Virtual Networks:

For Storage and Data

or

Untangling the Virtual Server Spaghetti Pile

Howard Marks
Chief Scientist
hmarks@DeepStorage.net

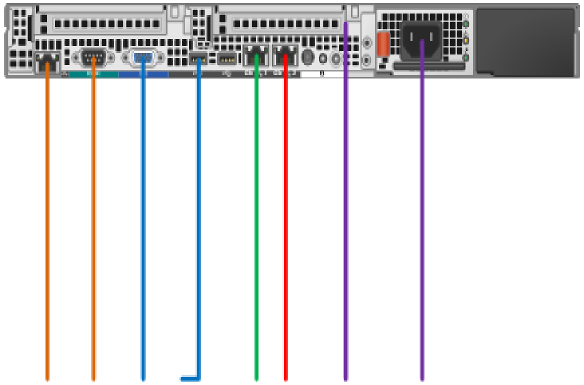
SearchStorage.com

DeepStorage.net

Our Agenda

- Today’s Virtual Server I/O problem
- Why Bandwidth alone isn’t enough
- Isn’t FCoE the Answer?
- PCI Virtual I/O
- Fibre Channel Virtualization
- Available Solutions
  - “Virtual NICs”
  - External I/O virtualization

The I/O Problem



Admins dedicate links to functions:

- User LAN
- Fibre Channel
- vCenter
- vMotion
- DMZ

- With redundancy, that’s 8 cables
- But still just 2 Gbps for user traffic
  - Some experts recommend 1 Gbps/core
- 1u and blades don’t have enough slots

The Storage Problem

- Admins want to access LUN directly
  - More than 2 TB VMDK limit
  - No copy for P2V
  - Allow fall back to physical server
  - Snapshots for Dev, etc.
- vMotion changes WWN, MAC

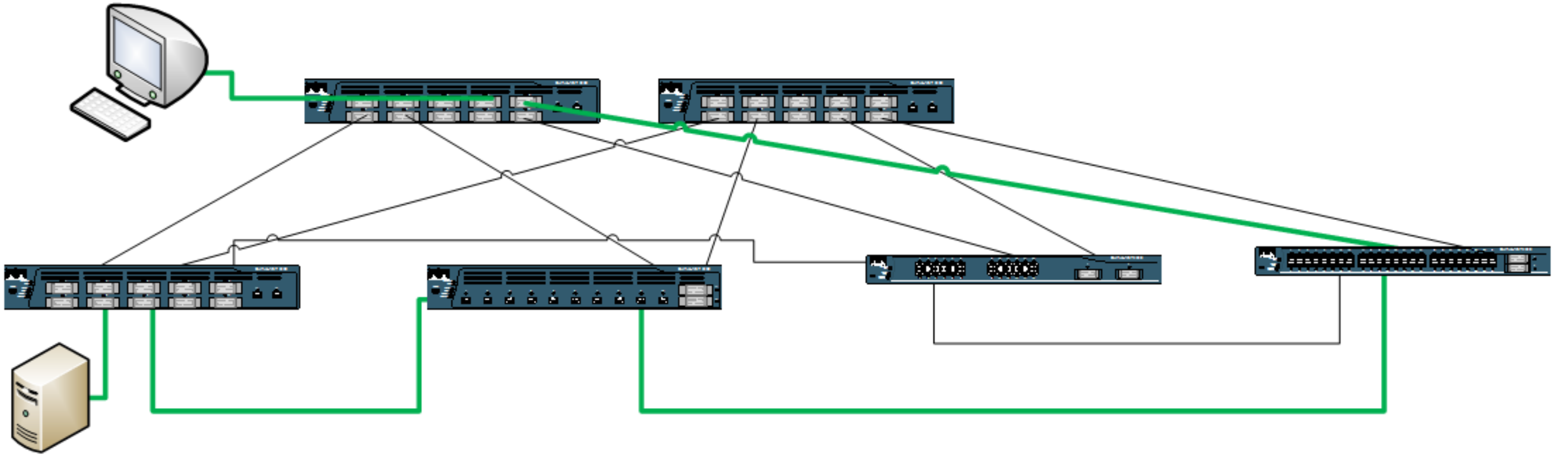
Enter NPIV

- One FC HBA can have multiple virtual N-port IDs
- NPIV moves with VM
- VMware support limited
  - No WWN visibility to guest

