Computer Architecture

Spring 2016

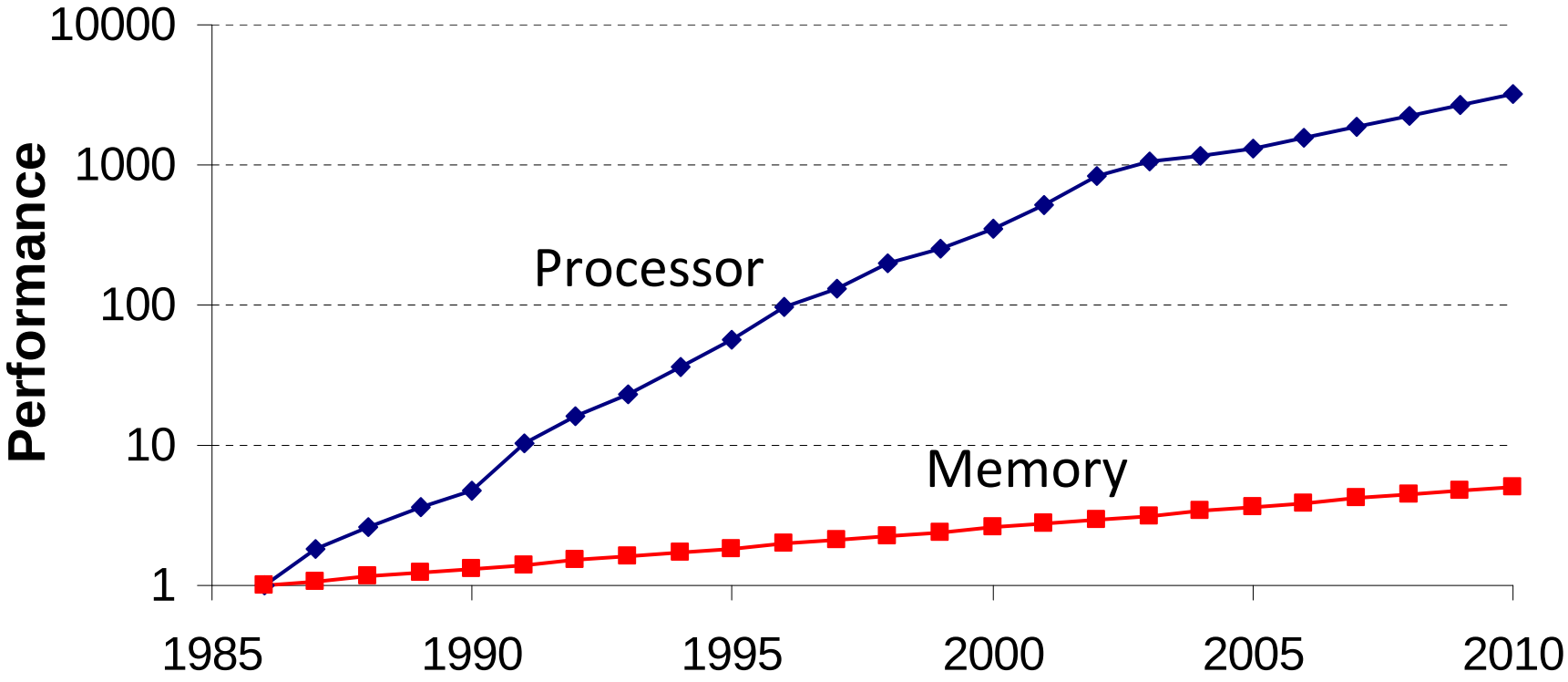
Lecture 06: Memory Hierarchy & Caches I

Shuai Wang

Department of Computer Science and Technology

Nanjing University

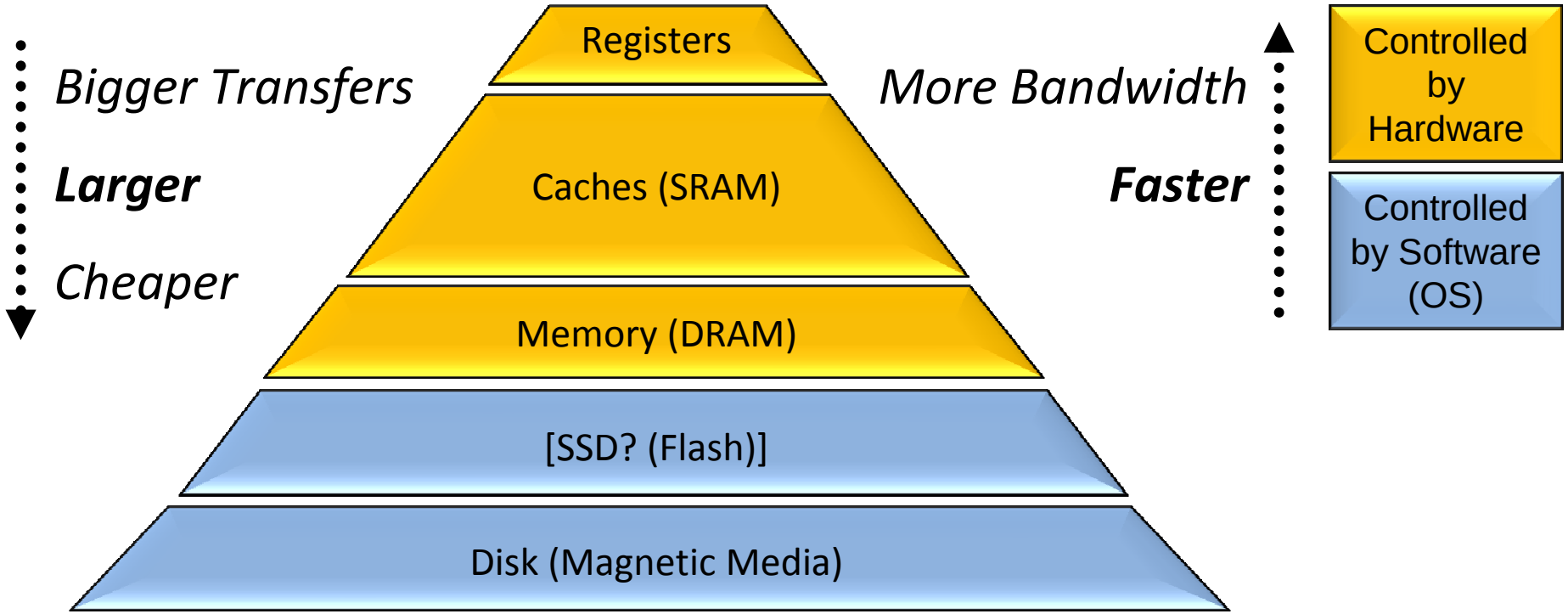
Motivation



- Want memory to appear:
  - As fast as CPU
  - As large as required by all of the running applications

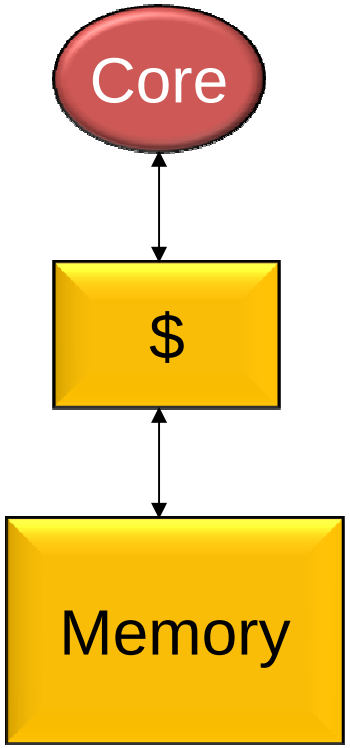
Storage Hierarchy

- Make common case fast:
  - Common: temporal & spatial locality
  - Fast: smaller more expensive memory



Caches

- An ***automatically managed*** hierarchy
- Break memory into *blocks* (several bytes) and transfer data to/from cache in blocks
  - *spatial locality*
- Keep recently accessed blocks
  - *temporal locality*



