2. 简单回答下列问题。

(1) 冯·诺依曼计算机由哪几部分组成？各部分的功能是什么？采用什么工作方式？

(2) 摩尔定律的主要内容是什么？

(3) 计算机系统的层次结构如何划分？计算机系统的用户可分哪几类？每类用户工作在哪个层次？

(4) 程序的 CPI 与哪些因素有关？

(5) 为什么说性能指标 MIPS 不能很好地反映计算机的性能？

3. 假定你的朋友不太懂计算机，请用简单通俗的语言给你的朋友介绍计算机系统是如何工作的。

4. 你对计算机系统的哪些部分最熟悉，哪些部分最不熟悉？最想进一步了解细节的是哪些部分的内容？

5. 若有两个基准测试程序 P1 和 P2 在机器 M1 和 M2 上运行，假定 M1 和 M2 的价格分别是 5000 元和 8000 元，表 1.2 给出了 P1 和 P2 在 M1 和 M2 上所用的时间和指令条数。

表 1.2　P1 和 P2 在 M1 和 M2 上所用的时间和指令条数

程序	M1		M2	
	指令条数	执行时间	指令条数	执行时间
P1	200×10^6	10 000ms	150×10^6	5000ms
P2	300×10^3	3ms	420×10^3	6ms

请回答下列问题。

(1) 对于 P1，哪台机器的速度快？快多少？对于 P2 呢？

(2) 在 M1 上执行 P1 和 P2 的速度分别是多少 MIPS？在 M2 上的执行速度又各是多少？从执行速度来看，对于 P2，哪台机器的速度快？快多少？

(3) 假定 M1 和 M2 的时钟频率各是 800MHz 和 1.2GHz，则在 M1 和 M2 上执行 P1 时的平均时钟周期数 CPI 各是多少？

(4) 如果某个用户需要大量使用程序 P1，并且该用户主要关心系统的响应时间而不是吞吐率，那么，该用户需要大批构进机器时，应该选择 M1 还是 M2？为什么？（提示：从性价比上考虑）

(5) 如果另一个用户也需要购进大批机器，但该用户使用 P1 和 P2 一样多，主要关心的也是响应时间，那么，应该选择 M1 还是 M2？为什么？

6. 若机器 M1 和 M2 具有相同的指令集，其时钟频率分别为 1GHz 和 1.5GHz。在指令集中有 5 种不同类型的指令 A～E。表 1.3 给出了在 M1 和 M2 上每类指令的平均时钟周期数 CPI。

表 1.3　在 M1 和 M2 上每类指令的平均时钟周期数 CPI

机器	A	B	C	D	E
M1	1	2	2	3	4
M2	2	2	4	5	6

请回答下列问题。

(1) M1 和 M2 的峰值 MIPS 各是多少？

(2) 假定某程序 P 的指令序列中，5 类指令具有完全相同的指令条数，则程序 P 在 M1 和 M2 上运行时，哪台机器更快？快多少？在 M1 和 M2 上执行程序 P 时的平均时钟周期数 CPI 各是多少？

7. 假设同一套指令集用不同的方法设计了两种机器 M1 和 M2。机器 M1 的时钟周期为 0.8ns，机器 M2 的时钟周期为 1.2ns。某个程序 P 在机器 M1 上运行时的 CPI 为 4，在 M2 上的 CPI 为 2。对于程序 P 来说，哪台机器的执行速度更快？快多少？

8. 假设某机器 M 的时钟频率为 4GHz，用户程序 P 在 M 上的指令条数为 8×10^9，其 CPI 为 1.25，则 P 在 M 上的执行时间是多少？若在机器 M 上从程序 P 开始启动到执行结束所需的时间是 4s，则 P 占用的 CPU 时间的百分比是多少？

9. 假定某编译器对某段高级语言程序编译生成两种不同的指令序列 S1 和 S2，在时钟频率为 500MHz 的机器 M 上运行，目标指令序列中用到的指令类型有 A、B、C 和 D 四类。四类指令在 M 上的 CPI 和两个

