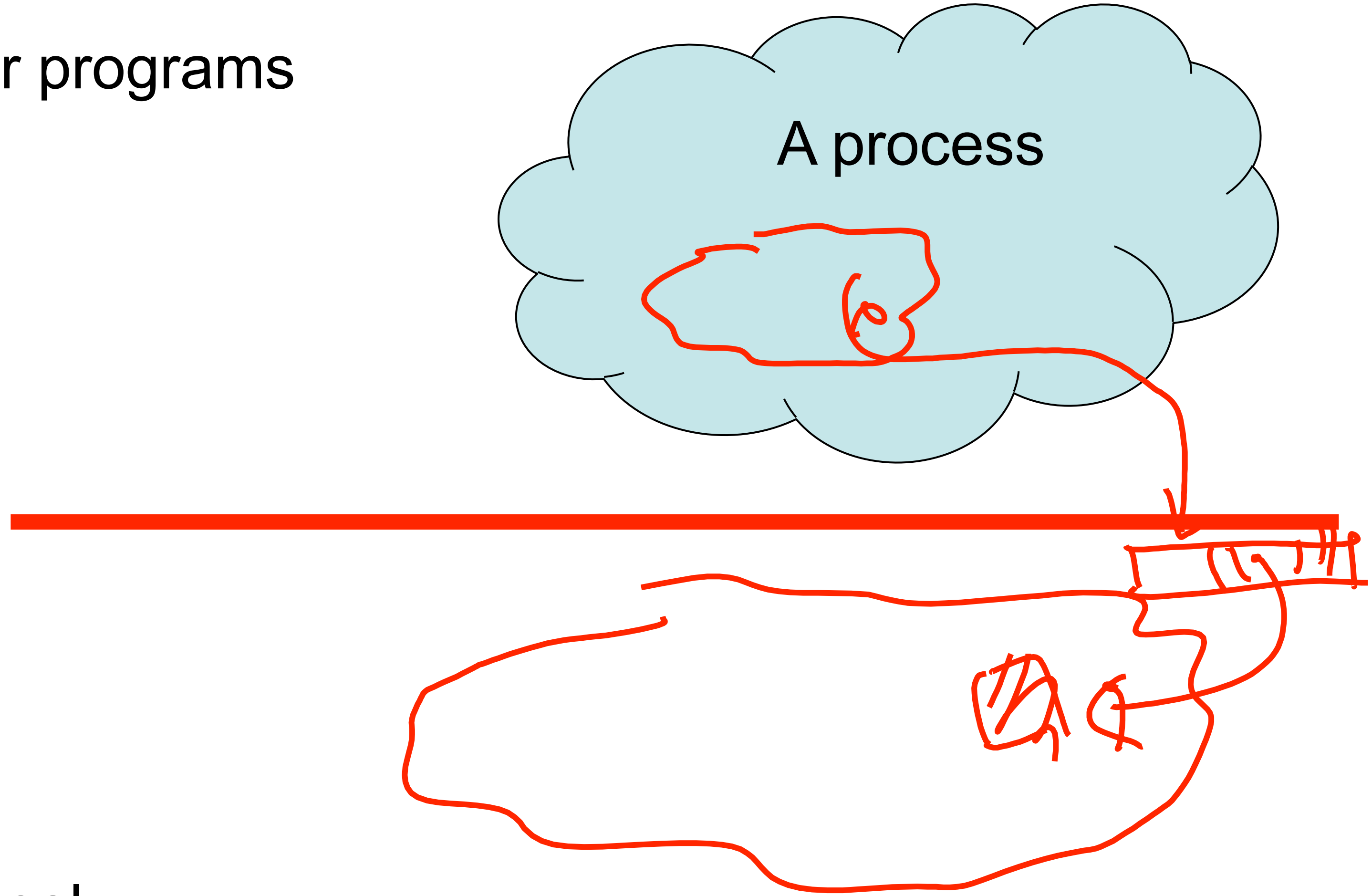
Storage 4: Consistency and Copies

- Learning Objectives
  - Define cache consistency
  - Determine which caches are consistent and which are not
  - Evaluate the efficacy of a cache: Average access time

