武汉大学

**海量存储技术结题报告**

学 院： 计算机科学技术

专业班级： 硕士5班

学生姓名： 陈雄

目 录

信 息 I

1 评估基于SSD的RAID系统的可靠性 1

1.1 定量评估模型QRM 1

1.2计算可靠性的方法 - IRM 2

2 Hybrid RAID的设计和提出 4

2.1 Hybrid RAID的设计和说明 4

3.2 如何再次提高Hybrid RAID的可靠性 5

3 评估Hyrid RAID 6

3.1 实验设置的简单介绍 6

3.1 安全时间的长短和死亡时间 7

3.2 安全时间的长短和死亡时间 8

4 总结 10

信 息

论 文 名：Hybrid RAID: A Solution for Enhancing the Reliability of SSD-Based RAIDs

论文来源：IEEE TRANSACTIONS ON MULTI-SCALE COMPUTING SYSTEMS, VOL. 3, NO. 3, JULY-SEPTEMBER 2017

术语解释：

1. SSD： 固态硬盘

2. RAID: 独立磁盘冗余阵列

3. MTTDL: 发生不可恢复数据丢失的期望时间

4. HDD: 机械硬盘

5. BER: 位错误率

6. ECC：错误检查和纠正

**论文背景和问题**

RAID系统由n个数据磁盘组成，m个奇偶校验磁盘最多可容纳m个磁盘故障。奇偶校验单元被放置在专用磁盘上或均匀分布在所有磁盘上。RAID4和RAID5，这两个结构由n个数据磁盘和一个奇偶校验磁盘组成，因此，他们只能容忍一个磁盘故障。RAID4的奇偶校验磁盘在其中的一个专用磁盘上，RAID5中，数据磁盘之间均匀分配奇偶校验单元，P / E周期以几乎相同的比例分配到磁盘之间。

SSD在此不做介绍。尽管SSD由于没有移动/机械元件，而提供比HDD更大的平均数据丢失时间（MTTDL）。但SSD也有其不足，与HDD不同的是，当P / E周期数量增加时，SSD会受到永久和短暂故障的威胁。作为永久性故障，当SSD达到P / E极限时（被称为写入耐力），不能提供期望的存储可靠性。但是HDD中不存在这样一个问题。此外，BER和P / E周期之间存在直接的依赖关系，通过增加P / E周期的数量，BER会呈指数增长。

虽然在RAID系统中应用SSD的好处很多，但是也们带来了新的问题。该文仅仅研究可靠性问题。磁盘阵列中存储奇偶校验单元的分配如何影响RAID的可靠性？通过在RAID4中将一个磁盘用于奇偶校验，奇偶校验磁盘施加更多数量的P / E周期;结果表示一个磁盘（即奇偶校验磁盘）的故障脆弱性要比其他磁盘高很多。而通过在RAID5中的所有磁盘之间均匀地分配奇偶校验单元，所有磁盘都施加相同数量的P / E周期。当P / E周期很小时， BER也较低，因此所有磁盘的可靠性都很高。但是，通过增加阵列中所有磁盘上的P / E周期数，当所有磁盘中的P / E周期数达到耐久极限时，阵列的可靠性将受到极大的威胁。在这种情况下，两个磁盘同时出现故障的可能性很大。在第一张磁盘恢复过程中第二张磁盘出现故障时发生。

SSD的RAID中的数据磁盘的可靠性是通过应用奇偶校验码来实现的。 一般来说，在RAID中分配奇偶校验码单元有两种策略：

1）在所有磁盘之间均匀分配奇偶校验单元;

2）在所有磁盘之间不均匀地分配奇偶校验单元；

这两种策略带来了不同的影响，均匀分配时，阵列中的SSD都已一个相同的速率磨损，最终同时故障的概率变得很大，风险也十分巨大。非均匀分配的时候，其中一个P/E周期最多的SSD会首先发生故障。无论哪种分配，都降低系统的可靠性。

