Hot Chips 21 (August 2009)

Innovation Envelope: Hot Chips in Blades

Kevin Leigh, Ph.D.
Distinguished Technologist
BladeSystem Architect Lead

Contributors

Collaborators

- Norm Jouppi (Ph.D., Director/Fellow, ExaScale Computing Labs)
- Partha Ranganathan (Ph.D., Distinguished Technologist, ExaScale Labs)
- Dwight Barron (Fellow, ISS CTO office)
- Dave Koenen (Network Architect, ISS)
- Chuck Hudson (Network Architect, ESS Blade)
- Paul Congdon (CTO/Fellow, ProCurve Network)

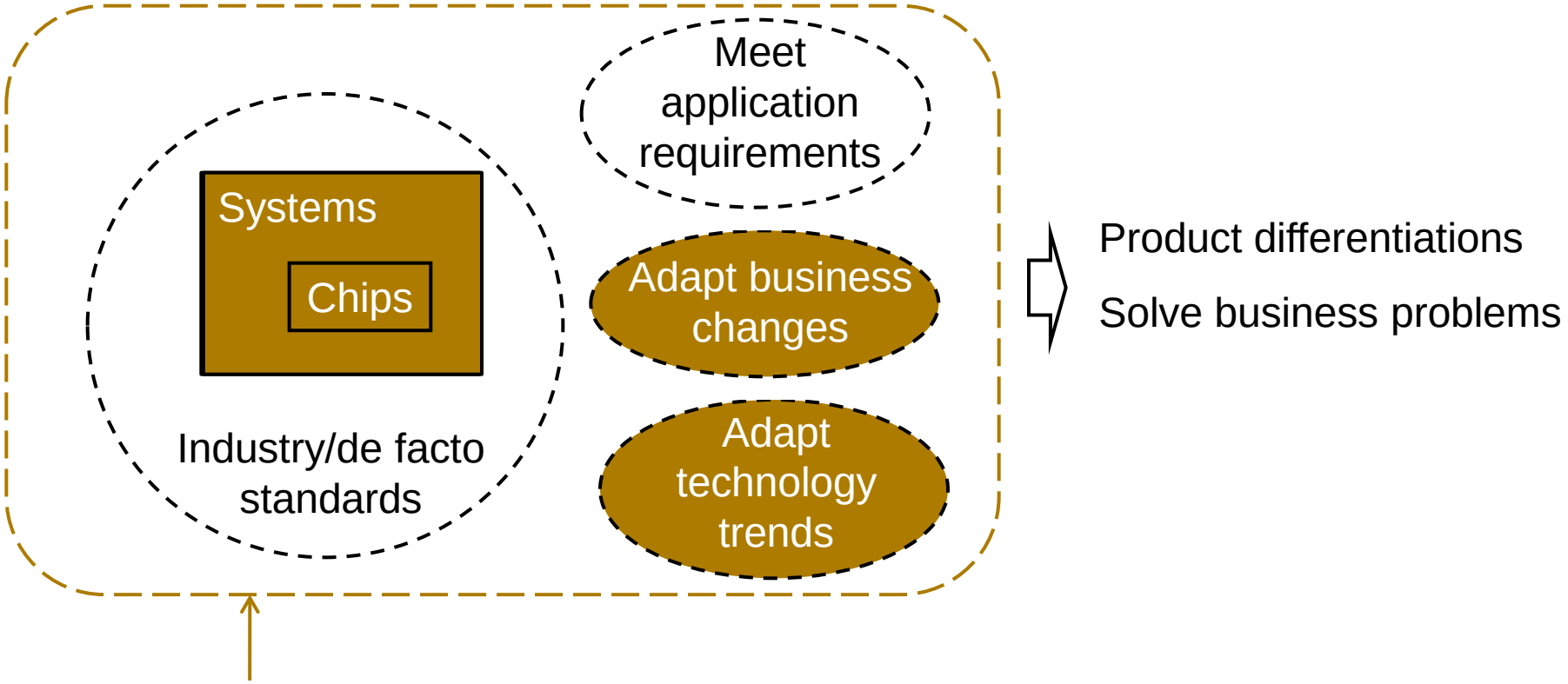
Reviewers

- Christos Kozyrakis (Ph.D., Assistant Professor, EE & CS, Stanford Univ.)
- Rob Elliott (Storage Architect, ISS Platform and Technology)
- Siamak Tavallaei (Distinguished Technologist, ISS)
- Mike Krause (Fellow, ISS CTO office)
- Gene Freeman team (ISS Platform and Technology)

Purpose of this talk

- To illustrate how synergy between system and chip innovations can lead to system and chip product differentiation features important to users
- Will use three chip innovation case studies in blade Ethernet networking
- Describe future chip innovation opportunities in blade architecture

