I. Semiconductor Main Memory

A. Common Properties
   1. Two stable (or semistable) states used to represent 0 and 1.
   2. Capable of being written into (at least once), to set the state.
   3. Capable of being read to sense state.

B. RAM, short for random access memory, but all semiconductor memory is random access.
   1. Can both read and write data quickly and easily.
   2. Is volatile
   3. DRAM = dynamic RAM that stores data as a charge on capacitors.
      a. Read causes discharge of the capacitor, so it must be restored.
      b. Capacitors are actually analog, so 1’s and 0’s are based on threshold values.
      c. Operations to read a data bit from a DRAM storage cell
         1) The sense amplifiers are disconnected.
         2) The bit-lines are precharged to exactly equal voltages that are in between high and low logic levels.
         3) The precharge circuit is switched off.
         4) The desired row's word-line is then driven high to connect a cell's storage capacitor to its bit-line, and the sense amplifiers are connected to the bit-lines.
         5) All storage cells in the open row are sensed simultaneously, and the sense amplifier outputs latched. A column address then selects which latch bit to connect to the external data bus.
         6) While reading of columns in an open row is occurring, current is flowing back up the bit-lines from the output of the sense amplifiers and recharging the storage cells.
         7) When done with reading all the columns in the current open row, the word-line is switched off to disconnect the storage cell capacitors (the row is "closed") from the bit-lines. The sense amplifier is switched off, and the bit lines are precharged again.

4. SRAM = static RAM that use logic elements.
   a. No need for refreshing
   b. Somewhat faster than DRAM.
   c. Used for cache, both on and off chip.

5. SRAM has three states.
   a. Standby (the circuit is idle) = if the word line (WL) is not asserted, the access transistors M5 and M6 disconnect the cell from the bit lines. The two cross-coupled inverters formed by M1 – M4 will continue to reinforce each other as long as they are connected to the supply.
   b. Reading procedure. Assume that the content of the memory is a 1, stored at Q.
      1) Precharging both the bit lines to a logical 1, then asserting the word line WL, enabling both the access transistors.
      2) The values stored in Q and $\bar{Q}$ are transferred to the bit lines by leaving BL at its precharged value and discharging BL through M1 and M5 to a logical 0 (i.e. eventually discharging through the transistor M1 as it is turned on because the Q is logically set to 1). On the BL side, the transistors M4 and M6 pull the bit line toward $V_{DD}$, a logical 1 (i.e. eventually being charged by the transistor M4 as it is turned on because $\bar{Q}$ is logically set to 0).
The BL and $\overline{BL}$ will have a small difference between them that reaches a sense amplifier. The higher the sensitivity of sense amplifier, the faster the speed of read operation is.

c. Writing procedure

1) Applying the value to be written to the bit lines. If we wish to write a 0, we would apply a 0 to the bit lines, i.e. setting $\overline{BL}$ to 1 and BL to 0. This is similar to applying a reset pulse to an SR-latch, which causes the flip flop to change state. A 1 is written by inverting the values of the bit lines.

2) WL is then asserted and the value that is to be stored is latched in. Note that the reason this works is that the bit line input-drivers are designed to be much stronger than the relatively weak transistors in the cell itself, so that they can easily override the previous state of the cross-coupled inverters.

C. Types of ROM = read-only memory

1. Used for microprogramming, system programs, function tables, BIOS.
2. Fabricated ROMs have a relatively large fixed cost, whether one or thousands of copies made. No room for error.
3. PROM = programmable ROM may be written only once, but done electronically.
4. EPROM = erasable programmable ROM can be read and re-written. Uses one transistor per bit. Must be erased totally using ultraviolet radiation for 20 minutes, before being re-written electrically.
5. EEPROM = electrically erasable programmable ROM can be written, but write takes sever hundred microseconds per byte. More expensive and less dense than EPROMs.
6. Flash memory like EEPROM, but faster, can erase sections at a time, and only one transistor per bit.

D. Chip Logic

1. Each chip contains a square array of memory cells connected with vertical and horizontal lines.
   a. Horizontal line connects to the Select terminal of each cell in its row.
   b. Vertical line connects to the Data-In/Sense terminal of each cell in its column, and terminate in a MUXed bit driver or sense amplifier.
   c. When read, the values are fed back in to the bit driver/refresh circuitry.
   d. Address can be time multiplexed to reduce the number of address pins needed by half. Needs a row address select (RAS), and a column address select (CAS) to provide timing to the chip. Each new pin quadruples the range of addresses.
2. Key design issue is number of bits of data that may be accessed at a time, e.g. 16-bit words, or 1-bit.

