第一题：

直接相联 L1 Cache（容量多于 4 个块），块 A，B，C 和 D 均映射到同一组中。处理器按如下地址流访问 L1 Cache：

A B B A C B C D A

1. 根据 3C 模型填写下表

	A	B	B	A	C	B	C	D	A
命中									
强制缺失									
容量缺失									
冲突缺失									

2. 假设为该 L1 Cache 配置了一个 Victim Cache，且 Victim Cache 只能容纳 1 块。那么将有多少次读请求被发往 L2 Cache?

3. 假设为该 L1 Cache 配置了一个 Victim Cache，且 Victim Cache 能容纳 N 块（如果 N>1，将采用 LRU 替换策略）。那么使发往 L2 Cache 的读请求次数最少的 N 的取值是多少？此时发往 L2 Cache 的最少读请求次数是多少?

答案：

1.

	A	B	B	A	C	B	C	D	A
Hit			*						
Compulsory miss	*	*			*			*	
Capacity miss									
Conflict miss				*		*	*		*

2. 6 次

	L1 set	VC
A: miss, L2 read	A	
B: miss, L2 read	B	A
B: hit	B	A
A: hit (swap)	A	B
C: miss, L2 read	C	A
B: miss, L2 read	B	C
C: hit (swap)	C	B
D: miss, L2 read	D	C
A: miss, L2 read	A	D

3. N = 3，最少向 L2 Cache 发出 4 次读请求（全部是强制缺失引发的）

下图括号中表示的是 LRU 的计数值

	L1 set	VC		
A: miss, L2 read	A	(0)	(1)	(2)
B: miss, L2 read	B	(1)	(2)	A (0)
B: hit	B	(1)	(2)	A (0)
A: hit (swap)	A	(1)	(2)	B (0)
C: miss, L2 read	C	(2)	A (0)	B (1)
B: hit (swap)	B	(2)	A (1)	C (0)
C: hit (swap)	C	(2)	A (1)	B (0)
D: miss, L2 read	D	C (0)	A (2)	B (1)
A: hit (swap)	A	C (1)	D (0)	B (2)

第二题：

假设某存储系统由 L1 和 L2 两级 Cache 构成，其中 L1 的参数为直接映射、容量为 256B、块大小为 16B、命中时间为 1 个时钟周期，L2 的参数为直接映射、容量为 4KB、块大小为 256B、命中时间为 10 个时钟周期。访问该存储系统的地址流如下：

r 0000FC00

r 0000FC20

r 0000FC40

r 0000FC60

r 0000FC80

r 0000FCA0

r 0000FCC0

r 0000FCE0

r 0000FD00

r 0000FD20

假设 L1 和 L2 Cache 初始状态均为 0，从主存中读取一个 256B 的块需要 100 个时钟周期，将上述地址流循环执行 10 次，请回答下列问题

1. 请填写下表，其中第一列表示读次数，中间 4 列是不同类型的读缺失次数

	reads	read misses				miss rate
		Compulsory	Capacity	Conflict	Total	
L1 cache:						
L2 cache:						

2.上述地址流循环执行 10 次的存储器平均访问时间 MAAT 是多少?

答案:

1.

KEY:

M(c1): compulsory miss
M(c2): conflict miss
H = hit

	L1 cache				L2 cache			
	tag	**index**	**1st iter: miss?**	**other iter: miss?**	**tag**	**index**	**1st iter: miss?**	**other iter: miss?**
0000FC00	0000FC	0	M(c1)	M(c2)	0000F	C	M(c1)	H
0000FC20	0000FC	2	M(c1)	M(c2)	0000F	C	H	H
0000FC40	0000FC	4	M(c1)	H	0000F	C	H	
0000FC60	0000FC	6	M(c1)	H	0000F	C	H	
0000FC80	0000FC	8	M(c1)	H	0000F	C	H	
0000FCA0	0000FC	A	M(c1)	H	0000F	C	H	
0000FCC0	0000FC	C	M(c1)	H	0000F	C	H	
0000FCE0	0000FC	E	M(c1)	H	0000F	C	H	
0000FD00	0000FD	0	M(c1)	M(c2)	0000F	D	M(c1)	H
0000FD20	0000FD	2	M(c1)	M(c2)	0000F	D	H	H

	reads	read misses				miss rate
		Compulsory	Capacity	Conflict	Total	
L1 cache:	100	10	0	36	46	0.46
L2 cache:	46	2	0	0	2	2/46 = 0.043

2.

