SDC
STORAGE DEVELOPER CONFERENCE
SNIA SANTA CLARA, 2015

NVMe over Fabrics

Wael Noureddine, Ph.D.
Chelsio Communications Inc.

Overview

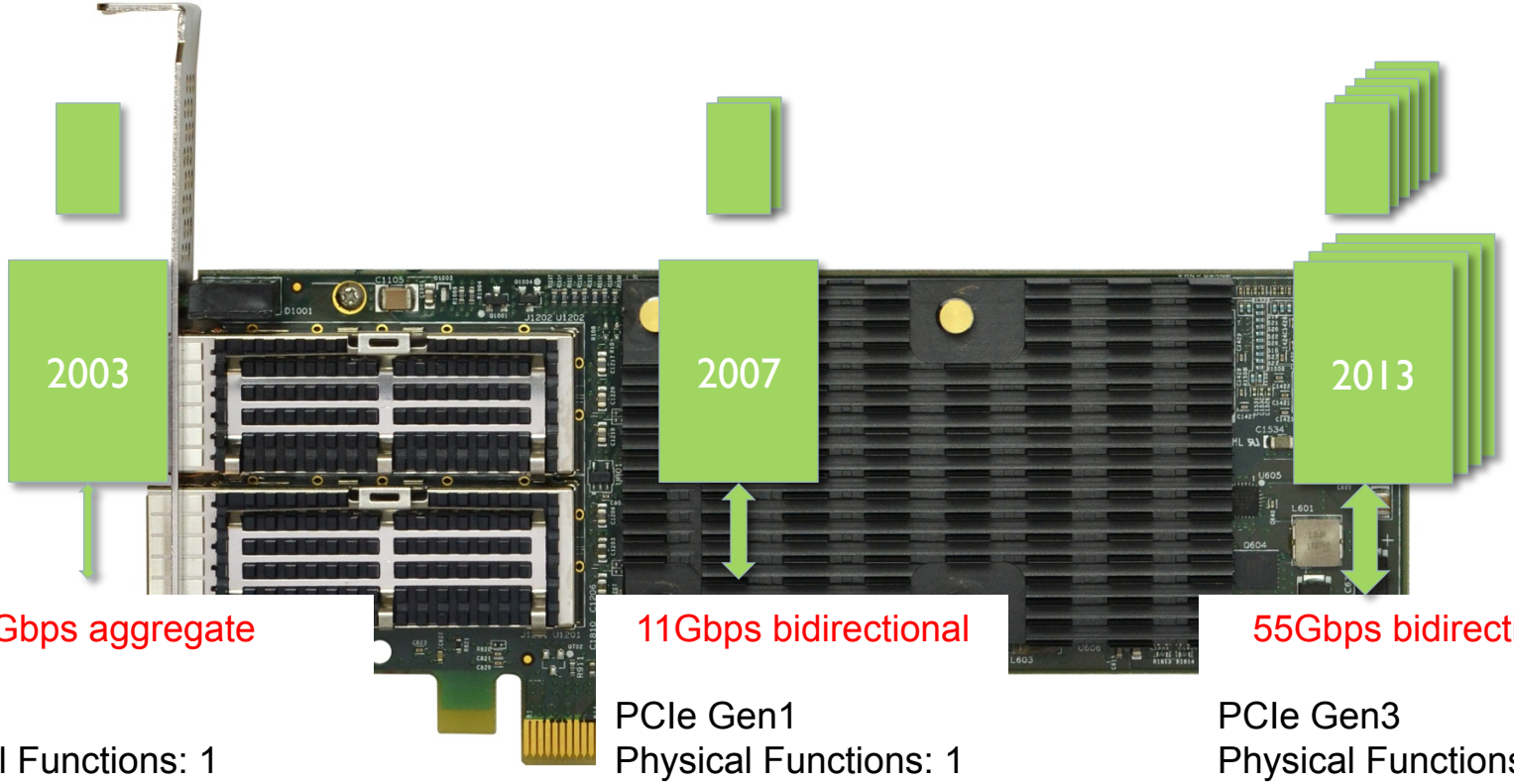
- Introduction
- NVMe/Fabrics and iWARP
- RDMA Block Device and results
- Closing notes

NVMe Devices

- Accelerating pace of progress
  - Capacity
  - Reliability
  - Efficiency
  - Performance
    - Latency
    - Bandwidth
    - IOPS – Random Access

Samsung Keynote, Flash Memory Summit, Santa Clara, August'2015

A Decade of NIC Evolution



7Gbps aggregate

PCI-X
Physical Functions: 1
DMA Queues: 1
Interrupt: Line
Logical NICs: 1
Lock-based Synchronization

11Gbps bidirectional

PCIe Gen1
Physical Functions: 1
DMA Queues: 8
Interrupt: MSI
Logical NICs: 8
Up to 8 cores

