Memory System and Arithmetic Unit

Group Number: 12, Members: Abhishek Deshmukh, Ankit Kaul, Chirag Mehta, Varsha Sankar

***Abstract*—This paper outlines the design and analysis of a pipelined 64-byte Static Random Access Memory (SRAM) system and an 8-bit adder in 45-nm FreePDK technology. The data is written to the SRAM array using registers, read sequentially from the array, delivered to the adder system through the interconnect and added for n cycles. The system is designed to operate at a maximum supply of 0.8V with a target clock frequency of 1GHz. Optimizations were performed for key system parameters like power, delay, read and write margins, etc. This system is capable of operating at a scaled voltage of 0.7V with a maximum operating frequency of 1.35 GHz.**

I. OVERALL DESIGN AND TARGET

The objective is to design a digital integrated system consisting of a 64-byte SRAM and an 8-bit adder. The data is written to the SRAM array using registers, read sequentially from the array, delivered to the adder system through the interconnect and added for n cycles. This is accomplished by using a 16x32 SRAM array, where 512 bits of data are stored as 64 8-bit words (16 rows x 4 words/row).

Each word of the SRAM is accessed by a 6-bit address (A0-A5) written into an address register. The A0-A3 bits specify the row address and the A4-A5 bits specify the column address of the SRAM array. A 4:16 decoder is used to decode the row address and a 2:4 decoder decodes the column address.

The clock signal and individual output signals of the row decoder are given as inputs to 16 different AND gates, the outputs of which are termed as word lines (WL). The WL logic ensures that all the word lines go low at the falling edge of the clock signal, thus preventing. The output of the column decoder is used to drive the column multiplexers which selects the 8 bit data to be written to or read from the SRAM array. The data to be written into the SRAM array is obtained from the WRITE register, and the data read from the SRAM array is stored in the READ register.

The data read into the READ register is transferred to the input register of the 8-bit adder through a highly capacitive long interconnect (150fF). The adder sequentially adds the 8-bit data coming from the input register with the stored value in the accumulator register. The accumulator register is initialized to zero for the first add operation. The result is stored back in the accumulator register and the add operation continues for n clock cycles.

II. METHODOLOGY

A. Design Methodology

1) Decoder: The 4:16 row decoder and 2:4 column decoder are designed for minimum delay, at the same time satisfying the power requirements. Using the method of logical effort, the NAND2INV topology for a 2:4 decoder was found to provide the minimum delay.

The 4:16 decoder was designed using five 2:4 decoders. One 2:4 decoder serves as the 'Master' and the other four as 'Slaves'. The output of the Master decoder is used to turn on the Slave decoders, which ensures that only one Slave and the Master are active at any given time. This design thus significantly reduces the transistor count and the overall power consumption.

2) Registers: Positive edge-triggered Master-slave flip-flops were used in all the registers. The hold time of this circuit topology is zero and hence the circuit is designed to satisfy the set-up time and clock-to-Q delay requirement.

The clock-to-Q delay of the MS flipflop is essentially a function of the load at the output. The circuit was designed and tested for an inverter load and it was ensured throughout the operation that the MS flipflop sees an inverter load at the output. The nMOS and pMOS in the transmission gates are sized in accordance with the ratio of the electron and hole mobilities. This is done to ensure that the resistance offered by both nMOS and pMOS and the delays through them are equal. The width of the pMOS is set to be twice the width of the nMOS as the mobility of electrons is roughly twice that of holes.

To determine the set-up time for the flipflop, the rising edge of the input D was progressively brought near the rising edge of the clock pulse, until D is currently sampled at the output or until we get a constant clock-to-Q delay. The minimum value of the time between the rising (or falling) edge of the data D and the rising edge of the clock pulse for which the data is correctly sampled at the output is the set-up time for the flip-flop.

3) WL Logic: The purpose of the WL generation logic is to ensure that during the negative half cycle of the clock signal, none of the word lines are at a high logic level. Thus, the WL generation logic block was built by 'AND'ing the outputs of the row decoder with the CLOCK signal, for each word line.

4) SRAM Cell: The SRAM Cell uses 6 transistors for storing one bit of data. Access to the cell is enabled by the word line, which controls two access transistors that are shared during read and write operations. Two bit lines are used to transfer the stored signal and its inverse to and from the READ and WRITE registers.

The SRAM cell needs to be sized as small as possible to achieve a smaller cell area. Reliable operation of the cell, however, imposes some sizing constraints. An initial starting size of Wpun = Wmin = 90nm, Waccess = 1.5Wpun, and Wpdn = 2Wpun is assumed. The read and write margin targets were achieved for this initial sizing itself. To minimize the area, further optimizations were performed to reduce the transistor sizes, while still meeting the target specifications. The final sizing for optimal performance uses W = 90nm for all pull-up, pull-down and access transistors.

