Direct-Mapped Cache

Cache Capacity

Number of data words a cache can hold

System designer must choose what is kept in cache as it is smaller than what is available in main memory.

Remember: Cost of cache limits what we can store!

Fetching

If data is already in cache, we have a hit. Data is available.

- Otherwise, processor fetches data from main memory and places it in the cache for future use.
- *Key:* Cache must replace old data

Questions:

- What data is held in the cache?
- How is data found?
- What data is replaced to make room for new data when the cache is full?

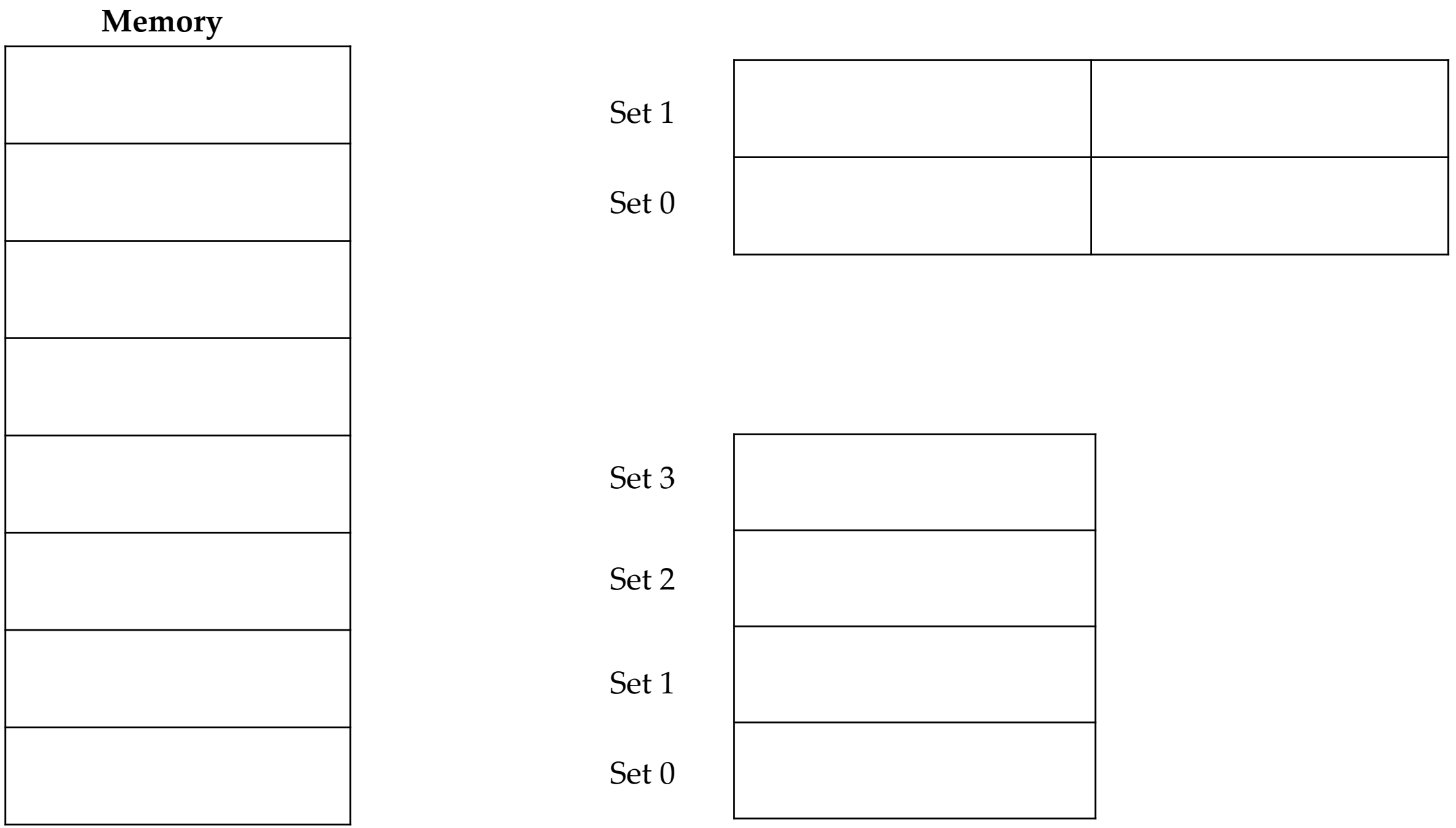
How is Data Found?

Relationship between the address of data in main memory and location of data in the cache is called the mapping.

