CrossBow:
Network Virtualization and Resource Control

Sunay Tripathi
Solaris Core Technology group
Sunay.Tripathi@sun.com

Real Scenarios

Financial Services

- Trading house starts offering free financial information to attract customers
- Brokerage customers start complaining that trading site slows down
- The paying customers start deserting

Large ISP

- ISP wants to deploy virtual systems on same physical machines
- ISP sells each virtual system at different price levels to its customers
- Any virtual instance can overwhelmed the shared networking resource

Enterprise Computing

- A large company uses a workgroup server for day to day as well as critical traffic
- IT Ops doing non critical stuff started a backup over the network
- Users doing time critical work can't get bandwidth to do their job

What Happened?

- Critical services are overwhelmed by non-critical services, traffic types, or virtual systems
- No usable mechanism available for fairness, priority and resource control for networking bandwidth

CrossBow Target Markets

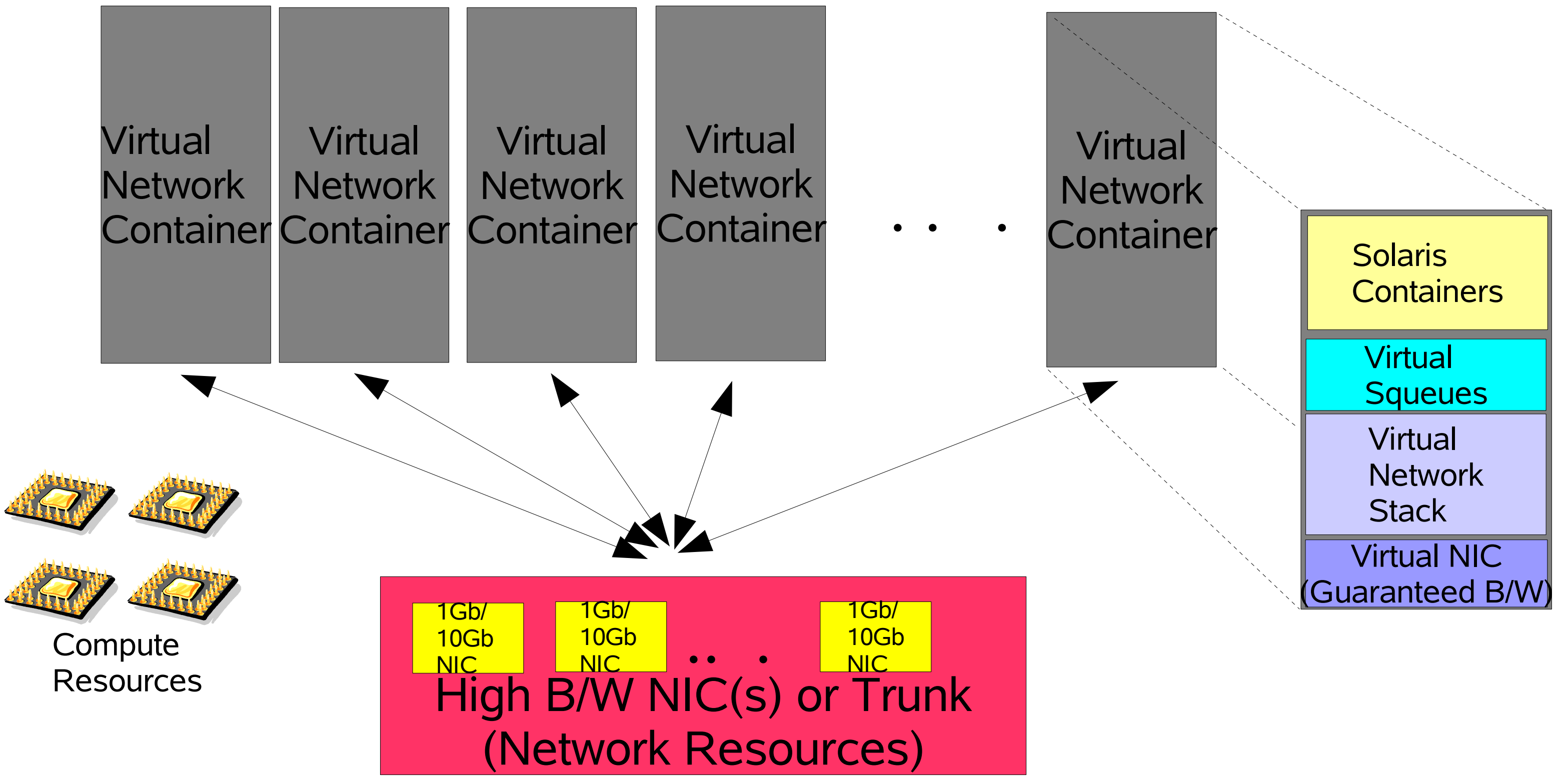
- Server/OS/Network consolidation
  - Cost of managing servers in an enterprise
    - Per physical machine
    - Per OS instance
    - Per port
- Traditional QOS markets
    - Application/Service level B/W control
    - Diffserv
- Horizontally Scaled markets
    - Enforcement of common policies through the server farm
    - Sharing of common B/W to a blade chasis