为了提高SSD的可靠性，并评估，该文章做出了两点贡献，可概括如下：

1）提出了一种称为混合RAID（Hybrid RAID）的基于SSD的可靠结构。

2）提出了一个定量模型来评估Hybrid RAID的可靠性。

我觉得，Hybrid RAID它的核心思想可以简单描述为兼顾平均分配和不均衡分配的好处。可以把他分为两个阶段：

1. 达到预定数量的P / E周期之前

2. 达到预定数量的P / E周期之后

在阶段1和阶段2的交替点作者称之为“切换点”，在阶段1均匀地在磁盘间分配奇偶校验单元，在阶段2转换为非均匀分配。

# 1 评估基于SSD的RAID系统的可靠性

## 1.1 定量评估模型QRM

MTTDL 是估算存储系统可靠性的传统指标，其他文献提出了DALOPY来作为可靠性衡量的其他标准。但是SSD有自己的特性，比如一个磁盘阵列中有若干个SSD，每个SSD的BER都不同，并且一个SSD又有不同的块（Blocks）每个Block的BER又不同。作者调查分析了，MTTDL适合长期可靠性估计，单不适合短期可靠性。SSD的RAID与基于HDD的RAID的可靠性是不同的，这是由于为此引入了两个度量之间的依赖关系：

a）安全时间的持续时间

b）死区时间的变化

本文也试图解决SSD-Based RAID 在不同的BER下的可靠性，因此提出了一个定量可靠性模型来评估其可靠性，该模型称之为QRM。

QRM引入了下列四个区域：

1. 安全区域：在这个区域内，SSD的误码率接近于零，因此SSD的可靠性可以假设为非常接近于该区域。当一个RAID系统的所有SSD被放置在这个区域时，系统的可靠性可以假定为1。

2. 半安全区：在这个区域，SSD的误码率不是零，但仍然足够低，可以接受。所以SSD的可靠性足够高，可以容忍读取，写入和擦除干扰错误。当至少一个SSD进入半安全区域时，系统的可靠性将小于1。通过增加离开安全区域进入半安全区域的SSD数量，RAID系统的可靠性将会降低。

3. 不安全的区域：在这个区域，固态硬盘的误码率是可以考虑的，固态硬盘的可靠性要低于上述两个区域。通过增加该区域的SSD数量，RAID系统的可靠性迅速下降。强烈建议在最坏的情况下，每个模拟时刻最多只有一个SSD放置在这个区域。为了最大限度地减少由读取，写入和擦除干扰引起的故障，进入不安全区域的SSD应该尽快离开该区域。当一个SSD离开这个区域时，它进入死区，应该换一个新的区域。

4. 死区：在这个区域，固态硬盘的P / E周期数达到了耐力极限。这个区域的SSD不再可靠，应该换一个新的。由于替换惩罚，同时在该区域内放置多个SSD是很重要的。这些区域的状态图如下图1所示。

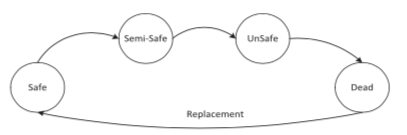


图1.1 QRM四区域循环状态图

这四个区域只是定义了四个不同的可靠性等级，用于计算系统的可靠性。作者提出的Hybrid RAID的过程中并没有使用到不安全和半安全区域。 这两个区域可以被视为放置在安全区域和死区之间的相同区域。

作者又针对上述四个区域又提出了两个指标：

1．安全时间的持续时间：此度量指示RAID系统的所有SSD放置在安全区域的时间持续时间。 第一个SSD离开安全区域的时间表示安全时间的持续时间。 由于SSD在安全区的可靠性很高（大约为1）; 因此，只要RAID系统的固态硬盘停留在该区域，系统就可以实现高可靠性。

2．死区时间的变化：这个度量指示了SSD死区时间的差异。 由于死亡/不安全区域SSD的可靠性非常低，在此区域的SSD数量应该尽可能的小。 当比较多的SSD在这些区域的时候，同时失败的概率更高。 如果死亡时间之间的差异变得足够高，同时失败的可能性就会下降

