**1.** 假设有一条长流水线，仅仅对条件转移指令使用分支目标缓冲。假设分支预测错误的开销为4个时钟周期，缓冲不命中的开销为3个时钟周期。假设：命中率为90%，预测精度为90%，分支频率为15%，没有分支的基本CPI为1。

①求程序执行的CPI。

②相对于采用固定的2个时钟周期延迟的分支处理，哪种方法程序执行速度更快？

**参考答案：**

解：（1）程序执行的CPI = 没有分支的基本CPI（1） + 分支带来的额外开销

分支带来的额外开销是指在分支指令中，缓冲命中但预测错误带来的开销与缓冲没有命中带来的开销之和。

分支带来的额外开销= 15% \* (90%命中×10%预测错误×4 + 10％没命中×3)= 0.099

所以，程序执行的CPI ＝ 1 ＋ 0.099 = 1.099

（2）采用固定的2 个时钟周期延迟的分支处理CPI = 1 + 15%×2 = 1.3

由（1）（2）可知分支目标缓冲方法执行速度快。

**2.** 计算机系统中有三个部件可以改进，这三个部件的部件加速比为：

部件加速比1=30； 部件加速比2=20 部件加速比3=10

1. 如果部件2和部件3的可改进比例均为30%，那么当部件 1的可改进比例为多少时，系统加速比才可以达到10？

②如果三个部件的可改进比例分别为20%、10%和30%，三个部件同时改进，那么系统中不可加速部分的执行时间在总执行时间中占的比例是多少？

**参考答案：**

解：（1）在多个部件可改进情况下，Amdahl定理的扩展：



已知S1＝30，S2＝15，S3＝15，Sn＝10，F1＝0.3，F2＝0.3，得：



得F3＝0.36，即部件3的可改进比例为36%。

（2）设系统改进前的执行时间为T，则3个部件改进前的执行时间为：（0.3+0.3+0.2）T = 0.8T，不可改进部分的执行时间为0.2T。

已知3个部件改进后的加速比分别为S1＝30，S2＝20，S3＝10，因此3个部件改进后的执行时间为：



改进后整个系统的执行时间为：Tn = 0.045T+0.2T = 0.245T

那么系统中不可改进部分的执行时间在总执行时间中占的比例是：



3. 设指令流水线由取指令、分析指令和执行指令3个部件构成，每个部件经过的时间为△t，连续流入12条指令。分别画出标量流水处理机以及ILP均为4的超标量处理机、超长指令字处理机、超流水处理机的时空图，并分别计算它们相对于标量流水处理机的加速比。

参考答案：

解：标量流水处理机的时空图：



执行完12条指令需T1＝14△t。

超标量流水处理机与超长指令字处理机的时空图：

超标量流水处理机中，每一个时钟周期同时启动4条指令。执行完12条指令需T2＝5△t，相对于标量流水处理机的加速比为：





超长指令字处理机中，每4条指令组成一条长指令，共形成3条长指令。执行完12条指令需T3＝5△t，相对于标量流水处理机的加速比为：



超流水处理机的时空图：



超流水处理机中，每1/4个时钟周期启动一条指令。执行完12条指令需T4＝5.75△t，相对于标量流水处理机的加速比为：



4. 设一条指令的执行过程分成取指令、分析指令和执行指令三个阶段，每个阶段所需的时间分别为△t、△t和2△t 。分别求出下列各种情况下，连续执行N条指令所需的时间。

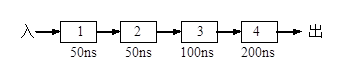
1. 只有“取指令”与“执行指令”重叠；
2. “取指令”、“分析指令”与“执行指令”重叠。

