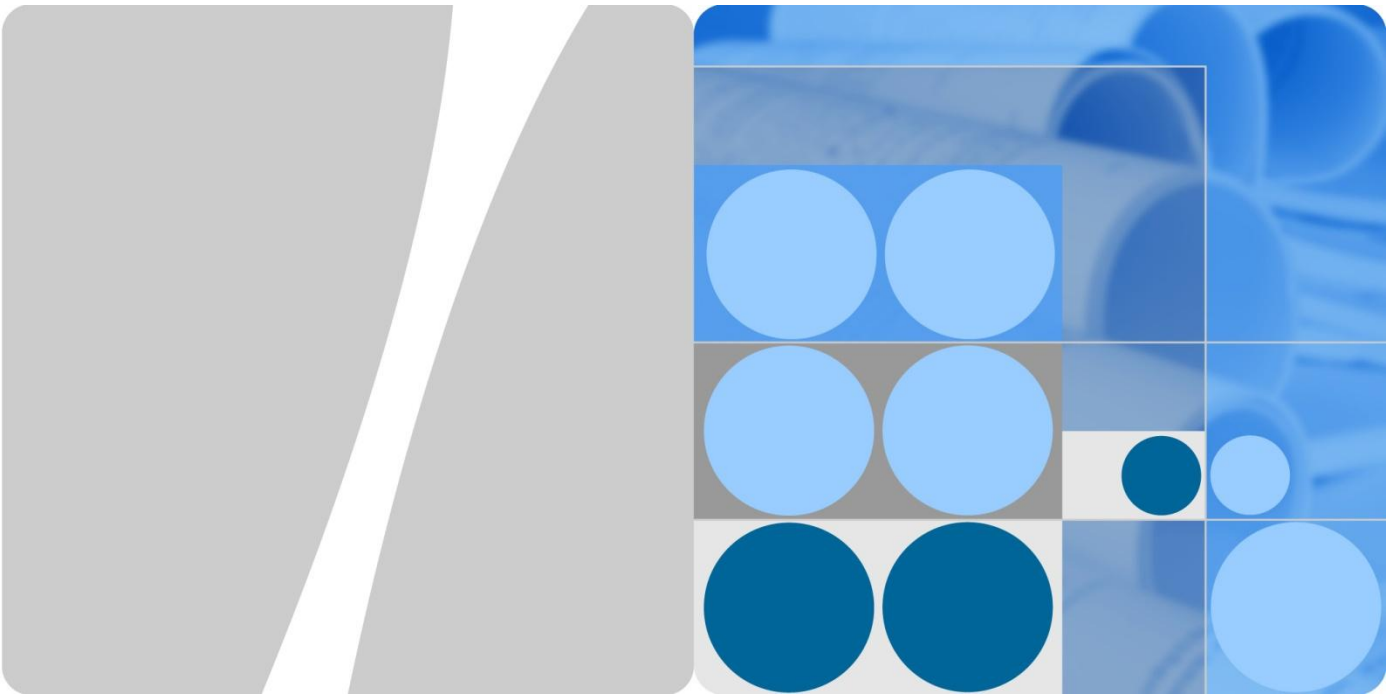
资料编码



OceanStor Dorado 全闪存阵列技术白皮书

文档版本 V1.1

发布日期 2014-03

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

目 录

1 总体概述/Executive Summary

通过对多个企业客户的数据中心进行深入分析和调研，华为公司发现当前的数据中心普遍面临两个问题亟待解决。

第一个问题是，伴随着高性价比的 x86 服务器技术的快速提升，虚拟化技术逐渐从高端服务器市场走向普通企业，各种规模的企业开始对其数据中心基础架构进行虚拟化，进而摆脱传统的“烟囱式”架构所带来的各种问题。这种基础架构虚拟化给客户带来服务器物理硬件利用率提升、IT 管理复杂度降低、数据中心运营成本降低等诸多好处的同时，也给客户带来了 I/O 搅拌效应。当各种不同类型的应用系统被基础架构虚拟化所屏蔽，各种不同特点的 I/O 被搅拌在一起，不仅使后端存储阵列因为所接收的 I/O 模型变得随机而性能降低，而且导致传统存储阵列难以针对上层业务系统进行优化。传统存储阵列逐渐变成了虚拟化基础架构的性能瓶颈，直接限制了虚拟化基础架构的效益最大化。

第二个问题是，为了消除传统存储阵列的性能瓶颈，很多企业客户选择增加机械硬盘数量来增加 IOPS。这不仅让客户在存储容量上产生了大量的浪费（CAPEX 的增加），也在数据中心的空间、能耗等方面引入了较大的运维费用（OPEX 的增加）。基础架构虚拟化带来的收益，大部分都被传统存储阵列所吞没。更为重要的是，堆叠机械硬盘这种方式，仅仅提升了传统存储阵列的 IOPS，却无法有效降低存储阵列的 I/O 响应时延，无法从根本上改善业务系统的性能。

部分企业客户对基础架构虚拟化和堆叠机械硬盘所带来的问题有着相对深刻的理解，于是采用阵列内配置固态硬盘和机械硬盘分级的方式来解决上述问题。可以说，在 2012 年以前，受限于固态硬盘的单位容量价格较高，这种阵列内分级的方式对于大多数企业客户而言，是性价比最高的选择。但是随着固态硬盘价格的不断下降、以及各种固态存储技术的不断成熟，企业客户的最佳选择逐渐演变为固态存储阵列。

华为公司针对企业客户所面临的各种问题，推出了 OceanStor Dorado 系列高性能固态存储阵列。本文档重点描述固态存储（尤其是闪存）带来的存储技术变革、固态存储阵列的客户关注点，并结合 Dorado 系列产品的特点对一些常见应用场景进行介绍。

2 引言/Introduction

根据 SNIA（Storage Networking Industry Association，全球网络存储工业协会）的相关文献，业界对固态存储（Solid State Storage）的定义为：使用硅晶半导体技术，而不是借助于机械旋转对磁碟、光碟或者磁带进行操作，从而实现数据存取的存储方式。

根据该定义，内存（RAM）、闪存（Flash）、相变存储（Phase Change Memory，后续简称为 PCM）等，都被称为固态存储。事实上，固态存储很早就被用于企业高价值数据的存储。在固态硬盘（Solid State Drive，后续简称为 SSD）出现以前，一些企业已经采用全内存阵列对其核心实时数据进行存储，以满足实时计算的存储性能需求。

随着闪存技术的不断成熟，以及价格逐渐变得可接受，部分存储阵列厂商于 2008 年左右开始在他们的阵列中引入基于闪存的 SSD，以提升整个阵列的性能。

随着存储业界对闪存技术的理解不断深入，人们发现，传统存储阵列是针对机械硬盘（Hard Disk Drive，后续简称 HDD）的各种特点进行设计的，如果简单地将 SSD 插入到传统存储阵列中，虽然可以立即获得一定程度上的性能提升，但是并不能充分发挥闪存的优势和规避闪存的缺点。于是，在 2010 年前后，市场上逐渐出现了各种形态的全闪存阵列，并宣称是针对闪存进行设计和开发的存储阵列。

伴随着形态各异的全闪存阵列逐步走向市场，一些传统存储阵列厂商也快速响应，推出满配 SSD 的存储阵列型号，并对外宣称拥有全闪存阵列。实际上，这种通过在传统存储阵列中满配 SSD 的所谓“全闪存阵列”，只是传统存储阵列厂商的市场手段。一时间，全闪存存储阵列市场鱼龙混杂，使得人们难辨真伪。

全闪存阵列的基础是闪存和 SSD，只有充分理解了闪存的特点和 SSD 的设计方式，才能理解全闪存阵列与传统阵列的不同。

本章节对闪存介质和固态硬盘进行技术性介绍。

2.1 传统阵列的问题

传统阵列极大地扩展了 HDD 的可靠性、性能，在不可靠的 HDD 基础上向外提供了可靠的存储服务。随着企业级存储应用的日趋复杂，传统阵列逐渐地在一些方面显露出不足：

- 可靠性。受限于 HDD 的机械部件，单块 HDD 的年失效率难以继续有效降低，这阻碍了存储阵列可靠性的进一步提升。
- 性能。传统阵列依靠堆积大量的 HDD 来获取较高的 IOPS，这种方式只能解决 IOPS 的问题，但是并不能有效降低 I/O 时延；同时，随着数据中心虚拟化的不断扩展，传统阵列所接收到的 I/O 模型越来越随机，越来越难以做针对性的优化。
- 成本。企业一般通过堆积 HDD 来获取所需的 IOPS，实际上，这导致了很多额外的容量成本，以及为了支撑这些大量 HDD 的空间、能耗成本。

闪存的出现，以及近年来成本大幅度降低，为解决上述问题提供了可能。

2.2 闪存介绍

并非只有基于闪存（Flash）介质的存储方式和技术才被称为固态存储。基于内存（RAM）、相变存储（PCM）等介质的存储方式和技术，均被称为固态存储。

当前，因为闪存在价格、容量、可靠性等多方面达到了相对领先的平衡，因此被广泛应用于固态存储领域。

2.2.1 概念和原理

闪存是一种非易失性的半导体存储器件。所谓“非易失性”，是指在断电情况下仍能保持所存储的数据信息。

目前市场上常见两种闪存类型：NOR 闪存和 NAND 闪存。这两种不同类型的闪存，因为使用方式上的差异，被用于不同的领域。NOR 闪存常用于存放系统启动程序，在嵌入式设备中较为常见；NAND 闪存主要用于数据存储，SSD 中使用的就是 NAND 闪存。

