White Paper

An Integrated End-to-End Data Integrity Solution to Protect Against Silent Data Corruption

Abstract

This white paper describes how T10 PI prevents silent data corruption, ensuring that incomplete and incorrect data cannot overwrite good data. Without T10 PI, data corruption events may result in system downtime, lost revenue, or lack of compliance with regulatory standards.

August 2012

EMC²

Part Number H11058

Table of contents

Introduction

T10 Protection Information (T10 PI), previously known as Data Integrity Field (DIF), is an important standard that supports the industry's commitment to end-to-end data integrity validation. T10 PI prevents silent data corruption, ensuring that incomplete and incorrect data cannot overwrite good data. Without T10 PI, data corruption events may result in system downtime, lost revenue, or lack of compliance with regulatory standards.

As the industry leader in enterprise data protection and availability, EMC® Symmetrix® intends to be the first enterprise storage array to join with Emulex and Oracle in implementing end-to-end T10 PI.

The data protection information generated by Oracle Automatic Storage Management (ASM) is validated first by the host operating system, then by the Emulex LightPulse 8 Gb Fibre Channel Host Bus Adapter (HBA, model number LPe12000-E), and finally by EMC Symmetrix VMAX® 40k storage array with EMC Enginuity™ version 5876.82.57 or later, ensuring protection through the I/O stack.

Purpose

The purpose of this document is to provide information regarding the addition of T10 PI on EMC VMAX series products and the results of the joint testing effort of EMC, Oracle, and Emulex.

Scope

This document focuses on the initial release of the EMC, Oracle, and Emulex joint T10 PI solution.

Audience

This document is intended for those seeking a method to overcome silent data corruption and enhancing the integrity of their data stored on EMC storage.

Technology Overview

This section provides information on the three components used to achieve end-to-end data integrity, each discussed briefly in this section:

- Unbreakable Enterprise Kernel for Oracle Linux
- Emulex LightPulse 8 Gb Fibre Channel HBA
- EMC Symmetrix VMAX Family

Unbreakable Enterprise Kernel for Oracle Linux

For the implementation discussed in this White Paper, the Unbreakable Enterprise Kernel [kernel-uek-2.6.39-200.24.1.el6uek] for Oracle (also available as part of Oracle Linux 6.3 as a default kernel) is recommended.

Unbreakable Enterprise Kernel contains many new features that are relevant to Oracle Linux running in the data center, including data integrity features.

Unbreakable Enterprise Kernel, including the data integrity features, is provided under the GNU General Public License (GPL) and is available to anyone in both binary and source form. As of this writing, binary versions of the kernel are provided via Unbreakable Linux Network (ULN) and Oracle's public yum server.

Subsequent releases of Oracle Linux will include Unbreakable Enterprise Kernel as an option on the installation media, which can be downloaded for free from edelivery.oracle.com/linux. Existing Oracle Linux support customers receive full support for this kernel as part of their existing support subscriptions.

Bug fixes and security errata are delivered via ULN and announced through the el-errata mailing list.

Emulex LightPulse 8 Gb Fibre Channel HBA

The Emulex LightPulse 8 Gb FC HBAs (model numbers LPe12000-E and LPe12002-E), with the Emulex BlockGuard feature, is a key component in the Oracle's Data Integrity solution. BlockGuard ensures that data corruption events do not go undetected as data traverses the system, from the operating system and application to the disk array storing valuable data. The PCI Express 2.0 Emulex HBA includes BlockGuard, which provides T10 Protection Information (T10 PI) and Oracle Data Integrity Extensions (DIX).

As part of the overall ecosystem deployment, once the Oracle Database Application creates data in memory, the ASM generates protection information which the Oracle Linux kernel then forwards to the Emulex HBA using Data Integrity Extensions (DIX).

The Emulex HBA verifies that the data, protection information, and target location match and then interleaves the data and protection information and transmits 520-byte sectors to the storage.

At this point, the Emulex HBA has completed its job. Now, the storage array controller, followed by the disk drive firmware, verifies that the data, protection information, and target location match. If a successful I/O completion ensues, it is then reported back to the application.

