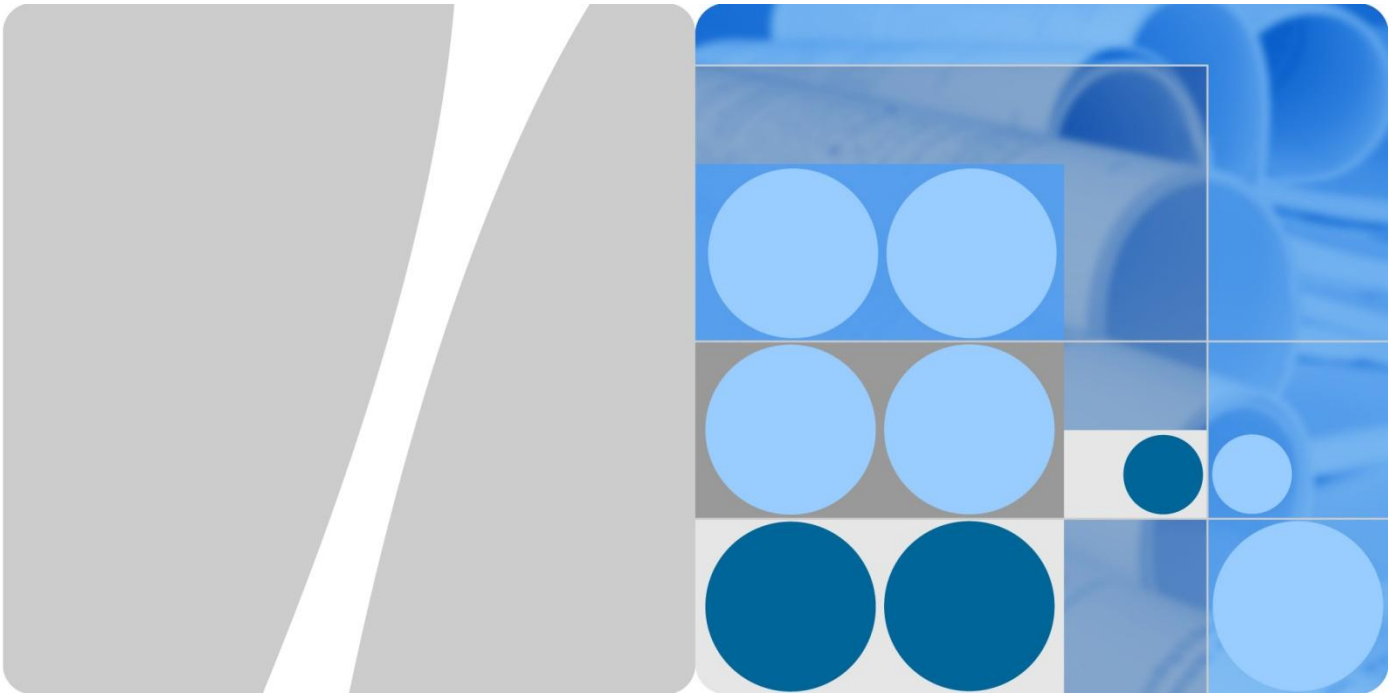


存储系统端到端数据完整性保护

技术白皮书

文档版本　V1.1

发布日期　2013-11-20

ORACLE　EMULEX　HUAWEI

华为技术有限公司

商标声明


和其他华为商标均为华为技术有限公司的商标。

本文档提及的其他所有商标或注册商标，由各自的所有人拥有。

注意

您购买的产品、服务或特性等应受华为公司商业合同和条款的约束，本文档中描述的全部或部分产品、服务或特性可能不在您的购买或使用范围之内。除非合同另有约定，华为公司对本文档内容不做任何明示或默示的声明或保证。

由于产品版本升级或其他原因，本文档内容会不定期进行更新。除非另有约定，本文档仅作为使用指导，本文档中的所有陈述、信息和建议不构成任何明示或暗示的担保。

华为技术有限公司

地址： 深圳市龙岗区坂田华为总部办公楼 邮编：518129

网址： http://enterprise.huawei.com

修改记录

修改记录累积了每次文档更新的说明。最新版本的文档包含以前所有文档版本的更新内容。

文档版本 V1.1 (2013-11-20)

