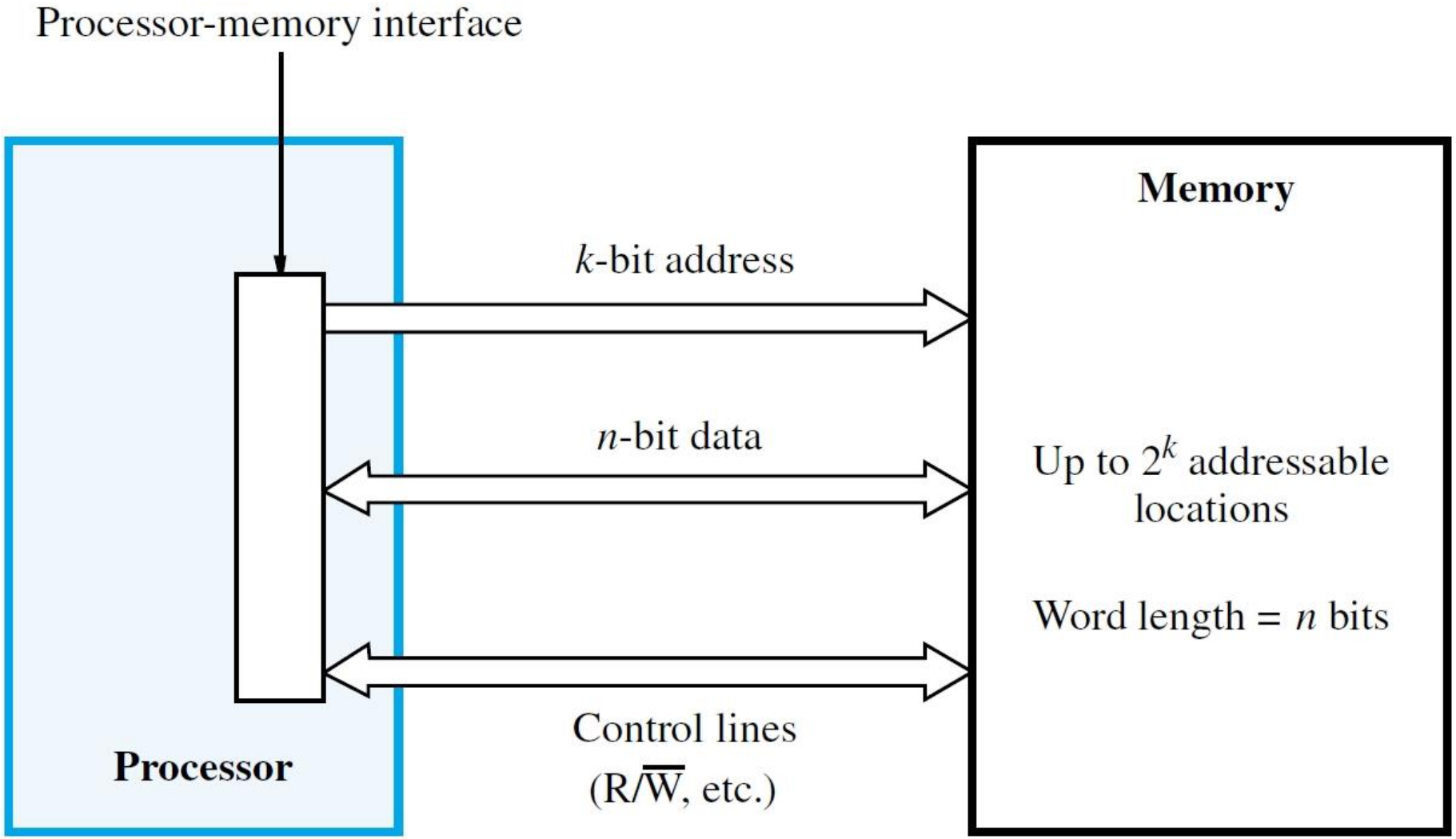
Chapter 8
The Memory System
Part I

Chapter Outline

- Basic memory circuits & memory organization
- Memory technology

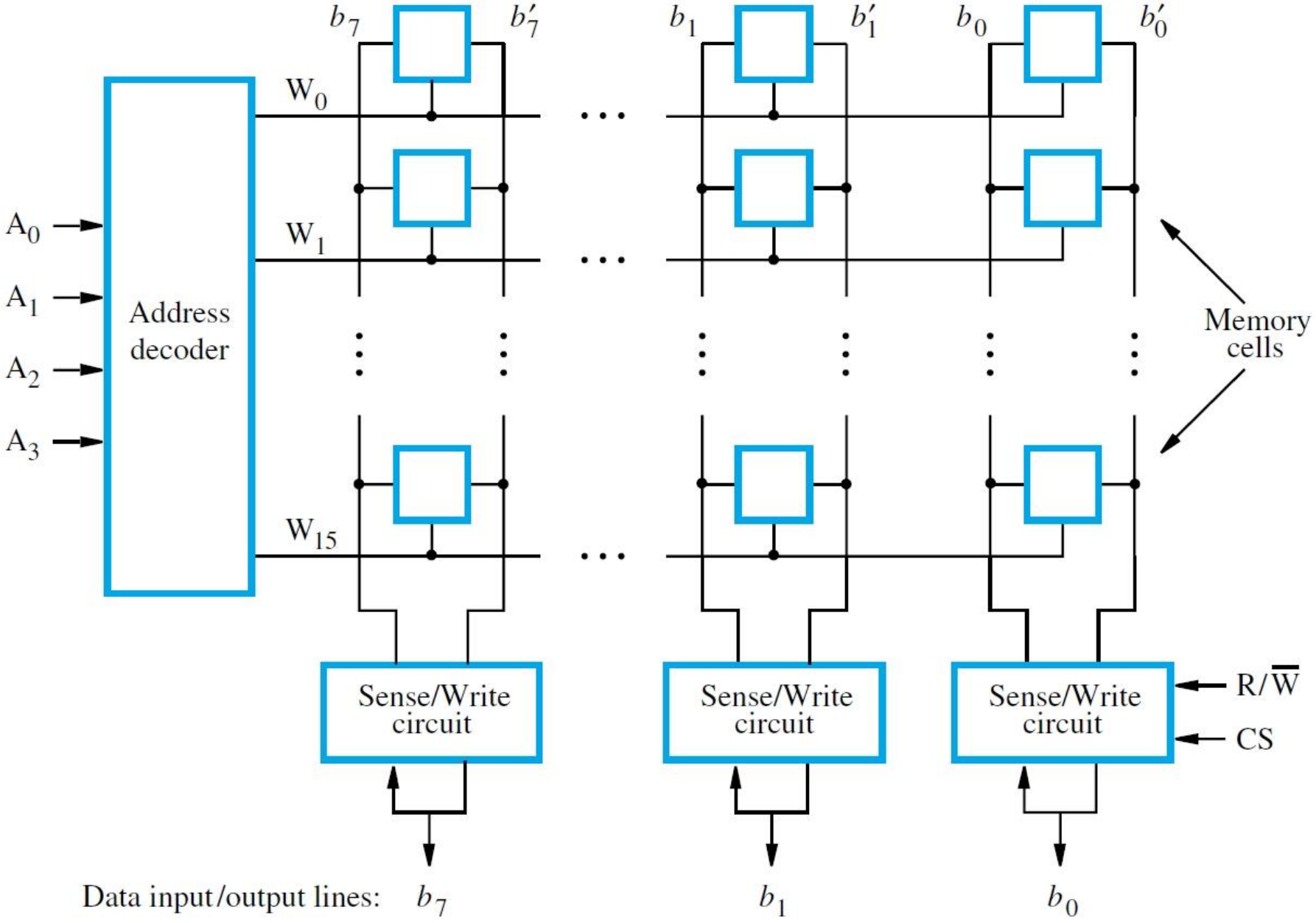
Basic Concepts

- Access provided by processor-memory interface
- *Address and data lines*, and also *control lines* for command (Read/Write), timing, data size
- *Memory access time* is time from initiation to completion of a word or byte transfer
- *Memory cycle time* is minimum time delay between initiation of successive transfers
- *Random-access memory (RAM)* means that access time is same, independent of location


Processor-memory interface
Memory
k-bit address
n-bit data
Up to 2^k addressable locations
Word length = n bits
Control lines
(R/W, etc.)
Processor

Semiconductor RAM Memories

- Memory chips have a common organization
- *Cells* holding single bits arranged in an array
- *Words* are rows; cells connected to *word lines* (cells per row ≠ bits per processor word)
