体系结构第四章课后练习

第一题：

B.5 [10/10/10/10/] <B.2>你正要采用一个具有以下特征的处理器构建系统：循序执行，运行频率为1.1 GHz，排除存储器访问在外的CPI为0.7。只有载入和存储指令能从存储器读写数据，载入指令占全部指令的20%，存储指令占5%。此计算机的存储器系统包括一个分离的L1缓存，它在命中时不会产生任何代价。I缓存和D缓存都是直接映射，分别为32 KB。I缓存的缺失率2%，块大小为32字节，D缓存为直写缓存，缺失率为5%，块大小为16字节。D缓存上有一个写入缓冲区，消除了绝大多数写入操作的停顿，占总写入操作的95%。512 KB写回、统一L2缓存的块大小为64字节，访问时间为15ns。它由128位数据总线连接到L1缓存，运行频率为266 MHz，每条总线每个时间周期可以传送一个128位字。在发往此系统L2缓存的所有存储器引用中，其中80%的引用无须进入主存储器就可以得到满足。另外，在被替换的所有块中，50%为脏块。主存储器的宽度为128位，访问延迟为60ns，在此之后，可以在这个宽128位、频率为133MHz的主存储器总线上以每个周期传送一个字的速率来传送任意数目的总线字。

a. [10] <B.2>指令访问的存储器平均访问时间为多少？

b. [10] <B.2>数据读取的存储器平均访问时间为多少？

c. [10] <B.2>数据写入的存储器平均访问时间为多少？

d. [10] <B.2>包括存储器访问在内的整体CPI为多少？

第二题：

B.8 [20/20/15/25] <B.3>LRU替换策略基于以下假定：如果最近访问地址A1的频率低于地址A2，那么未来再次访问A2的时机要早于A1。因此，为A2指定了高于A1的优先级。试讨论，当一个大于指令缓存的循环连续执行时，这一假定为什么不成立。例如，考虑一个全相联128字节指令缓存，其块大小为4个字节（每个块可以正好容纳一条指令）。此缓存使用LRU替换策略。

a. [20] <B.3>对于一个拥有大量迭代的64字节循环，渐近指令缺失率为多少？

b. [20] <B.3>对于大小为192字节和320字节的循环，重复(a)部分。

c. [15] <B.3>如果缓存替换策略改为最近使用最多（MRU）（替换最近访问最多的缓存行），以上三种情景（64、192、320字节的循环）中的哪一种情景将因为这一策略而受益？

d. [25] <B.3>提出执行性能可能优于LRU的更多替换策略。

第三题

2.11 [12/15] <2.2>考虑在L2缓存缺失时使用关键字优先和提前重启动。假定L2缓存的容量为1 MB、块大小为64字节、填充路径宽16字节。假定能够以每4个处理器周期16个字节的速度写入L2，从存储器控制器接收前16个字节块的时间为120个周期，每从主存储器接收另外16个字节的块需要16个周期，也可以直接将数据传送给L2缓存的读取端口。忽略向L2缓存发送缺失请求及向L1缓存传送被请求数据的周期数。

a. [12] <2.2>在使用、不使用关键字优先和提前重启动时，为L2缓存缺失提供服务分别需要多少个周期？

b. [15] <2.2>你是否认为关键字优先和提前重启动对于L1缓存或L2缓存更重要一些，哪些因素影响着它们的相对重要性？

第四题

2.12 [12/12] <2.2>在直写L1缓存与写回L2缓存之间设计一个写缓冲区。L2缓存写数据总线的宽度为16 B，可以每4个处理器周期向一个独立缓存地址执行一次写操作。

a. [12] <2.2>每个写缓冲区项目应当为多少字节？

b. [15] <2.2>如果所有其他指令可以与存储指令并行发射，块存在于L2缓存中，在通过执行64位存储指令将存储器置零时，使用一个合并写缓冲区来代替非合并缓冲区，在稳定状态下可以得到什么样的加速比？

c. [15] <2.2>对于采用阻塞缓存与非阻塞缓存的系统，可能出现的L1缺失对于所需写缓冲区项目的个数有什么样的影响？

HW06

EX1

> 本题为《计算机体系结构：量化研究方法（第五版）》第五章案例研究 5.6 题

假定图 5-24 的缓存内容和表 5-12 中实现方式 1 的定时参数。以下代码序列在基本协议和练习 5.5 的新 MESI 协议中的总停顿周期为多少？假定不需要互连事务的状态转换不会导致额外的停顿周期。

a.

P0: read 100
P0: write 100<–40

b.

P0: read 120
P0: write 120<–60

c.

P0: read 100
P0: read 120

d.

P0: read 100
P1: write 100<–60

e.

P0: read 100
P0: write 100<–60
P1: write 100<–40

EX2

> 本题为《计算机体系结构：量化研究方法（第五版）》第五章案例研究 5.10 题

目录式也议比监听式协议的可扩展性更强，因为它们会向那些拥有块副本的 节点发送显式请求和失效消息，而监听式协议则向所有节点广播所有请求和失效消息。考虑图 5-25 所示的八处理器系统，假定所有未显示缓存拥有失效块。对于下面的每个序列，确认哪些节点（芯片/处理器）接收每个请求和失效消息。

a.

P0,0: write 100<–80

b.

P0,0: write 108<–88

c.

P0,0: write 118<–90

d.

P1,0: write 128<–98

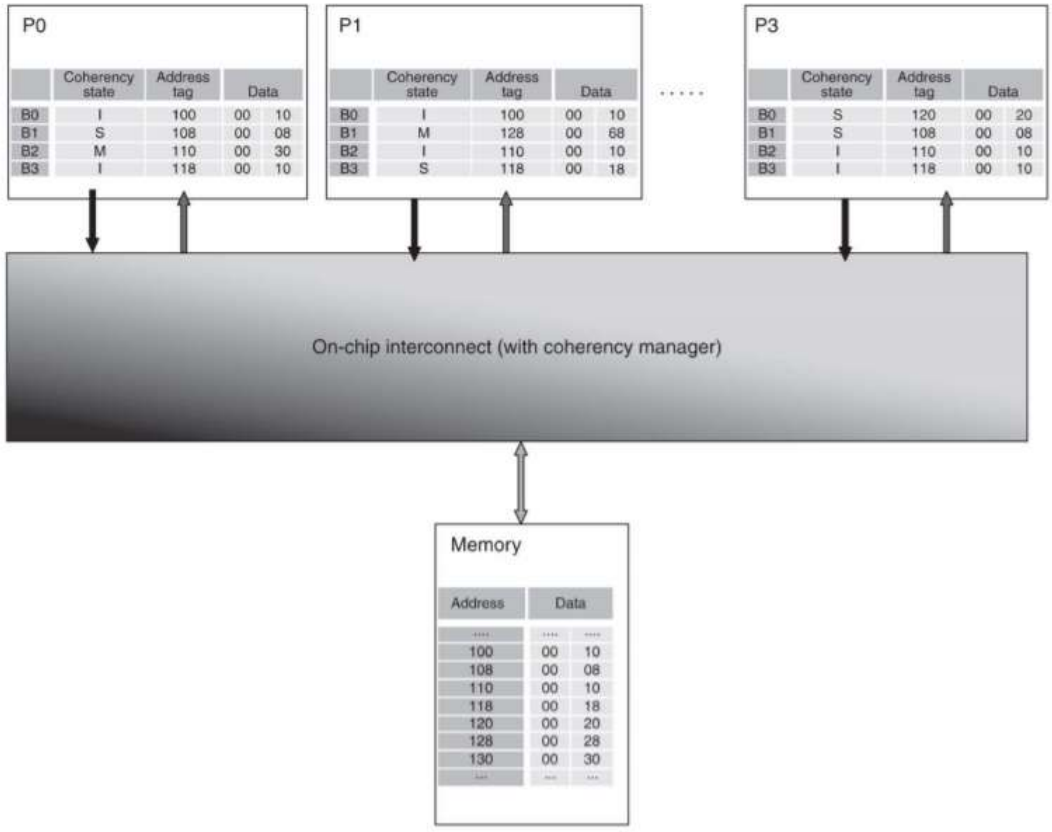


图 5-24

表 5-12：

Parameter	Implementation 1	Implementation 2
N_{memory}	100	100
N_{cache}	40	130
$N_{invalidate}$	15	15
$N_{writeback}$	10	10

图5-25：

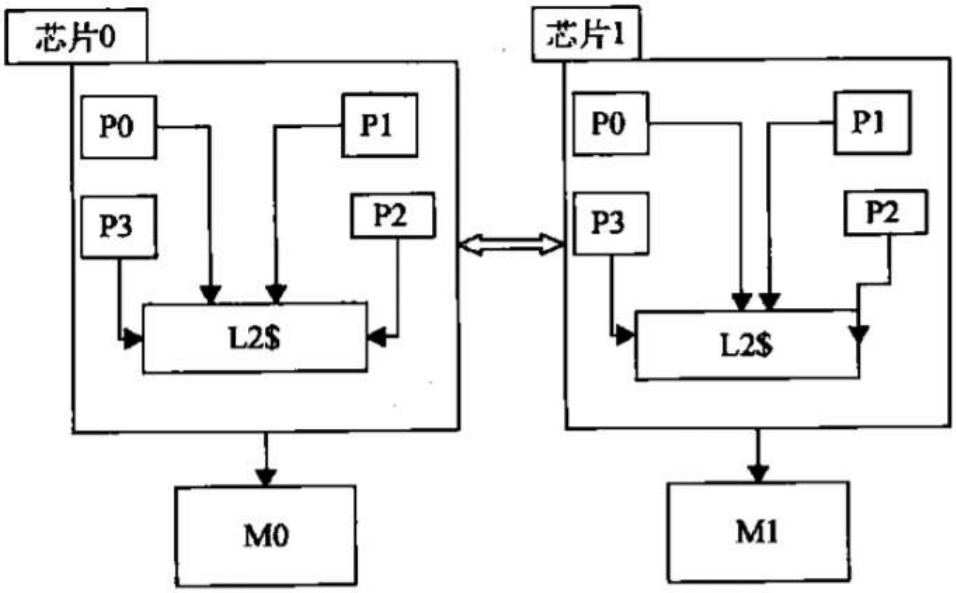
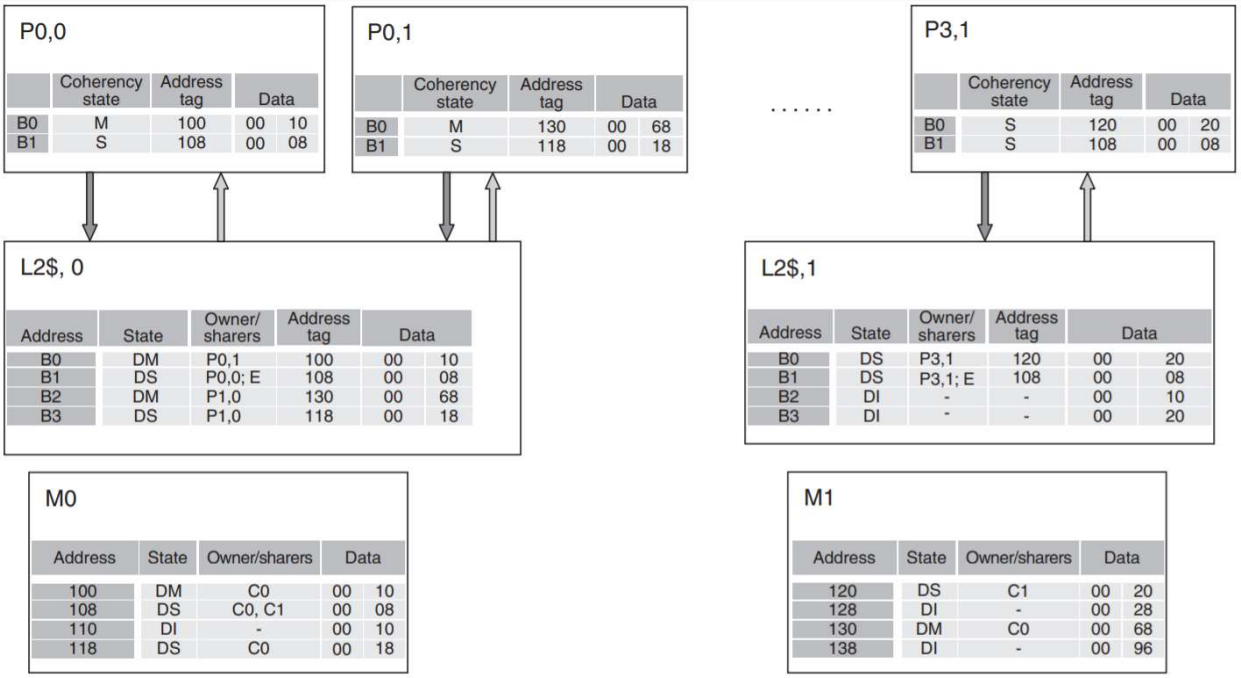


图5-26：



P0,0

	Coherency state	Address tag	Data	
B0	M	100	00	10
B1	S	108	00	08

P0,1

	Coherency state	Address tag	Data	
B0	M	130	00	68
B1	S	118	00	18

P3,1

	Coherency state	Address tag	Data	
B0	S	120	00	20
B1	S	108	00	08

L2$, 0

Address	State	Owner/ sharers	Address tag	Data	
B0	DM	P0,1	100	00	10
B1	DS	P0,0; E	108	00	08
B2	DM	P1,0	130	00	68
B3	DS	P1,0	118	00	18

L2$,1

Address	State	Owner/ sharers	Address tag	Data	
B0	DS	P3,1	120	00	20
B1	DS	P3,1; E	108	00	08
B2	DI	-	-	00	10
B3	DI	-	-	00	20

M0

Address	State	Owner/sharers	Data	
100	DM	C0	00	10
108	DS	C0, C1	00	08
110	DI	-	00	10
118	DS	C0	00	18

M1

Address	State	Owner/sharers	Data	
120	DS	C1	00	20
128	DI	-	00	28
130	DM	C0	00	68
138	DI	-	00	96

计算机体系结构第六章作业

1. 在以下的循环中，找出所有真相关、输出相关和反相关。通过重命名来消除输出相关和反相关。

```
for (i=0; i<100; i++) {
    A[i] = A[i] * B[i];        /* S1 */
    B[i] = A[i] + c;           /* S2 */
    A[i] = C[i] * c;           /* S3 */
    C[i] = D[i] * A[i];        /* S4 */
}
```

2.

[10/20/20/15/15] <4.2>考虑以下代码，它将两个包含单精度复数值的向量相乘：

```
for (i=0;i<300;i++) {
        c_re[i] = a_re[i] * b_re[i] - a_im[i] * b_im[i];
        c_im[i] = a_re[i] * b_im[i] + a_im[i] * b_re[i];
}
```

假定处理器的运行频率为 700MHz，•最大向量长度为 64。载入/存储单元的启动开销为 15 个时钟周期，乘法单元为 8 个时钟周期，加法/减法单元为 5 个时钟周期。

a. [10] <4.2>这个内核的运算密度为多少？给出理由。

b. [20] <4.2>将此循环转换为使用条带挖掘的 VMIPS 汇编代码。

c. [20] <4.2>假定采用链接和单一存储器流水线，需要多少次钟鸣？每个复数结果值需要多少个时钟周期（包括启动开销在内）？

d. [15] <4.2>如果向量序列被链接在一起，每个复数结果值需要多少个时钟周期（包含开销）？

3.

[10/15] <4.4>假定有一种包含 10 个 SIMD 处理器的 GPU 体系结构。每条 SIMD 指令的宽度为 32，每个 SIMD 处理器包含 8 个车道，用于执行单精度运算和载入/存储指令，也就是说，每个非分岔 SIMD 指令每 4 个时钟周期可以生成 32 个结果。假定内核的分岔分支将导致平均 80%的线程为活动的。假定在所执行的全部 SIMD 指令中，70%为单精度运算、20%为载入/存储。由于并不包含所有存储器延迟，所以假定 SIMD 指令平均发射率为 0.85。假定 GPU 的时钟速度为 1.5 GHz。

a. [10] <4.4>计算这个内核在这个 GPU 上的吞吐量，单位为 GFLOP/s。

b. [15] <4.4>假定我们有以下选项：

(1) 将单精度车道数增大至 16。

(2) 将 SIMD 处理器数增大至 15（假定这一改变不会影响所有其他性能度量，代码会扩展到增加的处理器上）。

(3) 添加缓存可以有效地将存储器延迟缩减 40%，这样会将指令发射率增加至 0.95，对于这些改进中的每一项。

吞吐量的加速比为多少？

3.

[10] <4.4>假定一个虚设 GPU 具有以下特性：

- 时钟频率为 1.5 GHz；
- 包含 16 个 SIMD 处理器，每个处理器包含 16 个单精度浮点单元；
- 片外存储器带宽为 100 GB/s。

不考虑存储器带宽，假定所有存储器延迟可以隐藏，则这一 GPU 的峰值单精度浮点吞吐量为多少 GFLOP/s？在给定存储器带宽限制下，这一吞吐量是否可持续？

第一题

3.2 [10] <1.8、3.1、3.2>思考一下延迟数目到底意味着什么——它们表示一个给定函数生成其输出结果所需要的时钟周期数，没有别的意思。如果整个流水线在每个功能单元的延迟周期中停顿，那么至少要保证任何一对“背靠背”指令（生成结果的指令后面紧跟着使用结果的指令）正确执行。但并非所有指令对具有这种“生成者/使用者”的关系。有时，两条相邻指令之间没有任何关系。如果流水线检测到真正的数据相关，而且只会因为这些真数据相关而停顿，而不会仅仅因为有某个功能单元繁忙就盲目停顿，那表 3-22 代码序列中的循环体需要多少个时钟周期？在代码中需要容纳所述延迟的时候插入<stall>。（提示：延迟为+2 的指令需要在代码序列中插入两个<stall>时钟周期。可以这样来考虑：一条需要一个时钟周期的指令的延迟为 1+0，也就是不需要额外的等待状态。那么延迟 1+1 就意味着 1 个停顿周期，延迟 1+N有 N个额外停顿周期。

表3-22 练习3.1至练习3.6的代码与延迟

				超过一个时钟周期的延迟
Loop:	LD	F2,0(RX)	存储器LD	+4
I0:	DIVD	F8,F2,F0	存储器SD	+1
I1:	MULTD	F2,F6,F2	整数ADD，SUB	+0
I2:	LD	F4,0(Ry)	分支	+1
I3:	ADDD	F4,F0,F4	ADDD	+1
I4:	ADDD	F10,F8,F2	MULTD	+5
I5:	ADDI	Rx,Rx,#8	DIVD	+12
I6:	ADDI	Ry,Ry,#8		
I7:	SD	F4,0(Ry)		
I8:	SUB	R20,R4,Rx		
I9:	BNZ	R20,Loop		

第二题

3.14 [25/25/25] <3.2、3.7>在这个练习中，我们研究如何利用软件技术从一个常见的向量循环中提取指令级并行（ILP）。下面的循环是所谓的 DAXPY 循环（双精度 aX加 Y），它是高斯消元法的核心运算。下面的代码实现 DAXPY 运算 Y=aX+Y，向量长度为 100。最初，R1 被设置为数组 X的基地址，R2 被设置为 Y的基地址：

```
      DADDIU   R4,R1,#800  ; R1 = upper bound for X
foo:  L.D      F2,0(R1)    ; (F2) = X(i)
      MUL.D    F4,F2,F0    ; (F4) = a*X(i)
      L.D      F6,0(R2)    ; (F6) = Y(i)
      ADD.D    F6,F4,F6    ; (F6) = a*X(i) + Y(i)
      S.D      F6,0(R2)    ; Y(i) = a*X(i) + Y(i)
      DADDIU   R1,R1,#8    ; increment X index
      DADDIU   R2,R2,#8    ; increment Y index
      DSLTU    R3,R1,R4    ; test: continue loop?
      BNEZ     R3,foo      ; loop if needed
```