文档内容更新如下：

第 4 章 方案验证

修改了表 4-1 测试环境中的操作系统版本号如下。

Kernel 从“Oracle Linux OLE6U3 2.6.39-400.203.4.el6uek.di.x86_64”修改为“Oracle Linux OLE6U3 2.6.39-400.207.0.el6uek.di.x86_64”。

SCSI driver 从“8.3.7.10.7p”修改为“8.3.7.26.3p”。

全文

修改了华为存储产品名称描述。

从“华为 OceanStor 高端存储系统”修改为“华为 OceanStor 18000 系列高端存储系统”。

文档版本 V1.0 (2013-09-23)

第一次正式发布。

目 录

1 摘要

本白皮书描述了使用 DIX 和 T10 Protection Information（T10 PI）的端到端数据完整性保护方案，用于防止数据存储中的静默数据破坏。白皮书同时给出了有关该方案基于 Oracle 数据库、Oracle Linux UEK、Emulex LightPulse® Fibre Channel 光纤通道主机总线适配器（HBA）及华为 OceanStor 18000 系列高端存储系统的验证信息。

关键字：静默数据破坏　　数据完整性　　端到端　　DIX　　T10 PI

2 简介

2.1 静默数据破坏

数据在读、写、传输、存储过程中，要经过多个部件、多种传输通道和复杂的软件处理，如果数据被破坏，可能会导致数据错误。若错误无法被立即检测出来，而是当后续应用在访问所保存的数据时才发现，叫做静默数据破坏。由于错误没有在发生时被发现，可能错失最佳的修复时机，最终将导致关键数据错误、系统宕机等严重后果。

造成静默数据破坏的环节可能有以下多种，包括：

- 硬件错误

  内存、CPU、硬盘、数据传输链路等。

- Firmware 错误

  HBA、硬盘等。

- 软件 bug

  操作系统、库、驱动程序、应用软件等。

欧洲粒子物理研究所（CERN）对约 8.7TB 的数据进行了连续 5 周的实际测试和统计，一共发现了 22 个静默数据破坏，平均每 1500 个文件就有 1 个静默数据破坏。另据威斯康辛大学、多伦多大学和 NetApp 在对存储系统中的 153 万块硬盘进行的连续 41 个月的实际业务运行检测，发现其中 3078 个 SATA 盘和 760 个 FC 盘的数据访问过程中出现了静默数据破坏问题。

2.2 相关技术

本章介绍业界两种主流的防止静默数据破坏的技术：T10 PI 和 DIX。

T10 保护信息（T10 PI）标准

ANSI T10 定义了一种通过对每个数据块加入保护信息（PI：Protection Information）也曾被称作数据完整性域（DIF：Data Integrity Field）的方法来保护数据完整性。在 T10 PI 标准中，每个逻辑扇区被扩充了 8 字节的保护信息，包括 2 字节的 Logical Block Guard，2 字节的 Logical Block Application Tag 和 4 字节的 Logical Block Reference Tag。

图2-1 T10 PI 格式

Byte \ Bit	7	6	5	4	3	2	1	0
0	USER DATA							
•••								
n - 1								
n	(MSB)	LOGICAL BLOCK GUARD						
n + 1								(LSB)
n + 2	(MSB)	LOGICAL BLOCK APPLICATION TAG						
n + 3								(LSB)
n + 4	(MSB)	LOGICAL BLOCK REFERENCE TAG						
•••								
n + 7								(LSB)

其中，Logical Block Guard 为 16 位的 CRC 值，用于验证数据有效性；Logical Block Application Tag 是应用层自定义的保护信息，Logical Block Reference Tag 是用于检查数据地址有效性的保护信息。

DIX

T10 PI 只包含了从主机 HBA 卡通过存储阵列到硬盘的数据保护，Oracle 和 Emulex 主导开发了 Data Integrity Extension（DIX）技术，将数据完整性保护扩充到了从应用层到 HBA。DIX 使用和 T10 PI 一样的 8 字节数据完整性信息作为数据校验字段。不同的是，DIX 中使用了 IP Checksum 作为 Logical Block Guard，以降低主机 CPU 的计算开销。

