T5 Offloaded iSCSI with T10-DIX

High Performance End-to-End Data Integrity for Ethernet SANs

Executive Summary

There are many sources of data integrity errors within storage networks and devices. The T10 (SCSI) standards committee developed the Data Integrity Field specification (DIF) to provide protection within storage disks by extending the size of disk sectors by 8 bytes, which contain protection information, including a 16 bit CRC. This integrity metadata enables corrupted write and read requests to be detected and the requests aborted and retried, thereby preventing silent data corruption and providing recovery options. The T10-DIF protection domain can be contained to the disk level, or extended to the storage network controller (HBA), covering the SAN interconnect.

T10-DIX (Data Integrity eXtensions) expands the range of the T10-DIF protection to cover the remaining part of the path, which is the portion between the HBA and the application or operating system. The protection information is thus appended to data blocks exchanged on this portion of the path to ensure similar levels of data integrity end-to-end.

This paper presents the features and shows how performance and efficiency are married to a solid data protection framework that makes T5 an ideal device for storage and other demanding applications. Chelsio's T5 ASIC is a full featured converged network controller that supports both T10-DIX and T10-DIF functionality for both iSCSI and FCoE traffic (512B and 4KB data block sizes). The T5 controller benefits from a decade of experience shipping high performance storage adapters and includes numerous data integrity protection features, with multiple levels of overlapped checks across the data processing pipeline.

Terminator 5

Sitting at the boundary of server and network, a typical Network Interface Card (NIC) forms the boundary of the protection domain offered by the network MAC layer (e.g. Ethernet CRC), and most often also offloads checksum validation for the routing and transport layers (e.g. IP and TCP). In doing so, it effectively terminates the protection offered by these various checks.

Unless the NIC implements explicit mechanisms to protect the data as it transitions to the host memory, this leaves the data exposed to bit errors and other risks to its integrity. For instance, so called “soft” bit errors can lead to random bit flips in data within an ASIC. As silicon process geometries shrink, the incidence of these bit errors, caused by radiation and other internal disturbances, increases. Other sources of errors include interface noise and interference, or even programming errors.

While some may be resilient to data errors, the vast majority of applications and storage in particular, require very high level of data protection, with no tolerance to silent data corruption.

T5 leverages more than a decade of experience building offload network processors, and includes mechanisms to protect data as it transitions through the chip, or is stored in external memory. It also is the first high performance network controller to support T10-DIF/DIX for both iSCSI and FCoE, allowing SCSI block level end-to-end data protection.

Overlapped Protection

The main datapath in T5 is designed such that at any point in time, there are at least 2 overlapped levels of protection, rising in some cases to 5 overlapped layers. With T10-DIF/DIX support, the protection is extended end-to-end, uninterrupted. The following figure shows the T10-DIF field format in relation with the data sector.

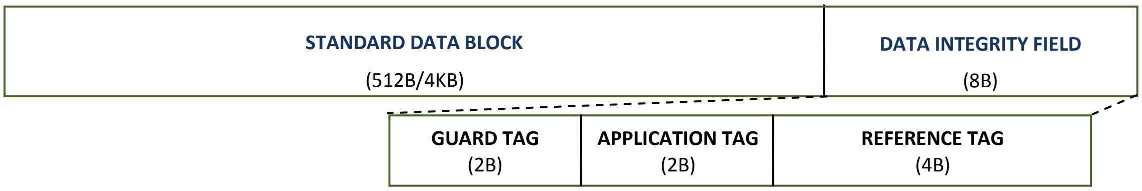

Figure 1 - T10-DIF/DIX Fields

Chelsio's T5-based adapters provide data integrity protection with T10-DIF support from target backend to initiator SCSI mid-layer. The T10-DIF capability is supported by Chelsio's T5 iSCSI target in conjunction with Chelsio's iSCSI initiator. The T5 provides data integrity protection in two modes:

a. Local mode: In this mode data integrity is performed between the T5 adapter and operating system. PI data is generated and inserted in receive, and verified then removed in transmit.
b. End-to-End mode: In this mode, data integrity is performed between Target backend and Initiator mid-layer. PI data is verified and separated from data in receive, or verified and merged with data in transmit. The steps can include replacing a checksum with CRC or vice versa.

