PCI Express and IOV: Maximizing Multi-Processor Systems

Shreyas Shah

PLX Technology, Inc

Agenda

- PCI Express: From PCIe 1.1 to PCIe 2.0
- Data Centers Issues
- Hypervisors
- SR-IOV
- MR-IOV
  - ✓ Reduction in Latency
  - ✓ Higher Performance
  - ✓ Lower Cost
- Multi Processor Systems
- Q & A
- Conclusion

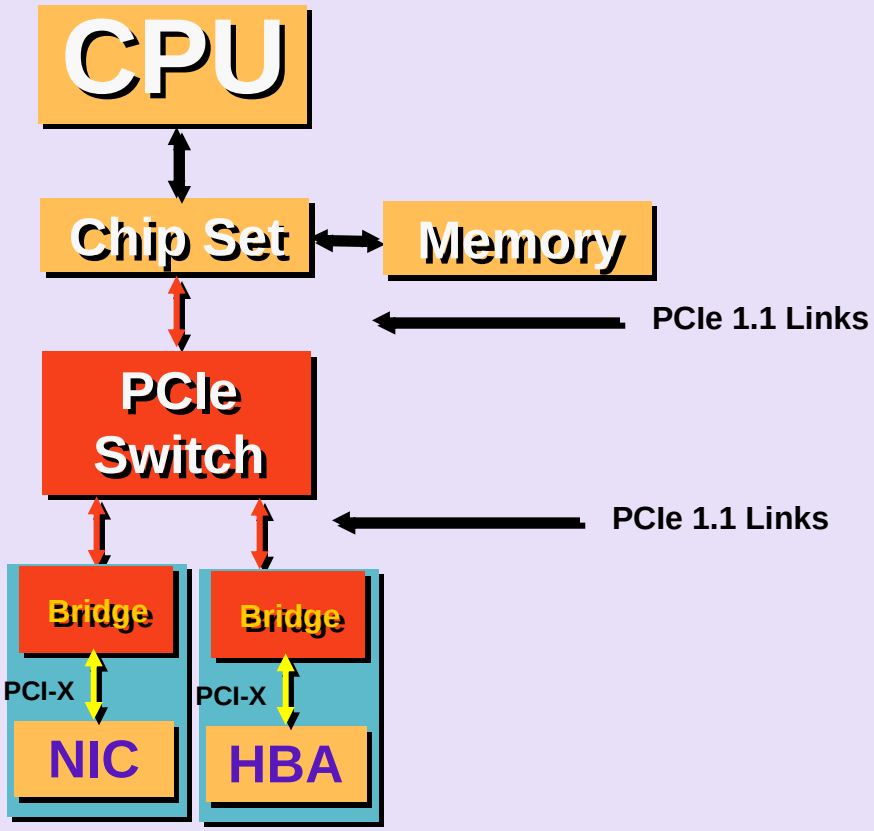
Agenda

- PCI Express: From PCIe 1.1 to PCIe 2.0
- Data Centers
- Hypervisors
- SR-IOV
- MR-IOV
  - ✓ Reduction in Latency
  - ✓ Higher Performance
  - ✓ Lower Cost
- Multi Processor Systems
- Q & A
- Conclusion

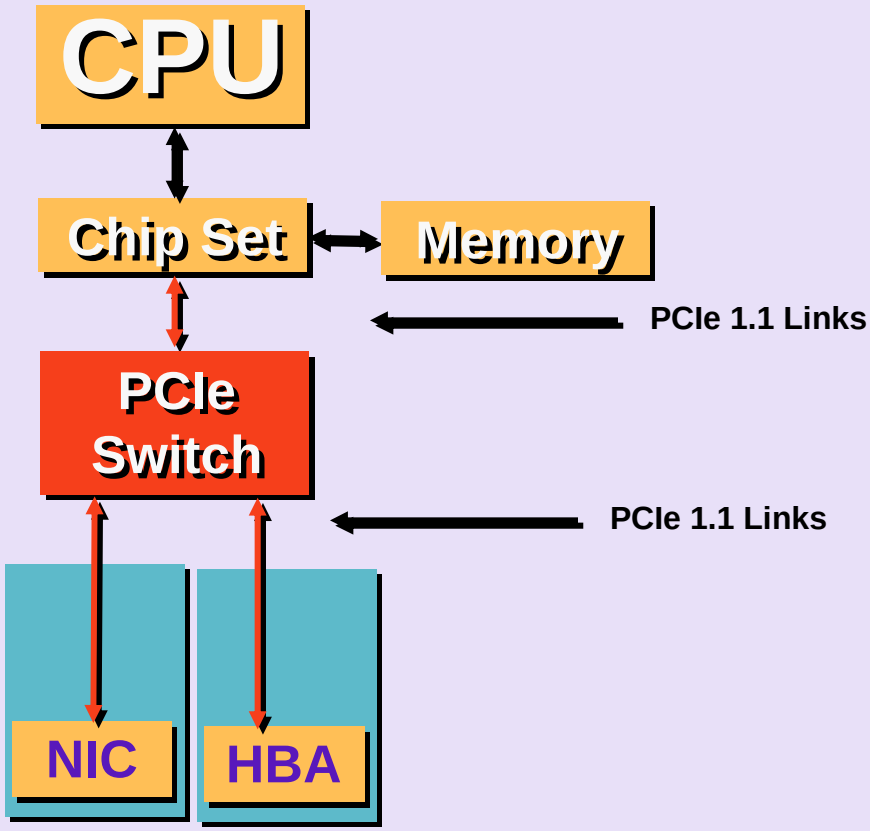
PCI Express: PCIe 1.1 to PCIe 2.0

- PCIe 1.1 Signaling rate: 2.5 Gbps with 8b/10b
- Replace PCI and PCI-X with serialized clock and data for higher performance and scalability
- PCIe 2.0 signaling rate: 5 GT/s with 8b/10b
- Connect the switches at higher speed

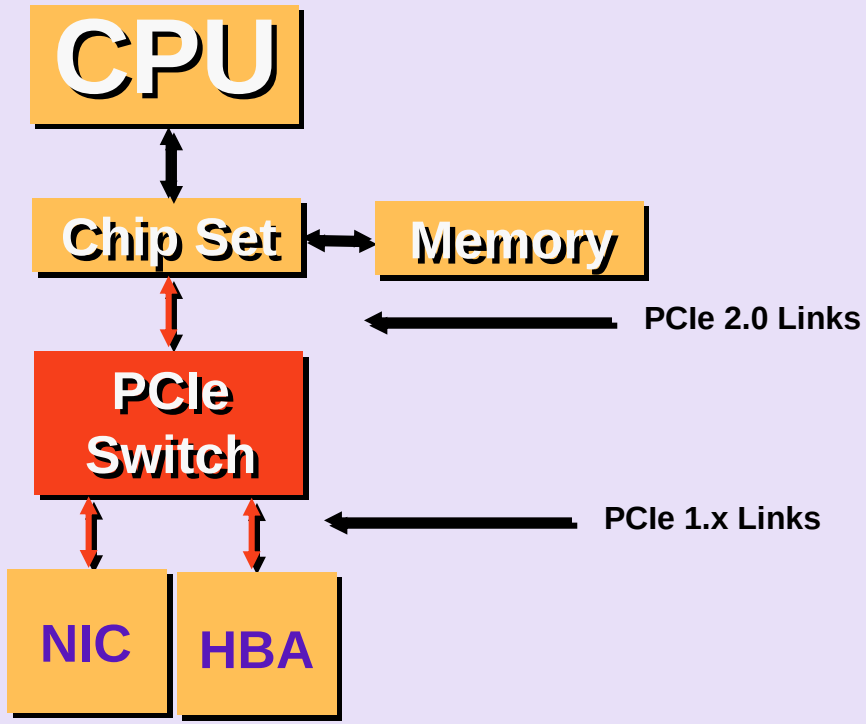
PCI Express Evolution: Bridged to PCI-X cards



