

Reading 18 (No Reading 17)

19.1 Cache basics: Part 1

- A processor's chip has limited space, a processor's memory commonly exists on a separate chip (or chips).
 - ↳ Access between chips is slower than within a chip.
 - ↳ So for faster access, a processor may have a small on-chip memory, called a cache.
 - ↳ The cache stores data that will be used frequently. (think of it as a favorites contact list)
 - ↳ cache is usually SRAM.
- Initially, a cache is empty.
 - ↳ For later reads, the system checks the cache first; if an item's copy exists in the cache (a hit), the system quickly reads that copy; else (a miss) the system first copies the item from memory to cache.
- A hit may require only 1 clock cycle, while a miss may require more.
- A direct-mapped cache directly maps memory addresses to cache addresses using a subset of address bits (called an index), storing the remaining the bits (called the tag) in the cache entry.

Processor
access

01000	55			
01110	207			
01000	55	<i>Fast (hit)</i>	Tag	Cache address
01000	55		010	00
10100	99		101	00

Cache addr (Index)	Cache	
	Data	Tag
00	99	101
01		
10	207	011
11		

Data memory

000	00	
00001		
00010		
00011		
001	00	
00101		
00110		
00111		
010	00	55
01001		
01010		
01011		
011	00	
01101		
01110		207
01111		
100	00	
10001		
10010		
10011		
101	00	99
10101		
10110		
10111		
110	00	
11001		
11010		
11011		
111	00	
11101		
11110		42
11111		

• Multiple addresses can map to the same cache address, but only one can be copied into cache at a time.