End-to-end Data integrity Protection in Storage Systems
Technical Whitepaper

Issue V1.1
Date 2013-11-20
Change History

Changes between document issues are cumulative. The latest document issue contains all the changes made in earlier issues.

Issue 02 (2013-11-20)

This issue is the second official release, which incorporates the following changes:

Chapter 4 Verification

The OS version in Table 4-1 Test Environment is modified as below.

The kernel was changed from “Oracle Linux OLE6U3 2.6.39-400.203.4.el6uek.di.x86_64” to “Oracle Linux OLE6U3 2.6.39-400.207.0.el6uek.di.x86_64”.

The SCSI driver was changed from “8.3.7.10.7p” to “8.3.7.26.3p”.

ALL

The name of Huawei Storage Product was changed from “Huawei OceanStor Enterprise Storage System” to “Huawei OceanStor 18000 Series Enterprise Storage System”.

Issue 01 (2013-09-23)

This issue is the first official release.
## Contents

1 Abstract ........................................................................................................................................ 1
2 Introduction .................................................................................................................................. 2
   2.1 Silent Data Corruption ........................................................................................................... 2
   2.2 Technologies to Prevent Silent Data Corruption .............................................................. 2
3 Solution ....................................................................................................................................... 4
   3.1 Introduction for Components ............................................................................................... 4
   3.2 End-to-End data integrity ..................................................................................................... 5
4 Verification .................................................................................................................................... 7
   4.1 Environment and Configuration .......................................................................................... 7
   4.2 Testing Methods .................................................................................................................... 8
   4.3 Conclusion ............................................................................................................................ 8
5 Glossary ....................................................................................................................................... 9
6 References .................................................................................................................................... 10
Abstract

This whitepaper describes an end-to-end data integrity protection solution from the application to the storage array using DIX and T10 Protection Information (T10 PI), which prevents silent data corruption in complex storage systems. A validation of this solution based on Oracle database, Oracle Linux with Unbreakable Enterprise Kernel, Emulex LightPulse® Fibre Channel Host Bus Adapters (HBAs), and Huawei OceanStor 18000 Series Enterprise Storage System is also provided in this whitepaper.

Keywords: silent data corruption data integrity end-to-end DIX T10 PI
2.1 Silent Data Corruption

Data reading, writing, and transmission from applications to storage systems, through a number of components, physical channels and complex software processing, may result in data errors if blocks of data become corrupted. This is known as silent data corruption, when the data errors are not detected when they occur but are discovered later when applications attempt to access the stored data. Since the error is not found when it occurs which is the best time to repair the data, it can lead to critical data errors, system downtime or any other unexpected serious consequences.

There are a variety of areas where silent data corruption occurs, including:

- **Hardware:** Memory, CPU, disk, transmission channels, etc.
- **Firmware:** HBA, RAID controller, disk, etc.
- **Software:** OS, library, device driver, application software, etc.

The European Organization for Nuclear Research (CERN) tested 8.7 TB of data for five consecutive weeks. The result shows that a total of 22 silent data corruption events occurred, that is, one silent data corruption event per 1500 files on average. According to other statistics from the University of Wisconsin, University of Toronto, and NetApp, 3078 SATA disks and 760 Fibre Channel disks out of 1.53 million SATA disks and Fibre Channel disks encountered silent data corruption in data access during a service operation test that lasted for 41 consecutive months.

2.2 Technologies to Prevent Silent Data Corruption

**T10 Protection Information (T10 PI)**

ANSI T10 defines a method to protect data integrity by adding Protection Information (PI) that also known as Data Integrity Field (DIF) to each block data. In the T10 PI standard, each logical block is extended to add an 8-bytes protection information, which includes a 2-byte Logical Block Guard, a 2-byte Logical Block Application Tag and a 4-byte Logical Block Reference Tag.
Figure 2-1 User data and protection information format in T10 PI Standard

<table>
<thead>
<tr>
<th>Byte</th>
<th>Bit</th>
<th>7</th>
<th>6</th>
<th>5</th>
<th>4</th>
<th>3</th>
<th>2</th>
<th>1</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>n - 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>n</td>
<td>(MSB)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>n + 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(LSB)</td>
</tr>
<tr>
<td>n + 2</td>
<td>(MSB)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>n + 3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(LSB)</td>
</tr>
<tr>
<td>n + 4</td>
<td>(MSB)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(LSB)</td>
</tr>
</tbody>
</table>

Logical Block Guard is 16-bit CRC that validates the data, Logical Block Application Tag is the information that is assigned by the application and Logical Block Reference Tag is the address information that validates the locale of the data.

