High performance 100G iSCSI Storage Solution

Chelsio T6 iSCSI Throughput, IOPS, Latency and CPU Utilization

Executive Summary

This paper highlights Chelsio 100G iSCSI offload's support for remote storage access over a standard, time-tested and reliable Ethernet infrastructure. The results demonstrate Chelsio T6's exceptional performance with 100Gb line-rate throughput numbers and IOPS of more than 2.8 Million at 4K I/O size. With half of the processing resources of Target machine free even at maximum usage, and latency delta of only 15 µs between remote and local storage access, Chelsio's T6 provides significant performance and efficiency gains to datacenters with power hungry applications.

The Chelsio iSCSI Offload Solution

The Terminator 6 (T6) ASIC from Chelsio Communications, Inc. is a sixth generation, high performance 1/10/25/40/50/100Gbps unified wire engine which offers storage protocol offload capability for accelerating both block (iSCSI, FCoE) and file (SMB, NFS, Object) level storage traffic. Chelsio iSCSI Offload solution runs at 100Gb and beyond, and will scale consistently with Ethernet evolution. Chelsio's proven TCP Offload Engine (TOE), offloaded iSCSI over T6 enjoys a distinct performance advantage over regular NIC. The T6 unified wire engine offers PDU iSCSI offload capability in protocol acceleration for both file and block-level storage (iSCSI) traffic.

Test Results

The following graphs plot the READ, WRITE IOPS, Throughput and Target CPU Usage of Chelsio T62100-CR adapter with and without Digest, using Ram disk and SSD as storage array. The results are collected using **fio** tool with I/O size used varying from 4K to 512K bytes and an access pattern of random READs and WRITEs.

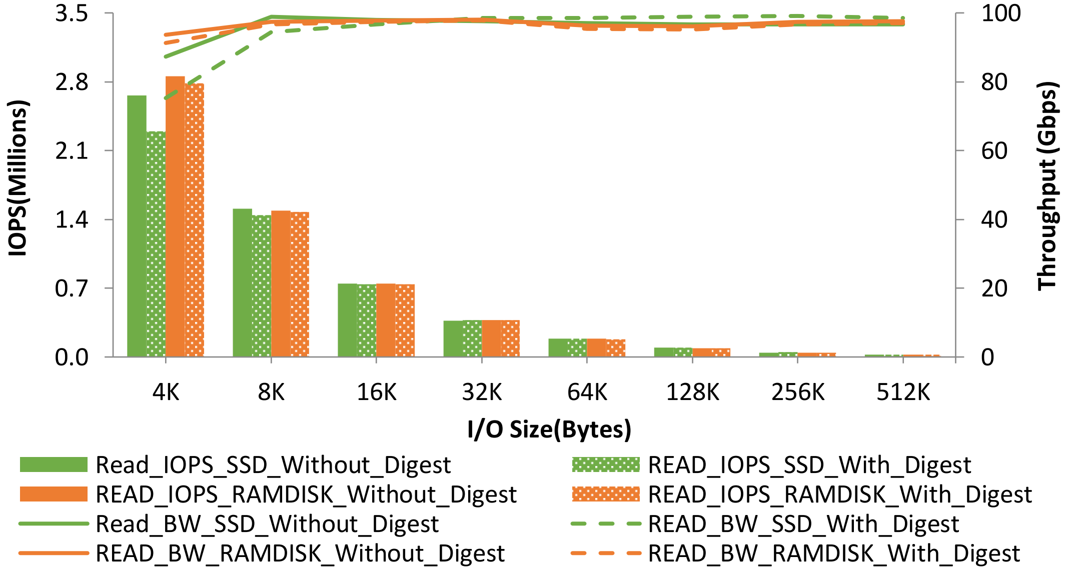


Figure 1 – READ Throughput & IOPS vs. I/O size

T6 iSCSI solution delivers line-rate READ throughput and 2.8 Million IOPS (at 4K I/O size) with both SSDs and Ram disks. The performance is similar with or without Digest, indicative of an efficient processing path.

