# CrossBow:
# A vertically integrated QoS stack

**Sunay Tripathi, Nicolas Droux, Thirumalai Srinivasan,
Kais Belgaied, Venu Iyer**

**Aug 21st. 2009**

**Sigcomm WREN 2009, Barcelona**

**Sunay Tripathi,**
**Distinguished Engineer, Sun Microsystems Inc**
**Sunay.Tripathi@Sun.Com**

# Issues in Host based QoS solutions

- Performance
  - Additional Classification/Queuing for all packets
  - QoS layers typically high up in the stack (bulk of the work already done)
  - Packet needs to be DMA'd into the system before any policy can be applied
- QoS layers are typically bump in the stack
- Management complexities

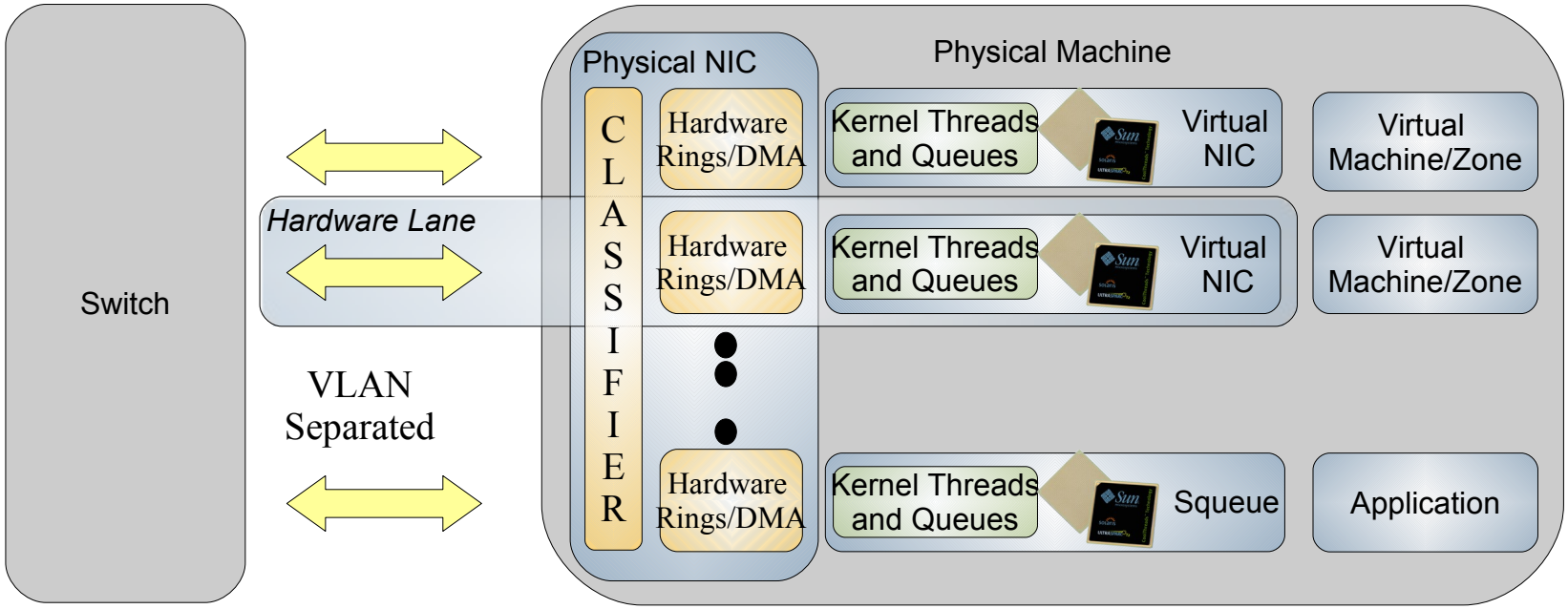
# Crossbow: Solaris Networking Stack

- 8 years of development work to achieve
  - Scalability across multi-core CPUs and multi-10gigE bandwidth
  - Virtualization, QoS, High Availibility designed in
  - Exploit advanced NIC features
- Key Enabler for
  - Server and Network Consolidation
  - Resource partitioning
  - Cloud computing

# Crossbow "Hardware Lanes"

Ground-Up Design for multi-core and multi-10GigE

- Linear Scalability using '*Hardware Lanes*' with dedicated resources
- Network Virtualization and QoS designed in the stack
- More Efficiency due to '*Dynamic Polling and Packet Chaining*'



# Hardware Lanes and Dynamic Polling

- Partition the NIC Hardware (Rx/Tx rings, DMA), kernel queues/threads, and CPU to allow creation of "Hardware Lane" which can be assigned to VNICs & Flows
- Use Dynamic Polling on Rx/Tx rings to schedule rate of packet arrival and transmission on a per lane bassis
- Effect of dynamic polling

***Mpstat (older driver)***

| intr | ithr | csw | icsw | migr | smtx | srw | syscl | usr | sys | wt | idl |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 10818 | 8607 | 4558 | 1547 | 161 | 1797 | 289 | 19112 | 17 | 69 | 0 | 12 |

***Mpstat (GLDv3 based driver)***

| intr | ithr | csw | icsw | migr | smtx | srw | syscl | usr | sys | wt | idl |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 2823 | 1489 | 875 | 151 | 93 | 261 | 1 | 19825 | 15 | 57 | 0 | 27 |

- Use Dynamic polling for B/W partitioning and isolation without any support from switches and routers