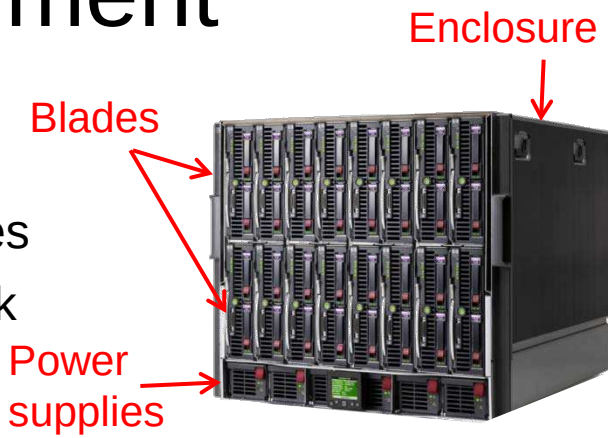
Synergistic innovation environment



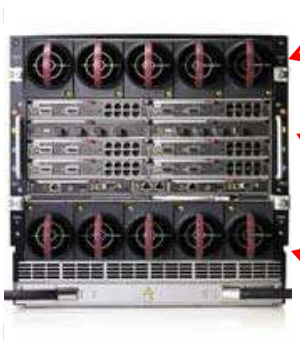
- Innovation envelope
  - foster system and chip innovations for multiple product generations
  - enable useful innovations to solve real-world problems
  - enable differentiations while complying necessary standards

Blade environment

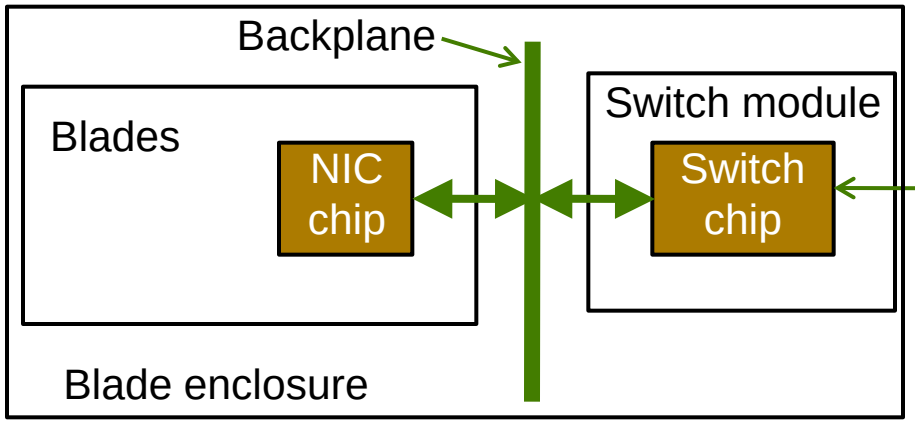
- Typically consists of
  - Server/storage/IO blades
  - 1st level or edge network switches
  - Backplanes
  - Power/cooling modules
  - Management module
  - Enclosure to house the above
- Blade designs vary in the way tradeoffs are made at design time for
  - Cost, scalability, flexibility, adaptability, …ity
- BladeSystem c-Class
  - Scalable architecture [1][2]



Front view



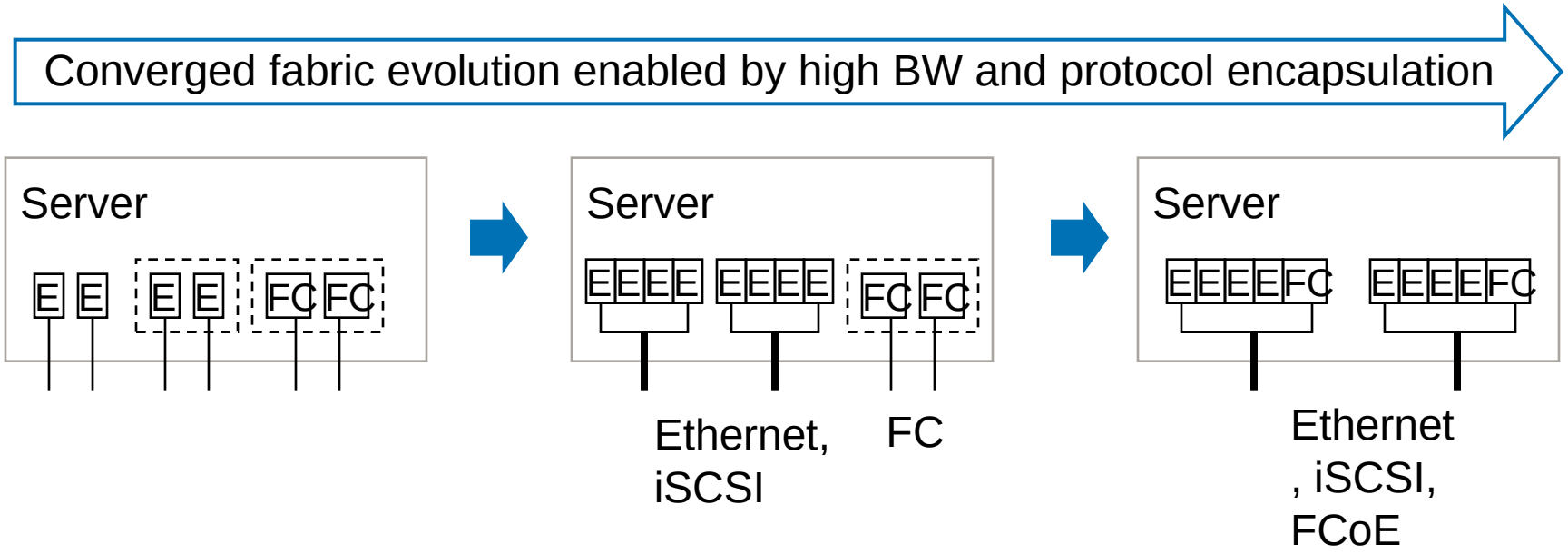
Rear view



Side view

Integrated nature within blade environment opens up more opportunities for innovations

