12th ANNUAL WORKSHOP 2016

INFINIBAND VIRTUALIZATION UPDATE

Liran Liss

IBTA MgtWG

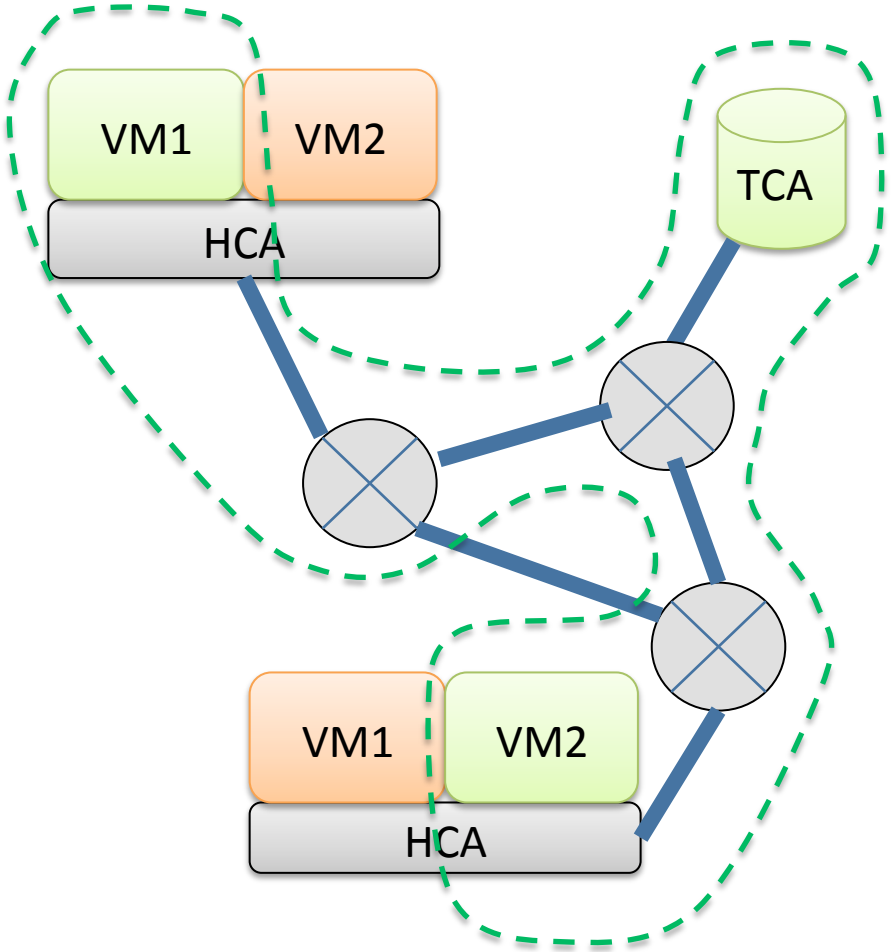
[April 4th, 2016]

AGENDA

- Infiniband Virtualization goals
- Virtual HCAs and ports
- Packet relay
- Verbs
- Subnet management
- Subnet administration
- Performance management
- Software implications

