简述RAID磁盘阵列技术

2017-09-21 运维之美

前言

RAID，独立硬盘冗余阵列（RAID, Redundant Array of Independent Disks），简称磁盘阵列。其基本思想就是把多个相对便宜的硬盘组合起来，成为一个硬盘阵列组，使性能达到甚至超过一个价格昂贵、容量巨大的硬盘。

根据选择的版本不同，RAID比单颗硬盘有以下一个或多个方面的好处：增强数据集成度，增强容错功能，增加处理量或容量。另外，磁盘阵列对于电脑来说，看起来就像一个单独的硬盘或逻辑存储单元。

RAID基础知识

基本原理

RAID （ Redundant Array of Independent Disks ）即独立磁盘冗余阵列，通常简称为磁盘阵列。简单地说， RAID 是由多个独立的高性能磁盘驱动器组成的磁盘子系统，从而提供比单个磁盘更高的存储性能和数据冗余的技术。 RAID 是一类多磁盘管理技术，其向主机环境提供了成本适中、数据可靠性高的高性能存储。 SNIA 对 RAID 的定义是 ：一种磁盘阵列，部分物理存储空间用来记录保存在剩余空间上的用户数据的冗余信息。当其中某一个磁盘或访问路径发生故障时，冗余信息可用来重建用户数据。磁盘条带化虽然与 RAID 定义不符，通常还是称为 RAID （即 RAID0 ）。

RAID 的初衷是为大型服务器提供高端的存储功能和冗余的数据安全。在整个系统中， RAID 被看作是由两个或更多磁盘组成的存储空间，通过并发地在多个磁盘上读写数据来提高存储系统的 I/O 性能。大多数 RAID 等级具有完备的数据校验、纠正措施，从而提高系统的容错性，甚至镜像方式，大大增强系统的可靠性， Redundant 也由此而来。

RAID 的两个关键目标是提高数据可靠性和 I/O 性能。磁盘阵列中，数据分散在多个磁盘中，然而对于计算机系统来说，就像一个单独的磁盘。通过把相同数据同时写入到多块磁盘（典型地如镜像），或者将计算的校验数据写入阵列中来获得冗余能力，当单块磁盘出现故障时可以保证不会导致数据丢失。有些 RAID 等级允许更多地 磁盘同时发生故障，比如 RAID6 ，可以是两块磁盘同时损坏。在这样的冗余机制下，可以用新磁盘替换故障磁盘， RAID 会自动根据剩余磁盘中的数据和校验数据重建丢失的数据，保证数据一致性和完整性。数据分散保存在 RAID 中的多个不同磁盘上，并发数据读写要大大优于单个磁盘，因此可以获得更高的聚合 I/O 带宽。当然，磁盘阵列会减少全体磁盘的总可用存储空间，牺牲空间换取更高的可靠性和性能。比如， RAID1 存储空间利

用率仅有 50% ， RAID5 会损失其中一个磁盘的存储容量，空间利用率为 (n-1)/n 。

磁盘阵列可以在部分磁盘（单块或多块，根据实现而论）损坏的情况下，仍能保证系统不中断地连续运行。在重建故障磁盘数据至新磁盘的过程中，系统可以继续正常运行，但是性能方面会有一定程度上的降低。一些磁盘阵列在添加或删除磁盘时必须停机，而有些则支持热交换（ Hot Swapping ），允许不停机下替换磁盘驱动器。这种高端磁盘阵列主要用于要求高可能性的应用系统，系统不能停机或尽可能少的停机时间。一般来说， RAID 不可作为数据备份的替代方案，它对非磁盘故障等造成的数据丢失无能为力，比如病毒、人为破坏、意外删除等情形。此时的数据丢失是相对操作系统、文件系统、卷管理器或者应用系统来说的，对于 RAID 系统来身，数据都是完好的，没有发生丢失。所以，数据备份、灾 备等数据保护措施是非常必要的，与 RAID 相辅相成，保护数据在不同层次的安全性，防止发生数据丢失。

RAID 中主要有三个关键概念和技术：镜像（ Mirroring ）、数据条带（ Data Stripping ）和数据校验（ Data parity ）。镜像，将数据复制到多个磁盘，一方面可以提高可靠性，另一方面可并发从两个或多个副本读取数据来提高读性能。显而易见，镜像的写性能要稍低， 确保数据正确地写到多个磁盘需要更多的时间消耗。数据条带，将数据分片保存在多个不同的磁盘，多个数据分片共同组成一个完整数据副本，这与镜像的多个副本是不同的，它通常用于性能考虑。数据条带具有更高的并发粒度，当访问数据时，可以同时对位于不同磁盘上数据进行读写操作， 从而获得非常可观的 I/O 性能提升 。数据校验，利用冗余数据进行数据错误检测和修复，冗余数据通常采用海明码、异或操作等算法来计算获得。利用校验功能，可以很大程度上提高磁盘阵列的可靠性、鲁棒性和容错能力。不过，数据校验需要从多处读取数据并进行计算和对比，会影响系统性能。 不同等级的 RAID 采用一个或多个以上的三种技术，来获得不同的数据可靠性、可用性和 I/O 性能。至于设计何种 RAID （甚至新的等级或类型）或采用何种模式的 RAID ，需要在深入理解系统需求的前提下进行合理选择，综合评估可靠性、性能和成本来进行折中的选择。

RAID 思想从提出后就广泛被业界所接纳，存储工业界投入了大量的时间和财力来研究和开发相关产品。而且，随着处理器、内存、计算机接口等技术的不断发展， RAID 不断地发展和革新，在计算机存储领域得到了广泛的应用，从高端系统逐渐延伸到普通的中低端系统。 RAID 技术如此流行，源于其具有显著的特征和优势，基本可以满足大部分的数据存储需求。总体说来， RAID 主要优势有如下几点：

(1) 大容量

这是 RAID 的一个显然优势，它扩大了磁盘的容量，由多个磁盘组成的 RAID 系统具有海量的存储空间。现在单个磁盘的容量就可以到 1TB 以上，这样 RAID 的存储容量就可以达到 PB 级，大多数的存储需求都可以满足。一般来说， RAID 可用容量要小于所有成员磁盘的总容量。不同等级的 RAID 算法需要一定的冗余开销，具体容量开销与采用算法相关。如果已知 RAID 算法和容量，可以计算出 RAID 的可用容量。通常， RAID 容量利用率在 50% ~ 90% 之间。

(2) 高性能

RAID 的高性能受益于数据条带化技术。单个磁盘的 I/O 性能受到接口、带宽等计算机技术的限制，性能往往很有 限，容易成为系统性能的瓶颈。通过数据条带化， RAID 将数据 I/O 分散到各个成员磁盘上，从而获得比单个磁盘成倍增长的聚合 I/O 性能。

(3) 可靠性

可用性和可靠性是 RAID 的另一个重要特征。从理论上讲，由多个磁盘组成的 RAID 系统在可靠性方面应该比单个磁盘要差。这里有个隐含假定：单个磁盘故障将导致整个 RAID 不可用。RAID 采用镜像和数据校验等数据冗余技术，打破了这个假定。 镜像是最为原始的冗余技术，把某组磁盘驱动器上的数据完全复制到另一组磁盘驱动器上，保证总有数据副本可用。 比起镜像 50%的冗余开销 ，数据校验要小很多，它利用校验冗余信息对数据进行校验和纠错。 RAID 冗余技术大幅提升数据可用性和可靠性，保证了若干磁盘出错时，不 会导致数据的丢失，不影响系统的连续运行。

(4) 可管理性

实际上， RAID 是一种虚拟化技术，它对多个物理磁盘驱动器虚拟成一个大容量的逻辑驱动器。对于外部主机系统来说， RAID 是一个单一的、快速可靠的大容量磁盘驱动器。这样，用户就可以在这个虚拟驱动器上来组织和存储应用系统数据。 从用户应用角度看，可使存储系统简单易用，管理也很便利。 由于 RAID 内部完成了大量的存储管理工作，管理员只需要管理单个虚拟驱动器，可以节省大量的管理工作。 RAID 可以动态增减磁盘驱动器，可自动进行数据校验和数据重建，这些都可以 大大简化管理工作。

