Supporting High Performance Molecular Dynamics in Virtualized Clusters using IOMMU, SR-IOV, and GPUDirect

Andrew J. Younge
School of Informatics & Computing
Indiana University
Bloomington, IN 47408
Email: ajyounge@indiana.edu

John Paul Walters
Information Sciences Institute
University of Southern California
Arlington, VA 22203
Email: jwalters@isi.edu

Geoffrey C. Fox
School of Informatics & Computing
Indiana University
Bloomington, IN 47408
Email: gcf@indiana.edu

Abstract—Cloud infrastructure-as-a-Service paradigms have recently shown their utility for a vast array of computational problems, however many scientific computing applications have been slow to adapt to virtualized cloud frameworks. This is due to performance impacts of virtualization technologies, coupled with the lack of advanced hardware support necessary for running many scientific appications at scale. By using KVM virtual machines that can leverage both Nvidia GPUs and InfiniBand, we show that molecular dynamics simulations with LAMMPs and HOOMD run at near-native speeds. This experiment also illustrates how cloud infrastructure can support the latest parallel computing paradigms, including both MPI+CUDA and new GPUDirect RDMA functionality.

While these experiments do go beyond a single-node, their early implementation is limited to only 4 nodes due to the lack of feasible resources. Currently efforts are under way to scale the deployment to hundreds of cores and 32 GPUs within the FutureGrid testbed, which we look to demonstrate at Supercomputing 2014 in November.

I. INTRODUCTION

At present we stand at the inevitable intersection between High Performance Computing (HPC) and clouds. Various platform tools such as Hadoop and MapReduce, among others, have already percolated into data intensive computing within HPC [1]. In addition, there are efforts to support traditional HPC-centric scientific computing applications in virtualized cloud infrastructure. There are a multitude of reasons for supporting parallel computation in the cloud [2], including features such as dynamic scalability, specialized operating environments, simple management interfaces, fault tolerance, and enhanced quality of service, to name a few. The growing importance of supporting advanced scientific computing using cloud infrastructure can be seen by a variety of new efforts, including the NSF-funded XSEDE Comet resource at SDSC [3].

Nevertheless, there exists a past notion that virtualization used in today's cloud infrastructure is inherently inefficient. Historically, cloud infrastructure has also done little to provide the necessary advanced hardware capabilities that have become almost mandatory in supercomputers today, most notably advanced GPUs and high-speed, low-latency interconnects. The result of these notions has hindered the use of virtualized environments for parallel computation, where performance must be paramount.

A growing effort is currently underway that looks to systematically identify and reduce any overhead in virtualization technologies, so far with relative success [?], [4]. Thus, we see a consitantly diminishing overhead with virtualization, not only with traditional overheads [] but also with HPC workloads. While virtualization will almost always include some additionaly overhead in relation to its dynamic abilities, the eventual goal for supporting HPC in virtualized environments is to minimize any overhead wheneve rpossible.

To advance the placement of HPC applications on virtual machines, new efforts are emerging foscusing specificaly on key HPC hardware. By leveraging new virtualization tools such as IOMMU device passthrough and SR-IOV, we can now support the same common HPC hardware such as the latest Nvidia Tesla GPUs [?] as well as InfiniBand fabric [?].

Recent advances in hypervisor performance [4] coupled with the newfound availably of HPC hardware in virtual machines analogous to the most powerful supercomputers used today, we see can see the formation of a high performance cloud infrastructure. While our previous advances in this area have focused on single-node advancements, it is now imperative to ensure real-world applications can also operate at scale.

To start, we demonstrate running two molecular dynamics simulations, LAMMPS and HOOMD, in a virtual infrastructure complete with both Kepler GPUs and QDR Infini-Band. Both LHOOMD and LAMMPS are used extensively in some of the world's fastest supercomputers and represent a key simulation example that HPC supports today. We show that these applications are able to run at near-native speeds within a completely virtualized environment, demonstrating small performance impacts usually acceptable by many users. Furthermore, we illustrate the ability of such a virtualized environment to support cuttinge edge software tools such as RDMA GPUDirect, demonstrating cutting-edge technologies

are indeed possible in a virtualized environment.

Following these efforts, we hope to ensure upstream infrastructure projects such as OpenStack [**?**] are able to make effective and quick use of these features, allowing users to build private cloud infrastructure to support high performance distributed computational workloads.

II. RELATED WORK

NOTE: first introduce virtualization, and I/O featuresets. Then enable GPUs. Then bring in SR-IOv and infiniBand. Finally, discuss applications and GPUDirect.

Virtualized and cloud-based HPC has been extensively studied with mixed results [4]. However, hypervisor advances, and recent hardware support for virtualization, suggest that additional study is needed.

