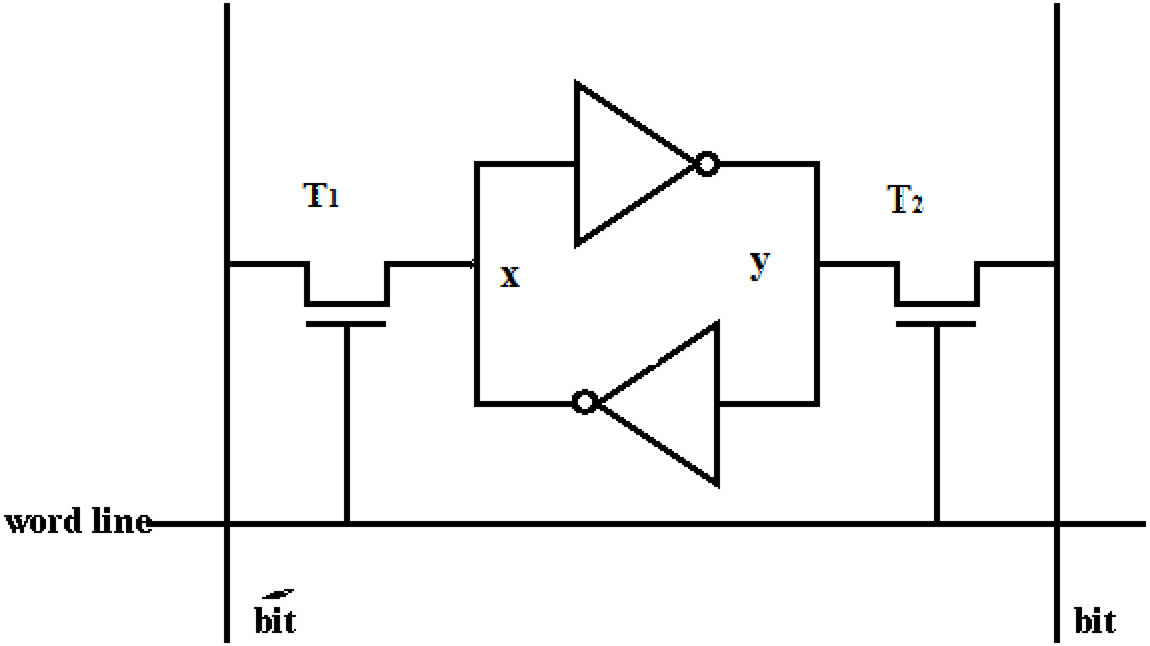
Computer Organization & Architecture

Static memories

- Able to retain their state as long as power is supplied
- Two inverters are cross-connected to form a latch
  - Latch is connected to two bit lines by transistors T1 andT2.
  - Transistors act as switches that can be opened or closed under control of word line
  - When word line is at ground level, transistors are turned off and latch retains its state
  - Ex., if logic value at point X is 1 and at point Y is 0, this state is maintained as long as signal on word line is at ground level
  - Assume that this state represents the value 1

Static memories



Static memories

- Read Operation:
  - In order to read state of SRAM cell, word line is activated to close switches T1 and T2
  - If cell is in state 1, signal on bit line *b* is high and signal on bit line *b’* is low
  - Opposite is true if the cell is in state 0
  - *b* and *b*’ are always complements of each other
  - Sense/Write circuit at end of two bit lines monitors their state and sets corresponding output accordingly

Static memories

- Write Operation:
  - Sense/Write circuit drives bit lines *b* and *b'*, instead of sensing their state
  - It places appropriate value on bit line *b* and its complement on *b'* and activates word line
  - This forces the cell into corresponding state, which the cell retains when word line is deactivated

