作业2

给定某内存系统，你的任务是确定其cache属性，包括块大小、相联度、cache大小、替换策略。已知这些cache属性值的范围是：

块大小：8, 16, 32, 64, or 128 Bytes

相联度：2-, 4-, or 8-way

Cache大小：4K or 8KB

替换策略：LRU or FIFO

在这个系统上，你唯一可以收集的统计数据是每次执行一系列内存访问后缓存命中率。以下是你观察到的情况：

序列	访问的内存地址(先->后)	命中率
1	0 16 24 25 1024 255 1100 305	2/8
2	31 65536 65537 131072 262144 8 305 1060	3/8
3	262145 65536 4	2/3

假设最开始时cache为空，3个序列连续访问（即序列1结束后不清空cache，马上开始序列2）。

推测cache的块大小、相联度、cache总大小、替换策略都是什么？并解释推断过程。

作业2

块大小：8, 16, 32, 64, or 128 Bytes
相联度：2-, 4-, or 8-way
Cache大小：4K or 8KB
替换策略：LRU or FIFO

0-> 00000
16->10000
24->11000
25->11001
1024->10000000000
255->11111111
1100->10001001100
305-> 100110001
31->11111
65536->10000000000000000
65537->10000000000000001
131072->100000000000000000
262144->1000000000000000000
8->1000
1060->10000100100
262145->10000000000000000001
4->100

序列	访问的内存地址(先->后)	命中率
1	0 16 24 25 1024 255 1100 305	2/8
2	31 65536 65537 131072 262144 8 305 1060	3/8
3	262145 65536 4	2/3

Cache Block Size： 16B
序列1中的缓存命中率是2/8。这意味着有2次命中。根据缓存块大小，我们可以推断出Block Size为16时，24,25命中，其他块大小都不是2次命中。

作业2

块大小：8, 16, 32, 64, or 128 Bytes
相联度：2-, 4-, or 8-way
Cache大小：4K or 8KB
替换策略：LRU or FIFO

0-> 00000
16->10000
24->11000
25->11001
1024->10000000000
255->11111111
1100->10001001100
305-> 100110001
31->11111
65536->10000000000000000
65537->10000000000000001
131072->100000000000000000
262144->1000000000000000000
8->1000
1060->10000100100
262145->10000000000000000001
4->100

序列	访问的内存地址(先->后)	命中率
1	0 16 24 25 1024 255 1100 305	2/8
2	31 65536 65537 131072 262144 8 305 1060	3/8
3	262145 65536 4	2/3

Associativity：2
序列2中的缓存命中率是3/8，这意味着有3次命中。
我们已经知道缓存块大小是16B，因此有4位偏移。
序列2中访问地址31会命中(和16在同一个block)，因为缓存块不会被替换。
序列2中访问地址305会命中，因为缓存块不会被替换。
序列2中访问地址65537会命中，因为缓存块不会被替换。
因此，其他所有访问应该都会失效。
序列2中访问地址65536、131072和262144会失效，因为这些地址没有属于任何之前访问过的缓存块。
地址65536、131072和262144的index和0相同 (index最多9bits)，会被映射到set 0中。
地址8未命中，表明其block被替换，因为它的缓存块映射到set 0，所以说明set 0大小必然小于4，因此关联度一定是2。

作业2

块大小：8, 16, 32, 64, or 128 Bytes
相联度：2-, 4-, or 8-way
Cache大小：4K or 8KB
替换策略：LRU or FIFO

0-> 00000
16->10000
24->11000
25->11001
1024->10000000000
255->11111111
1100->10001001100
305-> 100110001
31->11111
65536->10000000000000000
65537->10000000000000001
131072->100000000000000000
262144->1000000000000000000
8->1000
1060->10000100100
262145->10000000000000000001
4->100

序列	访问的内存地址(先->后)	命中率
1	0 16 24 25 1024 255 1100 305	2/8
2	31 65536 65537 131072 262144 8 305 1060	3/8
3	262145 65536 4	2/3

Cache Size：无法判断

作业2

块大小：8, 16, 32, 64, or 128 Bytes
相联度：2-, 4-, or 8-way
Cache大小：4K or 8KB
替换策略：LRU or FIFO