PCI Express Evolution: Native PCIe 1.1 cards



PCI Express Evolution: 5GT/s links to 2.5GT/s cards



Agenda

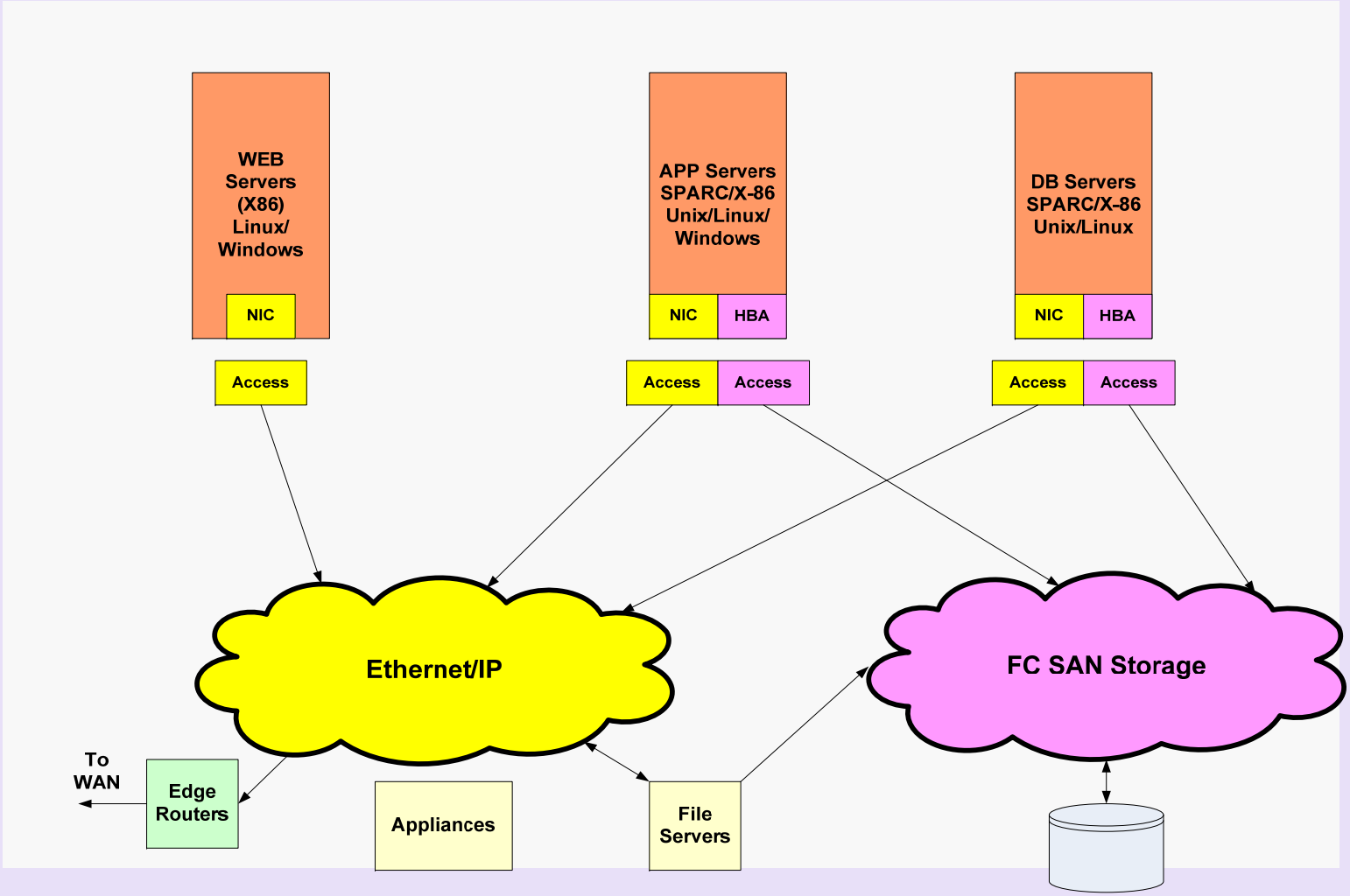
- PCI Express: From PCIe 1.1 to PCIe 2.0
- Data Centers Issues
- Hypervisors
- SR-IOV
- MR-IOV
  - ✓ Reduction in Latency
  - ✓ Higher Performance
  - ✓ Lower Cost
- Multi Processor Systems
- Q & A
- Conclusion

Data Center Issues

- Lower utilization of Servers
- Networked Storage Connection (Fibre Channel) limited to ~30% of servers
- Higher TCO
- Higher OPEX
  - ✓ Inflexibility of moving applications between servers
  - ✓ Management Cost $\propto$ Number of Systems in DC

Traditional Data Centers

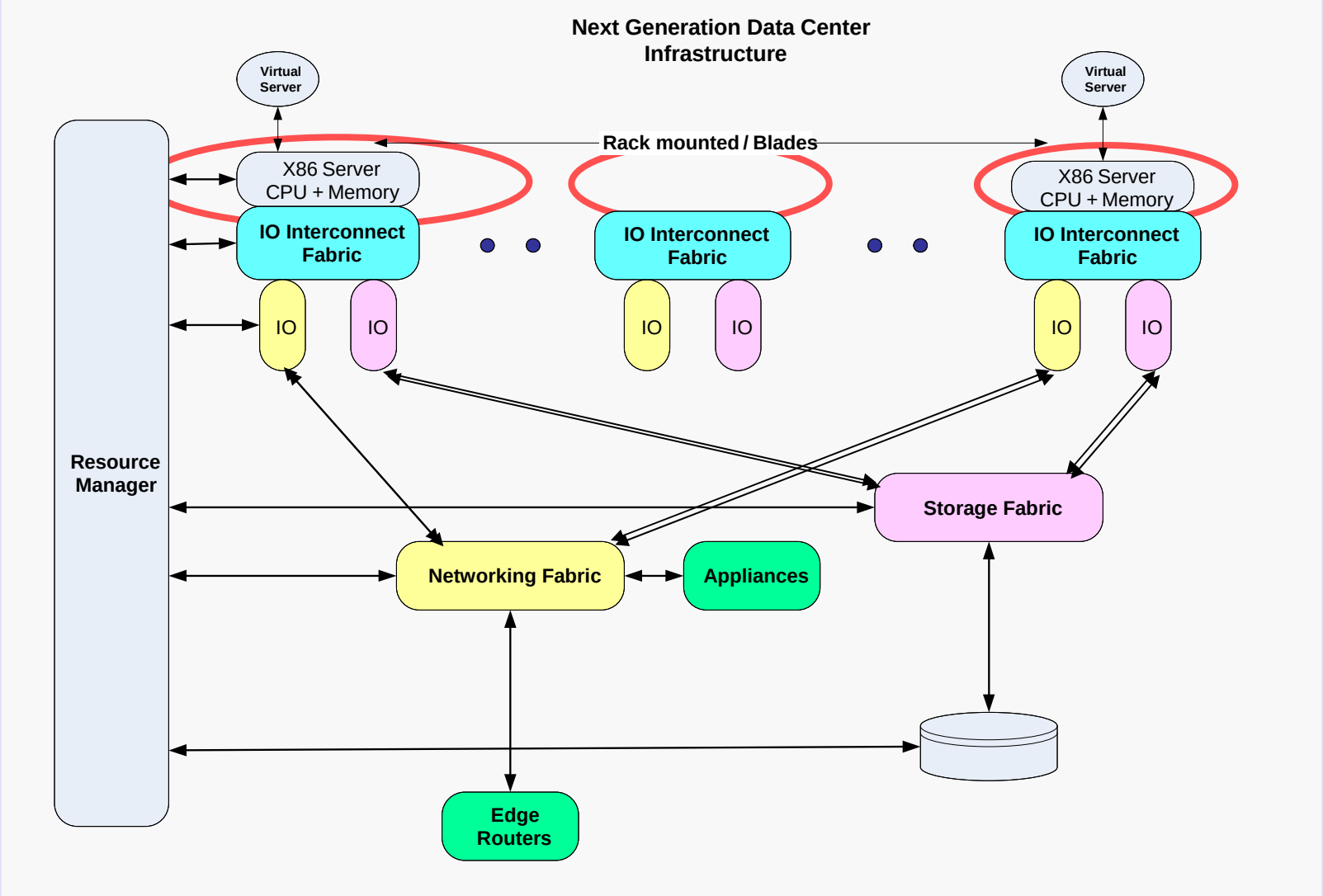
Three Tier Architecture



Data Center Trends

- Fluid infrastructure – Utility computing
  - Decouple Applications from infrastructure
  - Applications can be spawned on any server with privileges maintained
- Virtualized environment
  - ✓ Higher utilization of servers and infrastructure
- Access layer Fabric convergence
  - ✓ Reduction in TCO, Power and OPEX

Next Generation Data Center

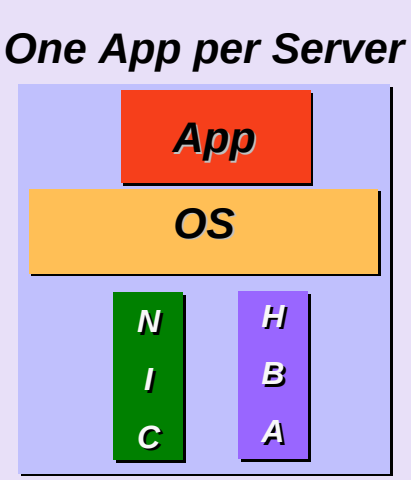


What Do We Have?

