RDMACM APM Support

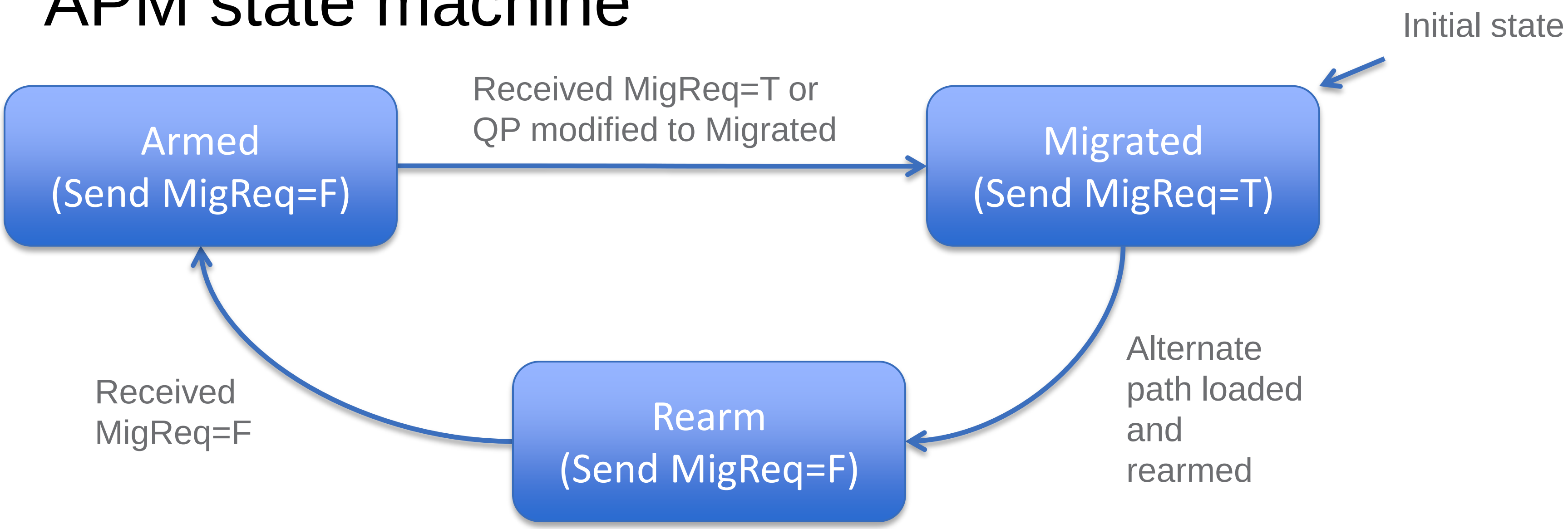
Liran Liss, Mellanox Technologies
March, 2012

Agenda

- APM recap
- APM suggestions
- APM “bonding” model
- Active-side state machine
- Example
- Kernel implementation
- User implementation
- Future enhancements

Automatic Path Migration (APM)

- A mechanism to allow connected QPs or EE-contexts to migrate to an alternate path without loosing the connection
  - For failover or load-balancing purposes
- APM state machine



Current CM Specification

- APM related messages
  - – REQ (active)
    - May include an alternate path
  - – REP (passive)
    - REP.failover_accepted=0 denotes that passive-side approved it
  - – LAP (active)
    - Load alternate path
  - – APR (passive)
    - Alternate path response
- Active side is always the initiator

Current CM Specification

- Short-comings
  - Active side cannot know when a new port on the passive side has joined the subnet
  - Active side must register for event forwarding to learn about the state of passive-side ports
    - SA must maintain state for each connection in the network
  - Passive side cannot notify active side on desired changes to the alternate path
    - Load balancing
    - Address mappings, e.g., a change in the IP address of an IPoIB interface

APM Suggestions (SWG7322 – Errata)

- Motivation: allow passive-side to suggest alternate paths
  - Scalable, immediate reaction to local link-state changes on both ends
  - Take into account passive-side information when determining alternate paths
- Benefits
  - Achieve scalable APM-HA with multi-port HCAs

APM Suggestions

- Simple request-response communication
  - Passive requests, active acknowledges
- Avoid changes to existing APM mechanism
  - Suggestions are only hints
  - Actual changes done by standard LAP/APR
- Changes to REQ message
  - ‘SAP supported’ bit
- New SAP message (passive)
  - Suggest alternate path
  - Message format similar to LAP but conveys only local information

