RDMA基础和UDISK实践

Agenda

- RDMA概念
- RDMA编程接口
- RDMA在UDisk的实践

RDMA – what is it?

- **R**emote
  - data transfers between nodes in a network
- **D**irect
  - no Operating System Kernel involvement in transfers
  - everything about a transfer offloaded onto Interface Card
- **M**emory
  - transfers between user space application virtual memory
  - no extra copying or buffering
- **A**ccess
  - send, receive, read, write, atomic operations

RDMA Benefits

- High throughput
- Low latency
- High messaging rate
- Low CPU utilization
- Message boundaries preserved
- Asynchronous operation

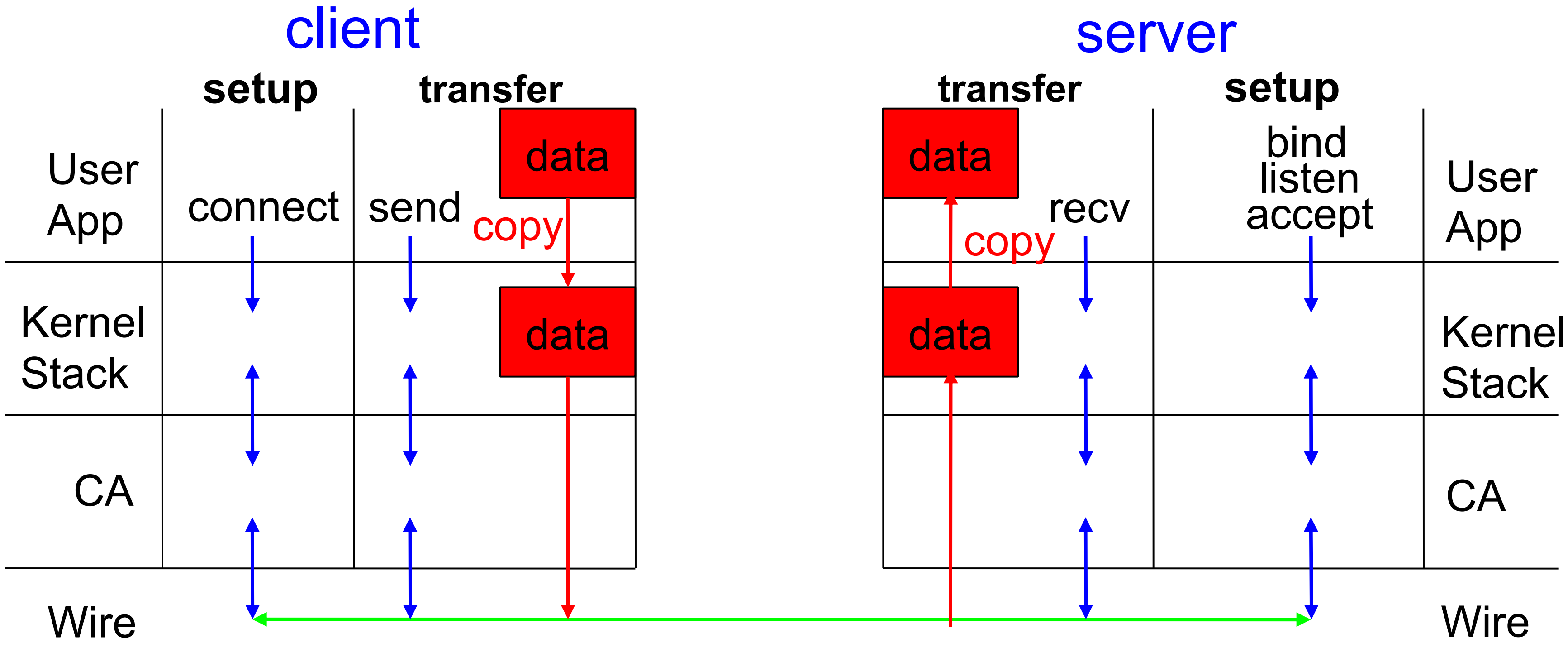
How RDMA differs from TCP/IP

- “zero copy” – data transferred directly from virtual memory on one node to virtual memory on another node

