RDMA Reset Flows

11TH ANNUAL
INTERNATIONAL
OPENFABRICS SOFTWARE
DEVELOPERS' WORKSHOP

Introduction

- Device resets are a necessity
  - PCI errors
  - Device errors
  - Unresponsive device
  - Device isolation
  - Hot unplug
  - Driver removal

- Challenges
  - Device is stateful
    - Resources
    - Connection state
    - WR IDs
  - ULPs hold direct references to device resources
  - User-space references
    - Direct references via uverbs
    - Indirect references via ucma

Solution Overview

- Kernel abortive shutdown
  - Raise “fatal” event
  - Put device in “error” mode
  - Unregister device
    - Ensures all ULPs release resources in dependency order

- User-space abortive shutdown
  - Dispatch “fatal” event
  - “Zombify” device

Error State

- Flush pending WQEs in SW
- Return immediate failure for new Post_Send/Recv()’s
- Verbs processing
  - All APIs that release resources successfully processed
    - HW state assumed to be reset
    - Release all associated SW resources
  - Remaining APIs (allocate/modify resources) return an immediate error

Zombie Devices

- A SW device that behaves in “error state”
  - Has no references to HW
  - Does not force applications to exit
- uverbs support (for supporting providers)
  - Disassociates HW from existing uverbs context
  - Frees all resources in IDR trees
  - Returns EIO for all system calls
- ucma support
  - Destroy underlying RDMA IDs
  - Mark ucma_context as closed
    - Avoid duplicate closing when App releases RDMA ID
  - Return EIO for all other system calls
- No change required in umad/ucm

Provider Support

- In kernel driver
  - Implement disassociate_ucontext()
    - For example
      - Remove MMIO mappings to device
      - Free related resources
      - Notify user-space driver
  - Implement “error mode”

- In user-space driver
  - Implement “error mode”

Status

- Scheduled for Linux 4.3
  - – uverbs
  - – ucma
  - – ConnectX-3 provider (mlx4_ib)

Future Work and Discussion

- ib_uverbs
  - Graceful abort
    - Allow grace period for apps to close their references
    - But delays reset completion…
- librdmacm
  - Respond to RDMA_CM_EVENT_DEVICE_REMOVAL
    - Refresh device list
- Maintain ULP state during transient errors
  - Introduce new IB client ops:
    - stop() – release all references to HW resources
    - start() – re-create HW resources
  - Called in ULP load order
- User-space provider support

Thank You

OPENFABRICS
ALLIANCE

11TH
ANNUAL
INTERNATIONAL
OPENFABRICS SOFTWARE
DEVELOPERS' WORKSHOP