OPENFABRICS ALLIANCE

12th ANNUAL WORKSHOP 2016

INFINIBAND ROUTER PREMIER

Mark Bloch, Liran Liss

Mellanox Technologies

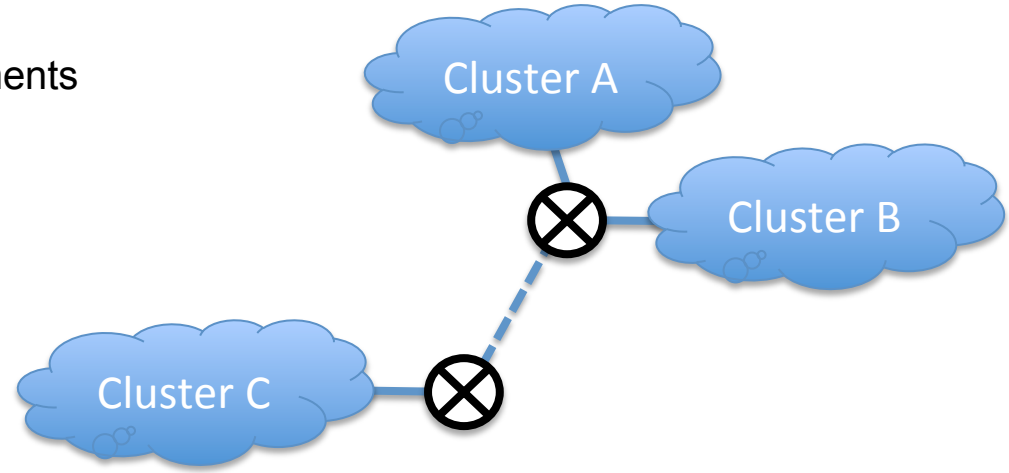
[April 7th, 2016]

AGENDA

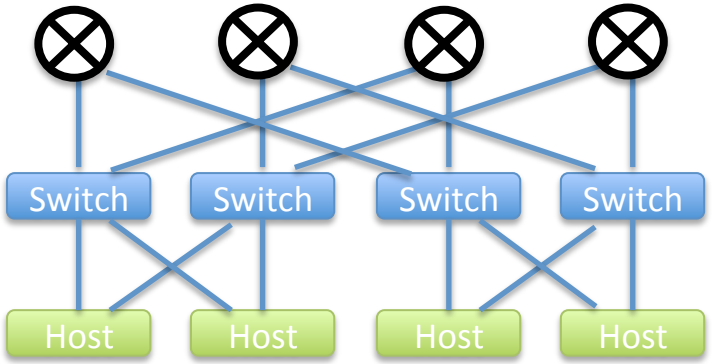
- **Why routing? Why now?**
- **Infiniband routing**
- **Host stack**
- **IB and IP(oIB) addressing**
- **Supporting arbitrary IPoIB subnets**
  - IPoIB vs. RDMACM
  - IBACM
- **IB routers and HPC**
  - Preliminary results

WHY ROUTING?

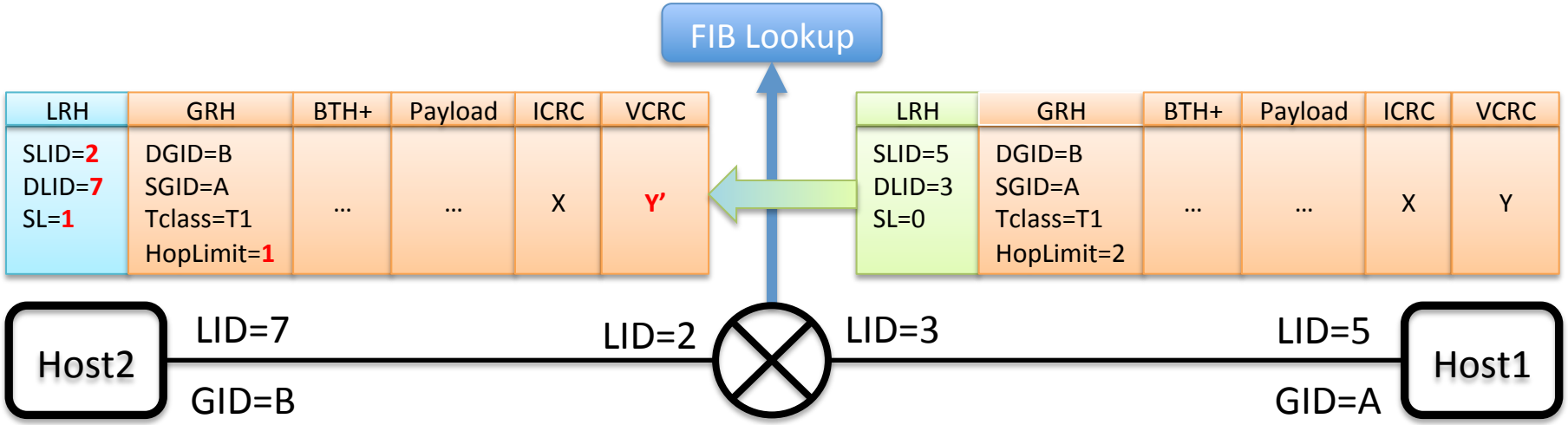
- **What changed?**
  - Complex RDMA systems and deployments
    - Interconnected appliances
    - Interconnected clusters
    - Inter data center connections
  - Exascale is here
    - 100Ks of nodes



- **Routing requirements**
  - Isolation
    - Local failures should not affect whole fabric
  - Consolidation
    - Interconnect resources provided by different IB “islands”
  - Scaling
    - Scale up addressable endpoints
    - Maintain bi-sectional bandwidth and latency characteristics of switches