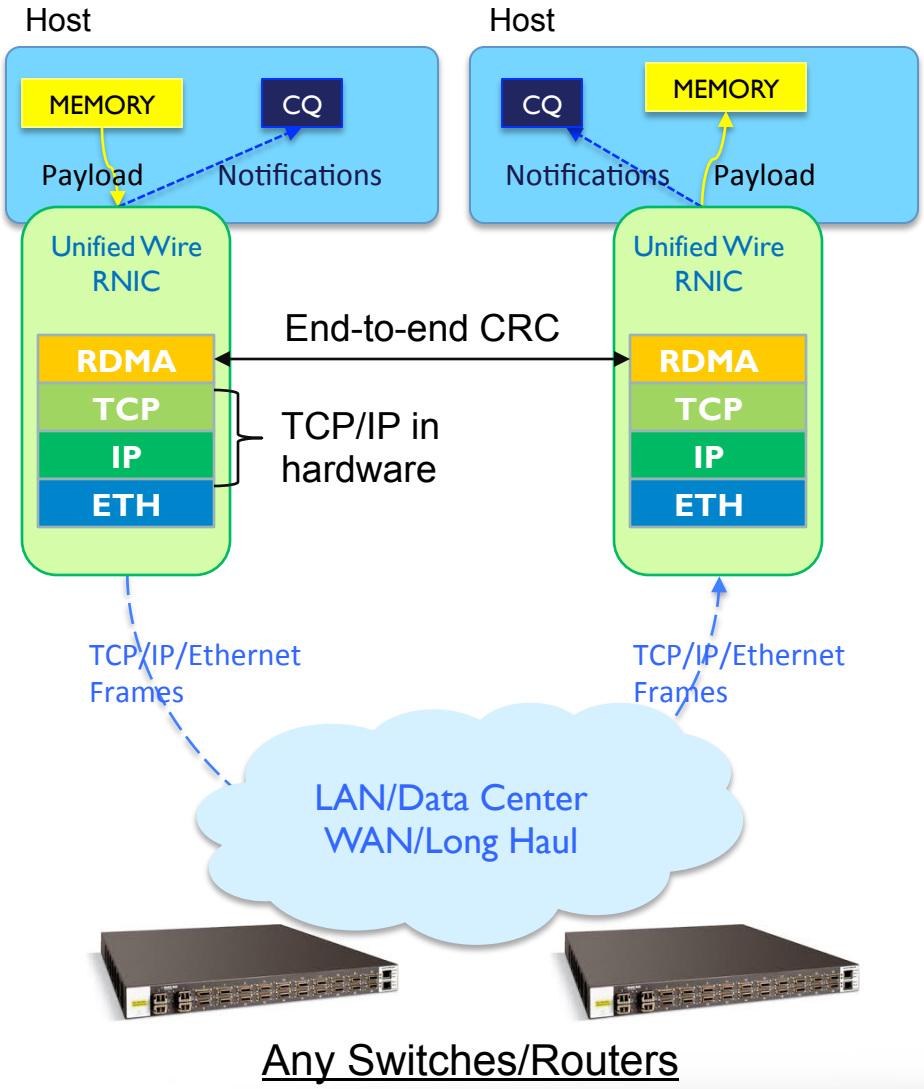
55Gbps bidirectional

PCIe Gen3
Physical Functions: 8
DMA Queues: 64K
Interrupt: MSI-X
Logical NICs: 128
Lock-free user space I/O

NVMe – Non-Volatile Memory Express

- Modern PCIe interface architecture
  - Benefit from experience with other high performance devices
  - Scale with exponential speed increases
- Standardized for enhanced plug-and-play
  - Register set
  - Feature set
  - Command set
- Parallels to high performance NIC interface
  - 64K x deep queues
  - MSI-X
  - Lock-free operation
- Security
- Data integrity protection

