Simplify network configuration for VMs by harmonizing multiple Bridging, QOS, DCB and CNA implementations

Shyam Iyer

Background..

- Ethernet Network controllers
  - Network driver(registered a netdev device)
  - VLAN configuration
    - vconfig
  - Remote Management
    - IPMI etc
  - Link management
    - Ethtool
  - MTU configuration
    - ifconfig etc
  - Other tools
    - iproute2

Background..

- Storage controller cards
  - SCSI low-level driver(registered as a scsi host device)
  - Link types
    - SCSI link
    - SAS link
    - FC link
  - Controller Management
    - Each vendor maintained management API/library
      - Used stubs in /sys/ or interacted via ioctls

Ethernet evolution

- Traditional Ethernet
  - Best-effort
  - Lossy

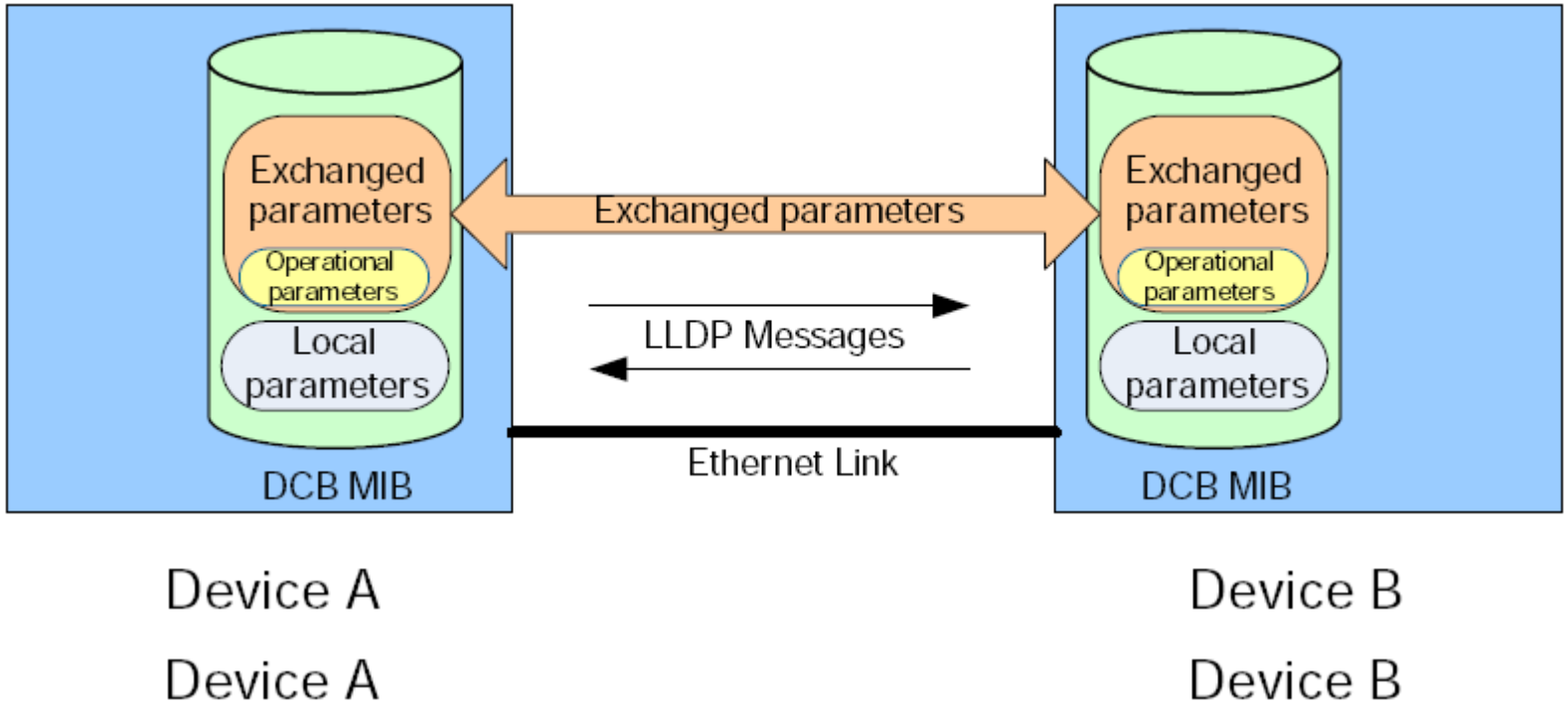
  But a few things changed..
- Ethernet Roadmap advanced faster than Fibre Channel roadmap
- Ethernet started becoming increasingly the medium of choice for the bandwidth hungry storage world

