本次作业采用和上次一样的评分标准，根据批改结果来看，同学们的掌握情况还不错，有不少满分的同学。本答案仅供参考，如果你认为某些题目的答案不是这样，或者你认为有更好的解决方案，欢迎随时联系助教。

Cache 设计（1）

假设有一个“Cache-主存“层次存储系统。Cache 为 8 块，主存为 64 块，试分别对于以下三种情况，计算访存块地址为 55 时对应的标识（tag），如果存在索引（index）请计算索引：

（1）全相连

（2）2 路组相连

（3）4 路组相连

（4）直接映射

主要问题：

1. 有大量同学忘记了计算 Tag

2. 有一些同学认为四种方案的 Tag 相同，但这一点显然是不正确的，因为如果四种方案的 Tag 相同，那么全相连 Cache 和直接映射的 Cache 的硬件开销就一样了（因为需要比较的位数相同）。

3. 有些同学不太清楚标记（tag）的定义，请在课件上复习相关的内容。

4. 本题中直接讨论的就是块地址，所以地址中没有 offset。

5. 本题的同学们的表示方法五花八门（有二进制，有十进制，有十六进制）由于题目中并没有表示方法的要求，因此每一种表示都被认为是正确的。不过相对而言在计算机系统设计问题中使用二进制和十六进制更为规范一些（可能有同学会问为什么题面中使用十进制表示 55，因为这里的 55 是具有实际含义的，即从 0 号块开始数的第 56 块内存，可能有一些同学被这个题面中的这种表述形式所误导了）

解：

55 对应的二进制为$(110111)_2$

连接类型	标识	索引
全相连	$(110111)_2$	无
2 路组相连	$(1101)_2$	$(11)_2$
4 路组相连	$(11011)_2$	$(1)_2$
直接映射	$(110)_2$	$(111)_2$

Cache 设计（2）

假设一台计算机具有如下特性

1. 97%的访存都在 Cache 中命中

2. 块大小为两个字，且不命中时整个块都被调入

3. CPU 发出访存请求的速率为10^9字/秒

4. 20%的存储访问为写访问

5. 存储器的最大流量为10^9字/秒（包括读和写）

6. 主存每一次只能读或者写一个字

7. 在写回法中，Cache 块中的脏块（被修改过的块）比率保持为 25%

现欲给该计算机增加一台外设，为此需要首先知道主存的带宽已经使用了多少。试对于以下情况分别分析主存带宽的平均使用比例

（1）写直达 Cache+写分配策略

（2）写直达 Cache+写不分配策略

（3）写回法 Cache+写分配策略

（4）写回法 Cache+写不分配策略

主要问题：

本题有一定的难度，首先需要同学将带宽占用问题抽象为一个访存次数计算的问题，接下来需要对于各类写策略有较准确的理解，同时还应考虑到各种可能的情况。答案中针对每一种情况给出了详细的分析，本题出错的同学请仔细阅读一下，有不理解的或者你认为不正确的地方欢迎随时给助教发邮件。

解：

我们将针对内存中一个字的读写定义为一次访存

1. 写直达+写分配

Cache 访问命中，有两种情况，可能是读命中，此时不需要访问主存，也可能是写命中，直接更新 Cache 和主存，访存 1 次。

Cache 访问失效，也可分为两种情况，可能是读失效，此时需要将主存中的块调入 Cache 中，需要访问主存储器 2 次（因为需要将整个块都调入到 Cache 中对应需要对两个字进行读操作），如果是写失效，那么需要一方面将失效的块调入内存中，对应访存 2 次，同时需要将结果写入到内存中，对应访存 1 次，所以合计 3 次访存。

由此我们有如下表格：

访问命中	访问类型	频率	访存次数
Y	读	97%*80%	0
Y	写	97%*20%	1
N	读	3%*80%	2
N	写	3%*20%	3

由此我们可以计算出一次存储访问实际对应的平均主存储器访问次数为：

$$0.97 * 0.8 * 0 + 0.97 * 0.2 * 1 + 0.03 * 0.8 * 2 + 0.03 * 0.2 * 3 = 0.26$$

由此我们得到主存储器带宽占比为：

$$0.26 * \frac{10^9}{10^9} * 100\% = 26\%$$

2. 写直达+写不分配

Cache 访问命中，有两种情况，可能是读命中，此时不需要访问主存，也可能是写命中，直接更新 Cache 和主存，访存 1 次。

Cache 访问失效，也可分为两种情况，可能是读失效，此时需要将主存中的块调入 Cache 中，需要访问主存储器 2 次，如果是写失效，由于不需要分配所以只需将结果写入到内存中，对应访存 1 次。

由此我们有如下表格：

访问命中	访问类型	频率	访存次数
Y	读	97%*80%	0
Y	写	97%*20%	1
N	读	3%*80%	2
N	写	3%*20%	1

由此我们可以计算出一次存储访问对应的平均主存储器访问次数为：

$$0.97 * 0.8 * 0 + 0.97 * 0.2 * 1 + 0.03 * 0.8 * 2 + 0.03 * 0.2 * 1 = 0.248$$

由此我们得到主存储器带宽占比为：

$$0.248 * \frac{10^9}{10^9} * 100\% = 24.8\%$$

3. 写回法+写分配

Cache 访问命中，有两种情况，可能是读命中，此时不需要访问主存，也可能是写命中，直接更新 Cache，对应访存次数为 0

Cache 访问失效时，由于是采用写分配的方法，因此无论读写请求都需要换出一个块，这样针对这个块是否是脏块（被修改过的块）也可分为两种情况，如果不是脏块，那么只需直接将内存中的块调入 Cache，对应的访存为 2 次，如果是脏块，那么首先需要将脏块写回主存，对应访存 2 次，然后再将需要的块调入 Cache，对应访存 2 次，合计访存 4 次