Technology trends: Converging fabrics



- 10GbE BW makes sense to consolidate GbE & encapsulate other protocols
- Low-latency 10GbE fabrics further enable RDMA (RoCEE [9])

Case studies: Three innovations in series

Time

#1 For non-disruptive network connectivity

- Problems: Switch count explosion
  Server & network admin domains overlap
- Solution: **Virtual Connect**
- HP shipping products

#2 For more efficient use of network bandwidths

- Problem: Under-provisioned or under-utilized ports
- Solution: **Flex-10**
- HP shipping products

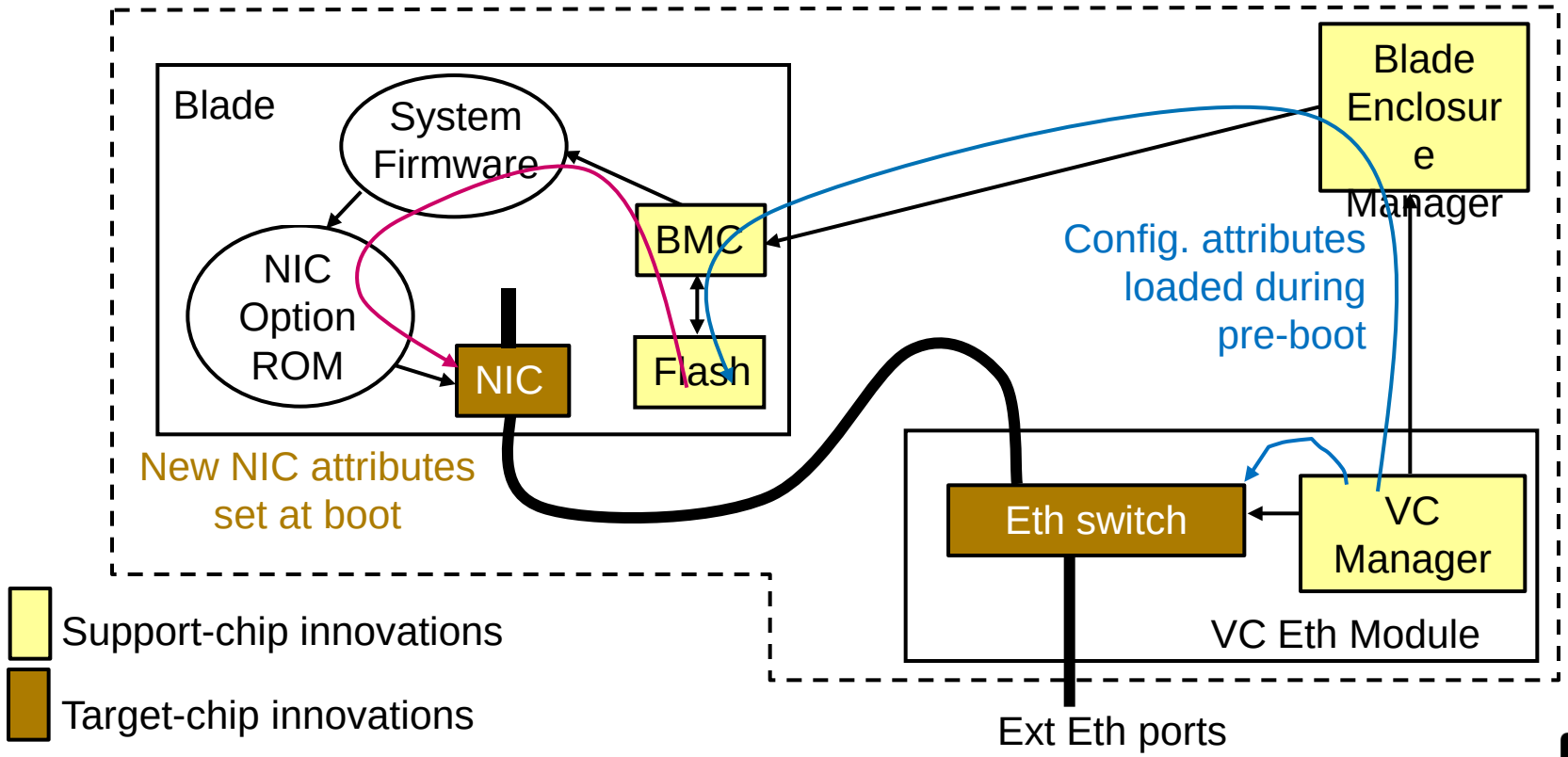
#3 For consistent datacenter-class networks

