Written by arstercz
on November 21, 2016

raid 控制器对系统性能的影响

raid 控制器介绍

raid 控制器的初衷设计是为了提高系统的性能, 通过管理将一组磁盘当做一个逻辑的单元进行处理, 进而使得系统更加稳定, 而且也具备了容错特性. 大多数企业的基础设施都以 raid 控制器作为保护磁盘数据的解决方案, 并提供更高的读写IO. 更多介绍见 intel 官方站点: intel-raid-controllers 不过在一般业务系统中, 我们通常使用 raid 控制器来满足数据的一致性需求. 最常见的则是在数据库系统中, 在是否使用 raid 以及使用那种 raid 的情况下, 数据的读写以及数据的一致性都会有很大的差别, 不同的 raid 级别见 Standard_RAID_levels.

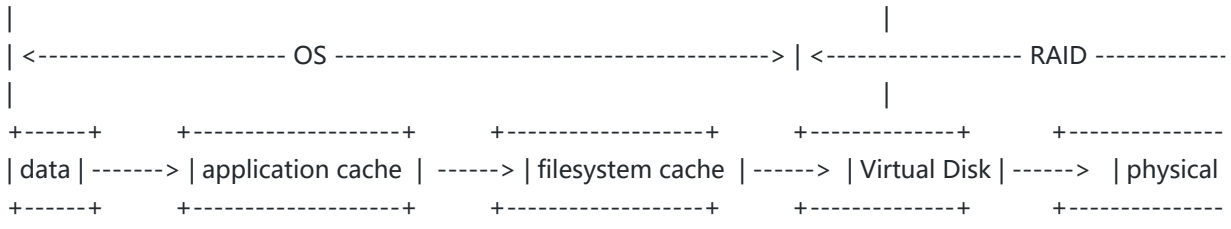
不同策略对系统的影响

这里我们主要以 dell perc 系列的 raid 控制器介绍不同策略对系统性能以及数据一致性的影响. 主要通过以下几方面介绍不同策略对系统性能的影响:

数据如何写到磁盘上
raid write policy
disk cache
read ahead policy
I/O policy
strip size

数据如何写到磁盘上

首先我们了解下数据如何落到磁盘上, 对于应用程序而言, 其产生的数据如果要落到磁盘介质上大致需要经过以下的流程:



上面的流程假定主机已经做好了 RAID. 我们将整个流程分为两部分, 第一部分为系统处理过程, 第二部分为 RAID 处理过程. 应用程序产生数据, 为了性能考虑可能会将数据放到自身的 cache 中, 以便后续的批量处理, 之后数据经过文件系统, 同样文件系统为了性能考虑不会立即刷新数据, 可能对程序写的数据进行缓存. 之后数据到达虚拟磁盘(VD, 通过 RAID 创建后的块设备). 当然程序也可能为了数据安全着想, 直接使用 direct io 的方式写数据, 这种情况下数据会跳过程序缓存和文件系统缓存而直接到达虚拟磁盘. 数据到达虚拟磁盘后就脱离了应

用和操作系统的范围, 进入到 RAID 控制器的范围内, 经过 RAID 的处理后最终使得数据到达物理磁盘. 下文则主要介绍 RAID 控制器的具体策略以及数据在不同策略下的处理过程.

raid write policy (写策略)对性能的影响

raid 控制器中不同的写策略在写的性能上也有很大的差别, 目前主要有两种可用的模式 – WB(write back) 和 WT(write through) .

1. write back

在该模式下, raid 控制器在数据加载到控制器的缓存后就会立即给 I/O 请求返回确认信息, 应用程序可以继续工作而不用等待数据被写到物理磁盘上. write back 模式在大多数情况下都能够提供更好的写性能. 当然写性能提高的情况下也会有不好的方面, 如果在该模式下突然断电, raid 缓存中的数据可能就有丢失的风险, 一般运营商机房中的 UPS(备用电源) 则能缓解数据丢失的风险, 基于这个原因重要业务的服务器都应该放在有 UPS 系统的机房中. raid 卡的 BBU(backup battery status) 特性则提供了另一种保护措施, 在主机异常断电的情况下, 即便没有 UPS, 只要 BBU 电池及电量没有异常, raid 卡里缓存的数据也不会立即丢失. 另外 BBU 在以下情况下会使得 write back 模式切换到 write through 模式:

1. 没有充满电
2. 正在自动校准(auto learn)

raid 卡默认情况下会启用自动校准模式以记录 BBU 电池的放电曲线, 以便控制器了解电池的状态, 这个过程可能需要12小时左右, 期间会禁用 write back 模式来保证数据的完整性, 这也会造成系统性能的降低. 校准(learn cycle) 分为以下步骤:

1. 控制器给 BBU 电池充满电
2. 开始校准, 对 BBU 电池进行放电
3. 放电完成后, 完成校准, 并重新充电, 充到电池的最大电量.