10 Gig Ethernet to the Rescue?

- 10 Gig provides bandwidth, but:
- We still need some isolation
  - vMotion could flood the net
  - Infected user machines could create a DDoS attack
  - VMware still wants separate nets

Then There’s Spanning Tree



- Enables just 1 path
- Takes seconds to converge on link state change
- Can choose stupid links
- Net admins end up wasting time to manage

Layer 2 Multipath

- Two proposed standards:
  - Shortest Path Bridging (SPB) (IEEE 802.1aq)
  - Transparent Interconnection of Lots of Links (TRILL) (IETF)
- Both use IS-IS routing protocol to learn topology
- Enable multiple paths
- Favor shorter paths
- Converge faster

FCoE?

- Embeds FCP in 10 GigE
- FC traffic has higher priority
- Requires Lossless DCB (CEE/DCE)
- FC part ready, DCB standards still under discussion
- FCoE alone still not enough isolation

Lossless Network?

- Really about congestion management
- Fibre Channel uses hop by hop buffer credits
- Ethernet relies on higher layer protocols (TCP)
    - But timeouts add latency
    - TCP throttles back

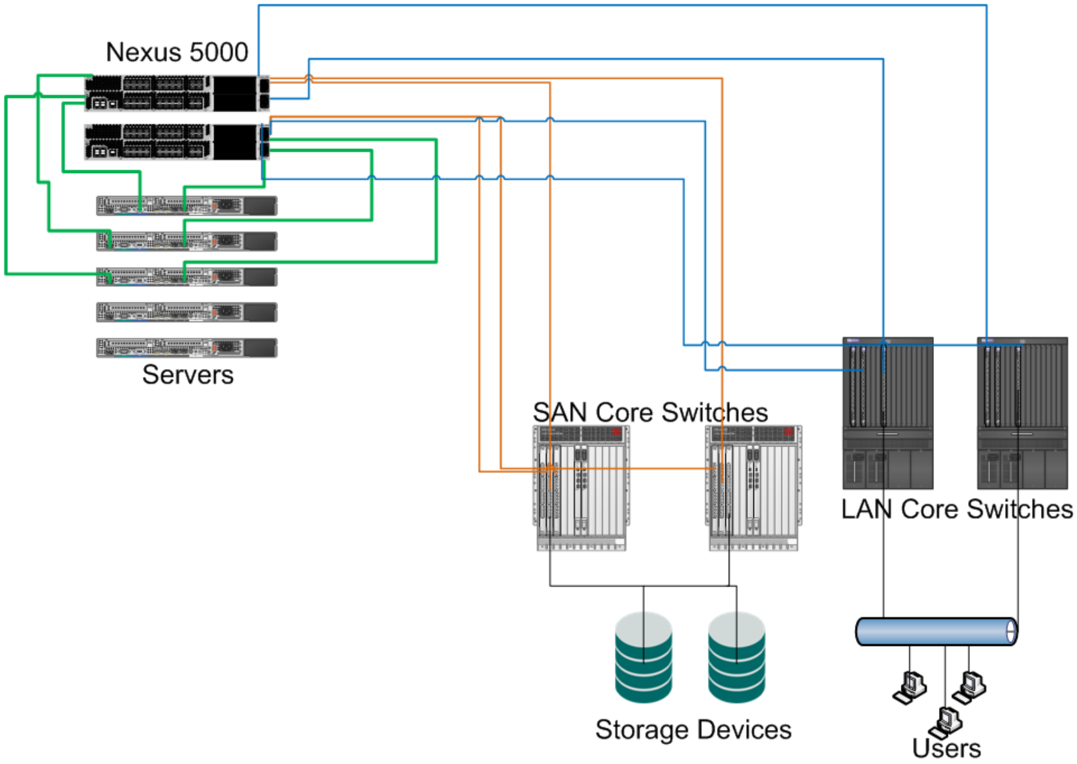
Data Center Bridging

- Extensions to make Ethernet "Lossless"
    - Due to congestion
- Per Priority Pause (802.1Qbb)
    - Flow control for each of 8 priorities
- Enhanced Transmission Selection (802.1Qaz)
    - Priority groups
- Data Center Bridging Exchange (802.1Qaz)
- Congestion Notification
    - Makes flow control end to end (optional)

