白皮书

企业高端存储：安全可信

赞助者：华为

William Zhang

July 2013

IDC 观点

IT 世界正在从“以信息为中心”向“以数据为中心”转变，越来越多的企业认识到数据对于企业的价值。数据损坏或丢失给企业带来的损失是无法想象的，严重时甚至会导致企业破产。因此，企业在选择数据存储系统时，首要考虑的因素就是系统的可靠性，特别是针对关键性应用，对存储系统的可靠性要求更加苛刻。因此，企业都会选择采用高端存储系统来承载核心应用。据 IDC 统计，2012 年，售价高于 10 万美元的硬盘存储系统在中国市场的增长率高达 43.3%，是整个硬盘存储系统市场平均增长率（19.8%）的两倍多。

如何选择一款可靠性较高的存储系统十分重要，IDC 建议 IT 部门关注以下几个层面：

- **存储系统硬件架构可靠性。**存储系统采用全冗余架构，系统的关键部件如：控制器、电源、风扇和组网等采用冗余架构设计，确保不存在单点故障，不会影响系统的可靠性。
- **存储系统数据存储可靠性。**数据存储必须具备端到端的数据保护能力，从故障自检测、故障预处理、故障快速修复和修复后数据完整性检测等多个方面来确保数据存储的可靠性。
- **存储系统业务应用可靠性。**存储系统能够支撑多个关键应用的负载，具备业务优先级监测与管理功能，同时具备高级功能，如快照、镜像、远程复制等确保业务连续性的解决方案。

IDC 中国：中国北京西长安街 88 号首都时代广场 611 室 P.+86.10.8391.7777 F. +86.10.8391.3600 www.idc.com.cn

关于本白皮书

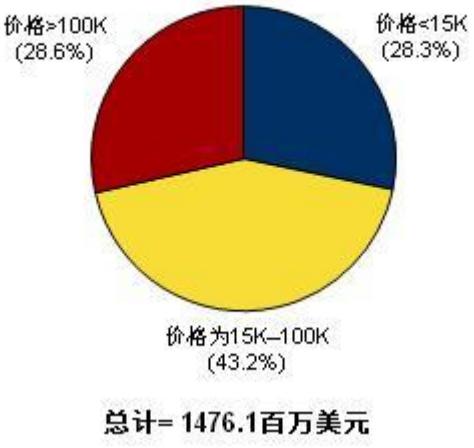
在本白皮书中，我们将探究华为 OceanStor 18000 系列高端存储系统的安全可信架构，深入了解 18000 系列高端存储产品在整机设计、系统架构、数据存储和业务运行保护等方面的设计，以及如何充分满足用户安全可信的需求，最大化提升用户的价值。

市场综述

据 IDC 统计，2012 年，中国的存储市场规模（按厂商销售额统计）达到 1476.1 百万美元，同比增长 19.8%，其中价格区间在 10 万美元以上的存储系统的市场规模为 421.9 百万美元，占整体市场份额的 28.6%，年均增长率高达 43.3%。详细信息，如图 1 和 2 所示。

