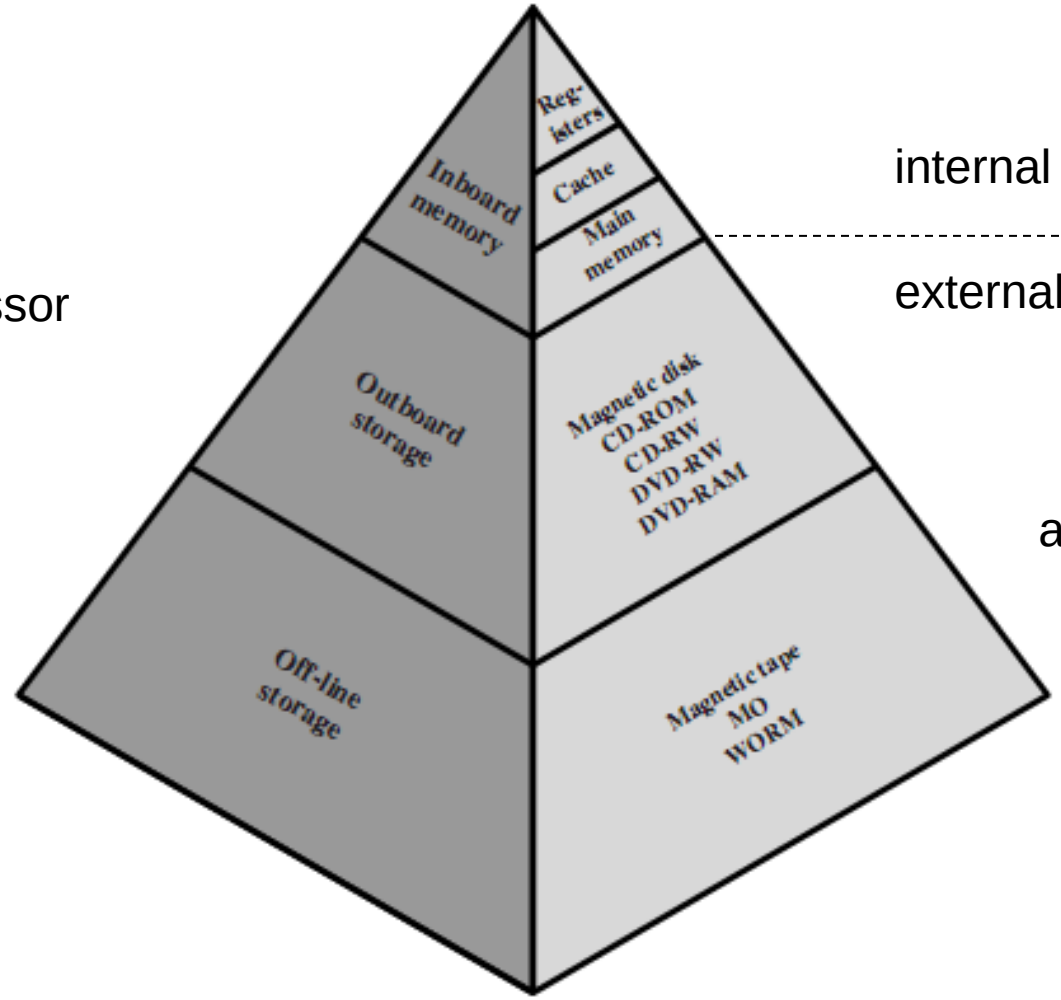
Memory Basics

CO 2206 Computer Organization

Topics

- Memory hierarchy
- Key characteristics
- Memory chip structure
- SRAM vs DRAM
- Memory mapping
- Memory address decoding

Memory Hierarchy



distance to processor

access time

capacity

internal

external

cost per bit

access frequency

Key Characteristics

- Organization
- Location
  - Processor
  - Internal (main)
  - External (secondary)
- Capacity
  - Word size
  - Number of words
- Unit of Transfer
  - Word
  - Block
- Performance
  - Access time
  - Cycle time
  - Transfer rate
- Physical Type
  - Semiconductor
  - Magnetic
  - Optical
  - Magneto-Optical
- Physical Characteristics
  - Volatile/non-volatile
  - Erasable/non-erasable

Characteristics - 1

- Organization – physical arrangement of bits to form words
- Word – natural unit of organization of memory determined by the size of the data bus
  - Most computer use words that are multiples of byte in length
  - Capacity of memory usually in bytes
- Access time – time it takes to perform a read or write operation

Characteristics - 2

- Memory cycle time – access time plus any additional time required before a second access can commence
- Transfer rate – the rate at which data can be transferred into or out of a memory unit. For random-access memory, it is equal to 1/(cycle time).
- Volatile – content will be lost once electrical power is removed
- Non-volatile – content will remain even when electrical power is removed

