13th ANNUAL WORKSHOP 2017

BUILDING A BLOCK STORAGE APPLICATION ON OFED - CHALLENGES

Subhojit Roy, Tej Parkash, Jeetendra R Sonar, Storage Engineering

[March 28th, 2017]

AGENDA

Introduction

- Setting the Context (SVC as Storage Virtualizer)
- SVC Software Architecture overview

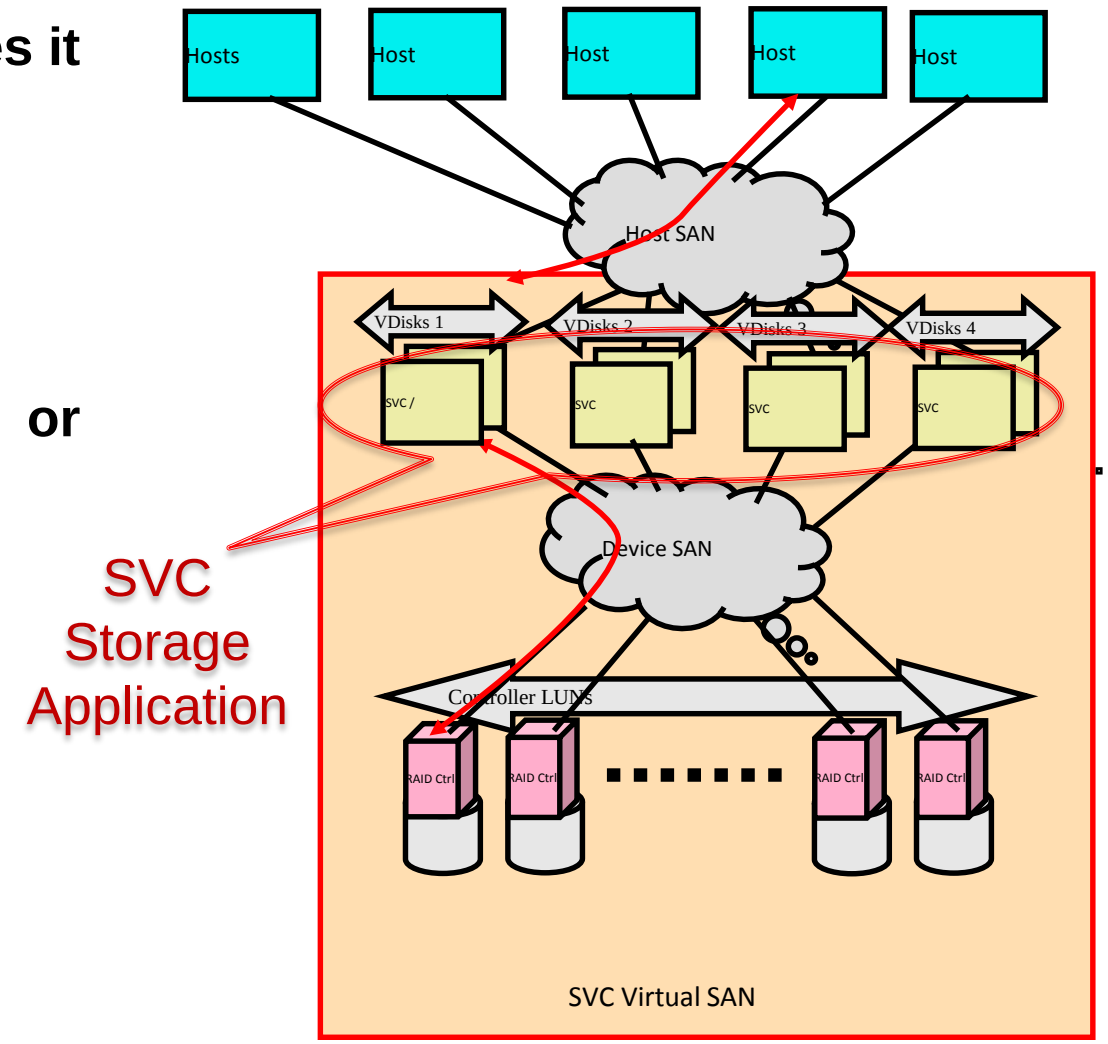
Challenges

- Queue Pair states
- RDMA disconnect behavior
- RDMA connection management
- Query and modify Queue Pair attributes
- Large DMA memory allocation
- Query Device List
- Conclusion