图 1

中国存储市场规模，2012

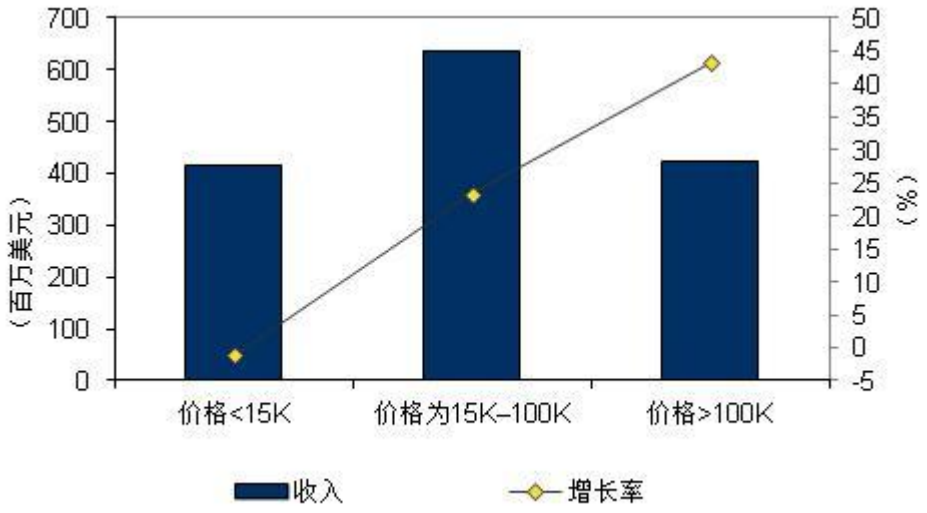


来源: IDC, 2013

 ©2013 IDC 中国

图 2

中国存储市场规模及增长率，按价格区间细分，2012



来源: IDC, 2013

华为 OCEANSTOR 1800 系列高端存储系统解决方案

系统架构安全可信

全冗余系统架构设计

华为 OceanStor 18000 系列高端存储系统采用全局冗余架构设计，从电源、风扇、控制器等基础设备层面，到互联冗余组网层面和跨机柜的全冗余业务交换层面，保证了整系统全冗余，任意一个组件、模块、乃至设备的单点故障都不会影响业务的正常运行。

Smart Matrix 系统架构

18000 系列高端存储采用智能矩阵式系统架构 Smart Matrix Architecture（如图 3 所示），它是一种基于 PCI–E2.0 全冗余全交换的系统架构，以互相冗余的双平面的 PCI–E 高速矩阵交换模块为数据存储/交换的核心，任意一台 PCI–E 高速矩阵交换模块故障不会影响数据读写的安全，从而保障了业务连续性。在 PCI–E 高速矩阵交换模块中实现数据交换有一个巨大的优势，它可以极大地减少协议转换的时延，实现全局资源无阻塞互连，获取更加高效的数据交换，提供高达 192GB/s 的内部交换带宽，而时延只有 300ns，仅为业界同样条件下的 1/5。18000 系列的存储控制器支持 scale–out 扩展，最多可支持 16 个控制器。

18000 系列的缓存为全局共享，这也是 Smart Matrix 架构设计的独特价值。主机卷在存储系统上划分为多个单元组成的 LUN，每个单元归属于不同的控制器，每个控制器都有自己的缓存并且用于数据读写加速，所有控制器共享全局缓存。因此，系统内所有的缓存可以为同一个主机卷加速，全面提升数据的缓存命中率，串行化达到最低，系统获得了理论最高加速比，缩短了业务响应时间，存储更快速响应业务需求，极大的加快了数据处理的速度。同时有效避免了传统存储的缓存独占和全局缓存争用从而影响业务的正常运行的问题，使系统更加稳定可靠。

图 3

Smart Matrix 架构

来源: 华为, 2013

九烈度抗震认证

在整机设计方面，华为存储通过了信息产业通信设备抗震性能质量监督检验中心的九烈度抗震验证，这是目前国内唯一一家通过此项验证的专业存储厂商。18000 系列继承了华为存储在整机设计方面的可靠性设计，确保在重大地震的情况下数据不丢失，可有效防御 50 年内 90%以上的地震危害。

数据存储安全可信

XVE（Extreme Virtual Engine）存储操作系统

18000 系列采用 XVE（Extreme Virtual Engine）存储操作系统，XVE 的核心理念是全虚拟化：全虚拟化内核、全虚拟化 RAID（RAID 2.0+）和全虚拟化资源池，通过虚拟化的底层存储实现资源的均衡分配，避免瓶颈的产生，使业务运行更稳定；同时虚拟化资源池上承载的 Smart 系列、Hyper 系列软件和虚拟卷管理软件，使存储在高效的管理数据的同时支持高级数据保护等特性，更好的支持业务容灾。XVE 操作系统全虚拟化的设计，使业务运行更稳定，符合华为 OceanStor 18000 系列高端存储系统安全可信的设计理念。