Constraint: Easy to use; No performance penalty

Network Virtualization

- Create virtual stacks over 1Gb and 10Gb NICs based on protocol, service, or container
- Requirements:
  - Specify priority and/or bandwidth relative to other virtual stacks on the system
  - Be able to choose protocol layers and any tuning specific to the virtual stack
  - Virtual stacks isolated from each other (for both resources and security purposes)

Virtual Network Stack


Virtual Network Container
Virtual Network Container
Virtual Network Container
Virtual Network Container
. . .
Virtual Network Container
Solaris Containers
Virtual Squeues
Virtual Network Stack
Virtual NIC (Guaranteed B/W)
Compute Resources
1Gb/ 10Gb NIC
1Gb/ 10Gb NIC
. . .
1Gb/ 10Gb NIC
High B/W NIC(s) or Trunk (Network Resources)

Technical Obstacles

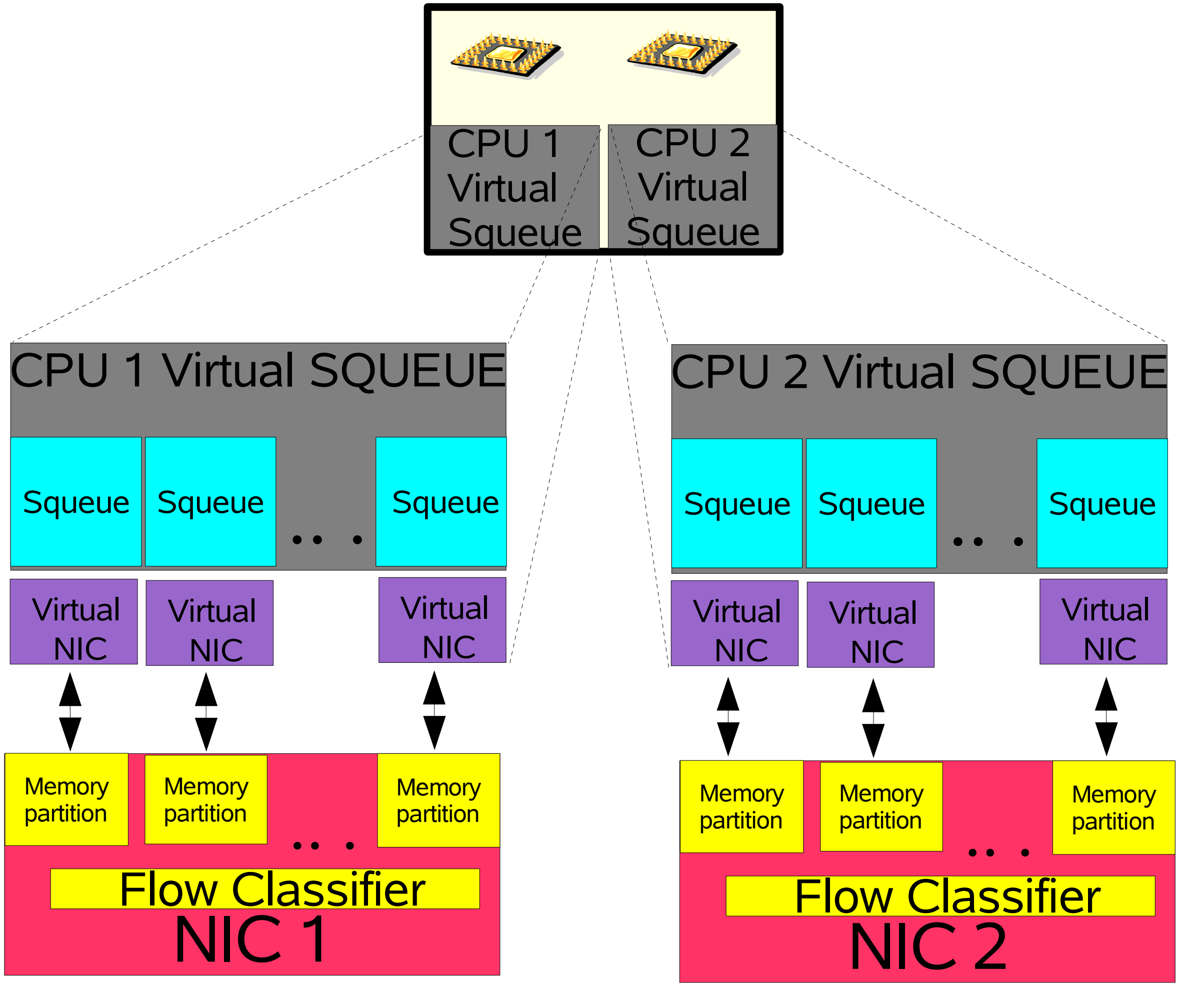
- Obstacles to achieving network virtualization:
  - Network processing in interrupt context
  - Anonymous packet processing in kernel
  - Common queues
- Performance can be degraded by the extra processing to enforce fairness, resource control or network virtualization
- No isolation for flows

