OPENFABRICS ALLIANCE

13th ANNUAL WORKSHOP 2017

VALIDATING ROCEV2 IN THE CLOUD DATACENTER

Sowmini Varadhan, Santosh Shilimkar

Oracle Corporation

[March 31, 2017]

AGENDA

- RoCEv2 overview and requirements background
- Validation objectives; environment description
- Testing methodology (work in progress), preliminary findings, lessons learned
- Configuration challenges
- Debugging challenges
- Standardization areas, observability tools, DCBX extensions, Netconf/YANG/XML based configuration models

ROCE OVERVIEW

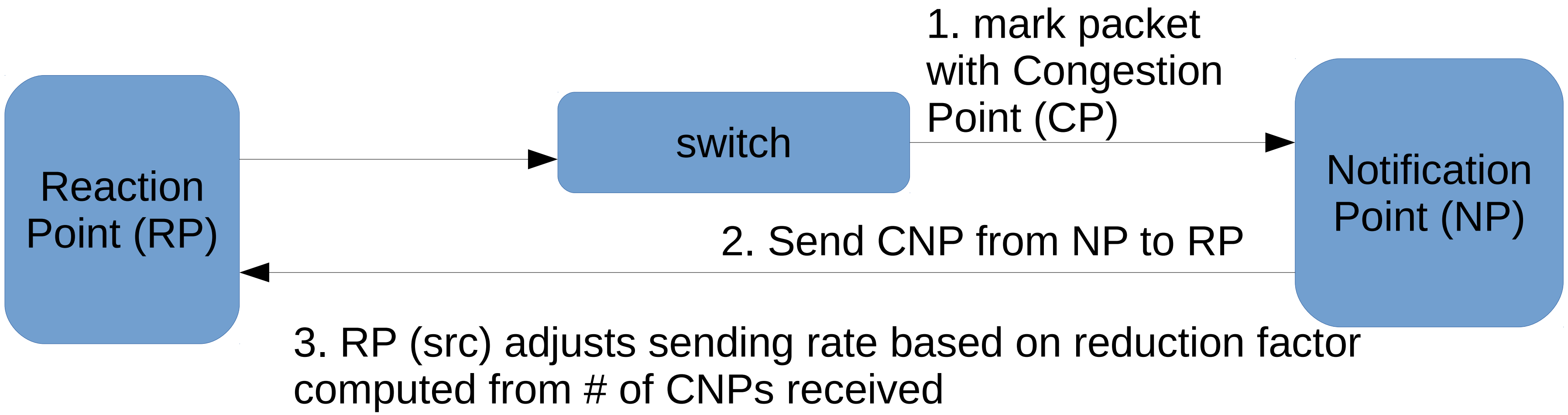
- **Datacenters are commonly built with commodity ethernet switches/NICs using TCP/IP**
- **RoCEv2 allows applications to tunnel IB frames over UDP/IP, and obtain benefits of RDMA over commodity ethernet.**
- **RDMA needs a lossless network: no packet loss due to buffer overflow at switches**
- **RoCEv2 environments achieve this through ethernet Priority based Pause Flow Control (IEEE 802.1qbb)**
  - RDMA traffic is assigned a unique priority
  - end-hosts and switches are configured to treat that priority as a lossless class
  - If the switch sees impending packet drop for a lossless class, it sends a PFC frame for that priority to flow-control sender
- **Priority based PFC only pauses the congested priority; it avoids the head-of-line blocking with the older IEEE 802.3X based PFC, where all flows were blocked when one flow congested the link**

CONSTRAINTS OF PFC

- **Known constraints with priority based PFC in the datacenter, e.g., see https://www.microsoft.com/en-us/research/publication/rdma-commodity-ethernet-scale/**
  - PFC works hop-by-hop, so there is a propogation delay if there are multiple hops between src and destination
  - Livelock seen for BUM (broadcast/unknown-unicast/multicast) traffic: many cloud operators are therefore reluctant to enable PFC in the CLOS
- **To mitigate some of these effects, flow-based congestion management is possible via DCQCN (Datacenter QCN) . DCQCN is similar to ECN used for TCP/IP networks**
- **DCQCN allows the sender to react to queue-lengths at intermediate senders, and flow-control the sending rate, reducing the number of PFC pause frames**
- **PFC pause frames are still the last line of defense, but in most cases, DCQCN adjusts the flow rate to existing traffic congestion.**