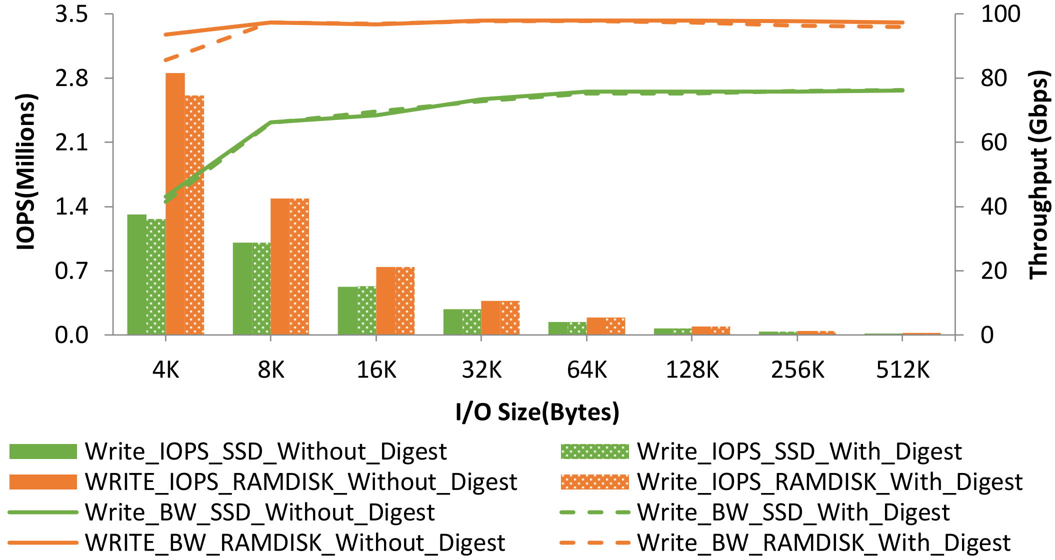


Figure 2 - WRITE Throughput & IOPS vs. I/O size

T6 iSCSI solution delivers line-rate WRITE throughput and 2.8 Million IOPS (at 4K I/O size) with Ram disks. With SSDs, the numbers are limited by number of SSDs in the test topology. *Further Performance tuning is in progress.*

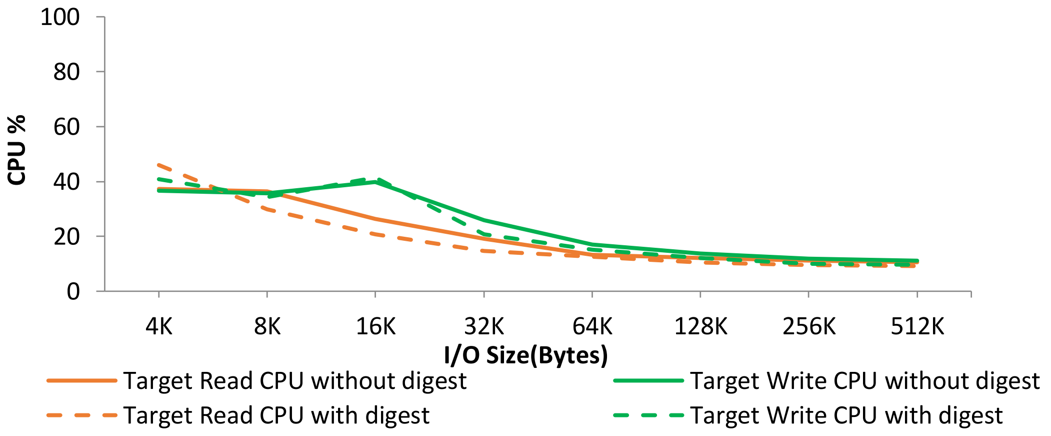


Figure 3 – Target % CPU vs. I/O size (Ram disk)

CPU savings on the Target is considerable for both READ and WRITE operations, peaking at less than 50% and gradually diminishing as I/O size increases.

The following graph presents WRITE, READ latency results of T6 iSCSI solution with SSD as storage array. The results are collected using the **fio** tool with I/O size varying from 512 to 64K bytes with an access pattern of random READs and WRITEs.

