NIC drivers in Crossbow

Oliver Yang
Software Engineer
Sun Mircosystem, Inc.

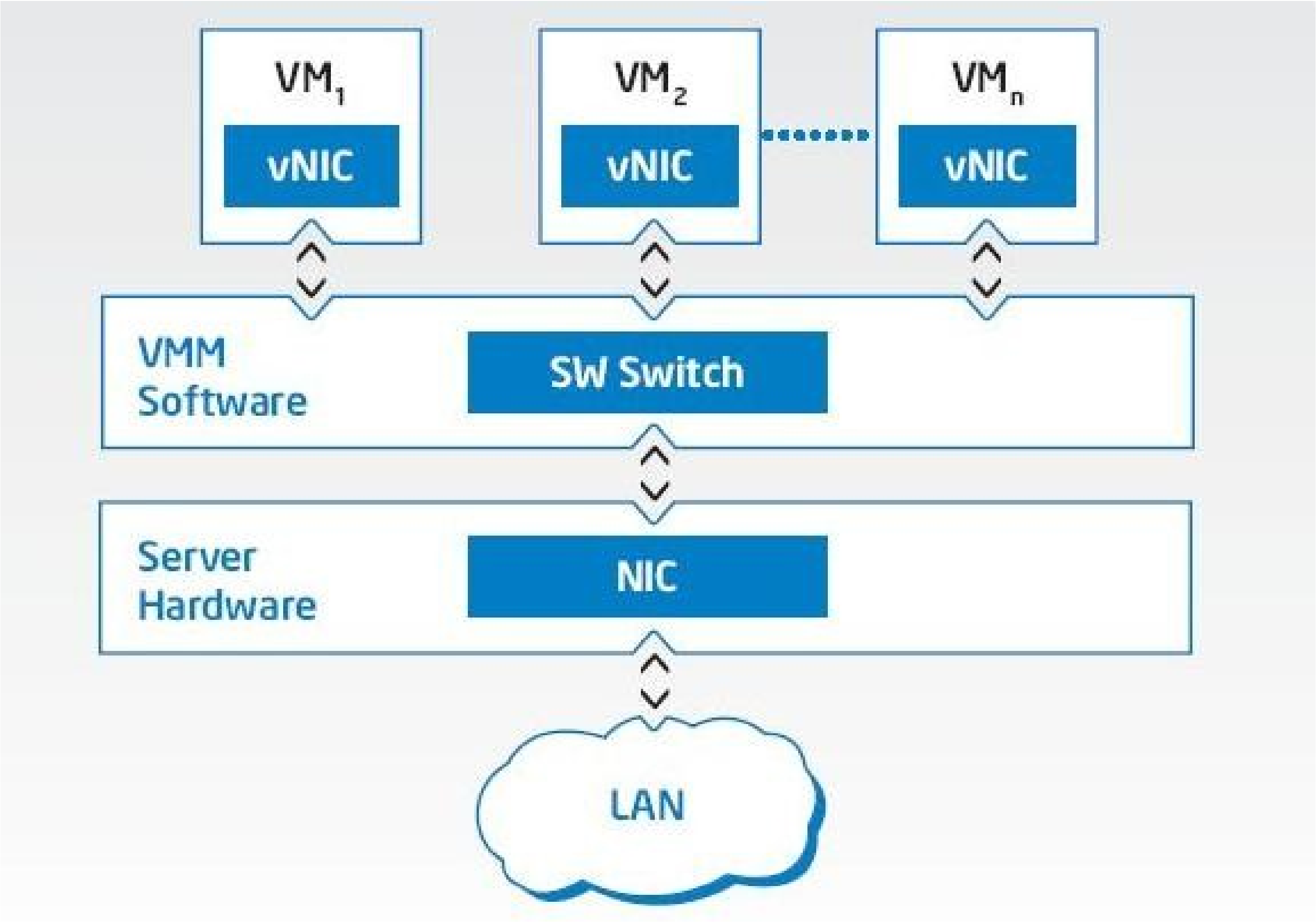
Agenda

- VMDq - Connectivity Virtualization
- Crossbow&NIC Drivers Overview
- VMDq Implementation In Igb
- References