不论是哪种类型的闪存，其基本原理相同：使用三端器件作为存储单元，分别为源极、漏极和栅极，主要利用电场的效应来控制源极与漏极之间的通断；在栅极与硅衬底之间增加了一个浮置栅极，浮置栅极可以存储电荷，利用电荷存储来存储记忆。

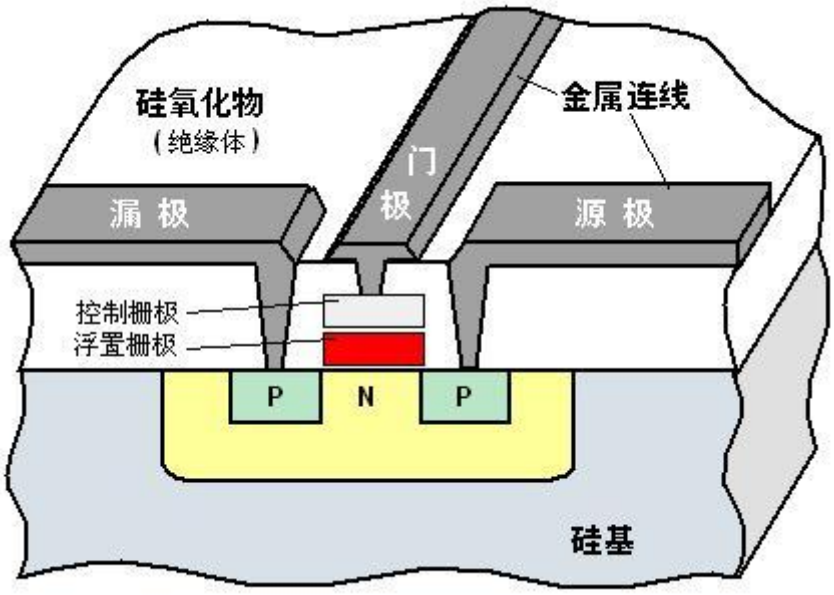


图 2-1 闪存存储单元示意

图 2-1 所示的闪存存储单元，可表示 1bit 的数据：向浮置栅极中注入电荷表示写入了‘0’，将电荷释放掉以后表示‘1’。

释放浮置栅极的电荷，从而使之变成‘1’，这个动作被称为“擦除”。

向浮置栅极注入电荷，从而使之变成‘0’，这个动作被称为“编程”。

2.2.2 技术特点

因为 NAND 闪存被广泛应用于 SSD 及全闪存阵列中，所以本文档后续在没有明确注明的情况下，均使用“闪存”一词指代“NAND 闪存”。

■ 闪存组织结构

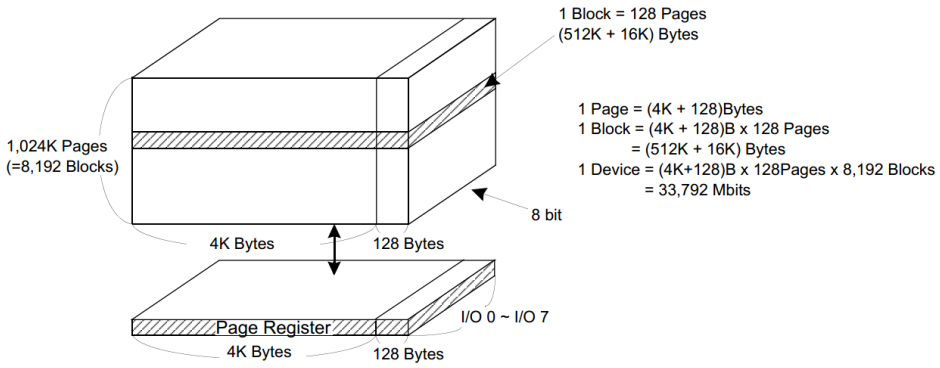


图 2-2 NAND 闪存颗粒组织结构

- 如图 2-2 所示，闪存颗粒内部，一般由成千上万个大小相同的块（block）所组成；块大小一般为数百 KB 到数 MB。

- 每一个块的内部，又分为若干个大小相同的页（page）；页的大小一般为4KB 或者 8KB。

■ 数据写入

- 向闪存中写入数据时，只能以页为粒度进行写入。
- 如果闪存中某个页已经被写入了数据，那么不能向这个页中再次直接写入数据，只能在这个页的数据被清空以后才能再次写入。
- 闪存进行数据清空的粒度是块，即一次清空动作会将一个块的数据全部抹除。清空动作对应着闪存的擦除动作，即擦除了一个块的数据后，这个块中所有的 bit 位都变成了 1。
- 写入动作对应着闪存的编程动作，将数据写入页时，将特定的 bit 位从 1 变成 0，就使得这个页保存了相应的数据。
- 闪存就工作在这样的“擦除”和“编程”循环中，一次这样的循环，被称为一次擦写（Program/Erase，简称为 P/E）。
- 闪存中每个块的 P/E 次数有限；当某个块的 P/E 次数达到上限后，就无法保证能够继续有效地存取数据。
- 闪存的 P/E 次数与很多因素相关，本文档后续谈及的 P/E 次数，均默认为在本文档编写的时间点，主流闪存的情况。

■ 数据读出

- 闪存中保存的数据，经过一段时间后，可能存在若干 bit 位的错误。如果直接将页中读出的数据返回给上层业务，就可能造成业务失败。
- 为了保证返回给上层业务的数据是正确有效的，闪存内部预留了部分空间用于保存业务数据的 ECC（Error Correcting Code，纠错码）。每当读取数据时，控制器会使用相应的 ECC 对这些数据进行错误检查和纠正。
- 受限于控制器的计算能力，ECC 的纠错范围有限，只能在页面数据中出现 bit 位错误的数量不超过一定的上限时才有效。当前主流的 ECC 纠错能力一般是 24bit/1KB，即每 1KB 数据（包含业务数据和 ECC 校验数据）内出现了 bit 位错误不超过 32 个时，控制器可以通过计算的方式得出正确有效的业务数据。
- 当某个页中的 bit 位错误数超过控制器的计算能力后，该页的业务数据无法被正确读出，此时便产生一个 UNC（Uncorrectable）错误。
- UNC 错误只能被更高层级的 RAID 机制所修复。

■ 闪存分类

- 目前常见的闪存，可分为 SLC、MLC、TLC 几类。
- 图 2-1 展示了闪存颗粒内部的一个存储单元，英文称之为 cell。

- SLC 是 Single Level Cell 的缩写，表示 NAND 闪存内部的每个 cell 只能存储一个 bit 位的信息，即使用两个电平值来分别表示 0、1。
- MLC 是 Multi Level Cell 的缩写，原本表示 NAND 闪存内部的每个 cell 可以保存多个 bit 位的信息，但是实际使用中，MLC 特指每个 cell 保存两个 bit 位的信息，即使用四个电平值来分表表示 00、01、10、11。
- TLC 是 Triple Level Cell 的缩写，表示 NAND 闪存内部的每个 cell 可以保存三个 bit 位的信息，即使用八个电平值来分别表示 000、001、010、011、100、101、110、111。
- MLC 又可以细分为 eMLC（Enterprise MLC，企业级 MLC）和 cMLC（Consumer MLC，消费级 MLC）。
- eMLC 和 cMLC 在本质上相同，只是在工厂制造处理过程中，厂商把通过了某些数据完整性和耐用度测试之后、确认可以达到较高 P/E 次数的 MLC 定义为 eMLC，余下的就成为 cMLC。
- 业界一般直接使用 MLC 来指代 cMLC，本文档使用 cMLC，以便与 eMLC 明显区分。
- 这几种不同类型的闪存，在容量、P/E 次数、价格等多个方面存在较大差异，如表格 2-1 所示：

表格 2-1 闪存类型对比

	单位体积的容量	P/E 次数	单位容量的价格
SLC	小	约 100,000 次	高
eMLC	适中	约 10,000 次	中
cMLC	适中	约 1,000~3,000 次	低
TLC	大	约 500~1,000 次	很低