DCQCN OVERVIEW

- **Quantized Congestion Notification (QCN) enables flow-level congestion at L2**
    - Flows are defined using src/dst mac address and the flow-id field
    - Switch computes congestion metric (based on instantaneous queue size) and sends feedback to the source of the arriving packet
    - Source uses feedback to adjust sending rate
- **DCQCN extends this for IP routed networks (L3) and builds on ECN mechanisms used for TCP/IP (RFC 3168, DCTCP)**



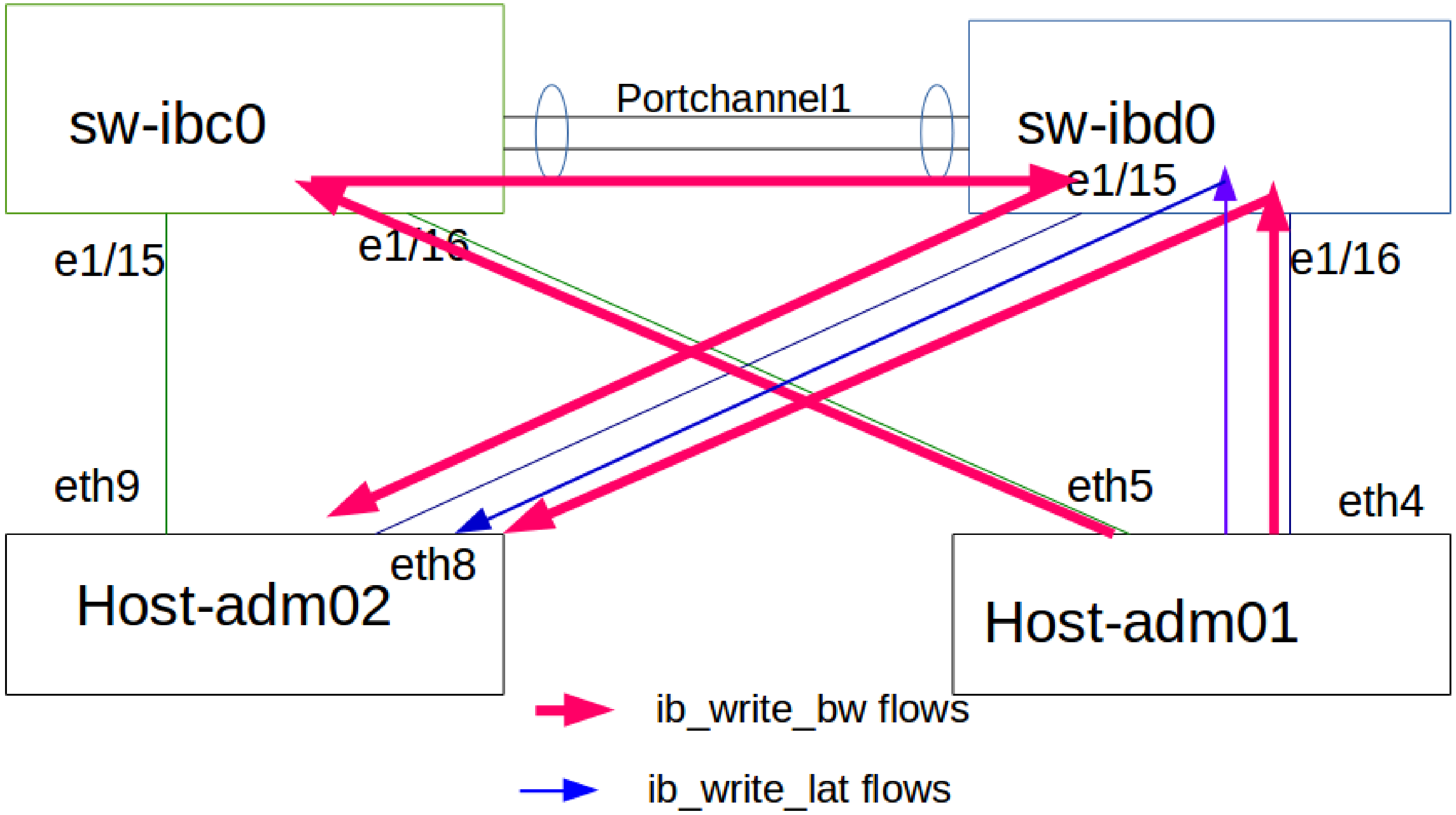
OBJECTIVES OF VALIDATION

- **RoCEv2 is standardized, so should have interoperability between vendors**
- **When transitioning from RDMA-IB to RoCE, need to figure out how to provision the system: set up PFC, ECN, virtual lanes so that RDMA semantics expected by the application are available in RoCE.**
- **Objectives:**
  - validate ease-of-use, perf profiles, configuration management, monitoring/observability in RoCE
  - Identify any areas where interoperability/standardization needs improvement in RoCE

ENVIRONMENT DESCRIPTION

- **End systems: MOFED with Connect-X4 NICS**
- **Switches: MLNX-OS (Spectrum)**
- **Testing methodology: basic validation**
  - ib_tools: all permutations of b/w and latency testing for ib_read, ib_write, ib-send were done. With a single qpair, ensure that B/W is between 92-96 Gbps. Latency (with a single switch between the ib src/dst) is about 3-4 µsecs
  - Verification of UDP header: src port selection based on qpair number, VLAN header verification for various choices of user-priority
  - VLAN-based, vs DSCP-based, priority marking of packets. Verification of CNP generation
- **Advanced validation in progress: permutations with varying flow-types, varying priorities. Behavior under congestion, resilience to impending buffer-overflow, graceful recovery from link/path failure.**

BENEFIT OF PRIORITY BASED PFC: TWO-PORT → ONE-PORT TEST