Static memories

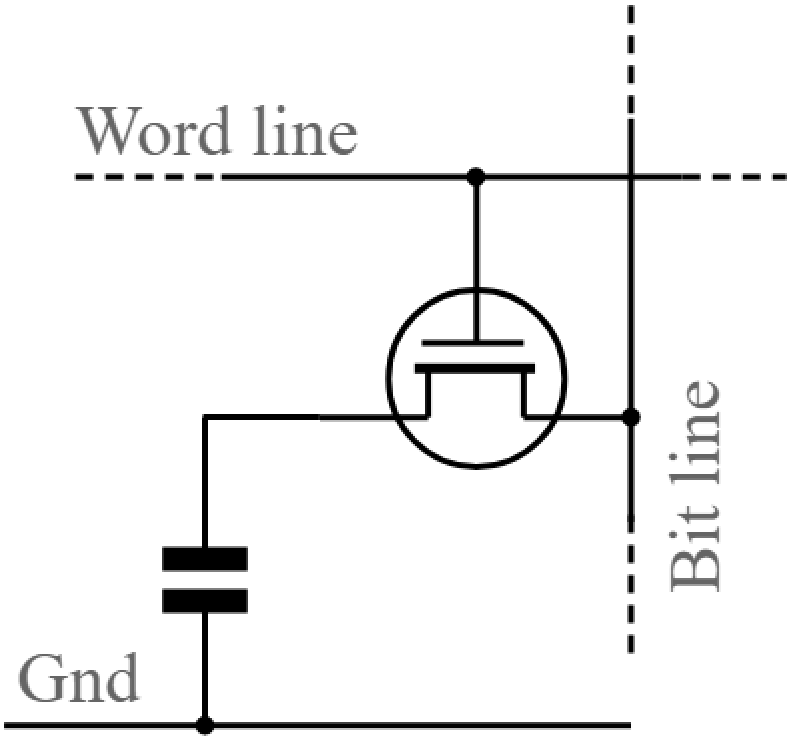
- Continuous power is needed for cell to retain its state
  - If power is interrupted, the cell’s contents are lost
  - Said to be volatile memories
- Static RAMs can be accessed very quickly
  - Access times on the order of a few nanoseconds are found in commercially available chips
  - Used in applications where speed is of critical concern

Dynamic RAMs

- Static RAMs are fast, but their cells require several transistors
- Less expensive and higher density RAMs can be implemented with simpler cells
  - Memories that use such cells are called dynamic RAMs (DRAMs)
  - But, do not retain their state for a long period, unless they are accessed frequently for Read or Write operations

Dynamic RAMs

- Information is stored in a dynamic memory cell in form of a charge on a capacitor
  - But this charge can be maintained for only tens of milliseconds
- Since cell is required to store information for a much longer time:
  - Its contents must be periodically refreshed by restoring capacitor charge to its full value
  - This occurs when contents of cell are read or when new information is written into it