3 解决方案

3.1 方案组件介绍

本节将介绍方案中涉及到的三个组件：

- Oracle 的 UEK 内核
- Emulex 的 LightPulse FC 主机接口卡
- 华为 OceanStor 18000 系列高端存储系统

Oracle UEK Linux 内核

Oracle 的 Unbreakable Enterprise Kernel（UEK），包括了 Oracle 的数据库、中间件和硬件设计团队合作开发的许多优化内容，从而确保为要求最高的企业业务负载提供稳定性和最佳性能。UEK 包括了对 T10 PI 和 DIX 的支持，提供从应用到硬盘的端到端完整性。

本白皮书推荐使用 kernel-uek-2.6.39-400.207.0.el6uek 以及后续版本。

Emulex LightPulse 光纤通道 HBA

Emulex LightPulse HBA 提供了防止静默数据破坏的端到端数据保护特性。Emulex LightPulse 8Gb 光纤通道（FC）HBA 系列的 LPe12000、LPe12002、LPe12004 及 Emulex LightPulse 16Gb 光纤通道 LPe16000 和 LPe16002 均提供 BlockGuard™ Data Integrity 特性。此外，新的 LPe16000B 和 LPe16002B 第五代（16GFC）PCIe 3.0 HBA 采用了 T10 PI 高性能卸载技术，可通过在硬件中进行检查来提高性能，将 CPU 资源节约出来用于其它处理任务。

本白皮书中的方案验证使用了 Emulex 8Gb Lpe12002 型号的 HBA。

华为 OceanStor 18000 系列高端存储系统

华为 OceanStor 18000 系列高端存储系统是华为存储的高端旗舰产品系列，包括 OceanStor 18500、OceanStor 18800、OceanStor 18800F 几种型号，致力于为企业级数据中心提供安全可信、弹性高效的核心存储解决方案。

OceanStor 18000 系列高端存储系统支持 T10 PI 标准的端到端数据完整性。在从应用到存储的端到端完整性方案中，Oracle ASMLib 生成的数据保护信息，经 Emulex

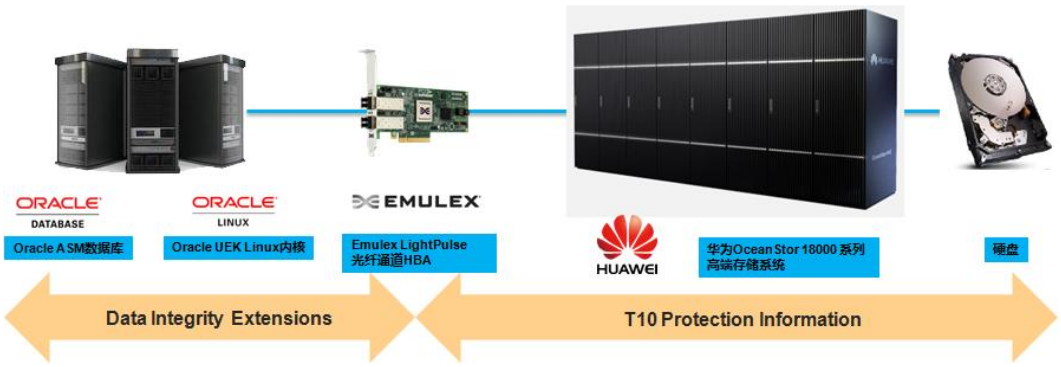
HBA 卡传输到阵列后，OceanStor 18000 系列高端存储系统将基于 T10 PI 标准对数据进行全 I/O 路径的端到端保护。存储系统对数据进行基于 T10 PI 的完整性校验，在数据访问过程中发现完整性错误时，会立即进行修复。

