SRAM Register File for a 16-bit RISC processor

Akhilesh Gautam
Student
16, 2414 Bishop St.,
Ann Arbor, MI 48105
+1.734.565.966
akgautam@umich.edu

Michael Krug
Student
2281 Bonisteel Blvd
DuderstadtCenter
AnnArbor
mjkrug@umich.edu

Abishek Madhavan
Student
2281BonisteelBlvd
DuderstadtCenter
Ann Arbor MI 48109
abishek@umich.edu

Joe Harman
Student
2281BonisteelBlvd
DuderstadtCenter
Ann Arbor MI 48109
jgharman@umich.edu

ABSTRACT
Register Files(RF) are commonly used memory modules in RISC architectures to store and access data . SRAM based register files are implemented to provide higher performance and better robustness. In this paper, we describe the design of a SRAM based Register File in a 16 bit RISC Processor which provides fast access times for both read and write and also ensures the robustness of the RF.

Categories and Subject Descriptors
B.3.1 [**Memory Structures**]: Semiconductor Memories – Static Random Access Memory (SRAM)

B.3.2 **[Memory Structure]** : Design Styles – *Small Capacity, Multiport Memory*

B.3.3 **[Performance and Reliability]** : *General*

General Terms
Performance, Design, Reliability.

Keywords
8-T SRAM, bitcell, Read Robustness, Writability, CK (System Clock), LSB(Least Significant Bit), MSB(Most Significant Bit)

1. INTRODUCTION
The SRAM RF is required to do 2 simultaneous read and 1 write operations in a single clock cycle These two read operations can be at different addresses or at the same address. The worst case would be a read followed by a write operation at one particular address. Instead of a 3-port SRAM which is area intensive in terms of layout, we decided to implement a 2-port (dual port) SRAM, and used it as 3-port SRAM by sharing one of the ports for both read and write operations and the other is purely dedicated to read.

2. SRAM ARCHITECTURE
The SRAM is triggered by positive edge of CK. The Address and Data busses are settled before the CK rising edge. Write enable Signal (WE) is also asserted at the CK rising edge, in the case of an instruction requiring a writeback to the RF. At the CK rising edge SRAM starts two read operations and contents of the SRAM stored at corresponding locations are presented on output buses corresponding to both the ports after a nominal read-access time of 496ps. In the case when write is asserted by the instruction, the write operation is done after the read has completed and it takes a nominal time of 933ps to write the contents to the corresponding location in the SRAM. The SRAM requires CK to be high for the full read access time. The architecture of SRAM allows WE signal to be deactivated at any time after CK falls. After the CK falling edge SRAM takes some time to recover. Total time from CK rising edge up to the recovery is the cycle time of SRAM.

The capacity of the SRAM RF is 16 words in size. The outputs of the SRAM bitcells are connected to a voltage mode sense amplifier (SA). SRAM core is organized as a mux2 column based structure, i.e. instead of having 16 rows and 16 colums of bitcells we have 8 rows and 32 columns. This bitline cap is reduced to give speed enhancement as well as to match the sense amplifier layout width with the width of 2 bitcells. Out of the 4 bits of address, LSB is used in column decoding and 3 MSBs are used in row decoding. 3 MSBs if address decode the 8 wordlines (for both the ports) and this decoding is done inside the controller so that the SRAM gets 8 decoded wordlines for each of the ports settled at the CK rising edge.

When the CK is low all bitlines are precharged. At the CK rising edge precharge is disabled and a word-line clock pulse is triggered which is gated with the 8 wordlines. The desired wordlines (for both the ports) are selected after some delay from the CK positive edge. The two bitline pairs of the columns are connected to the sense amplifiers through a mux controlled by the LSB of the address. As soon as the sense amp sees sufficient differential voltage, a sense enable pulse is triggered. The sense amps amplify the differential voltage for their corresponding bits. The output of the sense amps is latched before the SAen pulse is disabled. This allows the proper data to be sent to the output after a read has occurred.

After the read operation has completed, a delay circuit allows the WE signal to be sent to disable the precharge of the write bitlines and activate the write driver. The write driver overwrites the selected bitcells through both write bitlines. After that when both CK and WE go low, the precharge of both of the read bitlines and the write bitlines are enabled. Hence, in the case of only writes on the 2^{nd} port, a dummy read occurs before the write operation is triggered.

Therefore, the read operation is self timed and the SRAM read is dependent on CK being high. However, the write operation is not self timed and is also independent of the CK, hence we provide the user with an asynchronous write capability.