## 1.2计算可靠性的方法 - IRM

计算之前需要预先设定SSD 中Block的P / E周期数量，针对块的P / E周期的数量为每个盘定义具有不同BER的k个不同的BER水平。由于P / E周期分布在SSD的Blocks之间，所以块可以属于k个不同的BER级别中的每一个。第i级的块数由bi表示。所有的双块被认为是相同的误码率。

简单的说，IRM分三步走：

1. 根据一个SSD上每个Block的BER计算Block的可靠性

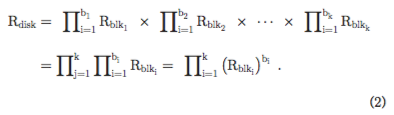
2. 根据Blocks的可靠性计算单个SSD的可靠性

3. 计算关于SSD和保护级别的RAID的可靠性

每个Block的BER水平计算如下公式一

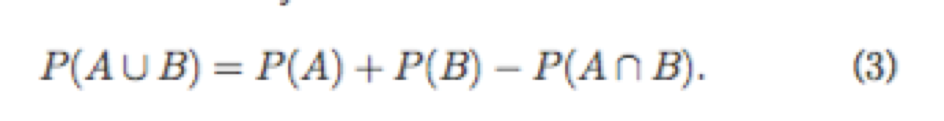


计算具有K个不同BER的SSD的可靠性计算公式二



将Block的BER连乘就得到了一个SSD的BER水平，每个SSD的BER水平的连城就得到了若干个BER水平的磁盘的可靠性计算表达式。Rdiski是第i个磁盘的可靠性。

文中提到了这样一个公式：



但是不太懂这个公式表达了什么意思。个人猜测是，参考RAID4和RAID5，推测N+1个SSD的RAID要想正常工作要满足两个条件,第一个是所有的SSD都正常工作，第二个是，只有一个SSD坏了，但是这两个条件从客观上说是独立的，并且不会同时发生的也就是P(A∩B)不可能发生，正好和文中提到的P(A∩B)=0符合。

在计算可靠性时候。需要考虑两个问题：

阵列中每个SSD的BER

不同操作时间的SSD的可靠性，文中称之为时变可靠性

其中第一个已经在计算公式中体现，第二个作者通过划分时间步长来解决。

# 2 Hybrid RAID的设计和提出

## 2.1 Hybrid RAID的设计和说明

Hybrid RAID提供了更高的可靠性，因为他改变了一般的奇偶分布策略。Hybrid RAID降低了同时发生的故障，并且在系统工作过程中通过改变奇偶校验分配策略推迟了SSD的首次故障。为什么说是推迟了首次故障呢，因为改变了奇偶分配策略，SSD的写入次数一开始被均匀分配，不存在某个被磨损的更多。

Hybrid RAID的主要工作如下：

1. 只要阵列中没有SSD离开安全区域，奇偶校验单元在所有SSD中均匀分布（类似于RAID5）。 在安全区域，所有SSD的可靠性足以提供RAID系统中的高可靠性。

2. 当一个SSD离开安全区域时，奇偶校验分配策略发生变化，并且奇偶校验单元被存储在特定的SSD中，导致这个SSD相对于其他盘快速磨损。 受害盘是在进入不安全区域的RAID系统的所有SSD中随机选择的。 这意味着所有这些盘在操作过程中被选为受害盘的概率一样。 并且受害盘将在达到耐力极限时被新盘替换。

3. 当阵列中的所有盘被新盘替换时，系统再次切换到偶数分配奇偶校验。 系统运行的其余部分重复上述过程（从第一步开始）。

作者设计的Hybrid RAID也面临另外一个难题，那就是如何确定一个最佳的时机去改变奇偶分配策略，这个时机被称为切换点。在本文中，作者选择第一张盘离开安全区的时间作为交换点。

并且根据实验的仿真结果表明，这一点确实为Hybrid RAID提供最高的可靠性。

根据切换点，操作的持续时间可分为两个步骤：切换前阶段和切换后阶段。在第一步中，奇偶校验单元在阵列的所有盘之间均匀分布，Hybrid RAID提供与RAID5相同的可靠性。如前所述，这一步的持续时间在本文中被称为安全区。系统保持在安全区域的时间越长，系统的可靠性就越高。在第二步中，通过不均匀地分配奇偶校验单元，增加了磁盘的P / E周期数的方差，因为不同的磁盘磨损速率不一样，方差必然变大。磁盘的P / E周期数量的变化是显示同时失败的概率的指标。方差越大，同时失败的概率越小。