0-> 00000
16->10000
24->11000
25->11001
1024->10000000000
255->11111111
1100->10001001100
305-> 100110001
31->11111
65536->10000000000000000
65537->10000000000000001
131072->100000000000000000
262144->1000000000000000000
8->1000
1060->10000100100
262145->10000000000000001
4->100

序列	访问的内存地址(先->后)	命中率
1	0 16 24 25 1024 255 1100 305	2/8
2	31 65536 65537 131072 262144 8 305 1060	3/8
3	262145 65536 4	2/3

替换策略：FIFO

- 缓存块大小是16B
- 缓存是2路关联的

序列3中的缓存命中率是2/3，这意味着有2次命中。
使用LRU策略时，只有序列3中访问地址262145会命中。使用FIFO策略时，序列3中访问地址262145和4会命中。
因此，缓存采用了FIFO策略。

作业3

某指令集支持8-bit虚拟内存地址，物理内存一共128 bytes，每个物理页16 bytes。页表使用一级结构，全部放在内存中。初始的时候内存布局如下：

物理页	存储内容
0	Empty
1	Virtual Page 13
2	Virtual Page 5
3	Virtual Page 2
4	Empty
5	Virtual Page 0
6	Empty
7	Page Table

采用容量为3个entry的TLB对地址转换结构进行缓存，TLB采用LRU替换策略（按照物理页访问的热度实现LRU）。初始时，TLB的内容为virtual page 0, 2和13对应的物理页号。对于下面的访问序列（virtual pages）：

0，13，5，2，14，14，13，6，6，13，15，14，15，13，4，3

（a）TLB的命中率为多少？

（b）全部访问结束后，TBL里面的内容是什么？

（c）全部访问结束后，页表的内容是什么？

作业3

某指令集支持8-bit虚拟内存地址，物理内存一共128 bytes，每个物理页16 bytes。页表使用一级结构，全部放在内存中。初始的时候内存布局如下：

物理页	存储内容
0	Empty
1	Virtual Page 13
2	Virtual Page 5
3	Virtual Page 2
4	Empty
5	Virtual Page 0
6	Empty
7	Page Table

采用容量为3个entry的TLB对地址转换结构进行缓存，TLB采用LRU替换策略（按照物理页访问的热度实现LRU）。初始时，TLB的内容为virtual page 0, 2和13对应的物理页号。对于下面的访问序列（virtual pages）：

0，13，5，2，14，14，13，6，6，13，15，14，15，13，4，3

（a）TLB的命中率为多少?

0(hit)，13(hit)，5(miss), 2 (miss), 14(miss, 分配page 0), 14(hit), 13(miss), 6(miss, 分配page 4), 6(hit), 13(hit), 15(miss, 分配page 6), 14(miss), 15(hit), 13(hit), 4(miss，分配page 5), 3(miss, 分配 page 2)

所以命中率为7/16

作业3

某指令集支持8-bit虚拟内存地址，物理内存一共128 bytes，每个物理页16 bytes。页表使用一级结构，全部放在内存中。初始的时候内存布局如下：

物理页	存储内容
0	Empty
1	Virtual Page 13
2	Virtual Page 5
3	Virtual Page 2
4	Empty
5	Virtual Page 0
6	Empty
7	Page Table

采用容量为3个entry的TLB对地址转换结构进行缓存，TLB采用LRU替换策略（按照物理页访问的热度实现LRU）。初始时，TLB的内容为virtual page 0, 2和13对应的物理页号。对于下面的访问序列（virtual pages）：

0，13，5，2，14，14，13，6，6，13，15，14，15，13，4，3

（b）全部访问结束后，TBL里面的内容是什么？

4, 13, 3

作业3

某指令集支持8-bit虚拟内存地址，物理内存一共128 bytes，每个物理页16 bytes。页表使用一级结构，全部放在内存中。初始的时候内存布局如下：

物理页	存储内容
0	Empty
1	Virtual Page 13
2	Virtual Page 5
3	Virtual Page 2
4	Empty
5	Virtual Page 0
6	Empty
7	Page Table

采用容量为3个entry的TLB对地址转换结构进行缓存，TLB采用LRU替换策略（按照物理页访问的热度实现LRU）。初始时，TLB的内容为virtual page 0, 2和13对应的物理页号。对于下面的访问序列（virtual pages）：