5) Read Circuit and Sense Amplifier: A current latch based sense amplifier is used to sense the relatively small differential voltage across the bit-lines of the SRAM cell and amplify the voltage to recognizable logic levels so the data can be interpreted properly. A NAND gate based latch further stabilizes the outputs of the sense-amplifier by sharpening the transitions.

The transistors are sized to create a larger current difference on both sides, in order to get a lower bit-differential required to correctly sense the output. The minimum bit differential for the sense amplifier was determined under 10% mismatch between the left side and right side transistors so as to provide worst case currents through the two branches. External stimuli were applied to the two bit-lines and The minimum difference between the two bit-line voltages for which the sense amplifier correctly sensed the output was determined. The designed sense amplifier is able to detect a minimum bit differential of 6mV while providing a critical read delay of 241.3psecs.

6) Write Circuit: The write circuit allows the data from a write register to be written into the SRAM cells. A write enable (WE) signal controls the operation of two transmission gates, which are followed by driver circuits. The WE signal is externally applied to the circuit. The driver circuits help drive an estimated 30fF load at the bit-lines.

7) Column Multiplexers: The 4:1 column MUX's are designed using transmission gates. The outputs of the column decoders are the select lines for the column multiplexers, which select the correct 8-bit data to be written to or read from the SRAM array. The nMOS and pMOS in the transmission gates are sized in a manner similar to that used for the registers.

8) Adder Circuit: An 8-bit adder circuit is built by cascading multiple 1-bit adders. A 1-bit Carry-Look-Ahead adder is constructed using Manchester Carry chain. The outputs are a function of 3 intermediate signals- 'generate', 'propagate' and 'kill'. The advantage of using the Manchester Carry chain is that only one out of the three intermediate signals are active for any combination of the input signals. As a result, the 'CARRY' output is made available simultaneously at the outputs of all the 1-bit adders, except when the carry needs to propagate through the stages for some combination of the input. This significantly reduces the delay of the adder circuit.

9) Inverter Chain Buffers: When the signals are fed from the row decoder into the word lines of the SRAM cells, highly capacitive loads are seen at the output of the WL logic block. Also, the SRAM READ register needs to drive a long capacitive interconnect to reach the INPUT register of the adder. Buffer chains are required in order to drive such large loads.

To determine the number of inverters 'N' and the upsizing factor 'u', the input capacitance of the unit inverter is required. Two unit inverters were connected in cascade. Applying a clock pulse at the input, the rise time of the output at the intermediate node (output of the first inverter) was measured. The finite rise time of the voltage at this node is primarily due to the capacitance offered by the second inverter. This capacitance is the input capacitance (Cin) of the unit inverter. The second inverter was replaced with a capacitor, whose capacitance was varied to obtain the same rise time, as in step 1. The value of Cin thus obtained was approximately equal to 0.416 fF for a unit inverter. Using this value for Cin, the number of inverters 'N' and the upsizing factor 'u' were determined for driving the SRAM word lines and the long interconnect.

10)Layout: The main objective of layout is to minimize the area while creating a modular system in order to minimize the coupling capacitances. The layout of a single SRAM cell was created first, which was then flipped and rotated to create a 2x2 array. This was further merged to produce a 4x4 array and the process was repeated to to create the complete modular 16x32 array. The SRAM peripherals - Precharge, Column MUX and Transmission gates of the Write circuit and the Word line logic and Buffer chain were also laid out taking into consideration the pitch matching. The total area efficiency of the SRAM was found to be 64.256%. Parasitic extraction was carried out considering one row and one column of the SRAM Array and Post layout simulations for delay and cell margin were performed.

B. Performance Metrics

The performance metrics for all the components and peripherals in the system were analyzed. The table below lists down the simulated values for the various performance parameters.