虽然从偶数分布式奇偶校验到非均匀分布式奇偶校验的切换过程会给系统带来一定的性能开销。由于这种切换在每个磁盘操作期间只发生一次，所以在执行性能评估时可以忽略。并且为了提高系统的稳定性，这点性能开销是非常值得的。

## 3.2 如何再次提高Hybrid RAID的可靠性

文中所述，可以通过在Hybrid RAID结构中采用不同的方法来提高Hybrid RAID的可靠性。 所有方法的共同点是，运作的持续时间分为两个阶段，不同的平价分配政策。 第一步，只要所有的磁盘都在安全区域，奇偶校验单元在所有磁盘之间均匀分布。 当一张磁盘离开安全区域时，意味着系统不像以前那样可靠，所以需要更多的保护方法来容忍更多的故障。

论文提供更高的可靠性方法如下：

1. 在第切换前操作中为所有磁盘应用更强大的ECC。 单纯的增加每个磁盘上的P / E周期的数量，在读取，写入和擦除操作期间，会导致该磁盘的BER增加。 通过采用更强大的ECC，SSD可以承受更多的故障，提高系统的可靠性。 为了保证数据磁盘和奇偶校验磁盘的一致性，奇偶校验和数据磁盘都采用更强大的ECC进行保护。

2. 在系统中采用RAID4H。在更换死奇偶盘后选择奇偶盘的策略中，RAID4H与RAID4不同。 在RAID4H中，奇偶校验磁盘并不总是相同的磁盘。 在RAID4H中选择P / E周期数较多的磁盘作为奇偶校验盘。

不太懂RAID4H 中为何选择P / E周期数较多的盘作为奇偶校验的盘。

# 3 评估Hyrid RAID

作者通过仿真实验评估了Hybrid RAID的可靠性，并与Diff-RAID以及传统RAID（即RAID4，RAID5）进行了比较。 为了比较不同RAID在P / E周期数量方面的可靠性，在基于SSD的DiskSim模拟器中实现了所有RAID系统（即RAID4，RAID5，Diff-RAID和Hybrid RAID）。

## 3.1 实验设置的简单介绍

DiskSim仿真器中使用的实验设置参数如下表1所示。存储系统包含5个磁盘阵列，大小为2 MB。

四个磁盘保存数据，一个磁盘存储奇偶校验。在比较中是通过上文中提出的定量标准完成的，即安全时间的持续时间，SSD死区时间的变化。

不详细的介绍作者在仿真器中的设置

为了简单起见，块的持久性考虑了非常小的值（在DiskSim模拟器中SSD\_MAX\_ERA\_SURES参数设置为50）。有了这个小的限制，每个磁盘在短时间内达到耐力极限，加快了仿真时间。为了承认这个假设的正确性，对于不同的擦除限度，即50,100,500和1,000重复实验。如图2所示，对于不同的删除限制，安全区的百分比几乎保持不变。

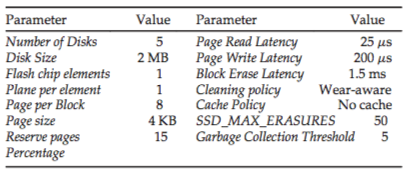


表3.1 实验仿真器配置表

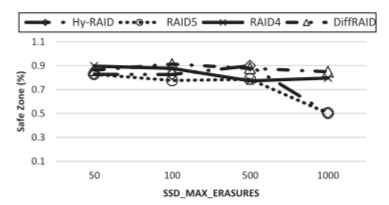


表3.2 不同擦除限制的安全区域百分比数

## 3.1 安全时间的长短和死亡时间

安全时间的长短和死亡时间是评估所提方法可靠性的两个关键指标。 在文章的评估中，如果该磁盘的剩余生命周期高于最大生命周期的30％，则假定块处于安全区域中。根据实验设置，观察结果表明，当考虑到磁盘的使用寿命小于30％时，系统的可靠性仍然非常接近（显然，对于其他工作负载和其他系统配置，这个值将会改变）。 系统可以停留在安全区的时间越长，系统中的安全时间越长。 安全时间越长，系统越可靠。 系统中第一次死亡的时间决定了系统中第一次更换磁盘的时间。

