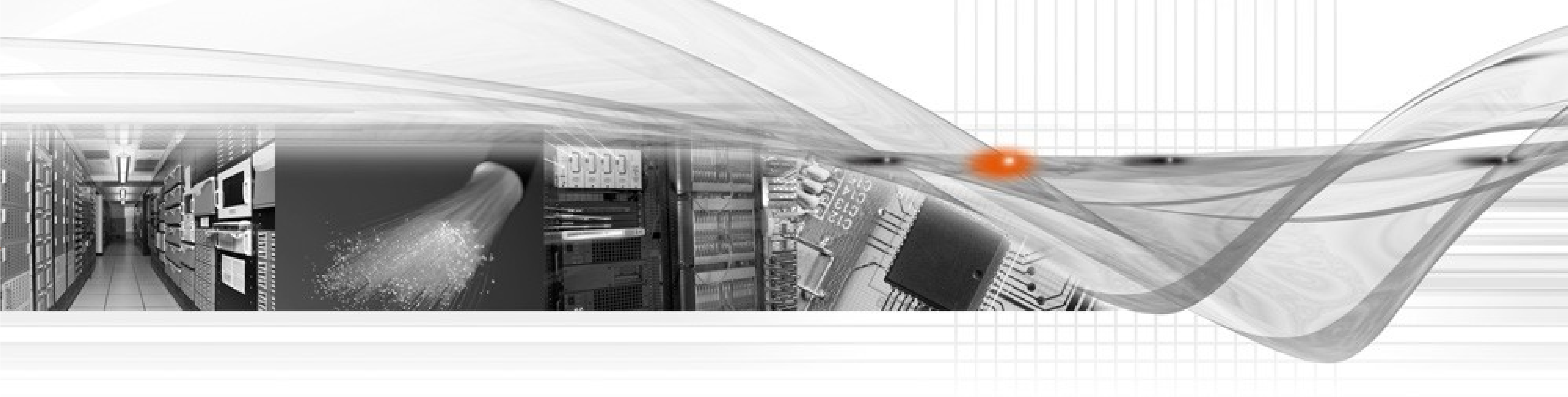


OPENFABRICS ALLIANCE

RoCE Update

Liran Liss, Mellanox Technologies

March, 2012

Agenda

- RoCE Ecosystem
- QoS
- Virtualization
- High availability
- Latest news

RoCE in the Data Center

- Lossless configuration recommended
- Network configuration options
  - Enable global pause

```
(config) # interface ethernet 1/x flowcontrol send on
(config) # interface ethernet 1/x flowcontrol receive on
```

  - Enable PFC

```
(config) # dcb priority-flow-control priority 3-4 enable
```

  - Enable priority tagging to work without VLANs

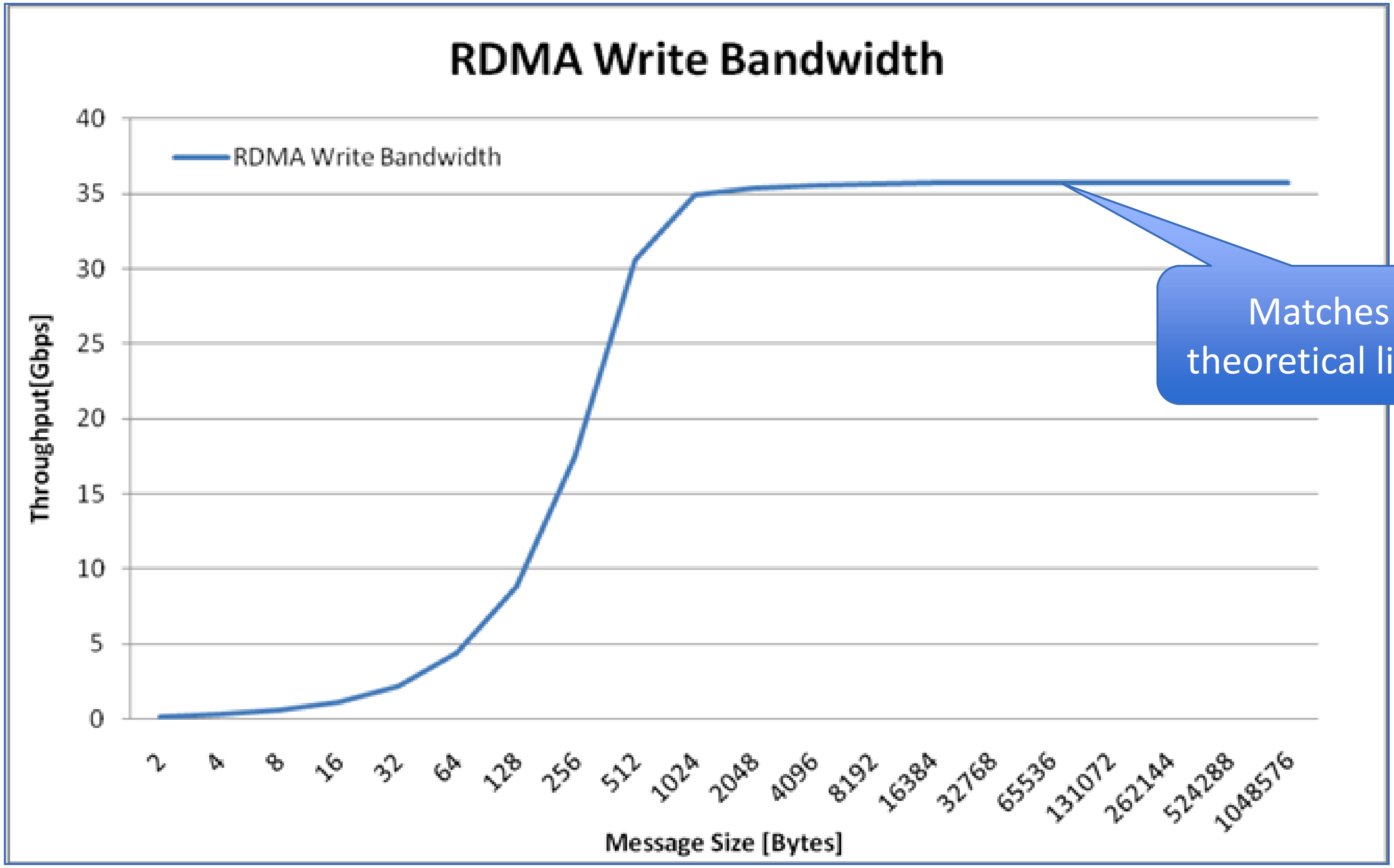
```
(config) # interface ethernet 1/x switchport mode access-dcb
```