- Cells in columns connect to *bit lines*
- Sense/Write circuits are interfaces between internal bit lines and data I/O pins of chip
- Typical control pin connections include Read/Write command and chip select (CS)

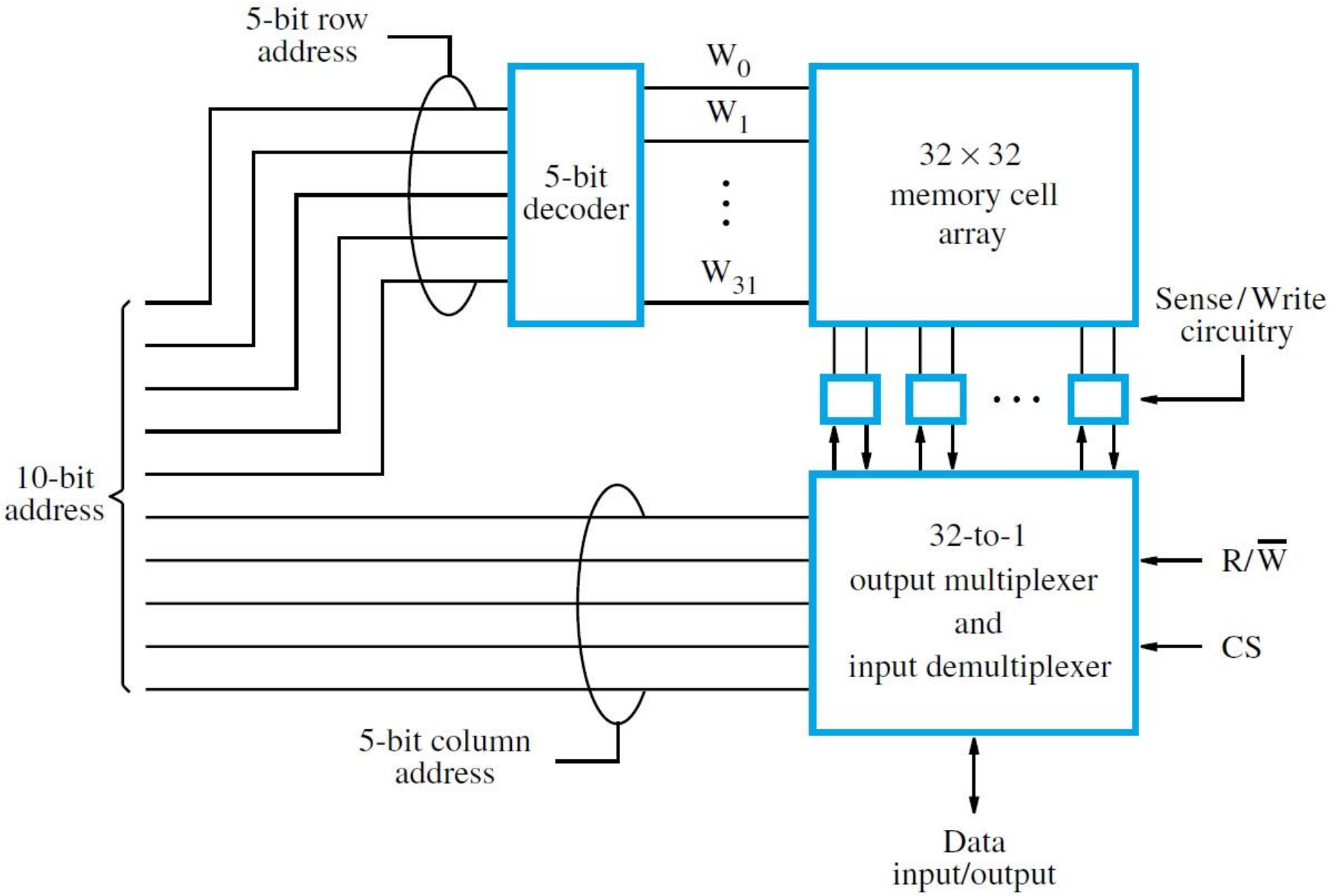

b_7
b'_7
b_1
b'_1
b_0
b'_0
W_0
W_1
W_{15}
A_0
A_1
A_2
A_3
Address decoder
Memory cells
Sense/Write circuit
Sense/Write circuit
Sense/Write circuit
R/$\overline{W}$
CS
Data input/output lines: b_7
b_1
b_0

Internal Organization and Operation

- Example of 16-word $\times$ 8-bit memory chip has *decoder* to select word line from 4-bit address
- Two complementary bit lines for each data bit
- External source provides stable address bits, and asserts chip-select input with command
- For Read operation, Sense/Write circuits transfer data from selected row to I/O pins
- For Write operation, Sense/Write circuits transfer data from I/O pins to selected cells