The Crossbow Architecture

- Divide NIC memory, DMA channels, etc and use a flow classifier to build a virtual stack on each H/W partition
- Each Virtual NIC is owned by the FireEngine Squeue's which independently switch the VNIC between interrupt & polling mode
- Rate of packet arrival from a VNIC is independently controlled by the Squeue owning the VNIC

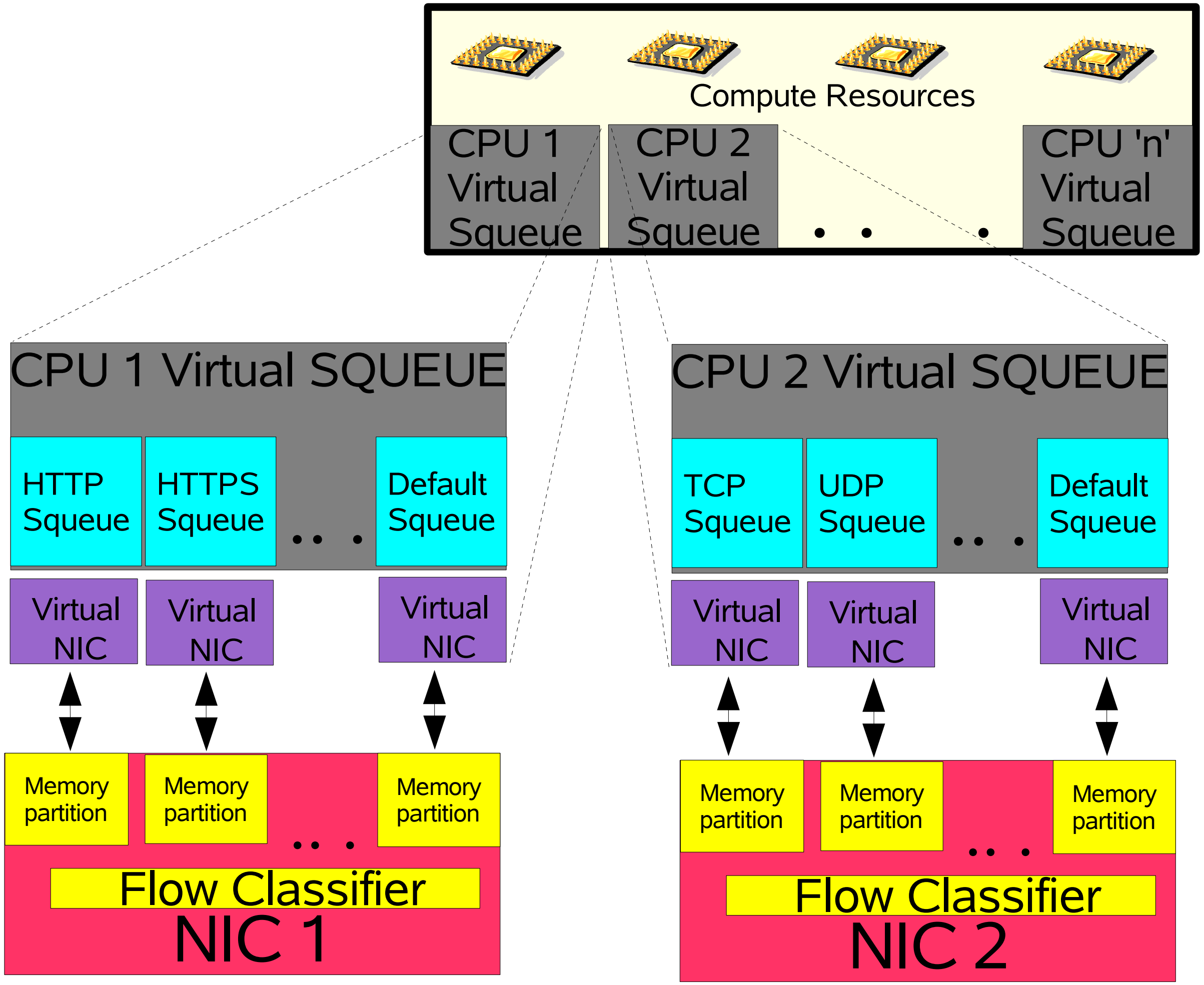
Virtual Stacks – Services & Protocols

The VNICs are in the control path only. The data link layer is bypassed when delivering packets to IP



The Squeue switches the MSI interrupt per stack between interrupt and polling mode and controls the rate of packet arrival for that stack

