Chapter 6
SRAM Mega Cell Design for Digital Applications

6.1 Introduction

The primary objective of this chapter is to present the application of silicon nanowire technology on a first large-scale digital mega cell design: an SRAM. The detailed steps of generating accurate BSIMSOI SPICE models from vertically-grown SNTs with undoped bodies and dual work function metal gates were already discussed in Chapter 3. This chapter uses the SNT device models in the design and analysis of a 16×16 SRAM block and reports the circuit simulation results and electrical data.

6.2 Brief Description of Transistor Design and Modeling

Both NMOS and PMOS transistors used in this chapter are enhancement type with undoped, cylindrical silicon bodies constructed perpendicular to the substrate as shown in Chapter 1 and then resumed in Chapter 3. Source/Drain (S/D) contacts are assumed to have ohmic contacts and 1.5 nm thick gate oxide. Device simulations are performed using Silvaco's three-dimensional ATLAS device simulation environment with a 1 V power supply voltage. Half of the device is constructed in a two-dimensional platform and then rotated around the *y*-axis to create a three-dimensional cylindrical form for simulations. The device radius is changed from 2 to 20 nm while its effective channel length is varied between 10 and 65 nm. The device simulator used low and high-electric field mobility models, concentration dependent Shockley-Read-Hall recombination model, Arora's lattice temperature model, Selberherr's impact ionization model, and Fermi statistics. Quantum mechanical effects are included using density gradient method.

The device design process starts by determining individual metal gate work functions for each NMOS and PMOS SNT that produces 300 mV threshold voltage.

A. Bindal, S. Hamedi-Hagh, *Silicon Nanowire Transistors*,
DOI 10.1007/978-3-319-27177-4_6

Once the gate work function for each transistor is determined, the body radius and effective channel length of both SNTs are simultaneously changed until each device reveals minimum static and dynamic power dissipations but exhibits the fastest transient times. This design process has ultimately produced an SNT body of 2 nm radius and 10 nm channel length. The BSIMSOI device models are based on these particular device dimensions and they given in Chapter 3 and they are used for all the circuit simulations in this chapter.

6.3 SRAM Design

This section demonstrates the overall SRAM architecture including the core, address decoder, read/write data-path and the self-time circuits. Power dissipation as well as read and write access times under different ambient conditions are also discussed.

6.3.1 SRAM Architecture

The SRAM architecture in this chapter consists of an SRAM core for data storage and retrieval, self-timed circuits for controlling read and write operation sequences, and an address decoder to decode a 4-bit wide address in order to produce 16 Word Lines (WL) as shown in Fig. 6.1.

To write into the SRAM block, WrEn is set high while RdEn is set low. This precharges the SRAM core for a write operation, enables the address decoder, and validates the input data. To read from the SRAM block, WrEn is set low and RdEn is set high. This setting enables the precharging circuit and the address decoder, but it also produces a strobe pulse to activate the sense amplifier to speed up the read operation. When both WrEn and RdEn are set low, no precharge pulse is produced to conserve power; the input data and address become invalid; only the prior output can be read from DataOut port.

6.3.2 SRAM Core

The 16×16 SRAM core consists of 16 identical columns, each of which includes 16 rows of six-transistor (6 T) memory cells for bit storage, a precharging circuit, a sense amplifier for read operation, a write circuit, and an output latch/buffer as shown in Fig. 6.2.

The input–output characteristics of the 6 T memory cell are shown in Fig. 6.3. The inverter threshold voltage is approximately 470 mV even though the ON