Any mismatch detected by the HBA (or storage array and disk drive) causes the I/O to abort and the error is passed up the stack, preventing bad data from being written. Protection information is transmitted to read requests and the ASM verifies I/O before signaling completion to the application.

Lastly, when using legacy storage, protection information exchange is dynamically negotiated and automatically enabled between the application and HBA.

EMC Symmetrix VMAX Family

As the industry leader in enterprise data protection and availability, the EMC Symmetrix VMAX Family is the first enterprise storage array to join with Emulex and Oracle Linux in implementing end-to-end T10 PI. The data protection information generated by the Oracle ASM is validated by the Oracle Linux operating system, then passed on to the EMC Host Bus Adapter (HBA) and the EMC VMAX storage array, ensuring protection through the I/O stack.

The EMC Symmetrix VMAX Family consists of the VMAX 40K, VMAX 20K, and the VMAX 10K.

VMAX 40K

The VMAX 40K is built for Hybrid Cloud environments and provides the industry's highest levels of consolidation, performance and scalability.

VMAX 20K

The VMAX 20K is built for performance, consolidation, and automation for demanding virtual data center environments.

VMAX 10K

The VMAX 10K is the most affordable multi-controller array, and built for performance and efficiency to consolidate applications in virtual environments

Providing End-to-End Integrity

When fully implemented, end-to-end data integrity consists of components that support Data Integrity Extensions (DIX) and T10 Protection Information (T10 PI).

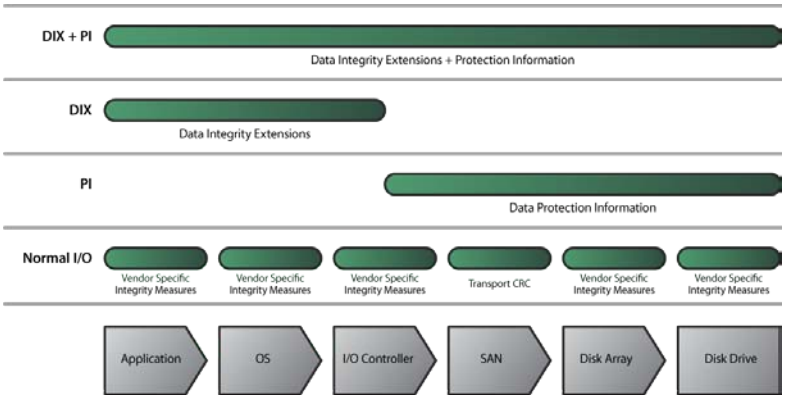

Figure 1. Achieving end-to-end data integrity

When writing date, end-to-end data integrity consists of the following steps:

1. The Oracle ASM library adds integrity metadata for each 512-byte sector as it is written to memory.
2. The integrity metadata is attached to the I/O request and passed through the layers in the operating system kernel to the Emulex driver.
3. The Emulex 8 Gb Fibre Channel adapter collects the information from memory buffers, verifies the data integrity, merges the data and integrity metadata, and sends out 520-byte sectors.
4. The EMC Symmetrix VMAX array firmware, Enginuity 5876.82.57, verifies the integrity metadata, and writes to disk.
5. The disk drive firmware verifies the integrity metadata before committing the data to physical media.

These steps are completed in reverse when reading data.

Solution Verification

Test Environment

The test environment consisted of the following, as shown in Figure 2 on page 7:

- Intel x86_64 based server with an Emulex LPe12002-E and firmware 2.01a10 installed
- Oracle Linux 6.3 with UEK kernel version 2.6.39-200.24.1.el6uek and the in kernel Emulex driver (lpfc) version 8.3.5.68.6p
- JDSU Fibre Channel analyzer

- SAN consisting of Brocade Fibre Channel switches
- VMAX 40k with Enginuity 5876.82.57

Configuration

The following figure depicts the configuration used to verify the solution.