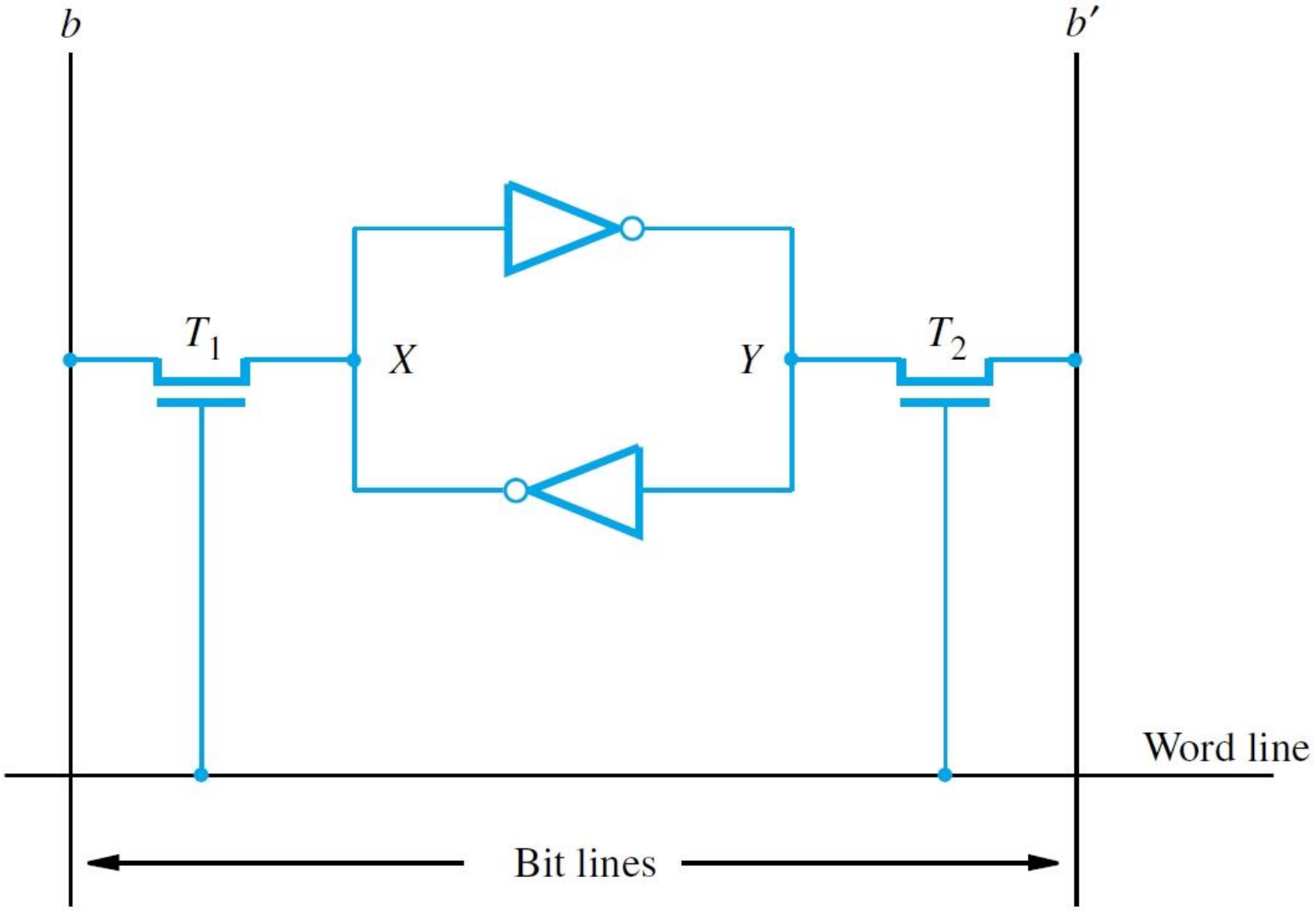
More on Chip Organization

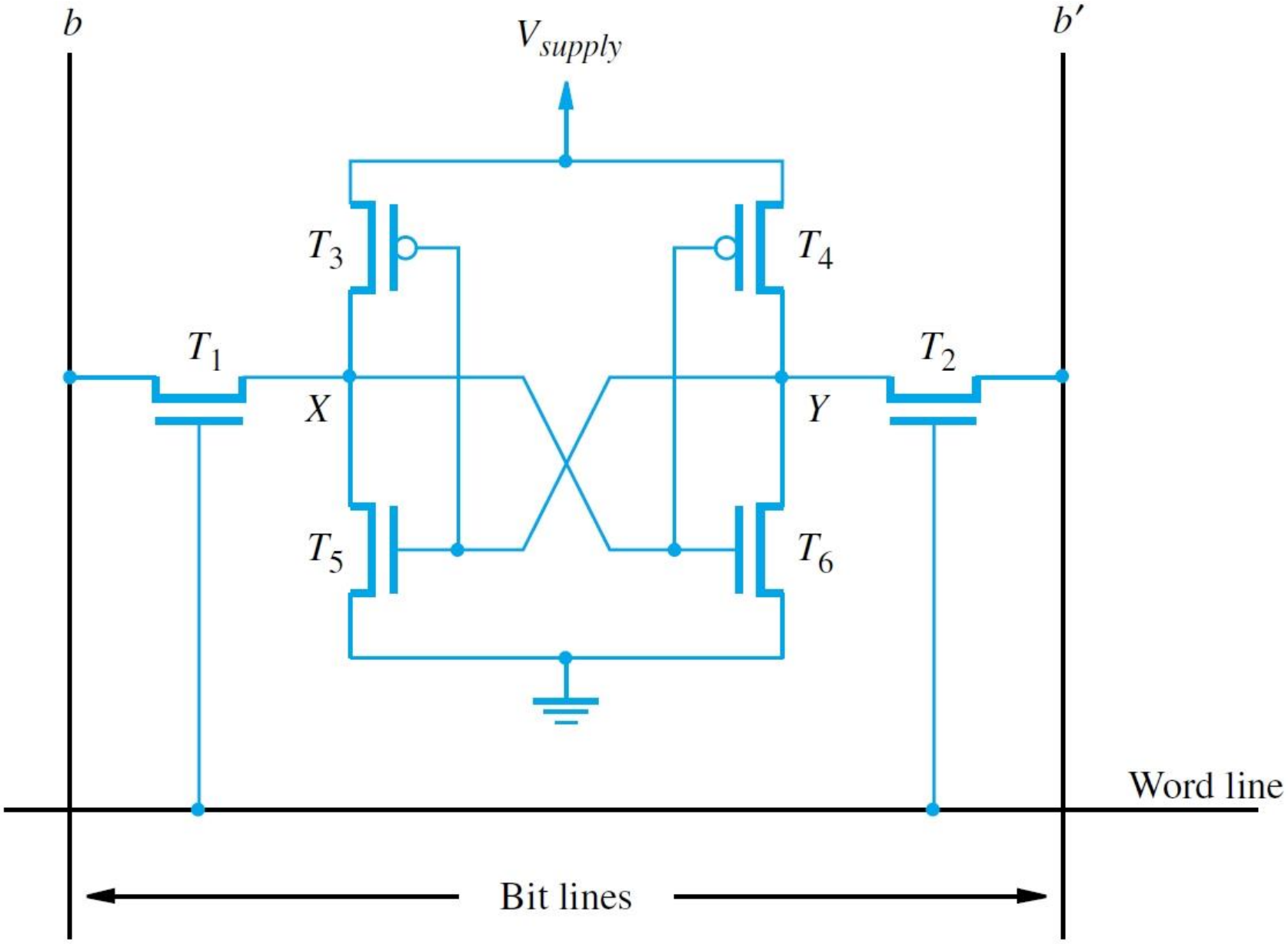
- Preceding 16×8 chip has only 128 storage cells
- With power/ground, need 4+8+2+2=14 pins
- Larger chips: similar organization, more pins
- For example, a chip with 1024 cells could be organized as 128×8 (7+8+2+2=19 pins) or alternatively as 1024×1 (10+1+2+2=15 pins)
- Pin count dictates cost, so consider 1024×1
- Use 32×32 array & divide address bits to have 5 upper bits for row, 5 lower bits for column


