可容忍四节点故障的紧凑型二维 RAID 6 级阵列

Jehan-François Pâris 计算机科学系 休斯敦大学 休斯敦，TX 77204-3010，美国
jfparis@uh.edu

Thomas Schwarz, SJ 计算机科学系 马凯特大学 密尔沃基，WI 53233，美国
thomas.schwarz@marquette.edu

扩展摘要

I. Introduction

里德-所罗门码 [1] 已广泛用于保护数据，防止因设备故障而造成数据丢失。例如，BackBlaze 云备份服务依靠一种 17/20 RS 码来确保用户数据的完整性[2]。该服务会将每个传入文件拆分为 17 个大小相同的分片，并计算 3 个校验分片，因此可从这 20 个分片中的任意 17 个重建文件内容。这种方法的主要优势是空间开销极低，因为 20 个分片中只有 3 个包含冗余数据。

虽然里德-所罗门码恰好是保护 n 个存储单元中的内容免受其中 k 个单元故障影响时空间效率最高的技术，但对于由多个条带组成、每个条带包含 n 个单元且各自配有 k 个校验单元的情况，则并非如此。例如，众所周知，二维 RAID 阵列 [3] [4] [5][6] 对数据的保护能力优于具有相同存储容量和相同空间开销的多个 RAID 6 级阵列，能够抵御更多三重和四重单元故障。

与此同时，现有的二维 RAID 阵列无法保护数据免受所有四重单元故障的影响。十年或二十年前，这或许不是什么大问题，但我们认为，现在是时候研究能够容忍所有四重单元故障和大多数五重单元故障、且不会发生数据丢失的二维阵列了。

我们在此研究首个满足这些标准的二维 RAID 阵列。这些阵列由 n 个 RAID 6 级校验条带构成，每个条带包含两个校验节点，并将数据节点放置在这些 n 个条带的 $n(n-1)/2$ 个交点上。我们评估了这些阵列应对五重和六节点故障的能力，并得出结论：与具有相同存储容量和相同空间开销的多个 n-out-of-~~n~~-k 阵列相比，它们能够容忍多得多的五重和六节点单元故障。此外，我们还展示了如何通过向阵列添加两个超级校验单元来消除所有致命五重故障。

II. 二维独立磁盘冗余阵列

二维 RAID 阵列（即二维校验阵列）最早由 Schwarz [3] 和 Hellerstein 等人 [4]进行研究。此外，Lee 还为一种磁盘阵列组织方式申请了专利，该组织方式在一个维度中及时进行校验更新，而在第二个维度中延迟进行校验更新 [6]。