RoCE in the Data Center

- Set up matching host configuration
  - Enable PFC
    - Manually by dcbtool or lldptool-pfc or automatically by DCBX (via lldpad)
  - Call rdma_set_option() with RDMA_OPTION_IP_TOS to determine UP in RoCE connections
    - Sets UP = ip_tos[7:5] (precedence bits)

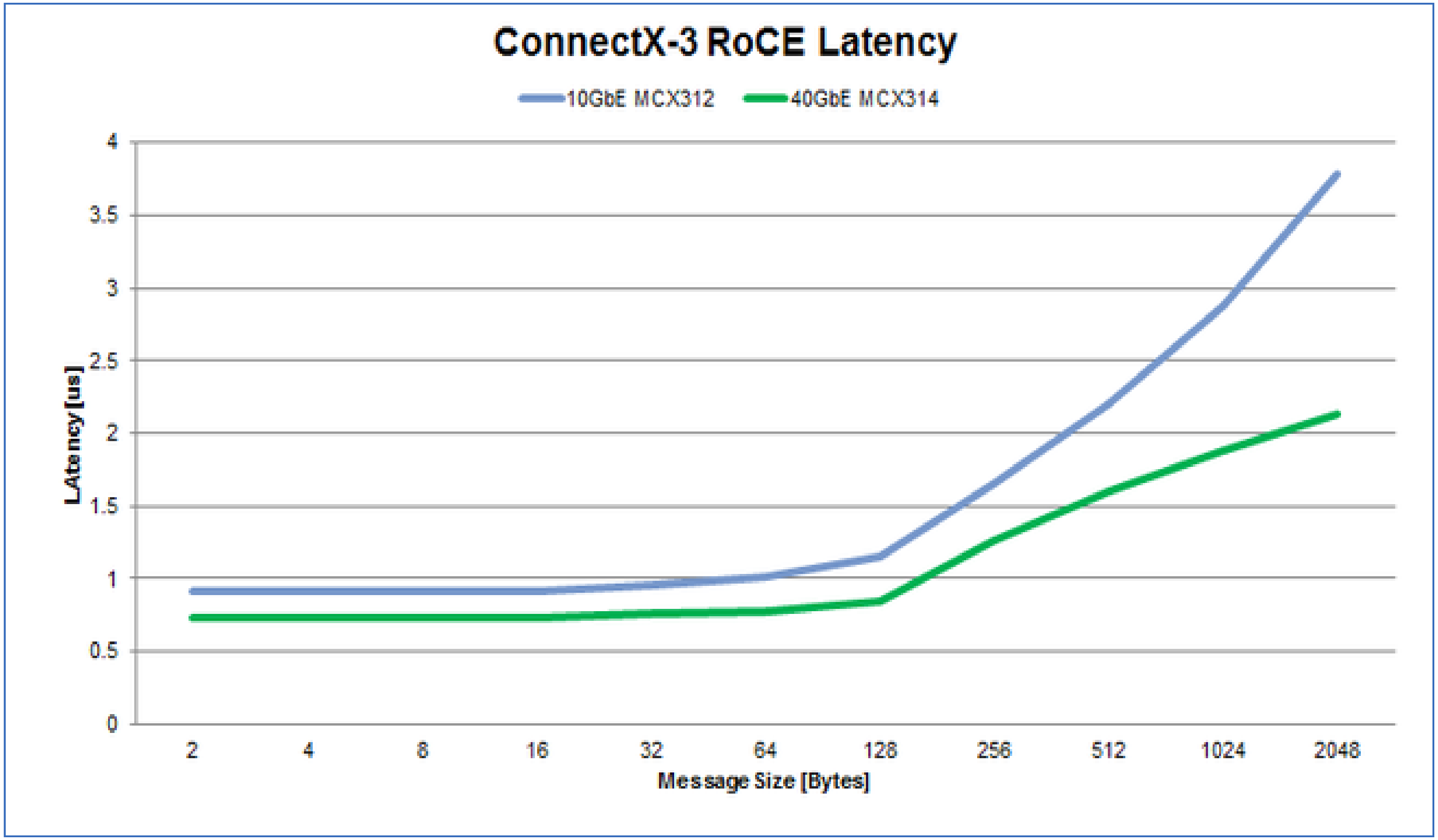
PFC (or global pause) is all it takes to get RoCE working well!

