CACHE MEMORY & SRAM/DRAM TIMING

An introduction to cache memories and their architecture

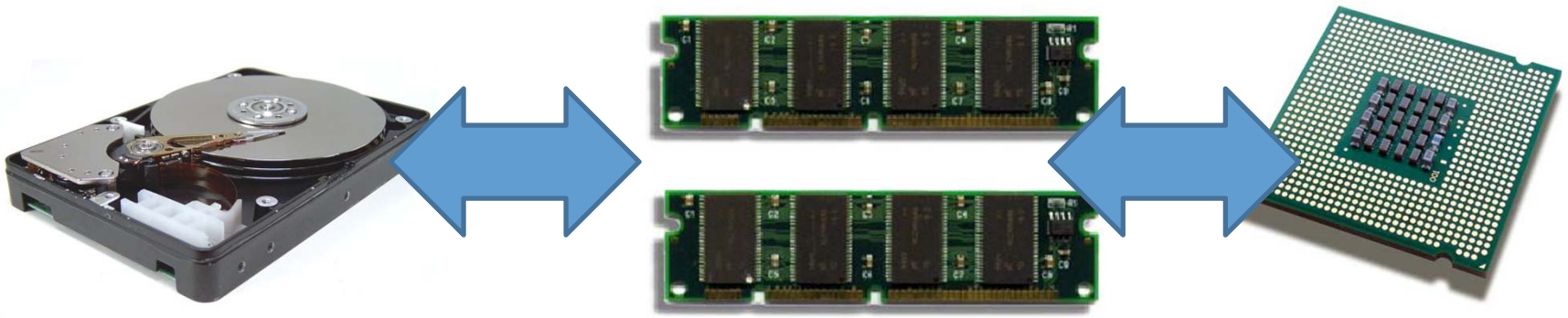
Microprocessor and Microcontrollers Course – Isfahan University of Technology – Mohammad Sadegh Sadri

Storages

- CPU needs memory
  - Program
  - Data
- Memories
  - Volatile memory
    - RAM
  - Non-Volatile memory
    - Hard disk
    - Flash

Storage Speed

- RAM is faster than
  - Hard disk and flash
- Access to RAM is easier
  - Compared to hard disk and flash



Basic Idea

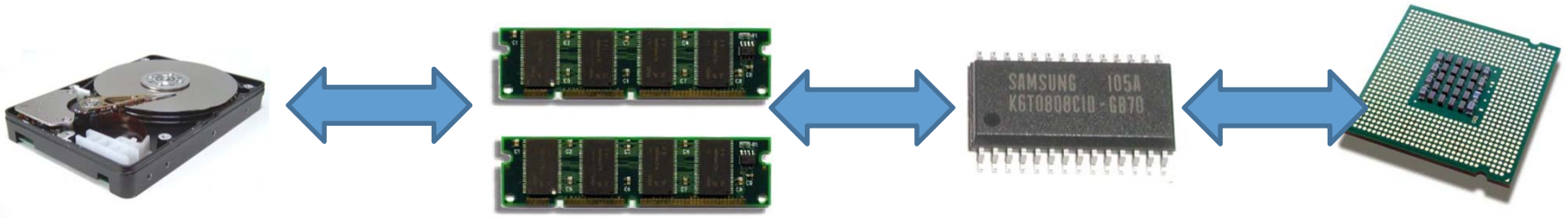
- Hard disk is slow
- Access to it is difficult
- So :
  - Put a memory between hard disk and CPU
  - Now move Mostly Used Data and Instructions to this memory
  - Allow CPU to read from/write to this memory

Basic Idea Expansion (1)

- DRAM memory
  - Large
  - Slow
- SRAM memory
  - Fast
  - Small
- Idea
  - Put a fast SRAM memory between DRAM and CPU
  - Store Mostly Used Data and Instructions on this SRAM memory
  - Allow CPU to use this memory

Basic Idea Expansion (2)

- When CPU needs an instruction
- Look into fast SRAM memory and check
  - If the instruction exist → Give it to CPU
  - If the instruction does not exist →
    - Read it from large DRAM
    - Give it to CPU
    - Update the contents of SRAM memory



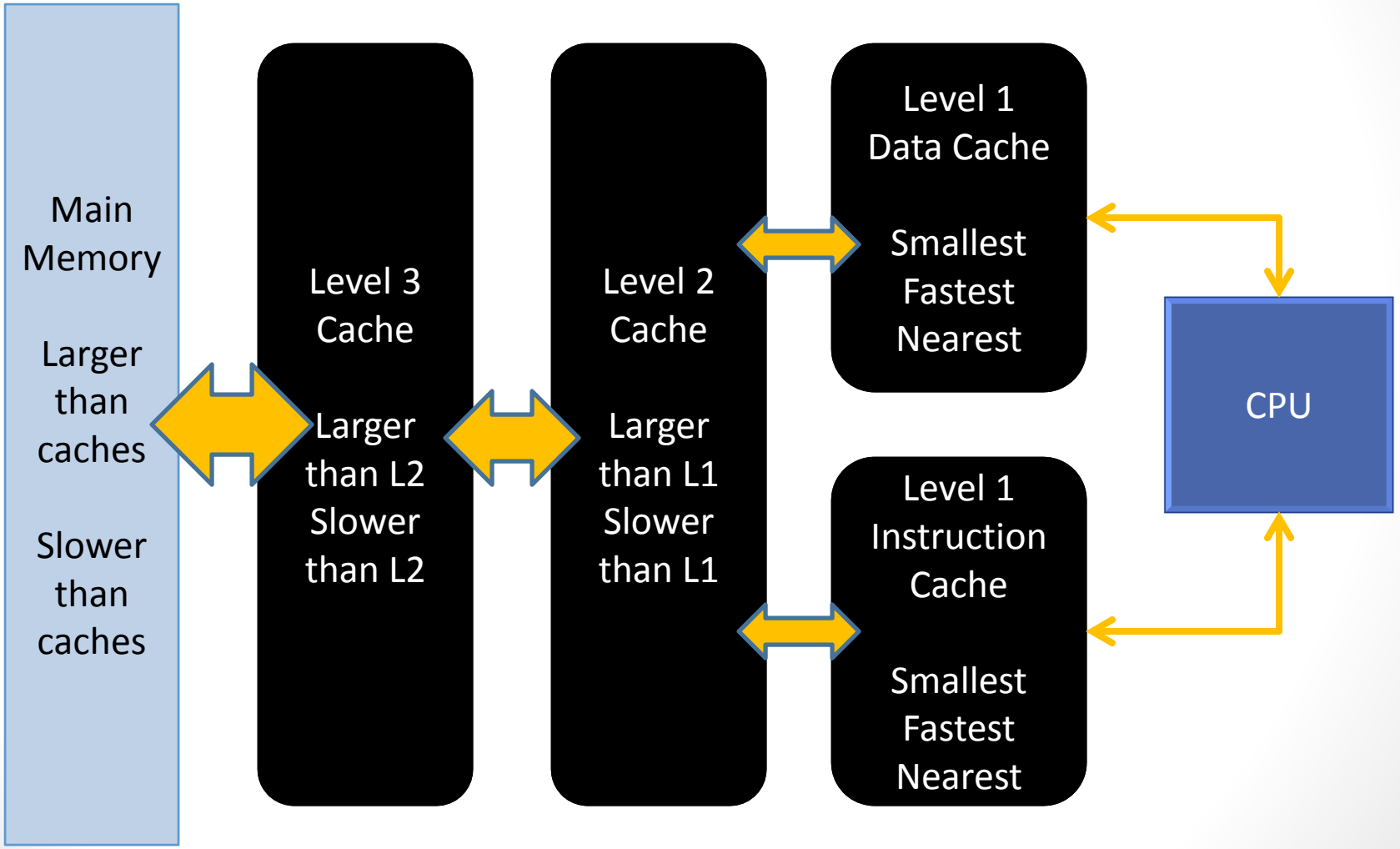
Basic Idea Expansion (3)

