Secure I/O Device Sharing among Virtual Machines on Multiple Hosts

Cheng-Chun Tu, Chao-tang Lee, and Tzi-cker Chiueh

ISCA'13 Tel-Aviv, Israel

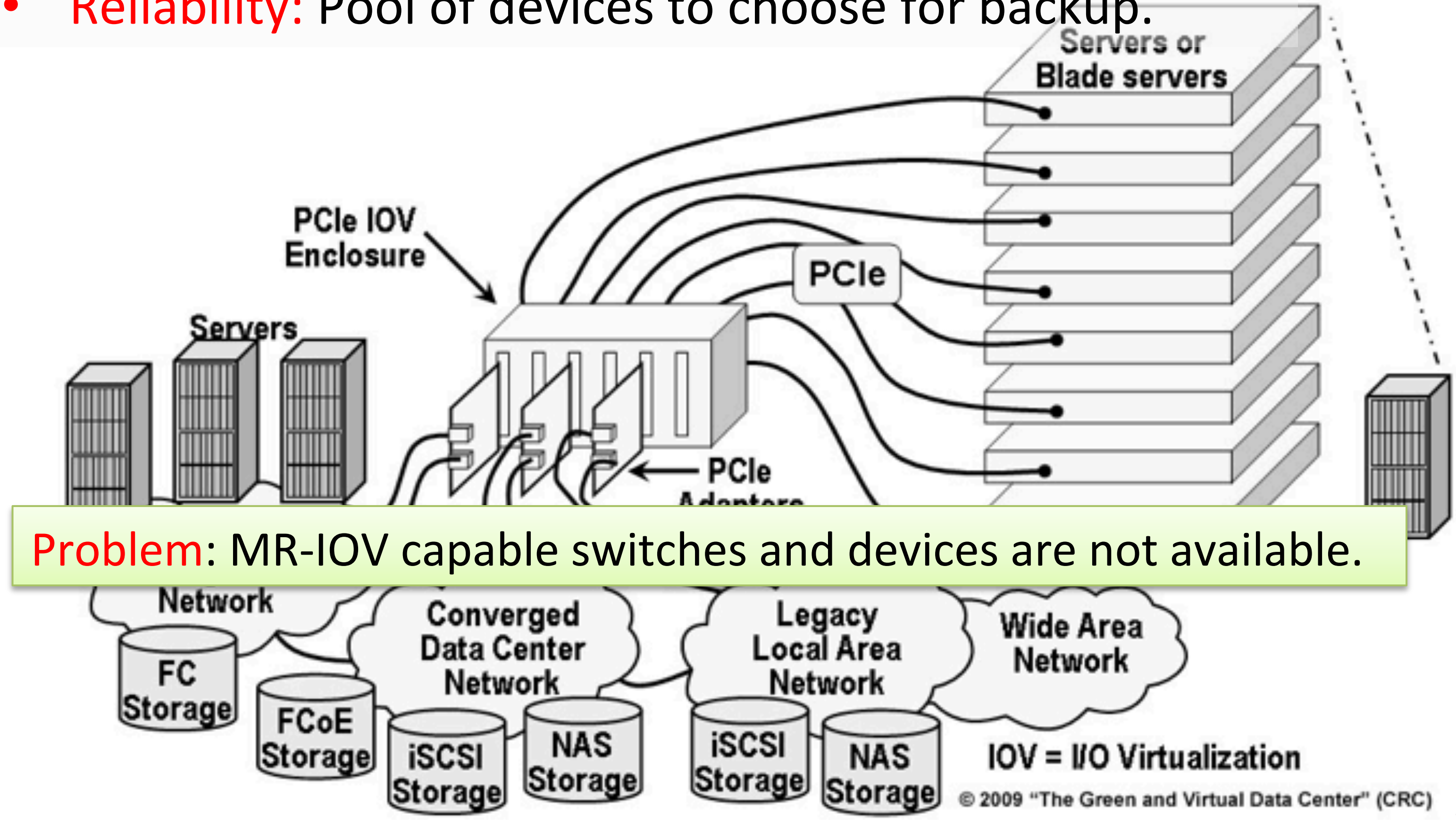
June 25, 2013

工業技術研究院 Industrial Technology Research Institute

STONY BROOK UNIVERSITY

I/O Device Disaggregation

- Reduce Cost: 1 device instead of each one per server.
- Increase Utilization: Accessible from every server.
- Flexible and Scalable: Easy to add/remove
- Reliability: Pool of devices to choose for backup.

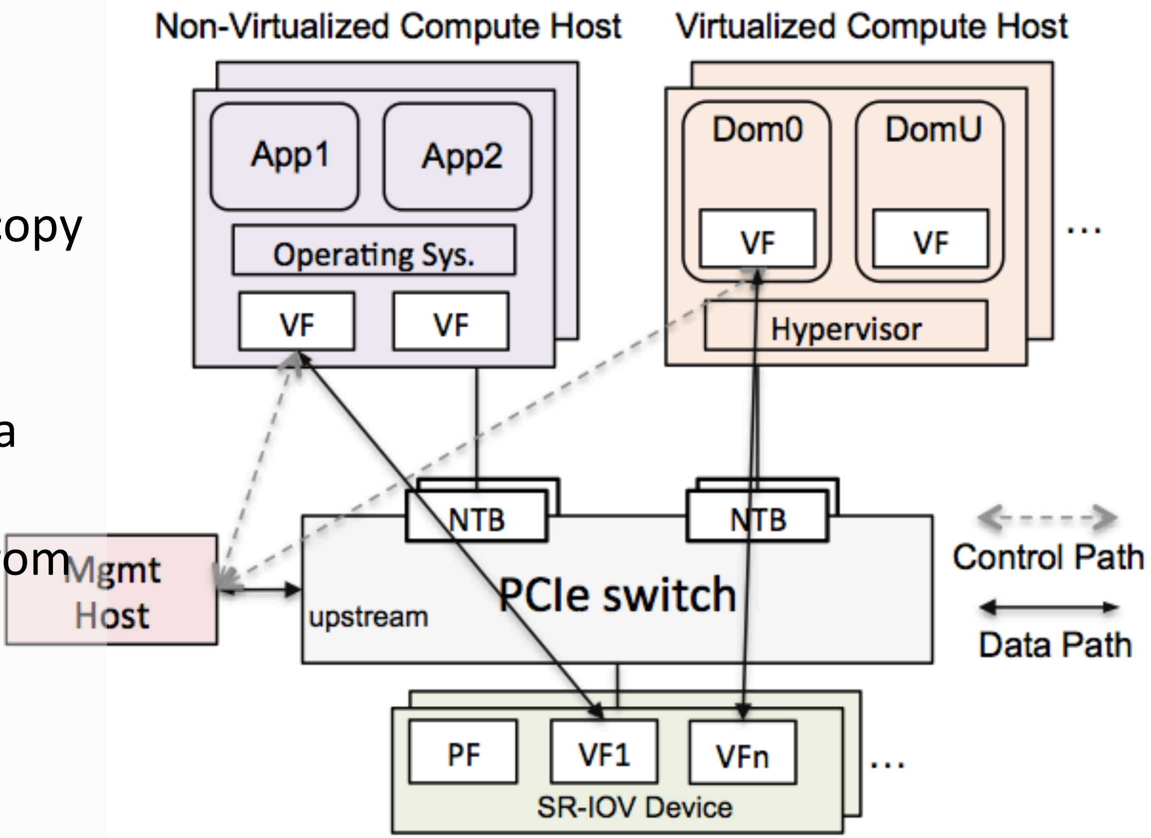


© 2009 "The Green and Virtual Data Center" (CRC)

Problem: MR-IOV capable switches and devices are not available.

