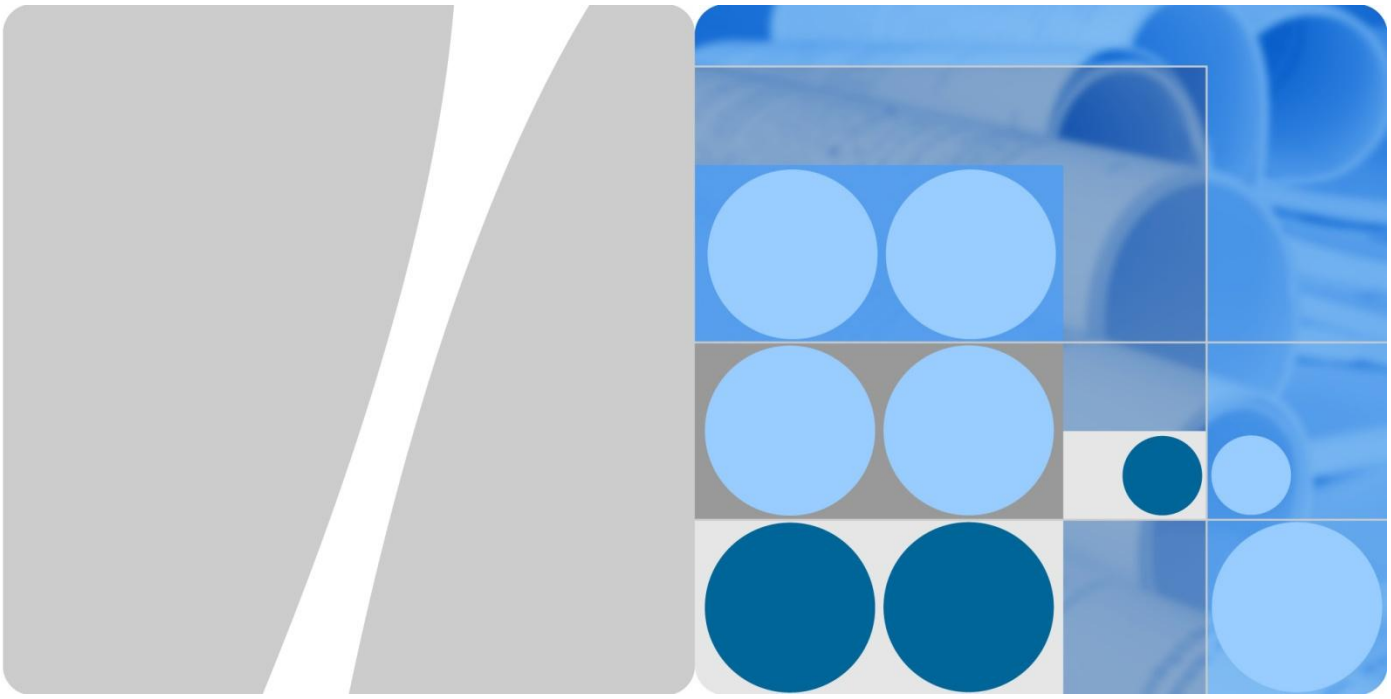
资料编码



OceanStor T 系列技术白皮书 (V200R002)：融易可靠 智能高效

文档版本 V1.0

发布日期 201402

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://www.huawei.com

客户服务邮箱： support@huawei.com

客户服务电话： 0755-28560000 4008302118

客户服务传真： 0755-28560111

修订记录/Change History

日期	修订版本	描述	作者

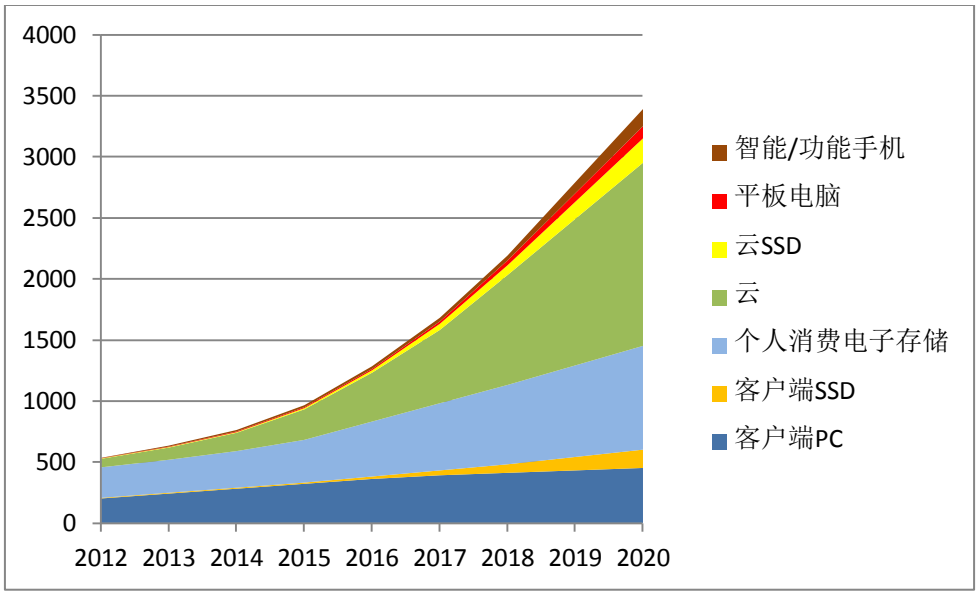
目 录

1 概述

信息技术从诞生发展到在生活和工作中无处不在，经历了大型机萌芽，小型机应用，个人电脑普及和桌面互联网，如今正在逐步步入移动互联的时代。应用环境的变化催生了数据的大爆炸。根据 Gartner 的统计结果，在小型机应用阶段，全世界大约生产了 2.6EB 的数据量；到个人电脑普及阶段，数据量增长到 15.8EB；而上一个桌面互联网阶段，数据量几乎翻了 3 倍，达到了 54.5EB；现在的移动互联时代，将会产生高达 1800EB 的数据。一个问题不禁摆在了人们的面前：爆炸的仅仅是容量吗？

首先，数据的来源更加的多样化。云会逐渐打破个人电脑和消费电子两强的局面，成为最大的数据产生源。未来数据来源的预计参考图表 1-1

图表 1-1 数据应用来源预测



产生数据的应用在不断变化，带来数据类型随之变化。关键业务数据（如数据库等）量持续增长，但在整个数据容量的比例却急剧减小；企业办公数据迅速增长，电子邮件、大媒体文件等数据一度暂居整个数据容量的最大比例；随着个人数据迅猛增长，媒体娱乐等消费产生的数据迅速替代了企业办公数据在整个数据容量比例中的老大位置。1993 年，关键业务和企业办公产生的数据各占 50%，个人数据几乎为零；2002 年左右，企业办公产生的数据占有 70%，关键业务数据占有 20%；而到了 2010 年后，个人数据占据 50%的比例，企业办公占有 40%，关键业务数据所占比例仅有 10%。

这些来自不同数据源的不同类型数据，对数据存储介质的性能、可靠性、成本等要求是多种多样的。关键业务要求的是高性能、高可靠的存储设备，而个人娱乐数据强调的却是低成本。如此矛盾的需求却常常要求在同一套存储设备中得到满足。这些新的趋势对中端存储提出了新的挑战。新一代的中端存储需要具备以下新的特质，才能不断跟上信息时代的脚步：

1. 具备融合、简约、智能的高性价比系统架构
2. 满足用户多变的存储使用需求
3. 灵活数据规划与管理
4. 实用多样的功能特性

华为技术有限公司全力打造的 OceanStor T 系列存储系统正是以融易可靠和智能高效为其设计理念，在一款产品中充分考虑时代发展趋势对存储阵列提出的需求，采用全新的软件平台，以强大的灵活扩展、智能的资源管理能力，最大化保护用户投资，提升用户价值。

2 “融易可靠，智能高效”的定义

OceanStor T 系列存储系统秉承统一存储的理念，在实现了文件级数据、块级数据和存储协议融合统一的基础上，以业界领先的存储资源虚拟化技术，超高可靠的软硬件设计架构，智能高效的存储资源调度算法，多种方式的 QoS 保障机制， 为用户提供了高性能、全方位的解决方案，使用户投资收益比最大化，能够满足大型数据库 OLTP/OLAP，高性能计算，数字媒体，因特网运营，集中存储，备份，容灾，数据迁移等不同业务应用的需求。