Word line
Bit line
Gnd

Dynamic RAMs

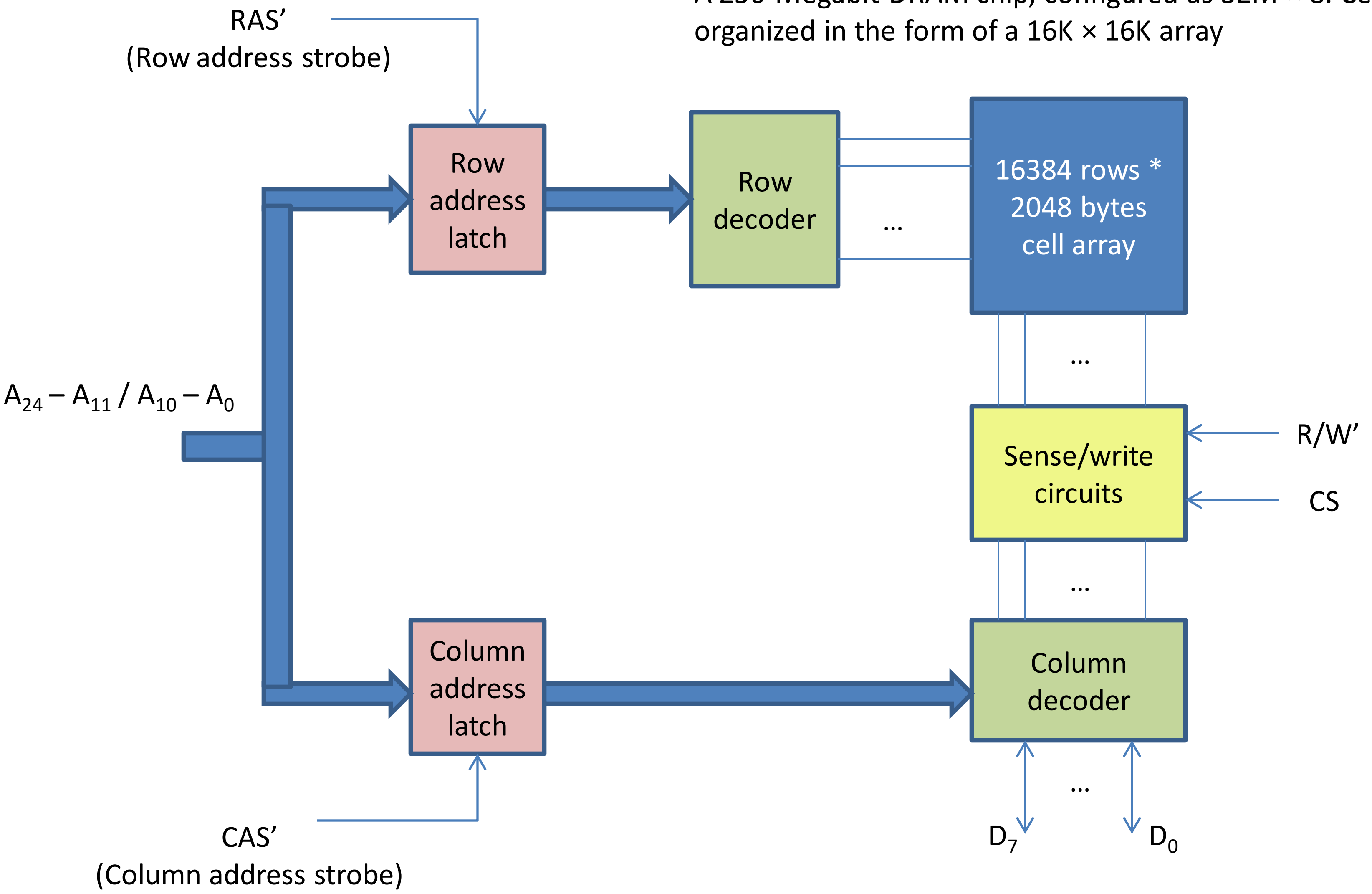
- To store information in this cell:
  - Transistor is turned on and an appropriate voltage is applied to the bit line
  - This causes a known amount of charge to be stored in capacitor
  - After transistor is turned off, charge remains stored in capacitor, but not for long; capacitor begins to discharge
  - This is because transistor continues to conduct a tiny amount of current, measured in pico amperes, after it is turned off
  - Information stored in cell can be retrieved correctly only if it is read before charge in capacitor drops below some threshold value

Dynamic RAMs

- During a Read operation:
  - Transistor in a selected cell is turned on
  - Sense amplifier connected to bit line detects whether charge stored in capacitor is above or below threshold value
  - If charge is above threshold, sense amplifier drives bit line to full voltage representing logic value 1
    - Capacitor is recharged to full charge corresponding to logic value1
  - If charge in capacitor is below threshold value, sense amplifier pulls bit line to ground level to discharge capacitor fully
  - Reading contents of a cell automatically refreshes its contents
    - Since word line is common to all cells in a row, all cells in a selected row are read and refreshed at same time

Internal organization of dynamic memory chip

A 256-Megabit DRAM chip, configured as 32M × 8. Cells are organized in the form of a 16K × 16K array



