An Oracle and Emulex White Paper
December 2010

Preventing Silent Data Corruption Using Emulex Host Bus Adapters and Oracle Linux

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 and Emulex to prevent silent data corruption.

Oracle and Emulex have been early leaders in enhancing data integrity are continuing that effort. In 2007, Emulex Oracle, LSI and Seagate announced the Data Integrity Initiative (DII). DII was established with the goal of developing an end-to-end data integrity solution. DII is still actively pursuing this goal.

In addition to participation in DII, 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 in conjunction with Emulex and exposes key 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 are now enabled across the entire software stack. This helps reduce system downtime and provides cost savings to end users.

This code is now available to all customers as part of Oracle Linux with the new Unbreakable Enterprise Kernel announced in September 2010. Unbreakable Enterprise Kernel is based on a stable 2.6.32 Linux kernel and includes optimizations developed in collaboration with Oracle's Database, Middleware and Hardware engineering teams to ensure stability and optimal performance for the most demanding enterprise workloads. This kernel also includes the data integrity features.

This program provides data integrity checking between Oracle Database applications and Emulex 8Gb/s LightPulse® Host Bus Adapters (HBAs). Future products will add data integrity checking between adapters and disk drives on a storage array, enabling full end-to-end data integrity.

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.

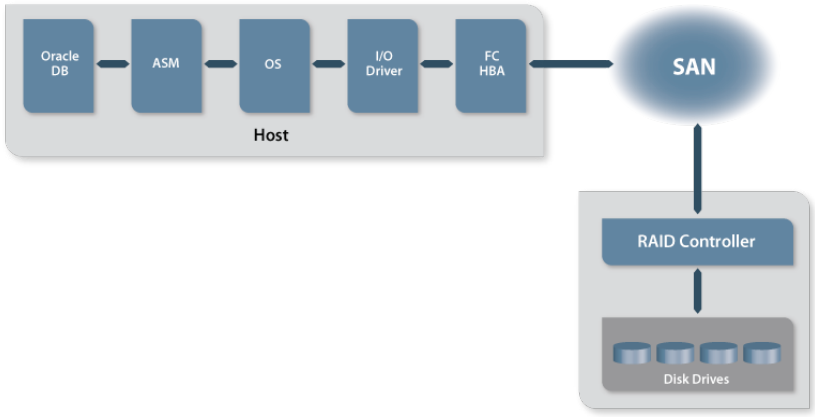

Figure 1. Software and hardware components in the I/O stack.

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 recent 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 recent study 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.

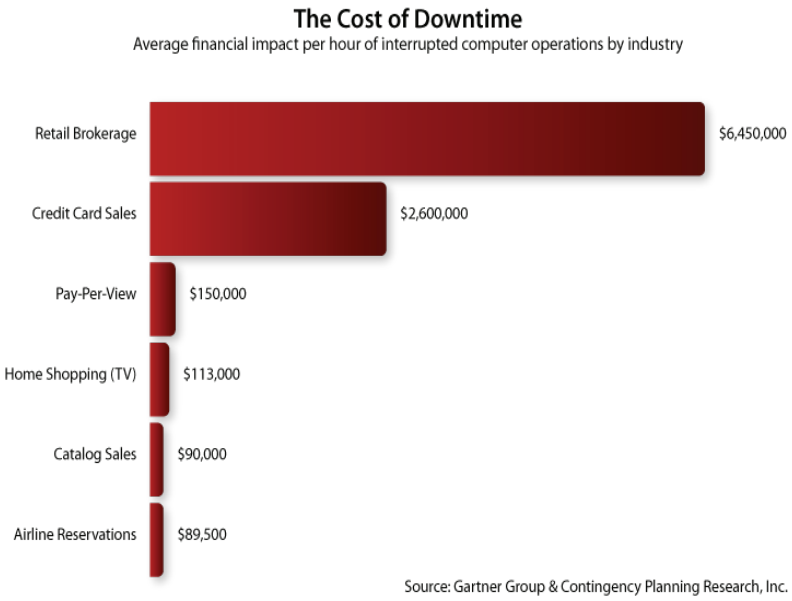

Figure 2. The hourly cost of downtime can be directly related to silent data corruption.

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.