- When CPU wants to store/read a data
  - Look into fast SRAM memory
  - Check if any of SRAM memory locations has been associated with this memory address
    - If yes: store the data into this location (or read data from this location)
    - If no: store data into main DRAM memory
      - And also in SRAM memory

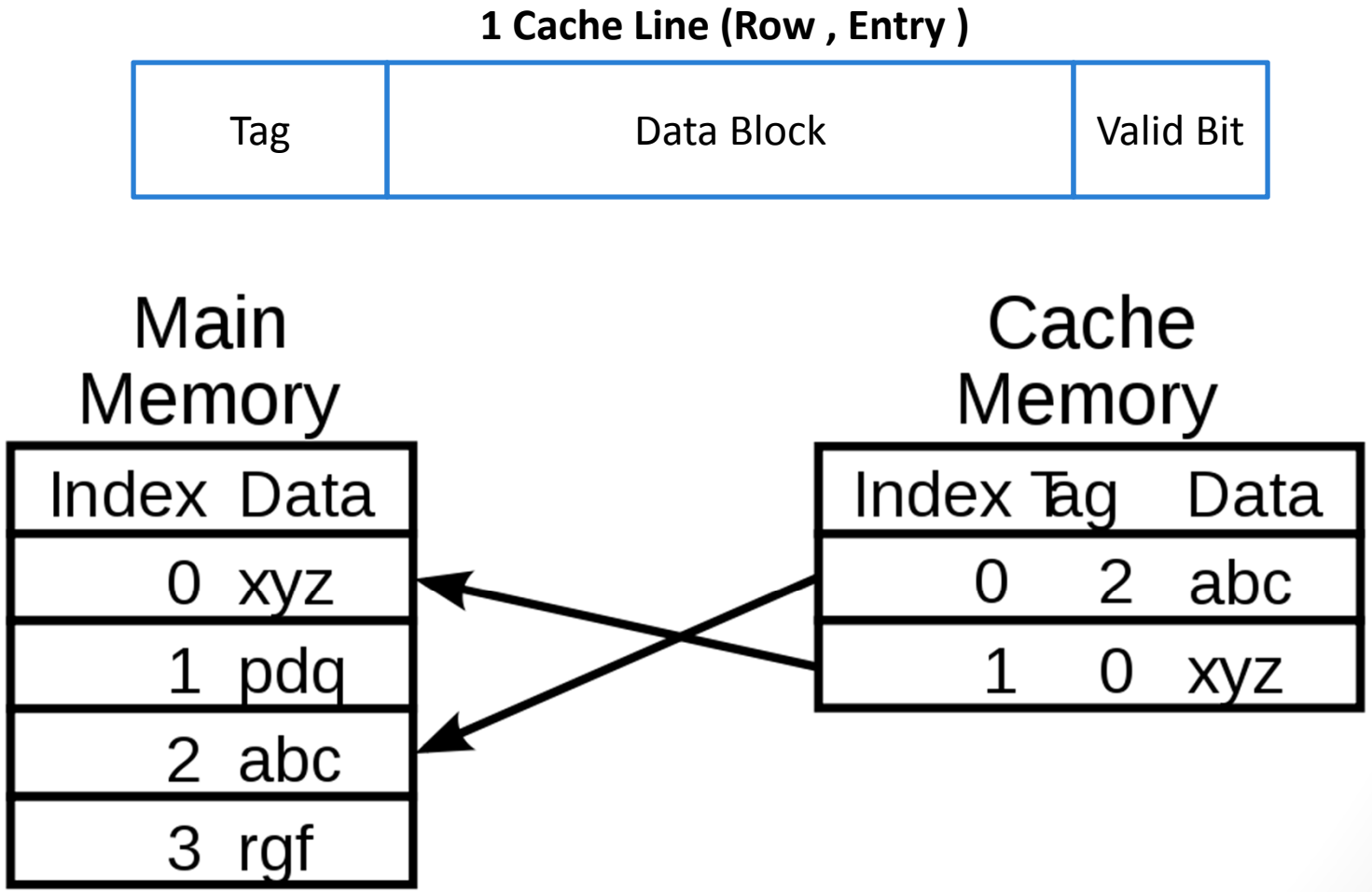
Cache Memory

- A “faster” small memory
- Stores frequently used data
- To shorten access time
- Caches are every where:
  - CPU cache
  - Hard disk cache
  - Operating system cache
  - …