Ladon: Software-based MR-SRIOV

- Standard PCIe switch / SR-IOV endpoint:
  - Without MR-IOV capability
  - Native driver / zero copy
- Native Performance:
  - Control path goes to a central authority
  - Direct data path to/from devices
- Secure Sharing:
  - Access control using existing technologies



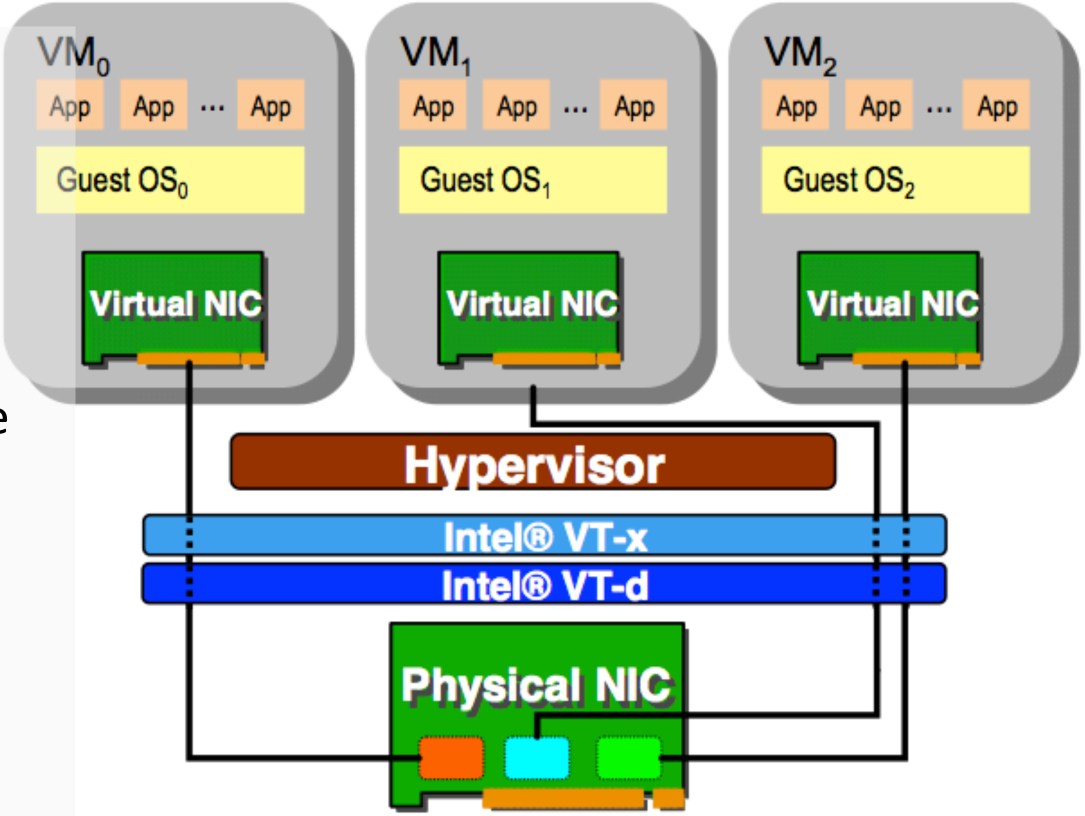
Outline

- Introduction
  - I/O Virtualization (IOV)
  - SR-IOV v.s MR-IOV
  - NTB (Non-Transparent Bridge)
- Ladon
  - Architecture
  - Secure Mechanism
- Performance
  - Throughput / Latency

STONY BROOK UNIVERSITY

I/O Virtualization (IOV)

- Direct communication:
  - Direct assigned to VMs
  - Hypervisor bypassing
- Physical Function (PF):
  - Configure and manage the SR-IOV functionality
- Virtual Functions (VFs):
  - Lightweight PCIe function
  - With resources necessary for data movement

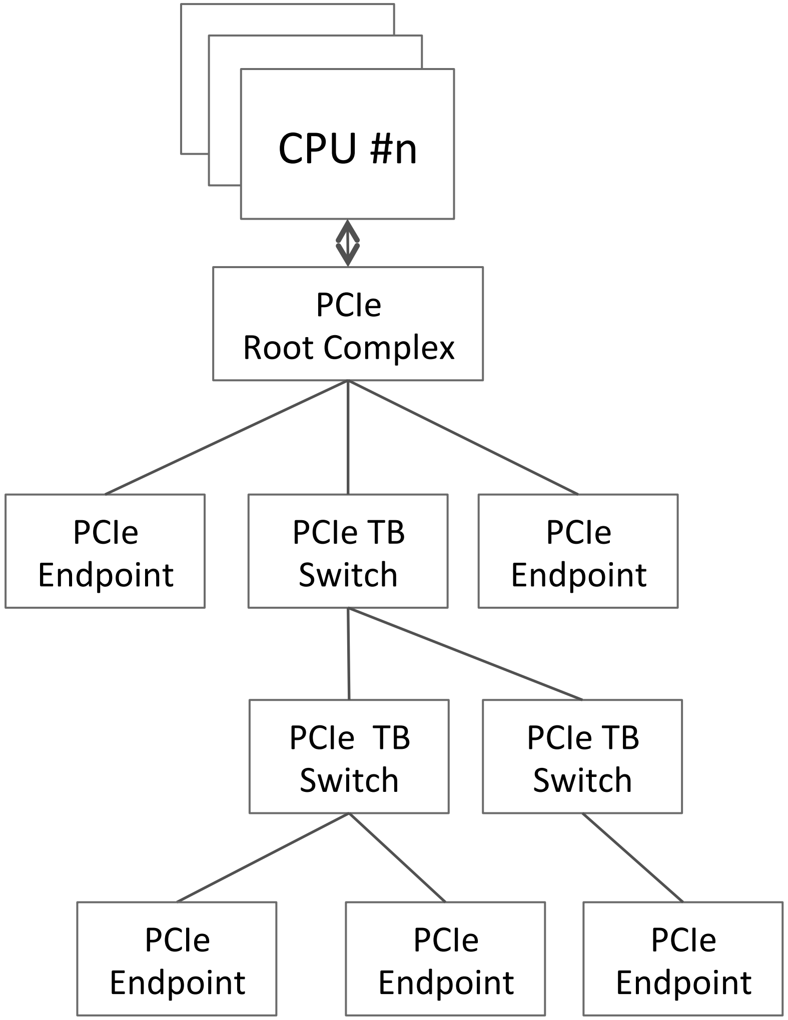


Applicable inside a single host (SR -> Single Root)

