6. 因为组索引越靠前的位数变化的范围越小，可以更好地利用局部性原理，提高缓存的命中率。而标签则需要存储的信息较多，需要更多位数来表示。

7. 使地址的组索引和块内偏移的总位数与虚拟内存系统的页偏移位数相同，可以使缓存系统的映射方式与虚拟内存系统的映射方式相同，方便管理与协调。

8 1) 存储系统平均访问延时 = 缓存命中时间 + 缓存缺失时间 X 缓存缺失率 = 1 + 110 X 3% + 105 X (1-3%) = 4.05个周期

2) 每组8个块64字节，64kB的L1缓存可以缓存1024个块，可以覆盖整个数组。由于完全随机访问，每个块的访问概率相等，因此平均访问延时 = 1 + 105 X 3% = 4.15个周期

3) 局部性原理好坏直接影响了缓存的命中率和缺失率，进而影响处理器访存的性能。当访问随机性很高时，缓存的效果就不如预测，当程序的访问具有局部性时，缓存会取得更好的效果。

4) 程序的平均缓存命中率需要高于3%，才能让存储系统在使用L1时获得性能收益。

9.

	组数量	组索引位数 Bit	标签位数 Bit	偏移位数 Bit
1	32	5	21	6
2	8	3	23	6
3	1	0	26	6
4	256	8	18	6
5	64	6	19	7
6	256	8	18	6
7	64	6	20	6
8	32	5	20	7

10(1) 系统A的平均内存访问时间 = 缓存缺失时间 X 缓存缺失率 + 缓存缺失代价 X 缓存缺失率 + 缓存命中时间 = 0.22 + 100p1 + 110p2 > 0.52 + 100p2，即 p1 > 5p2 - 1.1

(2) 系统A的平均内存访问时间 = 缓存命中时间 + 缓存缺失时间 X 缓存缺失率 + 缓存缺失代价 X 缓存缺失率 x_k = 0.22 + 0.52k + 100p1k + 110p2k，而系统B的平均内存访问时间为 0.52 + 0.52k + 100p2k，若前两者相等，则 p1 > 1.75p2 - 0.3

11. 直接映射，缓存容量为 16 X 64 ÷ 8 = 128字节。
2路组相联，缓存容量为 2X16 X64 ÷ 8 = 256字节，块替换次数为1。
4路组相联，缓存容量为 4X16 X64 ÷ 8 = 512字节，块替换次数为0。
同理8路组相联，缓存容量为1024字节，块替换次数为0。

12. 缓存A为2路组相联，总共可存放8个块，缓存B为直接映射，总共可存放16个块。
优化前，数组A在缓存A中的缺失次数为640，缓存B中的缺失次数为1600。
优化后，数组A在缓存A中的缺失次数为128，缓存B中的缺失次数为2510。

13. 将内层循环中访问A数组的方式修改为按照行优先，即将A[j][i]修改为A[i][j]，可以更好地利用行主元局部性，减少缓存缺失次数。

14.(1) 优化前，总共发生256次缓存缺失；
优化后，缓存缺失次数减少到128次，因为每次循环只需要缓存一行或一列(16字节)，因为块大小为32字节，所以每个块可以缓存2个元素，故需128个块。

(2) 优化前，总共发生256次缓存缺失；
优化后，缓存缺失次数减少到0次；因为全相联缓存可以缓存所有数据，不会发生替代。

(3) 对于块大小32字节的直接映射缓存，每个块可以包含2个元素；缓存所需块数为 16X4=64个，故优化前后均需64块。

15

	input数组				output数组			
	列0	列1	列2	列3	列0	列1	列2	列3
行0	miss	miss	miss	miss	miss	miss	miss	miss
行1	hit	hit	miss	miss	hit	hit	miss	miss
行2	hit	hit	miss	miss	hit	hit	miss	miss
行3	hit	hit	miss	miss	hit	hit	miss	miss