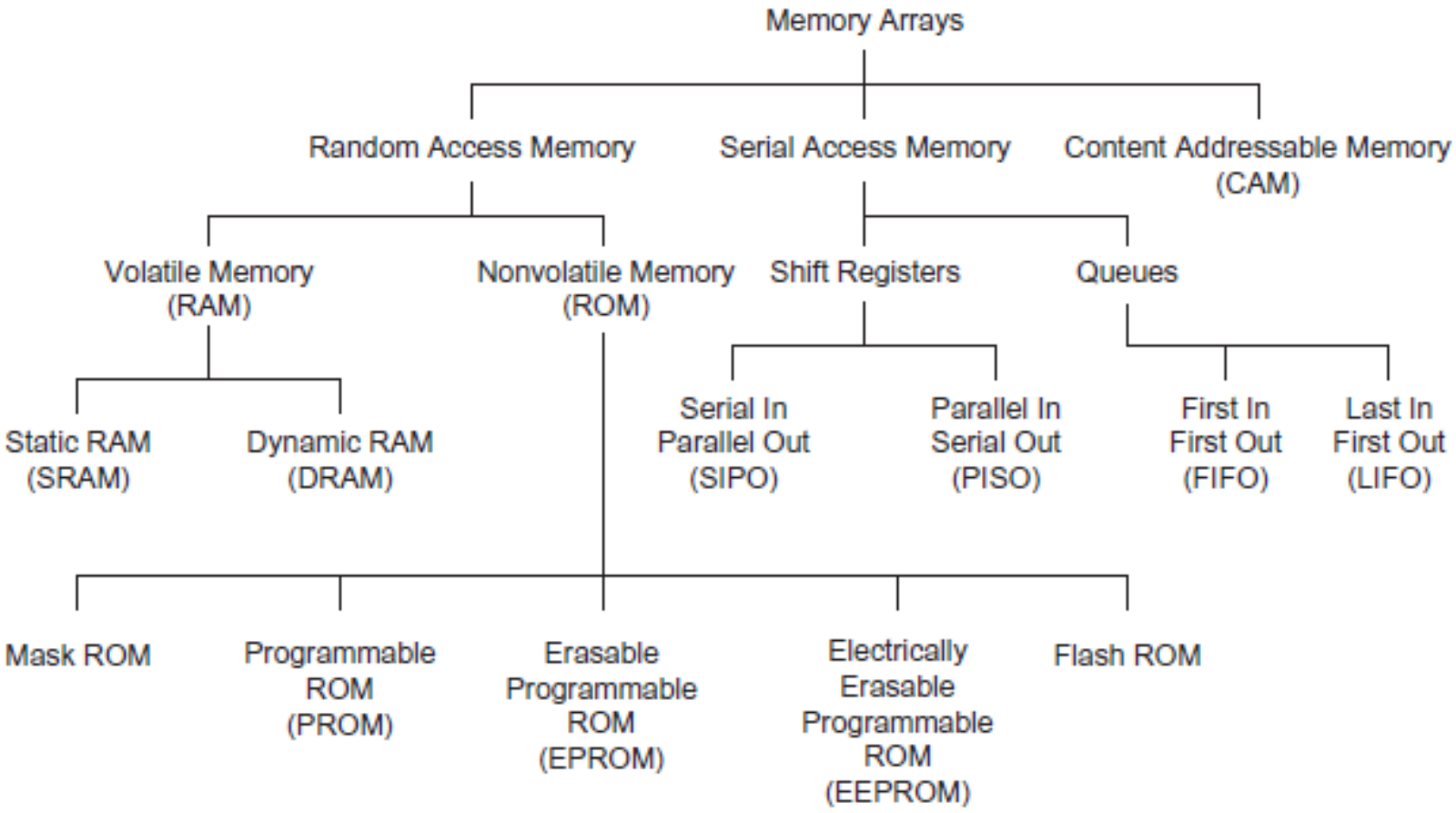
Memory

An Overview


Memory Arrays
Random Access Memory
Serial Access Memory
Content Addressable Memory (CAM)
Volatile Memory (RAM)
Nonvolatile Memory (ROM)
Shift Registers
Queues
Static RAM (SRAM)
Dynamic RAM (DRAM)
Serial In Parallel Out (SIPO)
Parallel In Serial Out (PISO)
First In First Out (FIFO)
Last In First Out (LIFO)
Mask ROM
Programmable ROM (PROM)
Erasable Programmable ROM (EPROM)
Electrically Erasable Programmable ROM (EEPROM)
Flash ROM

Static RAMs use a memory cell with internal feedback that retains its value as long as power is applied. It has the following attractive properties:

- Denser than flip-flops
- Compatible with standard CMOS processes

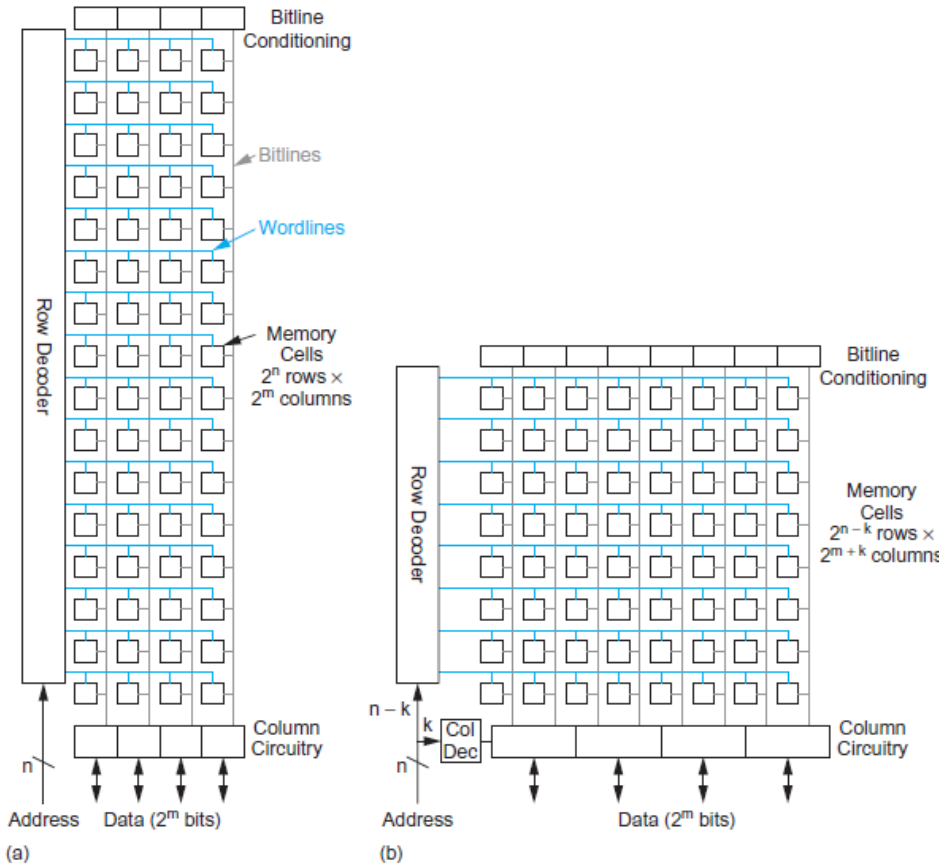
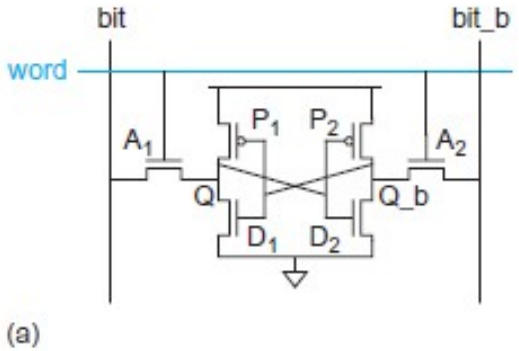


FIGURE 12.2 Memory array architecture

- Faster than DRAM
- Easier to use than DRAM

For these reasons, SRAMs are widely used in applications from caches to register files to tables to scratchpad buffers. The SRAM consists of an array of memory cells along with the row and column circuitry. This section begins by examining the design and operation of each of these components. It then considers important special cases of SRAMs, including multiported register files, large SRAMs and subthreshold SRAMs.

12.2.1.1 Read Operation Figure 12.4 shows a SRAM cell being read. The bitlines are both initially floating high. Without loss of generality, assume Q is initially 0 and thus Q_b is initially 1. Q_b and *bit_b* both should remain 1. When the wordline is raised, *bit* should be pulled down through *driver* and *access* transistors $D1$ and $A1$. At the same time *bit* is being pulled down, node Q tends to rise. Q is held low by $D1$, but raised by current flowing in from $A1$. Hence, the driver $D1$ must be stronger than the access transistor $A1$. Specifically, the transistors must be ratioed such that node Q remains below the switching threshold of the $P2/D2$ inverter. This constraint is called *read stability*. Waveforms for the read operation are shown in Figure 12.4(b) as a 0 is read onto *bit*. Observe that Q momentarily rises, but does not glitch badly enough to flip the cell.



