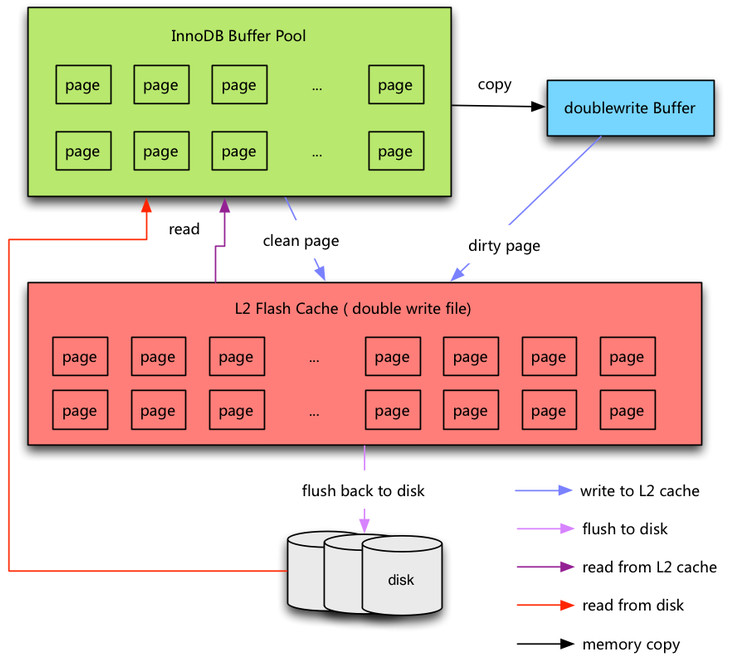
# L2 Cache简介

固态硬盘（Solid State Disk, SSD）的出现给数据库的工作方式带来了深远的影响。InnoSQL 5.5.8版本支持InnoDB Secondary Buffer Pool。通过SSD设备来做辅助缓冲池。该技术在大规模读的应用环境中有着非常不错的性能表现，但是在TPC-C这类测试中，可能就显的比较平庸。因为其只是辅助缓冲池，而非真正意义的Flash Cache。

InnoSQL 5.5.13开始支持InnoDB L2 Cache，以提升IO性能，进而提升数据库的TPS。虽然同样采用SSD作为一个Cache，但是两者的实现方式完全不同。

## 架构设计

InnoSQL L2 Cache架构上采用全局哈希表来管理Cache中缓存的page，基于Round方式来进行缓存数据写入和脏数据回刷，回刷线程为独立的后台线程。流程上主要包括缓存数据写入、脏数据向磁盘回刷、缓存数据读命中和脏数据备份等，架构和流程框图如下所示：



### 数据写入

在InnoDB存储引擎中，为了避免half-written的问题，采用了Doublewrite的设计方式。其工作方式为：在刷新128个页前（最多128个页），先顺序地将128个页写入到Doublewrite中，在确保写入后则进行页的实际写入。

InnoSQL L2 Cache接管了原先Doublewrite的工作，并扩大了原先Doublewrite的尺寸。例如，用户可以拥有一个300G的Doublewrite。因为Doublewrite的特性，对于SSD的写入都是顺序的。当页刷新到L2 Cache后建立对应的Hash表，那么之后读也可能会命中到L2 Cache，因为刷新完成的脏页非常可能从缓冲池中移出。此外，当写入L2 Cache后，并不像Doublewrite那样去马上完成页的刷新操作，而是通过后台的L2 Cache Flush Thread来完成这个操作。L2 Cache中可能存在一个页的多个副本，当从L2 Cache中刷回磁盘时，如果这个页已经有新的版本了，那老版本的页将不进行刷新，这个操作称为merge write。

L2 Cache还会将BP中换出的clean page有选择的写入到Cache中，存在两种情况，换出的page不存在Cache中和已在Cache中，如果是前一种情况，那么该page会写入到Cache，若为后一种情况，则会基于所设置的阈值参数选择是否将本次换出的page重新写入。

L2 Cache基于Round的写入是顺序的，将SSD文件从偏移位置0开始写到innodb\_flash\_cache\_size，完成一遍写入后，会重新开始从偏移位置0写入，这样的算法符合SSD的特性，即充分利用了SSD的高速离散读取性能又避免了离散写入带来的性能下降和写入寿命问题。

