on 2 socket machines (typical of today's HEC nodes) and on 4 socket machines. Using these mechanisms, a Northwestern Ph.D. thesis on how to automatically control them was completed.

- (Lead) Section 6.3, "Widened communication support", specifically integration with overlay networks and networks on chip. We developed a version of our VNET overlay network system that is embedded in Palacios itself and thus has the potential to perform at speeds needed in HEC. VNET can provide a simple, layer 2 (Ethernet) abstraction over different underlying network hardware, thus allowing virtual machine portability, mobility, and the possible connection of HEC resources and cloud/datacenter resources. By the end of the project, VNET was demonstrated on 1 and 10 Gbit Ethernet, Infiniband, and the Cray Seastar network. Along with the University of New Mexico, we carefully studied the issues of actually achieving high throughput and low latency in these kinds of environments, leading to a succession of papers, as well as a University of New Mexico Ph.D. thesis on this topic. A closely related Northwestern Ph.D. thesis on applying fast networking technology to the problem of tracking content across a supercomputer was also completed. We considered the interaction of networks on chip and virtualization analytically. However, we were unable to find a current hardware NOC or simulation that allowed for OS/VMM-level configuration, which limited empirical study.

- (Lead) Section 6.6 "A bridge between the VMM and micro-architectural simulation". We developed an integration of the Palacios VMM and the GEM5 microarchitectural simulator as described in the proposal. The proof of concept implementation is available as a set of patches on the v3vee.org web site. A core element of this work was considerable enhancement of Palacios's VM checkpoint capability (as well as VM migration), which has broad utility. The other element was bidirectional translation between Palacios and GEM5 checkpoints. This proved to be considerably more difficult than anticipated, both for practical reasons (GEM5's limited documentation of their checkpoint format), and for more fundamental reasons, namely that translating from an ISA-level checkpoint (e.g., Palacios) to a microarchitectural checkpoint (e.g. GEM5) means that microarchitectural state needs to be constructed from scratch. Not only may this be impossible in some circumstances, even if we construct "correct" microarchitectural state, it can lead to misleading performance characteristics once the simulator executes (e.g., empty cache, etc). Our conclusion is that the apparent benefits of a VMM/microarchitectural simulator integration (which had never been attempted before) do not hold up in practice.

- (Lead) Section 6.7 "Binary translation for instruction set extensions and fault injection", specifically binary translation. Our effort here lead to the GEARS framework, which is primarily concerned with code injection into the guest, including injection of trusted code. GEARS is included within the Palacios codebase. We also developed a basic capability for instruction interception and emulation and used it to develop a proof-of-concept system that emulates Intel hardware transactional memory on systems that do not support it.

- (Lead) Section 6.8 "Enabling vertical runtime profiling", specifically monitoring capabilities. We developed interfaces to the Intel and AMD hardware performance counters (the PMU), and to the Intel RAPL power monitoring/control hardware. In addition, the GEARS framework allows interception and interposition on system calls. Finally, we demonstrated tracking of both memory access patterns within the node, and memory content across nodes. These functionalities are available within the Palacios codebase.

- (Participant) Section 6.9 "Compile-time composability and run-time extensibility". We contributed to this effort through the development of the GEARS framework, which allows for directly modifying the guest kernel and user space. We also made our own contributions to the Palacios codebase configurable, so Palacios can be built with or without them. Finally, we added extensive Palacios and guest run-time configuration scripts.

Project Activities

The project had five partner institutions (Northwestern, University of New Mexico, University of Pittsburgh (via Northwestern subcontract), Sandia National Labs, and Oak Ridge National Lab). The institutions worked together closely, holding a weekly project meetings and weekly all-day software development meetings, both by teleconferencing. In addition, we held several joint "hackfest" activities where we physically gathered at a site and worked together to achieve certain milestones, such as software releases. These events occurred at the Northwestern, University of New Mexico, Sandia National Labs, and Oak Ridge National Lab.

Within the 2011-12 timeframe, Dinda spent six months on sabbatical at the University of New Mexico and Sandia National Labs. This allowed for a close direct working relationship among the PIs at the three sites, and our students. Dinda's student Kyle Hale finished a nine month stay at New Mexico, including a summer internship at Sandia. Dinda served as a committee member for a University of New Mexico student on the project, and University of New Mexico PI Patrick Bridges served as one for a Northwestern student on the project.