Fast page mode

- When DRAM is accessed, contents of all cells in selected row are sensed, but only 8 bits are placed on data lines
  - This byte is selected by column address
- A simple addition to circuit makes it possible to access other bytes in same row without having to reselect row
  - Each sense amplifier also acts as a latch
  - When a row address is applied, contents of all cells in selected row are loaded into the corresponding latches
  - Then, it is only necessary to apply different column addresses to place different bytes on data lines

Fast page mode

- All bytes in selected row can be transferred in sequential order:
  - By applying a consecutive sequence of column addresses under control of successive CAS signals
- A block of data can be transferred at a much faster rate than can be achieved for transfers involving random addresses
  - Block transfer capability is referred to as **fast page mode** feature
  - A large block of data is often called a page

Synchronous DRAMs

- DRAMs whose operation is synchronized with a clock signal are known as synchronous DRAMs (SDRAMs)
  - Cell array is the same as in asynchronous DRAMs
- Distinguishing feature of an SDRAM is use of a clock signal
  - Availability of which makes it possible to incorporate control circuitry on chip that provides many useful features
  - Ex., SDRAMs have built-in refresh circuitry, with a refresh counter to provide addresses of rows to be selected for refreshing
  - As a result, dynamic nature of these memory chips is almost invisible to user

Synchronous DRAMs

- Address and data connections of an SDRAM may be buffered by means of registers
- Internally, the Sense/Write amplifiers function as latches, as in asynchronous DRAMs
    - Read operation causes contents of all cells in selected row to be loaded into these latches
    - Data in latches of selected column are transferred into data register, thus becoming available on data output pins

