EMC Symmetrix：潜在的数据丢失问题

北京深思计算机系统集成技术有限公司 李凤章 博士

概述

多年以来，IBM和其他存储分系统供应商以及独立业界咨询顾问都提醒大家，某些磁盘分系统中的高速缓存设计，在单一主缓存部件发生故障时非常有可能造成数据丢失现象。由于数据丢失有可能对业务造成潜在的严重影响，所以在客户考虑选择磁盘分系统时应该对该问题予以高度的重视。本文着重讨论了EMC Symmetrix磁盘存储分系统中的缓存设计问题。

出现的问题是什么?

磁盘分系统高速缓存是用于提高磁盘分系统性能的电子存储器，在缓存器中通常要保存主机常读取数据的拷贝，从而避免磁盘旋转和磁盘磁头移动所造成的机械延迟。另外，高速缓存还常用于保存主机写往分系统的新的或变化数据，以避免写请求时的相同延迟。

对写请求进行缓存的磁盘存储分系统要完成以下操作:

1. 从主机接收数据;
2. 在缓存中保存数3/4;
3. 立即向主机报告写请求已经完成，而此时数据还未写入磁盘。

完成以上步骤后，主机应用重新开始其他的处理工作。新的或变化的数据在后续的"下载"(destaging)流程中再写入磁盘。由于新的或变化数据在其下载到磁盘之前要保存在缓存器中，所以保护缓存中的数据不受缓存故障影响就非常重要，这就如同要使用RAID来保护磁盘的数据不受磁盘故障影响一样重要。

保护缓存中的数据不丢失的重要性与响应主机读请求时对从磁盘读入缓存器的数据进行拷贝以提高性能的重要性不同，其原因在于即使是在缓存发生故障时，数据也可以直接从磁盘读入。而当数据从主机写入分系统的缓存但还未从缓存下载到磁盘时，若此时发生缓存故障，那么数据将会丢失。数据丢失后需要使用如恢复备份拷贝或重新运行批处理工作或在线交易等方法才能得到恢复。而存储分系统本身不能完成或帮助完成这些特殊的工作。

当考虑到以下一些常遇到的情况时，丢失数据的影响以及恢复工作的难度都将比原来更大。

(1)若丢失的数据中包括有生产数据或关键业务数据时(对照不重要的数据如临时测试数据或非常容易重建的数据)。

(2)若相互依赖的文件保存在相同的存储分系统上时。例如，若数据同时从数据库和其相关的记录中消失，那么该记录对恢复工作将没有任何价值。

(3)若保存在缓存器中的写入数据量非常大时。在Symmetrix系统中，驻留在缓存器中待下载到磁盘上的新/变化数据有可能在GB量级。

(4)若数据同时从多个磁盘卷、多个应用或多个处理器系统中丢失时。

正是数据丢失对业务和IS部门所造成的潜在影响产生了保护缓存器中新/变化数据的需求。若某公司花费资金使用RAID技术来保护磁盘卷上的数据，那么他没有理由将同样的数据放在甚至临时放在一个或多个故障点就有可能造成数据丢失的存储资源上。

在缓存器故障情况下予以保护

缓存器是一种电子存储器，为了保护缓存器中的数据安全，必须考虑各种有可能遇到的问题。有可能造成缓存器数据丢失的潜在问题包括:

▲ 电源故障

▲ 位错误

▲ 缓存器部件故障

以下将详细讨论这些类型的问题。

▲电源故障

缓存存储器使用的是半导体技术，需要持续供电才能保持数据。为了保护"写缓存器"－即从主机写操作接收数据来改善写性能的缓存器，缓存器设计人员使用电池来保护缓存器中的数据不受外部电源掉电的影响，这就是所谓的"非易失性"。Symmetrix和其他供应商的缓存分系统通常都为写缓存器设置了电池保护。