Figure: Intel® 82599 SR-IOV Driver Companion Guide

Single Root (SR) Architecture

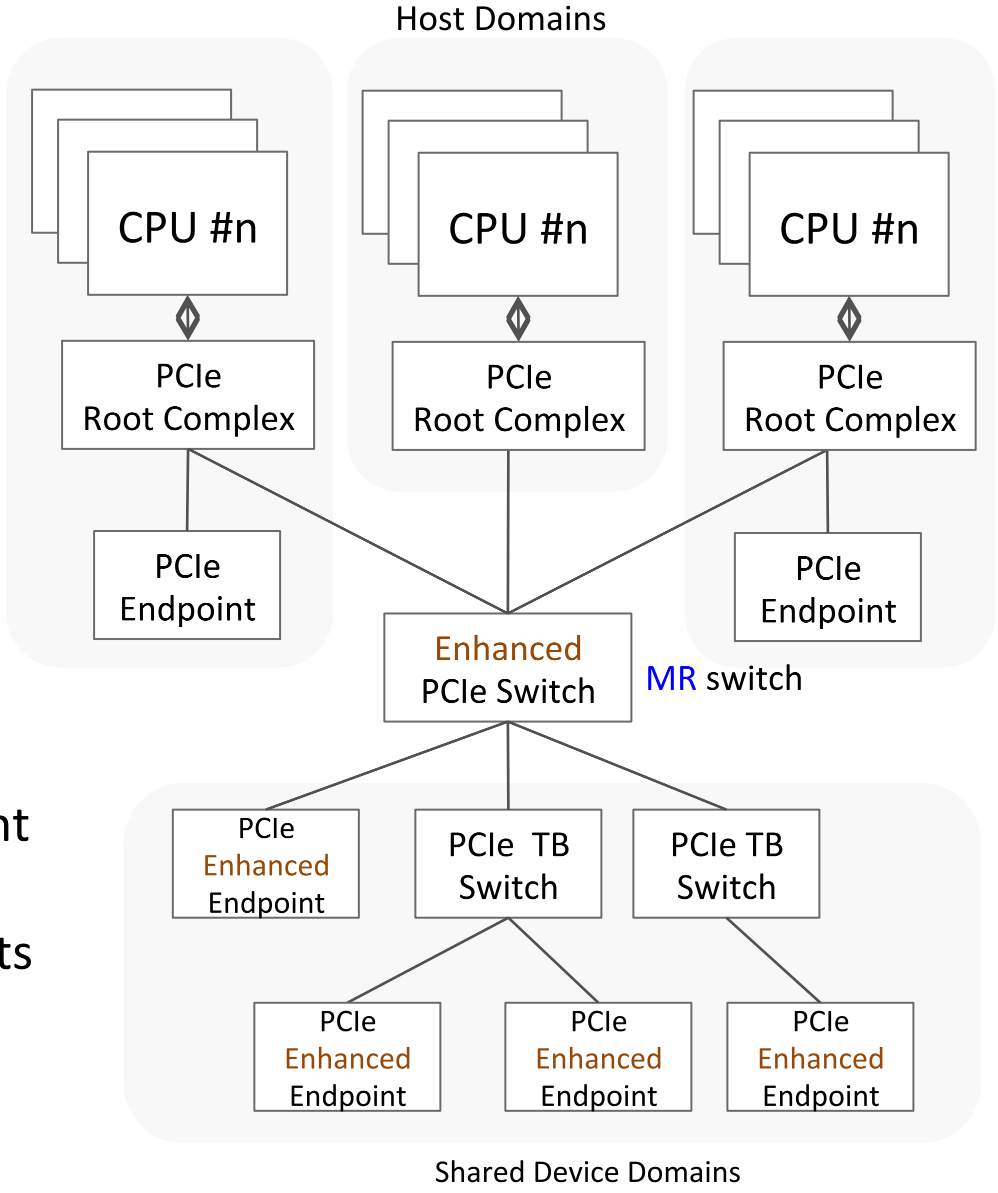
- Multi-CPU, one root complex hierarchies
  - Today's motherboard architecture
  - Single PCIe hierarchy
- Single Address Domain
  - BIOS/System software probes topology
  - Partition and allocate resources
- Each device owns a range (s)of physical address
  - BAR addresses, MSI-X, and device ID



TB: Transparent Bridge

Multi Root (MR) Architecture

- Interconnect multiple hosts
  - Devices shared in the cluster
- Enhanced switch/ endpoints
  - New switch silicon
  - New endpoint silicon
  - Management model
- Solution
  - NTB as isolation element between domains
  - Extend SR-IOV endpoints to multiple hosts



Non-Transparent Bridge (NTB)

- Isolation of two hosts or memory domains
  - Host stops PCI enumeration at NTB-D.
  - Yet allow status and data exchange

- Translations between domains
  - PCI device ID:

    Querying the ID lookup table (LUT)
  - Address:

    From primary side and secondary side

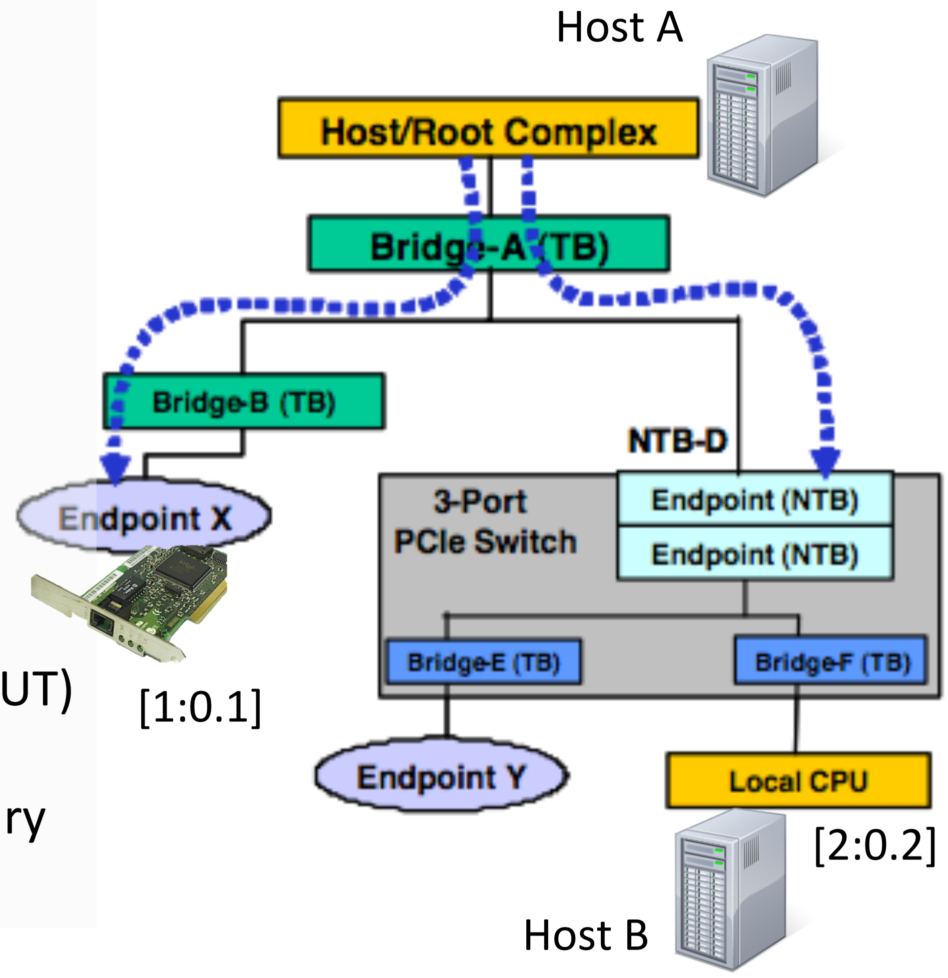


Figure: Multi-Host System and Intelligent I/O Design with PCI Express

NTB Address Mapping

- NTB address translation:
  - <the primary side to the secondary side>
- Configuration:
  - Addr0 at primary side's BAR window to Addr1 at the secondary side
- Example:
  - Addr0 = 0x8000 at BAR4 from HostA
  - Addr1 = 0x10000 at HostB's DRAM
- One-way Translation:
  - Read/write at Addr0 (0x8000) == read/write Addr1
  - Read/write at Addr0 does not translated to Addr1