融易：协议统一，管理统一

- SAN 和 NAS 存储协议的 统一，在同一套存储系统内可以支持结构化和非结构化数据。支持 iSCSI、FC、NFS、CIFS、HTTP、FTP 等多种存储网络和协议
- 采用 RAID2.0+存储资源虚拟化技术，能够将物理硬盘虚拟化为许多小的存储单元，实现更加精细化的存储空间管理。

融易：横向的资源扩展

- OceanStor T 系列存储系统具备出色的可扩展性，它支持多种硬盘类型和主机接口模块、支持控制器从 2 个平滑扩展至 4 个(OceanStor S5500T*/S5600T/S5800T/S6800T)、在线扩容技术使存储池可以在线新增硬盘，轻松扩容存储池。同时，主机接口模块密度也处于业界领先水平，从而带来了出色的高可扩展性。

  *OceanStor S5500T 仅每控制器 16GB 缓存规格支持扩展至 4 个控制器

可靠：降低故障，快速自愈

- 采用 RAID2.0+存储资源虚拟化技术，提升数据重构速度，最快可达 2TB/小时的数据重构量；
- 采用硬盘坏道修复技术，可以自动修复硬盘坏道，使硬盘故障率降低 50%，延长了硬盘的使用周期；
- 采用多重硬盘防护专利技术，从振动、腐蚀等各方面满足 DC G1 到 DC GX 级别条件下可靠运行。

可靠：秒级的容灾周期

- T 系列存储系统采用创新的多时间戳（Timestamp）缓存技术，在进行复制和同步时，直接从生产端读取相应时间戳分片的数据复制到灾备端，降低了时延，保证最小同步周期缩短到 3 秒。

智能：自动数据分级

- 自动分析单位时间内存储数据访问频率，根据分析结果自动将存储数据迁移到不同性能的硬盘中（高性能层硬盘存储活跃数据；性能层硬盘存储热点数据；容量层硬盘存储冷数据），获得最优的综合性能并且降低单位 IOPS 成本。

智能：自动精简配置

- 使存储空间能够根据需要自动扩展，而不必像传统方式那样一次性将存储空间全部分配出去，因此只需要配置少量硬盘即可开展业务，后续再根据存储空间使用情况新增硬盘，从而降低初次购买成本和 TCO。

高效：服务质量保障

- 可根据业务数据的一系列特征进行分类（每一种分类代表一种应用），并能够对于每一种分类设置优先级和性能目标，从而将合适的资源提供给合适的业务，达到充分利用存储资源的目的。

高效：缓存分区优化

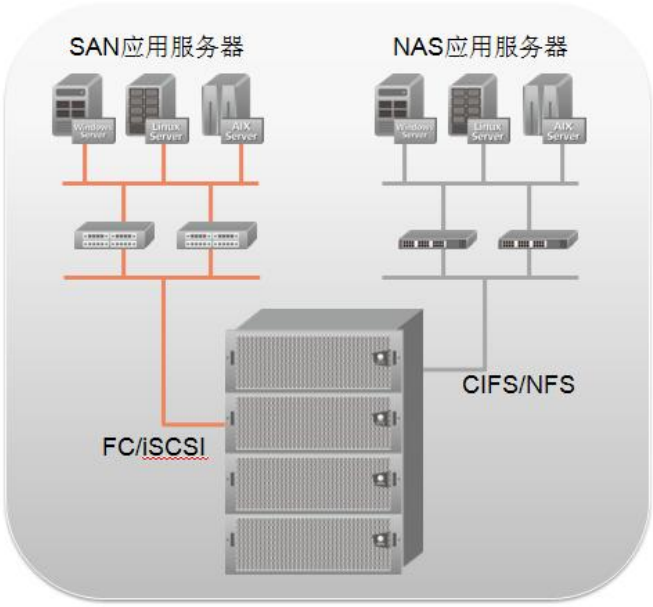
- 通过对系统核心资源的分区，保证关键应用的服务性能。用户可以配置不同大小的缓存分区，系统将保证该分区中业务应用所占用的缓存数量，并根据实际情况自动调整不同分区中的主机端并发数量，从而保证位于该分区中的业务应用的服务性能。

3 融易：协议统一，管理统一

3.1 存储协议的统一提供

OceanStor T 系列存储系统的阵列控制器和文件引擎基于统一硬件平台开发，可同时提供 IP SAN、FC SAN 和 NAS 组网，并支持 iSCSI、FCP、NFS、CIFS、HTTP、FTP 协议硬件形态统一。如**错误！未找到引用源。**所示。

图表 3-1 T 系列协议与组网的融合统一



表格 3-1 T 系列协议规格

协议名称	规格项	规格说明
FC 协议	协议支持	FCP、FC-SW、FC-PH、FC-PI
	中断聚合	支持，默认关闭
	端口自适应（速率/拓扑）	速率：8Gb/4Gb/2Gb 拓扑：Fabric/Loop/P2P
iSCSI 协议	协议支持	IPv4、IPv6
	端口自适应（速率/拓扑）	速率：1Gbps，10Gbps
	iSCSI CHAP 认证	单向 CHAP 认证，主机发起
	端口聚合类型	动态链路汇聚(IEEE802.3ad)
	巨型帧	支持，配置 MTU 大小范围为 1500～9216（bit）
CIFS 协议	协议支持	SMB1.0

	共享类型	Homedir、normal
	Normal 共享个数	256
	Homedir 共享文件系统个数	16
	Homedir 共享链接数	3000
	Homedir 共享活动链接数	800
NFS 协议	协议支持	V2、V3
	共享链接数	800
FTP 协议	本地用户数量	1000
	共享链接数	800
Http 协议	协议支持	V1.0

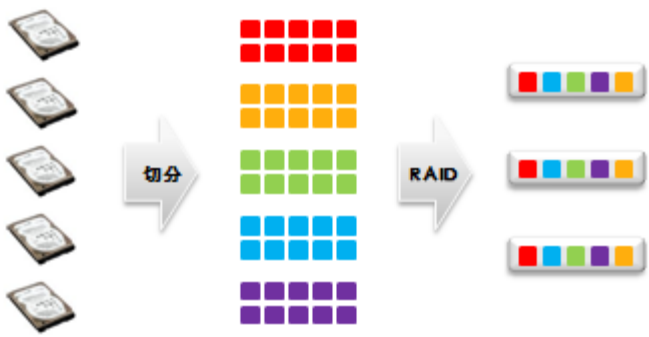
3.2 存储资源的统一管理

OceanStor T 系列存储系统结合专用的存储操作系统，针对传统 RAID 的缺点，设计的一种满足存储技术虚拟化架构发展趋势的全新的 RAID 技术（RAID2.0+）。该技术变传统固定管理模式为两层虚拟化管理模式，在底层块级虚拟化（Virtual for Disk）硬盘管理的基础之上，实现了上层虚拟化（Virtual for Pool）的高效资源管理。

RAID2.0+采用底层硬盘管理和上层资源管理两层虚拟化管理模式，在系统内部，每个硬盘空间被划分成一个个小粒度的数据块，基于数据块来构建 RAID 组，使得数据均匀地分布到存储池的所有硬盘上，同时，以数据块为单元来进行资源管理，大大提高了资源管理的效率。

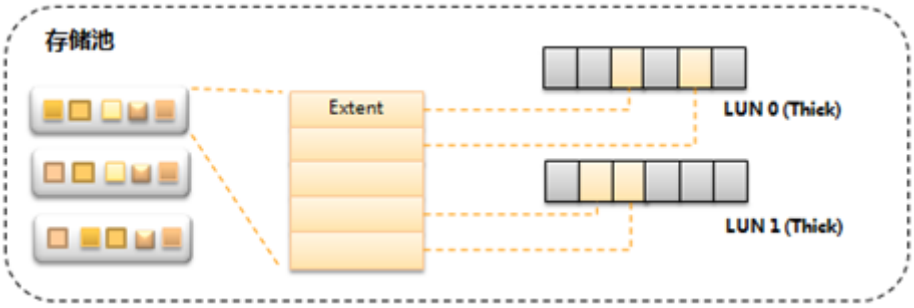
1) T 系列存储系统支持三种不同类型（SSD、SAS 和 NL-SAS）的硬盘，每种类型的盘可以组成一个 Tier 层。在 Tier 层中，每个硬盘被切分成固定大小的数据块（Chunk，也叫 CK），每个 Chunk 的大小为 64MB。T 系列存储系统通过随机算法，将不同硬盘的 Chunk（CK）按照 RAID 算法组成 Chunk Group（CKG）；