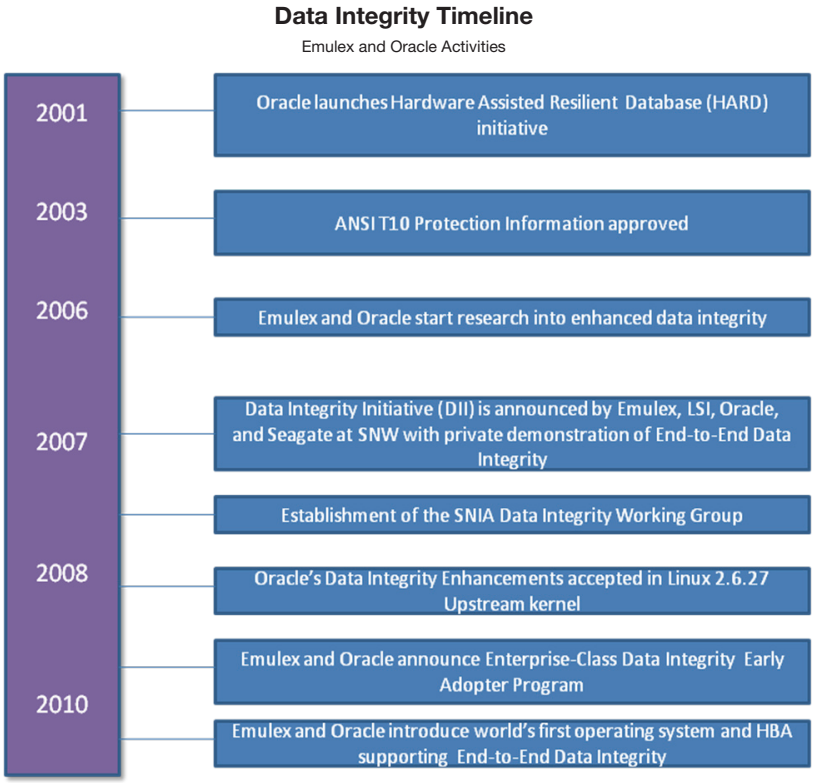

Figure 3. Emulex and Oracle working together to advance data integrity.

Hardware Assisted Resilient Data

In November 2001, Oracle launched the Hardware Assisted Resilient Data (HARD) initiative. HARD was the first industry initiative aimed at detecting and preventing silent data corruption and many leading storage vendors are members of this initiative.

HARD extends Oracle's database data validation algorithms into the storage stack to proactively prevent writes of physically and logically corrupt blocks. Storage vendors who are part of the HARD initiative can use the Oracle validation algorithms to validate Oracle blocks while performing write operations in their respective products.

The HARD solution was first made available with Oracle Database 9i.

Data Integrity Initiative

Recognizing the need to develop a complete data integrity solution for databases and other applications, Emulex, LSI, Oracle, and Seagate formed the Data Integrity Initiative (DII). Based on this initiative, the SNIA Data Integrity Working Group (DITWG) was established. Additional vendors are actively participating in the SNIA working group.

DII directly addresses 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 related to the DII initiative:

- 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.

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 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.

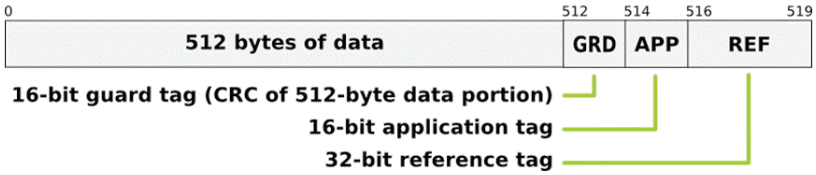

Figure 4. 520-byte sector containing 512 bytes of data followed by 8-byte PI tuple.

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

Storage vendors are currently not shipping PI-capable arrays, so the initial support from Oracle/Emulex will not support PI. However, it is expected that PI-capable arrays will be introduced in the next year, which will be compatible with Emulex adapters that are available now. Emulex and Oracle will be adding PI support in future releases that implement full end-to-end data integrity protection.

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.