Multi-Level Caches

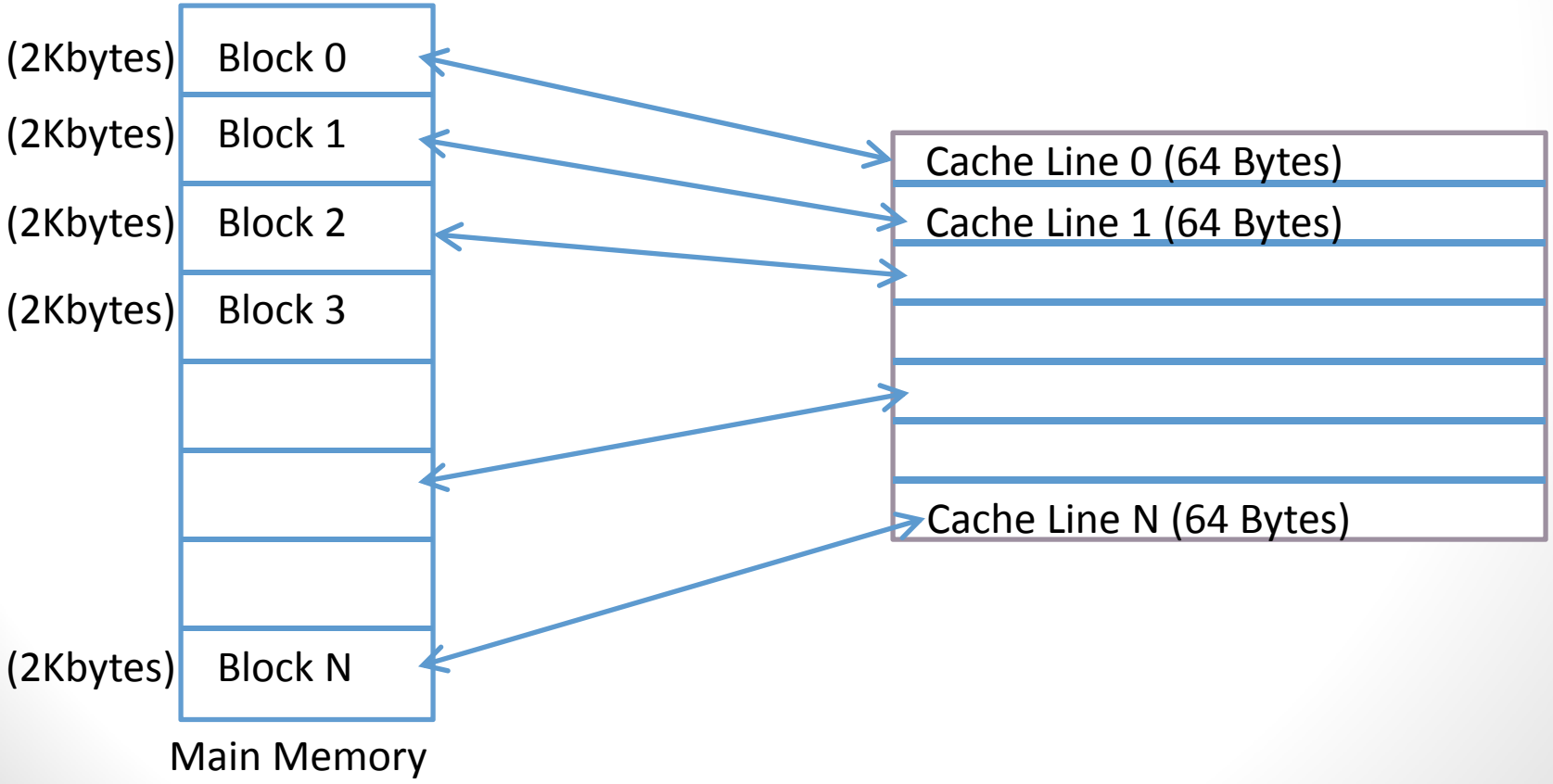


Basic Cache Operation

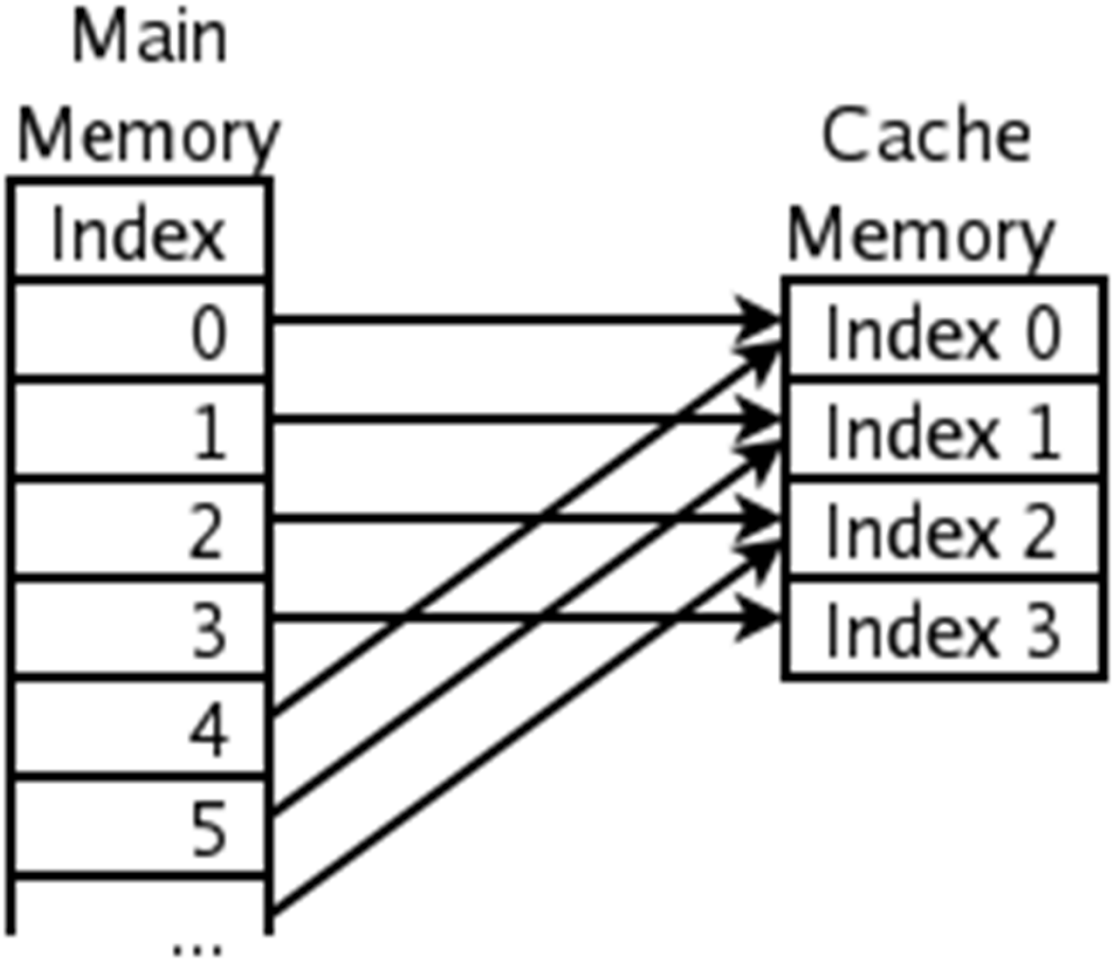


Direct Mapped Cache

- Main memory is divided into blocks
- Each block can be cached in only one cache row
- Fasted search speed (best hit time, worst hit rate)



Direct Mapped Cache



Each location in main memory can be cached by just one cache location.

Direct Mapped Cache

- Can some times be very bad!
- Suppose both of A, B and C are stored in the same memory block in the following code

```
for ( i = 0 ; i < 10; i++ )
    A[i] = B[i] + C[i];
```

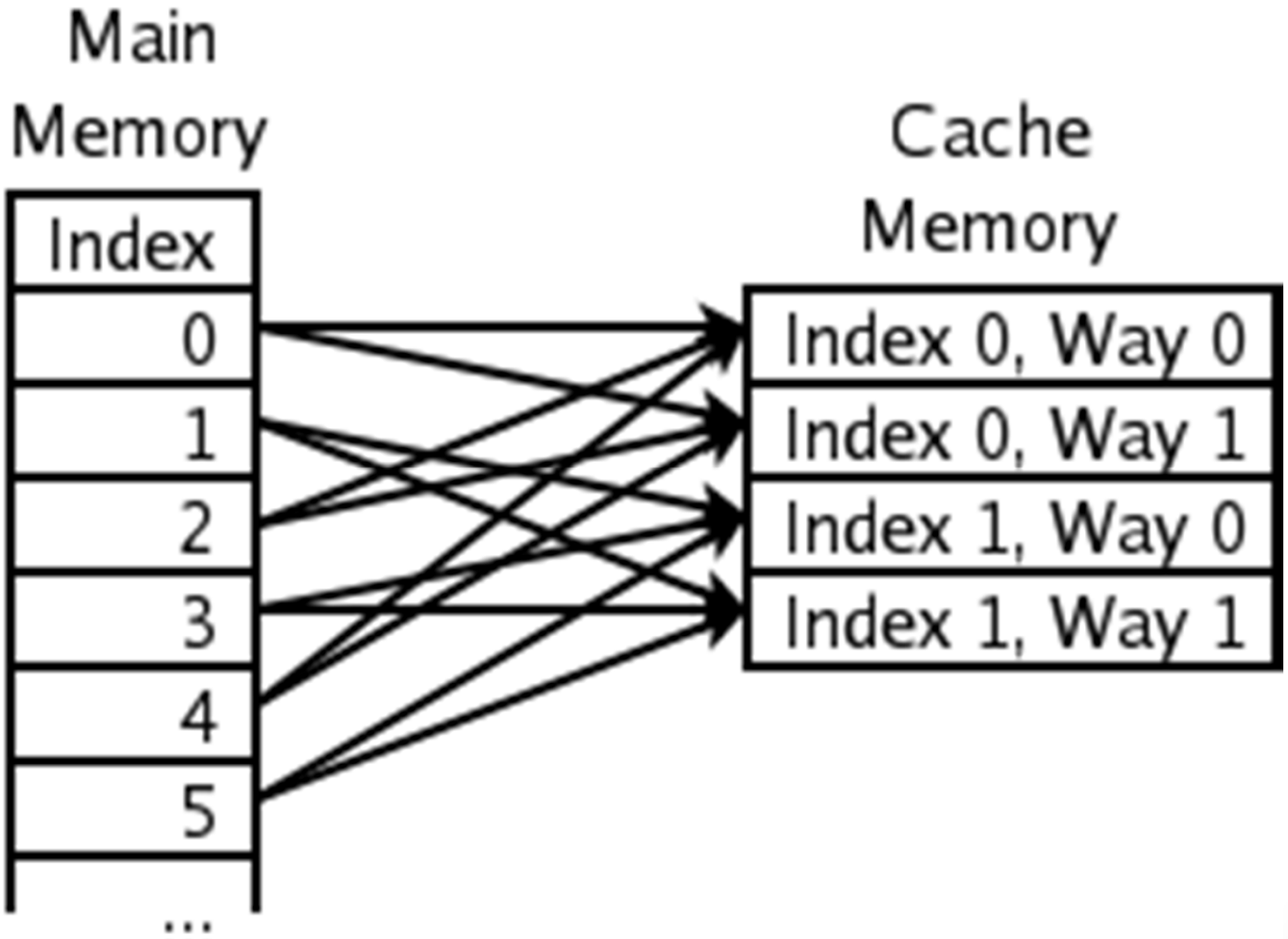
Fully Associative Cache

- Each main memory location
  - Can be cached in all of the cache memory lines
- When CPU wants to access a memory location
  - All of the cache lines should be searched first
- Slowest search (worst hit time)
- Best hit rate
- There are some techniques to speed up the search