VMDq (Virtual Machine Device queues)

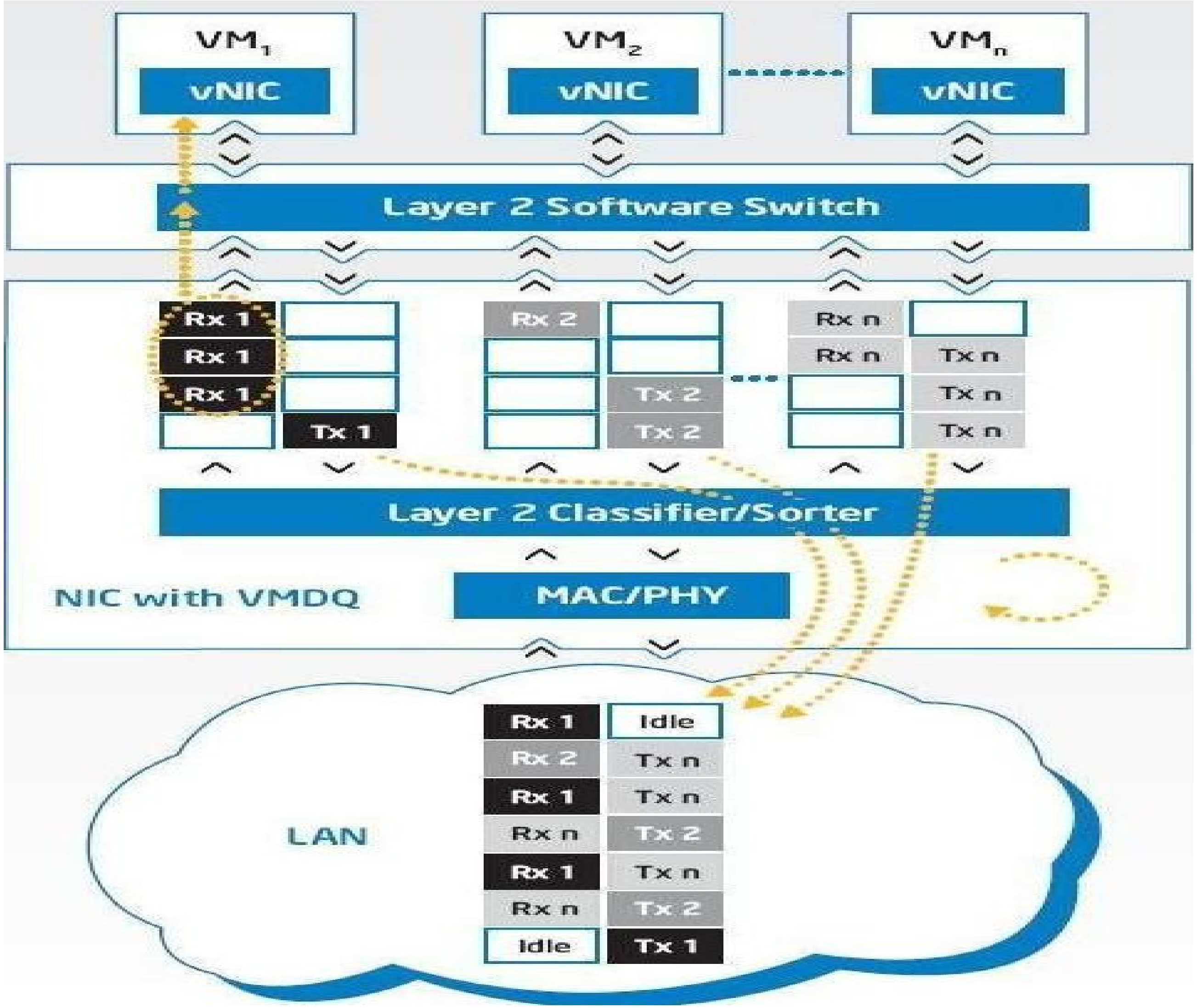
- Connectivity Virtualization & Off-load Technology
  - Improving networking performance
  - Reducing CPU utilization
- Take advantage of following hardware features
  - Multiple RX/TX device queues
  - Multiple MSI-X interrupts
  - Multiple MAC addresses
  - Layer 2 classifications based on MAC/VLAN
  - Layer 3, 4 fanout hashing - RSS(Receive Side Scaling)