指令序列所用的各类指令条数如表 1.4 所示。

表 1.4　A、B、C、D 四类指令在 M 上的 CPI 和两个指令序列所用的各类指令条数

	A	B	C	D
各指令的 CPI	1	2	3	4
S1 的指令条数	5	2	2	1
S2 的指令条数	1	1	1	5

请回答：S1 和 S2 各有多少条指令？CPI 各为多少？所含的时钟周期数各为多少？执行时间各为多少？

10. 假定机器 M 的时钟频率为 1.2GHz，某程序 P 在机器 M 上的执行时间为 12 秒。对 P 优化时，将其所有的乘 4 指令都换成了一条左移两位的指令，得到优化后的程序 P′。已知在 M 上乘法指令的 CPI 为 5，左移指令的 CPI 为 2，P 的执行时间是 P′执行时间的 1.2 倍，则 P 中有多少条乘法指令被替换成了左移指令被执行？

2. 简单回答下列问题。

(1) 计算机内部为何要采用层次化存储体系结构？层次化存储体系结构如何构成？

(2) SRAM 芯片和 DRAM 芯片各有哪些特点？各自用在哪些场合？

(3) CPU 和主存之间有哪两种通信定时方式？SDRAM 芯片采用什么方式和 CPU 交换信息？

(4) 为什么在 CPU 和主存之间引入 cache 能提高 CPU 访存效率？

(5) 为什么说 cache 对程序员来说是透明的？

(6) 什么是 cache 映射的关联度？关联度与命中率、命中时间的关系各是什么？

(7) 为什么直接映射方式不需要考虑替换策略？

(8) 为什么要考虑 cache 的一致性问题？读操作时是否要考虑 cache 的一致性问题？为什么？

(9) 什么是物理地址？什么是逻辑地址？地址转换由硬件还是软件实现？为什么？

(10) 什么是页表？什么是快表(TLB)？

(11) 在存储器层次化结构中，“cache—主存”、“主存—磁盘”这两个层次有哪些不同？

3. 已知某机主存容量为 64KB，按字节编址。假定用 1K×4 位的 DRAM 芯片构成该存储器，要求回答以下问题。

(1) 需要多少个这样的 DRAM 芯片？

(2) 主存地址共多少位？哪几位用于选片？哪几位用于片内选址？

(3) 画出该存储器的逻辑框图。

4. 假定用 64K×1 位的 DRAM 芯片构成 256K×8 位的存储器，要求回答以下问题。

(1) 所需芯片数为多少？画出该存储器的逻辑框图。

(2) 若采用异步刷新方式，每单元刷新间隔不超过 2ms，则产生刷新信号的间隔是多少时间？若采用集中刷新方式，则存储器刷新一遍最少用多少个读写周期？

5. 假定用 8K×8 位的 EPROM 芯片组成 32K×16 位的只读存储器，要求回答以下问题。

(1) 数据寄存器最少应有多少位？　　(2) 地址寄存器最少应有多少位？

(3) 共需多少个 EPROM 芯片？　　(4) 画出该只读存储器的逻辑框图。

6. 某计算机中已配有 0000H～7FFFH 的 ROM 区域，现在再用 8K×4 位的 RAM 芯片形成 32K×8 位的存储区域，CPU 地址线为 A0～A15，数据线为 D0～D7，控制信号为 R/$\overline{W}$(读写)、$\overline{MREQ}$(访存)。要求说明地址译码方案，并画出 ROM 芯片、RAM 芯片与 CPU 之间的连接图。假定上述其他条件不变，只是 CPU 地址线改为 24 根，地址范围 000000H～007FFFH 为 ROM 区，剩下的所有地址空间都用8K×4 位的 RAM 芯片配置，则需要多少个这样的 RAM 芯片？

7. 假定一个存储器系统支持四体交叉存取，某程序执行过程中访问地址序列为 3，9，17，2，51，37，13，4，8，41，67，10，则哪些地址访问会发生体冲突？