3.2 端到端数据完整性

方案概述

DIX 保护了从主机应用到主机 HBA 的数据完整性，T10 PI 保护了从主机 HBA 到硬盘介质的数据完整性，DIX 和 T10 DIF 一起组成了从主机应用到存储硬盘的端到端数据完整性保护。见图 3-1。

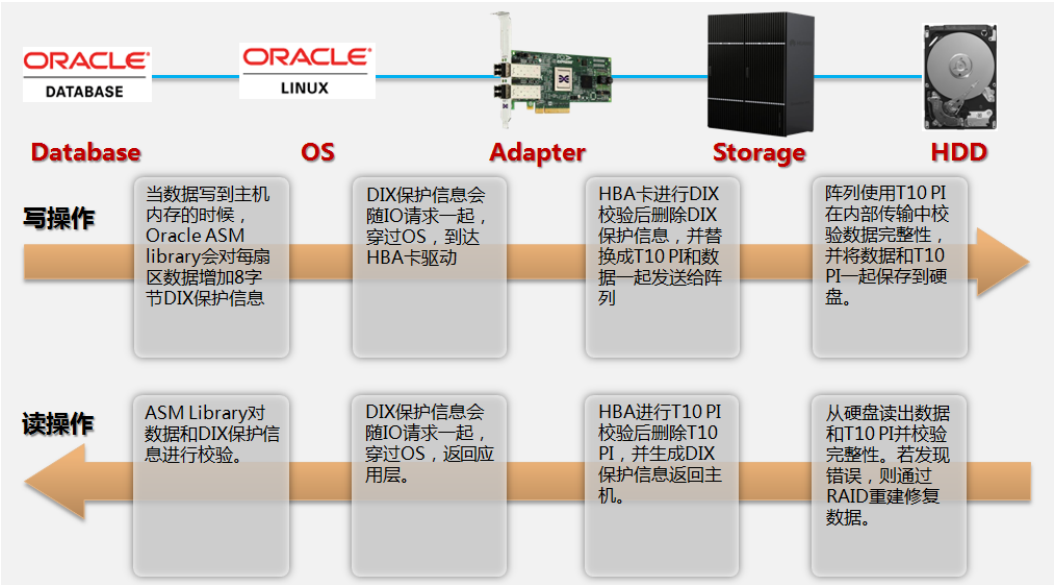
图3-1 端到端数据完整性解决方案



流程和方法

图 3-2 描述了端到端数据完整保护的流程。

图3-2 端到端数据完整性保护流程



以上流程构成了从主机应用层-OS-HBA-SAN-硬盘的端到端数据完整性保护流程。在整个流程中，从应用到存储系统中的一切潜在错误源，包括应用软件错误、驱动错误、内存错误、接口卡错误、传输链路错误、硬盘错误可能导致的数据破坏都得到了充分的检测和修复。

4 方案验证

4.1 环境与配置

测试环境

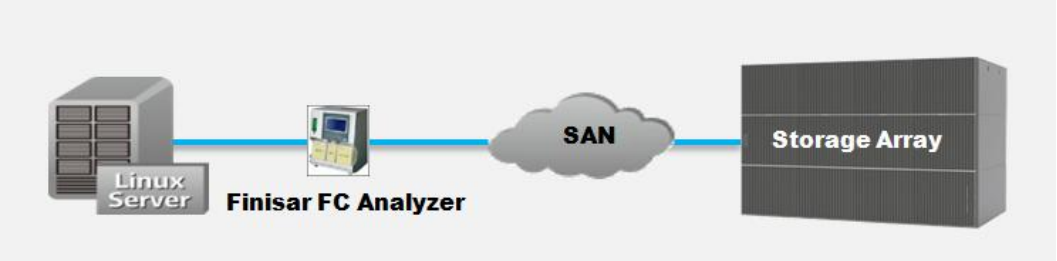
测试环境如表 4-1 所示。

表4-1 测试环境

主机服务器	Huawei Tecal RH2285 (Intel x86_64 based server)
数据库	Oracle 11gR2 11.2.0.3
操作系统	Kernel: Oracle Linux OLE6U3 2.6.39-400.207.0.el6uek.di.x86_64 SCSI driver: 8.3.7.26.3p
主机接口适配器	Model: Emulex LPe12002-M8 FW Version: 2.01A11 (U3D2.01A11)
存储阵列	Huawei OceanStor 18500 V100R001
测试工具	● ASMIO (for running IO and error injection) ● Finisar Fibre Channel analyzer (for jammer injection)
多路径软件	● Huawei UltraPath V100R006 ● Linux DM 多路径