PACKET RELAY



- **Packets with HopCount < 2 are discarded**
- **Tclass is preserved**
  - May be used to map incoming SL to outgoing SL
- **Partitions are global**
  - In/Out-bound P_Key enforcement in routers is optional
- **Routers may support multiple paths for a given DGID**
  - Via different next-hop routers or LMC
  - Identical GRH:FlowLabel values indicate packets for which ordering is important
- **Ordering must be maintained per <in-port, out-port, SL>**

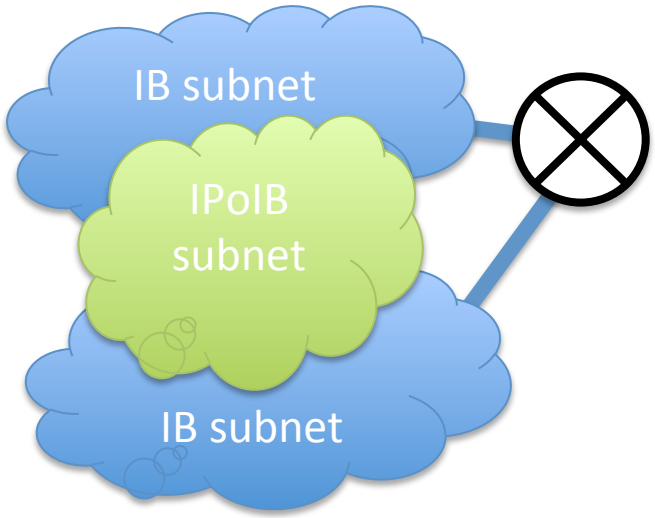
ROUTER MANAGEMENT

- **Specified**
  - Router NodeType
  - Each SM manages the router ports discovered in its own subnet
  - Endpoints obtain paths to remote destinations by querying the local SA
    - SA determines next-hop router
  - Communication management

- **Unspecified**
  - Router manager and agent entities
  - Routing MADs, methods, and attributes
  - Endpoint local interface selection

HOST STACK TODAY

- **Path queries**
  - Use standard path queries to obtain paths to remote nodes
  - If PathRecord.HopCount > 0
    - GRH is specified by AH attributes
- **Raw verbs**
  - Modify QP with AH attributes specifying a GRH
  - Create an AH with AH attributes specifying a GRH
- **AF_INET / AF_INET6 address resolution**
  - Local IPoIB interface selected by IP stack
  - SGID extracted from local interface HW address
  - DGID extracted from neighbor HW address
    - **Assumption: single IPoIB subnet spans the whole IB fabric**



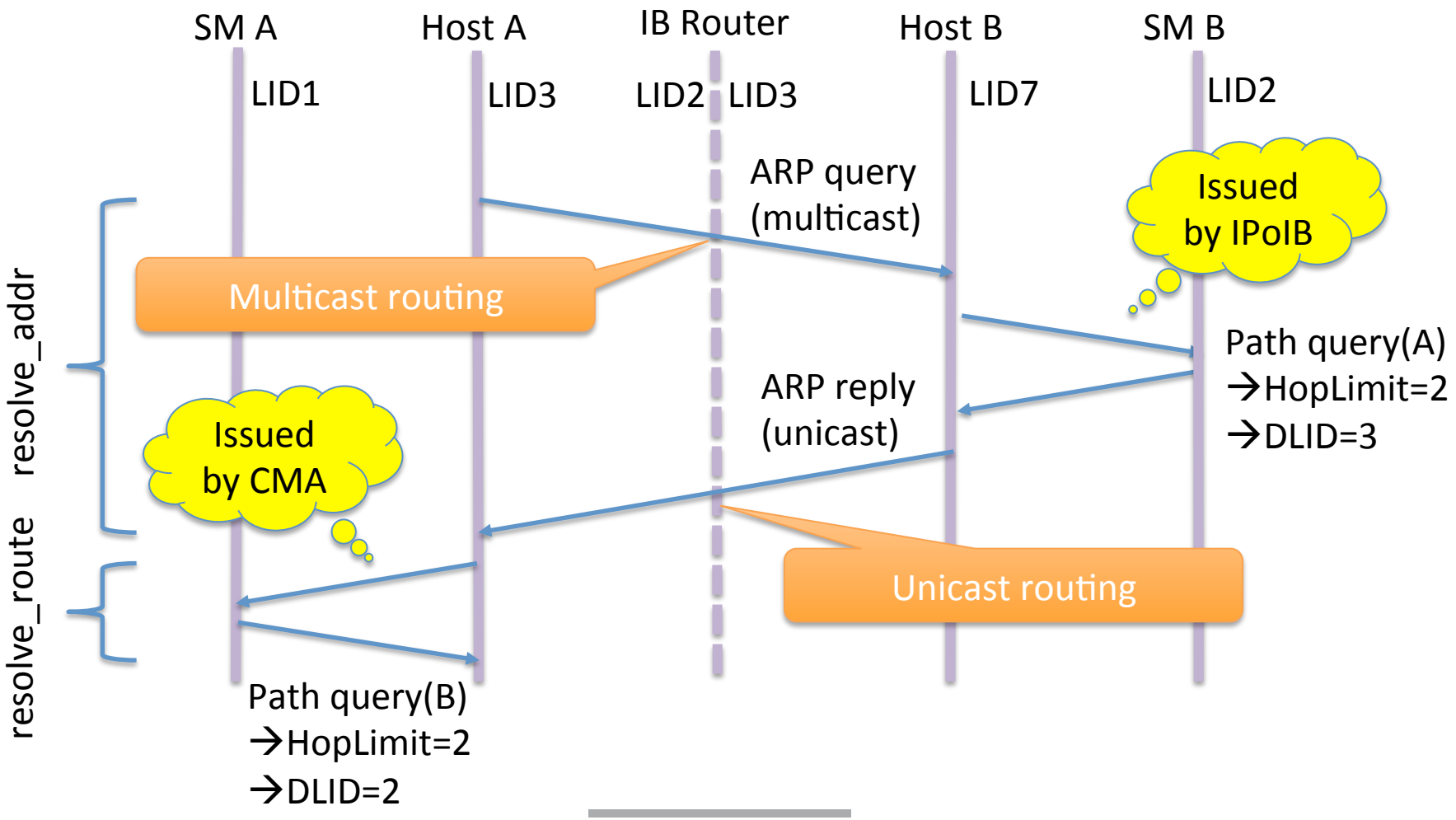
HOST STACK TODAY (CONT.)

