1、解：

页大小为 4KB=2^{12}B，页内地址为[11:0]位。

cache 容量 64KB=2^{16}B，地址范围为[15:0]；cache 块大小为 32B=2^{5}B，地址范围为[4:0]。

1）直接相联：

cache 索引位数为地址的[15:5]，需要页着色的是地址[15:12]，共 4 位；

2）二路组相联：

cache 索引位数为地址的[14:5]，需要页着色的是地址[14:12]，共 3 位；

3）四路组相联：

cache 索引位数为地址的[13:5]，需要页着色的是地址[13:12]，共 2 位。

2、解：

$MissPenalty_{L1}=HitTime_{L2} + MissPenalty_{L2} \times MissRate_{L2}$

1）直接相联：

$MissPenalty_{L1}=4 + 60 \times 25\% = 19$ 个时钟周期

2）2 路组相联：

$MissPenalty_{L1}=5 + 60 \times 20\% = 17$ 个时钟周期

3）4 路组相联：

$MissPenalty_{L1}=6 + 60 \times 15\% = 15$ 个时钟周期

3、解：

1）MissRate = CacheMissOps/MemOps

每 1000 条指令中 load/store 指令的条数是 1000×(26%/74%)=351

32KB 指令 cache MissRate 为 0.0015；

32KB 数据 cache MissRate 为 38/351=0.1083

指令 cache 和数据 cache 各 32KB 的 MissRate = (1.5 + 38)/(1000+351) = 0.0292

64KB 一体 cache 的 MissRate = 40/(1000+351) = 0.0296

所以指令 cache 和数据 cache 各 32KB 组织方式缺失率更低

2）

指令 cache 和数据 cache 各 32KB 的 AMAT = (1000+1.5×100+38×100)/(1000+351)=3.66

64KB 一体 cache 的 AMAT = (1000+351+40×100)/(1000+351)=3.96

4、解：

地址访问序列（单位：字）	**访问类型**	
	直接相联	**2 路组相联**
0	强制缺失	强制缺失
1	命中	命中
2	命中	强制缺失
3	命中	命中
4	强制缺失	强制缺失
15	强制缺失	强制缺失
14	命中	命中
13	命中	强制缺失
12	命中	命中
11	强制缺失	强制缺失

10	命中	命中
9	命中	强制缺失
0	命中	命中
1	命中	命中
2	命中	命中
3	命中	命中
4	命中	命中
56	强制缺失	强制缺失
28	强制缺失	强制缺失
32	强制缺失	强制缺失
15	容量缺失	命中
14	命中	命中
13	命中	冲突缺失
12	命中	命中
0	容量缺失	冲突缺失
1	命中	命中
2	命中	命中
3	命中	命中

5、答：

包含式 cache 关系中一级 cache 的内容是二级 cache 内容的真子集；非包含式 cache 关系中一级 cache 的内容与二级 cache 内容无交集。

1）一级 cache 缺失后查找二级 cache：

包含式 cache 关系中，若二级 cache 命中则直接返回，若二级 cache 也缺失，则从更低层次的存储中取回所需数据，填入二级 cache 并返回至一级 cache。在此过程中，二级 cache 无论命中与否均存在一次读访问过程，若不命中，则存在一次替换操作（一次读访问过程读出被替换 cache 块，一次写访问过程写入回填 cache 块）。

非包含式 cache 关系中，若二级 cache 命中则在将命中的 cache 块返回给一级 cache 的同时将该 cache 块从二级 cache 中无效，若二级 cache 也缺失，则从更低层次的存储中取回所需数据，直接返回至一级 cache。在此过程中，二级 cache 无论命中与否均存在一次读访问过程，二级 cache 命中时存在一次写访问过程，不命中时则无其它访问过程。

2）一级 cache 替换出的 cache 块：

包含式 cache 关系中，替换出的 cache 块若不是脏块，则数据不用写回二级 cache，若是脏块，则直接写回二级 cache 的对应位置。在此过程中，若替换出的不是脏块，则一级 cache 只需要将替换出的地址通知一致性管理部件即可，一级 cache 与二级 cache 间无数据通信发生。

非包含式 cache 关系中，替换出的 cache 块无论是否是脏块，都要写入二级 cache。在将此 cache 块写入二级 cache 时，可能先要替换出一个 cache 块。在此过程中，一级 cache 与二级 cache 间有数据通信发生。

3）硬件实现复杂性方面，略。

4）多核之间一致性维护：（此处考虑二级 cache 被多个处理器核共享的情况）

如果采用基于目录的协议：

包含式 cache 关系中，因所有一级 cache 的内容均在二级 cache 中有备份（可能数据是旧值），所以二级 cache 处可以获得足够的信息来维护多核间的 cache 一致性。而在非包含式 cache

关系中，一级和二级 cache 上发生的所有 cache 块操作的事件的消息都要传递给目录进行维护。
如果采用基于侦听的协议：
包含式 cache 关系中，每个核向外广播的无效请求或更新请求也要传递给二级 cache。而在非包含式 cache 关系中，每个核向外广播的无效请求或更新请求则不必传递给二级 cache。

6、答：
1）几乎没有空间局部性和时间局部性：数据库查询操作中被查询的数据
2）较好的空间局部性，几乎没有时间局部性：音视频编解码等流式应用被处理的数据
3）较好的时间局部性，很少的空间局部性：以指针链表、树、图等数据结构开发的应用程序。
4）较好的空间和时间局部性的应用：针对 cache 优化后的矩阵乘法、大量的科学计算、压缩解压缩、加解密、文字处理

7、答：
降低 cache 缺失代价的技术：1）读优先；2）关键字优先；3）写合并；4）Victim Cache；5）多级 cache。（降低原因参考教材 P191~192，略）
降低 cache 缺失率的技术：1）增加 cache 容量；2）增加 cache 块大小；3）提高 cache 相联度；4）软件优化。（降低原因参考教材 P188~189，略）

8、答：
1） 直接相联：每个内存块只能映射到 cache 的唯一位置；
全相联：每个内存块可以映射到 cache 的任一位置；
组相联：每个内存块对应 cache 的唯一的一个组，但可以映射到组内的任一位置。
2） 直接相联：根据访问地址将唯一映射位置上的 cache 块读出，将地址高位与 tag 比较看是否相等；
全相联：将 cache 中所有 cache 块读出，将地址高位与每一项的 tag 比较，看哪一项相等；
组相联：根据访问地址将唯一映射的组内的所有 cache 块读出，将地址高位与每一项的 tag 比较，看哪一项相等。
3）LRU、Random、FIFO。
4）写穿透策略和写回策略；写分配策略和写不分配策略。