工業技術研究院
Industrial Technology
Research Institute

Next Frontier of Data Center Hardware Architecture

Tzi-cker Chiueh

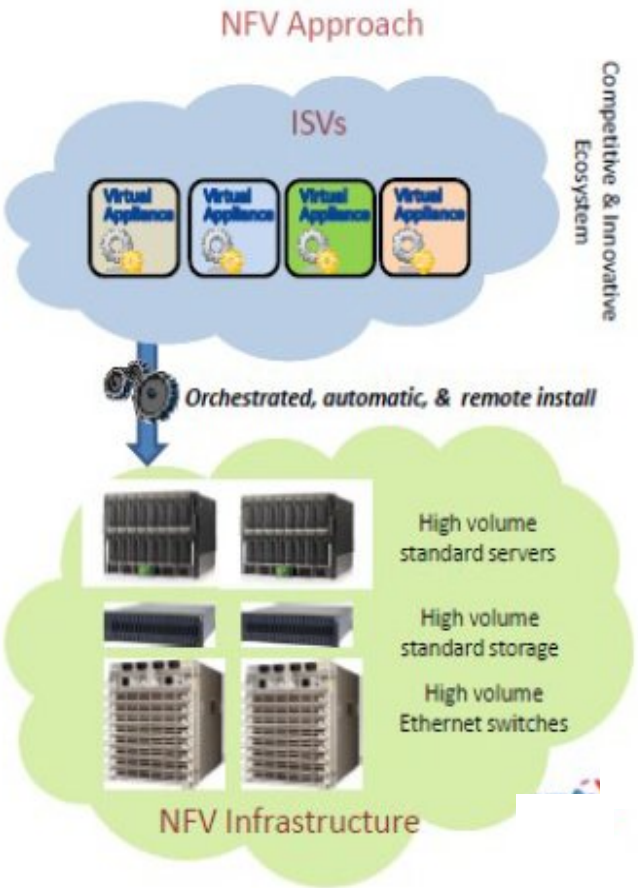
Information and Communications Laboratories

Emerging Data Center Requirements

- Disaggregated Rack Architecture (DRA)
  - Independent power domain and upgrade cycle for each resource type
  - Hardware as a service: How to partition hardware resources among physical servers at run time rather than at manufacture time?
- Hyper-Convergence Architecture
  - Combining compute, network and storage virtualization solutions into one box so that they are managed and scaled out as an indivisible unit
- Network Function Virtualization (NFV)
  - A virtualization platform for running NFV applications that supports low-latency and real-time network packet processing
- Virtual Mobility Infrastructure (VMI)
  - A scalable computing platform for running Android applications in the cloud