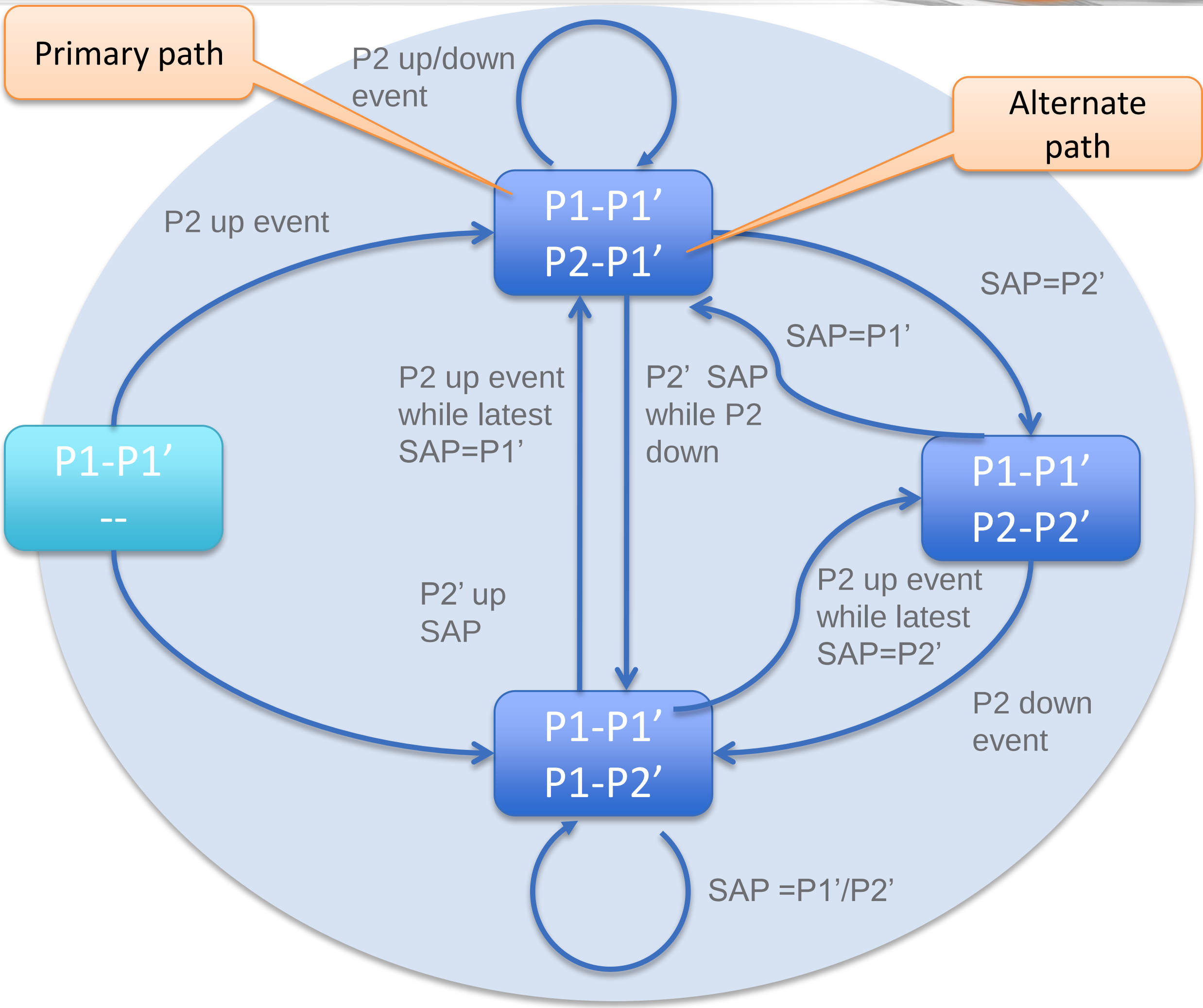
APM Suggestions

- New SPR message (active)
  - Suggest path response
  - Message format similar to APR