40GE RoCE is Here

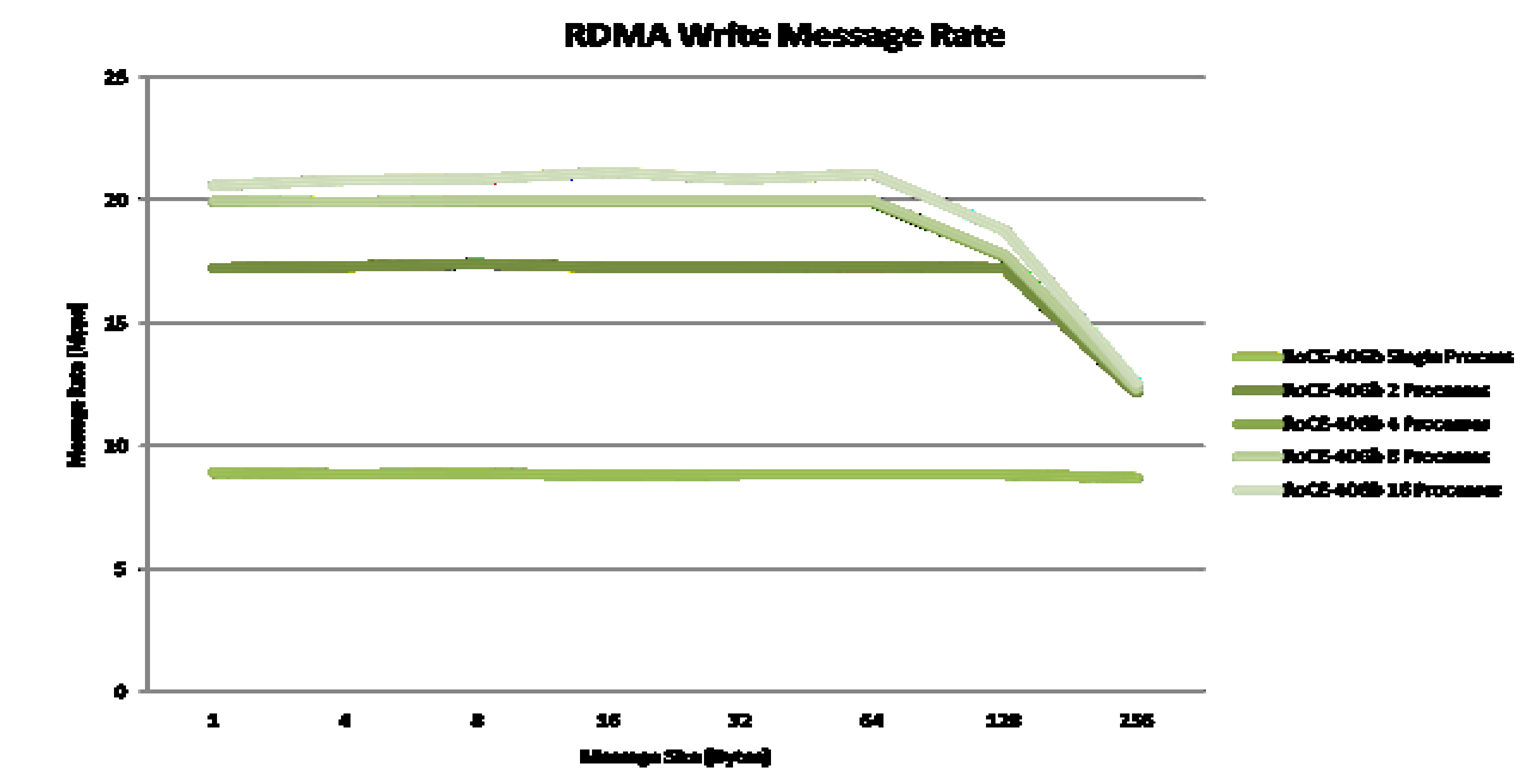


Matches theoretical limit

40GE RoCE is Here

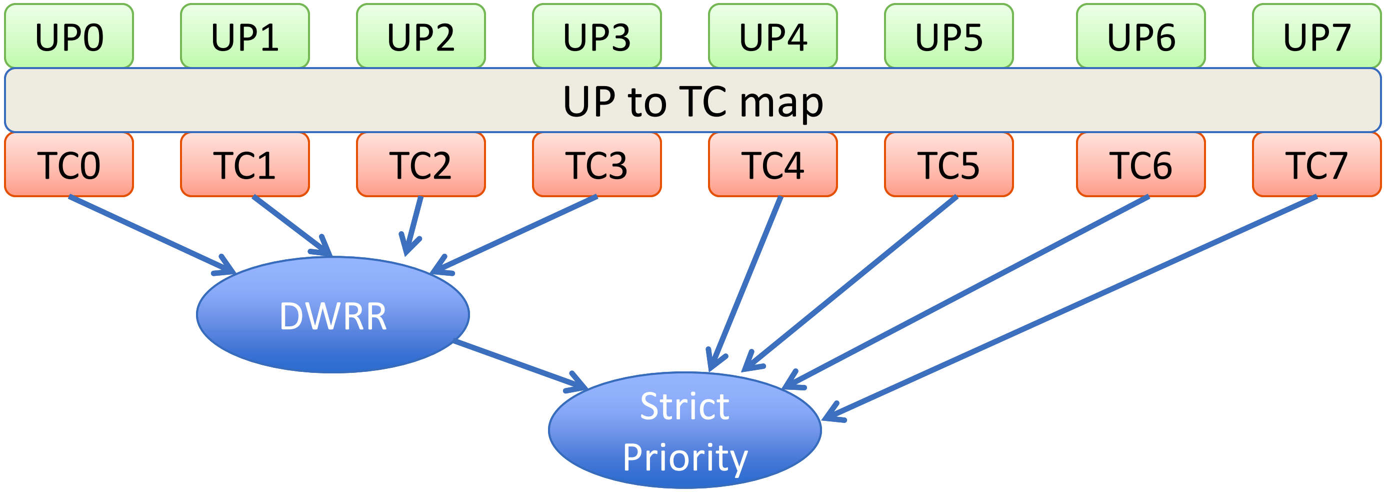


40GE RoCE is Here



Enhanced Transmission Selection (ETS)

- Provides BW guarantees to traffic classes (TCs) assigned for enhanced transmission selection
  - Denoted by a percentage of the BW remaining after transmitting from TCs subject to strict-priority or credit-based-shaper algorithms
- Designates User-Priority (UP) to TC mappings
- Host configuration: dcbtool/lldptool-ets



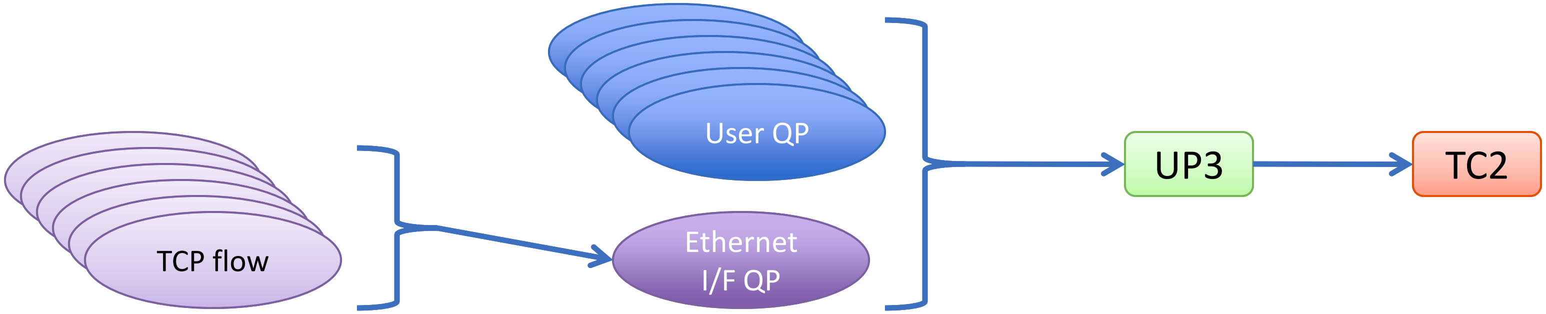
QoS Matters

		QoS OFF RFS / Multi-Tx OFF		QoS OFF RFS / Multi-Tx ON		QoS ON RFS / Multi-Tx ON	
#TCP STREAM	#TCP_RR	Latency [us]	Total BW [Gbps]	Latency [us]	Total BW [Gbps]	Latency [us]	Total BW [Gbps]
0	1	10.1	0	10.7	0	10.5	0
20	1	10548	8934	37.0	9187	12	9330