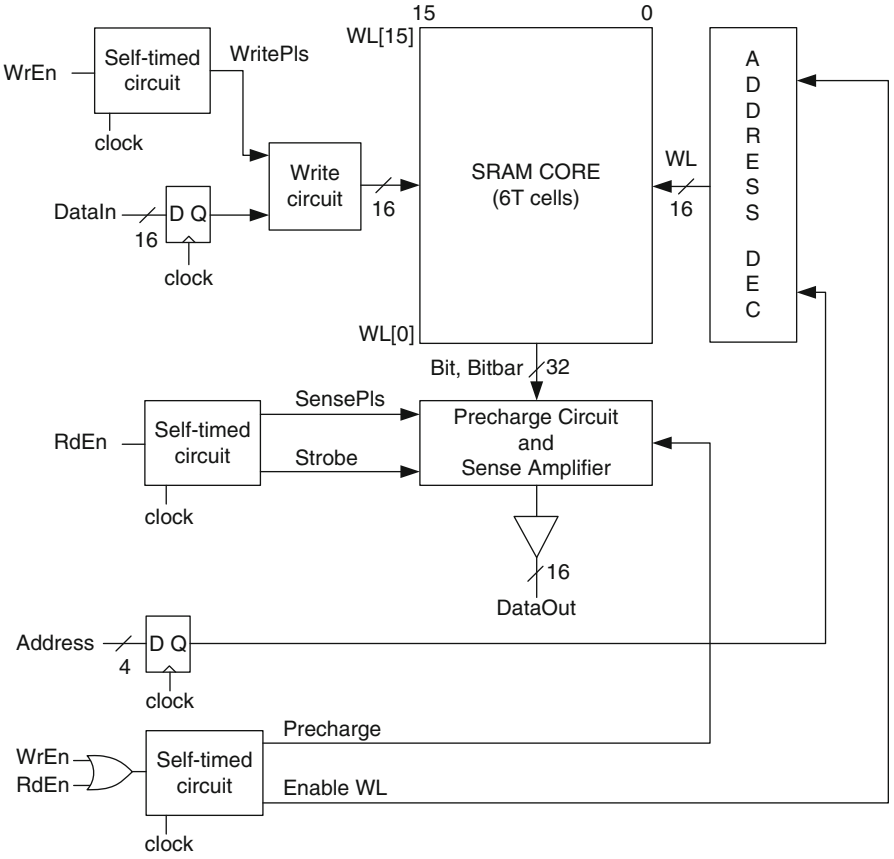

Fig. 6.1 16 × 16 SRAM architecture

current value of the NMOS SNT is twice as large as the ON current of the PMOS SNT. The low and high noise margin figures are 360 mV and 440 mV, respectively.

A typical write operation is initiated with a high WrEn signal which produces an active-low precharge pulse immediately after the positive edge of clock as shown in Fig. 6.4. Note that the positive clock edge corresponds to the origin of each graph. Figure 6.4 does not represent the actual waveforms obtained during a write operation; however, it shows the complete write sequence and data validation checks under nominal conditions. During precharge pulse, both Bit and Bitbar lines are pulled up to 1 V prior to accessing a memory cell. The voltage level on Sense and Sensebar nodes is immaterial for a write operation since the PMOS transistors, S1 and S2, are turned off. The active-low WritePulse and the active-high EnableWL signals must be generated following the input data and address for validation, respectively. The combination of EnableWL and valid address generates a WL pulse which turns on the pass-gate transistors, N1 and N2, to access a 6 T cell.

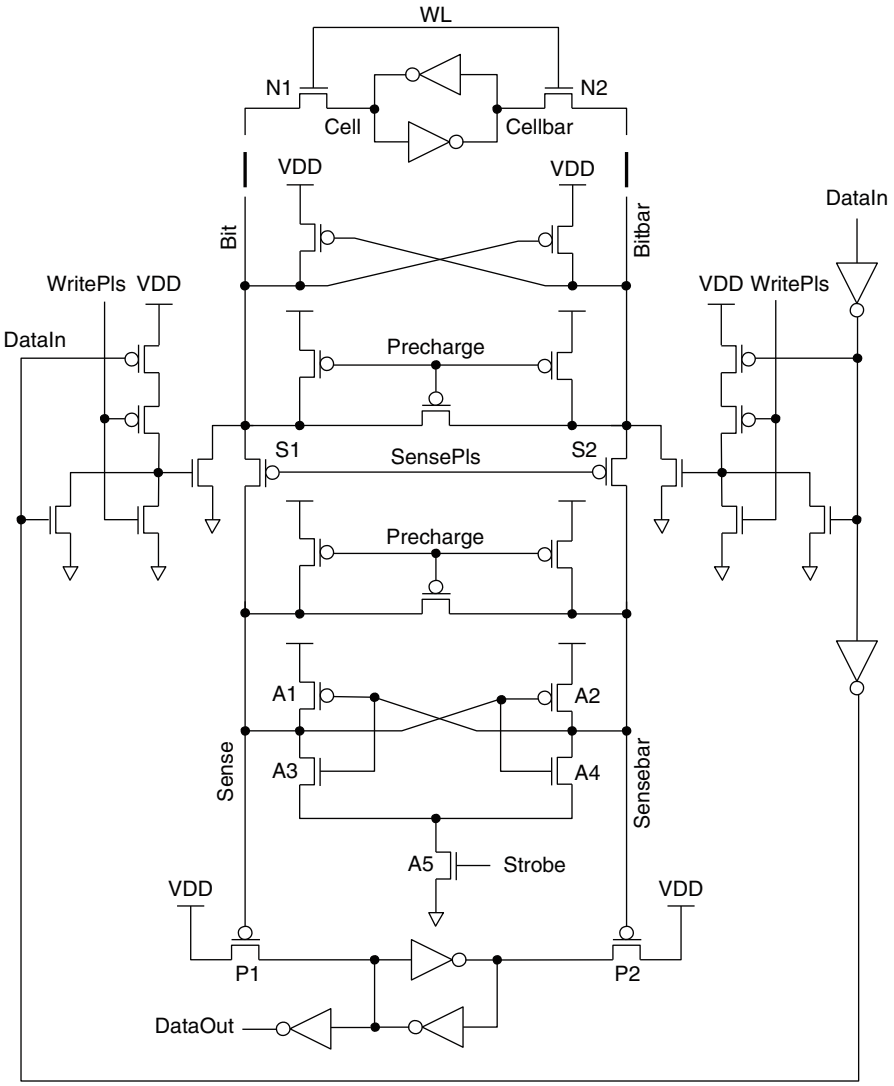

Fig. 6.2 An SRAM column consisting of 6 T cell, precharge, read and write circuits, and output buffer

