Towards a High Performance Virtualized IaaS Deployment

Andrew J. Younge[1,2], John Paul Walters[1], Jinwoo Suh[1], Dong-In D. Kang[1],
Youngsam Park[1], Stephen P. Crago[1], Geoffrey C. Fox[2]
[1] Information Sciences Institute, University of Southern California
3811 North Fairfax Drive, Suite 200, Arlington, VA 22203 U.S.A.
[2]Pervasive Technology Institute, Indiana University
2729 E 10th St., Bloomington, IN 47408, U.S.A.
Email corresponding author: ayounge@isi.edu

***Abstract*—Scientific computing endeavors have created clusters, grids, and supercomputers as high performance computing (HPC) platforms and paradigms. These resources focus on peak performance and computing efficiency, thereby enabling scientific community to tackle non-trivial problems on massively parallel architectures. Meanwhile, efforts to leverage the economies of scale from data center operations and advances in virtualization technologies have created large scale Clouds. Such Infrastructure-as-a-Service (IaaS) deployments provide mechanisms for handling millions of user interactions concurrently or organizing, cataloging, and retrieving mountains of data in a way that allows users to specify a custom computing environment tailored to their needs. Combining concepts from both supercomputing and clouds will enable users to leverage the performance of HPC applications with the ease and availably in clouds, creating an ideal environment for many scientific computing applications. This work proposes building a heterogeneous, high performance IaaS using the OpenStack software concentrating on virtualization performance, heterogeneous hardware and GPUs, high speed interconnects, advanced scheduling and high performance storage to create a novel IaaS supercomputing system.**

I. Introduction

For many decades, the HPC community has led the march for the best possible computational performance, striving for absolute speed made which is possible through the application of Moore's Law. This pursuit can be seen through the focus on cutting edge architectures, high-speed, low-latency interconnects, parallel languages, and dynamic libraries, all tuned to maximize computational efficiency. Performance has been the keystone of HPC since its conception, with many ways to evaluate performance. The Top500 supercomputer ranking system, which has been running for over 20 years, is a hallmark to performance's importance in supercomputing. Many other benchmarks targeting the HPC realm exist to provide additional information about supercomputing importance. This includes HPCC and SPEC benchmarking suites, as well as the new Graph500 list focusing on data intensive performance [1], [2]. Using these benchmarks to compare traditional supercomputers to IaaS deployments represent the key mechanism to evaluating the success of a HPC cloud.

The goal of this work is not to identify the differences between supercomputers and clouds, but rather to illustrate the need for a cohesive architecture that blends the two distinct paradigms together. We propose the creation of a heterogeneous, high-performance IaaS Cloud architecture leveraging the OpenStack project [3]. If such a system is successful, it will lower the barrier of entry for scientists and enable a new wave of scientific computation that is otherwise missed. Furthermore, such an architectural model will improve efficiency within data centers, increasing the potential for fine-grained computation, leading to reduced resource expenditures and/or increased resource utilization.

II. Strategy

To bring a heterogeneous, high performance cloud cyberinfrastructure to fruition, a few focal areas are needed to ensure the success of such an endeavour. Specifically, we outline 5 focal points; hypervisor performance, heterogeneous hardware, high speed interconnect support, advanced scheduling mechanisms, and distributed storage; all wrapped in custom IaaS cloud architecture based on OpenStack.

A. Virtualization Performance

One of the biggest concerns surrounding the use of hypervisors in production supercomputing resources is the performance overhead. While overhead was considerable in the past, recent hardware and hypervisor advancements have brought the overhead to near-native performance for many cases [4]. Furthermore, new methods for providing virtualized environments have proven to provide near-native performance through chroot environments, particularly using LXC.

B. Heterogeneous Hardware & GPUs

While the supercomputing industry moves from petascale to exascale, systems are continually leveraging accelerator cards, typically general purpose graphical processing units (GPUs). These GPUs are able to provide 1-2 orders of magnitude increase in processing power compared to CPUs while only adding a linear increase in power usage. As such, the performance-to-watt ratio for GPUs are very high compared to CPUs, a factor that's increasingly important as we move to multi-megawatt supercomputers and exascale.

Historically, providing such add-on accelerators has not been possible in a virtualized environment until recently. Current work utilizes PCI passthrough functionality in hypervisors to provide such hardware directly to a VM. Through this, GPUs become available in Openstack, a cloud IaaS project. Currently nVidia Tesla GPUs are supported through LXC [5], with a more pronounced Xen integration is forthcoming. This will enable supercomputing-class hardware in the more desirable cloud environment.