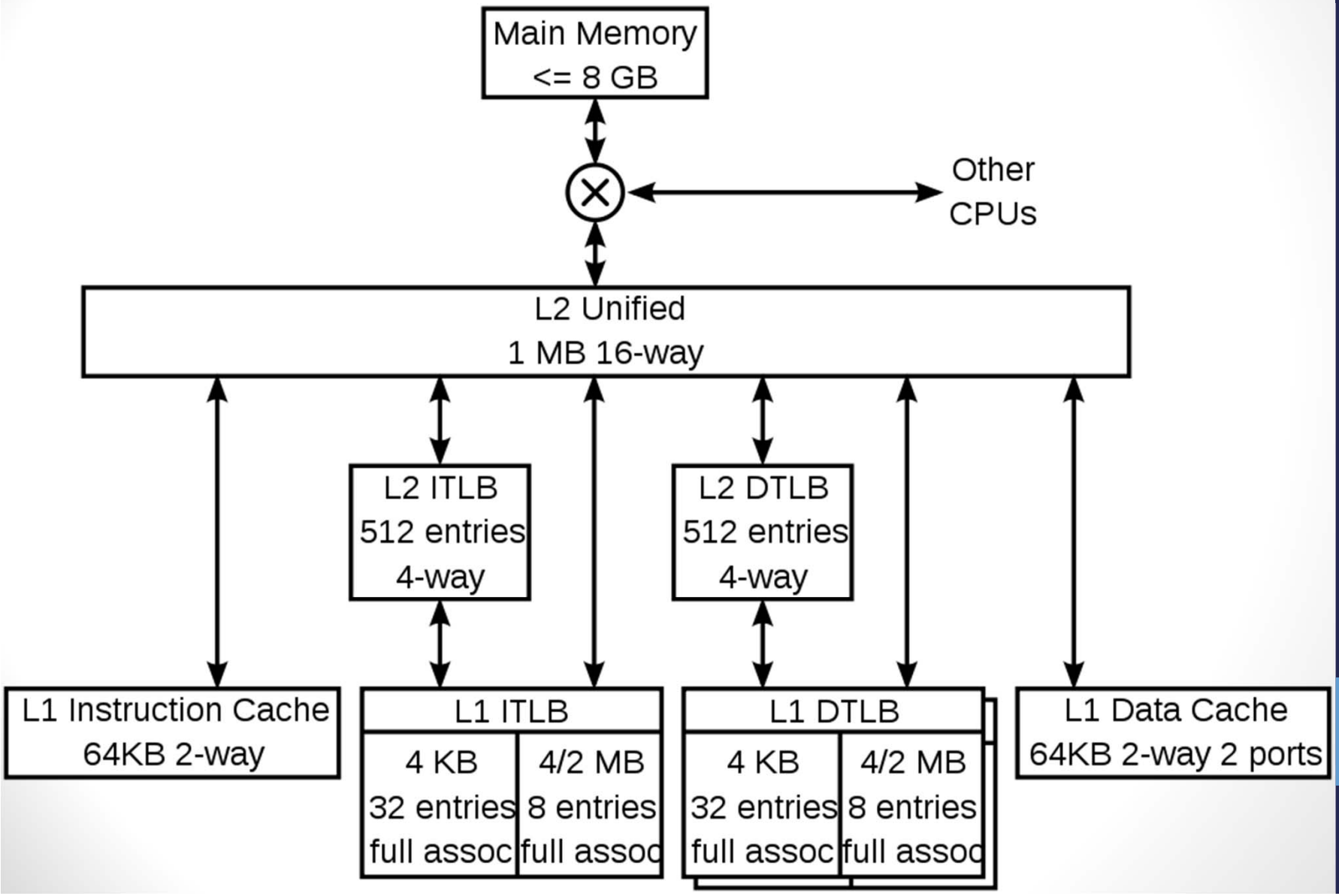
Set-Associative Cache

- Cache will be divided into sets
  - Each set is 2,4,8 or … cache lines
- The content of each memory block
  - Can be cached in any of cache lines in a specific set
- Doubling the associativity
  - For example from direct mapped to 2-way
  - Has same effect on hit rate as: doubling the cache size
- Trade off between hit rate, and hit time