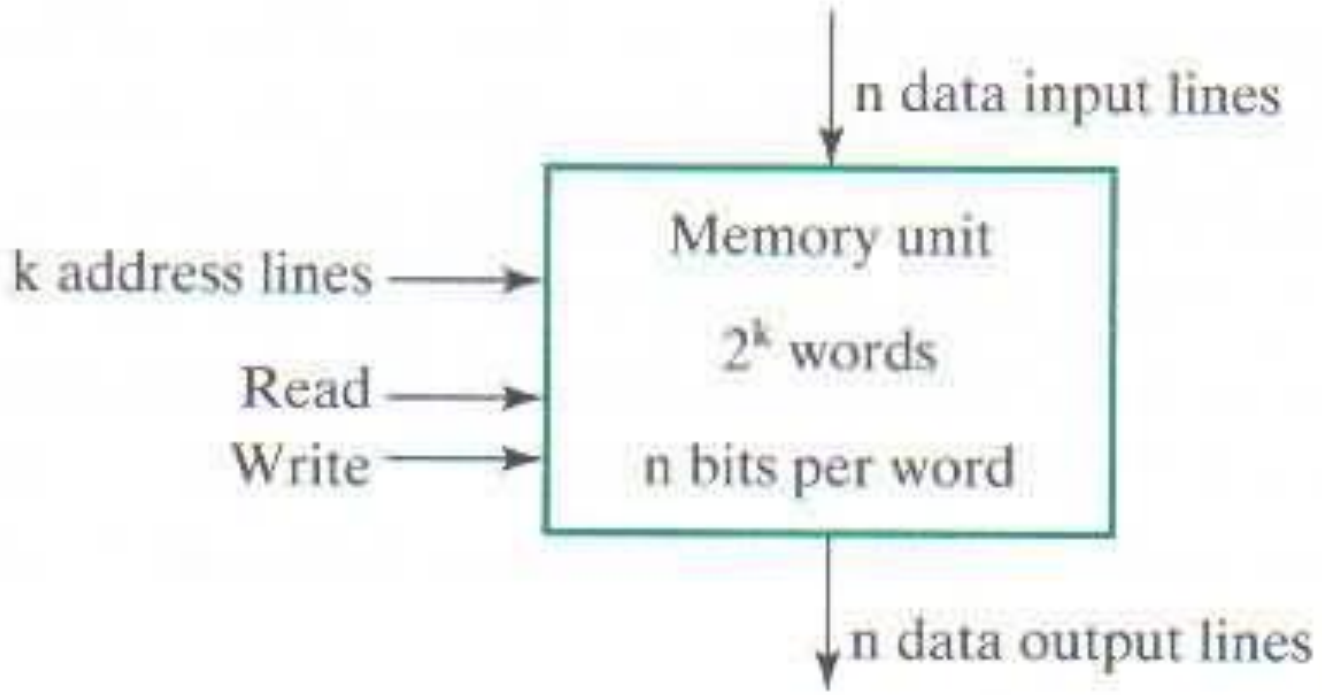
Definitions

- *Memory* – a collection of binary storage cells together with associated circuits needed to transfer info into and out of the cells
- *Write* – process of storing new info in memory
- *Read* – process of transferring stored info out of memory

Types of Main Memory

- Main memory is random access consisting of memory cells that can be accessed to transfer info to or from any desired location, with the access taking the same time regardless of the location. Two major types:
  - Read Only Memory (ROM) is non-volatile memory. It can be read only under normal memory access.
  - Random Access Memory (RAM) is volatile memory. It can be read and written under normal memory access.
- Conceptually, ROM is a RAM. However, the terminology RAM has been defined for read/write volatile memory