图表 3-2 Chunk 和 CKG 组织方式



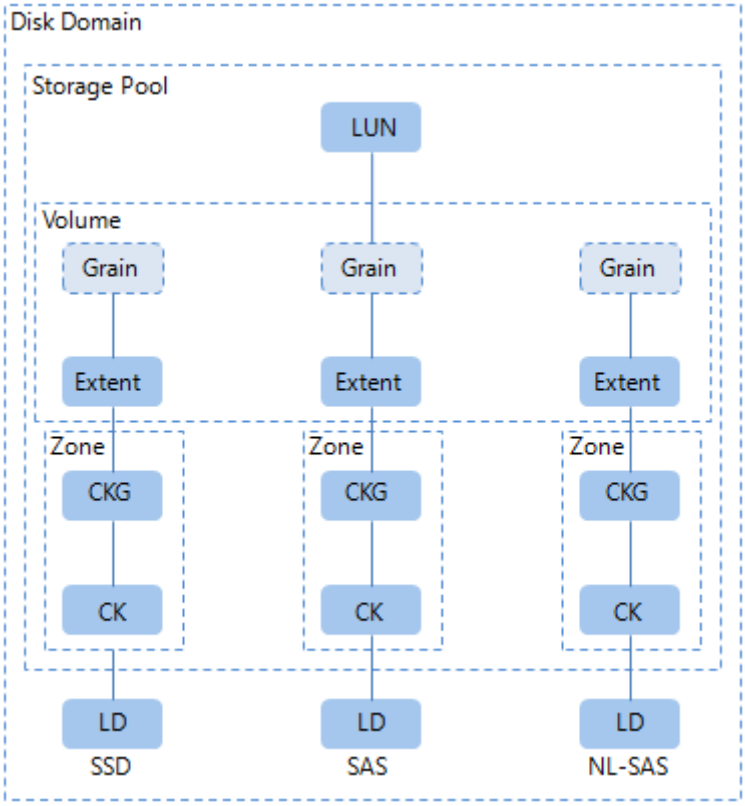
2) CKG 被划分为固定大小的逻辑存储空间（Extent），每个 Extent 的大小可以在 1MB、2MB、4MB、8MB、16MB、32MB 和 64MB 中任选其一，默认为 4MB。Extent 是构成 LUN 的基本单位。

图表 3-3 LUN 的数据结构



RAID2.0+的实现框架如图表 3-4 所示。

图表 3-4 RAID2.0+数据结构



- 硬盘域（Disk Domain）由一个或多个层级的硬盘组成，不同层级支持不同类型的硬盘：构成高性能层的 SSD 硬盘，构成性能层的 SAS 硬盘和构成容量层的 NL-SAS 硬盘。
- 各存储层的硬盘被划分为 64MB 固定大小的 Chunk（CK）。
- 每一个存储层的 Chunk（CK）按照用户设置的“RAID 策略”来组成 Chunk Group（CKG），用户可以为存储池（Storage Pool）中的每一个存储层分别设置“RAID 策略”。

- Chunk Group（CKG）将会被切分为更小的 Extent。Extent 作为数据迁移的最小粒度和构成 Thick LUN 的基本单位，在创建存储池（Storage Pool）时可以在“高级”选项中进行设置，默认 4MB。

  若干 Extent 组成了卷（Volume），卷（Volume）对外体现为主机访问的 LUN（这里的 LUN 为 Thick LUN）。在处理用户的读写请求以及进行数据迁移时，LUN 向存储系统申请空间、释放空间、迁移数据都是以 Extent 为单位进行的。例如：用户在创建 LUN 时，可以指定容量从某一个存储层中获得，此时 LUN 由指定的某一个存储层上的 Extent 组成。在用户的业务开始运行后，存储系统会根据用户设定的迁移策略，对访问频繁的数据以及较少被访问的数据在存储层之间进行迁移（此功能需要购买 SmartTier License）。此时，LUN 上的数据就会以 Extent 为单位分布到存储池的各个存储层上。

- 在用户创建 Thin LUN 时，存储系统还会在 Extent 的基础上再进行更细粒度的划分（Grain），并以 Grain 为单位映射到 Thin LUN，从而实现对存储容量的精细化管理。

虚拟池化设计，降低存储规划管理难度

目前主流存储系统拥有成百，甚至上千块不同类型的硬盘已经非常普遍，如果使用传统 RAID 技术，对于管理员来说，意味着不仅需要管理数量众多的 RAID 组，而且需要针对每一个应用，对每一个 RAID 组进行周密的性能、容量规划，在当今这样一个变化迅速的时代，要作到准确预估 IT 系统生命周期内业务的发展趋势以及与之对应的数据增长量级几乎是一项不可能实现的目标，这使得管理员不得不经常面临存储资源分配不均等一系列管理问题，大大增加了管理的复杂度。

使用 RAID2.0+技术的 T 系列存储系统，采用了领先的虚拟化技术，对存储资源进行池化设计，管理员只需要维护少量的存储资源池，所有的 RAID 配置在创建存储池时自动配置完成，同时，系统会自动根据制定的策略来智能管理和调度系统资源，大大降低了规划和管理的难度。

增加 LUN 所跨硬盘数，大幅提升单 LUN 性能

服务器计算能力的不断发展和越来越多的主机应用（数据库、虚拟机等）对存储的性能、容量、灵活性都提出了更高的要求，传统 RAID 组受到硬盘数的限制，容量小、性能差且难以扩展，已经越来越无法满足业务的需求。当主机对一个 LUN 进行密集访问时，只能访问到有限的几个磁盘，容易造成磁盘访问瓶颈，导致磁盘热点。

RAID2.0+技术支持由几十甚至上百块硬盘组成一个大的存储资源池，LUN 基于存储池创建，不再受限于 RAID 组磁盘数量，宽条带化技术能够让单个 LUN 上的数据分布到很多不同的磁盘上，避免了磁盘热点，使得单 LUN 性能和容量都得到了大幅提升。如果当前存储的容量无法满足要求时，只需要简单向硬盘域中增加硬盘就可以完成存储池和 LUN 的动态扩容，提升了磁盘的容量利用率。

空间动态分布，灵活适应业务变化

RAID2.0+基于业界领先的块虚拟化技术实现，卷上的数据和业务负荷会自动均匀分布到存储池所有的物理硬盘上，借助于智能的 Smart 系列效率提升套件，T 系列存储系统能自动根据业务所需的性能、容量、冷热数据等因素在后台进行智能调配，灵活地适应企业业务的快速变化。

4 融易：横向的资源扩展

T 系列存储系统采用领先的 TurboModule 技术，具备出色的可扩展性。TurboModule 包括 3 个技术：模块热插拔，前后端 I/O 模块灵活配比以及高密度 I/O 模块设计和高密度接口。

- **模块热插拔：**T 系列存储系统支持全冗余的硬件设计，控制器、电源、风扇、一体化 BBU、硬盘、I/O 模块等冗余部件均支持在线热插拔。

  其中 I/O 模块的热插拔技术是 T 系列在高扩展方面最独特的设计。T 系列产品可根据业务的增加，在线扩展 I/O 模块，不需另外增加交换设备即可增加可用端口的数量，有效降低成本；在维护方面，如果 I/O 模块出现故障，可在不中断业务的情况下进行在线更换，保证系统的可靠性以及业务的连续性。