假定功能单元的延迟如下表所示。假定在 ID 阶段解决一个延迟为 1 周期的分支。假定结果被完全旁路。

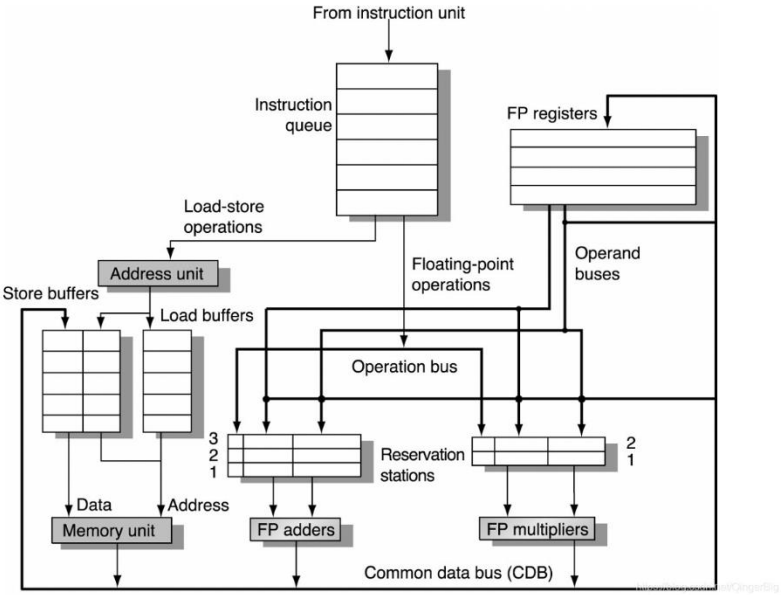
产生结果的指令	使用结果的指令	延迟（单位：时钟周期）
浮点乘	浮点ALU运算	6
浮点加	浮点ALU运算	4
浮点乘	浮点存储	5
浮点加	浮点存储	4
整数运算和所有载入	任何指令	2

a. [25] <3.2>假定一个单发射流水线。说明在编译器未进行调度以及对浮点运算和分支延迟进

行调度之后，该循环是什么样的，包括所有停顿或空闲时间周期。在未调度和已调度情况下，结果向量 Y 中每个元素的执行时间为多少个时钟？为使处理器硬件独自匹配调度编译器所实现的性能改进，时钟频率应当为多少？（忽略加快时钟速度会对存储器系统性能产生的影响。）

b. [25] <3.2>假定一个单发射流水线。根据需要对循环进行任意次展开，使调度中不存在任何停顿，消除循环开销指令。必须将此循环展开多少次？给出指令调度。结果中每个元素的执行时间为多少？

第三题



如图所示,假设浮点加法执行需要 2 个周期,浮点乘法需要 3 个周期,功能单元完全流水化。采用 Tomasulo 算法运行下列指令,写出第 7 个周期 Reservation Stations 和 Register result status 的状态。

ADD.D F4,F0,F8

MULT.D F2,F0,F4

ADD.D F4,F4,F8

MULT.D F8,F4,F2

第四题

3.19 [10/5] <3.9>考虑分支目标缓冲区，正确条件分支预测、错误预测和缓存缺失的代价分别为 0、2 和 2 个时钟周期。考虑一种区分条件与无条件分支的分支目标缓冲区设计，而条件分支存储目标地址，对于无条件分支则存储目标指令。

a. [10] <3.9>当缓冲区中发现无条件分支时，代价为多少个时钟周期？

b. [10] <3.9>判断对于无条件分支进行分支折合所获得的改进。假定命中率为 90%，无条件分支频率为 5%，缓冲区缺失的代价为两个时钟周期。这样可以获得多少改进？对于这一改进来说，必须达到多高的命中率才能提供性能增益？

第五题

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Δt,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。