Legacy NIC

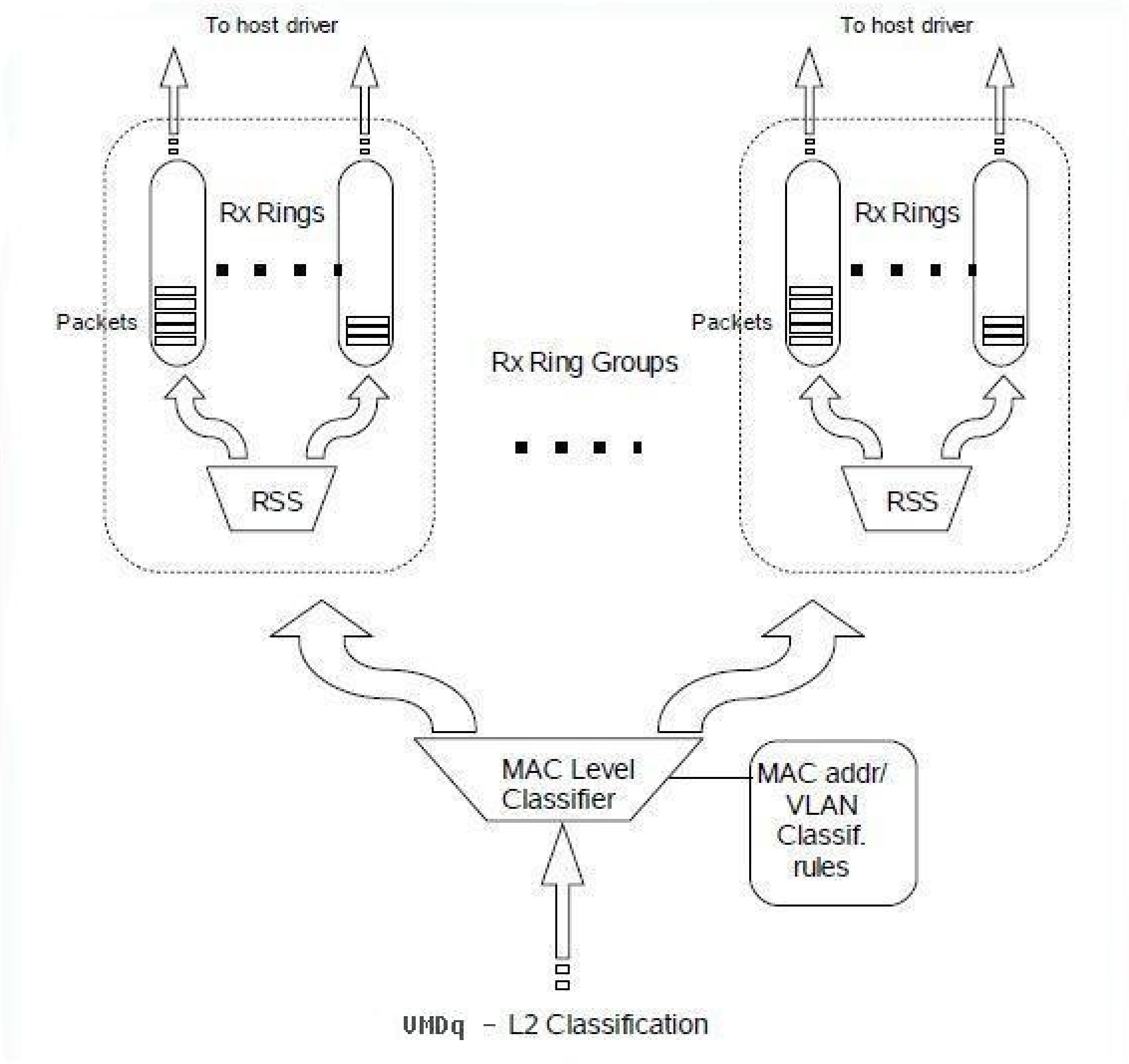


VM_1
vNIC
VM_2
vNIC
VM_n
vNIC
VMM Software
SW Switch
Server Hardware
NIC
LAN

NIC With VMDq – L2 Classification



Ring Ring Groups – VMDq + RSS