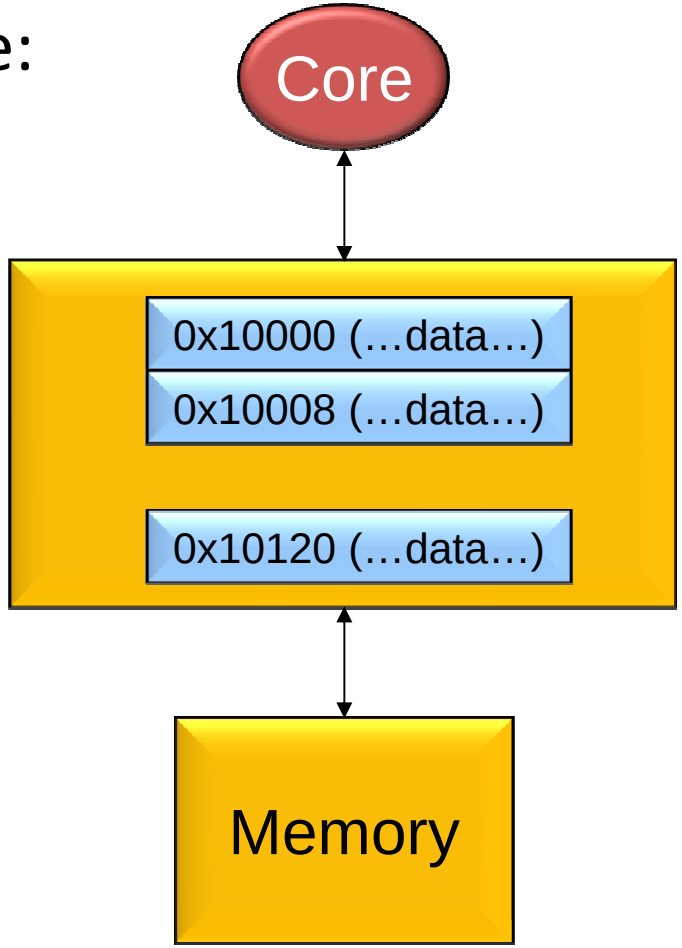
Cache Terminology

- *block* (*cache line*): minimum unit that may be cached
- *frame*: cache storage location to hold one block
- *hit*: block is found in the cache
- *miss*: block is not found in the cache
- *miss ratio*: fraction of references that miss
- *hit time*: time to access the cache
- *miss penalty*: time to replace block on a miss

Cache Example

- Address sequence from core: (assume 8-byte lines)

0x10000	Miss
0x10004	Hit
0x10120	Miss
0x10008	Miss
0x10124	Hit
0x10004	Hit



AMAT (1/2)

- Very powerful tool to estimate performance
- Average memory access time (AMAT) =

  Hit time + Miss rate x Miss penalty
- If …

  cache hit is 10 cycles (core to L1 and back)

  memory access is 100 cycles (core to mem and back)

- Then …

  at 50% miss ratio, avg. access: 10+0.5×100 = 60

  at 10% miss ratio, avg. access: 10+0.1×100 = 20

  at 1% miss ratio, avg. access: 10+0.01×100 = 11

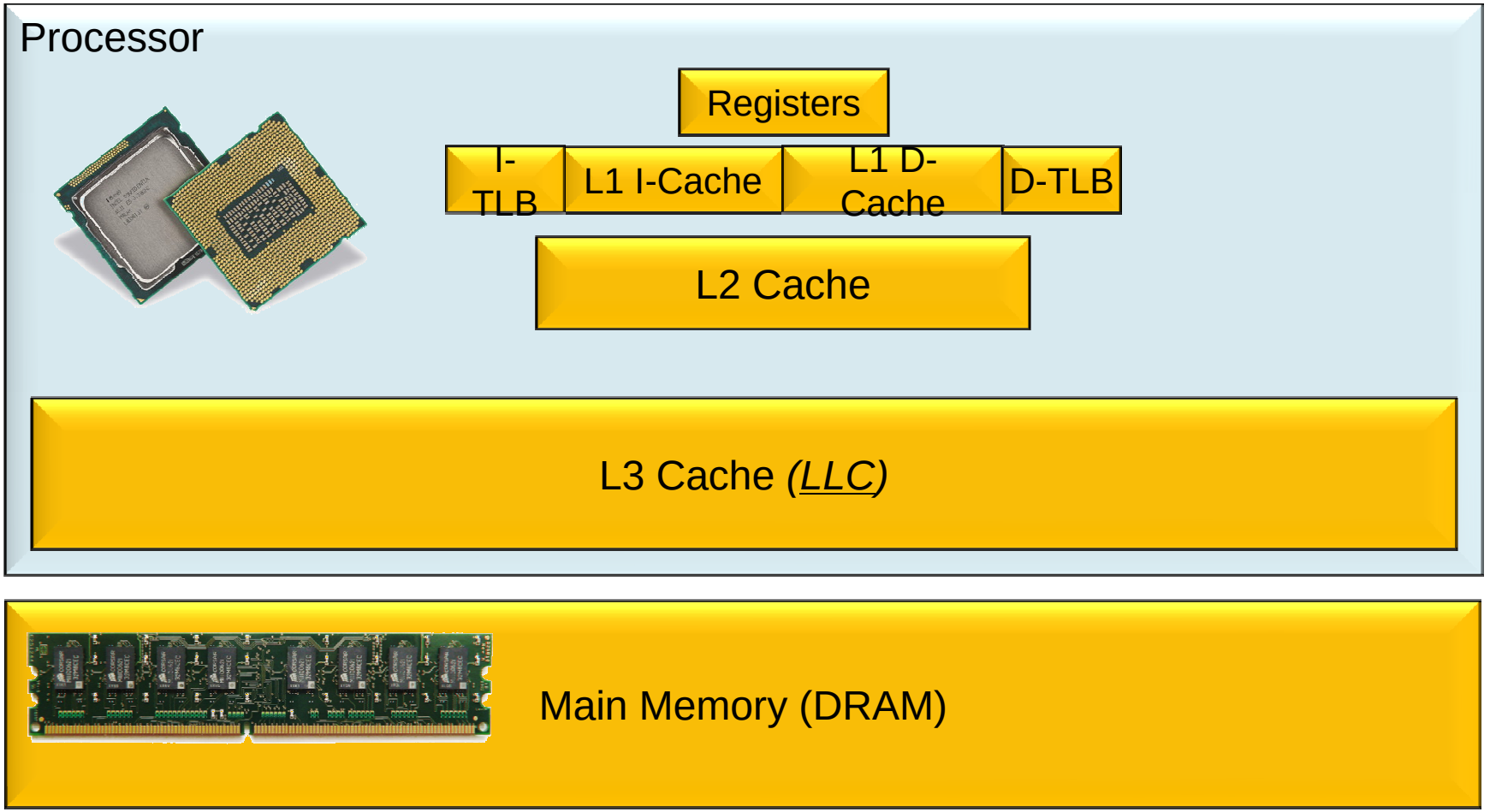
AMAT (2/2)