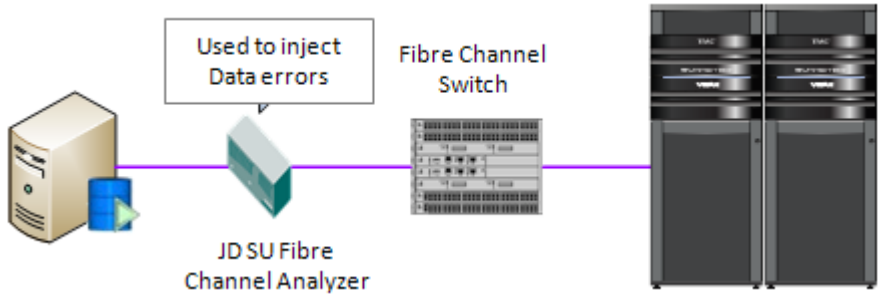

Figure 2. Testing configuration

Figure 3 explains how the stack works.

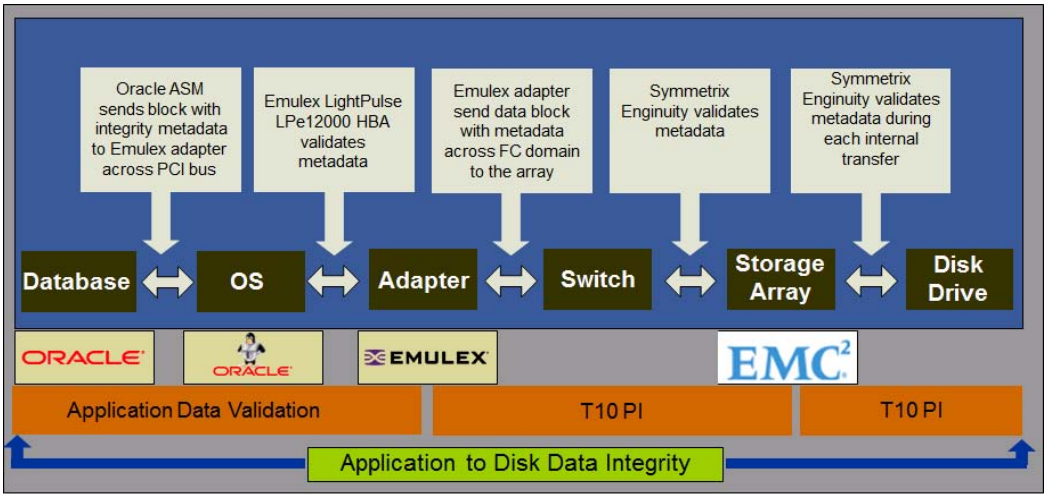

Figure 3. How the stack works

Method

EMC has tested the EMC/Oracle/Emulex T10 PI solution end-to-end in EMC's E-Lab™ and Symmetrix development labs. Testing included fault insertion through SAN events such as cable pulls, host failures, and storage failures, as well as the insertion of corrupted data between the host and target using the JDSU Fibre Channel analyzer. Additionally, corrupted data was inserted throughout the operating system stack to ensure compliancy with the specification. Target, initiator, and the operating system stack were monitored to ensure the T10 PI specification was met and that silent data corruption did not occur.

Conclusion

The stack as tested met the T10 PI specification, preventing the possible occurrence of silent data corruption. The insertion of faulty data on "the wire" was detected throughout the stack and appropriate responses were logged.

Actions were taken by the respective end-points (depending on the direction the errant data was injected) during both reads and writes to the array as well as when sent up the OS stack by injecting data errors within the operating system stack.

References

Refer to EMC Online Support website (registration required) for the following EMC documentation, at https://support.EMC.com:

- *EMC Host Connectivity Guide for Linux*
- *EMC Solutions Enabler Symmetrix Array Controls CLI V7.4 Product Guide*

Refer to the following website for Oracle Linux documentation:

- www.oracle.com

Refer to the following website for Oracle documentation:

- *Unbreakable Enterprise Kernel R.2 for Oracle Linux*
  www.oracle.com/linux

Refer to the following website for Emulex documentation:

- Emulex Product website for more information on Emulex-branded HBAs, at www.emulex.com/products/fibre-channel-hbas.html
- Emulex-EMC website for more information on EMC-branded HBAs , at www.emulex-emc.com