▲位错误

缓存器通常包括多个存储器模块(即存储器芯片)来保存组成数据的位（1或0）。例如，4 GB的缓存器至少要包含320亿的数据位（每字符或字节8位，乘以40亿字节）。

一位出现错误可能有多种原因，有可能某位的电路被保持住，总是保持相同的值，或者某位的值由于电子干扰而突然反相。

为了防止数据出现位错误，缓存器设计人员在缓存器设计中为数据增加了额外的位，称之为纠错码(也称为错误校验和修正)，即ECC。写入缓存器的数据被划分为数据位组，每一组我们简单地称之为"数据块"，由存储分系统为每一个数据块添加ECC位。例如，对于64位（8字节）的数据，增加8个ECC位来产生72位的数据块。数据块中的各位保存在不同的存储器模块上，这样数据块中的某些位出现问题，甚至是整个存储模块发生故障时，都能通过对可读取数据和相同数据块中ECC位的处理，利用算法重建正确的数据。

Symmetrix缓存器提供了对1位和2位误码的ECC校正，可以检测（但不能校正）3位或更多位的错误。Symmetrix采用了一种内部缓存器"擦洗"方法来周期性地扫描缓存器，并对检测出的误码进行校正，这样减小了误码积累超过ECC纠错能力的可能性。擦洗技术的有效性取决于一些参数如缓存器大小、时序关系以及检测出的误码属性。当然，若数据块中的误码数量超过了ECC的校正能力，数据肯定要丢失（除非使用了其他的保护机制）。

有些说法称Symmetrix的缓存器设计中采用了将一组数据位（例如一个字节）与附加位（称为奇偶校验位）一起保存的方法。当保存数据时，对奇偶校验位进行设置以反映数据各位总和的奇偶情况(取决于所使用的奇偶位设计方法，但从逻辑上讲技术是相同的)。当以后读取各位时，对各位的总数再进行计算，若总数是奇数，而奇偶校验位为偶数（反之亦然），那么可以肯定出现了问题。

不管Symmetrix缓存器是否使用了奇偶校验，基于奇偶校验的设计比使用ECC也没有提供更多的数据保护能力。若数据组中的某一位不能读取(不仅仅是误码)，并且知道该位的位置，那么从理论上讲，奇偶位加上其他可读取的位可以重建有一位不能读取的数据组。但是，若两位或更多的位不能读取，奇偶检验对恢复数据将没有任何帮助。若所有位都可读，但有一位是错误的，奇偶校验可以判断出现了一位错误，但不能断定是哪一位，这样，数据还是不能恢复（即丢失）。若多位发生错误，奇偶校验甚至不能判断是否出现了误码（例如，若1个1变成了0，一个0变成了1，但总值未变）。由此看来，奇偶检验只对某些误码情况提供了误码检测能力，只提供了有限的基本纠错能力或没有纠错能力。

▲缓存器部件故障

非易失缓存器（即缓存器得到电池的保护以防止外部电源掉电所造成的数据丢失）和ECC是计算机行业普遍采用的缓存器保护技术。例如，这些技术在IBM的分系统缓存器中已应用了多年。但正如上述所论述的，这些技术被业界普遍认为不具备足够保护数据不受所有类型缓存器故障影响的能力。

下面再考虑一下Symmetrix中缓存器的封装。Symmetrix缓存器将存储模块安装在存储卡上，然后将一些存储卡安装在存储板上（EMC的业务代表和业界咨询顾问有时在讨论存储板是不提及存储卡），卡和板上包含多种有源部件。数据块中的各位就分布保存在这些模块、卡和板上。

您可能还记得Symmetrix缓存器的ECC能够校正数据块中的1位和2位误码，因此，为了保证单缓存器卡或板的故障不会造成2位以上的数据丢失，数据块中的各位应该分散保存 在足够大量的卡和板上。但是Symmetrix中存储板的数量很少，有的存储卡的数量还不到数据块中各位数量的一半。例如，某些Symmetrix型号仅包含有4个缓存板。由此，任何一块卡或板上都至少包含数据块中的两位，这样，若某一块卡或板（或多个缓存卡或板）发生故障，势必带来数据丢失难以恢复的问题，所有保存在故障卡或板上的新／变化数据将彻底丢失。

数据块跨各存储部件分散保存的原因还包括：这种方法对提高性能也有帮助，主要是因为数据可以跨多个部件进行并行读取。这只是从性能上来考虑，而对数据保护

没有什么明显的优点。将数据分散在多个卡或板上进行保存的方法在业界使用得非常普遍，IBM一直使用这种方法将数据分散保存，但除了认为对性能有好处外，从来没有宣称能明显改进数据保护能力。

据报道，有一些EMC销售代表宣称跨多个存储卡和板来保存各位是一种“条纹化”技术，提供了类似RAID的缓存保护能力。很明显，这是一个错误的比喻，确实是有些RAID技术跨多个磁盘驱动器来进行数据条纹化，但磁盘的RAID技术是保证当整个磁盘驱动器发生故障时不影响从其他剩余的磁盘中重建故障磁盘中的数据。而与之相反，Symmetrix缓存器的设计思想与之不同，并且一个缓存卡或缓存板的故障要导致保存在故障部件中新变化数据的丢失(RAID磁盘保护或者使用镜向来制作数据的另一份拷贝，或者使用与ECC截然不同的逐位奇偶校验方法)。

独立的磁盘驱动器使用ECC来防止没有RAID保护能力的单个驱动器出现位错误，因为当出现整个磁盘故障时，它对客户没有丝毫帮助，所以一般使用RAID技术来帮助保护出现整个磁盘故障时的数据。

为了在实际应用时提供类似RAID的保护能力，存储分系统缓存器设计人员和制造商使用第二个独立的缓存器来保存新和变化数据的拷贝，这能防止出现缓存器部件如卡和板故障所带来的数据丢失现象以及防止出现超过ECC校正能力的多位误码。这种设计方法有时称为"镜向缓存器"或"缓存镜向"或"双写缓存器"。大多数实现写缓存能力的磁盘分系统都采用了这种设计方法。

当对保存在缓存器中的以降低读I/O时间为目的的磁盘数据进行保护时，缓存镜向没有太多的好处，其主要原因是若出于某种原因，缓存器中的数据丢失时，数据可以从磁盘中读出。在这种情况下，由于缓存器故障不会引起数据丢失问题。但缓存镜向对保护缓存器中的写数据，即新的或变化数据而言非常关键。

确实是可以快速下载从而减少了出现问题的机会吗?

据称Symmetrix中的数据可以快速从缓存中下载到磁盘，从而极大减少了出现问题的机会。下面让我们来进一步分析一下。

首先，该说法不说明问题不存在，只是说明快速的下载可以降低数据丢失的可能性。同时，这默认了由于缓存器的故障会出现数据丢失问题的事实。

所有的计算机流程，从指令执行到一个I/O的时间都可以描述为"快"。考虑发送给大型的磁盘分系统的每秒数百数千的I/O中，绝大多数为写I/O的情况，此时新的/变化数据保存在缓存器中，等待处理器发出I/O已经完成的指令。若缓存器内容的瞬像在某一时刻得到，有可能就有一些记录正在等待下载。

在Symmetrix中，等待从缓存器下载的数据要通过内部总线传到磁盘控制器上，然后再到UltraSCSI总线，最终传到磁盘驱动器的缓冲器中。因为很多Symmetrix磁盘配置成RAID 1(镜向)，数据需要移动到两个磁盘控制器和(有可能两个)UltraSCSI总线，特别是一条UntraSCSI总线一次仅能处理一个数据传输，所以下载的写操作将被同一条总线上的其他到磁盘的读或写操作所阻碍，由此，下载的速度受到了UltraSCSI一次一个协议的限制。

从性能的角度来考虑，当向主机发送完I/O结束指令后立即启动下载的方法在效率方面也存在缺陷，其主要原因是这种方法没有利用某些活动文件的优势，而在活动文件中有些数据将得到重复和经常性的更新(例如活动数据库索引)。有些存储分系统的缓存器设计利用了这些优势，其实现方法是在很短的时间内接收对相同数据的连续写入，而不需要逐个下载；特别是在缓存器中相同数据的多个更新数据还能互相覆盖，这样可以降低内部分系统的开销。

很多业界资深分析公司也对EMC产品存在的问题进行了评论。以下对他们的意见进行了汇总。

▲ Gartner集团

EMC的业务代表有时引用Gartner集团的文章当作其Symmetrix缓存器没有单点故障的证据，该文名称为："RAID：永不需要说抱歉"，发表于1996年9月20日。但通过对该文的仔细阅读，您就会发现它并没有说明Symmetrix缓存器能够防止缓存器故障造成数据丢失。该文对缓存器保护有两个主要论点：

1. 出现缓存器故障时，Symmetrix需要处理器操作系统重发数据。

[评论：这似乎仅适用于在主机被通知I/O操作已经结束之前发现错误时。在Symmetrix分系统中，当数据写入缓存器但在其下载到磁盘之前通知主机I/O操作结束，所以，在主机已经得到通知但在数据下载到磁盘前发生了缓存器故障使数据的拷贝不可读取，那么将会发生数据丢失现象。此时似乎不可能再返回处理器，请求它再重

发丢失的数据，处理器也不会在得到I/O结束通知后再保留数据以备分系统"改变想法"]。

2. 即使是双缓存器设计也不是完美无缺的。

[评论：这当然是正确的，没有任何事情是完美无缺的。例如，一种未知的微码缺陷就有可能在给定的分系统中产生一种错误，使系统不能从镜向缓存器中找到数据的其他拷贝。但是对于计算机中的每一种数据保护方法都存在相同的争论，即：它是否出于有意的原因而造成潜在的问题。

IBM认为应该谨慎地设计缓存器，把安全性与构建的实用性和经济性一样重要地来考虑。有些人可能会问："为什么IBM和其他主要存储制造商如日立和富士以及Data General CLARiiON，即使是认为写数据镜向不能提供巨大好处的情况下还仍然花费成本来增加写数据镜向措施？"。确实是，几乎所有主要的缓存磁盘分系统制造商（除了EMC）都提供了双缓存器设计数据保护方法，它不仅用于高可用性一直是最重要客户需求的与S/390连接的存储分系统，也同样用于连接中档服务器的存储分系统。这种方法当然会为缓存器设计增加某些成本和复杂性，但若其没有真正的巨大价值，绝不会有如此多的分系统采用这种保护方法]。

而Gartner集团令人信服地清楚说明了双缓存器的需要：

"只有写缓存器得到了镜向?以及实现了非易失，才能保证数据的完整性"。

▲ Giga信息集团

"?镜向写缓存器?对保证关键业务环境下的性能和可用性非常必要"（Anders Lofgren，Giga信息集团的前分析专家，引自信息周刊文章"理解光纤通道"，1997年12月8日）。

最后说明的一点是，Symmetrix中包括了一种标准方法来向客户报告在出现缓存器故障时的数据丢失问题，该报告并不包括数据本身，只是指出数据已经写入但却丢失的磁盘位置（磁盘位置在缓存器外的独立存储器内维护）。在Symmetrix中提供这种报告机制使大家自然而然的提出一个问题：为什么Symmetrix还为不可能发生的情况提供报告功能呢？

"真正的客户测试"是否能验明实际情况

安装有存储分系统的客户考虑验证他们存储产品中的缓存器是否能抵制单缓存器部件故障的影响。供应商有可能或不可能支持或认可该测试，而离开供应商的支持，客户自己不可能完成该测试。若得到供应商的支持，完成结论性测试的意义当然非常重大，但结论性测试却常常难以完成，其主要原因包括：

1. 只切断分系统的交流电源似乎还不够，因为写缓存器大都采用上述的电池供电来保护不受交流供电故障的影响。掉电与部件故障情况不同。

2. 还可以采用使缓存器板失效的方法来验证在性能下降的情况下分系统是否能继续工作。但请注意，这只是使板失效而不是上面所讨论过的成为Symmetrix单故障点的缓存器板上的独立缓存器卡故障，存储卡一般焊接在板上，因此不可能在存储卡故障的条件下开展测试。

3. 供应商服务代表能够在存储板或卡失效前下载数据，也能在部件失效前使部件以一种可控的方式离线。

4. 由于时序的原因，在存储板失效时，缓存器中不会有写数据。

结论

只有客户自己才能估计出Symmetrix磁盘分系统由于缓存器故障所造成的数据丢失对业务带来的财政和运营方面的影响。当IS部门在评估磁盘分系统时，应该认真考虑缓存器故障所带来的潜在数据丢失问题，以及确定所购买的存储产品是否包含有足够安全的缓存器设计，这有可能是一种谨慎的业务决策方法。■