5-bit row address
5-bit decoder
W_0
W_1
W_{31}
32 × 32 memory cell array
Sense/Write circuitry
10-bit address
32-to-1 output multiplexer and input demultiplexer
R/$\overline{W}$
CS
5-bit column address
Data input/output

Static RAMs and CMOS Cell

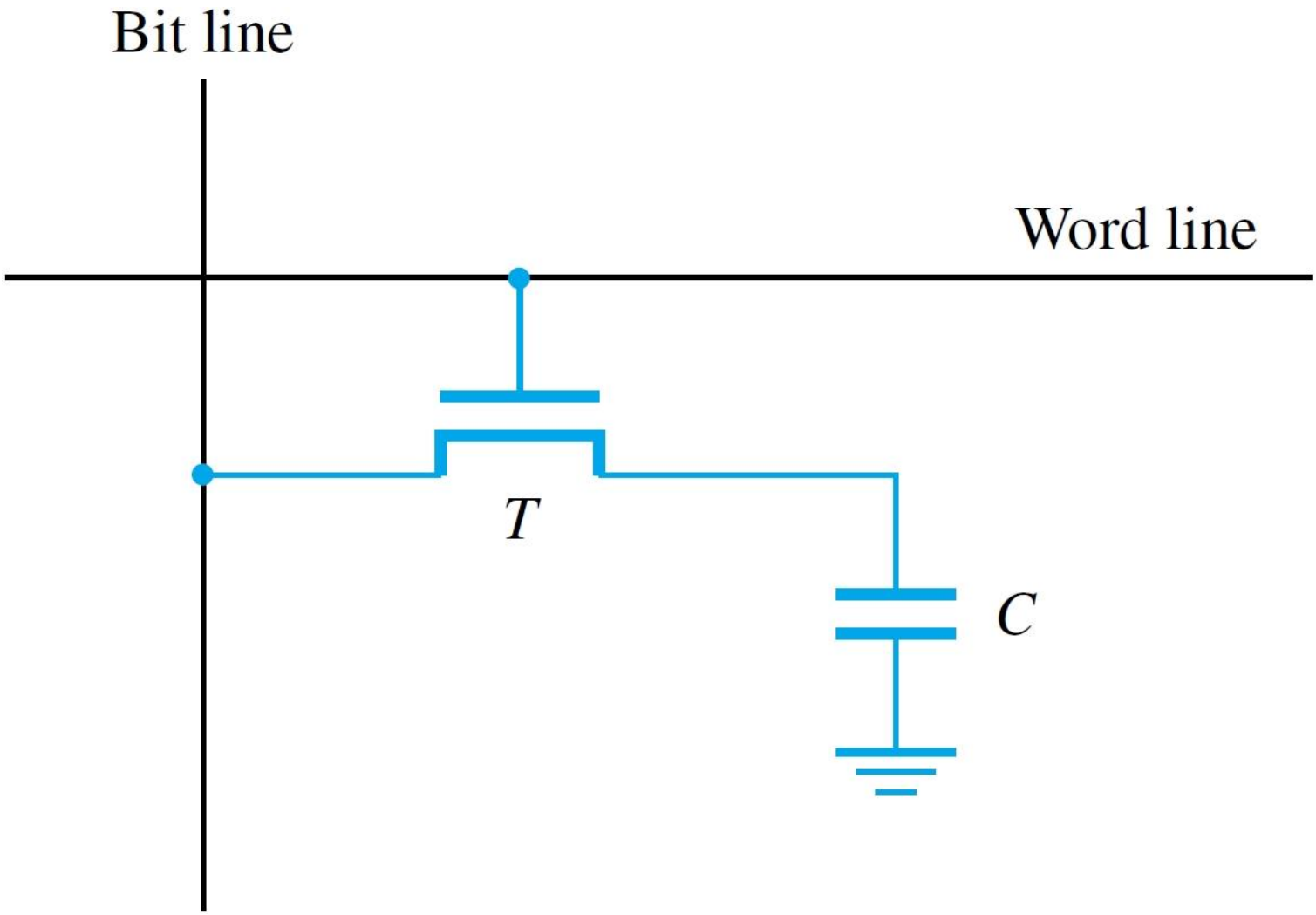
- Static memories need power to retain state; usually has short access time (few nanosecs.)
- A *static RAM cell* in a chip consists of two cross-connected inverters to form a *latch*
- Chip implementation typically uses *CMOS* cell whose advantage is low power consumption
- Two transistors controlled by word line act as switches between the cell and the bit lines
- To write, bit lines driven with desired data


b
b′
T1
X
Y
T2
Word line
Bit lines


b
b′
V_{supply}
T_3
T_4
T_1
T_2
X
Y
T_5
T_6
Word line
Bit lines

Dynamic RAMs

- Static RAMs have short access times, but need several transistors per cell, so density is lower
- Dynamic RAMs are simpler for higher density and lower cost, but access times are longer
- Density/cost advantages outweigh slowness
- Dynamic RAMs are widely used in computers
- Cell consists of a transistor and a capacitor
- State is presence/absence of capacitor charge
- Charge leaks away and must be *refreshed*