Table 1. Performance Metrics and Design Parameters of all Peripherals

Component	Parameter	Value
Row Decoder	Power	45 uW
	Delay	134.2 ps
Column Decoder	Power	603.4 nW
	Delay	18.42 ps
Register	Setup Time	22.2 ps
	Hold time	0 ps
	Clock to Q Delay	35.34 ps
Interconnect Driver	Number of Stages	6
	Sizing Factor	2.67
	Power	13.52 uW
	Delay	131.32 ps
Column MUX	Delay	15.01 ps
	Power	549.3 nW
Write Circuit	Critical Delay	143.119 ps
	Power	46.25 uW
Bit Line Driver	Number of Stages	4
	Sizing Factor	3.1
Word Line Driver	Number of Stages	4
	Sizing Factor	3.1
Read Circuit	Critical Delay	241.3 ps
	Power	1.574 uW
SRAM Cell Array	Total Read Delay	126.3 ps
	Total Write Delay	326.4 ps
	Pre-charge Delay	500 ps
	Power	413.95 uW
Adder	Power	18.9 uW
	Delay	$<$ 1 ns

Table 2. SRAM Cell Sizing and Performance Metrics before extraction

Parameter	Value
Read Voltage	120.136 mV
Trip Voltage	363.23 mV
Read Margin	243.096 mV
Read Current	12.84 uA
Write Margin	278 mV
Access Time	14.02 ps
Pull-up P-mos Size	90n/50n
Access N-mos Size	90n/50n
Pull-down N-mos Size	90n/50n

C. Signal Synchronization

Signal synchronization for the system involved synchronizing the address and data with the clock for the ADDRESS and WRITE registers respectively, and generating the Precharge and Sense Amplifier Enable (SAE) signal for the Sense Amplifier.

The 6-bit address and the 8-bit data are input to the ADDRESS and DATA registers. The ADDRESS register latches and transfers the first four bits (A0-A3) to the row decoder and the next two bits (A4-A5) to the column decoder at the rising edge of the clock pulse. In order to satisfy the set-up time requirement for the registers, the address and data must be applied sufficiently before the set-up time to ensure that the registers correctly latch the data.

The write operation is controlled by the independent Write Enable (WE) signal, which is an external input to the system. The SAE signal to the sense amplifier is obtained by ANDing WE', CLK' and the ORing of all 4 column select signals. This is because the read operation should be performed when the clock is low, write operation is complete and at least one of the bit lines are selected by the column MUX.

The Precharge signal is also a function of the clock. For this design, Precharge has been configured to be equivalent to the CLOCK. During the write operation, the read circuits and adders are disabled to prevent loss of data. During the read operation, WE disables the write circuit.

D. Testing Methodology

Testing for functionality and performance was performed for each individual components/peripherals and for all intermediate integration stages. The basic building blocks were tested for all possible input combinations with a certain capacitive load at the output.

The row and column decoders were individually tested by applying external stimuli as pulses to ensure that the correct output is selected by the decoder. The entire combination of row decoder, WL logic and buffer chain was tested with a 30fF capacitor load. This ensured three operations: the correct word line is selected by the decoder, the WL drops to zero at the clock falling edge and the decoder is able to drive a highly capacitive load at the output.

The adder and accumulator circuits were tested by first applying a constant value at the input, adding it with the value stored in the accumulator and updating the accumulator value with the result of the addition. Once the addition operation was successfully verified, the adder was tested for the worst case input condition '01010101' and '11111111', wherein the carry was made to propagate all the way from the input to the output. The adder delay under worst case input condition was determined to ensure proper functionality.

The sense amplifier was tested under symmetric conditions by externally generating a bit differential greater than the minimum bit differential derived for the amplifier and observing the signals and delay at the output of the stabilizing latches.

The SRAM array was tested by generating a read/write sequence of Write-0, Write-1, Read-1 to the same cell. Appropriate inputs were externally applied to the circuit to turn ON/OFF the precharge circuits and sense amplifier. 30fF capacitors were connected to the bit-lines to simulate the expected parasitic capacitors and also test the driver chain at the output of the write circuit.

An exhaustive testing was performed on the complete system by changing the input patterns to both the ADDRESS and DATA registers. This ensured that different cells of the SRAM array are accessed every time and different data are written to and read from the array. The complete system was also tested with the data switching between '00000000' and '11111111' with a carry input of 1. This combination proves to be the worst case condition as all the data bits are toggling plus the carry is propagated all the way from the input to the output.

III. RESULTS

The figure below shows the output of the full system for one possible combination.

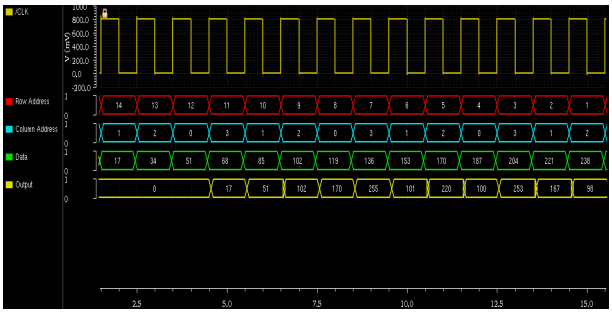

Fig. 1. Full system output

The full system is divided into three stages and analyzed for power and maximum operating frequency. Stage 1 (memory) includes the ADDRESS register, the decoders, the pre-charge, read and write circuits and the SRAM array. Stage 2 (interconnect) includes the interconnect driver chain from the READ register to the INPUT register of the adder. Stage 3 (adder) includes the CLA adder and the accumulator register.