2-Way Set Associative Cache



AMD Athlon 64 Cache

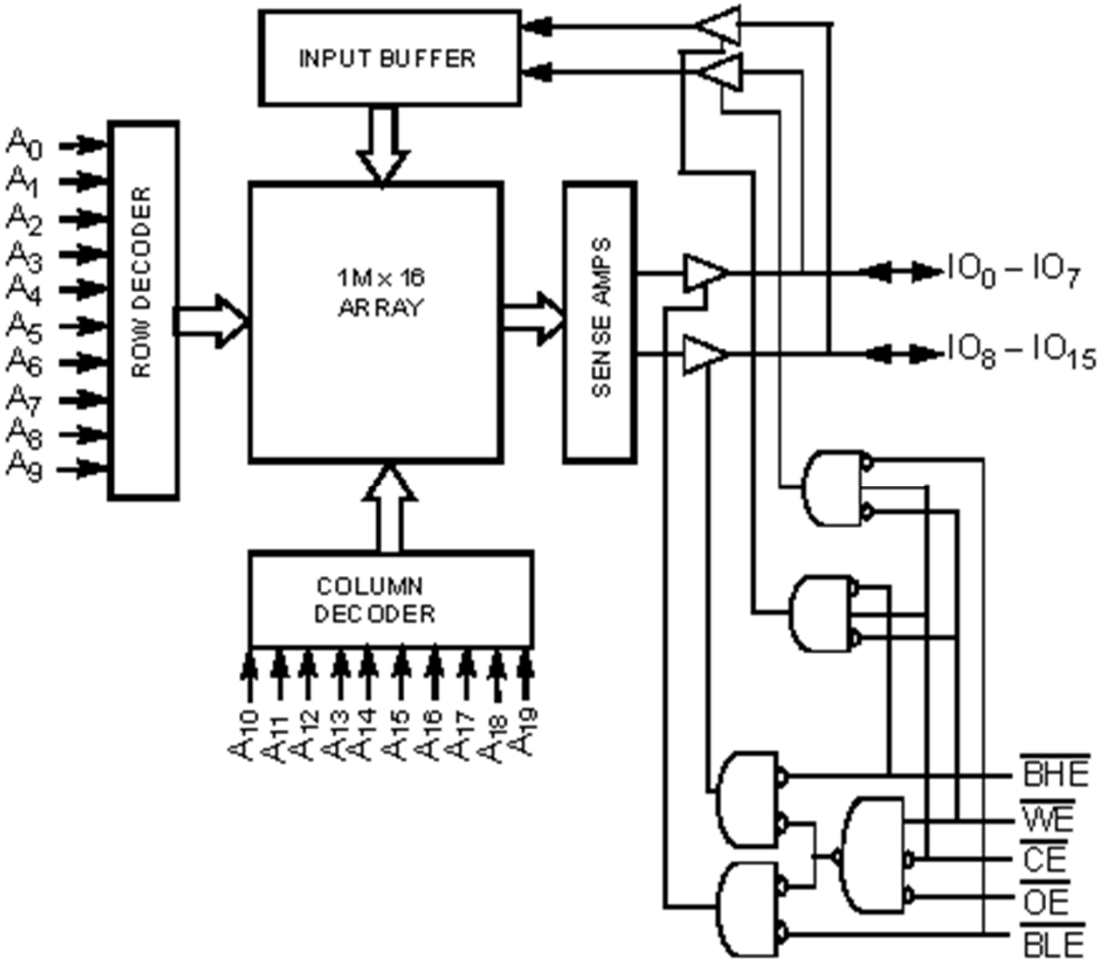


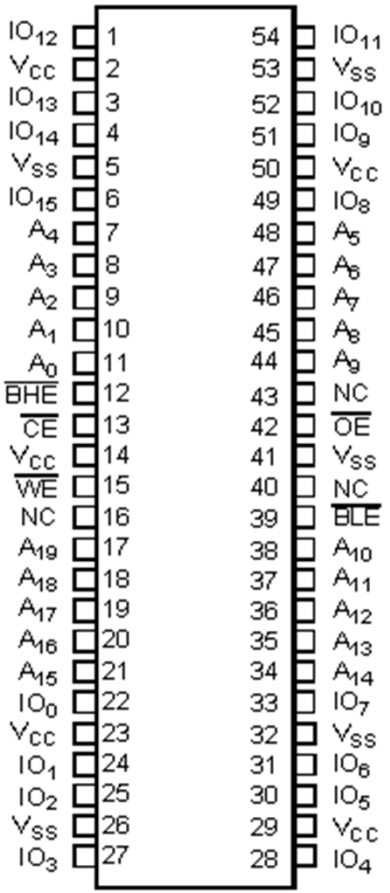
Detailed operation of DRAM and SRAM memories

DRAM VS. SRAM

Asynchronous SRAM

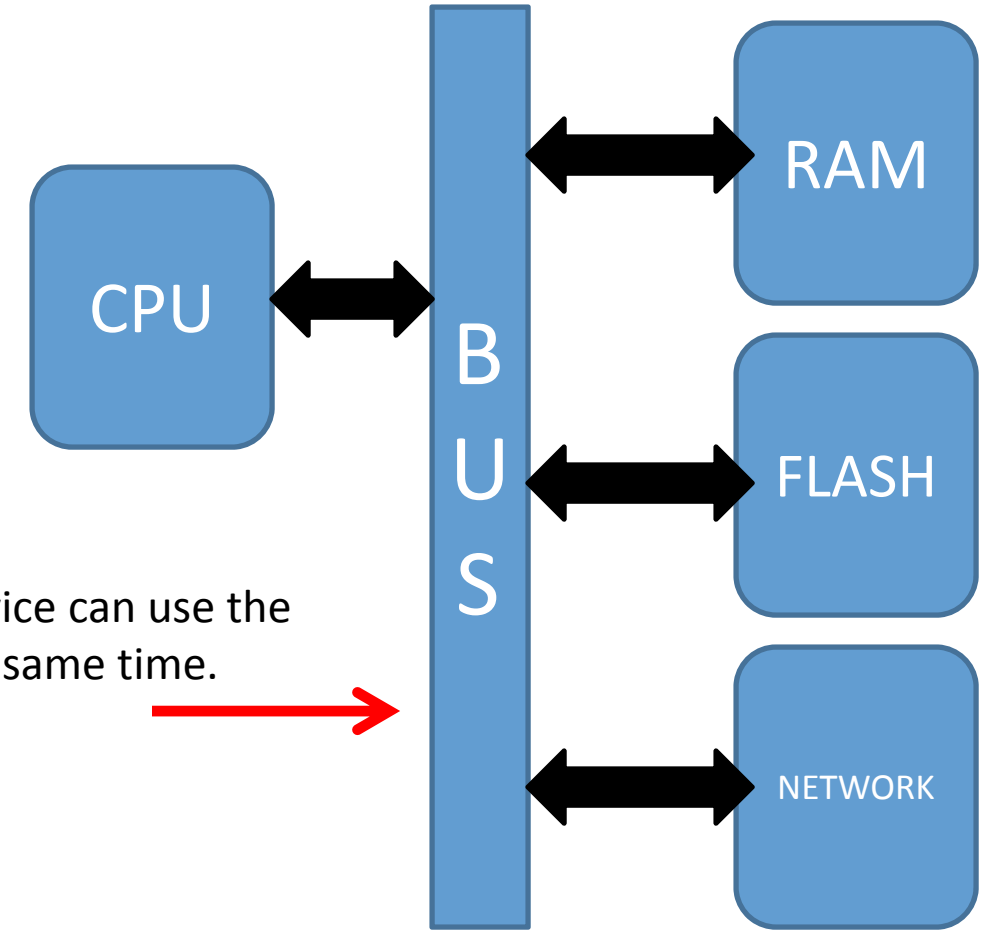
- Example : CYPRESS CY7C10612DV33 (16Mbits : 1M*16, 10ns)





What is chip enable?

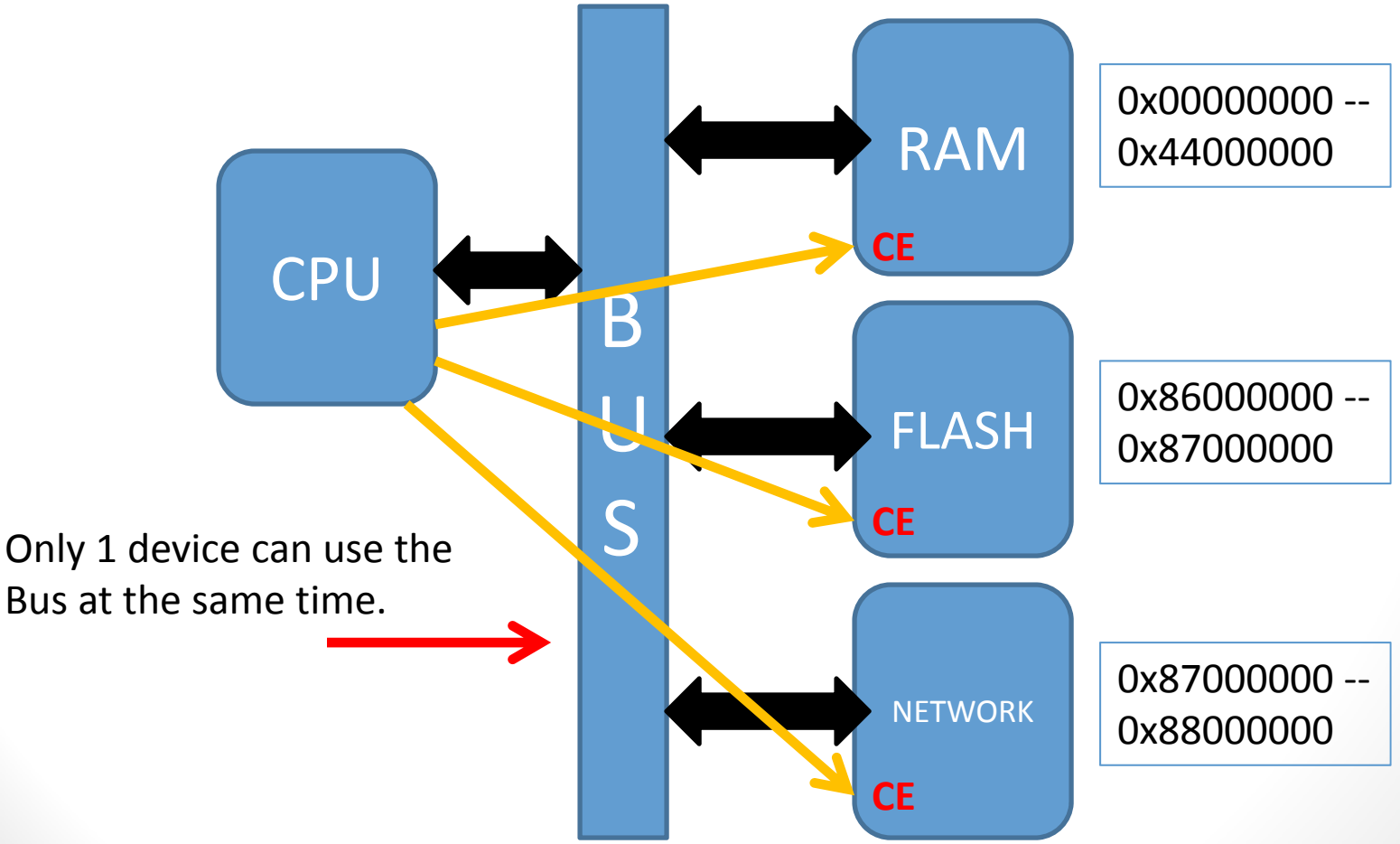
- Usually CPU is connected to multiple devices on the same bus



