**CHAPTER 4**

**CACHE ARCHITECTURE**

Cache memory is random access memory that a computer microprocessor can access more quickly than it can access regular RAM. It is a high speed memory in the CPU that is used for faster access to data. It provides the processor with the most frequently requested data and increases the performance. In this chapter basic components of a cache memory array and concepts of 6T and 4T cache memories are discussed.

**4.1 COMPONENTS OF SRAM ARRAY**

SRAM arrays are created by replicating the basic storage cell and adding the necessary peripheral circuitry. The basic architecture of SRAM memory array comprises of the following components.

**Dual core:** It consists of two core regions of storage cells that share central word line circuits. Memory cells are tiled to produce the left and right core regions. Width of core region is multiple of word size.

**Row decoder:** The output of a centrally located decoder provides word-line signals to the storage cells. The address word specifies a particular row, which is then driven high. The access transistors of the selected row cells are turned on, permitting the read or write operations to take place. The location of the circuitry allows a single decoder to be used for both left and right memory cores.

**Column decoder:** To choose a particular word in a row, a group of column decoder circuits is added, to select a particular set of columns in the matrix. Mux sections are driven by column decoder to steer the selected bit, bit-bar groups.

**Precharge circuit:** This circuit at the top is controlled by a clock signal that is used to synchronize the operation and data flow. During hold state these circuits are active and maintain constant voltages on bit and bit-bar lines. During read and write operations precharge circuitry is deactivated.

**Read/ Write circuitry:** It performs the following functions

* Directing the data flow into the array during a write operation or out of array during a read operation.
* Connecting the read and write circuits to the bit and bit-bar lines of every column.
* Providing amplifiers to detect and amplify the outputs during a read operation.

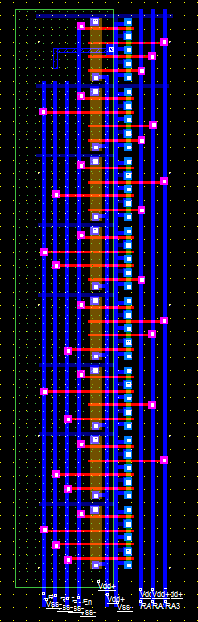
**4.2 6T SRAM CACHE**

Conventionally, cache memory is implemented using 6T SRAM cells. Here, layout of cache memory of size 8x16 is implemented using Microwind tool. Storage cell is made up of 6 transistors. A bi-stable latch circuit is used. Dual core is made by replicating 6T SRAM cells.

**Pre-charge circuitry:** A PMOS transistor is used which is activated during hold state which pulls bit and bit-bar lines to VDD. It is deactivated during read and write operations.

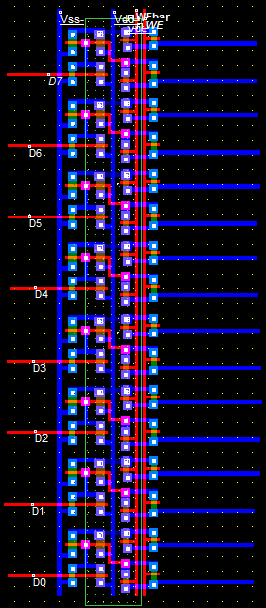
**Column decoder:** Since we have only two sets of words, we require one address bit to activate them. This address bit acts as column select signal. Transmission gates are used for this purpose as they provide strong “1” and “0” values.

**Row decoder:** Since we have 8 rows, to activate them we require 3 address bits. These three address bits along with enable signal are ANDed to activate the respective word-line that activates the corresponding row. Layout of row decoder circuitry is shown in Fig (4.1).



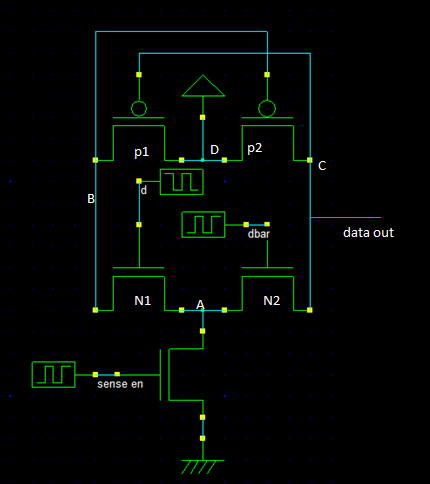
**Fig (4.1) Layout of Row decoder circuit**

**Write circuitry:** The input bits are inverted and buffered to provide complimentary pairs (d and dbar). When write enable control bit has a value WE=1, transmission gates act as closed switches connecting the data pairs to bit and bit-bar lines. Every bit pair is fed to the appropriate locations. Layout of write circuitry is shown in Fig (4.2).



**Fig (4.2) Layout of write circuit**

**Read circuitry:** It consists of a sense amplifier which detects the values on bit and bit-bar lines, amplifies the value and transfers it onto the output line. When read-enable and sense-enable signals are high sense amplifier read is activated and the value is read. Sense amplifier circuitry is shown in Fig (4.3).



**Fig (4.3) Schematic of 6T Sense amplifier**

**Working:**

When sense enable goes high, node A is grounded. When the data d=“1” and dbar =“0” the transistor N1 is ON and transistor N2 is OFF. Since node B is grounded via transistor N1, the ground potential at node B will activate transistor P2. Now node C will be pulled to VDD through P2. Now the data is read though node C. When the data d=“0” and dbar =“1” the transistor N2 is ON and transistor N1 is OFF. Node C is grounded via transistor N2. The data is read though node C

When sense enable is “0”, node A will be in high impedance state. Irrespective of the voltages on d and dbar, node C will be high impedance.