Fig 2 shows the power dissipation in the three stages of the system. Stage 1 is found to be the power bottleneck as it dissipated the maximum power out of all the three stages. Stage 1 dissipated 439uW or nearly 93 percent of the overall system power dissipation, which was 471.2 uW at an operating

voltage of 0.8V. This is predominantly caused by the power dissipation in the pre-charge circuit because the transistors in the pre-charge circuit are sized relatively high. Although only one cell is accessed at a time, all the highly capacitive bit-lines are charged during the clock OFF period.

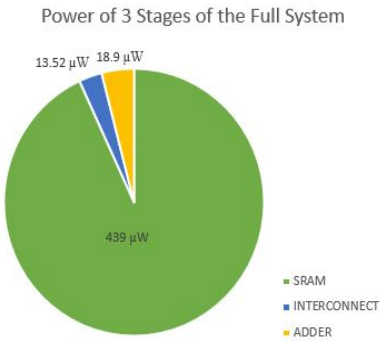

Fig. 2. Power Dissipation in different stages

The maximum operating frequency of the system at nominal power supply voltage of 0.8V was found to be 1.9GHz. Stage 1 is found to be the frequency bottleneck of the system as the WL logic and the SRAM array fail at significantly high frequencies. The circuit's functionality and performance at high frequencies allows us to scale down the supply voltage to reduce the overall power consumption. Figure shows the curve of maximum operating frequency versus the supply voltage. It can be seen that the circuit has the potential to run at a scaled supply voltage of 0.7V at a maximum operating frequency of 1.375GHz.

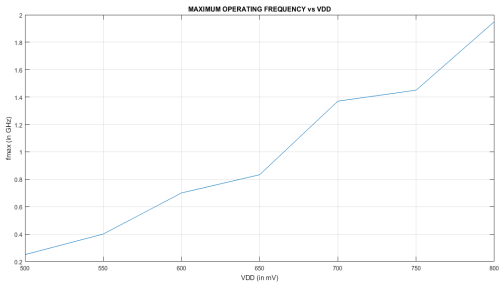

Fig. 3. Maximum operating frequency vs VDD

A comparison between the SRAM parameters before and after extraction is presented below.

Table 3. Comparison of SRAM Parameters before and after extraction

Parameter	Before Extraction	After Extraction
Trip Voltage	363.23 mV	363.23 mV
Read Voltage	120.136 mV	118.893 mV
Read Margin	243.096 mV	244.337 mV
Write margin	278 mV	278 mV
Access Time	214.02 ps	1194.1 ps
Total Read delay	626.3 ps	714.4 ps
Total Write Delay	326.4 ps	124.9 ps

IV. DISCUSSIONS

Certain key limitations can be seen in the system design. An ideal clock is used throughout the project. Since only a part of the design was laid out and not the entire system, inefficiencies arising out of clock overlap, clock skew and jitter have not been taken into account.

Due to the finite delay between the CLOCK signal and the data arriving at the input of the Word Line logic, the WL Logic captured the previous state of the decoder for a small period of time before capturing the right state, resulting in glitches at the Word Line output. This required the use of decoders with 'ENABLE' signal, which is CLOCK delayed by a certain amount. This increases the fan-in of the decoder gates and subsequently the delay of the stage. Better synchronization methods can be developed to eliminate the glitch and reduce the delay.

The designed CLA adder has the capability of generating only an 8-bit sum at the output. The accumulator register in the adder system also has no provisions to reset the data. This can sometimes cause unwanted data to be present in the accumulator during the first three CLOCK cycles, resulting in undesired system operation. This introduces the scope for future work, wherein higher bit adder systems can be designed with reset mechanisms to add and accumulate larger sums.

V. CONCLUSIONS

A 64-byte Static Random Access Memory and Arithmetic unit was designed at the transistor level and tested for power-performance optimization. The complete SRAM array, along with some peripherals, was laid out to verify the functionality. The performance parameters of individual components, peripherals and integrated system were simulated and recorded. Using the layout, parasitic elements such as coupling capacitances, wire capacitances and wire resistances were modelled and simulated using the extracted netlist from the layout.

The integrated system successfully meets the target operating frequency of 1GHz at a nominal supply voltage of 0.8V, while satisfying all other target specifications.