The Memory Hierarchy

CS520

Dept. of Computer Science
Univ. of New Hampshire

create illusion of very large memory

that can be accessed with high speed

why illusion?

because high-speed (i.e. low access time)

memory is expensive

technology	access time	cost
static RAM	1x	100x
dynamic RAM	10x	1x
disk	1Mx	.01x

How can this possibly work?

programs exhibit locality

they access a relatively small portion of their address space in any small interval of time

temporal locality

if a location is referenced, it will tend to be referenced again soon

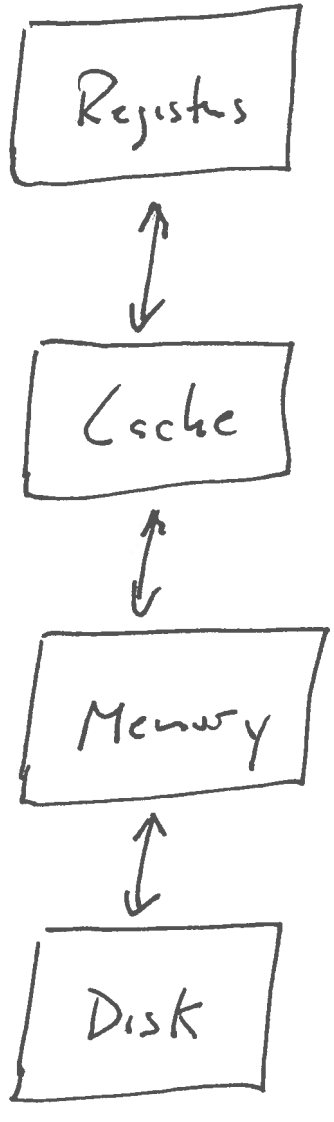
spatial locality

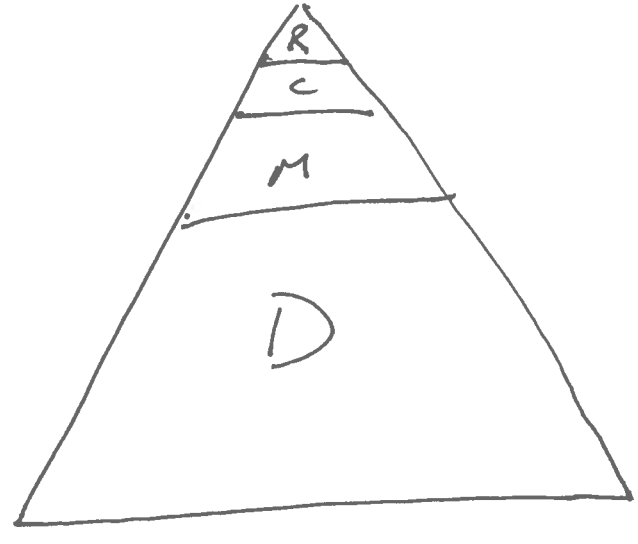
if a location is referenced, locations whose addresses are close by will tend to be referenced soon

```
for (i = 0; i < N; i++) {
    sum = sum + a[i];
}
```

So

place locations in a hierarchy in which least frequently accessed locations are lowest in the hierarchy and most frequently accessed locations are highest in the hierarchy


Registers
Cache
Memory
Disk


R
C
M
D

spatial locality causes systems to treat

blocks of contiguous memory locations

as a single unit

hit rate

fraction of accesses satisfied at an upper level

miss penalty

time to satisfy access that misses at upper level

Key problems

determining where a location is in the hierarchy

moving locations up and down in the hierarchy