- “kernel bypass” – no operating system involvement during data transfers

- asynchronous operation – threads not blocked during I/O transfers

TCP/IP transfer



client

setup

transfer

User App

Kernel Stack

CA

Wire

connect

send

data

copy

data

server

transfer

setup

data

recv

copy

data

bind
listen
accept

User App

Kernel Stack

CA

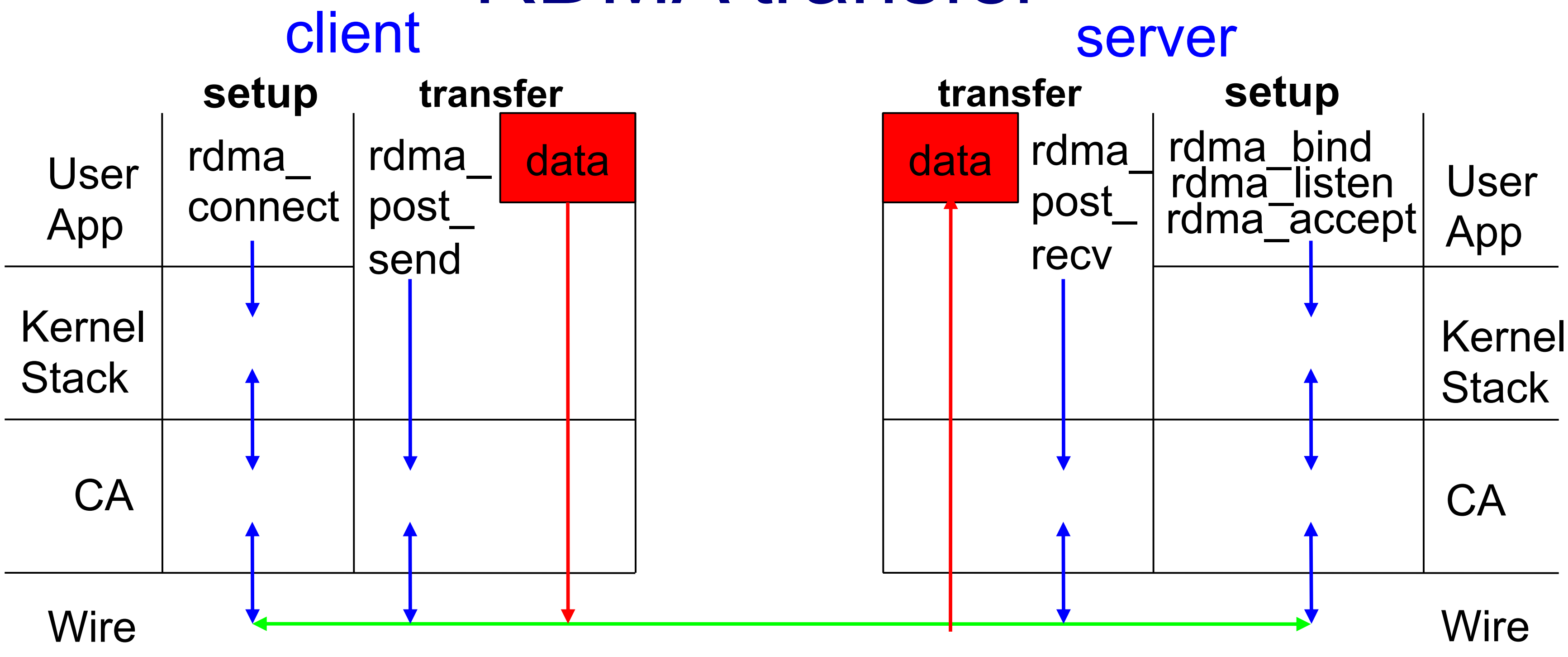
Wire

blue lines: control information

red lines: user data

green lines: control and data

RDMA transfer

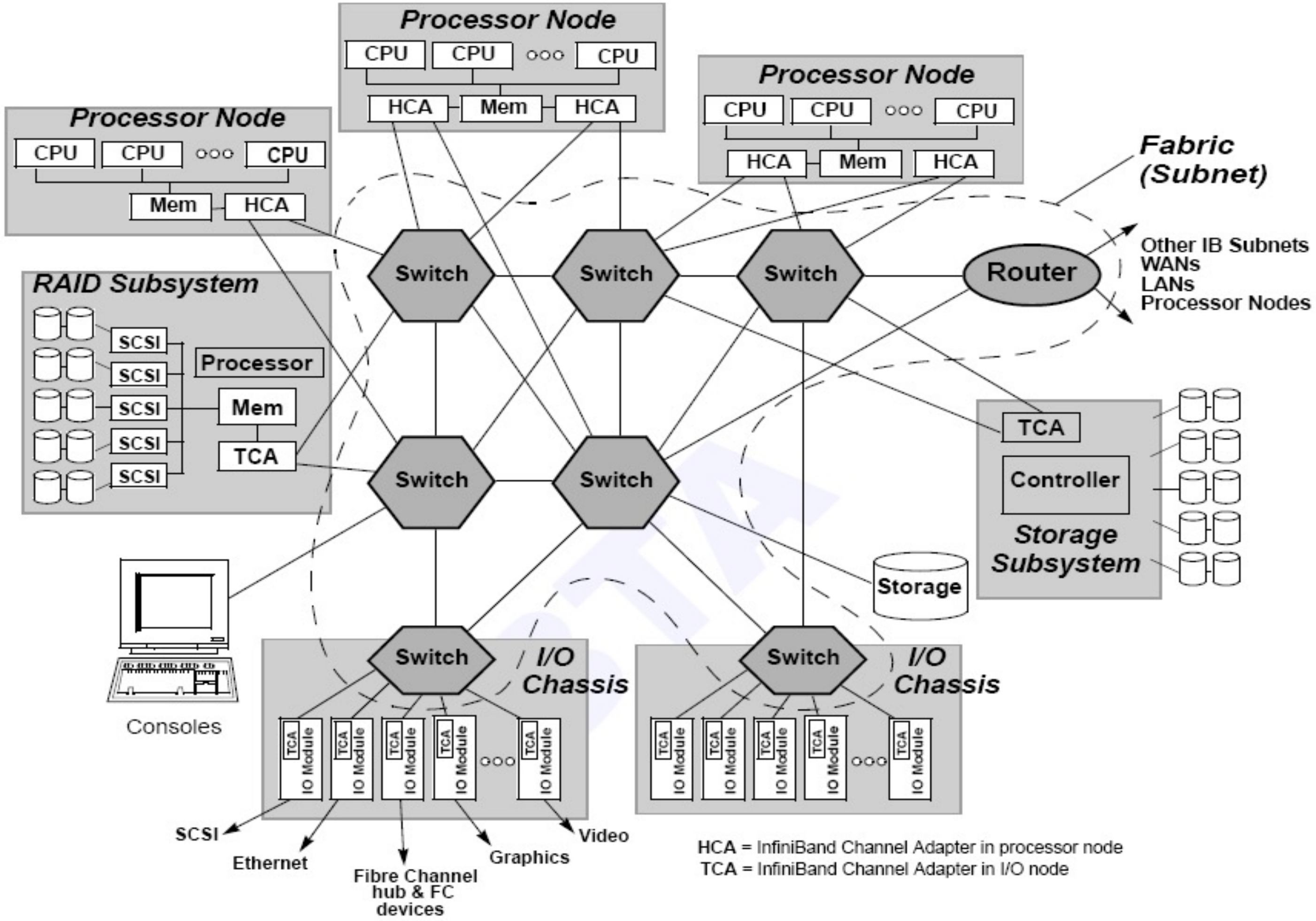


blue lines: control information

red lines: user data

green lines: control and data

Infiniband architecture overview



HCA = InfiniBand Channel Adapter in processor node
TCA = InfiniBand Channel Adapter in I/O node