T10-DIF READ and WRITE Operations

READ Operation

The Target backend sends 520-bytes of sector data with PI (Protection Information) to the Chelsio Target driver. The sector data and PI can be interleaved or separated. The Target driver forwards the sector data to the Target T5 adapter. The adapter verifies the PI and merges it with sector data. The 520-bytes sector data is then sent to Initiator T5 adapter. The adapter verifies the PI, separates it from sector data, and then transmits both of them to Chelsio Initiator.

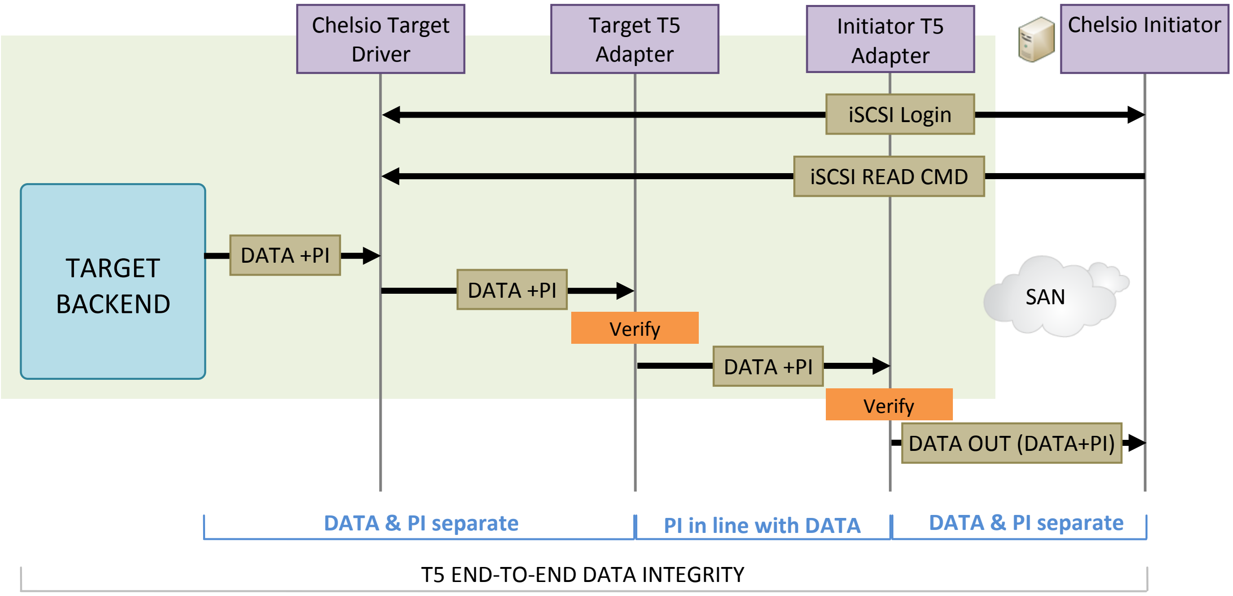

Figure 2 - T10-DIF I/O READ through T5 adapter

WRITE Operation

The Chelsio Initiator receives the separated 8-bytes PI and 512-bytes sector data from the SCSI mid-layer and forwards them to Initiator T5 adapter. The adapter verifies PI and merges it with sector data. The 520-bytes interleaved sector data is then forwarded to the Chelsio Target T5 adapter. The adapter verifies the PI, separates it from sector data, and sends both of them to the Chelsio Target driver. Finally, the target driver delivers the 512-bytes sector data and 8-bytes PI to the target backend.