(a)

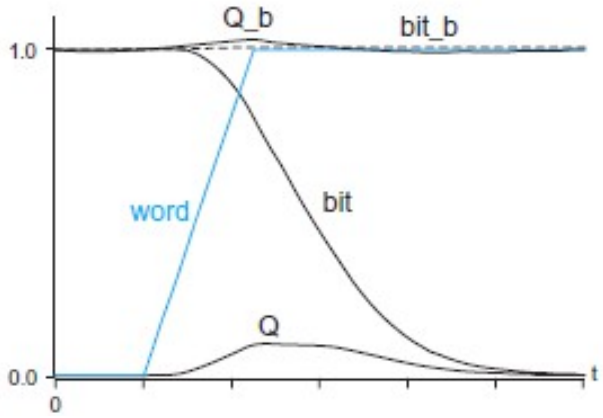
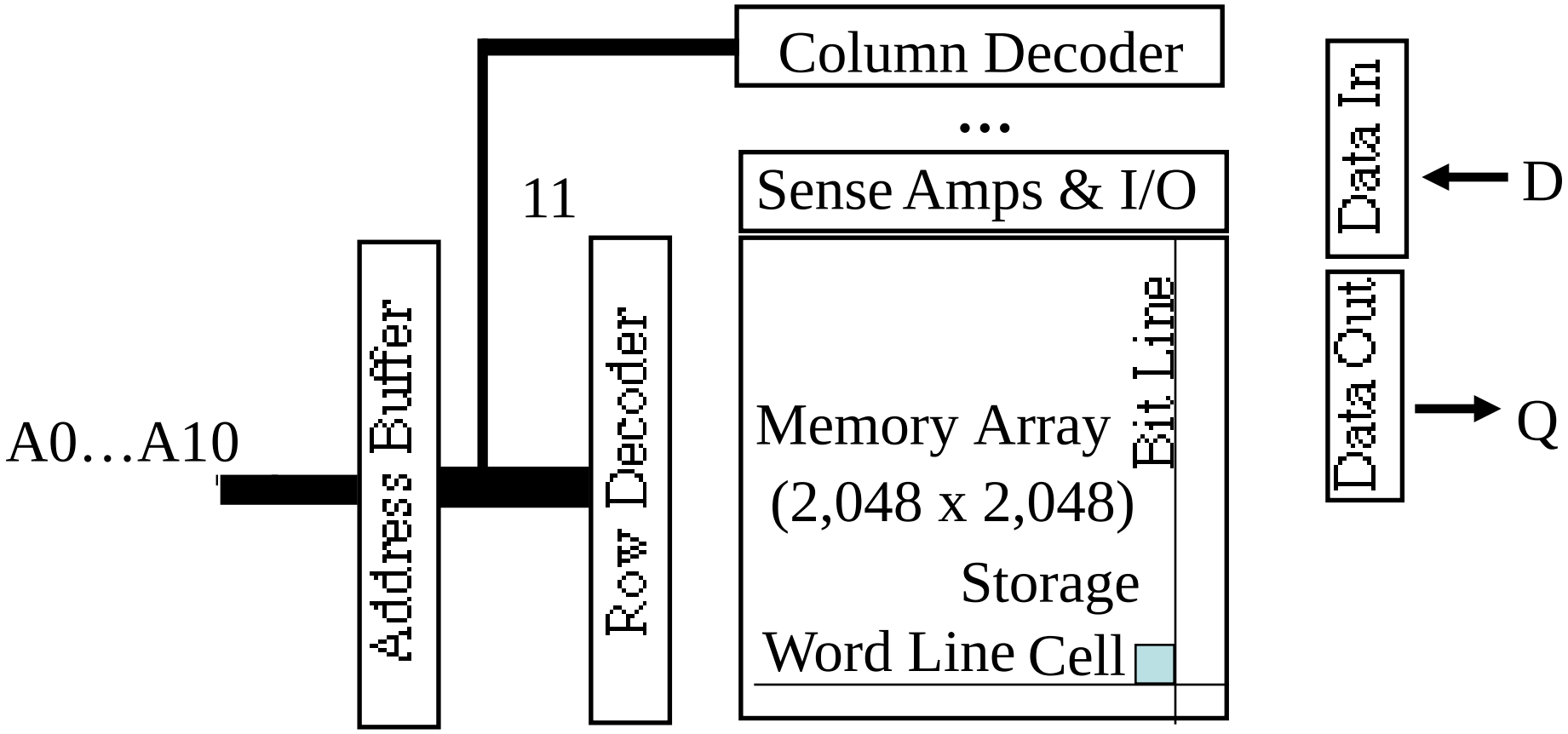


Figure 12.5 shows the same cell in the context of a full column from the SRAM. During phase 2, the bitlines are precharged high. The wordline only rises during phase 1; hence, it can be viewed as a _q1 qualified clock (see Section 10.4.6). Many SRAM cells share the same bitline pair, which acts as a distributed dual-rail footless dynamic multiplexer. The capacitance of the entire bitline must be discharged through the access transistor. The output can be sensed by a pair of HI-skew inverters. By raising the switching threshold of the sense inverters, delay can be reduced at the expense of noise margin. The outputs are dual-rail monotonically rising signals, just as in a domino gate.