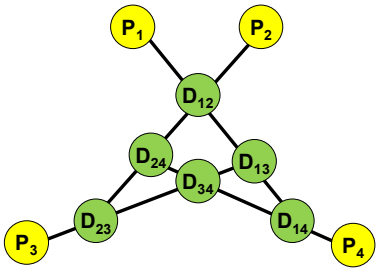

图1. 一个包含六个数据节点和四个校验节点的小型二维 RAID阵列。

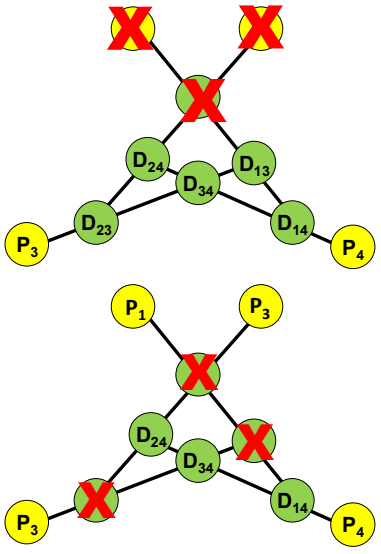

图 2. 两种三重致命故障。

二维数组同时包含数据节点和校验节点。这些节点被组织成校验条带，每个校验条带由若干数据节点和一个校验节点组成，后者包含这些数据节点内容的异或（XOR）结果。此外，

1. 每个数据节点必须属于两个不同的校验条带。2. 两个不同的校验条带最多只能共享一个数据节点。

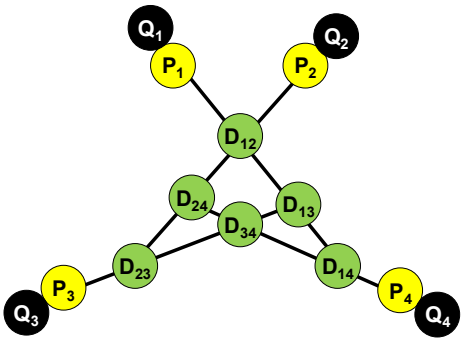

图3. 一个小型完整二维 RAID 6 级阵列。现在，每个校验条带都有两个校验节点。

例如，考虑图1所示的二维 RAID 阵列，并注意到其中所有校验条带彼此相交。因此，其中的四个校验节点能够保护六个数据节点的内容，使其免受所有双节点故障的影响。更一般地，任何具有 n 个校验节点的完整二维阵列最多可包含 n（$n-1$）个 /2 数据节点。

虽然完整二维 RAID 阵列比其他二维 RAID 阵列组织方式更紧凑，但在发生三节点故障时也更容易丢失数据。如图 2 所示，完整二维 RAID 阵列在以下任一情况下都会发生数据丢失：

1. 三个故障节点由一个数据节点及其两个校验条带上的校验节点组成。2. 三个故障节点分别位于三个不同校验条带的三个交点上。

这两种脆弱性中，后一种影响最为严重，因为它涉及更多节点，而且随着阵列规模扩大，组合数量会呈爆炸式增长。Paris 等人指出了这一局限[10]，并得出结论：消除这类故障的最佳方法是将校验磁盘数量加倍，并将每个校验条带改为 RAID 6 级条带。他们没有继续研究这一方案，因为其他二维 RAID 阵列能够以更低的成本容忍所有三节点故障。

我们认为，现在是重新探讨这一主题的时候了，因为与十二或十三年前相比，我们如今对容错能力的要求更高。因此，当年还只是个新奇事物的方案，如今已成为解决实际问题的潜在方案。

III. Complete Two-Dimensional 独立磁盘冗余阵列 6 级 Arrays

一个完整的二维 RAID 6 级阵列由 n 条相互交叉的 RAID 6 级校验条带[3] [8] 组成，$n(n-1)/2$ 个数据节点位于n 校验条带的交叉处。

这些n 校验条带中的每一条都包含 $n-1$ 个数据节点和两个校验节点，因此总共有 $2n$ 个校验节点。举例来说，图3所示的小型阵列有4个校验节点、6个数据节点，每条校验条带包含3个数据节点。（我们预计实际阵列会大得多，校验开销也会低得多。）