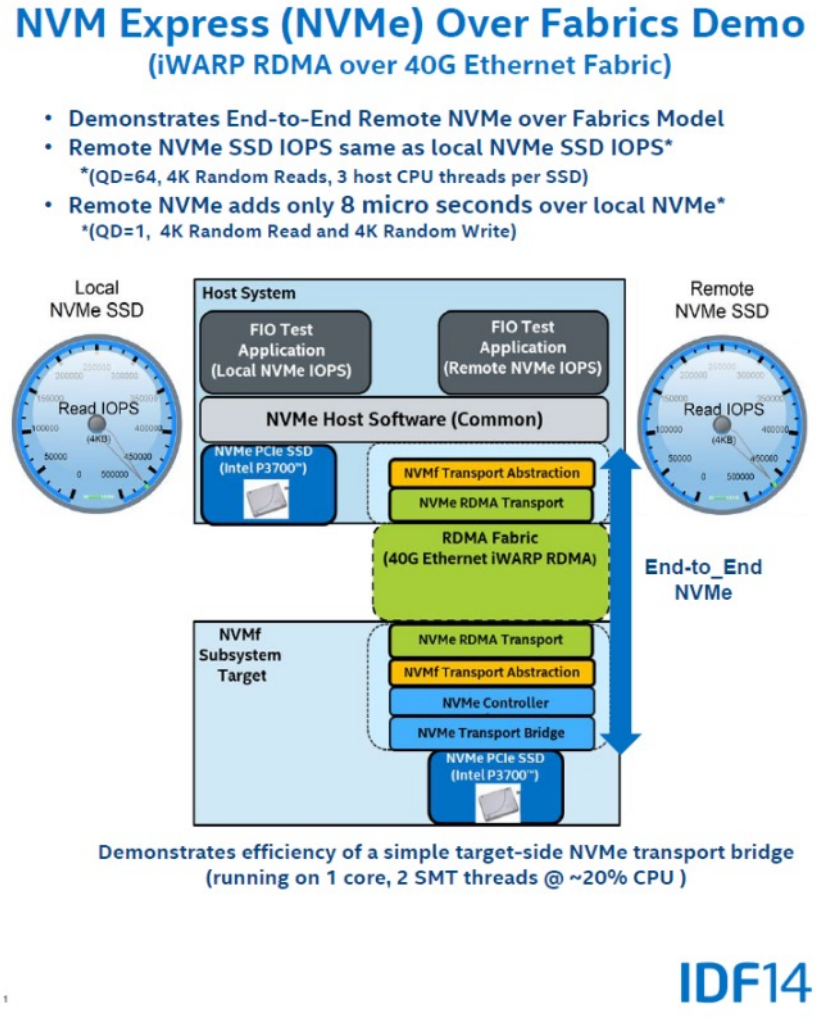
iWARP RDMA over Ethernet



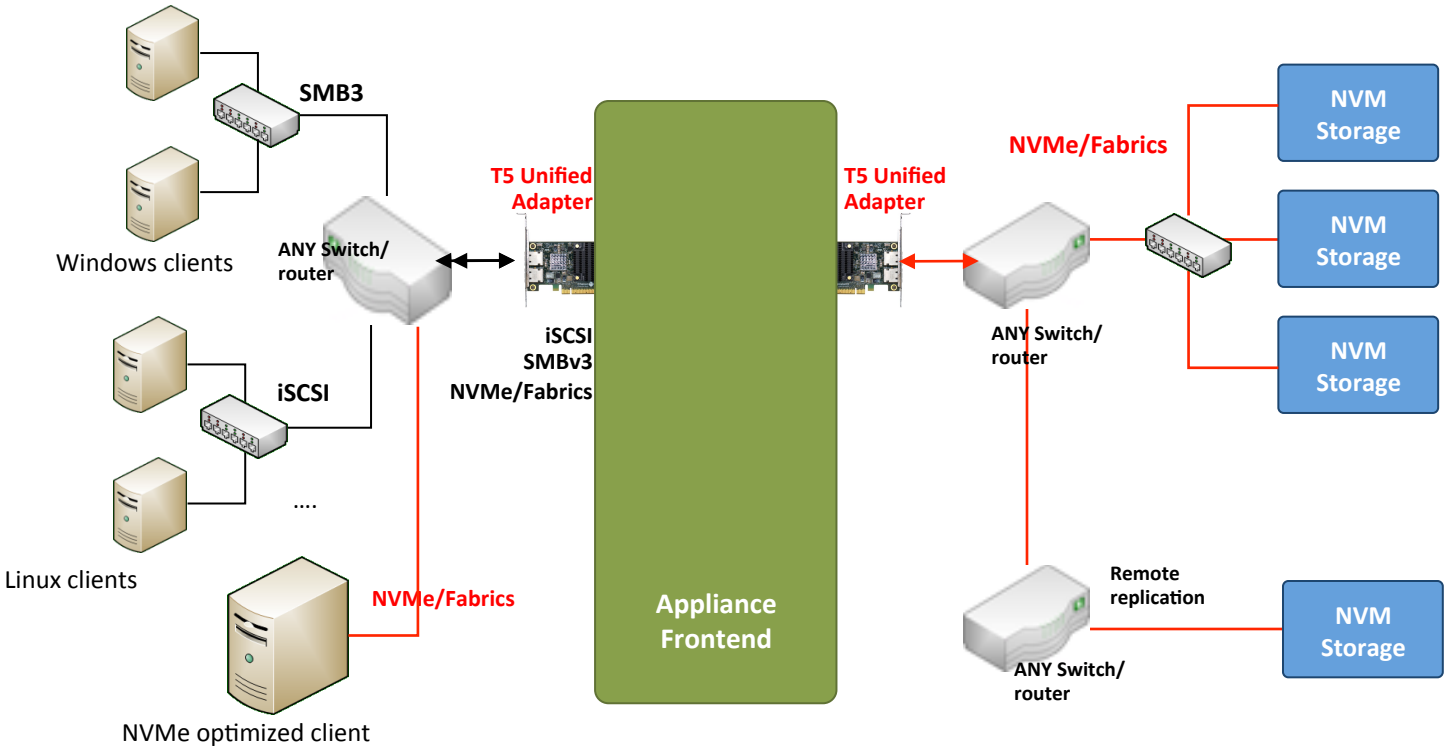
- Open, transparent and mature
  - IETF Standard (2007)
- Plug-and-play
  - Standard and proven TCP/IP
  - Built-in reliability congestion control, flow control
  - Natively routable
- Cost effective
  - Regular switches
  - Same network appliances
  - Works with DCB but NOT required
  - No need for lossless configuration
- No network restrictions
  - Architecture, scale, distance, RTT, link speeds
- Hardware performance
  - Exception processing in HW
  - Ultra low latency
  - High packet rate and bandwidth

NVMe over Fabrics

- Address PCI reach and scalability limits
- Extend native NVMe over large fabric
- iWARP RDMA over Ethernet first proof point
  - Intel NVMe drives
  - Chelsio 40GbE iWARP
  - Compared to local disk
    - Less than 8usec additional RTT latency
    - Same IOPS performance