### 数据回刷

当从L2 Cache将页刷回磁盘时，目前的算法为：

1、当L2 Cache中待刷新的页，小于innodb\_flash\_cache\_write\_cache\_pct \* innodb\_flash\_cache\_size / 16384时，不刷新页；

2、当L2 Cache中待刷新的页，大于innodb\_flash\_cache\_write\_cache\_pct \* innodb\_flash\_cache\_size / 16384，小于innodb\_flash\_cache\_do\_full\_io\_pct \* innodb\_flash\_cache\_size / 16384时，刷新10%的innodb\_io\_capacity / innodb\_flash\_cache\_io\_capacity页；

3、否则，刷新innodb\_io\_capacity / innodb\_flash\_cache\_io\_capacity个页。

该算法的目的是为了L2 Cache尽可能的cache大量的写入页，以此来提高merge write的比例。

### 缓存命中

当所需访问的page不在BP中时，在进行磁盘读取操作前会基于page的（space,offset）查找L2 Cache哈希表，若该page已经缓存在Cache中，则只需从Cache将其读取即可，大大提高了IO性能。InnoSQL 5.5.30-v4开始，将用于保护哈希表的锁从mutex改为rwlock，提供了并发访问哈希表的能力，提高了读Cache的并发性。

# L2 Cache特性说明

## 特性介绍

InnoDB L2 Cache特点简介如下：

1、读写Cache，可以缓存buf pool（BP）换出的clean page，同时也可以缓存BP回刷磁盘的dirty page，page写到SSD是顺序化的，没有随机写；

2、使用全局哈希表进行索引，快速定位page在SSD上的位置，尽可能地减少查找开销，提高读性能；

3、使用一个刷新线程异步且自适应地将SSD中的dirty page刷回磁盘，可以合并一个页的多个版本，只做一次刷回，大大减少了实际页写磁盘次数；

4、支持压缩，可以缓存InnoDB的压缩页；

5、支持在线备份，提供fcbck2databck配合xtrabackup使用；

6、重启后可快速恢复关闭前的缓存热点状态；

7、系统crash后能较快速恢复（与L2 Cache文件大小有关）；

8、预热时间大幅缩短。

从InnoSQL 5.5.30-v4开始，L2 Cache还支持以下特性：

1、可将缓存数据的SSD block管理单位设置为1KB/2KB/4KB/8KB/16KB，尽可能高效得利用Cache空间，减少浪费；

2、添加了L2 Cache压缩功能，用于增加L2 Cache能够缓存的数据量；

3、减小L2 Cache中的锁粒度，提高读缓存page的并发性；

4、重定义备份文件格式，增加备份文件中数据校验功能；

5、在SHOW ENGINE INNODB STATUS中添加L2 Cache更多的显示信息。

## 使用方法

### 开启和关闭

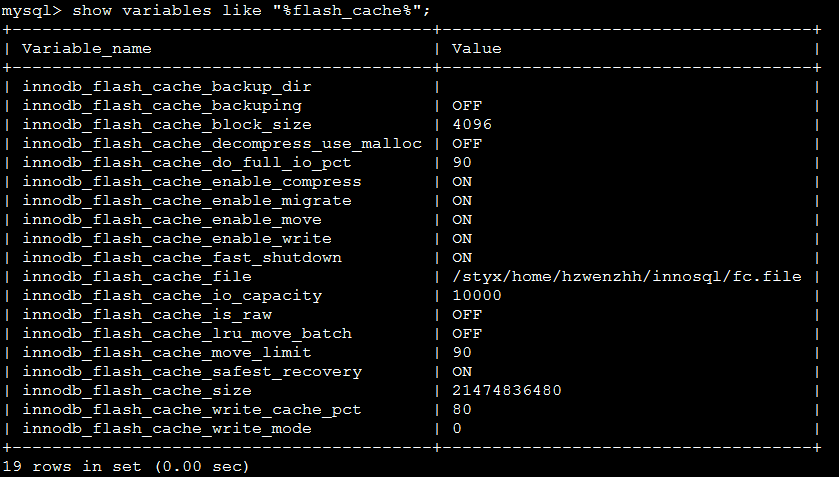
若用户在配置文件中配置了L2 Cache相关的参数，那么其会随着InnoDB一起启动，在InnoDB启动时，会判断是创建一个新的L2 Cache还是加载或者恢复已有的L2 Cache。判断的依据是数据库的datadir下是否存在可用的flash\_cache.log和flash\_cache.dump。