- 目前，SLC 和 eMLC 主要应用于企业级市场，cMLC 主要应用于消费级市场，TLC 尚未广泛应用。

2.3 SSD 介绍

本文档均描述基于闪存的 SSD。图 2-3 展示了一块 SSD 内部的构成。

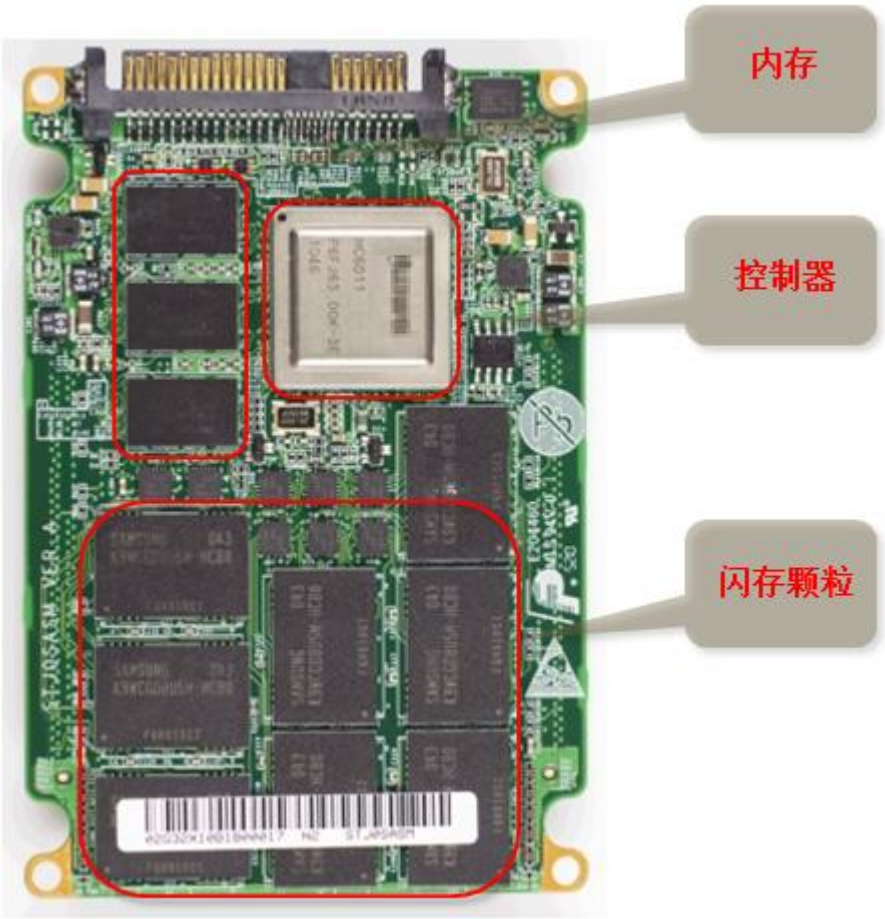


图 2-3 SSD 内部构成

一块 SSD，由控制器、内存、闪存颗粒等单元所组成。绝大多数厂商所生产的 SSD 在外观尺寸、接口规范、数据访问方法等方面均与 HDD 保持一致，从而使之可以被直接应用于 HDD 所能使用的任何场景。

控制器提供了外部主机接口、内部闪存管理接口，并通过内嵌的 CPU 来运行 SSD 固件；SSD 固件管理着主机可见的存储地址空间、闪存物理空间、垃圾回收、磨损均衡等。华为自研 SSD 所采用的控制器，是华为具有完全知识产权、自行设计开发的 ASIC 芯片，对外提供 6Gb/s SAS 2.0 双端口接口。

内存用于运行 SSD 固件，并保存在地址空间虚拟化所需要的各种表项。

多颗闪存颗粒分布在 SSD 的电路板上，共同为 SSD 提供存储空间。

SSD 没有 HDD 的音圈马达、悬臂等机械部件，因此抗震性极佳，更为重要的还在于可以真正做到多并发和低时延，使其 IOPS 可以超过 HDD 两个数量级以上。

2.3.1 地址空间虚拟化

在 2.1.2 节，提到了闪存有 P/E 次数的限制。为了避免热点区域被大量写入数据而导致对应的闪存物理区域快速达到 P/E 次数上限而失效，同时也为了加快 SSD 对写请求的响应速度，SSD 在设计时采用了地址空间虚拟化的方式。

所谓“地址空间虚拟化”，是指：

1、 SSD 的 LBA（Logical Block Address，逻辑块地址）到 PBA（Physical Block Address，物理块地址）的映射，并非固定的映射，而是可以随时更改的；

2、 SSD 内部的最小空间管理粒度是页，每个页都有唯一的一个编号，这就是 PBA；

3、 SSD 内部维护了一张映射表，记录了 LBA 到 PBA 的映射关系；

4、 每当有数据写入 SSD 时，SSD 都会选择一个或者多个干净的页来保存这些新写入的数据，同时让映射表记录下新的映射关系；所谓“干净”，是指这些页经历过一次擦除，同时尚未经历编程；

5、 经过上面的地址空间虚拟化之后，使用者反复向一块 SSD 的相同区域写入数据，实际上是写到了该 SSD 的不同物理区域；

2.3.2 容量冗余

为了避免部分闪存损坏所导致的整块 SSD 失效，SSD 在设计时都会做到容量冗余。举例，一块标称为 100GB 的 SSD，内部实际的闪存物理容量一般都做到了 110GB 以上。

超过标称容量的部分，与标称容量的比值，被称为冗余比。一般而言，冗余比越大，则 SSD 的可靠性、寿命、性能就越好。

Dorado 系列固态存储阵列，全部采用华为自研的 SSD。这些 SSD 的冗余比均做到了 28%，完全达到企业级要求。表 2-2 显示了部分华为自研 SSD 的标称容量和物理容量。

表 2-2 华为自研 SSD 容量冗余情况

标称容量	物理容量	冗余比
100 GB	128 GB	28%
200 GB	256 GB	28%
400 GB	512 GB	28%

2.3.3 垃圾回收

地址空间虚拟化和容量冗余，除了可以有效避免部分闪存失效导致整块 SSD 失效外，还能够为 SSD 的垃圾回收提供交换空间，保证 SSD 的性能平稳。

地址空间虚拟化在避免了对相同物理区域反复擦写的同时，也引入了垃圾数据和垃圾页。相关解释如下：

1、 主机访问 LBA 100，写入数据 AA，不妨假设 SSD 内部使用编号为 401 的页来保存数据 AA；

2、 过了一段时间，主机再次访问 LBA 100，写入数据 BB，地址空间虚拟化机制会让数据 BB 被写入到另外一个页，不妨假设这个页的编号为 623；

3、 经历了上述两个写操作，页 401 保存了数据 AA，页 623 保存了数据 BB，且页 401 保存的数据是无用的，页 623 中保存的数据才是有用的；

4、 页 401 中保存的数据被称为垃圾数据；

5、 根据闪存自身需要先擦后写的特点，页 401 不能直接写入新的数据，在完成擦除动作以前，页 401 就被称为垃圾页；

垃圾回收（Garbage Collection，简称 GC）的目的就是擦除垃圾数据，让垃圾页变成可以写入数据的干净页。

在 SSD 的内部实现中，垃圾回收一般是后台任务，不断监控着 SSD 内部的块和页的使用情况，当垃圾页过多的时候，就会：

1、 选择那些包含了较多垃圾页的块，将其中的有效数据迁移到其他块的干净页；

2、 对那些不再包含有效数据的块执行擦除操作；

3、 将执行了擦除操作的块放入可用资源池中，以供新的数据写入；

通过上述垃圾回收的动作，可以看出，垃圾回收在不断净化 SSD 的同时，也在 SSD 内部制造了更多的写操作，从而使得闪存所真正承载的写入数据量超过了主机写入的数据量，这种现象被称为写放大。

当业务模型固定时，闪存真正承载的写入数据量和主机写入的数据量这两者的比值，不会随着时间的推移而变化，而是以个固定数值为中心点小范围波动，我们称这个中心点为写放大系数。

写放大系数与冗余比、SSD 固件中的各种算法等相关，值越小越好。华为自研 SSD，在全随机小 IO 业务场景下，写放大系数约为 2.5；在顺序大 IO 业务场景下，写放大系数约为 1.1。这样的写放大系数，达到了业界主流的水准。

2.3.4 磨损均衡

仅仅具备地址空间虚拟化和容量冗余，并不足以完全避免部分闪存块相比其他闪存块提前达到 P/E 次数上限。为了让所有的闪存块尽量均衡地擦写，确保避免出现部分块达到 P/E 次数上限而其他块还剩余较多 P/E 次数的情况，必须引入磨损均衡技术。

磨损均衡（Wear Leveling，简称 WL）所做的事情，就是记录每一个块的 P/E 次数，然后在需要擦除或者写入数据时，尽量选择那些 P/E 次数相对较少的块。经过了磨损均衡的 SSD，使用寿命可以得到最大化。

磨损均衡分为动态磨损均衡和静态磨损均衡两种方式。

■ 动态磨损均衡

所谓动态磨损均衡，是指该类磨损均衡由主机 IO 所触发。

当主机发起一个写请求时，SSD 需要找一个或者多个干净页供新数据的写入。此时，动态磨损均衡算法会生效，尽量选择那些 P/E 次数相对较少的块来提供干净页。

■ 静态磨损均衡

所谓静态磨损均衡，是指该类磨损均衡是 SSD 内部自主发起的。

考虑这样的极端场景，一块 100GB 容量的 SSD 装满了用户数据，其中 99GB 是冷数据，用户写入后再没有更新，剩下 1GB 是热数据，用户在频繁更新。

对于这 99GB 的冷数据，如果不进行特殊的处理，那么就会一直占用至少 99GB 固定的物理闪存区域，于是剩下的空间就留给 1GB 热数据进行反复的擦写。显而易见，一段时间后，99GB 被固定占用的闪存区域的 P/E 次数就会变得明显低于剩下的闪存区域。这样的不均衡容易导致 SSD 提前失效。

静态磨损均衡可以解决上述问题：通过记录每个块的 P/E 次数，识别出那些磨损相对轻微、且长期没有产生垃圾页的块（即保存了冷数据的块），主动将其中的有效数据迁移到磨损相对严重的块中，从而达到保持整个 SSD 磨损均衡的效果。

2.3.5 坏块管理

在 SSD 的使用过程中，尽管有各种机制和算法来尽量延长使用寿命，但是闪存的损坏依然是不可避免的。容量冗余为解决闪存损坏的问题提供了基础保证。

闪存的损坏，是以页为粒度的，即一个块内部包含了若干的页，可能有的页处于正常状态，另外的一些页处于损坏状态。

实际上，当某个块内部出现了多个损坏的页时，这个块的其他页大都处于损坏的边缘。鉴于此，SSD 的固件一般以块为粒度来管理闪存的损坏情况：当发现某个块内部、因为损坏而无法读出数据的页的数量超过某个阈值时，则判定这个块已损坏，随后就将这个块中的有效数据迁移到其他可用的块中，并将这个损坏了的块标记为损坏，自此不再用于保存任何业务数据。