	No TCP streams		With 20 TCP streams			
			QoS OFF	QoS OFF	QoS ON strict sched	QoS ON ETS 99:1
#RRs	RFS/MQ OFF	RFS/MQ ON	RFS/MQ OFF	RFS/MQ ON	RFS/MQ ON	RFS/MQ ON
1	9.0	10.8	10134.2	37.9	12.0	12.0
2	11.9	10.7	14063.6	38.7	11.9	12.2
4	12.3	11.1	9137.2	50.8	12.5	12.8
6	12.9	11.2	12329.0	48.3	12.6	12.6
8	13.9	13.2	16261.5	41.2	15.0	15.0
10	15.2	14.5	12115.6	52.3	16.2	16.3
20	20.4	21.3	11455.8	48.6	23.4	23.2

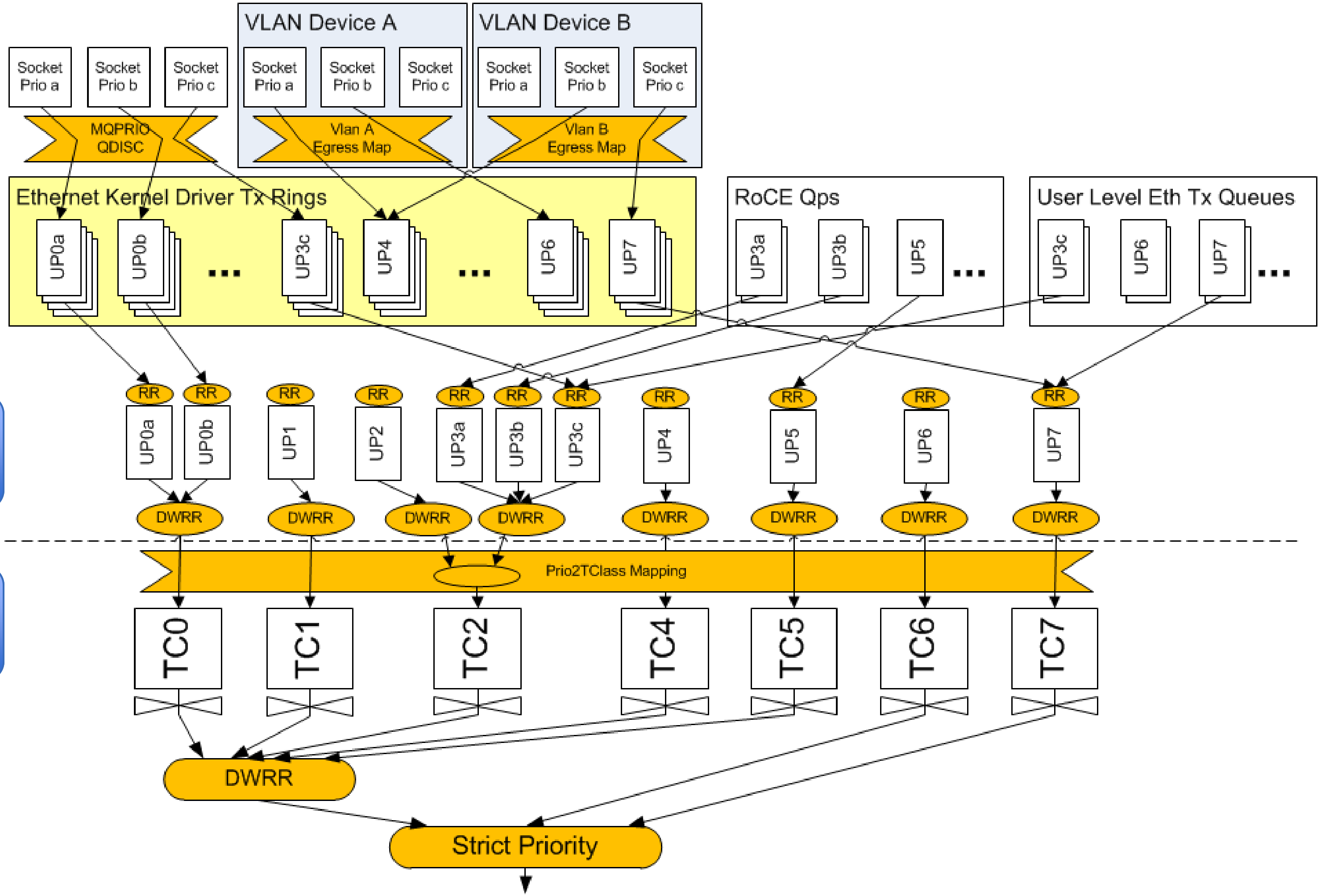
Granular QoS

- End-to-end network QoS may not be enough
  - Some applications may require more than 8 (Ethernet) / 15 (Infiniband) QoS levels
  - HW-level QoS under host admin control
  - Control over scheduling of application HW queues



- Solution: add another scheduling hierarchy level
  - QoS within a UP

The Complete Picture



Host Admin

Network Admin

Configuration (example)

```
# ethtool -f eth2
Fine-grain QoS for eth2:
Total number of QoS queues: 128
Current fine-grain QoS settings:
    UP0 0:100
    UP1 0:20 1:80
    UP2 0:100
    UP3 0:100
    UP4 0:50 1:30 2:20
    UP5 0:100
    UP6 0:100
    UP7 0:100

#ethtool -F eth2 up3 10, 40, 50 up1 100

# ethtool -f eth2
Total number of QoS queues: 128
Current fine-grain QoS settings:
    UP0 0:100
    UP1 0:100
    UP2 0:100
    UP3 0:10 1:40 2:50
    UP4 0:50 1:30 2:20
    UP5 0:100
    UP6 0:100
    UP7 0:100
```