While WL is active, the valid input data is written into the cell. WritePulse is terminated as soon as the input data is latched in the memory cell which alters the voltage levels at Cell and Cellbar nodes.

A typical read operation is initiated by a high RdEn signal which also produces a precharge pulse to pull up Bit, Bitbar, Sense and Sensebar lines to 1 V as shown in Fig. 6.5. Again, this graph does not show actual waveforms but reveals the complete

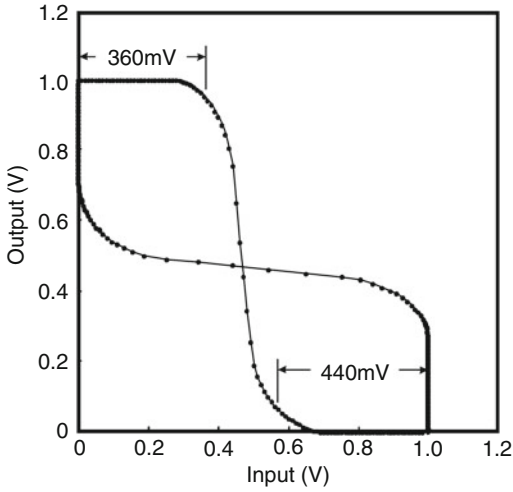

Fig. 6.3 Input–output characteristics of a 6-transistor (6 T) memory cell

read sequence and data validation checks under nominal conditions. Identical to the write operation, the active-high EnableWL signal is generated to validate an input address and turn on the pass-gate transistors of the memory cell. When the differential voltage between the Bit and Bitbar nodes reaches approximately 50 mV, the Strobe pulse is produced to activate the sense amplifier. During this period, the active-low SensePulse signal is also generated to turn on the transistors, S1 and S2, to transfer the charge from the Bit (Bitbar) node to the Sense (Sensebar) node. The ratio of the capacitance at the Bit (Bitbar) and Sense (Sensebar) nodes determines the initial voltage level at the Sense (Sensebar). The sense amplifier utilizes the higher transconductance values of the transistors, A1 and A4 (rather than A2 and A3), and pulls down the Sensebar node towards 0 V while sustaining 1 V at the Sense node. Consequently, P2 transistor turns on while P1 transistor stays off, allowing the LatchInbar and the DataOut nodes to reach 1 V. The Strobe and WL pulses are terminated when the Sensebar node reaches approximately 0 V and the contents of the memory cell are successfully stored at the LatchIn and Latchinbar nodes, respectively.

6.3.3 Address Decoder

The address decoder generates 16 WL signals for each row of the SRAM core. When enabled by the EnableWL, the valid 4-bit input address is decoded to activate only one row of the SRAM core. A disabled address decoder produces 0 V to all of its outputs. Subsequently, none of the rows is turned on for either read or write operation.

