技术 Technology

针对数据库优化

应用背景

现代 CPU 的计算能力在过去 20 年在以超越摩尔定律的速度在发展，速度提高了近 1 千倍。而基于磁记录的机械硬盘的存储 IO 读写速度在这期间增长了不到 10 倍。在 高性能的数据库系统中,性能瓶颈逐渐主要集中在存储 IO。这份技术白皮书描述了原 生 PCIe SSD 的架构和性能优势，包含数十微秒级别的读写延迟和高达百万 IOPS ; 突破了传统数据库系统中的存储 IO 瓶颈, 为大幅度提高数据库的响应速度和实时业 务处理能力提供了坚实的平台。

现代数据库系统包含传统的关系数据库和新兴的非关系数据库，应用包括 OLTP，数据 仓库等。数据库应用的特殊性对存储系统提出了严格的要求：

- 数据完整性。写入的数据在掉电或系统崩溃时必须保证写入介质中，因此数据库应用会非常频繁地 fsync 或使用直接写入（direct write）。这大大降低了系统写缓冲的效能，对写延时提出了非常高的要求。
- 小块离散访问。数据库系统对数据文件的访问是小块（典型为 8k~16k）的，离散的，近 似随机的，大大降低了系统预读和读缓存的效能，对随机读延时有非常高的要求。
- 高并发度。IOPS 应随并发度的增加线性增长，保证性能不会恶化。
- 高可用性。短期离线或性能降低均是不可接受的。

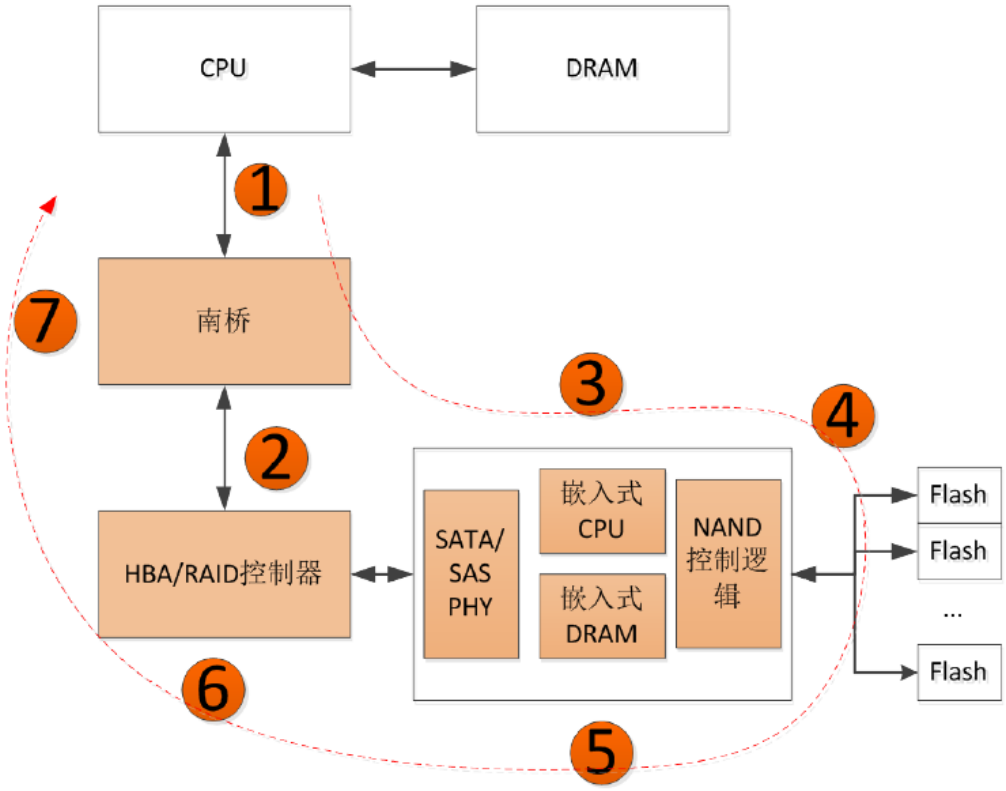
现有解决方案及不足

当前,很大一部分的数据中心架构解决方案都是基于磁记录硬盘(HDD)而设计并优化。 传统的 HDD 由于受机械臂寻址读写的限制，随机读取数据时需要磁头机械定位，即使 高端 15K 基于 SAS/FC 接口的 HDD 的每秒钟完成的随机读写次数 IOPS 通常限制在 200 次左右。