Nvidia GPUs comprise the single most common accelerator in the June 2014 Top 500 List. In previous work, we have shown that GPUs achieve up to 99% of their bare metal performance when passed to a virtual machine using PCI passthrough [5]. This work demonstrated PCI passthrough performance across a range of hypervisors and GPUs, but was limited to single node performance. Similarly, Infiniband SR-IOV (Single Root I/O Virtualization) has been evaluated within the context of microbenchmarks [6], [7], but performance for real applications is not yet well-understood.

Recent work recent work has focused on single-node performance. In [**?**], we've shown how the latest Kepler GPUs from Nvidia with Sandy-Bridge Intel Xeon CPUs can perform at near-native performance running various workloads across wide range of hypervisors. Furthermore, advanced configuration of SR-IOV enabled Infiniband fabric has taken shape, with recent research showing up to a 30% reduction latency [**?**].

Using two molecular dynamics tools, LAMMPS [8] and HOOMD [9], we demonstrate a high performance *system*. That is, we combine PCI passthrough for Nvidia Kepler-class GPUs with QDR Infiniband SR-IOV and show that high performance molecular dynamics simulations are achievable within a virtualized environment. For the first time, we also demonstrate Nvidia GPUDirect technology within such a virtual environment. Thus, we look to not only illustrate that virtual machines provide a flexible high performance infrastructure for scaling scientific workloads including MD simulations, but also that the latest HPC features and programming environments are also available in this same model.

III. EXPERIMENTAL SETUP

Our test environment is composed of 4 servers each with a single Nvidia Kepler-class GPU. Two servers are equipped with K20 GPUs, while the other two servers are equipped with K40 GPUs. Each server is composed of 2 Intel Xeon E5-2670 CPUs, 48GB of DDR3 memory, and Mellanox ConnectX-3 QDR Infiniband. CPU sockets and memory are split evenly between two NUMA nodes. For these experiments, both the GPUs and Infiniband adapters are attached to NUMA node 1 and both the guest VMs and the base system utilized identical software stacks.

CentOS 6.4 with a 3.13 upstream Linux kernel was used as the host OS with the KVM hypervisor. The native bare-metal base system and all guest VMs are composed of a CentOS 6.4 installation with a 2.6.32-358.23.2 stock kernel, MVAPICH 2.0 GDR, and CUDA version 5.5. Each guest was allocated 20 GB of RAM and a full socket (8 cores) as well as a single InfiniBand virtual function and 1 Kepler GPU per VM.

In order to effectively test MD simulations in LAMMPS and HOOMD beyond single-node tests, ronment. Bespin includes 4 blades, each with 2 Intel Xeron E5-2670 CPUs, 48Gb DDR3 memory, Mellanox ConnectX3 QDR InfiniBand cards, and a mixture of Nvidia Kepler series K20 and K40 GPUs. While the effective experimental hardware allocation remains relatively low compared to production runs of either application, it does allow for a useful evaluation at a larger scale than preivously evaluated as well as a valued extrapolation to larger resources.
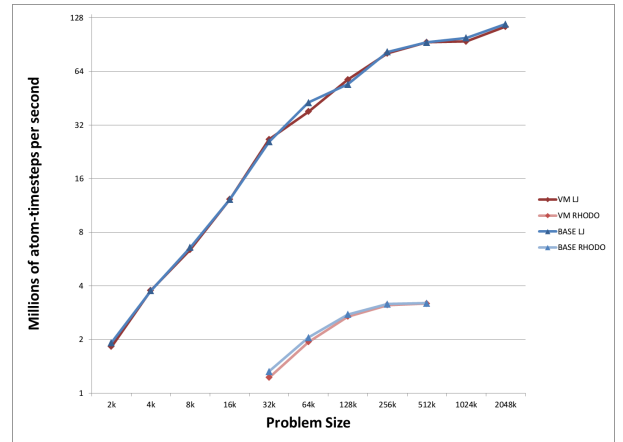
IV. RESULTS

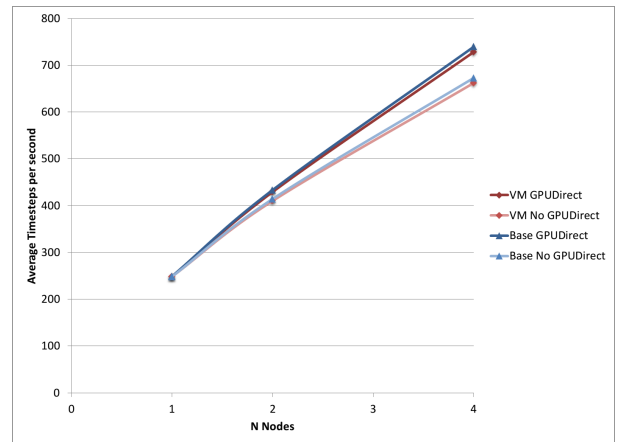Fig. 1.   LAMMPS RHODO & LJ Performance

Fig. 2.   HOOMD LJ Performance with 256k Simulation

Figure 1 shows two of the most common LAMMPS algorithms used; the Lennard-Jones potential (LJ) and the Rhodopsin protein in solvated lipid bilayer benchmark