INTRODUCTION

SETTING THE CONTEXT (SVC AS STORAGE VIRTUALIZER)

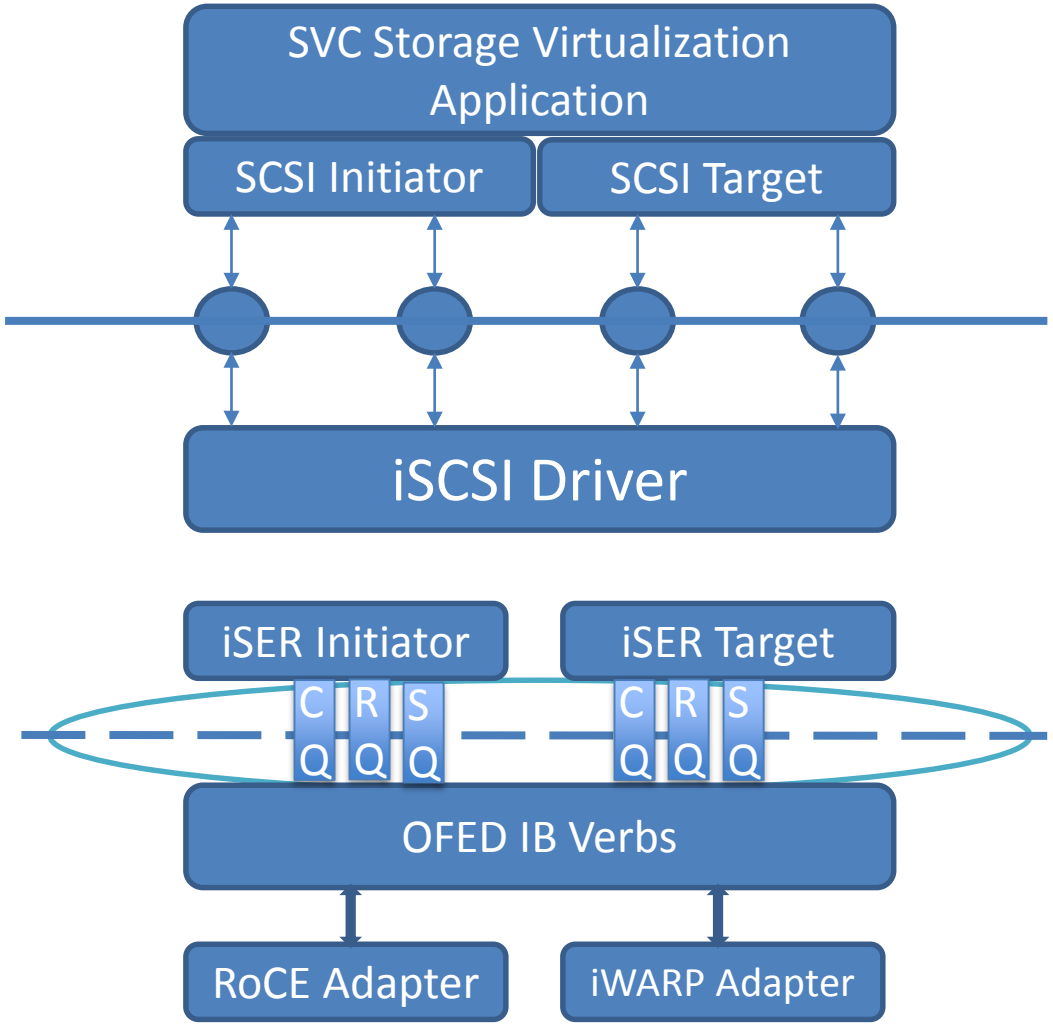
- **SVC pools heterogenous storage and virtualizes it for the host**
- **iSER Target for Host**
- **iSER Initiator for Storage Controller (FLASH or HDD)**
- **Clustered over iSER for high availability**
- **Supports both RoCE and iWARP**
- **Supports 10/25/40/50/100G bandwidths**