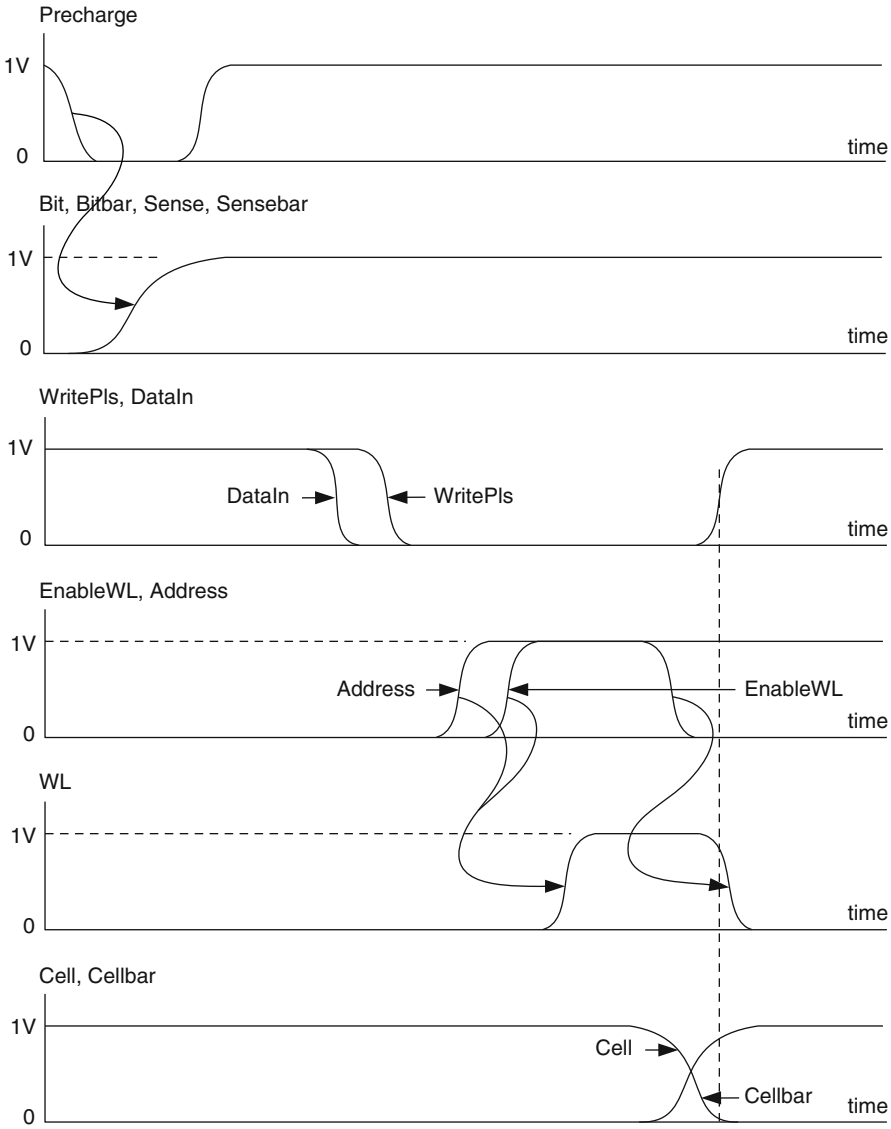

Fig. 6.4 A typical WRITE sequence

6.3.4 Self-Timed Circuits

Self-timed circuits produce either active-high or active-low signals for controlling the data flow sequence during a read or write operation. Figure 6.6 shows a typical self-timed circuit that generates an active-high pulse. Both the origin and duration of the pulse can be determined with respect to the positive edge of clock by isolated