组网配置

图4-1 组网图



4.2 测试方法

华为可靠性与兼容性实验室验证了华为/Oracle/Emulex 的 DIX+T10 PI 端到端数据完整性保护方案。测试包括超过 I/O 负载测试，以及错误注入测试，包含 ASMIO 模拟错误发生在 ASMLib、HBA、目标器，和使用协议分析仪在传输链路上注入错误。

测试覆盖以下多路径配置。

- 无多路径
- 华为 UltraPath 多路径
- Linux DM 多路径

测试覆盖以下 I/O 应用。

- ASMIO 测试
- RDBMS 应用测试

4.3 结论

上述测试均经过华为可靠性和兼容性实验室严格验证通过。

- 所有负载测试用例运行 40 小时以上无错误发生。
- Jammer 注入测试结果表明，HBA 和存储阵列分别校验出 FC 链路上的读方向 I/O 错误和写方向 I/O 错误，并在日志中记录了相应的 SCSI 错误码。
- ASMIO 注入测试表明，应用程序库的错误将由应用层校验出并恢复，HBA 的错误将由 HBA 校验出并返回给操作系统最终由 ASMIO 重试恢复，目标器的错误将由阵列校验出并返回给主机最终由 ASMIO 重试恢复。

华为可靠性与兼容性实验室验证通过了本端到端数据完整性方案，确认其符合目前最严格的端到端数据完整性验证的 DIX+T10 PI 行业标准要求。

5 缩略语

缩略语清单

英文缩写	英文全称	中文全称
ASM	Automatic Storage Management	自动存储管理
ASMLib	ASM Library	ASM 库
CERN	European laboratory for particle physics （=[法]Conseil Européen pour la Recherche Nucléaire）	欧洲粒子物理研究所
CPU	Central Processing Unit	中央处理器
CRC	Cyclic Redundancy Check	循环冗余码校验
DIF	Data Integrity Field	数据完整性域
DIX	Data Integrity Extension	数据完整性扩展
DM	Device Mapper	设备映射
FC	Fibre Channel	光纤通道
HDD	Hardware Disk Drive	硬盘驱动器
I/O	Input/Output	输入/输出
OS	Operating System	操作系统
PCIe	Peripheral Component Interconnect Express	外设组件快速互连技术
PI	Protection Information	保护信息。特指 T10 定义的数据完整性信息格式
RAID	Redundant Arrays of Inexpensive Disks	廉价冗余磁盘阵列

RDBMS	Relational Database Management System	关系型数据库管理系统
SAN	Storage Area Network	存储区域网络
SCSI	Small Computer System Interface	小型计算机系统接口
UEK	Unbreakable Enterprise Kernel	坚不可摧的企业级内核。Oracle 推出的具有高性能高可靠性的 Linux 内核

6 参考资料

出版物

[1] T10/1799-D, Information technology -SCSI Block Commands – 3 (SBC-3) Revision30, 21 February 2012

[2] L.N. Bairavasundaram, G.R. Goodson, B. Schroeder, A.C. Arpaci-Dusseau, and R. Arpaci-Dusseau, “An Analysis of Data Corruption in the Storage Stack,” in *Proceedings of the 6th USENIX Symposium on File and Storage Technologies (FAST ’08)*, San Jose, California, Feb. 2008.

链接

了解华为存储的更多信息请参考：

http://enterprise.huawei.com/cn/products/itapp/storage/index.htm

了解华为在企业 ICT 市场的更多信息，请访问：

http://enterprise.huawei.com

了解 Oracle Linux 的更多信息，请参考：

http://www.oracle.com/linux

了解 Emulex HBA 产品的更多信息，请参考：

http:// www.emulex.com/products/fibre-channel-hbas.html