FCoE Switches are Special

- Include Fibre Channel Forwarder
  - Implements FC by hop congestion control w/Pause
- Naming server

Top of Rack FCoE (The Cisco Model)



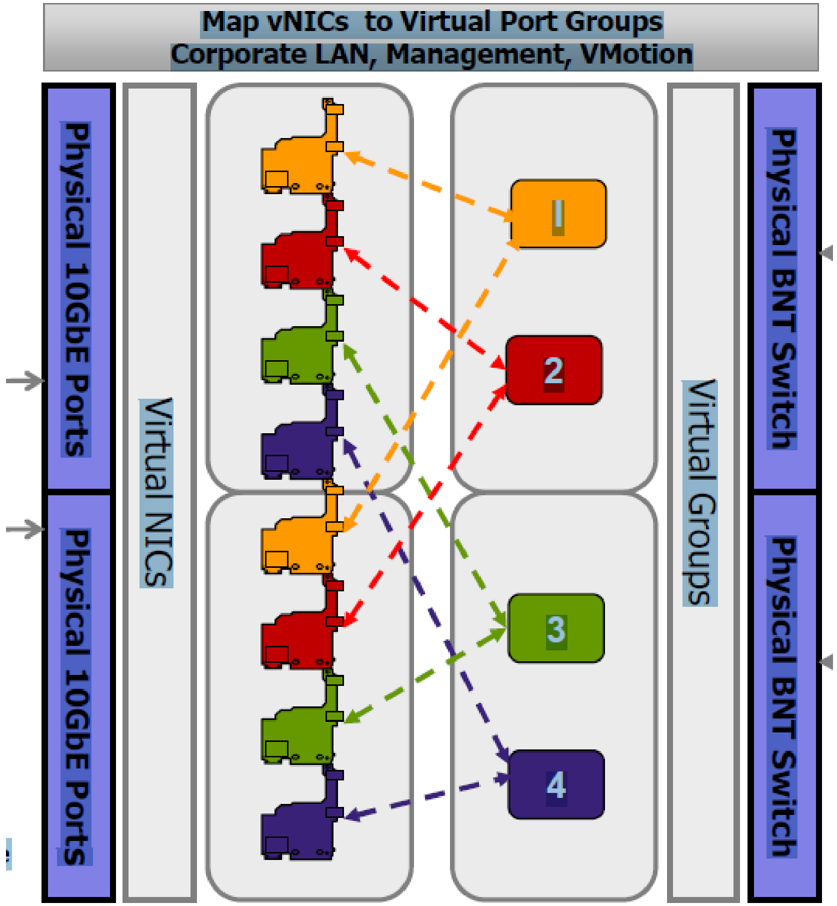
FCoE and DCB

- FCoE just needs Pause
    - We also recommend for iSCSI
- Converged networks need Qaz
- Multihop either needs FCF in every switch or end-to-end congestion control

FCoE Markets

- Brocade, Emulex and QLogic pushing CNAs
  - You need 20 yrs to develop FC
- Switch market opening
  - Cisco, Brocade
  - Blade Networks, HP w/QLogic
  - HP H3C
  - More coming
  - DCB and multi-hop issues