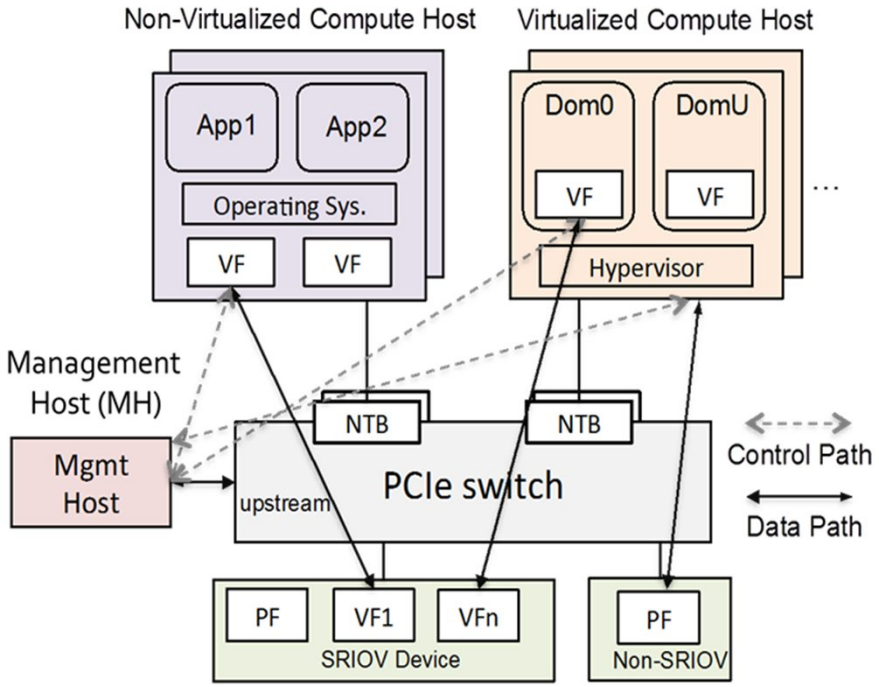
Network Function Virtualization

- Virtualization is the linchpin of 5G wireless communication architecture
- Telecom functions run in VMs on a virtualized IT infrastructure
- Could be applied to access networks as well as core networks



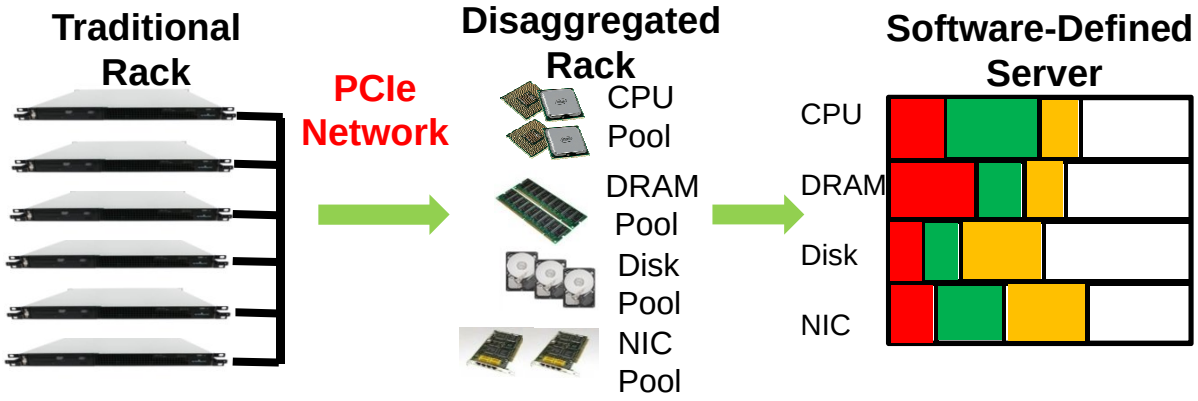
ITRI Rack Scale Architecture

- ITRI RSA is a disaggregated rack architecture that uses PCIe to connect machines within a rack and Ethernet to connect to other racks, and enables
  - Low-latency and high-bandwidth intra-rack communications
    - PCI Express Gen3: 8 lanes = 128 Gbps
  - Direct access to any remote memory/NIC/disks within a rack



Key Advantages

- Disaggregated Rack Architecture (DRA)



- Hyper-Convergence Architecture
  - Low-cost, low-power and high-performance connectivity for east-west traffic: a saving of $1K USD per server
  - Low-latency mirroring for NVMe
- Network Function Virtualization (NFV)
  - VMs directly access network interfaces
  - Network interface interrupts are delivered directly to VMs