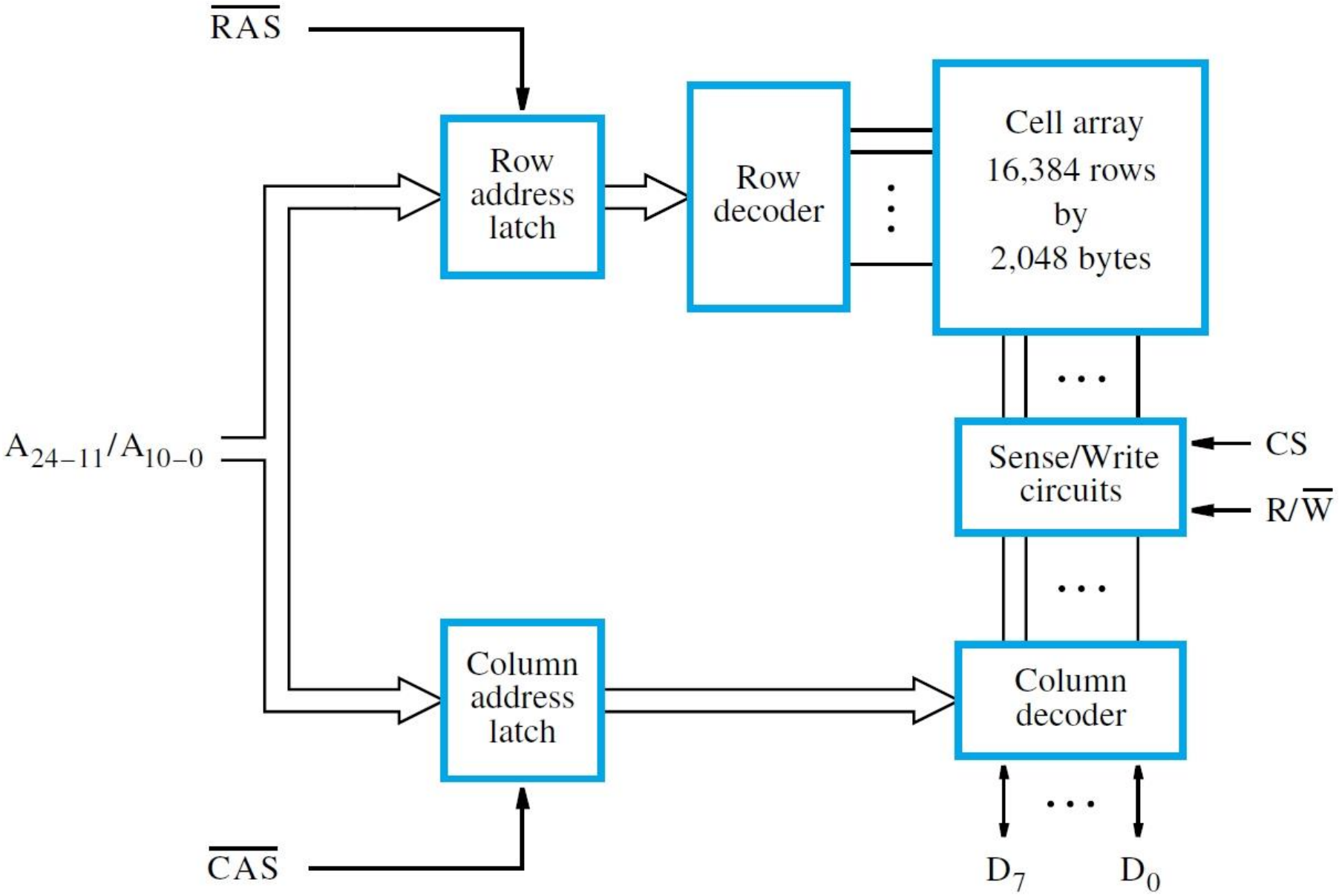

Bit line
Word line
T
C

Dynamic RAM Chip Operation

- Reflects general principles of chip operation
- For Read, charge from cells in selected row is checked by *sense amplifiers* on bit lines
- 1 or 0 if charge is above or below threshold
- Action of sensing the bit lines also causes refresh of charge in all cells of selected row
- For Write, access the row and drive bit lines to alter amount of charge in subset of cells
- Refresh rows periodically to maintain charge

More on Dynamic RAM Chips

- Consider 32M × 8 chip with 16K × 16K array
- 16,384 cells per row organized as 2048 bytes
- 14 bits to select row, 11 bits for byte in row
- Use *multiplexing* of row/column on same pins
- Row/column *address latches* capture bits
- Row/column *address strobe signals* for timing (asserted low with row/column address bit*s)*
- Asynchronous DRAMs: delay-based access, external controller refreshes rows periodically