RoCE - RDMA over Converged Ethernet

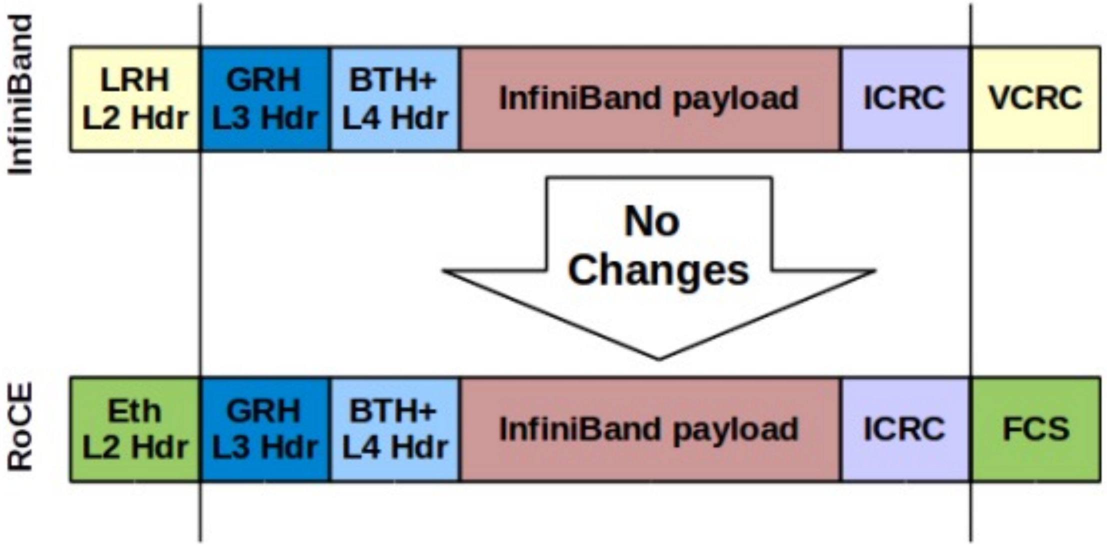


Figure 13 – Differences between IB and RoCE frames [20]

RoCEv2 – Based on UDP

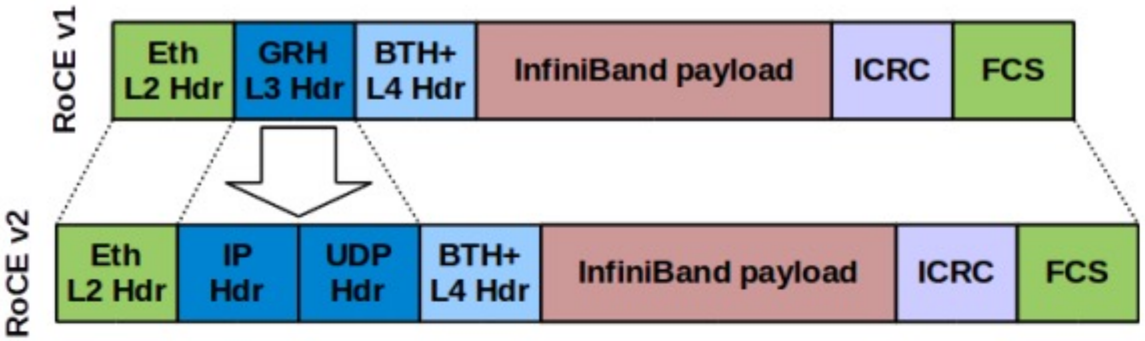
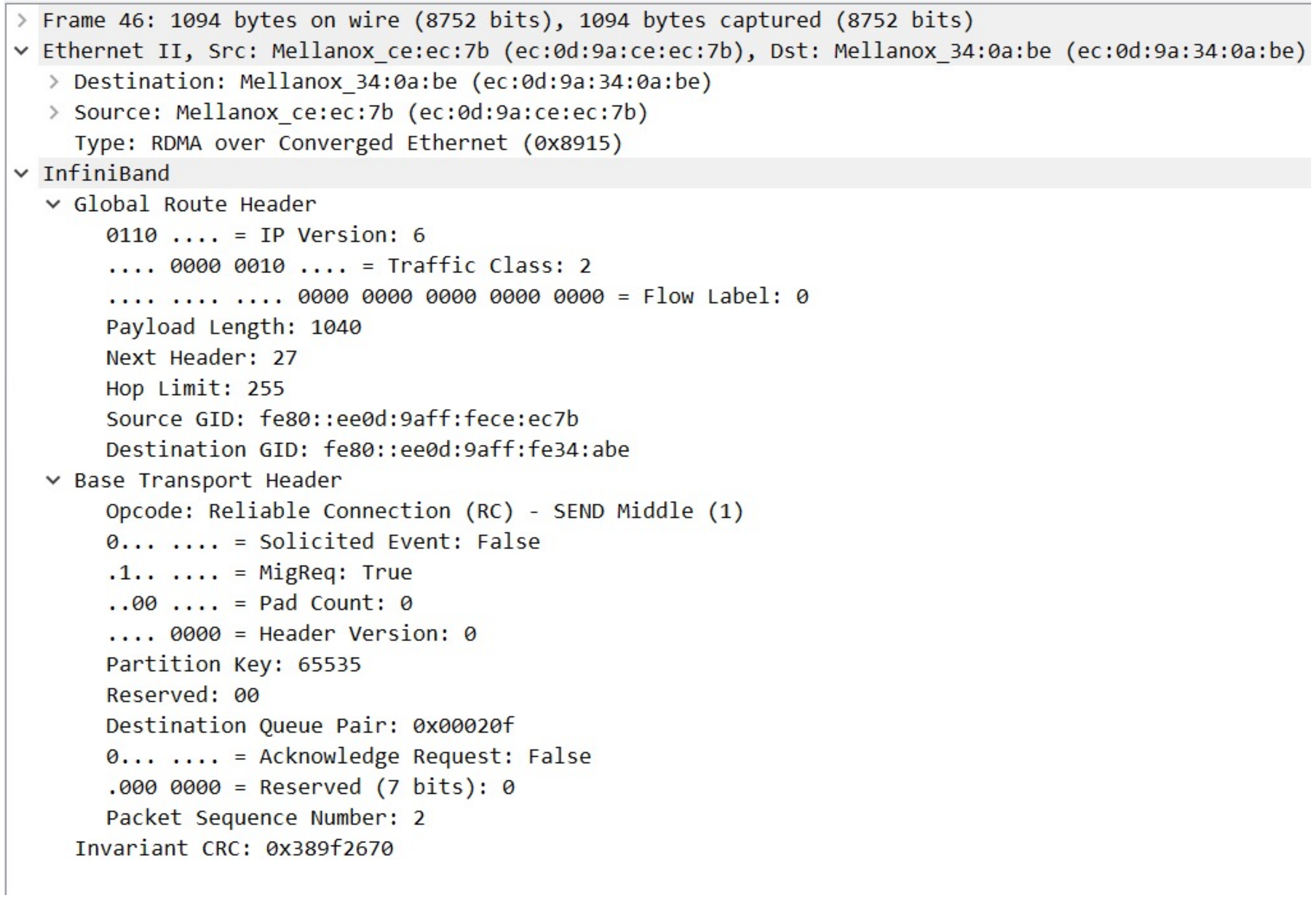