- Problem: "Internal" network traffic hidden from network administrators
  Inefficient to support rich networking functions
- Solution: **VEPA**
- Work in progress

Series of system and chip innovations

Pre-Boot Configuration Environment (PCE)

- A basic mechanism to enable Virtual Connect, Flex-10 & VEPA
  - An automated reprogramming mechanism of HW attributes [3]
  - Leveraged /amended industry standard methods (PCI Firmware 3.0 spec ECN [4], DMTF SM CLP [11])

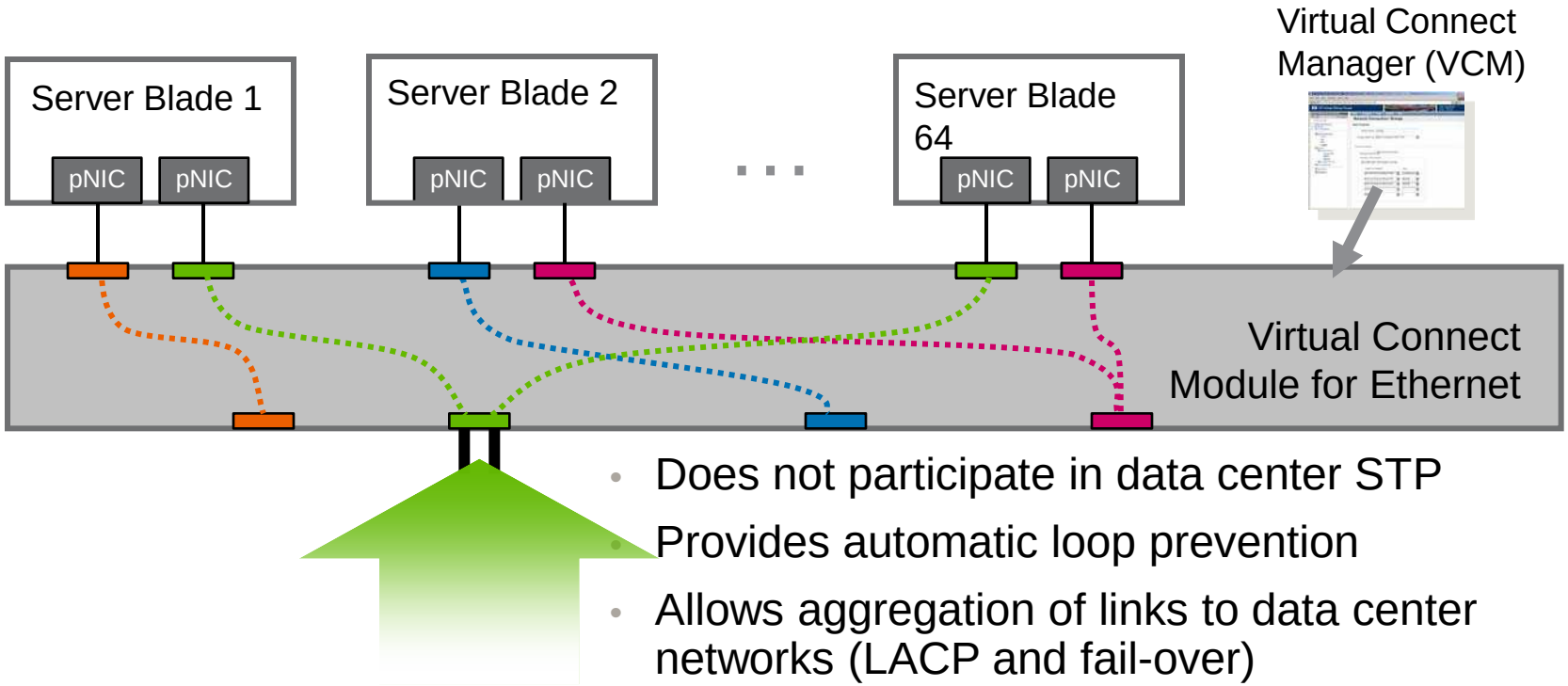


Innovation #1: Virtual Connect

- Problem for embedding switches in blade enclosures
  - network switch count explosion
  - server/network administration domain overlap
- Solution: Virtual Connect (VC) [5]
  - Make switches transparent to the network admins
  - Only NIC firmware (no NIC or switch chip hardware) changes
  - Changed enclosure and VC Ethernet module management firmware
- Features & benefits
  - PCE is transparent to device drivers, OS and applications
  - VC is transparent to core switches
  - Enables server administrators to manage VC modules
  - Migrating applications

How Virtual Connect works?