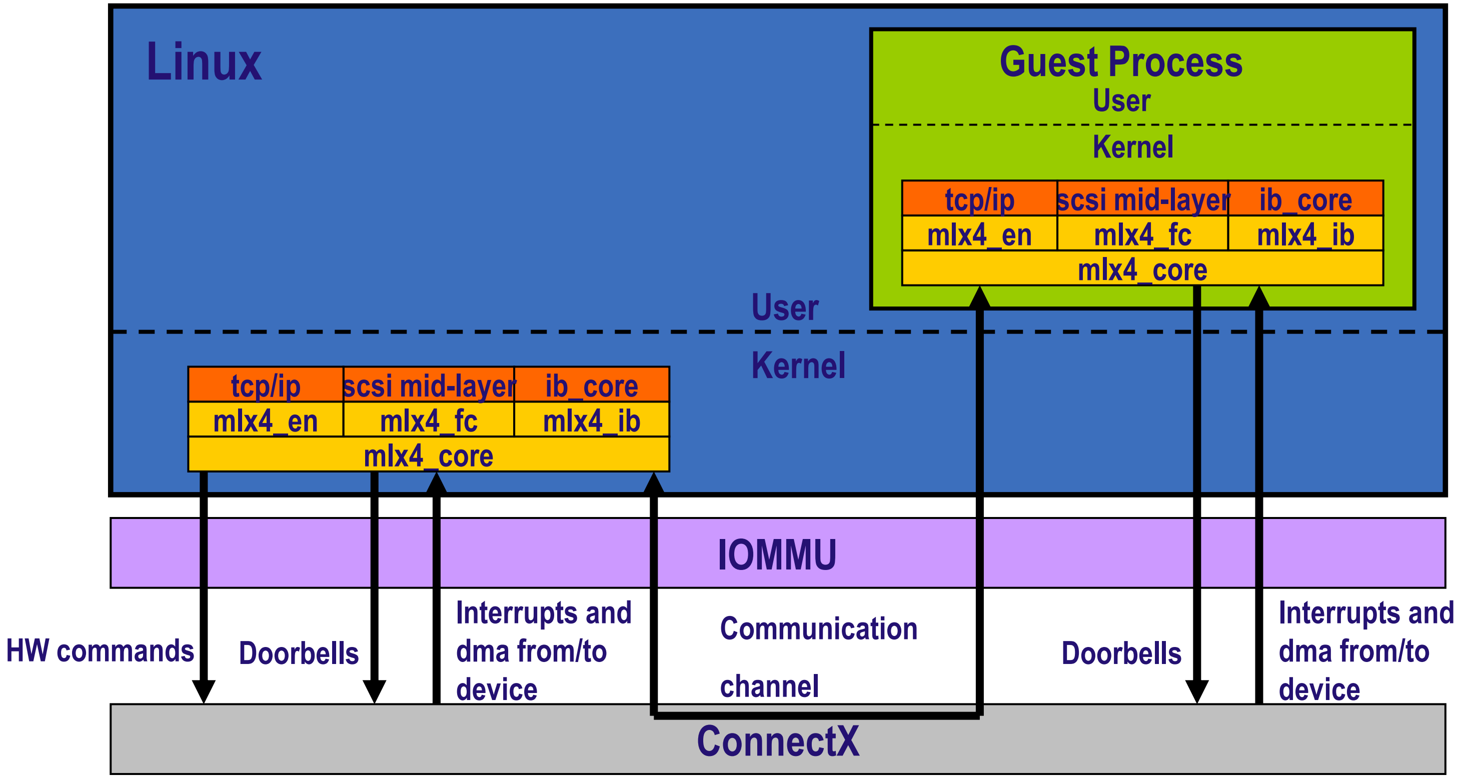
Possible APIs

- Sockets
  - – High bits of SO_PRIORITY option
- RDMACM
  - – Add equivalent RDMA_OPTION_PRIORITY
- Verbs
  - – High bits of ‘SL’ field

RoCE SRIOV

- Each VF exposes a NIC + RoCE device
  - RoCE shares the NIC MAC address
  - GID table entries populated accordingly
- HW virtual switch settings apply to both
  - MAC assignment / enforcement
  - VLAN enforcement
  - Default / allowable priorities
  - Rate limiting
- Same drivers for Hypervisor and Guest
  - To be released in MLNX_OFED-2.0