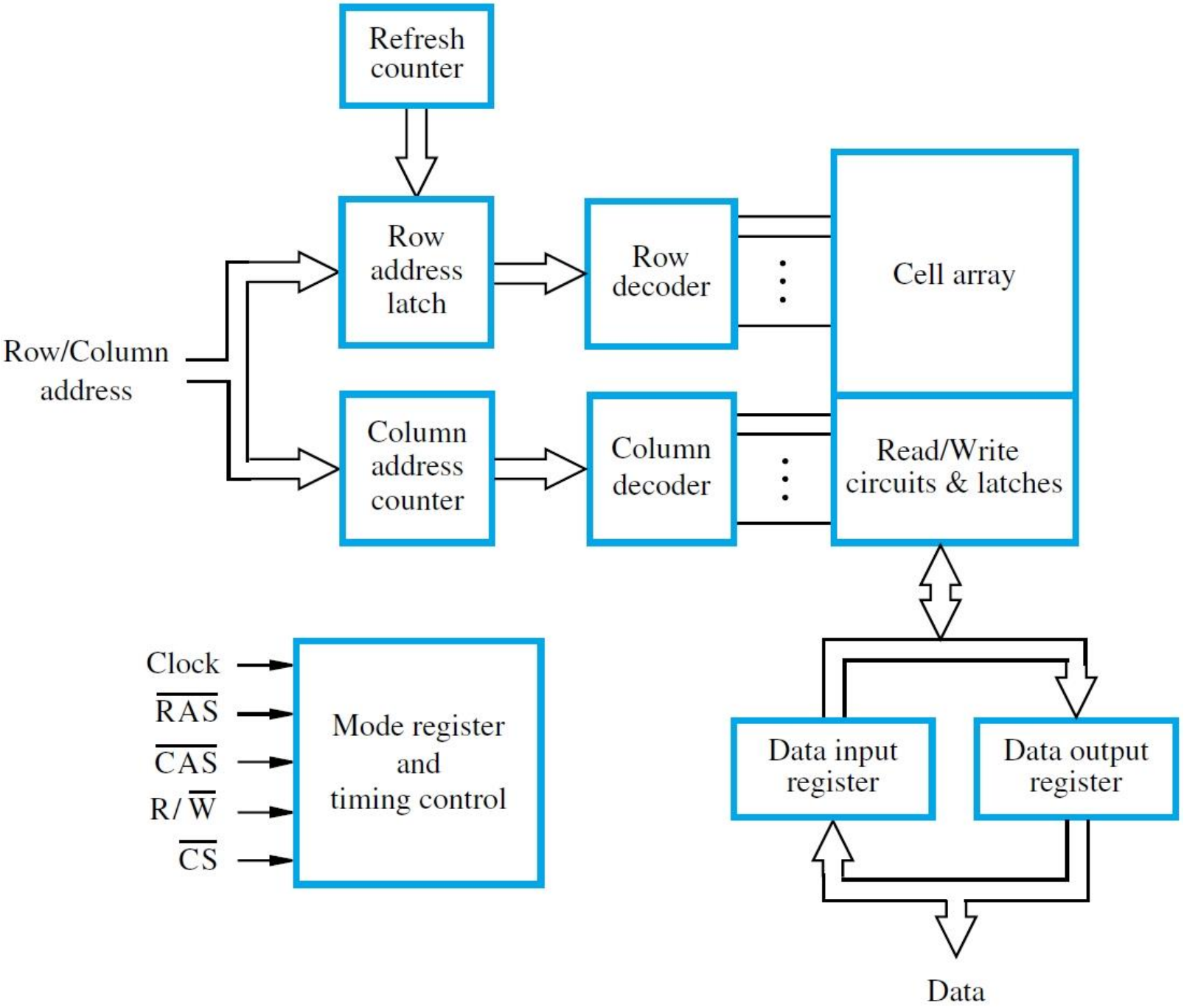

RAS
Row address latch
Row decoder
Cell array 16,384 rows by 2,048 bytes
A24–11/A10–0
Sense/Write circuits
CS
R/W
Column address latch
Column decoder
CAS
D7
D0

Fast Page Mode

- In preceding example, all 16,384 cells in a row are accessed (and also refreshed as a result)
- But only 8 bits of data are actually transferred for each full row/column addressing sequence
- For more efficient access to data in same row, *latches* in sense amplifiers hold cell contents
- For consecutive data, just assert CAS signal and increment column address in same row
- This fast page mode is useful in *block transfers*

Synchronous DRAMs

- In early 1990s, DRAM technology enhanced by including clock signal with other chip pins
- More circuitry also added for enhancements
- These chips are called synchronous DRAMs
- Sense amplifiers still have latching capability
- Additional benefits from internal buffering and availability of synchronizing clock signal
- Internal row counter enables built-in refresh instead of relying on external controller

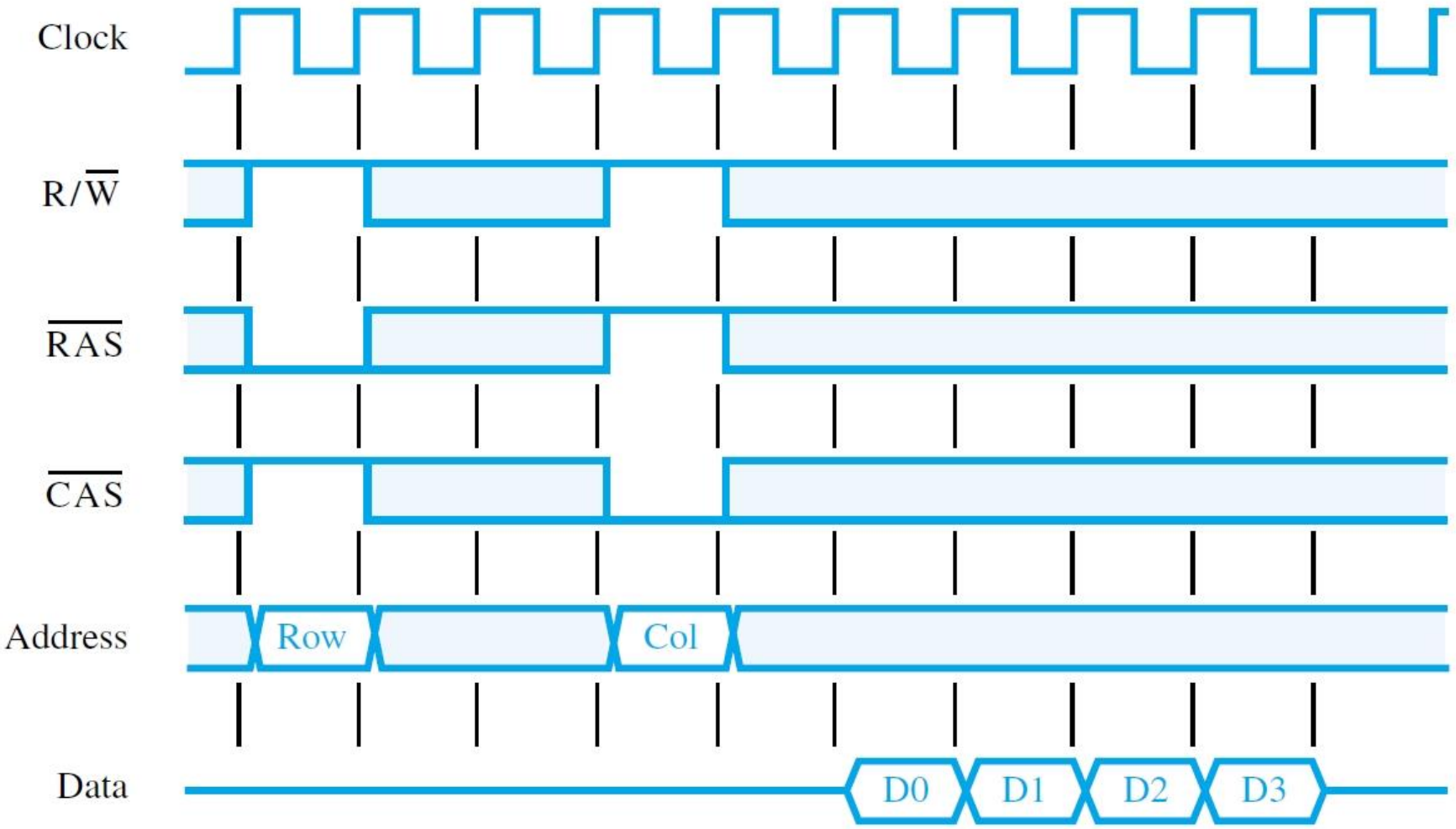

Refresh counter
Row address latch
Row decoder
Cell array
Row/Column address
Column address counter
Column decoder
Read/Write circuits & latches
Clock
RAS
CAS
R/W
CS
Mode register and timing control
Data input register
Data output register
Data

SDRAM Features

- Synchronous DRAM (SDRAM) chips include *data registers* as well as address latches
- New access operation can be initiated while data are transferred to or from these registers
- Also have more sophisticated control circuitry
- SDRAM chips require power-up configuration
- Memory controller initializes *mode* register
- Used to specify *burst length* for block transfers and also to set delays for control of timing