For each year of the project, Dinda taught a Northwestern graduate course in operating systems design and implementation. This project-oriented course focused on the Palacios codebase and served both as a springboard for work within the project, and also an attractor for competent graduate and advanced undergraduate students.

Palacios Development: A unifying activity in the project was the continued development and enhancement of the Palacios VMM. This continued throughout the project and was where the majority of time was spent by Northwestern. During the course of the project a milestone release (1.3) was made that included SMP multicore support on Intel and AMD, I/O MMU, the VNET/P overlay network, and a range of other enhancements. Work beyond this was included in the development branch of the project, which is also publicly accessible via our web site, which includes a web interface to our git repository. This work included NUMA support, overlay networking enhancements, the GEARS toolchain, Intel hardware transactional memory emulation, and performance and power monitoring mechanisms. As of the end of the project (August 2014), all elements of the work described above were included in this codebase and available for all.

VNET/P: A major effort involved incorporating our VNET overlay network for virtual machines into the Palacios VMM and investigating how to make it fast enough for HEC uses. This effort spanned several years, a Ph.D. thesis (Zheng Cui, University of New Mexico), and major efforts on both our and University of New Mexico's part. The end result was, I believe, the fastest demonstrated software-based overlay network at the time of its publication, and probably still is. 10 Gbit and higher speeds were demonstrated on Ethernet, Infiniband, and Seastar hardware.

GEARS and Guarded Modules: Northwestern designed and implemented the Guest Examination and Revision Services (GEARS) extensions to Palacios. GEARS provides the ability to manipulate the guest kernel and user space through code injection, environment modification, and system call interposition. GEARS allows us to create VMM-based services that have components implemented in the guest itself, even in an uncooperative guest. In our initial paper on GEARS, we demonstrated two such services: an MPI acceleration service for VMs that are collocated, in which MPI intercepts are injected into the user-level MPI program, and an enhanced VNET/P service, in which the forwarding component of VNET/P is injected into the guest kernel as a driver. We also prototyped a system call migration service, that allows the VMM to determine on which core a guest's system calls are executed. Guarded Modules is functionality that extends GEARS to securely support trusted code. With Guarded Modules, we can inject a piece of the VMM itself into the guest, and allow it, and only it, to run with full privilege. We demonstrated how to do so for a device driver for a network interface.

Palacios/GEM5 Connectivity: Northwestern developed a proof-of-concept integration of the GEM5 microarchitectural simulator and the Palacios VMM. This is available as patches from our web site. We were able to take a checkpoint in Palacios, translate it into a Gem5 compatible form, and then resume the translated checkpoint in GEM5. The reverse (GEM5 to Palacios) was also achieved. Migration can be trigged by a hypercall, or the execution of a special instruction. As far as we are aware, ours is the first such integration. As we described earlier, however, we reached the conclusion that such an integration is likely inherently complex enough to limit its merits.

Transactional Memory Emulation: Intel's TSX instruction set adds hardware transactional memory to the x86. We developed a technique to emulate the memory accesses of an instruction stream without requiring full instruction emulation, and only requiring minimal instruction decoding. Based on this general technique, we implemented an emulation of Intel's TSX within Palacios. It was about 60 times faster than Intel's simulator. The code is available within Palacios.

Performance, Power, and System Monitoring: Northwestern developed and made available in Palacios tools for collecting hardware performance monitoring data (PMU data) on Intel and AMD processors, as well as power monitoring/capping data on Intel processors (RAPL data). We also prototyped DVFS control within Palacios. Using GEARS, we protoyped acquisition of system calls from the guest, including the ability to selectively choose which calls to acquire to reduce overhead.

Adaptive Virtualization Policy on NUMA machines: A Ph.D. thesis at Northwestern (Chang Bae) developed VMM-based inference mechanisms and adaptation policies for using the NUMA mechanisms within Palacios to optimize guests for performance, power, or energy. The work exhaustively considers 2 socket NUMA machines (common HEC nodes today) and the mechanisms of virtual core mapping, NUMA memory mapping, and core scheduling, for under-, at-, and over-subscription of resources. A prototype system, NAVAR, built on this analysis and Palacios demonstrated the feasibility of applying the mechanisms and policies in a real system.