关键技术

> 镜像

镜像是一种冗余技术，为磁盘提供保护功能，防止磁盘发生故障而造成数据丢失。对于 RAID 而言，采用镜像技术 典型地 将会同时在阵列中产生两个完全相同的数据副本，分布在两个不同的磁盘驱动器组上。镜像提供了完全的数据冗余能力，当一个数据副本失效不可用时，外部系统仍可正常访问另一副本，不会对应用系统运行和性能产生影响。而且，镜像不需要额外的计算和校验，故障修复非常快，直接复制即可。镜像技术可以从多个副本进行并发读取数据，提供更高的读 I/O 性能，但不能并行写数据，写多个副本会会导致一定的 I/O 性能降低。

镜像技术提供了非常高的数据安全性，其代价也是非常昂贵的，需要至少双倍的存储空间。高成本

限制了镜像的广泛应用，主要应用于至关重要的数据保护，这种场合下数据丢失会造成巨大的损失。另外，镜像通过 “ 拆分 ” 能获得特定时间点的上数据快照，从而可以实现一种备份窗口几乎为零的数据备份技术。

数据条带

磁盘存储的性能瓶颈在于磁头寻道定位，它是一种慢速机械运动，无法与高速的 CPU 匹配。再者，单个磁盘驱动器性能存在物理极限， I/O 性能非常有限。 RAID 由多块磁盘组成，数据条带技术将数据以块的方式分布存储在多个磁盘中，从而可以对数据进行并发处理。这样写入和读取数据就可以在多个磁盘上同时进行，并发产生非常高的聚合 I/O ，有效提高了整体 I/O 性能，而且具有良好的线性扩展性。这对大容量数据尤其显著，如果不分块，数据只能按顺序存储在磁盘阵列的磁盘上，需要时再按顺序读取。而通过条带技术，可获得数倍与顺序访问的性能提升。

数据条带技术的分块大小选择非常关键。条带粒度可以是一个字节至几 KB 大小，分块越小，并行处理能力就越强，数据存取速度就越高，但同时就会增加块存取的随机性和块寻址时间。实际应用中，要根据数据特征和需求来选择合适的分块大小，在数据存取随机性和并发处理能力之间进行平衡，以争取尽可能高的整体性能。

数据条带是基于提高 I/O 性能而提出的，也就是说它只关注性能， 而对数据可靠性、可用性没有任何改善。实际上，其中任何一个数据条带损坏都会导致整个数据不可用，采用数据条带技术反而增加了数据发生丢失的概念率。

数据校验

镜像具有高安全性、高读性能，但冗余开销太昂贵。数据条带通过并发性来大幅提高性能，然而对数据安全性、可靠性未作考虑。数据校验是一种冗余技术，它用校验数据来提供数据的安全，可以检测数据错误，并在能力允许的前提下进行数据重构。相对镜像，数据校验大幅缩减了冗余开销，用较小的代价换取了极佳的数据完整性和可靠性。数据条带技术提供高性能，数据校验提供数据安全性， RAID 不同等级往往同时结合使用这两种技术。

采用数据校验时， RAID 要在写入数据同时进行校验计算，并将得到的校验数据存储在 RAID 成员磁盘中。校验数据可以集中保存在某个磁盘或分散存储在多个不同磁盘中，甚至校验数据也可以分块，不同 RAID 等级实现各不相同。当其中一部分数据出错时，就可以对剩余数据和校验数据进行反校验计算重建丢失的数据。校验技术相对于镜像技术的优势在于节省大量开销，但由于每次数据读写都要进行大量的校验运算，对计算机的运算速度要求很高，必须使用硬件 RAID 控制器。在数据重建恢复方面，检验技术比镜像技术复杂得多且慢得多。

海明校验码和 异或校验是两种最为常用的 数据校验算法。海明校验码是由理查德·海明提出的，不仅能检测错误，还能给出错误位置并自动纠正。海明校验的基本思想是：将有效信息按照某种规律分成若干组，对每一个组作奇偶测试并安排一个校验位，从而能提供多位检错信息，以定位错误

点并纠正。可见海明校验实质上是一种多重奇偶校验。异或校验通过异或逻辑运算产生，将一个有效信息与一个给定的初始值进行异或运算，会得到校验信息。如果有效信息出现错误，通过校验信息与初始值的异或运算能还原正确的有效信息。

常见RAID类型

> 常见5种RAID类型对比，n位磁盘数量。

RAID 等级	RAID0	RAID1	RAID5	RAID6	RAID10
别名	条带	镜像	分布奇偶校验条带	双重奇偶校验条带	镜像加条带
容错性	无	有	有	有	有
冗余类型	无	有	有	有	有
热备盘	无	有	有	有	有
读性能	高	低	高	高	高
随机写性能	高	低	一般	低	一般
连续写性能	高	低	低	低	一般
需要磁盘数	n≥1	2n (n≥1)	n≥3	n≥4	2n(n≥2)≥4
可用容量	全部	50%	(n-1)/n	(n-2)/n	50%

RAID 等级

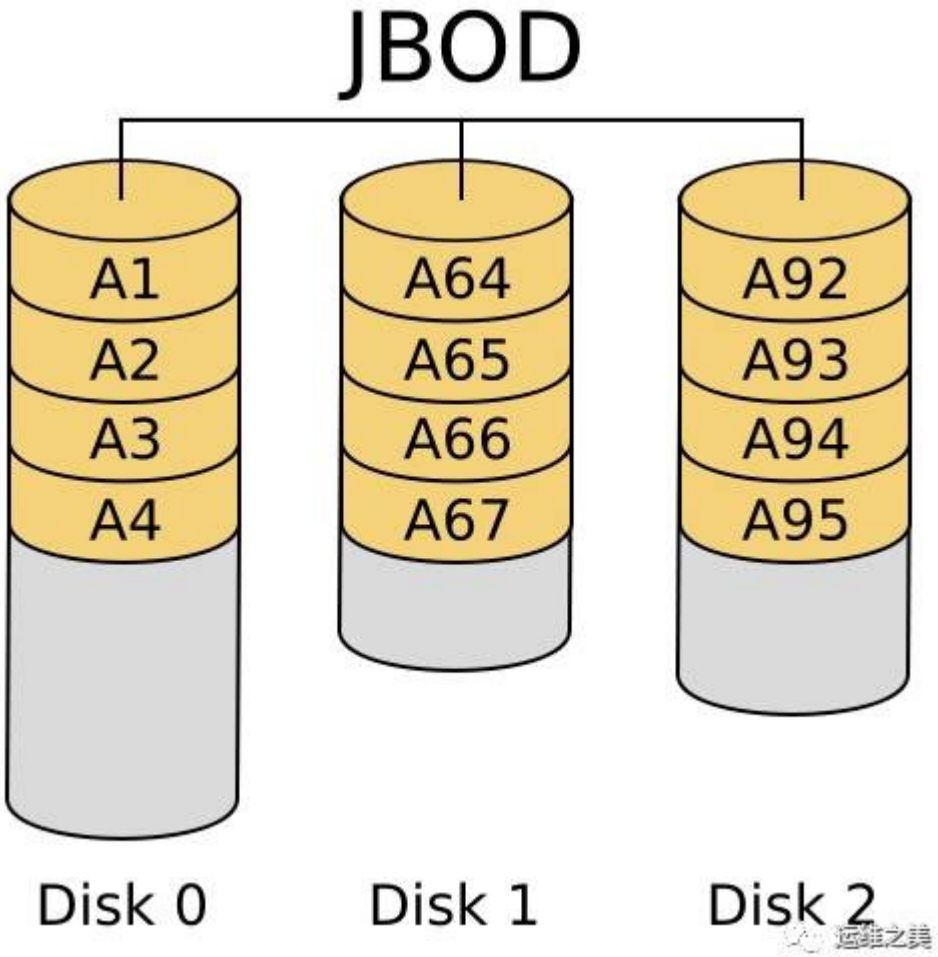
标准 RAID 等级

SNIA 、 Berkeley 等组织机构把 RAID0 、 RAID1 、 RAID2 、 RAID3 、 RAID4 、 RAID5 、 RAID6 七个等级定为标准的 RAID 等级，这也被业界和学术界所公认。标准等级是最基本的 RAID 配置集合，单独或综合利用数据条带、镜像和数据校验技术。标准 RAID 可以组合，即 RAID 组合等级，满足 对性能、安全性、可靠性要求更高的存储应用需求。

JBOD

JBOD （ Just a Bunch Of Disks ）不是标准的 RAID 等级，它通常用来表示一个没有控制软件提供协调控制的磁盘集合。 JBOD 将多个物理磁盘串联起来，提供一个巨大的逻辑磁盘。 JBOD 的