SVC ARCHITECTURE OVERVIEW

Architecture characteristics

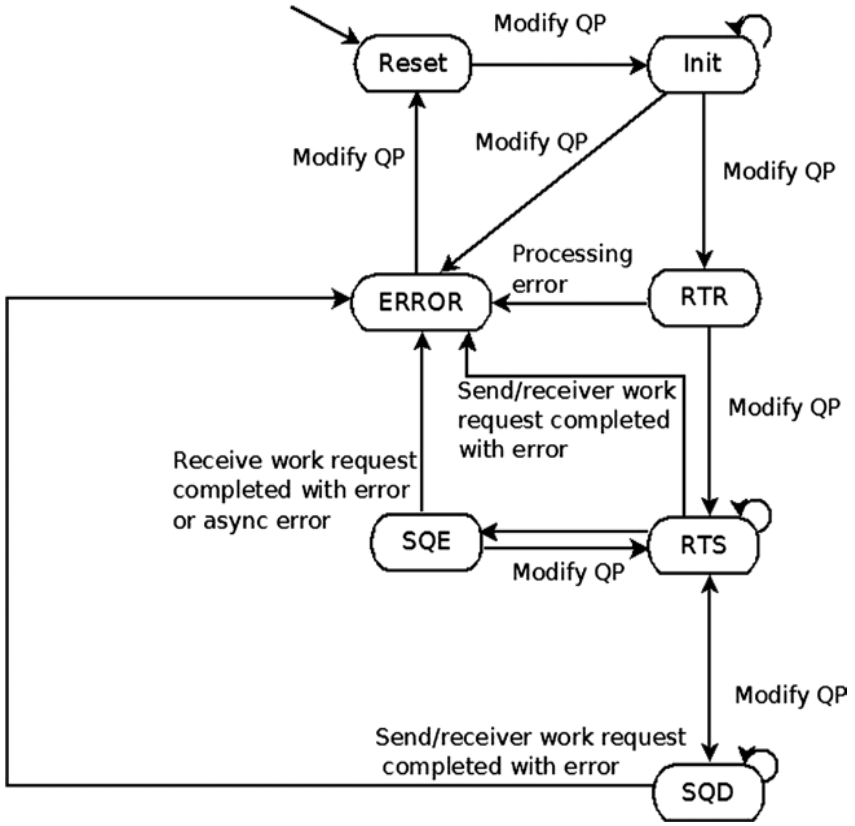
- SVC application runs in user space
- iSER and iSCSI drivers in kernel space
- Lockless architecture (Per CPU port handling)
- Polled mode IO handling
- Supports RoCE and iWARP
- Vendor Independent (Mellanox, Chelsio, Qlogic, Broadcom, Intel etc.)
- Dependence on OFED kernel IB Verbs



CHALLENGES

QUEUE PAIR STATES

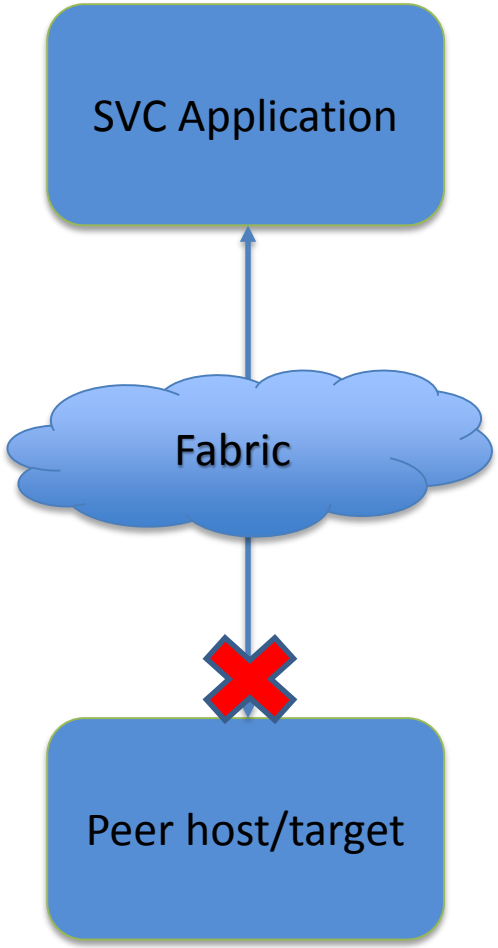
- **Goal**
  - Control number of retries and retry timeout during network outage
- **Actual behavior**
  - State transition differs across RoCE and iWARP e.g iWARP does not support SQD state
- **Expectation**
  - Transition QP to SQD state to modify QP attributes
  - `ib_modify_qp()` must transition QP states as per state diagram shown
  - All state transition must be supported by both RoCE and iWARP
- **Work Around**
  - No work around found
  - Exploring vendor specific possibilities



