HPC in Virtualized Platforms – Current Status

Dr. R. Chandramouli (Mouli)
mouli@nist.gov

(Information Technology Lab, NIST, USA)

High Performance Computing Security Workshop

March 27-28, 2018

NIST Gaithersburg MD, USA

HPC in Virtualized Platforms – Current Status

- **Virtualized Platforms – State of Deployment**

- **Features in virtualized platforms enabling HPC artifacts**
  - Leveraging PCI Passthrough for HPC (Slide 5)
  - Leveraging Single Root I/O Virtualization (SR-IOV) for HPC (Slide 7)
  - Leveraging GPUDirect RDMA for Virtualized Clusters (Slide 13)

- **HPC Artifacts for Virtualized Infrastructures – Summary**

Virtualized Platforms – State of Deployment

- Widespread deployment in large enterprise data centers as well as those used for offering Infrastructure-as-a-Service cloud services

- Limited use of virtualized platforms for high performance scientific computing applications. Reasons are:
  - Overhead due to Hypervisor functions – e.g., Context switching between Host Mode and Guest Mode, Device Emulation function etc.
  - Lack of support for necessary hardware in virtualized hosts or clusters

Features in virtualized platforms enabling HPC artifacts -1

- **Passthrough Approach for Device Virtualization**: In this approach, PCI and PCIe devices are directly assigned to a VM. (Hence also called PCI Passthrough)
- These devices have direct memory access (DMA) capability
- To prevent subversion of memory isolation by these DMA-capable devices, virtualized platforms have hardware support in the form of I/O Memory Management Unit (IOMMU) for validating and translating all device access to host memory

Features in virtualized platforms favoring HPC artifacts -2

- **Leveraging PCI Passthrough Approach for HPC** :
- All leading hypervisor offerings have the capability to leverage the IOMMU feature in their hosted platform.
- It is well-known that the modern GPU using PCI interface is a highly data-parallel processor, optimized to provide very high floating point arithmetic throughput for scientific problems with a single program multiple data model.
- Hence by using these GPUs using PCI passthrough, virtualized platforms supporting IOMMU can run VMs with significant computational power.

Features in virtualized platforms enabling HPC artifacts -3

- **Single Root I/O Virtualization (SR-IOV)**:
- A special kind of PCI device is one that has self-virtualizing capability.
- These devices have interfaces that can export a set of virtual functions (VFs) corresponding to a physical function (PF).
- The hypervisor then can assign these VFs to multiple guest VMs, while it retains control of the PF.
- The virtualization and multiplexing capabilities of these hardware devices stem by their conformance to a specification called Single Root I/O Virtualization (SR-IOV)

Features in virtualized platforms enabling HPC artifacts -4

- **Leveraging Single Root I/O Virtualization (SR-IOV) for HPC**
- The virtualization and multiplexing capabilities in SR-IOV enabled hardware provides higher performance and greater control than software solutions.
- One example of SR-IOV implementation is a SR-IOV enabled 10Gb Ethernet adapters being used in some virtualized platforms that provides high performance 10Gb TCP/IP connectivity within VMs.

Communication Requirements of Parallel HPC Applications – A digression

- In almost all parallel HPC applications, the interconnect fabric that enables fast and efficient communication between processors is a central requirement for achieving good performance [3].
- Two Requirements for Interconnect fabric:
  - High Bandwidth – for distributed processors to share large amounts of data across the system
  - Low Latency – ensuring quick delivery of small message communications.

Features in virtualized platforms favoring HPC artifacts -5

- **GPU-GPU Communication Requirements for virtualized cluster**:
- Merely using GPUs in individual virtualized hosts and having a virtualized cluster with hosts interconnected using the high speed Infiniband network is insufficient to guarantee the communication requirements of some scientific applications
- The way GPUs use the network to transfer data between them (i.e., interaction between the GPU and the Infiniband network) has a bearing on the QoS parameters for communication.

Features in virtualized platforms favoring HPC artifacts -6

- **Earlier Implementations of GPU-GPU Communication** :
- GPUs use pinned memory in the host memory (to write their messages) to enhance DMA performance by eliminating the need for intermediate buffers
- Due to the lack of coordinated mechanisms, the message passing libraries (MPI) of Infiniband could not directly uses the contents of the pinned memory used by GPUs to perform RDMA (Remote Direct Memory Access) - technique by which a message can be directly placed in a remote node's memory thereby avoiding intermediate copies.

Features in virtualized platforms favoring HPC artifacts -7

- **A new model for GPU-GPU Communication - GPUDirect**:
- An improved model would involve the development of a mechanism for performing DMA operations of GPUs and RDMA operations of Infiniband adapters directly between GPUs and bypass the host entirely.
- Such an interface can potentially allow RDMAs from one GPU device directly to another GPU on a remote host.

Features in virtualized platforms favoring HPC artifacts -8

- **A new model for GPU-GPU Communication – GPUDirect …contd**:
- An intermediate solution under this model can use the host memory for the data transaction (between GPU and Infiniband network card) but eliminates host CPU involvement by having the GPUs and Infiniband adapters share the same pinned memory.
- This new hardware/software mechanism is called GPUDirect and enables higher GPU-based cluster efficiency – by eliminating the need for CPU involvement in the data transfer.

Features in virtualized platforms favoring HPC artifacts -9

- **Leveraging GPUDirect for Virtualized Clusters**
- Using the latest GPUDirect RDMA feature, an Infiniband adapter can directly access GPU bypassing host memory altogether
- The use of GPUDirect RDMA enables creating a high performance virtualized cluster with each virtualized host consisting of many GPUs, Infiniband network adapters with GPUDirect RDMA support performing RDMA operations to place messages directly in the GPU memory of the remote node.

HPC Artifacts for Virtualized Infrastructures - Summary

- GPU Passthrough - enables hosting of high performance VMs
- SR-IOV – enables high speed communication adapters for multiple VMs and
- GPUDirect RDMA - enables high performance communication for virtualized clusters
- Empirical results by running several HPC applications – finite element computation, several types of simulations on virtualized platforms/clusters using above features have yielded encouraging results.

References

Ramaswamy Chandramouli, "*Comprehensive Security Assurance Measures for Virtualized Server Environments*", a chapter in a book titled "*Information Security and Privacy: Status and Prospects*" – Springer Verlag, Berlin, Germany (to be published in June 2018).

Ramaswamy Chandramouli, "*Security Recommendations for Hypervisor Deployment on Servers*", NIST Special Publication SP 800-125A, January 2018.

Andrew Young, John Walters, Stephen Crago, and Geoffrey Fox, "*Supporting High Performance Molecular Dynamics in Virtualized Clusters using IOMMU, SR-IOV, and GPUDirect*", VEE '15, March 2015, Istanbul, Turkey.

.

Contact Details & Questions

- Contact Details:

  Dr. Ramaswamy Chandramouli

  Computer Security Division – Information Technology Lab

  National Institute of Standards and Technology

  (301) -975-5013 – mouli@nist.gov

- Questions (?):