DRAM logical organization (4 Mbit)



- Square root of bits per RAS/CAS

Technologies for DRAM Memories

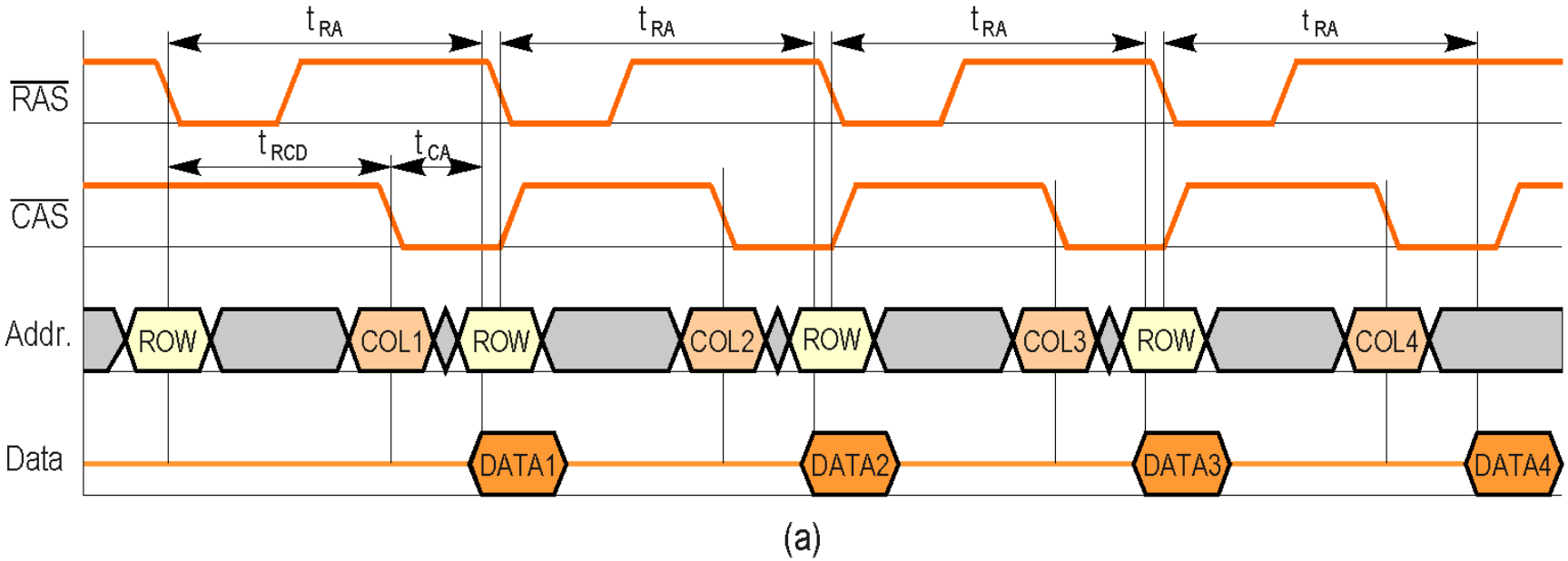
- Technologies for DRAM Memories
  - Categories of DRAM Memories
  - SDRAM
  - DDR SDRAM
  - DDR2 SDRAM
  - DDR3 SDRAM
  - DDR4 SDRAM