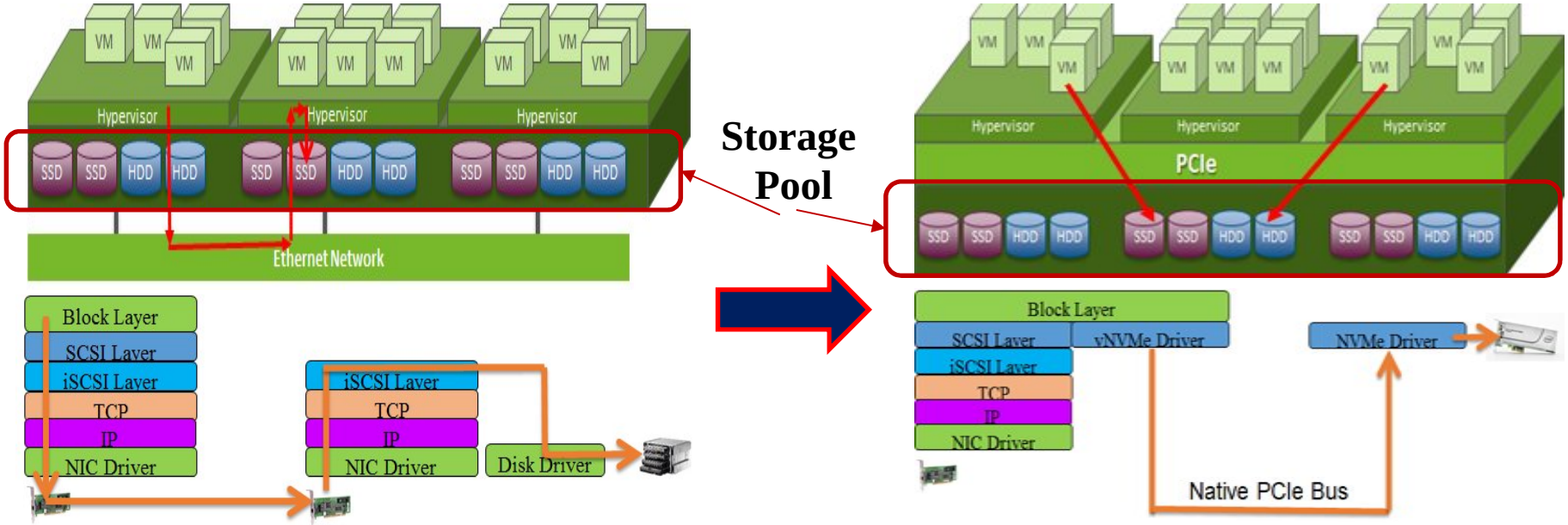
Direct Access to Remote Resources

- Virtual Functions on an SRIOV NIC are assigned to different VMs on a single host
- Software-based MRIOV: Virtual Functions on an SRIOV NIC are assigned to different VMs on multiple hosts
- Host 1 can access non-SRIOV disks attached to Host 2
  - RAID over disks from multiple hosts
  - Almost identical performance as RAID over local disks
- Host 1 can access the main memory of Host 2
  - Only works for multiprocessor chipset
  - 50 nsec vs. 5 usec
  - Must be made non-cacheable when used as DMA targets