- Propagates blades' network physical addr to "external" ports
  - Reprogram network HW attributes via PCE (e.g., MAC addresses)
  - HW address flow through VC module (used to be "switch")

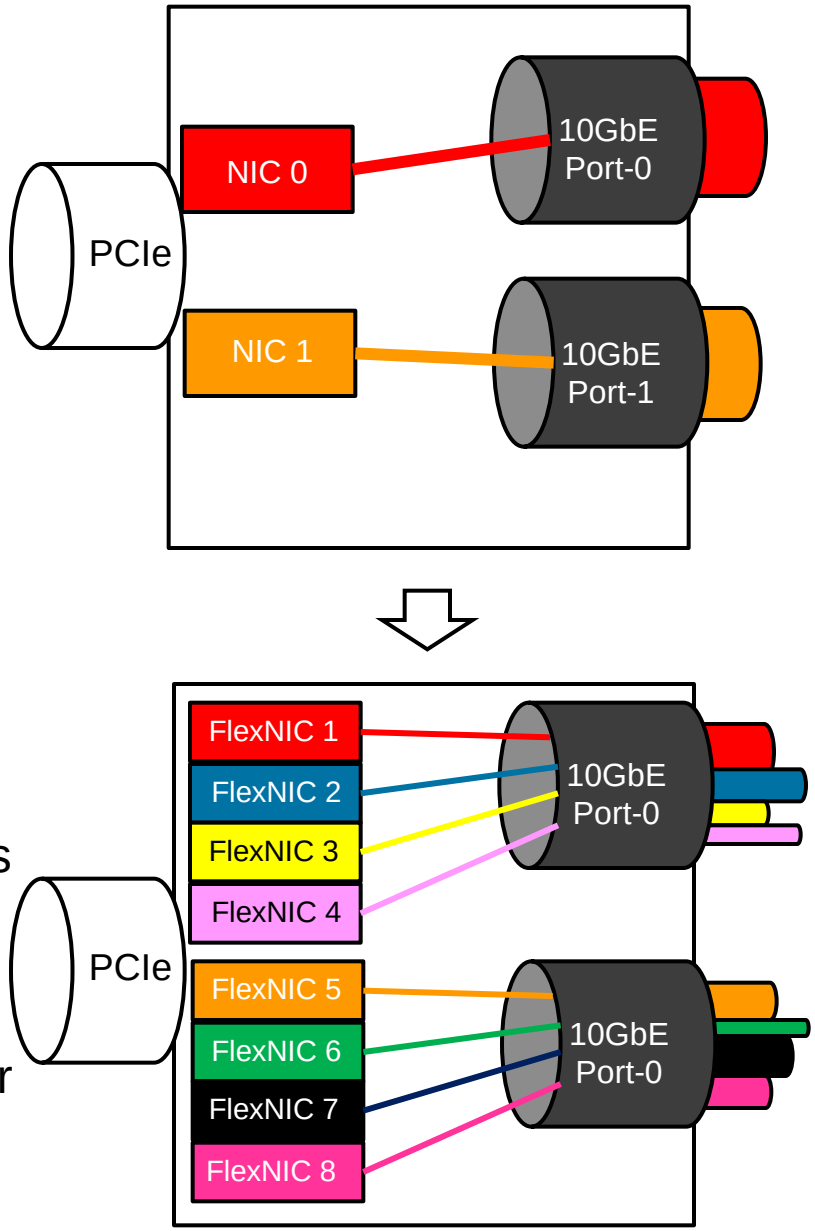


- Does not participate in data center STP
- Provides automatic loop prevention
- Allows aggregation of links to data center networks (LACP and fail-over)
- Supports VLAN tagging on egress or pass-through of VLAN tags
- Supports Link Layer Discovery Protocol (LLDP)

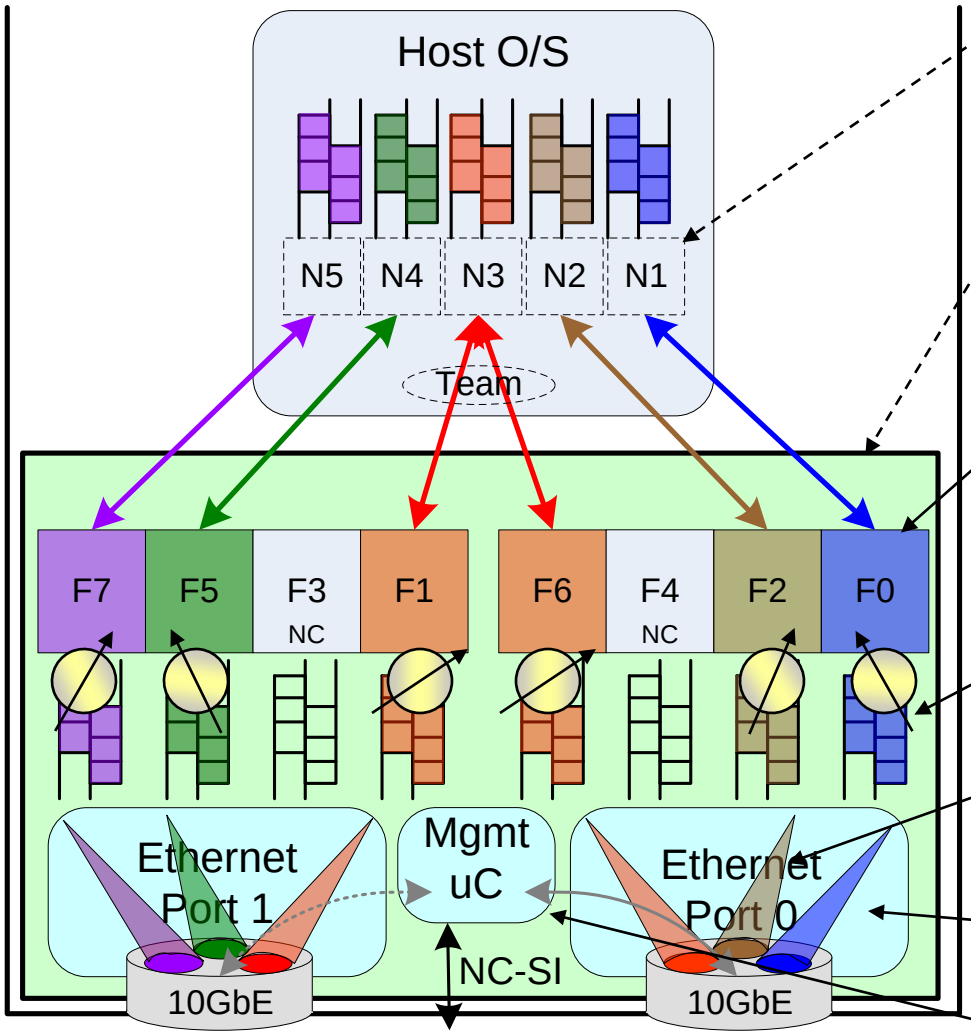
From outside of an enclosure,
VC-Enet uplinks look a lot like regular server connections