- One App per Server
- Mid Tier and DB Tier have two I/O devices per server (Ethernet and FC HBA)
- Other servers have local storage (Hard Disk) inside server

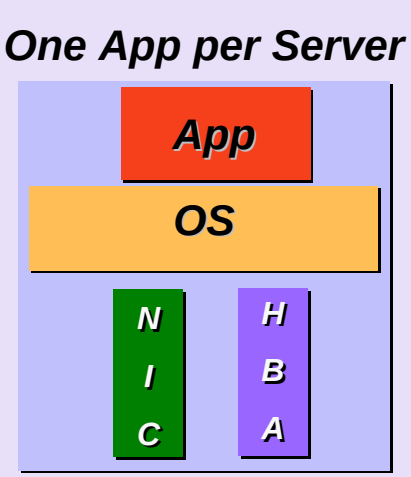
- Infrastructure not flexible ….. Application migration takes weeks ….

Today in Data Centers …



- Inflexible infrastructure
- Applications tied with Server and IO hardware
- Server scheduled maintenance takes Saturday/Sunday nights/mornings
- New server hardware takes 3 to 4 hours to set up and configure

Today in Data Centers …



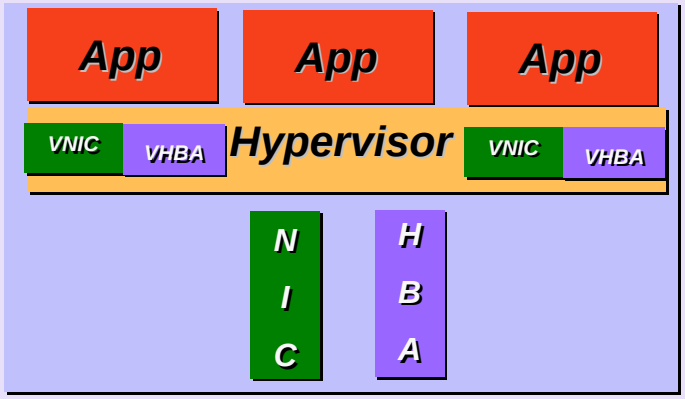
- Per server Number of connections ~ 8/10
- 2 Data, 2 Storage, 2 Management, 2 KVM, 2 low latency high speed fabric
- Infrastructure over-provisioned to meet peak demand of applications
- Stays idle for rest of the time...

Agenda

- PCI Express: From PCIe 1.1 to PCIe 2.0
- Data Centers Issues
- Hypervisors
- SR-IOV
- MR-IOV
  - ✓ Reduction in Latency
  - ✓ Higher Performance
  - ✓ Lower Cost
- Multi Processor Systems
- Q & A
- Conclusion

Hypervisors inside Servers

Multiple Apps per Server



Hypervisors

- Run multiple Guest OS's (Multiple Apps) on physical servers
- I/O devices are being virtualized within Hypervisors
- Next Generation devices include SR-IOV that incorporates IOV in hardware

Limitations of Hypervisors

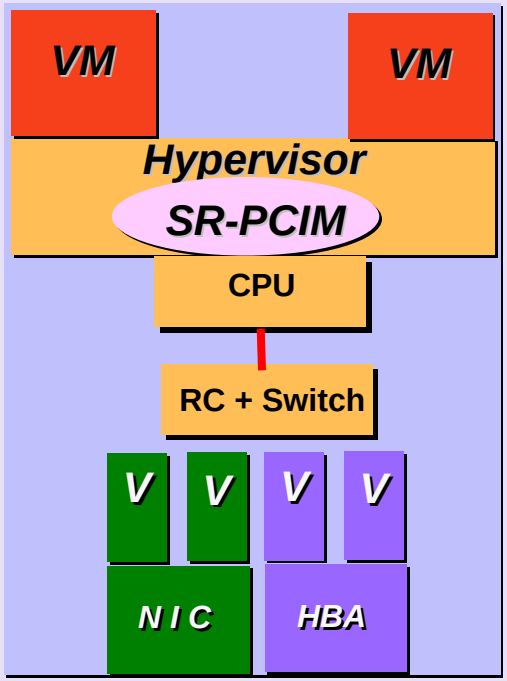
- Software based queuing – Scalability and Performance
- Software based QoS per VM – Scalability and Performance
- Can not scale across the servers
  - ✓ No distributed Hypervisors from Hypervisor leaders...
- I/O devices can not be shared across servers

Agenda

- PCI Express: From PCIe 1.1 to PCIe 2.0
- Data Centers Issues
- Hypervisors
- SR-IOV
- MR-IOV
  - ✓ Reduction in Latency
  - ✓ Higher Performance
  - ✓ Lower Cost
- Multi Processor Systems
- Q & A
- Conclusion

SR-IOV

Multiple Apps per Server



IO virtualization in hardware – SR IOV

SR-PCIM boots through base function

Assign VM with VFs

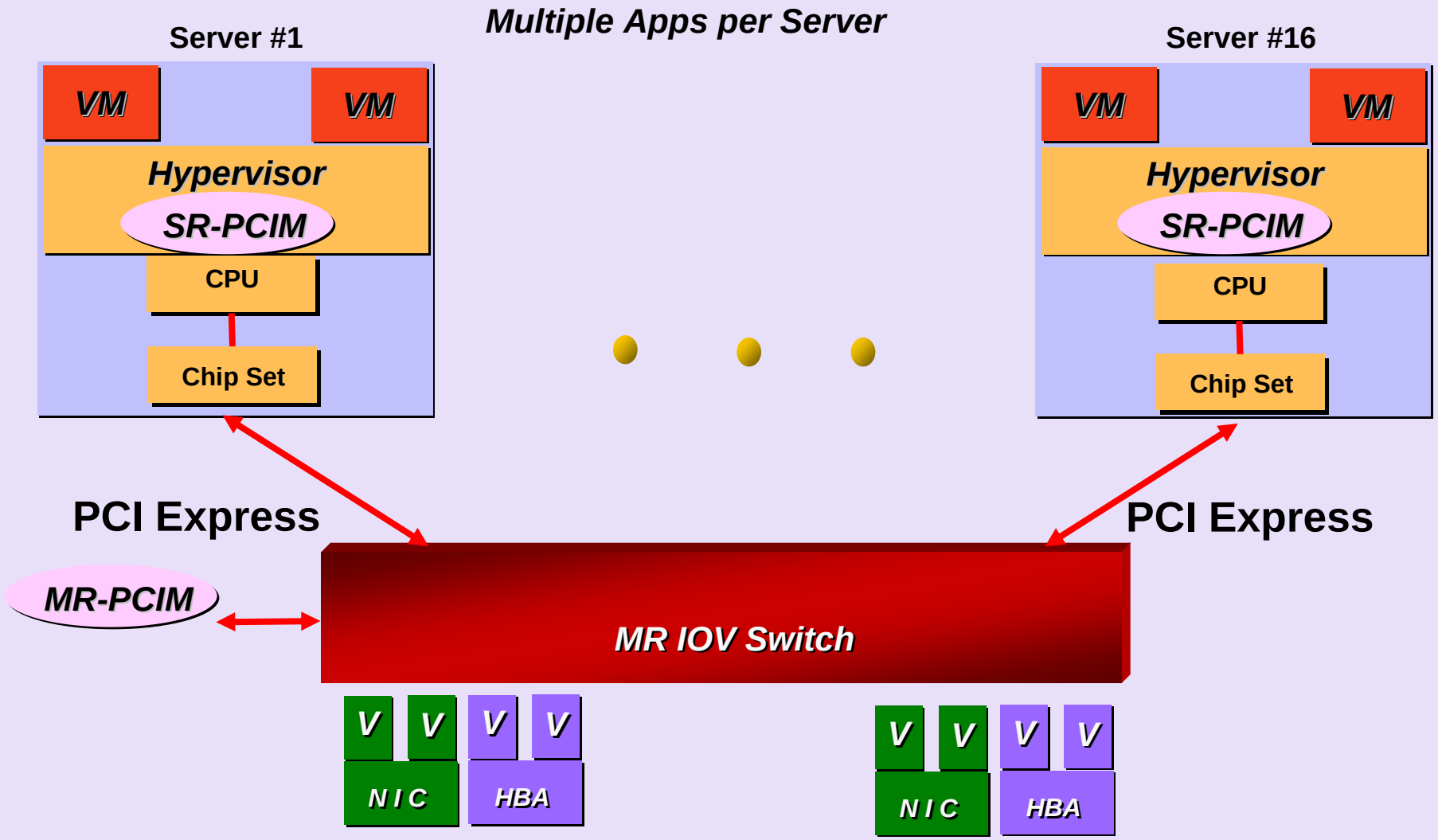
SR-IOV (Cont'd)