RoCE SRIOV

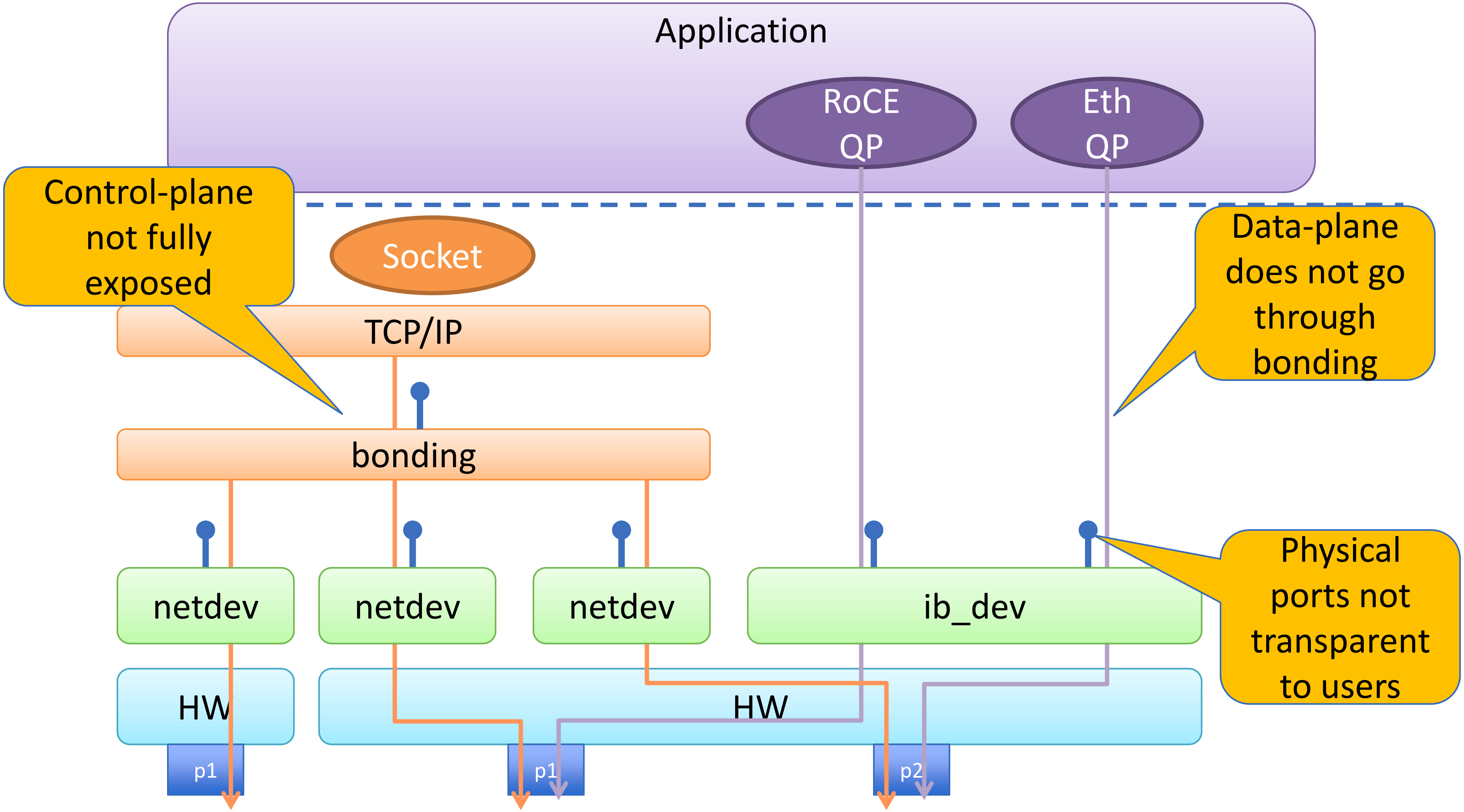

Linux
Guest Process
User
Kernel
tcp/ip
scsi mid-layer
ib_core
mlx4_en
mlx4_fc
mlx4_ib
mlx4_core
User
Kernel
tcp/ip
scsi mid-layer
ib_core
mlx4_en
mlx4_fc
mlx4_ib
mlx4_core
IOMMU
HW commands
Doorbells
Interrupts and dma from/to device
Communication channel
Doorbells
Interrupts and dma from/to device
ConnectX

HA/LAG Solutions Today

- Linux Bonding
  - Active-Backup, transmit/receive load-balancing, 802.3ad
  - Applies only to network interface traffic
- RDMACM bonding support
  - Active-backup only
  - Binds to active RDMA device upon connection establishment
- APM
  - Currently IB only but could be enabled for RoCE
  - Applicable for RC and RD EECs
- In middleware
  - Not generic

No single solution fits all!

Bonding as Unified Solution?



Control-plane not fully exposed

Data-plane does not go through bonding

Physical ports not transparent to users

Device LAG Support

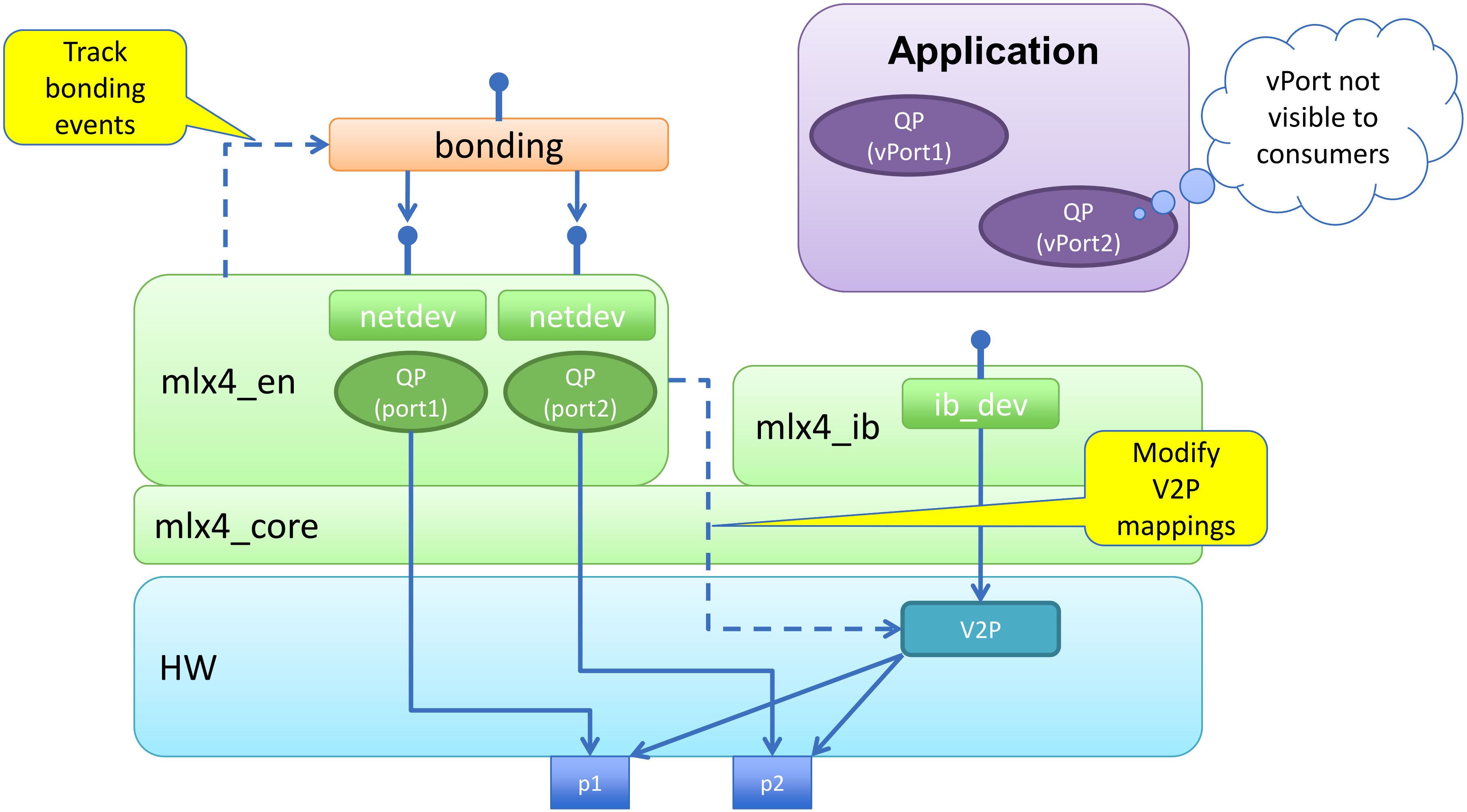
- Manage LAG/HA data-plane by the device
  - A single “bonded” port is exposed to the OS
  - Load-balancing and failover handled below the Verbs
- Bonding modes:
  - 802.3ad
  - Active-Backup (fail-over MAC disabled)
- Complete, coherent HA management for all protocols
  - Ethernet interface
  - RoCE kernel ULPs and user applications
    - Including RC QPs – no need for APM !!!
  - Raw Ethernet QPs