Quest for DRAM Performance

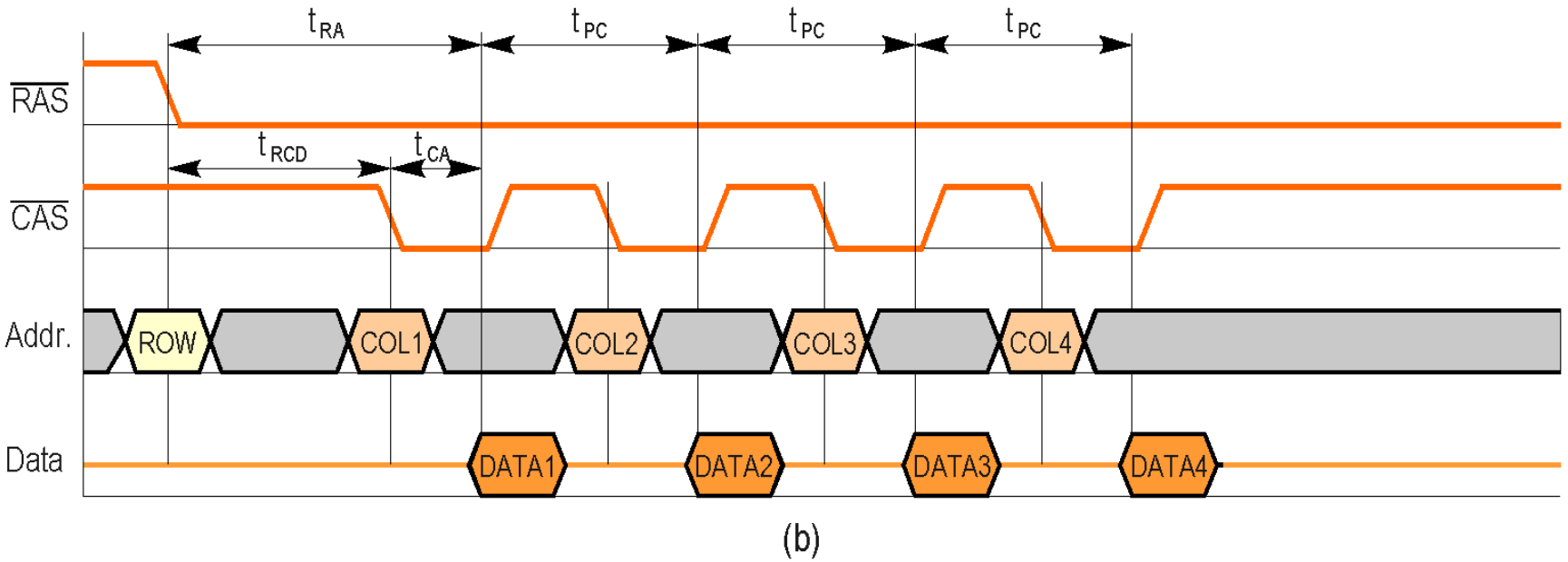
1. Fast Page mode
   - Add timing signals to allow repeated accesses to row buffer without another row access time
   - Easy, since array buffers 1024-2048 bits for each access
2. Synchronous DRAM (SDRAM)
   - Add clock signal to DRAM interface so repeated transfers don't pay overhead to synchronize with controller
3. Double Data Rate (DDR SDRAM)
   - Transfer data on rising and falling edges of clock, doubling peak data rate
   - DDR2 lowers power by dropping voltage from 2.5 to 1.8 volts, and offers higher clock rates: up to 400 MHz
   - DDR3 drops to 1.5 volts and raises clock rates to 800 MHz

- Improved bandwidth, not latency

Four memory accesses in a row without page mode



Four memory accesses in a row with page mode



(b)

Parameters of DRAM Memories (1)

- *Access time* (t_A): the time between the placement of the row address and the availability of the requested word
- *Cycle time* (t_M): the minimum time between sequential read operations
  - $t_M > t_A$
- *RAS Precharge time* (t_{RP}): the time needed to write back the memory contents and to activate the *RAS* signal

Parameters of DRAM Memories (2)

Operations performed for a read access:

- The processor sends the address of data
- The memory controller determines the row and column addresses of the data
- The memory controller sends the row address and asserts the *RAS* signal
  - The row address decoder selects the row in which the data are stored, or reads the entire row and stores it in a buffer

Parameters of DRAM Memories (3)

- *Row Access time* (t_{RA}), also called *access time* (t_A) or *random access time*: the time between the assertion of the *RAS* signal and the selection of the row or the presence of data in the output buffer

- The memory controller sends the column address and asserts the *CAS* signal
  - *RAS-to-CAS delay* (t_{RCD})

Parameters of DRAM Memories (4)

- *Column Access time* (t_{CA}): the time from the activation of the *CAS* signal after which the requested data will be available
- For synchronous memories the *CAS latency* (t_{CL}) is used; expressed as an integer number of clock cycles

- The selected data are sent from the output buffer to the data bus

Parameters of DRAM Memories (5)

- *Peak bandwidth* of a DRAM module: the amount of data transferred at maximum rate for a given memory bus configuration
  - Ignores the initial time necessary to fetch the data from the DRAM module
- *Sustained bandwidth*: takes into account an initial access to the memory, followed by a four-word transfer at the maximum rate

Page mode (1)