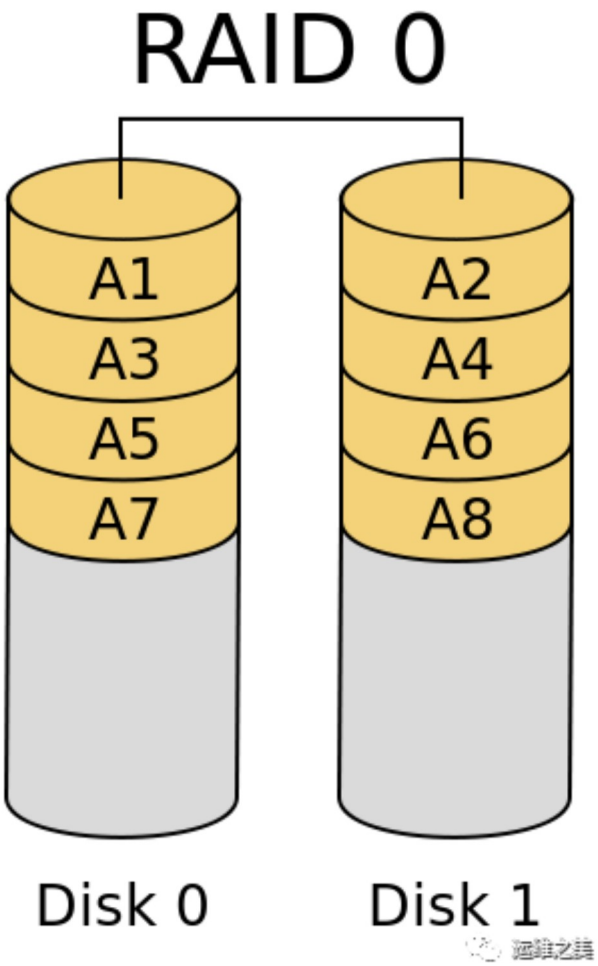
数据存放机制是由第一块磁盘开始按顺序往后存储，当前磁盘存储空间用完后，再依次往后面的磁盘存储数据。 JBOD 存储性能完全等同于单块磁盘，而且也不提供数据安全保护。它只是简单提供一种扩展存储空间的机制， JBOD 可用存储容量等于所有成员磁盘的存储空间之和。目前 JBOD 常指磁盘柜，而不论其是否提供 RAID 功能。



RAID0

RAID0 是一种简单的、无数据校验的数据条带化技术。实际上不是一种真正的 RAID ，因为它并不提供任何形式的冗余策略。 RAID0 将所在磁盘条带化后组成大容量的存储空间，将数据分散存储在所有磁盘中，以独立访问方式实现多块磁盘的并读访问。由于可以并发执行 I/O 操作，总线带宽得到充分利用。再加上不需要进行数据校验， RAID0 的性能在所有 RAID 等级中是最高的。理论上讲，一个由 n 块磁盘组成的 RAID0 ，它的读写性能是单个磁盘性能的 n 倍，但由于总线带宽等多种因素的限制，实际的性能提升低于理论值。

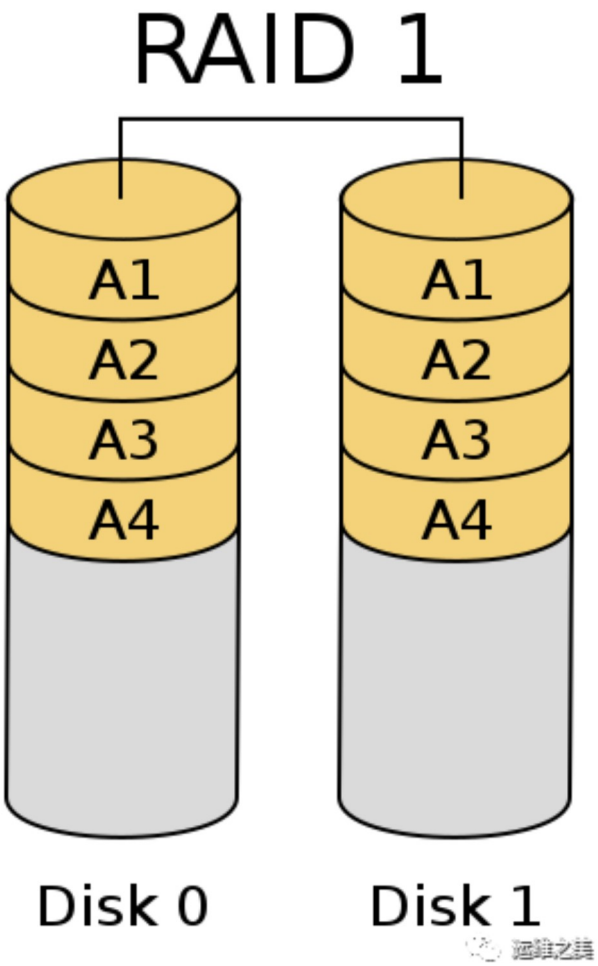
RAID0 具有低成本、高读写性能、 100% 的高存储空间利用率等优点，但是它不提供数据冗余保护，一旦数据损坏，将无法恢复。 因此， RAID0 一般适用于对性能要求严格但对数据安全性和可靠性不高的应用，如视频、音频存储、临时数据缓存空间等。



RAID1

RAID1 称为镜像，它将数据完全一致地分别写到工作磁盘和镜像 磁盘，它的磁盘空间利用率为 50% 。 RAID1 在数据写入时，响应时间会有所影响，但是读数据的时候没有影响。 RAID1 提供了最佳的数据保护，一旦工作磁盘发生故障，系统自动从镜像磁盘读取数据，不会影响用户工作。

RAID1 与 RAID0 刚好相反，是为了增强数据安全性使两块 磁盘数据呈现完全镜像，从而达到安全性好、技术简单、管理方便。 RAID1 拥有完全容错的能力，但实现成本高。 RAID1 应用于对顺序读写性能要求高以及对数据保护极为重视的应用，如对邮件系统的数据保护。