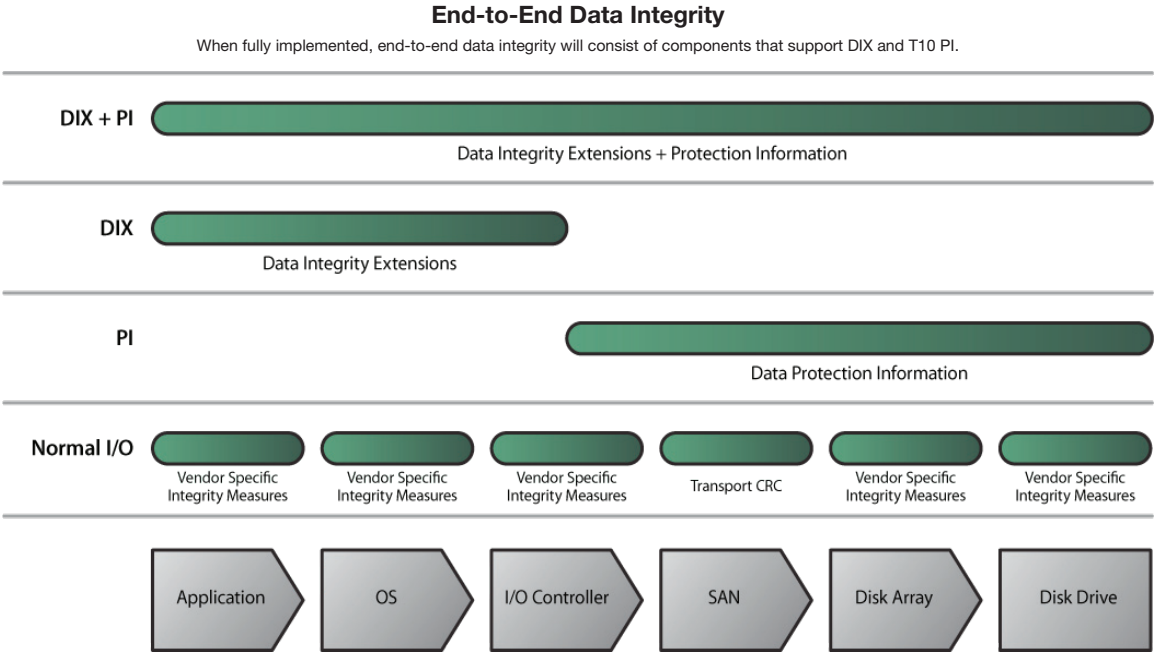

Figure 5. DIX and PI provide end-to-end data integrity.

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 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 distribution as well as Emulex HBAs.

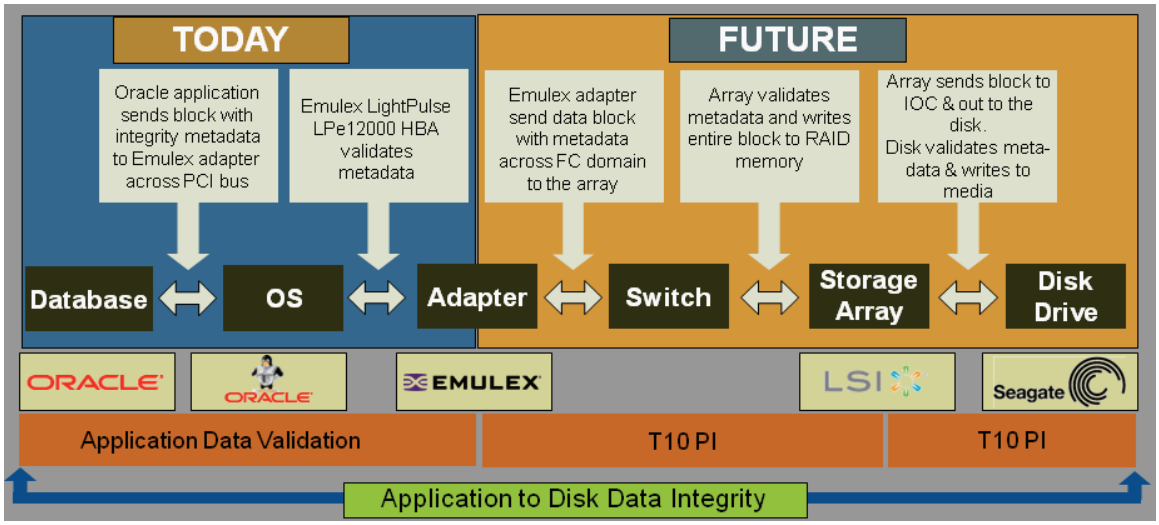

Figure 6. Current and future support for enhanced data integrity.

In September 2010, Oracle introduced the Unbreakable Enterprise Kernel for Linux as a recommended kernel to deploy with Oracle Linux 5 or Red Hat Enterprise Linux 5. In addition to performance improvements for large systems, Unbreakable Enterprise Kernel contains many new features that are relevant to Linux running in the data center, including support for T10 PI and data integrity.

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 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.

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

As an active member, and current chair, of the SNIA T10 PI technical workgroup 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. Initial T10 PI support is provided in the Emulex lpfc device driver included in the Oracle Unbreakable Enterprise Kernel and available from Emulex for early customer deployments. Emulex will work with these early customers to test this functionality in their environments.

---

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

World Headquarters 3333 Susan Street, Costa Mesa, CA 92626 +1 714 662 5600
Wokingham, UK +44 (0) 118 977 2929 | **Munich, Germany** +49 (0) 89 97007 177
Paris, France +33 (0) 158 580 022 | **Beijing, China** +86 10 68499547
Tokyo, Japan +81 3 5325 3261 | **Bangalore, India** +91 80 40156789

Connect with Emulex

twitter.com/emulex friendfeed.com/emulex bit.ly/emulexlinks bit.ly/emulexfb

www.emulex.com

11-0544 · 11/10