0，13，5，2，14，14，13，6，6，13，15，14，15，13，4，3

（c）全部访问结束后，页表的内容是什么？

物理页0-7对应的虚拟页分别是 14, 13, 3, 2, 6, 4, 15, Page table

作业5

假设某处理器有一个2-bits 的Global History Register（GHR），由所有的分支语句共享，其初始值为00（表示Not Taken）。每个Pattern History Table Entry（PHTE）包括一个2-bits的饱和计数器，其含义如下：

00 - Strongly Not Taken

01 - Weakly Not Taken

10 - Weakly Taken

11 - Strongly Taken

假设下面的代码运行在该处理器上。该代码含有两个分支语句（B1和B2）。

```
for (int i = 0; i < 1000000; i++) { /* B1 */
                                    /* TAKEN PATH for B1 */
    if (i % 3 == 0) {               /* B2 */
        j[i] = k[i] -1;             /* TAKEN PATH for B2 */
    }
}
```

作业5

```
for (int i = 0; i < 1000000; i++) { /* B1 */
                                    /* TAKEN PATH for B1 */
    if (i % 3 == 0) {               /* B2 */
        j[i] = k[i] -1;             /* TAKEN PATH for B2 */
    }
}
```

（a）有没有可能发生前5次循环所有分支预测全部错误？如果可能，列出每个PHTE可能的初始值（Not Taken用N表示，Taken用T表示）。

PHT Entry	Value
TT	01
TN	00
NT	01
NN	00或01

可能发生，前5次循环分支实际跳转的情况是：TT TN TN TT TN
给上面的跳转情况填加编号 T1 T2 T3 N4 T5 N6 T7 T8 T9 N10
对于GHR=NN，只在初始的时候为NN，其余都不是，因此，只有T1观察到了NN，所以，NN对应的PHTE初始值只能是00或者01，这样才会将T1预测为N；
对于GHR=TT，T3, N4, T9和N10都观察到了该pattern。如果将TT对应的PHTE设置为01，那么可以保证T3，N4，T9和N10都预测错误；
对于GHR=TN，T5和T7观测到了该pattern。如果要保证T5和T7都是预测错误，那么TN对应的PHTE必须为00；
对于GHR=NT，T2，N6和T8观测到了该pattern。为了保证这几个都预测错误，NT对应的PHTE为01即可；

作业5

```
for (int i = 0; i < 1000000; i++) { /* B1 */
                                    /* TAKEN PATH for B1 */
    if (i % 3 == 0) {               /* B2 */
        j[i] = k[i] -1;             /* TAKEN PATH for B2 */
    }
}
```

（b）当系统达到稳定状态之后（很多次循环之后），该分支预测器的准确率是否可以达到100%？如果可以达到，对PHT的初始值的设置是否有特殊要求？

PHT Entry	Value
TT	01
TN	00
NT	01
NN	00或01

稳态之后分支的真实模式为TTTNTN，可以推算，无论如何设置PHT的初始值，都不可能达到100%准确率。
TTTN连续2次出现，如果前面1次预测正确，后面一次必然也预测成T，不会预测成N！

作业8

GPU的利用率通常定义为处于busy状态的GPU 核心(PE，单个计算单元)占所有GPU核心的比例。考虑下面的代码片段。每个thread执行循环中的1次迭代（包含6条指令）。假设数组A、B、C已经在寄存器中（不需要从内存读入）。该GPU的一个warp包含64个threads，该GPU包含64个核心。假设数组B的每个元素的绝对值都小于10。

```
for (i = 0; i < 1024; i++) {
  A[i] = B[i] * B[i];
  if (A[i] > 0) {
    C[i] = A[i] * B[i];
    if (C[i] < 0) {
      A[i] = A[i] + 1;
    }
    A[i] = A[i] - 2; }
}
```

1.执行该代码段需要多少个warps?

2.执行整个代码段，最大的GPU利用率可能是多少?

3.获得最大的GPU利用率时，数组B的值有何特征?

4.执行整个代码段，最小的GPU利用率可能是多少?