RAID 2、3、4

RAID2、3、4较少实际应用，它们大多只在研究领域有实作。

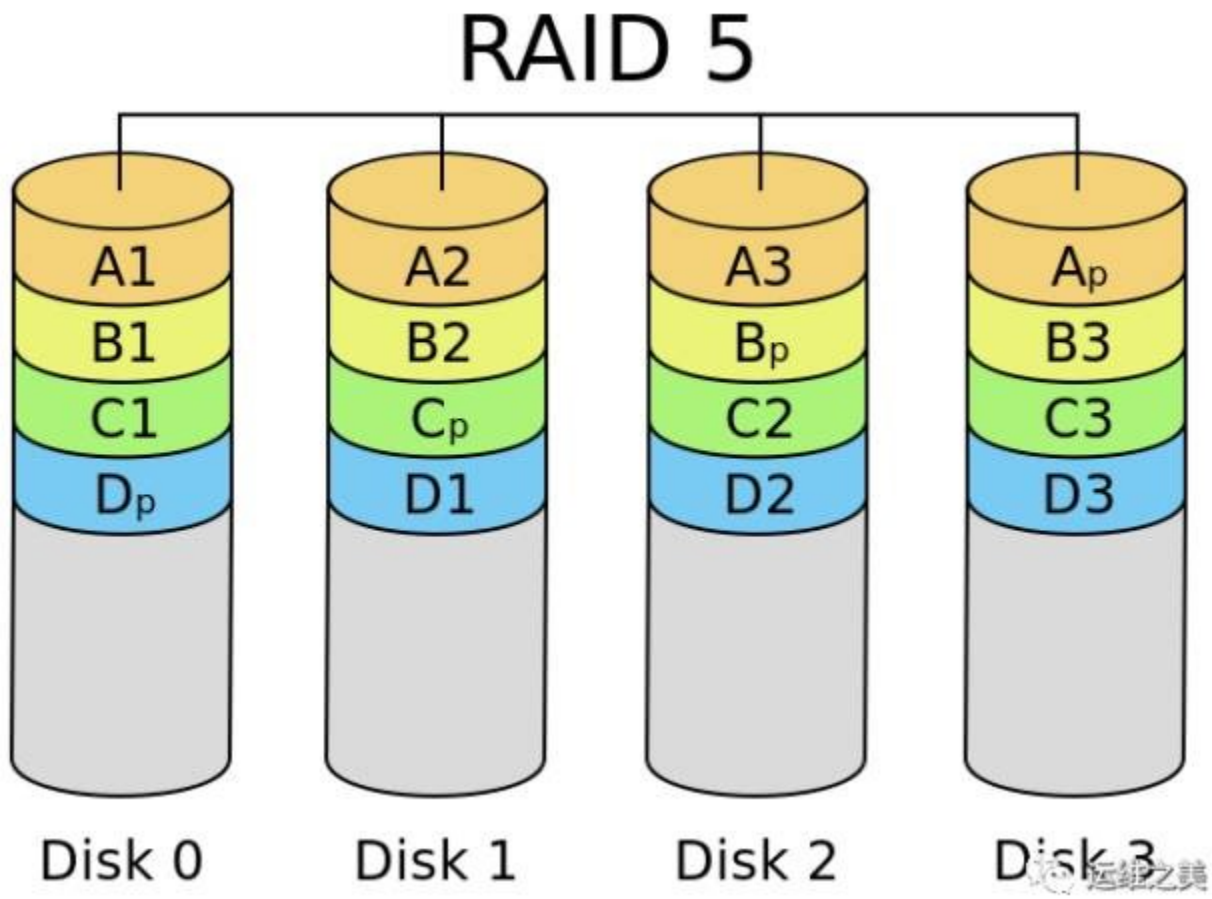
RAID5

RAID5 应该是目前最常见的 RAID 等级，它的原理与 RAID4 相似，区别在于校验数据分布在阵列中的所有磁盘上，而没有采用专门的校验磁盘。对于数据和校验数据，它们的写操作可以同时发生在完全不同的磁盘上。因此， RAID5 不存在 RAID4 中的并发写操作时的校验盘性能瓶颈问题。另外， RAID5 还具备很好的扩展性。当阵列磁盘 数量增加时，并行操作量的能力也随之增长，可比 RAID4 支持更多的磁盘，从而拥有更高的容量以及更高的性能。

RAID5 的磁盘上同时存储数据和校验数据，数据块和对应的校验信息存保存在不同的磁盘上，当一个数据盘损坏时，系统可以根据同一条带的其他数据块和对应的校验数据来重建损坏的数据。与其他 RAID 等级一样，重建数据时， RAID5 的性能会受到较大的影响。

RAID5 兼顾存储性能、数据安全和存储成本等各方面因素，它可以理解为 RAID0 和 RAID1 的折中方案，是目前综合性能最佳的数据保护解决方案。 RAID5 基本上可以满足大部分的存储应用需

求，数据中心大多采用它作为应用数据的保护方案。

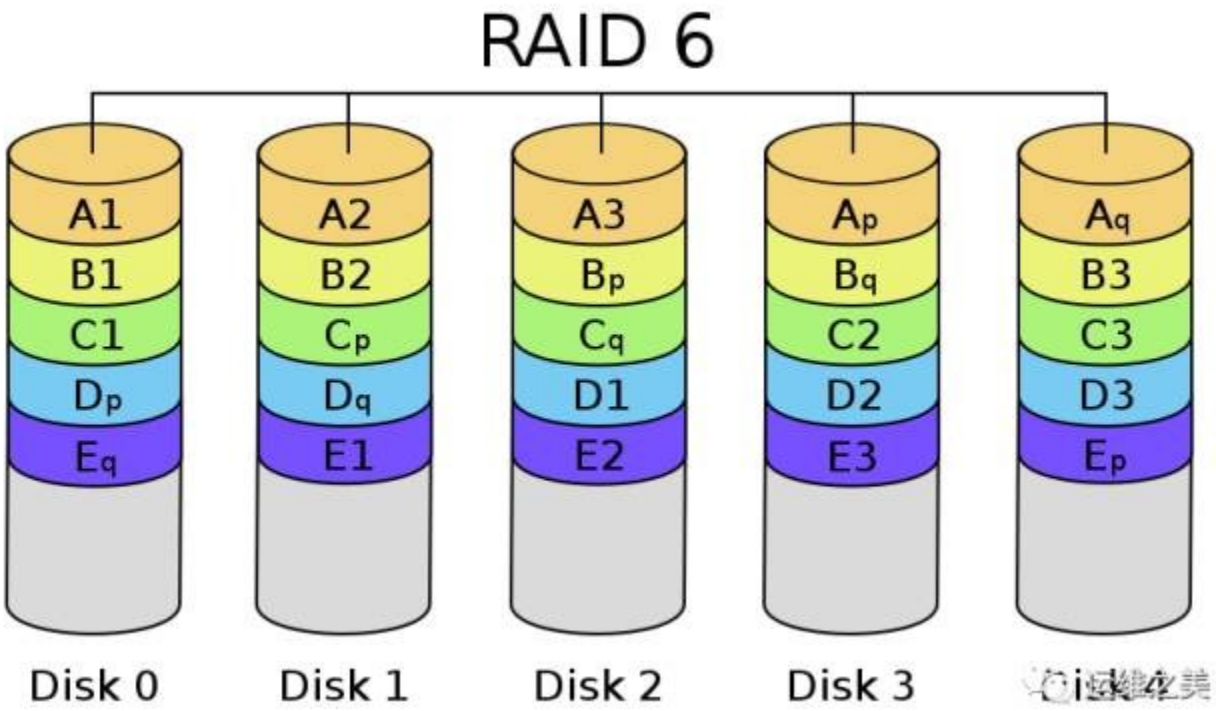


RAID6

前面所 述的各个 RAID 等级都只能保护因单个磁盘失效而造成的数据丢失。如果两个磁盘同时发生故障，数据将无法恢复。 RAID6 引入双重校验的概念，它可以保护阵列中同时出现两个磁盘失效时，阵列仍能够继续工作，不会发生数据丢失。 RAID6 等级是在 RAID5 的基础上为了进一步增强数据保护而设计的一种 RAID 方式，它可以看作是一种扩展的 RAID5 等级。

RAID6 不仅要支持数据的恢复，还要支持校验数据的恢复，因此实现代价很高，控制器的设计也比其他等级更复杂、更昂贵。 RAID6 思想最常见的实现方式是采用两个独立的校验算法，假设称为 P 和 Q ，校验数据可以分别存储在两个不同的校验盘上，或者分散存储在所有成员磁盘中。当两个磁盘同时失效时，即可通过求解两元方程来重建两个磁盘上的数据。

RAID6 具有快速的读取性能、更高的容错能力。但是，它的成本要高于 RAID5 许多，写性能也较差，并有设计和实施非常复杂。因此， RAID6 很少得到实际应用，主要用于对数据安全等级要求非常高的场合。它一般是替代 RAID10 方案的经济性选择