若datadir下两个文件均不存在，那么InnoDB会创建一个新的L2 Cache，若flash\_cache.dump不存在，说明上一次L2 Cache（msyqld）未正常关闭，需要通过flash\_cache.log和innodb\_flash\_cache\_file两个文件将L2 Cache恢复到可用状态。若两个文件均存在，那么从L2 Cache正常关闭时所生成的flash\_cache.dump文件中将Cache元数据加载到内存即可完成快速启动。

L2 Cache只能随InnoDB一起关闭，若用户临时不想使用L2 Cache，可以关闭L2 Cache的写入功能，使BP的脏数据回刷走原有的Doublewrite流程，并BP中换出的page直接丢弃。

## 参数配置和使用

本文档主要针对InnoSQL 5.5.30-v4版本中的L2 Cache进行使用介绍。该可配置的参数如下：



与之前版本相比，InnnoSQL 5.5.30-v4版本中的L2 Cache暂时少了两个warmup参数：innodb\_flash\_cache\_warmup\_table和innodb\_flash\_cache\_warmup\_file。将在后续的版本开发中重新加入。各参数的作用及配置介绍如下：

### innodb\_flash\_cache\_file

L2 Cache文件所在的文件路径，默认为空。若没有手工创建，则InnoSQL在启动时，会根据innodb\_flash\_cache\_size的大小自动生成该文件。因为L2 Cache较大，生成文件需要较长的时间，建议用户自动生成该文件。在Linux操作系统下，可以根据以下方式来创建L2 Cache文件，例如：

root@db-51:/mnt/ddb/2# dd if=/dev/zero of=ib\_fcfile bs=16384 count=63356

63356+0 records in

63356+0 records out

1038024704 bytes (1.0 GB) copied, 1.0725 s, 968 MB/s

root@db-51:/mnt/ddb/2# chown -R mysql:mysqls ib\_fcfile

上述通过Linux自带的dd命令产生了一个大小为1G的L2 Cache文件，最后不要忘记将L2 Cache文件的权限设置为数据库用户的权限。

### innodb\_flash\_cache\_size

该参数决定本次L2 Cache运行的缓存大小，注意该大小必须小于等于L2 Cache文件本身的大小。目前L2 Cache创建后不允许改变大小，请确保在mysqld重启前后，该值保持不变，除非想删除flash\_cache.log和flash\_cache.dump重新创建L2 Cache。默认为0，表示不配置L2 Cache。

### innodb\_flash\_cache\_block\_size

该参数从Innosql-5.5.30-v4版本开始支持，用于设置L2 Cache缓存数据的SSD块管理单位，可以配置的值有1024/2048/4096/8192/16384（byte）。请确保在mysqld重启前后，该值保持不变，如果要启用下文所述的压缩功能，请将该参数设置为小于16KB的值。默认值为4096，静态参数，不能动态改变其大小。

### innodb\_flash\_cache\_is\_raw

该参数可将基于固态硬盘的裸设备作为L2 Cache。该值默认为OFF。裸设备的制作与配置方法如下：

root@xxx:# modprobe raw

root@xxx:# raw /dev/raw/raw1 /dev/sdc1

root@xxx:# chown mysql:mysql /dev/raw/raw1

vi /etc/my.cnf

[mysqld]

innodb\_flash\_cache\_file = /dev/raw/raw1

innodb\_flash\_cache\_is\_raw = 1

### innodb\_flash\_cache\_warmup\_table

该参数控制L2 Cache的预热，在Innosql-5.5.30-v4版本中暂时禁用。第一次启用L2 Cache时，用户可以指定将表中全部或部分page读取到L2 Cache中，因为是顺序读，因此可极大的提高L2 Cache的预热速度，比使用记录级的select方式预热更好。该参数可以设置为：

innodb\_flash\_cache\_warmup\_table=tpcc.\* innodb\_flash\_cache\_warmup\_table=tpcc.warehouse:tpcc.orders innodb\_flash\_cache\_warmup\_table=tpcc.\*:user.\*

tpcc.\*表示预热tpcc架构下的所有表。冒号用来分区预热的各个表。

当预热的文件大于L2 Cache大小时，L2 Cache会预热innodb\_flash\_cache\_size大小的文件，然后停止。在启动时，可以从.err文件中观察到：

111103 11:35:19 InnoDB: start to warm up tablespace tpcc.history to flash cache.