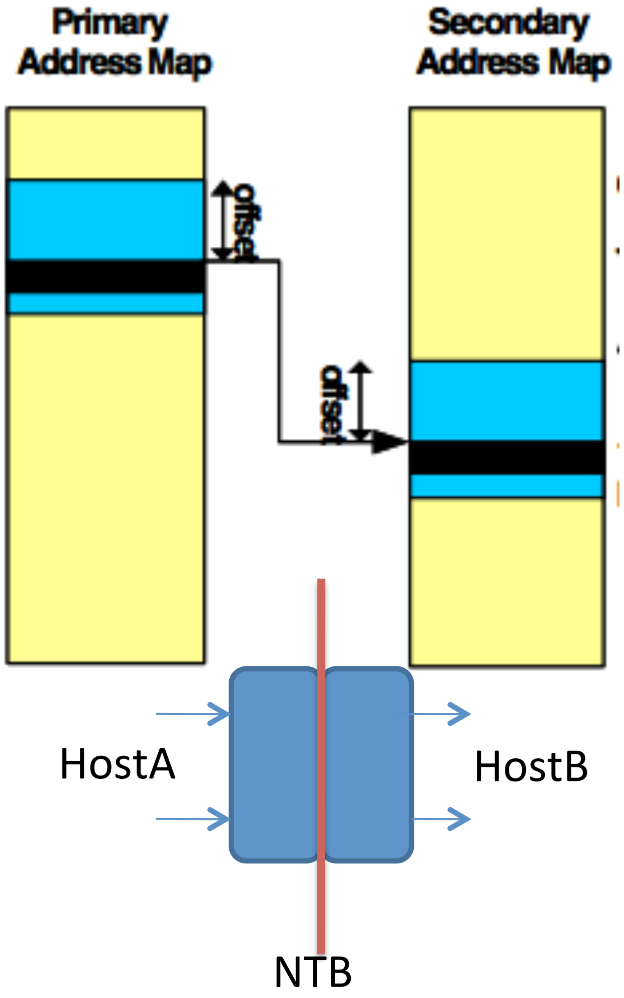


Figure: Multi-Host System and Intelligent I/O Design with PCI Express

A software-based MR-SRIOV

ARCHITECTURE OF LADON

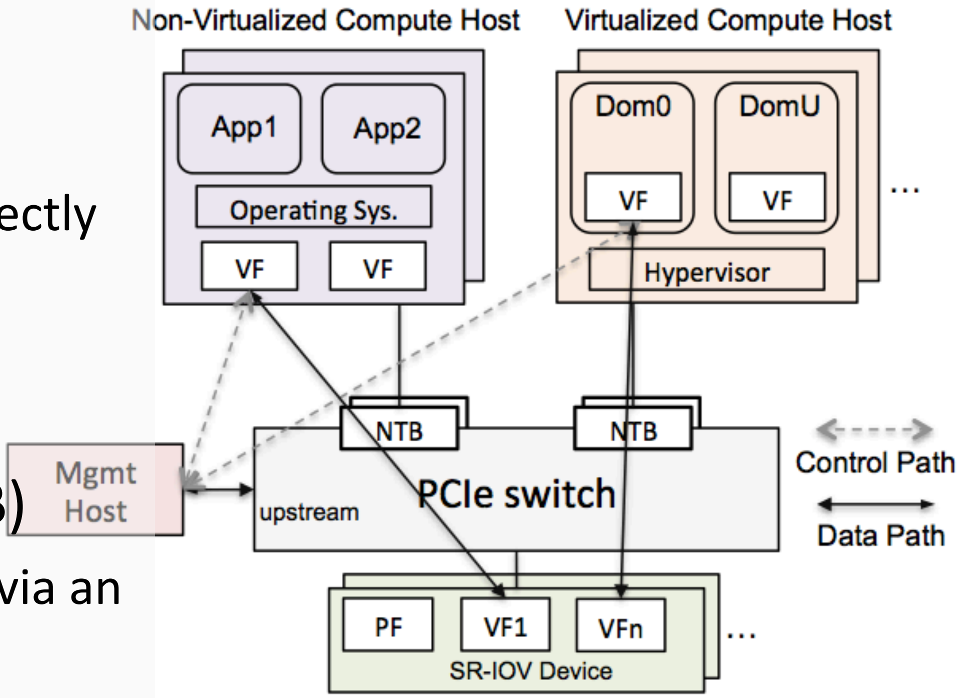
Ladon: a hundred-heads dragon
who guards the Garden of Hesperides

STONY BROOK UNIVERSITY

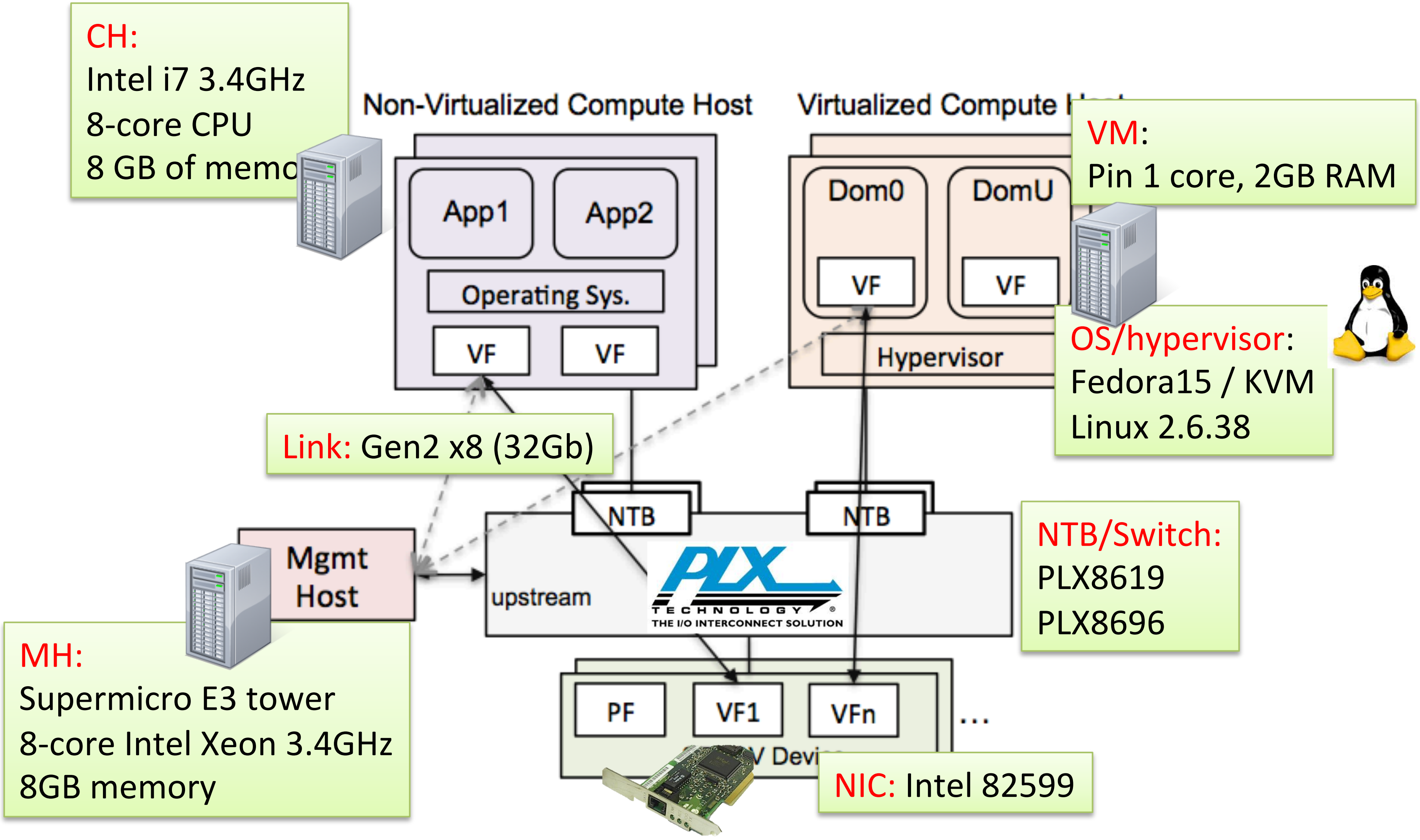
工業技術研究院
Industrial Technology Research Institute

System Components

- Management Host (MH)
  - Manage the shared I/O devices
- Compute Host (CH)
  - Non-virtualized host OS can directly access VF
  - Virtualized host with VMs can directly access VF
- Non-Transparent Bridge (NTB)
  - Each CH connects to the fabric via an NTB
- PCIe Switch & SR-IOV device
  - PF and multiple VFs