E. Chip Packaging

1. Pins are: address (half the number if multiplexed), data, power supply ($V_{cc}$), ground ($V_{ss}$), chip enable/select (CE or CS) based on MSB of address.
   a. For EPROM also program voltage for writing ($V_{pp}$).
   b. For RAM also write enable (WE), read/output enable (OE).
   c. For multiplexed, row address select (RAS), and column address select (CAS).
   d. For synchronous, clock input (CLK), bank selection (BAn).
2. An 8-Mbit chip organized as a1M x 8 chip will have 8 data pins. A 16-Mbit chip organized as 4M x 4 will have only 4 data pins.

F. Module Organization has arrays of chips, with the high order bits of the address used to assert CE to the appropriate row of chips.

G. Interleaved Memory = chips are grouped together to form 2, 4, or 8 independent memory banks with consecutive addresses stored in different banks. Then it is possible that 2, 4, or 8 requests could be serviced simultaneously if the addresses sought are in different banks.

II. Advanced DRAM Organization

A. Synchronous DRAM (SDRAM) = a system clock determines when memory chips will do things.
   1. Burst mode eliminates address setup time and row and column precharge time after the first access so a series of data bits can be clocked out rapidly after the first bit has been access. Great for sequential data like word processing, and multimedia.
   2. When CS, RAS, CAS, WE, BAn are all low, then the addresses A0-A9 set the burst length (1, 2, 4, or 8) in the mode register, and the CAS latency (2 or 3 cycles).
B. Rambus DRAM (RDRAM) used a special bus to address up to 320 RDRAM chips that is rated at 1.6GBps on its 16 data lines. Rather than RAS, CAS, etc, the bus is used to specify the type of operation and number of bytes.
C. DDR (Double-data-rate) SDRAM = can send data on both the rising edge and falling edge of the clock pulse, sets the minimum read or write to 2 consecutive words. Up to 200MHz clock, and reduced voltage from 3.3V to 2.5V.
   1. DDR2 doubles the minimum read or write to 4 consecutive words with up to 400MHz clock.
   2. DDR3 sets the minimum read or write to 8 consecutive words with up to 800MHz, and 1.5 V.
   3. DDR4 keeps minimum at 8 consecutive words, but the frequency rate can be 2133MHz to 4266MHz with 1.2V
D. Cache DRAM = integrates a small SRAM cache (16 Kb) onto a generic DRAM chip.
Operations to read a data bit from a DRAM storage cell

1. The sense amplifiers are disconnected.
2. The bit-lines are precharged to exactly equal voltages that are in between high and low logic levels. The bit-lines are physically symmetrical to keep the capacitance equal, and therefore the voltages are equal.
3. The precharge circuit is switched off. Because the bit-lines are relatively long, they have enough capacitance to maintain the precharged voltage for a brief time. This is an example of dynamic logic.
4. The desired row's word-line is then driven high to connect a cell's storage capacitor to its bit-line. This causes the transistor to conduct, transferring charge between the storage cell and the connected bit-line. If the storage cell's capacitor is discharged, it will greatly decrease the voltage on the bit-line as the precharge is used to charge the storage capacitor. If the storage cell is charged, the bit-line's voltage only decreases very slightly. This occurs because of the high capacitance of the storage cell capacitor compared to the capacitance of the bit-line, thus allowing the storage cell to determine the charge level on the bit-line.
5. The sense amplifiers are connected to the bit-lines. Positive feedback then occurs from the cross-connected inverters, thereby amplifying the small voltage difference between the odd and even row bit-lines of a particular column until one bit line is fully at the lowest voltage and the other is at the maximum high voltage. Once this has happened, the row is "open" (the desired cell data is available).
6. All storage cells in the open row are sensed simultaneously, and the sense amplifier outputs latched. A column address then selects which latch bit to connect to the external data bus. Reads of different columns in the same row can be performed without a row opening delay because, for the open row, all data has already been sensed and latched.
7. While reading of columns in an open row is occurring, current is flowing back up the bit-lines from the output of the sense amplifiers and recharging the storage cells. This reinforces (i.e. "refreshes") the charge in the storage cell by increasing the voltage in the storage capacitor if it was charged to begin with, or by keeping it discharged if it was empty. Note that due to the length of the bit-lines there is a fairly long propagation delay for the charge to be transferred back to the cell's capacitor. This takes significant time past the end of sense amplification, and thus overlaps with one or more column reads.
8. When done with reading all the columns in the current open row, the word-line is switched off to disconnect the storage cell capacitors (the row is "closed") from the bit-lines. The sense amplifier is switched off, and the bit lines are precharged again.