一般来说，SSD 固件发现坏块的手段包括两种：主机 IO 触发、内部巡检。

上述工作在 SSD 内部被称为坏块管理。

2.3.6 SSD 寿命

地址空间虚拟化、容量冗余、垃圾回收、磨损均衡、坏块管理等一系列机制和措施，保证了 SSD 使用寿命的最大化。

一般而言，SSD 的使用寿命可以直接反映在写入数据量上。

表 2-3 华为自研 SSD 支持写入数据量

SSD 类型	支持的最大写入数据量	
	随机	顺序

100GB SLC	5 PB	13PB
200GB SLC	10 PB	25.6PB
200GB eMLC	1PB	2.5PB
400GB eMLC	2PB	5 PB

需要指出的是，根据操作系统和存储业界主流厂商的分析和统计，企业级市场上，每块硬盘平均每天写入的数据量远小于 50GB。

所以，SSD 的使用寿命，对于各种企业客户而言，不会带来任何问题。

3 解决方案/Solution

随着闪存介质单位容量价格的逐渐下降，完全基于闪存进行构建的存储阵列进入了人们的视野。今天，已经有很多企业宣称提供全闪存阵列。这些全闪存阵列，因为不再受到 HDD 的限制，因而在外观形态上各不相同，有的是一体化的盒子，有的看上去就是传统阵列满配 SSD。事实上，有的企业，确实就是在传统阵列中将 HDD 简单替换为 SSD，就宣布支持全闪存阵列。

实际上，SSD 的 I/O 响应时延比 HDD 低出 2 个数量级，量变进而引起质变，因此针对 HDD 所设计的传统阵列，是无法充分发挥 SSD 的优势的。

本章节对全闪存阵列进行介绍，主要涉及全闪存阵列与传统阵列的异同、华为自研全闪存阵列 Dorado 的基本特性、全闪存阵列的客户价值和实用性等。

3.1 Dorado 系列全闪存阵列

Dorado 是华为自研的全闪存阵列，包括阵列控制器硬件、软件、SSD，均为华为自主研发，具有高可靠、高性能、易使用、易维护的特点。

图 3-1 Dorado 标识

图 3-1 展示了 Dorado 系列全闪存阵列的标识，是一只快速前进的剑鱼。

Dorado 是拉丁语中的剑鱼，寓意存储海洋中遨游速度最快的鱼。实际上，Dorado 系列全闪存阵列，在业界公认的性能基准测试中，在 IOPS 和时延等方面均取得了优异的成绩。

3.1.1 Dorado5100

Dorado5100 是一款超高性能的全闪存阵列，配置灵活，覆盖面广。

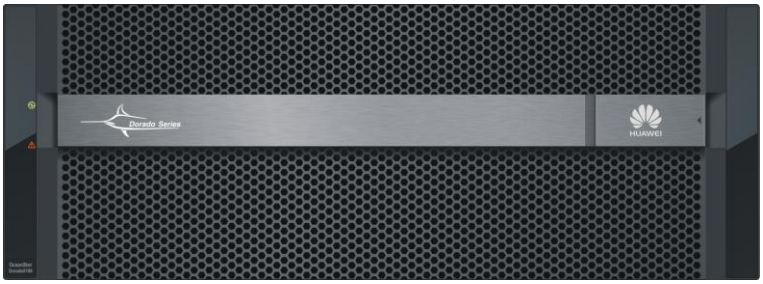


图 3-2 Dorado5100 控制框正面（带面罩）

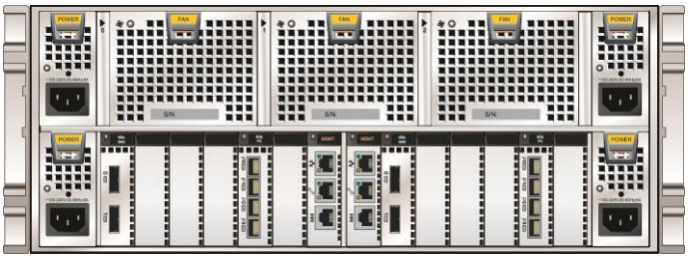


图 3-3 Dorado5100 控制框背面

图 3-4 Dorado5100 硬盘框

Dorado5100 的简要规格见表 3-2。

表 3-2 Dorado5100 规格

形态	独立控制框，高度 4U，支持若干接口卡 硬盘框高度 2U，带 24 个 2.5 寸 SSD 槽位 硬盘框满框配置相同型号的 SSD 进行打包销售 一台 Dorado5100 最多级联 4 个相同规格的硬盘框 所有有源器件（包括控制器、电源、风扇等）均冗余并支持现场更换
SSD 支持	SLC：100GB、200GB eMLC：200GB、400GB
容量	SLC：2.4TB ～ 19.2TB eMLC：4.8TB ～ 38.4TB
主机接口	8Gb FC、10Gb ETH（iSCSI）
性能	SPC-1 IOPS™：600,052.49 @ 1.09ms

高级特性	快照、远程复制

3.1.2 Dorado2100 G2

Dorado2100 G2 是 Dorado2100 的升级换代，除了性能有了大幅度提升外，还增加了若干高级特性。

图 3-5 Dorado2100 G2 控制框正面（带面罩）

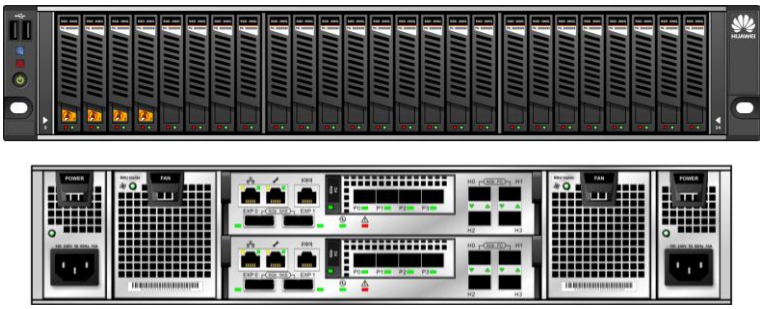


图 3-6 Dorado2100 G2 控制框正反面（无面罩）

图 3-7 Dorado2100 G2 硬盘框

从图片中可以看到，Dorado2100 G2 采用盘控一体的设计，支持级联硬盘框，且每框硬盘的数量从 24 盘位增加到 25 盘位，这主要是为了方便配置热备盘。

Dorado2100 G2 的简要规格见表 3-3。

表 3-3 Dorado2100 G2 规格

形态	控制框高度 2U，带 25 个 2.5 寸 SSD 槽位 硬盘框高度 2U，带 25 个 2.5 寸 SSD 槽位 控制框和硬盘框满框配置相同型号的 SSD 进行打包销售 一台 Dorado2100 G2 最多级联 3 个相同规格的硬盘框 所有有源器件（包括控制器、电源、风扇等）均冗余并支持现场更换
SSD 支持	SLC：100GB、200GB eMLC：200GB、400GB
容量	SLC：2.5TB ～ 20.0TB

	eMLC：5.0TB ～ 40.0TB
主机接口	8Gb FC、10Gb ETH（iSCSI，带 TOE 功能）、40Gb InfiniBand
性能	SPC-1 IOPS™：400587.11 @ 0.75ms
高级特性	Thin provisioning、全局磨损均衡、VMware VAAI

3.2 客户价值

全闪存阵列的出现，旨在帮助客户改善现有数据中心的处理能力、提升客户业务竞争力、降低 TCO，并应用于一些传统存储阵列所无法满足的应用场景。

3.2.1 降低 TCO

目前，全闪存阵列比传统存储阵列的单位容量价格高出数倍，这导致部分客户在投资时，简单地认为全闪存阵列比传统存储阵列贵。实际上，全闪存阵列因为时延低、性能高、空间和能耗需求低，且无需通过堆积硬盘来获取高 IOPS，在生命周期内的运维支出也显著低于传统存储阵列，从而使得全闪存阵列的总体拥有成本低于传统的高性能存储阵列。

表 3-4 对比了两款传统存储阵列 H、I 和 Dorado2100 G2 的各项成本。

表 3-4 传统存储阵列和全闪存阵列 TCO 对比

	传统存储阵列 H	传统存储阵列 I	Dorado2100 G2
硬盘配置	10K RPM 300GB SAS HDD 共 896 块	15K RPM 300GB SAS HDD 共 230 块	400GB eMLC SSD 共计 100 块
物理容量	268,800GB	69,000GB	40,000GB
SPC-1 IOPS™	109,986.41	82,496.08	~250,000
时延	～5ms	~7ms	～2ms
售价 （含三年维保）	$484,985.78	$361,416.00	~$310,000
容价比 （$/GB）	1.80	5.24	~7.75
性价比 （$/SPC-1 IOPS™）	4.41	4.38	~1.24
机架空间	3 个机架	16U	8U
典型功耗	~13KW	~3.3KW	~1.5KW
第一年运营支出	$42,000	$5,600	$2,800

第二年运营支出	$42,000	$5,600	$2,800
第三年运营支出	$42,000	$5,600	$2,800
TCO	$610,985.78	$53,216.00	~$318,400