Only 1 device can use the Bus at the same time.

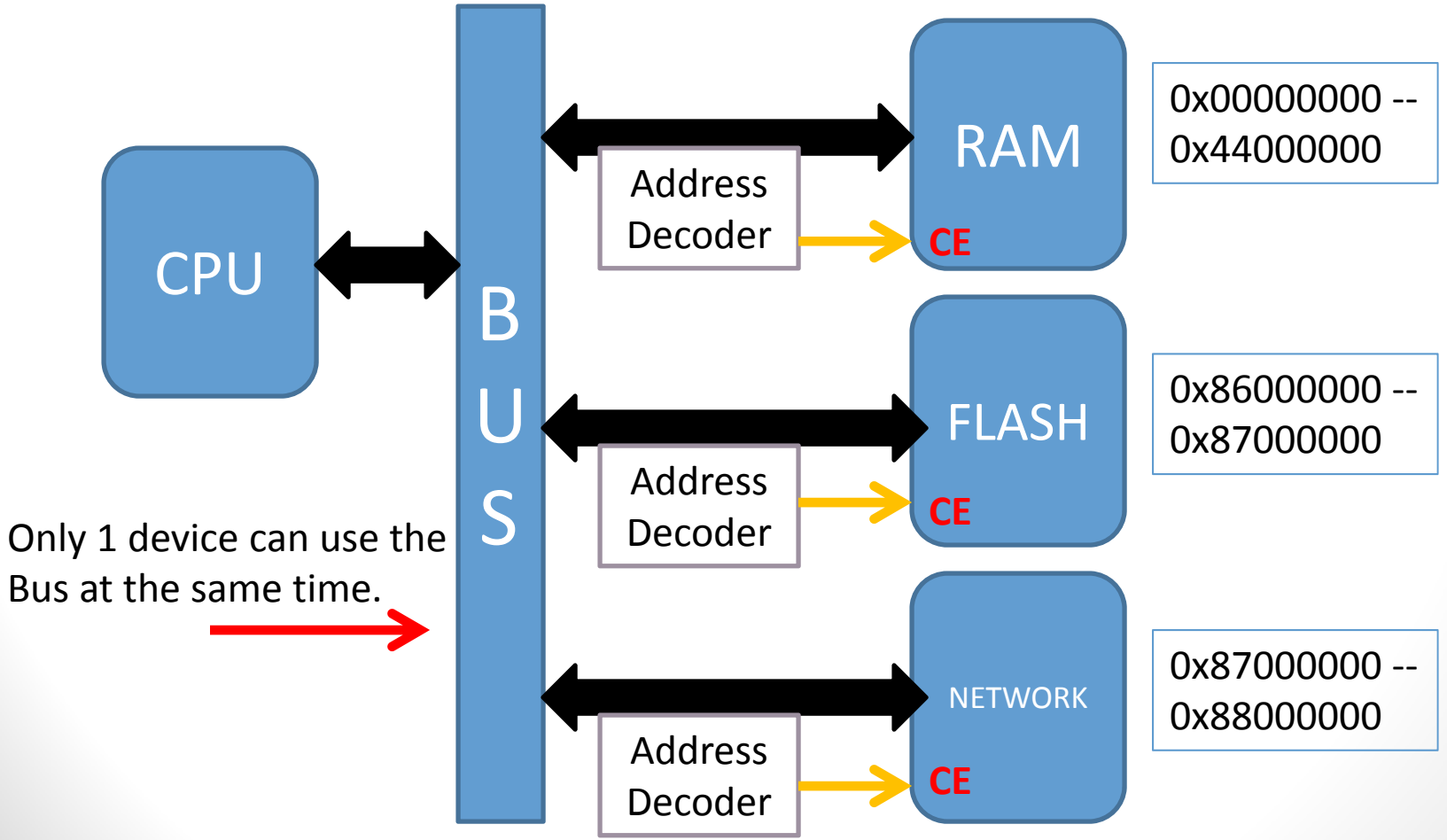
What is chip enable? (2)