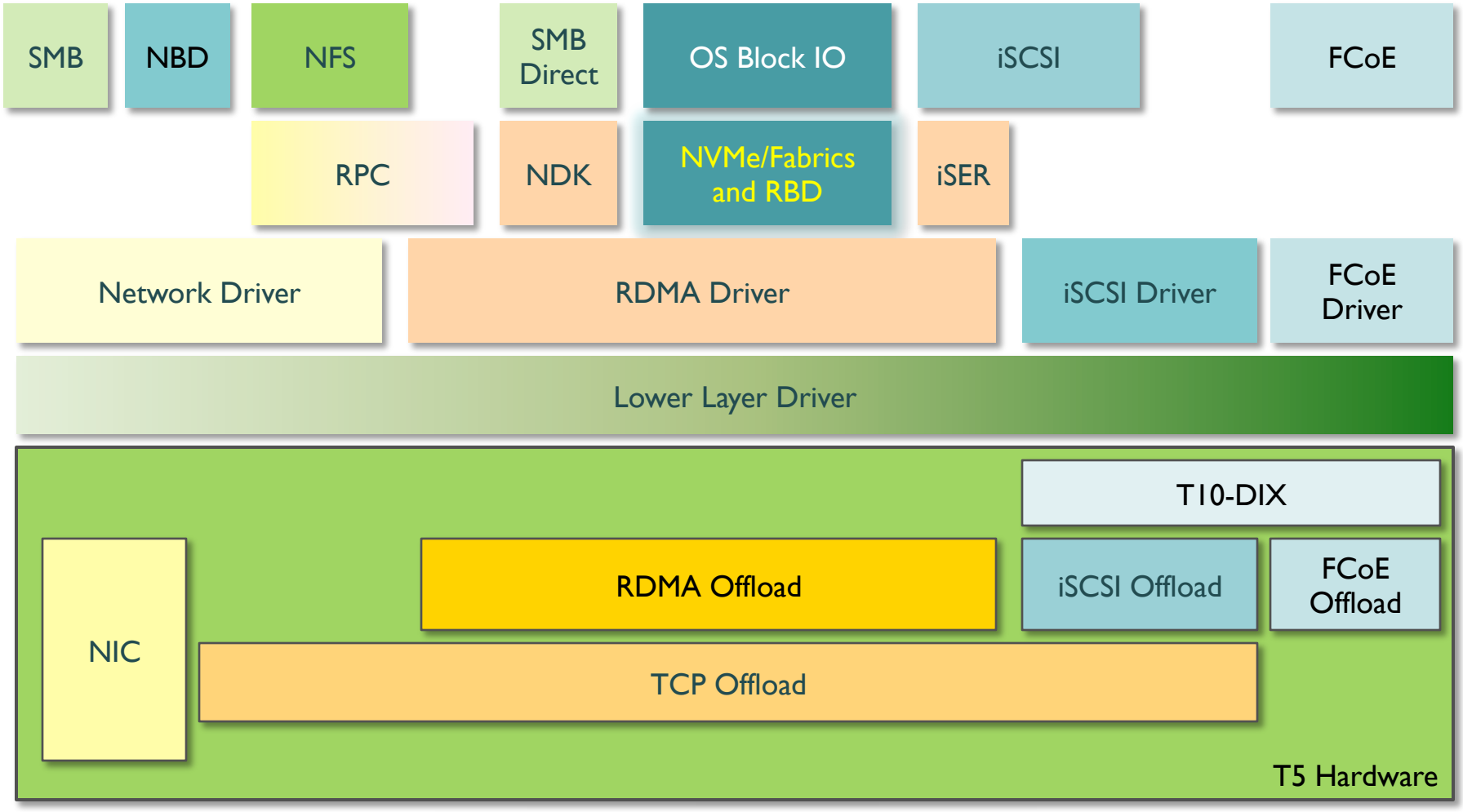


NVMe/Ethernet Fabric with iWARP

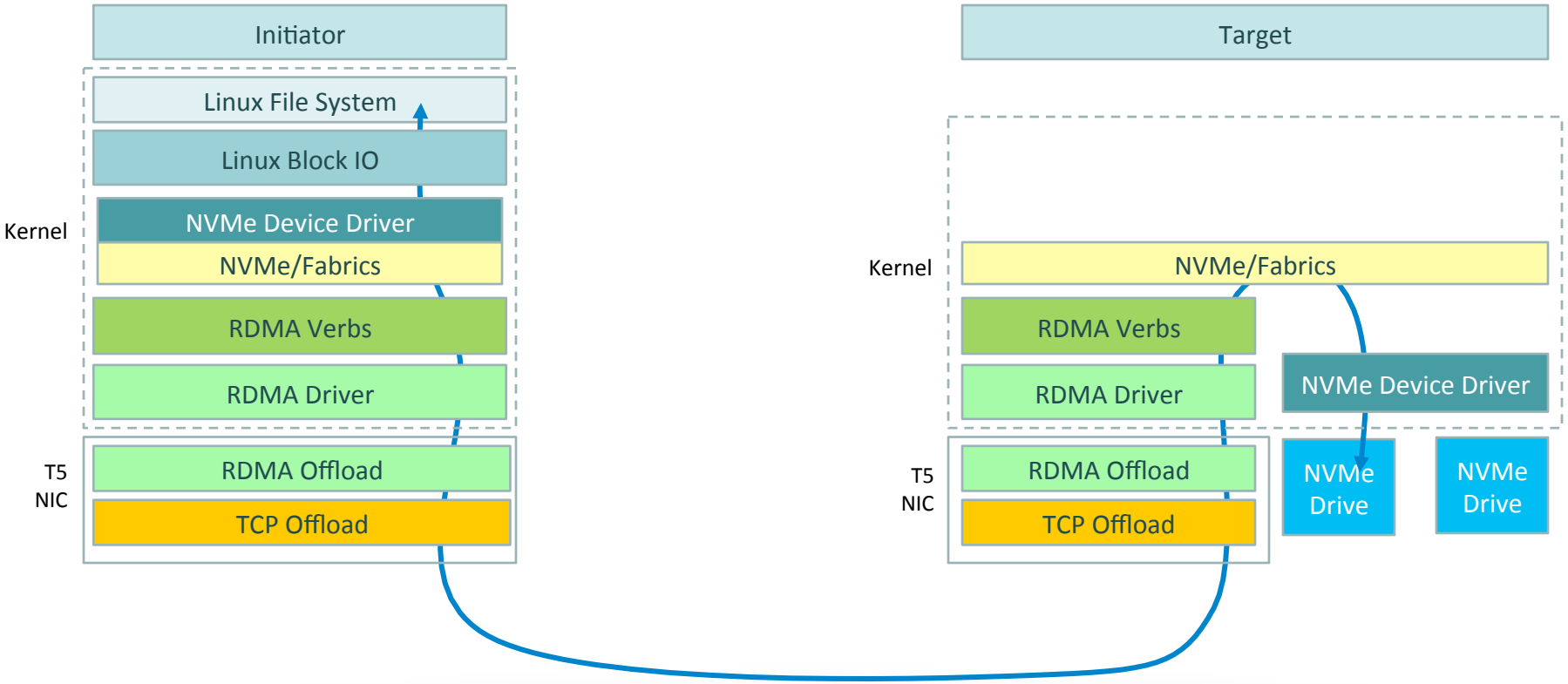


T5 supports iSCSI, SMBDirect and NVMe/Fabrics offload simultaneously