- **AF_IB address resolution**
  - SGID must be provided by either rdma_bind_addr() or rdma_resolve_addr()
    - Used to locate local IB port
    - Choosing local port based on DGID:subprefix doesn't apply to routers !!!!
  - DGID provided by rdma_resolve_addr()

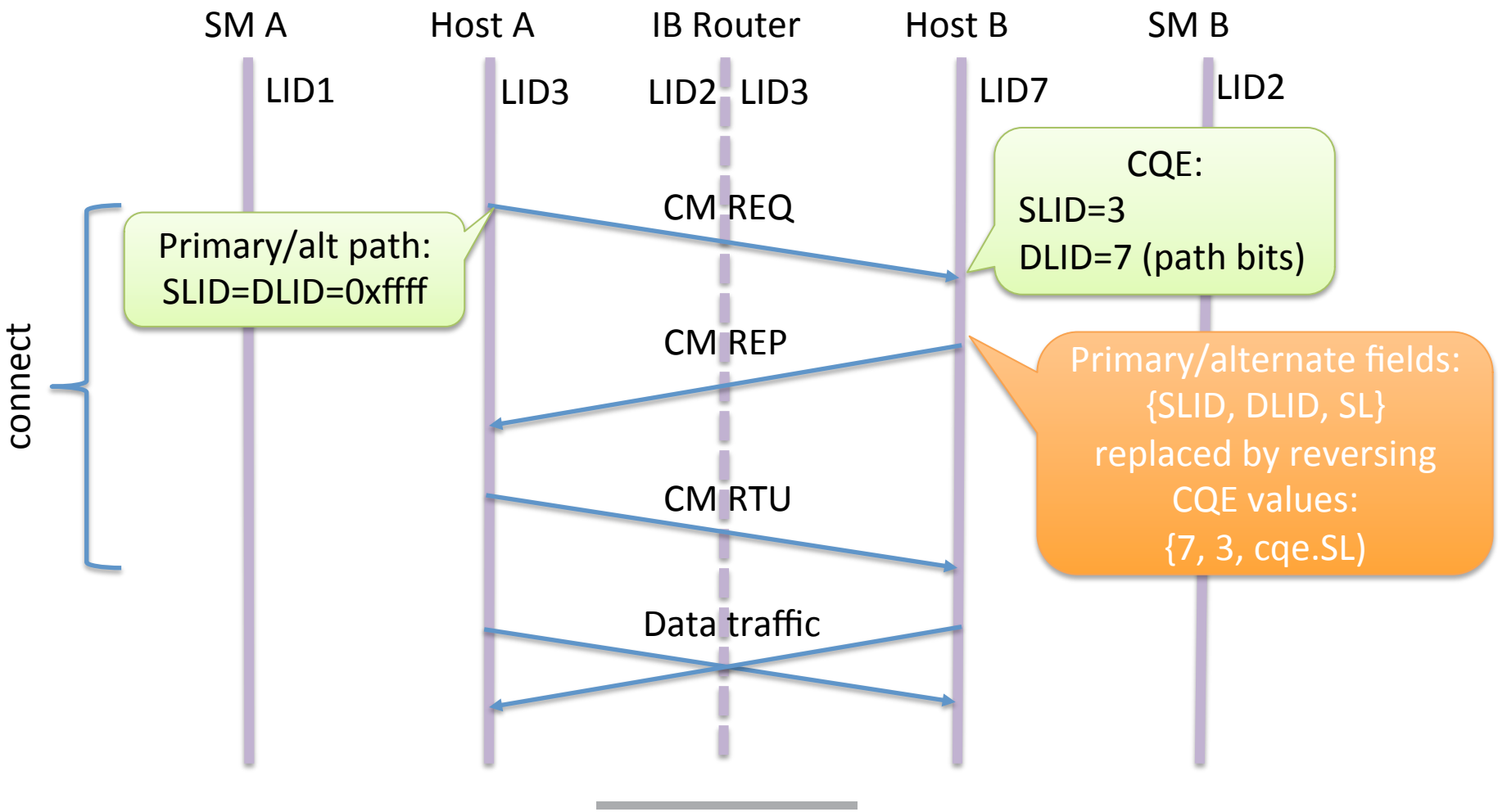
- **Connection Management**
  - No standard way to obtain remote path attributes required in CM REQ
  - On active side: set SLID = DLID = 0xffff
  - On passive side
    - If SLID == 0xffff
      - Set SLID←CQE.SLID (router LID)
    - If DLID = 0xffff
      - Set DLID ← CQE.path-bits
  - Otherwise, no change in 3-way handshake

Routing management is transparent to host stack

AF_INET - PUTTING IT ALL TOGETHER (1/2)