Prototype Implementation



CH:
Intel i7 3.4GHz
8-core CPU
8 GB of memo

VM:
Pin 1 core, 2GB RAM

OS/hypervisor:
Fedora15 / KVM
Linux 2.6.38

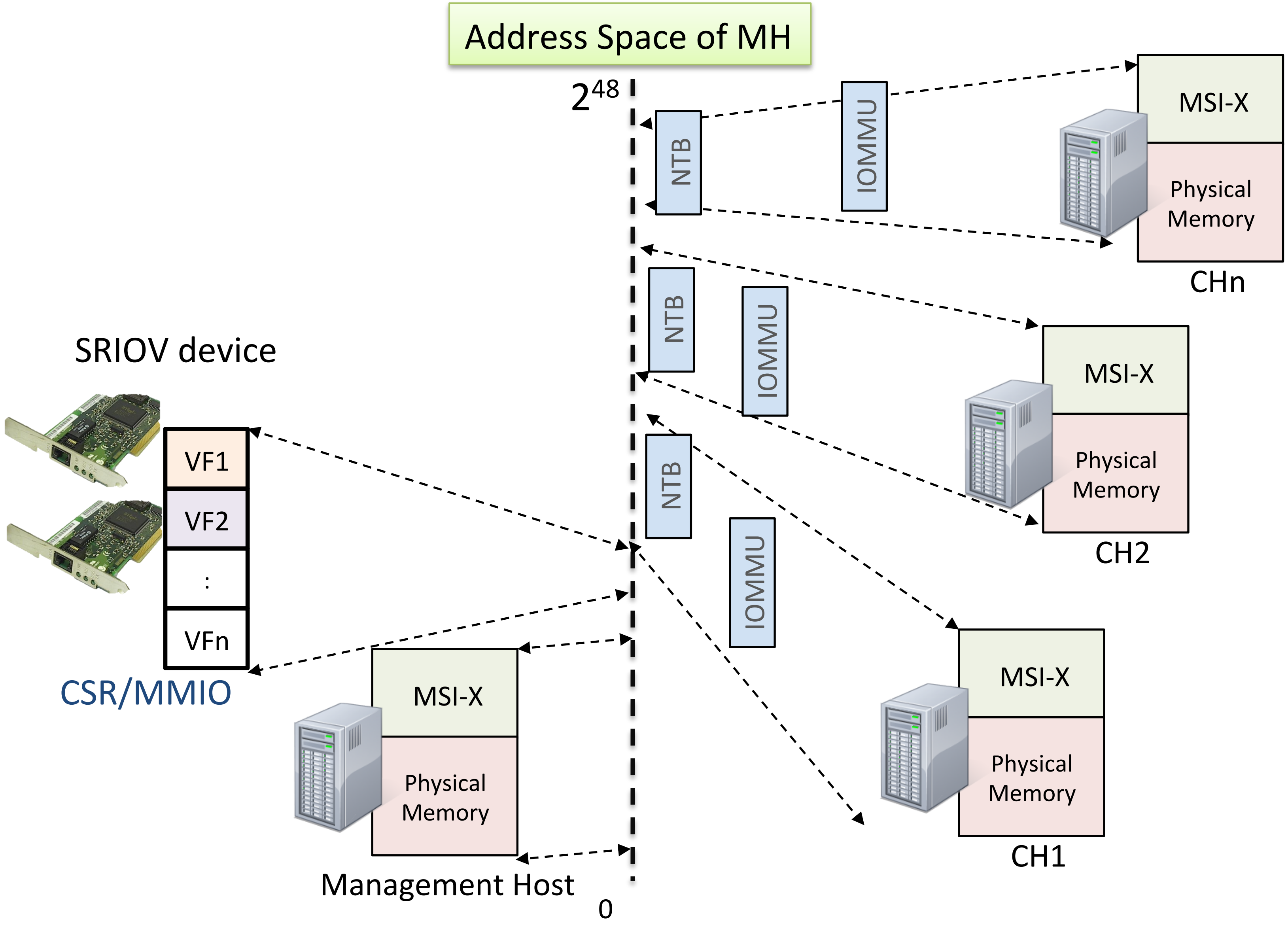
Link: Gen2 x8 (32Gb)