Efficient Block Transfers

- Asynchronous DRAM incurs longer delay from CAS assertion for *each* column address
- Synchronous DRAM reduces delay by having CAS assertion *once* for initial column address
- SDRAM circuitry increments column counter and transfers consecutive data automatically
- Burst length determines number of transfers
- Consider example with burst length of 4, RAS delay of 3 cycles, CAS delay of 2 cycles

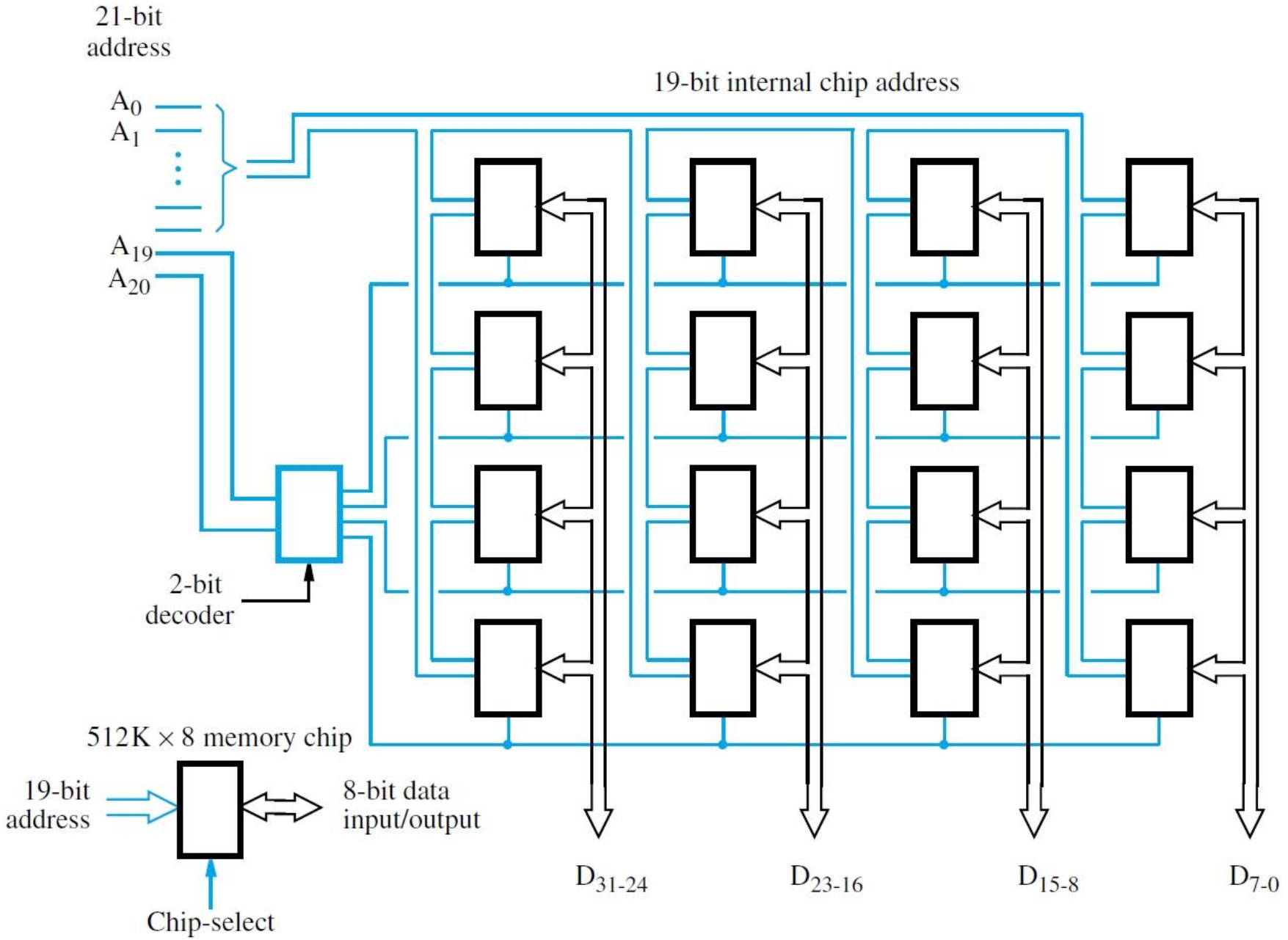

Clock
R/W
RAS
CAS
Address
Row
Col
Data
D0
D1
D2
D3

Double-Data-Rate (DDR) SDRAM

- Early SDRAMs transferred on rising clock edge
- Later enhanced to use rising and falling edges
- *Doubles* effective rate *after* RAS/CAS assertion
- Requires more complex clock/control circuitry
- Internal array access not significantly faster
- How can transfer rate be effectively doubled?
- *Interleave* consecutive data across two arrays
- Switch between arrays for each clock edge

Structure of Larger Memories

- Internal chip organization has been discussed
- Larger memories combine multiple chips
- How are these chips connected together?
  - Consider an example based on static memory
  - Memory size is 2M words, each 32 bits in size
  - Implement with 512K $\times$ 8 static memory chips
  - 4 chips for 32 bits, 4 groups of 4 chips for 2M


21-bit
address
19-bit internal chip address
A_0
A_1
A_{19}
A_{20}
2-bit
decoder
512K × 8 memory chip
19-bit
address
8-bit data
input/output
Chip-select
D_{31-24}
D_{23-16}
D_{15-8}
D_{7-0}

Address Decoder and Tri-state Pins

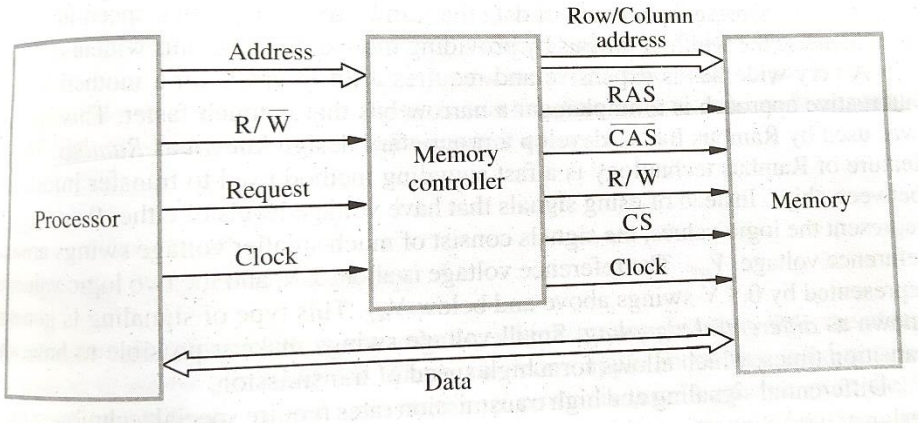
- 2M word-addressable memory needs 21 bits
- Each chip has only 19 address bits ($2^{19} = 512K$)
- Address bits A_{20} and A_{19} select one of 4 groups
- Outputs of 2-bit *decoder* drive chip-select pins
- Only the selected chips respond to request