- Generalizes nicely to any-depth hierarchy

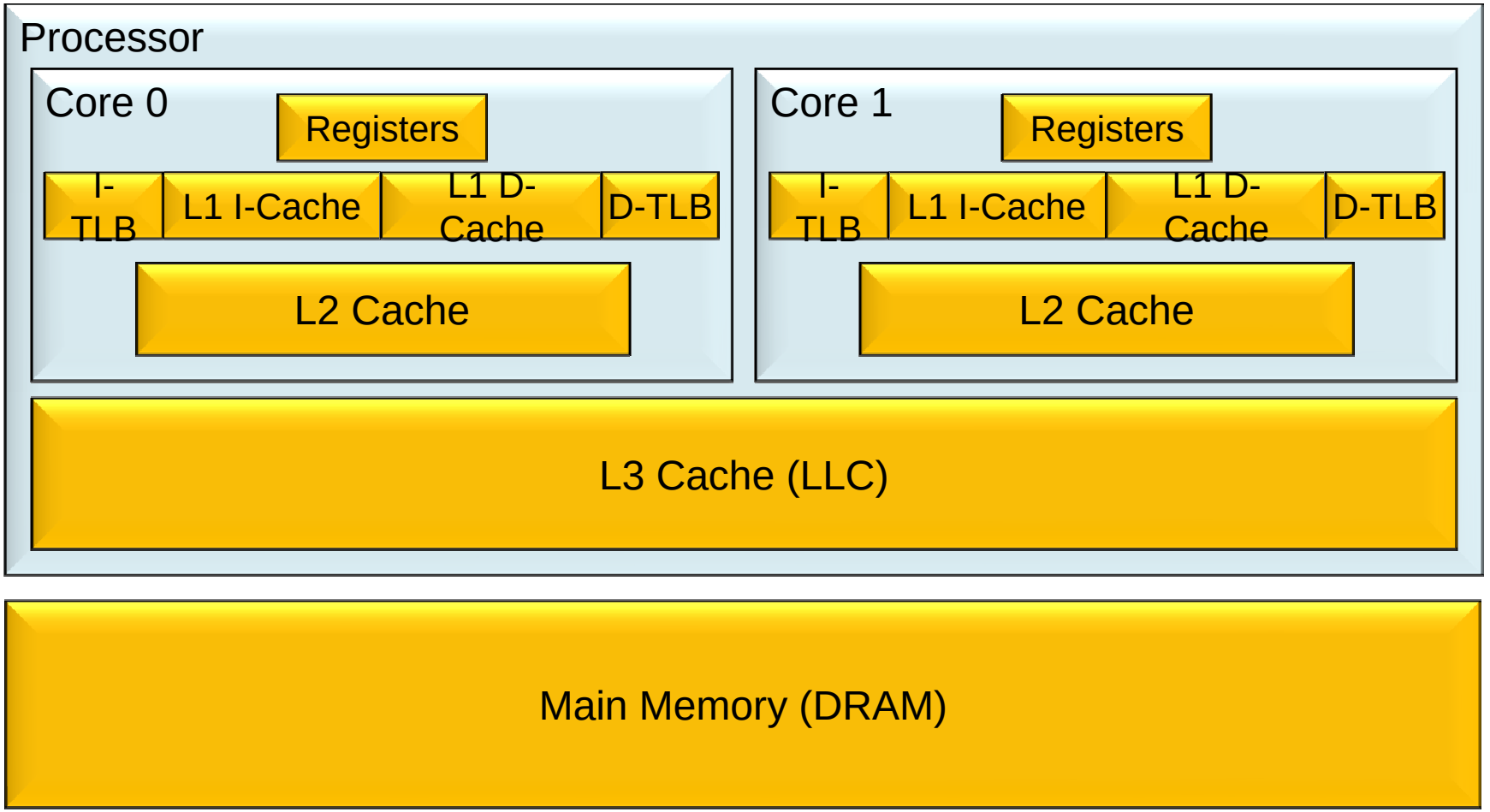
- If …
  L1 cache hit is 5 cycles (core to L1 and back)
  L2 cache hit is 20 cycles (core to L2 and back)
  memory access is 100 cycles (core to mem and back)