RESULTS OF TWO-PORT → ONE-PORT TEST

- **Three cases were examined:**
  - When ib_write_lat was running without any other competing flows (baseline)
  - ib_write_lat and ib_write_bw flows running at the same priority (4)
  - ib_write_lat at a different prio (3) than ib_write_bw
- **Experiments were done both with, and without, ECN enabled at the switches**
- **Objective: investigate the effect of head-of-line blocking on the ib_write_lat flow**
- **Reported ib_write_lat latency**
  - **Baseline latency was 3.22 microseconds**

Priority used for flows: (ib_write_lat , ib_write_bw)	**ECN enabled on switches**	**ECN disabled on switch**
(4, 4)	5.67 µsecs	15.48 µsecs
(3, 4)	4.28 µsecs	5.64 µsecs

- **Conclusions:**
  - Priority based PFC helps reduce HOLB-delays for the ib_write_lat flow
  - ECN significantly helps mitigate latency degradation within a given priority

SIGNIFICANCE OF DCQCN

- **Experiments show that ECN is very important for managing congestion gracefully, without having to fall back to PFC**
- **DCQCN needs standardization across vendors to ensure interoperability**
- **DCTCP/DCQCN is based on the assumption that ECN is based on instantaneous queue occupancy. RFC 3168 is based on the assumption that ECN is based on average queue occupancy**
  - RFC 3168 based flow control at the sender is much more conservative and targets long internet paths in a wide-area network. Faster convergence than DCTCP
  - DCTCP assumes that the flows are microbursts, with little statstical multiplexing: a single flow can dominate a given path. Can achieve both high throughput and low delay, but slower convergence time [https://people.csail.mit.edu/alizadeh/papers/dctcp-sigcomm10.pdf]
  - Need some BCP guidance for different RoCE traffic patterns needed in this space.
  - Standardized tunables to administratively manage algorithms for reaction to CNP