  And thus a standards race ensued…
  - DCE = Data Centre Ethernet
  - CEE = Convergence Enhanced Ethernet
  - DCB = Data Centre Briding

Data Centre Bridging

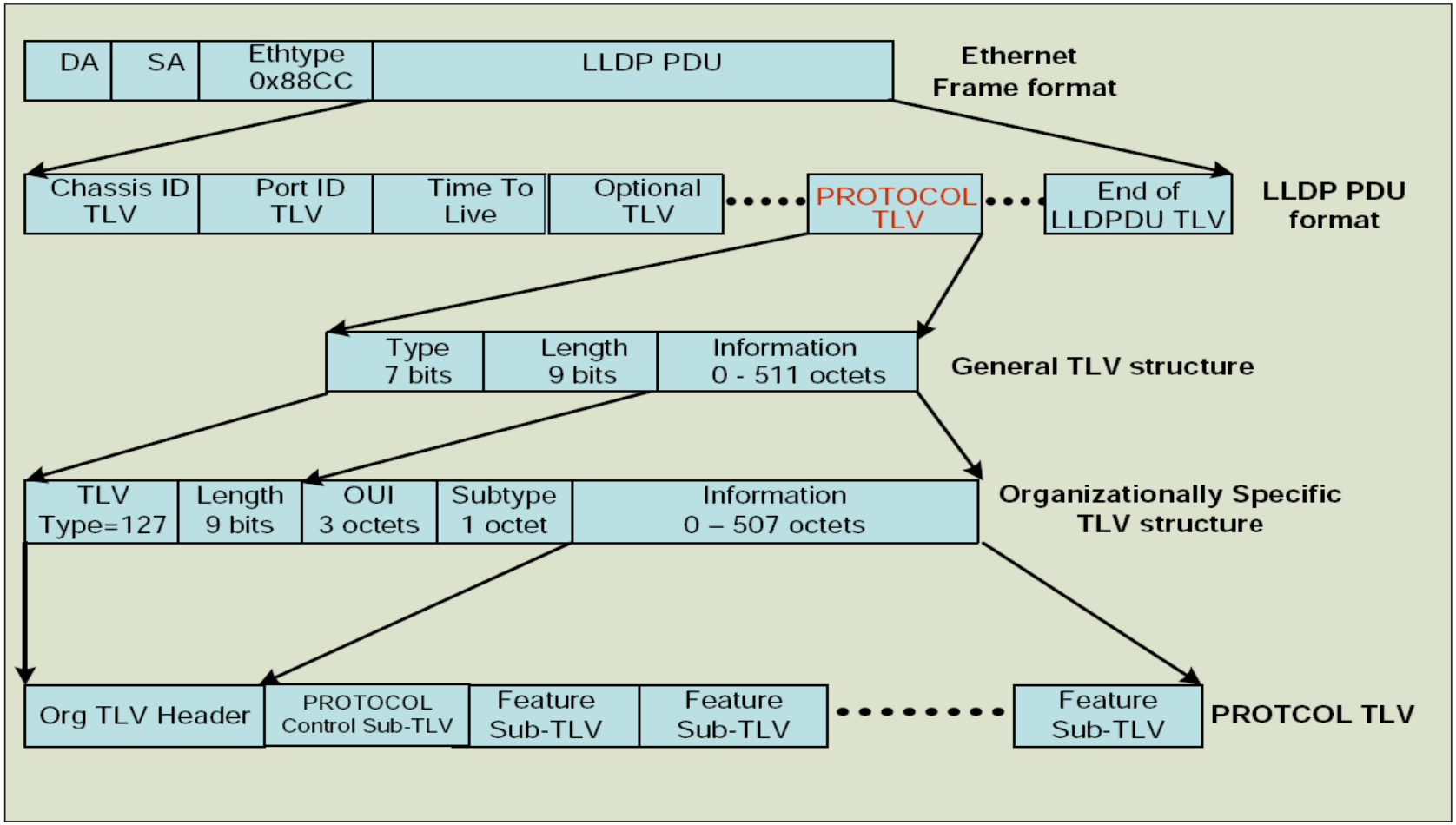
- Part of the IEEE 802.1 Working Group
- 802.1 Qau
  - Congestion Notification
- 802.1Qaz
  - Enhanced Transmission Selection
- 802.1Qbb
  - Priority-based Flow control
- DCBX
  - Protocol to exchange configuration/capabilites of above features