RAID 2.0+：数据重构速度提升 20 倍

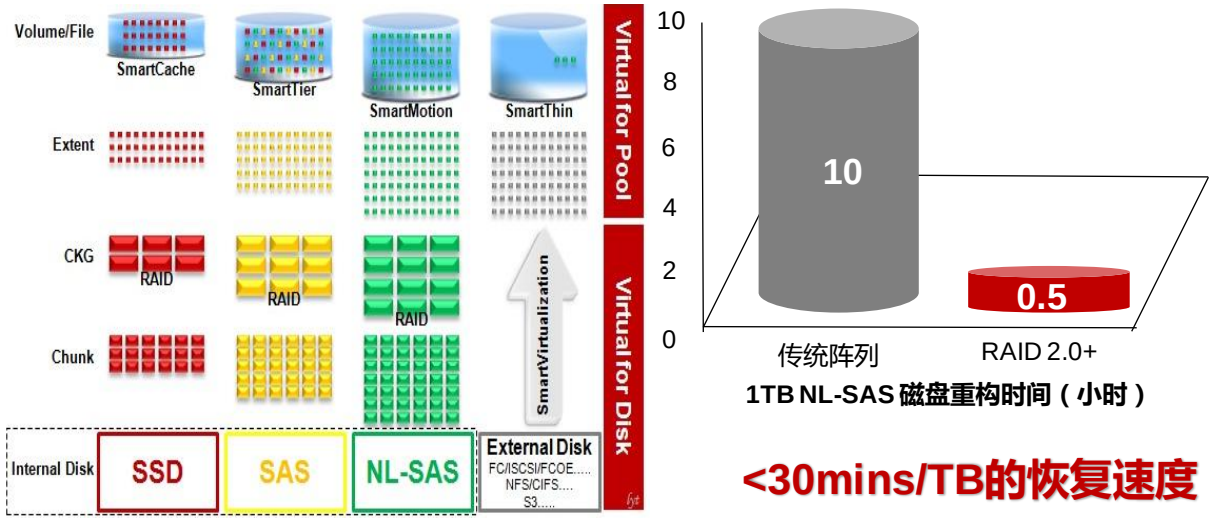
华为公司首创的 RAID 2.0+技术是完全的存储虚拟化管理模式，它提出了两层虚拟化的理念，通过底层虚拟化实现硬盘读写与基础数据保护，通过上层虚拟化实现融合的存储资源池与智能资源调度以及高级数据保护。如图 4 所示，针对底层存储介质，18000 系列高端存储系统构建底层硬盘虚拟化（Virtual for disk），系统内部物理硬盘，按介质的性能不同，分为三个存储层，通过其它方式接入系统的作为外部存储层，整体对外呈现为一个大的资源池；将系统内部各个硬盘空间切分成 64MB 大小的逻辑块（Chunk），将来自不同硬盘上的逻辑块（Chunk）按 RAID 组成逻辑块组（CKG）；将逻辑块组（CKG）切分成更细粒度逻辑块（Extent），按需将 1–N 个更细粒度逻辑块（Extent）组成卷（Volume）/文件（File）。创建 RAID 时，不再基于固定的硬盘，而是先将每个物理硬盘虚拟化为若干个逻辑块（Chunk），再以多个硬盘的逻辑块（Chunk）按一定的算法关系构建 RAID。当一个逻辑块故障时，重构的仅是一个逻辑

 ©2013 IDC 中国

块（Chunk）大小的数据，1–3 秒即可完成重构。当一个物理硬盘故障时，重构的仅是多个有实际数据的逻辑块（Chunk），同时更多的目标硬盘参加重构。使得每 TB 数据重构时间小于 30 分钟，数据重构速度相对传统 RAID 提高了 20 倍。并且由于重构是分布进行的，对每块硬盘的压力都极小，大幅降低了重构过程对业务的影响。

图 4

RAID 2.0+两层虚拟化逻辑结构：数据重构速度提升 20 倍



来源: 华为, 2013

如图 5 所示,在 RAID 2.0+的重构时间测试中，采用 64 个 2TB NearLine SAS 盘构建一个 Storage Pool，Pool 中创建 50 个大小为 1TB LUN，LUN 已格式化完成，并写入数据。 此时我们拔出一块硬盘，18000 系列启动重构，数据写入 Storage Pool 的热备空间。