- **前后端口 I/O 模块灵活配比：**T 系列每控制器支持最多达 6 块 I/O 模块，可以根据业务类型灵活选择前后端的模块配比。

- **高密度 I/O 模块与高密度接口设计：**T 系列支持 4Gb FC、8Gb FC、1GE、10GE、4*6Gb SAS 五种接口类型；前后端接口数最高可达 64 个，可以最大程度节省用户前期采购成本和后期维护成本。

此外，T 系列支持 2.5 寸和 3.5 寸两种规格的硬盘，支持 SAS、NL SAS 和 SSD，以及 2U、4U 两种不同规格的硬盘框，可根据不同业务进行最优选择。

随着社会的进步和业务的发展，不断增加、累积的企业数据对存储系统提出了更高的要求。传统的双控架构的中端存储往往无法跟上其数据增长的步伐，出现存储性能成瓶颈。T 系列存储系统秉承弹性高效的设计理念，允许双引擎（每个引擎包括两个控制器）的横向扩展，为企业业务提供强大的性能支撑。

T 系列存储系统每引擎内的两个控制器之间通过 8 lane PCIe2.0 镜像通道互连，引擎间的每个控制器使用 QSFP 光缆分别与 2 个冗余的 DSW（Data Switch，数据交换单元）交换平面通过 4 lane PCIe2.0 进行互连，在 DSW 中实现数据交换。成熟的 PCIe2.0 全交换互连架构为 T 系列存储系统多个控制器之间提供了无阻塞的专用网络，使得每个控制器都能够访问系统范围内的所有资源；同时，由于减少了协议转换的延时，可以实现更加高效的数据交换。

PCI-E 数据交换机具有高带宽、低延迟的特点，是各个引擎之间相互连接和通信、实现控制器间控制信息流和业务数据流交换的关键设备。扩展时采用两台数据交换机，两者采用 Active-Active 的工作模式，提高了整个控制信息流和业务数据流的交换带宽和可靠性。

图表 4-1 PCI-E 数据交换机示意图

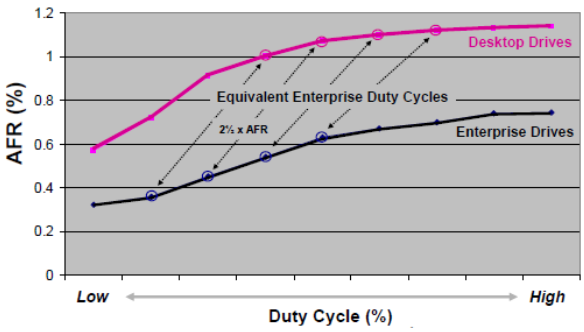
5 可靠：降低故障，快速自愈

5.1 降低故障

自动负载均衡，降低整体故障率

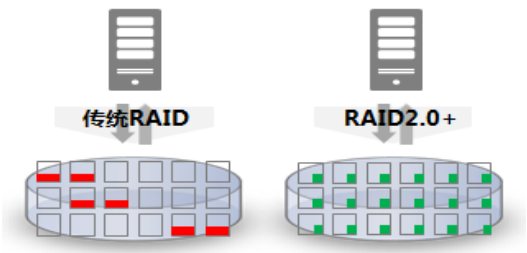
传统 RAID 存储系统中一般会有多个 RAID 组，每个 RAID 组中包含几块到十几块硬盘。由于每个 RAID 组的业务繁忙程度不同，导致硬盘的工作压力不均，部分硬盘存在热点，根据 SNIA 的统计数据，热点盘的故障率会明显增高。如图表 5-1 所示，其中 Duty-Cycle 即忙闲度，指的是硬盘工作时间占总上电时间的比例，AFR 为硬盘年故障率，不难看出，Duty-Cycle 高时的硬盘年故障率几乎是低时的 1.5～2 倍。

图表 5-1 磁盘忙闲度与故障率的关系



RAID2.0+技术通过块虚拟化实现了数据在存储池中硬盘上的自动均衡分布，避免了硬盘的冷热不均，从而降低了存储系统整体的故障率。

图表 5-2 RAID2.0+技术下的数据分布



专利设计保障，增强系统适应性

在结构件抗振设计领域，T 系列存储系统从硬盘、风扇、机箱和滑道等多个领域采用了领先设计，保证产品通过国家信息产业通信设备抗震性能质量监督检验中心 9 烈度抗震权威认证，成为唯一达到《电信设备抗地震性能检测规范》YD5083 最高抗震等级要求的专业存储系统：

1） 硬盘单元的振动隔离：托架内侧增加粘弹性材料吸收硬盘自身旋转振动能量；紧固螺钉处的粘弹性垫圈有效隔离外界线性振动能量；控制相邻磁盘转向降低磁盘谐振（华为发明专利：中国 200910221868.3）；

2） 风扇振动的多级隔离：热塑粘弹性材料风扇安装钉，覆盖硬盘敏感振动频率；风扇、支架与机箱之间的垂直、水平多级减振，降低 40%以上来自于风扇的振动；

3） 高强度机箱及硬盘滑道设计：机箱硬盘部位的双层式结构，强度增强 20%以上，有效保证各硬盘槽位尺寸一致性；压铸锌基合金的滑道材料，良好的耐冲击性降低机箱向硬盘振动的传递放大作用。

在硬件防腐蚀设计领域，T 系列存储系统联合多个供应商，在系统多个模块采取了防腐蚀工艺，保证系统满足在数据中心空气污染物等级 DC G1~GX 级别全场景下的正常运行：

1） 联合硬盘厂商开发出硬盘防腐蚀工艺：ENIG/SPV（化镍浸金/锡封），有效提升了硬盘在污染环境中的寿命和可靠性；

2） 通过防腐蚀工艺结合温升、电压分布设计进行局部防护，有效提升了控制器在污染环境中的寿命和可靠性；

3） 通过在线腐蚀监控设备（华为发明专利：中国 201210519754.9）提前预警数据中心腐蚀风险，并可通过专业测试设备快速(72hr)量化数据中心腐蚀等级；

4） 通过机框加装防腐蚀过滤器（华为发明专利：中国 201110314472.0）解决机框级腐蚀问题；通过化学过滤方案设计解决机房级腐蚀问题。

5.2 快速自愈

故障自检自愈，保证系统可靠性

T 系列存储系统针对硬盘采用了多重故障容错设计，具有硬盘在线诊断、DHA（Disk Health Analyzer，硬盘故障诊断与预警）、坏道后台扫描、坏道修复等多种可靠性保障，RAID2.0+技术会根据热备策略自动在硬盘域中预留一定数量的热备空间，用户无需进行设置，当系统自动检测到硬盘上某个区域不可修复的介质错误或整个硬盘发生故障时，系统会自动进行重构，将受影响的数据块数据快速重构到其他硬盘的热备空间中，实现系统的快速自愈合。

1） DHA（Disk Health Analyzer）硬盘故障诊断与预警：硬盘作为存储系统中一个重要的机械部件，经过长时间的不间断工作运行后会出现部件老化，故障率会随着时间呈上升趋势。T 系列存储系统的硬盘健康度分析子系统通过建立硬盘故障模型，对整个系统的硬盘关键指标进行监控，利用先进的算法，评估运行硬盘的健康度，根据硬盘的应用场景，设置合适的阈值。当硬盘的健康度值低于既定阈值时，进行硬盘替换，做到风险提前预防；

2） 坏道后台扫描：硬盘在正常读写的过程中，会出现某些扇区数据不可读的现象，这些不可读的扇区就是常说的坏道 LSE（Latent Sector Errors，潜在扇区错误）。当硬盘的扇区出现坏道时，硬盘不会自动将这些坏扇区信息告知主机进行修复，只有在读写的时候才能发现坏道。T 系列存储系统的硬盘坏道后台扫描可以在不影响业务和硬盘自身可靠性的前提下，根据硬盘的物理参数，在一定的扫描周期内，设置合适的扫描策略，在最短的时候发现 LSE 并进行修复，从而降低数据丢失的风险；

3） 硬盘坏道修复：坏道修复可以在发现坏道后，对其进行读重试或利用 RAID 重构出该坏道对应扇区的数据并下发写命令，进而利用硬盘自身的重映射功能恢复，可以避免全盘失效和重构。