如果第二或第三阶段被中断，重新校准的任务会停止，而不会重新执行. DELL 机器默认 90 天校准一次.

2. write through

在该模式下, 不会使用 raid 控制器的缓存来加速写IO请求. 大多数情况下该模式都慢于 write back 模式, 因为应用程序需要等待数据被写到物理磁盘上. 不过该模式在 raid 级别为 raid 0 或 raid 10 的时候, 如果是顺序写则能提供最高的写带宽(write bandwidth).

disk cache

磁盘缓存策略决定了硬盘的写缓存是否开启. raid控制器设置为 write through 模式的时候, 磁盘缓存策略对系统的写性能影响很大, 想想磁盘没有缓存的时候, 每个写操作都要等待数据写到物理磁盘中, 如果是较多的随机写性能肯定会很差; 相反如果 raid 控制器为 write back 模式, 磁盘缓存是否开启对系统的写性能影响很小, 大多场景下都可以忽略影响. 另外磁盘缓存开启的时候, 如果突然断也会存在丢失数据的风险. 从这个角度看, UPS 备用电源真是至关重要, 重要数据的主机应该放到有 UPS 支持的机房中. raid 卡的 BBU 特性则不能保护磁盘缓存的数据.

read ahead policy

read ahead 策略决定 raid 控制器读取数据的时候是只读取一个块(block)的数据还是整个条带(strip)的数据, 该设置对 raid 卡读的性能影响很大.

1. No Read Ahead

在应用发出请求后 raid 控制器仅读取一个块的数据, 在业务随机读取较多的场景下可以选用该策略.

2. Always Read Ahead

应用发出请求后 raid 控制器会读取整个条带的数据放到 raid cache 中, 每个 read 操作会消耗较多的资源. 在这种模式下, 对主要是顺序读取的业务会有很好的性能提升.

3. Adaptive Read Ahead

该策略下由 RAID 控制器根据读请求的类型自行调整是否预取. 该模式结合了上述两种策略的有点. 所以如果不清楚读取的类型或者业务存在顺序读取和随机读取的请求, 则推荐使用自适应模式.

I/O Policy

RAID 控制器的 I/O 策略决定了是否保留 raid cache 中的数据, 如果应用的请求读的都是同一块数据, raid cache 则直接返回数据, 可以减少很多应用访问的时间. 目前 RAID 控制器的策略主要有两种方式:

1. Direct I/O

直接从磁盘读取数据, 不使用 raid cache 功能, 比起 Cached I/O 要慢很多. 大多数场景下可以使用该策略. 文件系统或应用程序都有自己的缓存, 要不要使用 raid 级别的 cache 便显得不那么重要.

2. Cached I/O

在该种策略下, raid 控制器会把读和写请求先从 raid cache 中获取和更新. 如果有很多读操作都请求同一块数据, 则 raid cache 直接返回数据给读操作. 如果是写操作, 性能则根据 raid write policy 的不同而不同. 另外 MegaCli 可以配置当 BBU 损坏的时候启用 Cache 功能, 默认为 off. `

strip size

条带大小决定了数据怎么分布到硬盘中, 同样由于硬盘数量的原因, 条带大小也可能决定了一次 I/O 请求需要访问多少磁盘. 通常条带大小较大(512 KB 或 1MB)的话就特别适合顺序读取较多的业务, 因为单块读取的数据更多. 如果业务类型是随机访问, 系统性能则依赖于访问数据块的大小和分配的条带大小的值, 比如数据库系统中, 单个记录是 16KB(比如 MySQL 一个页的大小默认是 16KB), 条带大小是也是 16 KB 的情况下, raid 控制器就能很好的提高系统的性能.