- Virtualized I/O shared across virtual machines on single physical server
- SR-IOV Advantages over Hypervisor based IOV
  - ✓ Improves performance and scalability compared to software based solutions
  - ✓ Security between virtual interfaces enforced in hardware
  - ✓ Application QoS – Mission critical and regular applications can share same physical infrastructure with guaranteed performance

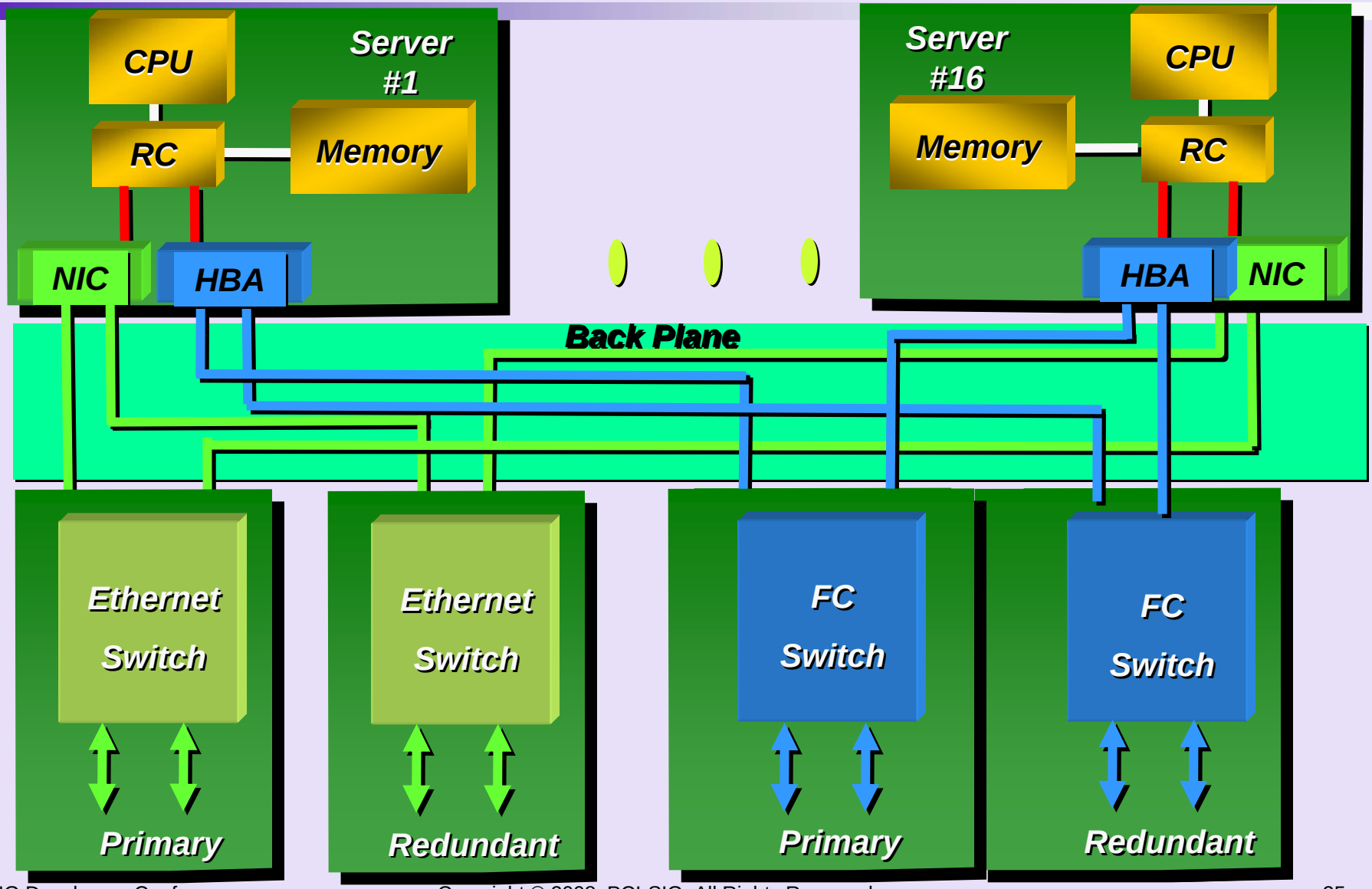
Agenda

- PCI Express: From PCIe 1.1 to PCIe 2.0
- Data Centers Issues
- Hypervisors
- SR-IOV
- MR-IOV
  - ✓ Reduction in Latency
  - ✓ Higher Performance
  - ✓ Lower Cost
- Multi Processor Systems
- Q & A
- Conclusion