Virtual NIC

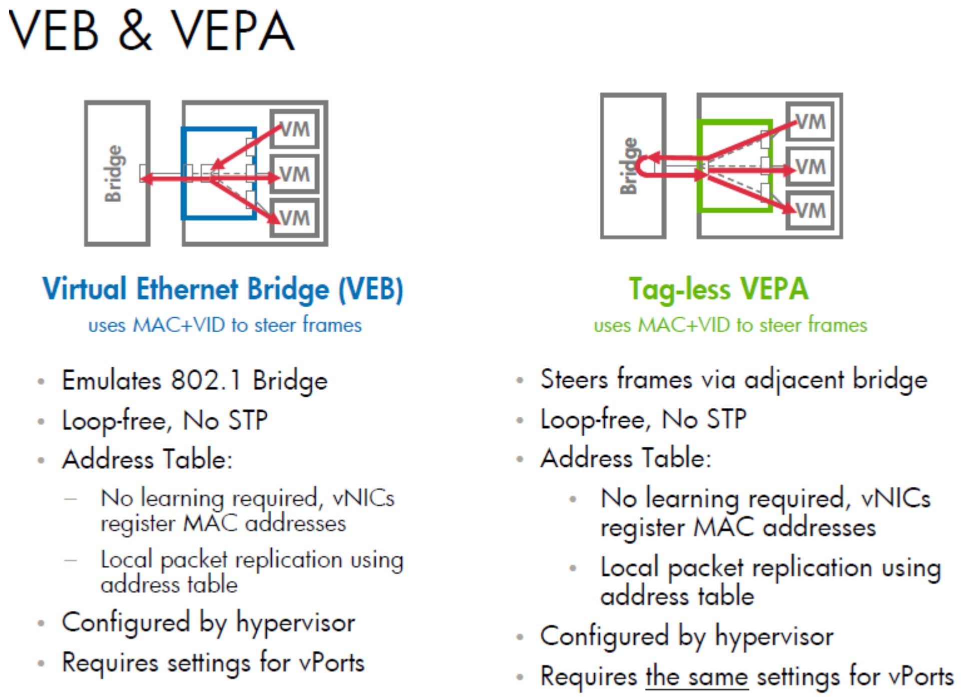


- One physical port multiple virtual NICs
- Each vNIC with MAC address, etc.
- vNIC/Switch manage bandwidth per vNIC
- Segregation via vLAN tags
- HP Flex-10, Neterion, QLogic, Emulex, etc.

PCI I/O Virtualization

- Standards from PCI SIG
- Cards share functions with processes
- SR-IOV
  - Multiple VMs on 1 PC (PCIe Root)
- MR-IOV
  - Across multiple systems
  - Also defines connectors, cables, etc.

Virtual Bridges



VEB & VEPA

Virtual Ethernet Bridge (VEB)

uses MAC+VID to steer frames

- Emulates 802.1 Bridge
- Loop-free, No STP
- Address Table:
  - No learning required, vNICs register MAC addresses
  - Local packet replication using address table
- Configured by hypervisor
- Requires settings for vPorts

Tag-less VEPA

uses MAC+VID to steer frames

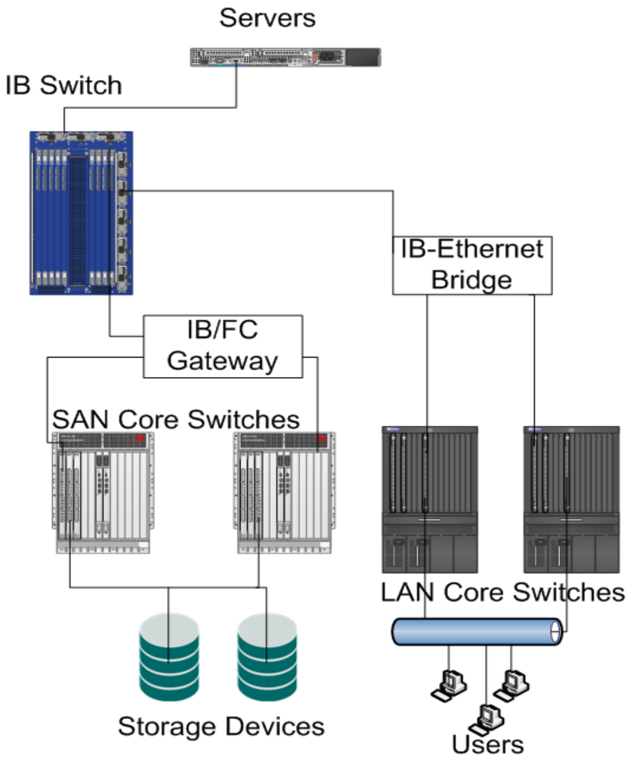
- Steers frames via adjacent bridge
- Loop-free, No STP
- Address Table:
  - No learning required, vNICs register MAC addresses
  - Local packet replication using address table
- Configured by hypervisor
- Requires the same settings for vPorts

