一次 load 或 store 操作访问 Cache 的命中时间都要增加一个时钟周期，32KB 的指令 Cache 的失效率为 0.39%，32KB 的数据 Cache 的失效率为 4.82%，64KB 的混合 Cache 的失效率为 1.35%。又假设采用写直达策略，且有一个写缓冲器，并且忽略写缓冲器引起的等待。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，哪种 Cache 的失效率更低？两种情况下平均访存时间各是多少？

解：（1）根据题意，约 75%的访存为取指令。

因此，分离 Cache 的总体失效率为：（75%×0.15%）＋（25%×3.77%）＝1.055%；

容量为 128KB 的混合 Cache 的失效率略低一些，只有 0.95%。

（2）平均访存时间公式可以分为指令访问和数据访问两部分：

平均访存时间＝指令所占的百分比×（读命中时间＋读失效率×失效开销）＋数据所占的百分比×（数据命中时间＋数据失效率×失效开销）

所以，两种结构的平均访存时间分别为：

分离 Cache 的平均访存时间＝75%×（1＋0.15%×50）＋25%×（1＋3.77%×50）

＝（75%×1.075）＋（25%×2.885）＝1.5275

混合 Cache 的平均访存时间＝75%×（1＋0.95%×50）＋25%×（1＋1＋0.95%×50）

＝（75%×1.475）＋（25%×2.475）＝1.725

因此，尽管分离 Cache 的实际失效率比混合 Cache 的高，但其平均访存时间反而较低。分离 Cache 提供了两个端口，消除了结构相关。

5.11 给定以下的假设，试计算直接映象 Cache 和两路组相联 Cache 的平均访问时间以及 CPU 的性能。由计算结果能得出什么结论？

（1）理想 Cache 情况下的 CPI 为 2.0，时钟周期为 2ns，平均每条指令访存 1.2 次；

（2）两者 Cache 容量均为 64KB，块大小都是 32 字节；

（3）组相联 Cache 中的多路选择器使 CPU 的时钟周期增加了 10%；

（4）这两种 Cache 的失效开销都是 80ns；

（5）命中时间为 1 个时钟周期；

（6）64KB 直接映象 Cache 的失效率为 1.4%，64KB 两路组相联 Cache 的失效率为 1.0%。

解：　平均访问时间＝命中时间＋失效率×失效开销

平均访问时间 $_{1\text{-路}}$=2.0+1.4% *80=3.12ns

平均访问时间 $_{2\text{-路}}$=2.0*(1+10%)+1.0% *80=3.0ns

两路组相联的平均访问时间比较低

CPU_{time}=（CPU $_{执行}$+存储等待周期）*时钟周期

CPU $_{time}$=IC（CPI $_{执行}$+总失效次数/指令总数*失效开销）　*时钟周期

=IC（（CPI $_{执行}$*时钟周期）+（每条指令的访存次数*失效率*失效开销*时钟周期））

CPU $_{time\ 1\text{-}way}$=IC(2.0*2+1.2*0.014*80)＝5.344IC

CPU $_{time\ 2\text{-}way}$=IC(2.2*2+1.2*0.01*80)＝5.36IC

相对性能比：$\dfrac{CPU_{time-2way}}{CPU_{time-1way}}=5.36/5.344=1.003$

直接映象 cache 的访问速度比两路组相联 cache 要快 1.04 倍，而两路组相联 Cache 的平均性能比直接映象 cache 要高 1.003 倍。因此这里选择两路组相联。

5.12 假设一台计算机具有以下特性：

（1）95%的访存在 Cache 中命中；
（2）块大小为两个字，且失效时整个块被调入；
（3）CPU 发出访存请求的速率为 10^9 字/s；
（4）25%的访存为写访问；
（5）存储器的最大流量为 10^9 字/s（包括读和写）；
（6）主存每次只能读或写一个字；
（7）在任何时候，Cache 中有 30%的块被修改过；
（8）写失效时，Cache 采用按写分配法。