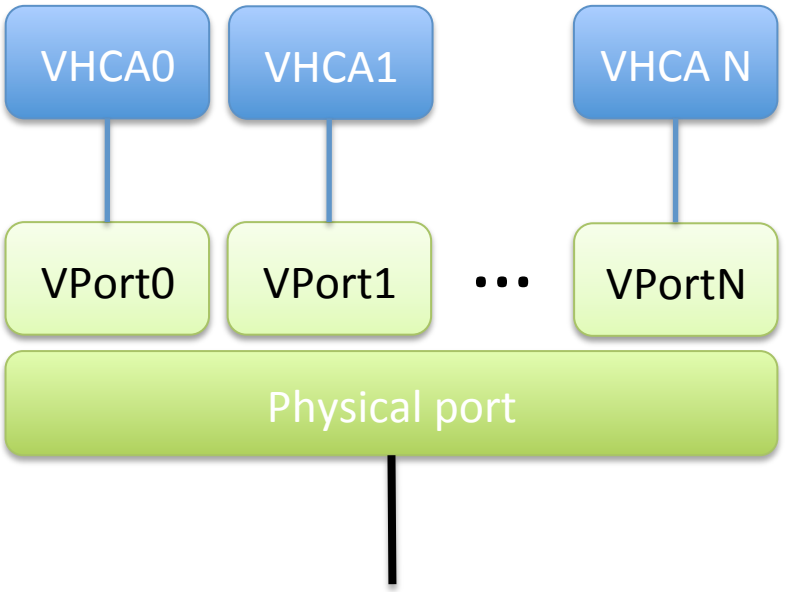
INFINIBAND VIRTUALIZATION GOALS

- **Scalable**
  - Multiple virtual endpoints
  - Efficient use of fabric resources
- **Explicit**
  - Virtual endpoints are visible to subnet management
- **Simple**
  - Management
  - Implementation
- **Backward compatible**
  - Interoperable with legacy nodes
  - Interoperable with legacy SM
    - Fall back to non-virtualizaed mode



VIRTUAL HCAS AND PORTS

- **Virtual HCA**
  - Independent consumer interface
    - Resources
    - Namespace
- **Virtual Port**
  - Provides connectivity to a VHCA
  - Light-weight transport endpoint
  - Share physical link



VPORT PROPERTIES

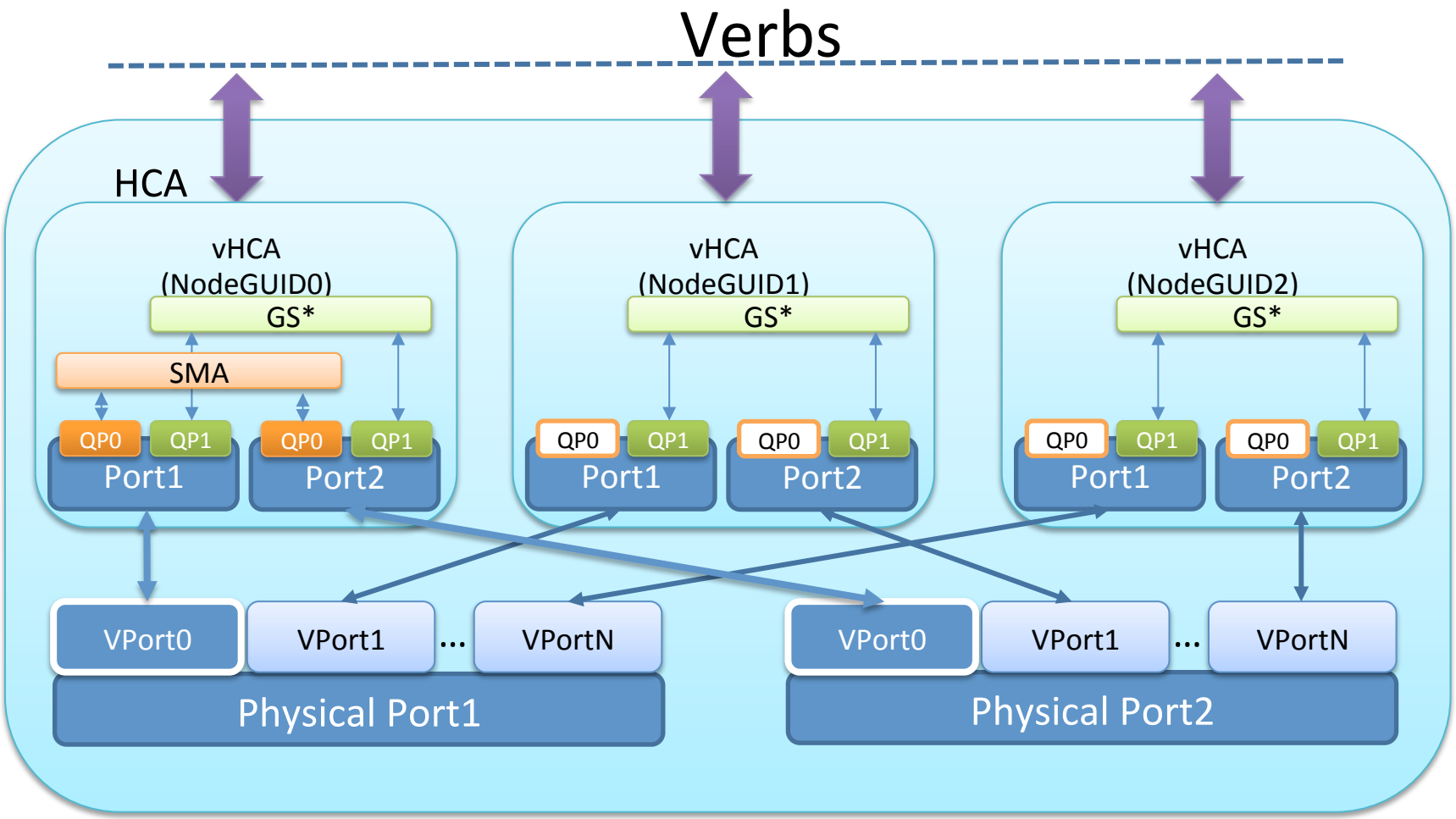
- **Per VPort**
  - GID Table
  - P_Key Table
  - (Logical) PortState
  - Capability Mask
  - P_KeyViolations counter
  - Q_KeyViolations counter
  - LID (optional)
  - Profile
  - SL mask
- **Shared by physical port**
  - LID, LMC, SL2VL, VL arbitration, etc.