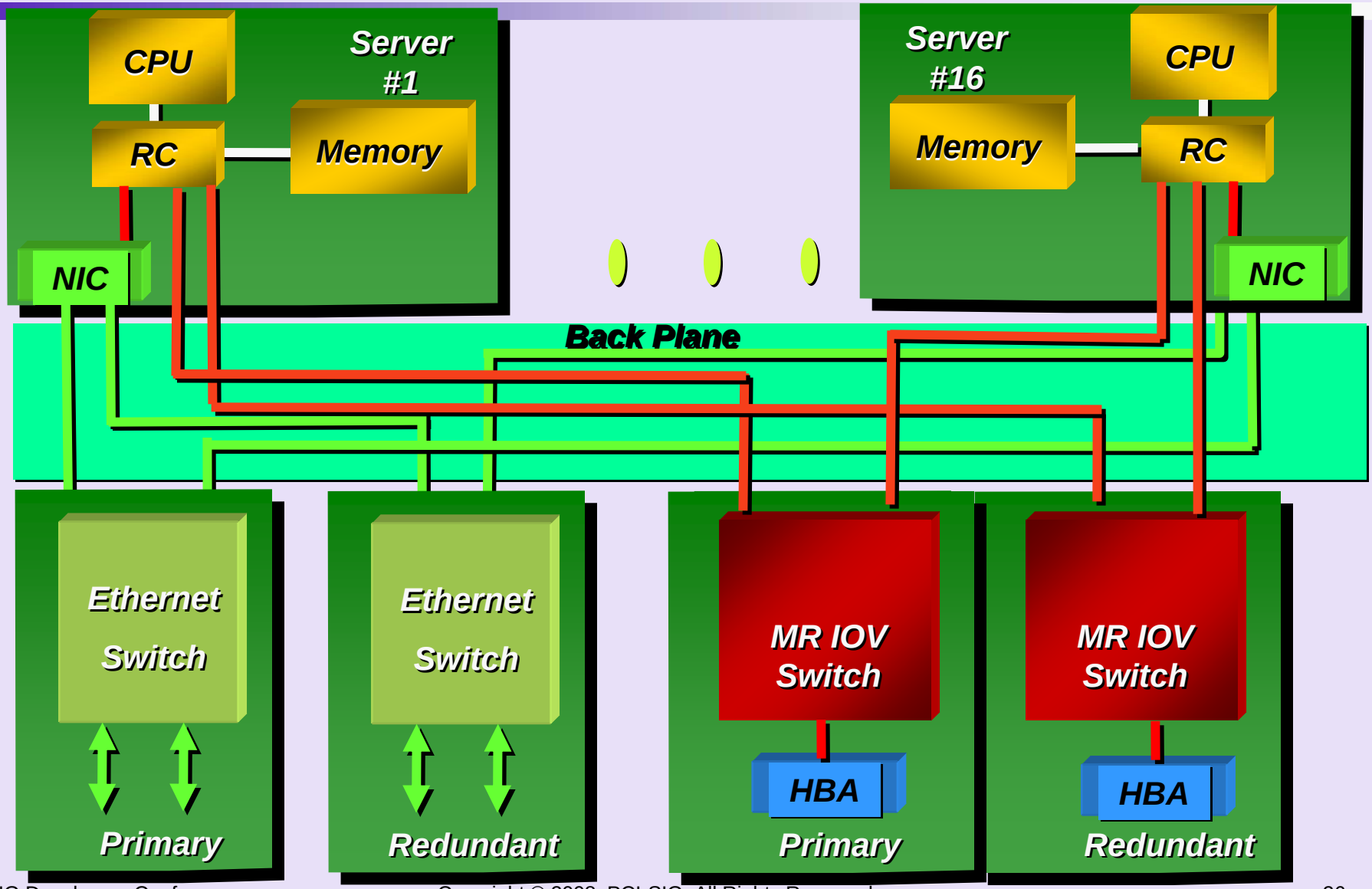
MR-IOV -- Basics



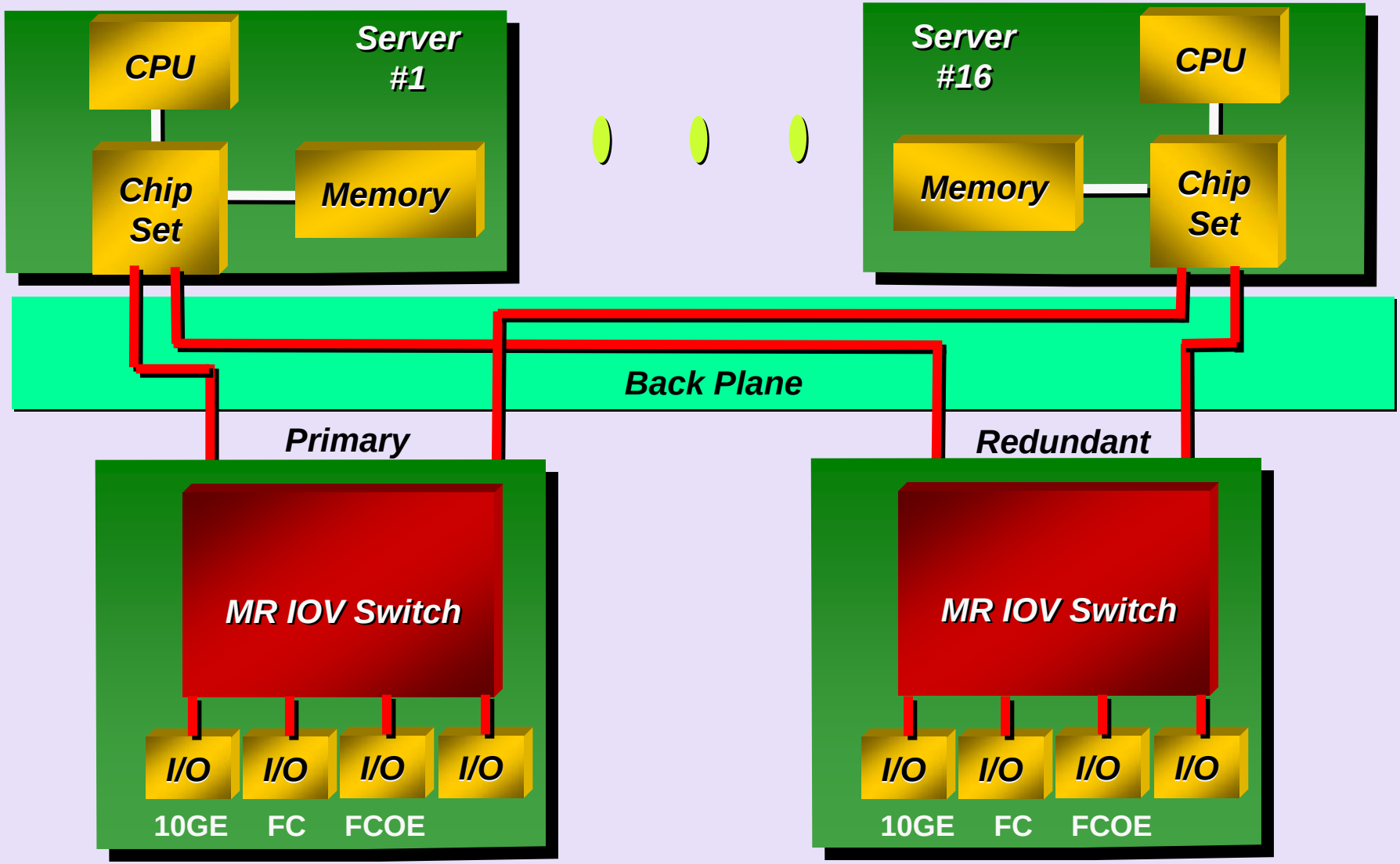
Network Based Back Plane



MR-IOV Network Based Back Plane



MR-IOV Blade Servers



MR-IOV

- MR-IOV de-couples servers and IO devices
- Servers and IO devices can scale independently
- Do not require Access switches
- NICs and HBAs connect to director/core switches
- No Ethernet/FC switch management in access layer

MR-IOV

- MR-IOV offer all the advantages of SR-IOV compared to Hypervisor based IOV
- Virtualized IO interfaces shared across multiple physical servers – Cost and Power
- MR-IOV offers significant other advantages not offered by SR-IOV
  - ✓ Next Slide….

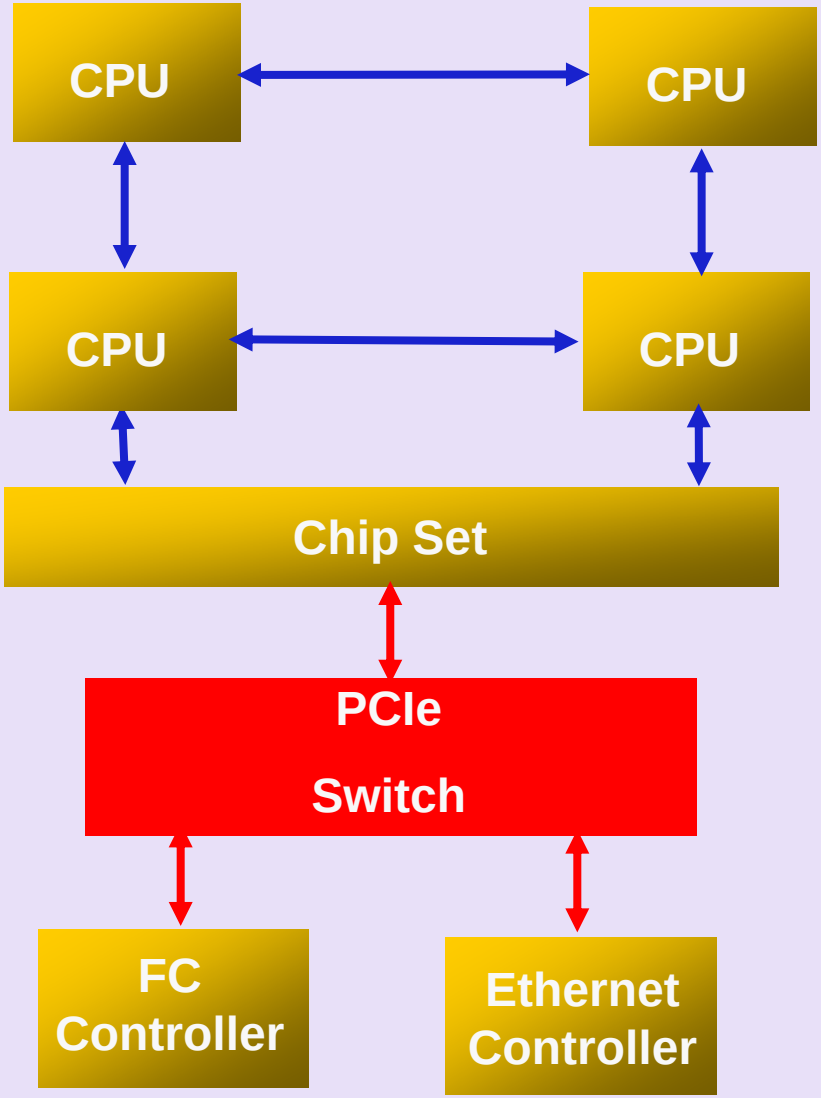
Additional MR-IOV Advantages

- Reduction in cost
  - ✓ Per server – Shared IO – Cost and Power
  - ✓ Per system
- Reduction in Power consumption
  - ✓ Per server
  - ✓ Reduction in access switches
- Efficient usage of network IO devices
- Reduction in latency
- Statistical multiplexing at later stage – Higher IO performance per server