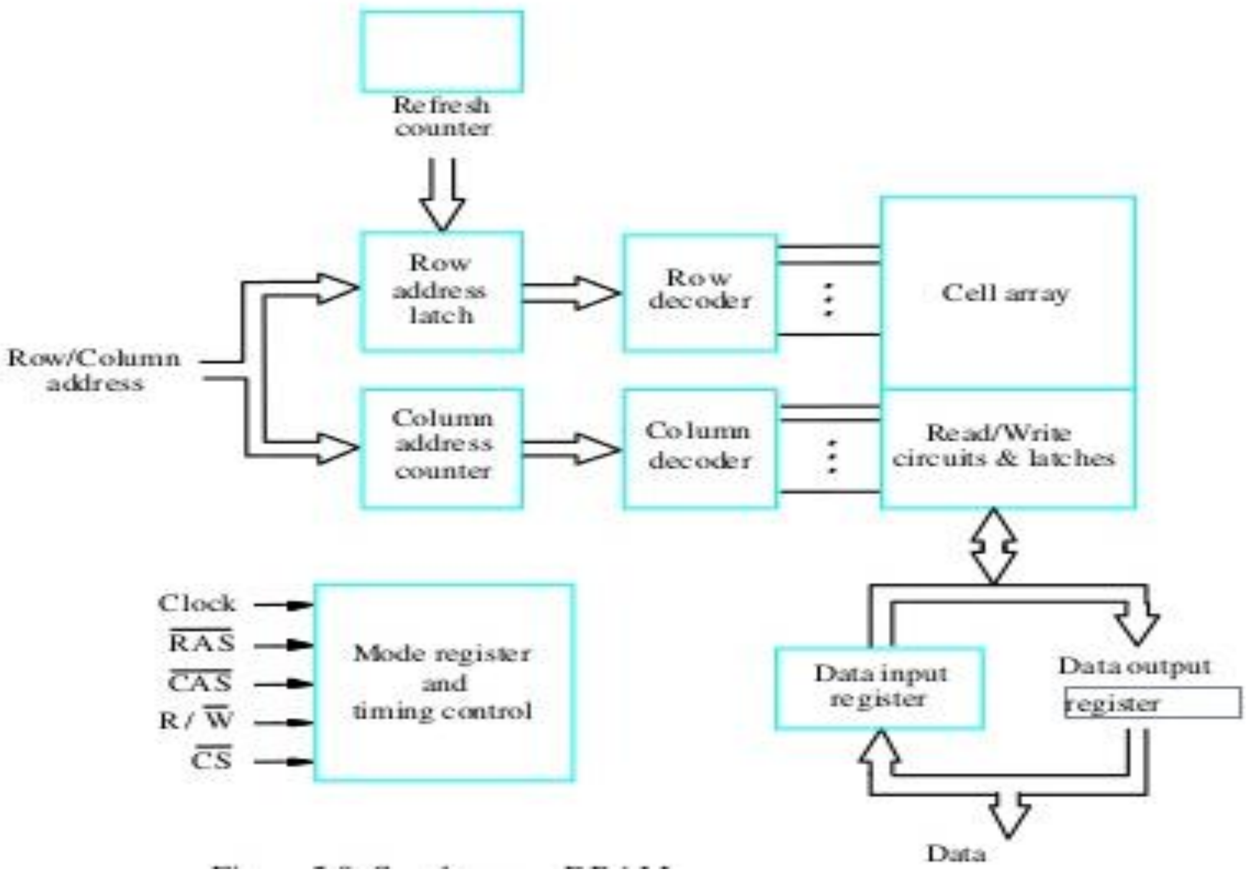
Synchronous DRAMs

- Buffer registers are useful when transferring large blocks of data at very high speed
  - Isolate external connections from chip’s internal circuitry
  - Becomes possible to start a new access operation while data are being transferred to or from registers