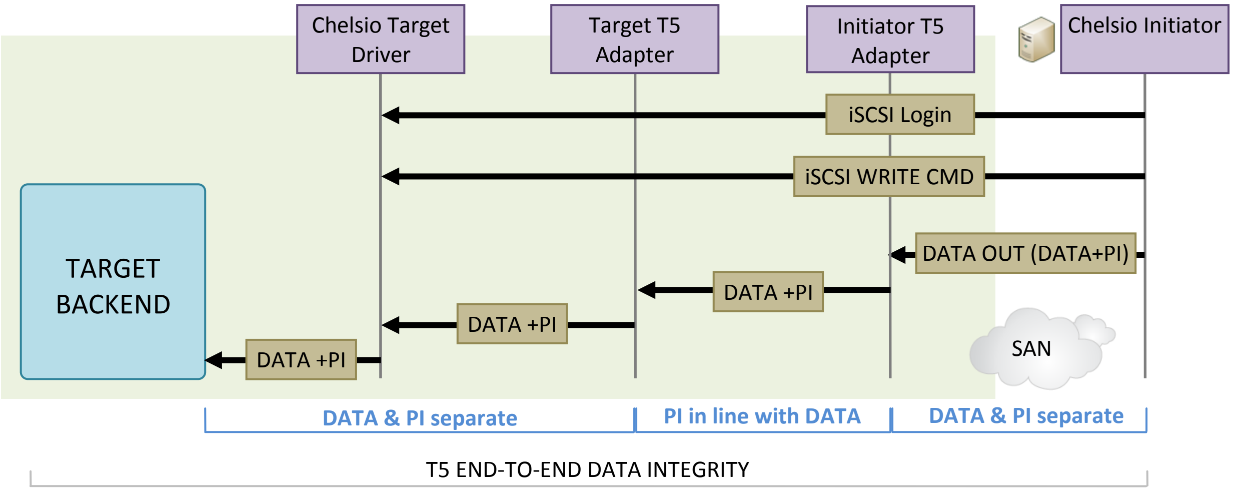

Figure 3 - T10-DIF I/O WRITE through T5 adapter

T10-DIX READ and WRITE Operations

READ Operation

The Target backend sends 520-bytes of sector data with PI to the Chelsio Target driver, which is eventually sent to the Target T5 adapter. The sector data and PI can be interleaved or separated. The Target T5 adapter verifies the PI, performs a WRITE_STRIP operation on the 520-bytes of data and discards the 8-byte of PI. The adapter then forwards the 512-bytes sector data to the iSCSI initiator over the iSCSI inter-link.

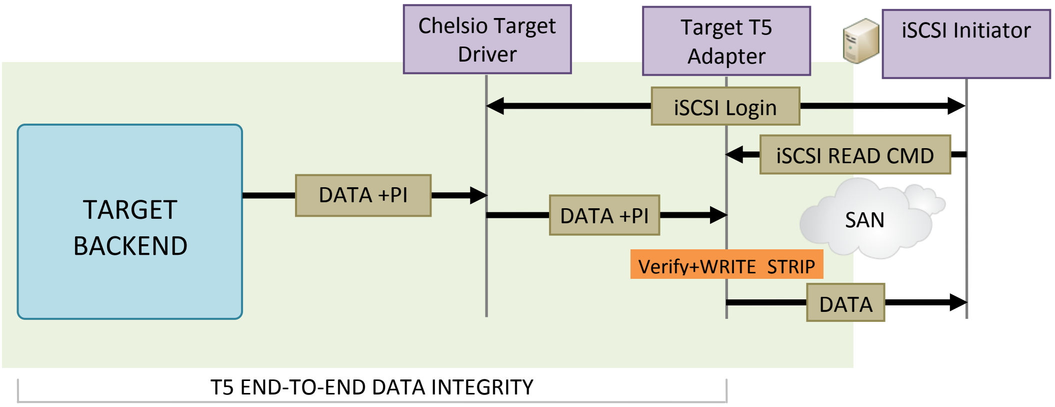

Figure 4 - T10-DIX I/O READ through T5 adapter

WRITE Operation

The iSCSI initiator sends data to the Target T5 adapter over the iSCSI inter-link in multiples of 512-byte sectors without any PI. The adapter generates 8-byte PI and performs a READ_INSERT operation on the 512-byte sector data by merging the data with newly generated PI. The 520-byte sector data interleaved with protection information is then sent to the Chelsio Target driver which delivers the 520-bytes of sector data to the Target backend. The sector data and PI can be interleaved or separated.