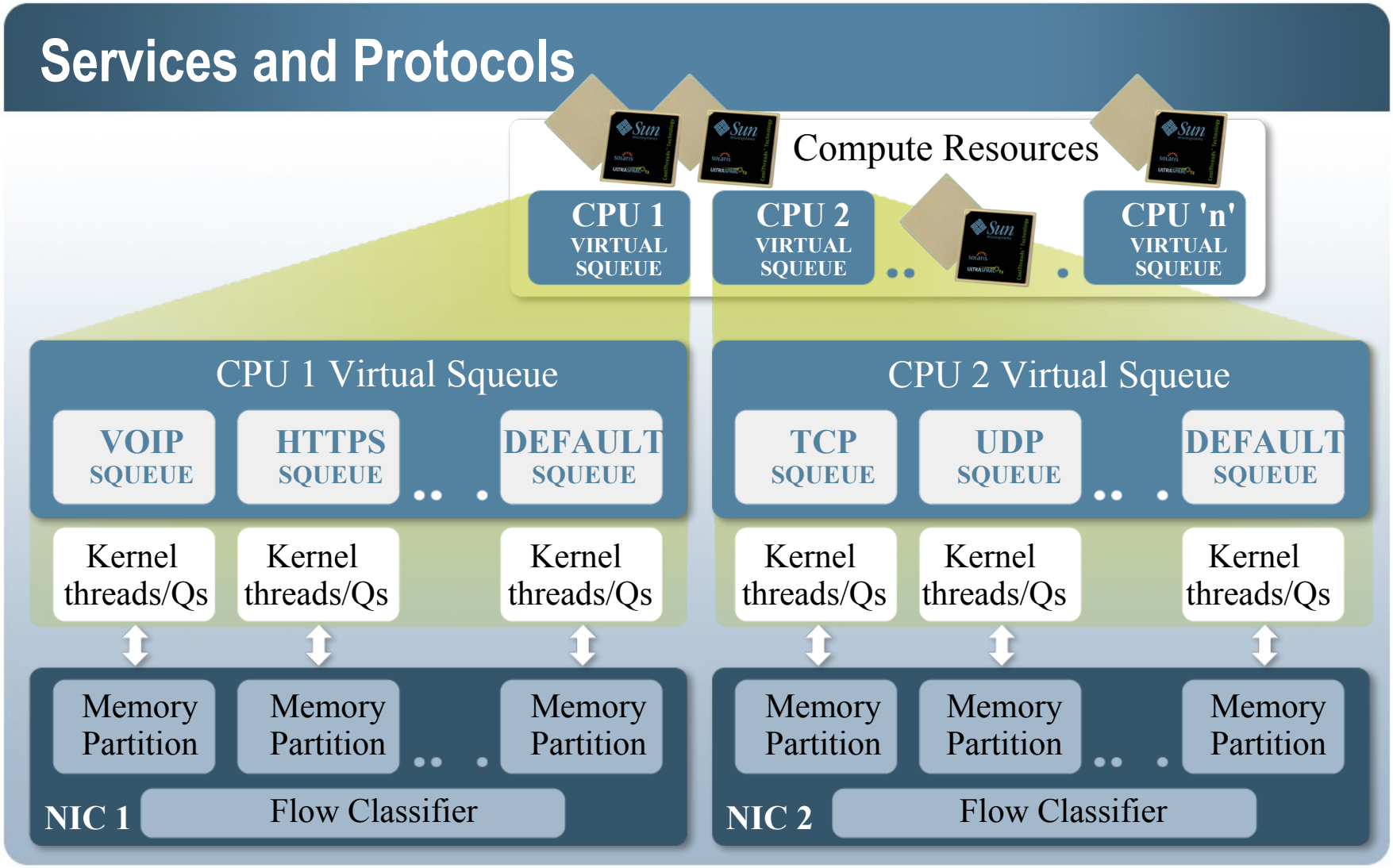
~75% Fewer Interrupts

~85% Fewer Context Switches

~85% Fewer Mutexes

~15% More CPU Free

# Crossbow *Flows*: Service Virtualization



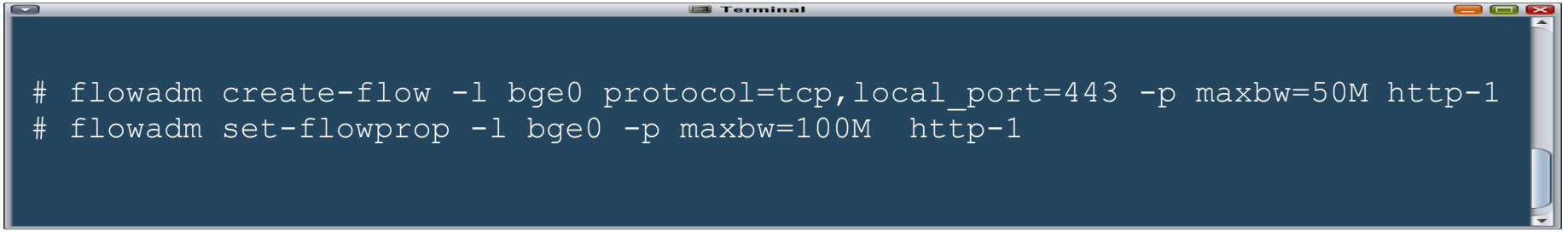
# Crossbow *Flows*

Crossbow *Flows* based on:

- Services (protocol + remote/local ports)
- Transport (TCP, UDP, SCTP, iSCSI, etc)
- Remote and local IP addresses
- Remote IP Subnets
- DSCP labels