Agenda

- PCI Express: From PCIe 1.1 to PCIe 2.0
- Data Centers Issues
- Hypervisors
- SR-IOV
- MR-IOV
  - ✓ Reduction in Latency
  - ✓ Higher Performance
  - ✓ Lower Cost
- Multi Processor Systems
- Q & A
- Conclusion

SMP Servers

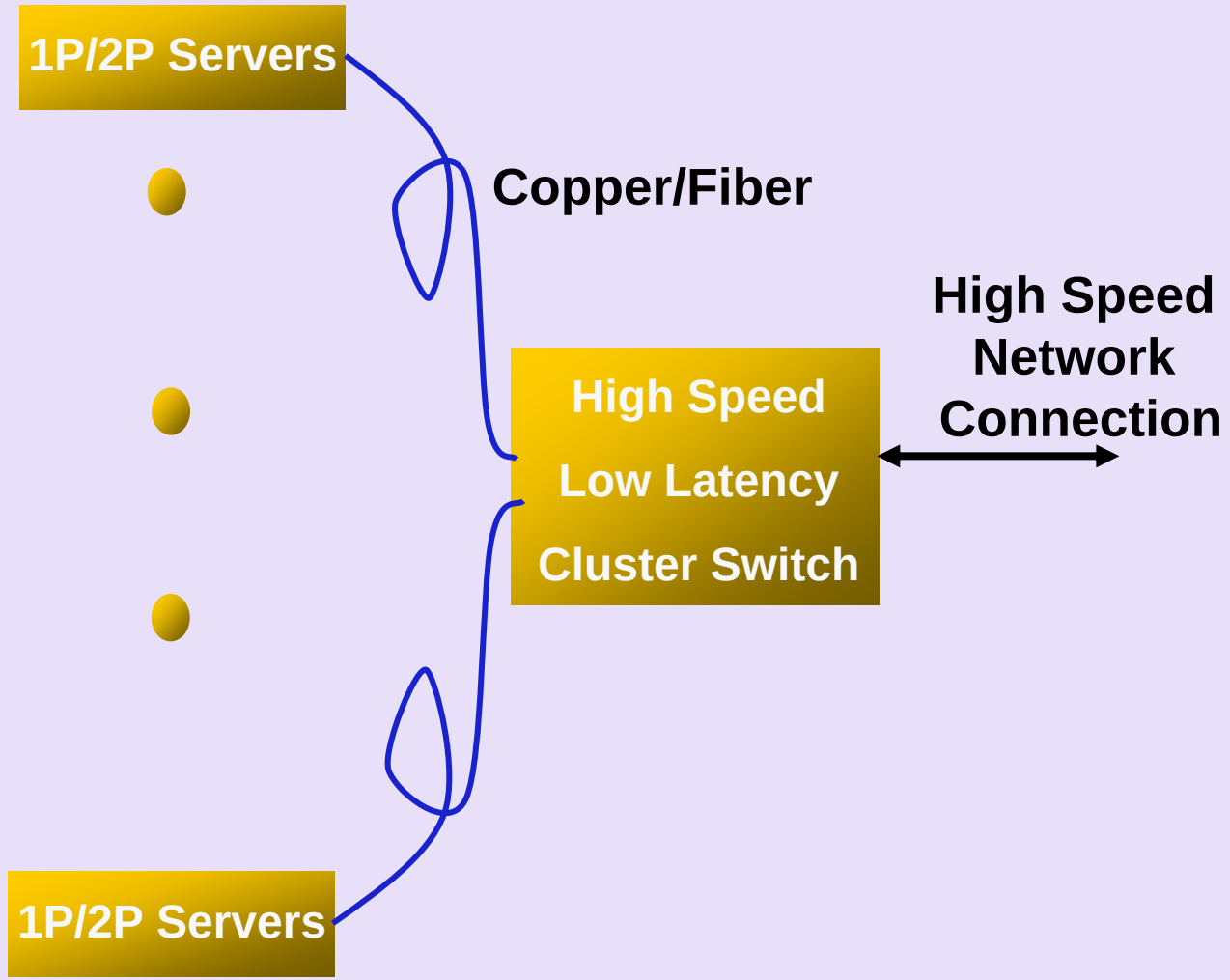


- Process to Process ~ 750ns (Any CPU to any CPU)
- Chipset latency ~ 200ns
- Network controller ~ 140ns

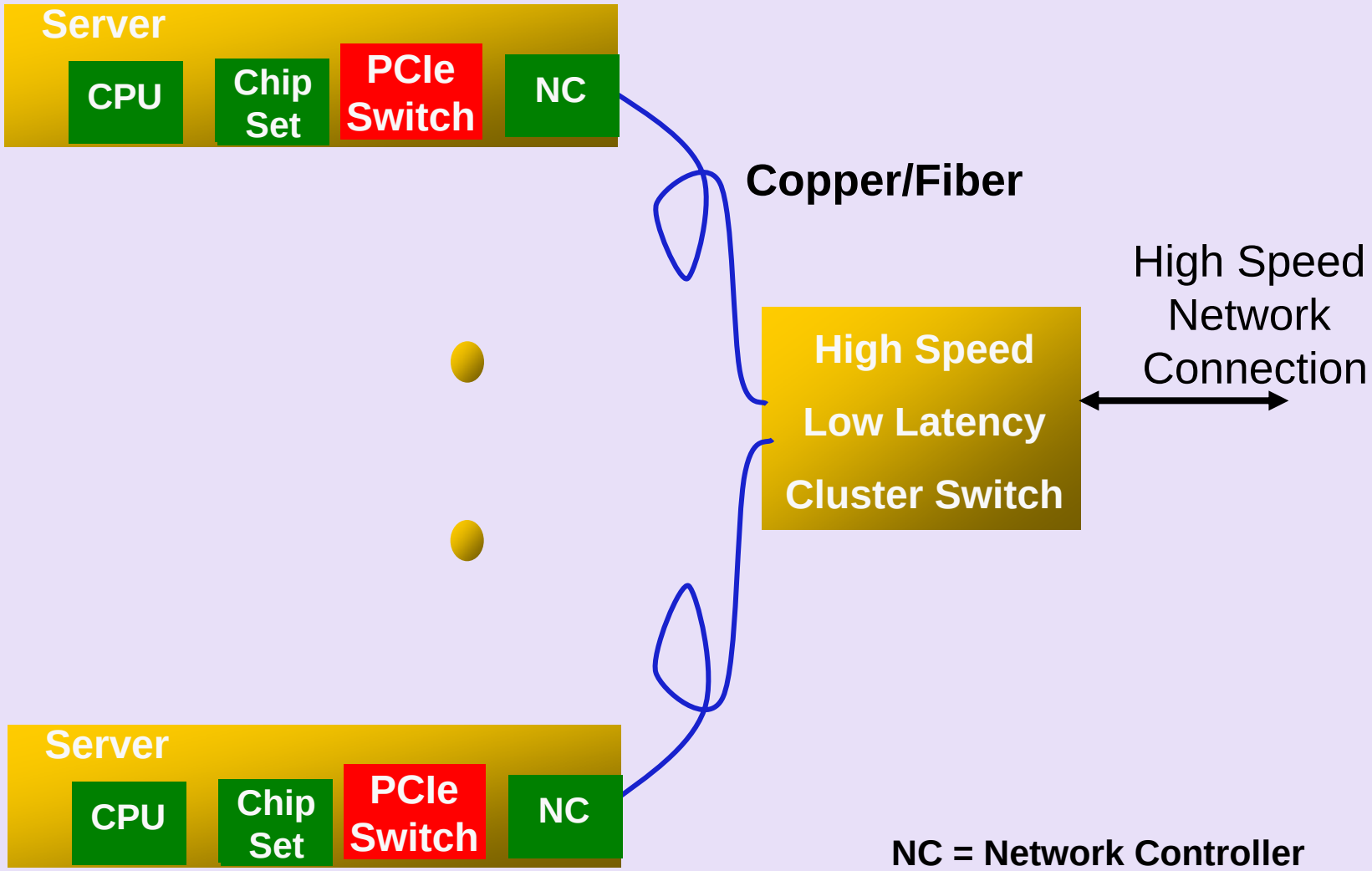
The lower the switch latency the higher the I/O performance