NTB/Switch:
PLX8619
PLX8696

MH:
Supermicro E3 tower
8-core Intel Xeon 3.4GHz
8GB memory

NIC: Intel 82599

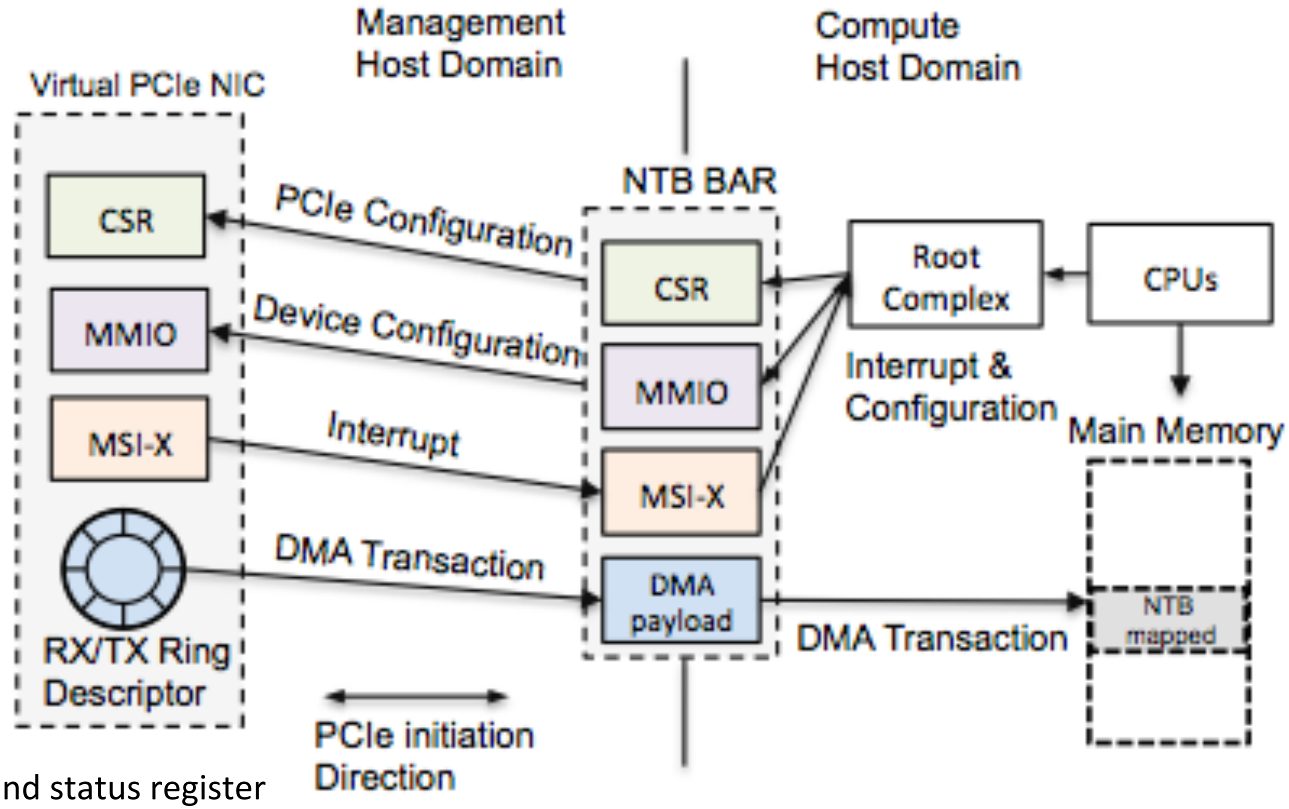
Global Address Space Allocation



Address Space of MH

Per-Virtual NIC Configuration

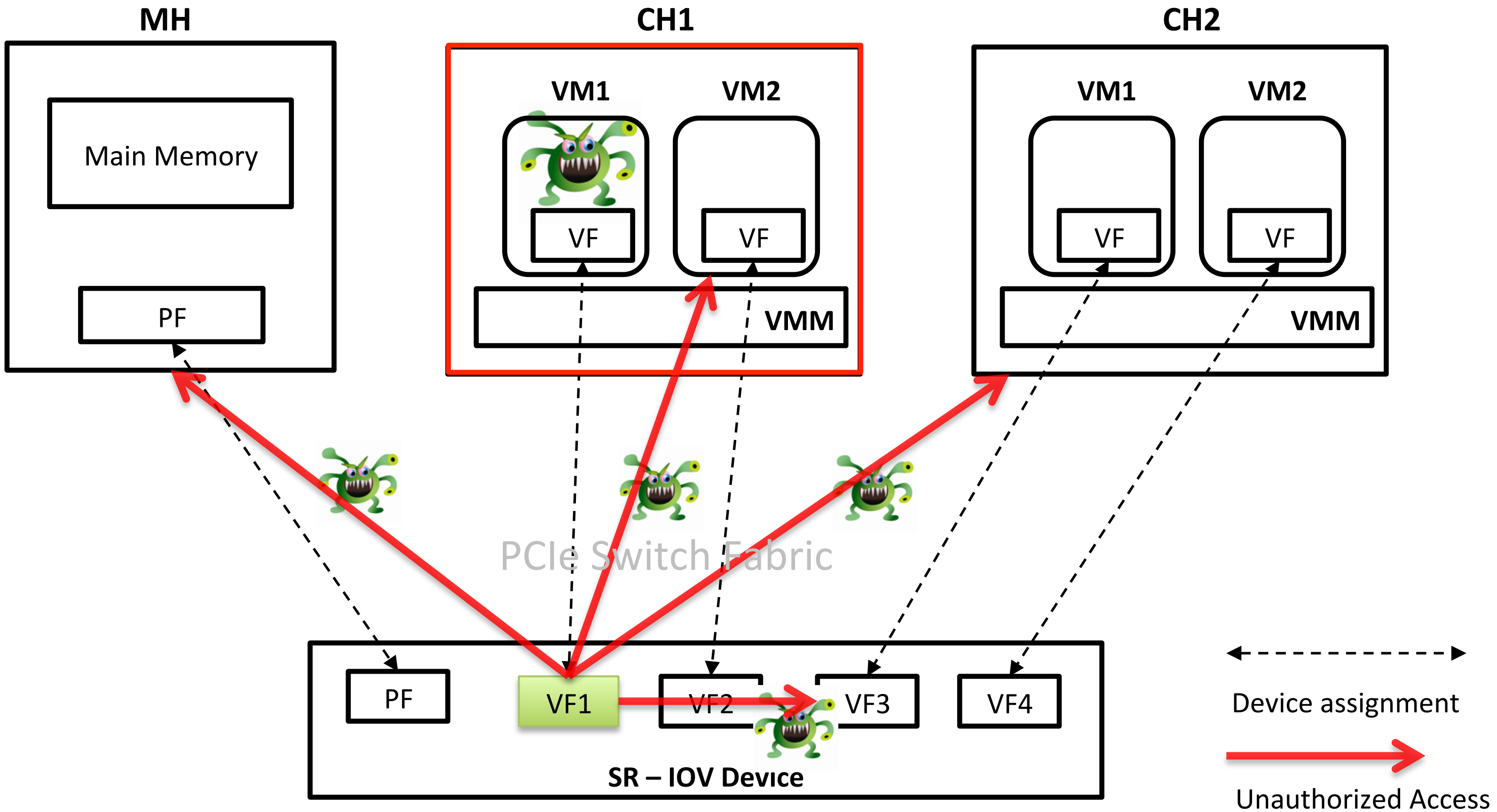
- Virtual NIC is backed by a VF of an SRIOV NIC
- Identify the virtual NIC's CSR, MMIO, MSI-X and DMA payload area
- Install mappings in the BARs of the NTB port



CSR: control and status register

Security Threats

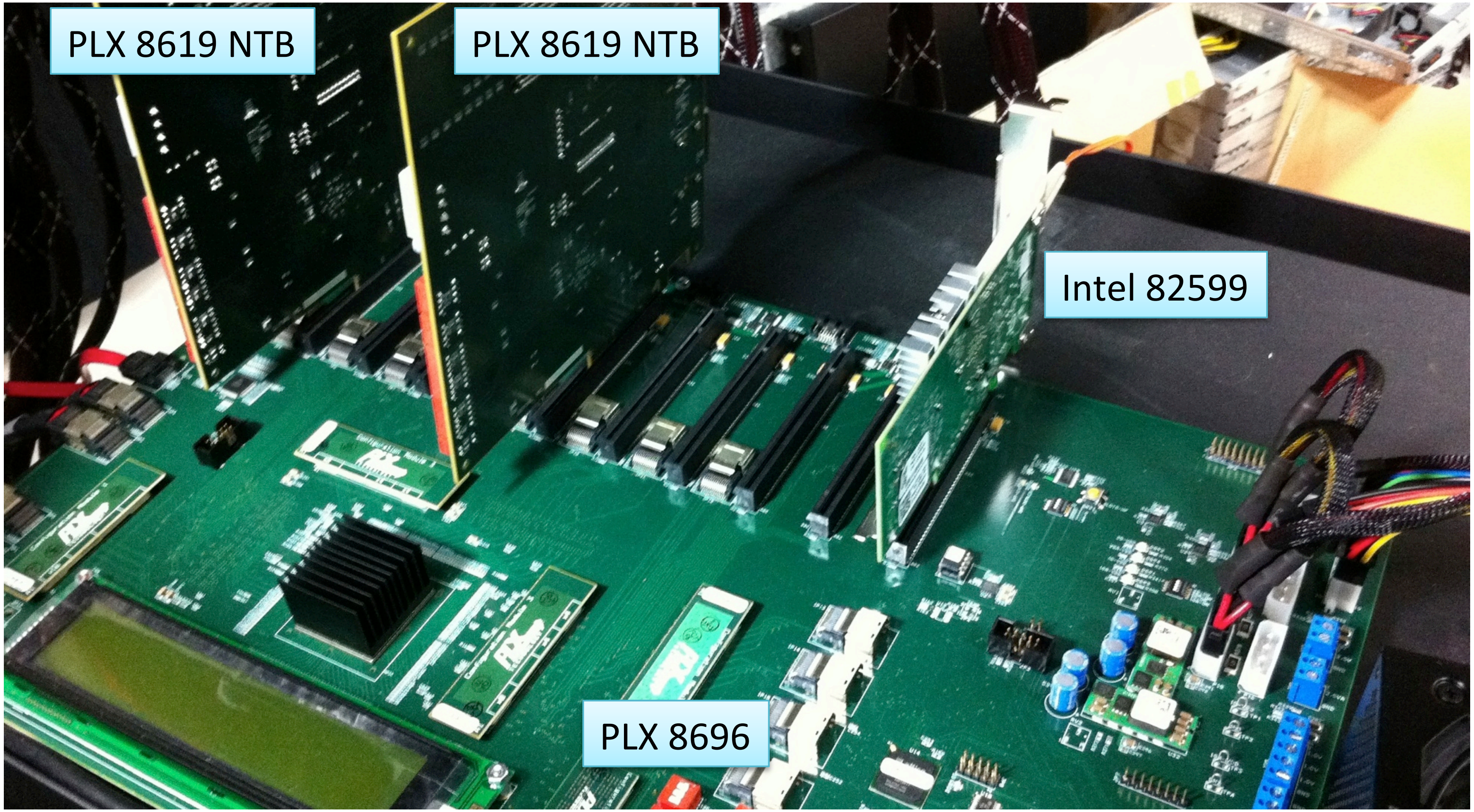
Consider malicious VM and malicious CH



VF1 is assigned to VM1 in CH1, but it can screw multiple memory areas.