111103 11:35:19 InnoDB: warm up table tpcc.history to space: 21 offset 512.(100%)

111103 11:35:19 InnoDB: flash cache is full, warm up stop.

111103 11:35:20 InnoDB: flash cache warm up finish.

### innodb\_flash\_cache\_write\_cache\_pct

该参数控制dirty page刷新的算法。当L2 Cache中脏块比例小于该值时，L2 Cache不回刷dirty page到磁盘。这个阶段可称为cache阶段。该参数在InnoSQL 5.5.30-v4之前版本默认值为50，从InnoSQL 5.5.30-v4版本开始默认为80。

### innodb\_flash\_cache\_do\_full\_io\_pct/ innodb\_flash\_cache\_io\_capacity

该参数控制dirty page刷新的算法。当L2 Cache中脏块比例大于该值时，在InnoSQL 5.5.30-v4之前版本会刷新innodb\_io\_capacity个dirty page，本版本开始会刷新innodb\_flash\_cache\_io\_capacity个dirty page到磁盘上，小于该值时则刷新10%\*innodb\_io\_capacity/innodb\_flash\_cache\_io\_capacity个dirty page到磁盘上。该值越大则可以得到越高的merge write比率。但是如果太大时，BP中替换出或者刷出的page要写入L2 Cache时，则可能会遇到需要等待，该参数默认值为90。

由于innodb\_flash\_cache\_io\_capacity关系到每次刷新dirty page的数量，所以需要根据所选择的设备性能合理设置该参数，默认为10000。

### innodb\_flash\_cache\_fast\_shutdown

该参数控制L2 Cache在关闭时是否需要回刷Cache中的dirty page到磁盘，开启表示在关闭时不需要回刷dirty page，默认开启。

### innodb\_flash\_cache\_enable\_write

该参数控制L2 Cache的Write Cache功能，在备份时使用，默认值为1，表示将L2 Cache作为Write Cache。0表示不将L2 Cache作为Write Cache，BP中LRU换出的clean page不会写入到L2 Cache，BP刷出的dirty page走doublewrite流程写入磁盘，L2 Cache中老版本的page会逐渐被清除。

在xtrabackup备份时，该参数需设置为0，并在备份结束后迅速的设置为1。其余情况都不应设置为0。

### innodb\_flash\_cache\_enable\_migrate

该参数控制L2 Cache是否缓存从BP中换出的clean page，L2 Cache支持将BP中换出的page缓存到SSD上，这样相当于扩展了BP的空间。该功能由参数innodb\_flash\_cache\_enable\_migrate控制，用于将从BP换出，且L2 Cache中未缓存的page缓存到Cache中。该参数默认开启。

### innodb\_flash\_cache\_enable\_move/ innodb\_flash\_cache\_move\_limit

innodb\_flash\_cache\_enable\_move参数控制L2 Cache是否缓存从BP中换出，且L2 Cache中已缓存的page，将其写入到Cache的新位置上，防止刚从BP中换出的page被很快从L2 Cache中替换出去，有将BP次一级的热点数据集中到Cache中的功能。该参数默认开启。

但开启该参数后并不代表所有从BP中换出，且L2 Cache中已缓存的page都会写到Cache的新位置上，还需要满足innodb\_flash\_cache\_move\_limit所设置的阈值条件。innodb\_flash\_cache\_move\_limit表示会对fc->write\_off前进方向上Cache大小的innodb\_flash\_cache\_move\_limit%范围内符合条件的块执行move操作，将其写入到fc->write\_off位置上。防止fc->write\_off的推进使得对应的page被替换出L2 Cache，该参数的默认值为90。

### innodb\_flash\_cache\_lru\_move\_batch

该参数控制L2 Cache在进行move/migrate时，是否将从BP替换出的符合写入到L2 Cache的page先缓存到开辟的内存缓冲区中，带缓冲区满后再将其打包写入Cache中。该参数默认关闭。

### innodb\_flash\_cache\_enable\_compress

该参数控制L2 Cache是否启用压缩，从InnoSQL 5.5.30-v4版本开始支持。L2 Cache不重复压缩已经在InnoDB层面进行了压缩的page，启用该功能后，会在page写入L2 Cache时进行判断，若是未压缩，则会先对其进行压缩，在写入。这样可以减小一个page所占据的Cache空间，从而使固定大小的Cache能够缓存更多的page，测试显示对于sysbench oltp应用，压缩比（压缩前后数据大小对比）在25%左右，对于blogbench应用，压缩比在50%左右。该参数默认开启。