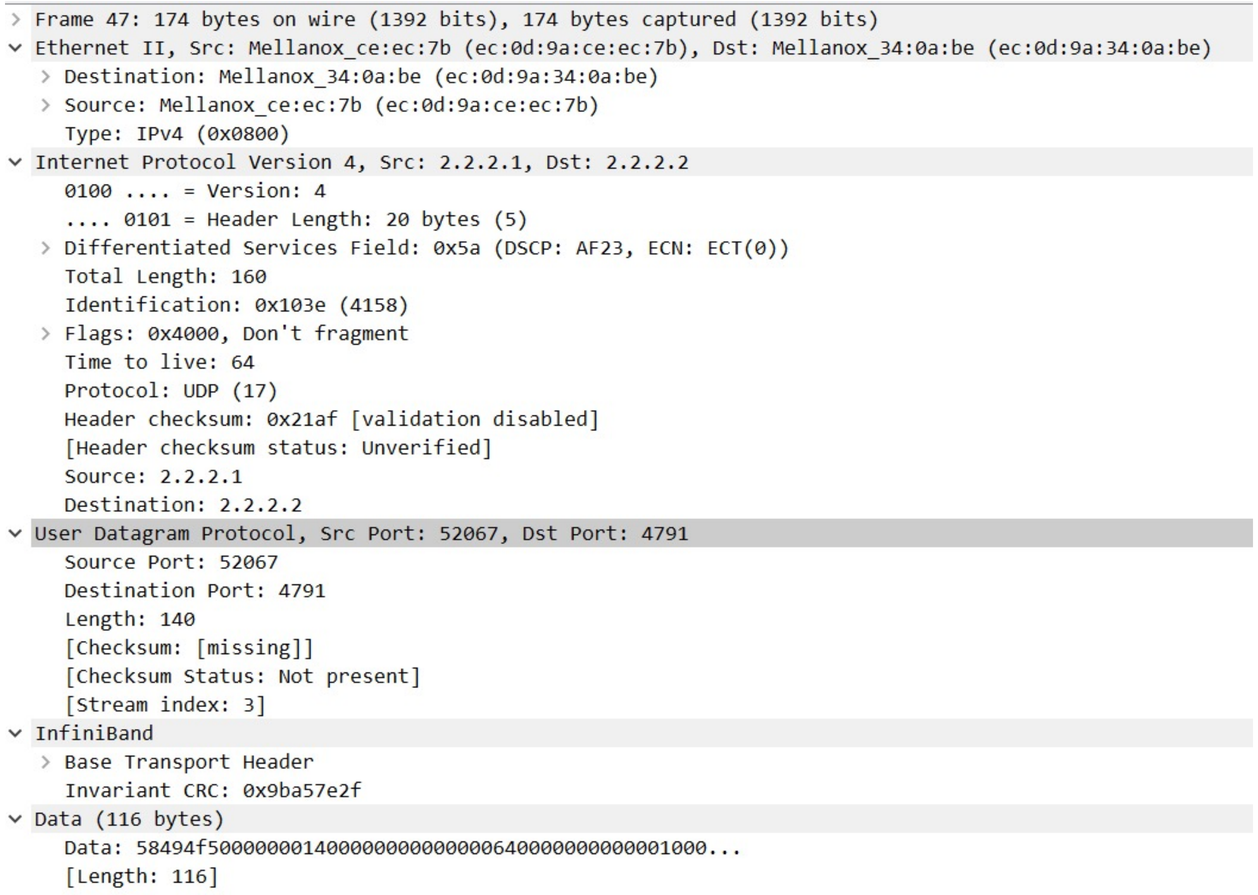
Figure 14 – Differences between RoCE v1 and v2 frames [21]

RoCE packet



```
> Frame 46: 1094 bytes on wire (8752 bits), 1094 bytes captured (8752 bits)
⌄ Ethernet II, Src: Mellanox_ce:ec:7b (ec:0d:9a:ce:ec:7b), Dst: Mellanox_34:0a:be (ec:0d:9a:34:0a:be)
  > Destination: Mellanox_34:0a:be (ec:0d:9a:34:0a:be)
  > Source: Mellanox_ce:ec:7b (ec:0d:9a:ce:ec:7b)
    Type: RDMA over Converged Ethernet (0x8915)
⌄ InfiniBand
  ⌄ Global Route Header
      0110 .... = IP Version: 6
      .... 0000 0010 .... = Traffic Class: 2
      .... .... .... 0000 0000 0000 0000 0000 = Flow Label: 0
      Payload Length: 1040
      Next Header: 27
      Hop Limit: 255
      Source GID: fe80::ee0d:9aff:fece:ec7b
      Destination GID: fe80::ee0d:9aff:fe34:abe
  ⌄ Base Transport Header
      Opcode: Reliable Connection (RC) - SEND Middle (1)
      0... .... = Solicited Event: False
      .1.. .... = MigReq: True
      ..00 .... = Pad Count: 0
      .... 0000 = Header Version: 0
      Partition Key: 65535
      Reserved: 00
      Destination Queue Pair: 0x00020f
      0... .... = Acknowledge Request: False
      .000 0000 = Reserved (7 bits): 0
      Packet Sequence Number: 2
    Invariant CRC: 0x389f2670
```