- Usually CPU is connected to multiple devices on the same bus

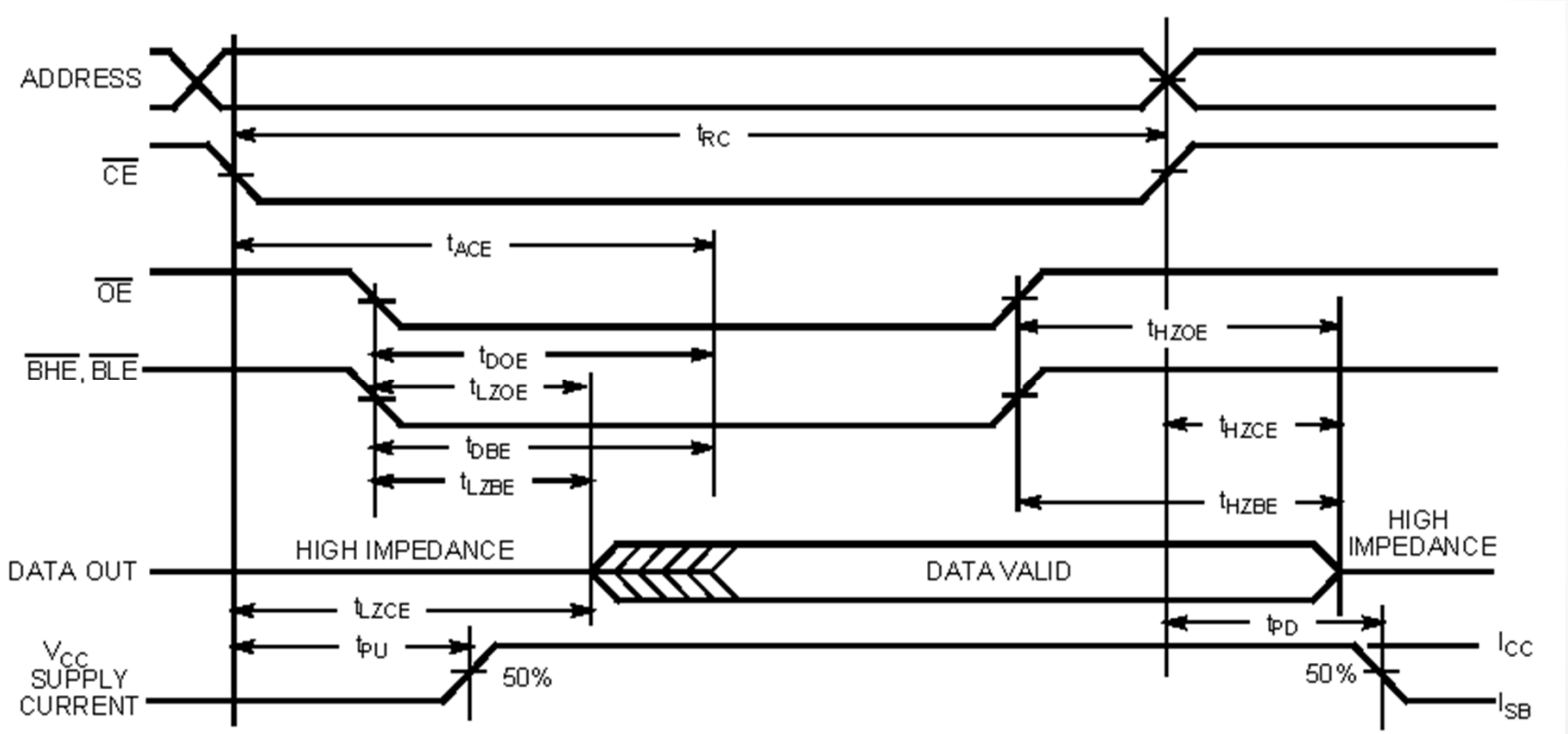


What is chip enable? (3)

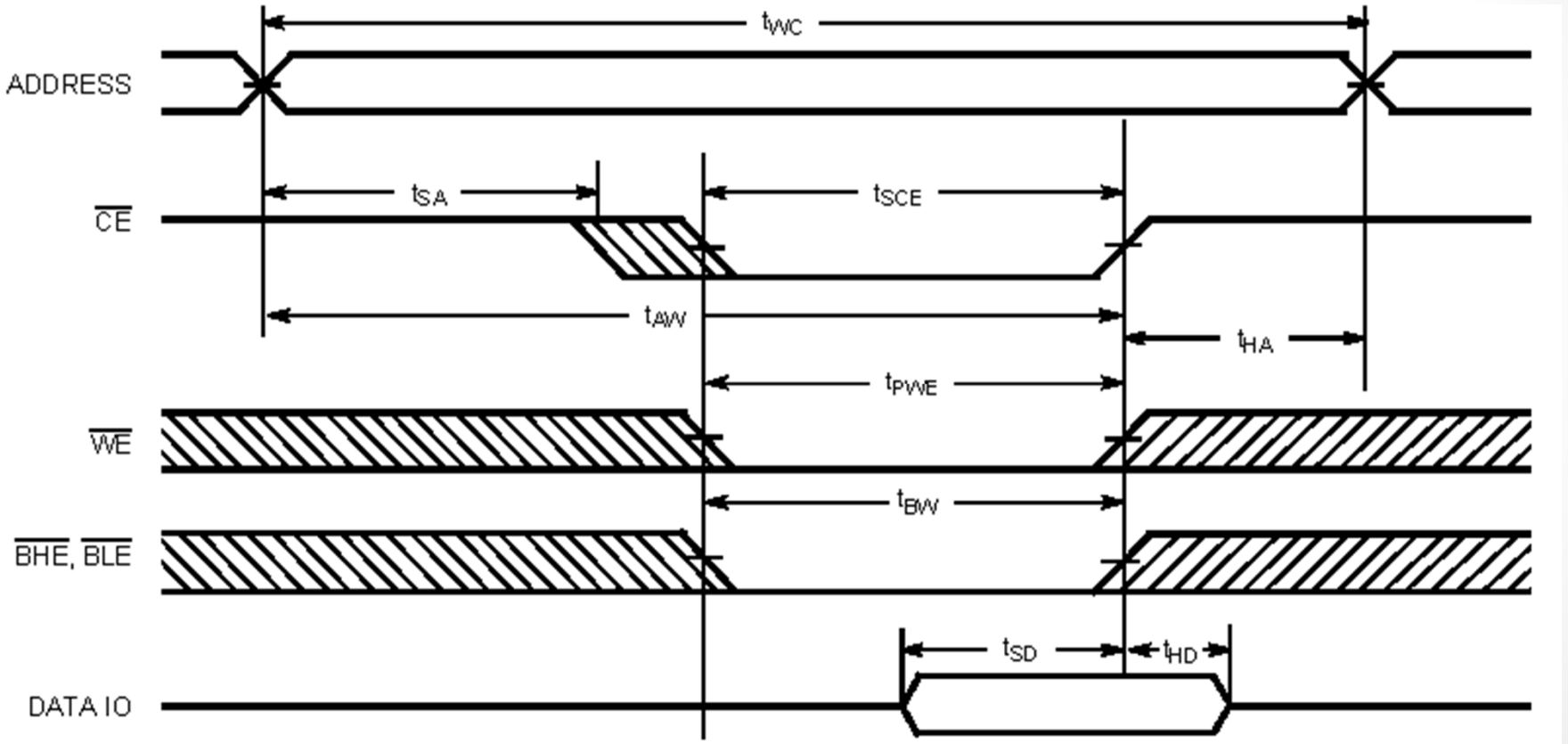
- Usually CPU is connected to multiple devices on the same bus