- Shared data connections need *tri-state* circuits
- When a chip is not selected (i.e., CS input is 0), its I/O pins are *electrically disconnected*

Dynamic/Expandable Memory Systems

- DRAM chip capacity has grown over time with 2G bits/chip available now, and more in future
- Individual DRAM chips grouped into a module with aggregate capacity of 4 Gbytes or more
- Module socket interface is standardized – SIMM and DIMM
- Enables simple upgrades to increase capacity by replacing smaller modules with larger ones
- Printed-circuit board can have many sockets with common lines for address and data

Memory Controller



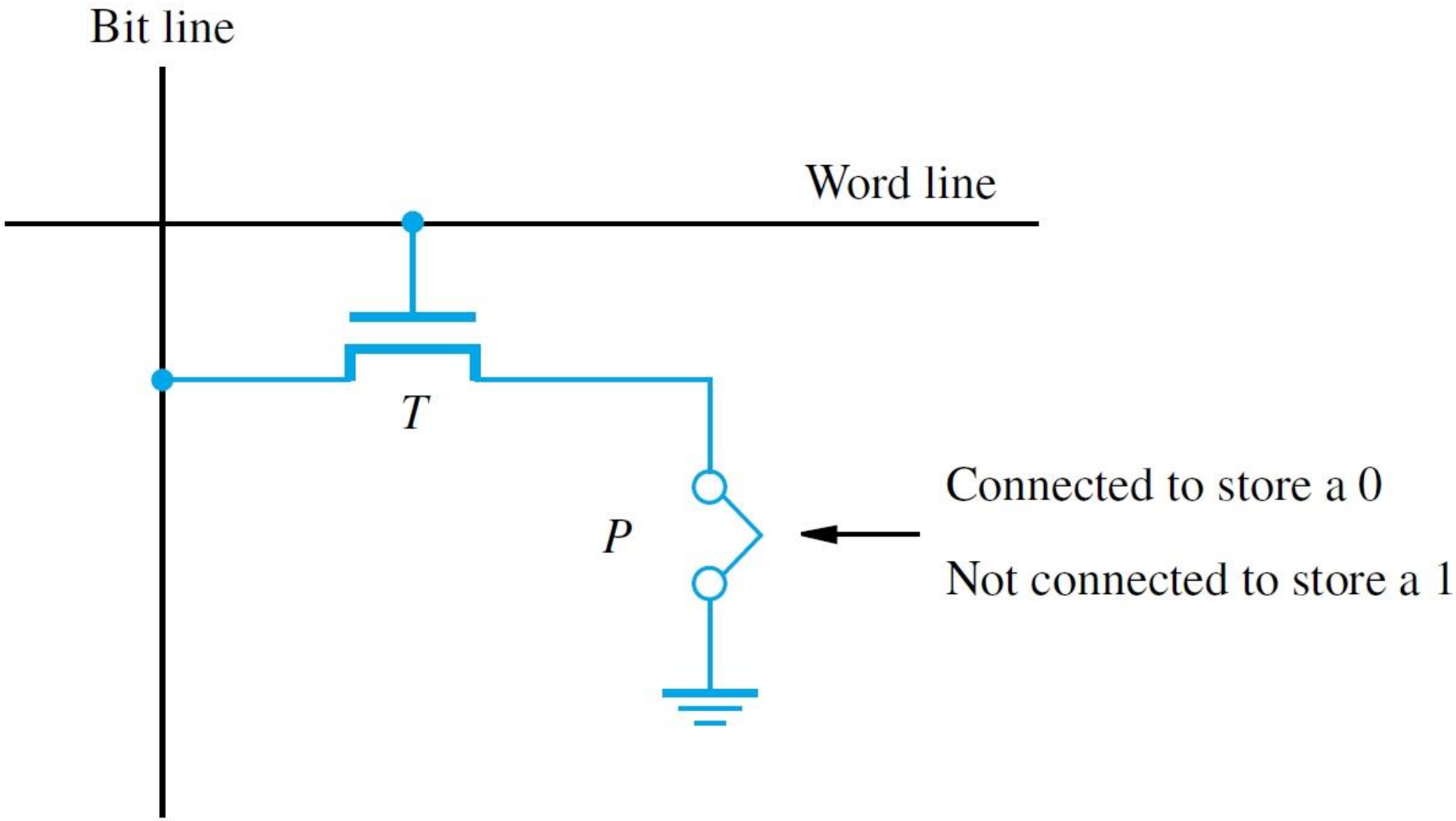
- Processor issues all address bits together, but DRAM chips need row/column multiplexing
- A *memory controller* handles this task and also asserts control signals for proper timing
- A large main memory is implemented using multiple DRAM modules sharing data lines
- Controller decodes high-order address bits to assert chip-select signal of only one module
- Other modules turn off their tri-state outputs

Read-only Memories

- Static and dynamic RAM chips are *volatile*, (information retained only when power is on)
- Some applications require information to be retained in memory chips when power is off
- Example: computers without disk drives
- Read-only memory (ROM) chips provide the nonvolatile storage for such applications
- Special writing process sets memory contents
- Read operation is similar to volatile memories

Basic ROM Cell

- A *read-only memory (ROM)* has its contents written only *once,* at the time of manufacture
- The basic ROM cell in such a memory contains a single transistor switch for the bit line
- The other end of the bit line is connected to the power supply through a resistor
- If the transistor is connected to ground, bit line voltage is near zero, so cell stores a 0
- Otherwise, bit line voltage is high for a 1


Bit line
Word line
T
P
Connected to store a 0
Not connected to store a 1

PROM, EPROM, and EEPROM

- Cells of a *programmable ROM (PROM)* chip may be written after the time of manufacture
- A fuse is burned out with a high current pulse
- An *erasable programmable ROM (EPROM)* uses a special transistor instead of a fuse
- Injecting charge allows transistor to turn on
- Erasure requires UV light to remove all charge
- An *electrically erasable ROM (EEPROM)* can have individual cells erased with chip in place

Flash Memory

- Flash memory is based on EEPROM cells
- For higher density, Flash cells are designed to be erased in larger blocks, not individually
- Writing individual cells requires reading block, erasing block, then writing block with changes
- Greater density & lower cost of Flash memory outweighs the inconvenience of block writes
- Widely used in cell phones, digital cameras, and solid-state drives (e.g., USB memory keys)

Sections to Read (From Hamacher's Book)

- Chapter on Memory System
  - All sections and sub-sections
    - Till Section 8.3.5

    (i.e., the entire handout shared in class ending in Flash Memory)