图 5

RAID 2.0+:数据重构测试结果

来源: 华为, 2013

从日志看出，数据重构完成一共耗时 22 分 15 秒，平均重构速度为 1229MB /s。按照此速度推算，1TB 数据的重构时间约为 15 分钟。

RAID 2.0+的另一层虚拟化是上层资源虚拟化，底层存储资源对上层来说是一个大的资源池，提供给主机操作系统的 Volume/File 基于资源池（Pool）创建，资源池内的资源由 Extent 组成，Extent 作为逻辑地址，映射关系的基本单位是十分灵活的，而且支持动态调整，不再受限于单个 RAID 内硬盘数量有限的问题，配合华为 Smart 系列数据存储管理软件，实现灵活高效的数据管理，在资源池内实现数据的智能流动，包括智能精简配置（SmartThin）、智能数据分级（SmartTier）、智能数据迁移（SmartMotion）、智能异构虚拟化（SmartVirtualization）、智能服务质量控制（SmartQoS）、智能缓存分区（SmartPartition）满足最大化的自动负载均衡、最大化硬盘资源利用率、最大化容量资源利用率和提升存储管理的效率。对管理员来说，在做存储系统规划时，只需要计算出当前业务总容量和性能的需求，并预留一定比例增长余量即可；在存储系统的具体配置过程中，只需要做简单的资源划分，系统将根据业务的实际使用情况自动调整 Extent 的配比，以满足业务各单元的容量和性能需求；当业务发展导致系统预留余量资源不足时，系统会自动提示管理员添加资源，而添加资源的过程仅需插入硬盘，并将硬盘加入到相应的硬盘资源池（Disk Pool）中，然后，系统将在后台自动调整 Extent 的分布，重新实现全局均衡。

数据自检测自修复

华为 OceanStor 18000 系列高端存储系统从故障自检测、故障预处理、故障快速修复、修复后数据完整性检测等多个方面，全流程构建端到端的数据保护，确保数据存储的安全可信。存储系统周期性收集和统计硬盘信息，根据硬盘的运行时间、内部错误统计值、硬盘 IO 模型进行 DHA（Disk Health Analysis）硬盘健康度分析，根据分析结果生成一个包含分值信息的文档。IT 维护人员可以根据分值，便捷的判断硬盘是否处于正常状态。例如：100 分为满分，硬盘状态正常；60 分为不合格，需要更换硬盘。当硬盘处于异常状态时，存储系统会启动数据自修复功能。如存在硬盘坏道时，系统会自动启动坏道修复；如存在慢盘或者即将失效时，系统会启动数据预拷贝，将数据转移到健康的硬盘上。拷贝完成后，系统会自动检测已拷贝的数据与原数据是否一致。18000 系列应用硬盘自检测自修复技术，提前预测了即将发生的故障，在故障发生前对数据进行保护，有效避免了很大一部分故障的发生，提高了数据存储的安全可信。

数据完整性检测

在经过数据自检测自修复和 RAID 2.0+快速数据重构等一系列处理后，18000 系列还具备数据完整性检测机制，采用符合 T10 标准的 PI 技术保证数据的完整性。PI 指存储系统通过自动在每个扇区数据后加上八个字节的数据完整性字段来实现数据完整性检测，检查从主机 HBA 卡通过 SAN 光纤向硬盘写入数据，以及从硬盘读取数据的完整性。该字段跟随用户数据一起参与各种转发、传输并最终存储到磁盘介质中，数据被主机应用重新读出前，系统会通过 PI 检查数据的正确性和完整性，并通过数据冗余（如 RAID）修复数据，保证用户数据的可靠性。数据从应用到主机 HBA 的路径上，主机侧具有同样的数据保护，确保用户在受到影响前可以自动的检测和恢复错误，这就是 DIX（Data Integrity Extensions）。DIX 延伸了 PI 的保护范围，而 PI+DIX 则实现了从应用到硬盘的端到端的数据保护。

存储业务安全可信

保证关键业务的服务质量和性能

SmartQoS 全面控制 CPU、内存等存储系统资源，使用户可以针对不同的业务划分不同的优先级（以 LUN 为单位），保证关键业务能够获取足够多的存储资源，系统优先