快速重构，改善双盘失效率

纵观近 10 年硬盘的发展，其容量的增长远远快于性能的进步，现在 4TB 的高容量磁盘在当前的企业和消费市场已经非常普遍，而 5TB 的高容量磁盘也将在 2014 年 Q2 会出现，即便是专门针对企业市场的高性能 SAS 磁盘，也已经达到了 1.2TB 的容量。

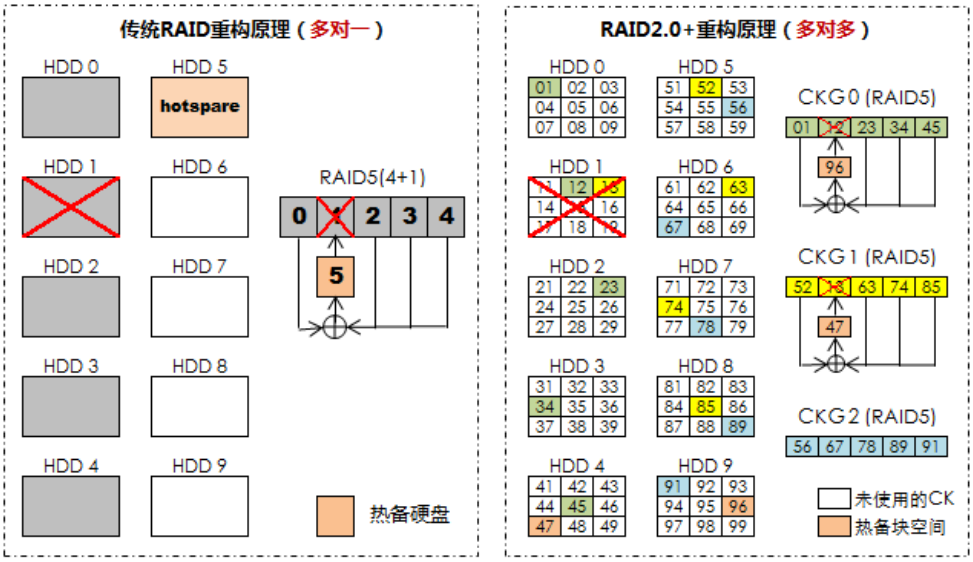
容量的增长使得传统 RAID 不得不面临一个严重的问题：10 年前重构一块硬盘可能只需要几十分钟，而如今重构一块硬盘需要十几甚至几十个小时。越来越长的重构时间使得企业的存储系统在出现硬盘故障时长时间处于非容错的降级状态，存在极大的数据丢失风险，存储系统在重构过程中由于业务和重构的双重压力导致数据丢失的案例也屡见不鲜。

基于底层块级虚拟化的 RAID2.0+技术由于克服了传统 RAID 重构的目标盘（热备盘）性能瓶颈，使得重构数据流的写带宽不再成为重构速度的瓶颈，从而大大提升了重构速度，降低了双盘失效的概率，提升了存储系统的可靠性。

图表 5-3 是传统 RAID 和 RAID2.0+两种技术重构原理的对比：

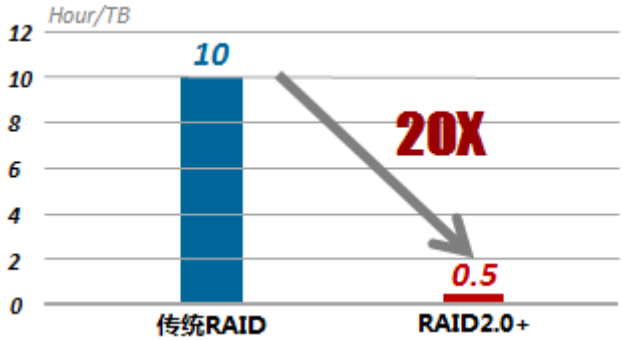
- 左图传统 RAID 中，HDD0～HDD4 五块硬盘创建 RAID5，HDD5 为热备盘，当 HDD1 故障后，HDD0、HDD2、HDD3、HDD4 通过异或算法将重构的数据写入 HDD5 中；
- 在右图的 RAID2.0+示意图中，当 HDD1 故障后，故障盘 HDD1 中的数据按照 CK 的粒度进行重构，只重构已分配使用的 CK（图中 HDD1 的CK12和CK13），存储池中所有的硬盘都参与重构过程，重构的数据分布在多块硬盘中（图中的 HDD4 和 HDD9）

图表 5-3 传统 RAID 和 RAID2.0+重构原理对比



由于 RAID2.0+技术在重构方面的巨大优势，使得 T 系列存储系统在重构方面与传统阵列相比具有明显的优势，是采用传统 RAID 的存储系统与采用 RAID2.0+的 T 系列存储系统在采用 NL-SAS 高容量磁盘环境中重构 1TB 数据所需时间的对比。

图表 5-4 传统 RAID 和 RAID2.0+重构时间对比



6 可靠：秒级的容灾周期

T 系列存储系统异步远程复制采用了创新的多时间戳（Timestamp）缓存技术，缓存中的数据和与缓存交互的 IO 都携带时间戳（Timestamp）信息，在进行复制和同步时，直接从主 LUN Cache 中读取相应时间戳分片的数据复制到从 LUN，降低了时延，同时降低了传统异步远程复制快照对性能的影响，因此同步周期可以缩短为秒级。

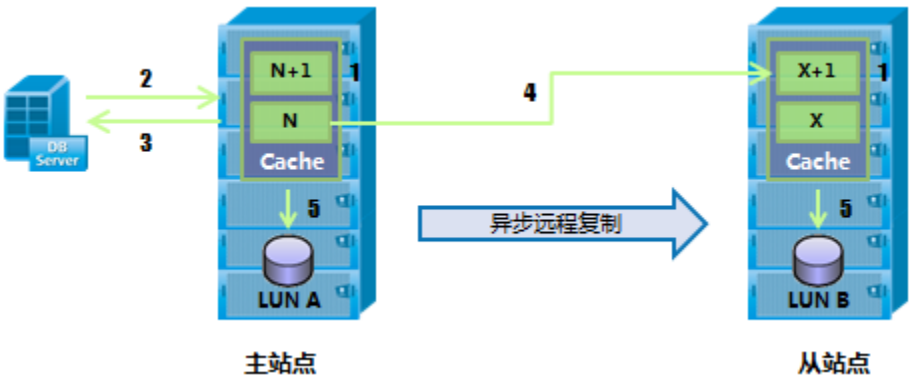
由于异步远程复制主 LUN 上的数据更新不是立即同步到从 LUN 的，所以 RPO 取决于用户设置的同步周期，用户可以根据应用场景设置不同的同步周期（范围是 3s～1440min）。其实现原理如下：

1) 当主站点的主 LUN 和远端复制站点的从 LUN 建立异步远程复制关系以后，会启动一个初始同步，将主 LUN 数据全量拷贝到从 LUN；

2) 如果在初始同步时主 LUN 收到生产主机写请求，只会将数据写入主 LUN；

3) 初始同步完成后，从 LUN 数据状态变为一致，然后开始按照下面的流程进行 I/O 处理：

图表 6-1 多时间戳缓存技术下的异步远程复制



① 每当间隔一个同步周期（由用户设定，范围为 3s～1440min），系统会自动启动一个将主站点数据增量同步到从站点的同步过程（如果同步类型为手动，则需要用户来触发同步）。每个复制周期启动时在主 LUN（LUN A）和从 LUN（LUN B）的缓存中产生新的时间戳分片（TP_{N+1} 和 TP_{X+1}）；

② 主站点接收生产主机写请求；

③ 主站点将写请求的数据写入缓存时间戳为 TP_{N+1} 的分片中，立即响应主机写完成；

④ 同步数据时，读取前一个周期主 LUN（LUN A）缓存中时间戳为 TP_N 分片的数据，复制写入从 LUN（LUN B）缓存中时间戳为 TP_{X+1} 的分片中；

⑤ 同步数据完成后，按照刷盘策略将主 LUN（LUN A）和从 LUN（LUN B）缓存中时间戳为 TP_N 和 TP_{X+1} 分片的数据下盘，等待下一个同步的到来。