VPORT TYPES

- **VPort0**
  - Privileged, backward-compatible to non virtualization aware environments
- **Other VPorts**
  - Non-privileged

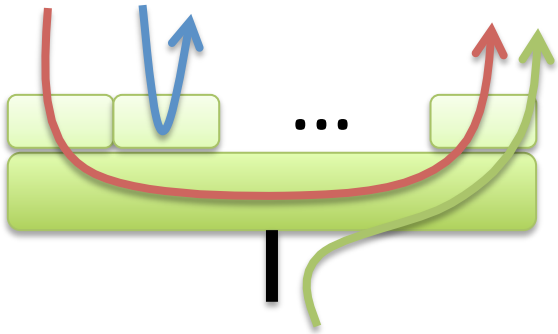
	VPort0	VPortN; N>0
GID table	Mirrors physical port	Independent
P_Key table	Mirrors physical port	Independent
Capabilities	Mirrors physical port	Independent
SMP traffic	Yes	No
Raw Ethertype traffic	Yes	No
Raw IPv6 traffic	Yes	No
GMP traffic	Yes	Yes

OVERALL PICTURE



PACKET RELAY (SHARED LID)

- **Unicast**
  - Packets are relayed to VPorts according to their DGID
  - VPort0 receives default traffic
    - Packets whose DGID does not match any GID Table
    - Packets without a GRH



- **Multicast**
  - Delivered to any QP attached to the packet MGID

- **Loopback**
  - Within VPort if either
    - (DGID matches VPort.GIDTable) && (DLID == Port.LID)
    - Loopback indicator is set (either on QP or Address Handle)
  - Within physical port if
    - (DLID == Port.LID)

- **Extension to LID-assigned VPorts is straightforward**

VERBS

- **OpenHCA**
  - Returns a handle to a VHCA
  - Regardless of whether Virtualization was enabled by the SM

- **QueryHCA**
  - CA attributes pertain to VHCA resources
  - The following Port Attributes correspond to the associated VPort
    - PortState
    - P_Key and GID Tables
    - P_Key and Q_Key violation counters
    - CapabilityMask bits
  - GRH-required Indicator

VERBS

- **ModifyHCA**
  - The following Port Attributes correspond to the associated VPort
    - Optional shutdown port indicator
    - Q_Key Violation counter reset bit
    - CapabilityMask bits (IsSM applicable only to VPort0)
  - Optional InitType value (VPort0 only)

- **Asynchronous events**
  - Affiliated events and errors are delivered to the corresponding VHCA
  - Unaffiliated asynchronous events and errors
    - PortActive issued when VPort PortState transitions to Active
    - PortError issued when VPort PortState transitions from Active to another state
    - PortChange event issued when
      - VPort GID or P_Key tables change
      - Physical PortInfo fields change (e.g., MasterSM LID)
    - ClientReregsitration issued when SM triggers it on either the VPort or physical port

SUBNET MANAGEMENT

- **Virtualization support**
  - Indicated by a PortInfo:CapabilityMask2 bit – IsVirtualizationSupported

- **VirtualizationInfo Attribute**

Component	Access	Description
VPortCap	RO	Maximum supported VPorts
VPortIndexTop	RO	Top index of enabled VPort
VirtualizationEnable	RW	Enable VPort traffic
VClientReregister	RW	Client reregister for all VPorts
VPortStateChange	RW	Set by SMA whenever any VPort transitions to/from the Down state
VirtualizationRevision	RO	Local SMA virtualization revision
CapabilityMask	RO	Optional virtualization capabilities

SUBNET MANAGEMENT (CONT.)