Use Case: Hyper-Converged Storage

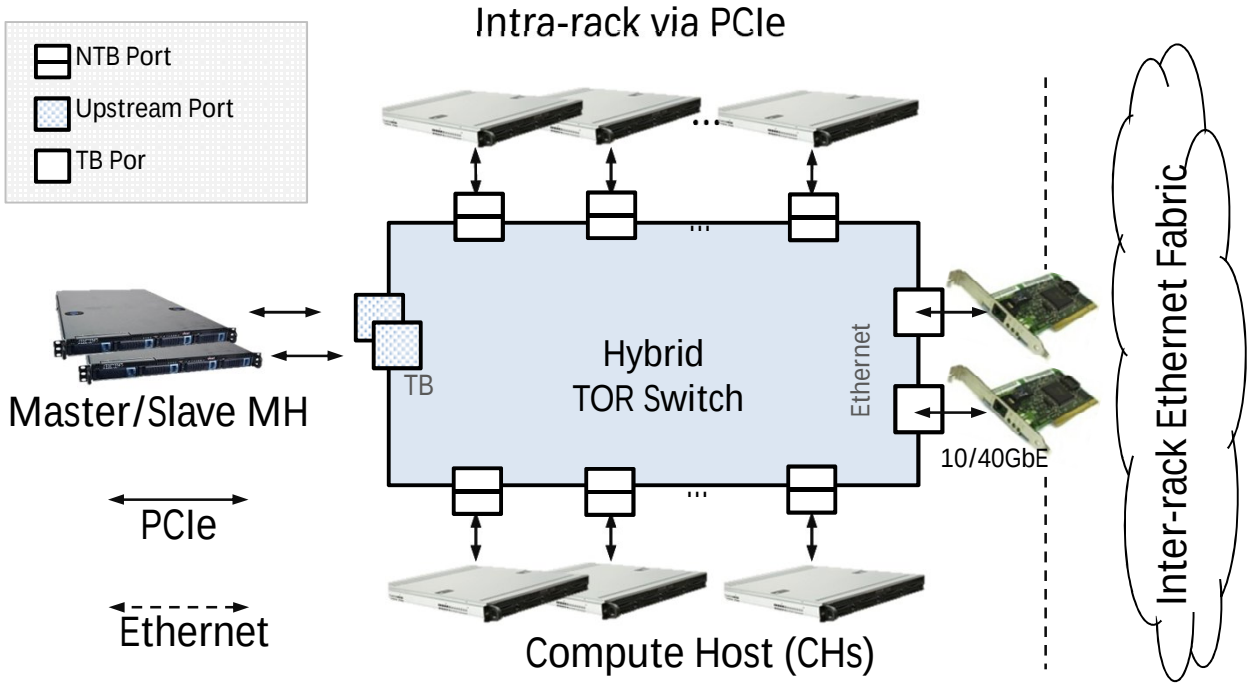
- iSCSI ➔ TCP/IP ➔ Ethernet ➔ TCP/IP ➔ iSCSI ➔ SCSI

- SCSI
- The same command for locally attached storage as for remote storage



It Gets Better: PCIe for Communication

- HRDMA: Hardware-based Remote DMA between servers
  - Direct read and write remote memory through DMA
- EOP: Ethernet Over PCIe for socket-based communications
  - Intra-rack via PCIe using HRDMA
  - Inter-rack via shared Ethernet NICs
- Hybrid TOR switch: Ethernet and PCIe ports



Comparison

Criterion	ITRI	Intel
Intra-rack communication	PCIe	Ethernet
Inter-rack communication	Ethernet	Ethernet
Bandwidth of intra-rack communication	++	++
Latency of intra-rack communication	++	+
Run-time server re-sizing	Yes	No
Network interface resource pooling	+	--
Hard/Solid-state disk resource pooling	+	--
Memory resource pooling	+	--
Cost of intra-rack connectivity	++	--

Comparison between PCIe and Ethernet

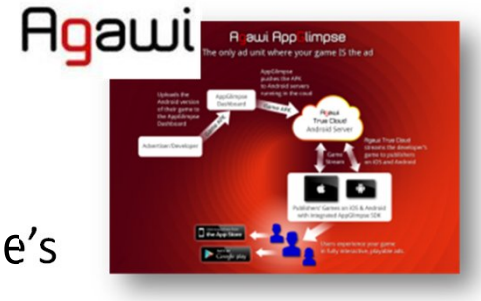
Unit: USD

	Maximum Bandwidth	Adapter Cost per port	Switch Cost per port	Total Cost per port	Source
PCIe (Gen 3)	128Gbps	0	150 900 / 6 ports	**150** **(0+150)**	Avago
10GE	10Gbps	200	100 1600 / 16 ports	**300** **(200+100)**	Amazon
25GE	25Gbps	340	480 7800 / 16 port	**820** **(340+480)**	reseller
40GE	40Gbps	400	580 7000 / 12 ports	**980** **(400+580)**	reseller
100GE	100Gbps	1200	600 19600 / 32 ports	**1800** **(1200+600)**	reseller

Virtual Mobility Infrastructure

Why VMI?