Innovation #2: Flex-10

- Problems
  - Inefficient use of bandwidth, volume space, power, etc.
  - Too many under-provisioned 1GbE NICs and switch ports
  - Under-utilized 10GbE NIC and switch ports
- Solution
  - Partition a 10GbE port into multiple logical ports with programmable bandwidth [6]
  - Relatively easy NIC chip hardware changes (Same package with ~20% more gates)
- Features & benefits
  - Essentially replaces multiple 1GbE NICs for VM apps
  - Transparent to the edge switches



How Flex-10 works (with Typical OS)



Separate NIC driver instances and easily add/remove

Pre-Boot Configuration of PCIe Functions by BIOS

FlexNIC parameters:
- Disable/enable
- Outer VLAN Tag ID
- Min/max B/W & QoS priority
- Function type (LAN, iSCSI)

Each Func has Q-set & Intrpt

Outer VLAN steering to Qs & insertion/removal of tags

QoS Priority Flow Control

In-band mgmt to NC-SI or uC

Chip modifications

NC-SI: Network Controller Side-Interface

How Flex-10 works (with VMM)

Direct DMA from NIC Rx Qs

VM physical NIC (pNIC):
- NetQueue & VMQ Support
- Shared Control & Setup
- Provides Addr Translation
- Software Switch for bcasts