8. 现代计算机中，SRAM 一般用于实现快速小容量的 cache，而 DRAM 用于实现慢速大容量的主存。以前超级计算机通常不提供 cache，而是用 SRAM 来实现主存(如 Cray 巨型机)，请问：如果不考虑成本，你还这样设计高性能计算机吗？为什么？

9. 对于数据的访问，分别给出具有下列要求的程序或程序段的示例。

(1) 几乎没有时间局部性和空间局部性。

(2) 有很好的时间局部性，但几乎没有空间局部性。

(3) 有很好的空间局部性，但几乎没有时间局部性。

(4) 空间局部性和时间局部性都好。

10. 假定某计算机主存地址空间大小为 1GB，按字节编址，cache 的数据区(不包括标记、有效位等)有 64KB，块大小为 128 字节，采用直接映射和直写(write-through)方式。请回答以下问题。

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

(2) cache 的总容量为多少位？

11. 假定某计算机的 cache 共 16 行，开始为空，块大小为一个字，采用直接映射方式，按字编址。CPU 执行某程序时，依次访问以下地址序列：2，3，11，16，21，13，64，48，19，11，3，22，4，27，6 和 11。要求，

(1) 说明每次访问是命中还是缺失，试计算访问上述地址序列的命中率。

(2) 若 cache 数据区容量不变，而块大小改为 4 个字，则上述地址序列的命中情况又如何？

12. 假定数组元素在主存按从左到右的下标顺序存放。试改变下列函数中循环的顺序，使得其数组元素的访问与排列顺序一致，并说明为什么修改后的程序比原来的程序执行时间短。

```
int sum_array ( int a[N][N][N])
{
    int i, j, k, sum=0;
    for (i=0; i<N; i++)
        for (j=0; j<N; j++)
            for (k=0; k<N; k++) sum+=a[k][i][j];
    return sum;
```

13. 分析比较以下三个函数中数组访问的空间局部性，并指出哪个最好，哪个最差？

```
#define N 1000
typedef struct {
      int vel[3];
      int acc[3];
  } point;
point p[N];
void clear1(point * p, int n)
{
  int i, j;
  for (i=0; i<n; i++) {
      for (j=0; j<3; j++)
          p[i].vel[j]=0;
      for (j=0; i<3; j++)
          p[i].acc[j]=0;
  }
```

```
#define N 1000
typedef struct {
      int vel[3];
      int acc[3];
  } point;
point p[N];
void clear2(point * p, int n)
{
  int i, j;
    for (i=0; i<n; i++) {
      for (j=0; j<3; j++) {
          p[i].vel[j]=0;
          p[i].acc[j]=0;
      }
    }
```

```
#define N 1000
typedef struct {
      int vel[3];
      int acc[3];
  } point;
point p[N];
void clear3(point * p, int n)
{
  int i, j;
    for (j=0; j<3; j++) {
      for (i=0; i<n; i++)
          p[i].vel[j]=0;
      for (i=0; i<n; i++)
          p[i].acc[j]=0;
    }
```

14. 以下是计算两个向量点积的程序段。

```
float dotproduct (float x[8], float y[8])
{
    float sum=0.0;
    int i,;
    for (i=0; i<8; i++) sum+=x[i] * y[i];
    return sum;
}
```

要求：

(1) 试分析该段代码中访问数组 x 和 y 的时间局部性和空间局部性，并推断命中率的高低。

(2) 假定该段程序运行的计算机中数据 cache 采用直接映射方式，其数据区容量为 32 字节，每个主存块大小为 16 字节。假定编译程序将变量 sum 和 i 分配给寄存器，数组 x 存放在 00000040H 开始的 32 字节的连续存储区中，数组 y 紧跟在 x 后进行存放。试计算该程序中数据访问的命中率，要求说明每次访问时 cache 的命中情况。

(3) 将上述(2)中的数据 cache 改用 2-路组相联映射方式，块大小改为 8 字节，其他条件不变，则该程序数据访问的命中率是多少？

(4) 上述(2)中条件不变的情况下，如果将数组 x 定义为 float[12]，则数据访问的命中率又是多少？

```
⋮
for(i=0; i<10000; i++)
    for(j=0; j<128; j=j+s)
        c=a[j];
```