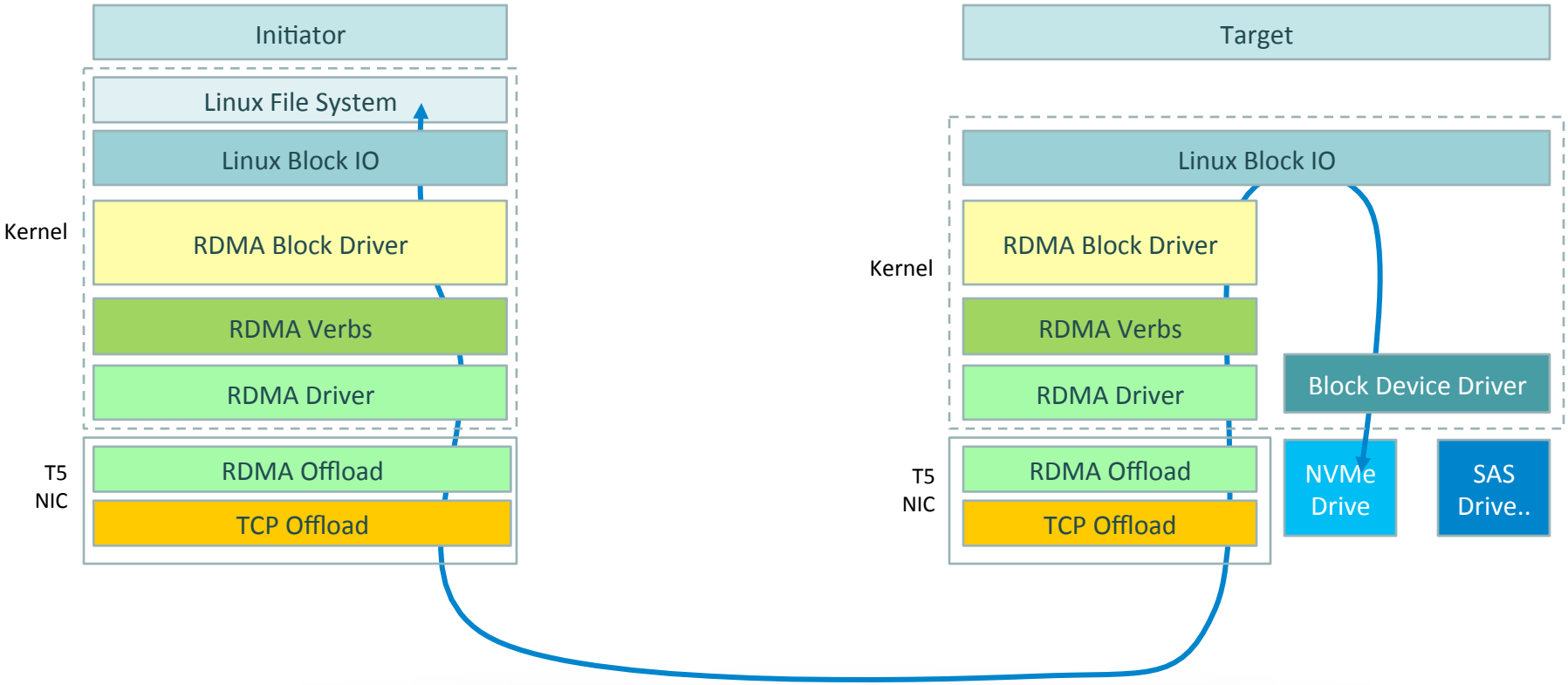
T5 Storage Protocol Support



NVMe/Fabrics Layering



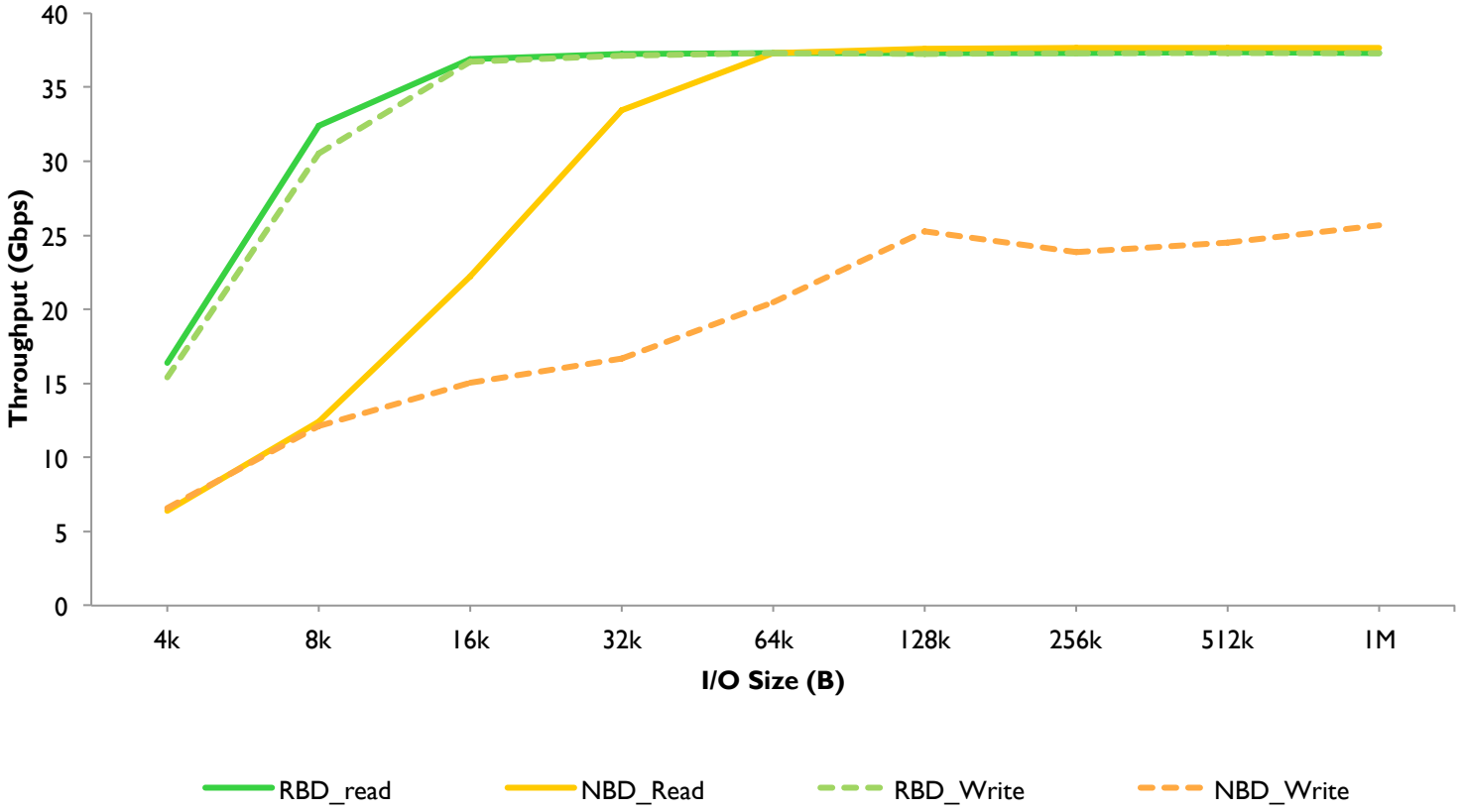
RBD Layering



Generic RDMA Block Device (RBD)