Following attributes can be set on each *Flow*

- B/W limits
- Priorities
- CPUs

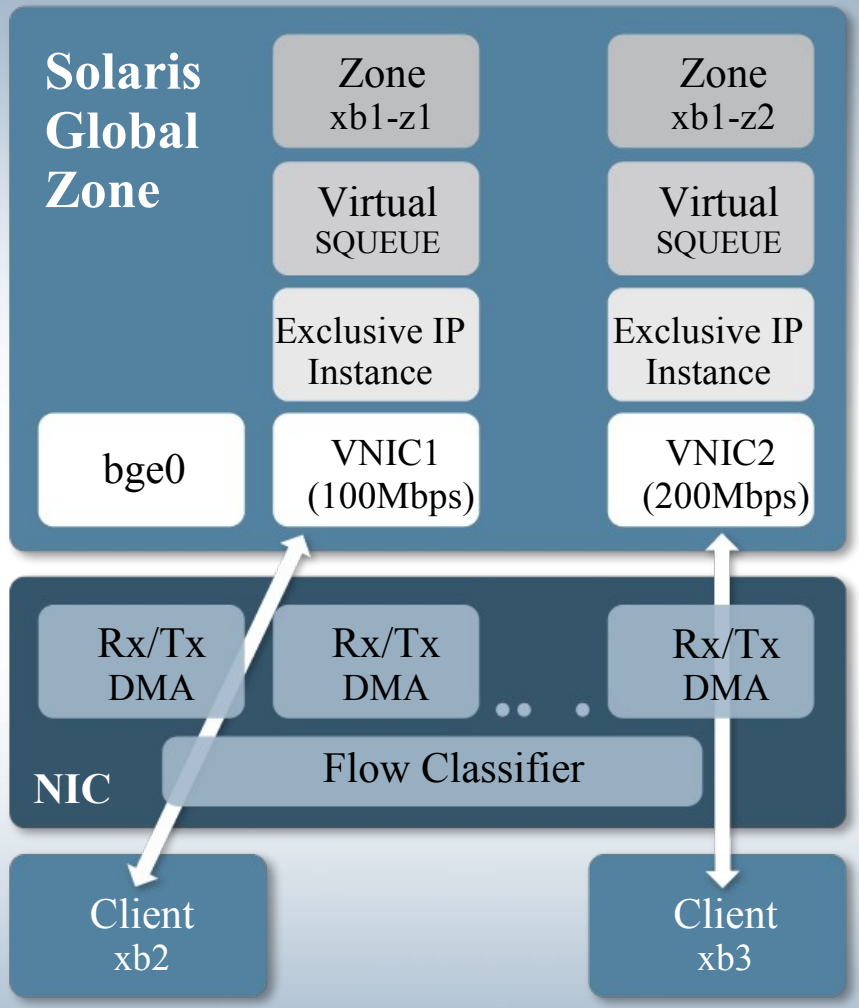


# Virtual Network Containers

## Virtualization
- Flows
- Virtual NICs & Virtual Switches
- Virtual Wire

## Resource Control
- Bandwidth Partitioning
- NIC H/W Partitioning
- CPUs/pri assignment

## Observability
- Real time usage for each Link/flow
- Finer grained stats per Link/flow
- History at no cost



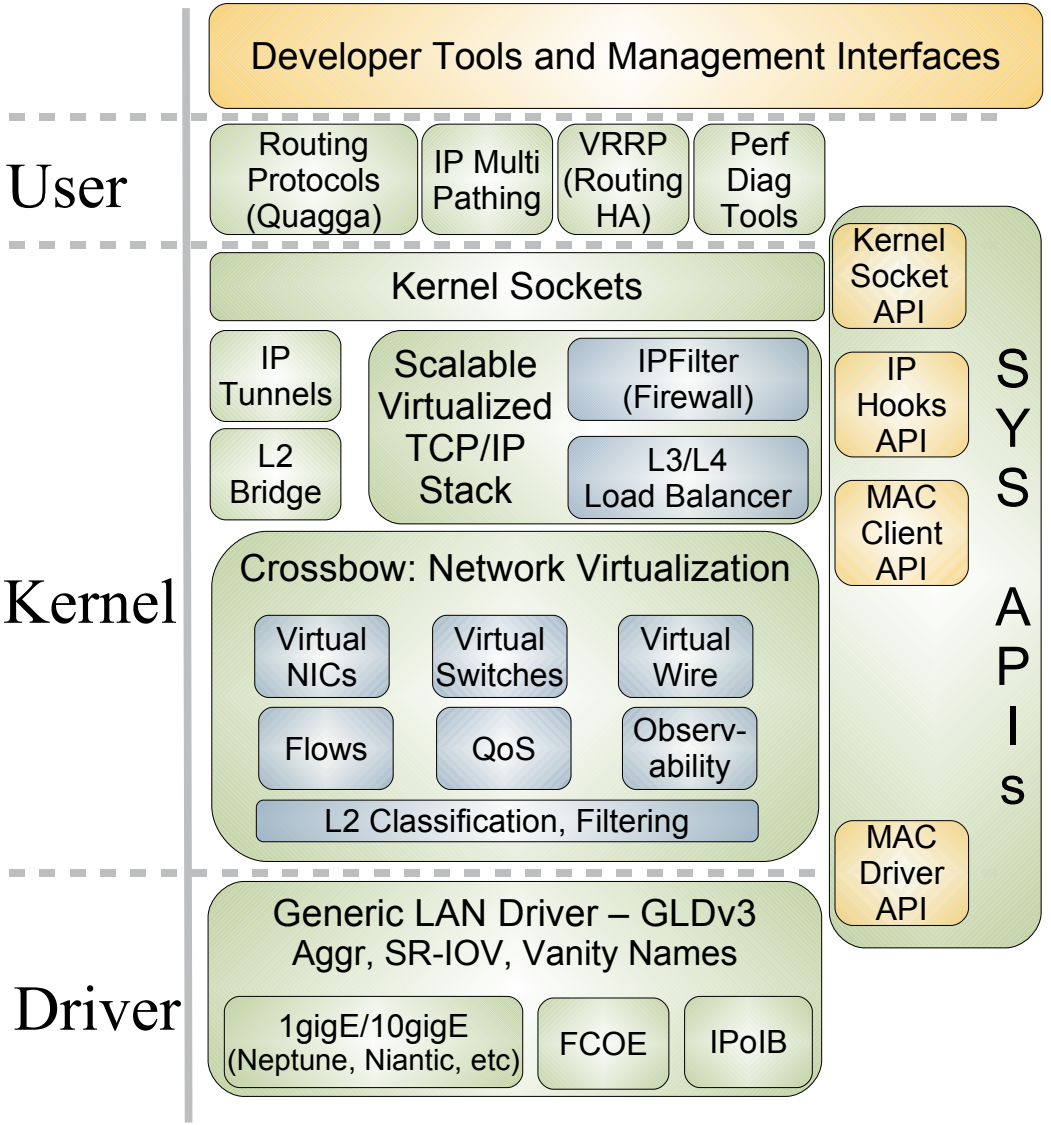
# Defense against DOS/DDOS

- DDOS have the ability to cripple entire server farms and all services offered by them
- Only the impacted services or virtual machine takes the hit instead of the entire grid
- Under attack, impacted services start all new connections under lower priority flow with limited bandwidth
- Connections transition to appropriate priority stacks after application authentication
- IDS systems can use Crossbow APIs to create '0' B/W flows based on remote IP addresses or subnets of the attackers and minimize their impact

# BACKUP

# Solaris Core Network Functionality

- Networking Services
  - Routing Protocols using Quagga
  - L3/L4 Load Balancer kernel modules
  - IP Firewall (IPFilter)
  - DNS, DHCP, NTP, SIP, VOIP, etc
- Scalable & Virtualized Network Stack
  - Kernel Socket & Socket Filter
  - Modernized TCP/IP Stack
  - QoS: B/W limits, Priorities, CPU bindings
  - IP Multi Pathing (IPMP)
  - IP Tunneling
  - Defense against DDoS attacks
- Crossbow: Virtual Networking
  - VNICs, VSwitches, VWire
  - Service Virtualization (Flows)
  - L2 Services: Classification, Filtering
- Generic LAN Driver v3 – GLDv3
  - Aggregation
  - Vanity Names
  - Drivers (1GbE and 10GbE, FCoE, IPoIB)