Often, researchers have special hardware requirements that are particular to their project. Many researchers also require direct access to hardware without a hypervisor, which is typically not possible in a Cloud environment. Through the use of OpenStack, we will provide both Tilera and x86 resources through the use of bare-metal operating system provisioning in a trusted environment. Bare metal provisioning is performed through a unique implementation of nova-compute which provisions a network-booted image directly to the hardware of the individual resources. This creates illusion of a virtualized environment with the ability of native runtime with no potential overhead.

C. Advanced Interconnects

One of the key distinctions between current supercomputers and clouds is the lack of a tightly coupled interconnect network within clouds. These interconnects, represented typically as proprietary networks or InfiniBand, offer high bandwidth ranging from 10 to 80 Gbs with latencies that are many times less than typical Gigabit Ethernet based solutions. Furthermore, many of today's supercomputers are built using 10GbE networks, and the performance of such networks within VMs is often missed. Today's supercomputing applications leverage both the low latency and high bandwidth interconnects to build large scale distributed applications capable of advanced communication with minimal overhead.

Recently two technologies have become available to enable virtualzied environments to leverage the same interconnects as many supercomputers; hardware-assisted I/O virtualization (VT-d or IOMMU) and Single Root I/O Virtualization (SR-IOV). With VT-d, we can pass PCI-based hardware directly to a guest Virtual Machine, thereby removing the overhead of communicating with the host OS through emulated drivers. Instead, we can give a VM direct access to InfiniBand card and use drivers without emulation or modification to achieve near-native I/O performance for the first time. Second, with the use of SR-IOV, we can create multiple virtual functions (VFs), which can be assigned directly to a given VM. This enables IaaS providers to leverage InfiniBand interconnects for applications that leverage MPI, RDMA, or even IP while also providing a guaranteed QoS based on VF configuration.

D. Scheduling

In a traditional cloud computing service such as Amazon's EC2, the user has a limited range of options when requesting virtual machine instances such as amount of memory, number of cores, and amount of local storage. To support a wider range of heterogeneity for high performance computing on OpenStack, a user must be able to request an instance with a configuration of computing resources that meets the user's needs in a simple way. The user can choose it through the user interface by using predefined instance types just the same way as simple instance types. The magic behind the scene is that the instance types are extended using a database table that can be populated with any keywords to include more requirements such as types and number of accelerators. The requirements are relayed to a scheduler that chooses the computing resources that meet all request. As implemented in our software stack, if *cg1.small* instance type maps to a configuration that includes one virtual CPU and one GPU, then by using *cg1.small*, a user can choose the predefined configuration.

We are extending the scheduler further to meet high performance computing. One of them is proximity scheduler that chooses a set of hosts that are close each other such that the provisioned set of hosts can provide low latency and high throughput among the virtual machines on them.

E. High Performance Distributed Storage

There are an increasing number of scientific problems that require a data-centric model for enabling science in the newfound data deluge of the 21st century. This will require integration of large scale, high performance storage systems that can span large numbers of resources in a way that integrates naturally with computation. Leveraging distributed high performance storage systems such as Lustre or GPFS within a virtualized environment can potentially enable a new class of scientific applications. Furthremore, providing such services in a Cloud IaaS deployment would ease the learning curve for researchers and increase utilization drastically. For instance, a mechanism to attach VMs to a Lustre storage service as if it was a cloud block storage device would enable a wide range of new HPC applications on the cloud, and work is under way to make this happen. Furthermore, we have an opportunity to improve the performance of live migration within IaaS deployments using a Lustre distributed file system to host VM images directly, leading to a wide range of advancements in cloud deployments, from efficient resource scheduling to advanced application resiliency.

III. Conclusion

There exists an opportunity for the best of both worlds in supercomputing and clouds. Through the use of the OpenStack IaaS framework, we provide high performance supercomputing resources for advanced parallel computation that the greater scientific computing community is already familiar with, while simultaneously enabling a whole new class of services that will usher in a new wave of data centric science.

References

[1] P. Luszczek, D. Bailey, J. Dongarra, J. Kepner, R. Lucas, R. Rabenseifner, and D. Takahashi, "The hpc challenge (hpcc) benchmark suite," in *Proceedings of the 2006 ACM/IEEE conference on Supercomputing*. Citeseer, 2006, pp. 11–17.

[2] M. Anderson, "Better benchmarking for supercomputers," *Spectrum, IEEE*, vol. 48, no. 1, pp. 12–14, 2011.

[3] K. Pepple, "Revisiting openstack architecture: Essex edition," Webpage, Feb 2012. [Online]. Available: http://ken.pepple.info/openstack/2012/02/21/revisit-openstack-architecture-diablo/

[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, July 2011.

[5] S. Crago, K. Dunn, P. Eads, L. Hochstein, D.-I. Kang, M. Kang, D. Modium, K. Singh, J. Suh, and J. Walters, "Heterogeneous cloud computing," in *Cluster Computing (CLUSTER), 2011 IEEE International Conference on*, sept. 2011, pp. 378 –385.