Security Guarantees: Summary

- Inter-host
  - A VF can only access the CH it belongs to.
  - Accessing other hosts is blocked by other host’s LUT & IOMMU
- Intra-host
  - A VF assigned to a VM can only access to memory assigned to the VM.
  - Accessing other VMs is blocked host’s IOMMU
- Inter-VF / inter-device
  - A VF can not write to other VF’s registers (MMIO).
  - Isolate by IOMMU in MH
- Compromised CH
  - Not allow to touch other CH’s memory nor MH



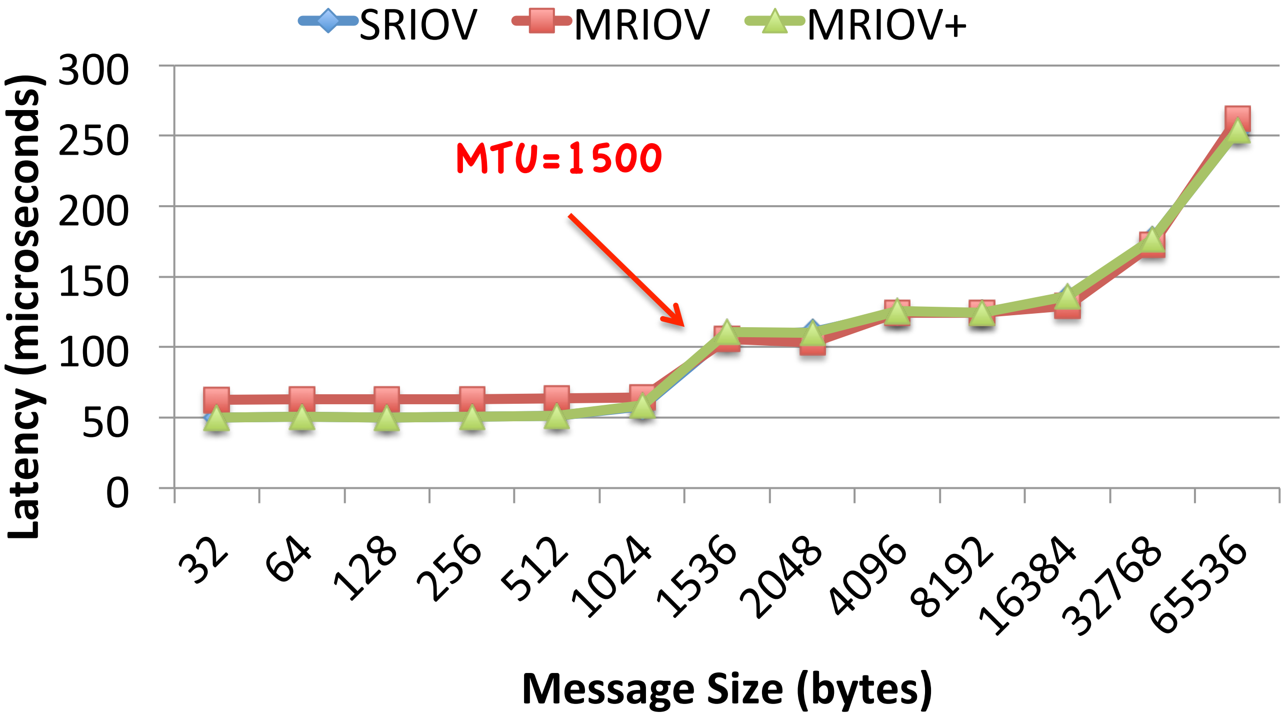
Let’s see some performance numbers!

EVALUATION

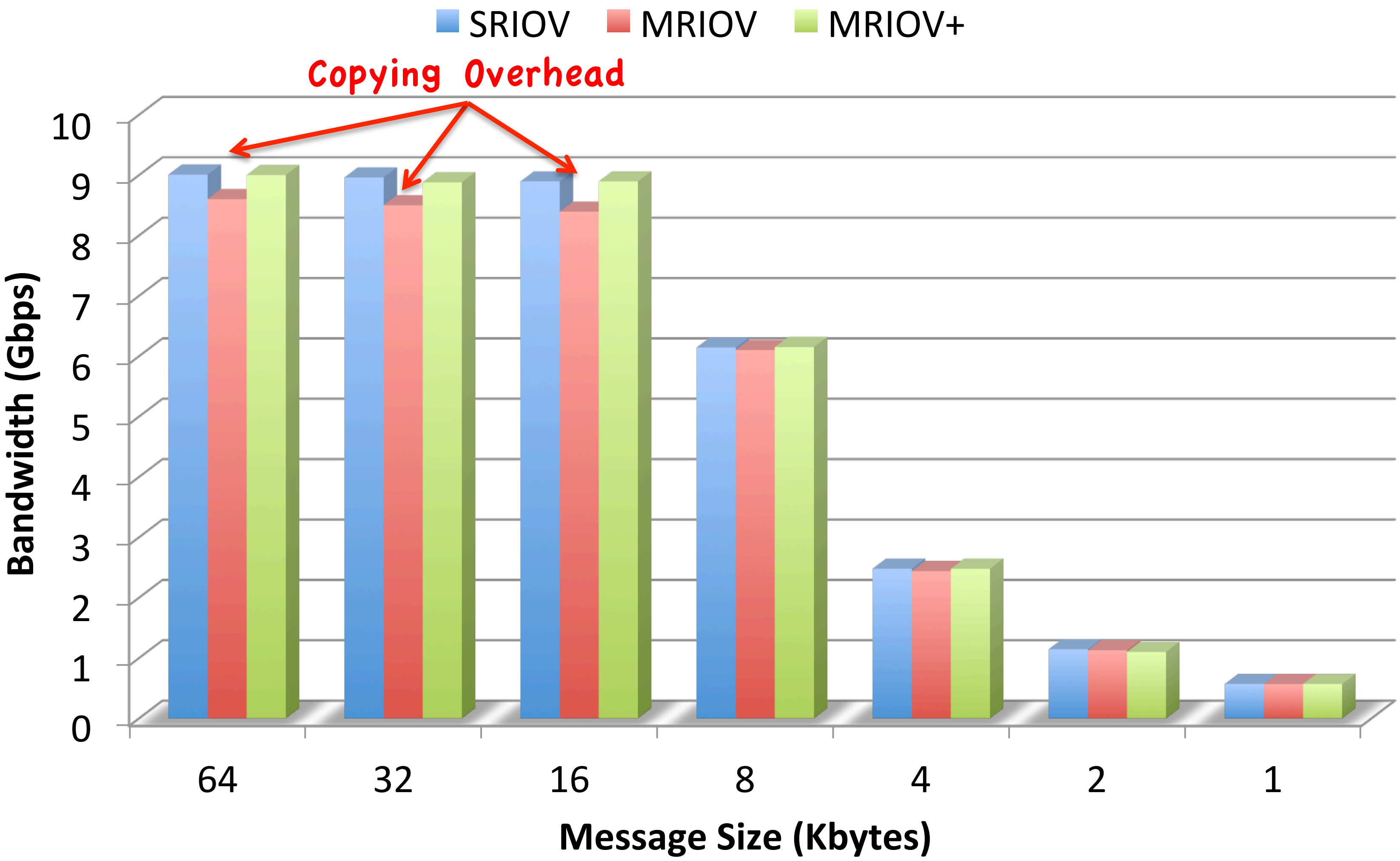
Video Available at : http://youtu.be/B_-GesOjkG0

Latency

- SRIOV: between VM in MH and remote test host
- MRIOV: between VM in CH and remote test host
- MRIOV+: between VM in CH and remote test host
  with zero-copy optimization



TCP RX Throughput



Conclusion

- A scalable PCIe-based system interconnect architecture
  - Decouple devices from CPU/memory in the motherboard
  - Open up a new design space for compute cluster

