# Routing Verification Tools

ibutils – e.g. ibdmchk

infiniband-diags – e.g. ibsim, etc.

Dave McMillen

# What do you verify?

- Did it work?
- Is it deadlock free?
- Does it distribute routes as expected?
- What happens when pieces break?

# Simulate and Automate

- ibsim and friends allow real programs to be used
- Symmetry is very helpful
- opensm produces files with useful information
- ibdmchk is needed to answer tough questions
- Routing “failures” are often position dependent

# What If?

- You should really know what the routing engines do
- … But …
- You want to see what a routing engine does with a topology
- You want to know if a change will break things
- You want to know how routes distribute when parallel links are added/removed

# Caveats

- Simulation is only for MAD traffic
- Timing under simulation is very different
- Be aware of version differences

# Routing Verification Tools

## Discussion?

Dave McMillen

# IB Multicast

## What are the uses?

## IPoIB and arp activity

Dave McMillen

# What is Multicast?

- One source sending to any number of destinations
- Datagram Service Only
- Normally messages, can be RDMA Write
- Multicast Create/Join required to get subnet routing
- Standard Infiniband packet delivery

# What is Different About IB Multicast?

- High performance
- Data loss in fabric only on hardware error (mostly)
- Receipt into QP queues
- Destinations are MLIDs instead of LIDs

# Why Use IB Multicast?

- Multiple destinations need the same information
- Status / Statistics updates
- Well known address
- Distributed and/or Parallel Servers
- ARP (specific case of above)
- Fault tolerance
- Potentially deep queues
- Pretty good idea of delivery deadline
- But don’t forget it is datagram service

# IB Multicast

## Discussion?

Dave McMillen

# Partitions

## What are they?

## What applications need partitions?

## Dave McMillen

# Partitions and Normal Activity

- P_Key is a 16 bit value specifying a partition
- A collection of endnodes with the same P_Key in their P_Key Tables are referred to as being *members of a partition*, or *in a partition*.
- The high-order bit of the partition key is used to record the type of membership in a partition table: 0 for Limited, and 1 for Full.
- Limited members cannot accept information from other Limited members, but communication is allowed between every other combination of membership types.
- Two P_Keys have special meaning: the default partition key (0xFFFF ), and the invalid partition keys (low-order 15 bits are all zero).
- The maximum number of entries the P_Key Table can hold must be >= to one and <=to 65535.
- You might only have one, but there are always partitions.

# Partitions and Subnet Management

Every IBA port has a QP dedicated to subnet management. This is QP0. QP0 has special features that make it unique compared to other QPs.

- QP0 is permanently configured for Unreliable Datagram class of service.
- Each port of an IBA device has a QP0 that sends and receives packets.
- QP0 is a member of all partitions (i.e., can accept any packet specifying any partition).
- Only subnet management packets (SMPs) are valid
- Traffic for QP0 (i.e., SMPs) exclusively uses VL15, which is not subject to link-level flow control.

# Partitions and General Services

Every IBA channel adapter has a QP dedicated to general fabric services. This is QP1. QP1 has special features that make it unique compared to other QPs.

- QP1 is permanently configured for Unreliable Datagram class of service.
- Each port of an IBA device has a QP1 that sends and receives packets.
- QP1 is a member of all of the port's partitions (i.e., can accept any packet specifying a P_Key contained in the port's P_Key table).
- Only management datagrams (MADs) are valid
- Traffic for QP1 does not use VL15

# Where are partitions used?

- VLANs are mapped to partitions
- VLANs only exist under IPoIB
- Isolate different sets of attachments
  1) Grouped by system
  2) Grouped by interface
  3) Arbitrary
- Security in the sense of no accidental connections
- QoS
- Multicast domains are different
- IPoIB bonding in one partition can only be active/passive but if the interfaces are in different partitions it can be active/active. Note individual connections do not use both paths at the same time.

# How are partitions defined?

- Overlapping is allowed
- Each partition is defined by either a complete list of all GUIDs participating, or the special keyword “ALL”
- --Pconfig, -P, or partition_config_file option
- Default=0x7fff,ipoib:ALL=full;
- 0 for Limited, and 1 for Full results in 0xffff
- Partition flags are:

1) ipoib - indicates that this partition may be used for IPoIB, as result IPoIB capable MC group will be created.
2) rate=<val> - specifies rate for this IPoIB MC group (default is 3 (10GBps))
3) mtu=<val> - specifies MTU for this IPoIB MC group (default is 4 (2048))
4) sl=<val> - specifies SL for this IPoIB MC group (default is 0)
5) scope=<val> - specifies scope for this IPoIB MC group (default is 2 (link local))

# Caveats

- All of the parts you care about need to handle partitions
- Most code has only ever been run in default partition
- VLAN mapping may constrain choices

# Partitions

## Discussion?

Dave McMillen