High Performance IO – *What is it and can today's systems and applications really take advantage it?*

May 10, 2012
Tom Ambrose

Who am I?

- **Tom Ambrose**
- **Sr. Director of Engineering – Systems, Technology, & Architecture at Emulex Corporation**
- **B.S. ECE from Carnegie Mellon University**
- **15+ years in ASIC design/management**
- **5 years in current Architecture role**

Agenda

- Emulex Overview
- Adaptor Types
- Today’s adaptors: Features and Performance
- Architecture: Queues and QoS
- Topics for Investigation
- References

Emulex Corporate Overview

Corporate Facts

- Founded in 1978
- Based in Costa Mesa, CA
- Employees Approx. 960
- Strong Financials
- 2011 Revenue $496 MM
- FQ3:12 Revenue $125 MM
- FQ3:12 Net Cash $339 MM
- NYSE Symbol - ELX

Host Server Products

- Fibre Channel SAN
- Enhanced Ethernet Solutions
- Converged Networking
- Virtualization Infrastructure
- Connectivity Management
- Data Center Proven
- Remove Server Management
- Performance Analysis

Embedded Storage

- System-level embedded I/O
- Resilient high capacity disk solutions
- FC-SAS JBOD conversion preserves FC backend
- High throughput, solution oriented silicon
- Trusted abstraction layer software

Interface Card Types

- **HBAs**
  - *Host Bus Adapter* is usually used to describe FC, SAS, and SATA interface cards
- **NICs**
  - *Network Interface Controller* (NIC) is usually used to describe Ethernet LAN interface cards
- **HCA**
  - *Host Channel Adapter* (HCA) is usually used to describe Infiniband interface cards
- **CNAs**
  - *Converged Network Adapter* is usually used to describe Ethernet interface cards that support BOTH network and storage traffic

Today’s Adaptors - Features

- **PCIe gen3 x8 - 64Gbs to/from the host**
- **10G Ethernet moving to 40G**
  - Servers transitioning from 1G to 10G now, then to 40G, then to 100G
- **Fibre Channel 8G moving to 16G**
  - Storage Area Networks (SANs) for high performance
- **Many stateless offloads**
  - Checksums, IPv4/IPv6, LSO/LRO, RSS, HDS,, etc.
- **Multi-protocol stateful offloads**
  - FCoE, iSCSI, TCP, RDMA
  - Connections, sessions, logins, outstanding IOs, etc.
- **Side band management interface**
  - Configuration, inventory, management traffic pass-thru
- **Low power support**
  - PCIe low power/sleep state; Side band management still runs
  - Energy Efficient Ethernet

Today’s Adaptors – More Features

- **Data Integrity – T10PI, minimizing SDC from soft errors**
- **Enhanced Ethernet support**
  - Allowing lossless and lossy traffic classes to be define
  - Priority Flow Control, Congestion Notification, Quality of Service
  - Needed to meet FCoE requirements
- **Supporting server virtualization – many functions, many queues**
  - PCI-SIG Single Root IO Virtualization (SR-IOV)
    - Allowing Virtual Machines to have their own PCIe functions
  - IEEE Virtual Ethernet Bridging (VEB) and Virtual Ethernet Port Aggregation (VEPA)
    - Supports forwarding of traffic to/from VMs on the same host
    - Includes MC/BC replication, access control lists, promiscuous modes, etc.
- **Multiple levels of private networks**
  - Virtual LAN tags
  - Coke and Pepsi on the same physical network each with separate dept. networks (Marketing, Engineering, Finance)