VN-Tag

- Cisco’s approach
- Tags packets with virtual port info
- Upstream switch makes forwarding decisions

InfiniBand Gateways

- 40 Gbps, low latency connect to servers
- Gateways/Bridges to 10 GigE and FC
- Players:
  - Mellanox
  - Voltaire
- Frankly best for HPC
  - Already use IB



InfiniBand Gateways

Pros

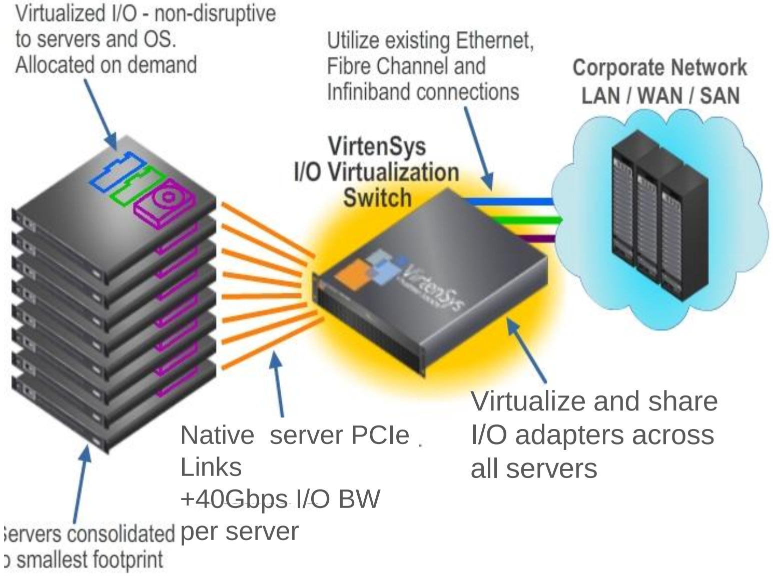
- High bandwidth, low latency for server-to-server traffic
- High port density switches for end of row

Cons

- Yet another network to run
- Drivers needed

Switched PCI

- Low cost PCIe extender in server
- Std cards in TOR switch
- Not just FC 10 Gig
- RAID controllers for shared DAS
- 16 servers/4 slots typical
- NextIO, Virtensys
- Aprius uses 10 GigE



Native server PCIe Links
+40Gbps I/O BW per server

Virtualize and share I/O adapters across all servers

Switched PCI

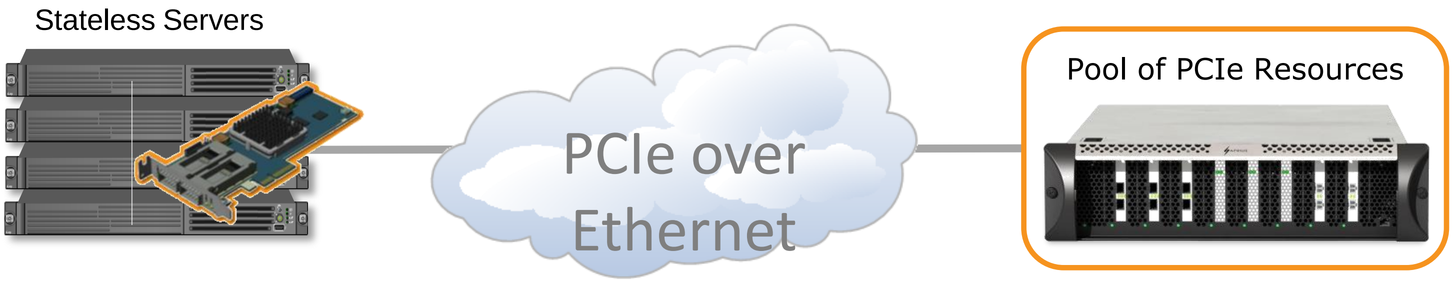
Pros

- Low cost per server
- Low power cable adapter in server
- Can share RAID
- Can share other I/O cards
- Uses standard drivers