# Virtual NIC (VNIC) & Virtual Switches

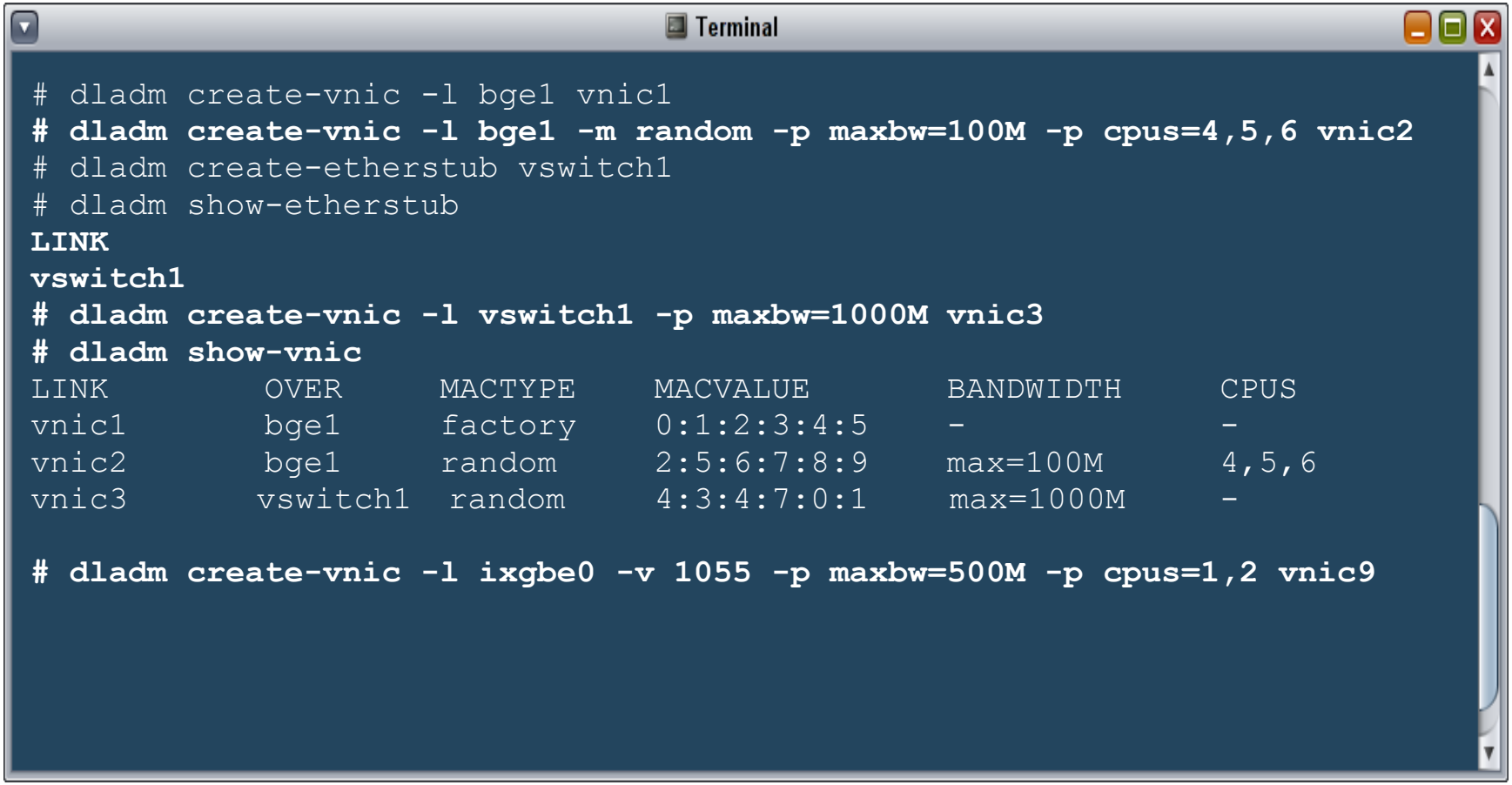
## Virtual NICs

- Functionally physical NICs:
  - IP address assigned statically or via DHCP and snooped individually
  - Appear in MIB as separate '*if*' with configured link speed shown as '*ifspeed*'
  - VNICs can be created over Link Aggregation on can be assigned to IPMP groups for load balancing and failover support
- VNICs Can have multiple hardware lanes assigned to them
- Can be created over physical NIC (without needing a Vswitch) to provide external connectivity with switching done in NIC H/W
- VNICs have configurable link speed, CPU and priority assignment
- Standards based End to End Network Virtualization
  - VLAN tags and Priority Flow Control (PFC) assigned to VNIC extend Hardware Lanes to Switch
- No configuration changes needed on switch to support virtualization