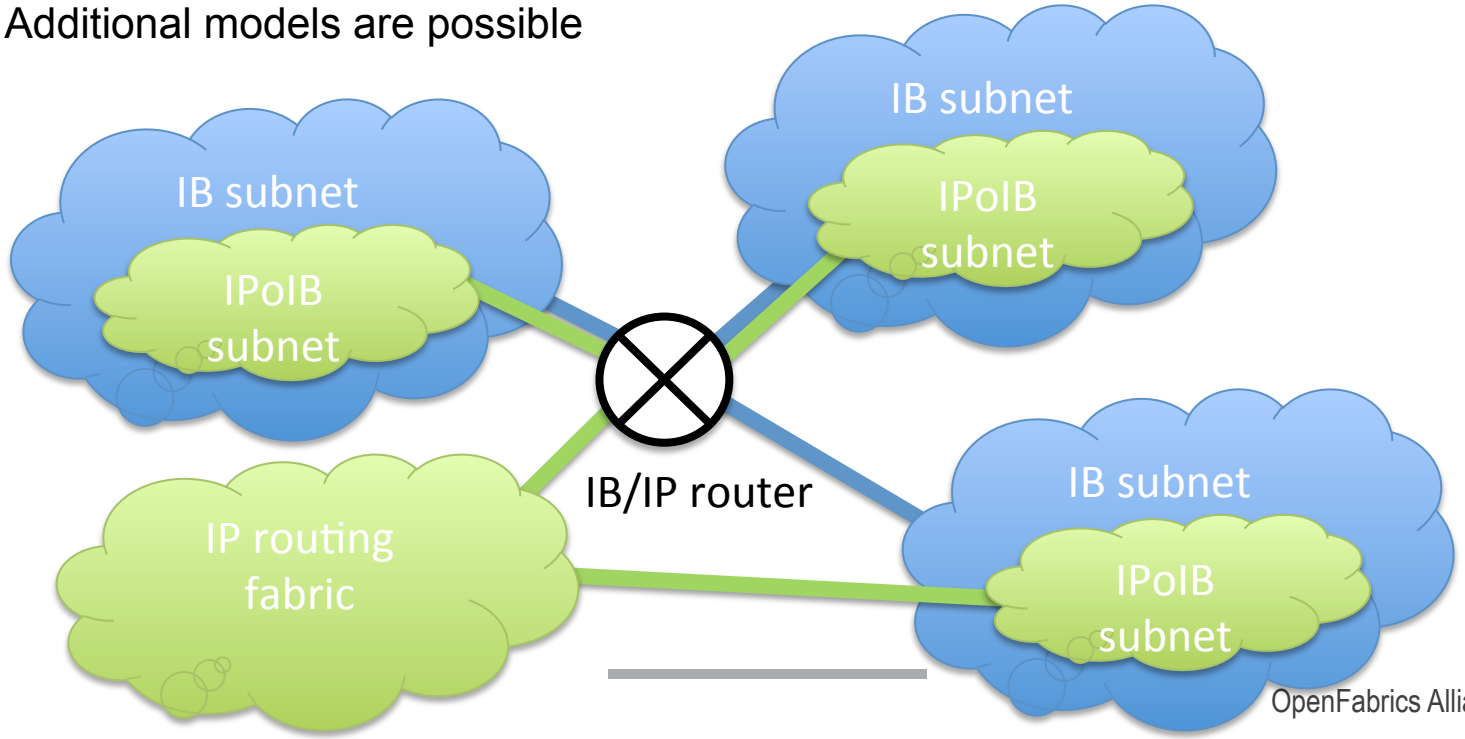


AF_INET - PUTTING IT ALL TOGETHER (2/2)