现在，我们来考察该组织提供的数据丢失保护级别。令 $N = 2n + n(n-1)/2$ 表示阵列中的节点总数：

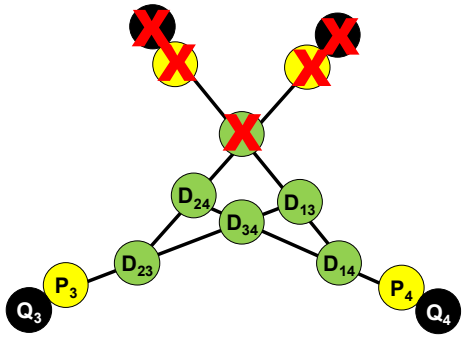

图 4. 六种可能的致命五重故障之一。

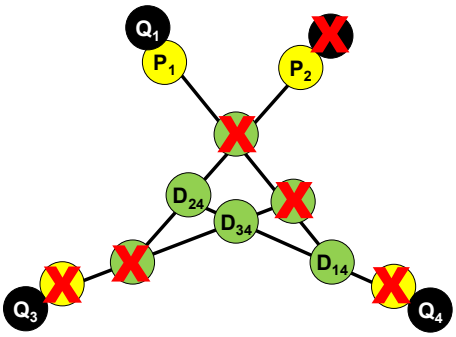

图5. 三个数据节点在相同的三个校验条带上发生故障，且这三个校验条带中各有一个校验节点发生故障所导致的致命故障。

1. 完整的二维 RAID 6 级阵列最多可容忍四个节点同时故障，而不会发生任何数据丢失。

2. 如图 4 所示，唯一会导致数据丢失的五重故障，是任意一个数据节点与其两个校验条带上的四个校验节点同时发生故障。在 $\binom{N}{5}$ 种可能的五节点故障中，此类故障有 $n(n-1)/2$ 种。因此，只有一部分

$$\alpha = \frac{n(n-1)}{2\binom{N}{5}}$$

五重故障会导致数据丢失。随着阵列规模增大，这一比例会迅速下降：在数据节点超过 45 个、校验节点超过 20 个的阵列中，导致致命后果的五重故障不到全部五重故障的 2.10^{-6}。

3. 以下三种不同的六节点故障配置会导致数据丢失：

a. 任意一个数据节点、其所在的两个校验条带上的四个数据节点，以及任意其他节点发生故障：此类致命六元组故障共有 $(N-5)\ n(n-1)/2$ 种。b. 如图5所示，三个不同校验条带上的三个数据节点，以及这三个校验条带中各一个发生故障的校验节点发生故障：此类致命六元组故障还有 $2^3\binom{n}{3}$ 种。

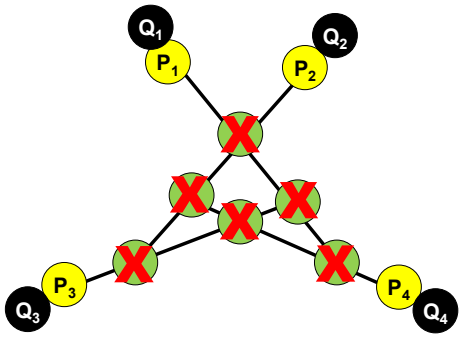

图6. 占据四个校验条带六个交点的六个数据节点发生致命故障。

c. 如图6所示，占据四个校验条带全部六个交点的六个数据节点发生故障：此类致命六元组故障还有 $\binom{n}{4}$ 种。

因此，所有六节点故障中会导致数据丢失的比例为：

$$\beta = \frac{(N-5)\frac{n(n-1)}{2} + 2^3\binom{n}{3} + \binom{n}{4}}{\binom{N}{6}}$$

同样，随着阵列规模增大，这一比例会迅速下降：在拥有 45 个以上数据节点和 20 个校验节点的阵列中，所有六节点故障中致命故障的比例不到 3.10^{-5}。

IV. 比较研究

我们希望比较完整的二维 RAID 阵列所提供的数据丢失保护水平，与采用相同数量校验节点、保护相同数量数据节点免受所有可能的四节点故障影响的单维 m-out-of-m+k 组织所实现的保护水平。我们选择了一种由 $n/2$ 个条带组成的组织，每个条带包含 $n-1$ 个数据节点和四个校验节点。因此，每个条带构成一个独立的 n–1-out-of-n+3

这种组织能够在不发生任何数据丢失的情况下容忍：

1. 最多四个节点同时故障。2. 五个节点同时故障，但五个故障节点全部位于同一条带中的情况除外。在 $\binom{N}{5}$ 种可能的五节点故障中，有 $\frac{n}{2}\binom{n-1}{5}$ 种属于此类。因此，只有一部分

$$\alpha' = \frac{n\binom{n-1}{5}}{2\binom{N}{5}}$$

所有五重故障中，都会导致数据丢失。

3. 六个节点同时发生故障，但五个故障节点全部位于同一条带的情况除外。在所有 $\binom{N}{6}$ 种可能的六节点故障中，有 $\frac{n}{2}\binom{n-1}{5}(N-5)$ 种失败。因此，只有一部分：

