**作业（高速缓存、虚拟内存）**

1. **直接映射：**

为了将内存数据块放置到高速缓存中，可以将内存地址分几部分看待：



Offset bits是块内偏移， Index bits用于标记该内存数据块在高速缓存中的组号（set number）, 剩下的tag bits是该数据块的标记。

如果高速缓存采用直接映射（direct mapped）采用写直达（write through）的更新策略，那么高速缓存中每一行包含的内容为：cache data block, tag, valid bit， 不需要包含dirty bit.

**根据已知信息，请填写下表中的空格部分：**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Address size(bits)内存地址的长度，寻址到byte | Cache size | Block Size  数据块大小 | tag bits  tag 位数 | Index bits  Index 位数 | Offset bits  块内偏移位数 | Bits per row  每一行的总位数 |
| 16 | 4KB | 4B | 4 | 10 | 2 | 32+4+1 |
| 32 | 32KB | 16B | 17 | 11 | 4 | 146 |
| 32 | 64KB | 16B | 16 | 12 | 4 | 145 |
| 64 | 2048KB | 128B | 43 | 14 | 7 | 1068 |

**2、组相联映射：**

假设某计算机的主存地址空间大小为64MB，采用字节编址方式。其cache数据区容量为4KB，采用4路组相联映射方式、LRU替换和回写（write back）策略，块大小为64B。请问：

1. **主存地址字段如何划分？要求说明每个字段的含义、位数和在主存地址中的位置。**

块大小为64B，说明块内偏移量为6位。

Cache的大小为4KB，而块大小为64B，故Cache中有4KB/64B=64个块。由于采取4路组相联映射方式，可知每组中有16个块，也就是说组号需要4位。

主存地址空间大小为64MB，也就是26位。低6位是块内偏移量，中间4位是组号，则高16位为Tag位。

1. **该cache的总容量（不仅包括数据区容量）有多少位？**

Cache的每一行包括16位Tag位，1位有效/无效位，2位LRU位，1位dirty位和共64B的数据位，而一共有64行，因此一共有64×(16+1+2+1+64×8)=34048位=4256B。

**3、代码分析与高速缓存的性能：**

一个二路组相联映射的高速缓存（ 2-way associative cache ）容量为128 bytes ，每个高速缓存块大小为32 字节（32 bytes per block）. long long型数据的长度为 8个字节（8 bytes）.假定table数组的起始地址为 0x0. **以下代码的高速缓存失效率(miss rate)为多少？**

**int i;**

**int j;**

**long long table[4][8];**

**for (j = 0; j < 8; j++)**

**for (i = 0; i < 4; i++)**

**{ table[i][j] = i + j; }**

块大小为32字节，说明主存地址的低5位是块内偏移位；cache大小为128字节，说明cache内一共包含了4个块，而使用二路组相联映射，说明一组里有两个块，组号有一位。

由于table数组为long long型，一个数据的长度为8个字节，故4个数据正好填满低5位，8个数据正好填满低6位。也就是说，table[i][j]和table[i+1][j]的低六位是完全相同的。因此，访问table[0][0]时，第一次cache miss并将table[0][0-7]放入cache的第0组第0块；访问table[1][0]时，第二次cache miss，由于组号完全相同，因此将table[0][0-7]放入cache的第0组第1块；而访问table[2][0]时，第三次cache miss，由于组号完全相同，且该组所有的块都已经满了，所以将table[2][0-7]放入cache的第0组第0块而替换掉table[0][0-7]；访问table[3][0]时，第三次cache miss，由于组号完全相同，且该组所有的块都已经满了，所以将table[3][0-7]放入cache的第0组第0块而替换掉table[1][0-7]。

之后访问table[0][1]，之前虽然table[0][0-7]被放入过cache，但是它被table[2][0-7]替换过了，因此仍然是cache miss，再一次用table[0][0-7]放入cache的第0组第0块而替换掉table[2][0-7]；依次进行下去，始终发生cache miss。直到访问到table[0][4]，组号变为1，但是其实还是一样的情况，由于之后访问的数据组号也都是1，所以之后仍然不可能命中。

因此cache miss rate 是100%。

**4、平均存储器访问时间（Average Memory Access Time：AMAT）**

AMAT 是内存访问的平均（expected）时间 , 可以用以下公式来估算：

AMAT = hit\_time + miss\_rate × miss\_penalty

* Hit\_time: cache hit 时，访问cache 所花的时间
* Miss\_rate: 高速缓存的失效率
* miss penalty：当发生cache miss时，需要花的额外的访存时间 ，所以一次 cache miss 需要花费 (hit\_time + miss\_penalty) 的时间

**假设高速缓存系统的属性如下，求AMAT是多少？**

a) L1$ hits in 1 cycle (local miss rate 25%)

b) L2$ hits in 10 cycles (local miss rate 40%)

c) L3$ hits in 50 cycles (global miss rate 6%)

d) Main memory hits in 100 cycles (always hits)

Global miss rate和 Local miss rate 的定义请参考如下描述：

Global miss rate：

– the fraction of references that miss some level of a multilevel cache

– misses in this cache divided by the total number of memory accesses generated by the CPU

Local miss rate – the fraction of references to one level of a cache that miss

Local Miss rate L2$ = L2$ Misses / L1$ Misses

= L2$ Misses / total\_L2\_accesses • L2$ local miss rate

即75%的概率L1$ hits；剩下25% L1$ misses时，有40%的概率L2$ misses，也就是说25%×40%=10%的概率L2$ hits；L3$的global miss rate 6%，说明在剩下的15%中，9%的概率L3$ hits，6%的概率main memory hits。

