Scalable System Virtualization in High Performance Computing Systems

Patrick Bridges\textsuperscript{1}, Peter Dinda\textsuperscript{2}, John Lange\textsuperscript{3}, Kevin Pedretti\textsuperscript{4}, Stephen Scott\textsuperscript{5}, Zheng Cui\textsuperscript{1}, Lei Xia\textsuperscript{2}, Michael Levenhagen\textsuperscript{3}, and Ron Brightwell\textsuperscript{4}
\textsuperscript{1}Scalable Systems Lab, Department of Computer Science, University of New Mexico, Albuquerque, NM, bridges@cs.unm.edu
\textsuperscript{2}Department of Electrical and Computer Engineering, Northwestern University, Evanston, IL, pdinda@northwestern.edu
\textsuperscript{3}Department of Computer Science, University of Pittsburgh, Pittsburgh, PA, jacklange@cs.pitt.edu
\textsuperscript{4}Sandia National Laboratories, Albuquerque, NM, ktpedre@sandia.gov
\textsuperscript{5}Oak Ridge National Laboratory, Oak Ridge, TN, scottsl@ornl.gov

Palacios: Scalable Virtualization for HPC Systems

Virtualization is a power technology with diverse applications in high-performance systems. The adoption of virtualization in HPC systems can only occur, however, if it has minimal performance impact in the most demanding uses of the machines, specifically running the capability applications that motivate the acquisition of petascale and exascale systems. Virtualization cannot succeed in HPC systems unless the performance overheads are truly minimal and, importantly, that those overheads that do exist do not compound as the system and its applications scale up.

The VEE Project (http://v3vee.org), a collaboration between UNM, Northwestern, Pittsburgh, Sandia, and Oak Ridge developed the Palacios Virtual Machine monitor to provide scalable virtualization for HPC systems. Palacios provides minimal-overhead virtualization for HPC systems, with less than 5% application performance overheads when running on large-scale Cray supercomputing systems. Building on these results, we have conducted a wide range of cutting edge HPC research described below.

Exascale Design Space Exploration

One key application of virtualization is to provide a platform for exploring the design of hardware and software for next-generation exascale computing platforms. In particular, virtualization allows existing platforms to emulate future platforms before they are built, allowing system designers to understand the impact of different hardware and software design choices on application performance.

To enable this, we are adding system emulation capabilities to Palacios and integrating external architectural simulators, particularly the Sandia SST simulation toolkit with Palacios. This allows Palacios to quickly emulate modest architectural changes with minimal slowdown, or to call out to external device and processor simulators to evaluate the performance of more radical design changes while still quickly emulating system features that do not require costly simulations.

Using this capability, Palacios provides a high-speed processor model to SST network simulations, improving the speed of network simulations. In particular, Palacios speeds up SST network simulations by more than an order of magnitude on pure networking tests and by a factor of 30 more on application tests.

We are currently working on further enhancing Palacios’s simulation capabilities using time dilation techniques and Palacios and SST’s checkpoint mechanisms. This will allow Palacios and SST to emulate radically different processor architectures, for example many-core and heterogeneous core systems, provide even greater design space exploration flexibility to system hardware and software architects.

Sandia CTH application with virtualized Catamount guest on Palacios up to 4096 Red Storm nodes

Simulation Result Comparison
Gem5 vs. Palacios VM
**VNET: A High-performance Overlay Network for Bridging Cloud and HPC Systems**

Virtual overlay networks allow HPC applications to span both tightly-coupled supercomputing and distributed cluster resources. However, to achieve the application performance that the tightly-coupled resources are capable of, it is important that the overlay network not introduce significant overhead relative to the native hardware. Previous virtual overlay network systems, for example the VNET/U system, were not able to provide the communication performance necessary to allow scalable HPC applications to span both HPC and Cloud systems.

To address this, we developed the VNET/P high-performance virtual overlay network system in Palacios. VNET/P is a layer 2 virtual networking system that has negligible latency and bandwidth overheads in 1–10 Gbps networks. With recent enhancements such as early interrupt injection along with the use of a noise-free host OS such as Kitten, VNET/P can achieve passthrough-level application and microbenchmark performance on 10 Gbps and Infiniband networks, while still maintaining the full flexibility of a virtual overlay network, as shown in the figure.

**Acknowledgements**

The development of Palacios has been funded by grant from the U.S. Department of Energy Office of Science, Advanced Scientific Computing Research, under award number DE-SC0005050, program manager Sonia Sachs, and from grant CNS-0709168 and CNS-0707365 from the National Science Foundation.

**Publications**