Block Diagram of Memory Chip

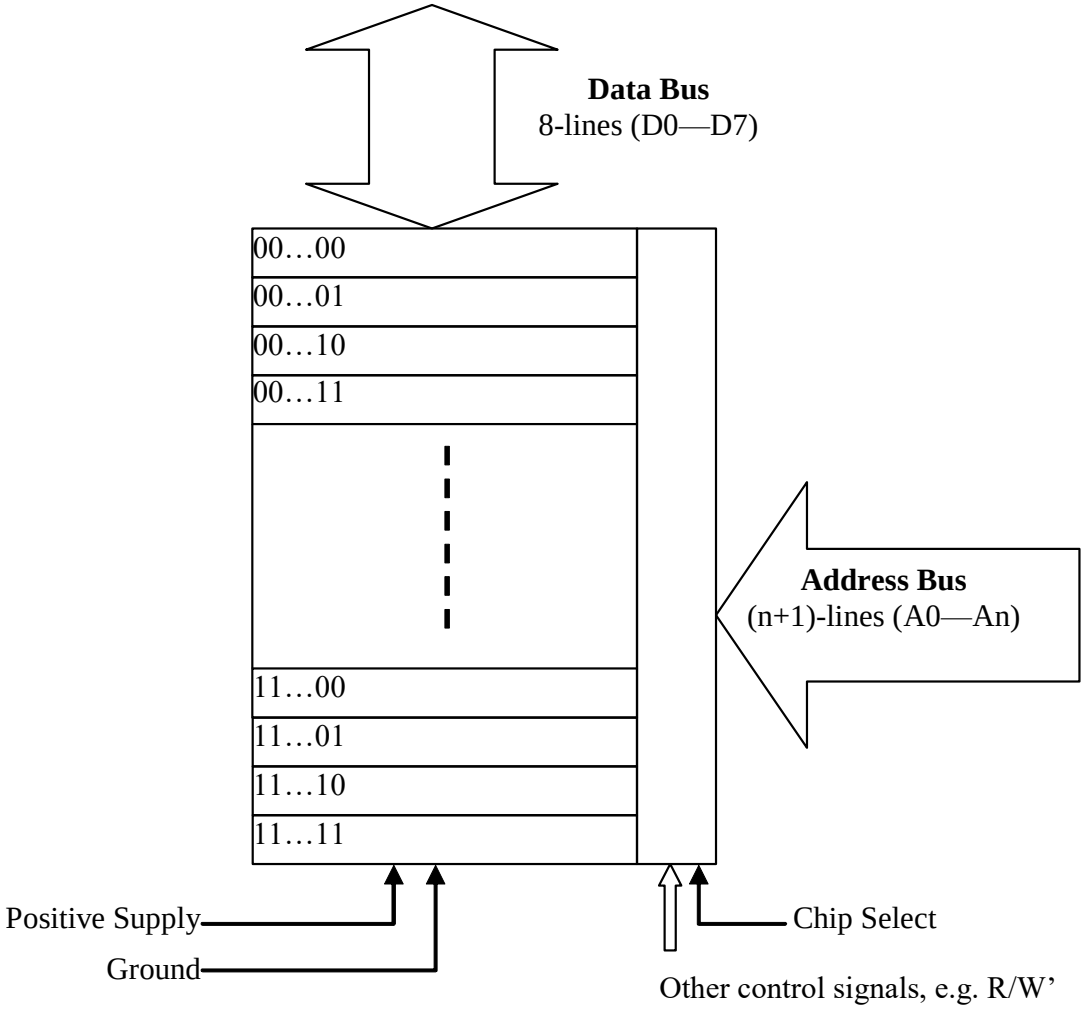


Memory Chip Structure

- *n* data input lines supply information to be stored
- *n* data output lines supply information coming out
- *k* address lines specify particular word chosen
- Control inputs specify direction of transfer
  - *Write* input causes data to be transferred into memory
  - *Read* input causes data to be transferred out of memory
  - Instead of having separate Read and Write inputs, most ICs provide at least
    - a Chip Select that selects the chip to be read or written to and
    - a Read/Write that determines the particular operation
- *Address* – Identification number assigned to each word in memory

Block Diagram of Memory Chip

Elaborated



Write and Read Operations

- *Write* steps:-
  1. CPU applies data bits that must be stored in memory to data input lines
  2. CPU applies binary address of the desired word to address lines
  3. CPU activates the Write input
  4. Memory receive data from data bus
- *Read* steps:-
  1. CPU applies binary address of the desired word to address lines
  2. CPU activates the Read input
  3. Memory place data on data bus; CPU receive data

Control Inputs to a Memory Chip

Chip select CS	Read/$\overline{\text{Write}}$ R/$\overline{\text{W}}$	Memory operation
0	×	None
1	0	Write to selected word
1	1	Read from selected word

RAM: SRAM & DRAM

- IC RAM may be either *static* or *dynamic*
- *Static* RAM (SRAM) consists of internal latches that store binary info. Stored info remains valid as long as power is applied to RAM
- *Dynamic* RAM (DRAM) stores the binary info in the form of electric charges on capacitor. Capacitors must be periodically recharged by refreshing the DRAM.

