Notes for 10/14/09

Cost ↑ Speed ↑ Size ↓	Registers
	Cache
	DRAM
	HDD
	Magnetic Disk

Caching—create illusion that our memory is cheap, fast, and large, i.e., have our cake and eat it too.

Caching works because of spatial and temporal locality, i.e., because you tend to access nearby memory addresses around the same time.

There is spatial and temporal locality of instructions. Aside from jumps, you tend to execute instructions one after the other. In loops, you also tend to execute instructions nearby in memory around the same time.

There is spatial and temporal locality of data. Consider cases like iterating through an array.

Processors tend to have separate caches for instructions and data—i-cache and d-cache. This is because their caching is best optimized separately because they have different access patterns.

Cache coherency—when you have multiple caches, you can have multiple copies of the same data on different caches, so you want to ensure that you are not operating on stale data in any cache.