Data Integrity Extension (DIX)

T10 PI only contains data integrity protection from the host bus adapter (HBA) through the storage array to disk. Oracle and Emulex jointly developed the Data Integrity Extension (DIX), which extends data integrity from the application to HBA. DIX also uses the same 8-bytes for each block as protection information in the same format as T10 PI. The difference is that DIX uses IP checksum for the Logical Block Guard to reduce the overhead on the host CPU.
3 Solution

3.1 Introduction for Components

This section provides information about the components as follows, which are used to achieve end-to-end data integrity.

- Oracle Linux with Unbreakable Enterprise Kernel
- Emulex LightPulse Fibre Channel HBA
- Huawei OceanStor 18000 Series Enterprise Storage System

Oracle Linux with Unbreakable Enterprise Kernel

Oracle Linux with Unbreakable Enterprise Kernel, which 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. Oracle’s Unbreakable Enterprise Kernel also supports the T10 Protection Information Model and the Data Integrity Extensions, providing the complete data integrity checking from application to disk is part of Oracle Linux and fully supported by Oracle.

This whitepaper recommends kernel-uek- 2.6.39-400.207.0.el6uek and the later versions.

Emulex LightPulse Fibre Channel HBA

Emulex LightPulse HBAs provides end-to-end data integrity for protection against silent data corruption events. The BlockGuard™ Data Integrity feature is available with Emulex LightPulse 8Gb Fibre Channel (FC) HBAs- models LPe12000, LPe12002 and LPe12004, and LightPulse Gen 5 (16GFC) LPe16000 and LPe16002. In addition the new LPe16000B and LPe16002B Gen 5 (16GFC) PCIe 3.0 HBAs, feature T10 PI with high-performance offload, improving performance through execution of checks in hardware, to free-up the CPU for other processing.

The model Lpe12002 of 8Gb FC HBAs are used to verify the solution in this whitepaper.

Huawei OceanStor 18000 Series Enterprise Storage System

Huawei OceanStor 18000 Series Enterprise Storage System is Huawei’s high-end flagship product series. This series consists of several models, OceanStor 18500, OceanStor 18800 and OceanStor 18800F, which provide trusted and flexible storage solution for critical appliances of enterprise data centers.
Huawei OceanStor 18000 Series Enterprise Storage System support T10 PI standard data integrity. In the end-to-end data integrity solution, the data protection information generated by the Oracle ASMLib, crossing the operating system, passes through the Emulex HBA to the Huawei OceanStor 18000 Series Enterprise Storage System, ensuring the protection through the I/O path. The storage system verifies data integrity based on T10 PI and rectifies data integrity errors as soon as the errors are identified in data access.

3.2 End-to-End data integrity

Overview

DIX and T10 PI together provide end-to-end data integrity protection from the application, across operating system kernel, HBA and SAN, to the disks, as shown in Figure 3-1.

Figure 3-1 end-to-end data integrity protection solution

Process and Method

Figure 3-2 describes the end-to-end data integrity protection flow.

Figure 3-2 End-to-end data integrity protection flow
The above process contains the whole end-to-end data integrity protection flow from host application through operating system, HBA, SAN, and Storage Array to disks. Throughout the process, from the application to the storage system, potential errors such as application software errors, driver errors, memory errors, interface card errors, transmission link errors and disk errors that may cause data corruption have been fully detected and repaired.
4 Verification