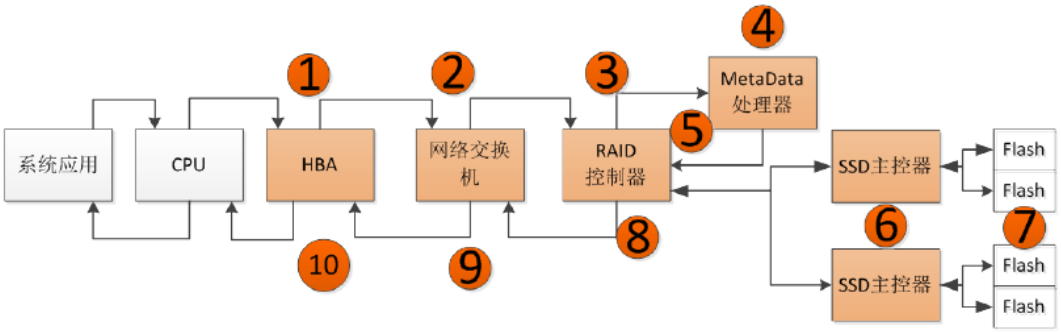
SSD 的出现为突破性能瓶颈带来了可能：伴随着 SSD 技术的成熟，SSD 开始大规模地被数据中心所采用。为了兼容已有的数据中心架构，很多 SSD 采用了与 HDD 相同的接口， 如 SAS(Serial Attached SCSI)或者 FC (Fiber Channel),并在物理形式上与 HDD 相 似，如 2.5 英寸或 3.5 英寸。采用传统硬盘接口的 SSD 直接在现有的系统中取代 HDD 可以在一定的程度上缓解存储系统的瓶颈，改善系统效能。普通的 SSD 采用模拟 HDD 的方式在兼容性上有一定的优势，但在性能上有非常大的限制:

- 复杂的数据链路导致高 IO 延迟
- SSD 性能受嵌入式 CPU 限制
- 冗余的 IO 调度

如下图所示：



尤其是在应用了存储区域网（SAN）后，由于在访问存储介质中所需经过的多道协议和 场景转换，导致延迟损失更加严重，并不能在最大程度上发挥闪存的性能。

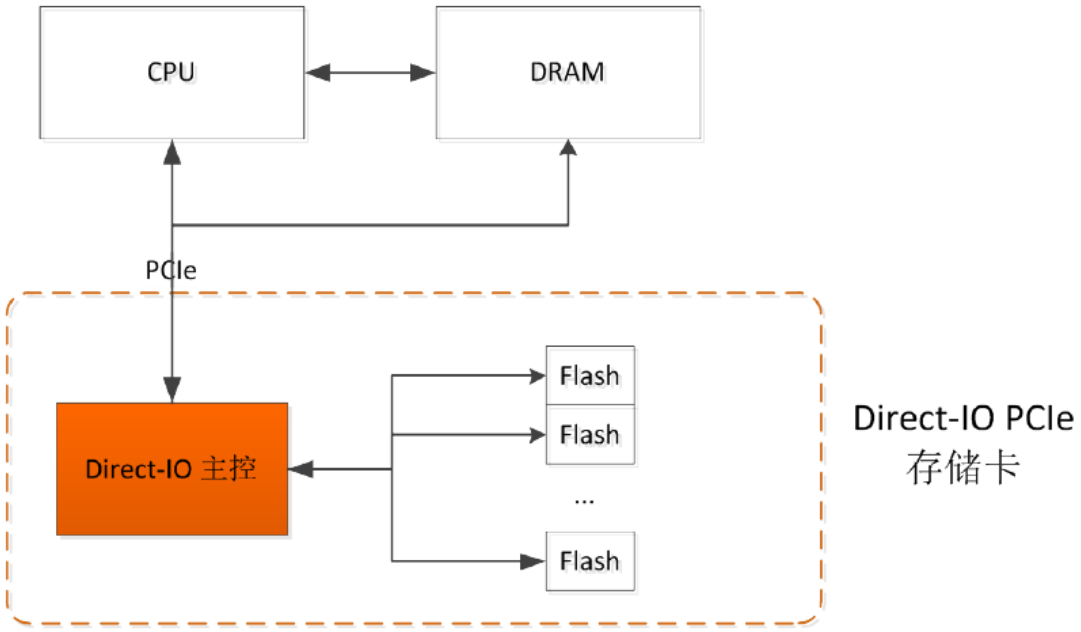


Shannon Direct-IO PCIe 固态存储卡

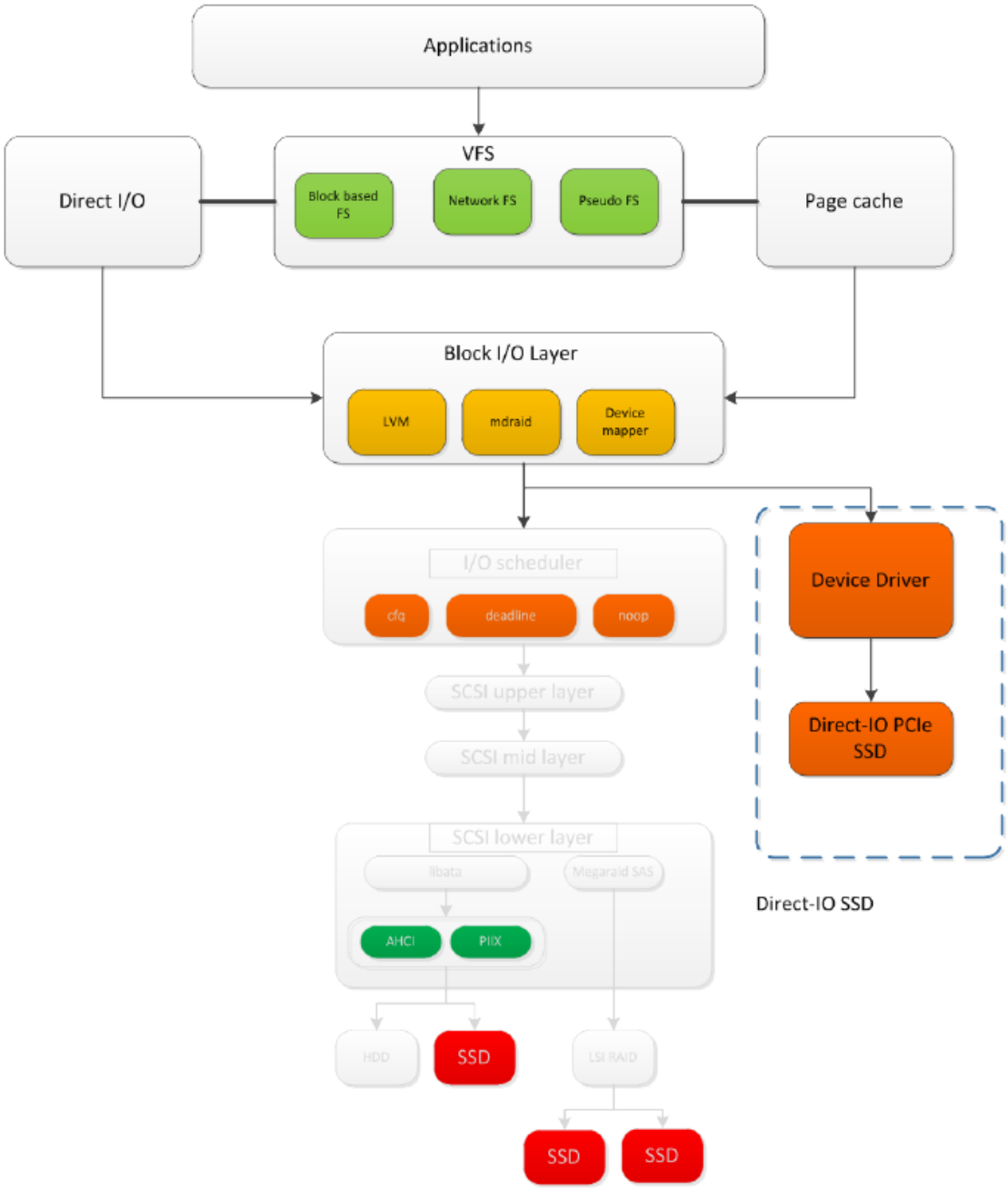
Shannon Direct-IO PCIe 固态存储卡基于原生 PCIe 接口，将 NAND 闪存直接接入 PCIe 总线而实现应用对闪存的直接读写访问。由于去除了传统 SSD 的协议转换及 RAID 控制器，网络协议转换等等系统开销，数据的存储读写延迟 大为削减。在 Direct-IO PCIe SSD 中，闪存的