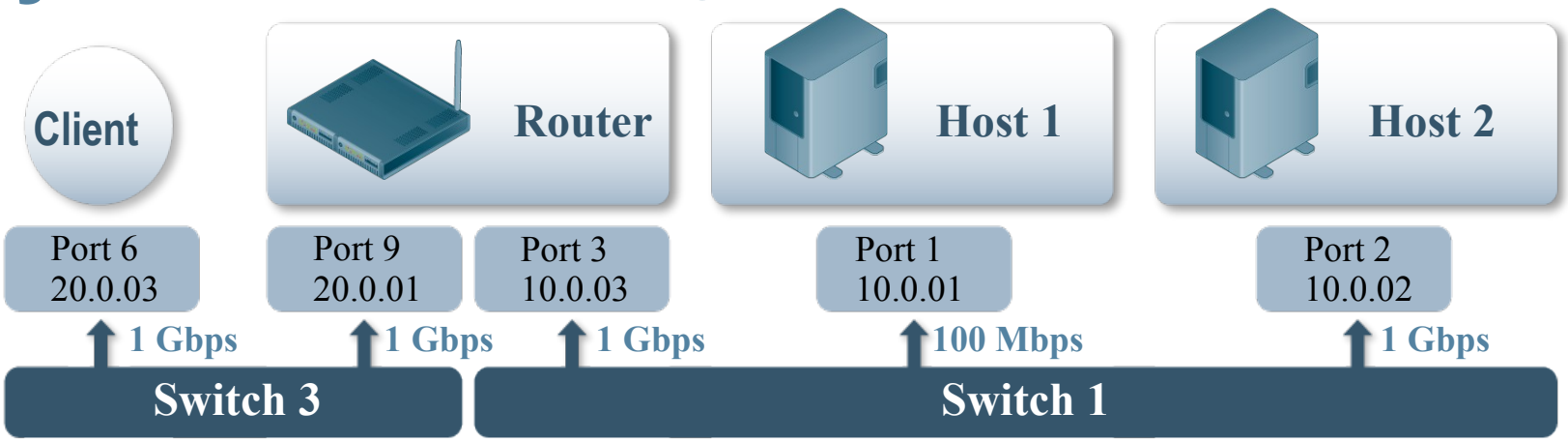
## Virtual Switches

- Can be created to provide private connectivity between Virtual Machines

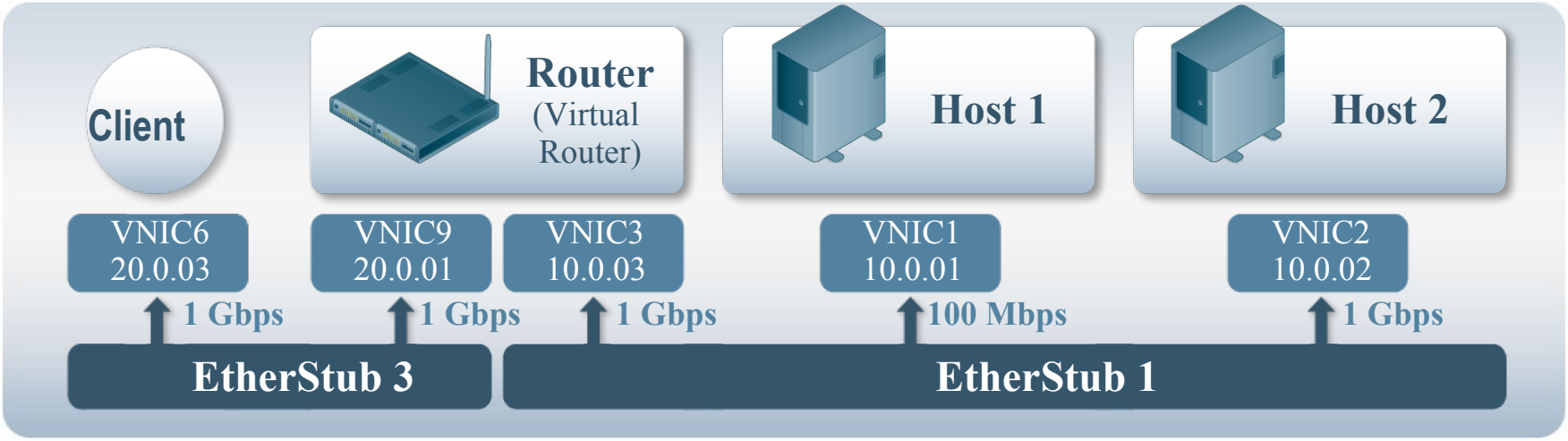
# Virtual NIC & Virtual Switch Usage



# Physical Wire w/Physical Machines

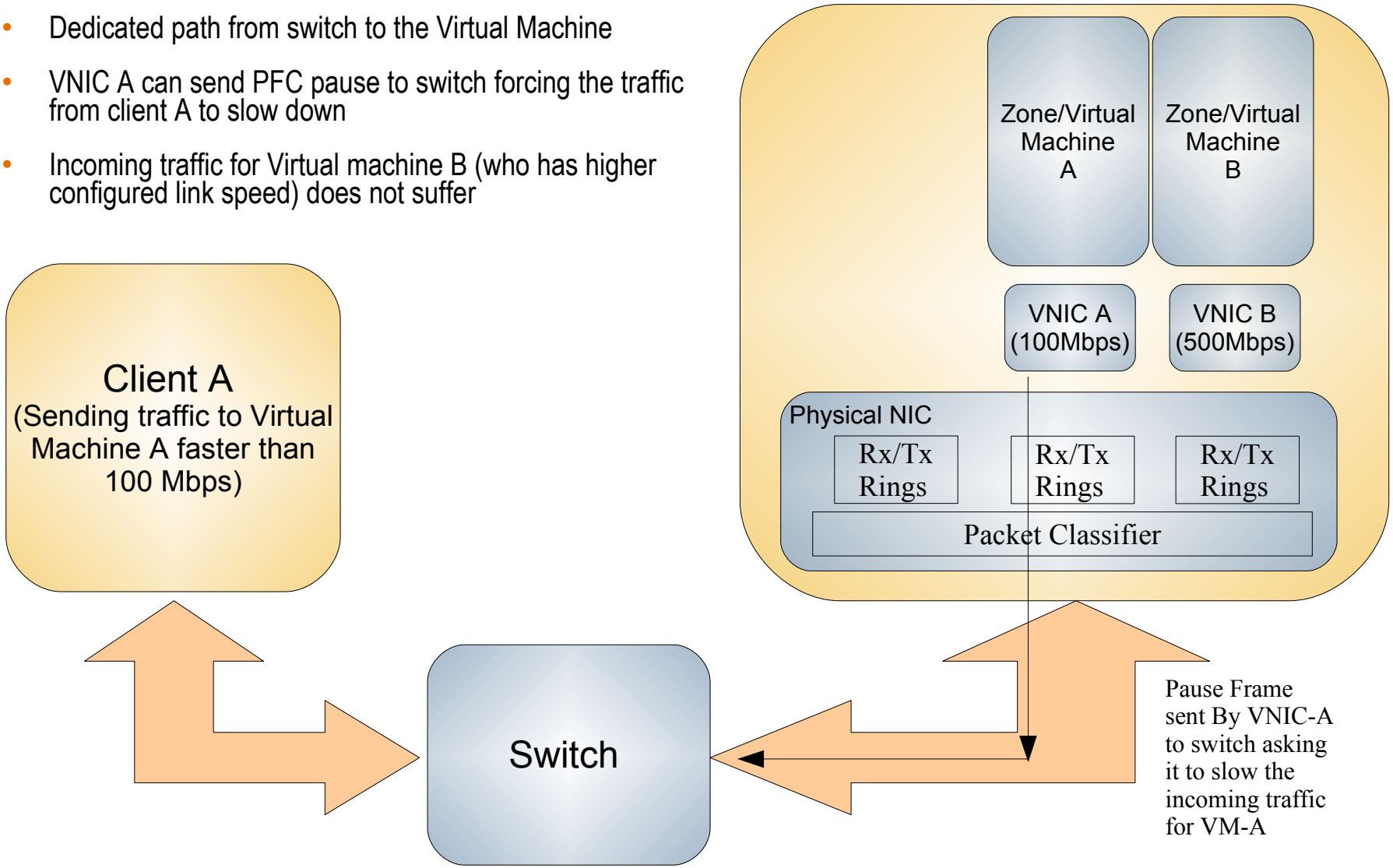



# Virtual Wire w/Virtual Network Machines

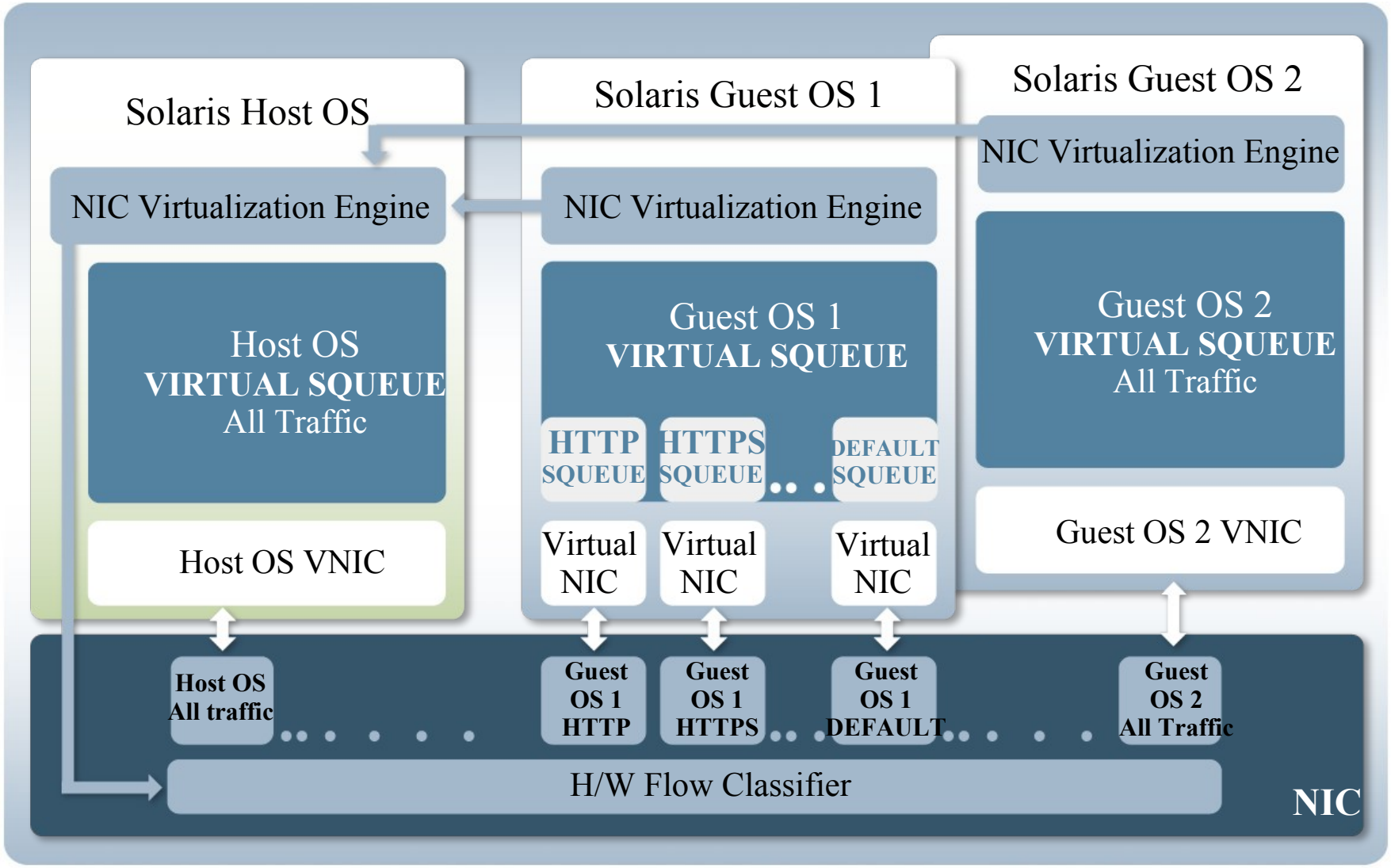


# Crossbow extends H/W Lanes to Switch

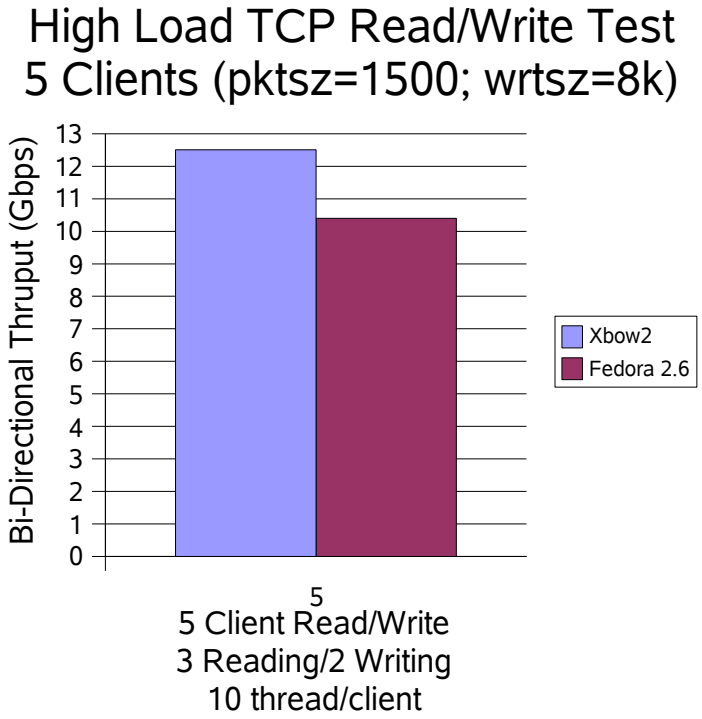
- Dedicated path from switch to the Virtual Machine