Virtual Stacks – Services & Protocols



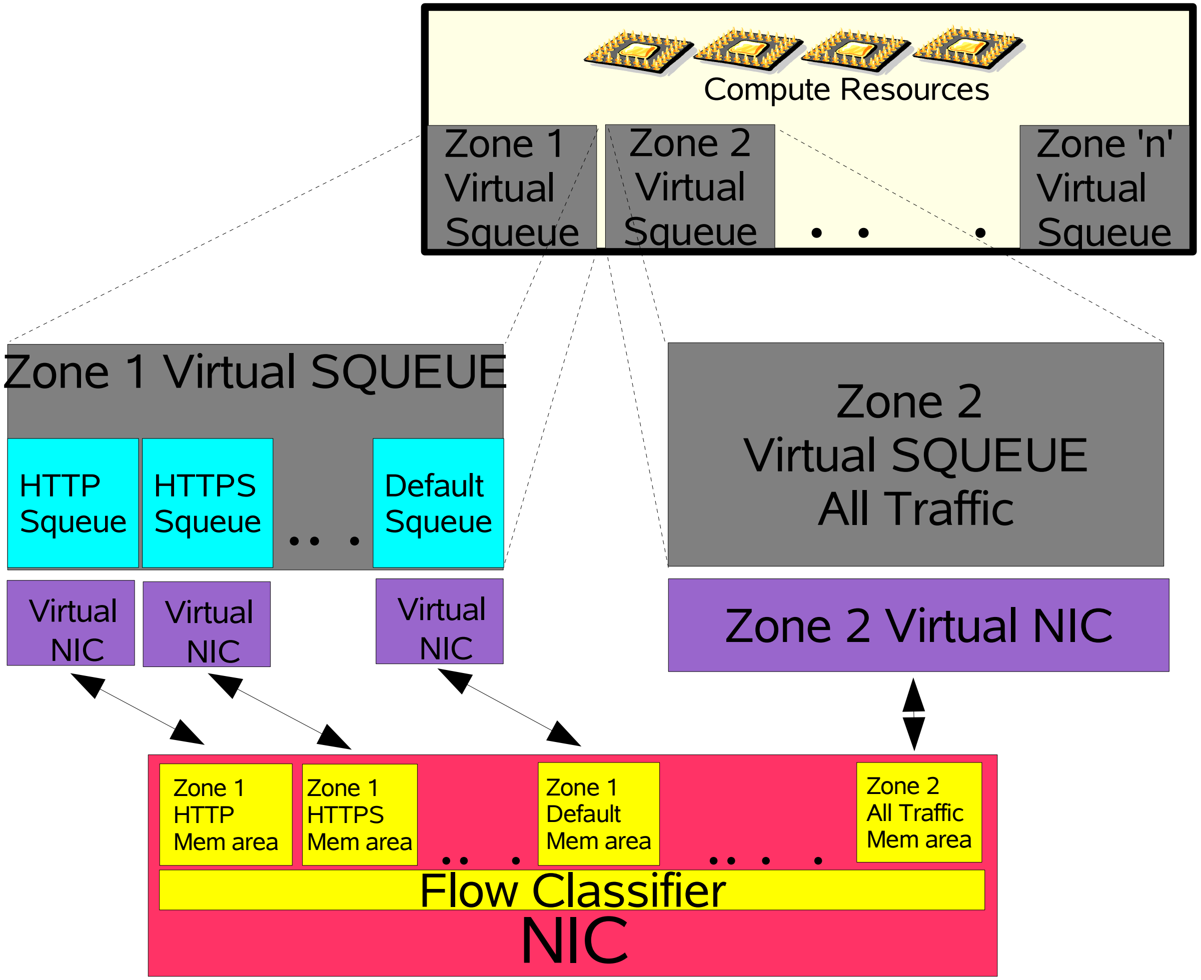
The VNICs are in the control path only. The data link layer is bypassed when delivering packets to IP

The Squeue switches the MSI interrupt per stack between interrupt and polling mode and controls the rate of packet arrival for that stack

Virt. Stack per container

- Each Solaris container has its own virtual stack
- When container is created, the B/W, priority and number of possible virtual stacks within the container is specified
- The Container administrator can configure the allocated virtual stacks to its own taste
- Each Container can have its own routing table, firewall, etc and tune it according to its requirement