Cons

- Low switch port count
- Low slot density

Aprius I/O over Ethernet Approach



NIC/CNA
FC HBA
iSCSI HBA
SAS/SATA
Flash SSD
GPU/NPU
Accelerators
Utility I/O

DA	SA	802.1Q Tag	Type=PCIeOE	PCIeOE Header	PCIe Data (TLP)	CRC32

Standard 802.1Q Frame

PCIeOE

Host Initiator Logic

- HW logic extends the host's PCIe topology and encapsulates PCIe over Ethernet

PCIe over Ethernet

- Low latency protocol for data traffic and resource discovery/management

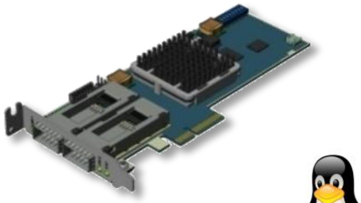
Virtualized Resource Pools

- Platform for shared I/O
- Software management of resources and hosts
- Any PCIe-based I/O

I/O Gateway Specs

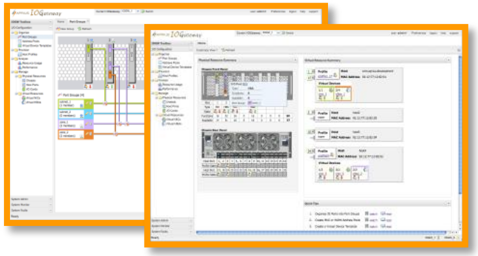
Aprius Host Initiator

- Multiple virtual PCIe slots per host
- No change to server software
- 2 x 10 GbE interface (QSFP)



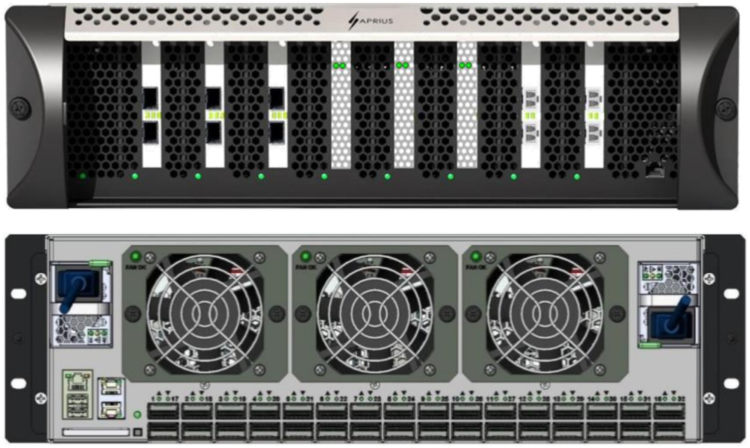
Management

- PCI manager & PF drivers
- CLI and Web-based GUI
- vSphere plug-in
- SNMP-based management



Shared PCIe I/O slots

- 8 full-height slots
- x8 PCIe 1.1 or x4 PCIe 2.0
- Accepts standard I/O adapters
- 480 Gbps low-latency fabric



High Availability

- Redundant power/cooling
- Hot-plug I/O slots
- Hot-plug cabling

PCIeOE Host Ports

- PCIe over 10 G Ethernet
- 32 servers @ 10 Gbps
- 16 servers @ 20 Gbps

Demonstrated Cards

- Intel 10G NIC, SR-IOV; E10G42BTDA ('82599')
- Exar/Neterion 10G NIC, SR-IOV; X3100
- QLogic FC HBA, 4 Gb, Dual port; QLE2562
- LSI MegaRAID SAS HBA, 6 Gb, 4 internal SAS ports, non-SR-IOV; 9260-8i