(Rhodo), both running with the GPU package across 8 cores per GPU. Here we see that both benchmarks scale remarkably well in the virtualized KVM guest environment. Compared to the base system performance, the VMs running LAMMPs acheive 96.7% and 99.3% efficiency for LJ and Rhodo, respectively when running across all nodes. These low overheads illustrate the utility of running LAMPS on cloud infrastructure, and also hold promise for other hybrid MPI + CUDA applications to also scale well in a virtualized environment.

In Figure 2 we show the performance of a Lennard-Jones liquid simulation with 256K particles running under HOOMD. HOOMD includes support for CUDA-aware MPI implementations via GPUDirect. The MVAPICH 2.0 GDR implementation enables a further optimization by supporting RDMA for GPUDirect. From Figure 2 we can see that HOOMD simulations, both with and without GPUDirect, perform very near-native. The GPUDirect results at 4 nodes achieve 98.5% of the base system's performance. The non-GPUDirect results achieve 98.4% efficiency at 4 nodes. Further, GPUDirect RDMA shows a clear advantage over the non-GPUDirect implementation, achieving a 9% performance boost in both the native a virtualized experiments.

V. Conclusion

With the advent of cloud infrastructure, the ability to run large-scale parallel scientific applications has become possible but limited due to both performance and hardware availability issues. In this work we show that advanced HPC-oriented hardware such as the latest Nvidia GPUs and InfiniBand fabric are now available within a virtualized infrastructure. Our results find MPI + CUDA applications run at near-native performance compared to traditional non-virtualized HPC infrastructure, with just an averaged 1.9% and 1.5% overhead for LAMMPs and HOOMD, respectively. Moving forward, we show the utility of GPUDirect RDMA for the first time in a cloud environment with HOOMD. Effectively, we look to pave the way for large-scale virtualized cloud Infrastructure to support a wide array of advanced scientific computation commonly found running on many supercomputers today.

References

[1] S. Jha, J. Qiu, A. Luckow, P. K. Mantha, and G. C. Fox, "A tale of two data-intensive paradigms: Applications, abstractions, and architectures," in *Proceedings of the 3rd International Congress on Big Data*, 2014.

[2] M. Armbrust, A. Fox, R. Griffith, A. D. Joseph, R. Katz, A. Konwinski, G. Lee, D. Patterson, A. Rabkin, I. Stoica, and M. Zaharia, "A view of cloud computing," *Commun. ACM*, vol. 53, no. 4, pp. 50–58, Apr. 2010.

[3] M. Norman and R. Moore, "NSF awards 12 million to sdsc to deploy comet supercomputer," Web page, 2013.

[4] A. J. Younge, R. Henschel, J. T. Brown, G. von Laszewski, J. Qiu, and G. C. Fox, "Analysis of Virtualization Technologies for High Performance Computing Environments," in *Proceedings of the 4th International Conference on Cloud Computing (CLOUD 2011)*. Washington, DC: IEEE, 2011.

[5] J. P. Walters, A. J. Younge, D.-I. Kang, K.-T. Yao, M. Kang, S. P. Crago, and G. C. Fox, "GPU-Passthrough Performance: A Comparison of KVM, Xen, VMWare ESXi, and LXC for CUDA and OpenCL Applications," in *Proceedings of the 7th IEEE International Conference on Cloud Computing (CLOUD 2014)*. Anchorage, AK: IEEE, 2014.

[6] J. Jose, M. Li, X. Lu, K. C. Kandalla, M. D. Arnold, and D. K. Panda, "SR-IOV support for virtualization on infiniband clusters: Early experience," in *Cluster Computing and the Grid, IEEE International Symposium on*, 2013, pp. 385–392.

[7] M. Musleh, V. Pai, J. P. Walters, A. J. Younge, and S. P. Crago, "Bridging the Virtualization Performance Gap for HPC using SR-IOV for InfiniBand," in *Proceedings of the 7th IEEE International Conference on Cloud Computing (CLOUD 2014)*. Anchorage, AK: IEEE, 2014.

[8] S. Plimpton, P. Crozier, and A. Thompson, "Lammps-large-scale atomic/molecular massively parallel simulator," *Sandia National Laboratories*, 2007.

[9] J. Anderson, A. Keys, C. Phillips, T. Dac Nguyen, and S. Glotzer, "Hoomd-blue, general-purpose many-body dynamics on the gpu," in *APS Meeting Abstracts*, vol. 1, 2010, p. 18008.