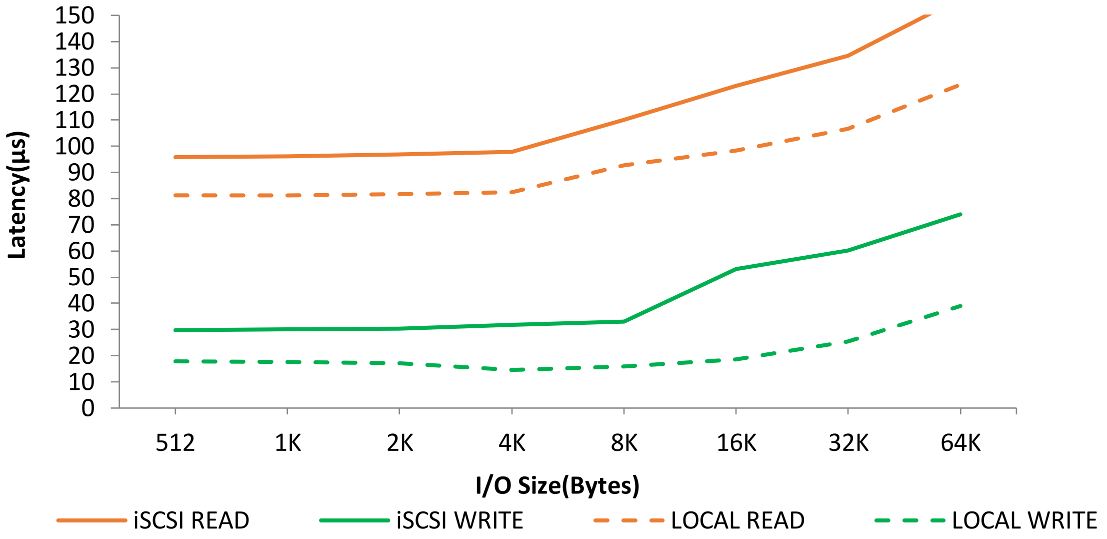
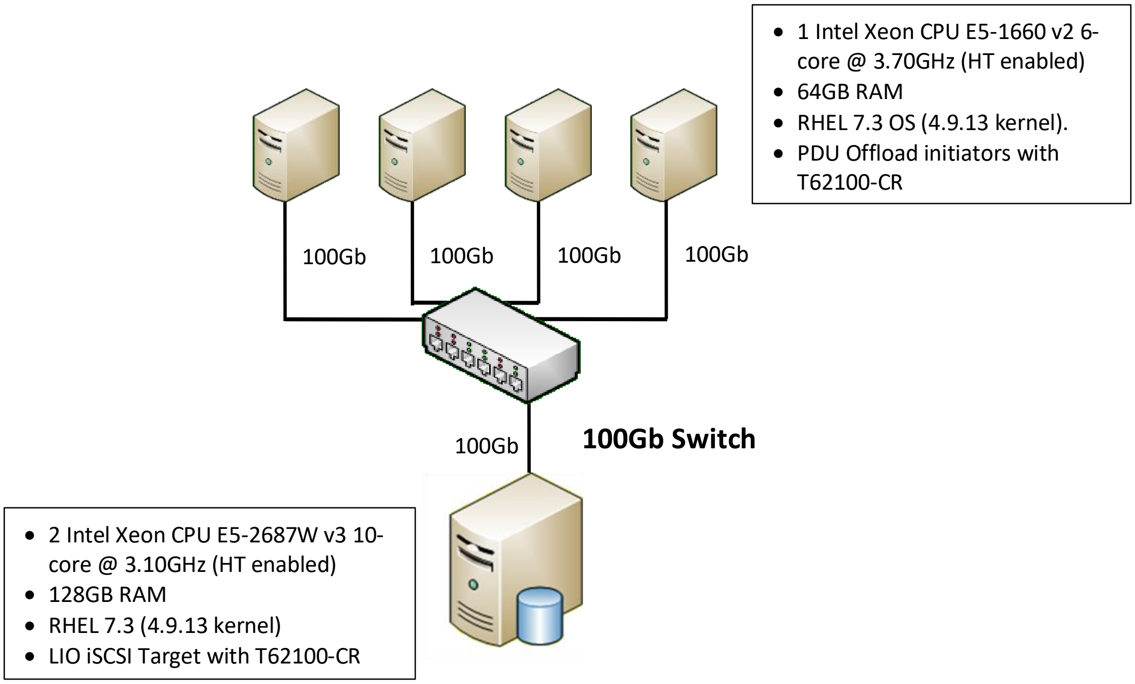


Figure 4 – Latency vs. I/O size (SSD)

Chelsio's T6 iSCSI solution offers negligible remote storage access latency compared to local, with 17 µs delta for WRITE and 15 µs for READ at 4K I/O size.