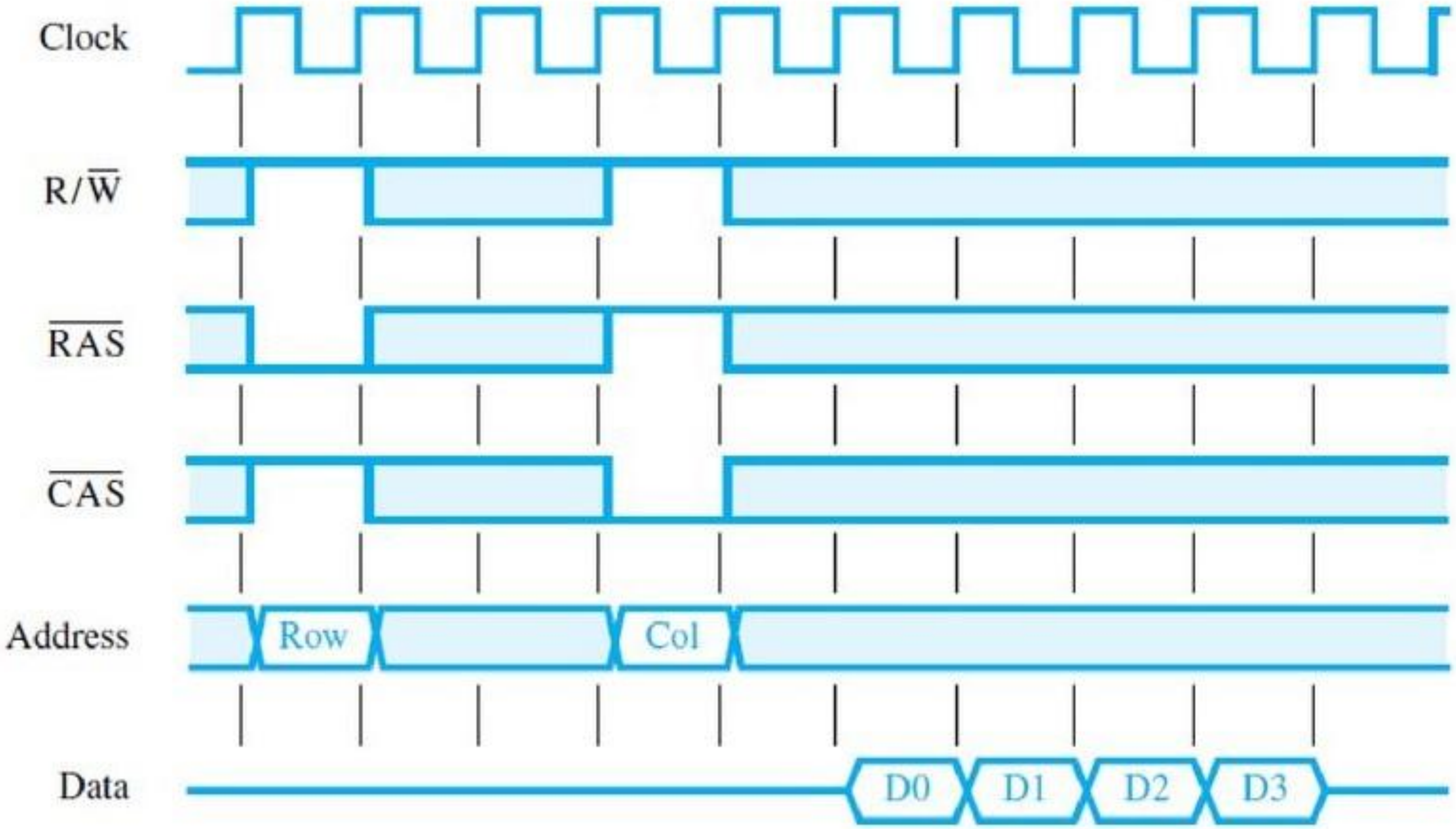
Synchronous DRAMs

- SDRAMs have several different modes of operation:
  - Can be selected by writing control information into a mode register
  - Ex., burst operations of different lengths can be specified
- It is not necessary to provide externally-generated pulses on the CAS line to select successive columns
  - Necessary control signals are generated internally using a column counter and clock signal
  - New data are placed on data lines at rising edge of each clock pulse

Synchronous DRAMs



Synchronous DRAMs



Burst read of length 4 in an SDRAM

Synchronous DRAMs

- Synchronous DRAMs can deliver data at a very high rate:
  - Because all the control signals needed are generated inside the chip
- Today’s SDRAMs operate with clock speeds that can exceed 1 GHz

Latency and bandwidth

- Data transfers to and from main memory often involve blocks of data
  - Speed of these transfers has a large impact on performance of a computer system.
- Memory access time is not sufficient for describing memory's performance when transferring blocks of data
- During block transfers, **memory latency** is amount of time it takes to transfer first word of a block

Latency and bandwidth

- Time required to transfer a complete block depends also on rate at which successive words can be transferred and on size of block
- Time between successive words of a block is much shorter than time needed to transfer first word
  - Ex., in timing diagram, access cycle begins with assertion of RAS signal
  - First word of data is transferred five clock cycles later
  - Thus, latency is five clock cycles
  - If clock rate is 500 MHz, then latency is 10 ns
  - Remaining three words are transferred in consecutive clock cycles, at rate of one word every 2 ns

Latency and bandwidth

- A useful performance measure is number of bits or bytes that can be transferred in one second
  - This measure is often referred to as **memory bandwidth**
- It depends on speed of access to stored data and on number of bits that can be accessed in parallel
- Rate at which data can be transferred to or from memory depends on bandwidth of system interconnections
  - Interconnections used always ensure that bandwidth available for data transfers between processor and memory is very high

Double-data-rate SDRAM

- In quest for improved performance, faster versions of SDRAMs have been developed
  - Take advantage of fact that a large number of bits are accessed at same time when a row address is applied
  - Various techniques are used to transfer these bits quickly to pins of chip
- To make best use of available clock speed, data are transferred externally on both rising and falling edges of clock
  - Memories that use this technique are called **double-data-rate SDRAMs (DDR SDRAMs)**

Double-data-rate SDRAM

- Several versions of DDR chips have been developed
  - Earliest version is known as DDR
  - Later versions, called DDR2, DDR3, and DDR4, have enhanced capabilities
  - Offer increased storage capacity, lower power, and faster clock speeds
  - Ex., DDR2 and DDR3 can operate at clock frequencies of 400 and 800 MHz, respectively
  - They transfer data using effective clock speeds of 800 and 1600 MHz, respectively