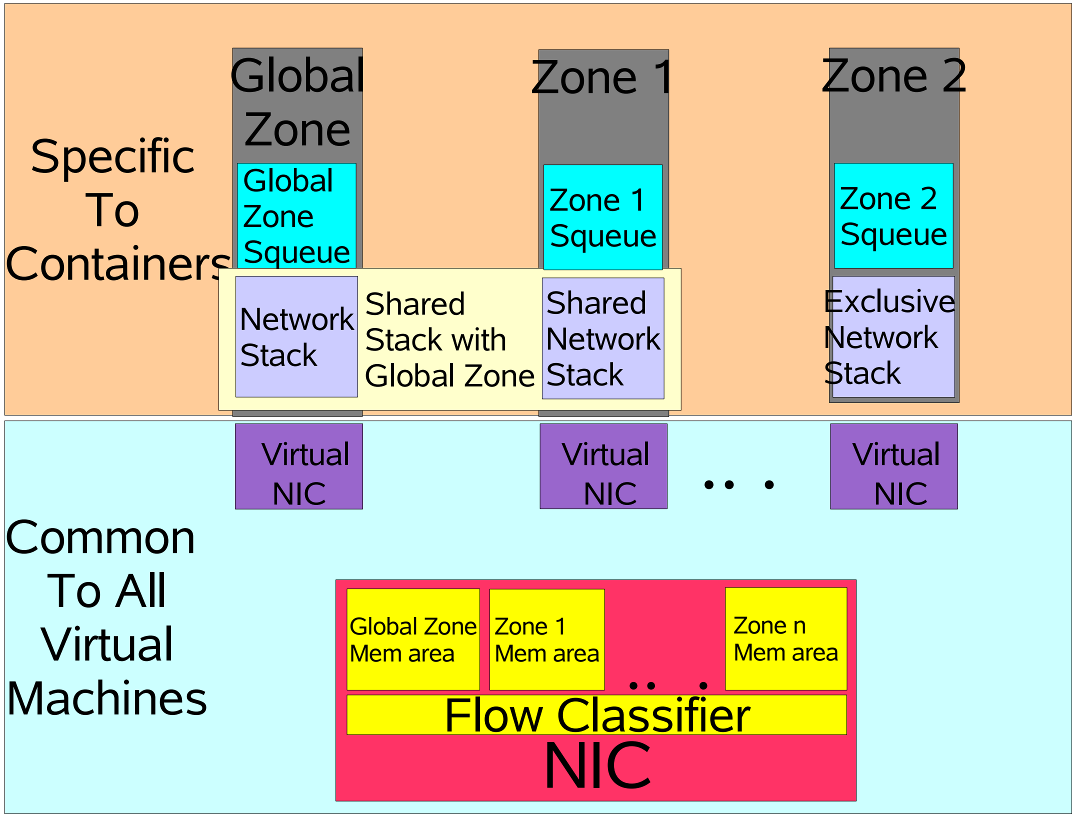
Virtual Stacks - Containers



The VNICs are in the control path only. The data link layer is bypassed when delivering packets to IP

The Squeue switches the MSI interrupt per stack between interrupt and polling mode and controls the rate of packet arrival for that stack