Independent NIC Functions

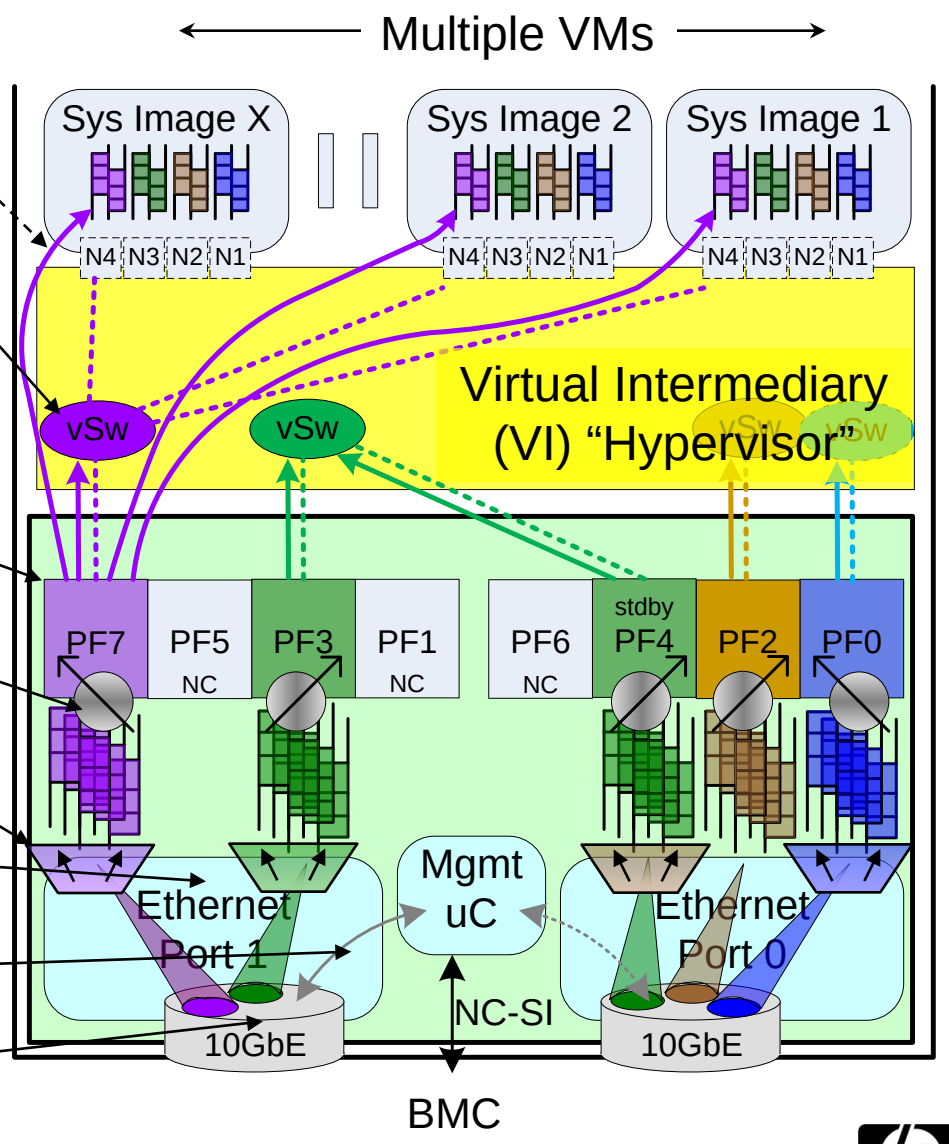
Egress B/W Control per PF

MAC Addr Steering to Rx Qs

Offload Engine per PF

In-Band Mgmt for NC-SI or uC

VLAN Steering to PF & Insertion/Removal



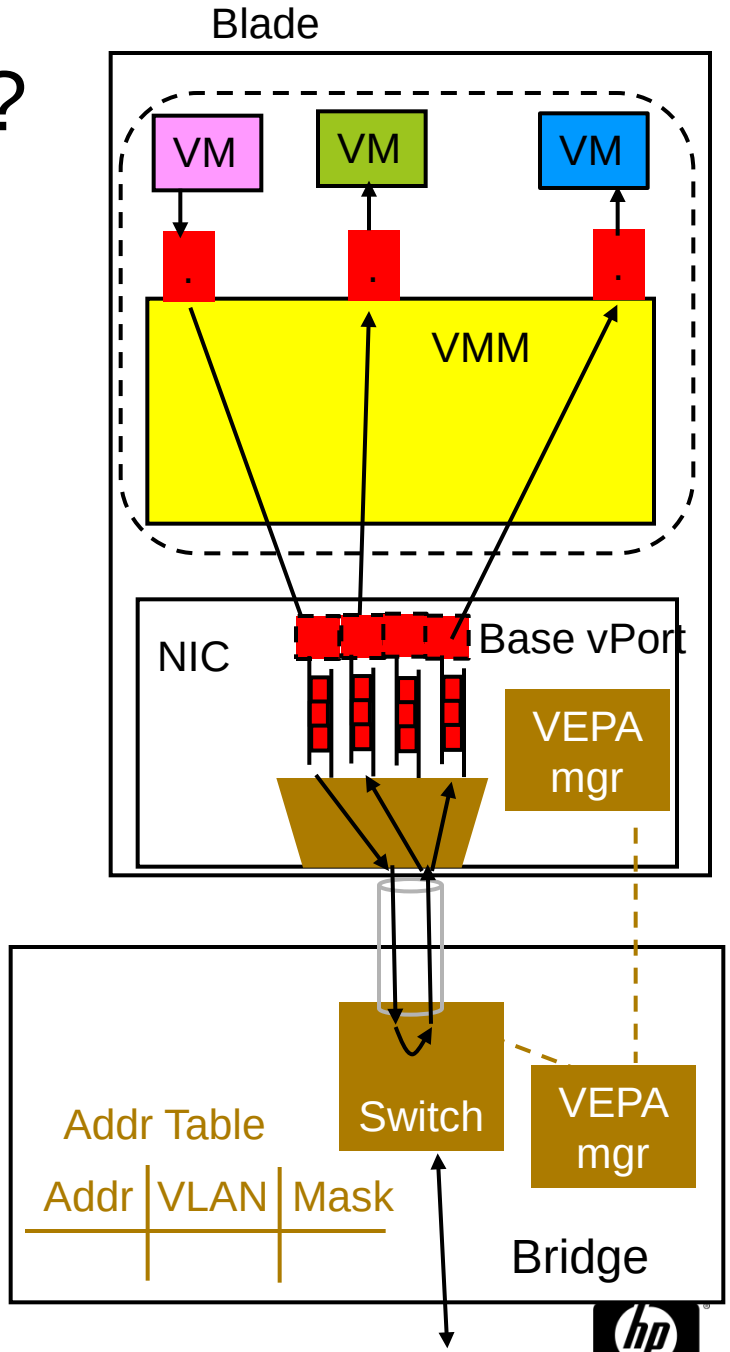
Chip modifications

Innovation #3: VEPA

- Problem
  - VMs communicate among them via vNICs and vSwitches within hypervisor
  - VM network traffic are not visible and managed by external switches
  - Users want more control on VM and PM network traffic
  - Too many advanced network functions in each server causing performance and management problems
- Solution
  - VEPA (Virtual Ethernet Port Aggregation) [7]
  - Enables hairpin forwarding on a per-port basis when a port aggregator is attached to a bridge port
- Features & benefits
  - Complimentary to PCI-SIG SR-IOV (Single-root I/O Virtualization) [10]
  - Transparent to Ethernet frame format and existing bridges
  - VM & PM network traffic are visible and managed at the network edge
  - VMs benefit rich edge switch features (ACLs, private VLANs, security)

How Tag-Less VEPA works?