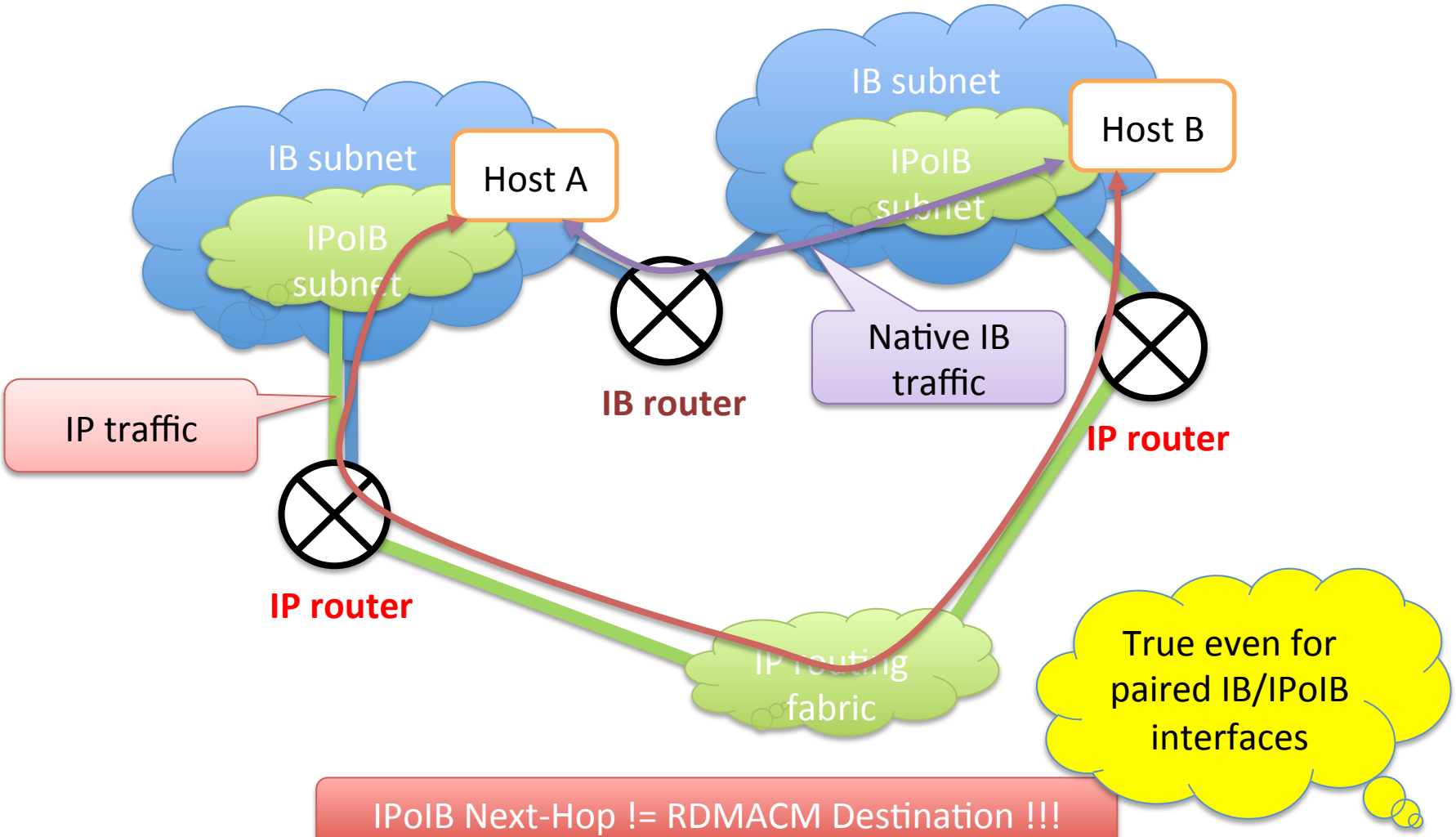


IB ROUTING AND IP(OIB) ADDRESSING

- **IP can be used to**
  - Select local interface
  - Determine SGID
  - Determine next-hop DGID (for IP connectivity)
  - Resolve ServiceIDs within proper network namespace
- **This does not mandate a global IPoIB subnet**
  - Additional models are possible



ARBITRARY IPOIB SUBNETS

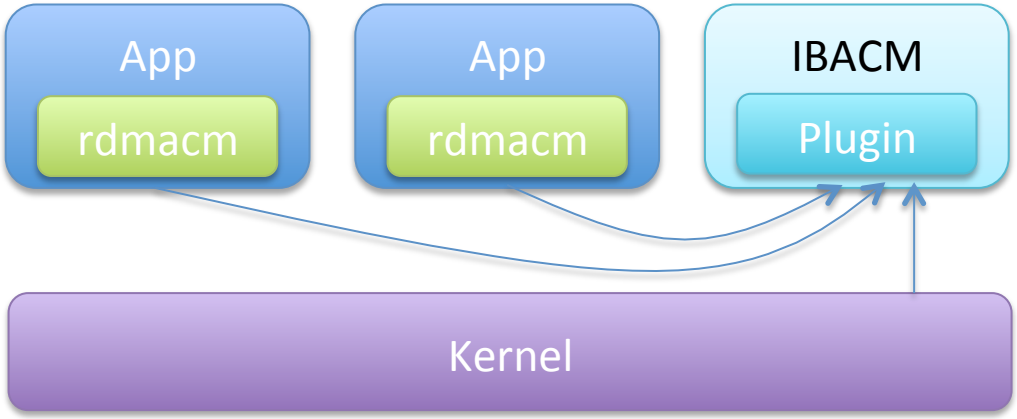


ARBITRARY IPOIB SUBNETS (CONT.)

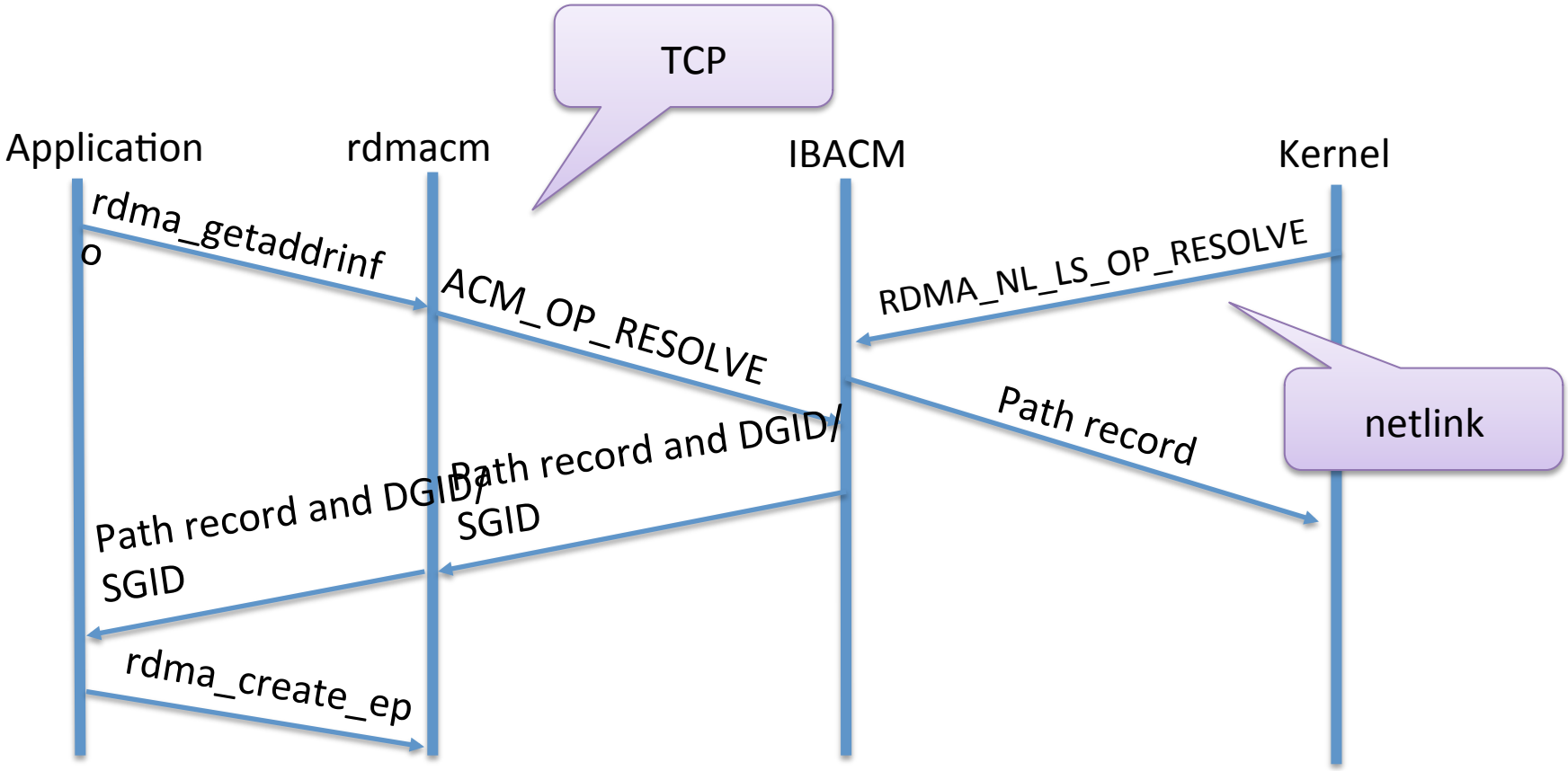
- **Global IPoIB**
  - Neighbor (ARP table) holds HW address of peer node
  - CMA may derive peer GID from HW address
- **Multiple IPoIB subnets**
  - Neighbor holds HW address of the next-hop IP router
  - CMA needs to resolve remote IP to peer GID
- **Global IP→GID resolution is not a kernel task**
- **Solution: use IBACM daemon**