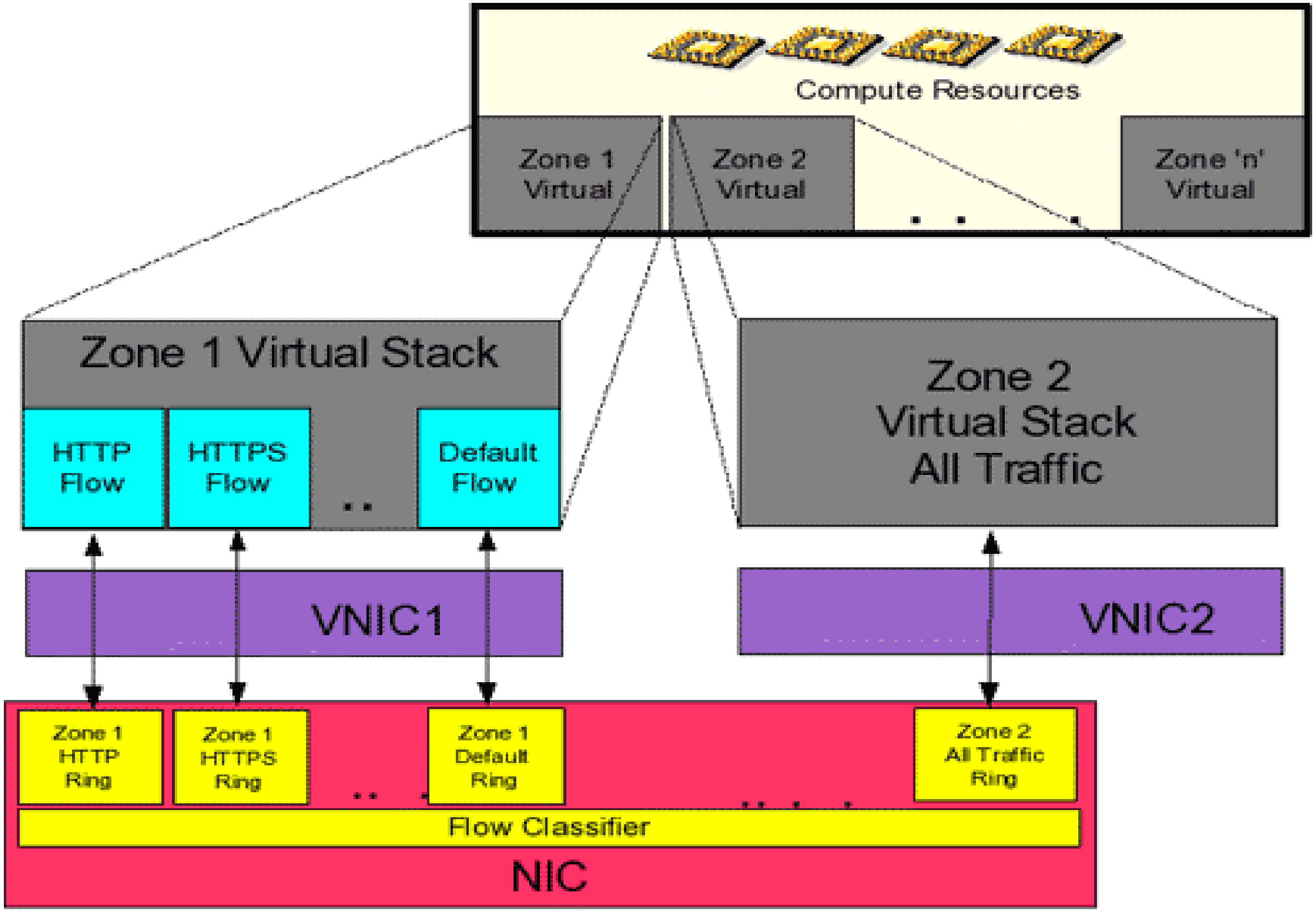


Agenda

- VMDq - Connectivity Virtualization
- Crossbow&NIC Drivers Overview
- VMDq Implementation In igb
- References

Footnote position, 12 pts.