- ✓ Most of the enterprise workload demands high-performance I/O

Scale Out Architecture



Scale Out Architecture



Scale Out Architecture

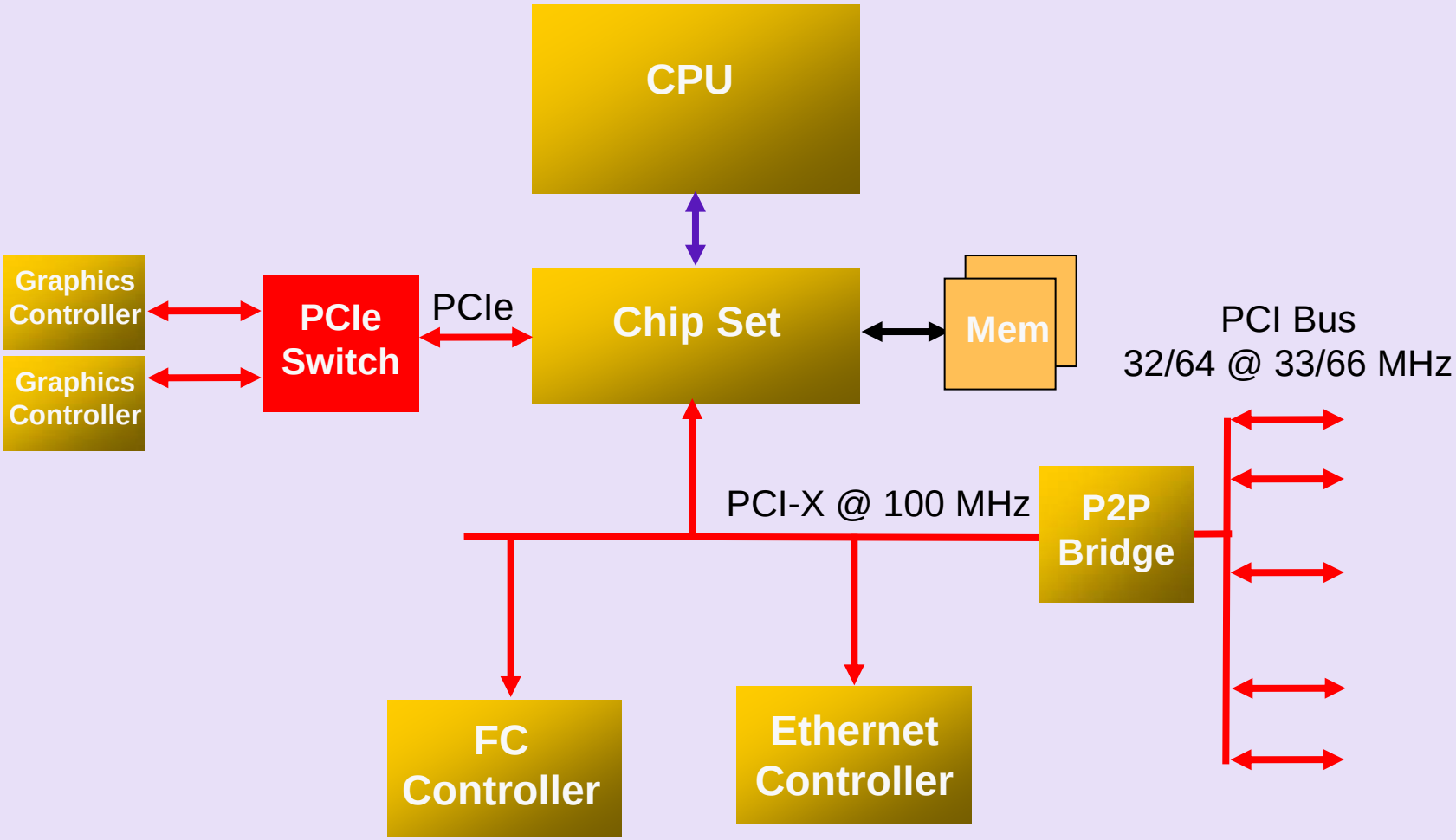
PCIe switch latency budget

- Process2Process – 2 x (Chipset Latency + Network controller latency) – Cluster Switch Latency
  - = 1.2µs – 2(200ns + 140ns) - 200ns
  - = 320ns For Two PCIe switches
  - = 160ns/ PCIe Switch

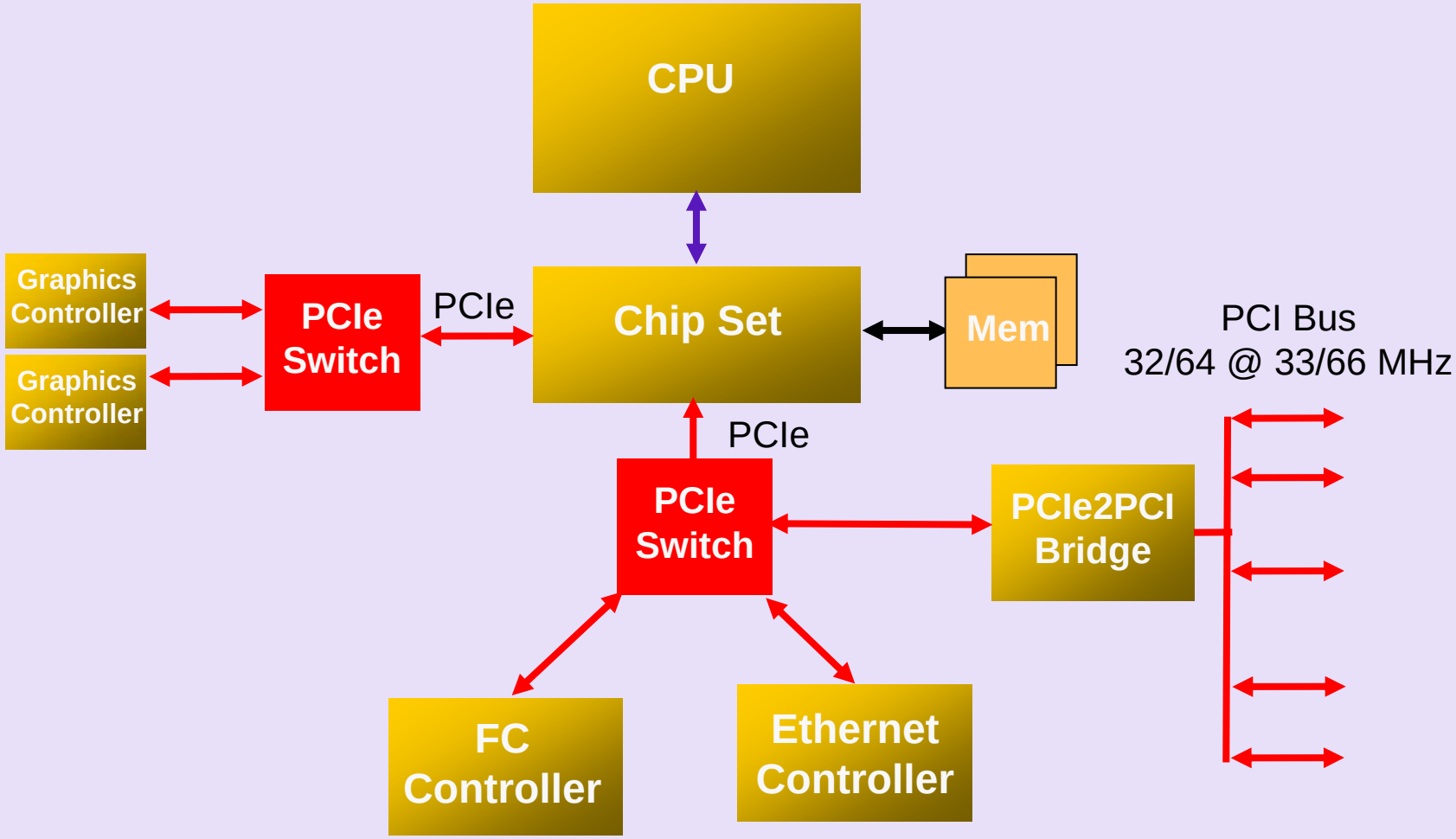
Agenda

- PCI Express: From PCIe 1.1 to PCIe 2.0
- Data Centers Issues
- Hypervisors
- SR-IOV
- MR-IOV
  - ✓ Reduction in Latency
  - ✓ Higher Performance
  - ✓ Lower Cost
- Multi Processor Systems – Server Architectures
- Q & A
- Conclusion