- If a sequence of memory accesses have the same row address: it is sufficient to transfer the row address once
  - An entire row of data (*page*) is read out and stored in an internal buffer page mode
  - For a subsequent access to the same page, only a column address needs to be transferred
  - There is no need to restore the page data at every access to a word

Page mode (2)

- In page mode, the *RAS* signal is maintained active for the duration of a sequence of transfers
  - The *CAS* signal is toggled in the normal way
- *Page mode cycle time* (t_{PC})
  - For a memory with t_{RA} = 60 ns, a typical value is t_{PC} = 35 ns

Categories of DRAM Memories

- Almost all types of DRAM have the same initial latency to access the first word of memory (50 .. 60 ns)
- Various techniques are used to perform the sequential read operations after the first word of memory is read
- Types of DRAM memories:
  - With asynchronous interface
  - With synchronous interface
  - Protocol-based

DRAM Memories with Asynchronous Interface (1)

- The internal operations are assigned minimum time values
- If a clock pulse occurs prior to the minimum time, another clock pulse must occur the performance is limited
- Enhancing the performance: increasing the number of bits per access; overlapping various operations; eliminating some operations

DRAM Memories with Asynchronous Interface (2)

- Using wider I/O ports
  - Additional I/O pins needed the cost increases
  - The current drawn increases the speed is reduced
- Overlapping various operations
- Eliminating some internal operations
  - FPM (*Fast Page Mode*)
  - EDO (*Extended Data Out*)
  - BEDO (*Burst Extended Data Out*)

DRAM Memories with Synchronous Interface (1)

- The waiting periods for the processor are eliminated
- The DRAM latches some information from the processor under control of the system clock: addresses, data, and control signals
- The system clock is the only timing signal that needs to be provided to the memory
- The inputs are simplified

Protocol-based DRAM Memories

- The previous memory categories have separate address, data and control lines
  - This may limit the operating speed
- Protocol-based DRAM memories implement the address, data, and control signals on the same high-speed bus
  - Rambus DRAM
  - SLDRAM (*SyncLink* DRAM)

Principle of SDRAM Memory

- Differences to the asynchronous DRAM:
  - Uses a multi-bank architecture (2 or 4 banks per module)
  - Can operate in burst mode for 2 bits, 4 bits, 8 bits, or a page
  - The control method
    - Synchronous DRAM is controlled by commands placed on the bus interpreted on the rising edge of the clock signal

Signals of an SDRAM memory

- *CLK* (*Clock*)
  - The rising edge of the clock signal initiates the command decoding and execution
  - An SDRAM module uses 2 or 4 clock lines
- *CKE* (*Clock Enable*)
  - Activates and deactivates the *CLK* signal
  - When the *CLK* signal is deactivated, the input buffers are turned off to save power

Signals of an SDRAM memory

- *CS* (*Chip Select*)
- *RAS*, *CAS*, *WE*
  - Same function as for asynchronous DRAM memories
- *DQ* (*Data*)
- *DQM* (*DQ Mask*)
  - Used to control the data lines
- *A* (*Address*)
- *BA* (*Bank Address*)