TABLE I. COMPARED 两种组织方式中致命五重故障和六节点故障的数量

数量 of 节点	致命五重故障		致命六元组故障	
	2-D RAID 6	1-D 编码	2-D RAID 6	1-D 编码
14	6	42	87	378
27	15	378	465	8,316
44	28	1,848	1,498	72,072
65	45	6,435	3,630	386,100
90	66	18,018	7,425	1,531,530
119	91	43,316	13,559	4,938,024
152	120	93,024	22,820	13,674,528
189	153	183,141	36,108	33,69,7944
230	190	336,490	54,435	75,710,250
275	231	584,430	78,925	157,796,100
324	276	968,760	110,814	309,034,440
377	325	154,3815	151,450	574,299,180

$$\beta' = \frac{n\binom{n-1}{5}(N-5)}{2\binom{N}{6}}$$

所有六节点故障会导致数据丢失。

表 I 展示了比较结果。可以看出，与我们完整的二维 RAID 6 级阵列相比，一维组织中的致命五重故障和致命六节点故障明显更多。事实上，与一维组织相比，我们的方案遭遇的致命五重故障减少了 87% 至 99.99%，致命六节点故障减少了 77% 至 99.98%。

V. 可OSSIBLE 扩XTENSIONS

由于所有致命五重故障都涉及校验节点，添加超奇偶校验节点便能以简单且经济高效的方式保护阵列内容，避免所有致命五重故障。

顾名思义，超奇偶校验节点充当常规校验节点的校验节点 [7]。在我们的方案中，我们将利用它们使阵列能够从两对校验节点发生故障的情况中恢复。实现这一目标最经济高效的方式，是将所有 P 奇偶校验节点或所有 Q 奇偶校验节点（但不能两者兼有）连接在一个 n 取 n+2 条带中。图 7 展示了这样一种组织方式。

在阵列中增加两个超奇偶校验节点，也会大幅减少致命六元组故障。回想一下，在发生六次故障时，所有导致数据丢失的情况都源于三种可能的六重故障配置：

1. 涉及任意一个数据节点及其所在两条校验条带上的四个校验节点的唯一致命六元组故障，还必须涉及两个超奇偶校验节点中的一个。此类六节点故障共有 $n(n-1)$ 个。

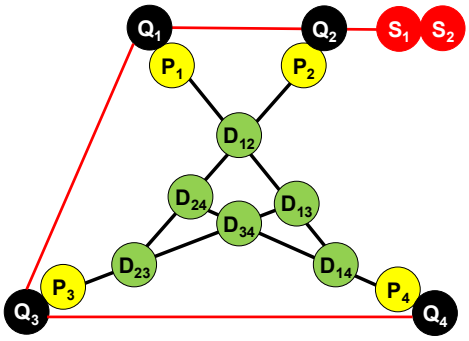

图 7. 一个小型完整二维阵列，包含两个超奇偶校验设备 S_1 和 S_2，，用于保护校验节点 Q_1 到 Q_4的内容。

2. 涉及同三条校验条带上的三个数据节点，并且这三条校验条带中各有一个校验节点发生故障的致命六元组故障，现在只有 $2\binom{n}{3}$ 个，因为发生故障的校验节点必须全部是 P 奇偶校验或全部是 Q 奇偶校验。3. 占据四条校验条带的六个交点的六个数据节点发生故障，仍然是致命故障。

因此，所有六节点故障中仍会导致数据丢失的比例为：

$$\beta_{sp} = \frac{2\binom{n}{2} + 2\binom{n}{3} + \binom{n}{4}}{\binom{N_{sp}}{6}}$$

其中 $N_{sp} = 2n + n(n-1)/2 + 2$。

如表 II 所示，对于节点数超过 46 的所有阵列，向阵列中添加两个超奇偶校验节点可消除 85% 的致命六元组故障。添加更多超奇偶校验节点对阵列韧性的影响微乎其微，因为阵列仍无法抵御六个数据节点同时发生故障的情况；这六个节点分别位于四条校验条带的交点上。