现欲给该计算机增添一台外设，为此首先想知道主存的频带已用了多少。试对于以下两种情况计算主存频带的平均使用比例。

（1）写直达 Cache；
（2）写回法 Cache。

解：采用按写分配

（1）写直达 cache 访问命中，有两种情况：

读命中，不访问主存；

写命中，更新 cache 和主存，访问主存一次。

访问失效，有两种情况：

读失效，将主存中的块调入 cache 中，访问主存两次；

写失效，将要写的块调入 cache，访问主存两次，再将修改的数据写入 cache 和主存，访问主存一次，共三次。上述分析如下表所示。

访问命中	访问类型	频率	访存次数
Y	读	95%*75%=71.3%	0
Y	写	95%*25%=23.8%	1
N	读	5%*75%=3.8%	2
N	写	5%*25%=1.3%	3

一次访存请求最后真正的平均访存次数=(71.3%*0)+(23.8%*1)+(3.8%*2)+(1.3%*3)＝0.35

已用带宽=$0.35\times10^9/10^9$=35.0%

（2）写回法 cache 访问命中,有两种情况：

读命中，不访问主存；

写命中，不访问主存。采用写回法，只有当修改的 cache 块被换出时，才写入主存；

访问失效,有一个块将被换出，这也有两种情况：

如果被替换的块没有修改过，将主存中的块调入 cache 块中，访问主存两次；

如果被替换的块修改过，则首先将修改的块写入主存，需要访问主存两次；然后将主存中的块调入 cache 块中，需要访问主存两次，共四次访问主存。

访问命中	块为脏	频率	访存次数
Y	N	95%*70%=66.5%	0
Y	Y	95%*30%=28.5%	0
N	N	5%*70%=3.5%	2
N	Y	5%*30%=1.5%	4

所以：

一次访存请求最后真正的平均访存次数=66.5%*0＋28.5%*0+3.5%*2+1.5%*4=0.13

已用带宽＝$0.13\times10^9/10^9$＝13%

5.13 在伪相联中，假设在直接映象位置没有发现匹配，而在另一个位置才找到数据（伪命中）时，不对这两个位置的数据进行交换。这时只需要 1 个额外的周期。假设失效开销为 50 个时钟周期，2KB 直接映象 Cache 的失效率为 9.8%，2 路组相联的失效率为 7.6%；128KB 直接映象 Cache 的失效率为 1.0%，2 路组相联的失效率为 0.7%。

（1）推导出平均访存时间的公式。

（2）利用（1）中得到的公式，对于 2KBCache 和 128KBCache，计算伪相联的平均访存时间。

解：

不管作了何种改进，失效开销相同。不管是否交换内容，在同一“伪相联”组中的两块都是用同一个索引得到的，因此失效率相同，即：失效率$_{伪相联}$＝失效率$_{2路}$。

伪相联 cache 的命中时间等于直接映象 cache 的命中时间加上伪相联查找过程中的命中时间*该命中所需的额外开销。

命中时间$_{伪相联}$＝命中时间$_{1路}$＋伪命中率$_{伪相联}\times1$

交换或不交换内容，伪相联的命中率都是由于在第一次失效时，将地址取反，再在第二次查找带来的。

因此 伪命中率$_{伪相联}$＝命中率$_{2路}$－命中率$_{1路}$＝（1－失效率$_{2路}$）－（1－失效率$_{1路}$）

＝失效率$_{1路}$－失效率$_{2路}$。交换内容需要增加伪相联的额外开销。

平均访存时间$_{伪相联}$＝命中时间$_{1路}$＋（失效率$_{1路}$－失效率$_{2路}$）×1

＋失效率$_{2路}\times$失效开销$_{1路}$

将题设中的数据带入计算，得到：

平均访存时间$_{2Kb}$=1+(0.098-0.076)*1+(0.076 *50) =4.822

平均访存时间 $_{128Kb}$=1+(0.010-0.007)*1+(0.007 *50) =1.353

显然是 128KB 的伪相联 Cache 要快一些。

5.14 假设采用理想存储器系统时的基本 CPI 是 1.5，主存延迟是 40 个时钟周期；传输速率为 4 字节/时钟周期，且 Cache 中 50%的块是修改过的。每个块中有 32 字节，20%的指令是数据传送指令。并假设没有写缓存，在 TLB 失效的情况下需要 20 时钟周期，TLB 不会降低 Cache 命中率。CPU 产生指令地址或 Cache 失效时产生的地址有 0.2%没有在 TLB 中找到。