说明

- ✓ 时间戳分片：在缓存中管理一段时间内写入数据的逻辑空间（数据大小没有限定）
- ✓ 在低 RPO 的应用场景下，异步远程复制周期很短，T 系列存储系统的缓存中能缓存多个时间戳分片中的全部数据；如果主机业务带宽或容灾带宽出现异常或故障，造成复制周期变长或中断，此时缓存中的数据会按照刷盘策略自动刷盘并进行一致性保护，复制时再从盘上进行读取。

支持镜像分裂、主从切换和故障快速恢复

异步远程复制拥有分裂、同步、主从切换和断开后恢复的功能。

分裂以后的异步远程复制，不会再进行周期性的同步，直到用户手动进行“同步”操作，然后按照制定好的同步策略（手动或自动）进行同步。

异步远程复制提供三种数据同步的方式（同步类型）供用户选择：

- ➢ 手动：用户需要手动进行主 LUN 和从 LUN 的数据同步。选择手动同步时，用户可以根据自己的意愿将数据更新到从 LUN，以此来决定从 LUN 的数据是哪一个时间点上主 LUN 的副本。
- ➢ 同步开始后定时等待：启动同步时开始计时，等待一个同步周期后再次启动同步并计时，即：在最近一次同步操作开始时，经过用户设置的“定时时长”，自动进行主 LUN 和从 LUN 的数据同步。
- ➢ 同步完成后定时等待：上一次同步完成以后再进行下一次同步周期的计时，即：在最近一次同步操作完成后，经过用户设置的“定时时长”，自动进行主 LUN 和从 LUN 的数据同步。

三种不同的同步类型应用于不同的场合，用户可以根据具体情况进行选择。

从 LUN 数据持续保护

异步远程复制支持对从 LUN 数据的持续保护，在从站点，主机对从 LUN 的读、写有权限控制，实现从 LUN 数据的完全保护。当同步中断时，可以将前一个 TP_X 周期的数据恢复到从 LUN，覆盖第 TP_{X+1} 个周期的数据，使从 LUN（LUN B）回退到最近一次同步开始前时间点的可用数据。

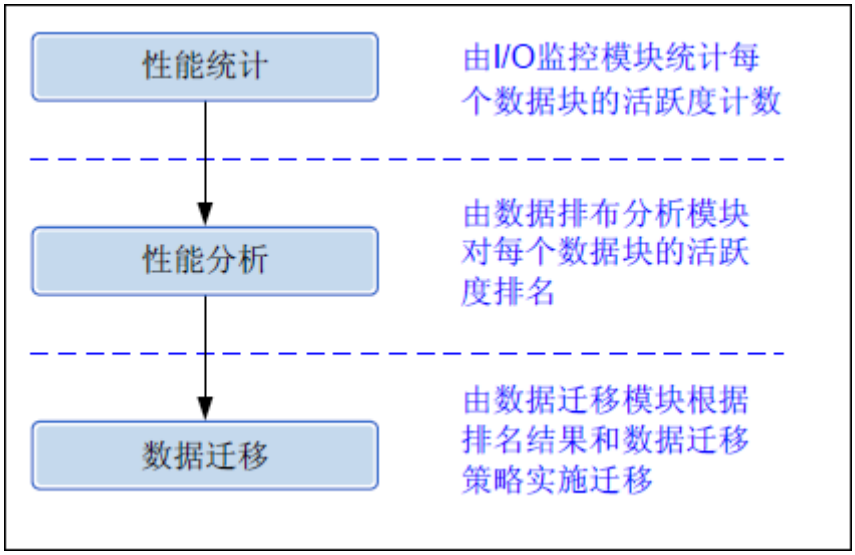
异步远程复制进行主从切换后，根据原从 LUN 数据的可用情况，决定是否对原从 LUN 进行数据恢复：如果原从 LUN 数据本身是可用的，则不必对原从 LUN 回滚；如果原从 LUN 数据不可用，则启动数据恢复，使原从 LUN 回退到最近一次同步开始前时间点的可用数据。整个恢复过程是在后台进行的，完成后会提示用户数据恢复完成。

7 智能：自动数据分级

T 系列存储系统支持华为公司自主研发的自动分级存储特性，简称 SmartTier。简单来说，SmartTier 在合适的时间，将合适的数据放置到合适的地方。SmartTier 提高存储系统性能并降低用户成本,满足企业对性能和容量的双重需求，避免历史数据占用昂贵的存储介质，保证企业有效投入，消除无用容量带来的能耗开销，降低企业 TCO，得到最优性价比。

SmartTier 进行 LUN 级别的智能化数据存放管理，将 LUN 数据按照 512KB~64MB 的粒度划分，该粒度被称为“数据迁移粒度 extent”。SmartTier 以“extent”为单位，统计和分析数据的活跃度，将不同活跃度的数据和不同特点的存储介质动态匹配，并通过数据迁移将活跃度高的“繁忙”数据迁移至具有更高性能的存储介质（如 SSD 硬盘），将活跃度低的“空闲”数据迁移至具有更高容量且更低容量成本的存储介质（如 NL-SAS 硬盘）。SmartTier 经历的性能统计、性能分析、数据迁移三个阶段，如图表 7-1 所示：

图表 7-1 SmartTier 处理数据的三个阶段



其中，性能统计、性能分析阶段，根据用户配置由存储系统自动完成，数据迁移阶段通过用户手动触发或根据用户配置的定时策略触发。

性能统计由存储系统的 I/O 监控模块完成。

SmartTier 允许自定义 IO 监控的时间段，在指定的时间段内，对写入和读取的数据进行统计。随着数据生命周期的推移，数据的活跃度会发生变化，存储系统根据两个 extent 的活跃度来判断一个数据块比另一个更“热”或更“冷”。每个 extent 的活跃度通过统计数据块的性能指标得出。

具体实现原理如下：

1. 在 IO 监控时段内，对每个下发的 IO 都会进行记录，为性能分析和性能预测提供用于分析的数据源。记录的信息以 extent 为单位，包括：读写访问频率、I/O 大小，I/O 顺序度等信息。
2. IO 监控模块采用全内存方案实时记录 extent 的 IO 访问情况，每个存储控制器最大可以监控 512TB 的存储池空间。
3. I/O 监控模块按天进行 IO 监控统计信息的加权操作，弱化历史业务对当前业务的影响。

性能分析由存储系统的数据排布分析模块完成。

性能分析阶段使用性能统计数据进行性能分析，分析结果会对存储池中的每个 extent 排名，排名由高至低，从同一个存储池中的最热 extent 开始，直到最冷 extent（排名仅在同一个存储池中进行），最终生成数据迁移方案。SmartTier 在数据迁移之前根据最近一次生成的数据迁移方案决定 extent 的迁移方向。

具体实现原理如下：

1. 以 I/O 监控模块生成的每个 extent 的性能统计信息作为输入，根据 pool 中各个 Tier 的容量，按照数据块的热度确定出每个 tier 放置 extent 的 IO 计数阈值（确定阈值时，按照最热的数据块放在最高性能层级的原则进行）。
2. 对大于阈值的 extent 进行排序，选出最热的 extent 优先进行迁移。
3. 在数据排布时，针对 SSD 的性能属性制定了相应的策略，对于 SSD 中变顺序的 extent 制定了主动下迁到 HDD 的策略。

数据迁移由存储系统的数据迁移模块完成。

数据迁移实现存储系统冷热数据的重新分布，使得随机热点数据尽可能多的分布在高性能层和性能层，冷点数据和顺序度高的数据分布在容量层，在满足业务性能需求的前提下，最大程度降低存储系统 TCO，为用户节约成本。

SmartTier 有 2 种迁移触发模式：手动触发迁移模式和定时触发迁移模式。手动触发的优先级高于自动迁移。手动触发迁移模式可以根据需要立即触发迁移，定时触发迁移模式是根据事先设定好的迁移开始时间和持续迁移时段来自动触发迁移，可以事先设置每周的哪些天的什么时间开始触发迁移及迁移允许持续的时长。

此外，SmartTier 可设置高中低三档迁移速度，进行迁移的动态控速。其中，低档迁移速度上限为 10MB/S，中档迁移速度上限为 20MB/S，高档迁移速度上限为 100MB/S。