Memory Content Tracking Across a Supercomputer: A Ph.D. thesis at Northwestern (Lei Xia) developed scalable mechanisms for tracking memory content across a parallel computer. The system he developed, named ConCORD, provides a best-effort global view of memory content, and supports a map-reduce-like framework, the content-aware service command, for writing services on top of it. The content-aware service command combines the best-effort knowledge in the ConCORD distributed hash table with the ground-truth knowledge of the VMMs, to make it possible to build faster, correct services such as distributed checkpointing with deduplication. We demonstrated this service.

Products and Technology Transfer

We now describe the software and papers developed at Northwestern during this project. All papers are publicly available from the v3vee.org web site and/or via their respective authors' sites, or the publishers. All software noted below, with the exception of the NAVAR and ConCORD thesis projects, is publicly available from the v3vee.org web site. Generally speaking, the software is included in the Palacios codebase mainline, although some takes the form of separate patches. Publicly available software is generally BSD licensed, with some components being GPL licensed for Linux compatibility. Papers are listed in order of publication, for publication dates within the overall project period. Tech reports are not included except where they are non-overlapping (dissertations).

Software

- *Palacios Virtual Machine Monitor.* This is the major software component developed in the project, and where most time was spent by Northwestern and the collaborating institutions. Palacios continues to be under active development within the DOE Hobbes project. The performance and power monitoring elements of the project are included in this line item. The Palacios/GEM5 integration patches are also included here.

- *VNET/P.* This is the high performance overlay networking software that is included in the Palacios codebase.

- *GEARS and Guarded Modules.* These are the Palacios components that allow guest manipulation, including for trusted code, as described earlier. They are included in the Palacios codebase.

- *MIME and RTME.* These are the Palacios components that provide general purpose memory reference emulation (MIME) and Intel hardware transactional memory emulation (RTME). Both are included in the Palacios codebase.

- *NAVAR.* This is the thesis system of Chang Bae. It implements run-time inference and adaptation for NUMA virtualization within Palacios, as described earlier.

- *ConCORD.* This is the thesis system of Lei Xia. It implements scalable memory content tracking across a parallel computer, as described earlier.

Papers

- J. Lange, K. Pedretti, P. Dinda, P. Bridges, C. Bae, P. Soltero, A. Merritt, *Minimal Overhead Virtualization of a Large Scale Supercomputer*, Proceedings of the 2011 ACM SIGPLAN/SIGOPS International Conference on Virtual Execution Environments (VEE 2011), March, 2011.

- J. Lange, P. Dinda, *SymCall: Symbiotic Virtualization Through VMM-to-Guest Upcalls*, Proceedings of the 2011 ACM SIGPLAN/SIGOPS International Conference on Virtual Execution Environments (VEE 2011), March, 2011.

- C. Bae, J. Lange, P. Dinda, *Enhancing Virtualized Application Performance through Dynamic Adaptive Paging Mode Selection*, Proceedings of the 8th International Conference on Autonomic Computing (ICAC 2011), June, 2011.

- J. Lange, P. Dinda, K. Hale, L. Xia, *An Introduction to the Palacios Virtual Machine Monitor---Version 1.3*, Technical Report NWU-EECS-11-10, Department of Electrical Engineering and Computer Science, Northwestern University, November, 2011.

- Y. Tang, L. Xia, Z. Cui, J. Lange, P. Dinda, P. Bridges, J. Li, *High Performance Virtual Network Embedding Virtual Machine Monitor*, Chinese Journal of Scientific Instrument, Volume 33, Number 5, pages 1195-1199, May, 2012.

- P. Bridges, D. Arnold, K. Pedretti, M. Suresh, F. Lu, P. Dinda, R. Joseph, J. Lange, *Virtual Machine-based Emulation of Future Generation High-performance Computing Systems*, International Journal of High Performance Computing Applications, Volume 26, Number 2, pages 125-135, May, 2012.

- L. Xia, P. Dinda, *A Case for Tracking and Exploiting Inter-node and Intra-node Memory Content Sharing in Virtualized Large-Scale Parallel Systems*, Proceedings of the 6th International Workshop on Virtualization Technologies in Distributed Computing (VTDC 2012), June, 2012.

- C. Bae, L. Xia, P. Dinda, J. Lange, *Dynamic Adaptive Virtual Core Mapping to Improve Power, Energy, and Performance in Multi-socket Multicores*, Proceedings of the 21st ACM Symposium on High-performance Parallel and Distributed Computing (HPDC 2012), June, 2012.