Data Centre Bridging – Exchange protocol



Source: http://www.ieee802.org/1/files/public/docs2008/az-wadekar-dcbx-capability-exchange-discovery-protocol-1108-v1.01.pdf

Data Centre Bridging TLV



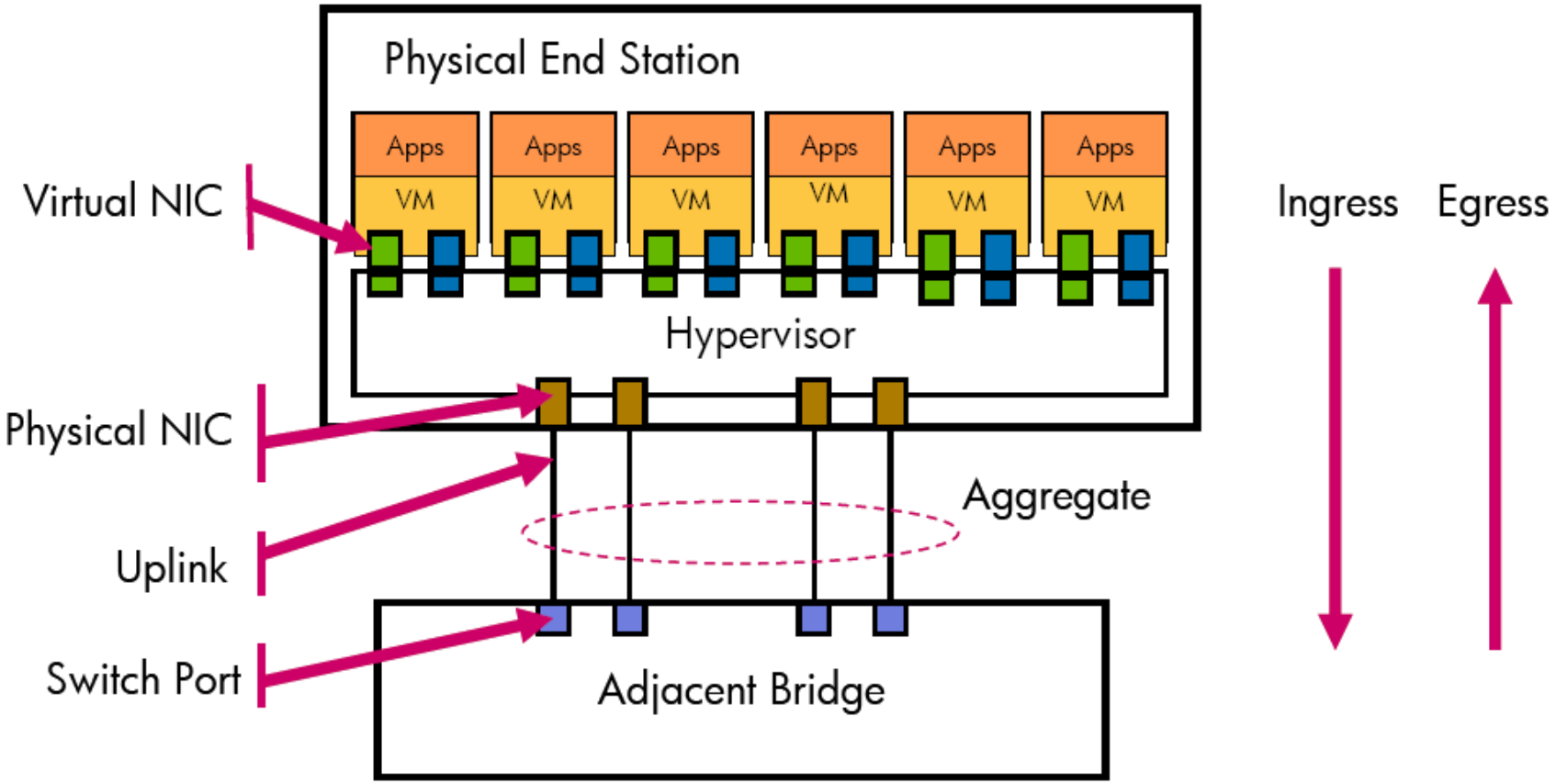
Source: http://www.ieee802.org/1/files/public/docs2008/az-wadekar-dcbx-capability-exchange-discovery-protocol-1108-v1.01.pdf