- VNIC A can send PFC pause to switch forcing the traffic from client A to slow down
- Incoming traffic for Virtual machine B (who has higher configured link speed) does not suffer



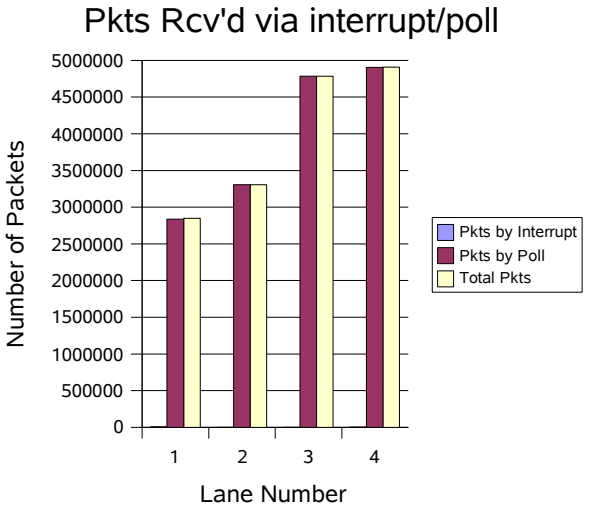
# Virtual Machines

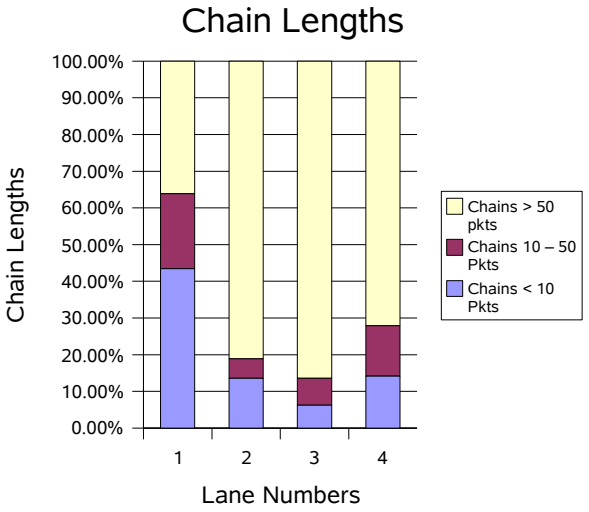


# Dynamic Polling: Effect on Throughput



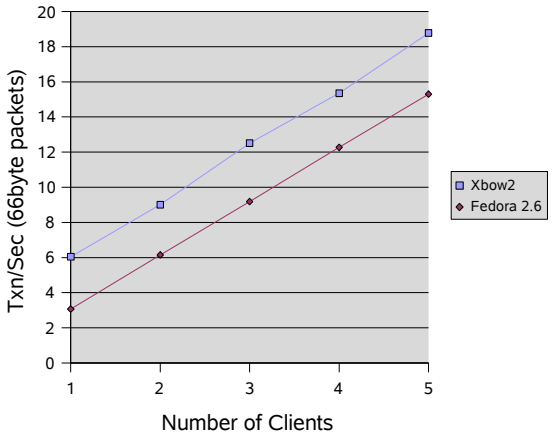

Config Details:
5 Client; 1 Server – 10GigE Links
3 Clients reading (10 thread each)
2 Clients writing (10 thread each)
All Client/Sever:
x4150 dual soc 8x2.8Ghz Intel CPU
10 GigE NIC – Intel Oplin (ixgbe)