Test Setup

The following sections provide the test setup and configuration details.



- 1 Intel Xeon CPU E5-1660 v2 6-core @ 3.70GHz (HT enabled)
- 64GB RAM
- RHEL 7.3 OS (4.9.13 kernel).
- PDU Offload initiators with T62100-CR

- 2 Intel Xeon CPU E5-2687W v3 10-core @ 3.10GHz (HT enabled)
- 128GB RAM
- RHEL 7.3 (4.9.13 kernel)
- LIO iSCSI Target with T62100-CR

Figure 5 -Test Setup

Network Configuration

The iSCSI setup consists of a target storage array connected to 4 iSCSI initiator machines through a 100GbE switch using single port on each system. MTU of 9000B is used.

Storage Configuration

In SSD scenario, the target is configured with 32 LUNs from 5 Samsung PM1725 1.6TB SSDs, each of 1GB size. In Ram disk scenario, the target is configured with 32 Ram disk LUNs, each of 1GB size.

- For throughput, IOPS and CPU Usage test, each iSCSI initiator connects to 8 targets.
- For latency test, 1 iSCSI initiator connects to 1 target.

Setup Configuration

Tunings

BIOS: Virtualization, c-state technology, VT-d, Intel I/O AT and SR-IOV disabled.
Kernel command line: `scsi_mod.use_blk_mq=Y`

Target Configuration

i. LIO Offload target driver was loaded:

```
[root@host~]# modprobe cxgbit
```

ii. Chelsio interface was assigned with IPv4 address and brought-up.

iii. CPU affinity was set

```
[root@host~]#t4_perftune.sh -n -Q iSCSIT
```

iv. Create 32 LUNs from the SSDs.

```
[root@host~]# pvcreate /dev/nvme0n2
[root@host~]# pvcreate /dev/nvme1n1
[root@host~]# pvcreate /dev/nvme2n1
[root@host~]# pvcreate /dev/nvme3n1
[root@host~]# pvcreate /dev/nvme4n1
[root@host~]# vgcreate VG_1 /dev/nvme0n1 /dev/nvme1n1 /dev/nvme2n1
/dev/nvme3n1 /dev/nvme4n1
[root@host~]# for i in `seq 1 32`;do lvcreate -i5 -L1G -n lun$i VG_1;done
```

v. Configured target using *targetcli*:

```
[root@host~]# for i in `seq 1 32 `; do targetcli /iscsi create iqn.2017-
01.org.linux-iscsi.target$i ; done
[root@host~]# for i in `seq 1 32 `; do targetcli /iscsi/iqn.2017-
01.org.linux-iscsi.target$i/tpg1/ set attribute authentication=0
demo_mode_write_protect=0 generate_node_acls=1 cache_dynamic_acls=1  ; done
[root@host~]# for i in `seq 1 32 `; do targetcli /iscsi/iqn.2017-
01.org.linux-iscsi.target$i/tpg1/portals/ delete ip_address=0.0.0.0
ip_port=3260 ; done
[root@host~]# for i in `seq 1 32 `; do targetcli /iscsi/iqn.2017-
01.org.linux-iscsi.target$i/tpg1/portals create ip_address=10.1.1.42
ip_port=3260 ; done
```

SSD

```
[root@host~]# for i in `seq 1 32 `; do targetcli /backstores/block create
dev=/dev/VG_1/lun$i name=lun$i ; done
[root@host~]# for i in `seq 1 32 `; do targetcli /iscsi/iqn.2017-
01.org.linux-iscsi.target$i/tpg1/luns create lun=0
storage_object=/backstores/block/lun$i  ; done
```

Ram disk

```
[root@host~]# for i in `seq 1 32 `; do targetcli /backstores/ramdisk/ create
nullio=false size=600M name=ramdisk$i ; done
[root@host~]# for i in `seq 1 32 `; do targetcli /iscsi/iqn.2017-
01.org.linux-iscsi.chelsio.target$i/tpg1/luns create lun=0
storage_object=/backstores/ramdisk/ramdisk$i  ; done
```

vi. Enable LIO Target Offload

```
[root@host~]# for i in `seq 1 32 `; do echo 1 >
/sys/kernel/config/target/iscsi/iqn.2017-01.org.linux-
iscsi.target$i/tpgt_1/np/10.1.1.42\:3260/cxgbit ; done
```