由此我们有如下表格：

访问命中	访问类型	是否是脏块	频率	访存次数
Y	读	/	97%*80%	0
Y	写	/	97%*20%	0
N	读/写	N	3%*75%	2
N	读/写	Y	3%*25%	4

由此我们可以计算出一次存储访问对应的平均主存储器访问次数为：

$$0.97 * 0.8 * 0 + 0.97 * 0.2 * 0 + 0.03 * 0.75 * 2 + 0.03 * 0.25 * 4 = 0.075$$

由此我们得到主存储器带宽占比为：

$$0.075 * \frac{10^9}{10^9} * 100\% = 7.5\%$$

4. 写回法+写不分配

Cache 访问命中，有两种情况，可能是读命中，此时不需要访问主存，也可能是写命中，直接更新 Cache，对应访存次数为 0。

Cache 访问失效时，如果是读访问失效，那么分析结果类似情况 3，如果是写访问失效，那么就没有将块调入 Cache 的需要，因此也不会有块换出，只需要修改主存储器即可，对应的访存次数为 1。

由此我们有如下表格：

访问命中	访问类型	是否是脏块	频率	访存次数
Y	读	/	97%*80%	0
Y	写	/	97%*20%	0
N	读	N	3%*75%*80%	2
N	读	Y	3%*25%*80%	4
N	写	N/Y	3%*20%	1

由此我们可以计算出一次存储访问对应的平均主存储器访问次数为：

$$0.97 * 0.8 * 0 + 0.97 * 0.2 * 0 + 0.03 * 0.75 * 0.8 * 2 + 0.03 * 0.25 * 0.8 * 4 + 0.03 * 0.2 * 1 = 0.066$$

由此我们得到主存储器带宽占比为：

$$0.066 * \frac{10^9}{10^9} * 100\% = 6.6\%$$

Cache 性能分析

假设一段计算机程序的存储访问中有 80%是指令访问，有 20%是数据访问（假设均为读访问），指令或数据 Cache 的命中时间为 1 个时钟周期，混合 Cache 的命中时间为 2 个时钟周期，不命中开销都为 50 个时钟周期。32KB 的指令 Cache 的不命中率为 0.39%，32KB 的数据 Cache 的不命中率为 4.82%，64KB 的混合 Cache 的不命中率为 1.35%。

试问指令 Cache 和数据 Cache 容量均为 32KB 的分离 Cache 和容量为 64KB 的混合 Cache 相比，哪一种 Cache 的不命中率更低？两种情况下的平均访存时间各是多少？

主要问题：

1. 本题题面描述中的不命中开销一词存在歧义，有同学将其理解为缺失损失，有同学将其理解为缺失时的访存时间（此时对应的缺失损失分别为 49 和 48），但相对来说将其理解为缺失损失较为合理，因为主存储器的访问延迟不应该受到 Cache 设计的影响，所以答案中的缺失损失就都按照 50 来计算。而按照另一种情况计算得到的结果为分离 Cache 和混合 Cache 的平均访存时间分别为 1.62524 和 2.648，判作业时这个答案也被视作是正确的。

2. 有同学在将混合 Cache 的指令命中时间视作了 1 个周期，这在设计上是可能的（因为按照 MIPS 五段流水 CPU 只有在数据访问和指令访问冲突时才需要插入一个气泡），但是和题意并不相符合，因此并没有被视作正确。

解：

分离 Cache 的不命中率如下

$$20\% * 4.82\% + 80\% * 0.39\% = 1.276\%$$

混合 Cache 的不命中率为 1.35%，因此分离 Cache 的不命中率更低

平均访存时间为：

$$\mathrm{T} = C_{Hit} + C_{miss} * R_{miss}$$

其中C_{Hit}为命中时间，C_{miss}为不命中时间，R_{miss}为不命中率，由此得到：

分离 Cache 对应的平均访存时间为

$$\mathrm{T} = 1 + 50 * 1.276\% = 1.638$$

混合 Cache 对应的平均访存时间为

$$\mathrm{T} = 2 + 50 * 1.35\% = 2.675$$

多级 Cache

某个应用运行在三级 cache 结构的系统上，一级 cache 的访问时间为 1，二级 cache、三级 Cache、memory 的访问时间分别为 10，20，100。该应用共发出 1000 次访存操作，一级 cache 缺失 50 次、二级 cache 缺失 20 次，三级 cache 缺失 5 次，则每一级 Cache 的全局缺失率和局部缺失率分别为多少？平均存储器访问时间为多少？

本题只要理解了全局缺失率和局部缺失率这两个概念的含义就基本不会出错。

解：

局部缺失率=该级 Cache 的缺失次数/到达该级 Cache 的总访问次数

全局缺失率=该级 Cache 的缺失次数/CPU 发出的总访存次数

由此得到：

一级 Cache 缺失率为（R_l 表示局部缺失率，R_g表示全局缺失率）：

$$R_l = \frac{50}{1000} = 5\%$$

$$R_g = \frac{50}{1000} = 5\%$$

二级 Cache 缺失率为（R_l 表示局部缺失率，R_g表示全局缺失率）：

$$R_l = \frac{20}{50} = 40\%$$

$$R_g = \frac{20}{1000} = 2\%$$

三级 Cache 缺失率为（R_l 表示局部缺失率，R_g表示全局缺失率）：

$$R_l = \frac{5}{20} = 25\%$$

$$R_g = \frac{5}{1000} = 0.5\%$$

平均访存时间为：

$$\mathrm{T} = 1 + 5\% * 10 + 2\% * 20 + 0.5\% * 100 = 2.4$$