APM “Bonding” with RDMACM

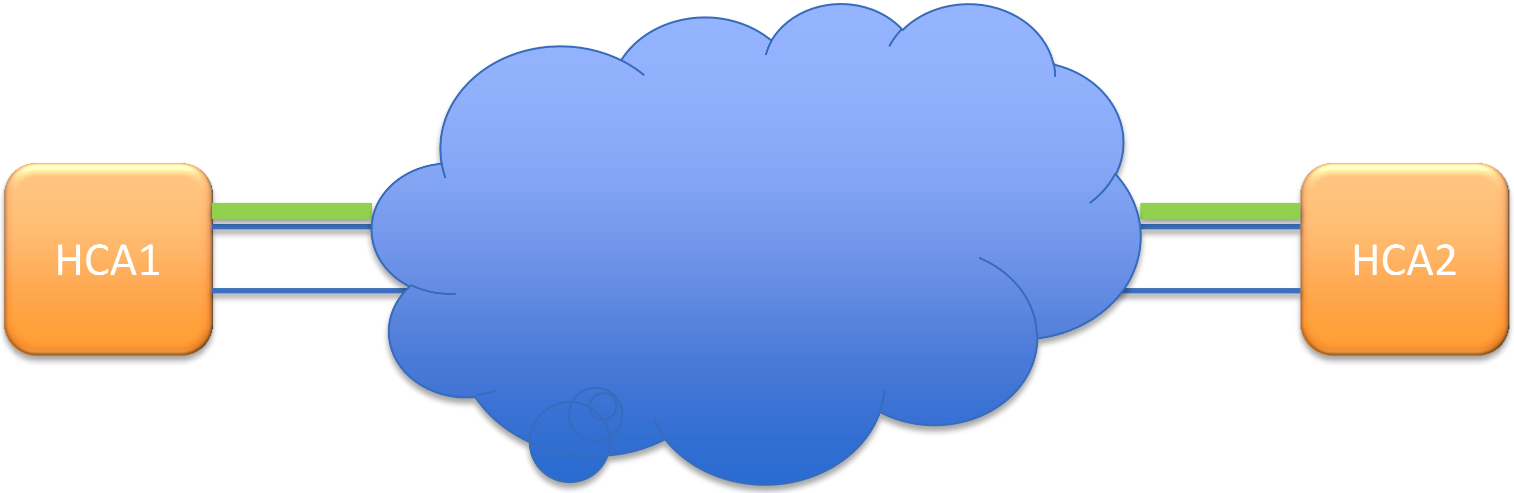
- Goal: handle HA automatically in RDMACM
  - Relieve applications from worrying about link-level HA
    - (Almost) transparent to applications
  - Supports kernel ULPs (e.g., SDP, RDS) and user-space applications
- Mimic behavior of an independent Active-Backup HA scheme at both ends
  - At any point in time, if a backup port exists, an alternate path will make use of it
- Automatically rearm connection whenever an alternate path exists