Initiator Configuration

i. iSCSI Initiator driver was loaded:

```
[root@host~]# modprobe cxgb4i
```

ii. Chelsio interface was assigned with IPv4 address and brought-up.

iii. CPU affinity was set:

```
[root@host~]# t4_perftune.sh -n -Q iSCSI
```

iv. Target was discovered using *cxgb4i* iface

```
[root@host~]# iscsiadm -m discovery -t st -p 10.1.1.42 -I <cxgb4i_iface>
```

v. Logged in to 8 targets from initiator1.

```
[root@host~]# for i in `seq 1 8`; do iscsiadm -m node -T iqn.2017-
01.org.linux-iscsi.target${i} -p 10.1.1.42 -I <cxgb4i_iface> -l; done
```

vi. Logged in to 8 targets from initiator2.

```
[root@host~]# for i in `seq 9 16`; do iscsiadm -m node -T iqn.2017-
01.org.linux-iscsi.target${i} -p 10.1.1.42 -I <cxgb4i_iface> -l; done
```

vii. Logged in to 8 targets from initiator3.

```
[root@host~]# for i in `seq 17 24`; do iscsiadm -m node -T iqn.2017-
01.org.linux-iscsi.target${i} -p 10.1.1.42 -I <cxgb4i_iface> -l; done
```

viii. Logged in to 8 targets from initiator4.

```
[root@host~]# for i in `seq 25 32`; do iscsiadm -m node -T iqn.2017-
01.org.linux-iscsi.target${i} -p 10.1.1.42 -I <cxgb4i_iface> -l; done
```

ix. fio tool was run on all 4 initiators for Throughput, IOPS and CPU Usage test.

```
[root@host~]# fio --rw=<randwrite/randread> --name=random --norandommap --
ioengine=libaio --size=400m --group_reporting --exitall --fsync_on_close=1 -
-invalidate=1 --direct=1 --
filename=/dev/sdb:/dev/sdc:/dev/sdd:/dev/sde:/dev/sdf:/dev/sdg:/dev/sdh:/dev
/sdi --time_based --runtime=30 --iodepth=64 --numjobs=16 --unit_base=1 --
bs=<IO_size> --kb_base=1000
```

x. fio tool was run on 1 initiator connected to 1 target for Latency test.

```
[root@host ~]# fio --rw=<randread/randwrite> --name=random --norandommap --
ioengine=libaio --size=400m --group_reporting --exitall --fsync_on_close=1 -
-invalidate=1 --direct=1 --filename=/dev/sdb --time_based --runtime=30 --
iodepth=1 --numjobs=1 --unit_base=1 --bs=<IO_size> --kb_base=1000
```

xi. To enable Digest on initiator machines, the following lines were added in *iscsid.conf*.

```
[root@host ~]# cat /etc/iscsi/iscsid.conf|grep Digest|grep -v "#"
node.conn[0].iscsi.HeaderDigest = CRC32C,None
node.conn[0].iscsi.DataDigest = CRC32C,None
```

Conclusion

This paper provided 100GbE iSCSI throughput, IOPS, Latency and %CPU performance results for Chelsio's T62100-CR adapter with and without Digest enabled. Numbers were collected for READ and WRITE operations using SSD and Ram disk as storage. The results show that T6 iSCSI solution:

- delivers 100Gb line-rate throughput performance for READ using both SSDs and Ram disks, and for WRITE using Ram disks. WRITE numbers with SSD are limited by number of SSDs used in the test.
- reaches more than 2.8 Million IOPS at 4K I/O for both READ and WRITE.
- delivers remote storage access delta latency of 17 µs for WRITE and 15 µs for READ operations, compared to local storage.
- uses less than 50% CPU on the target, freeing up processing resources for other memory intensive applications.
- does not show any performance degradation with Digest enabled.

Related Links

The Chelsio Terminator 6 ASIC
High Performance iSCSI at 100GbE
Chelsio T6 100G iSCSI Demonstration
100G iSCSI – A Bright Future for Ethernet Storage