RoCEv2 packet



```
> Frame 47: 174 bytes on wire (1392 bits), 174 bytes captured (1392 bits)
v Ethernet II, Src: Mellanox_ce:ec:7b (ec:0d:9a:ce:ec:7b), Dst: Mellanox_34:0a:be (ec:0d:9a:34:0a:be)
  > Destination: Mellanox_34:0a:be (ec:0d:9a:34:0a:be)
  > Source: Mellanox_ce:ec:7b (ec:0d:9a:ce:ec:7b)
    Type: IPv4 (0x0800)
v Internet Protocol Version 4, Src: 2.2.2.1, Dst: 2.2.2.2
    0100 .... = Version: 4
    .... 0101 = Header Length: 20 bytes (5)
  > Differentiated Services Field: 0x5a (DSCP: AF23, ECN: ECT(0))
    Total Length: 160
    Identification: 0x103e (4158)
  > Flags: 0x4000, Don't fragment
    Time to live: 64
    Protocol: UDP (17)
    Header checksum: 0x21af [validation disabled]
    [Header checksum status: Unverified]
    Source: 2.2.2.1
    Destination: 2.2.2.2
v User Datagram Protocol, Src Port: 52067, Dst Port: 4791
    Source Port: 52067
    Destination Port: 4791
    Length: 140
    [Checksum: [missing]]
    [Checksum Status: Not present]
    [Stream index: 3]
v InfiniBand
  > Base Transport Header
    Invariant CRC: 0x9ba57e2f
v Data (116 bytes)
    Data: 58494f500000001400000000000000640000000000001000...
    [Length: 116]
```

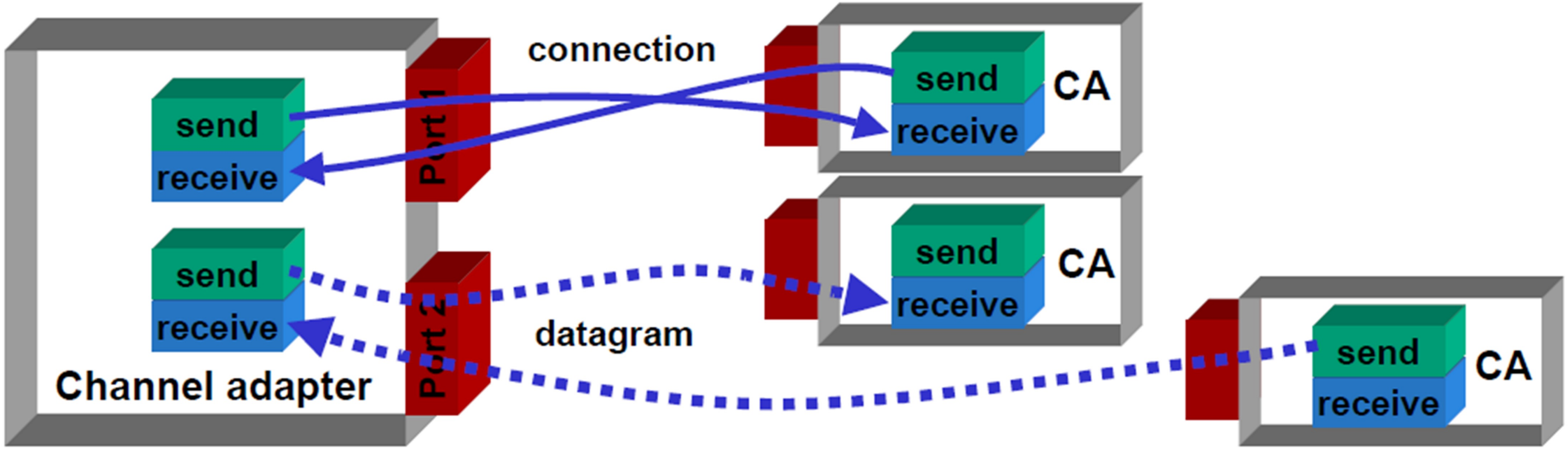
4K ping-pong test with 25Gbps Card

- Latency 8us –depth=1
- IOPS 700K –depth=128

相关名词

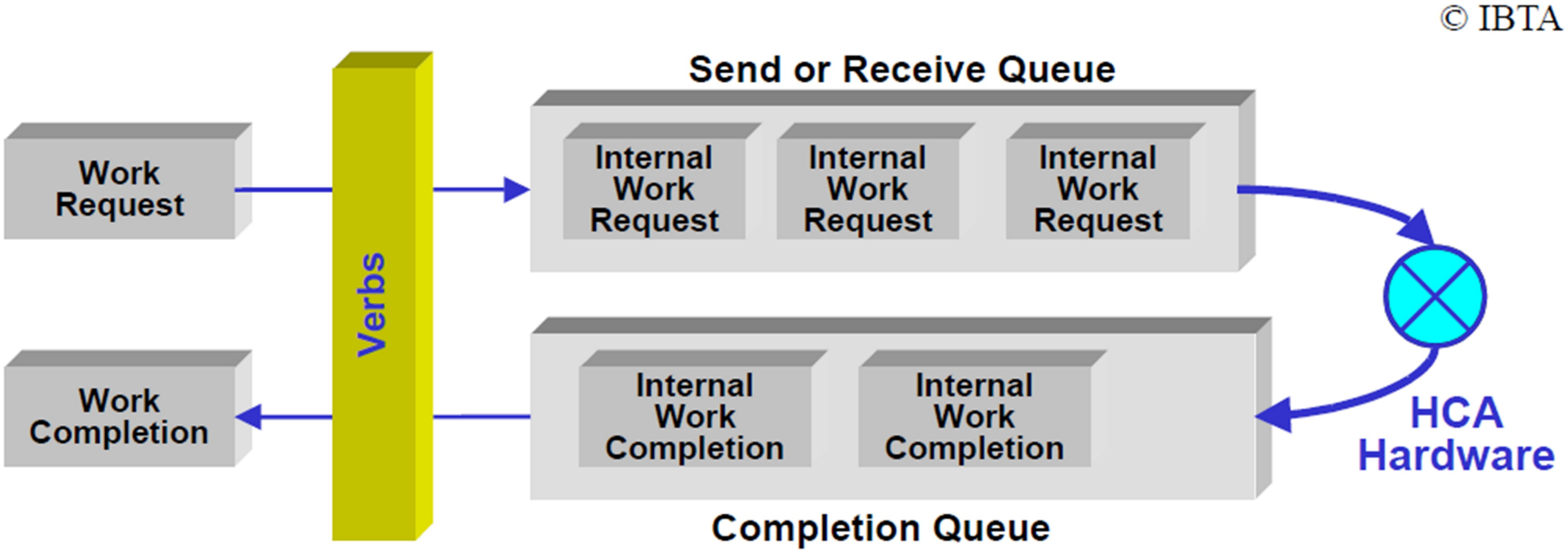
PD	Protection Domain	Glues queue pairs and memory regions
MR	Memory Region	Registered memory region that HCA can read from or write to. Contains R_Key and L_Key
QP	Queue Pair	Send / Receive work queue. Send or receive work requests are placed onto a queue pair
CQ	Completion Queue	Completion Queue. Completed work requests, so called work completions are placed onto a completion queue. Is associated with queue pair.
WR	Work Request	Either send or receive work request. Specifies action to be processed and will be put onto send or receive queue (QP). References scatter/gather element
SGE	Scatter/Gather Element	Defines address(es) in memory to read from or to write to. Must be given L_Key or R_Key to authenticate access to memory region
WC	Work Completion	After a work request has been completed the work completion delivers result