表 3-4 中的传统存储阵列 H 和 I，相关数据均来源于 SPC 官网上公开的 SPC-1 报告，以及产品官网提供的规格参数。这两款产品均在 2013 年初通过 SPC-1 基准测试认证，与本技术白皮书编写时间点接近，因此相关数据具备参考意义。

SPC-1 基准测试结果官方网页：
http://www.storageperformance.org/results/benchmark_results_spc1

表 3-4 中的 Dorado2100 G2，采用满配 100 块 400GB eMLC 的配置，共提供 40TB 的物理存储空间，可以满足绝大多数高性能存储的空间需求。

上述对比中，运营支出主要考虑空间费用和用电费用，考虑到该项因素存在较大的地域性差异，因此选择采用中国主流电信运营商关于机架租用的相关费用进行计算。

通过表格中的对比可以看出，传统存储阵列通过堆积硬盘的方式来获取较高的性能，这样的方式，不仅使得用户付出了大量额外的成本来购买并不需要的容量，而且所获得的性能，不论是从 IOPS 层面还是从时延层面，均无法与全闪存阵列相提并论。同时，在满足容量需求的前提下，Dorado2100 G2 这样的全闪存阵列的 TCO 实际上更低。而且，随着闪存价格不断下降，全闪存阵列将会更加具备竞争优势。

全闪存阵列相比传统存储阵列，可以有效帮助那些容量需求不是特别大、对存储性能有较高诉求的客户节省投资，降低这些客户在存储方面的 TCO。

3.2.2 提升客户业务竞争力

全闪存阵列不仅可以帮助客户节省投资，还有助于帮助客户解决一些使用传统存储阵列所无法解决的问题。

某建材零售商在业务扩展过程中，发现传统存储阵列的高时延，导致运行在服务器上的数据库的 IO wait 时间长期居高不下，进而导致整个业务系统的每笔交易的时延过大，以及整个系统的有效并发数量无法进一步提升。

通过尝试追加传统存储阵列，客户发现在相应的业务并发压力情况下，整个系统的性能并未得到有效改善。

为了解决该问题，该客户 IT 部门通过与系统集成代理商的交流，提供了两种方案：

- 对整个 IT 基础架构进行重新设计，包括服务器和存储。
- 在存储侧引入固态存储阵列作为主存。

经过初步评估，客户认为第一种方案的改动过大，且效果难以评估，于是首先尝试第二种方案。

通过在原有系统中引入 Dorado5100 作为数据主存储设备，客户发现事务处理等待时间降至原来的 18.3%，实际应用中每笔交易平均等待时间仅为原来的 20%。在模拟环境中，系统最大用户数提升了 20 倍。客户最终选择了 Dorado5100 来提升整个系统的业务处理能力，进而帮助自己达成业务扩张的目的。

实际案例证明，全闪存阵列不仅能够帮助客户在达成既定目标时有效节省 IT 投资，而且通过低时延特性，能够帮助客户完成一些在传统存储阵列上难以达成的目标，有效提升客户的业务竞争力。

该案例在 4.3.1 节进行详细介绍。

3.3 技术剖析

对全闪存阵列做一个定义：完全不依赖于传统机械硬盘，仅仅依靠闪存进行数据存取的存储阵列。

根据上述定义，再结合当前市场上已有的全闪存阵列产品，可以将全闪存阵列划分为三种类型：

- 封闭结构。一体化的盒子，内部结构形形色色，总体来说，这种形态的全闪存阵列更像是将 SSD 做成了机架式设备，即体积更大的 SSD。除了性能较高以外，几乎没有其他的存储特性，可维护性也较差。
- 传统结构。硬件看上去与传统存储阵列几乎无差异，但是系统软件针对闪存进行开发，不仅性能较高，而且具备各种存储相关的特性。
- 传统阵列满配 SSD。在传统存储阵列中，将 HDD 全部更换为 SSD。性能一般，但是功能特性一般较为丰富。

根据该定义，在一台传统存储阵列中，将 HDD 全部更换为 SSD，这样的阵列也可以被称为全闪存阵列。但是这样的全闪存阵列，并不能充分发挥闪存的各种优势。

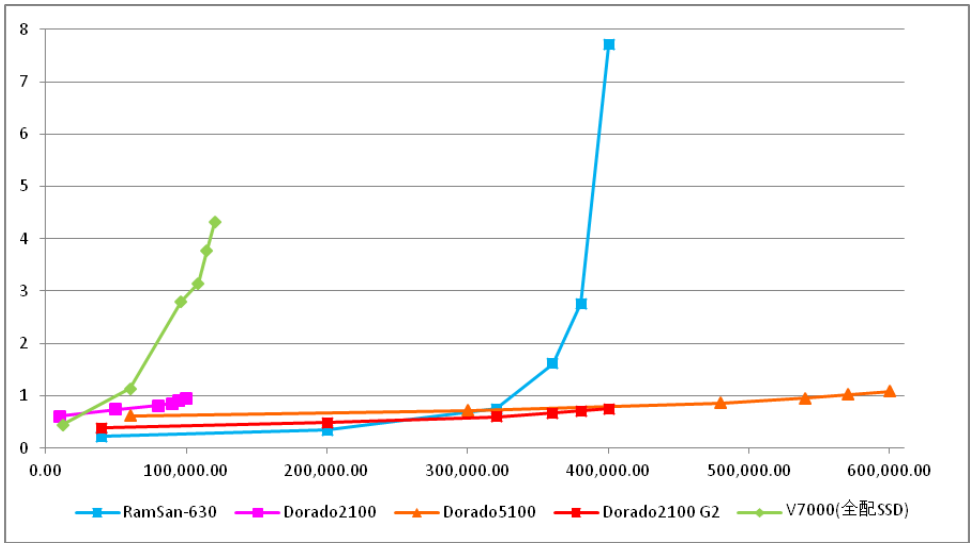


图 3-8 五款全闪存阵列对比

图 3-9 对比了五款型号的固态存储阵列的 SPC-1 IOPS™ 性能，横轴是 IOPS，纵轴是时延（毫秒），展示了存储阵列在不同 IOPS 压力下的时延分布情况。图 3-9 的数据来源于 SPC 官网。

SPC-1 是存储性能委员会（Storage Performance Council，简称 SPC）定义的一个 I/O 模型和测试基准，主要模拟 OLTP 和 OLAP 的 I/O 业务特征。SPC-1 测试基准在存储业界具有标杆意义。一般来说，中低端存储阵列的 SPC-1 IOPS™ 认证测试指标在 50K 以下，中高端一般在 50K～200K 左右，高端存储阵列一般在 200K～300K 左右。

图 3-9 中的五款全闪存阵列，分别归类于三种不同的类型：

- 封闭结构：RamSan-630
- 传统结构：Dorado2100、Dorado5100、Dorado2100 G2
- 传统阵列满配 SSD：V7000

通过对比，可以看到：

1、 Dorado5100、Dorado2100 G2、RamSan-630 都已经超过传统高端存储阵列的性能。

2、 在 IOPS 较小时，RamSan-630 这种封闭结构的全闪存阵列，时延略低于 Dorado 系列，但是当 IOPS 较高时，Dorado 依然保持了极低的时延，而 RamSan-630 的时延会出现暴增的现象。

3、 V7000 满配 SSD，这种类型的阵列，其时延保持线性增长，与 Dorado 相比，明显没有充分发挥出闪存的低时延优点。

4、 综合来看，只有 Dorado 这种传统结构的全闪存阵列，始终保持了闪存的低时延优点，比其余两种类型的全闪存阵列表现更加优异。

Dorado 这种传统结构的全闪存阵列，不仅通过改写阵列系统软件来实现了性能上的胜出，而且在可靠性、可维护性等方面，充分继承了传统存储阵列的精华，并针对闪存进行改进，从而可以实现更高的可靠性、可维护性。