由于压缩和解压需要额外消耗计算资源，开启该参数后，请合理配置计算资源。

### innodb\_flash\_cache\_decompress\_use\_malloc

该参数控制L2 Cache启用压缩时，读命中进行数据解压时所选择的解压缓冲区内存来源，关闭该参数时，选择从BP中分配内存页来作为解压缓存区，该方式最安全，因为总能够分配到所需的页框，但代价加大，正常情况下每次分配都会触发BP中page的换出，可能需要数十微秒的时间，同时也占据了BP的空间，减少了BP可用于缓存page的页框数目，在读较多场景下对数据库性能有一定影响；开启时，选择从系统内存动态分配解压缓冲区，这样会使数据库占用更多的系统内存，增大系统压力，但不会占据BP的内存空间，所以在多线程并发读时能够获得更好的性能。该参数默认关闭。

### innodb\_flash\_cache\_backup\_dir

L2 Cache的备份路径。若不进行设置，则默认在datadir目录下。注意该参数只指定备份路径，备份文件名都为ib\_fc\_backup。

### innodb\_flash\_cache\_backuping

L2 Cache dirty page备份开关，默认OFF。置为TRUE后会将L2 Cache中dirty page备份到文件ib\_fc\_backup。开始备份前，必须将innodb\_flash\_cache\_enable\_write设置为OFF。

### innodb\_flash\_cache\_write\_mode

该参数在InnoSQL 5.5.20-v3版本开始支持。用来区分L2 Cache的模式。0表示Write Back，1表示Write Through。

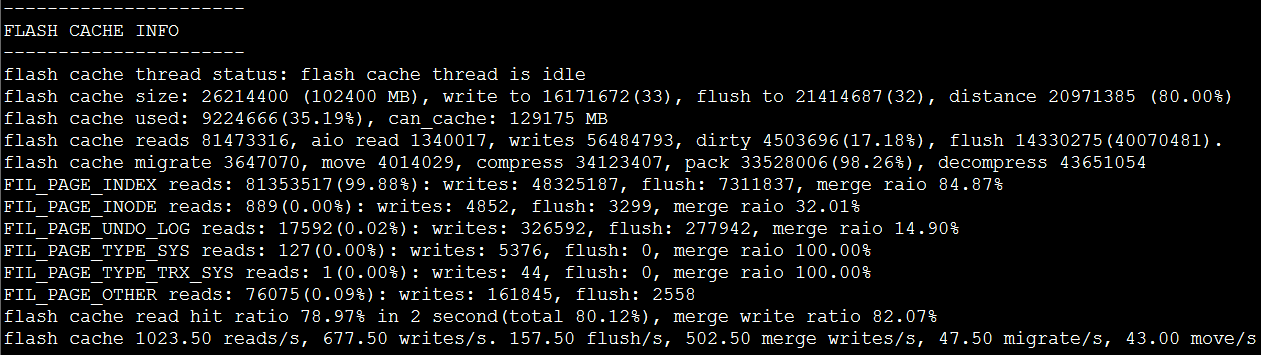
### innodb\_flash\_cache\_safest\_recovery

该参数在InnoSQL 5.5.30-v3版本开始支持。用于在L2 Cache阶段完成基于flash\_cache.log和innodb\_flash\_cache\_file数据恢复后，是否需要将从innodb\_flash\_cache\_file中恢复的数据与磁盘中的数据进行一一比对以确保数据的正确性。

## 状态观察

### SHOW ENGINE INNODB STATUS

在SHOW ENGINE INNODB STATUS的FLASH CACHE INFO下包含了L2 Cache全部状态信息，建议DBA通过该命令来观察flash cache的状态。



flash cache thread status: flash cache thread is idle

表示当前L2 Cache刷dirty page的线程正处于idle状态，此时不回刷dirty page，对应数据回刷算法第一点；当状态为"flusing small flash cache pages"时，对应第二点；当为"flusing full flash cache pages"时，对应第三点。

flash cache size: 26214400 (102400 MB), write to 14346110(33), flush to 19589107(32), distance 20971403 (80.00%)