Situation

New Chips on the block

- Ethernet Controllers
  - ethX
- Converged Network Adapters
  - ethernet function
  - iSCSI function
  - FCoE function
- Nic Partitioning
- SR-IOV

Bridge Networked VM configurations



Source: http://www.ieee802.org/1/files/public/docs2010/bg-joint-evb-0410v1.pdf

Bridging Modes

- VEB
  - Restrict VM to VM traffic to the switch
  - Hypervisor vSwitch
  - CNA vSwitch
  - SR-IOV
- VEPA
  - Push VM to VM traffic to the switch
  - Hypervisor vSwitch
  - CNA vSwitch
  - SR-IOV

Complication

Vendors Implementations..

- Nic vendors
  - Vlan management
    - vconfig
  - DCB management
    - dcbnl
  - Multiqueue
  - QOS implementation
    - Tc filters, qdisc

Vendors Implementions...

- iSCSI offload vendors
  - Vlan management
    - iSCSI iface configuration(passes details via a library to offload firmware)
    - Vconfig
- FC offload vendors
  - Vlan management
    - Sysfs attributes
    - Switch based operation

QoS implementations

- Per VM QoS configuration
  - Storage QoS configurations
    - Per VM I/O priority via blkio controller cgroup for I/O controllers. Libvirt can provides an api for the same.
    - Load balance an aggregated/multipathed I/O path to a LUN.
  - Network QOS configuration
    - Per VM network I/O priority can be configured via tc and lldpad
    - Bonding Modes ? (Does specification allow this)
    - Bonding multiqueues

Solution

Implementation outline..

- Make dcbnl the defacto kernel interface for communicating DCBX parameters
- Move net/dcb to a generic subsystem that is not netdev specific
- CNA drivers register dcb configuration related functions to dcbnl
- CNA/Offload drivers provide hooks to manage vlans via standard OS utilities. Support CNAs via iproute2
- Exporting storage, FCoE, CNA attributes – Arrive at a consensus.
- Harmonizing different QoS layers.
  - Set QoS policy via a libvirt XML configuration
  - Libvirt api reads policy and make appropriate configuration decisions.
  - Eg: A policy setting will make a decision to select the appropriate load-balancing algorithm for multipathed I/O path on a DCB network
    - Round-robin for equal priority traffic and queue-length or service time algorithm for unequal priority traffic.

References

- http://www.ieee802.org/1/files/public/docs2008/az-wadekar-dcbx-capability-exchar
- http://www.ieee802.org/1/files/public/docs2010/bg-joint-evb-0410v1.pdf
- http://www.ieee802.org/1/pages/dcbridges.html
- http://www.kernel.org/doc/ols/2009/ols2009-pages-297-302.pdf

Credits

•John Fastabend – Intel
•Chad Dupuis - Qlogic
•Debashis Dutt – Brocade
•Benjamin Li – Broadcom
•Mike Christie – Redhat
•Matt Domsch – Dell
•Gaurav Chawla - Dell

Thank You

Backup

- Congestion Notification (CN) provides end to end congestion management for protocols that do not already have congestion control mechanisms built in; e.g. Fibre Channel over Ethernet (FCoE). It is also expected to benefit protocols such as TCP that do have native congestion management as it reacts to congestion in a more timely manner.
- Priority-based Flow Control (PFC) provides a link level flow control mechanism that can be controlled independently for each priority. The goal of this mechanism is to ensure zero loss due to congestion in DCB networks.
- Enhanced Transmission Selection (ETS) provides a common management framework for assignment of bandwidth to traffic classes.
- A discovery and capability exchange protocol that is used for conveying capabilities and configuration of the above features between neighbors to ensure consistent configuration across the network. This protocol is expected to leverage functionality provided by 802.1AB (LLDP)

Source: http://www.ieee802.org/1/pages/dcbridges.html