Preventing Silent Data Corruption

Using Emulex Host Bus Adapters, EMC VMAX and Oracle Linux

An EMC, Emulex and Oracle White Paper
September 2012

Preventing Silent Data Corruption

Introduction

Database administrators face many challenges in developing and supporting applications that are often the lifeblood of their organization. One of the concerns is the integrity of data as it travels through the storage area network (SAN) between servers and storage arrays. When data corruption is undetected, or "silent", there can be serious consequences when the database attempts to use that data. Over the years, vendors have implemented many features to ensure data integrity. Database vendors have added logical integrity checks, server memory is protected by Error Correcting Code (ECC), PCI Express buses are protected by Cyclic Redundancy Check (CRC), storage area networks are protected by CRC, and storage arrays are protected through various error detecting and correcting techniques.

Even with these checks, increasing complexity of the data center environment and growth in storage have led to significant concerns about silent data corruption. This paper provides an overview of the concepts of data integrity and silent data corruption; how silent corruption can impact an organization; and a solution from Oracle, Emulex and EMC to prevent silent data corruption.

Oracle and Emulex have been early leaders in enhancing data integrity and are continuing that effort. Oracle announced the contribution of block I/O data integrity infrastructure code to Linux and the acceptance of this code into the 2.6.27 Linux kernel. This open source code was developed by Oracle and exposes data protection information to the Linux kernel. For the first time, subsystems can utilize crucial data integrity features that extend from applications to the Linux operating system to storage. Comprehensive data integrity capabilities can now be enabled across the entire software stack. This helps reduce system downtime and provides cost savings to end users.

This code is available to all customers as part of Oracle Linux with the Unbreakable Enterprise Kernel. UEK is based on version 3.0.16 of the Linux kernel and includes optimizations to ensure stability and optimal performance for the most demanding enterprise workloads.

Potential Data Integrity Problems

Physical Data Integrity

Figure 1 illustrates the I/O stack in a typical Oracle environment. Note that one of the key components is the Oracle Automatic Storage Management (ASM) subsystem, which is Oracle's preferred storage management solution for the Oracle database environment.

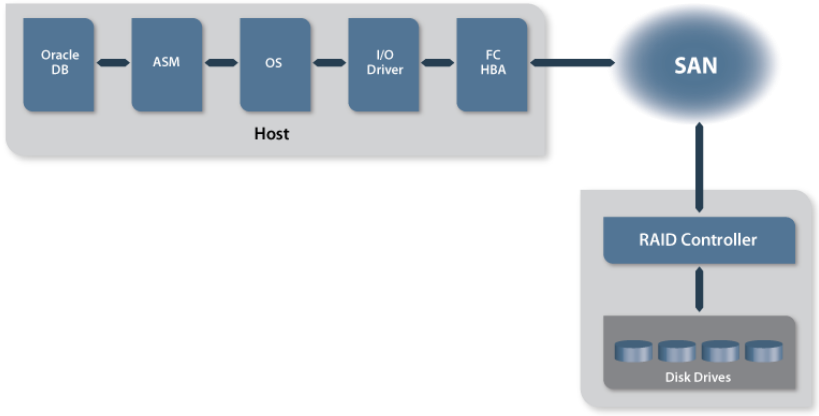

Figure 1. Software and hardware components in the I/O stack.

Physical data integrity relates to the integrity of data as it travels the I/O path between an application on a server and disk drives on a storage array.

Most devices on the I/O path, including all Emulex adapters, are designed to verify the integrity of the data as it passes through the device. However, previously there has been no mechanism for end-to-end data integrity checking from the database, to the operating system and server parts of the I/O path, the HBA, the storage array, and the disk drive to make sure the data written to the disk is the correct data.

The potential for problems in these areas has increased as data centers have moved to virtualized servers, multi-core processors and faster server buses. For example, operating systems have to deal with more complex memory mapping, which increases the potential for data to be corrupted with unusual "edge" conditions that are difficult to fully test.

Silent Data Corruption

Without an end-to-end protection technology, data corruption can go unnoticed until recovery is difficult and costly, or even impossible to perform. Without end-to-end integrity checking, these silent data corruptions can lead to unexpected and unexplained problems.

