Enabling Exascale Hardware and Software Design through Scalable System Virtualization

Patrick G. Bridges, University of New Mexico; Peter Dinda, Northwestern University;
Jack Lange, University of Pittsburgh; Kevin Pedretti, Sandia National Laboratories;
Stephen Scott, Oak Ridge National Laboratory

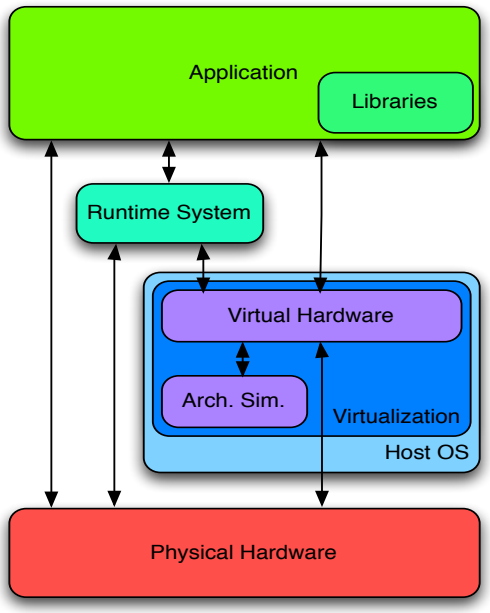
Overview

In this project, we are investigating system software tools to accelerate the development and use of exascale systems. In particular, we are developing new system virtualization techniques to enable the development of the hardware and software innovations needed to enable exascale systems. Virtualization techniques provide traction on a wide range of exascale design and development issues, as described below.

We are using virtualization to enable the design, development, and use of exascale systems. Virtualization allows new hardware and software features to be prototyped as extensions to a virtual machine monitor (VMM), making them immediately available for experimentation and use. Furthermore, it allows new system software and runtime stacks to be launched above production host operating systems without the need for dedicated system time. The VMM is the ideal place for measuring, monitoring, and evaluating these innovations.

Virtualization will also provide support for legacy applications and system software on future exascale systems. This enables flexibility in innovative system designs by ameliorating the barrier that past software investments present to the deployment of new systems.

Hardware and software researchers can also use virtualization to gain leverage on other exascale challenges. For example, virtualization eases system-level checkpointing and migration for resilience and layout optimization purposes. It also provides a level of indirection that can be used to optimize system software behavior even when system software is not easily modifiable due to software complexity or legal issues.



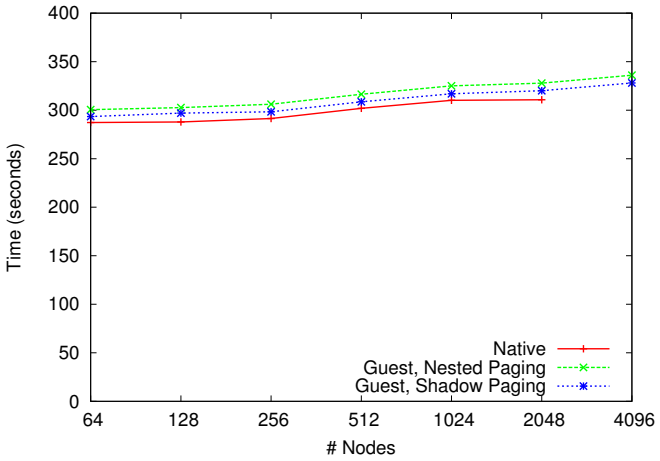
High-Level Approach

This project extends our long-term research that has led to the development of the Palacios HP VMM (Jack Lange 2011). Palacios provides minimal overhead, a self-contained code base, embedability in a wide range of host operating systems, and an open-source license. This makes it ideal both for our research and for other researchers to use to their advantage.

We plan to prototype several new hardware techniques in Palacios and integrate an existing architectural simulator to enable wide-ranging exascale hardware research. We will also add VMM-level measurement and monitoring tools to Palacios so that researchers can fully quantify the impact of their hardware and software changes on the entire system. Finally, we plan to integrate Palacios into production host operating systems (e.g. the Cray Linux

Environment) and to expand Palacios's support for current and future HPC platforms.

Current Status



Palacios currently provides minimal-overhead multi-core virtualization of multi-core guests on both commodity and supercomputer-class x86/64 hardware. In particular, we have demonstrated that Palacios can virtualize 4096 nodes of the ASC Red Storm supercomputer running the CTH shock hydrodynamics code with less than 5% overhead, as shown above. Palacios release 1.2 is currently available from http://www.v3vee.org, and release 1.3, which includes enhanced multi-core support and support for Linux as a host operating system, is planned for release in mid-March 2011.

Example Contributions

In addition to fundamental research to expand the capabilities and scalability of HPC system virtualization, we will demonstrate the ability of virtualization to accelerate exascale research. For example, we plan to use virtualization to prototype non-coherent global addressing and performance-heterogeneous many-core processors on current HPC cluster systems. We also are using virtualization to provide fault injection capabilities to study runtime systems and applications that can recover from or roll forward through hardware failures. System and network virtualization also form the basis for our ongoing work in adaptive parallel and distributed computing.

Collaboration Opportunities

We are strongly interested in collaborating with other researchers, including hardware, system software, runtime system, programming model, compiler, and application researchers. For example, we are interested providing Palacios support for new virtual and physical communication devices, novel memory models for scalability at the node and machine level, and instruction set extensions for high performance computing and communication. We to hope to form collaborations that would enable the large-scale evaluation of such ideas on current HPC systems, as well as drive the development of new virtualization features and system level features. We are also interested in working with vendors and system operators to make Palacios available on production systems to make virtualization capabilities broadly available to the community.

Additional Information

Additional information, including source code releases, development repositories, technical papers, and email discussion lists are available at http://www.v3vee.org.

Research Team

- University of New Mexico: Patrick G. Bridges (Lead PI), Dorian Arnold
- Northwestern University: Peter Dinda (PI) , Russ Joseph, Fabian Bustamante, and Jack Lange (U. Pittsburgh)
- Oak Ridge National Laboratory: Stephen Scott (PI), Geoffroy Vallée, and Thomas Naughton
- Sandia National Laboratories: Kevin Pedretti (PI) and Ron Brightwell

Bibliography

Jack Lange, Kevin Pedretti, Peter Dinda, Patrick G. Bridges, Chang Bae, Philip Soltero, and Alexander Merritt. "Minimal Overhead Virtualization of a Large-scale Supercomputer." *2011 ACM SIGPLAN/SIGOPS International Conference on Virtual Execution Environments (VEE 2011).* Newport Beach, CA, CA, 2011.