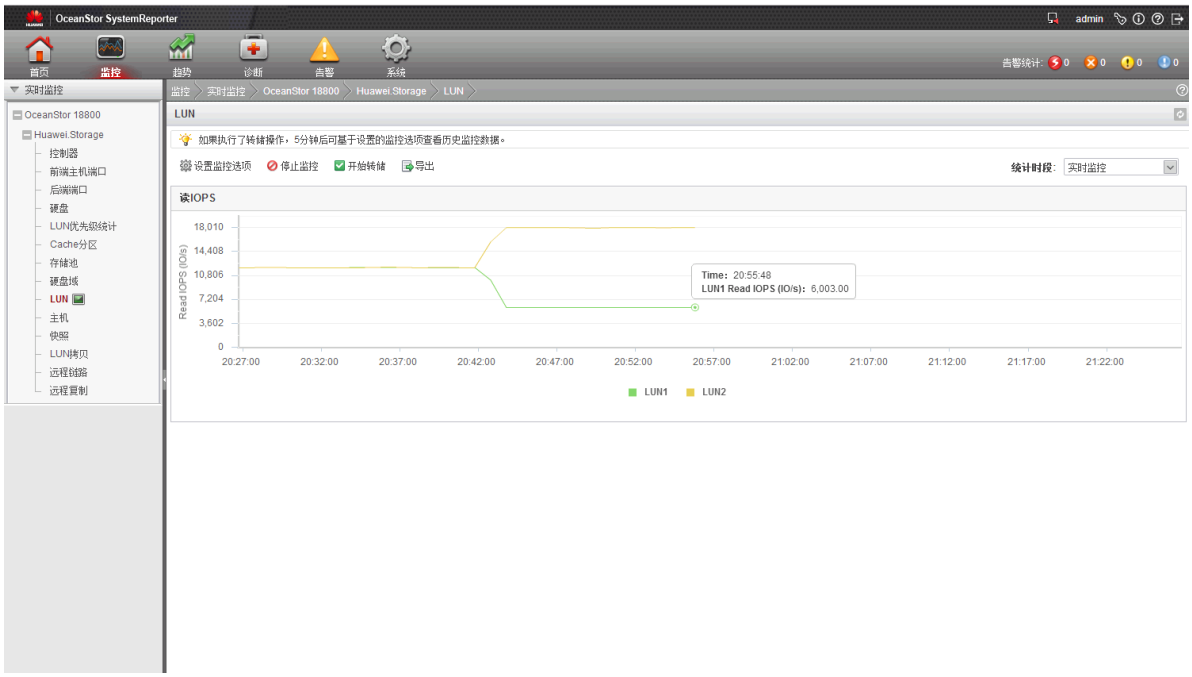
 ©2013 IDC 中国

响应高优先级的访问请求，从而保证了关键业务的稳定运行。18000 系列可针对不同业务设置优先级，也可通过限制 IOPS 和带宽的手段保证高优先级业务的资源请求。

如图 6 所示，在 SmartQoS 限定 IOPS 测试中，构建了一个 192 盘的存储资源池，在此资源池中建立了两个 2TB 的 LUN，归属同一个控制器，进行数据库业务模型（8∶2 读写比例，全随机，IO 延迟 10ms 以下）的测试，单个 LUN 的 IOPS 达到 12000，两个 LUN 的性能基本相同。此时，将 LUN1 的 IOPS 限制到 6000。当 SmartQos 策略生效后，LUN1 的 IOPS 逐渐降为 6000，LUN2 的 IOPS 上升到 18000。