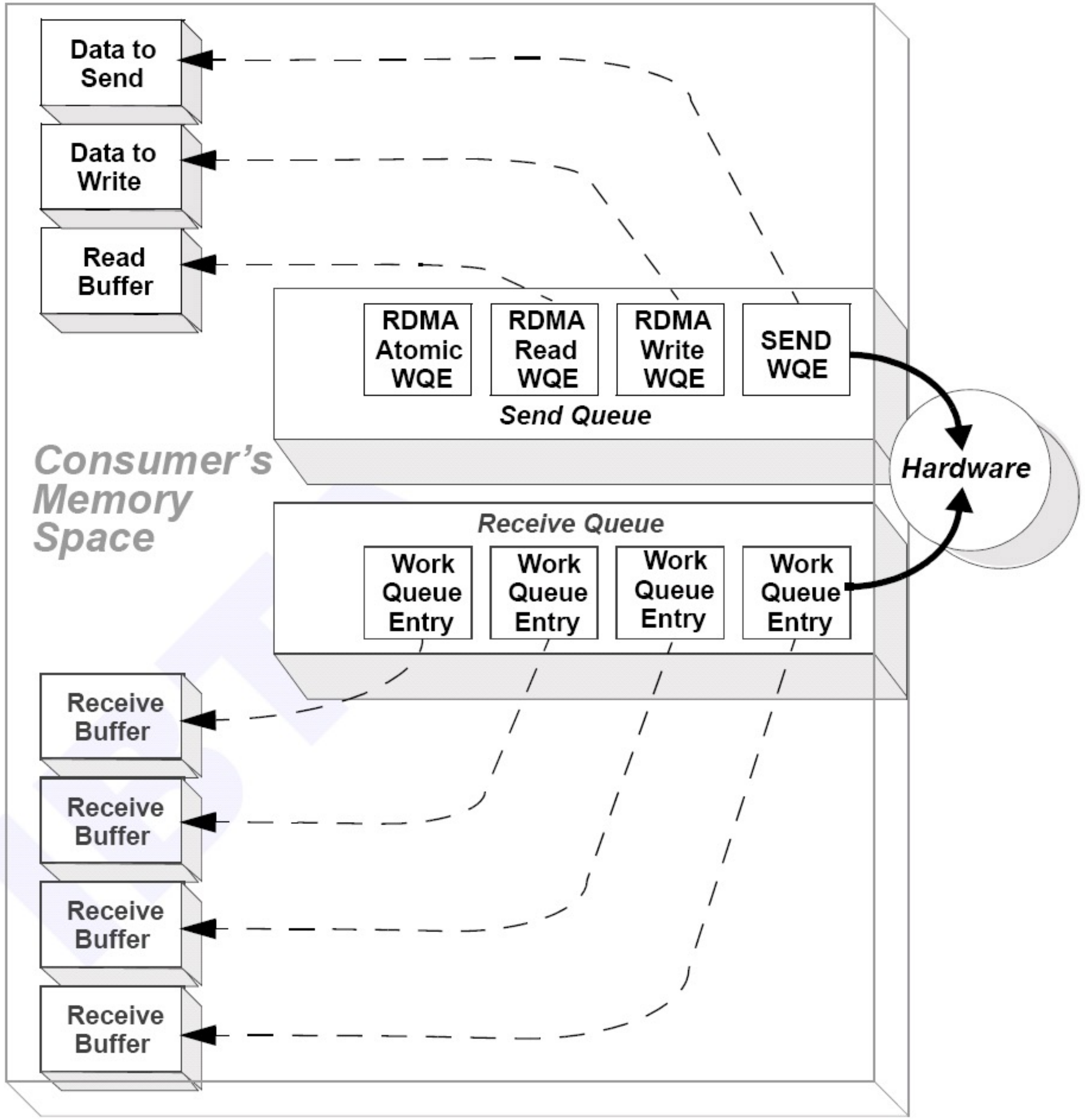
QueuePair



© IBTA

Send/Recv Queue/Completion Queue



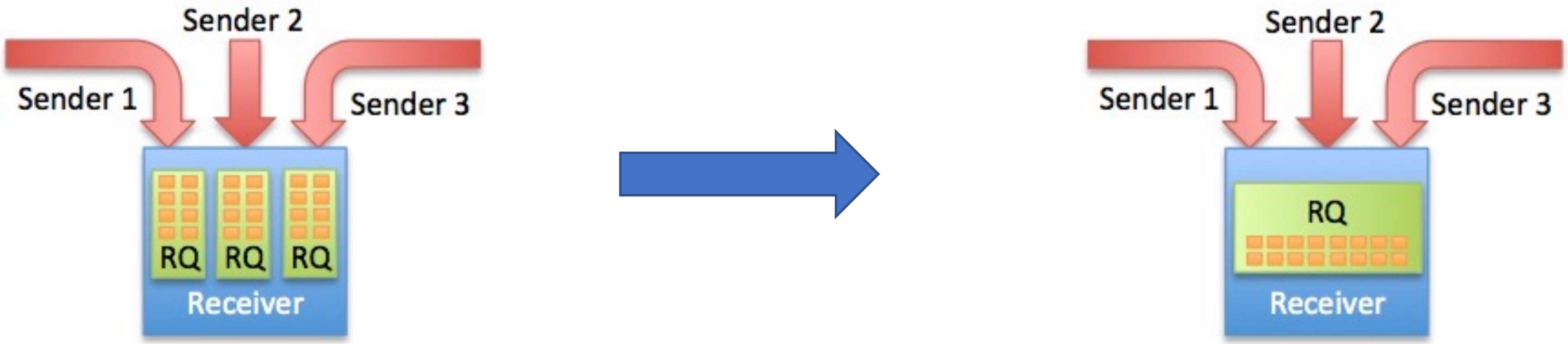

Data to Send
Data to Write
Read Buffer
RDMA Atomic WQE
RDMA Read WQE
RDMA Write WQE
SEND WQE
Send Queue
Consumer's Memory Space
Hardware
Receive Queue
Work Queue Entry
Work Queue Entry
Work Queue Entry
Work Queue Entry
Receive Buffer
Receive Buffer
Receive Buffer
Receive Buffer

Steps

- Create QueuePair
- Exchange QueuePair Info between Nodes
- Modify QueuePair state to RTS （via TCP）