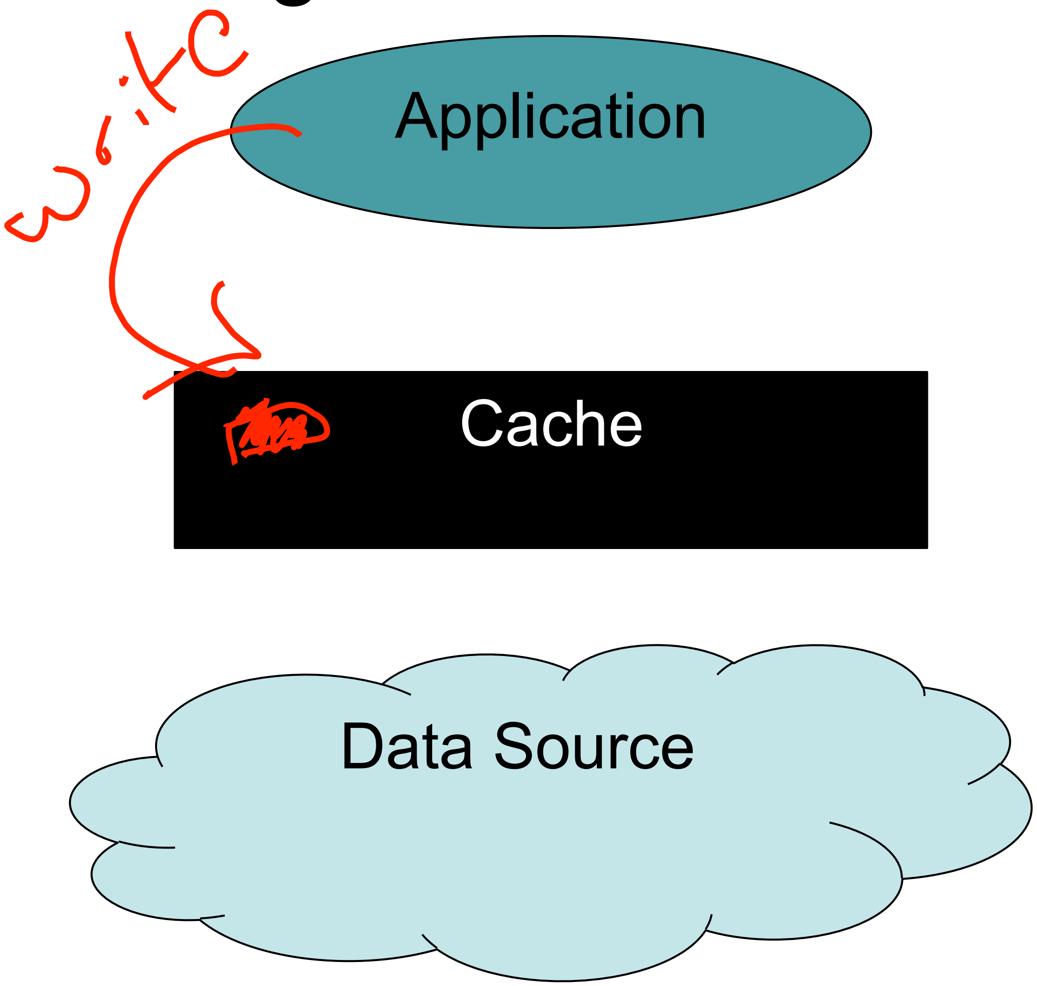
Where do we find caches?

User programs



Kernel

Writing to the cache: HIT



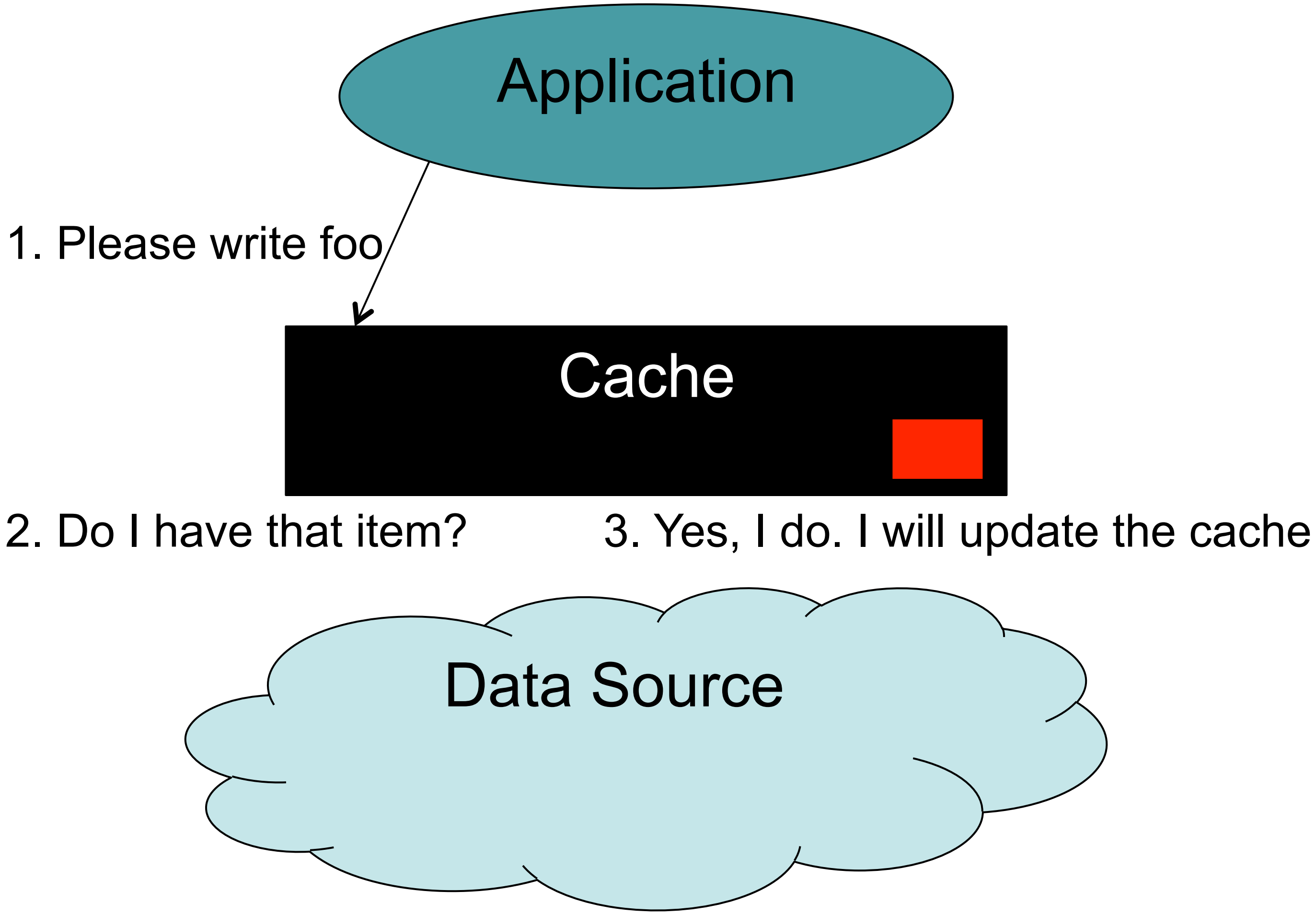
write

Application

Cache

Data Source

Write cache policy: Write Back



Application

1. Please write foo

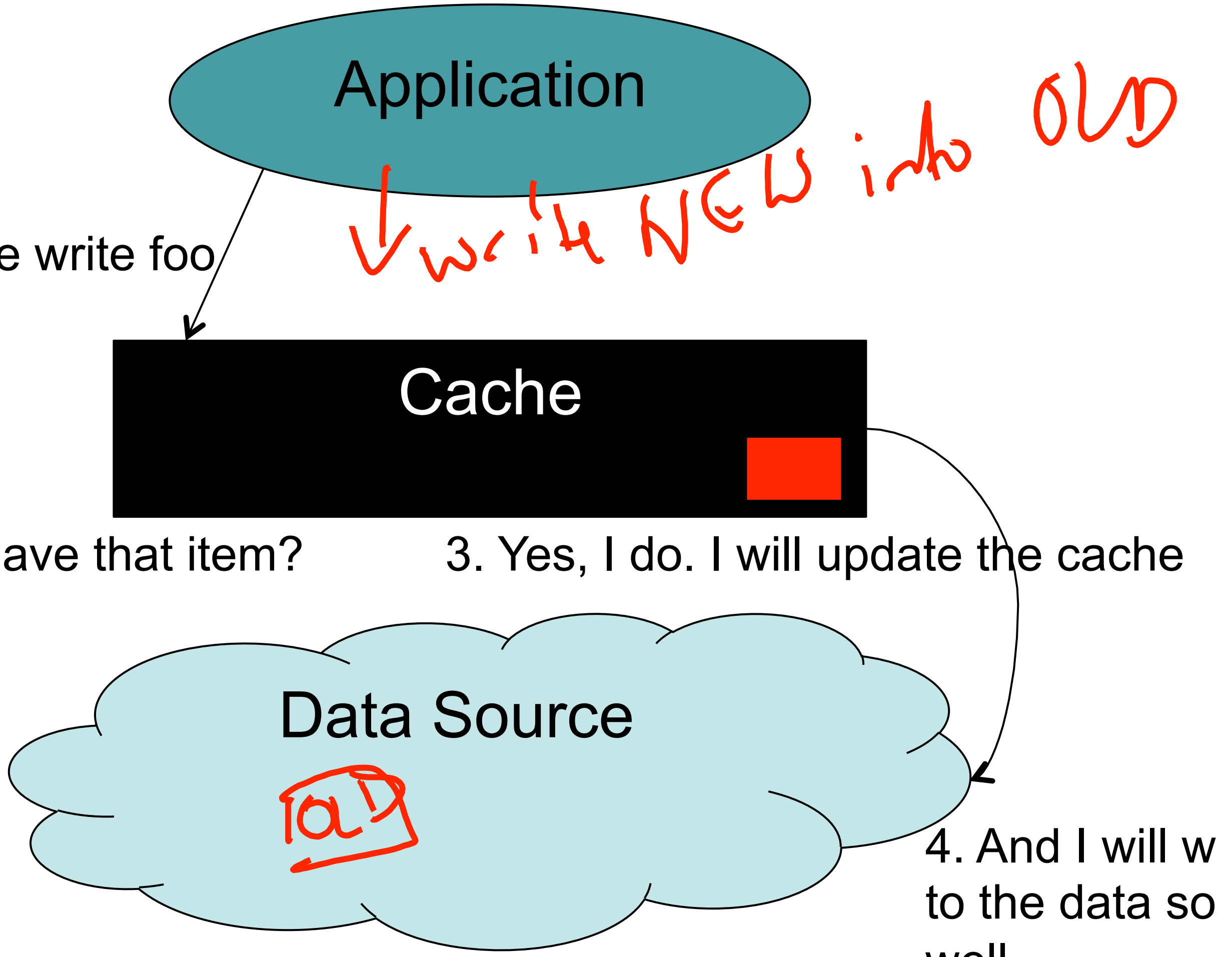
Cache

2. Do I have that item?

3. Yes, I do. I will update the cache

Data Source

Write cache policy: Write Through



write NEW into OLD

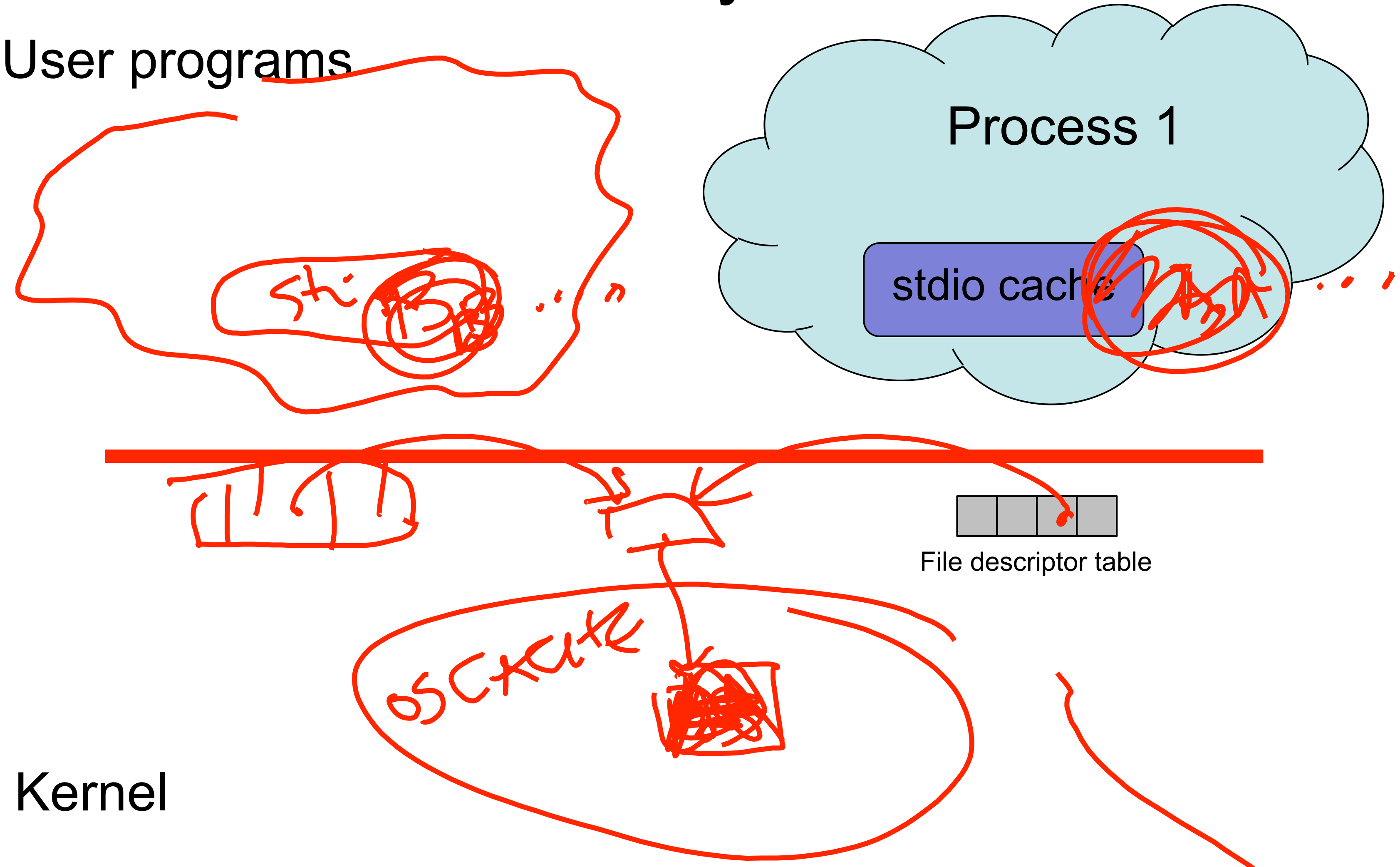
1. Please write foo

2. Do I have that item?

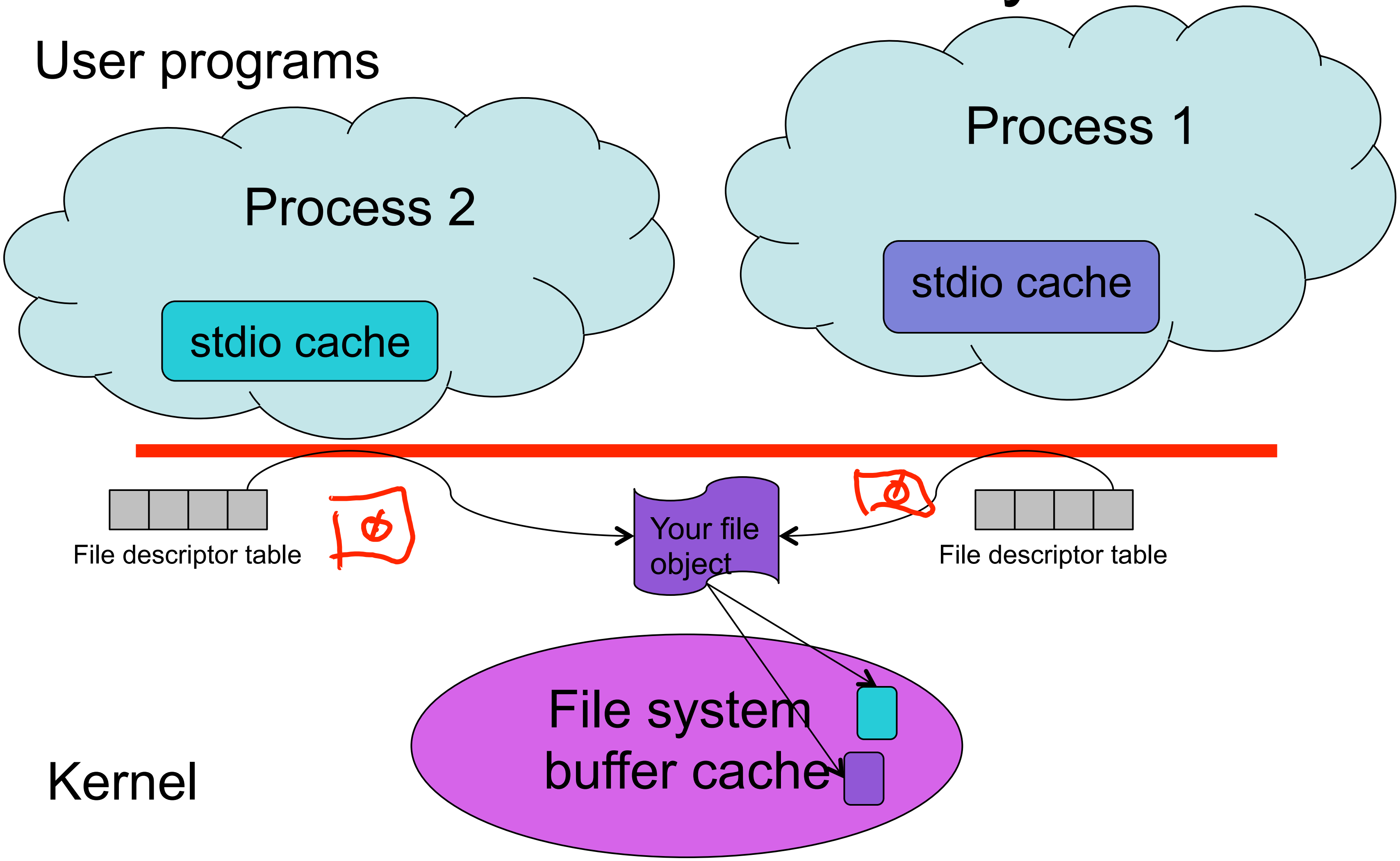
3. Yes, I do. I will update the cache

4. And I will write back to the data source as well.

Inconsistency in Action



Cache Consistency



User programs

Process 2

stdio cache

Process 1

stdio cache

File descriptor table

Your file object

File descriptor table

File system buffer cache

Kernel

Screen Capture

- Both stdio and syscall allow processes to write to files; you can run multiple copies, all of which will write to the same file (each of stdio and syscall write to different files).
- What happens to the output?
- Why?

Evaluating a Cache:

- We showed that we can use hit rate to compare cache replacement algorithms.
- But how do we compare fundamentally different caches?
- Assume it takes 1 time unit to access your cache.
- Which is better:
  - A data source that takes 100 time units, which will produce a 95% hit rate
  - A data source that takes 50 time units, but will produce an 85% hit rate?

Storage 4X

Evaluating a Cache:
Average Access Time

- Assume it takes 1 time unit to access your cache.
- Which is better:
  - A data source that takes 100 time units, which will produce a 95% hit rate
  - A data source that takes 50 time units, but will produce an 85% hit rate?

$95 * 1 + 5 * 100 = 5.95$

$85 * 1 + 15 * 50 = 85 + 7.50$

Evaluating a Cache:
Average Access Time

- Assume it takes 1 time unit to access your cache.
- How good would the 50 time unit cache need to be to produce the same performance as the 100 unit cache?
  - A data source that takes 100 time units, which will produce a 95% hit rate (avg access time = 5.95 units)
  - A data source that takes 50 time units, but will produce an 85% hit rate? (avg access time = 8.35 units)