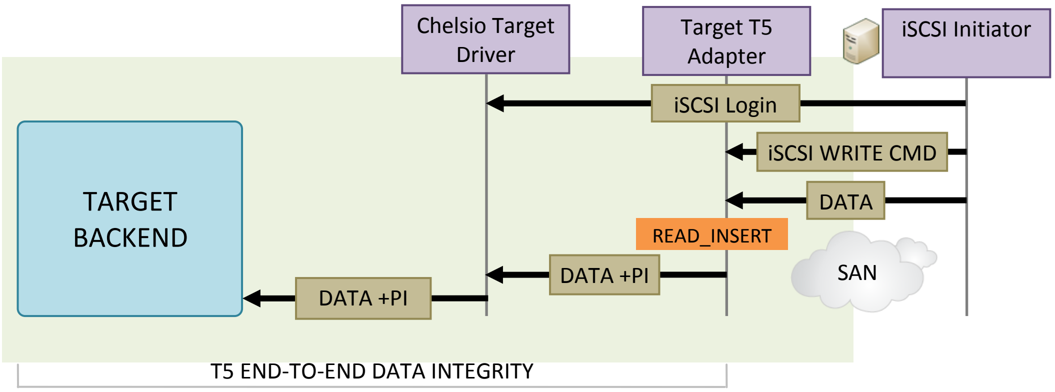

Figure 5 - T10-DIX I/O WRITE through T5 adapter

Test Results

The following graphs plot the READ and WRITE Throughput results, with and without DIF enabled, obtained by varying I/O sizes using the **fio** tool.

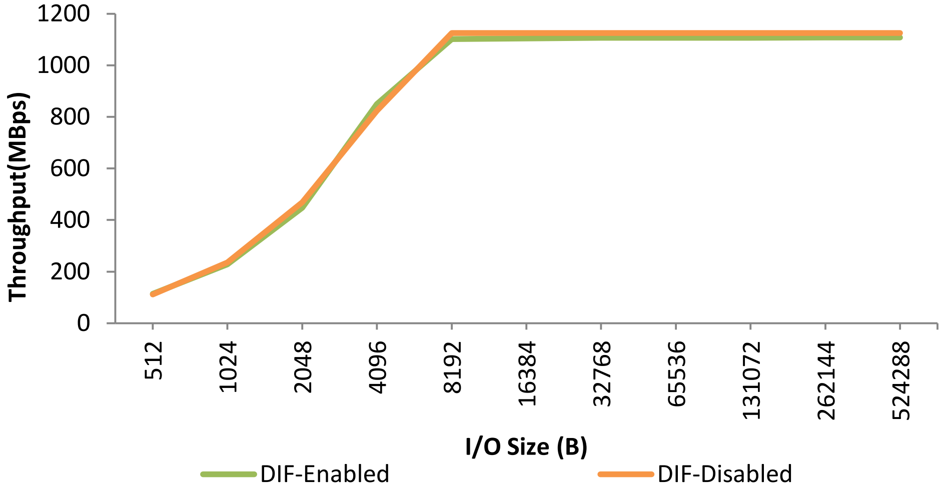

Figure 6 - T10-DIF READ Throughput vs. I/O size

As evident from the results above, Chelsio T5 achieves line rate throughput from I/O size 8 KB. The READ throughput with DIF enabled is same as that of DIF disabled. Hence, T5 can provide data integrity protection without sacrificing performance.