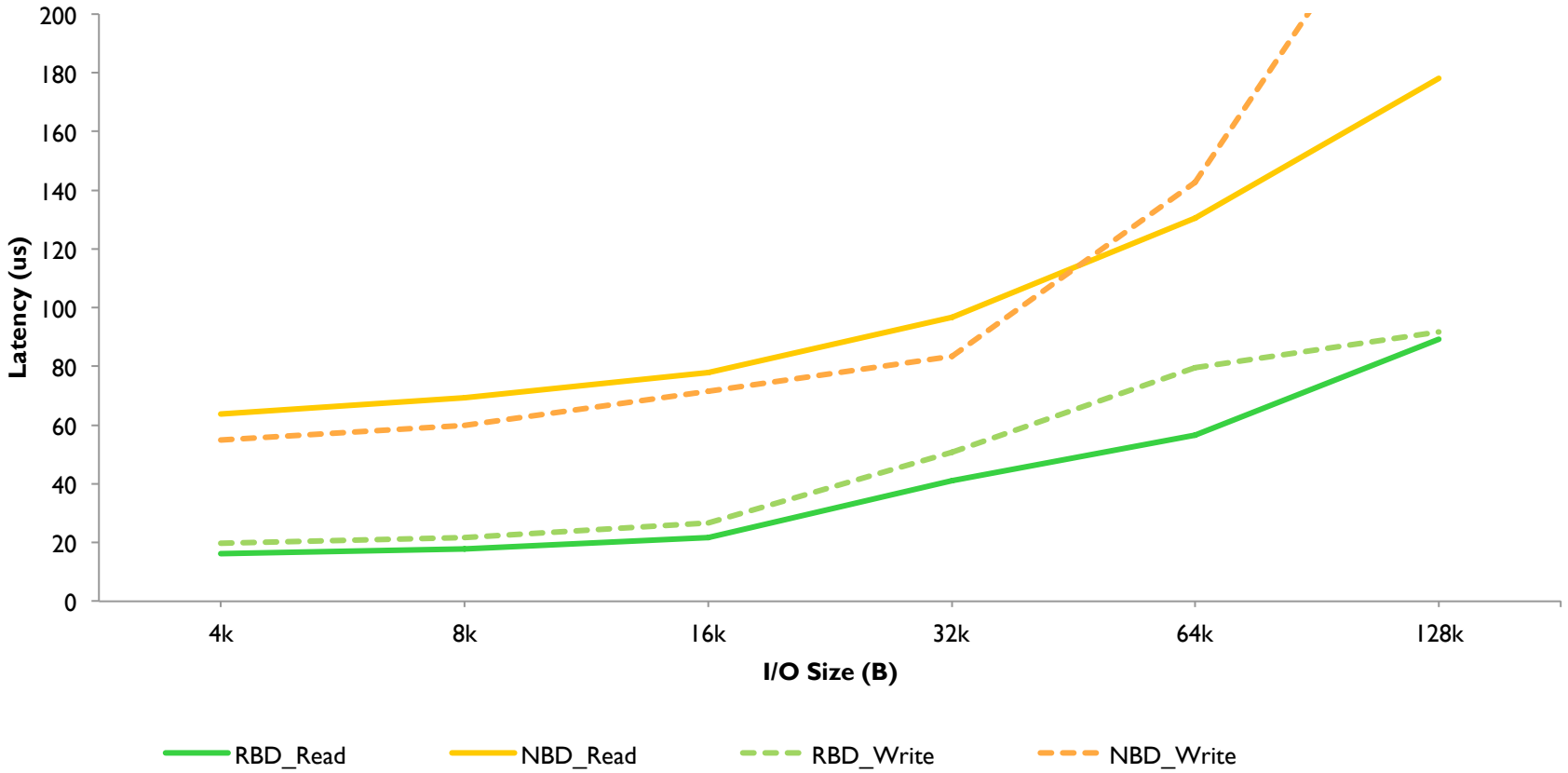
- Open Source Chelsio Linux driver
  - Part of Chelsio's Unified Wire SW
- Virtualizes block devices over an RDMA connection
- Target
  - Claims the physical device using blkdev_get_by_path()
  - Submits IOs from the Initiator via submit_bio()
- Initiator
  - Registers as a BIO device driver
  - Registers connected Target devices as a BIO device using add_disk()
- Current RBD Protocol Limits
  - Max RBD IO size is 128KB
  - 256 max outstanding IOs in flight
  - Initiator physical/logical sector size is 4KB
- Management via rbdctl command
  - Initiator provides character device command interface for adding/ removing devices
  - Simple command, rbdctl, used to connect, login, and register the target devices
- Examples
  - Target
    - Load `rbdt` module
  - Initiator
    - Load `rbdi` module
    - Add a new target: *`rbdct -n -a <ipaddr> -d /dev/nvme0n1`*
    - Remove an RBD device: *`rbdctl -r -d /dev/rbdi0`*
    - List connected devices: *`rbdctl -l`*

RBD vs. NBD Throughput – 40GbE



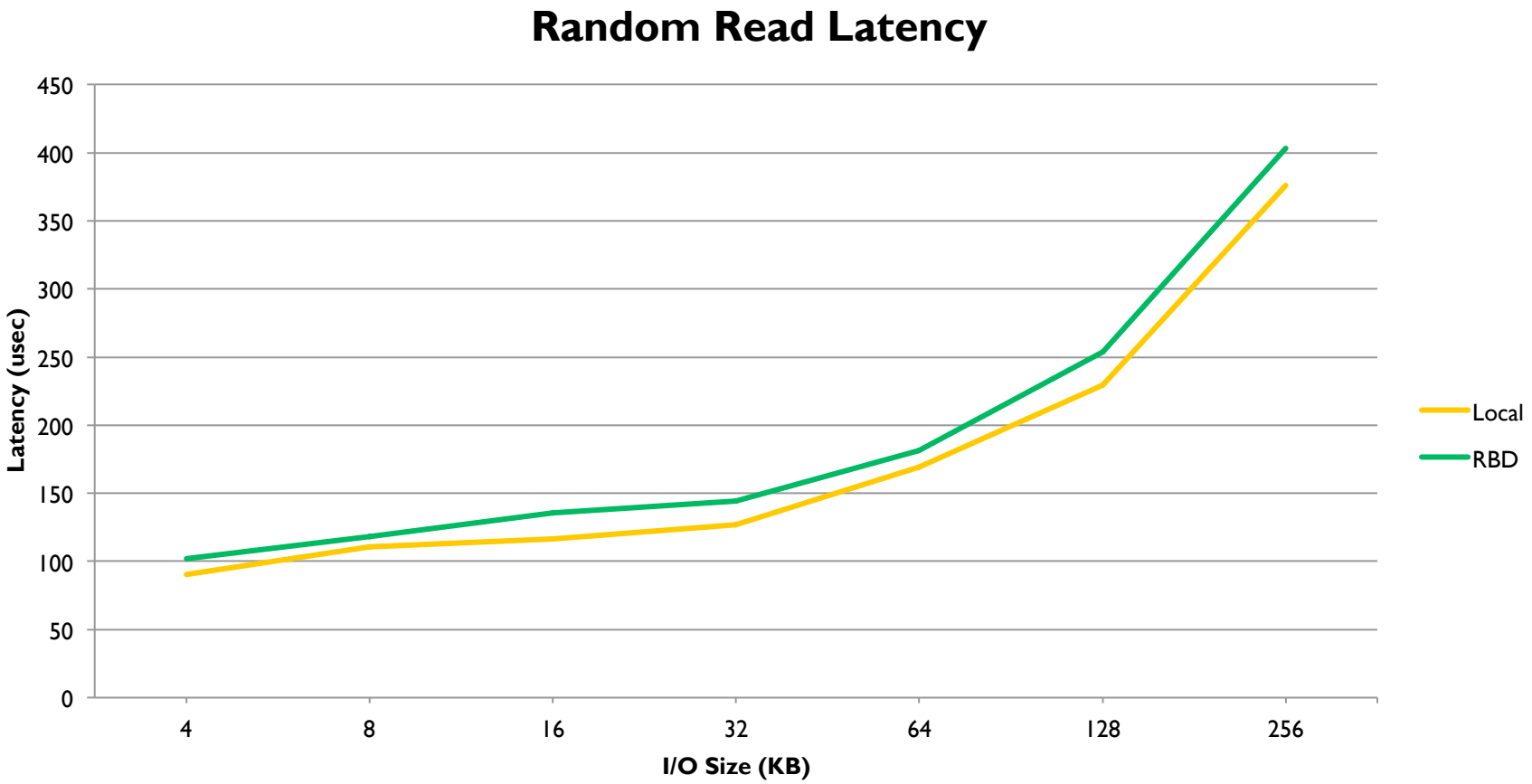
Intel Xeon CPU E5-1660 v2 6-core processor clocked at 3.70GHz (HT enabled) with 64 GB of RAM,
Chelsio T580-CR, RHEL 6.5 (Kernel 3.18.14), standard MTU of 1500B, using RAMdisks on target (4)
fio --name=<test_type> --iodepth=32 --rw=<test_type> --size=800m --direct=1 --invalidate=1 --fsync_on_close=1 --norandommap --group_reporting --ioengine=libaio --numjobs=4 --bs=<block_size> --runtime=30 --time_based --filename=/dev/rbdi0