所以AMAT=75%×1 cycle + 11 cycles×10% + 61 cycles×9% + 161 cycles×6% = 17 cycles

**5、虚拟存储器（Virtual Memory**）

程序中使用的内存地址是虚拟地址，一个虚拟地址（VA）可以看作两部分：虚页号、页内偏移（page offset），如下图中标示：

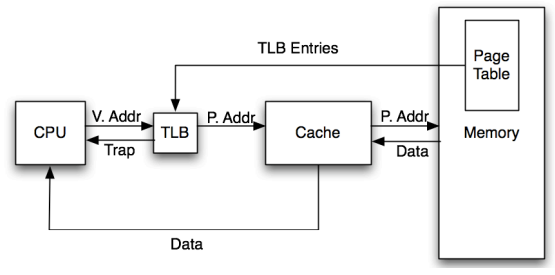


但事实上无论是数据还是指令都是存放在物理内存上的。一个物理地址（PA）也可以看作两部分：物理页号、页内偏移。如下图中标示：



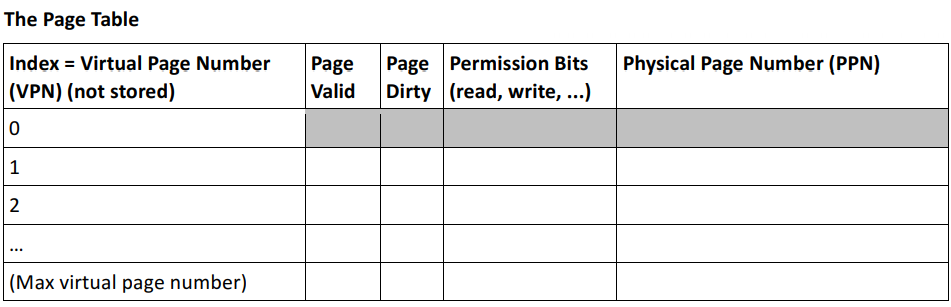
如果页大小是4KB=4096 bytes，那么page offset 就是12位。

将 VA 转换为 PA 会使用到快表（TLB） 和页表（Page Table）. 下图示意了TLB和页表（ｐ在内存访问时所处的位置。



每一个进程都有一个页表，页表存储在内存中，操作系统通过设置一个专用寄存器的值，告诉硬件页表在内存中的起始地址。每当切换进程时，操作系统会将要执行的进程的页表起始地址转载到这个专用寄存器中。

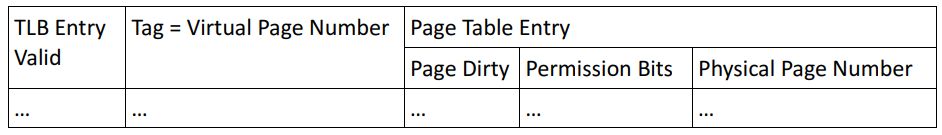
页表的结构一般如下：



每一个页表项（ page table entry）除了记录虚页号（VPN）和物理页号（PPN）的映射关系之外，还设置了有效位、脏位和权限位：

* “page valid” 有效位：用于标记该虚页是否在内存中；
* “page dirty” 脏位：操作系统需要知道，是否将该内存页更新到磁盘上；
* “permission bits”权限位：用于限制对该页进行某种操作。

快表（TLB）是页表的缓存（cache）, 假设TLB如果采用全相联映射（fully associative）机制，它的结构如下：



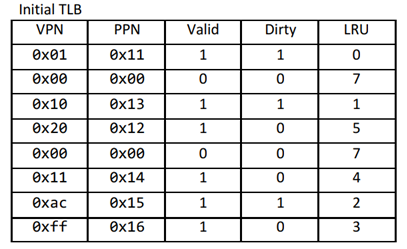
回答问题：

1. **如果页表地址寄存器中装入了新的值，TLB会发生什么操作？**

将TLB中的所有值全部置零

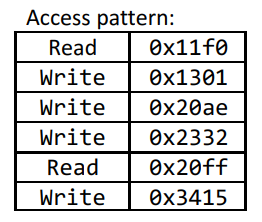
1. 某个处理器内存地址长度为16位，页大小为256 byte，TLB采用全相联映射，总共有8个TLB 表项， 并采用LRU替换机制（LRU位有3位，可以表示8个TLB表项的访问次序。如果LRU位的值为0，表示该页最近刚刚被访问）。

假设当前进程初始时TLB的内容如下，并假设该进程访问的所有页既可以读也可以写。



假设现在空闲的物理页是：0x17, 0x18, 0x19；

如果接下来，用标记出的访问模式（读或者写）对以下内存地址进行访问：



**请画出完成以上访问后，TLB的final state。**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| VPN | PPN | Valid | Dirty | LRU |
| 0x01 | 0x11 | 1 | 1 | 5 |
| 0x13 | 0x17 | 1 | 1 | 3 |
| 0x10 | 0x13 | 1 | 1 | 6 |
| 0x20 | 0x12 | 1 | 1 | 1 |
| 0x23 | 0x18 | 1 | 1 | 2 |
| 0x11 | 0x14 | 1 | 0 | 4 |
| 0xac | 0x15 | 1 | 1 | 7 |
| 0x34 | 0x19 | 1 | 1 | 0 |