4.1 Environment and Configuration

Test Environment

Table 4-1 Test Environment

<table>
<thead>
<tr>
<th>Host Server</th>
<th>Huawei Tecal RH2285</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(Intel x86_64 based server)</td>
</tr>
<tr>
<td>Database</td>
<td>Oracle 11gR2 11.2.0.3</td>
</tr>
<tr>
<td>OS</td>
<td>Kernel: Oracle Linux OLE6U3 2.6.39-400.207.0.el6uek ди.х86_64</td>
</tr>
<tr>
<td></td>
<td>SCSI driver: 8.3.7.26.3p</td>
</tr>
<tr>
<td>HBA</td>
<td>Model: Emulex LPe12002-M8</td>
</tr>
<tr>
<td></td>
<td>FW Version: 2.01A11 (U3D2.01A11)</td>
</tr>
<tr>
<td>Storage Array</td>
<td>Huawei OceanStor 18500 V100R001</td>
</tr>
<tr>
<td>Tools</td>
<td>ASMIO (for running IO and error injection)</td>
</tr>
<tr>
<td></td>
<td>Finisar Fibre Channel analyzer (for jammer injection)</td>
</tr>
<tr>
<td>Multipathing Software</td>
<td>Huawei UltraPath V100R006</td>
</tr>
<tr>
<td></td>
<td>Linux DM-Multipath</td>
</tr>
</tbody>
</table>
Network Configuration

Figure 4-1 Network Diagram

4.2 Testing Methods

The Huawei/Oracle/Emulex DIX+T10 PI end-to-end data integrity protection solution has been validated in Huawei’s Storage Reliability and Interoperability labs. Testing covered I/O workload for continuous running, and fault insertion including ASMIO injection tests for data corrupted in the ASMLib, HBA and target, and jammer tests for data corrupted between the host and target using Finisar Fibre Channel analyzer.

Multipathing configurations included in the tests are as follows:

- Without multipathing software
- Huawei UltraPath
- Linux DM-Multipath

Applications included in the tests are as follows:

- ASMIO
- RDBMS

4.3 Conclusion

All of the tests mentioned in this whitepaper passed in the strict verification of Huawei Reliability and Interoperability Labs.

- All workload tests have run continuously over 40 hours.
- In jammer tests, the read injection errors were detected by the HBA and the write injection errors were detected by the storage array. The corresponding SCSI errors were logged in messages.
- In ASMIO injection tests, library errors were detected by the application, HBA errors were detected by the HBA and Target errors were detected by the storage array. The corresponding SCSI errors were logged in messages and the failed I/Os were retried for recovery by the application.

Huawei Reliability and Compatibility Labs have approved this solution, certifying that it complies with Data Integrity Extensions (DIX) and T10 Protection Information (T10 PI), the strictest E2E data integrity standards.
## 5 Glossary

### Abbreviations List

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Full Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>ASM</td>
<td>Automatic Storage Management</td>
</tr>
<tr>
<td>ASMLib</td>
<td>ASM Library</td>
</tr>
<tr>
<td>CERN</td>
<td>[Fr]Conseil Européen pour la Recherche Nucléaire = European laboratory for particle physics</td>
</tr>
<tr>
<td>CPU</td>
<td>Central Processing Unit</td>
</tr>
<tr>
<td>CRC</td>
<td>Cyclic Redundancy Check</td>
</tr>
<tr>
<td>DIF</td>
<td>Data Integrity Field</td>
</tr>
<tr>
<td>DIX</td>
<td>Data Integrity Extension</td>
</tr>
<tr>
<td>DM</td>
<td>Device Mapper</td>
</tr>
<tr>
<td>FC</td>
<td>Fibre Channel</td>
</tr>
<tr>
<td>HDD</td>
<td>Hardware Disk Drive</td>
</tr>
<tr>
<td>I/O</td>
<td>Input/Output</td>
</tr>
<tr>
<td>OS</td>
<td>Operating System</td>
</tr>
<tr>
<td>PCIe</td>
<td>Peripheral Component Interconnect Express</td>
</tr>
<tr>
<td>PI</td>
<td>Protection Information</td>
</tr>
<tr>
<td>RAID</td>
<td>Redundant Arrays of Inexpensive Disks</td>
</tr>
<tr>
<td>RDBMS</td>
<td>Relational Database Management System</td>
</tr>
<tr>
<td>SAN</td>
<td>Storage Area Network</td>
</tr>
<tr>
<td>SCSI</td>
<td>Small Computer System Interface</td>
</tr>
</tbody>
</table>
6 References

Publications


Web sites

Refer to the following website for Huawei Storage Product:

Visit the following website for more information about Huawei enterprises service:
http://enterprise.huawei.com

Refer to the following website for Oracle Linux:
http://www.oracle.com/linux

Refer to the following website for Emulex HBAs: