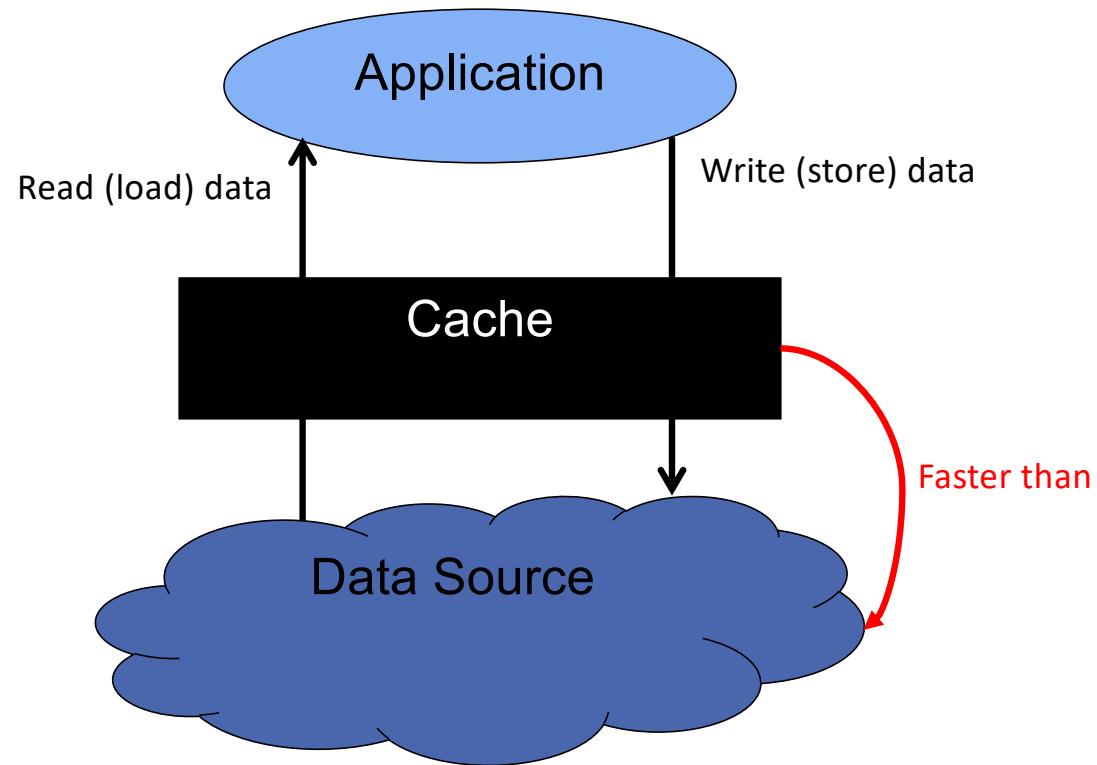# **Today**

- Learning Objectives
  - Define
    - Cache block
    - Cache slot
    - Cache hit/miss
    - Different miss types: Compulsory, capacity, conflict
- How we'll get there
  - Present an abstract model of a cache (applies to both hardware and software)
  - Use that model to introduce some basic terminology and introduce the kinds of design decisions we need to make in constructing caches.
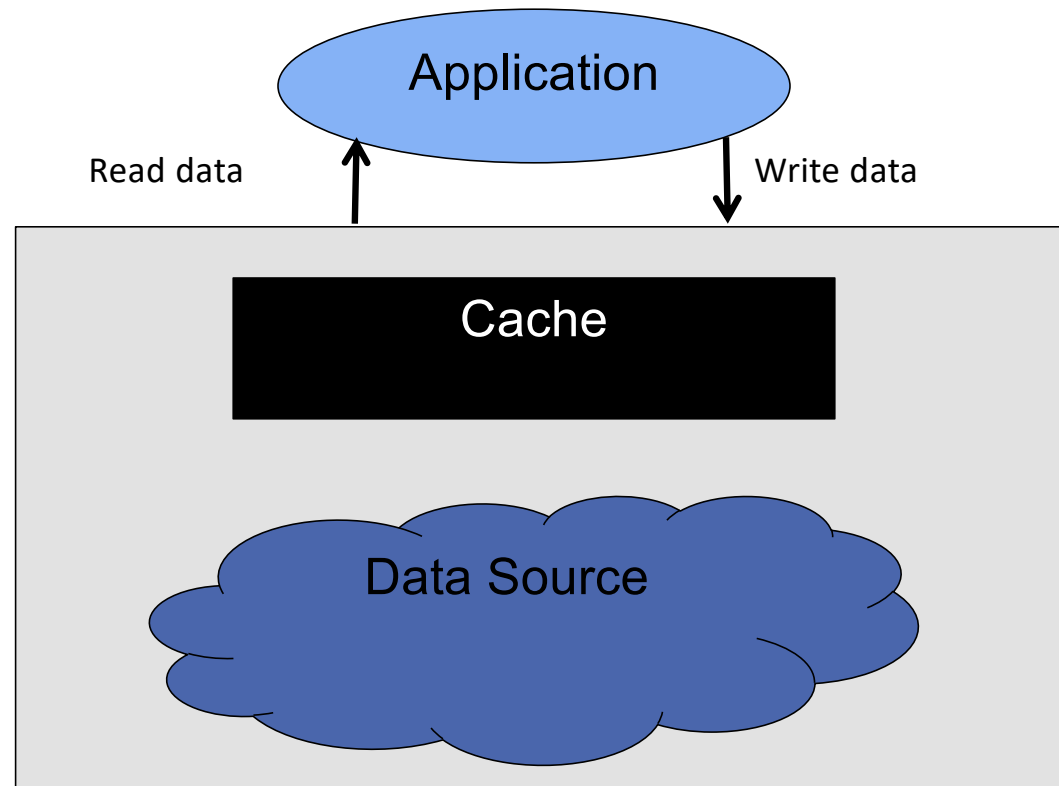- Reading
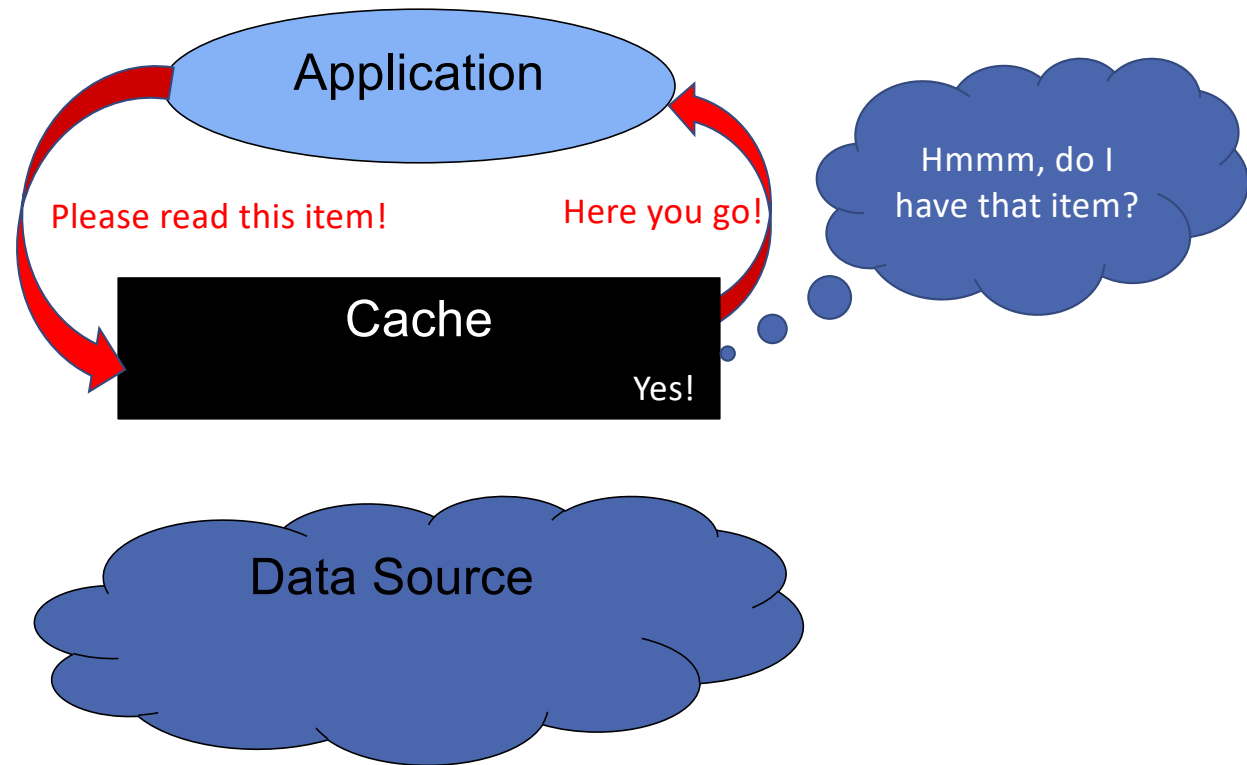 - 6.2, 6.3
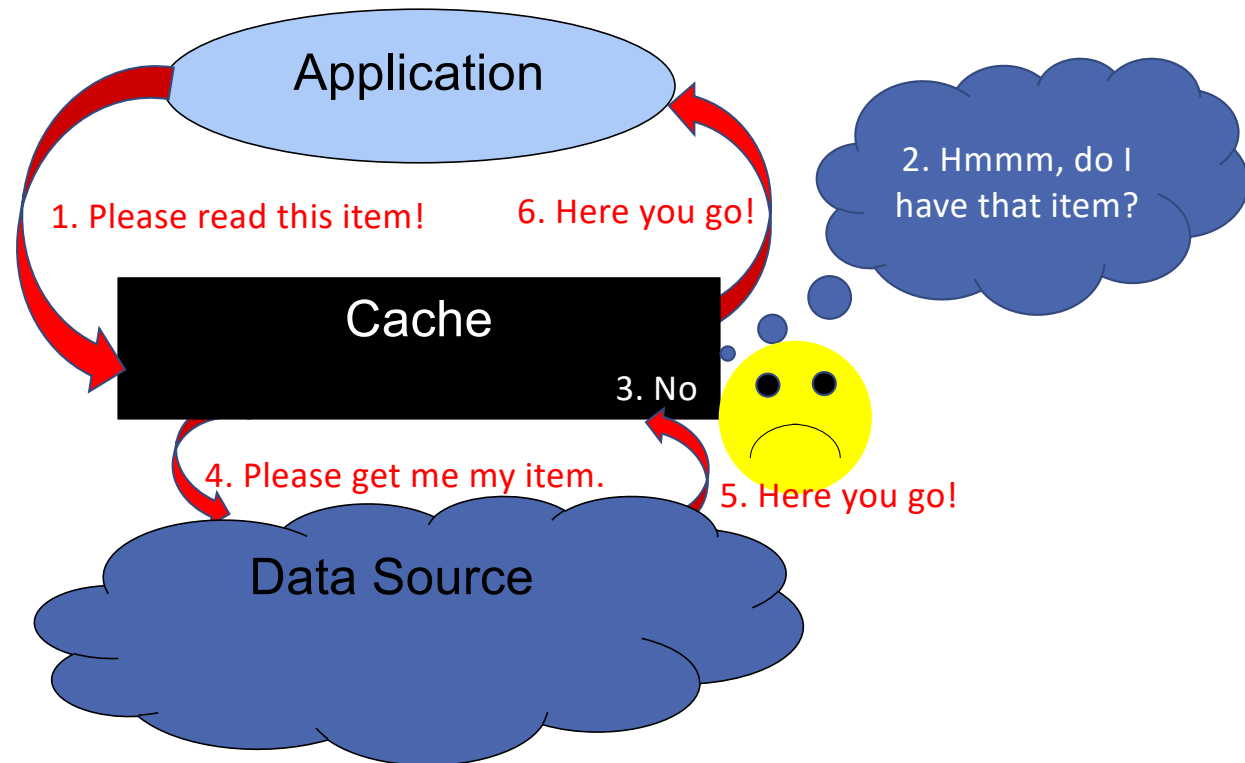
Applications Access Data