- L. Xia, Z. Cui, J. Lange, Y. Tang, P. Dinda, P. Bridges, *VNET/P: Bridging the Cloud and High Performance Computing Through Fast Overlay Networking*, Proceedings of the 21st ACM Symposium on High-performance Parallel and Distributed Computing (HPDC 2012), June, 2012. (**Best Paper Nominee**)

- B. Kocoloski, J. Lange, *Better than Native: Using Virtualization to Improve Compute Node Performance*, Proceedings of the 2nd International Workshop on Runtime and Operating Systems for Supercomputers (ROSS 2012), June, 2012.

- K. Hale, L. Xia, P. Dinda, *Shifting GEARS to Enable Guest-context Virtual Services*, Proceedings of the 9th International Conference on Autonomic Computing (ICAC 2012), September, 2012.

- B. Kocoloski, J. Ouyang, and J. Lange, *A Case for Dual Stack Virtualization: Consolidating HPC and Commodity Applications in the Cloud*, Proceedings of the third ACM Symposium on Cloud Computing, (SOCC 2012), October, 2012.

- Z. Cui, L. Xia, P. Bridges, P. Dinda, J. Lange, *Optimizing Overlay-based Virtual Networking Through Optimistic Interrupts and Cut-through Forwarding*, Proceedings of the ACM/IEEE International Conference on High Performance Computing, Networking, Storage and Analysis (SC'12) (Supercomputing), November, 2012.

- Z. Cui, P. Bridges, J. Lange, P. Dinda, *Virtual TCP Offload: Optimizing Ethernet Overlay Performance on Advanced Interconnects*, Proceedings of the 22nd ACM Symposium on High-performance Parallel and Distributed Computing (HPDC 2013), June, 2013.

- R. Brightwell, R. Oldfield, D. Bernholdt, A. Maccabe, E. Brewer, P. Bridges, P. Dinda, J. Dongarra, C. Iancu, M. Lang, J. Lange, D. Lowenthal, F. Mueller, K. Schwan, T. Sterling and P. Teller, *Hobbes: Composition and Virtualization as the Foundations of an Extreme-scale OS/R*, Proceedings of the 3rd International Workshop on Runtime and Operating Systems for Supercomputers (ROSS 2013), June, 2013.

- L. Xia, *ConCORD: Tracking and Exploiting Cross-Node Memory Content Redundancy in Large-Scale Parallel Systems*, Doctoral Dissertation, Northwestern University. Published as Technical Report NWU-EECS-13-05, Department of Electrical Engineering and Computer Science, Northwestern University, July, 2013.

- Z. Cui, *Enhancing HPC on Virtual Systems in Clouds through Optimizing Virtual Overlay Networks*, Doctoral Dissertation, Department of Computer Science, University of New Mexio, July, 2013.

- C. Bae, *Dynamic Adaptive Resource Management in a Virtualized NUMA Multicore System for Optimizing Power, Energy, and Performance*, Doctoral Dissertation, Northwestern University. Published as Technical Report NWU-EECS-13-07, Department of Electrical Engineering and Computer Science, Northwestern University, July, 2013.

- L. Xia, Z. Cui, J. Lange, Y. Tang, P. Dinda, P. Bridges, *Fast VMM-based Overlay Networking For Bridging the Cloud and High Performance Computing*, Cluster Computing, Volume 17, Number 1, pages 39-59, March 2014.

- M. Swiech, K. Hale, P. Dinda, *VMM Emulation of Intel Hardware Transactional Memory*, Proceedings of the 4th International Workshop on Runtime and Operating Systems for Supercomputers (ROSS 2014), June, 2014.

- L. Xia, K. Hale, P. Dinda, *ConCORD: Easily Exploiting Memory Content Redundancy Through the Content-aware Service Command*, Proceedings of the 23rd ACM Symposium on High-performance Parallel and Distributed Computing (HPDC 2014), June, 2014.

- K. Hale, P. Dinda, *Guarded Modules: Adaptively Extending the VMM's Privileges Into the Guest*, Proceedings of the 11th International Conference on Autonomic Computing (ICAC 2014), June, 2014.

Computer Modeling

This project does not involve computer modeling.