IBACM

- **IBACM assists in establishing IB connections**
- **Implemented as user-space daemon**
  - Plugin architecture for augmenting behavior and implementation
- **Provides**
  - Mapping of hostname/IP→path record for rdmacm
  - Path record lookups for the Kernel
- **Lookup results are cached for fast future access**

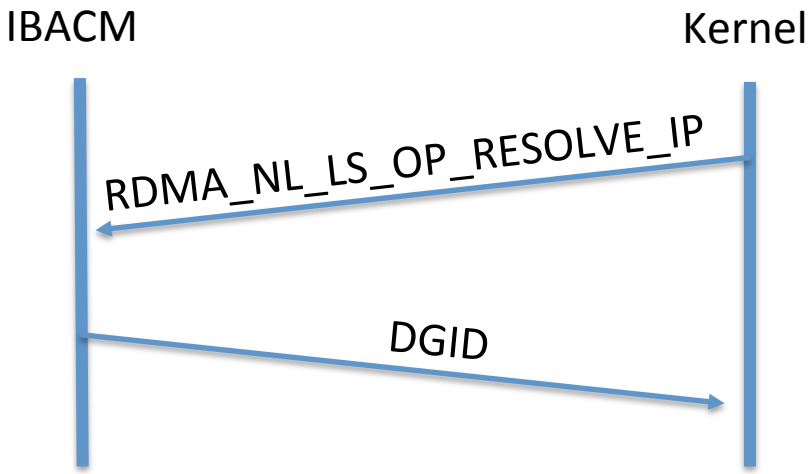


IBACM EXISTING FLOWS



IP→GID RESOLUTION FLOW

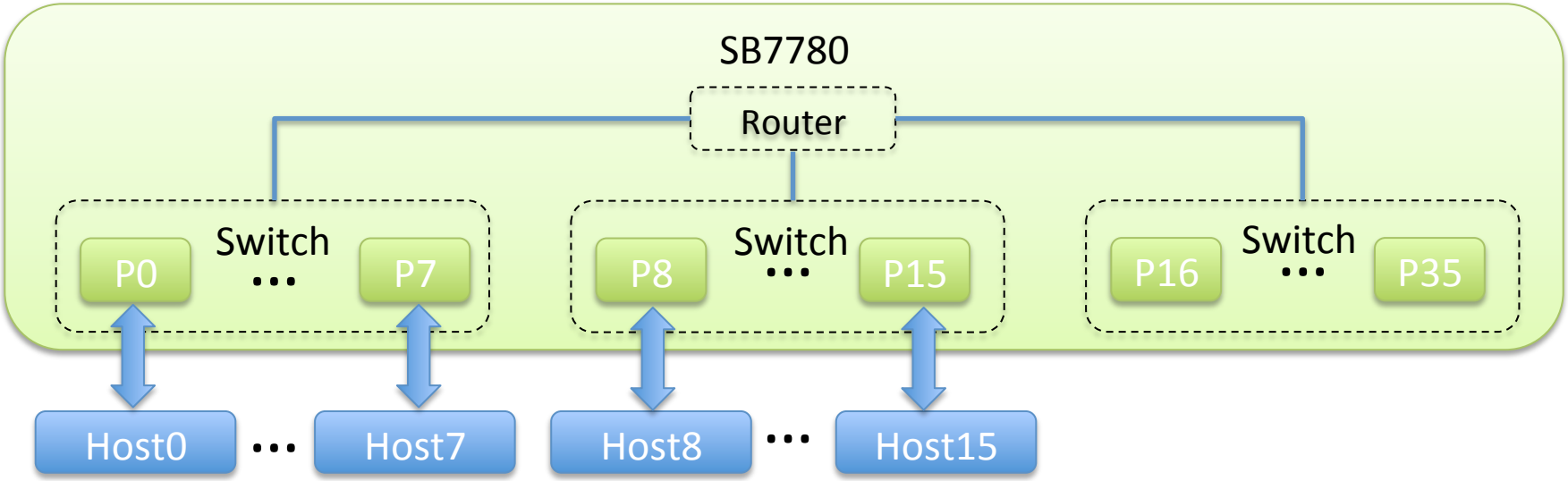
- **Kernel CMA**
  - If destination IP is in a non-adjacent IP subnet: obtain DGID from ibacm
  - Otherwise: fall back to neighbor lookup
- **RDMACM not changed**
  - Applications that obtain path via rdma_getaddrinfo() will use existing flow
  - Others will obtain remote GID and path from the kernel CMA