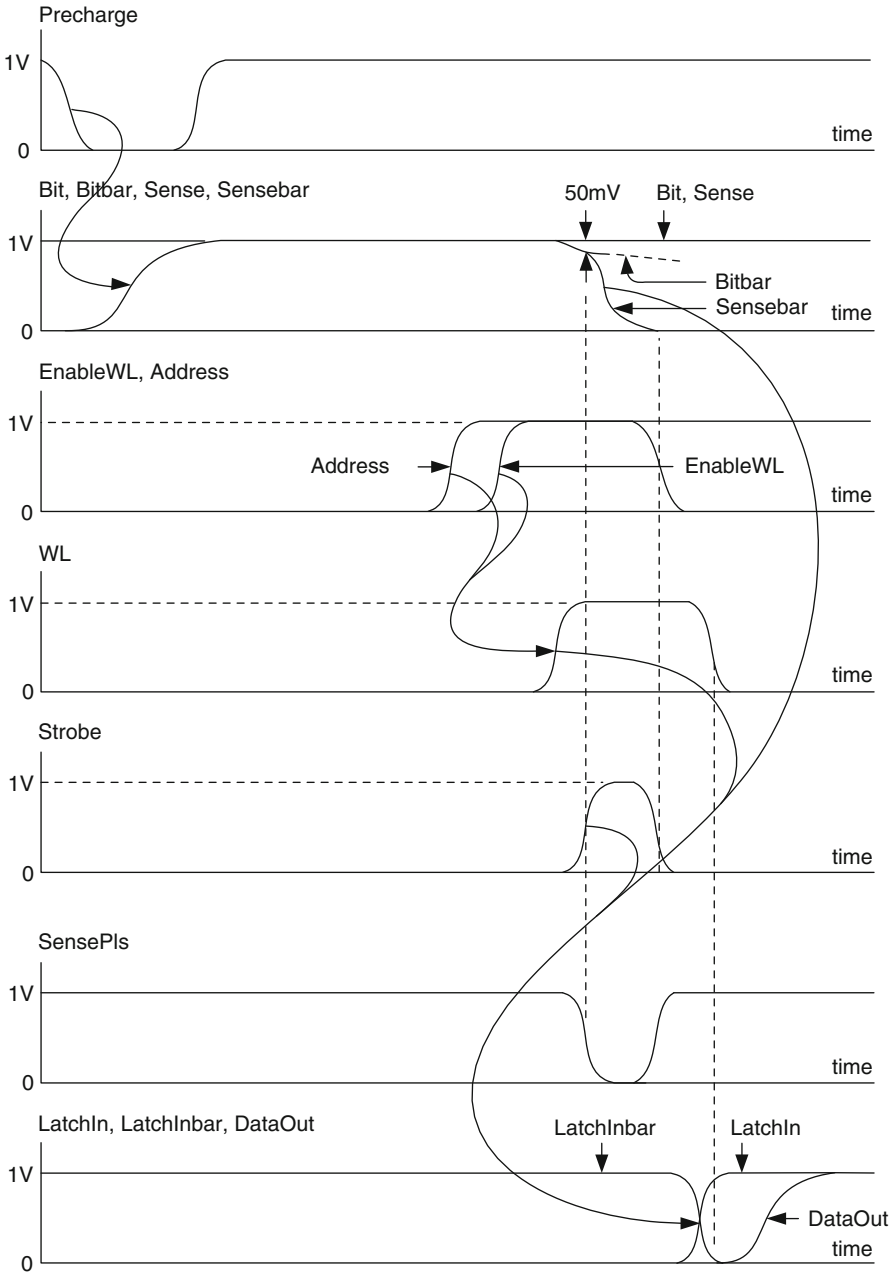

Fig. 6.5 A typical READ sequence

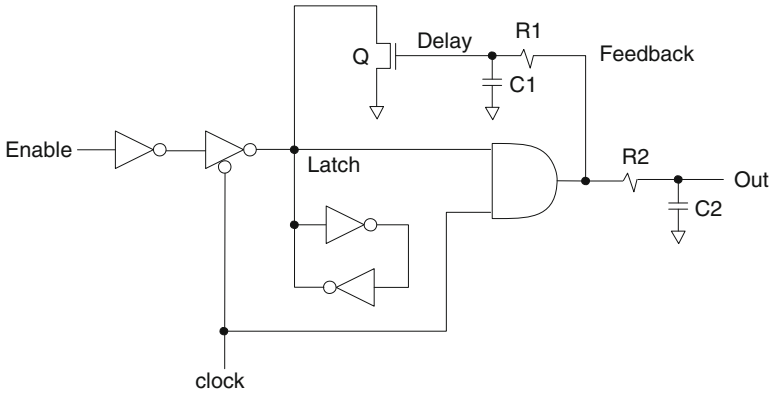

Fig. 6.6 Typical self-timed circuit producing an active-high pulse with respect to positive clock edge

Table 6.1 Time constant values of self-timed circuits

	R1 (MΩ)	C1 (aF)	R2 (MΩ)	C2 (aF)
Precharge	1.0	4.3	0.0	0.0
Enable	5.0	4.3	4.2	4.3
SensePls	9.2	4.3	7.8	4.3
Strobe	9.5	4.3	8.2	4.3
WritePls	4.0	4.3	2.6	4.3

RC circuits. An active-high Enable signal is latched in the self-timed circuit when clock is low. Since both the Feedback and Delay nodes are at 0 V, the NMOS transistor does not turn on; the voltage level at the Latch stays undisturbed at 1 V. However, as soon as clock goes high, the input tri-state inverter isolates the Latch node from the changes at the Enable input. The Feedback node goes high and pulls up the Delay node after a delay determined by R1 and C1. The NMOS transistor, Q, turns on and discharges the Latch node. Following this discharge, the Feedback node goes low and turns off Q after the same RC delay. The pulse generated at the Feedback node is reproduced at the Out node after a delay determined by R2 and C2. Therefore, R1 and C1 adjust the pulse duration while R2 and C2 establish the origin of this pulse with respect to the positive edge of clock. The values of R1, C1, R2 and C2 are listed in Table 6.1 for each self-timed circuit used in the SRAM block. The waveforms of the self-timed circuit at each critical node are shown in Fig. 6.7.