表示L2 Cache的大小；下一次page写入Cache的block偏移（14346110），已经写入的Round值（33）；下一次L2 Cache回刷线程进行dirty page回刷时的起始block偏移（19589107）已经回刷的Round值（32）；写入block和回刷block之间的距离（20971403）及相对L2 Cache大小的比值。

flash cache used: 9226702(35.20%), can\_cache: 129172 MB, io skip: 135

表示L2 Cache被使用的大小及占Cache总大小比值；Cache能够缓存的数据量大小（若未启用压缩，则该值与Cache大小一致，若启用，则为根据已压缩的page压缩比计算出的Cache能够缓存数据量的估计值，与所设置的innodb\_flash\_cache\_block\_size有关）；L2 Cache在进行写入操作时发现对应位置上的page正被读取的次数

flash cache reads 80162632, aio read 1317745, writes 55573388, dirty 4500495(17.17%), flush 14069132(39420621).

表示L2 Cache中发生的读、异步读、写和回刷的次数；dirty表示L2 Cache中dirty page的数目以及相对L2 Cache大小的比值。

flash cache migrate 3587578, move 3950871, compress 33573805, pack 32988726(98.26%), decompress 42951309

表示L2 Cache中发生的migrate和move的次数；尝试进行压缩的次数，压缩成功的次数及所占百分比；读命中压缩page并进行解压的次数。

FIL\_PAGE\_INDEX reads: 80044752(99.88%): writes: 47544462, flush: 7177947, merge raio 84.90%

FIL\_PAGE\_INODE reads: 880(0.00%): writes: 4779, flush: 3241, merge raio 32.18%

FIL\_PAGE\_UNDO\_LOG reads: 17315(0.02%): writes: 321285, flush: 272903, merge raio 15.06%

FIL\_PAGE\_TYPE\_SYS reads: 127(0.00%): writes: 5249, flush: 0, merge raio 100.00%

FIL\_PAGE\_TYPE\_TRX\_SYS reads: 1(0.00%): writes: 43, flush: 0, merge raio 100.00%

FIL\_PAGE\_OTHER reads: 74858(0.09%): writes: 159218, flush: 2516

各种类型的page的读、写、刷新和merge write数值及比例统计。

flash cache read hit ratio 77.63% in 2 second(total 80.12%), merge write ratio 82.07%

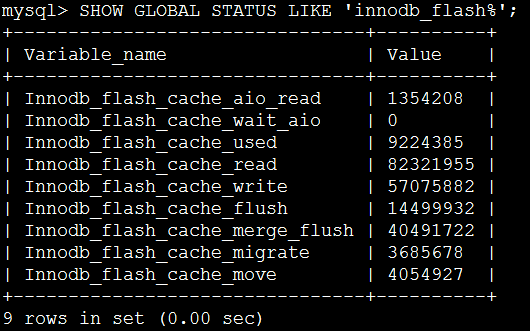
表示L2 Cache在过去2秒内的读命中率和平均命中率，merge write比例。

flash cache 1074.00 reads/s, 832.50 writes/s. 197.00 flush/s, 586.00 merge writes/s, 58.00 migrate/s, 55.00 move/s

表示L2 Cache在过去2秒内发生读、写、刷新、merge write、migrate和move的平均值。

### SHOW [GLOBAL] STATUS

当然也可以通过来获取L2 Cache部分状态信息：



Innodb\_flash\_cache\_read

通过L2 Cache读取得到页的次数，该数据可反映L2 Cache的命中率

Innodb\_flash\_cache\_write

写入L2 Cache中页的数量

Innodb\_flash\_cache\_flush

将L2 Cache中的dirty pages刷回磁盘的数量

Innodb\_flash\_cache\_merge\_flush

merge write的次数，通过该变量可得到merge write的效率

Innodb\_flash\_cache\_migrate

Clean page从InnoDB BP写入到L2 Cache的次数

Innodb\_flash\_cache\_move

L2 Cache中缓存的page，由于InnoDB BP相同（space, offset）的clean page换出而发生Cache位置更新的次数

Innodb\_flash\_cache\_used

L2 Cache中已经使用的页

Innodb\_flash\_cache\_aio\_read

通过L2 Cache进行AIO读取得到页的次数

Innodb\_flash\_cache\_wait\_aio

L2 Cache在进行写入操作时发现对应位置上的page正被读取的次数，可反映L2 Cache的繁忙程度