- VEPA Ports (vPorts) as vNICs to VM
  - PCIe Virtual Functions
  - Typ. NIC features (TCP checksum, RSS, LSO)
- Bridge ports configured for VEPA attach for hairpin forwarding mode
- VEPA manager aggregates configs of vPorts
  - MAC addr, multi-cast addr, VLAN tags
- Invokes by special Bridge mode negotiation
- Sends all outbound traffic to the physical port
- Forward/multicast/broadcast traffic using the Address Table
- No local bridging like Virtual Eth Bridge



Target-chip innovations

Summary

- Blade environment is a catalyst for innovations
  - Designed blade infrastructure to phase-in generations of useful innovations
  - Turned commodity systems and components into better solutions
- Illustrated a series of system and chip innovations with 3 case studies
  - **Virtual Connect** ➔ Decouples server and network admin domains
    - Enable blade deployment with minimum disruption [5]
  - **VC Flex-10** ➔ Efficient bandwidth partitioning
    - Lower CapEx: Reduce network HW ≤75% & HW costs ≤66%[8]
    - Lower OpEx: Reduce power usage & costs ≤56% [8]
  - **VEPA** ➔ Expose VM network traffic to edge network
    - Efficient traffic management and processing in edge network
    - Work in progress
      - VEPA proposal to IEEE [7]
      - Multichannel & Remote Services proposal to IEEE [12]
      - Published patches for Linux and Xen [13][14][15][16]

Closing remarks

- Standards are important, but are not sufficient to differentiate products
- Common design for the mass promotes volume but prevent differentiation
- Important to Synergistically innovate chips within system and solution contexts, striking the right tradeoff balances
- In addition, innovation envelope should encompass OS and VMM
- Chip innovation opportunities in
  - Addressing proc/memory packaging, perf., power/cooling challenges
  - Addressing inefficient overheads for NICs, especially for small message sizes
  - Exploiting new memory hierarchy levels (e.g., using flash devices)
  - Dealing with signaling rates >10Gbps across PCB and other media channels
  - Exploiting higher bandwidths (e.g., 40GbE, 100GbE)
  - Enabling new fabric applications, e.g., PCIe for more than local I/O
  - Enabling new storage systems and sub-systems

References

[1] HP, “HP BladeSystem c-Class Architecture,” Technology Brief, 2006.

[2] Leigh et. al., “General-purpose blade infrastructure for configurable system architectures,” Distributed and Parallel Databases, Volume 21, Issue 2-3, pp. 115-144, June 2007.

[3] Leigh et. al., “Pre-Boot Configuration Environment,” HP internal document, 2006.

[4] “PCI Firmware 3.0 Specification” and Option ROM CLP ECN, 2006. http://www.pcisig.com/specifications/conventional/pci_firmware

[5] “Virtual Connect Specification,” HP internal documents, 2007-2008.

[6] “Virtual Connect Flex-10 Specification,” HP internal document, August 2008.

[7] C. Hudson & P. Congdon, “Tag-less Virtual Ethernet Port Aggregator (VEPA) Proposal,” January 2009. http://www.ieee802.org/1/files/public/docs2009/new-dcb-hudson-tagless-vepa-0109.pdf

[8] “Potential Savings with HP Virtual Connect Flex-10 for Consolidated Blade Networking,” HP and HP Channel Partner Internal Use document, November 2008.

[9] D. Cohen et. al., “Remote Direct Memory Access over the Converged Enhanced Ethernet Fabric: Evaluating the Options,” Hot Interconnects 17, 2009.

[10] PCI-SIG, “Single-Root I/O Virtualization and Sharing,” 1.0 specification, September 2007.

[11] DMTF, “Server Management Command Line Protocol (SM CLP) specification,” V1.0, June 2005.

[12] P. Congdon (HP), C. Hudson (HP) & M. Wadekar (Qlogic), “Edge Virtual Bridging Proposed PAR,” July 2009. http://www.ieee802.org/1/files/public/docs2009/new-evb-hudson-PAR-Discussion-0709-v01.pdf

[13] [PATCH][RFC] net/bridge: add basic VEPA support, 2009. http://lkml.org/lkml/2009/6/15/415

[14] [Xen-devel] [PATCH][RFC] net/bridge: Add basic VEPA support to Xen Dom0c, 2009. http://lists.xensource.com/archives/html/xen-devel/2009-06/msg01041.html

[15] [PATCH][RFC] bridge-utils: add basic VEPA support, 2009. http://lkml.org/lkml/2009/6/15/417

[16] [Xen-devel] [PATCH][RFC] tools: Add basic VEPA support, 2009. http://lists.xensource.com/archives/html/xen-devel/2009-06/msg01042.html

Q&A