如果系统没有备份磁盘，则系统将由于第一个死区而失败。 在这些系统中，尽可能延迟第一次死亡是重要的。图3.3显示了磁盘数量对安全时间的影响。 在Hybrid RAID中，当阵列中的磁盘数量增加时，安全时间的持续时间会增加。 如图3.3a所示，当磁盘数量增加时，具有偶数奇偶分布的RAID5和Hybrid RAID在安全区域中保持较长时间。 通过增加磁盘的数量，写入/清除请求将在大量磁盘之间分配; 因此，每个磁盘在较长时间后达到耐力极限。

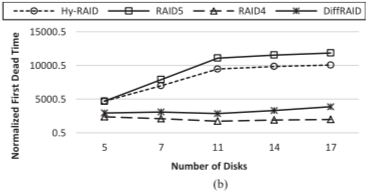
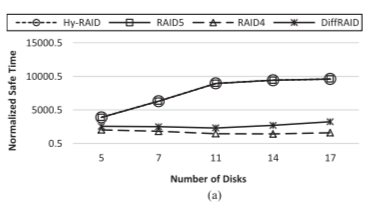


图3.3 不同磁盘数量的安全时间和第一次死亡时间比较图

图3.4a和3.4b示出了在安全时间以及归一化形式的死时间方面不同RAID之间的比较，即RAID4，Diff-RAID，RAID5和Hybrid RAID。 如图所示，RAID5和Hybrid RAID的区别可以忽略不计。

Hybrid RAID的安全时间是RAID4和Diff-RAID的2倍和1.6倍。 在RAID5和Hybrid RAID中，由于所有磁盘之间的均匀分配，所有磁盘都以相同的速率磨损，更长时间地保留在安全区域。

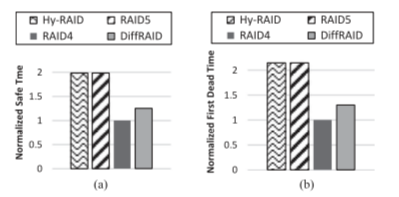


图3.4 不同RAID系统的安全时间和死时间变化图

## 3.2 安全时间的长短和死亡时间

死区时间的变化表示RAID系统中SSD之间P / E周期的分布。 一般来说，建议RAID系统的P / E周期变化很小，导致RAID中SSD的磨损相同。通过增加SSD中P / E周期的数量，达到其耐力极限，进入死/不安全区域。在这个区域，不建议再用相同的速度磨损SSD。如果不安全区域中有多个SSD，则同时发生故障的概率将会增加。

“死亡时间差异”一词表示本文评估中SSD死亡时间的差异。为了降低RAID系统中同时故障的概率，重要的是每个磁盘的死区时间比其他磁盘有更大的时间差。 换句话说，“死亡时间方差”越大，同时失败的概率越低。这个因素是比较RAID中P / E周期均匀分布的一个很好的指标。

RAID系统中固态硬盘之间的P / E周期分布越均匀，死区时间的变化越小。 实际上，从图3.5中可以清楚地看出，不均匀分布的奇偶校验RAID与均匀分布的奇偶校验RAID相比，具有显着较大的死区时间差异。 因此，本节的评估仅针对偶数分布式奇偶校验RAID，即Hybrid RAID和RAID5。

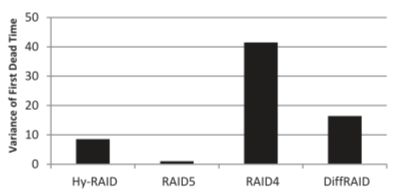


图3.5 不同磁盘数量的安全时间和第一次死亡时间比较图

对于不同数量的磁盘，Hybrid RAID和RAID5之间的死区时间差异进行了比较。 如图3.6所示，与RAID5相比，Hybrid RAID分别提供了2X到8X的较大的差异。 这种差异对于少量的磁盘是重要的，而对于大量的磁盘，这种差别会减小。 原因是，当磁盘数量增加时，P / E周期分布在大量磁盘中。 因此，在离开安全区域之后，在Hybrid RAID的后期出现第一次故障。 例如，当系统中有5个磁盘时，Hybrid RAID中SSD的首次死亡比RAID5大8倍，而对于14个磁盘的系统，这个差别是2倍。