图 6

SmartQoS 限定测试

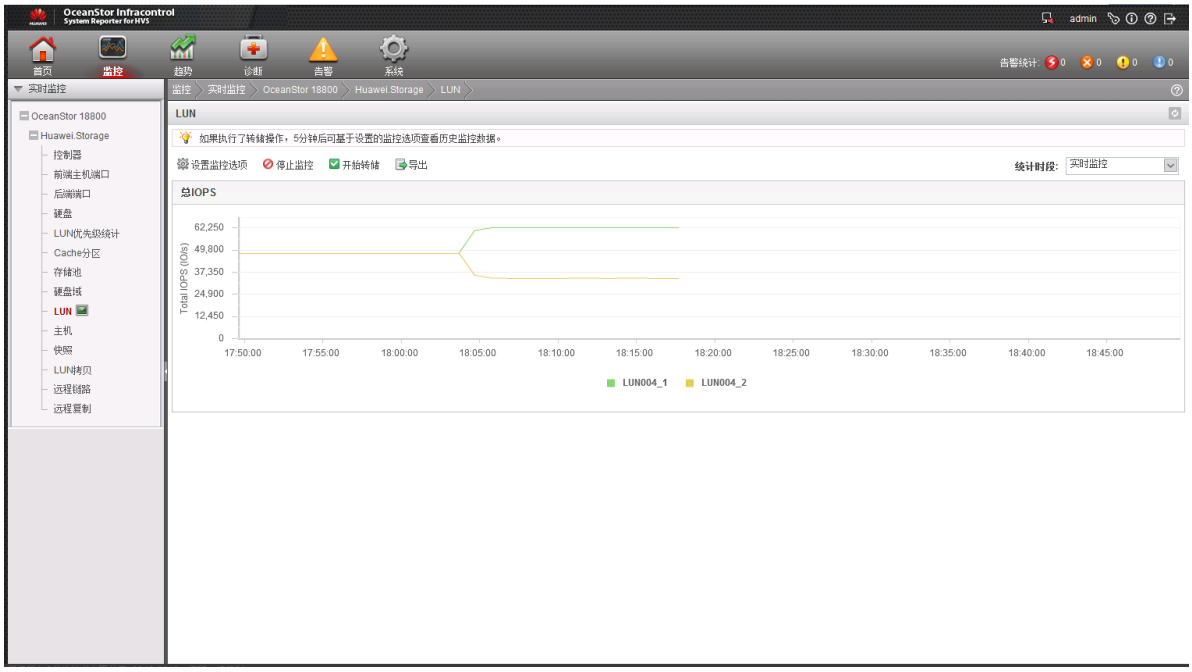


来源: 华为, 2013

如图 7 所示，在 SmartQoS 优先级测试中，在 192 盘的存储资源池中建立两个 100GB 的 LUN，归属为同一个控制器，进行数据库业务模型（8∶2 读写比例，全随机，IO 延迟 10ms 以下）的测试，单个 LUN 的 IOPS 为 48000。读写一段时间后，将 LUN1 设置为 IO 高优先级，LUN2 设置为 IO 低优先级。当 SmartQos 策略生效后，系统资源向高优先级 LUN 倾斜，LUN1 的 IOPS 达到 62000，LUN2 的 IOPS 下降到 34000。

图 7

SmartQoS 优先级测试



来源: 华为, 2013

SmartPartition 把物理缓存分割为大小不同的多个区域，每个分区之间彼此隔离，所有分区共享整个系统的缓存，用户可以为高优先级的业务提供专用的缓存资源，并且智能的调整主机并发和硬盘访问并发。确保在业务繁忙的时候，低优先级的业务不会抢占高优先级的缓存资源，从而保证高优先级的关键业务能够更高效的处理事务并保持持续稳定的运行。

核心业务容灾

华为 OceanStor 18000 系列高端存储系统具备完备的数据容灾解决方案，支持多种关键业务的容灾，包括 Oracle、SAP、ERP 等，其内部的 Hyper 系列高级数据保护特性实现 HyperSnap 快照、HyperClone 克隆、HyperCopy LUN 拷贝、HyperReplication/S 同步远程复制、HyperReplication/A 异步远程复制等功能，保障不同站点之间的数据完全一致。Ultra 系列容灾管理软件实现一键式容灾管理切换，摒弃了以往复杂的容灾任务管理，降低了人为误操作的可能，极大的提升了业务恢复的速度，业务连续性达到 99.9999%，保障了系统的可靠性和业务运行的安全可信。

0～5 秒——业界最短的 RPO，18000 系列通过在 Cache 中设置时间戳，将生产中心 Cache 中的数据周期性的同步到灾备中心。在容灾过程中如果发生链路中断再恢复的场景，存储系统将自动恢复进行重同步，且重同步会自动对从 LUN 生成保护，保证从 LUN 数据的一致性。

UltraAPM，华为公司独特的 UltraAPM 容灾管理平台解决方案摒弃了以存储为视角的容灾管理方式，改为以客户应用为视角，以应用为容灾的核心元素，将容灾相关的管理工作集中化、图形化、流程化，符合用户习惯，帮助客户更好地建设、维护、使用容灾系统。