基本原理：

1、 数据迁移模块根据迁移策略启动数据迁移。在用户定义的迁移时间段内，自动完成数据迁移；
2、 数据迁移模块根据数据排布分析模块生成的数据迁移方案，把数据按迁移粒度在不同的存储层之间移动，最终达到用户数据按冷热程度、顺序随机度重新排布的目的；
3、 数据迁移模块根据当前存储池的负载和用户设置的数据迁移速率进行迁移的动态控速；

4、 数据迁移时，extent 是迁移的最小单位，迁移过程中不影响业务数据的访问。每个 extent 的迁移，是从源 extent 中读取数据写入到目的 extent 的过程。在迁移过程中，读 IO 访问会从源 extent 中读取数据，写 IO 会同时写源 extent 和目的 extent。在迁移完成后会修改源 extent 和目的 extent 的元数据，修改完成后读写 IO 会访问目的 extent，源 extent 将被释放。

8 智能：自动精简配置

T 系列存储系统支持华为公司自主研发的自动精简配置特性，简称 SmartThin。在创建 LUN 的时候由用户选择分配一定的容量，在使用过程中采用“按需分配”的存储空间分配策略，提高存储资源使用效率，更大限度满足业务的实际要求。SmartThin 不会预先分配空间，而是将大于物理存储空间的容量形态呈现给用户，使用户看到的存储空间远远大于系统实际分配的空间。用户对这部分空间的使用实现“按需分配”的原则，即：用多少提供多少。如果用户的存储空间不足，可通过扩充后端存储单元的方式来进行系统扩容，整个扩容过程无需系统停机，对用户完全透明。

当出现数据容量超过预期的情况时，可以动态调整该 LUN 的空间。未使用的空间作为公共的空间可以分配给任何需要空间的 LUN。这样，不存在私有的一直不能被使用到的空间，提高了利用率和效能比。同时，动态空间调整提供了在线调整 LUN 空间大小的能力，可以做到扩容的同时不影响业务。

SmartThin 基于 RAID2.0+存储虚拟资源池创建 Thin LUN，即 Thin LUN 和传统的 Thick LUN 共存于同一个存储资源池中。精简 LUN（Thin LUN）是在精简池中创建的并可以映射为主机直接访问的逻辑单元。Thin LUN 的容量大小并不是实际的物理空间，而是一个虚拟值，只有在对 Thin LUN 进行真正 IO 读写时，才通过写时分配的策略从存储资源池中申请物理空间。

SmartThin 允许主机可感知容量大于 Thin LUN 实际存储空间。主机可感知容量指的是用户能够创建的 Thin LUN 大小，也即是 Thin LUN 创建成功后映射给主机，在主机侧显示的卷容量（逻辑虚拟空间）的大小；Thin LUN 实际存储空间指的是 Thin LUN 真实占用的存储池物理空间的大小。SmartThin 会向主机隐藏 Thin LUN 实际存储空间大小，而向主机提供 Thin LUN 的名义存储空间。

除此之外，SmartThin 支持创建大于存储池最大物理可用空间的 Thin LUN。例如存储池提供的最大物理空间是 2TB，但 SmartThin 支持创建大于 10TB 的 Thin LUN。

SmartThin 主要通过 Capacity-on-write 和 Direct-on-time 两种技术来响应主机对 Thin LUN 的读写操作。利用 Capacity-on-write 来进行写时空间分配，再使用 Direct-on-time 技术来进行读写重定向。

1. Capacity-on-write

当 Thin LUN 接收到主机写 IO 请求，首先会通过 direct-on-time 技术判断该写 IO 请求的逻辑存储区域是否已经分配了实际存储区域，如果尚未分配就会触发空间分配，分配的最小粒度叫 Grain, Grain 大小为 64KB，然后将数据写入到新分配的实际存储区域中。

2. Direct-on-time

由于采用了 Capacity-on-write 技术，数据的实际存储区域和逻辑存储区域的关系不再是按照确定的公式可以固定不变计算出来的，而是按照写时分配的原则随机映射确定的。所以在对 Thin LUN 进行读写时需要重定向实际存储区域和逻辑存储区域的关系，重定向依赖于映射表。映射表的主要作用是用来记录实际存储区域和逻辑存储区域的映射关系。在写过程中动态更新映射表，在读过程中查询映射表。因此，Direct-on-time 重定向操作也就分为读重定向和写重定向。

读重定向：Thin LUN 接收到主机读 IO 请求后，先查询映射表，如果该读 IO 的逻辑存储区域已分配对应的实际存储区域，则将该读 IO 的逻辑存储区域重定向到实际存储区域，然后从实际存储区域中读取到数据后，将该数据返回给主机；如果该读 IO 的逻辑存储区域尚未分配空间，则将该逻辑存储区域的数据置为全 0 返回给主机。

写重定向：Thin LUN 接收到主机写 IO 请求后，先查询映射表，如果该写 IO 的逻辑存储区域已分配对应的实际存储区域，则将该写 IO 的逻辑存储区域重定向到实际存储区域，然后将数据写入到实际存储区域中，并返回写成功给主机；如果该写 IO 的逻辑存储区域尚未分配空间，则通过 Capacity-on-write 技术操作。

SmartThin 支持分别针对单个 Thin LUN 和存储资源池的在线扩容，两种扩容方式均不会影响主机业务。

单个 Thin LUN 的扩容即增大 Thin LUN 的名义存储空间大小。在修改 Thin LUN 的名义存储空间大小后，SmartThin 会自动向主机提供新的 Thin LUN 名义存储空间大小。这样，在主机侧显示的卷容量（逻辑虚拟空间）的大小就是扩容后的大小。整个扩容过程中不涉及原有存储区域的调整，新写入的数据如果需要存储到新增的 Thin LUN 存储空间中，也通过 SmartThin 的写时分配机制从存储资源池中申请实际存储空间。

存储资源池的扩容是 RAID2.0+存储虚拟化技术本身提供的能力，可以在不影响主机业务的情况下增大存储空间容量，同时还通过 SmartMotion 软件功能将数据在整个存储池中的磁盘（包括新加入磁盘）上重新均衡。

SmartThin 功能支持标准 SCSI 命令(unmap)和零数据释放两种空间回收方式。两种方式的实现原理如下：

标准 SCSI 命令空间回收方式：在删除虚拟机等场景下，主机通过 SCSI 协议下发 unmap 释放命令，SmartThin 收到该命令后，通过 direct-on-time 查找到 Thin LUN 上需要释放的逻辑存储区域对应的实际存储空间，然后将该实际存储空间从 Thin LUN 中释放回存储池中，同时从映射表中删除相关映射记录；该空间回收方式需要主机的应用支持下发 unmap 命令（VMware，SF 和 Windows 2012 均支持该 SCSI 命令）。

零数据释放空间回收方式：当 SmartThin 接收到主机写 IO 的请求后，会判断该写 IO 请求中包含的数据块是否是全零，如果下发全零数据段的逻辑存储区域尚未分配实际存储区域，那么 SmartThin 直接返回写成功给主机，不再进行空间分配；如果下发全零数据位段的逻辑存储区域已有对应的实际存储区域，那么

SmartThin 会直接释放将该实际存储空间从 Thin LUN 中释放回存储池中，同时从映射表中删除相关映射记录，并返回写成功给主机。该方式不需要主机下发特殊的命令。

9 高效：服务质量保障

T 系列存储系统支持华为公司自主研发的服务质量保障特性，简称 SmartQoS。SmartQoS 能够对存储系统中的计算资源，缓存资源，并发资源以及硬盘资源的智能分配和调节，来满足多种不同重要性业务在同一台存储设备上的不同 QoS 要求。

SmartQoS 特性从以下三个方面来保证数据业务的服务质量：

- **IO 优先级调度技术**：通过区分不同业务的重要性来划分业务响应的优先级。在存储系统为不同业务分配系统资源的时候，优先保证高优先级业务的资源分配请求。在资源紧张的情况下，为高优先级的资源分配较多的资源，以此尽可能保证高优先级业务的服务质量。当前用户可以配置的优先级分为高、中、低三个等级；
- **IO 流量控制技术**：基于传统的令牌桶机制，针对用户设置的性能控制目标（IOPS 或者带宽）进行流量限制，通过 IO 流控机制，限制某些业务由于流量过大而影响其它业务；
- **IO 性能保证技术**：基于流量扼制的方式，允许用户为高优先级业务指定最低性能目标（最小 IOPS/带宽或最大时延），当该业务的最低性能无法保障时，系统内部通过对其他低优先级业务的 IO 逐级增加时延的方式来限制其流量，从而尽力保障该业务的最低性能目标。

