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-virtualized 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

<table>
<thead>
<tr>
<th></th>
<th>VPort0</th>
<th>VPortN; N&gt;0</th>
</tr>
</thead>
<tbody>
<tr>
<td>GID table</td>
<td>Mirrors physical port</td>
<td>Independent</td>
</tr>
<tr>
<td>P_Key table</td>
<td>Mirrors physical port</td>
<td>Independent</td>
</tr>
<tr>
<td>Capabilities</td>
<td>Mirrors physical port</td>
<td>Independent</td>
</tr>
<tr>
<td>SMP traffic</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Raw Ethertype traffic</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Raw IPv6 traffic</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>GMP traffic</td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>
OVERALL PICTURE

Verbs

HCA

vHCA (NodeGUID0)
GS*
SMA
QP0 QP1
Port1 Port2

vHCA (NodeGUID1)
GS*
QP0 QP1
Port1 Port2

vHCA (NodeGUID2)
GS*
QP0 QP1
Port1 Port2

vHCA
(NodeGUID0)

QP0 QP1
QP0 QP1
QP0 QP1
Port1 Port2 Port1 Port2

VPort0 VPort1 VPortN VPort0 VPort1 VPortN

Physical Port1 Physical Port2
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)
    - ClientReregistration issued when SM triggers it on either the VPort or physical port
**Virtualization support**
- Indicated by a PortInfo:CapabilityMask2 bit – IsVirtualizationSupported

**VirtualizationInfo Attribute**

<table>
<thead>
<tr>
<th>Component</th>
<th>Access</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>VPortCap</td>
<td>RO</td>
<td>Maximum supported VPorts</td>
</tr>
<tr>
<td>VPortIndexTop</td>
<td>RO</td>
<td>Top index of enabled VPort</td>
</tr>
<tr>
<td>VirtualizationEnable</td>
<td>RW</td>
<td>Enable VPort traffic</td>
</tr>
<tr>
<td>VClientReregister</td>
<td>RW</td>
<td>Client reregister for all VPorts</td>
</tr>
<tr>
<td>VPortStateChange</td>
<td>RW</td>
<td>Set by SMA whenever any VPort transitions to/from the Down state</td>
</tr>
<tr>
<td>VirtualizationRevision</td>
<td>RO</td>
<td>Local SMA virtualization revision</td>
</tr>
<tr>
<td>CapabilityMask</td>
<td>RO</td>
<td>Optional virtualization capabilities</td>
</tr>
</tbody>
</table>
### VPortInfo Attribute

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

- Attribute modifier provides VPort index
## VNodeInfo Attribute

<table>
<thead>
<tr>
<th>Component</th>
<th>Access</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>VPartitionCap</td>
<td>RO</td>
<td>Number of partitions</td>
</tr>
<tr>
<td>VLocalPortNum</td>
<td>RO</td>
<td>Port that received this SMP</td>
</tr>
<tr>
<td>VNumPorts</td>
<td>RO</td>
<td>Number of VHCA ports</td>
</tr>
<tr>
<td>VSystemImageGUID</td>
<td>RO</td>
<td>System Image GUID</td>
</tr>
<tr>
<td>VNodeGUID</td>
<td>RO</td>
<td>Node GUID</td>
</tr>
<tr>
<td>VPortGUID</td>
<td>RO</td>
<td>Port GUID at index 0</td>
</tr>
</tbody>
</table>

- Attribute modifier provides VPort index
### 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

<table>
<thead>
<tr>
<th>Component</th>
<th>Access</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>VPortStateBlock</td>
<td>RO</td>
<td>List of 128 VPortState elements</td>
</tr>
</tbody>
</table>

- Attribute modifier indicates block number
The following trap types are defined for VPorts:

<table>
<thead>
<tr>
<th>Trap</th>
<th>Name</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>1144</td>
<td>VPort Local Change</td>
<td>Informational</td>
</tr>
<tr>
<td>1146</td>
<td>VPort State Change</td>
<td>Urgent</td>
</tr>
<tr>
<td>1257</td>
<td>VPort P_Key Violation</td>
<td>Security</td>
</tr>
<tr>
<td>1258</td>
<td>VPort Q_Key Violation</td>
<td>Security</td>
</tr>
</tbody>
</table>

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

Traps 1257 and 1258 are VPort specific:
- Notice DataDetails indicates VPort index.
• **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
THANK YOU

Liran Liss

IBTA MgtWG