- Poll Complete Queue

RDMA in UDisk

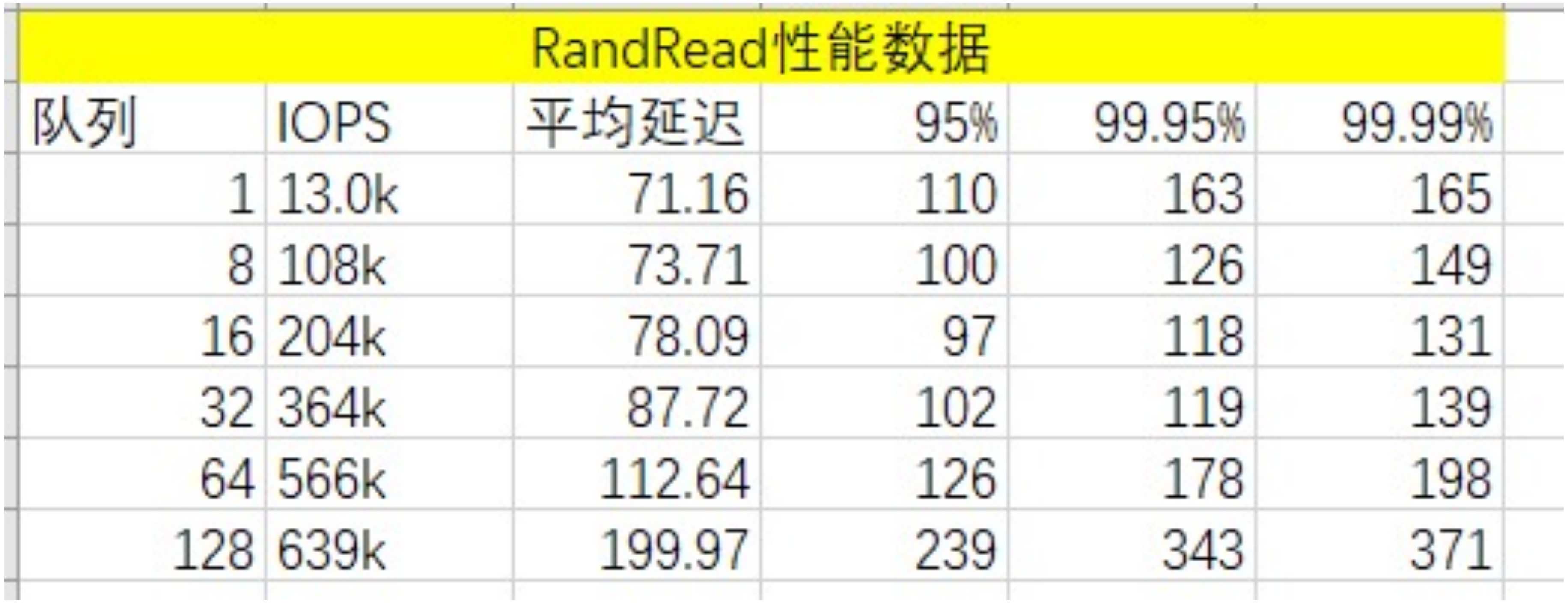
- RC QueuePair
- RoCEv2
- Share CQ – per thread/core
- SRQ – per thread/core



RDMA in UDisk

- Keep Alive
  - ➢ Do RDMA WRITE with 0 size data
- Large Packet
  - ➢ Split into small packets
  - ➢ Do RDMA READ – difficult when there is a failure
- Memory recycling on failure
  - ➢ Handle different kinds of WC error
  - ➢ Move qp to error state and wait all WC is done before destroy it
- Integration with Spdk
  - ➢ Create Memory Pool From spdk_dma_zmalloc