RAID 7

RAID 7并非公开的RAID标准，而是Storage Computer Corporation的专利硬件产品名称。

RAID 7的效能超越了许多其他RAID标准的实做产品，但也因为如此，在价格方面非常的高昂。

RAID 组合等级

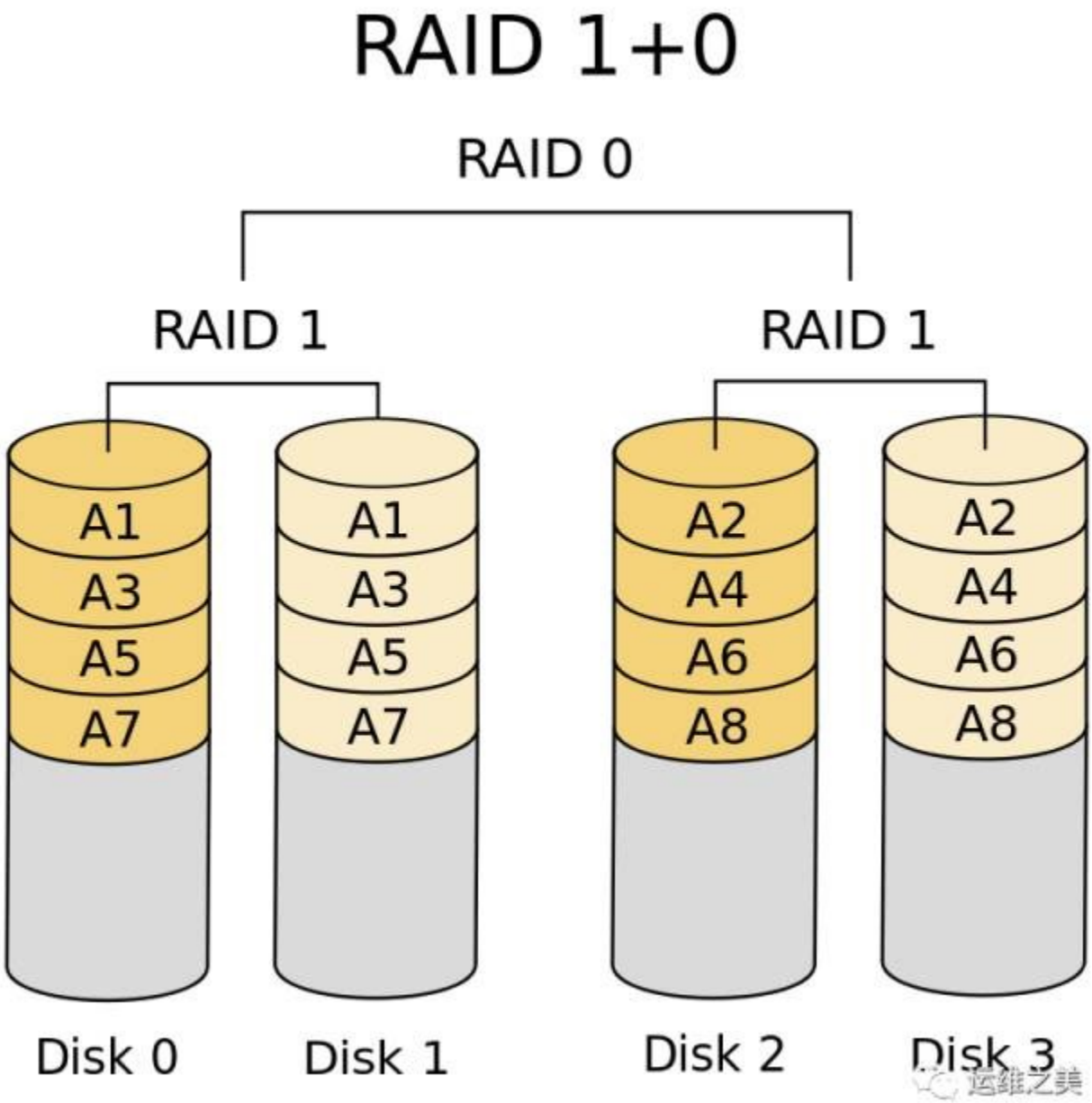
标准 RAID 等级各有优势和不足。自然地，我们想到把多个 RAID 等级组合起来，实现优势互补，弥补相互的不足，从而达到在性能、数据安全性等指标上更高的 RAID 系统。目前在业界和学术研究中提到的 RAID 组合等级主要有 RAID00 、 RAID01 、 RAID10 、 RAID100 、 RAID30 、 RAID50 、 RAID53 、 RAID60 ，但实际得到较为广泛应用的只有 RAID01 和 RAID10 两个等级。当然，组合等级的实现成本一般都非常昂贵，只是在 少数特定场合应用。

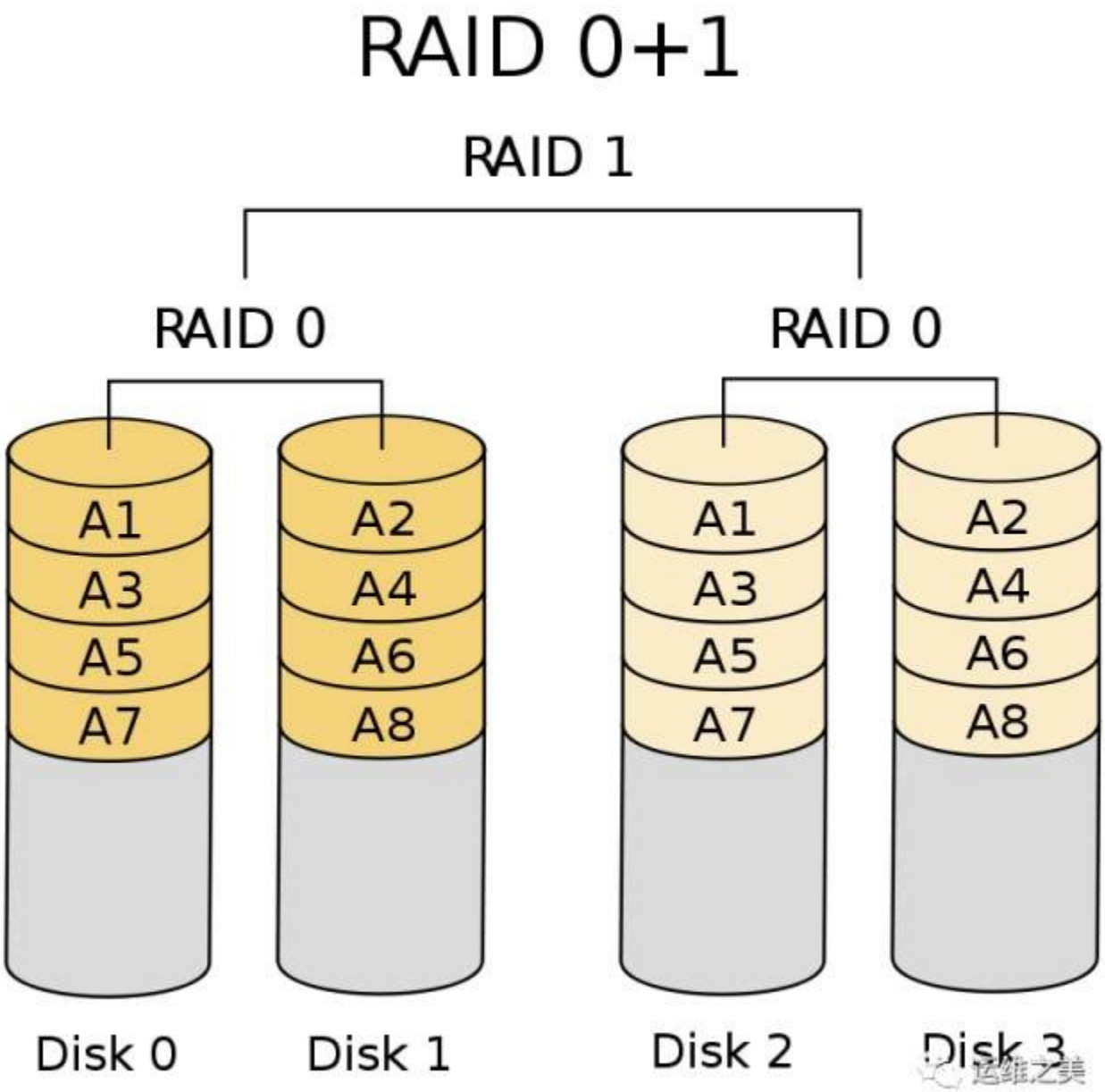
RAID10 和 RAID01

一些文献把这两种 RAID 等级看作是等同的，本文认为是不同的。 RAID01 是先做条带化再作镜像，本质是对物理磁盘实现镜像；而 RAID10 是先做镜像再作条带化，是对虚拟磁盘实现镜像。相同的配置下，通常 RAID01 比 RAID10 具有更好的容错能力。

RAID01 兼备了 RAID0 和 RAID1 的优点，它先用两块磁盘建立镜像，然后再在镜像内部做条带化。 RAID01 的数据将同时写入到两个磁盘阵列中，如果其中一个阵列损坏，仍可继续工作，保证数据安全性的同时又提高了性能。 RAID01 和 RAID10 内部都含有 RAID1 模式，因此整体磁盘

利用率均仅为 50% 。


RAID 1+0
RAID 0
RAID 1
RAID 1
A1
A3
A5
A7
A1
A3
A5
A7
A2
A4
A6
A8
A2
A4
A6
A8
Disk 0
Disk 1
Disk 2
Disk 3