快速读写性能可以得到最大程度上的利用，充分发挥 CPU 的性能 并最终提升应用的响应速度。

下图描述了 Direct-IO PCIe 固态存储卡的系统架构。Direct-IO 卡采用了原生 (native)PCIe 的主控器，绕开了传统的 SATA/SAS 存储协议，并充分利用主机 CPU 强大的 处理能力对闪存进行直接读写。



软件栈得到了极大的简化，如下图所示：



原生 PCIe 接入

Direct-IO 卡通过 PCIe 总线将 NAND 闪存直接接入系统 CPU。上层应用通过 DMA(Direct Memory Access)直接访问闪存完成数据的实时读写。Direct-IO 主要由两部分组成：软件驱动层和硬件主控层。软件驱动层运行在主机的操作系统内核，利用 Host 稳定可靠的 CPU 处理能力和高速内存完成对块设备 IO 及闪存的管理。 硬件层则完成 DMA 数据的编解 码、数据流控制及对闪存的并行化读写操作。

Direct-IO 卡对闪存的读写采用了类似于 CPU 对内存 DRAM 的读取存储方式， 去除了冗余的传统 SAS 或 SATA 存储协议，并完全避免使用可靠性相对较低的嵌入式 CPU 和 DRAM, 大大地减少了系统开销和可能失效点数目。SCSI 协议是为了兼容传统旋 转机械式硬盘的深度延迟而设计优化的存储协议。由于闪存并不存在机械硬盘的延迟 特性，使用 SCSI 协议管理读写闪存不但有意义而且会大幅度增加系统开销和延迟。 Direct-IO 卡以块设备的接口直接将闪存提供给文件系统，Volume Manager 和上层应用，其简洁的体系架构不仅最大程度上减低了系统延迟，而且降低 可能失效的节点，形成了高效和可靠的系统架构。

利用主机强大的 CPU/DRAM 功能

闪存由于其不可覆盖写的特性，SSD 需承担一定的管理工作，包括逻辑地址和物理地址 的相对转换，垃圾回收，磨损均衡，故障恢复等。普通 SSD 使用嵌入式 CPU 及其固件完 成以上管理，性能比当前的主机 CPU 低一个数量级。Direct-IO 卡借助于主机强大的 CPU 处理功能实现对闪存的管理，把逻辑物理映射表存储在主机的内存中。应用读写 闪存时对内存中的映射表直接查找,然后通过 Direct-IO 卡硬件完成对相应闪存页面读 取之后，数据将通过 DMA 直接返回。中间免除了嵌入式 CPU 的操作、协议开销及数据 的重复拷贝。由于去除了众多中间环节，Direct-IO 卡的读延迟仅在 Flash 固有的延时 基础上增加 15 微秒，写延迟更低至 10 微秒以内，这是基于嵌入式系统的 SSD 无法企及的。

高可靠高容错的数据存储

作为企业级的数据存储设备，Direct-IO 卡充分利用闪存的优点特性并弥补其不足，对 于存储在闪存中的数据采用了多种安全设计方案确保数据的安全性和持久可恢复性。 除了普遍使用的企业级纠错，数据随机化，动态磨损均衡等措施，还使用了 Flash RAID，端到端校验和和完善的突发断电保护机制，为用户提供了高容错数据恢复机制 和严格的数据完整正确性校验。具体详见 数据安全性第一白皮书

优异的性能

由于采用了以上所述的技术，Shannon Direct-IO PCIe 固态存储卡展现了优异的性能：

- 业界最低的读写延时
- 业界最高的随机 IOPS
- 业界最高的单主控闪存容量

具体性能详见 Direct-IO PCIe SSD g2

应用模式

Direct-IO PCIe 卡最可以有多种应用模式：

- 主存储模式
- 数据分区模式
- 缓存模式
- 分层模式

主存储模式

Direct-IO PCIe 存储卡采用的是标准块设备接口,可以直接作为主存储设备存放整个数 据库。Direct-IO 卡提供高达 50 万 IOPS 每秒(4KB)。相对于传统的 HDD/SSD 阵列有十倍 至百倍提高。而且数据延迟由数十毫秒至数百毫秒量级降至微秒量级。在容量方 面，Direct-IO 卡提供高达 3.2TB 数量级的存储容量,还可以通过软件 RAID 将多块卡绑定 以提供更高的容量和吞吐性能。

数据分区模式

如果数据库的规模较大，一种更经济的使用方法是将常用的数据如 TempDB，索引，日 志及常用表等放在 Direct-IO PCIe 存储卡上以加速访问，对一般非热点数据仍放在传 统阵列上以满足容量需求。