Rambus memory

- Rate of transferring data between memory and processor depends on both:
  - Bandwidth of memory and bandwidth of its connection to processor
  - One way for increasing bandwidth of this connection is to use a wider datapath
  - This requires more space and more pins, increasing system cost
  - Alternative is to use fewer wires with a higher clock speed
  - Used by Rambus technology

Rambus memory

- Rambus is a memory technology:
  - Achieves a high data transfer rate by providing a high-speed interface between memory and processor
- Key feature of Rambus technology is use of a differential-signaling technique to transfer data to and from memory chips
  - Signals are transmitted using small voltage swings of 0.1V above and below a reference value

Rambus memory

- Several versions of this standard have been developed
  - With clock speeds of upto 800MHz and data transfer rates of several gigabytes per second
- Rambus technology competes directly with DDR SDRAM technology
  - Each has certain advantages and disadvantages
  - A nontechnical consideration is that specification of DDRSDRAM is an open standard that can be used free of charge
  - Rambus is a proprietary scheme that must be licensed by chip manufacturers

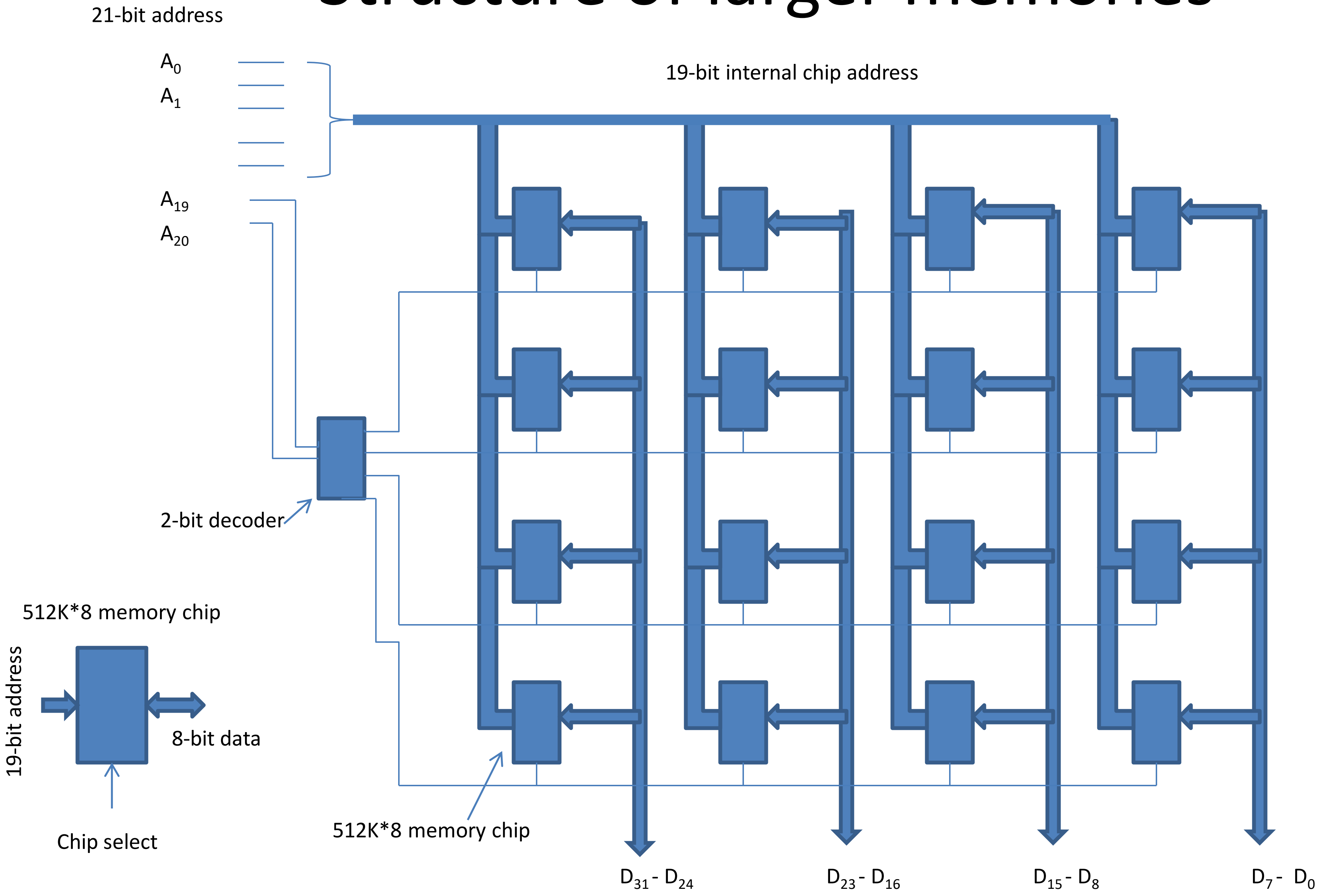
Structure of large memories: Static memories