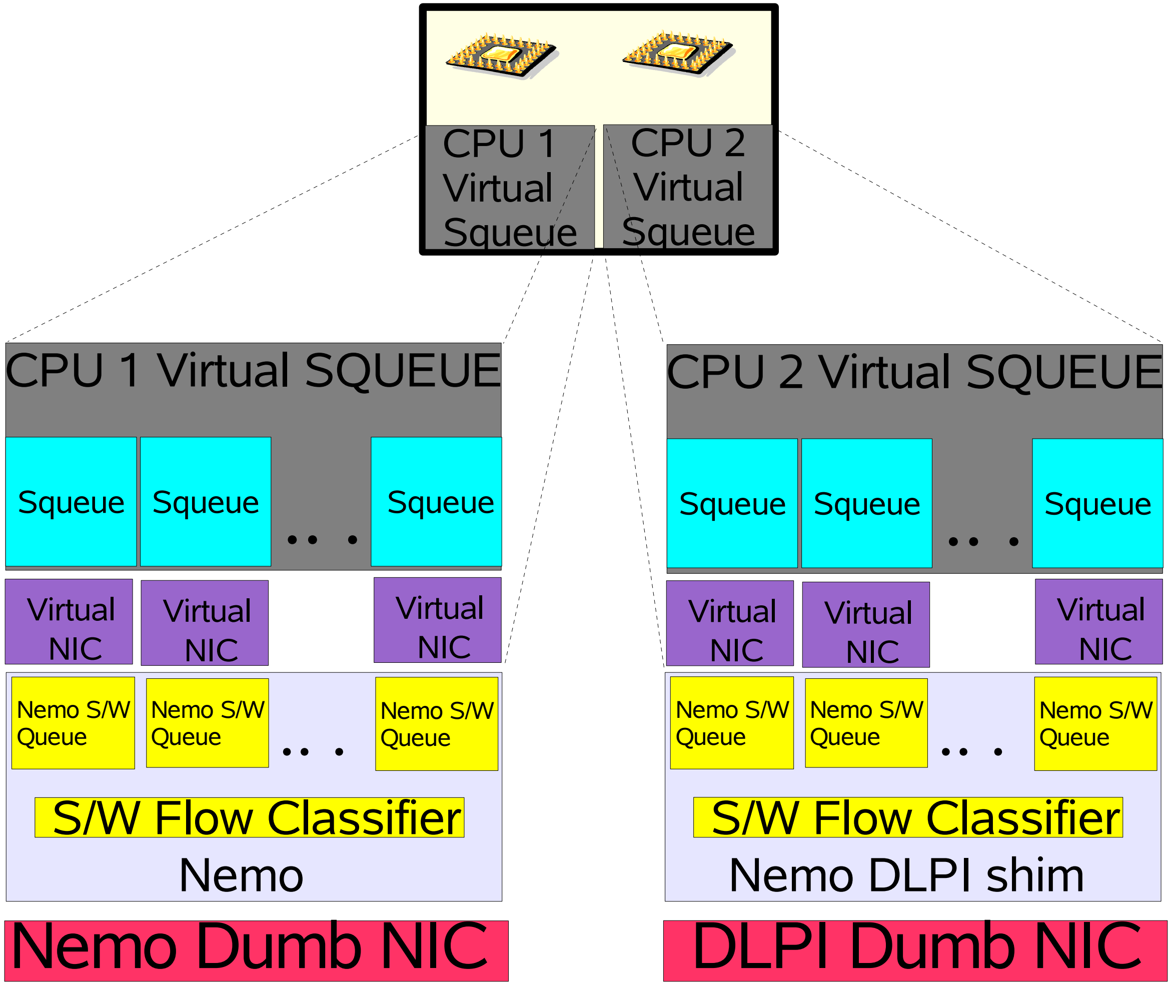
Virtualized Networking



Dumb NICs

- The architecture supports non Nemo NICs as well as Nemo NICs which don't have flow classification capabilities
- We simulate multiple queues or memory area in the Nemo layer using a S/W flow classifier
- Nemo provides a DLPI shim layer for non Nemo drivers
- All the general 1Gb and 10Gb NICs in future will support the flow classification and memory partitioning capability at no extra cost