消除这些致命六元组故障最显而易见的方法，是将阵列中所有 RAID 6 级条带替换为 m-out-of-m+3 条带。新阵列将包含 $n(n-1)$ 个 /2 数据节点和 $3n$ 个校验节点。它可以容忍其 $3n + n(n-1)$ 个 /2 节点中最多六个发生故障而不丢失数据，但仍会受到某些致命六元组故障的影响，即一个数据节点及其所在的两条数据条带上的六个校验节点同时发生故障。

VI. Conclusion

我们分析了首批能够容忍四重故障的二维独立磁盘冗余阵列的性能。这些阵列以 n 个 RAID 6 级校验条带为基础，每个条带包含两个校验分区 n 。研究表明，与存储容量和空间开销相同的多个n取n中的k阵列相比，这些阵列的致命五重故障次数减少了 87% 至 99.99%，致命六节点故障次数减少了 77% 至 99.98%。此外，向阵列中增加两个超奇偶校验节点，可消除所有致命五重故障，并消除 85% 的六节点故障。

无超奇偶校验节点与有超奇偶校验节点时的致命六元组故障数量。

无超奇偶校验节点		有超奇偶校验节点	
数量 of 节点	致命六元组失败次数	数量 of 节点	致命六元组失败次数
14	87	16	21
27	465	29	85
44	1,498	46	238
65	3,630	67	540
90	7,425	92	1,067
119	13,559	121	1,911
152	22,820	154	3,180
189	36,108	192	4,998
230	54,435	232	75,05
275	78,925	277	10,857
324	110,814	326	15,226
377	151,450	379	20,800

仍需进一步研究，通过完全枚举或蒙特卡罗方法，评估阵列能够容忍多少次致命七重故障。

References

[1] I. S. Reed 和 G. Solomon，《某些有限域上的多项式码》，SIAM Journal，第 8 卷，第 2 期，第 300–304 页，1960 年。[2] B. Beach，《BackBlaze 开源 Reed-Solomon 擦除码源代码》，https://www.backblaze.com/blog/reed-solomon/，6 月 16， 访问日 2015 [29, 2025] 期：1 月。[3] T. J. E. Schwarz，《磁盘阵列的可靠性与性能》，博士论文，计算机科学与工程系，加利福尼亚大学圣迭戈分校，1994 年。[4] L. Hellerstein、G. Gibson、R. M. Karp、R. H. Katz 和 D.A. Patterson，《大型磁盘阵列故障处理编码技术》，*Algorithmica*，12(3-4):182-208，1994 年 6 月[5] W. S. Lee，《一种二维存储阵列，在一个维度中采用即时奇偶校验，在第二个维度中采用延迟奇偶校验》，美国专利 #6675318 B1，2004 年。[6] T. Schwarz, SJ、D. D. E. Long 和 J.-F. Pâris，《双重奇偶校验磁盘阵列的可靠性》，发表于*19th* IEEE 太平洋地区可靠计算国际研讨会（PRDC 2013），加拿大不列颠哥伦比亚省温哥华，2013 年 12 月，第 108–117 页。[7] A. Wildani、T. J. E. Schwarz、E. L. Miller 和 D. D. E. Long，《防范归档系统中的罕见事件故障》，发表于17*th* IEEE 国际计算机与电信系统建模、分析与仿真研讨会（MASCOTS '09），英国伦敦，2009 年 9 月，第 246–256 页。[8] W. A. Burkhard、J. Menon，《磁盘阵列存储系统可靠性》，发表于 23*rd* 国际容错计算研讨会（FTCS-19），伊利诺伊州芝加哥，1993 年 6 月，第 432–441 （1993）[9] J.-F. Pâris、T. Schwarz, SJ、A. Amer 和 D. D. E. Long，《防范磁盘故障率意外升高对 RAID 阵列造成的影响》，发表于 20*th* IEEE 太平洋地区可靠计算国际研讨会（PRDC 2014），新加坡，2014 年 11 月，第 68–75 页。