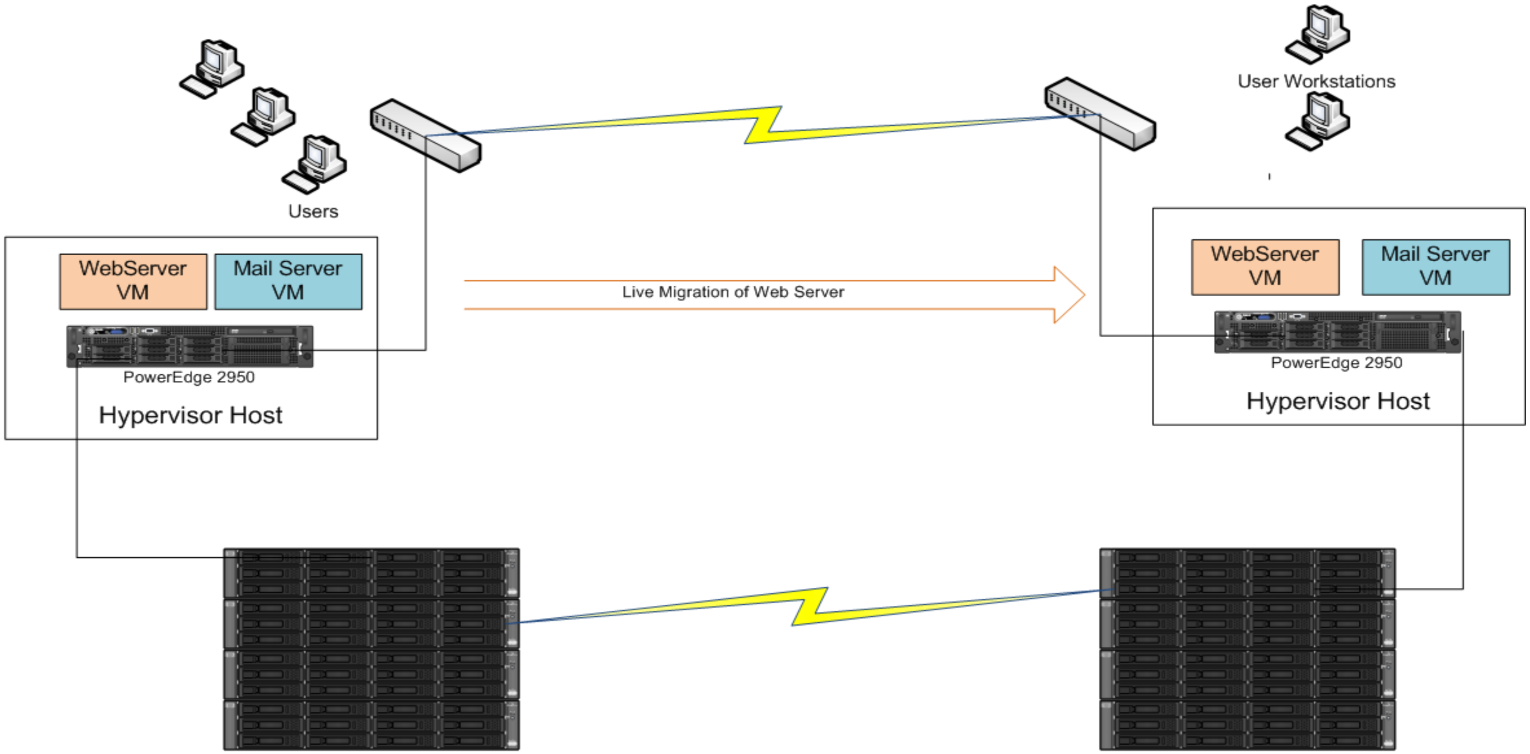
Xsigo’s Hybrid

- InfiniBand or 10 GigE in servers
- Switch w/PCIe slots
- Can add IB switch in between
- Standard drivers but limited choice of cards
  - 4 x 1 Gig Ethernet
  - 1 x 10 Gig Ethernet
  - 2 x 4 Gig FC (QLogic)