作业8

GPU的利用率通常定义为处于busy状态的GPU 核心(PE，单个计算单元)占所有GPU核心的比例。考虑下面的代码片段。每个thread执行循环中的1次迭代（包含6条指令）。假设数组A、B、C已经在寄存器中（不需要从内存读入）。该GPU的一个warp包含64个threads，该GPU包含64个核心。假设数组B的每个元素的绝对值都小于10。

```
for (i = 0; i < 1024; i++) {
  A[i] = B[i] * B[i];
  if (A[i] > 0) {
    C[i] = A[i] * B[i];
    if (C[i] < 0) {
      A[i] = A[i] + 1;
    }
    A[i] = A[i] - 2; }
}
```

1.执行该代码段需要多少个warps?

1024/64 = 16

作业8

GPU的利用率通常定义为处于busy状态的GPU 核心(PE，单个计算单元)占所有GPU核心的比例。考虑下面的代码片段。每个thread执行循环中的1次迭代（包含6条指令）。假设数组A、B、C已经在寄存器中（不需要从内存读入）。该GPU的一个warp包含64个threads，该GPU包含64个核心。假设数组B的每个元素的绝对值都小于10。

```
for (i = 0; i < 1024; i++) {
  A[i] = B[i] * B[i];
  if (A[i] > 0) {
    C[i] = A[i] * B[i];
    if (C[i] < 0) {
      A[i] = A[i] + 1;
    }
    A[i] = A[i] - 2; }
}
```

2.执行整个代码段，最大的GPU利用率可能是多少?

100%

作业8

GPU的利用率通常定义为处于busy状态的GPU 核心(PE，单个计算单元)占所有GPU核心的比例。考虑下面的代码片段。每个thread执行循环中的1次迭代（包含6条指令）。假设数组A、B、C已经在寄存器中（不需要从内存读入）。该GPU的一个warp包含64个threads，该GPU包含64个核心。假设数组B的每个元素的绝对值都小于10。

```
for  (i  =  0;  i  <  1024;  i++)  {
  A[i]  =  B[i]  *  B[i];
  if  (A[i]  >  0)  {
    C[i]  =  A[i] *  B[i];
    if  (C[i]  <  0)  {
      A[i]  =  A[i]  +  1;
    }
    A[i]  =  A[i]  -  2; }
}
```

3.获得最大的GPU利用率时，数组B的值有何特征?

对于每64个连续值，或者都是0，或者都是正数，或者都是负数。

作业8

GPU的利用率通常定义为处于busy状态的GPU 核心(PE，单个计算单元)占所有GPU核心的比例。考虑下面的代码片段。每个thread执行循环中的1次迭代（包含6条指令）。假设数组A、B、C已经在寄存器中（不需要从内存读入）。该GPU的一个warp包含64个threads，该GPU包含64个核心。假设数组B的每个元素的绝对值都小于10。

```
for (i = 0; i < 1024; i++) {
  A[i] = B[i] * B[i];
  if (A[i] > 0) {
    C[i] = A[i] * B[i];
    if (C[i] < 0) {
      A[i] = A[i] + 1;
    }
    A[i] = A[i] - 2; }
}
```

4.执行整个代码段，最小的GPU利用率可能是多少?

$(64 \times 2 + 1 \times 4)/(64 \times 6) = 132/384$

作业8

GPU的利用率通常定义为处于busy状态的GPU 核心(PE，单个计算单元)占所有GPU核心的比例。考虑下面的代码片段。每个thread执行循环中的1次迭代（包含6条指令）。假设数组A、B、C已经在寄存器中（不需要从内存读入）。该GPU的一个warp包含64个threads，该GPU包含64个核心。假设数组B的每个元素的绝对值都小于10。

```
for  (i  =  0;  i  <  1024;  i++)  {
  A[i]  =  B[i]  *  B[i];
  if  (A[i]  >  0)  {
    C[i]  =  A[i] *  B[i];
    if  (C[i]  <  0)  {
      A[i]  =  A[i]  +  1;
    }
    A[i]  =  A[i]  -  2; }
}
```

4.执行整个代码段，最小的GPU利用率可能是多少?

$(64 \times 2 + 1 \times 4)/(64 \times 6) = 132/384$

仅有1个thread通过了第1个分支语句，其他threads都没通过，因此，只有一个PE busy，其余都是空闲的。也就是，每连续64个值，有1个是负数，其余都是0。