- Consider a memory consisting of 2M words of 32 bits each using 512K×8 static memory chips
  - Each column figure implements one byte position in a word, with four chips providing 2M bytes
  - Four columns implement required 2M×32 memory
  - Each chip has a control input called Chip-select
  - When this input is set to 1, it enables chip to accept data from or to place data on its data lines
  - Data output for each chip is of tri-state type

Structure of large memories: Static memories

- Only selected chip places data on data output line
  - All other outputs are electrically disconnected from data lines
- 21 address bits are needed to select a 32-bit word in this memory
  - High-order two bits of address are decoded to determine which of the four rows should be selected
  - Remaining 19 address bits are used to access specific byte locations inside each chip in selected row
- R/W inputs of all chips are tied together to provide a common Read/Write control line

Structure of larger memories



Dynamic memory systems

- Dynamic RAMs, mostly of synchronous type, are widely used in memory units of computers
  - Because of their high bit density and low cost
  - Slower than static RAMs, but use less power and have considerably lower cost per bit
- Available chips have capacities as high as 2G bits
- To reduce number of memory chips needed:
  - A memory chip may be organized to read or write a number of bits in parallel
  - A 1-Gbit chip may be organized as 256M × 4, or 128M × 8

Dynamic memory systems

- Packaging considerations have led to development of assemblies known as memory modules
  - Each module houses many memory chips, typically in range 16 to 32, on a small board that plugs into a socket on the computer’s motherboard
  - Memory modules are commonly called SIMMs (Single In-line Memory Modules) or DIMMs (Dual In-line Memory Modules), depending on the configuration of the pins
  - Modules of different sizes are designed to use same socket
  - Total memory capacity is easily expanded by replacing a smaller module with a larger one, using same socket

Memory controller

- Address applied to dynamic RAM chips is divided into two parts
  - High-order address bits, which select a row in cell array, are provided first and latched into memory chip under control of RAS signal
  - Then, low-order address bits, which select a column, are provided on same address pins and latched under control of CAS signal
  - Since a typical processor issues all bits of an address at same time, a multiplexer is required
  - This function is usually performed by a **memory controller circuit**

Memory controller

- Controller accepts a complete address and R/W signal from processor
  - Under control of a *Request* signal which indicates that a memory access operation is needed
  - Forwards R/W signals and row and column portions of address to memory
  - Generates RAS and CAS signals, with appropriate timing
- When a memory includes multiple modules:
  - One of these modules is selected based on high-order bits of address
  - Memory controller decodes these high-order bits and generates chip-select signal for appropriate module

Memory controller

- Data lines are connected directly between processor and memory.
- Dynamic RAMs must be refreshed periodically
  - Circuitry required to initiate refresh cycles is included as part of internal control circuitry of synchronous DRAMs
  - However, a control circuit external to chip is needed to initiate periodic Read cycles to refresh cells of an asynchronous DRAM. Memory controller provides this capability.