Today’s Adaptors – High performance

- **1M+ IOs per second for storage protocols**

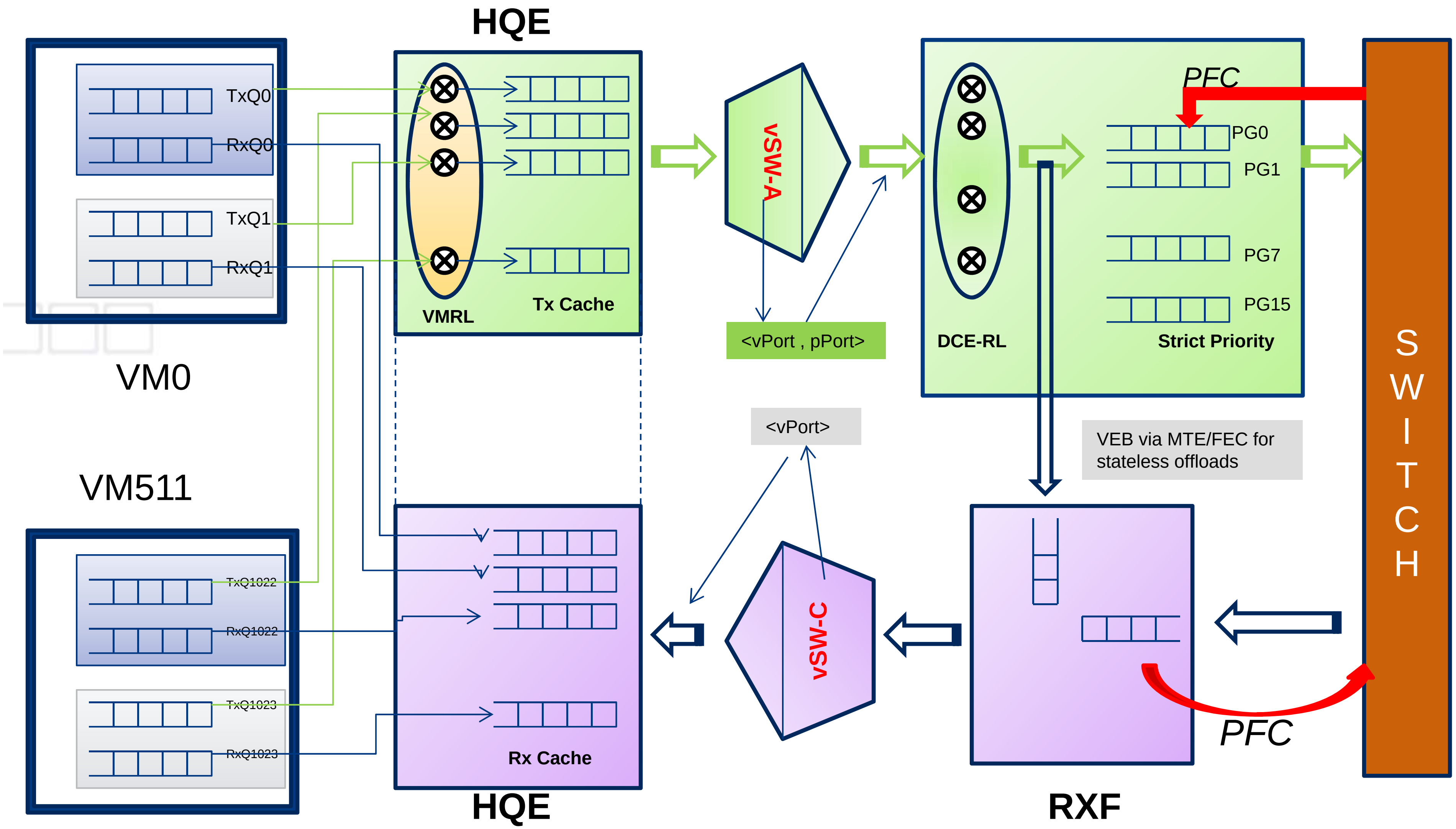
- **5M+ packets per second Ethernet**

- **Pushing IO latencies down**
  - NIC and RDMA to the low single digit uS range
  - Storage protocols to the single digit uS range

How do we do it?

- **Lots of logic gates**
- **Lots of RAMs and CAMs**
- **Lots of firmware**
- **Multiple processors per chip**
  - About a dozen!
  - 400-600MHz

QoS Servicing Overview



Investigation Topics – How can you help?

- **With today’s adaptors providing low IO latency and high throughput/bandwidth, Operating Systems and Applications need to scale to make use of it**
  - Balancing of host CPU cores between compute and IO
  - Optimizing IO paths through the application, driver, kernel, hypervisor, etc.
  - Design with IO performance in mind

Investigation Topics – How can you help?

- **With flash and other future storage technologies, the maximum response times are much lower than with rotating media**
  - Hierarchical storage vs. heterogeneous
  - Is the a way to predict when the storage will respond to an IO request allowing the adaptor pre-fetch the appropriate buffer entries (application, VM, etc.) to truly take advantage?
  - When working in a heterogeneous storage system, can IOs be scheduled to be speed matched and balanced to avoid head-of-line blocking, starvation, etc.?

Investigation Topics – How can you help?

- Adaptors can provide accurate (ns) time stamping of IOs than can be used for performance monitoring and tuning
  - The analysis and tuning is mostly a manual process today
  - Could this be used to for better/automated coordination of system tasks?

Networking and Storage- References

- **Industry Standards**
  - Fibre Channel. ANSI T10,T11,FCIA http://www.ansi.org, http://www.fibrechannel.org/, www.T10.org, www.T11.org
  - Ethernet: iSCSI, FCoE. IEEE. http://www.ieee.org
  - Infiniband: IBTA. http://www.infinibandta.org
- **Protocols**
  - SCSI. Small Computer Systems Interface.
  - FCP. Fibre Channel Protocol. http://www.t10.org/index.html
  - iSCSi. Internet SCSI. www.snia.org , http://en.wikipedia.org/wiki/ISCSI
  - FCoE. FC over Ethernet. T11. http://www.fibrechannel.org/, http://fcoe.com/, http://www.t11.org/fcoe
  - RoCE. RDMA over Converged Ethernet. http://www.ethernetalliance.org/

Questions?

Thank you!

Emulex
connects™
servers, storage and people
EMULEX CONNECTED

EMULEX®