我们在磁盘写操作持续繁忙的服务器上曾经碰到一个特殊的性能问题。每隔 30 秒，服务器就会遇到磁盘写活动高峰，导致请求处理延迟非常大（超过3秒）。后来上网查了一下资料，通过调整内核参数，将写活动的高峰分布成频繁的多次写， 每次写入的数据比较少。这样可以把尖峰的写操作削平成多次写操作。以这种方式执行的效率比较低，因为内核不太有机会组合写操作。但对于繁忙的服务器，写操 作将更一致地进行，并将极大地改进交互式性能。

下面是相关参数的调整：

一、2.6内核下

1、/proc/sys/vm/dirty_ratio

这个参数控制文件系统的文件系统写缓冲区的大小，单位是百分比，表示系统内存的百分比，表示当写缓冲使用到系统内存多少的时候，开始向磁盘写出数据。增大 之会使用更多系统内存用于磁盘写缓冲，也可以极大提高系统的写性能。但是，当你需要持续、恒定的写入场合时，应该降低其数值，：

echo '1' > /proc/sys/vm/dirty_ratio

2、/proc/sys/vm/dirty_background_ratio

这个参数控制文件系统的pdflush进程，在何时刷新磁盘。单位是百分比，表示系统内存的百分比，意思是当写缓冲使用到系统内存多少的时 候，pdflush开始向磁盘写出数据。增大之会使用更多系统内存用于磁盘写缓冲，也可以极大提高系统的写性能。但是，当你需要持续、恒定的写入场合时， 应该降低其数值，：

echo '1' > /proc/sys/vm/dirty_background_ratio

3、/proc/sys/vm/dirty_writeback_centisecs

这个参数控制内核的脏数据刷新进程pdflush的运行间隔。单位是 1/100 秒。缺省数值是500，也就是 5 秒。如果你的系统是持续地写入动作，那么实际上还是降低这个数值比较好，这样可以把尖峰的写操作削平成多次写操作。设置方法如下：

echo "100" > /proc/sys/vm/dirty_writeback_centisecs 如果你的系统是短期地尖峰式的写操作，并且写入数据不大（几十M/次）且内存有比较多富裕，那么应该增大此数值：

echo "1000" > /proc/sys/vm/dirty_writeback_centisecs

4、/proc/sys/vm/dirty_expire_centisecs

这个参数声明Linux内核写缓冲区里面的数据多“旧”了之后，pdflush进程就开始考虑写到磁盘中去。单位是 1/100秒。缺省是 30000，也就是 30 秒的数据就算旧了，将会刷新磁盘。对于特别重载的写操作来说，这个值适当缩小也是好的，但也不能缩小太多，因为缩小太多也会导致IO提高太快。

echo "100" > /proc/sys/vm/dirty_expire_centisecs

当然，如果你的系统内存比较大，并且写入模式是间歇式的，并且每次写入的数据不大（比如几十M），那么这个值还是大些的好。

5、/proc/sys/vm/vfs_cache_pressure

该文件表示内核回收用于directory和inode cache内存的倾向；缺省值100表示内核将根据pagecache和swapcache，把directory和inode cache保持在一个合理的百分比；降低该值低于100，将导致内核倾向于保留directory和inode cache；增加该值超过100，将导致内核倾向于回收directory和inode cache

缺省设置：100

6、/proc/sys/vm/min_free_kbytes

该文件表示强制Linux VM最低保留多少空闲内存（Kbytes）。 缺省设置：724（512M物理内存）

7、/proc/sys/vm/nr_pdflush_threads

该文件表示当前正在运行的pdflush进程数量，在I/O负载高的情况下，内核会自动增加更多的pdflush进程。

缺省设置：2（只读）

8、/proc/sys/vm/overcommit_memory

该文件指定了内核针对内存分配的策略，其值可以是0、1、2。

0， 表示内核将检查是否有足够的可用内存供应用进程使用；如果有足够的可用内存，内存申请允许；否则，内存申请失败，并把错误返回给应用进程。

1， 表示内核允许分配所有的物理内存，而不管当前的内存状态如何。

2， 表示内核允许分配超过所有物理内存和交换空间总和的内存（参照overcommit_ratio）。

缺省设置：0

9、/proc/sys/vm/overcommit_ratio

该文件表示，如果overcommit_memory=2，可以过载内存的百分比，通过以下公式来计算系统整体可用内存。

系统可分配内存=交换空间+物理内存*overcommit_ratio/100 缺省设置：50（%）

10、/proc/sys/vm/page-cluster

该文件表示在写一次到swap区的时候写入的页面数量，0表示1页，1表示2页，2表示4页。 缺省设置：3（2的3次方，8页）

11、/proc/sys/vm/swapiness

该文件表示系统进行交换行为的程度，数值（0-100）越高，越可能发生磁盘交换。

二、2.4内核下

通过修改文件/proc/sys/vm/bdflush实现。文件中的九个参数含义如下：

nfract：dirty缓冲在缓冲区中的最大百分比。超过这个值将bdflush进程刷新硬盘。当可用内存比较少的情况下，将引发大量的磁盘I/O。为了均衡磁盘I/O，可以保持一个比较低的值。

Ndirty：bdflush进程一次写入磁盘的最大dirty缓冲块数量。这个值比较大将导致I/O急剧增加，如果这个比较小，bdflush进程执行不够从而可能导致内存的瓶颈。

Dummy2 ：未使用

Dummy3：未使用

Interval：kupdated工作和刷新的最小频率，默认值是5秒。最小值是0秒最大值是600秒。

Age_buffer：缓冲数据写到磁盘之前操作系统等待的最大时间。默认值是30秒，最小值是1秒最大值是6000秒。

Nfract_sync：dirty缓存激活bdflush进程同步的百分比。默认值是60%。

Nfract_stop：dirty缓存停止bdflush进程的百分比。默认值是20%。

Dummy5：未使用

比如在一个写操作频繁的数据库服务器上设置：

10 500 0 0 50 30 10 0 0

----------华---丽---的---分---割---线----------

网上有很多都在问如何限制 cache 的大小，找了一轮都没有找到答案，其中一个方法就是修改 /proc/sys/vm/min_free_kbytes 这个文件，把它的值设置大一点，cache 就相应的受到限制，但是把这个值调大后会不会对系统有其它影响，暂时未知。欢迎大牛们提供点意见，谢谢！