缓存模式

Direct-IO PCIe 存储卡可以作为缓存与后端阵列组成集成的混合存储方案。采用 FlashCache 软件，常用热点数据被存储在 Direct-IO 存储卡中，从而可以大幅度的增 加数据访问速度，降低后端存储的负荷。

分层模式

如果应用的冷热数据转换速度较慢，可以使用较简单的自动分层（auto-tiering）技 术实现混合存储，以较低的代价达到缓存模式的便利程度。

总结

基于高容错的硬件和软件混合一体架构，Shannon Direct-IO 存储卡通过高可靠的数 据硬件读写通道和具有多重保护机制的软件栈设计充分地发挥 Flash 的读写性能潜 力，把数据存储的可靠性和可用性能提升到了一个全新的高度。Direct-IO 固态存储卡 直接通过 CPU 的 PCIe 总线提供给数据库应用高达接近 3GB/s 的随机数据访问带宽和 低至数十微秒的读写延迟，大幅度提多核 CPU 的利用率，从本质上加速数据库的处理 和响应速度。无论是数据库 OLTP，或应用报表，数据分析，或者 BI，Direct-IO 卡将大 幅度提升系统性能密度，加速应用速度和降低系统响应时间，并同时减低系统运行开销 （OPEX）和总拥有成本（TCO）。