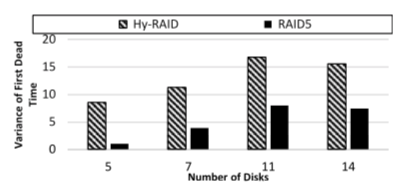


图3.6 不同磁盘数量的RAID5和Hybrid RAID的死区时间的变化

图3.7说明了时间间隔对死区时间方差的影响。从图中可以清楚的看出，当系统中磁盘数量较少时，两个RAID在大时间间隔的变化量的差异明显大于一个小的时间间隔。例如，如图3.7a所示，当系统中有5个磁盘时，Hybrid RAID的时间间隔大于RAID5的10倍，而这个变化在小时间间隔内可以忽略不计。另一方面，当阵列中有大量的磁盘时，两个RAID在大小时间间隔上的差异是微不足道的。如图3.7b所示，Hybrid RAID中不同时间间隔的死区时间的变化量比RAID5中的变化量大2倍左右。而且，当磁盘数量较少时，两个RAID在较大时间间隔上的差异比在RAID系统中存在大量磁盘时大得多（图3.7）。例如，当阵列中有少量的磁盘（即5个磁盘）时，对于较大的时间间隔（即1），Hybrid RAID的方差比RAID5高10倍。而对于大量相同时间间隔的磁盘，Hybrid RAID的性能仅比RAID5高出2倍。

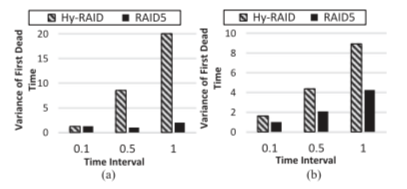


图3.7 不同时间间隔的RAID5和Hybrid RAID的死区时间的变化

# 4 总结

前一节所述，Hybrid RAID分别提高了偶数分布式奇偶校验RAID和不均匀分布式奇偶校验RAID在安全时间和死区时间方面的可靠性。最后总结一下文章评估的主要观察结果：

1. 只要是RAID5，Hybrid RAID就保持在安全区域，而Diff-RAID和RAID4则早早离开安全区域。 Hybrid RAID保持在比RAID4和Diff-RAID长1.9倍的安全区域。

2. 所提出的Hybrid RAID在安全时间上的改进对于大量磁盘而不是小磁盘来说是相当可观的。

3. 增加磁盘数量不会影响RAID4和Diff-RAID的安全时间，但是会延长RAID5和Hybrid RAID的安全时间。

4. 对于大量和少量磁盘，Hybrid RAID与RAID5相比，死区时间的变化范围为2X到8X。 因此，与RAID5相比，Hybrid RAID降低了同时故障的可能性。

5. 实际上，当磁盘容量增加时，从故障磁盘到新磁盘的数据恢复时间相当长。 Hybrid RAID通过推迟替换第一个磁盘来考虑这个事实。 它还可以防止同时更换磁盘，这可能会影响系统的可用性。

在本文中，已经提出Hybrid RAID作为基于SSD的存储系统的可靠RAID。提出的Hybrid RAID背后的主要思想是提高传统RAID的可靠性，即均匀分布和不均匀分布。 Hybrid RAID利用了均匀分布和不均匀分布的好处。

最初，Hybrid RAID在磁盘之间均匀分配奇偶校验单元，以增加安全时间并推迟第一次故障。然后，当P / E周期的数量达到预定数量的周期时，Hybrid RAID切换到不均匀的均衡分布。这一更改避免了系统中同时发生磁盘故障。本文还提出了一个可靠性模型（称为IRM），用于评估基于SSD的RAID的可靠性。