AAT with L2 cache $= HT_{L1} + MR_{L1} * [HT_{L2} + MR_{L2} * miss\_penalty]$
= (1 cycle) + (0.46) * [(10 cycles) + (2/46)*(100 cycles)]
= (1 cycle) + (46/100)*(10 cycles) + (46/100)*(2/46)*(100 cycles)
= (1 cycle) + (4.6 cycles) + (2 cycles)
= **7.6 cycles**

第三题：

变换下面的循环以充分发挥 Cache 的性能优势，并解释变换后程序性能改善多少?

假设变量 x，k 和 i 均被分配到寄存器中；每个元素 a[i]占 4 个字节；Cache 采用直接相联，容量 32B，块大小 4B

```
x = 0;
for (k = 1; k < 10; k++)
    for (i = 0; i < 128; i++)
```

x += k*a[i];

答案：

对于未变换的循环，Cache 的访问情况如下

a[0]-miss，a[1]-miss，a[2]-miss，…，a[127]-miss，a[0]-miss，a[1]-miss，

a[2]-miss，…，a[127]-miss . . .

缺失率 100%

变换后的代码为：

x = 0;

for (i = 0; i < 128; i++)

for (k = 1; k < 10; k++)

x += k*a[i];

此时访存地址流为：a[0], a[0], a[0], ..., a[1], a[1], a[1], ..., a[127], a[127], a[127],

此时，对于同样的 Cache，其访问情况为：

a[0] – miss; a[0] – hit; a[0] – hit; ..., a[0]—hit;

a[1] – miss; a[1] – hit; a[1] – hit; ..., a[1]—hit;

……

a[127]– miss; a[127] – hit; a[127] – hit; ..., a[127]—hit;

因此，缺失率为 1/9（11.1%）

第四题：

通常来说，cache 命中时间与 cache 容量成正比。假设访问主存需要 70ns，并且在所有指令中，有 36%的指令需要访存。下表是 P1 和 P2 两个处理器各自的 L1 cache 的参数。

	L1 Cache 容量	L1 Cache 缺失率	L1 Cache 命中时间
P1	2KB	8%	0.66ns
P2	4KB	6%	0.90ns

1. 假定 L1 cache 的命中时间决定了 P1 和 P2 的时钟周期，它们各自的时钟频率是多少?

2. P1 和 P2 各自的 AMAT (平均存储器访问时间）分别是多少?

3. 假定在没有任何存储器阻塞时基本的 CPI 为 1. 0, P1 和 P2 各自的总 CPI 分别是多少？哪个处理器更快?

对于下面上个问题，我们考虑在 P1 中增加 L2 cache，以弥补 L1 cache 容量的限制。在解决这些问题时，依然使用上表中 L1 cache 的容量和命中时间。表中二级 cache 的缺失率指的是它的局部缺失率。

L2 Cache 容量	L2 Cache 缺失率	L2 Cache 命中时间
1MB	95%（局部缺失率）	5.62ns

4. 增加 L2 cache 后，P1 的 AMAT 是多少？有了 L2 cache, AMAT 是更好还是更差了?

5. 假定在没有任何存储器阻塞时基本的 CPI 为 1.0, 增加 L2 cache 后，P1 的总的 CPI 是多少?

答案:

1. P1 的主频 = 1/P1 的命中时间 = 1.53GHz；P2 的主频 = 1/P2 的命中时间 = 1.11GHz

2. AMAT = 命中时间+缺失率×缺失代价

P1 的 AMAT = 0.66+0.08×70=6.26ns

P2 的 AMAT=0.9×0.06×70=5.1ns

3. CPI = $CPI_{无阻塞}+CPI_{阻塞}$

P1 的 CPI = 1+(1+0.36)×0.08×70ns/0.66ns = 12.54

P2 的 CPI = 1+(1+0.36)×0.06×70ns/0.9ns = 7.35

P2 处理器更快一些

4. P1 的 AMAT=0.66+0.08×(5.62+0.95*70) = 6.6ns > 6.26，AMAT 变差了

5. P1 的 CPI = 1+(1+0.36)×0.08×(5.62ns/0.66ns+0.95×70ns/0.66ns) = 12.9