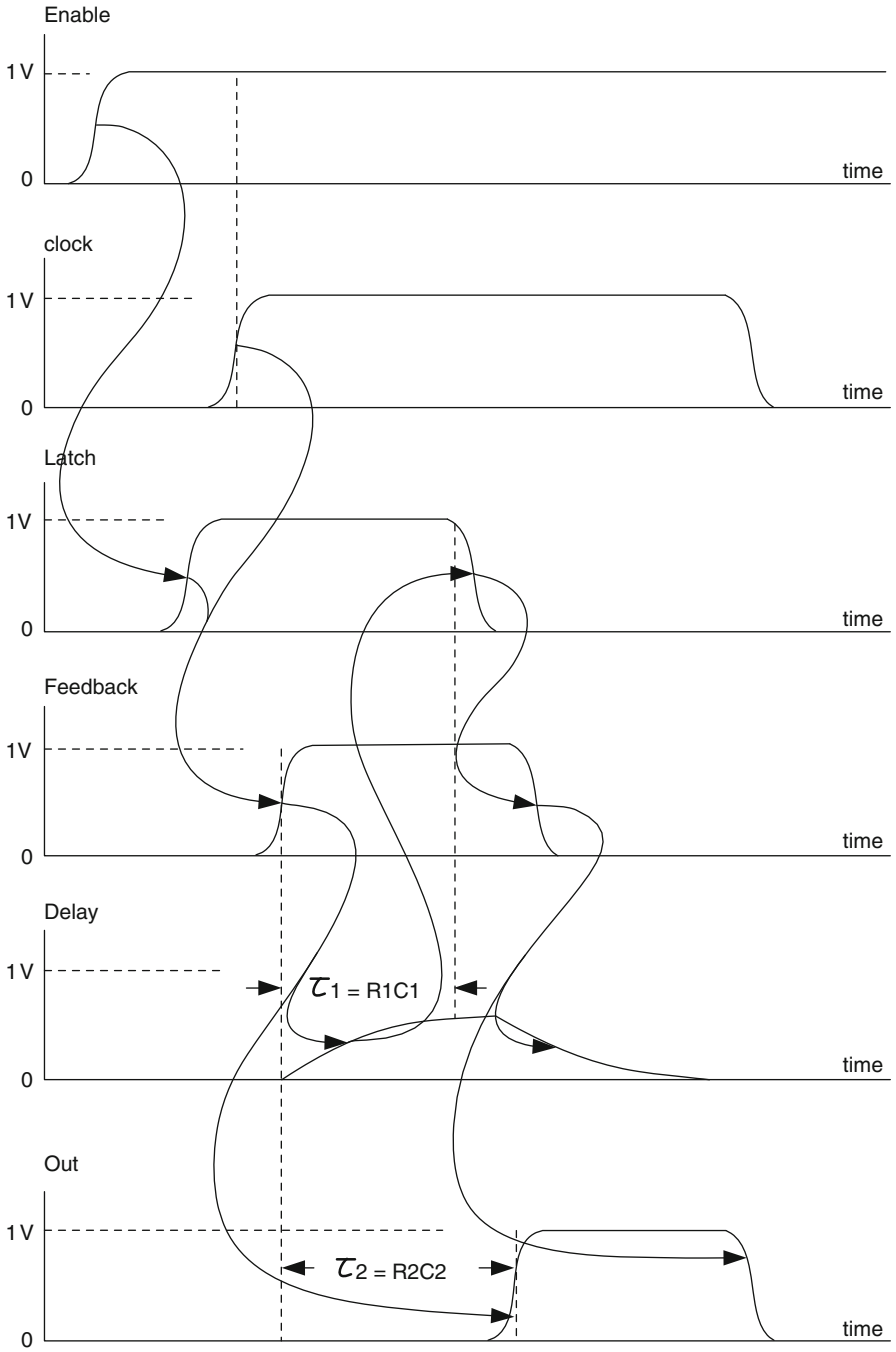

Fig. 6.7 Waveforms of a self-timed circuit

6.4 SRAM Characteristics

6.4.1 Parasitic Layout Extraction

6 nm wide and 1.4 aspect ratio (wire height to width) copper wires are used for interconnects. Since sub-10 nm range copper wire electrical characteristics do not exist in the literature, copper resistivity was extrapolated from Srivastava's model for 1.4 aspect ratio wires as discussed in Chapter 1, and 20 μΩ-cm resistivity was subsequently used to calculate the sheet resistance for 6 nm wide interconnects. Similarly, contact resistance was extrapolated from experimental data on 100 nm and larger via diameters and resulted in 18 Ω for each metal contact.

N-well and P-well contacts are formed with guard ring structures around each well periphery to reduce the NMOS and PMOS source extension resistances as discussed in Chapter 3. This scheme produces 2.3 kΩ N-well and 3.4 kΩ P-well extension resistances in series with the source terminal. Even though source extension resistance normally dominates over the combination of local interconnect and contact resistances, its magnitude is much smaller than the transistor channel resistance. The NMOS and PMOS equivalent channel resistances extracted from inverter rise and fall times are approximately 51.3 kΩ and 78.6 kΩ, respectively. Each resistance is approximately 22 times higher than the corresponding source extension resistance. The degradation in ON current due to source resistance becomes better than 8 % for both transistors. However, the guard ring structure increases the single transistor layout area to approximately 900 nm^2 and produces a gate-source capacitance, C_{GS}, of 10.8 aF. Nevertheless, the effect of source resistance has more impact on the transistor performance, and as a result transistors with guard rings are used throughout the SRAM design despite larger layout area. Besides the guard ring, another element that increases the SNT layout is the gate metal thickness. Minimum metal thickness has to be approximately 5 nm to form with moderate grain formation and continuous film coverage [1].