Async-SRAM Read Cycle



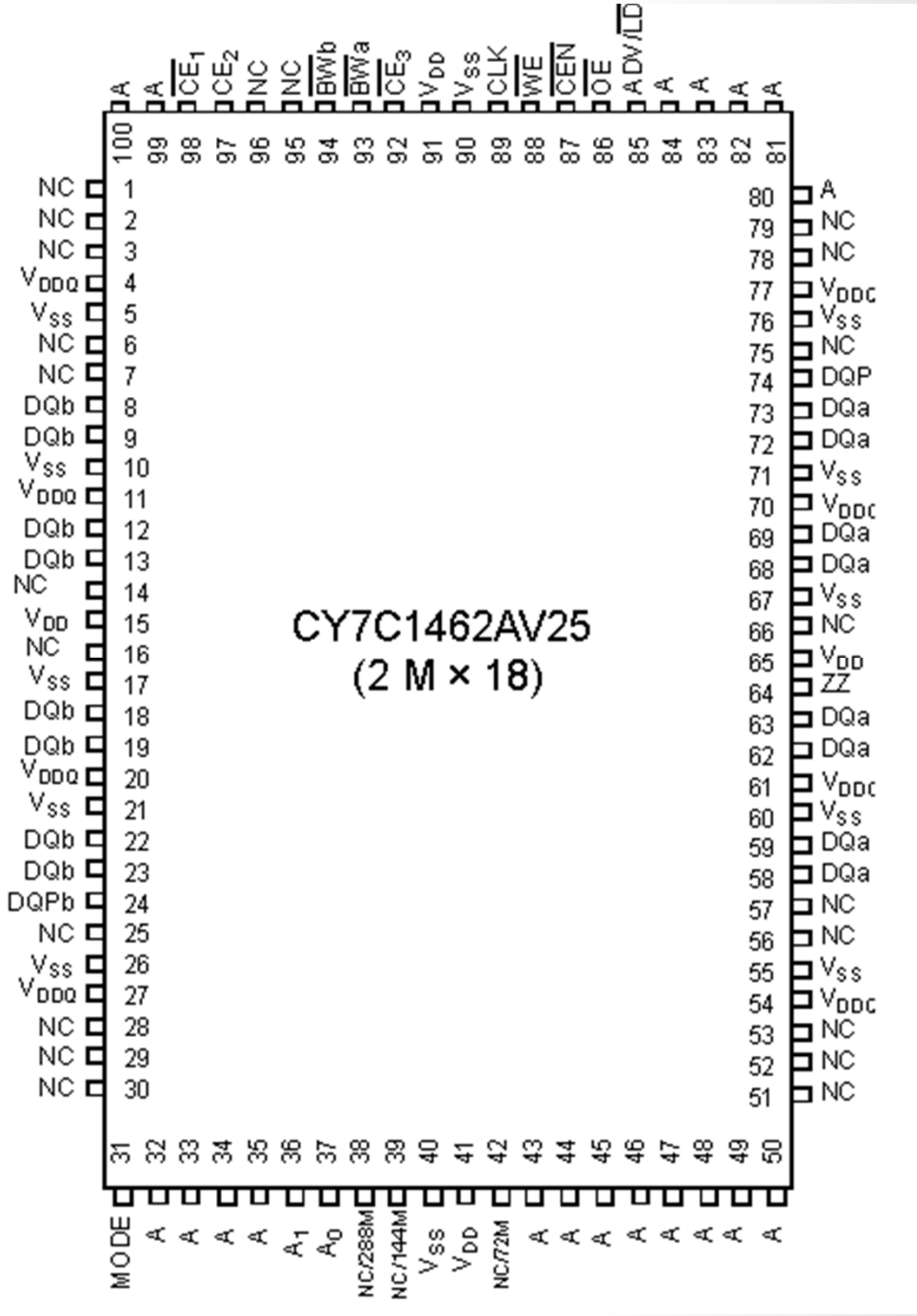
Async-SRAM Write Cycle



Synchronous SRAM

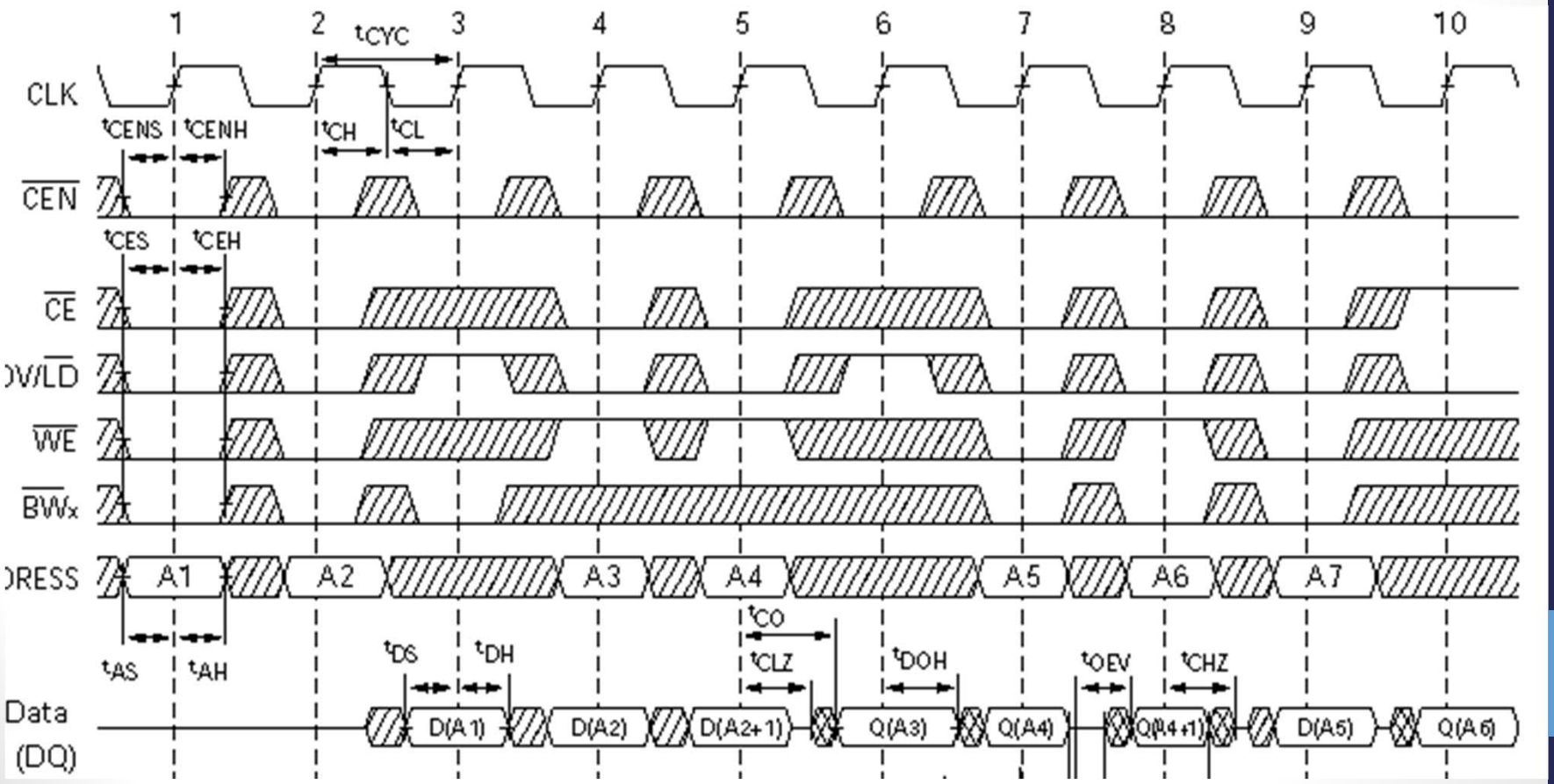
- Similar to Async SRAM
- But
  - Every Read and Write operation is done with Clock pulse edges
- Example:
  - Cypress CY7C1460
  - 36MBits – 2M*18
  - No-Bus-Latency Synchronous SRAM (NoBL Architecture)
- No-Bus-Latency:
  - There is no delay for switching between read and write operations

CY7C1462


CY7C1462AV25
(2 M × 18)
MODE
A
A
A
A
A1
A0
NC/288M
NC/144M
VSS
VDD
NC/72M
A
A
A
A
A
A
A
A
NC
NC
NC
VDDQ
VSS
NC
NC
DQb
DQb
VSS
VDDQ
DQb
DQb
NC
VDD
NC
VSS
DQb
DQb
VDDQ
VSS
DQb
DQb
DQPb
NC
VSS
VDDQ
NC
NC
NC
A
NC
NC
VDDQ
VSS
NC
DQP
DQa
DQa
VSS
VDDQ
DQa
DQa
VSS
NC
VDD
ZZ
DQa
DQa
VDDQ
VSS
DQa
DQa
NC
NC
VSS
VDDQ
NC
NC
NC
A
A
CE1
CE2
NC
NC
BWb
BWa
CE3
VDD
VSS
CLK
WE
CEN
OE
ADV/LD
A
A
A
A
100
99
98
97
96
95
94
93
92
91
90
89
88
87
86
85
84
83
82
81
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
80
79
78
77
76
75
74
73
72
71
70
69
68
67
66
65
64
63
62
61
60
59
58
57
56
55
54
53
52
51

Read/Write Timing

- Page 22, component's data sheet



DRAM Memory

- Access to DRAM memory is very complicated (compared to SRAM)
- DRAM accept commands (there is not just a simple WEn signal)
- DRAM accepts Address in 3-phases
  - Bank Address
  - Row Address
  - Column Address

Micron MT46V64M16

- Total number of : 1G bits
  - 4 Banks
  - 16M lines
  - Each line is 16Bits

Timings

- Page 53 Datasheet
- Page 64 Datasheet