Configure once, apply to all

Implementation

- Bind control-plane to bonding driver
  - Centralized place for LAG/HA configuration
  - Leverage bonding modes and options of existing code
- RoCE/Raw-Ethernet QP internal configuration
  - On Rx, may receive from both physical ports
  - On Tx, each QP is associated with a ***virtual*** port, which can map to any physical port
    - QPs are distributed between virtual ports for load balancing
    - Virtual ports are assigned to different physical ports if available
- Ethernet driver tracks bonding decisions
  - Modifies the Virtual-to-Physical (V2P) port mappings accordingly
  - Offloaded traffic is sent to mapped physical port

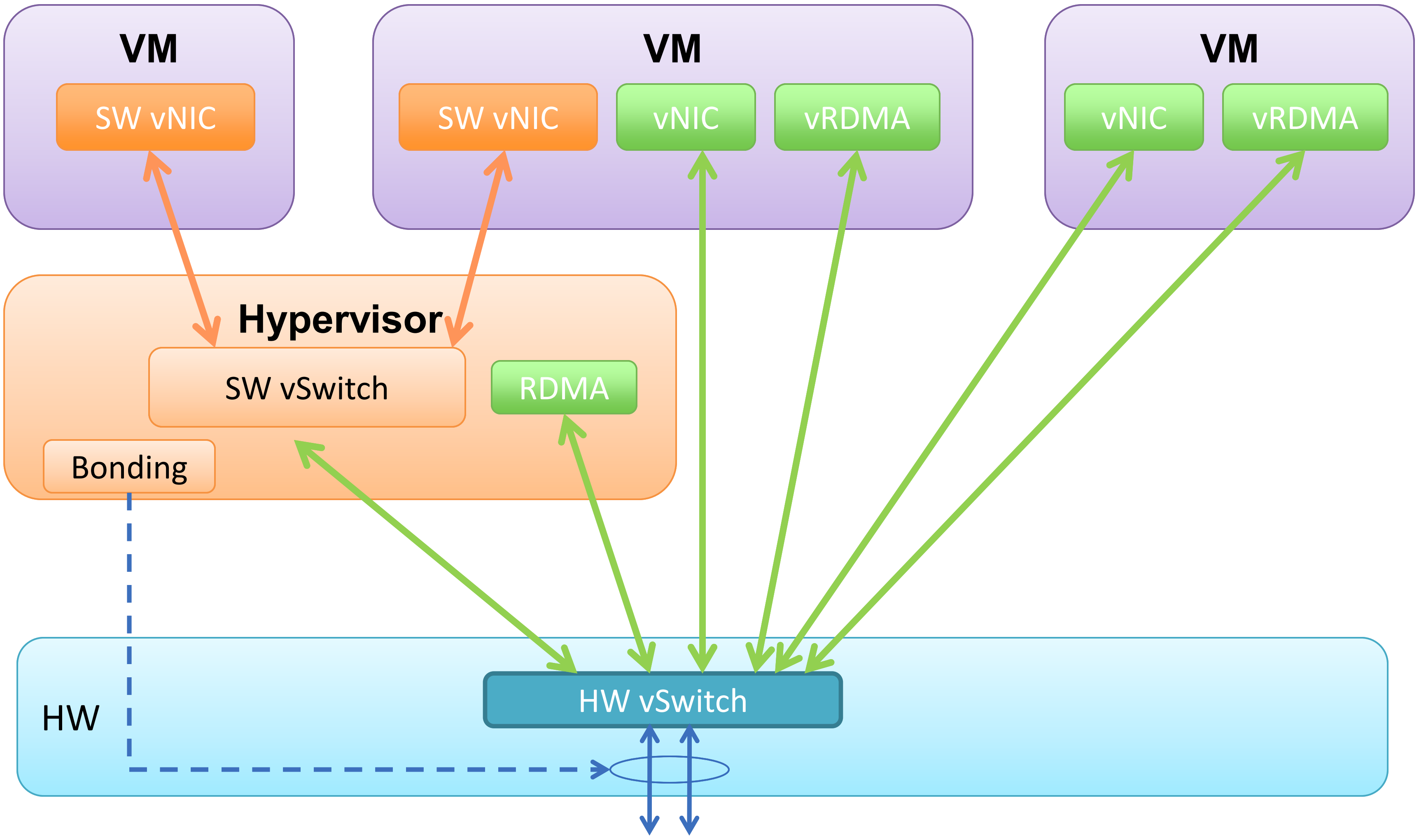
Implementation



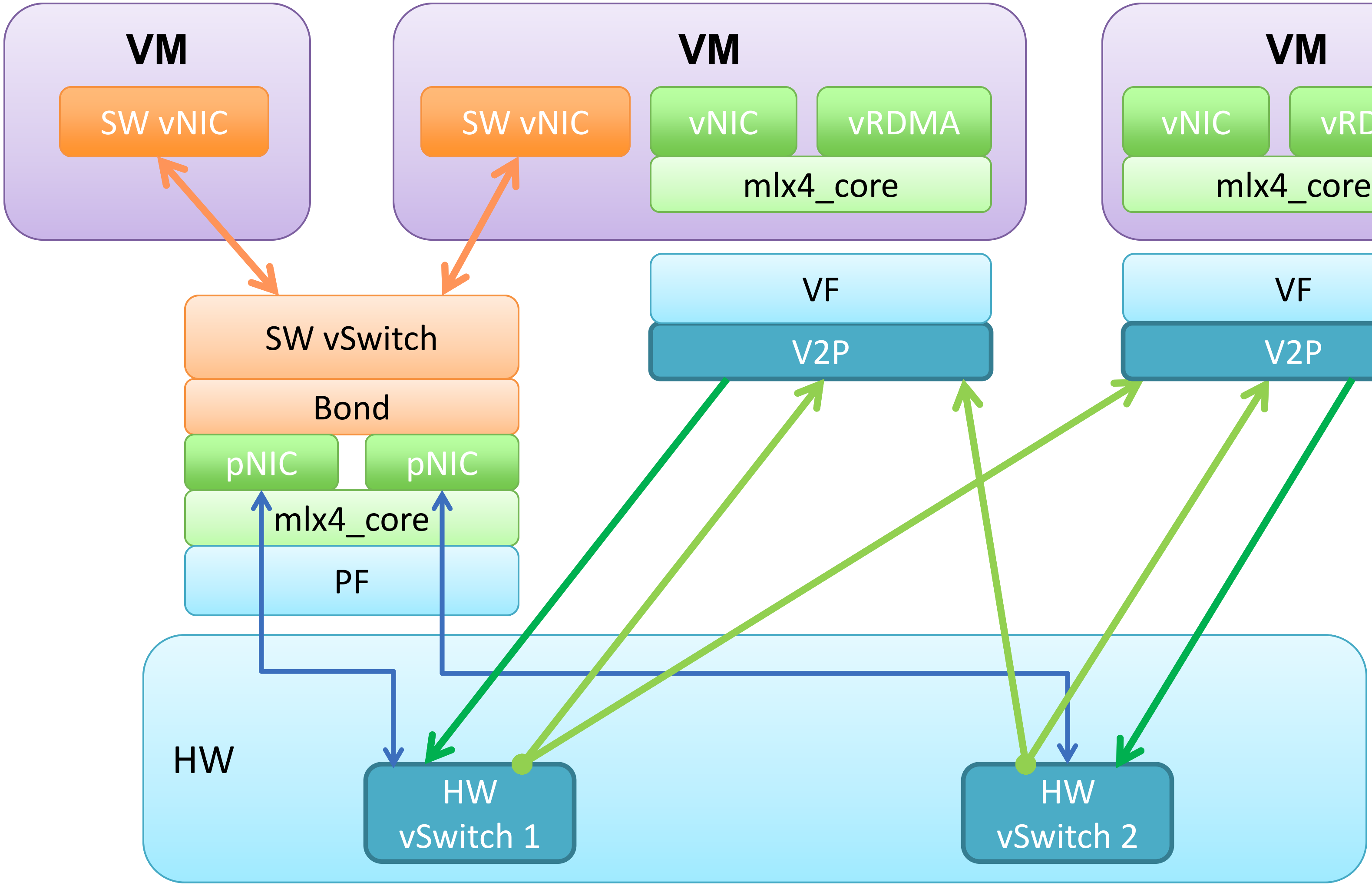
Extensions to SRIOV

- SW virtual switch configurations often have more than a single uplink
  - Uplinks are teamed for LAG/HA
  - vNICs have a single link
  - Teaming is accomplished transparently to the vNIC
- Same model can be used for SRIOV