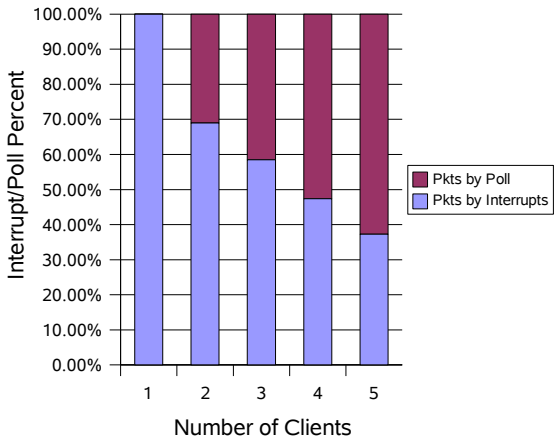



# Dynamic Polling: Effect on Latency

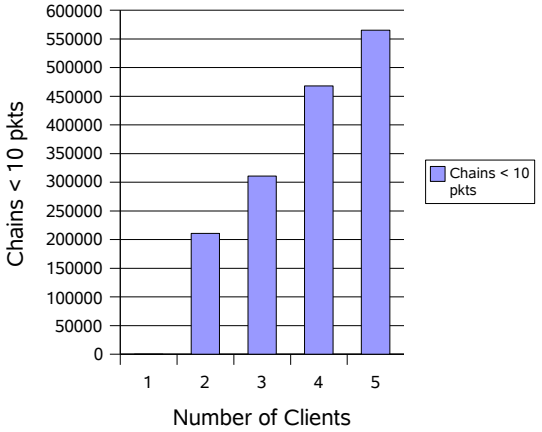
UDP 66byte pkt Low Load Latency Test



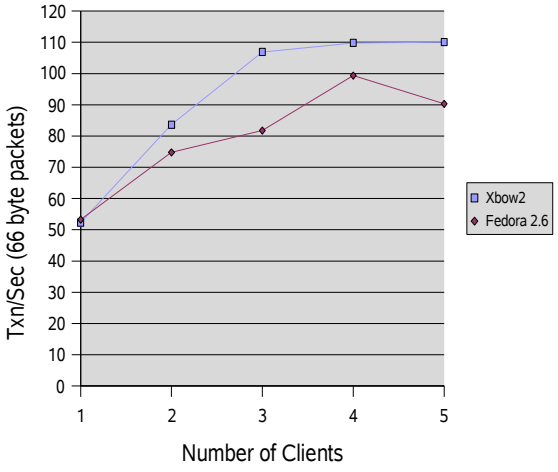

Pkts Received via Interrupt/Poll Ratio



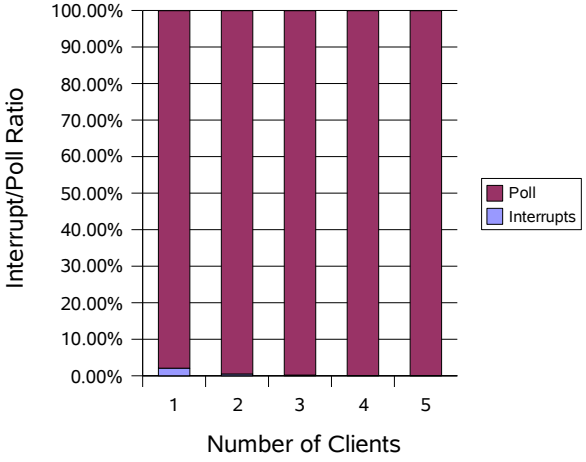

Number of Chains < 10 pkts



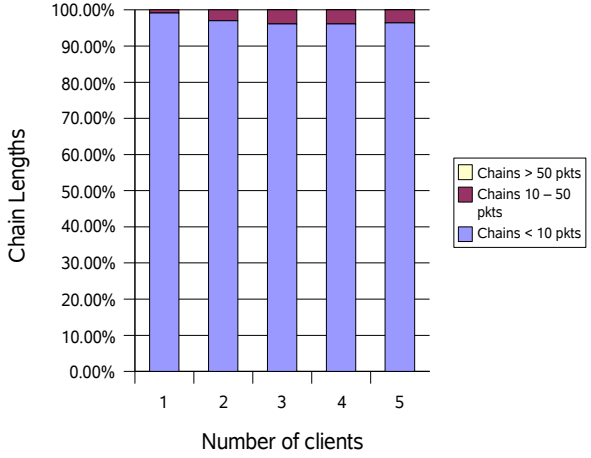

UDP 66byte pkt High Load Latency Test




Pkts Received via Interrupt/Poll Ratio



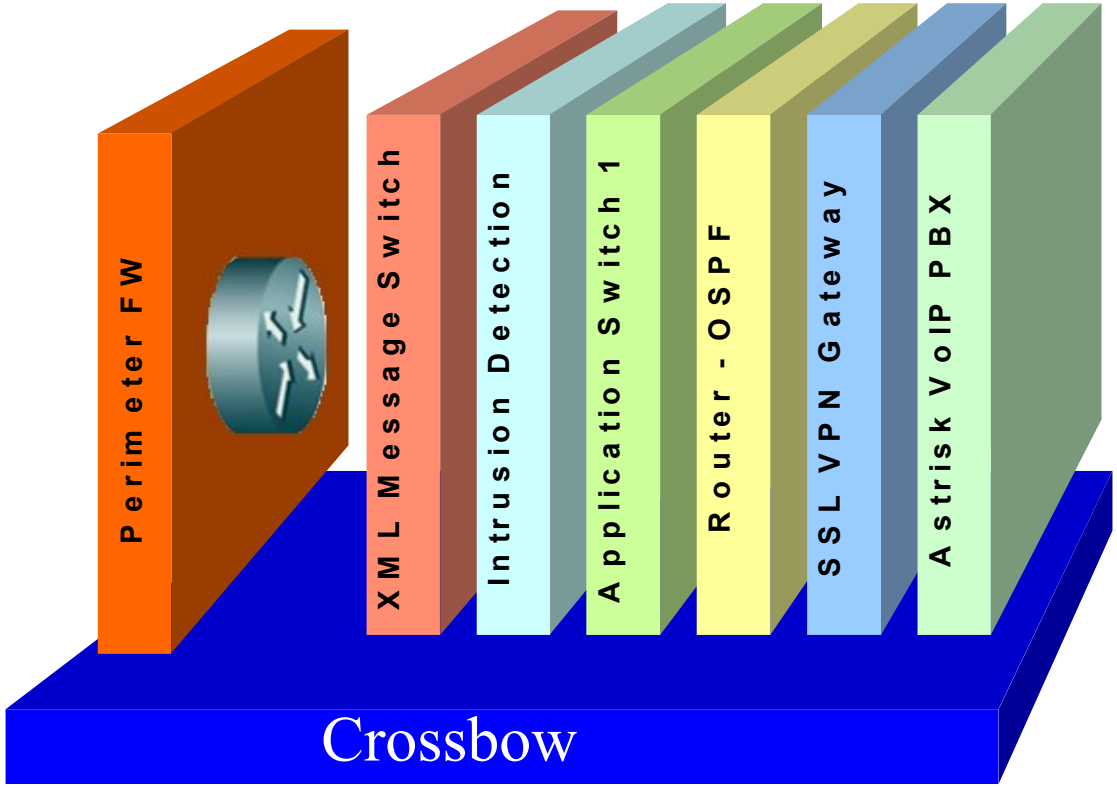

Pkt Chain Lengths



# Virtual Network Machines
## Networking as a Service (NaaS)

- Virtual Network Machines – Networking as a Service
  - Monetize via the subscription model in cloud using virtualized networking services like vRouter, vloadbalancer, vFirewall, vDHCPserver, vDNSserver, etc
- Virtualized Networking Services wrapped in Solaris Zone/Xen/VB running on dedicated Networking blades/appliance
  - Open Source Virtualized Networking Services
  - VNICs and Vswitches provide the virtualized ports similar to physical ports
  - Enable Virtual Networks with configurable link speeds using Virtual Wire
- Management for a Virtual Network Machines
  - Solaris command line
  - Cisco Style 'cli'
  - Web based