©2013 IDC 中国

图 8

使用了 18000 系列的 Oracle 数据库容灾测试

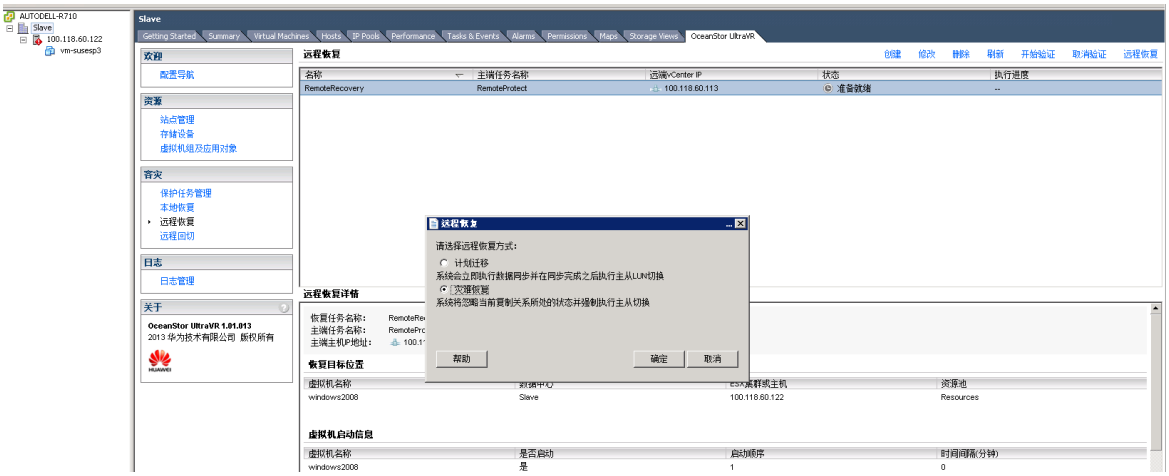
来源: 华为, 2013

如图 8 所示,本次测试中，生产端和灾备端使用两台华为 RH5885 V2 高性能服务器，并安装了 Oracle 数据库，使用华为 UltraAPM 软件结合 18000 系列高端存储系统进行数据库容灾测试。模拟生产端存储故障，进行容灾的切换。UltraAPM 的灾难恢复会对阵列侧远程复制做强制主从切换，让从 LUN 可读写，然后映射给灾备端应用主机，并自动启动 Oracle 数据库。测试结果显示，执行灾难迁移后，发现 Oracle 数据库已经在灾备端启动，数据与原生产端一致，容灾切换时间在 5 分钟左右。

UltraVR 是一款与虚拟化架构高度集成的容灾管理软件，可以对虚拟化环境中的虚拟机进行容灾设置与管理。UltraVR 支持 Vcenter 的深度集成，适配 VMware 虚拟机环境，配合华为存储设备的增值功能，可以提供：远程恢复、虚拟机本地恢复、容灾切换、容灾回切、容灾演练、一键恢复、数据验证、计划内迁移等功能。将纯手工操作变为系统按照设定好的流程自动执行，在不改变虚拟化基础架构的情况下满足用户的各种容灾需求，使虚拟机资源充分应用在应用服务器软件层面，提升数据保护的效率。UltraVR 虚拟化容灾管理软件支持 VMware 虚拟化平台、华为 FusionSphere 虚拟化平台，未来还将支持 Xen、Hyper–V 等虚拟化平台，用一套软件即可支持多个虚拟化平台的容灾管理工作。

图 9

使用了 18000 系列的 WMware 虚拟机容灾测试



来源: 华为, 2013

如图 9 所示,本次测试中，使用两台 RH5885 V2 高性能服务器做为 ESX 服务器，在 ESX 上建立多台 Windows 2008 R2 的虚拟机，底层挂载 18000 系列高端存储系统，虚拟机上运行了 Oracle 数据库。在主站点配置远程复制关系，灾备端存储不挂载。模拟生产端存储发生故障，强行进行系统业务切换。UltraVR 的灾难恢复会对阵列侧远程复制做强制主从切换，让从 LUN 可读写，然后映射给灾备端 ESXi Server，并自动启动虚拟机。测试结果显示，待容灾切换完成后，检查发现虚拟机已经正常启动，数据库的数据正常，和生产端一致，容灾切换时间在 5 分钟左右。