3.3.1 SSD 带来的问题

低时延带来的困扰

IOPS、I/O 并发、I/O 时延之间的关系是：$IOPS = \frac{I/O并发}{I/O时延}$ 。

一块高性能企业级 SAS HDD，4KB I/O 随机访问时延约为 5ms。

一块 SAS 接口 SSD，4KB I/O 随机访问时延约为 0.2ms。

如图 3-2 所示，假设将一块 HDD 和一块 SSD 分别接入两台相同配置的主机，然后这两台主机分别以单并发的压力下发 I/O。

图 3-9 HDD、SSD 直接与主机相连

图 3-10 左侧为 HDD，其 IOPS 为 1 / 5ms = 200 IOPS。

图 3-10 右侧为 SSD，其 IOPS 为 1 / 0.2ms = 5000 IOPS。

现在，将图 3-10 中的 HDD 和 SSD 分别插入两台相同配置的存储阵列控制器，再通过阵列接入到服务器，如图 3-11 所示。

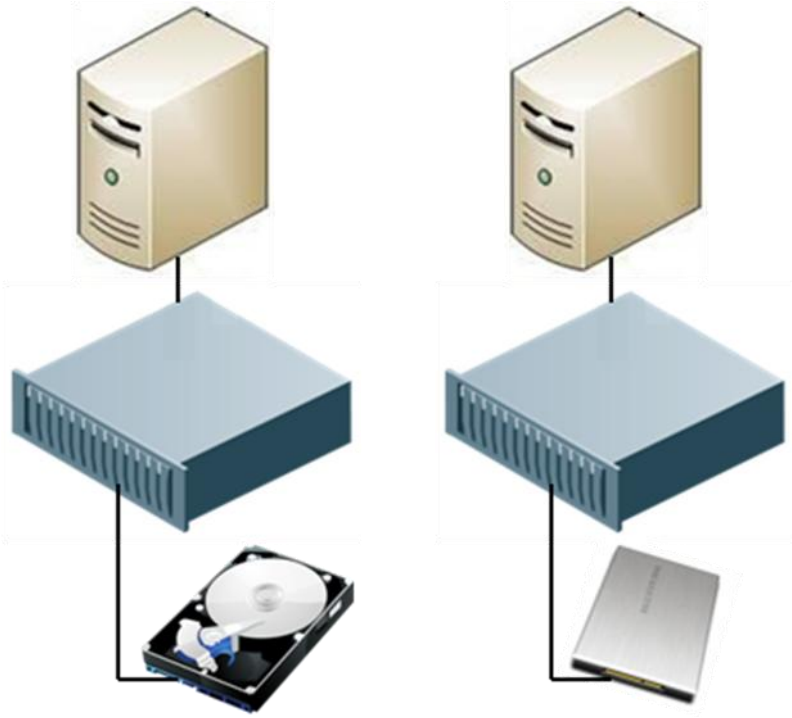


图 3-10 HDD、SSD 通过阵列与主机相连

存储阵列控制器会引入处理时延，一般而言，控制器引入的时延随压力的不同而变化，在单并发这种小压力情况下，引入的时延一般在 0.2ms 左右。

此时，主机看到的 IOPS 会发生变化：

图 3-11 左侧的控制器加上 HDD，其 IOPS 为 1 / (0.2ms + 5ms) = 192 IOPS。

图 3-11 右侧的控制器加上 SSD，其 IOPS 为 1 / (0.2ms + 0.2ms) = 2500 IOPS。

通过上述计算可以看到，HDD 插入存储阵列后，控制器引入的时延占总时延的比例很小，IOPS 性能几乎无降低，这也就是为什么可以用盘数乘以单盘性能来预估传统存储阵列的 IOPS 的原因。而将 SSD 插入存储阵列后，因为控制器引入的时延占总时延的比例较大，SSD 的性能仅发挥了一半。同时，这里也可以看到，对于全闪存阵列的性能，是不能单纯以 SSD 数量乘以 SSD 的性能来进行评估的。

简单地将传统存储阵列中的 HDD 更换为 SSD，这样所构建出来的全闪存阵列，并不能充分发挥 SSD 的高性能优势。

随机型业务和顺序型业务的性能差异

HDD 和 SSD，在应付不同类型的 I/O 业务时，表现差异很大，表 3-5 对比了两者在随机型业务和顺序型业务的性能差异。

表 3-5 HDD 和 SSD 在随机和顺序业务下的性能

	随机型业务（4KB 随机访问）	顺序型业务（512KB 顺序访问）

		IOPS	带宽（MB/s）	IOPS	带宽（MB/s）
HDD	读	～200	～0.8	～400	～200
	写				
SSD	读	～20,000	～80	～500	～250
	写	～60,000	～240	～600	～300

从表 3-5 中可以看到，HDD 的顺序型业务数据吞吐速率（即带宽）超过随机型业务数据吞吐速率的 200 倍以上，这也就是为什么传统存储阵列设计了非常了复杂的 CACHE 算法以尽量对主机下发的业务数据进行重组，再以顺序的方式去访问 HDD，进而提升整体的性能。

而对于 SSD 而言，顺序写业务数据吞吐速率不足随机写业务数据吞吐率的 4 倍，顺序读更是不足随机读的 2 倍。

这些数据表明，不同厂商针对 HDD 所设计的 CACHE 算法以及 I/O 调度算法等，并不一定适用于全闪存阵列，需要进行重新考虑。

性能瓶颈

对于传统存储阵列，HDD 是显而易见的性能瓶颈，这也正是为什么通过增加 HDD 就可以增加整个阵列的 IOPS 和带宽。同时，传统存储阵列的系统软件，都是围绕着如何消除 HDD 的性能瓶颈来设计和开发的。

而对于全闪存阵列，从表 3-1 中可以看到，现在单块 SSD 的 IOPS 就已经达到数万，一个常见的 24 盘位硬盘框的潜在最高 IOPS 就已经达到了一百万以上。所以全闪存阵列的性能瓶颈，一般位于阵列控制器，包括 CPU 处理能力、系统带宽能力、系统软件的设计和算法等等。

向关键路径要时间，向非关键路径要资源；取长补短，消除短板。这是做任何设计的基本原则之一。全闪存阵列相比传统阵列，性能瓶颈点已经发生了迁移，因此在软硬件设计上就应该有所不同，否则无法充分发挥闪存的高性能优势。

擦写次数限制

闪存的一个特点是 P/E 次数有限，而且对闪存的相关研究数据表明，闪存失效的概率随着 P/E 次数的增加而上升。

尽管 SSD 的失效率在到达 P/E 次数前是可以保证的，但是如果能够有效减少 P/E 次数、避免热点数据区域对固定的 SSD 进行频繁访问，还是可以进一步降低失效率，进而降低闪存阵列使用过程中出现各种故障的概率，提升整个闪存阵列的使用寿命。

Dorado 采用全局磨损均衡，来减少 P/E 次数、避免热点区域提前失效。

3.3.2 设计理念

Dorado 作为华为自研的全闪存阵列，在设计和开发时，坚持以下思路：

- 继承传统存储阵列的精华，尽量保持客户已有的存储阵列使用习惯。传统存储阵列在可靠性、可维护性等方面，已经积累了多年的经验，例如所有有源器件均冗余且可在线更换，类似于这样的特性，全部予以继承。
- 充分考虑 SSD 与 HDD 的巨大性能差异。SSD 的性能高出 HDD 两个数量级，这直接导致系统的瓶颈发生了迁移，因此以前针对 HDD 的各种性能设计必须被重新审视。
- 充分考虑 SSD 与 HDD 完全不同的故障模式。业界关于 SSD 的统计年失效率为 0.44%，而 HDD 的年失效率为 0.6%。尽管 SSD 已经比 HDD 相对可靠，但是 SSD 的故障模式与 HDD 并不相同，因此有针对的性的设计和开发，有利于进一步降低 SSD 在全闪存阵列中的故障率。

3.4 可靠性、寿命、性能

经过多年的发展，SSD 及固态存储阵列在可靠性、寿命、性能等方面均有了较大程度的发展，完全满足各类企业级存储应用需求。

3.4.1 可靠性

可靠性基础

可靠性分为狭义可靠性和广义可靠性。狭义可靠性是指设备不出故障的概率，广义可靠性包括设备不出故障的概率、设备出现故障后的修复时间、设备在较长的运行周期内的可用度等。

本文档的可靠性是指广义的可靠性。

■ **设备不出故障的概率**

业界一般使用 MTBF、 FIT、AFR 来衡量设备出故障的概率。

MTBF（Mean Time Between Failure），平均故障间隔时间，从可靠性方面标识故障发生的概率。指一个组件或设备的无故障运行平均时间，通常以小时为单位。MTBF 值越大，设备越可靠。

FIT（Failure In Time），失效率，某个器件在 10 亿个小时的工作时间内出错的总次数，或者 1000 个器件在 1 百万个小时的工作时间内出错的总次数，或者 1 百万个器件在 1000 个小时的工作时间内出错的总次数。某个复杂器件的 FIT，等于组成该复杂器件的各个基本器件的 FIT 之和。FIT 值越小，设备越可靠。

AFR（Annual Failure Rate），年失效率，一般而言，这是一个统计值，根据对大量样本进行统计得到的设备失效概率，一般用百分比来表示。AFR 值越小，设备越可靠。

MTBF、FIT、AFR 三者之间的关系是：

$$MTBF = 10^9 / FIT$$

$$FIT = (10^9 \times AFR) / (365 \times 24)$$

- **设备出现故障后的修复时间**

业界一般使用 MTTR 来衡量设备修复时间。

MTTR（Mean Time To Repair），平均修复时间，从可维护性方面标识故障的恢复能力，一般以小时为单位。指一个组件或设备从故障到恢复正常所需的平均时间，实质上指的是设备的容错能力。MTTR 值越小，设备容错能力越强。

- **设备在较长的运行周期内的可用度**

可用度表示设备的出勤率，计算公式为：

$$可用度\ A = MTBF / (MTBF + MTTR)$$

从该公式可以看出，增加 MTBF 或者减小 MTTR，都可以提高设备的可用度。

对于由多个独立器件（可单独修复或者现场更换）所组成的复杂设备，其可用度为所有独立器件的可用度之乘积，即：

复杂设备可用度 = 独立器件 1 可用度×独立器件 2 可用度×……

在企业级市场，大多数客户要求 99.999%的可用度，这意味着设备每年的宕机时间不超过 5 分钟。

SSD 可靠性

SSD 是全闪存阵列的主要部件。与 HDD 相比，SSD 消除了机械部件，因此故障模式也与 HDD 存在较大的差异。一般而言，SSD 的故障更加容易预测和管控，这也就是为什么业界统计的 SSD 年失效率明显低于 HDD 的主要原因。