Active-side State Machine

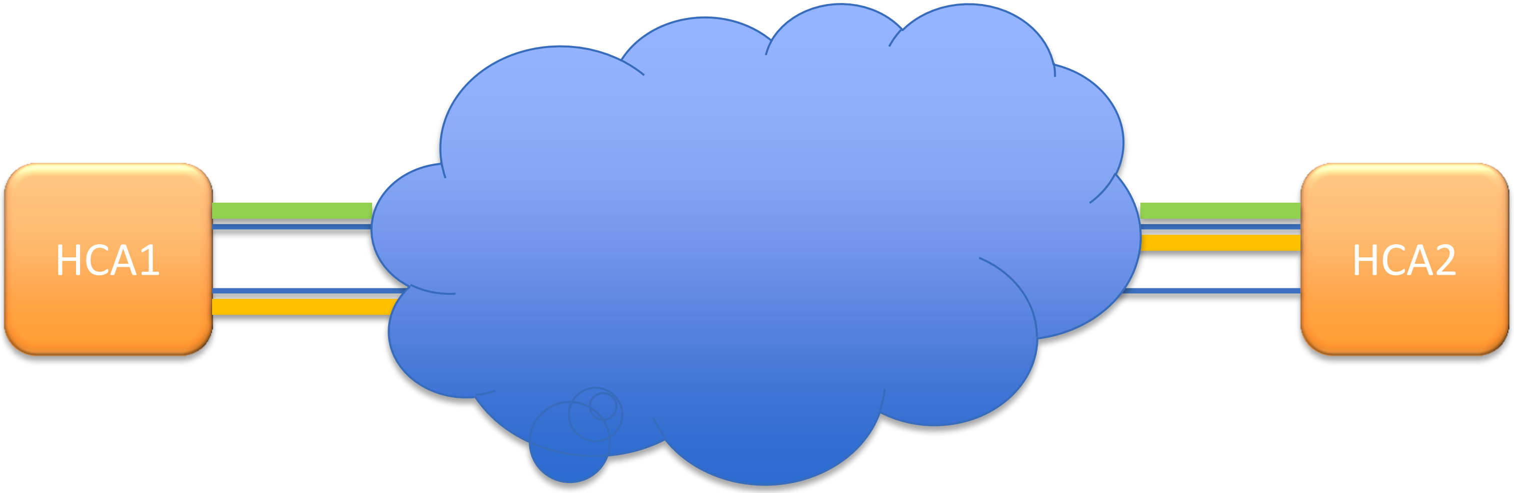


Automatic rearm upon APM event

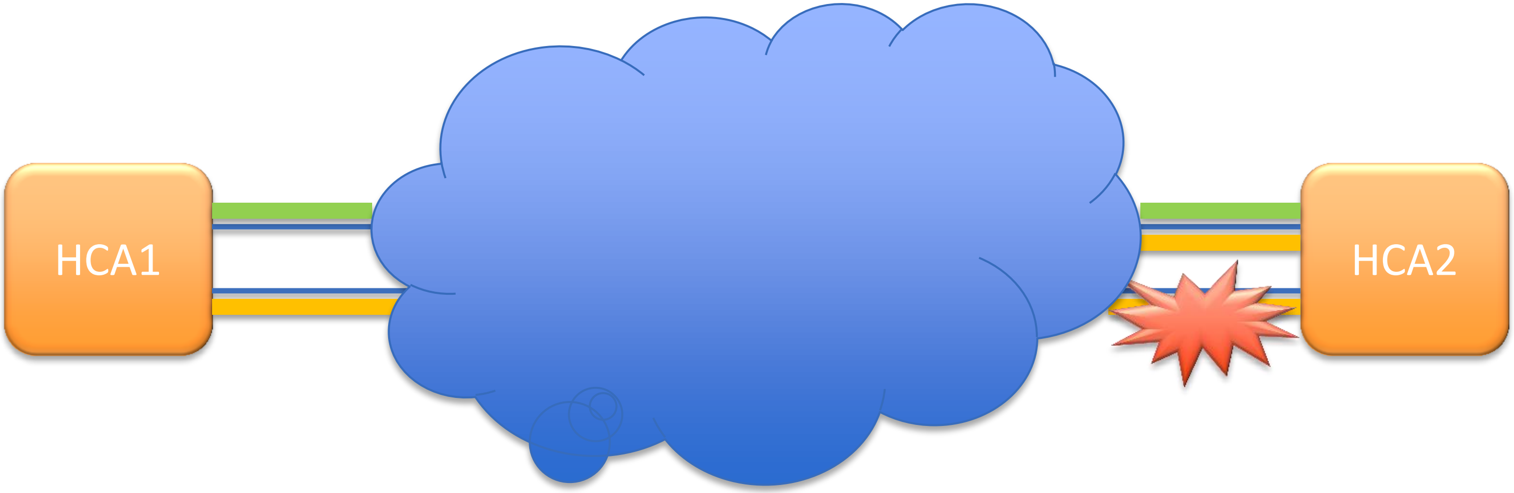
APM “Bonding” Example



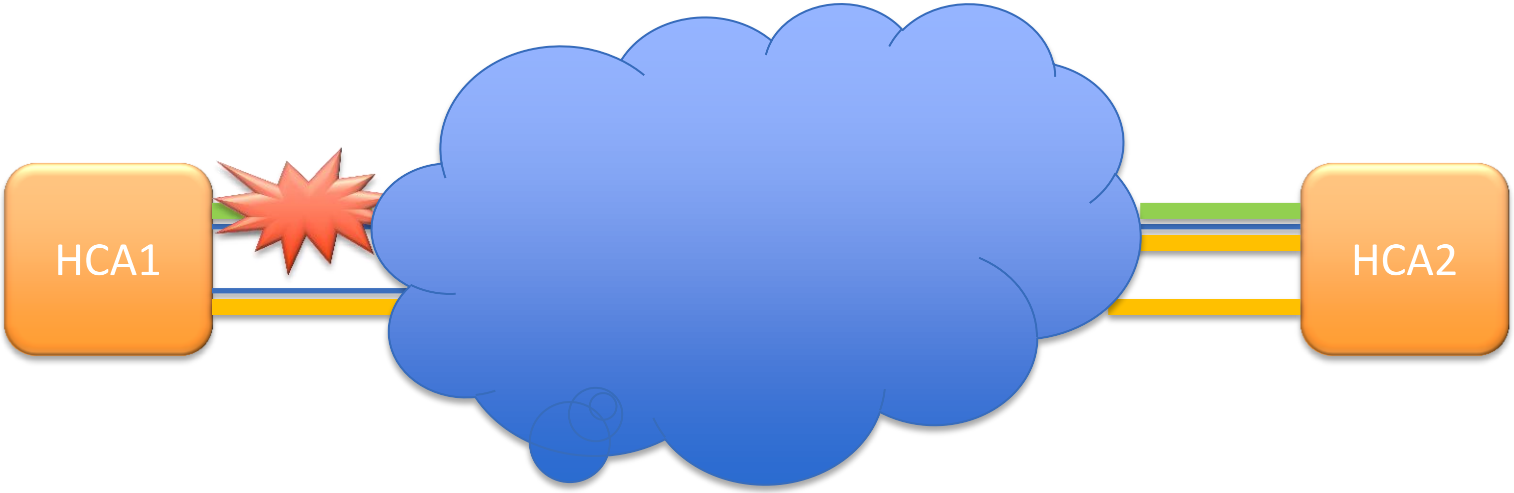
APM “Bonding” Example



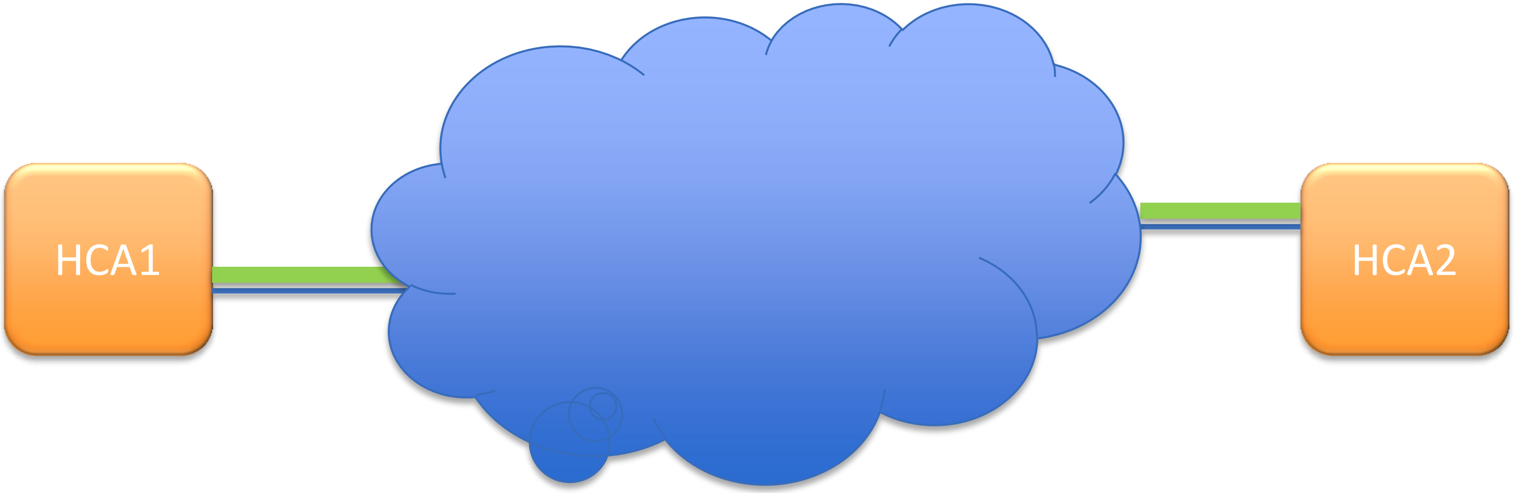
APM “Bonding” Example



APM “Bonding” Example



APM “Bonding” Example



Kernel Implementation

- API
  - rdma_enable_apm()
    - Called on both sides
- Initial connection based on existing RDMACM
  - Bonding driver should be used to select port
  - APM kicks in after the connections is established
- Asynchronous state machine
  - On established/port change/APM events and SAP
    - Reevaluate alternate path
    - If a better option exists
      - Active: send path query
      - Passive: send SAP

Kernel Implementation

- On successful path query response (active side only)
  - Send LAP
- On LAP/APR
  - Load alternate path
- On SPR (passive side only)
  - If status=retry, reschedule SAP

User Implementation

- API
  - rdma_set_option() with RDMA_OPTION_IB_APM
- Logic remains in kernel
- User-space commits QP state changes following events
  - ALT_ROUTE_RESOLVED
    - Used for updating the alternate path
  - ALT_ROUTE_ERROR
    - For informational purposes only
  - LOAD_ALT_PATH
    - Used for arming the alternate path
- Events delivered through rdmacm event channel
  - Application provides processing context while calling rdmacm_get_event()
  - Application must poll event channel continuously

Future Enhancements

- Extend to support LMC
- Track valid alternate path-records
- Strict guarantees rather than best-effort
- Load-balance different connections among ports to achieve Active-Active configurations

OPENFABRICS ALLIANCE

Thank You!