Referenced from book “Linux Kernel Networking - Implementation and Theory”

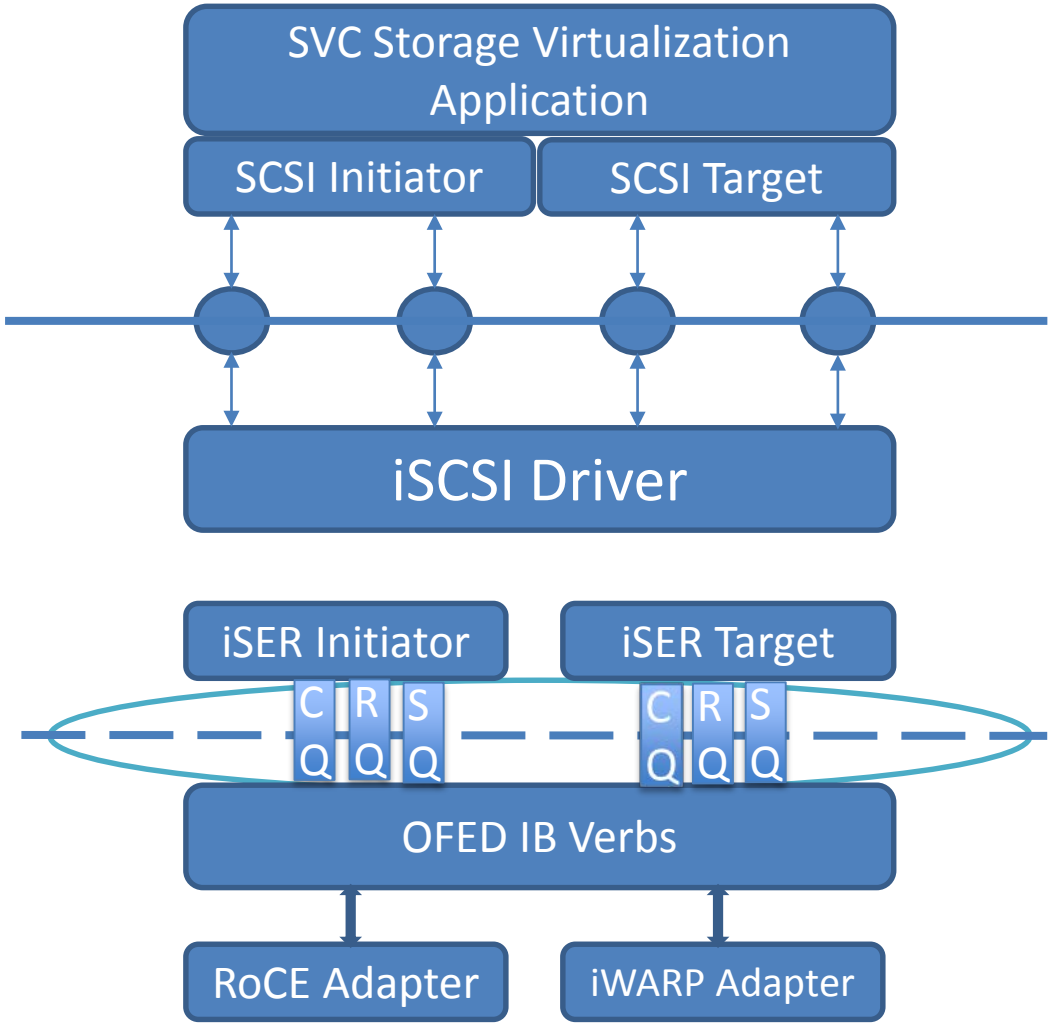
RDMA DISCONNECT BEHAVIOR

- **Goal/Observation**
  - QP cannot be freed before `RDMA_CM_EVENT_DISCONNECTED` event is received
  - There is no control over the timeout period for this event
- **Actual behavior**
  - Link down on peer system causes DISCONNECT event to be received after long delay
    - RoCE: ~100 Sec
    - iWARP: ~70 Sec
  - There is no standard mechanism (verb) to control these timeouts
- **Expectation**
  - RDMA disconnect event must exhibit uniform timeout across RoCE and iWARP
  - Timeout period for disconnect must be configurable
- **Work Around**
  - Evaluating vendor specific mechanism to tune CM timeout