SearchStorage.com

WAN vMotion

- Disaster Avoidance
- Data Center Load Balancing



WAN vMotion Issues

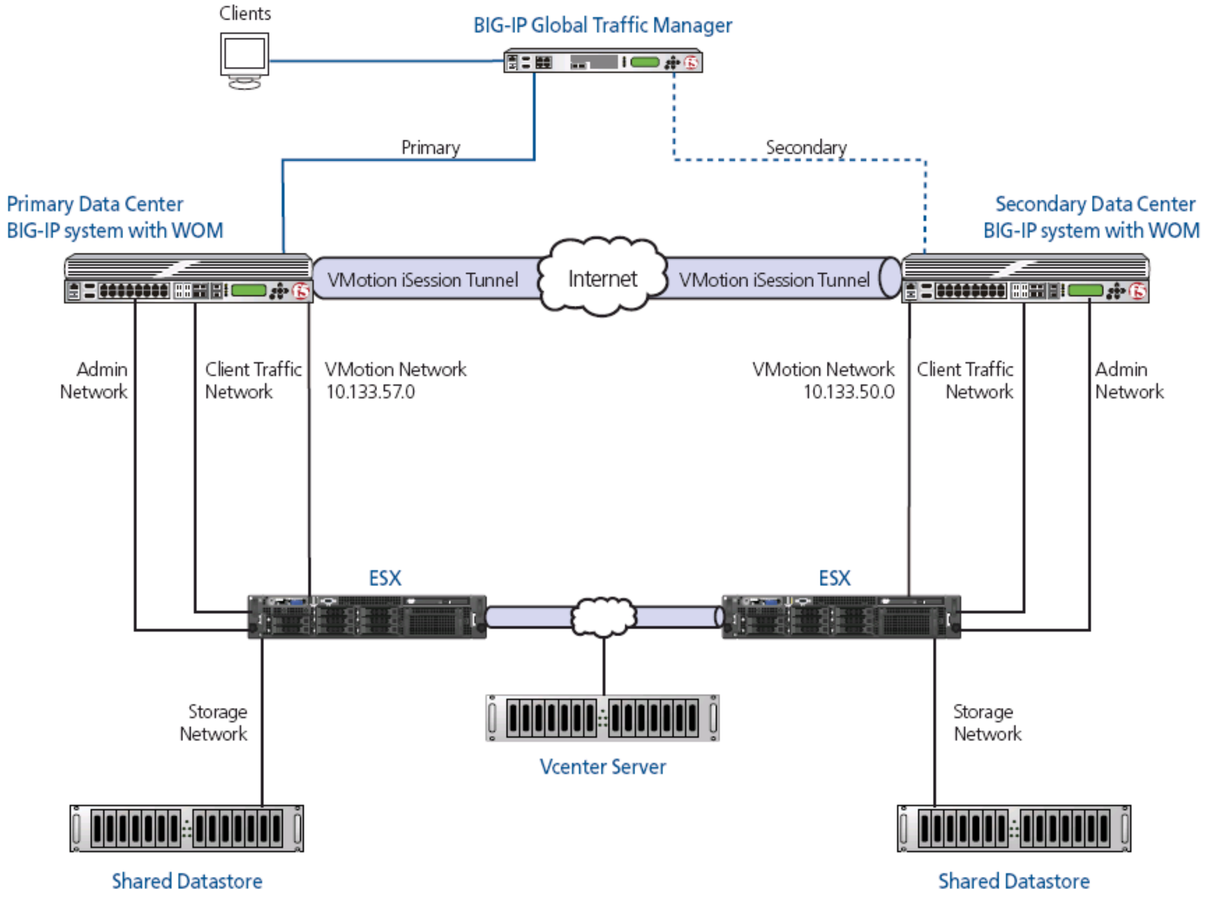
- Server IP moves with VM
  - Requires flat networks
- Speeding the transfer
  - WAN Acceleration
    - Riverbed, Cisco, NetEx, SilverPeak
- Storage Identity

Flat Net Multiple Locations

- Needs spanning tree filters, etc.
- May need L2 in L3 Tunnel
  - EtherIP
  - VPLS
  - Cisco OTV
- Load balancer too

SearchStorage.com

F5's Solution



Storage for WAN vMotion

- Storage vMotion too big
- Replicating anyway
- 2 Arrays, 1 LUN, 1 Identity
  - EMC Vplex
  - Compellent Live Volume
  - StorMagic

Choosing your solutions

- Politics are important
  - Who owns what
- FCoE
  - Large SAN management tools in place
  - Strong net team
- External Virtualization
  - Strong server team
  - High change rates

Thank you…

Howard Marks

Chief Scientist

hmarks@DeepStorage.net

SearchStorage.com

Some graphics courtesy Emulex, Virtensys, Xsigo, Mellanox, Voltaire