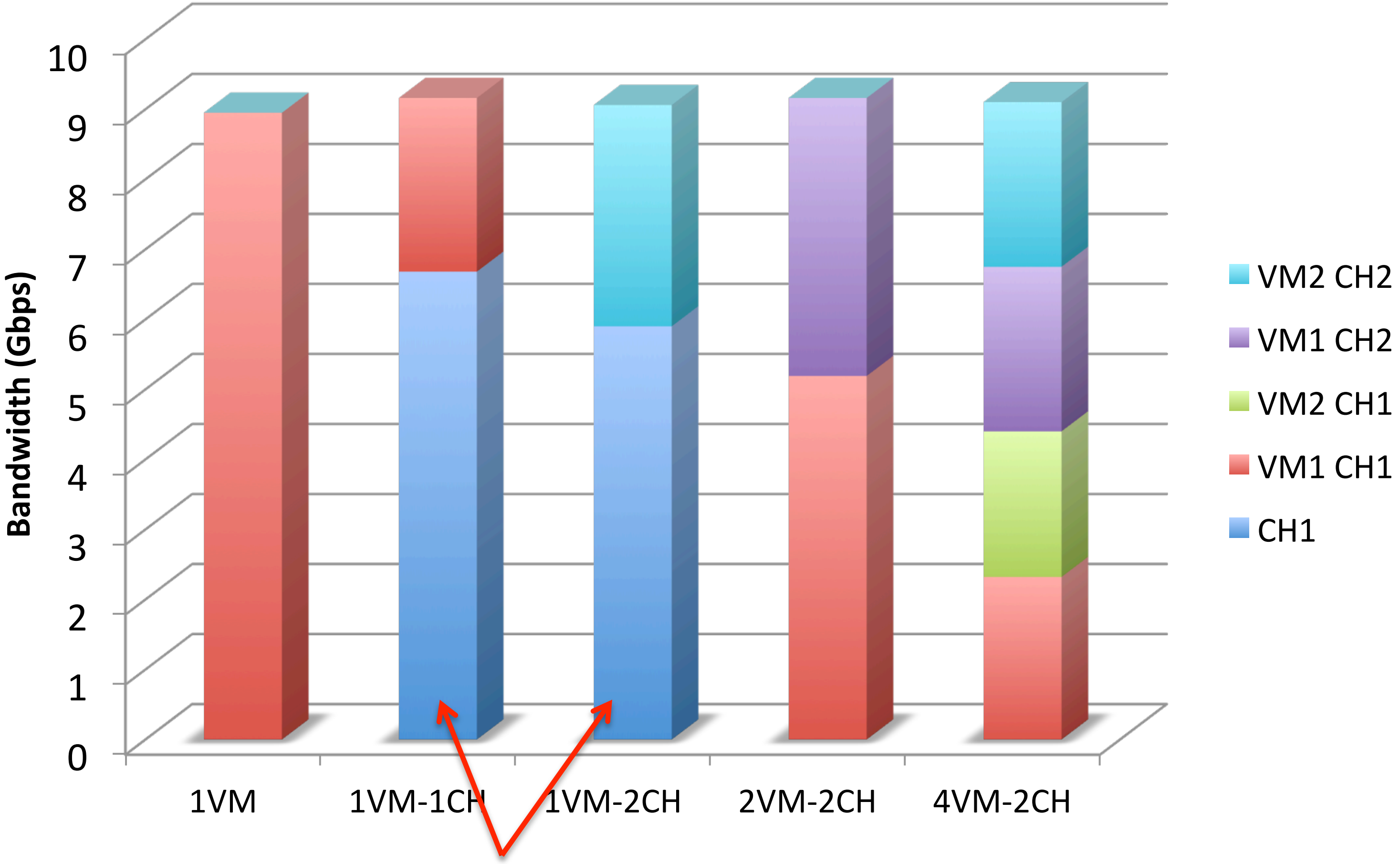
- A secure I/O device sharing scheme that
  - Leverages the address/ID translation schemes, such as EPT, BAR translation registers, IOMMU, LUT
  - Prevent unauthorized accesses to shared I/O devices

- A fully operational prototype
  - Demonstrates the feasibility and efficiency of the software-based MR-IOV
  - Virtually no throughput/latency penalty when compared with SRIOV.

Shalom!

THANK YOU

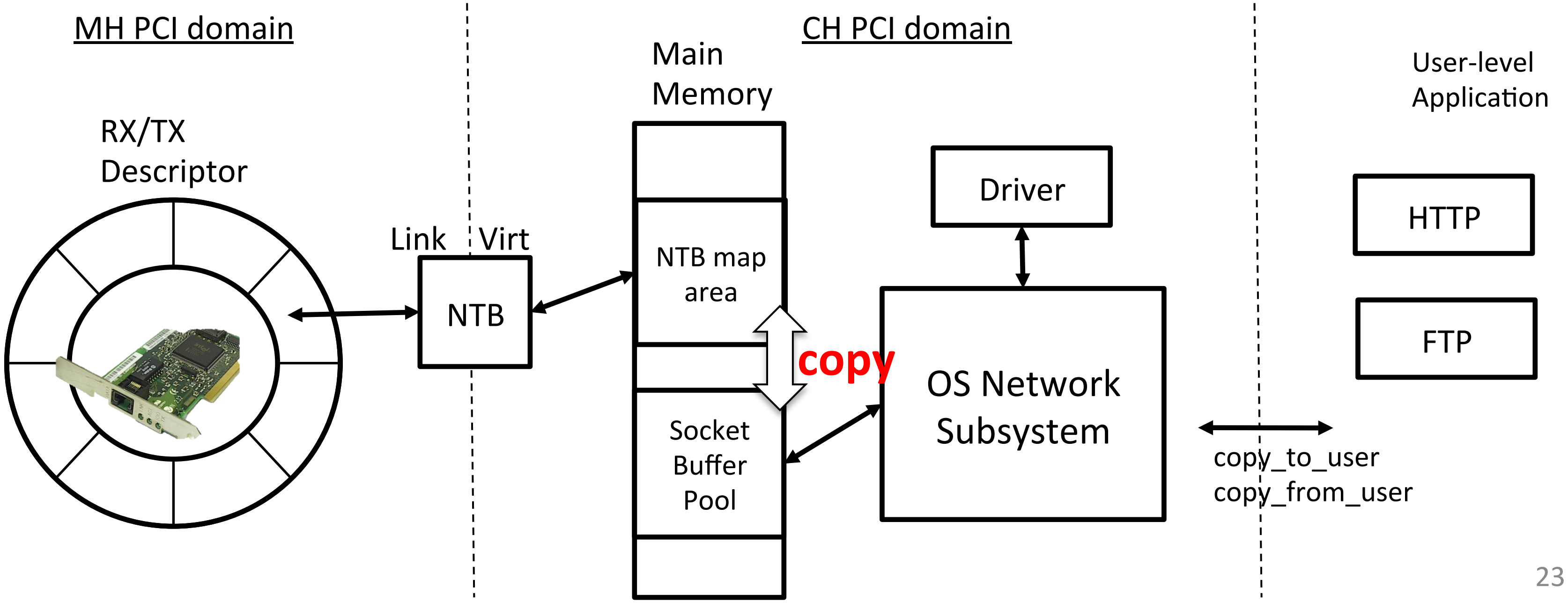
Multiple VMs and CHs



Little unfairness
between CH and VM

Optimization

- Zero driver modification
  - Driver on the CH only see CH’s physical address space.
  - Solution: trapping the DMA access API.
- Zero-copy
  - Limited BAR size support in BIOS (between 3G-4G)
  - Solution: Mapping the entire CH’s physical memory space to avoid copying.



Scalability / Compatibility

- How many CHs/VMs can Ladon support?
  - Each CH: 2 BARs for CSR and MMIO, 2 BARs for DMA and MSI-X
  - As many VMs as a CH could run
  - As many CHs as allowed by the PCIe switch
- How many SRIOV devices can Ladon shares?
  - As many as allowed by the PCIe switch
- Vendor-neutrality
  - Ladon only uses the basic features of NTB: address translation and LUT
  - Intel's on-board NTB: Xeon C5500/C3500