TROUBLE SHOOTING THE DATAPATH

- **Tcpdump is typically used in ethernet fabric for packet-level monitoring: works well to diagnose congestion issues for TCP.**
- **Can enable packet sniffing and have copies of data packets punted to the host stack (tcpdump)**
  - This was useful for checking UDP header and detecting a bug in udp source port selection
- **But RDMA poses challenge for packet-level monitoring: perf penalty for punting a copy of the packet to the host stack is usually very high.**
- **Control plane packets (Pause frames, CNP) are not passed up to tcpdump, vendor-specific hardware counters need to be relied upon**
  - Useful stats: # of pause/ECN frames, interval between pause/ECN generation, packet drops at ingress/egress queues, bytes and packets sent/received per port/priority

CONFIGURATION/OBSERVABILITY CHALLENGES

- **Switch config: PFC/QoS, buffer provisioning, Priority Groups, Traffic Class**
  - Per-port PFC/DCBX/ECN config, reserved/shared buffer allocation
- **Host config: enable/disable RoCEv2, PFC, DCQCN config**
- **Some vendors provide a rich feature-set for buffer management and managing priority based PFC and priority groups e.g.,**
  - https://community.mellanox.com/docs/DOC-2558
  - https://community.mellanox.com/docs/DOC-2673
- **Optimal parameters for provisioning buffers are not intuitive**
- **Complexity of the rich feature set means there is a steep learning curve for the system administrator. Easy to make a mistake in config, resulting in perf anomalies that are hard to debug**
- **some automatic configuration of PFC parameters for RoCEv2 possible via DCBX**

PROTOCOLS FOR NETWORK TOPOLOGY DISCOVERY

- **In Infiniband based networks, the Subnet Manager configures all the ports, endpoints.**
- **No analogous centralized configuration manager in RoCE.**
- **DCBX (Data Center Bridging Capability Exchange Protocol), aka IEEE 802.1qaz, allows some automatic configuration of PFC parameters for RoCEv2**
- **DCBX builds on top of LLDP (Link Layer Discovery Protocol), which is an IEEE L2 ethernet protocol for devices to advertise their identity, capabilities, neighbors and L2/L3 addresses to directly connected peers**
- **LLDP is frequently used to build the network topology graph.**

WHAT IS DCBX?

- **DCBX extensions to LLDP are defined by IEEE 802.1qaz. Adds TLV extensions to LLDP to share info about PFC related parameters**
- **Allows a node to do the following**
  - Peer capability discovery (PG, PFC)
  - Feature misconfiguration detection
  - Optional modification to local configuration based on config advertised by peer
- **Spectrum supported TLVs (per-port):**
  - PFC (Priority Classes and Priority Groups config),
  - AP (Application Priority),
  - Traffic shaping TLVs: ETS-Config, ETS-Recommended

CENTRALIZED CONFIGURATION MANAGEMENT

- **DCBX allows a configuration montioring entity to sniff for LLDP packets and figure out the configuration of directly connected peers.**
  - LLDP packets are sent at intervals of TxDelay (recommended default 30s) with fast updates as defined by IEEE 802.1 qbb when the local config changes.
- **For switches and hosts that are not directly connected, a centralized configuration management system would do the following:**
  - pull RDMA state information (ECN stats, Pause counters, Pause generation intervals, RDMA bytes/packets I/O stats, buffer status) via XML
  - Push configuration state to the nodes in the datacenter
- **Typically done using Netconf/YANG for Internet Protocols**
  - Controller would push/pull config information in XML. The netconf server at the target would translate the XML to native vendor-proprietary syntax

FUTURE WORK

- **Ongoing: performance evaluation for complex multi-flow cases, HA validation when routed path changes**
- **DCBX scaling and interoperability evaluation- reduce the amount of static/manual configuration in the datacenter**
- **RoCE standardization areas:**
  - DCQCN as a standard?
  - Netconf/YANG models to allow centralized RoCE configuration management from a controller?

OPENFABRICS ALLIANCE

13th ANNUAL WORKSHOP 2017

THANK YOU

Sowmini Varadhan, Santosh Shilimkar

Oracle Corp