Parasitic device and wire capacitances are calculated using ANSOFT's two-dimensional electrostatic solver. Because of the 2 nm apart concentric cylindrical surfaces between gate and source terminals, and the resultant C_{GS} of 10.8 aF, the effective input capacitance of a single transistor has increased from 4.3 aF (the gate oxide capacitance) to 15.1 aF. The gate-drain capacitance, C_{GD}, has stayed the same at 1.7 aF in both layout topologies. The metal-metal interconnect coupling capacitance dictates the total parasitic wire capacitance and is approximately equal to 0.05 aF/nm.

6.4.2 Read and Write Access Times

The 16×16 SRAM block is designed to operate between two extreme ambient conditions. The best-case condition is defined so that the transistor ON current reaches its highest value with a 20 % less threshold voltage, 20 % more supply

Table 6.2 The best, the nominal, and the worst ambient conditions

Temperature (°C)	Supply voltage (V)	NMOS V_T (mV)	PMOS V_T (mV)
0	1.2	213	225
27	1.0	266	281
125	0.8	319	337

voltage, and 0 °C operating temperature. Similarly, the worst-case condition causes the ON current to reach its lowest value with a 20 % more threshold voltage, 20 % less supply voltage, and 125 °C operating temperature. Table 6.2 tabulates the values of the best, nominal, and worst ambient temperatures as well as the threshold and the supply voltages.

Propagation delays through address and data paths vary more than 300 %, i.e. the propagation delay at the Address node decreases from 53 ps to 15 ps when the ambient condition is changed from the worst-case to the best-case condition. Similarly, the origin and duration of each control pulse change between 100 % and 600 % with respect to two extreme ambient conditions. For example, the precharge pulse shrinks from 22 ps to 10 ps, EnableWL from 85 ps to 18 ps, SensePulse from 134 ps to 23 ps, Strobe from 134 ps to 20 ps and WritePulse from 78 ps to 28 ps between the worst-case and the best ambient conditions. Therefore, two basic timing rules should always be checked prior to a read or a write operation: (a) an associated control pulse must always follow the data for validation; this rule must especially be verified during the worse-case read and the worst-case write operations, and (b) the precharge pulse must never overlap any data validation pulse.

The longest propagation delays through the address decoder, write and read circuits, and the duration of the control pulses including precharge, EnableWL, SensePulse, Strobe, and WritePulse are shown in Table 6.3a, b for the best and the worst-case conditions, respectively. Each control pulse follows the timing rules outlined above. The worst-case EnableWL is generated approximately 4 ns after the Address signal. The worst-case WritePulse is generated 4 ns after the termination of precharge pulse and terminated 5 ns after the data is written into the 6 T cell. Strobe is produced when the differential potential reaches 50 mV between the Bit and the Bitbar lines. Each control pulse is also well formed and free of superfluous spikes at the rising and falling edges. For example, the best-case WL pulse produced by the combination of the Address signal and EnableWL has 7.6 ps rise time, 6.3 ps fall time, and 15 ps pulse width. Even though its duration almost equals to the sum of rise and fall times, the pulse is well formed for accessing the data in a 6 T cell and its rail-to-rail transition is smooth in both rising and falling edges. The Strobe pulse produced under the best-case ambient conditions is another short pulse with 20 ps duration, 11.6 ps rise time, and 13.2 ps fall time. Nevertheless, it exhibits the same mature pulse characteristics as the WL signal.

Table 6.3 Best and worst-case propagation delay and validation pulse characteristics

	Prop delay (ps)	Pulse origin (ps)	Pulse duration (ps)
(a) The best-case propagation and validation pulse characteristics			
Address	15	–	–
DataIn	5	–	–
Precharge	–	4	10
Enable WL	–	39	18
SensePls	–	71	23
Strobe	–	71	20
WritePls	–	26	28
(b) The worst-case propagation and validation pulse characteristics			
Address	53	–	–
DataIn	21	–	–
Precharge	–	18	22
Enable WL	–	57	85
SensePls	–	92	134
Strobe	–	92	134
WritePls	–	44	78

Table 6.4 Write and read access times, average dynamic power dissipations

	Read access time (ps)	Write access time (ps)	P_{AVER} (read) (μW/col)	P_{AVER} (write) (μW/col)
Best case	78	49	20.7	15.7
Nominal	90	62	11.4	8.2
Worst case	133	98	6.9	5.0

6.4.3 Power Dissipation

Average power dissipation is measured as a function of the ambient condition during the read and the write cycles. The power dissipation measurements are taken at 500 MHz and with an output capacitance of 45 aF which corresponds to a fan-out of 3 transistors, each with an effective input capacitance of approximately 15 aF. Table 6.4 summarizes the average dynamic power dissipation per SRAM column at the best-case, nominal-case and worst-case ambient conditions for the read and the write operations. According to this table, a 16×16 SRAM dissipates maximum of 331.2 μW during read and 251.2 μW during write operations at 500 MHz. Since the transistor ON current is highest during the best ambient condition and it charges a fixed nodal capacitance, the average dynamic power dissipation is highest during the best-case ambient condition than any other condition at 500 MHz.