SRAM vs DRAM

- Refreshing the DRAM is done by cycling through the words every few milliseconds, reading and rewriting them to restore the decaying charge.
- DRAM offers reduced power consumption and larger storage capacity in a single memory chip – preferred choice
- SRAM is easier to use and has shorter read and write cycles.
- No refresh is required for SRAM

Main Memory Circuitry

- Main memory is made of ROM and RAM IC (chips), plus additional logic circuits.
- The logic circuits decode the address coming from the CPU to select the appropriate word to be transferred. Note that all devices share the common address, data and control buses of the CPU

Memory Space

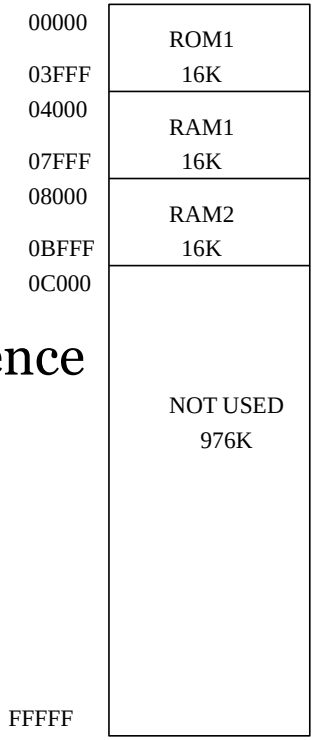
- Available (maximum) memory space (length) depends on address width
  - n-bit address gives 2^n word memory space
  - Intel 8086 has 20-bit address bus giving up to 1M memory space
- Memory space usually specified in bytes
  - Intel Core 2 Duo has 36-bit address and 64-bit data giving 64G word memory space
    - Each word is 8B (64-bit) giving up to 512GB memory space

Memory Mapping

- Ideally, we have one memory chip to occupy the whole available memory space, however
  - memory chip are not that huge
  - unrealistic to spend on space not required
  - a system employs more than one type of memory chips (RAM, ROM) and memory-mapped IOs need to be spliced into the memory space
- Hence, we have more than one memory chip (may be of same or different capacity) in a system
- Deciding which location in which memory chip is being referred to by the full address (on address bus) is called Memory Mapping
  - where the device (chip) maps into the memory space

Memory Map (an Example)

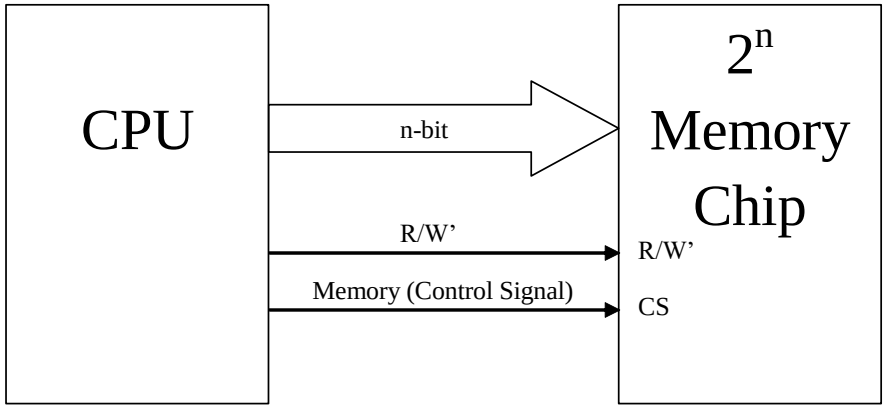
- 2^{14} = 16K memory chip
- Each memory has 14 address lines – directly connected
  - x – direct connection for internal selection
- Higher order bits (A14 – A19) are used for chip select, hence deciding the mapping address range



Device	Size	Pins	20-bit Address Bus					Address Range
ROM1	16K	14	0000	00xx	xxxx	xxxx	xxxx	00000 – 03FFF
RAM1	16K	14	0000	01xx	xxxx	xxxx	xxxx	04000 – 07FFF
RAM2	16K	14	0000	10xx	xxxx	xxxx	xxxx	08000 – 0BFFF