关于 SSD 自身可靠性的信息，请参考华为自研 SSD 技术白皮书。

硬件可靠性

Dorado 系列全闪存阵列均采用全冗余硬件设计，所有有源器件均做到了冗余备份，有效消除单点故障，且可在线更换。

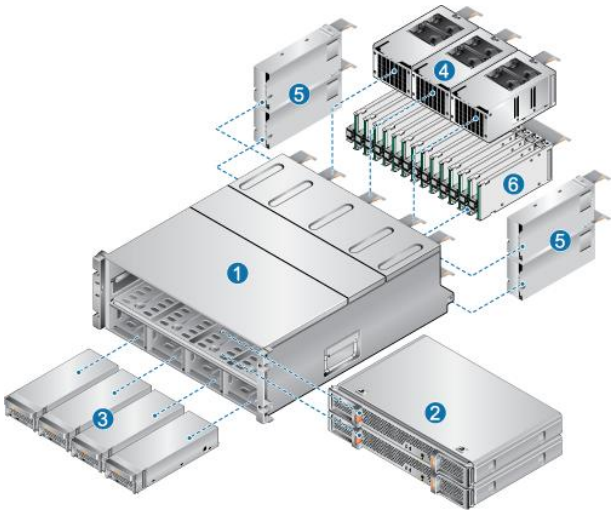


图 3-11 Dorado5100 控制框整体结构图

图 3-12 展示了 Dorado5100 的控制框整体结构，相关冗余情况如下：

- 编号 1——系统插框。无源设计，可靠性高。
- 编号 2——控制器。双控设计，支撑高可靠应用。现场可更换。
- 编号 3——BBU 模块。4×BBU 设计，充分支撑各种异常掉电处理流程，确保系统稳定运行。
- 编号 4——风扇模块。3 风扇强力散热，16 档位智能调速。
- 编号 5——电源模块。4×电源设计，有效防止各种线路故障，提供最可靠的电力保障。
- 编号 6——接口卡模块。可以根据需要，选配不同种类的接口卡，如 FC 接口卡、SAS 接口卡等。

软件可靠性

除了存储阵列常见的一些可靠性措施之外，如双控双活、RAID 保护、全局热备、在线升级等，Dorado 全闪存阵列还针对 SSD 特有的产品特点和故障模式进行了有针对性的可靠性提升。

通过对业界以及华为自己出货的大量 SSD 进行统计、失效分析，总结出导致 SSD 失效的两个原因：

- 闪存颗粒失效
- SSD 固件缺陷

关于闪存颗粒失效的问题，虽然无法完全消除闪存颗粒的失效率，但是通过阵列 RAID 保护和修复，可以有效降低闪存颗粒失效带来的 SSD 失效。

关于 SSD 固件缺陷，华为全闪存阵列 Dorado 所采用的自研 SSD，当前已经发展到第三代，而且每一代 SSD 都是在上一代的基础上进行演进式开发，极大程度上避免了上一代产品的各种设计与实现层面的缺陷。

Dorado 全闪存阵列通过与自研 SSD 进行深度配合，实现了以下一些提升可靠性的特性：

- **智能坏块修复**，从 SSD 内部直接获取所有失效的区域，利用阵列 RAID 进行数据修复；
- **全局容量冗余**，利用阵列内其他 SSD 的冗余空间来容忍个别 SSD 的多片闪存颗粒失效；
- **全局反磨损均衡**，通过全局磨损均衡来尽量延长所有 SSD 的使用寿命，通过全局反磨损均衡来避免同一个 RAID 组内多块 SSD 同时失效；
- **SSD 错时运行**，考虑到软件的很多错误是由于计数器溢出所导致，且这类错误在开发阶段难以发现，Dorado 实现了所有 SSD 的错时运行，让所有 SSD 的运行周期相互错开，避免计数器失效导致的批量故障；

业界对 SSD 的年失效率进行统计，年失效率约为 0.44%。而华为自研 SSD 通过与 Dorado 的深度配合，统计得到的 Dorado 全闪存阵列 SSD 年失效率仅为 0.29%，仅为业界平均水平的 65%左右。

可用度计算

下面对 Dorado 全闪存阵列的可用度进行分析和计算。表 3-6 提供了 Dorado 所采用的各种器件的可靠性数据。

表 3-6 Dorado 所采用各种器件可靠性数据

	单元名称	FITs	MTBF（小时）	MTTR（小时）	可用度%
控制框	控制器	2,500	400,000	0.5	99.99988
	控制框背板	150	6,666,666.7	4	99.99994
	风扇	1,000	1,000,000	0.1	99.99999
	电源	1,000	1,000,000	0.1	99.99999
	BBU	1,000	1,000,000	0.1	99.99999
硬盘框	级联模块	400	2,500,000	0.2	99.99999
	硬盘框背板	150	6,666,666.7	4	99.99994
	风扇	1,000	1,000,000	0.1	99.99999
	电源	1,000	1,000,000	0.1	99.99999
SSD	硬盘单元	331	3,021,148	1	99. 99997

Dorado 阵列的可靠性模型中共包含三部分：控制框、硬盘框、RAID 组，于是：

Dorado 可用度 = 控制框可用度×硬盘框可用度×RAID 组可用度

为简单起见，对超过 1+1 冗余的部件，统一按照 1+1 冗余进行计算。

■ **控制框可用度分析**

双控制器可用度 A1 = 1 -（1 - 控制器可用度）×（1 - 控制器可用度）= 99.99999%

控制框背板可用度 A2 = 99.99994%

双风扇可用度 A3 = 1 -（1 - 风扇可用度）×（1 - 风扇可用度）= 99.99999%

双电源可用度 A4 = 1 -（1 - 电源可用度）×（1 - 电源可用度）= 99.99999%

双 BBU 可用度 A5 = 1 -（1 - BBU 可用度）×（1 - BBU 可用度）= 99.99999%

控制框可用度 = A1×A2×A3×A4×A5 = 99.99990%

■ **硬盘框可用度分析**

级联模块可用度 A1 = 1 -（1 - 级联模块可用度）×（1 - 级联模块可用度）= 99.99999%

硬盘框背板可用度 A2 = 99.99994%

双风扇可用度 A3 = 1 -（1 - 风扇可用度）×（1 - 风扇可用度）= 99.99999%

双电源可用度 A4 = 1 -（1 - 电源可用度）×（1 - 电源可用度）= 99.99999%

硬盘框可用度 = A1×A2×A3×A4 = 99.99991%

■ **RAID 组可用度分析**

假设系统采用 5+1 的 RAID-5 配置，那么共有 C(6,2)种失效组合，于是：

RAID 组可用度 = 1 -（1 - 硬盘单元可用度）×（1 - 硬盘单元可用度）×C(6,2)= 99.99999%

■ **Dorado 全闪存阵列可用度分析**

以 Dorado5100 为例，假设配置为 1 个控制器、4 个硬盘框、96 个 SSD，共计 16 个 RAID 组，于是：

Dorado5100 可用度

= 控制框可用度×硬盘框可用度 ^{4}×RAID 组可用度 ^{16}

= 99.99970%

达到了 5 个 9 的企业级高可靠要求。

网络可靠性

■ **双交换组网与多路径软件**

Dorado 系列固态存储阵列采用双控双活模式，允许用户同时连接到两台交换机构成双交换组网，如图 5-1 所示，服务器到 Dorado 系列固态存储阵列形成了 4 条路径，互相冗余。

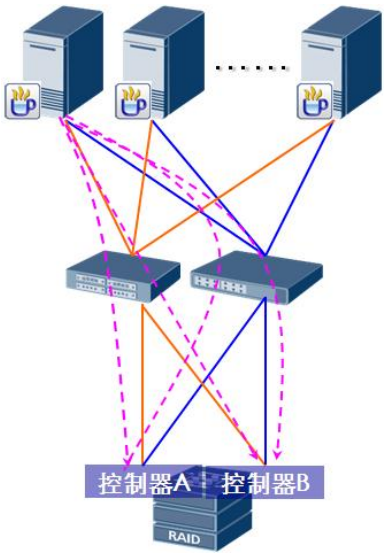


图 3-12 双交换组网示意图

Dorado 系列固态存储阵列采用华为多路径软件 UltraPath，当采用多路径组网时，UltraPath 安装在服务器上，为服务器访问硬盘阵列提供多条路径，以达到更高可靠性和性能。

UltraPath 主要功能如下：（1）避免操作系统看到多份相同的物理硬盘。（2）具备 failover 功能，当主路径出现故障后自动将业务切换到备用路径上，避免了因单点故障而造成业务中断。（3）具备 failback 功能，当主路径的故障解除或修复后自动将 I/O 传输路径重新切换回主路径上。（4）具备负载均衡功能，在所有可达路径上均衡分配 I/O 操作，并支持最短队列算法、最小负载算法和轮转算法调度 I/O 操作。

UltraPath 支持 Linux、AIX 和 Windows 等主流操作系统，支持 Hyper-V 和 XEN 等主流虚拟机。

- **虚拟快照技术**

Dorado 系列固态存储阵列的虚拟快照，支持生成源 LUN 在某个时间点上虚拟的一致性映像，在不中断正常业务的前提下，快速得到一份与源 LUN 一致的数据副本。副本生成之后立即可用，并且对副本的读写操作不再影响源数据。因此通过快照技术就可以解决如在线的备份、数据分析、应用测试等难题。

- **异步远程复制技术**

远程复制是数据镜像技术的一种，一般分为同步和异步两种执行模式，Dorado 系列固态存储阵列实现了异步模式。它能够在两个或多个站点维护若干个数据副本，利用长距离来避免灾难发生时的数据丢失。 异步远程复制会用到虚拟快照技术来提供瞬时数据集和失败还原点。