23. 假定一个虚拟存储系统的虚拟地址为 40 位，物理地址为 36 位，页大小为 16KB，按字节编址。若页表中有有效位、存储保护位、修改位、使用位，共占 4 位，磁盘地址不记录在页表中，则该存储系统中每个进程的页表大小为多少？如果按计算出来的实际大小构建页表，则会出现什么问题？

24. 假定一个计算机系统中有一个 TLB 和一个 L1 Data Cache。该系统按字节编址，虚拟地址 16 位，

15. 以下是对矩阵进行转置的程序段。

```
typedef int array[4][4];
void transpose(array dst, array src)
{
    int i, j;
    for (i=0; i<4; i++)
        for (j=0; j<4; j++)
            dst[j][i]=src[i][j];
}
```

假设该段程序运行的计算机中 sizeof(int)＝4，且只有一级 cache，其中 L1 data cache 的数据区大小为 32B，采用直接映射、回写方式，块大小为 16B，初始为空。数组 dst 从地址 0000C000H 开始存放，数组 src 从地址 0000C040H 开始存放。填写表 4.2，说明对数组元素 src[row][col]和 dst[row][col]的访问是命中(Hit)还是缺失(Miss)？若 L1 data cache 的数据区容量改为 128B 时，重新填写表 4.2 的内容。

表 4.2　题 15 用表

	src 数组				dst 数组			
	col＝0	col＝1	col＝2	col＝3	col＝0	col＝1	col＝2	col＝3
row＝0	Miss				Miss			
row＝1								
row＝2								
row＝3								

18. 假定某处理器可通过软件对高速缓存设置不同的写策略，那么，在下列两种情况下，应分别设置成什么写策略？为什么？

(1) 处理器主要运行包含大量存储器写操作的数据访问密集型应用。

(2) 处理器运行程序的性质与(1)相同，但安全性要求很高，不允许有任何数据不一致的情况发生。

19. 已知 cache 1 采用直接映射方式，共 16 行，块大小为一个字，缺失损失为 8 个时钟周期；cache 2 也采用直接映射方式，共 4 行，块大小为 4 个字，缺失损失为 11 个时钟周期。假定开始时 cache 为空，采用字编址方式。要求找出一个访问地址序列，使得 cache 2 具有更低的缺失率，但总的缺失损失反而比 cache 1 大。

20. 提高关联度通常会降低缺失率，但并不总是这样。请给出一个地址访问序列，使得采用 LRU 替换算法的 2-路组相联映射 cache 比具有同样大小的直接映射 cache 的缺失率更高。

21. 假定有三个处理器，分别带有以下不同的 cache。

cache 1：采用直接映射方式，块大小为一个字，指令和数据的缺失率分别为 4%和 6%。

cache 2：采用直接映射方式，块大小为 4 个字，指令和数据的缺失率分别为 2%和 4%。

cache 3：采用 2-路组相联映射方式，块大小为 4 个字，指令和数据的缺失率分别为 2%和 3%。

在这些处理器上运行同一个程序，其中有一半是访存指令，在三个处理器上测得该程序的 CPI 都为 2.0。已知处理器 1 和 2 的时钟周期都为 420ps，处理器 3 的时钟周期为 450ps。若缺失损失为(块大小+6)个时钟周期，请问：哪个处理器因 cache 缺失而引起的额外开销最大？哪个处理器执行速度最快？

22. 假定某处理器带有一个数据区容量为 256B 的 cache，其块大小为 32B。以下 C 语言程序段运行在该处理器上，设 sizeof(int)=4，编译器将变量 i，j，c，s 都分配在通用寄存器中，因此，只要考虑数组元素的访存情况。若 cache 采用直接映射方式，则当 s=64 和 s=63 时，缺失率分别为多少？若 cache 采用 2-路组相联映射方式，则当 s=64 和 s=63 时，缺失率又分别为多少？

```
int i, j, c, s, a[128];
```