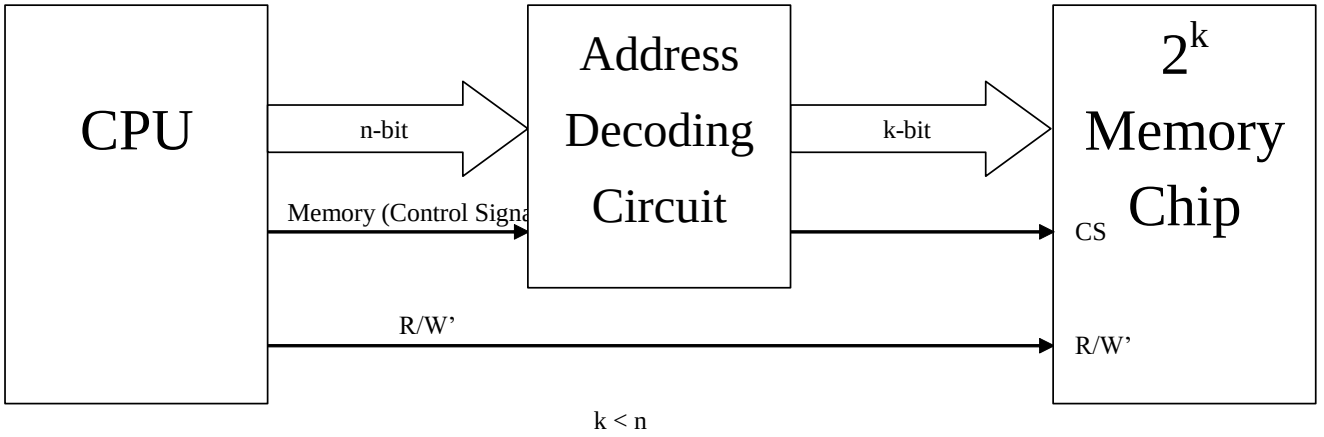
Memory Address Decoding - 1

- Address decoding is determining the correct location within the correct device from the address
  - It is achieved through the process of generating chip select (CS) signals from the address bus for each device in the system
- No decoding is required in one memory chip system (ideal case)



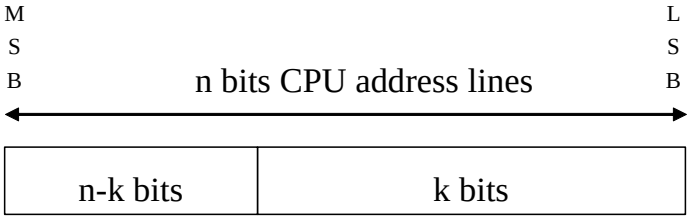
Memory Address Decoding - 2

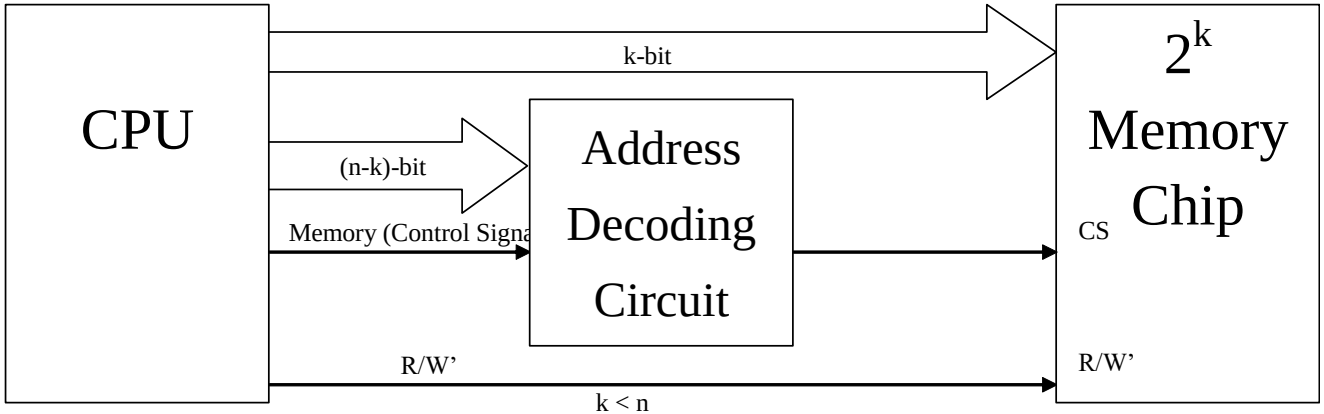
- In multiple memory chip system, or system using memory chip smaller than the memory space, address decoding logic circuit is used in between the CPU and memory devices



Address Decoding Format

- Address decoding is achieved by dividing the full address into two parts:
  - Lower significance k bits directly connect to the address lines of the memory chip to select the individual location within it
  - Upper significance (n–k) bits are used to generate CS signal for the memory chip





Address Decoding Illustration - 1

- Example: CPU has 20-bit address bus, 8-bit data bus, three 16KB memory chips to be used.
  - Total space: 2^{20} = 1MB or 1024KB
    - 20-bit: A0 – A19
    - 16KB = 2^{14} B, has 14-bit address lines
  - 14-bit: A0 – A13
    - Remaining 6-bit: A14 – A19 to be used for address decoding
    - A19 || A14 can take values from 000000b to 111111b (64 values)
    - Each 16KB memory chip uses one of the values within the above range
      - Can have up to 64 memory chips of 16KB
    - Values of A19 || A14 for the chip determine the start and end addresses for the chip