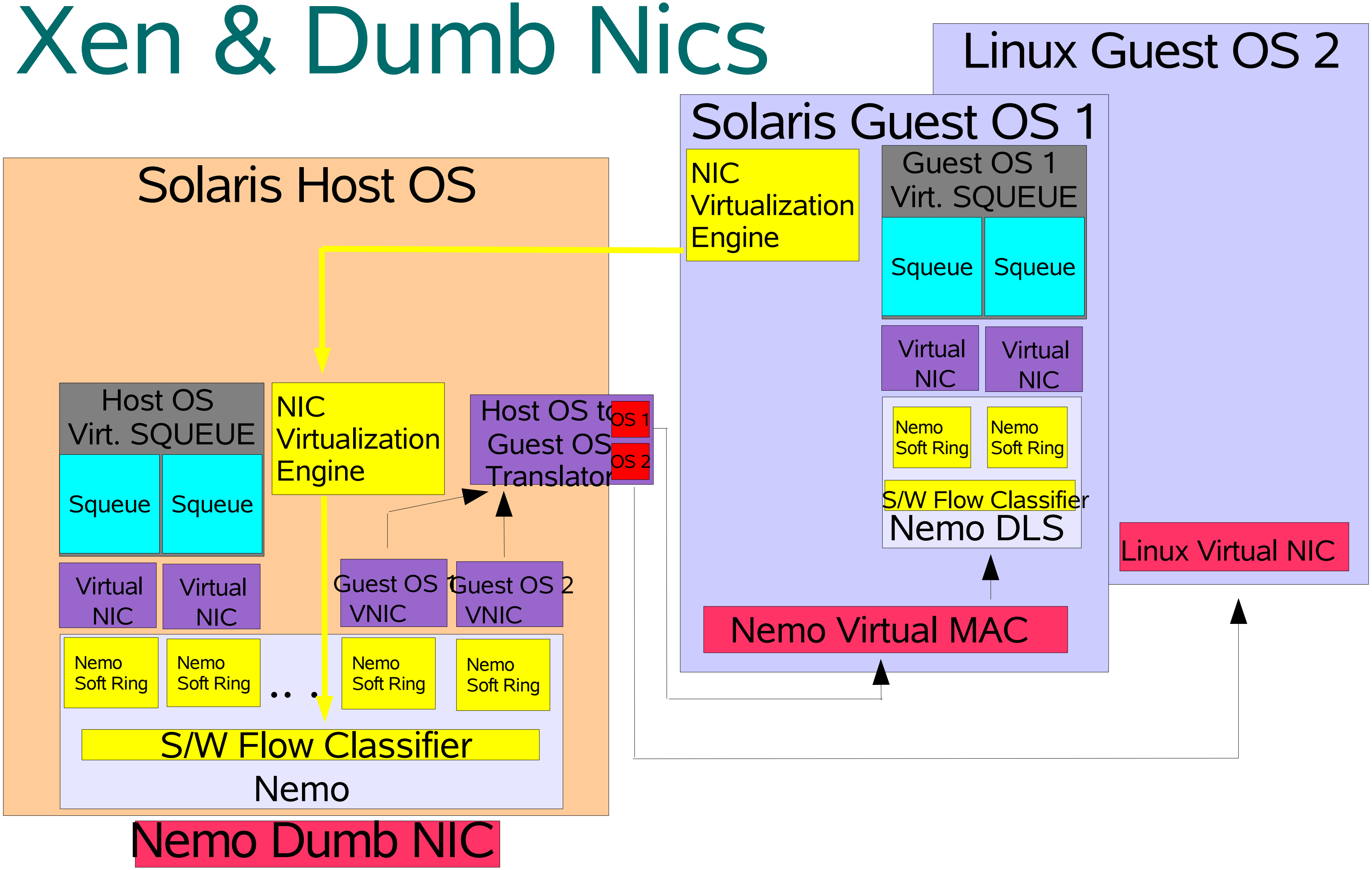
Virtual Stacks with Dumb NICs



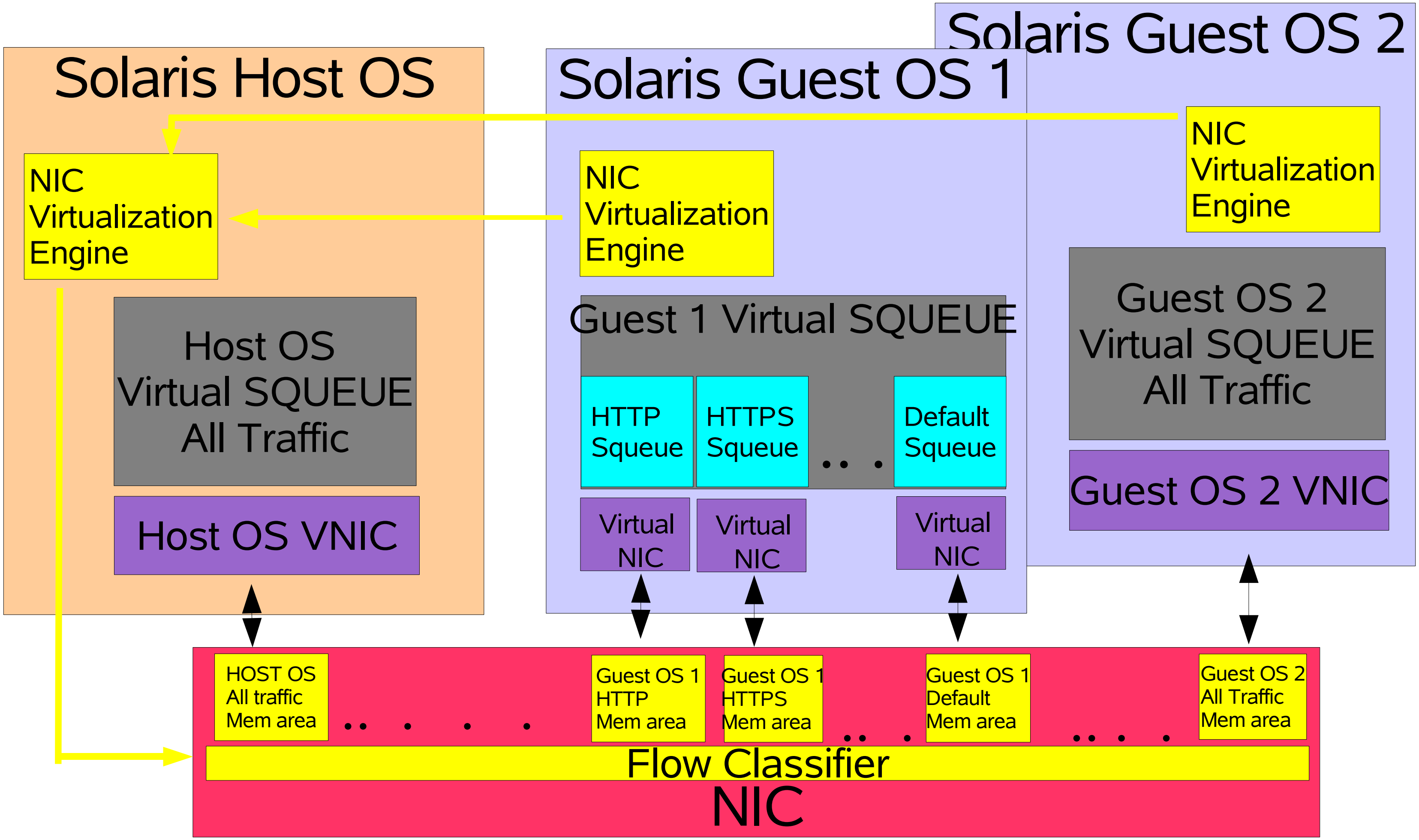
Soft vs Hard Virtualization

- Crossbow is evolving Solaris soft virtualization strategy
  - Containers provide the virtual application environment
  - Crossbow virtual stacks associated with Containers and CPU/Mem resource pool allow vertical partitioning on the machine
- Crossbow is complementary to Hard virtualization and allow network resource control for other virtual machines