Cache is organized into *S* sets, each holds one or more blocks of data.

Each memory address maps to exactly one set in the cache

- Some address bits determine which cache set contains data
- If a set holds more than one block, data may be kept in any of the blocks in the set.

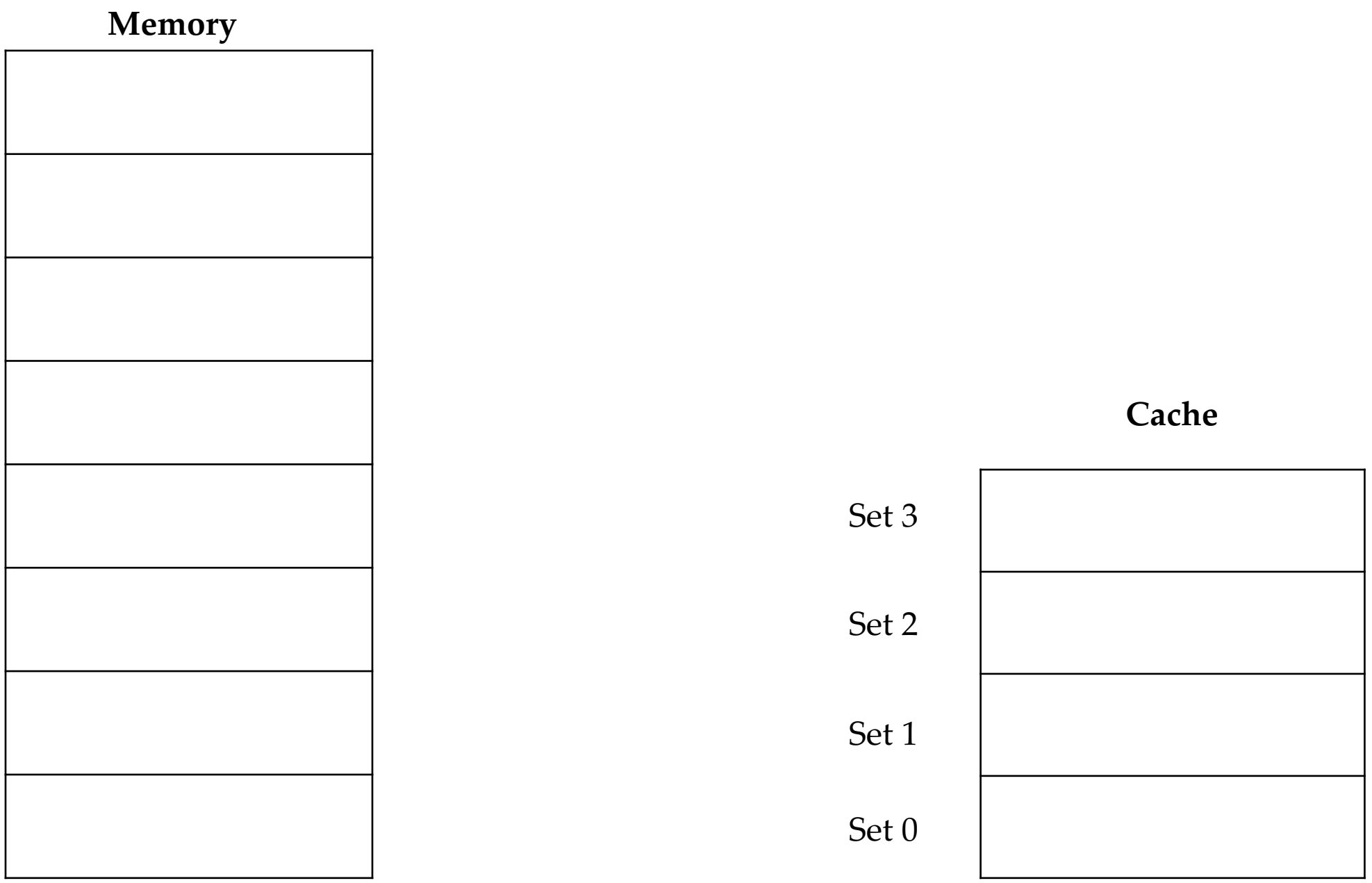

Memory
Set 1
Set 0
Set 3
Set 2
Set 1
Set 0

Cache Specifications

1. Capacity
   How much data can we store?
2. Number of Sets
   How many mappings between cache and main memory addresses?
3. Block Size
   Grab related words around a piece of data added to the cache.
4. Number of Blocks
   How many blocks can we store?
5. Degree of Associativity
   Different types of cache handle organization differently.