A PC Magazine article (published August 25, 2008) reported a real-life data corruption incident:

"Netflix monitors flagged a database corruption problem in its shipping system. Over the course of the day, we began experiencing similar problems in peripheral databases until our shopping system went down." The root cause was determined to be a faulty hardware component, but the problem was that the component "reported no detectable errors."

One of the areas where data corruption can occur is writing to disk drives. There are two basic kinds of disk drive corruption. The first is "latent sector errors", which are typically the result of a physical disk drive malfunction. An example would be a file system read error reported from a disk array. This type of corruption is usually detected by ECC or CRC in the I/O path and most often is corrected automatically.

The other type is silent corruption, which can happen without warning. There is no effective means of detection without end-to-end integrity checking.

A study[i] conducted by the University of Wisconsin, University of Toronto, and NetApp focused on silent data errors that occurred with disk drives. The study was conducted over a 41-month period and analyzed checksum errors on 1.53 million disk drives.

The study used file system-level disk block identity information to detect previously undetectable checksum errors. During the 41-month period, silent data corruption was observed on 0.86% of 358,000 nearline SATA drivers and .065% of 1.17 million enterprise-class Fibre Channel drives. While these percentages are low, they represent undetected or completely silent errors that could lead to lost or inaccurate data and significant downtime.

The software stack has also become more complex, making it more vulnerable to data corruption. Examples include bad buffer pointers and missing or misdirected writes. If data is corrupted in the software stack, the file system has limited means for detection and self-correction, resulting in the potential for silent data corruptions.

The Cost of Silent Data Corruption

It is difficult to put a dollar figure on the cost of data corruption because it depends on the industry, the type of application, and the circumstance.

Figure 2 (next page) shows the results of a Gartner Group study demonstrating the hourly cost of downtime by industry. The cost of downtime not only includes the administrative cost to bring a company back online, but also the productivity loss by those impacted and the business cost associated with system unavailability.

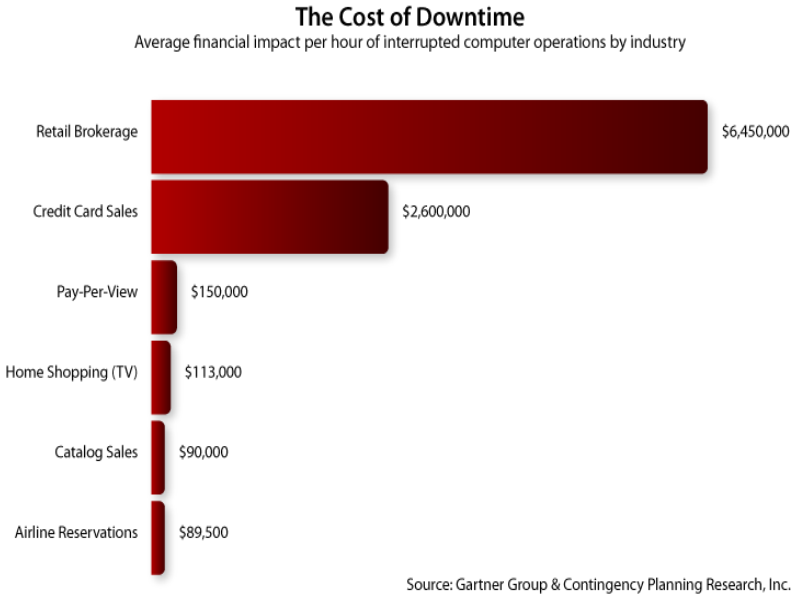

Figure 2. The hourly cost of downtime can be directly related to silent data corruption.

One conservative measure of data corruption cost is downtime, which does not take into the account the impact on business functions that rely on the data. If data corruption occurs, database and storage administrators must spend time recovering the database.

The typical steps are:

1. Reload the database or the affected tablespace from the last complete backup.
2. Replay the log entries until the database or the tablespace is fully recovered. Depending on the time interval from the backup and the number of log entry replays, this step can take minute, hours, or days.