IB ROUTERS AND HPC

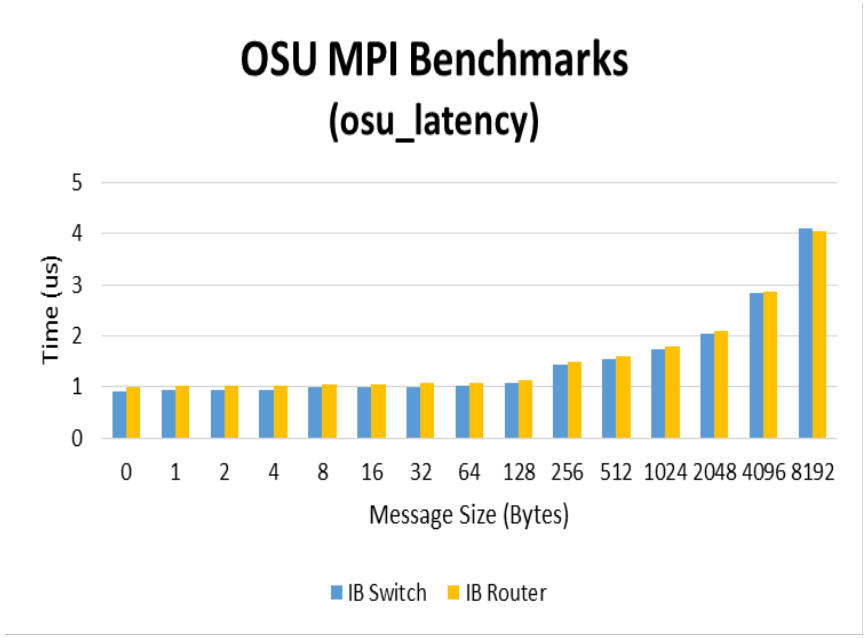
Preliminary results

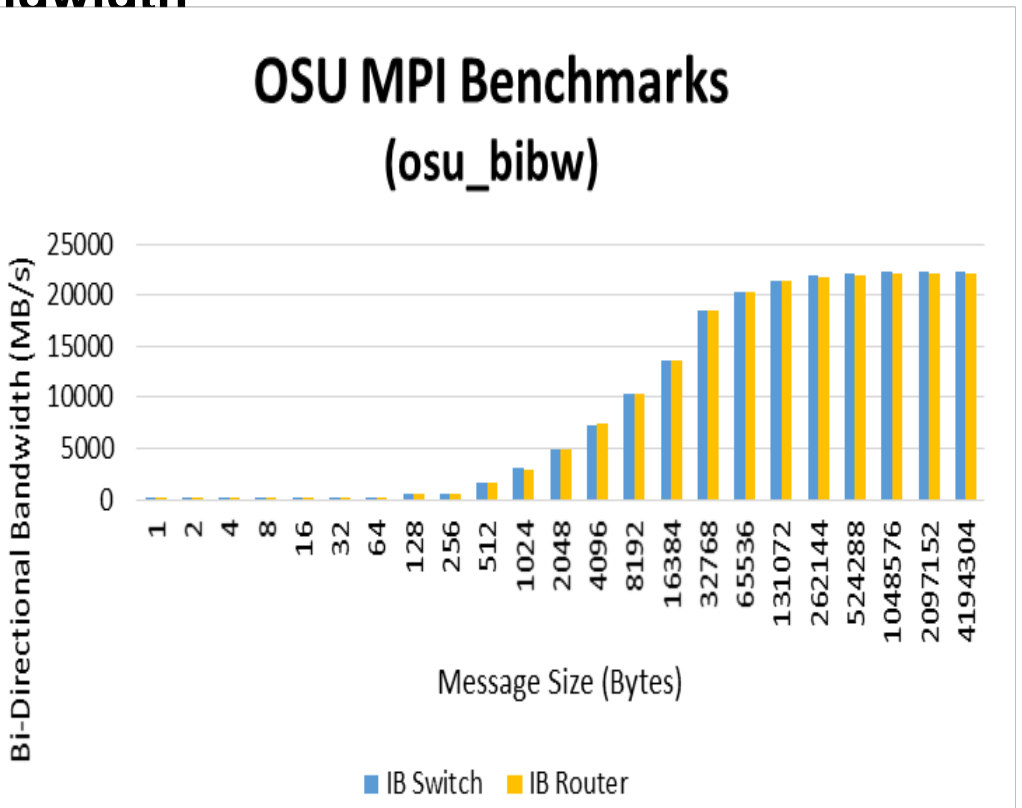
- **Configuration**
  - Mellanox SB7780 configurable, 36-port, EDR switch/router
  - Dell PowerEdge R720 16-node cluster
    - Dual-Socket 10-Core Intel E5-2680v2 @ 2.80 GHz CPUs
  - Vanilla OpenMPI 1.10.3a1
- **Test environment**
  - Compare single subnet vs. splitting ports across 2 subnets



OSU MPI BENCHMARKS

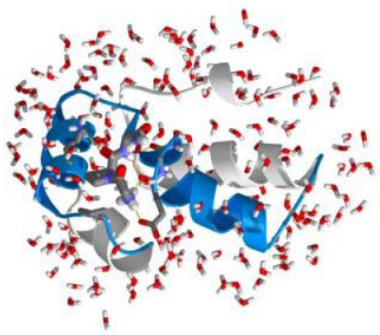
- 2 node MPI test
- ~50ns difference in latency
- No apparent difference in bandwidth