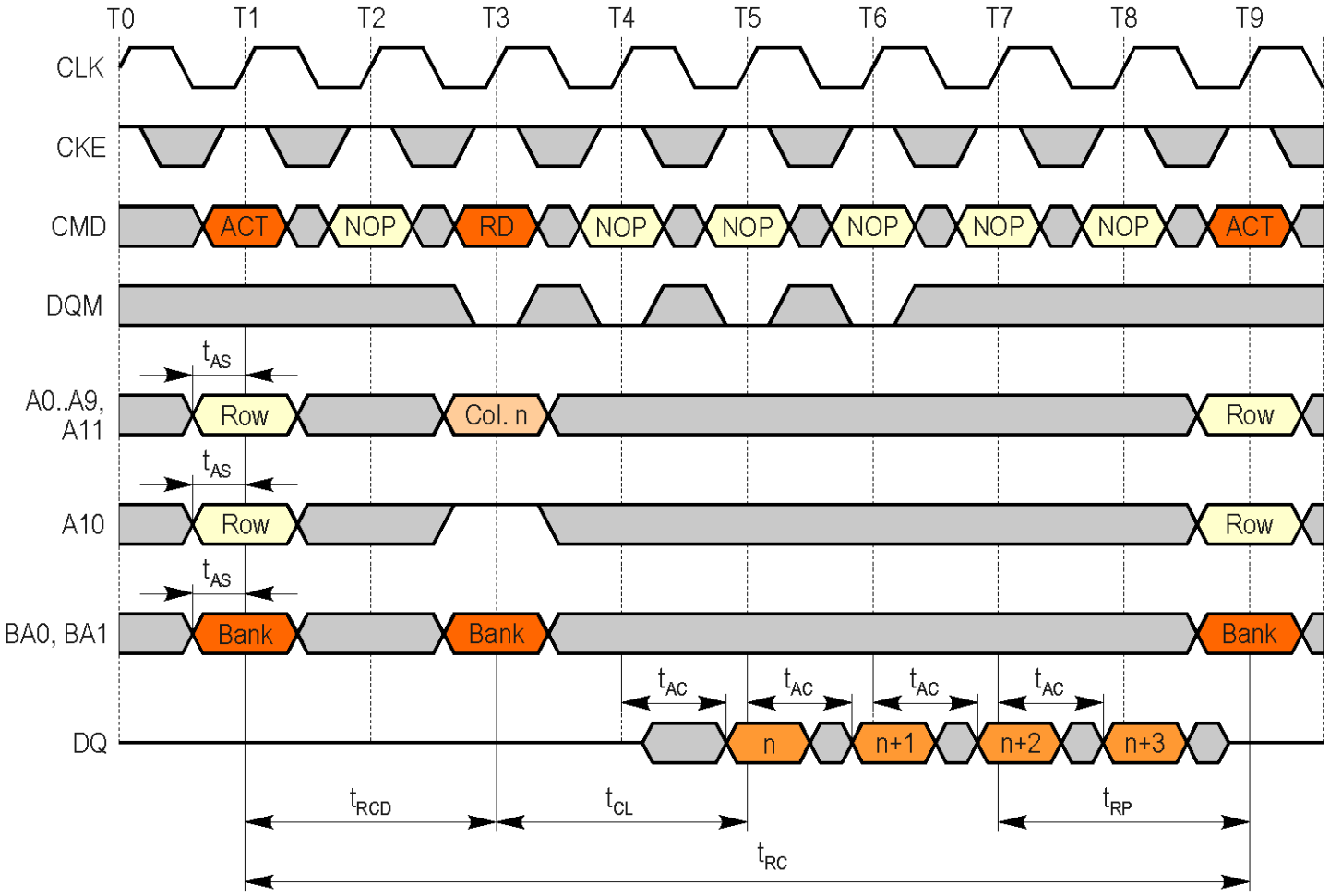
SDRAM Commands

- An SDRAM command is determined by a combination of the *CS*, *RAS*, *CAS*, *WE* signals
- *No Operation* (NOP)
  - Activates a memory chip and place it in the idle state
- *Activate* (ACT)
  - Selects a particular memory bank and activates a row in the selected bank
- *Read*, *Write*

SDRAM Commands

- *Read/Write with Auto Precharge*
  - Combine a read or write operation with an automatic precharge of an individual bank, without an explicit precharge command
  - Advantage: the precharge is performed at the earliest time within a burst transfer
  - Execution of a *Read with Auto Precharge* command

SDRAM Commands



SDRAM Commands

- *Burst Terminate*
  - Used to terminate burst transfers
- *Precharge Selected Bank*
  - Indicates to the active memory bank to recharge itself in order to be ready for the next access
- *Precharge All*
  - All banks are precharged at the same time

SDRAM Commands

- *Auto Refresh*
  - Refreshes the SDRAM array explicitly
- *Mode Register Set*
  - Loads the mode register with information on:
    - The burst length
    - The *CAS* latency
    - The order of accesses within a burst transfer