If the backup, archive logs or online redo logs are corrupted, database recovery can be painful and may be impossible.

If the corruption happened long ago, the chances are high that recent backups would also contain the corrupted data. Unless a good backup can be found, there will be a loss of valuable data.

To prevent downtime caused by data corruptions, some businesses use duplicate copies of the same database. For example, a large data center manager uses up to six duplicate copies of their database just for quick recovery from unexplained database downtime. This increases their data center operating and hardware costs significantly. Note that these duplicate standby databases cannot be storage-based clones because a primary database corruption could be propagated to the standbys. Therefore, any standby database must be a logical copy.

Silent Data Corruption Impact on Business Continuity

Businesses use remote replication to ensure mission critical Oracle databases run with minimum interruption if disaster recovery becomes necessary. Storage-based remote replication is commonly used because it is simple to set up and relatively easy to administer. However, if silent data corruption occurs, the same corruption can be replicated to the remote site. Therefore, it is very important for businesses to make sure that their replication configuration is free of silent data corruption.

Preventing Silent Data Corruption

Oracle recognized the need to prevent silent data corruption and insure data integrity. Oracle has taken the lead in defining and developing technologies to meet this objective. As shown in Figure 3, Emulex and Oracle have been working together to bring products to the market that support these technologies.

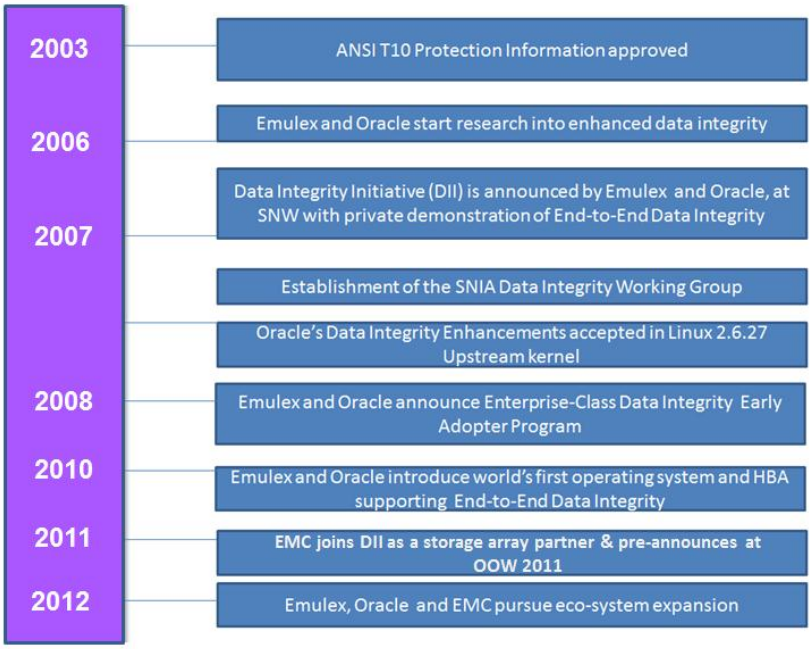

Figure 3. Emulex, EMC and Oracle working together to advance data integrity.

Data integrity enhancements directly address a problem like the Netflix example by monitoring and identifying data corruption events that can be caused by any hardware or software component in the data path. The key benefit is that data corruption events that were previously undetected are identified and flagged.

There are two key technologies involved:

- T10 Protection Information
- Data Integrity Extensions

Both are based on appending extra information to data that can be used to verify its integrity and prevent silent data corruption. This extra information is referred to as *integrity metadata*.

T10 Protection Information Standard

The Protection Information model (PI) is an extension of the SCSI Block Commands specification that was approved by the T10 technical committee. The PI model applies to communication between SCSI controllers and storage devices. For the purposes of this white paper, the SCSI controller would be an Emulex 8Gb/s or Emulex 16Gb/s HBA.

Data that resides on a hard disk is typically divided into 512-byte blocks or sectors. The T10 PI model defines the contents of an additional 8 bytes of information, increasing the sector size to 520 bytes. The additional bytes are used to store tags that can be used to verify the 512 bytes of data in the sector.