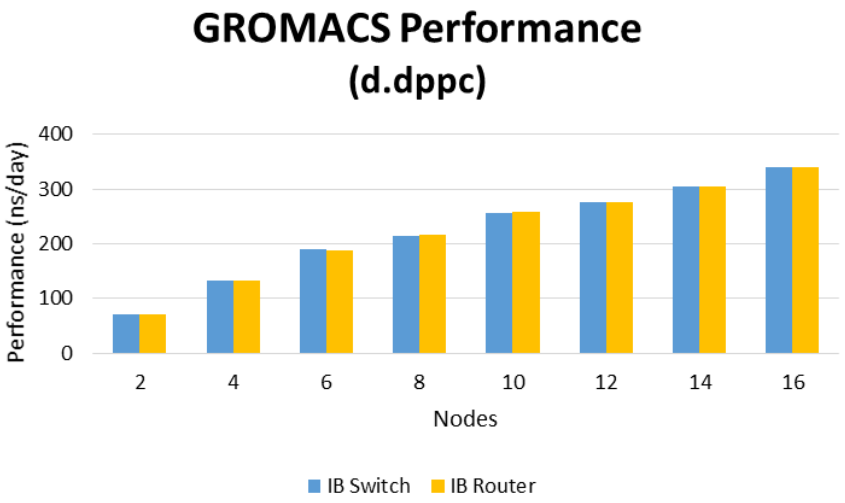


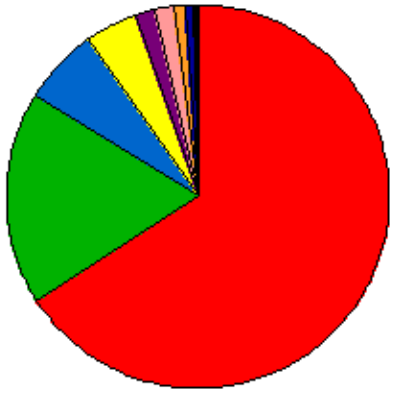


GROMACS APPLICATION

- **GROningen MAchine for Chemical Simulation**
  - Molecular dynamics simulation package
- **Run up to 16 nodes**
- **No apparent differences between switch/router**

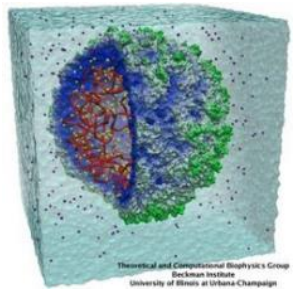


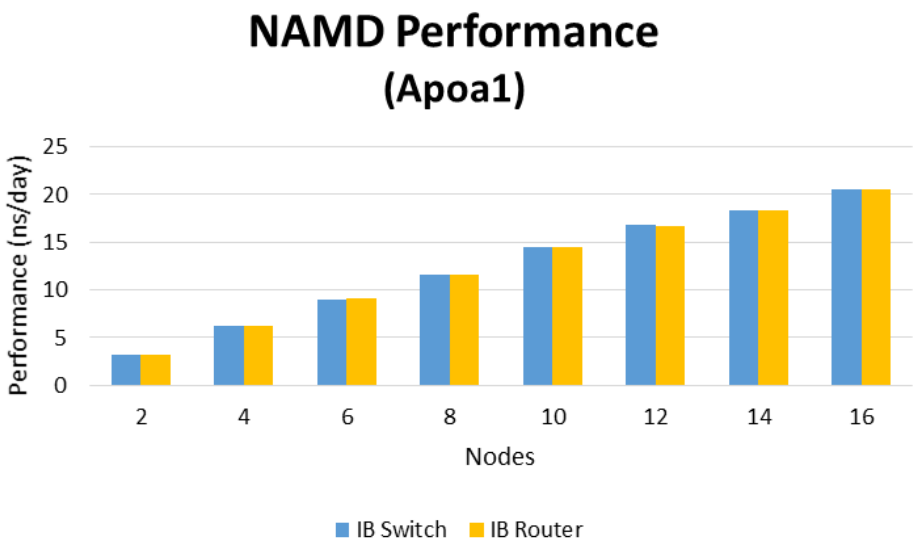


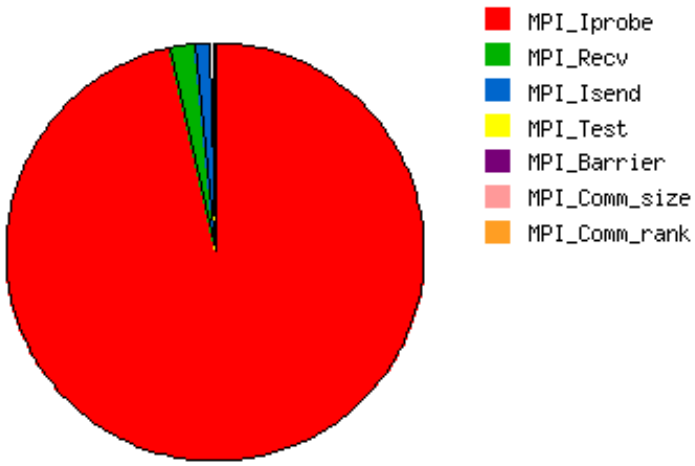


NAMD APPLICATION

- **Parallel molecular dynamics**
  - High-performance simulation of large biomolecular systems
- **Run up to 16 modes**
- **No apparent differences between switch/router**







OPENFABRICS
ALLIANCE

12th ANNUAL WORKSHOP 2016

THANK YOU

Mark Bloch, Liran Liss

Mellanox Technologies