6.4.4 SRAM Layout

The 16×16 SRAM block is composed of 16 columns of SRAM core, five self-timed circuits, and an address decoder as shown in Fig. 6.8. The total layout area of the block is approximately equal to 3.79 μm by 4.21 μm, which is merely 28 times larger than a 6 T cell in a 65 nm technology [2]. The SRAM core primarily occupies most of the layout area and is equal to 8.25 μm^2. The address decoder and self-timed circuits occupy 1.70 μm^2 and 1.97 μm^2, respectively.

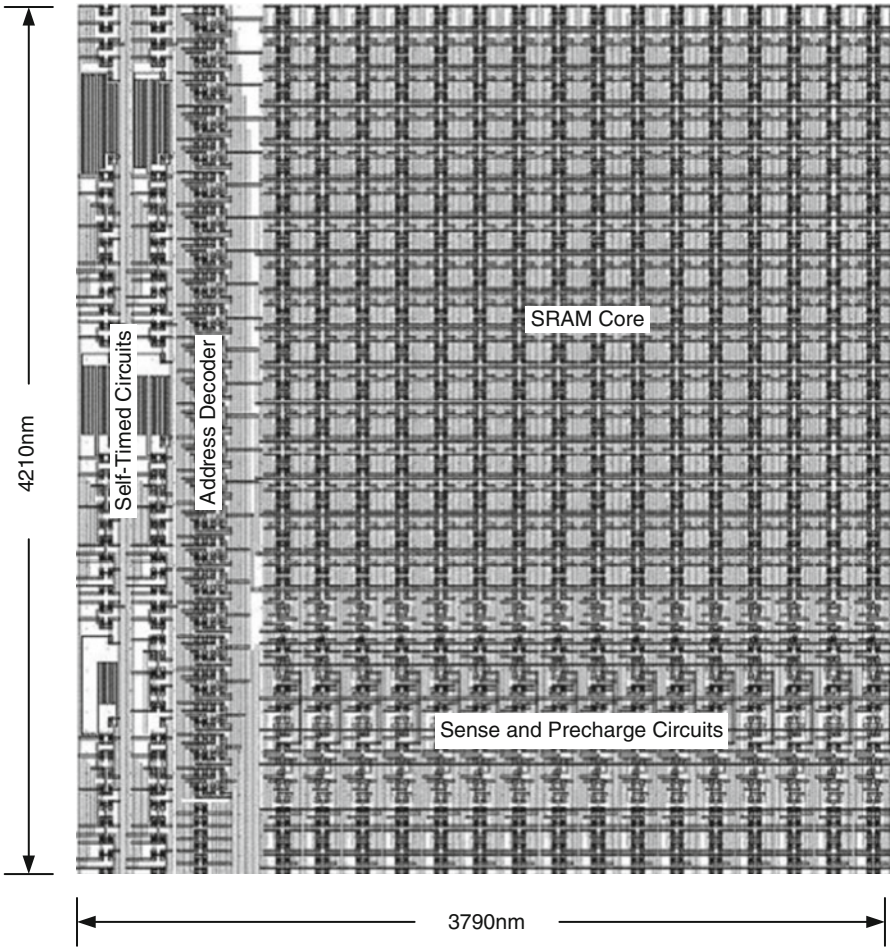

Fig. 6.8 The layout of the 16×16 SRAM block. SRAM core, address decoder, and self-timed circuits are shown on the layout

6.5 Summary

A 16 × 16 SRAM block is designed using silicon nanowire technology. In the first section of this chapter, an SRAM architecture including the SRAM core, the read and write circuits, the address decoder, and the self-timed circuits is described. The write and the read operation sequences are explained; waveforms illustrating typical data propagation and validation are presented for each case. In the second section, the performance and dynamic power dissipation figures are given. For example, the worst-case write and read access times for this SRAM block are 98 ps and 133 ps, respectively; the dynamic power dissipation is 20.7 μW per column during a read and 15.7 μW per column during a write operation at 500 MHz. The SRAM layout occupies approximately 16 μm^2.

References

1. Becker JS, Gordon RG (2003) Diffusion barrier properties of tungsten nitride films grown by atomic layer deposition from bis(tert-butylimido)bis(dimethylamido) tungsten and ammonia. Appl Phys Lett 82(14):2239–2241
2. Bai P et al (2004) A 65nm logic technology featuring 35nm gate lengths, enhanced strain, 8 Cu interconnect layers, low-k ILD and 0.57 μm2 SRAM cell. IEDM. pp 657–660