Application

Read (load) data

Write (store) data

Data Source
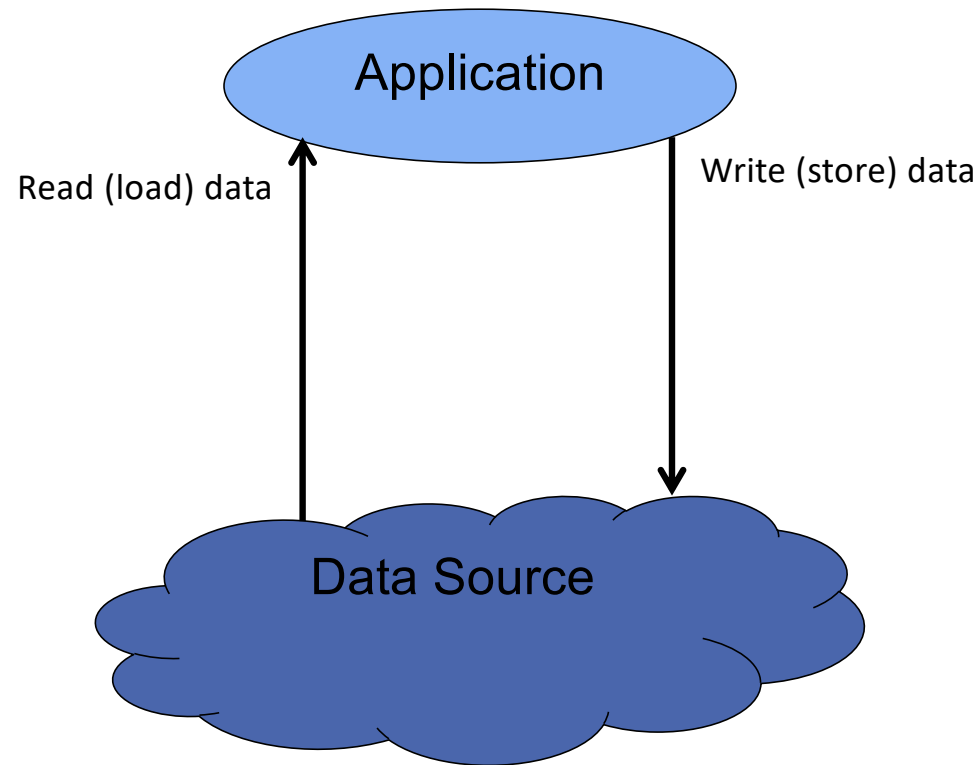
Applications Access Data

An Abstract Cache

Application

Read data

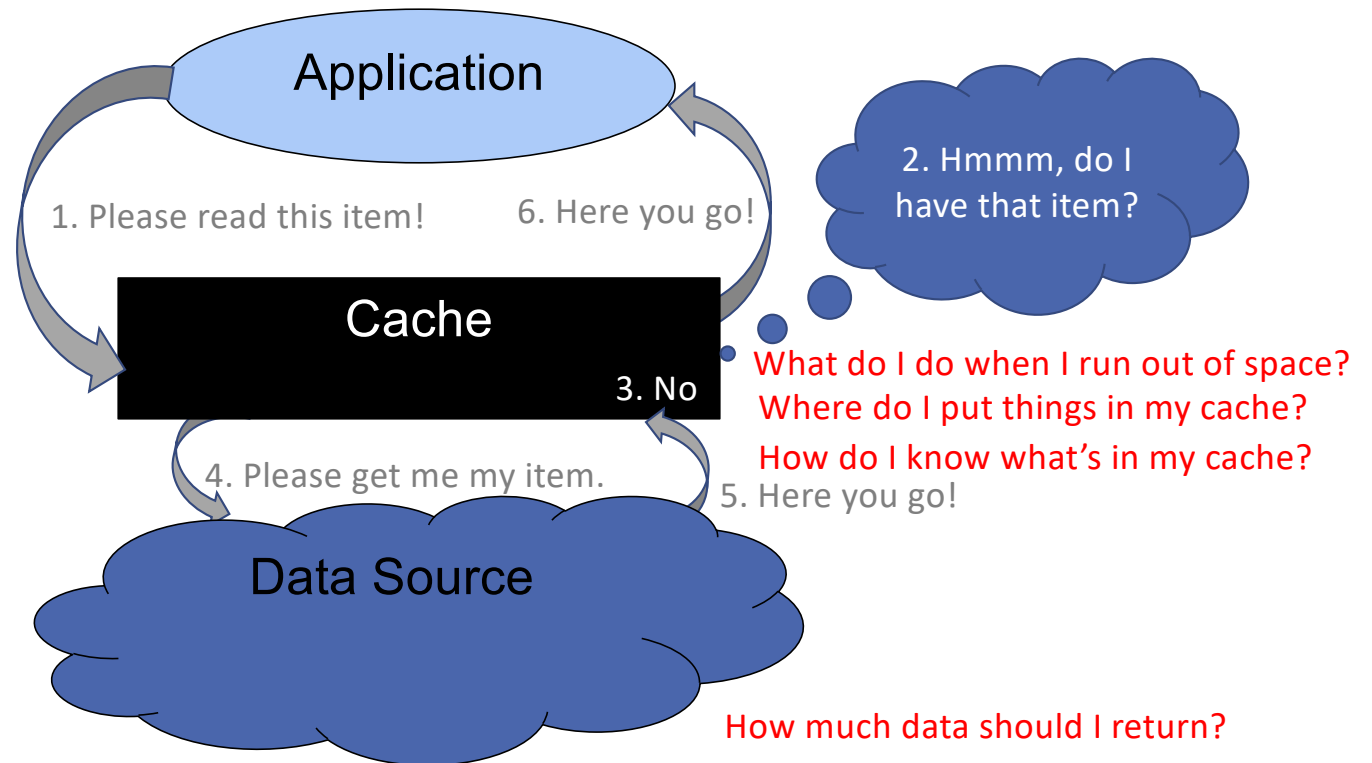Write data

Cache

Data Source

A Read Cache Hit

A Read Cache Miss

Types of Cache Misses

- **Compulsory**:
  - On first access to an object, you take a miss; there is little you can do about it.

- **Capacity**:
  - You are touching more data than can fit in the cache.
  - If your cache were larger, you would have fewer misses.

- **Conflict**:
  - In some (most) caches, there only a limited number of places in the cache in which you can put a particular piece of data.
  - Misses that occur because the particular place (or places) in which a piece of data must go are occupied are called conflict misses.

A Read Cache Miss – Decisions!!

Application

1. Please read this item!

6. Here you go!

2. Hmmm, do I have that item?

Cache

3. No

What do I do when I run out of space?

Where do I put things in my cache?

How do I know what's in my cache?

4. Please get me my item.

5. Here you go!

Data Source

How much data should I return?

Cache Design Decisions

- How much data should I return?   Blocksize or Linesize

- What do I do when I run out of space?   Eviction/Replacement Policy

- How do I know what's in my cache?

- Where do I put things in my cache?   ## Cache Organization

- How do I cache writes?   Cache sets
    Cache indices
    Cache tags

How much data should I return?

- Most (persistent) storage devices have a native data access and/or transmission size, e.g., disk block (4 KB).

- Caches also have a native size.
  - E.g., Broadwell we talked about on Wednesday has 64-byte cache lines.
  - Implication: Data moves in and out of caches in 64-byte chunks.

- Block size: the unit in which data is stored in a cache.
  - Broadwell caches: 64 bytes (block size called cache line size in HW caches)
  - File system caches: 4 KB
  - Object caches: size of the object