IO 优先级调度技术是以存储资源的分配和调度为设计出发点来实现的。存储系统在不同应用场景中性能取决于不同场景下对存储资源的消耗水平，因此只要实现了资源尤其是瓶颈资源的合理调度和分配，就能有效地对系统的性能产生影响。该技术通过监控对性能影响最大的并发，计算，缓存和硬盘四个资源的使用情况，并在出现资源瓶颈时进行资源调度的方式，尽可能满足高优先级的资源需求，较好地解决了关键业务在不同场景下的服务水平保证问题。

SmartQoS 特性的 IO 优先级调度技术主要针对存储系统 IO 路径上关键瓶颈资源进行调度，主要调度的资源包括**并发资源**，**计算资源**，**缓存资源**以及**硬盘资源**。调度策略基于用户配置的 LUN 的优先级来进行，不同的优先级对应不同的调度策略。LUN 的优先级由用户根据部署在该 LUN 上的业务重要性来指定，目前用户可以配置**高**，**中**，**低**三个优先级。

优先级调度通过控制前端主机并发，系统 CPU 资源，阵列内缓存资源，后端硬盘资源四种瓶颈资源的分配来达到控制每个调度对象在存储系统内部的响应时间。

- **前端并发资源**的优先级调度在存储系统前端进行，即针对主机的并发访问来进行控制。由于存储系统对主机的最大并发访问的承载能力是有限的，因此当系统达到最大并发数时，SmartQoS 会根据每个控制器上工作的不同优先级的 LUN 的个数来对每个优先级的最大并发数进行限制，限制的原则是保证优先级高的业务能获得更多并发数，保证业务量更大的业务获得更多并发数；
- **计算资源**的优先级调度主要通过控制 CPU 运行时间的分配来实现。SmartQoS 会根据高、中、低三个优先级各自的权重来分配每个优先级业务占用的 CPU 计算时间，当 CPU 成为系统性能瓶颈时，会通过优先级调度保证高优先级的业务获得更多的 CPU 计算时间；
- **缓存资源**的优先级调度主要通过控制缓存页面资源的分配来实现。SmartQoS 会根据每个优先级的权重来对不同优先级的页面分配请求进行调度，优先满足较高优先级业务的页面分配请求；
- **硬盘资源**的优先级调度主要通过控制 IO 的下盘顺序来实现。SmartQoS 会根据 IO 的优先级，在访问硬盘时让高优先级的 IO 优先访盘。当出现硬盘繁忙，大部分 IO 在硬盘侧出现排队时，通过引入基于硬盘资源的优先级调度机制，可以减小高优先级 IO 的排队时间，总体上减小高优先级 IO 的时延。

SmartQoS 特性的优先级调度技术基于 LUN 的优先级实现。因此每个 LUN 都有一个优先级属性，这个属性由用户配置并保存在数据库中，当一个 IO 从主机（SCSI 目标器）发送到阵列，这个 IO 将会根据其归属的 LUN 来获得这个优先级属性，并且在整个 IO 路径上携带这个优先级信息。

IO 流量控制技术通过限制存储系统中的一个 LUN 或者多个 LUN 的总体 IOPS 或者带宽，来达到限制系统中某些应用的性能，避免这些应用由于突发流量过大，影响系统中其它业务的正常性能。

IO 流量控制技术针对特定 LUN 上的数据业务限制可使用的数据处理资源，流控对象主要有以下两类：首先是需要限制的 IO 类型（限制读、限制写，或者同时限制读和写），其次是需要限制的流量类型（IOPS 或带宽），最终针对特定 LUN 得到一个流控限制的二元组（IO 类型，流量类型）。

SmartQoS 特性的 IO 流量控制技术根据确定好的二元组，通过 IO 分类来实现流量的限制。每个 IO 分类即对应一个流控组，即包含一定数量 LUN，并被设置了最大流量限制的 LUN 组。IO 分类流控的功能通过 IO 分类队列管理，令牌分发和出队控制几部分共同实现。

IO 时延控制技术通过保高限低端的方式来保证某些关键业务的最低性能要求。用户可以为高优先级的业务设置最低性能指标，当此业务的最低性能指标无法达成之后，系统会通过依次限制低优先级和中优先级业务的性能来保障设置了最低性能目标的高优先级业务的性能。

SmartQoS 通过给中、低优先级的业务逐步增加时延的方式来做到对于其性能的限制，为了避免对系统性能产生较大抖动，当逐步增加的时延达到最大时延时，将不再进行增加；同时，当需要保证最低性能的业务性能达到最低性能指标的 1.2 倍时，系统将逐步消除中、低优先级所增加的时延。

10 高效：缓存分区优化

T 系列存储系统支持华为公司自主研发的缓存分区优化特性，简称 SmartPartition。SmartPartition 的核心思想是通过对系统核心资源的分区，保证关键应用的性能。管理员可以针对不同的应用配置不同大小的缓存分区，系统将保证该分区中的缓存资源被该应用独占，并根据业务实际情况实时动态调配不同分区中的前后端并发，从而保证位于该分区的应用性能。SmartPartition 还可以与其他 QoS 技术（如智能服务质量控制 SmartQoS）相配合，从而达到更好的服务质量保证效果。

缓存从类型上分为读缓存和写缓存。读缓存的主要作用是通过读预取及数据保有提高主机读 IO 的命中率；写缓存的主要作用是通过合并、命中、排序等手段提高访盘性能。不同的业务对读写缓存大小有不同的要求，SmartPartition 支持用户单独为分区配置读、写缓存的大小，从而满足不同类型业务的要求。

由于目的不同，配置读写缓存对 IO 的处理流程也是不同的。对写 IO 的影响主要在缓存资源分配阶段，需要做分区主机并发的判断和写缓存数量的判断。选择在这个阶段处理的原因在于这是写过程的最初阶段，且该阶段中并没有实际占用存储缓存。对读 IO 的影响分为两个部分：第一个部分与写 IO 类似，需要判断分区主机并发是否满足，如果不满足，则需要将 IO 返回。由于读缓存的目的是控制读数据占用的缓存大小，而读缓存的大小是由读缓存淘汰过程控制的，所以第二个部分在淘汰过程中控制。如果分区内的读缓存没有达到设定值，此时需要以极慢的速度淘汰；反之，则需要快速淘汰，保证读缓存数量在设定值以内。

相对于主机应用，存储系统的处理资源是有限的。所以，存储需要限制系统内总的主机并发。针对每个分区，当然也需要控制其并发以保证服务质量。

SmartPartition 中分区的主机并发并不是固定的，而是基于多种因素，采用优先级加权算法计算出来的，这些因素包括：

- 上个统计周期内分区中活动的 LUN 的个数
- 上个统计周期内分区中活动的 LUN 的优先级
- 上个统计周期内各 LUN 的完成 IO 个数
- 上个统计周期内各 LUN 由于分区并发已到而返回主机的 IO 个数

通过这些因素的加权，可以实现系统内主机并发最大化利用的同时兼顾分区的服务质量保证。

当一个统计周期结束后，分区可能需要根据新的统计结果调整其并发。这个过程由 SmartPartition 逻辑控制，按一定步长逐步调整，以保证其调整尽量平滑，避免对主机造成强烈的性能波动。

与主机并发控制类似，后端并发控制的目的也是最大化利用系统资源的同时兼顾分区的服务质量保证。这个并发也是多种因素按优先级加权计算出来的，这些因素包括：

- 上个统计周期内分区中各种优先级 LUN 的脏数据数量
- 上个统计周期内分区中 LUN 的刷盘时延
- 上个统计周期内分区中 LUN 的实际刷盘并发

调整周期和调整方式也与主机并发控制类似，不再赘述。