（1） 在理想 TLB 情况下，计算均采用写回法 16KB 直接映象统一 Cache、16KB 两路组相联统一 Cache 和 32KB 直接映象统一 Cache 机器的实际 CPI；

（2） 在实际 TLB 情况下，用（1）的结果，计算均采用写回法 16KB 直接映象统一 Cache、16KB 两路组相联统一 Cache 和 32KB 直接映象统一 Cache 机器的实际 CPI；

其中假设 16KB 直接映象统一 Cache、16KB 两路组相联统一 Cache 和 32KB 直接映象统一 Cache 的失效率分别为 2.9%、2.2%和 2.0%；25%的访存为写访问。

解： CPI=CPI $_{执行}$+存储停顿周期数/指令数

存储停顿由下列原因引起：

- 从主存中取指令
- load 和 store 指令访问数据
- 由 TLB 引起

$$\frac{\text{存储停顿周期数}}{\text{指令数}}=\frac{\text{取指令停顿}}{\text{指令数}}+\frac{\text{数据访问停顿}+\text{TLB停顿}}{\text{指令数}}$$

$$\frac{\text{停顿周期数}}{\text{指令数}}=\frac{\text{存储访问}}{\text{指令数}}\times\text{失效率}\times\text{失效开销}$$

$$\frac{\text{存储停顿周期数}}{\text{指令数}}=\left(R_{\text{指令}}P_{\text{指令}}\right)+\left(f_{\text{数据}}R_{\text{数据}}P_{\text{数据}}\right)+\frac{\text{TLB停顿}}{\text{指令数}}$$

（1）对于理想 TLB，TLB 失效开销为 0。而对于统一 Cache，$R_{\text{指令}}=R_{\text{数据}}$

$P_{\text{指令}}$=主存延迟＋传输一个块需要使用的时间＝40＋32/4＝48（拍）

若为读失效，$P_{\text{数据}}$＝主存延迟＋传输一个块需要使用的时间＝40＋32/4＝48（拍）

若为写失效，且块是干净的，

$P_{\text{数据}}$＝主存延迟＋传输一个块需要使用的时间＝40＋32/4＝48（拍）

若为写失效，且块是脏的，

$P_{\text{数据}}$＝主存延迟＋传输两个块需要使用的时间＝40＋64/4＝56（拍）

CPI=1.5+[RP+(RP*20%)+0]

指令访存全是读，而数据传输指令 Load 或 Store 指令，

$f_{\text{数据}}*P_{\text{数据}}$＝读百分比*（$f_{\text{数据}}*P_{\text{数据}}$）＋写百分比*（$f_{\text{数据}}*P_{\text{干净数据}}$*其对应的百分比

＋$f_{\text{数据}}*P_{\text{脏数据}}$*其对应的百分比）

＝20%*（75％×48＋25％*（50％*48+50％*（48＋16））=50（拍）

代入上述公式计算出结果为：

配置	失效率	CPI
16KB 直接统一映象	0.029	4.4
16KB 两路统一映象	0.022	3.4
32KB 直接统一映象	0.020	3.2

（2） $$\frac{TLB\text{停顿}}{\text{指令数}}=\left(\frac{\text{存储访问次数}}{\text{指令数}}\times\frac{TLB\text{访问}}{\text{存储访问次数}}\right)\times TLB\text{失效率}\times TLB\text{失效开销}$$

将 $f_{\text{数据}}$（数据访问指令频率），R_t 和 P_t（分别是 TLB 的失效率和失效开销），R_c 和 P_w（分别是 Cache 的失效率和写回的频率）代入公式得：

TLB 停顿/指令数=$\{[1+f_{\text{数据}}]*[R_c(1+R_w)]\}R_tP_t$

其中，$1+f_{\text{数据}}$：每条指令的访问内存次数；$R_c(1+R_w)$：每次内存访问需要的 TLB 访问次数。

由条件得：TLB 停顿/指令数=$\{[1+20\%]*[R_c(1+25\%)]\}0.2\%\times20$

配置	失效率	理想 TLB 的 CPI
16KB 直接统一映象	0.029	4.0
16KB 两路统一映象	0.022	3.4
32KB 直接统一映象	0.020	3.2