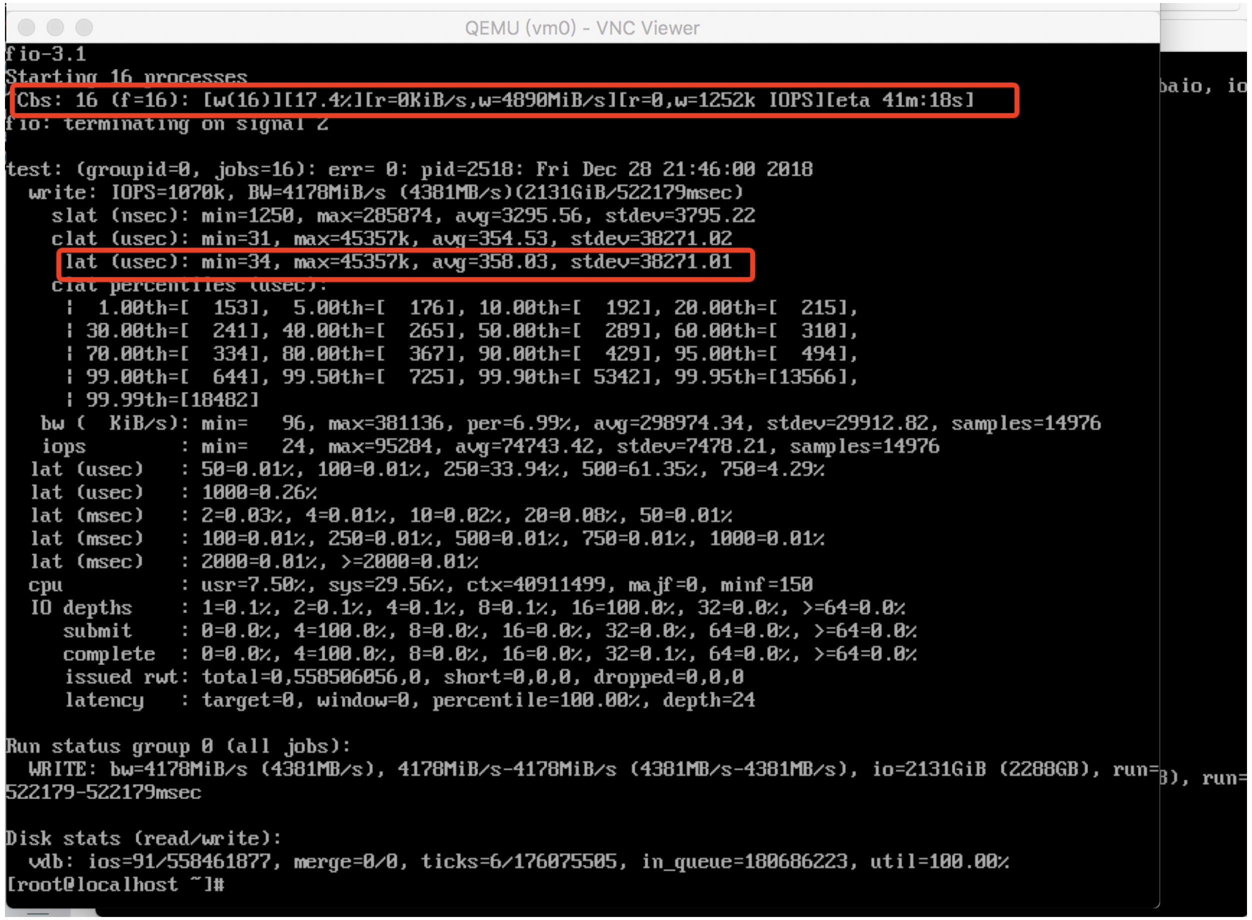
Udisk Performance –Single Chunk/Core



RandRead性能数据					
队列	IOPS	平均延迟	95%	99.95%	99.99%
1	13.0k	71.16	110	163	165
8	108k	73.71	100	126	149
16	204k	78.09	97	118	131
32	364k	87.72	102	119	139
64	566k	112.64	126	178	198
128	639k	199.97	239	343	371

Udisk Performance –Cluster

16*24 RandWrite
Lat 358us
IOPS 1252k



```
fio-3.1
Starting 16 processes
Cbs: 16 (f=16): [w(16)][17.4%][r=0KiB/s,w=4890MiB/s][r=0,w=1252k IOPS][eta 41m:18s]
fio: terminating on signal 2

test: (groupid=0, jobs=16): err= 0: pid=2518: Fri Dec 28 21:46:00 2018
  write: IOPS=1070k, BW=4178MiB/s (4381MB/s)(2131GiB/522179msec)
    slat (nsec): min=1250, max=285874, avg=3295.56, stdev=3795.22
    clat (usec): min=31, max=45357k, avg=354.53, stdev=38271.02
     lat (usec): min=34, max=45357k, avg=358.03, stdev=38271.01
    clat percentiles (usec):
     |  1.00th=[  153],  5.00th=[  176], 10.00th=[  192], 20.00th=[  215],
     | 30.00th=[  241], 40.00th=[  265], 50.00th=[  289], 60.00th=[  310],
     | 70.00th=[  334], 80.00th=[  367], 90.00th=[  429], 95.00th=[  494],
     | 99.00th=[  644], 99.50th=[  725], 99.90th=[ 5342], 99.95th=[13566],
     | 99.99th=[18482]
   bw (  KiB/s): min=   96, max=381136, per=6.99%, avg=298974.34, stdev=29912.82, samples=14976
   iops        : min=   24, max=95284, avg=74743.42, stdev=7478.21, samples=14976
  lat (usec)   : 50=0.01%, 100=0.01%, 250=33.94%, 500=61.35%, 750=4.29%
  lat (usec)   : 1000=0.26%
  lat (msec)   : 2=0.03%, 4=0.01%, 10=0.02%, 20=0.08%, 50=0.01%
  lat (msec)   : 100=0.01%, 250=0.01%, 500=0.01%, 750=0.01%, 1000=0.01%
  lat (msec)   : 2000=0.01%, >=2000=0.01%
  cpu          : usr=7.50%, sys=29.56%, ctx=40911499, majf=0, minf=150
  IO depths    : 1=0.1%, 2=0.1%, 4=0.1%, 8=0.1%, 16=100.0%, 32=0.0%, >=64=0.0%
     submit    : 0=0.0%, 4=100.0%, 8=0.0%, 16=0.0%, 32=0.0%, 64=0.0%, >=64=0.0%
     complete  : 0=0.0%, 4=100.0%, 8=0.0%, 16=0.0%, 32=0.1%, 64=0.0%, >=64=0.0%
     issued rwt: total=0,558506056,0, short=0,0,0, dropped=0,0,0
     latency   : target=0, window=0, percentile=100.00%, depth=24

Run status group 0 (all jobs):
  WRITE: bw=4178MiB/s (4381MB/s), 4178MiB/s-4178MiB/s (4381MB/s-4381MB/s), io=2131GiB (2288GB), run=
522179-522179msec

Disk stats (read/write):
  vdb: ios=91/558461877, merge=0/0, ticks=6/176075505, in_queue=180686223, util=100.00%
[root@localhost ~]#
```