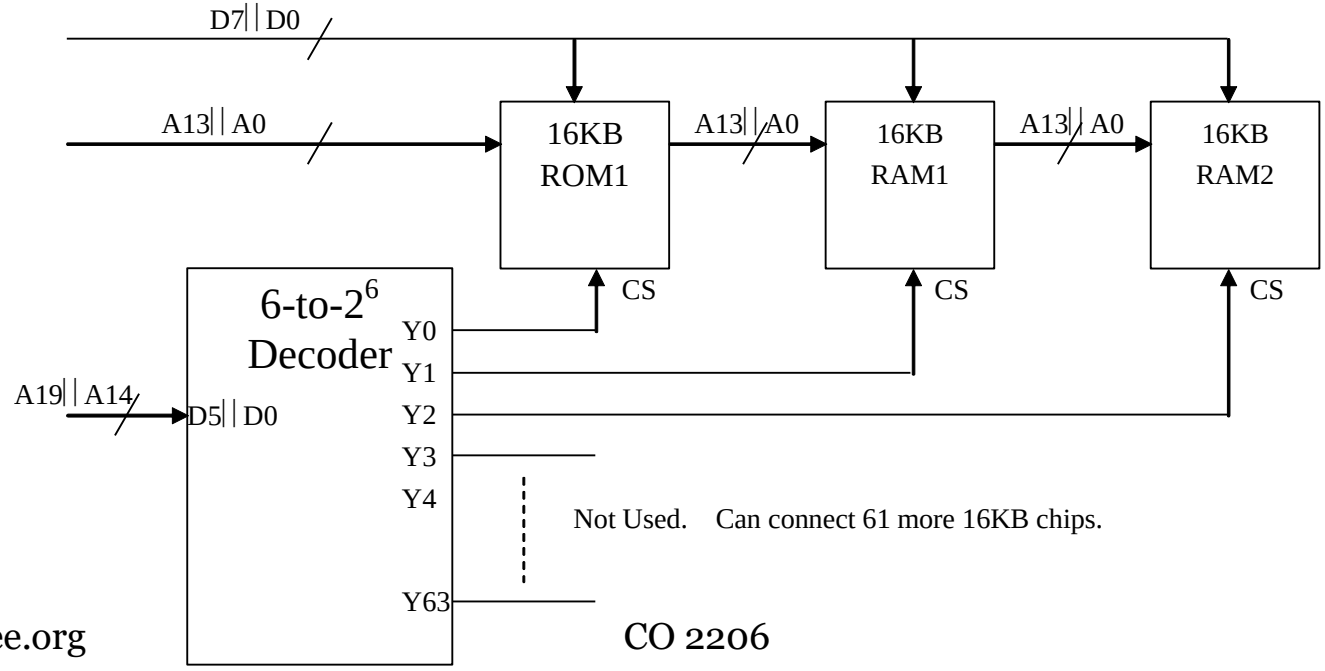
Address Decoding Illustration - 2

- Refer to example mapping below:
  - A19 || A14 = 000000b for ROM1 giving address range 00000h – 03FFFh
  - A19 || A14 = 000001b for RAM1 giving address range 04000h – 07FFFh
  - A19 || A14 = 000010b for RAM2 giving address range 08000h – 0BFFFh
- Choice of mapping can be different, e.g. we may want ROM1 to start at a different address

Device	Size	Pins	20-bit Address Bus						Address Range
			A19	A14	A13			A0	
ROM1	16K	14	0000	00	xx	xxxx	xxxx	xxxx	00000 – 03FFF
RAM1	16K	14	0000	01	xx	xxxx	xxxx	xxxx	04000 – 07FFF
RAM2	16K	14	0000	10	xx	xxxx	xxxx	xxxx	08000 – 0BFFF