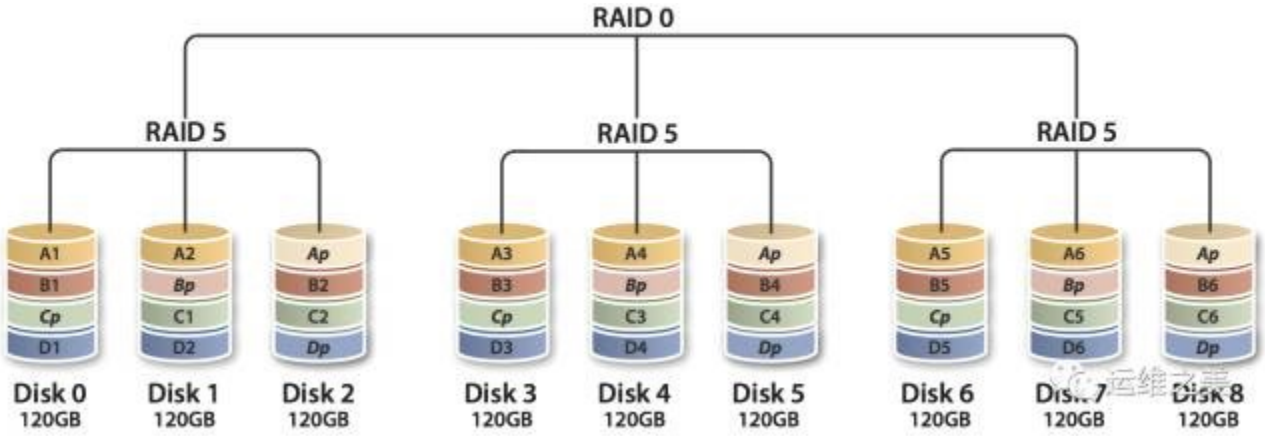
RAID 50

RAID 5与RAID 0的组合，先作RAID 5，再作RAID 0，也就是对多组RAID 5彼此构成Stripe访问。由于RAID 50是以RAID 5为基础，而RAID 5至少需要3颗硬盘，因此要以多组RAID 5构成RAID 50，至少需要6颗硬盘。以RAID 50最小的6颗硬盘配置为例，先把6颗硬盘分为2组，每组3颗构成RAID 5，如此就得到两组RAID 5，然后再把两组RAID 5构成RAID 0。

RAID 50在底层的任一组或多组RAID 5中出现1颗硬盘损坏时，仍能维持运作，不过如果任一组RAID 5中出现2颗或2颗以上硬盘损毁，整组RAID 50就会失效。

RAID 50由于在上层把多组RAID 5构成Stripe，性能比起单纯的RAID 5高，容量利用率比RAID5要低。比如同样使用9颗硬盘，由各3颗RAID 5再组成RAID 0的RAID 50，每组RAID 5浪费一颗硬

盘，利用率为(1-3/9)，RAID 5则为(1-1/9)。

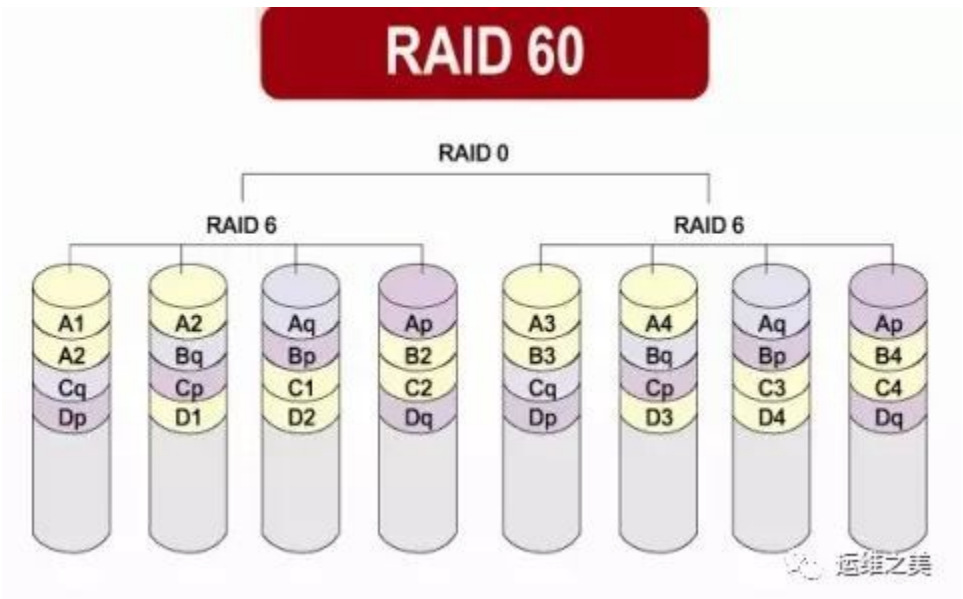


RAID 60

RAID 6与RAID 0的组合：先作RAID 6，再作RAID 0。换句话说，就是对两组以上的RAID 6作Stripe访问。RAID 6至少需具备4颗硬盘，所以RAID 60的最小需求是8颗硬盘。

由于底层是以RAID 6组成，所以RAID 60可以容许任一组RAID 6中损毁最多2颗硬盘，而系统仍能维持运作；不过只要底层任一组RAID 6中损毁3颗硬盘，整组RAID 60就会失效，当然这种情况的概率相当低。

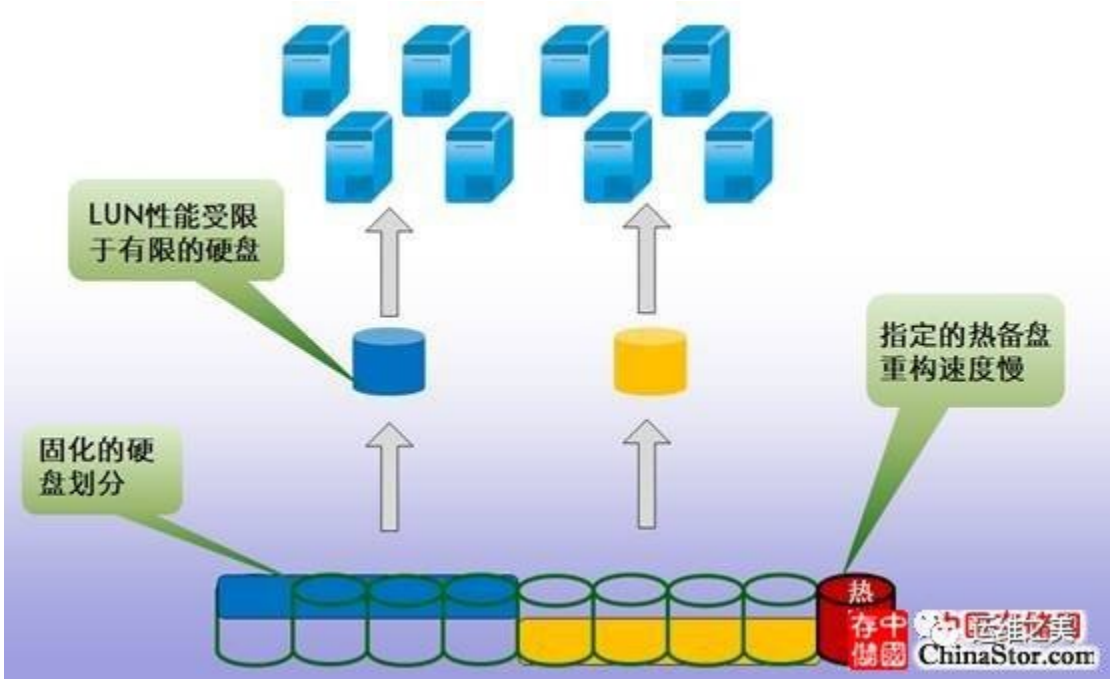
比起单纯的RAID 6，RAID 60的上层通过结合多组RAID 6构成Stripe访问，因此性能较高。不过使用门槛高，而且容量利用率低是较大的问题。