- Then …
  at 20% miss ratio in L1 and 40% miss ratio in L2 …
  avg. access: 5+0.2×(20+0.4×100) = 17

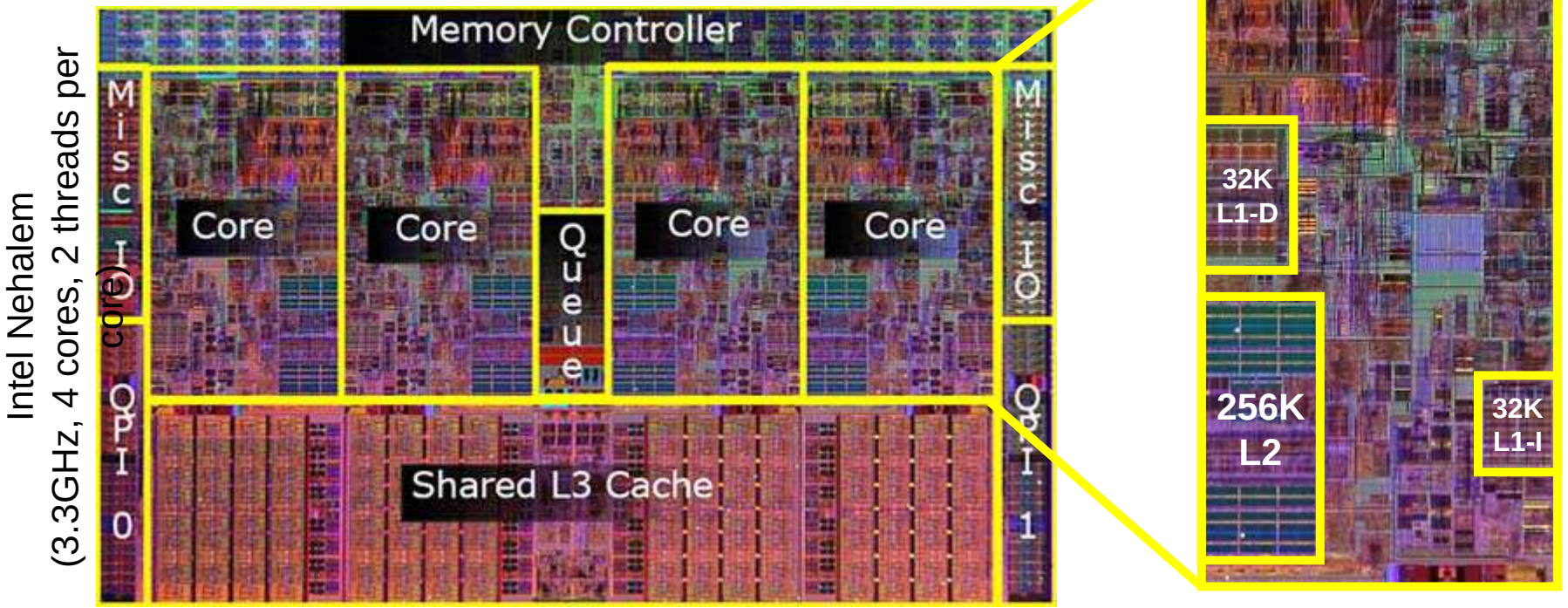
Memory Organization (1/3)