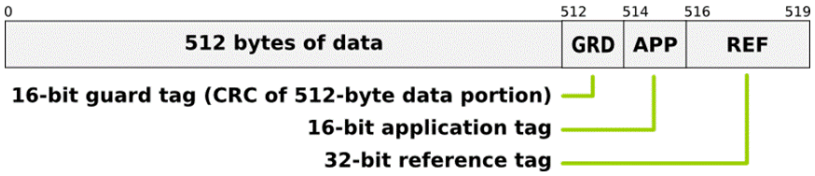

Figure 4. 520-byte sector containing 512 bytes of data followed by 8-byte PI tuple.

For data writes, Emulex 8Gb/s and 16Gb/s HBAs have the capability to generate the 8 bytes of integrity metadata that would be appended to the 512-byte sector received from the host operating system. A PI- capable storage array would then use the metadata to verify the integrity of the data before accepting it.

For reads, Emulex 8Gb/s and 16Gb/s HBAs have the capability to verify the integrity metadata written by a PI-capable storage array.

Data Integrity Extensions

Oracle has taken the lead with Emulex to define Data Integrity Extensions (DIX), which is a set of requirements for controllers to exchange metadata with a host operating system. DIX enables data integrity between an application and the controller, and the combination of T10 PI and DIX provides true end-to-end data integrity.

Although an application typically sees data as a contiguous buffer, the buffer is likely to be scattered in several areas of physical memory. Storage adapters use a scatter-gather list to know how to assemble the data buffer to be transferred.

To improve the detection of corrupted data, DIX is implemented with separate scatter-gather lists for the data buffer and the integrity metadata. This protects against some forms of stale data, incorrect pointers, and other possible operating system errors. If data integrity check fails, the data error is flagged and the data is not transmitted.

Oracle has defined an API for an integrity-capable block I/O layer for Linux and submitted it to kernel.org, along with changes that enable support for both PI and DIX. This integrity infrastructure was accepted into Linux 2.6.27.

To reduce the overhead on the host CPU, Oracle and Emulex are using a Transmission Control Protocol (TCP) checksum, which

requires less CPU overhead than CRC. Note that with DIX, the HBA does not generate the integrity protection data—that is now the responsibility of Oracle Automatic Storage Management.

End-to-End Data Integrity

When fully implemented, end-to-end data integrity will consist of components that support DIX and T10 PI.

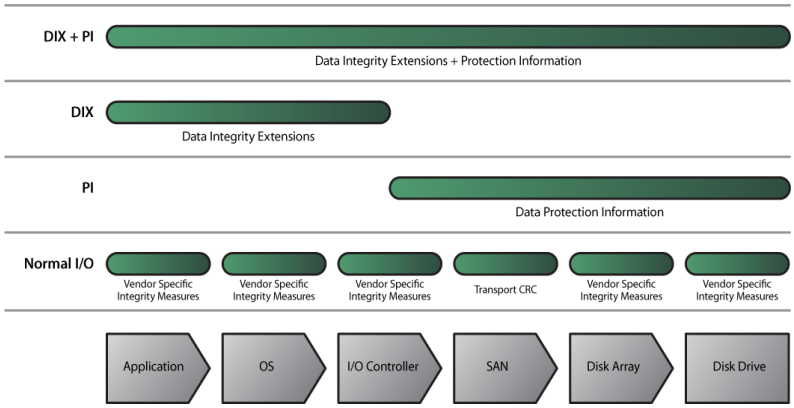

Figure 5. DIX and PI provide end-to-end data integrity.

End-to-end data integrity will consist of the following steps when writing data:

1. The Oracle ASM library adds integrity metadata for each 512-byte sector as it is written to memory.
2. The integrity metadata is attached to the I/O request and passed through the layers in the operating system kernel to the Emulex driver.
3. The Emulex adapter collects the information from memory buffers, verifies the data integrity, merges the data and integrity metadata, and sends out 520-byte sectors.
4. The EMC array firmware verifies the integrity metadata, and writes to disk.
5. The disk drive firmware verifies the integrity metadata before committing the data to physical media.