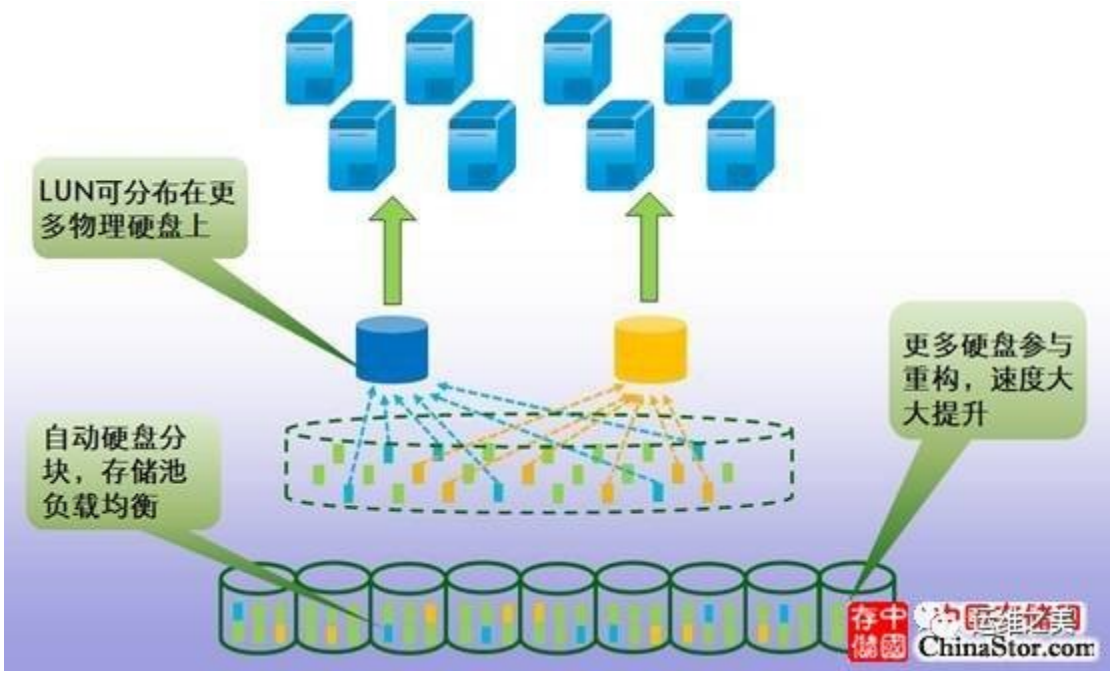
理解 RAID 2.0 和 RAID 2.0+

RAID 2.0（独立磁盘冗余数组2.0, Redundant Array of Independent Disks Version 2.0），为增强型RAID技术，有效解决了机械硬盘容量越来越大，重构一块机械硬盘所需时间越来越长，传统RAID组重构窗口越来越大而导致重构期间又故障一块硬盘而彻底丢失数据风险的问题。其基本思想就是把大容量机械硬盘先按照固定的容量切割成多个更小的分块（Chunk，通常为64MB），RAID组建立在这些小分块上，而不是某些硬盘上，我们称为分块组（Chunk Group）。此时硬盘间不再组成传统的RAID关系，而是组成更大硬盘数量的硬盘组（建议最大硬盘数量为96-120,不建议超过120块盘），每个硬盘上不同的分块可与此硬盘组上不同硬盘上的分块组成不同RAID类型的分块组，这样一个硬盘上的分块可以属于多个RAID类型的多个分块组。以这样的组织形式，基于 RAID2.0 技术的存储系统能够做到在一块硬盘故障后，在硬盘组上的所有硬盘上并发进行重构，而不再是传统 RAID 的单个热备盘上进行重构，从而大大降低重构时间，减少重构窗口扩大导致的数据丢失风险，在硬盘容量大幅增加的同时确保存储系统的性能和可靠性。RAID 2.0 并没有改变传统的各种RAID类型的算法，而是把RAID范围缩小到分块组上。因此，RAID2.0技术具备以下技术特征：

- 几个、几十个甚至上百个机械硬盘组成硬盘组；
- 硬盘组中的硬盘被分割成几十兆、上百兆的分块，不同硬盘上的分块组成的分块组 (Chunk Group)；
- RAID 计算在分块组 (Chunk Group) 内进行，系统不再有热备盘，而是被同一分块组内保留的热备块所代替。



基于传统RAID技术的存储阵列故障恢复机制



基于RAID 2.0技术的存储阵列故障恢复机制

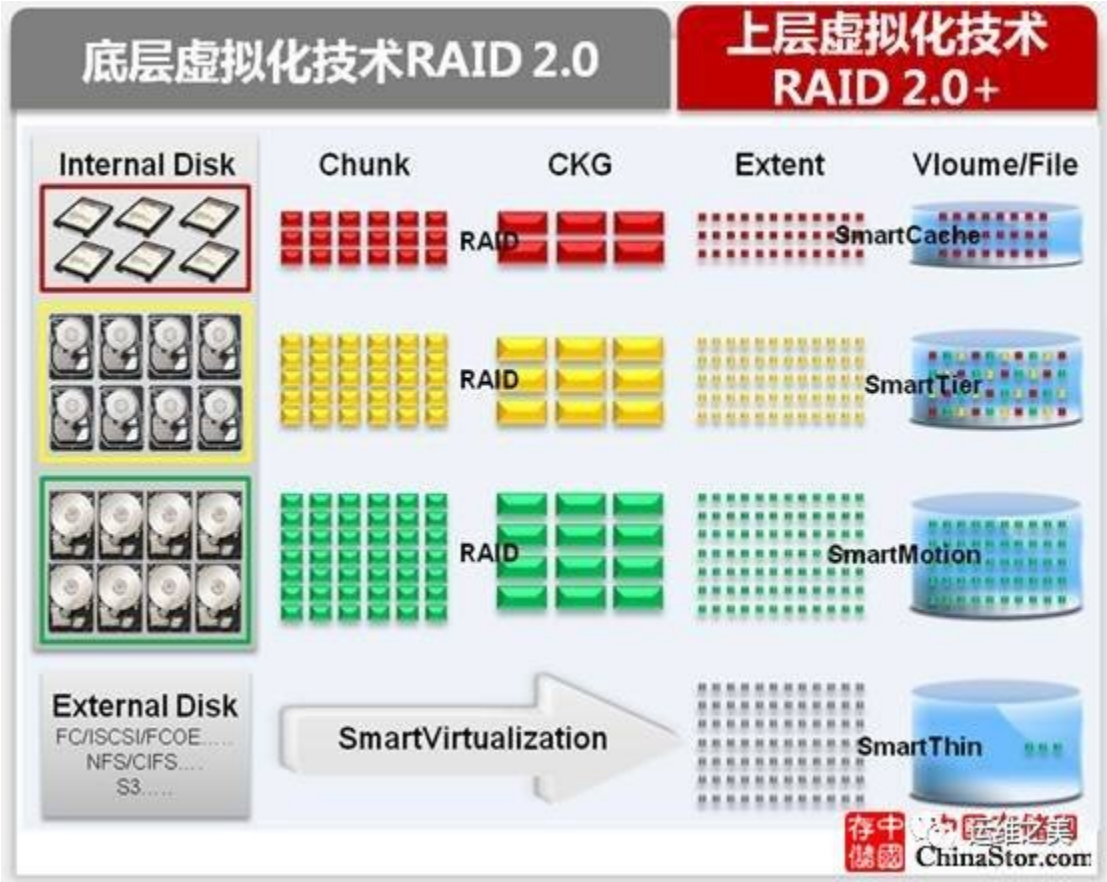
由于RAID 2.0系统中一块硬盘故障后，重构可以在同一硬盘组内其他所有硬盘保留的热备空间上并发进行，使用RAID 2.0技术的存储系统具备以下优势：

- 快速重构：存储池内所有硬盘参与重构，相对于传统RAID重构速度大幅提；
- 自动负载均衡：RAID 2.0使得各硬盘均衡分担负载，不再有热点硬盘，提升了系统的性能和硬盘可靠性；
- 系统性能提升：LUN基于分块组创建，可以不受传统RAID硬盘数量的限制分布在更多的物理硬盘上，因而系统性能随硬盘IO带宽增加得以有效提升；
- 自愈合：当出现硬盘预警时，无需热备盘，无需立即更换故障盘，系统可快速重构，实现自愈合。

RAID2.0+（独立磁盘冗余数组2.0, Redundant Array of Independent Disks Version 2.0+）在RAID 2.0 的基础上提供了更细粒度 (可以达几十KB粒度) 的资源颗粒，形成存储资源的标准分配及回收单位，类似计算虚拟化中的虚拟机，我们称之为虚拟块。这些容量单位一致的虚拟块构成了一个统一的存储资源池，所有应用、中间件、虚拟机、操作系统所需的资源可以在这个资源池中按需分配及回收。相对传统 RAID 系统，RAID2.0+ 技术实现了存储资源的虚拟化及预配置，存储资源的申请及释放完全自动化的通过存储池实现，而不再需要传统 RAID 阵列的RAID组创建，LUN创建，LUN格式化等耗时而容易出错的手工配置过程。因此，RAID 2.0+ 技术解决了虚拟机环境下，存储资源必须动态按需分配及释放的问题。