Big Picture of Crossbow



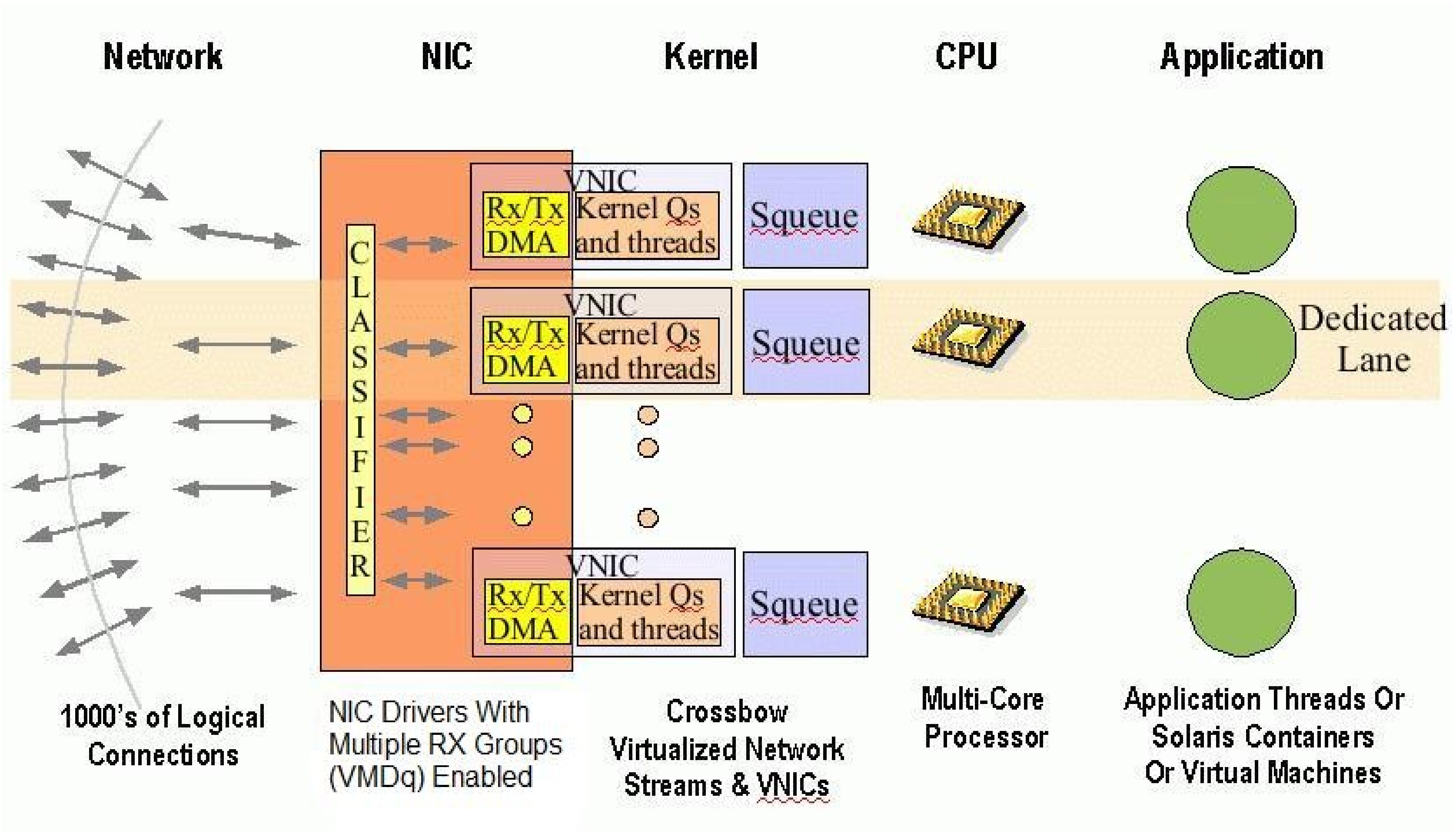
NIC Driver Frameworks In Crossbow

- Legacy Drivers without virtualization capabilities
  - > DLPI
  - > GLDv2
  - > GLDv3
- Virtualization Capable Drivers
  - > GLDv3 extensions for different NICs
    - – NICs that support MAC_CAPAB_RINGS capability
    - – NICs that support MAC_CAPAB_SHARES capability for guest domains direct access NIC's hardware resources
    - – NICs capable of L2, L3 and L4 classification