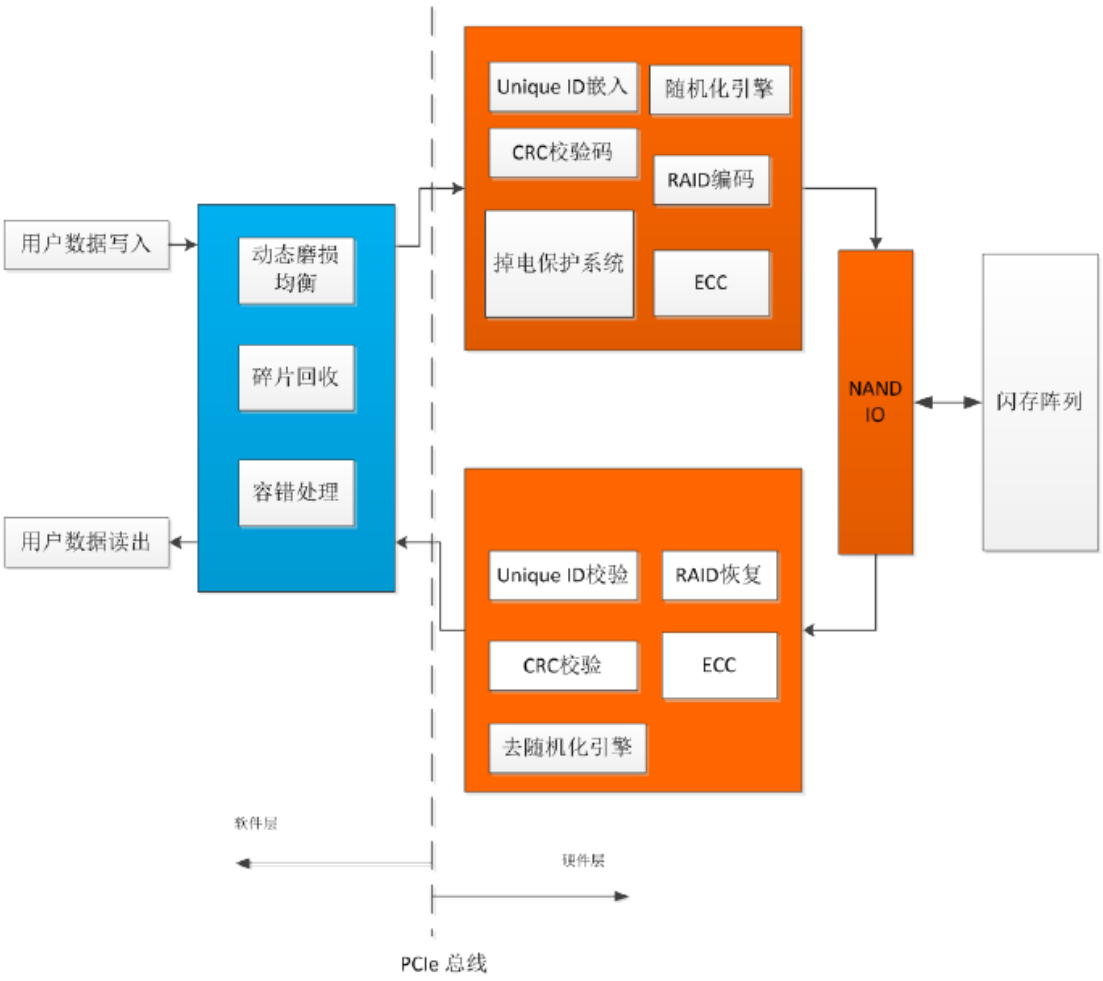
数据安全性第一

引言

Shannon Direct-IO PCIe 固态存储卡是专为解决数据中心及高性能服务器中的存储 IO 瓶颈，提升系统应用性能和响应速度而定制的一款高性能企业级存储设备。对于企 业级存储设备来说，数据的安全性是第一位的考虑，Shannon Direct-IO 卡基于精简的 混合系统架构，在提供卓越数据 IO 性能的同时，采用具有多重容错机制的高效算 法，使得应用数据在写入存储和读出恢复过程中都受到多重安全机制的保护及容错处 理。这份技术白皮书主要从体系架构，数据的安全存储，恢复及读写等几个方面 展开描述了 Direct-IO 固态存储卡在数据安全性方面的设计思想和应用。

精简的混合体系架构

Direct-IO 卡通过 PCIe 总线将 NAND 闪存直接接入系统 CPU。上层应用通过 DMA(Direct Memory Access)直接访问闪存完成数据的实时读写。下图描述了 Direct-IO 卡的体系架 构。Direct-IO 卡主要由两部分组成，软件驱动层和硬件主控层。软件驱动层运行在 主机的操作系统内核，利用主机稳定强大的 CPU 处理能力完成对块设备 IO 及闪存的管 理。硬件层则完成 DMA 数据的编解码、数据流控制及对闪存的并行化读写操作。



Direct-IO 卡对闪存的读写采用了类似于 CPU 对内存 DRAM 的读取存储方式，去除了冗 余的传统 SAS 或 SATA 存储协议，并完全避免使用可靠性相对较低的嵌入式 CPU 和 DRAM, 大大地减少了系统开销和可能失效点数目，形成了高效和可靠的系统架构。

数据的安全存储

下图描述了 Direct-IO 主控器和驱动软件的主要功能模块，其中包括 FTL(Flash Translation Layer)层,掉电保护系统,和 Flash RAID 等。控制器的各个模块相对独 立并形成功能互补关系。应用层的数据在写入，存储，读出及传送过程中受 到多重安 全机制的保护。其中，为保护数据在闪存中的可靠和持久的存储，Direct-IO 卡采用了 包括数据随机化，BCH 纠错码，动态磨损均衡等各项处理措施。

数据随机化

闪存对于某些特定的数据样式具有一定的敏感性：例如全 1 或全 0 的数据在闪存中 的存储可靠性就相对较差。值得指出的是,全 0 或全 1 或其它固定样式的数据在实际 应用中却经

常出现。Direct-IO 卡在用户数据写入时采用了独有的随机化引擎，将用户 数据完全动态随机化后再存入闪存介质，在最大程度上减低了对闪存的磨损，保正了 用户数据的存储最优可靠性。

BCH 纠错码

闪存在数据读写的过程中会出现偶发性的随机错误，其误码率与数据在写入闪存后 的存储时间和所在物理块经历的读写擦除次数有关。Direct-IO 卡采用业界主流的 19nm 的高品质闪存和高可靠的 40bit/1KB BCH 纠错码，充分满足 NAND 对纠错码的要求。

动态磨损均衡

为最大化系统的使用寿命，保证数据有效性，Direct-IO 卡采用了动态智能的磨损均衡 技术：去除由于热点数据造成的物理块热点，使得闪存中每一物理块的擦写次数都得到 接近充分的利用。同时，在系统设计中 Direct-IO 卡对磨损均衡算法做了全面优化， 以最小化由于磨损均衡造成的数据拷贝，从而最大程度上减小了写放大因子，进一步延 长了系统的使用寿命，增强了可靠性。