Xen & Dumb Nics



Crossbow Smart NIC & Xen


Solaris Host OS
NIC Virtualization Engine
Host OS Virtual SQUEUE All Traffic
Host OS VNIC
Solaris Guest OS 1
NIC Virtualization Engine
Guest 1 Virtual SQUEUE
HTTP Squeue
HTTPS Squeue
Default Squeue
Virtual NIC
Virtual NIC
Virtual NIC
Solaris Guest OS 2
NIC Virtualization Engine
Guest OS 2 Virtual SQUEUE All Traffic
Guest OS 2 VNIC
HOST OS All traffic Mem area
Guest OS 1 HTTP Mem area
Guest OS 1 HTTPS Mem area
Guest OS 1 Default Mem area
Guest OS 2 All Traffic Mem area
Flow Classifier
NIC

Defense against DOS/DDOS

- DDOS have the ability to cripple the entire grids 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 (limited resource) stack
- Connections transition to appropriate priority stacks after application authentication

Accounting & Capacity planning

- Finer grain accounting comes for free
- We can now do per squeue accounting to track usage by a container, service or protocol
- A userland daemon can pull the statistics out at fixed interval and do accounting etc.
- Running the virtual stacks without any resource control helps capacity planning

Admin Interfaces (***dladm***)

- Using ***dladm(1M)*** to create the Virtual NICs and set the attributes on them

  / Synopsis to create a VNIC */*

  dladm create-vnic -d dev_name [-m {factory|shared|random| value}] [-b bandwidth {-L|G}] [-p pri] -H vnic-key

  / Create a simple VNIC & let the system pick the mac address */*

  dladm create-vnic -d bge0 2

  / Create a VNIC with a guaranteed B/W of 600Mbps & priority Hi */*

  dladm create-vnic -d bge0 -m factory -b 600m -G -p high 1

  /* Use ifconfig or DHCP to configure IP addresses */

ifconfig and *dladm*

/ Use dladm to create a VNIC and ifconfig to configure it */*

```
# ifconfig -a
bge0: flags=1000843<UP,BROADCAST,RUNNING,MULTICAST,IPv4> mtu 1500 index 3
        inet 172.16.1.1 netmask ffff0000 broadcast 172.16.255.255
        ether 0:10:18:a:29:44
vnic1: flags=1000843<UP,BROADCAST,RUNNING,MULTICAST,IPv4> mtu 1500 index 4
        inet 192.16.1.1 netmask ffffff00 broadcast 192.16.1.255
        ether 0:10:18:c:77:55
```

Admin Interfaces (***netrcm***)

- Set the bandwidth related attributes for any service, protocol or virtual machine

```
/* Synopsis to add a flow with attributes to the system */
 netrcm add-flow  -d dev [-b bandwidth {-L | -G}] [-H] [-p pri]
     {[mac_addr = value] | [sap = value] | [ip_addr = value] |
     [{proto = TCP|UDP} [[local] port = value]]} flow-id

/* Create dedicated resources  around  HTTPS (port 443) service */
 netrcm add-flow -d bge0 -H proto = TCP local port = 443 https-1
/* Add a bandwidth guarantee as well to above */
 netrcm modify-policy -d bge0 -b 90% -G -p high https-1
```

Admin Interfaces (****cfg***)

- Alternatively, *cfg commands for virtual machines can be modified to take B/W, pri, phys/virt interfaces, IP addresses etc

```
zonecfg -z new_zone
zonecfg:new_zone> create
zonecfg:new_zone> net phys=bge1
zonecfg:new_zone> net virt=eth0
zonecfg:new_zone> net bw=30Mbps
zonecfg:new_zone> net pri=hi
zonecfg:new_zone> net ip_addr=a.b.c.d
```

- Similar mechanism for Xen/ldom etc
- Within a virtual machine, local admin can use dladm or netrcm to create more VNICs or policies

Open Issues: APIs & Stats

- APIs?
  - Administrative
  - stats
  - Alarms
- Statistics
  - Real Time Usage per VNIC/Flow
  - History
  - Billing & Accounting

Open Issues: Configuration

- Configuration files
  - Where are they stored?
  - Flat files or SMF properties
  - Hand editable

Open Issues: B/W Control

- Bandwidth limits & priority specification
  - Full duplex or half duplex
  - How many priority levels
  - How to specify bandwidth resource
    - Fixed limit
    - Guarantee
  - Units
    - Traditional units (Mbps, Gbps)
    - Shares
    - Percentages

CrossBow:

Network Virtualization and Resource Control

Sunay Tripathi
Solaris Core Technology group
Sunay.Tripathi@sun.com