RBD vs. NBD Latency – 40GbE



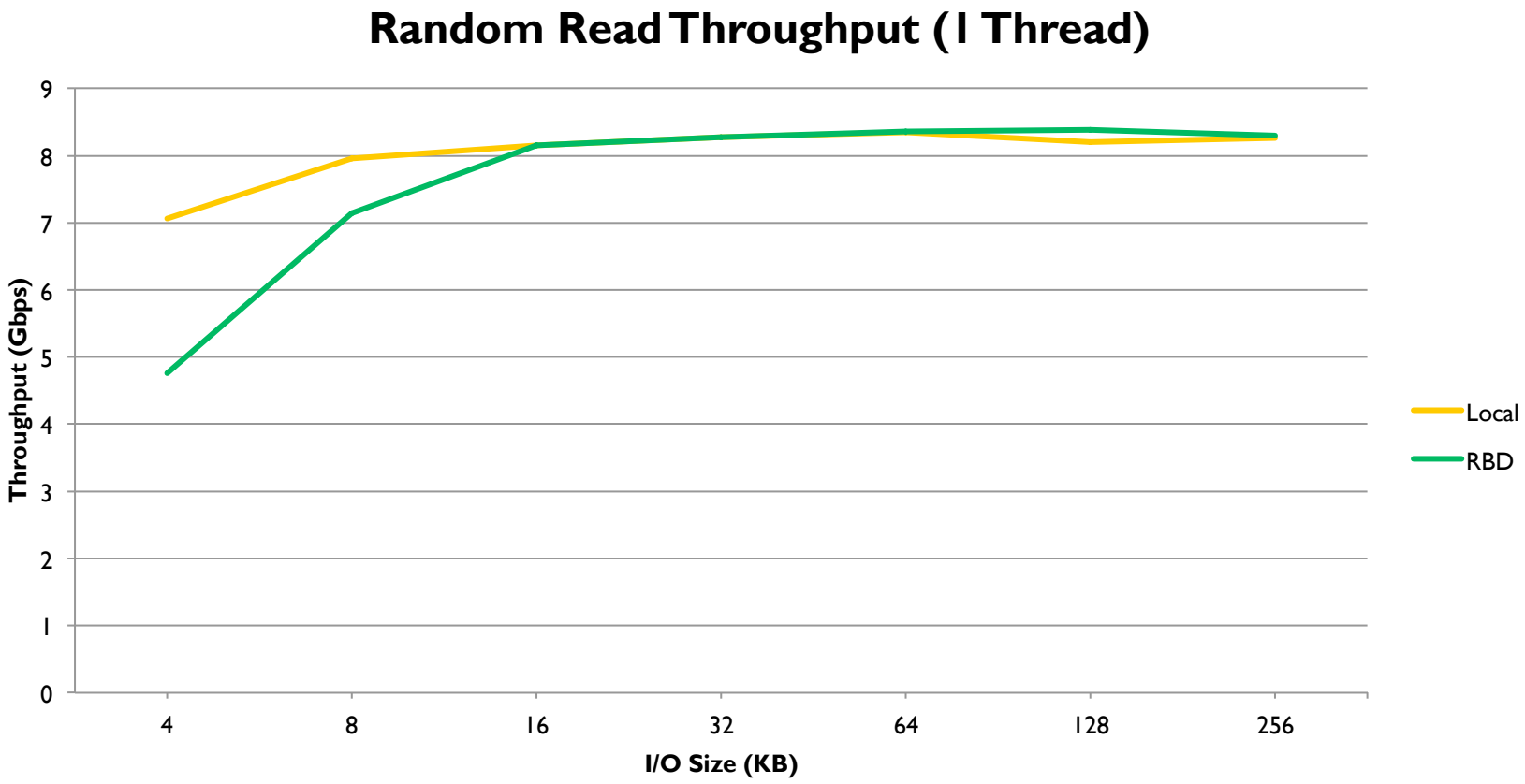
Intel Xeon CPU E5-1660 v2 6-core processor clocked at 3.70GHz (HT enabled) with 64 GB of RAM.
Chelsio T580-CR, RHEL 6.5 (Kernel 3.18.14), standard MTU of 1500B, using RAMdisk on target (1)
fio --name=<test_type> --iodepth=1 --rw=<test_type> --size=800m --direct=1 --invalidate=1 --fsync_on_close=1 --norandommap --group_reporting --ioengine=libaio --numjobs=1 --bs=<block_size> --runtime=30 --time_based --filename=/dev/rbdi0

SSD Latency Results



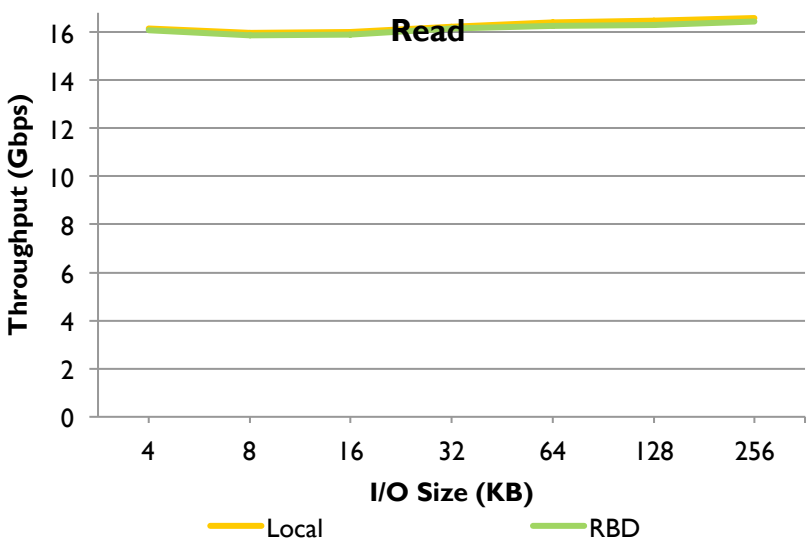
Supermicro MB, 16 cores - Intel(R) Xeon(R) CPU E5-2687W 0 @ 3.10GHz, 64GB memory
Chelsio T580-CR, RHEL6.5, linux-3.18.14 kernel, standard MTU of 1500B, through Ethernet switch, NVMe SSD target
fio-2.1.10, 1 thread, direct IO, 30 second runs, libaio IO engine, IO depth = 1, random read, write, and read-write.