数据监测与更新

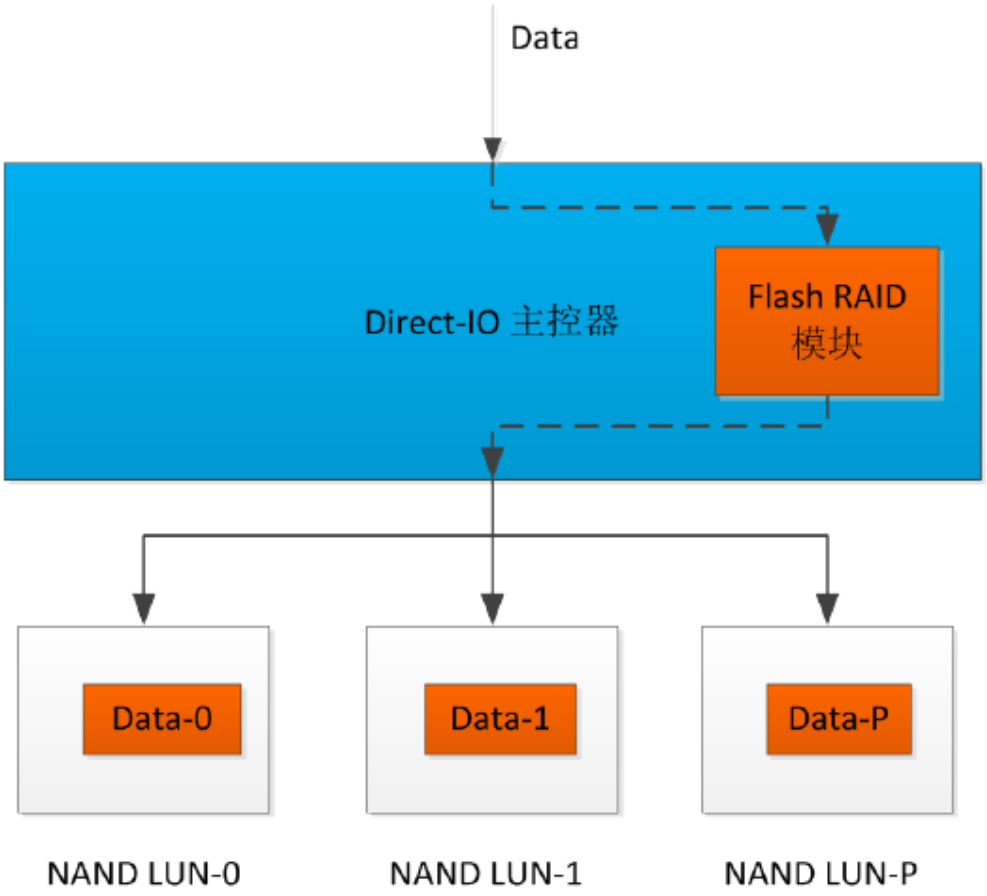
随着时间的推移，闪存单元中 Floating Gate 的电荷泄漏会造成存放在闪存中的数 据误码率上升，可靠性的下降。同时，应用对闪存物理页面的读写也会对其相邻页面的 数据造成读干扰和写干扰，导致误码率的上升。在 Direct-IO 的设计中，所有数据的 误码率会被实时监测，一旦某一页面的数据可靠性低于设定的预警值, 该数据页面将 被自动刷新。这一动态刷新机制在最大程度上减小了数据出错或丢失的可能性。由于 拷贝刷新的过程是在背景中执行的，这些操作对用户而言完全透明。

Flash RAID

在以上的安全存储措施之上，Direct-IO 卡中进一步采用了专门针对 Flash 应用而优化 设计的 Flash RAID 算法。在下列极端情况出现导致数据读取错误时仍能恢复数据：

- 随机出错比特数超过 ECC 的纠错范围。
- 物理页面读取错误。
- 闪存物理块失效。
- 整个物理 LUN 失效。

虽然这些事件发生的概率极小，但处理不当则会引发数据的丢失。利用 Flash RAID，Direct-IO 卡会在数据写入时基于多个用户数据包生成奇偶校验码并对其进行关联。 所生成的奇偶校验码将与相关联的用户数据一并写 入闪存，并分别存储在不同的 NAND LUN 中，如下图所示。



如数据由于任何原因读取错误，正确数据仍能从校验码和其他用户数据中恢复。如果 物理页面被判定为不再可靠，这些物理地址将会被标记成坏块，不再被使用。所有这 些坏块管理及出错处理 Direct-IO 卡自动完成，对于用户而言完全透明。