Block Diagram

Micron, 64 Mbit

Control Signals Sampled Synchronously

4 Banks

Mode Register

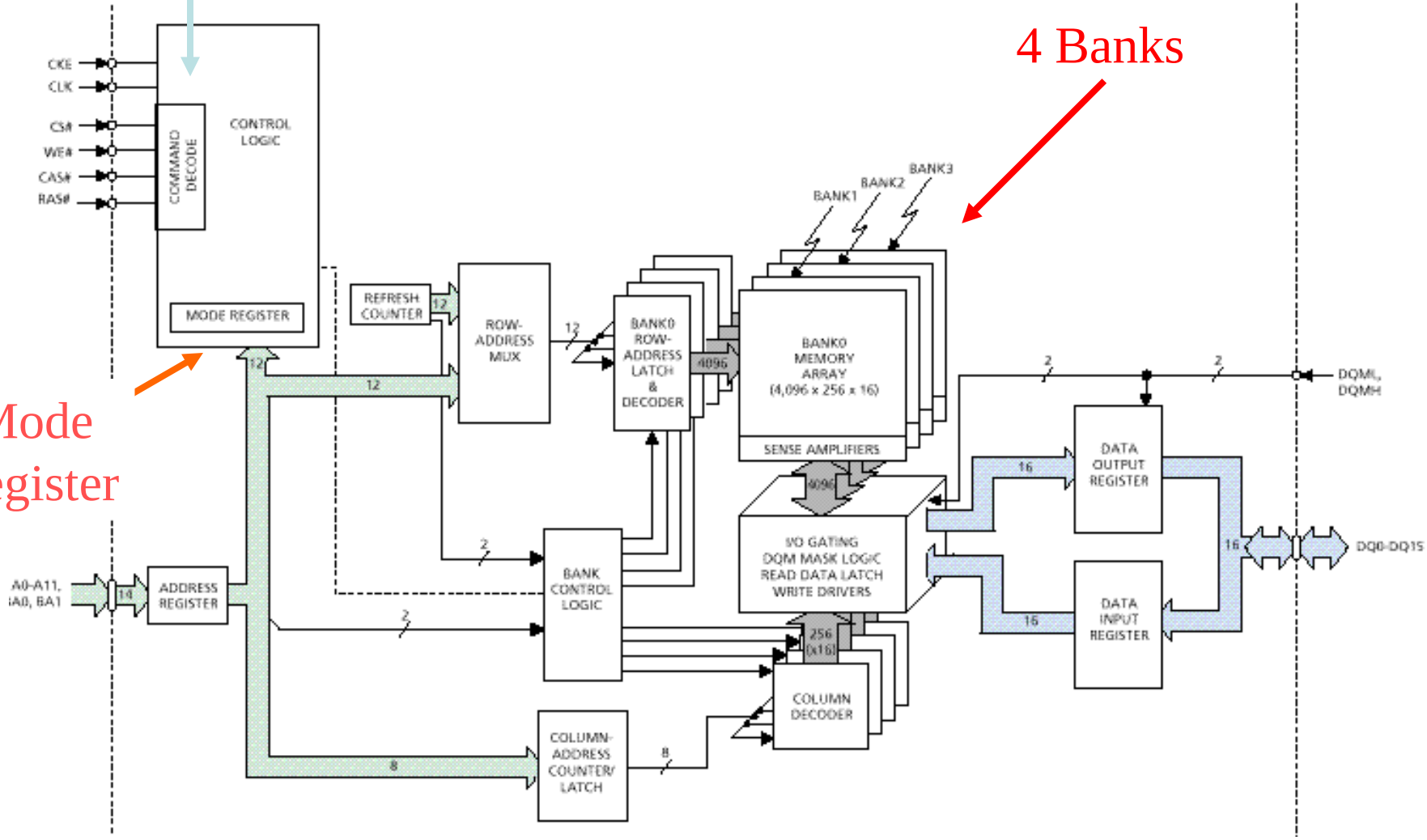


Table 1. Basic characteristics of memory technologies.

Characteristic	Description	Working memory (DRAM)	Storage memory (flash)
Capacity	How many bits of data it can store	Gbytes	Hundreds of Gbytes
Granularity	How much data is read/written in one access	Bytes	Kbytes
Read latency	How fast data can be read from memory	Nanoseconds	Microseconds
Write latency	How fast data can be written into memory	Nanoseconds	Microseconds
Retention	How long cells can store uncorrupted data	Microseconds	Years
Durability	How many writes on average before a cell fails	10^{15}	10^{4}-10^{5}

Table 2. Alternative memory technologies, their storage principles, and specific storage mechanisms.

Technology	Principle	Storage mechanism
Hybrid memory cube (HMC)	Electronic (charge-based)	Multiple layers of DRAM technology chip-stacked on top of a high-performance logic layer; trades total memory capacity for better performance
Vertical flash	Electronic (charge-based)	Charges are trapped in a floating-gate transistor; cells are vertically integrated (3D), significantly increasing density and allowing the use of larger cells
Phase-change memory (PCM)	Atomic (resistive)	Cell material can be crystalized or put into an amorphous state by controlled heating and cooling; material has different resistances when crystalline or amorphous
Memristors	Atomic (resistive)	Memristors use a thin film of materials such as titanium dioxide; applying high currents moves oxygen vacancies around the film, changing its resistance
Conductive-bridging RAM (CB-RAM)	Atomic (resistive)	Metal ions in a cell migrate when a current is applied and form a conductive path within a nonconductive material; this changes the resistance of the cell
Spin-transfer torque memory (STT-RAM)	Magnetic (resistive)	Cell includes a permanent and a floating ferromagnetic material; the polarity of the latter can be changed by a polarized electrical current, and their relative alignment can be determined by running a current through them and observing their resistance
Racetrack memory	Magnetic (resistive)	Multiple ferromagnetic domains share a smaller set of read/write ports; to be read or written, these domains have to be shifted into the port region