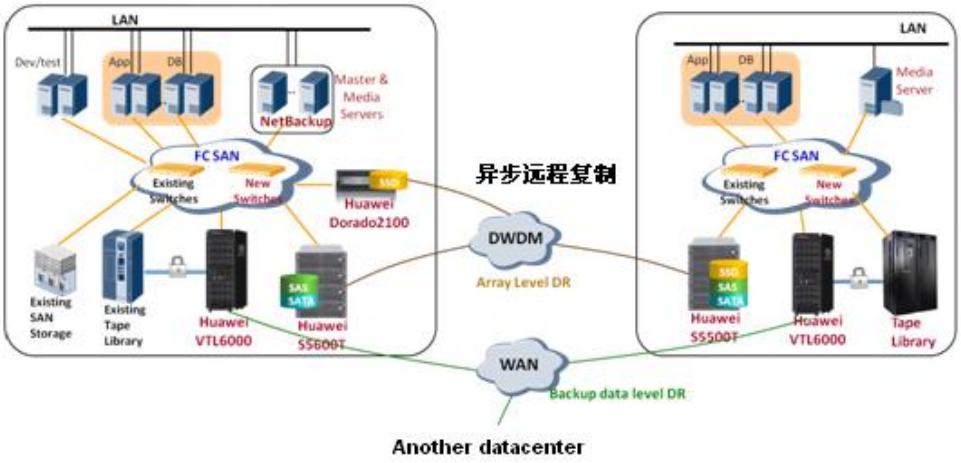


图 3-13 异步远程复制方案示意图

3.4.2 寿命

2.2.6 节对 SSD 的寿命做了简单介绍，从表 2-3 中可以看到，华为自研的各种 SSD 的使用寿命，在给定的极大压力负载下，均满足企业级的寿命要求。

业界对于 SSD 的使用负载，也有相关的定义。JEDEC（Joint Electron Devices Engineering Council），是具有全球领导地位的微电子产业的标准机构。作为一个全球性组织，JEDEC 并不隶属于任何一个国家或政府实体，其会员构成是跨国性的。目前，JEDEC 主要开发固态技术相关的开放性标准。

JEDEC 发布的 JES D218、JES D219A 标准，定义了测算 SSD 使用寿命的工作负载模型。根据该标准工作负载模型，华为自研 SSD 的使用寿命完全满足各种企业级应用的要求。

除了 SSD 自身的使用寿命可以得到保证外，Dorado 全闪存阵列也开发了相关特性，以提升整个阵列的使用寿命：

- **全局磨损均衡**，在整个阵列的层面实施磨损均衡，避免业务热点导致个别 SSD 提前磨损耗尽；

3.4.3 性能

目前，所有的 Dorado 全闪存阵列型号均参加了 SPC-1 基准测试认证。图 3-12 对比了 Dorado 系列全闪存存储阵列、一些厂商的中端存储阵列、高端存储阵列的 SPC-1 数据。

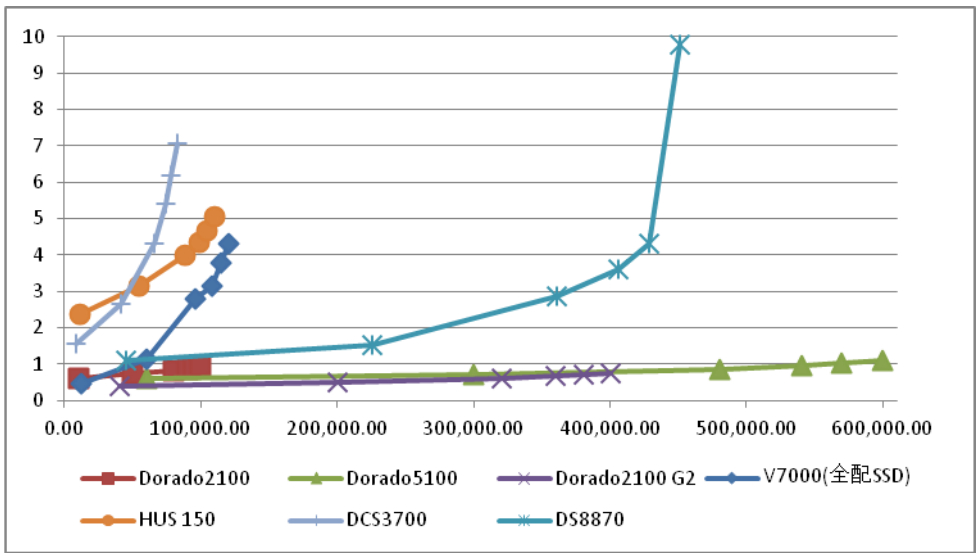


图 3-14 全闪存阵列和传统阵列 SPC-1 测试结果对比

图 3-15 的数据来源于 SPC 官方网站，横轴是 IOPS，纵轴是时延（ms），展示了存储阵列在不同 IOPS 压力下的时延分布情况。

从图中的对比可以看到：

1. 全闪存阵列针对闪存进行设计和开发，主要依靠低时延来提供高 IOPS，这与传统存储阵列依靠堆积硬盘来提供高 IOPS 的方式截然不同；
2. 在中端存储阵列中全部配置 SSD，虽然 IOPS 可以保持较高水平，但是时延很难与全闪存阵列相比较；
3. Dorado5100 和 Dorado2100 G2 的 IOPS 值，已经与高端存储阵列相当甚至高出一大截，同时时延一直保持在微秒级别，低出传统高端存储阵列 1～2 个数量级；

为了提供图 3-15 所示的低时延和高 IOPS，Dorado 系列全闪存阵列针对闪存的特点进行了专门的设计和开发：

- **改写 CACHE 算法**，3.1.1 节提及的 SSD 与 HDD 之间的性能差异，导致传统存储阵列上针对 HDD 所设计的各种 CACHE 算法不适用于 SSD，Dorado 选择了内存消耗较大、但是更简单、CPU 消耗更少的页表结构来组织 CACHE，并简化刷盘、淘汰算法，以空间换取时间和 CPU 利用率，进而支撑更低时延和更高 IOPS；
- **数据平面和管控平面从物理上分离**，充分利用硬件的各种特性对业务数据的处理进行加速，进而释放 CPU 以处理更高的 IOPS；
- **全局垃圾回收**，利用阵列的 cache，尽量缓存住主机的业务数据，轮流为所有的 SSD 提供时间窗口，任意 SSD 在给定的时间窗口内，不会收到写 I/O，从而可以很好地进行垃圾回收工作，减少对阵列的性能影响；（开发中）

4 推广/Experience

	特点	优势	不足	适用场景
传统阵列混插SSD	HDD 盘处理 IO 的方式是顺序大块读写，而绝大部分应用 IO 都是随机的。传统阵列针对 HDD 盘设计，缓存算法以及 IO 下盘算法必须确保将随机的 IO 整合成大块顺序 IO，带来算法复杂度高，时延高等问题。SSD 盘在传统阵列中被当做一类 HDD 处理，其擅长随机 IO 的特性得不到发挥，复杂的算法将其低时延的优势抹杀。	分级存储，综合成本较优	1.可以起到一定的应用加速效果，但是无法将 SSD 盘性能优势充分发挥。 2. 传统阵列中无法针对 SSD 做可靠性设计以及磨损均衡设计，SSD 的可靠性和寿命方面不如固态存储好。	少量应用加速分级存储
PCIe SSD	将 SSD 介质以 PCIe 插卡形式置于服务器中，是存储介质离 CPU 最近的方式。此种方式导致维护复杂，扩容麻烦，不可共享，外置存储的一切便利和可靠性设计均被抛弃。	目前已知最高性能的存储部署方式。	1.单点故障，可靠性低。 2. 维护需断业务停机开箱，非常不便 3. 无法扩容、无法共享。	可靠性低和维护、扩容、共享不便的特点决定此类产品仅适合不担心数据丢失的非核心应用系统以及分布式计算的系统。
全闪存阵列	SSD 置于专为此类介质设计的全闪存阵列中，充分发挥 SSD 性能，并最大限度确保可靠性和 SSD 寿命。提供可靠、易维护的高性能存储。	1. 高性能 2. 高可靠 3. 按性能计算成本低 4. 绿色节能	按容量计算成本较高。	适用高性能、高可靠性要求的核心业务系统。尤其是数据库应用

5 结论/Conclusion

华为公司始终致力于为用户提供高品质的存储产品及人性化的服务，Dorado 系列全闪存阵列产品始终秉承这一理念，为客户提供低时延、高 IOPS、高可靠、简单易用的产品，帮助客户最小化总体拥有成本，并最大化提升客户自身业务竞争力。

6 缩略语表/Acronyms and Abbreviations

表 6-1 缩略语清单

英文缩写	英文全称	中文全称
LUN	Logical Unit Number	逻辑单元号
RAID	Redundant Arrays of Independent Disks	独立硬盘冗余阵列
SCSI	Small Computer System Interface	小型计算机系统接口
SAS	Serial Attached SCSI	串行 SCSI
RAM	Random Access Memory	随机访问存储
PCM	Phase Change Memory	相变存储
SSD	Solid State Drive	固态硬盘
HDD	Hard Disk Drive	机械硬盘
MTBF	Mean Time Between Failure	平均无故障时间
FIT	Failure In Time	失效率
AFR	Annual Failure Rate	年失效率
MTTR	Mean Time To Repair	平均修复时间
OLTP	On-Line Transaction Processing	联机事务处理系统
OLAP	On-Line Analytical Processing	联机分析处理系统
CAPEX	CAPital EXpenditure	基本建设费用
OPEX	OPerating EXpense	运营成本
TCO	Total Cost of Ownership	总体拥有成本