**参考答案：**

① 连续执行N条指令所需的时间为：4△t＋3（N-1）△t＝（3N＋1）△t

1. 连续执行N条指令所需的时间为：4△t＋2（N-1）△t＝（2N＋2）△t

**5.**有一指令流水线如下所示



* 1. 求连续输入10条指令，该流水线的实际吞吐率和效率；

②该流水线的“瓶颈”在哪一段？请采取两种不同的措施消除此“瓶颈”。对于你所给出的两种新的流水线，连续输入10条指令时，其实际吞吐率和效率各是多少？

**参考答案：**

（1）







（2）瓶颈在3、4段。

* 变成八级流水线（细分）







* 重复设置部件

1

2

3-1

3-2

4-1

4-2

4-3

4-4







6. 动态多功能流水线由6个功能段组成，如下图：

其中，S1、S4、S5、S6组成乘法流水线，S1、S2、S3、S6组成加法流水线，各个功能段时间均为50ns，假设该流水线的输出结果可以直接返回输入端，而且设置有足够的缓冲寄存器，若以最快的方式用该流水计算：



1. 画出时空图；
2. 计算实际的吞吐率、加速比和效率。

**参考答案：**



7. 某向量处理机有16个向量寄存器，其中V0~V5中分别放有向量A、B、C、D、E、F，向量长度均为8，向量各元素均为浮点数；处理部件采用两条单功能流水线，加法功能部件时间为2拍，乘法功能部件时间为3拍。采用类似于CARY-1的链接技术，先计算（A+B）\*C，在流水线不停流的情况下，接着计算（D+E）\*F。

①求此链接流水线的通过时间？（设寄存器入、出各需1拍）

②假如每拍时间为50ns，完成这些计算并把结果存进相应寄存器，此处理部件的实际吞吐率为多少MFLOPS？

**参考答案：**

解：（1）在这里假设A＋B的中间结果放在V6中，（A＋B）×C地最后结果放在V7中，D＋E地中间结果放在V8中，（D＋E）×F的最后结果放在V9中。具体实现参考下图：



通过时间应该为前者（（A＋B）×C）通过的时间：

T通过= (1+2+1)+(1+3+1) =9（拍）

（2）在做完（A＋B）×C之后，作（C＋D）×E就不需要通过时间了。

V6←A＋B

V7←V6×C

V8←D＋E

V9←V8×F



8. 假设分支目标缓冲的命中率为90%，程序中无条件转移指令的比例为5%，没有无条件转移指令的程序CPI值为1。假设分支目标缓冲中包含分支目标指令，允许无条件转移指令进入分支目标缓冲，则程序的CPI值为多少？

**参考答案：**

解：设每条无条件转移指令的延迟为x，则有：

1＋5%×x＝1.1

x＝2

当分支目标缓冲命中时，无条件转移指令的延迟为0。

所以 程序的CPI ＝ 1 ＋ 2 × 5% ×(1 －90%) ＝1.01

9. 一台**32**个处理器的计算机，对远程存储器访问时间为**400**ns。除了通信以外，假设计算中的访问均命中局部存储器。当发出一个远程请求时，本处理器挂起。处理器时钟时间为**1**GHz，如果指令基本的IPC为**2**(设所有访存均命中Cache)，求在没有远程访问的状态下与有**0.2%**的指令需要远程访问的状态下，前者比后者快多少?

**参考答案：**

解:没有远程访问时，机器的CPI为 1/基本IPC=1/2=0.5

有0.2%远程访问的机器的实际CPI为

CPI＝基本CPI＋远程访问率×远程访问开销

＝0.5＋0.2%×远程访问开销

远程访问开销为 ：

远程访问时间/时钟周期时间＝400 ns/1 ns＝400个时钟周期

∴ CPI＝0.5＋0.2%×400＝1.3

因此在没有远程访问的情况下的计算机速度是有0.2%远程访问的计算机速度的1.3/0.5=2.6倍。

10. 简述Tomasulo算法的基本思想。

**参考答案：**

答：核心思想是：① 记录和检测指令相关，操作数一旦就绪就立即执行，把发生RAW冲突的可能性减小到最少；② 通过寄存器换名来消除WAR冲突和WAW冲突。寄存器换名是通过保留站来实现，它保存等待流出和正在流出指令所需要的操作数。