- An enormous number of applications running on Android and iOS devices
- What if we can run them on a common cloud infrastructure and expose their GUI to whatever smart devices out there?
  - From Deep Link to Application Streaming: Google's Agawi acquisition



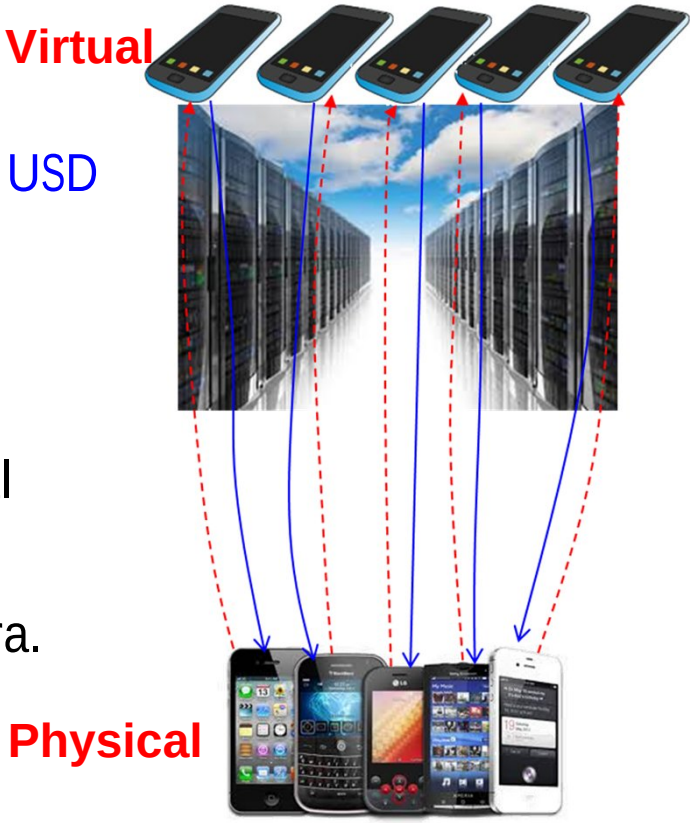
- Use cases:
  - Policy-based smartphone application management
  - Mobile security due to BYOD
  - Elimination of app installation hassle
  - Lower energy consumption on mobile device
  - Better app performance when they run on powerful servers
  - Trial of smartphone apps, e.g. smartphone games

Design Considerations

- Server Hardware
  - Cluster of ARM SOC-based devices
  - ARM-based server
  - X86 server
  - X86 server + GPU
- Operating system that apps run on
  - Android on ARM
  - Android on X86 (backed by Asus)
  - Android on IA (backed by Intel)
  - Android (iOS) emulator on Linux/Windows (OS X)
- Virtualization mode
  - Hardware Abstraction Layer: hypervisor
  - OS Layer: Linux container or Docker
  - Multi-programming
- Client device
  - Android
  - iOS
  - Windows
  - Firefox OS

Cloud Infrastructure for VMI

- Proposition: how to leverage the enormous R&D on smartphone SOC for VMI computation?
- Building block: smartphone SOC
  - MTK Helio 10-core MT6795 costs less than $50 USD
- Technical Challenges:
  - Multi-programming or Container on Android
    - Android is designed to run one frontend app
  - Redirection of sensor information from physical smartphone to virtual smartphone, e.g., Multi-touch, GPS, 3-axis accelerator and Camera.
  - Low-latency streaming of virtual frame buffer
  - Smartphone GPU virtualization



Summary

- ITRI rack-scale architecture features software-enabled MRIOV, which allows direct and secure access to all NIC, disks and memory resources within a rack
  - The best known approximation to the ideal of DRA
  - PCIe networking provides an inexpensive alternative for east-west traffic
  - Scalability limit
    - Number of lanes per server
    - Number of ports in PCIe switch chip (currently 96)
    - Could be improved with a mesh of PCIe switches
- VMI ushers in a world of install-less apps, and opens up a brand new dimension of extending the functionalities of future smartphones

Thank You!

Questions and Comments?

tcc@itri.org.tw