The steps will be done in reverse when reading data.

What's Available Today

Customers interested in experiencing the Oracle OS implementation of the T10 Protection Information Model standard for an operating system now have access to it through the Oracle Linux with the Unbreakable Enterprise Kernel when used in combination with Emulex HBAs and EMC Symmetrix VMAX storage.

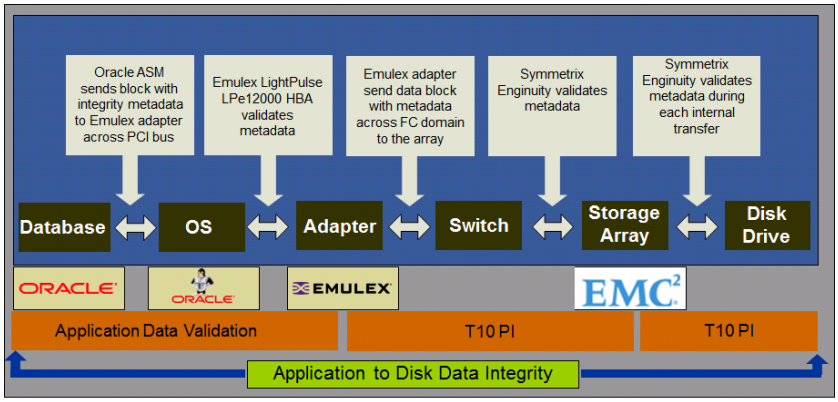

Figure 6. Current support for enhanced data integrity.

Unbreakable Enterprise Kernel

In addition to performance improvements for large systems, the Unbreakable Enterprise Kernel contains many new features that are relevant to Linux running in the data center, including support for T10 PI and the Data Integrity Extensions.

Oracle Linux with the Unbreakable Enterprise Kernel, including the data integrity features, is provided under the GNU General Public License (GPL) and is available to anyone in both binary and source form. Oracle Linux with the default Unbreakable Enterprise Kernel can be downloaded for free from edelivery.oracle.com/linux. Existing Oracle Linux support customers will receive full support for this kernel as part of their existing support subscriptions.

Bug fixes and security errata are delivered via ULN and announced via the el-errata mailing list.

Emulex LightPulse Fibre Channel HBAs

A key component in the T10 PI system is of course the I/O adapter.

Emulex has helped drive the evolution of the T10 PI functionality and promote the benefits this technology will bring. Emulex has worked closely with Oracle over the past couple of years to evolve the T10 PI functionality – both in the OS kernel support and in the supporting Emulex adapter software. T10 PI and DIX support is provided in the Emulex lpfc device driver included in the Oracle Unbreakable Enterprise Kernel.

Emulex is the proven FibreChannel HBA leader, shipping its 10th generation of FibreChannel HBA with the LPe16000B series HBA with full data integrity offload capability. EMC Qualified 8GFC

LPe1200x-E and 16GFC LPe1600x-E cards are also readily available.

EMC Symmetrix VMAX

As the industry leader in enterprise data protection and availability, the EMC Symmetrix VMAX Family is the first enterprise storage array to join with Emulex and Oracle Linux in implementing end-to-end T10 PI. The data protection information generated by the Oracle ASM is validated by the Oracle Linux operating system, then passed on to the EMC Host Bus Adapter (HBA) and the EMC VMAX storage array, ensuring protection through the I/O stack.
The EMC Symmetrix VMAX Family consists of the VMAX 40K, VMAX 20K, and the VMAX 10K.

VMAX 40K

The VMAX 40K is built for Hybrid Cloud environments and provides the industry's highest levels of consolidation, performance and scalability.

VMAX 20K

The VMAX 20K is built for performance, consolidation, and automation for demanding virtual data center environments.

VMAX 10K

The VMAX 10K is the most affordable multi-controller array, and built for performance and efficiency to consolidate applications in virtual environments

---

[i] L. Bairavasundaram, G. Goodson, B. Schroeder, A. Arpaci-Dusseau, R. Arpaci-Dusseau, "**An Analysis of Data Corruption in the Storage Stack** ", FAST08