端到端数据校验

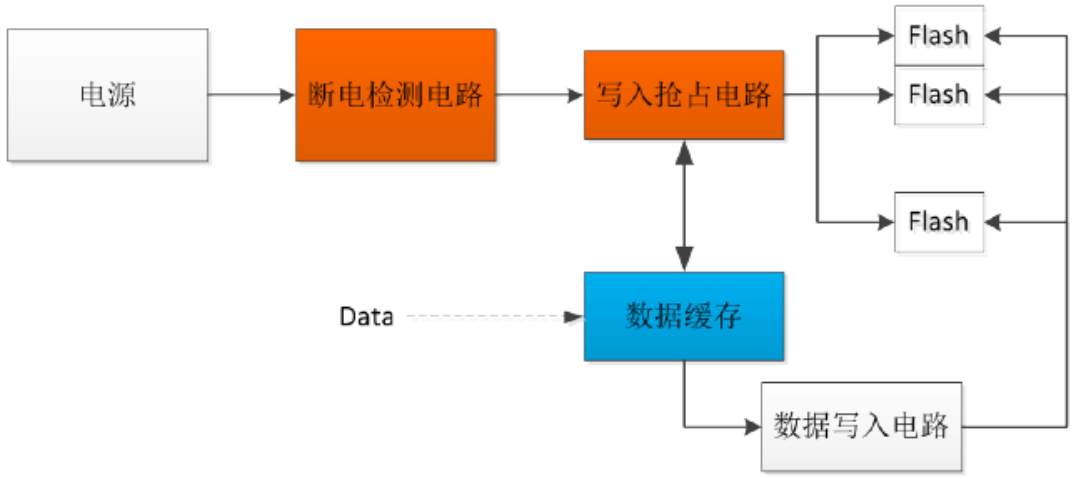
为保护误纠错，软件错误或数据流程内其他部分出错，Direct-IO 卡在数据写入时的第 一个环节产生 CRC 校验码，并将这一校验码与数据一并存入。同时，软件层也产生一 个唯一标识 ID, 并将该 ID 与数据一并存入闪存中。在数据读出时，CRC 码和唯一标识 ID 在不同层面独立校验，从统计意义上消除 ECC 的误纠错可能性，并且确保读出时数 据与写入时的完全一致性和完整实时性。

突发掉电数据保护

在突发掉电情况发生的时候，一个企业级的存储系统需要保证写缓冲区中的数据写入 到非易失的存储介质中，并在下一次开机后保证数据的完整可靠。在 Direct-IO 卡系 统设计中，我们采用了下列手段：

- 分布式的冗余元数据存储，保证在任意时刻断电都没有过多元数据需要存储。
- 元数据和用户数据协同存储，保证内部一致性。
- 使用快速的少量写缓冲区，降低紧急写入时的压力。
- 使用全硬件的电源监控电路，降低紧急写入的延时和风险。

下图描述了 Direct-IOTM 卡的突发断电数据保护机制。断电检测电路模块实时监控存 储卡的电源供给情况，一旦检测到电源供给产生突变，写入抢占电路会被启动。这时 所有其它的操作，例如读操作会被中断而数据通道中的数据会被迅速写入闪存之中, 保障数据存储安全完整性。



采用上述机制后，Direct-IO 卡的紧急写入过程至多不超过 5 毫秒，不需要使用超级电 容或者备份电池，最大程度上减少了系统维护开销，减少了失效点，从而增强了系统 可靠性和耐用性。

过热保护

如果系统过热，电路可能失效，Flash 中存储中的数据也有可能不稳定。在 Direct-IO 卡上安装了多个温度传感器，并设计了纯硬件的温度保护逻辑，在温度上升到第一警 戒点时，自动降低系统速度，以防止温度进一步升高。如果温度进一步升高至第二警 戒点时，所有读写操作暂停以防止电路过热损坏及影响 Flash 中数据有效性。同理，当 温度下降后，所有操作自动恢复进行。全过程无须用户干预，对于用户而言完全透明。

总结

基于高容错的硬件和软件混合一体架构，Shannon Direct-IO 存储卡通过高可靠的数 据硬件读写通道和具有多重保护机制的软件栈设计充分地发挥 Flash 的读写性能潜 力，把企业级的数据可靠性和可用性能提升到了一个全新的高度。例如,1.2TB 的 Direct-IO 卡提供高达近 16PB 的总数据写入容量。在正常使用环境下每天容纳多 达 6.5TB 的写入数据, 24x7, 持续 5

年。多重保护和智能预判有助于消除和缩减宕 机时间，延长系统寿命，从而减低系统运行开销（OPEX）和总拥有成本（TCO）。