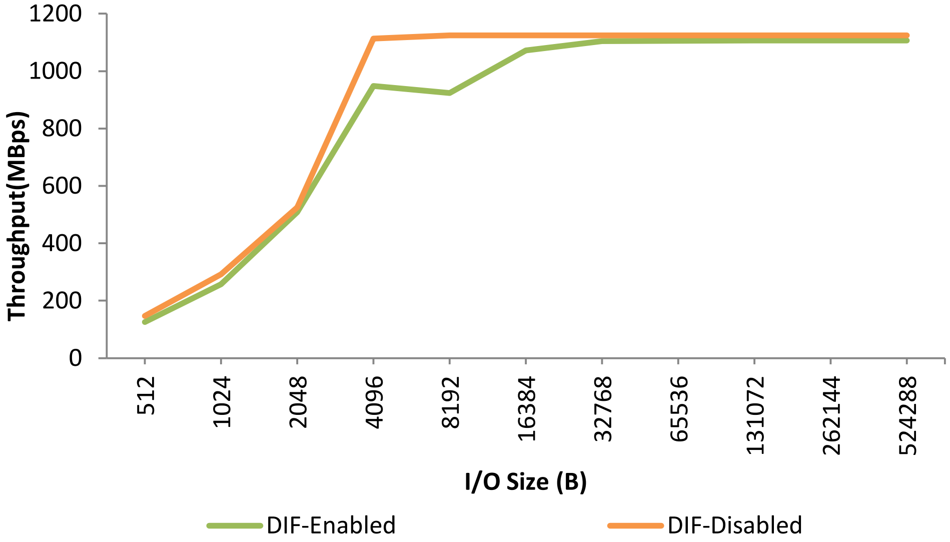

Figure 7 - T10-DIF WRITE Throughput vs. I/O size

The graph above shows that Chelsio T5's WRITE throughput reaches line rate at 16KB I/O size with DIF enabled.

Test Configuration

The following sections provide the test setup and configuration details.

Topology

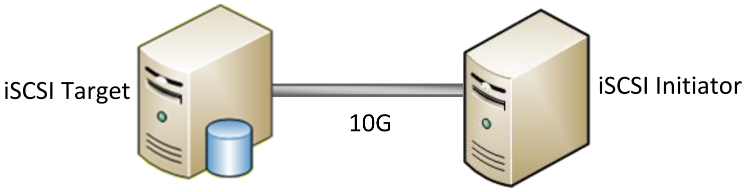

Figure 8 – Test Setup

Storage Topology and Configuration

The iSCSI setup consists of a target storage array connected back-to-back to an Initiator machine using single port. The storage array is configured with 2 Intel Xeon CPU E5-2687W v2 8-core processors running at 3.40GHz, 64GB of RAM, RHEL 6.5 (3.6.11 Kernel) operating system, T520-LL-CR adapter and Chelsio iSCSi Target driver. The storage array contains 4 iSCSI *ramdisk* targets.

The Initiator machine is configured with an Intel Xeon CPU E5-1660 v2 6-core processor clocked at 3.70GHz, 64GB of RAM, Fedora 17 operating system, T520-LL-CR adapter and Chelsio PDU offload initiator. Standard MTU of 1500B is used.

I/O Benchmarking Configuration

fio tool is used to assess the storage capacity of the configuration. The I/O sizes used varied from 512B to 512KB with an I/O access pattern of random READs and WRITEs.

Command Used

```
[root@host]# fio --name=randread --iodepth=32 --rw=randread --size=800m --
invalidate=1 --direct=1 --fsync_on_close=1 --norandommap --group_reporting --
ioengine=libaio --numjobs=1 --bs=16k --runtime=30 --time_based --
filename=/dev/sdx
```

Conclusion

This paper demonstrates how Chelsio T5 Unified Wire adapters provide end-to-end data integrity solution without any performance degradation. Chelsio T5's T10-DIF implementation effectively eliminates silent data corruption, and is part of an array of data integrity protection features that leverages 5 generations of storage focused silicon. The results show READ and WRITE throughput numbers with DIF-enabled reach line rate at 8KB and 16KB respectively.

Related Links

The Chelsio Terminator 5 ASIC
iSCSI at 40Gbps
iSCSI Heritage and Future
High Performance iSCSI for Virtual Machines