第6章 存储器层次架构

> 存储技术：

- DRAM和SRAM(断电丢失，Main Memory用的是DRAM)、ROM（断电不丢失，flash memory基于EEPROM，用于手机、SSD硬盘等）
- 磁盘存储
  - 构造：platter->surface->track->sector->bytes
  - 时间计算：
    - access time = seek time + rotational latency + transfer time
    - T/avg seek = 3~9ms
    - T/avg rotation = 1/RPM X 60s/1min (RPM- Revolution Per Minute)
    - T/avg transfer = 1/RPM X 60s/1min X 1/(average#sectors/track)
  - DMA(Direct Memory Access)
- SSD
- 速度 SRAM > DRAM > Disk
- 价格 SRAM > DRAM > SSD > 旋转磁盘

> 局部性（Locality）：

- 时间局部性：重复引用相同变量的程序
- 空间局部性：步长越小越好
- 对于指令：循环体越小，局部性越好

> 层次结构（Hierarchy）：

- Regs -> L1 cache(SRAM) -> L2 cache(SRAM) -> L3 cache(SRAM) -> Main Memory(DRAM) -> Local disks -> Remote(distributed file systems, Web serves)

> Cache Memories：

- General organization of cache (S, E, B, m).
- C = S x E x B (S=2^s B=2^b t=m-(s+b)
- cache: S set -> E line -> 1 valid bit & t tag bits & B bytes
- address: t bits Tag & s bits Set index & b bits Block offset
- Type:
  - Direct-Mapped Caches(E=1)
  - Set Associative Caches
  - Fully Associative Caches(E=C/B)
- write hit: write-through v.s. write-back
- write miss: write-allocate v.s. not-write-allocate
- i-cache d-cache / unified cache
- Performance:
  - miss rate
  - hit rate
  - hit time
  - miss penalty

Cache-Friendly Code

- Example:

```
for(i=0; i<M; i++)
  for(j=0; j<N; j++)
    sum += a[i][j];
```

```
for(j=0; j<M; j++)
  for(i=0; i<N; i++)
    sum += a[i][j];
```

Memory mountain