  - Only a single network interface is passed to the VM
    - Single VF with one link
  - No bonding configuration required in the VM
  - VF supports both vNIC and vRDMA

LAG Virtualization Model



Implementation

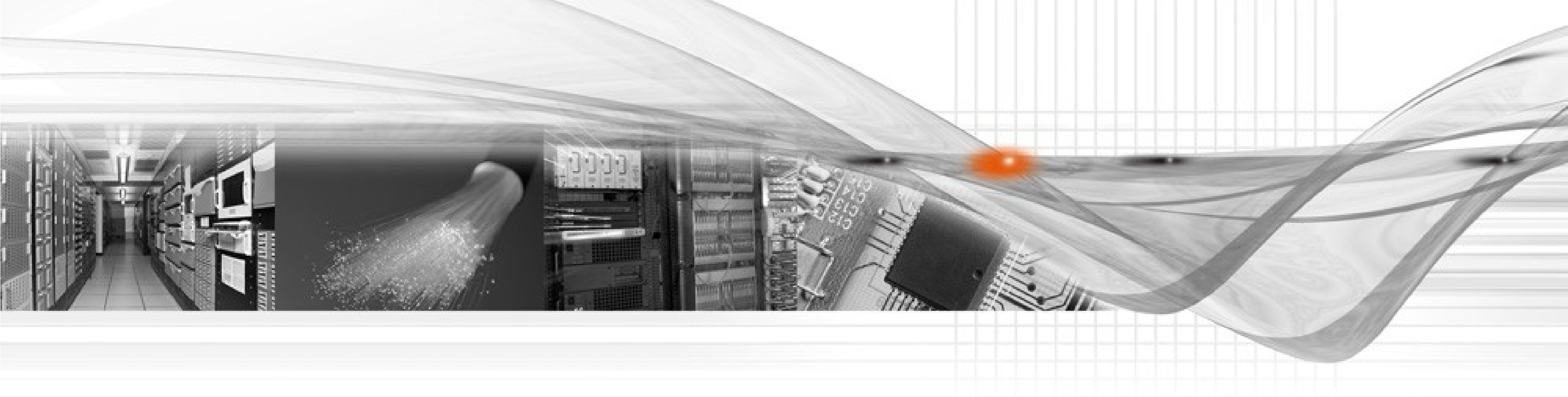


VF Port Assignment Options

- Distribute VFs between ports
  - Each VF is associated with a single port
  - On port failure, VF migrates with its MAC
  - Bonding modes: Active-Backup, 802.3ad
- Distribute traffic between ports within VF
  - Ethernet driver hashes traffic according to L2/3/4
  - Each RoCE/Raw-Ethernet QP is assigned to a single port (as in the non-virtualized case)
  - Bonding mode: 802.3ad

Latest News

- RoCE for Windows included in Mellanox WinOF 3.0
    - Submitted for Windows-8.0
- RoCE RDMA support in SMB-2.2 (Tom Talpey, OFA’12)
- VMware migration over RoCE (Bhavesh Davda and Josh Simons, OFA’12)
    - 36% improvement in vMotion time
    - >30% higher pre-copy BW
    - >90% reduction in CPU utilization



OPENFABRICS ALLIANCE

Thank You!