GLDv3 Extensions in Crossbow

- Obsolete interfaces
    - xxx_m_unicst/_add/_remove/_modify
    - MAC_CAPAB_RX_CLASSIFY in xxx_m_getcapab
- New interfaces
  > MAC_CAPAB_RINGS in xxx_m_getcapab
    - xxx_fill_ring for RX/TX rings registration
        - Callback for send interface
        - Callback for polling interface
        - Callbacks for RX interrupt enable/disable
        - Callbacks for RX/TX rings stat/stop
    - xxx_fill_group for RX/TX groups registration
        - xxx_addmac/xxx_remmac callbacks

VMDqs -> RX Groups -> VNICs



Driver Features in Crossbow

	E1000g	Bge	Igb	Ixgbe	Xge
Multiple TX rings	N	Y	Y	Y	Y
Multiple RX rings	N	Y	Y	Y	Y
Multiple MSI-X interrupts	N/A	N	Y	Y	Y
RX L2 (MAC)Virtualization	N	Y	Y	N	Y
RX L٢ (MAC+VLAN) Virtualization	N	N/A	N	N	N
RX L3,4 Fanout Hashing	N	N/A	Y	Y	N

Agenda

- VMDq - Connectivity Virtualization
- Crossbow&NIC Drivers Overview
- VMDq Implementation In Igb
- References

Footnote position, 12 pts.

Hardware Configurations - Zoar

- 4 RX + 4 TX rings
- 4 RX + 4 TX MSI-X interrupt vectors
- RX group - An abstraction for VMDq
  - 1 RX group, 4 RX rings per group
  - 2 RX groups, 2 RX rings per group
  - 4 RX groups, 1 RX ring per group
- RSS – L3/L4 fanout hashing
  - 1 RX group, fanout to 4 rings
  - 2 RX groups share 1 RSS, fanout to 2 rings for each group
  - 4 RX groups, RSS is NOT enabled

Receive Packets

- Interrupt mode
  - Per-ring RX interrupt handlers

    igb_intr_rx ->igb_rx->mac_rx
- Polling mode
  - Switches between polling mode and interrupt mode

    igb_rx_ring_intr_disable/igb_rx_ring_intr_Enable
  - Per-ring polling interfaces

    igb_rx_ring_poll ->igb_rx
- RX group and RSS initialization code
  - igb_add_mac/remove_mac
  - igb_setup_rss/igb_setup_mac_rss_classify

Send Packets

- Per-ring tx interfaces
  - igb_tx_ring_send
- Per-ring tx interrupt handlers
  - igb_intr_tx -> igb_tx_recycle_* -> mac_tx_ring_update
- About TX Logic
  - The number of TX groups is always 1
  - Round-Robin sending (Determined by upper layer)

Enable VMDq In Igb Driver

- Two tunables in igb.conf
  - mr_enable

    Enable multiple rx queues and tx queues

    Allowed values: 0, 1

    Default value: 1
  - rx_group_number

    The number of the receive ring groups

    Allowed values: 1, 2, 4

    Default value: 1

Agenda

- VMDq - Connectivity Virtualization
- Crossbow&NIC Drivers Overview
- VMDq Implementation In Igb
- References

Footnote position, 12 pts.

Documentations & Links

- Intel Docs
  - www.intel.com/design/network/products/lan/controllers/82598.htm
  - www.intel.com/technology/platform-technology/virtualization/vmdq_whitepaper.pdf
- Crossbow Docs
  - opensolaris.org/os/project/crossbow/Docs/
- Code Review
  - dlc.sun.com/osol/netvirt/downloads/20081009/Webrev_PhaseIII/webrev.PhaseIII/

Q&A

Oliver Yang
Oliver.Yang@sun.com