两地三中心解决方案：支持不同档位产品之间的数据流动

在生产中心、同城灾备中心和异地灾备中心分别部署不同档位的华为存储设备，通过远程复制功能使数据在生产中心和灾备中心之间流动，实现数据的远程保护和业务连续性。当生产中心发生灾难时，可在同城灾备中心进行主从切换并拉起业务，并保持与异地灾备中心的容灾关系。若生产中心和同城灾备中心均发生故障，可在异地灾备中心进行主从切换，并拉起业务，确保业务系统的持续运行。用户可以根据实际业务的需要制定不同的数据复制策略，复制过程可以根据业务特点灵活调整，极大的提高了容灾方案的灵活性；对客户而言，在生产中心和灾备中心部署不同档位的存储设备，实现远程数据容灾，打破了传统容灾系统中不同档位的存储系统不能互通的弊病，极大降低了容灾系统的建设成本，实现了超高的性价比，大幅提升了容灾方案的整体价值。

高达 32：1 的集中灾备

传统的容灾模式中只能实现 4：1 的复制，即 4 台存储设备到 1 台存储设备的复制。然而，当面临大型企业和政府部门众多分支机构需要集中容灾的需求时，容灾系统建设的成本和管理难度都将急剧上升。18000 系列高端存储系统针对该应用场景可实现 32：1 的复制模式，减少灾备中心存储设备的部署，降低容灾系统建设成本。同时结合华为容灾管理平台，简化大型容灾系统的管理和维护难度，开创了集约化容灾建设的新时代。

 ©2013 IDC 中国

支持第三方的存储系统

华为 OceanStor 18000 系列高端存储系统可以实现不同厂商、不同类型存储设备之间的异构整合，把其他厂商的存储设备当成华为 18000 系列的资源，统一进行管理，对外提供存储服务。同时配合 Smart 系列和 Hyper 系列软件，提升了存储设备的资源利用率，保护客户的既有投资，简化了用户管理，提高了系统的可靠性。异构整合支持异构存储设备内的数据迁入 18000 系列，改善原有数据的存储效能；也支持把 18000 系列上的数据迁移到异构存储设备中，将不活跃数据迁移至廉价存储上，降低总体拥有成本（TCO）。

华为 OceanStor 18000 系列高端存储系统面临的挑战与机遇

IDC 注意到，高端存储产品以其极高的性能、稳定性、可靠性受到金融、电信、政府等行业用户的青睐，主要应用于关键性业务，如核心计费系统、运营系统等。因此高端存储系统对厂商的研发实力要求很高，这个领域一直以来都以国际厂商为主，而华为是目前第一个国内自主研发推出高端存储产品的厂商。近两年，随着云计算、大数据等热门技术的快速发展，带动了高端存储的应用部署，华为选择这个时机进入高端存储市场是个不错的机会。

但同时，IDC 也注意到，华为在高端存储解决方案的推广方面将面临一些挑战，一个成熟的存储系统是需要经过长时间的市场考验的，能开发出产品只是其中的第一步，后续还需要在实际应用环境中不断的更新、完善产品，来满足复杂的 IT 环境的需求，尤其是高端存储系统主要面对的都是需求相对苛刻的关键性应用，所以对产品自身的要求会更高。华为的数据显示，18000 系列高端存储系统自 2012 年 9 月 5 日发布以来，已经成功部署超过 50 套，行业遍布政府、金融、电信和能源等领域，系统运行良好。华为预计到 2013 年底 18000 系列高端存储系统的出货量会超过百套。

结论

在可预见的未来，预计企业的 IT 预算增长水平有限，而数据泛滥却仍将继续。因此，IT 组织在确保关键数据安全性的同时还需要寻求高效、简化的解决方案，提高资金利用率。在本白皮书中，IDC 为 IT 部门概述了高端存储在系统架构、数据存储、存储业务三个方面安全可信的特性。除此之外，在评估存储解决方案时，IT 部门还应注意以下几点：

- 确保系统硬件架构安全可信的同时，应考虑存储系统的性能。
- 确保数据安全可信存储的同时，应考虑存储系统的效率。
- 确保灾备功能实现的同时，应考虑简化管理。

版 权 声 明

"如需向外界公布 IDC 资讯，包括用在广告、新闻发布、宣传资料等文件中，须经 IDC 相关地区级副总裁或该国分支总裁书面核准。该文件本身也应与 IDC 咨询一同提交。IDC 保留因任何原因而拒绝此类公开引用的权利。版权所有 2013 IDC。未经书面许可，不得复制"