RDMA CONNECTION MANAGEMENT

- **Goal**
  - Polled mode data path and Connection Management
- **Current mechanism**
  - No mechanism to poll for CM events. All RDMA CM events are interrupt driven
  - Current implementation involves deferring CM events to Linux workqueues
  - Application has no control over which CPU to POLL CM events from
- **Expectation**
  - Queues for CM event handling
- **Work Around**
  - Usage of locks add to IO latency



LARGE DMA MEMORY ALLOCATION

- **Observation**
  - Allocation of large chunks DMAable memory during session establishment fails
  - SVC reserves majority of physical memory during system initialization for caching
- **Current mechanism**
  - IB Verbs use `kmalloc()` to allocate DMAable memory for all the queues
- **Expectation**
  - IB Verbs must provide a means to allocate DMA-able memory from pre-allocated memory pool. e.g. in the following
    - `ib_alloc_cq()`
    - `ib_create_qp()`
- **Work Around Solutions**
  - Modified iWARP and RoCE driver to use pre-allocated memory pools from SVC

Type	Elements	Size	Total Size(KB)
SQ	2064	88	~177KB
RQ	2064	32	~64KB
CQ	2064	32	~64KB

Single Connection Memory requirement in Linux OFED Stack = ~297KB

retry_cnt

QUERY AND MODIFY QUEUE PAIR ATTRIBUTES

- **Goal/Observation**
  - Query and set QP parameters to control error recovery behavior
- **Actual behavior**
  - Unable to get and set QP parameters
  - iWARP does not support modify/query of all parameters defined in `ib_qp_attr()` e.g. field `rnr_retry`
- **Expectations**
  - `ib_query_qp()` and `ib_modify_qp()` should behave as documented
  - If QP parameters are specific to iWARP or RoCE, they must be documented
- **Work Around Solutions**
  - Evaluating vendor specific possibilities

```
enum ib_mtu              path_mtu;
enum ib_mig_state        path_mig_state;
u32                      qkey;
u32                      rq_psn;
u32                      sq_psn;
u32                      dest_qp_num;
int                      qp_access_flags;
struct ib_qp_cap         cap;
struct ib_ah_attr        ah_attr;
struct ib_ah_attr        alt_ah_attr;
u16                      pkey_index;
u16                      alt_pkey_index;
u8                       en_sqd_async_notify;
u8                       sq_draining;
u8                       max_rd_atomic;
u8                       max_dest_rd_atomic;
u8                       min_rnr_timer;
u8                       port_num;
u8                       timeout;
u8                       retry_cnt;
u8                       rnr_retry;
u8                       alt_port_num;
u8                       alt_timeout;
u32                      rate_limit;
```

Referenced from: Linux Kernel

QUERY DEVICE LIST

- **Observation**
  - No kernel verb to find list of rdma devices on system until RDMA session is established
  - Per device resource allocation during kernel module initialization
- **Current mechanism**
  - RDMA device available only after connection request is established by CM event handler
- **Expectation**
  - Need verb equivalent to `ibv_get_device_list()` in kernel IB Verbs
- **Work Around**
  - Complicates per port resource allocation during initialization

CONCLUSION

- **Initial indications of IO performance compared to FC – excellent!**
- **iSER presents an opportunity for high performance Flash based Ethernet data center**
- **Error recovery and handling is troublesome**
- **Mass adoption by storage vendors requires more work in OFED**
  - IB Verbs is not completely protocol independent
  - Proper documentation of RoCE vs iWARP specific differences
  - Definitive resource allocation timeout values (R_A_TOV equivalent in FC)
- **Same requirements applicable to NVMef**
- **Seeking right forum to address these requirements**

13th ANNUAL WORKSHOP 2017

THANK YOU

subhojit.roy@in.ibm.com, tprakash@in.ibm.com, jeesonar@in.ibm.com

[March 28th, 2017]

BACKUP SLIDES

FC V/S ISER LATENCY PERFORMANCE

IO Size	FC Latency (milliseconds)	iSER Latency (milliseconds)
Read_4k	0.107	0.072
Write_4k	0.185	0.222
Read_32k	0.121	0.100
Write_32k	0.224	0.267
Read_64k	0.183	0.150
Write_64k	0.299	0.342