Dell 系统 raid 信息查看

dell perc raid 型号详细信息见: dell-raid-controllers 通过上述链接的列表来看, 在实际使用中, 我们更关注 interface support , cache memory size , write back cache 和 raid level 四列信息, 当然在实际采购 raid 卡的时候越高性能的 raid 卡对应价格肯定也越高, 不过对于大多数应用服务器而言, 比如数据库服务, 我们可能更关注以下信息:

interface support: 接口读取数据的速度是否够快, 业务服务器的 sas 盘使用可能更多;
cache memory size: raid 卡是否支持缓存, 缓存有多大;
write back cache: raid 卡是否支持 write back 模式的写策略;
raid level: raid 卡支持的级别, 即通常我们讨论的 raid 级别;

raid level 级别中, 大多数的 raid 卡都支持通用的 raid 级别, interface 的速度差别都不大. 所以如果要采购 raid 卡的话, 该 raid 型号至少要支持缓存以及支持 write back 写策略, 缓存多大按预算的价格决定, 一般 512M 的也不错. 根据以前经历的事件, 在 MySQL 的从数据库中, 没有缓存的 raid 卡每秒仅更新 500KB 的 binlog, 换成 512M 缓存的 raid 卡后每秒可以更新 4MB 的binlog, 完美解决了主从延迟问题.

如何查看 dell 服务器的 raid 信息

以下以 Dell PowerEdge R720 服务举例说明 可以使用使用 MegaCli 命令查看 raid 卡的基本信息.

```
# MegaCli -AdpAllInfo -aALL
.....
Product Name    : PERC H710P Mini
.....
RAID Level Supported            : RAID0, RAID1, RAID5, RAID6, RAID10, RAID50, RAID60
Supported Drives                : SAS, SATA
.....
Memory Size     : 1024MB
.....
Stripe Size                     : 64kB
Flush Time                      : 4 seconds
Write Policy                    : WB
Read Policy                     : Adaptive
Cache When BBU Bad              : Disabled
Cached IO                       : No
.....
```

可以看到我们的 RAID 卡的型号为 H710P Mini，缓存大小为 1G, 目前的写策略为 Write back (WB)，没有启用 Cached IO 策略, 当 BBU 不正常的时候需要从 write back 切换到 write through . 查看 BBU 的状态, 如下现在充电是满的, 差不多90天会校准一次, 已经成功执行了校准操作:

```
# MegaCli -AdpBbuCmd -GetBbuStatus -aALL
BBU status for Adapter: 0
BatteryType: BBU
Voltage: 3927 mV
Current: 0 mA
Temperature: 34 C
Firmware Status: 00001000
Battery state:
GasGuageStatus:
  Fully Discharged        : No
  Fully Charged           : Yes
  Discharging             : No
  Initialized             : No
  Remaining Time Alarm    : No
  Remaining Capacity Alarm: Yes
  Discharge Terminated    : No
  Over Temperature        : No
  Charging Terminated     : No
  Over Charged            : No
Relative State of Charge: 100 %
Charger Status: Complete
Remaining Capacity: 272 mAh
Full Charge Capacity: 274 mAh
isSOHGood: Yes
Exit Code: 0x00
# MegaCli -AdpBbuCmd -GetBbuProperties -aALL
```

查看及设置 raid 卡写策略:

```
# ./MegaCli -LDGetProp -Cache -LAll -aAll
Adapter 0-VD 0(target id: 0): Cache Policy:WriteBack, ReadAdaptive, Direct, No Write Cache if bad BBU
Adapter 0-VD 1(target id: 1): Cache Policy:WriteBack, ReadAdaptive, Direct, No Write Cache if bad BBU
Adapter 0-VD 2(target id: 2): Cache Policy:WriteBack, ReadAdaptive, Direct, No Write Cache if bad BBU
Exit Code: 0x00

# ./MegaCli -LDSetProp WT -L0 -a0
Set Write Policy to WriteThrough on Adapter 0, VD 0 (target id: 0) success
Exit Code: 0x00
# ./MegaCli -LDGetProp -Cache -L0 -a0
Adapter 0-VD 0(target id: 0): Cache Policy:WriteThrough, ReadAdaptive, Direct, No Write Cache if bad BBU
Exit Code: 0x00
```

总结

基于 raid 控制器管理的磁盘阵列的性能受到不同策略因素的影响, 对系统性能的影响也不尽相同. 我们只有在理解了 raid 写策略, disk cache , I/O policy 以及 strip size 等不同的设置后,才能更好的判断以及处理系统的性能问题.

参考

dell-raid-controllers
Optimal-RAID-Performance
Disk_array_controller

← Top →