Cache Categorization

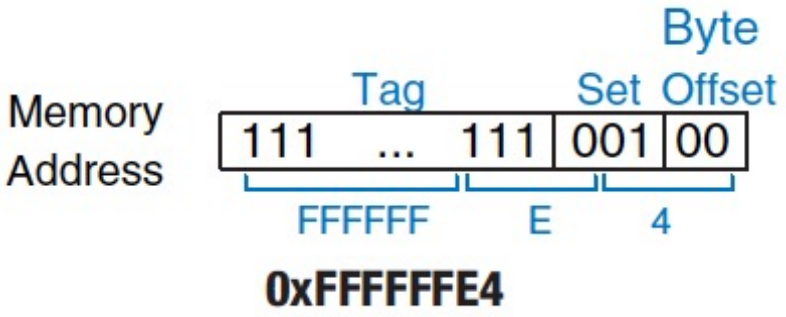
- Direct Mapped: each set contains only one block.
- N-way Set Associative Cache: each set contains N blocks
- Fully Associative Cache: only one set (B-way associative cache)

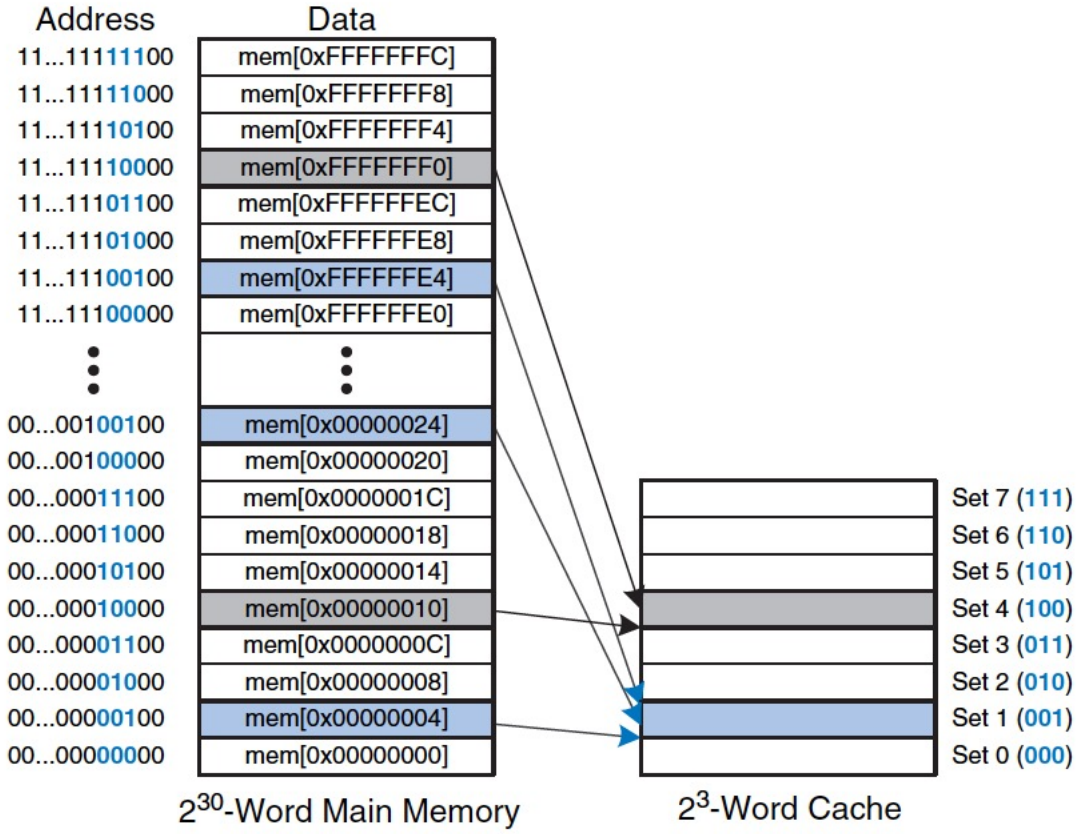

Memory
Cache
Set 3
Set 2
Set 1
Set 0

Direct Mapped Cache

One block in each set.

One-to-one mapping, but there are more blocks in memory than there are in cache.





Hardware View: Eight Entry SRAM

Cache accessed by the 32-bit address.

Load instruction reads the entry from the cache and checks tag and valid bits.

- If tag matches 27 MSB of the address and it is valid
- Cache hits and data is returned
- Otherwise it's a cache miss.