**4.3 4T SRAM CACHE**

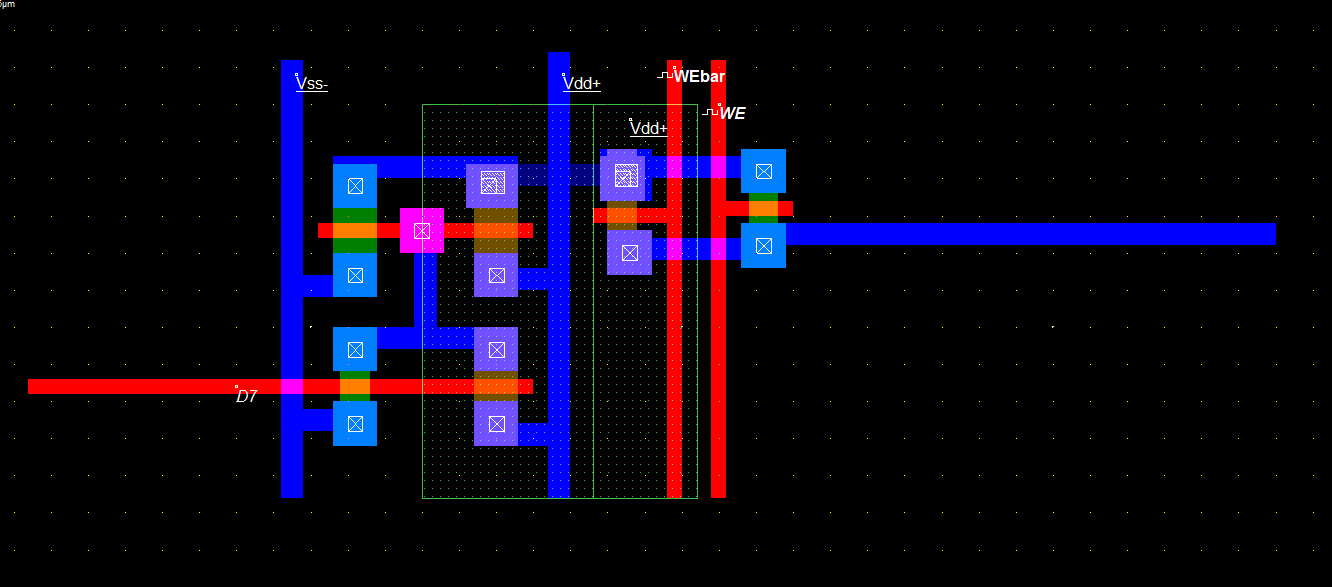
Cache memory is implemented using 4T SRAM cells. Here, layout of cache memory of size 8x16 is implemented using Microwind tool. Storage cell is made up of 4 transistors. Dual core is made by replicating 4T SRAM cells.

**Pre-charge circuitry:** A NMOS transistor and PMOS transistor are used which are activated during hold state to pull bit and bit-bar lines to GND and VDD respectively. They are deactivated during read and write operation.

**Row decoder:** Since we have 8 rows, to activate them we require 3 address bits. These three address bits along with enable signal are ANDed to activate the respective word-line that activates the corresponding row.

**Column decoder:** Since we have only two sets of words, we require one address bit to activate them. This address bit acts as column select signal. Transmission gates are used for this purpose as they provide strong “1” and “0” values.

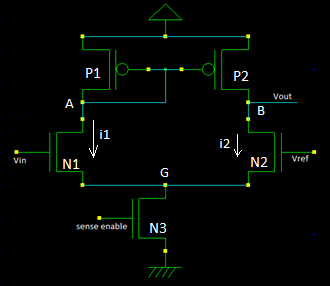
**Write circuitry:** The input bit d is buffered onto the bit-line. When write enable control bit has a value WE=1, transmission gates act as closed switches connecting the data to bit-line. Every bit is fed to the appropriate location. Unlike in 6T, here write operation is done using only BL so we need only one transmission gate.

****

**Fig (4.4) Layout of structural unit of 4T write circuitry**

The above structure is replicated to get a complete write circuitry.

**Read circuitry:** It consists of a sense amplifier which detects the values on bit-line, amplifies the value and transfers it onto the output line. When read enable and sense enable signals are high sense amplifier is activated and the value is read. A current mirror along with differential amplifier circuit is used. A reference voltage VREF is used. Any voltage above this value is detected as logic “1” and any voltage value below this value is detected as logic “0”.



**Fig (4.5) Schematic of 4T Sense amplifier**

**Working:** When sense enable is “1” it activates transistor N3 and the node G is grounded. Now if Vin > Vref that implies the gate to source voltage of N1 is greater that of gate to source voltage of N2. We know that sub-threshold current and the active current both are directly proportional to Vgs.

Ids α eVgs (sub-threshold region)

Ids α Vgs2 (active and saturation region)

Node A at the drain of N2 is pulled to ground faster than node B. When the node A reaches a voltage less than VTP of the transistor P2, it gets activated and pulls node B to VDD­.

When Vref >Vin, Igs of N2 is greater than that of N1 which pulls node B faster to ground.

**4.4 SUMMARY**

In this chapter basic architecture and components of SRAM array are discussed. Architecture of 6T and 4T SRAM caches is almost similar, except for the sense amplifier design. This is because of analog detection requirement in case of 4T.