基本思想：只要操作数有效，就将其取到保留站，避免指令流出时才到寄存器中取数据，这就使得即将执行的指令从相应的保留站中取得操作数，而不是从寄存器中。指令的执行结果也是直接送到等待数据的其它保留站中去。因而，对于连续的寄存器写，只有最后一个才真正更新寄存器中的内容。一条指令流出时，存放操作数的寄存器名被换成为对应于该寄存器保留站的名称（编号）。

11. 假定有一个处理机台数为p的共享存储器多处理机系统。设m为典型处理机每条指令执行时对全局存储器进行访问的平均次数。

设t为共享存储器的平均存取时间，x为使用本地存储器的单处理机MIPS速率。再假定在多处理机的每台处理机上执行n条指令。

①根据参数m，t，x，n和p，确定多处理机的有效MIPS速率。

②假设一台多处理机有p=32台RISC处理机，m=0.4，t=1us，要使多处理机的有效性能达到56MIPS，需要每台处理机的MIPS速率是多少（即x=？）？

③假设有p=32台CISC处理机用在上述多处理机系统中，每台处理机的x=2MIPS、m=1.6、t=1us，试问多处理机系统的有效MIPS速率是多少？

**参考答案：**

解：（1）有效MIPS速率=p\*x/(1+m\*x\*t)

（2）32\*x/(1－0.4\*x\*1)=56，得x=5.83

（3）有效MIPS速率=p\*x/(1+m\*x\*t)=32\*2/(1+1.6\*2\*1)=15.24

12. 假设对指令Cache的访问占全部访问的75%；而对数据Cache的访问占全部访问的25%。Cache的命中时间为1个时钟周期，失效开销为50 个时钟周期，在混合Cache中一次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0.15%，32KB的数据Cache的失效率为3.77%，64KB的混合Cache的失效率为0.95%。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？

**参考答案：**

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。

13. 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％。

**参考答案：**

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-路=2.0+1.4% \*80=3.12ns

平均访问时间2-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

CPUtime=（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-way=IC(2.0\*2+1.2\*0.014\*80)＝5.344IC

CPU time 2-way=IC(2.2\*2+1.2\*0.01\*80)＝5.36IC

相对性能比：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。

14. 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109字/s；
4. 25％的访存为写访问；
5. 存储器的最大流量为109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。

**参考答案：**

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=0.35×109/10 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＝0.13×10 9/10 9＝13%

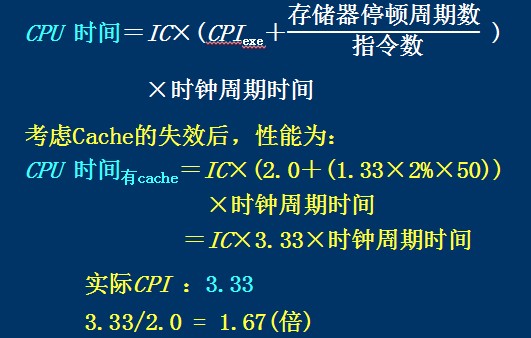
15. 降低Cache失效率有哪几种方法？简述其基本思想。

**参考答案：**

答：常用的降低Cache失效率的方法有下面几种：

1. 增加Cache块大小。增加块大小利用了程序的空间局部性。
2. 增加Cache的容量。
3. 提高相联度，降低冲突失效。
4. 伪相联Cache，降低冲突失效。当对伪相联Cache进行访问时，首先是按与直接映象相同的方式进行访问。如果命中，则从相应的块中取出所访问的数据，送给CPU，访问结束。如果不命中，就将索引字段的最高位取反，然后按照新索引去寻找“伪相联组”中的对应块。如果这一块的标识匹配，则称发生了“伪命中”。否则，就访问下一级存储器。
5. 硬件预取技术。在处理器提出访问请求前预取指令和数据。
6. 由编译器控制的预取，硬件预取的替代方法，在编译时加入预取的指令，在数据被用到之前发出预取请求。
7. 编译器优化，通过对软件的优化来降低失效率。
8. “牺牲”Cache。在Cache和其下一级存储器的数据通路之间增设一个全相联的小Cache，存放因冲突而被替换出去的那些块。每当发生不命中时，在访问下一级存储器之前，先检查“牺牲”Cache中是否含有所需的块。如果有，就将该块与Cache中某个块做交换，把所需的块从“牺牲”Cache 调入Cache。

16.假设Cache失效开销为50个时钟周期，当不考虑存储器停顿时，所有指令的执行时间都是2.0个时钟周期， Cache的失效率为2%，平均每条指令访存1.33次。试分析Cache对性能的影响。

解：

CPU时间也增加为原来的1.67倍。但若不采用Cache,则：

*CPI*＝2.0+50×1.33＝68.5

17. 考虑两种不同组织结构的Cache：直接映象Cache和两路组相联Cache，试问它们对CPU的性能有何影响？先求平均访存时间，然后再计算CPU性能。分析时请用以下假设：

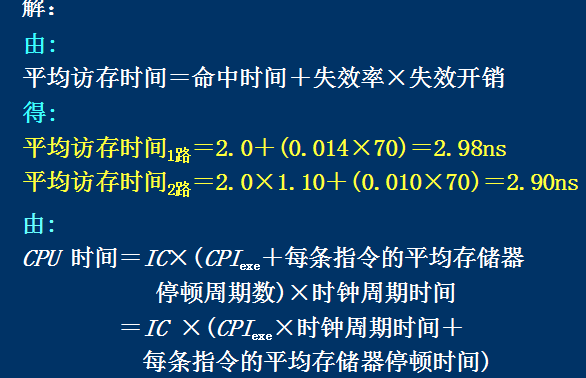
⑴ 理想Cache(命中率为100％)情况下的CPI为2.0，时钟周期为2ns，平均每条指令  
 访存1.3次。

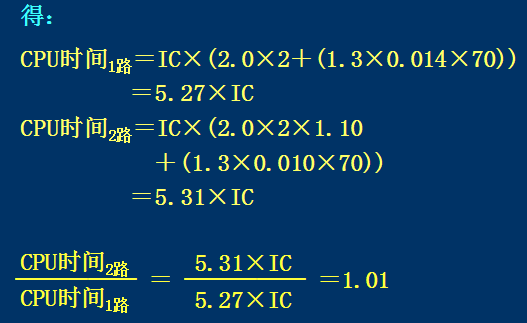
⑵ 两种Cache容量均为64KB，块大小都是32 字节。

⑶ 图5.10说明，在组相联Cache中，我们必须增加一个多路选择器，用于根据标识匹配结果从相应组的块中选择所需的数据。因为CPU 的速度直接与Cache命中的速度紧密相关,所以对于组相联Cache，由于多路选择器的存在而使CPU的时钟周期增加到原来的1.10倍。

⑷ 这两种结构Cache的失效开销都是70ns。在实际应用中，应取整为整数个时钟周期。

⑸ 命中时间为1个时钟周期，64KB直接映象Cache的失效率为1.4%，相同容量的两路组相联Cache的失效率为1.0%。





18. 假设一台计算机的I/O处理占10％，当其CPU性能改进，而I/O性能保持不变时，系统总体性能会出现什么变化？

* 如果CPU的性能提高10倍
* 如果CPU的性能提高100倍

解：假设原来的程序执行时间为1个单位时间。如果CPU的性能提高10倍，程序的计算（包含I/O处理）时间为：

(1 - 10%)/10 + 10% = 0.19

即整机性能只能提高约5倍，差不多有50％的CPU性能浪费在I/O上。

如果CPU性能提高100倍，程序的计算时间为：

(1 - 10%)/100 + 10% = 0.109

而整机性能只能提高约10倍，表示有90％的性能浪费在没有改进的I/O上了。