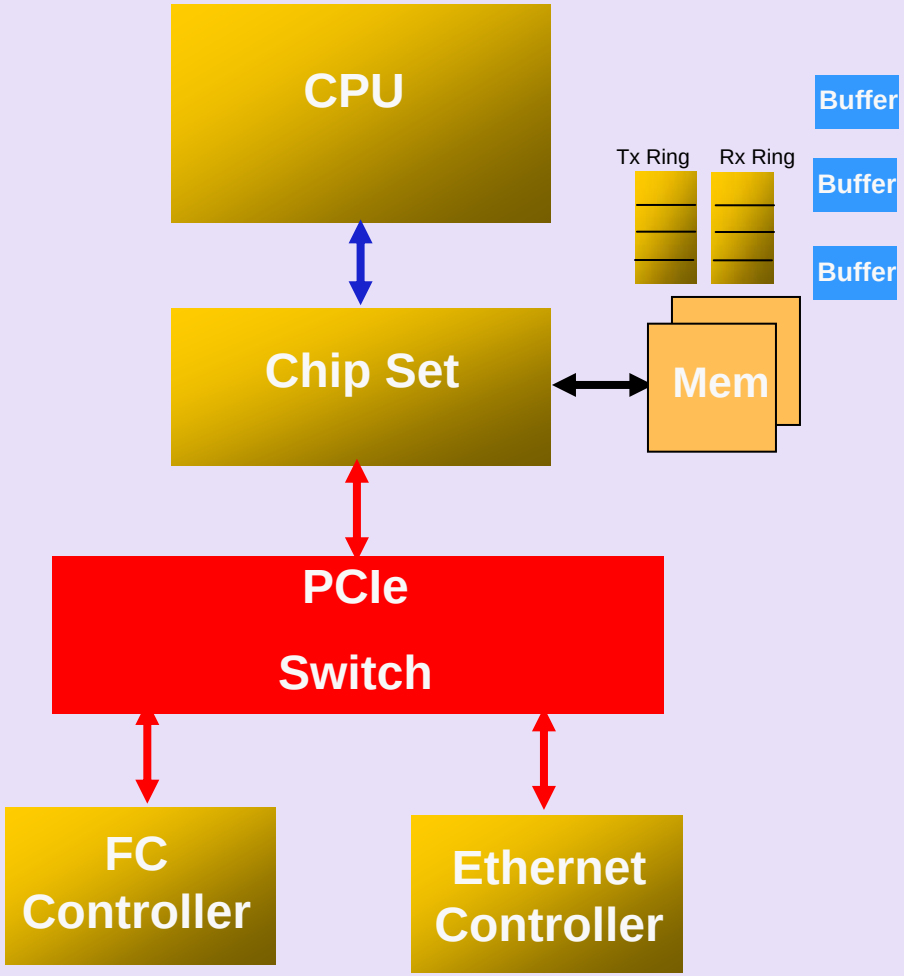
Example Servers w/GPUs



Example Servers w/GPUs



Example Servers and I/O Communication



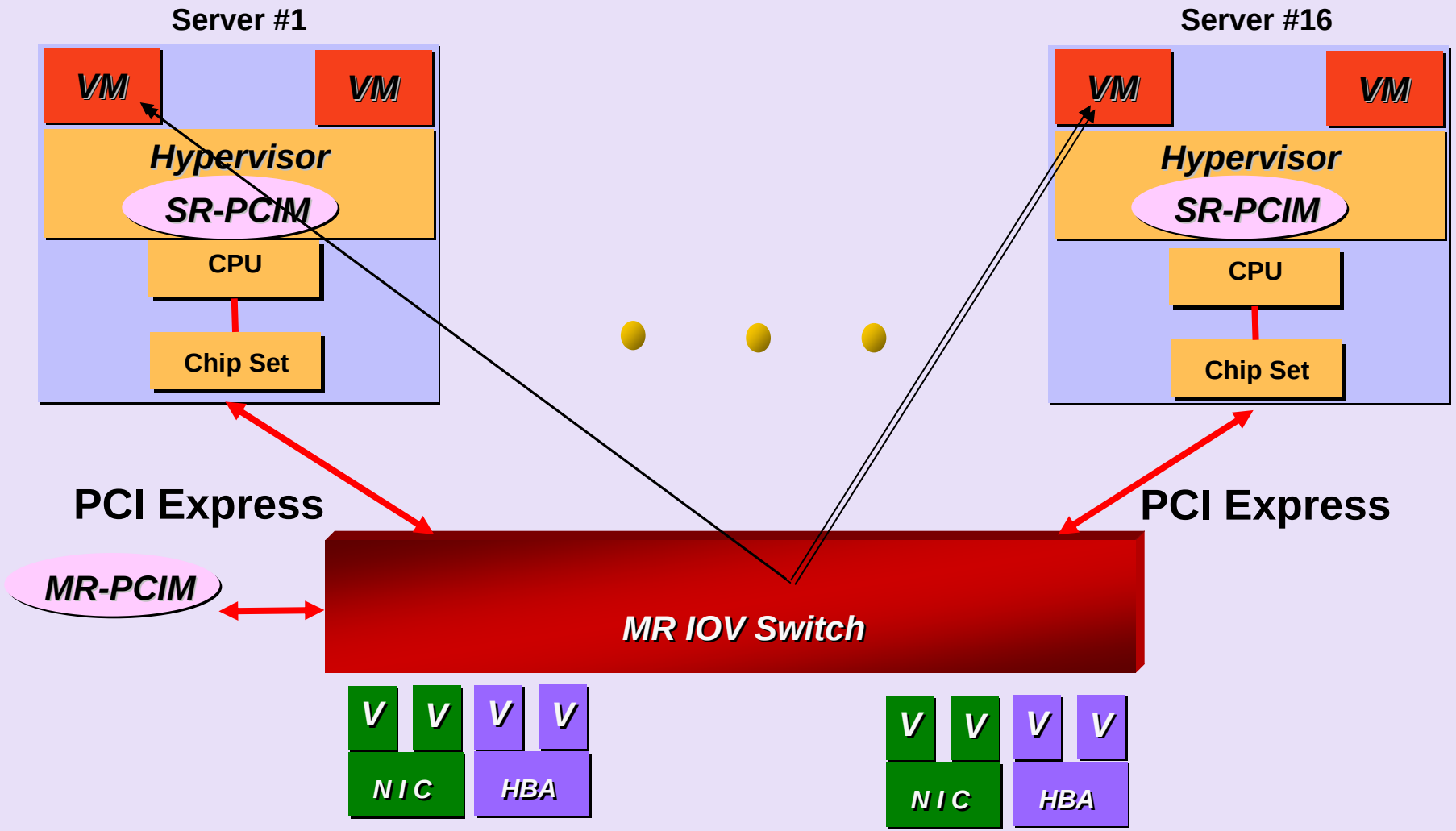
Agenda

- PCI Express: From PCIe 1.1 to PCIe 2.0
- Data Centers Issues
- Hypervisors
- SR-IOV
- MR-IOV
  - ✓ Reduction in Latency
  - ✓ Higher Performance
  - ✓ Lower Cost
- Multi Processor Systems
- Q & A
- Conclusion

Multi-Processor Systems

- Processors can communicate with each other through DMA engines as part of the chip set
- Processors access very high speed IO devices
- No Access switches
  - ✓ FC – Saves 2 us un-loaded latency per switch
  - ✓ Ethernet – 40 to 60 us (Loaded latency GE) and 300 ns for un-loaded 10GE latency per switch

Multi Processors Cont…



Summary

- Hypervisors solves server utilization problem
- MR-IOV solves
  - ✓ Servers Sharing of IO devices
  - ✓ Each server in data center == CPU, Memory and Chip set
  - ✓ Lower TCO
  - ✓ Lower OPEX
    - One can move application to any server
    - Reduction in management cost
  - ✓ Application abstracts out the server and I/O hardware => Higher utilization, lower cost and lower power

Thank you for attending the
PCI-SIG Developers Conference 2008

For more information please go to
www.pcisig.com