- **VPortInfo Attribute**

Component	Access	Description
VPortState	RW	(Logical) PortState
VPortClientReregister	RW	Per VPort ClientReregister bit
LIDRequired	RO	Assign a unique LID to this VPort
VGUIDCap	RO	Per VPort client reregister
VPortCapabilityMask	RO	Capabilities
P_KeyViolations	RW	Local SMA virtualization revision
Q_KeyViolations	RW	Optional virtualization caps
VPortLID	RW	LID for VPorts that require it
LIDByVPortIndex	RO	LID reference for VPorts without a LID
VPortProfileID	RO	Port profile
SLMask	RW	SL Mask

- Attribute modifier provides VPort index

SUBNET MANAGEMENT (CONT.)

- **VNodeInfo Attribute**

Component	Access	Description
VPartitionCap	RO	Number of partitions
VLocalPortNum	RO	Port that received this SMP
VNumPorts	RO	Number of VHCA ports
VSystemImageGUID	RO	System Image GUID
VNodeGUID	RO	Node GUID
VPortGUID	RO	Port GUID at index 0

  - Attribute modifier provides VPort index

SUBNET MANAGEMENT (CONT.)

- **VNodeDescription Attribute**
  - Format identical to NodeDescription
  - Attribute modifier provides VPort index
- **VPortGUIDInfo**
  - Format identical to GUIDInfo
  - Attribute modifier provides VPort index and block number
- **VPortPartitionTable**
  - Format identical to P_KeyTable
  - Attribute modifier provides VPort index and block number
- **VPortState Attribute**

Component	Access	Description
VPortStateBlock	RO	List of 128 VPortState elements

  - Attribute modifier indicates block number

SUBNET MANAGEMENT (CONT.)

- **The following trap types are defined for VPorts**

Trap	Name	Type
1144	VPort Local Change	Informational
1146	VPort State Change	Urgent
1257	VPort P_Key Violation	Security
1258	VPort Q_Key Violation	Security

- **Traps 1144 and 1146 aggregate changes for all VPorts**
  - SM must query the Port to detect which VPorts have changed their state

- **Traps 1257 amd 1258 are VPort specific**
  - Notice DataDetails indicates VPort index

SUBNET ADMINISTRATION

- **VPorts access the SA via MADs with GRH**
  - DGID must be refer to well-known SA GUID
- **Partition checks apply to VPort P_Key tables**
- **VPort GIDs may be provided in the following Attributes**
  - InformInfoRecord
  - ServiceRecord
  - PathRecord
  - MCMemberRecord
  - MultiPathRecord

PERFORMANCE MANAGEMENT

- **Providers per VPort counters**
  - Similar to the PortCounterExtended Attribute

- **Counters**
  - PortXmitData
  - PortRcvData
  - PortXmitPkts
  - PortRcvPkts
  - PortUnicastXmitPkts
  - PortUnicastRcvPkts
  - PortMultiCastXmitPkts
  - PortMultiCastRcvPkts
  - PortRelayErrors
    - Accounts for SL Mask and GRH violations

SOFTWARE IMPLICATIONS

- **Applications**
  - Use GRH in Address Handle attributes
- **Host stack**
  - Extend kernel port information to indicate when a GRH required
    - Used by SA code
  - SRIOV management APIs
    - Control VHCA identify, port state, and other properties
- **OpenSM**
  - Discover and initialize VPorts
  - React to VPort state changes following traps
- **Management tools**
  - Discover and list VPorts

OPENFABRICS
ALLIANCE

12th ANNUAL WORKSHOP 2016

THANK YOU

Liran Liss

IBTA MgtWG