在 RAID 2.0 基础上，RAID2.0+技术具备以下技术特征：

- 在 RAID 2.0 基础上，分块组（Chunk Group）被切分为容量从256KB到64MB的虚拟化存储颗粒（Extent）；
- 存储资源以以上颗粒为单位自动分配及释放；
- 可以以以上颗粒度为单位在存储池内或不同存储池间进行细粒度分级存储；
- 在系统通过扩展控制器扩展性能或容量后，可以通过自动化的迁移这些标准颗粒来达到负载均衡的目的。

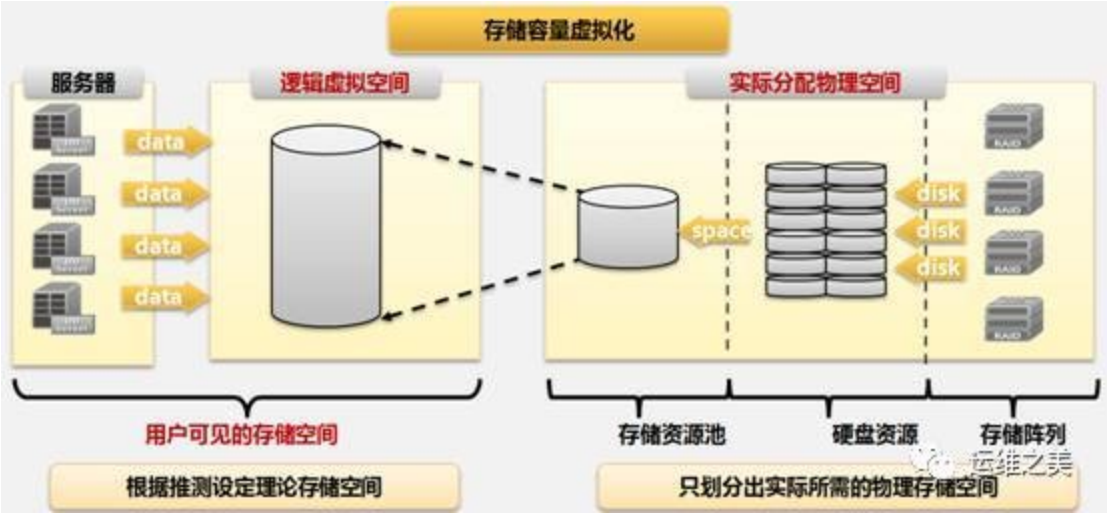


基于RAID 2.0+技术的存储阵列

技术优点

RAID2.0+技术主要用于实现系统资源的智能分配，满足虚拟机环境对存储的需求：

- 存储资源按需自动化分配及释放，满足了虚拟机对存储最本质的需求；



基于RAID 2.0+技术的存储容量虚拟化

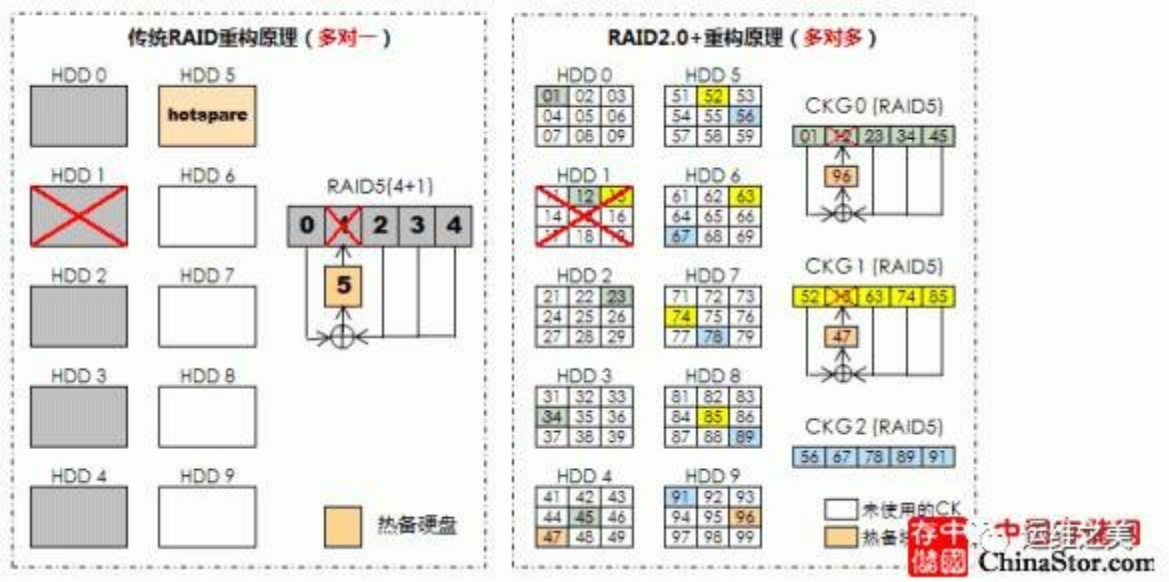
- 可根据业务实时情况，将不同数据分级存储，通过灵活调配SSD等高性能存储资源满足高性能业务需求；

基于RAID 2.0+技术的实时资源调配

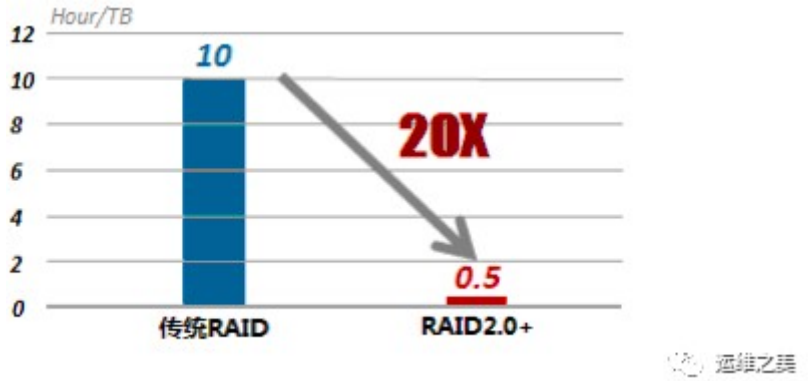
- 根据业务特点自动迁移数据，提高存储利用效率；

基于RAID 2.0+技术的自动数据迁移

RAID 2.0与传统RAID的对比



传统 RAID 和 RAID2.0+两种技术重构原理的对比



传统 RAID 和 RAID2.0+两种技术重构性能的对比

不适合使用RAID的场景

Hadoop集群中为何不使用RAID？

尽管建议采用RAID (Redundant Array of Independent Disk,即磁盘阵列) 作为 namenode 的存储器以保护元数据，但是若将 RAID 作为 datanode 的存储设备则不会给 HDFS 带来益处。HDFS 所提供的节点间数据复制技术已可满足数据备份需求，无需使用 RAID 的冗余机制。

此外，尽管 RAID 条带化技术 (RAID 0) 被广泛用户提升性能，但是其速度仍然比用在 HDFS 里的 JBOD (Just a Bunch Of Disks) 配置慢。JBOD 在所有磁盘之间循环调度 HDFS 块。RAID 0 的读写操作受限于磁盘阵列中最慢盘片的速度，而 JBOD 的磁盘操作均独立，因而平均读写速度高于最慢盘片的读写速度。需要强调的是，各个磁盘的性能在实际使用中总存在相当大的差异，即使对于相同型号的磁盘。针对某一雅虎集群的评测报告(http://markmail.org/message/xmzc45zi25htr7ry)表明，在一个测试(Gridmix)中，JBOD 比 RAID 0 快10%；在另一测试(HDFS写吞吐量)中，JBOD 比 RAID 0 快30%。

最后，若 JBOD 配置的某一磁盘出现故障，HDFS 可以忽略该磁盘，继续工作。而 RAID 的某一盘片故障会导致整个磁盘阵列不可用，进而使相应节点失效。

参考文档

http://www.google.com
https://wsgzao.github.io/post/raid/
http://blog.csdn.net/load2006/article/details/9176891
http://www.cnblogs.com/Richardzhu/p/5315064.html
http://www.chinastor.com/a/jishu/raid/010910b32015.html

如果你觉得内容很赞，还等什么？快快长按打赏吧，iOS的土豪们也是可以的哟！

更多精彩热文：

- 电子期刊下载 | 2017年04期
- 利用ngx_http_mirror_module实现流量镜像
- Redis高可用架构最佳实践
- MySQL 5.7多源复制实践

- MySQL 5.7并行复制实践

阅读原文