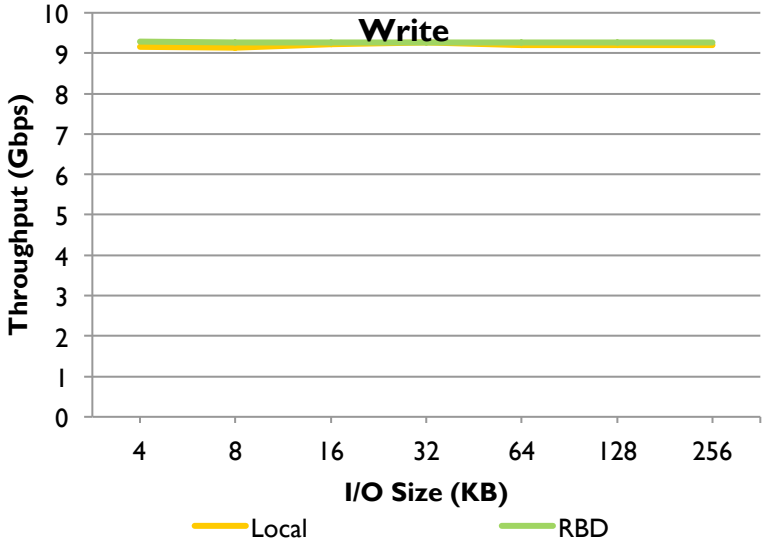
SSD Bandwidth Results



Supermicro MB, 16 cores - Intel(R) Xeon(R) CPU E5-2687W 0 @ 3.10GHz, 64GB memory
Chelsio T580-CR, RHEL6.5, linux-3.18.14 kernel, standard MTU of 1500B, through Ethernet switch, SSD target
fio-2.1.10, 1 thread, direct IO, 30 second runs, libaio IO engine, IO depth = 32, random read, write, and read-write.

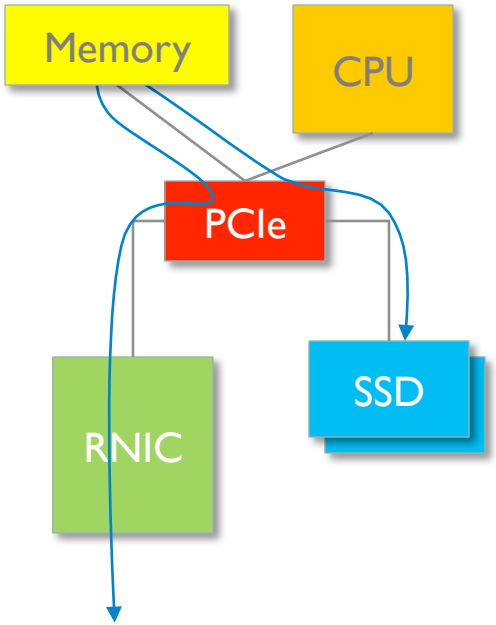
Local vs. RBD Bandwidth – 2 SSD





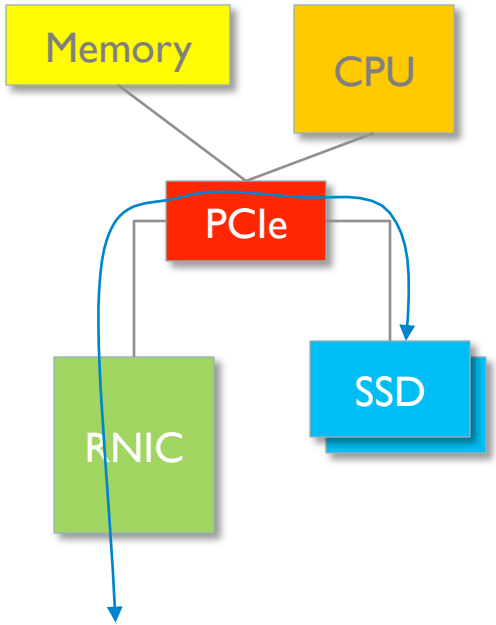
Intel(R) Xeon(R) CPU E5-2687W 0 @ 3.10GHz 2x6 cores 64GB RAM, 2xNVMe SSD
Chelsio T580-CR, CentOS Linux release 7.1.1503 Kernel 3.17.8, standard MTU of 1500B through 40Gb switch
fio --direct=1 --invalidate=1 --fsync_on_close=1 --norandommap --group_reporting --name=foo --bs=$2 --size=800m --numjobs=8 --ioengine=libaio --iodepth=16 --rw=randwrite --runtime=30 --time_based --filename_format=/dev/rbdi\$jobnum --directory=/

RDMA over Fabrics Evolution



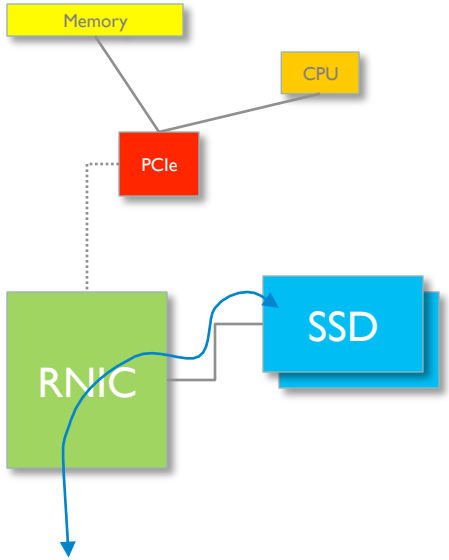
DMA to/from host memory
CPU handles command transfer

YESTERDAY



Peer-to-peer DMA through PCI bus
No host memory use

TODAY



Network integrated NVMe storage
No CPU or memory involvement

TOMORROW

Conclusions

- NVMe over Fabrics for scalable high performance and high efficiency fabrics that leverage disruptive solid-state storage
- Standard interfaces and middleware
  - Ease of use and widespread support
- iWARP RDMA ideal for NVMe/Fabrics
  - Familiar and proven foundation
  - Plug-and-play
    - Works with DCB but does not require it
    - No DCB tax and complexity
  - Scalability, reliability, routability
    - Any switches and any routers over any distance
    - Scales from 40 to 100Gbps and beyond
- NVMe/Fabrics devices with built-in processing power