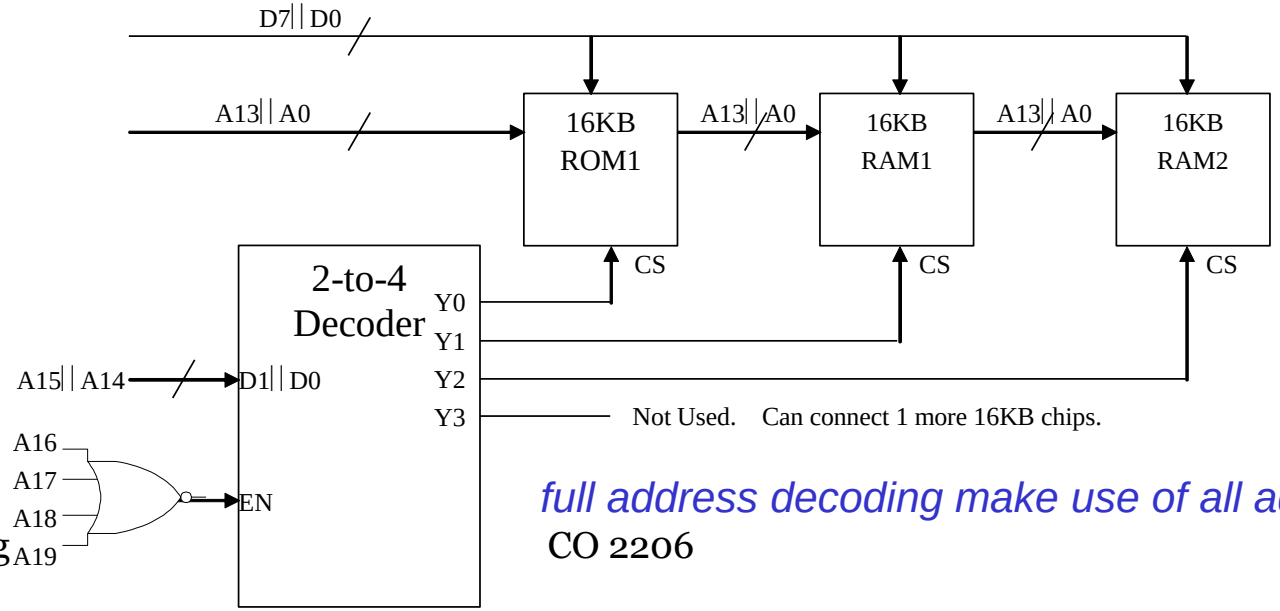
Full Address Decoding - 1

- Implementation of the address decoding in the previous slide can be done using a decoder
  - Input to decoder is A19||A14 (6-bit), hence a 6-to-2^6 decoder
  - Outputs from decoder connect to CS of the corresponding memory chips, e.g. ROM1 has A19||A14=000000b (0), hence output 0 of decoder connects to CS of ROM1



Full Address Decoding - 2

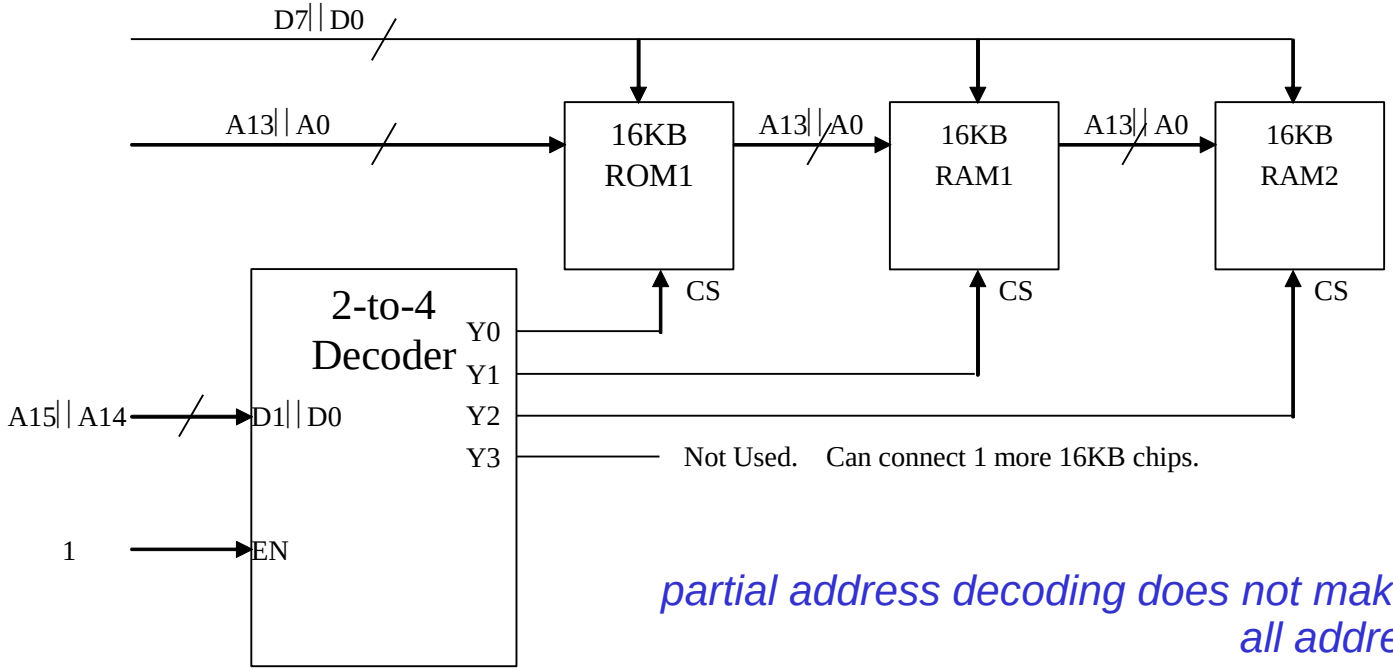
- If there are only 3 chips as in previous slide, it is not necessary to use 6x64 decoder
  - Only A15 || A14 (2-bit) are used for selecting the device, A19 ||A16 are always zero
  - Use 2x4 decoder
    - A15 || A14 are inputs to the decoder
    - A19 ||A16 are NORed to enable the decoder when they are 0s