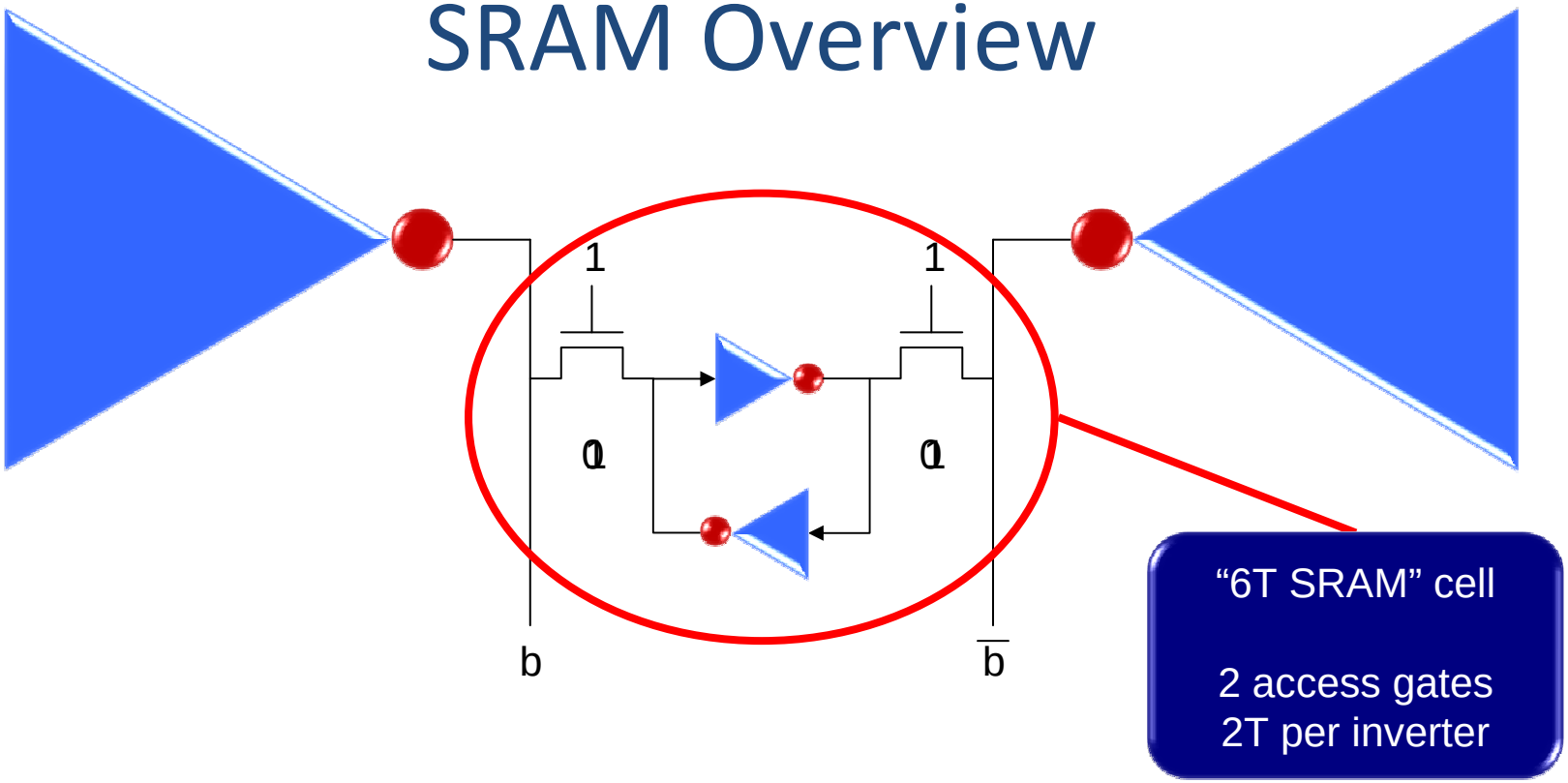
Memory Organization (2/3)



Memory Organization (3/3)



SRAM Overview



- Chained inverters maintain a stable state
- Access gates provide access to the cell
- Writing to cell involves over-powering storage inverters

8-bit SRAM Array