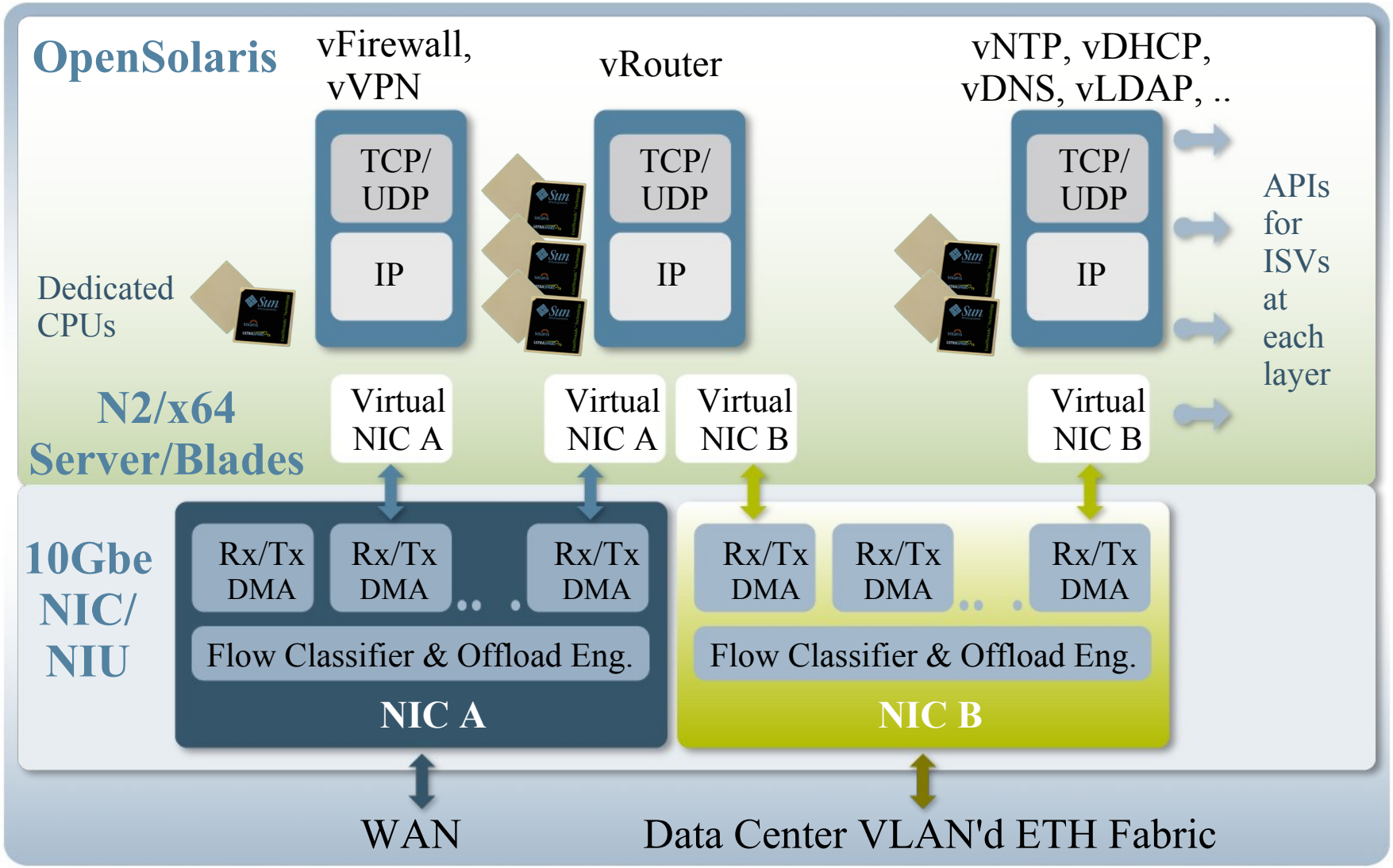
# Virtual Network Machine Appliance
## for the cloud



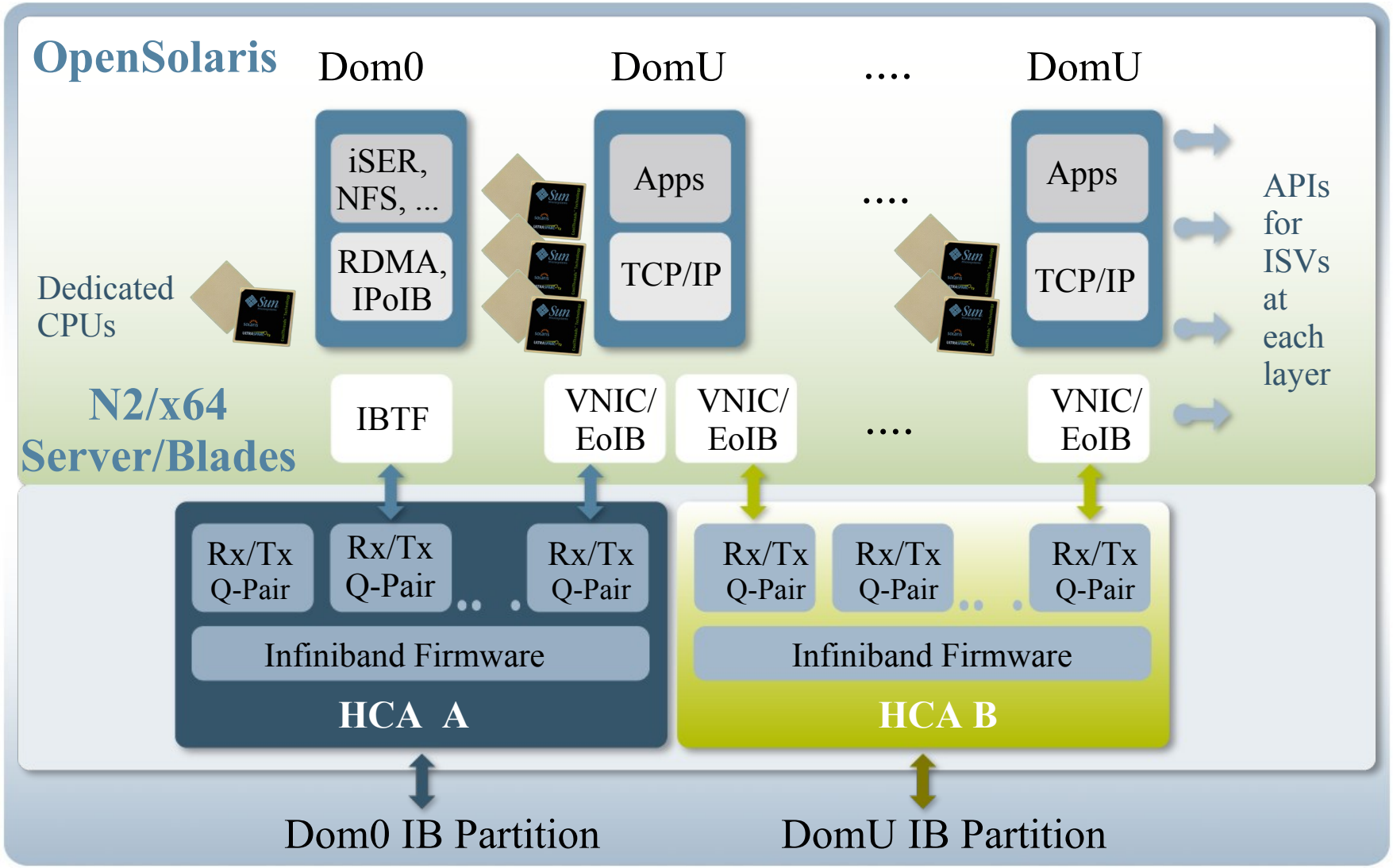

The network is the computer

# Networking as a Service (Naas)
## Subscription based or dedicated appliance

# Virtual Network Machines over 10Gbe



# Cloud Virtual Machines over 40Gbs IB



# Open Storage Networking

- Priority Based Flow Control (PFC)
  - 8 ethernet virtual lanes with their own pause mechanism
  - Extend the Crossbow H/W Virtualized Lanes to the switch
- Enhanced Transmission Selection (ETS)
  - Add Class of service support within the ethernet virtual lane
  - Extend the Crossbow flow based QoS to the switch
- Link Layer Discovery Protocol (LLDP) and Congestion notification (optional)
- PFC and ETS is useful in normal virtualization and server QoS scenarios
- PFC, ETS, and LLDP are necessary to implement Data Center Bridge Exchange protocol (DCBX) and FCOE

# Join Us...

- Our communities and projects are open on OpenSolaris.org:
  - CrossBow: http://opensolaris.org/os/project/crossbow
  - VNM: http://opensolaris.org/os/project/vnm
  - Networking: http://opensolaris.org/os/community/networking
- Where you will find:
  - Active discussions, design docs, FAQs, source code drops, preliminary binary releases, etc...