full address decoding make use of all address lines

Partial Address Decoding

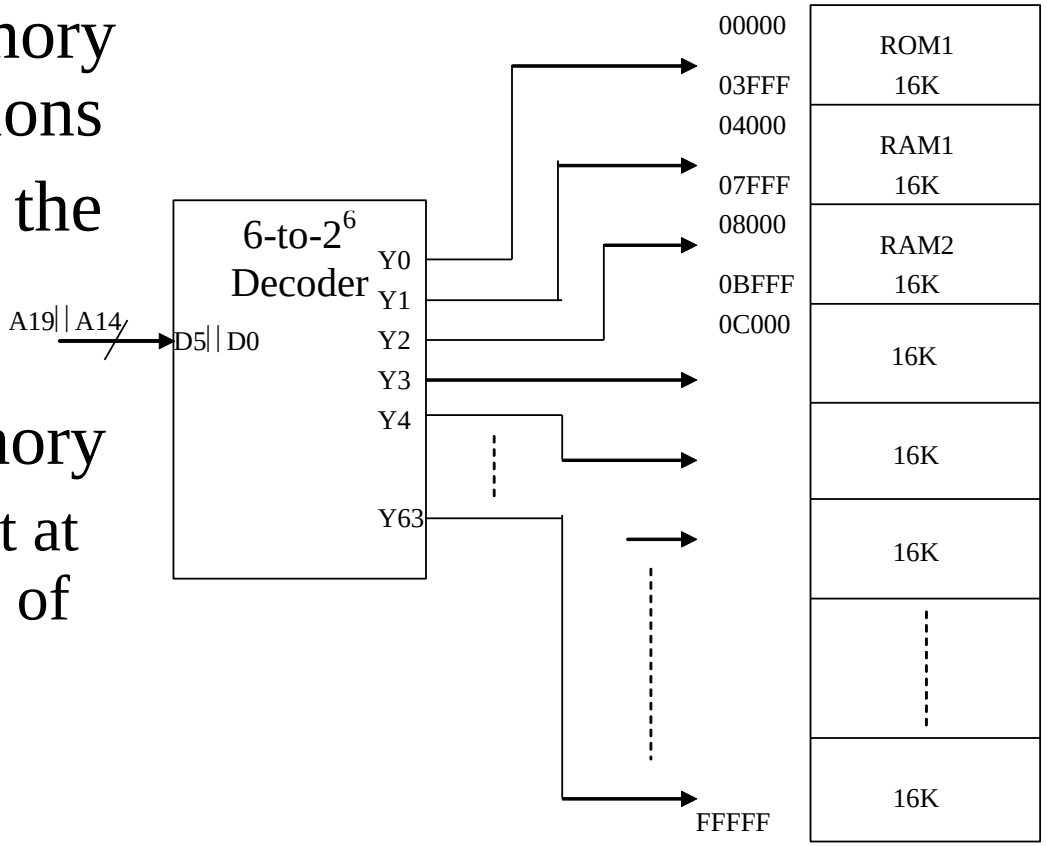
- If it is not important to fix A19||A16 to 0s, they can be ignored to give partial address decoding
  - Each memory location can be accessible with few different addresses



partial address decoding does not make use of all address lines

Address Decoding using Decoder: Summarized - 1

- Decoder divides memory space into equal sections
- Output connecting to the CS of the memory decides the start/end addresses of the memory
  - E.g. if ROM1 to start at 0C000h, connect Y3 of the decoder to CS of ROM1



Address Decoding using Decoder: Summarized - 2

- Address decoding for different sizes of memory chip in a system requires use of logic gates and decoders of different sizes
- EN and outputs of the decoder, and CS of the memory chip may be active-low
  - Same design concept, with care on inversion (e.g. use OR instead of NOR, use NOT when necessary) of the signals
- Address decoding for IO ports (registers) follow the same concept
  - IO ports has less address lines (few registers inside)
  - Only few lines going to the port, while most of the higher order bits are used for address decoding

Summary

- Memory hierarchy: distance to processor, access time, cost, capacity
- Key characteristics: as above, physical characteristics
- Memory chip structure: data lines, address lines, control signal
- SRAM vs DRAM: SRAM fast, costly; DRAM slower, lower cost, require refresh
- Memory mapping: partitioning memory space into different sections/memory chips
- Memory address decoding: determining which physical memory location (in which memory chip) from the given address