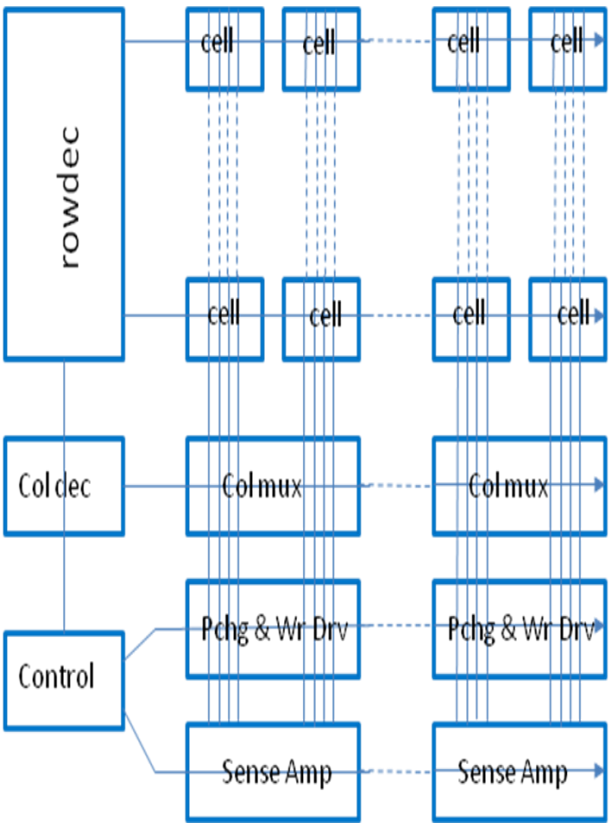


Figure 1: SRAM Register File Architecture

.

3. SRAM CIRCUITS

3.1 Bitcell

The SRAM is based on 8-T dual port bitcell. In the bitcell, we have 2 pairs of bitlines and 2 wordlines(wl1 and wl2). One bitline pair and one wordline is used only for read operation. So at the time of read, both wl1 & wl2 are selected and one bitline out of both bitline pairs is discharged. At the time of write, only wl2 is selected and one bitline, either bl2 or bl2b, out of the second bitline pair is pulled down by the write driver. For 2 simultaneous read operations, the PDN(pull-down network) has to be sized up for more robustness to prevent the internal nodes from flipping their stored bits. Hence, PDN is sized up relatively in comparison to the access transistor. PUN(pull up network) is kept weaker than the access transistor (Pass Gate) to ensure write robustness.

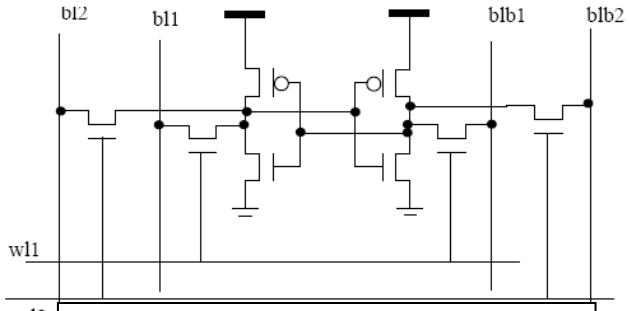


Figure 2: 8-T SRAM Bitcell

3.2 Sense Amplifier

A voltage mode sense ampilifer is used. Other sense amp designs such as charge transfer based sense amps, current mode sense amps were considered. However, the voltage mode sense amp is chosen due to higher robustness and minimum Vdiff voltage. The voltage mode sense amp was able to sense the signal at the highest speed. In CK low period, internal nodes of sense are precharged to VDD and are connected to bitlines. There is a phase equalizer on the precharge devices to ensure read robustness. In evaluation phase, when sufficient differential voltage appears between the bitlines, a Sense Amp Enable (SAEn) pulse is triggered to enable the sense amplifier to sense the voltage drop and amplify the voltage based on the difference. Bitlines are decoupled from internal nodes and differential voltage is amplified and latched before sense enable pulse falls down. Precharge of sense internal nodes starts when both CK and WE falls down. The circuit of the sense amp is shown below:

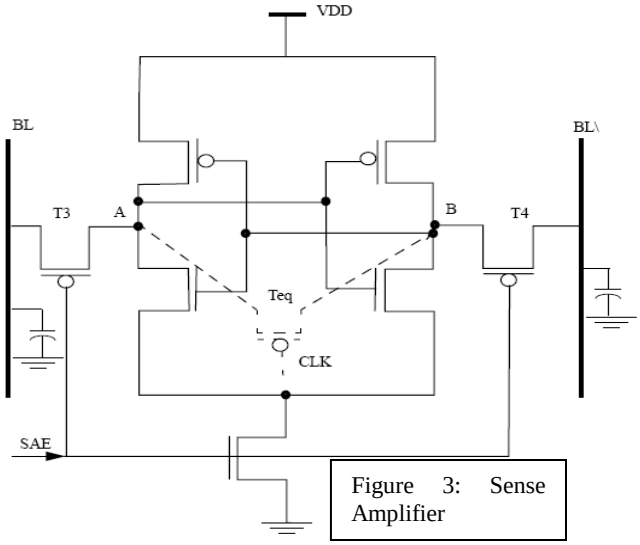
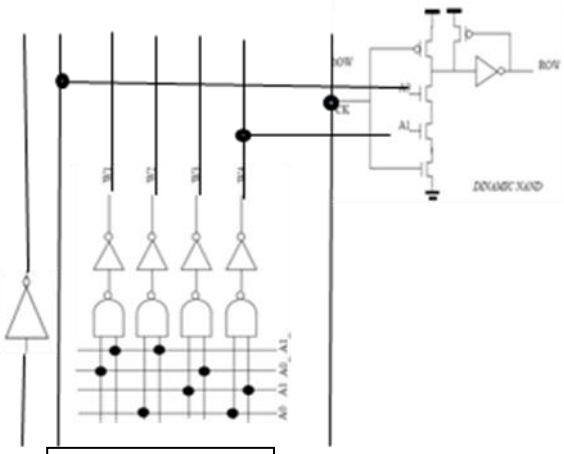


Figure 3: Sense Amplifier

3.3 Decoder

The LSB of the address determines the column decoding. Out of 3 MSBs, 2 bits are pre-decoded using a 2x4 static decoder. The final MSB is again decoded using a 1x2 decoder (an inverter). Then 4 and 2 predecoded lines are connected to a dynamic NAND decoder to finally decode 8 WLs.



. Figure 4: Decoder

3.4 Control

For wordline selection, a pulse is triggered by the rising clock, and then fed to the dynamic NAND decoder. For sense amp triggering, the generated pulse is matched with the rising delay of the bit line discharge to give a sufficient differential voltage. Its falling delay is matched with the sense amp reaction time plus output latching time. For write, a delayed signal is generated from the WE rising edge, matched to be greater than the read access time. This is done using a buffer chain with tristate drivers to ensure that write doesn't disturb the read operation.

4. ROBUSTNESS AND PERFORMANCE

For the read operation, the bump generated at the internal node of the bitcell is compared against its static noise margin. This is done with Monte-Carlo simulations to take process variations into account. In order to save on simulation time, speedpath schematics modeling the vertical and horiziontal RC delays were used to simulate the parasitic capacitances. The Gaussian distribution obtained for the two quantities are far apart.

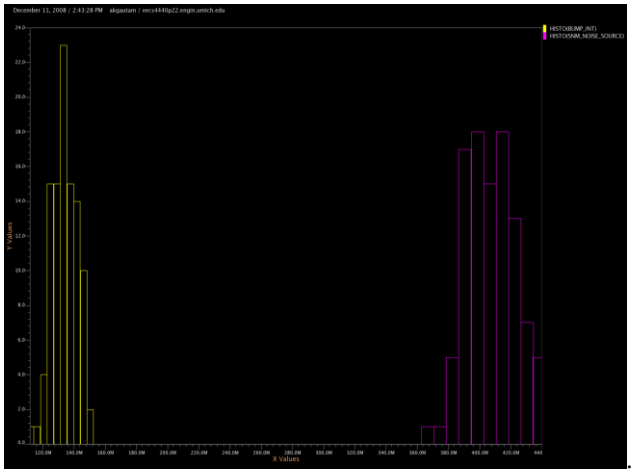


Figure 5: Histogram showing Bump voltage distribution

For write robustness, the write margin of the bitcell is measured under Monte-Carlo variations and ensured that nominal-3*sigma of write margin value is not too low.

Similarly, there is marginality in SRAM which states that delayed WE should not come before the read operation is completed. The two delays were measured with Monte-Carlo variations and their Gaussian distributions are not overlapping more over than the desired range.

The distribution of the read and write access times based on our simulations done using 0.13um technology, is as follows:

	Read Time[ps]	Write Time[ps]
Nominal	496.99	933.58
Sigma Variation	35.037	64.390

Hence, using the above values, the minimum clock period required for the overall system is 1.126ns and thus the frequency of the processor at max is 888Mhz. It is important to note all the Monte-Carlo runs were conducted at all process corners to ensure reliability in results.

Similarly, for read robustness with respect to different impacts process variation on selftime path and actual path, and also the bitline leakage through unselected rows, we have measured differential voltage appearing on bitline pairs at the time when sense is enabled, with Monte-Carlo variations. Also the offset requirement of the sense amp is measured with Monte-Carlo variations. Then it is ensured that the Gaussian spread of the two quantities are not overlapping.

Read access time of the SRAM obtained at default PVT is 497 ps. Then after that write operation is done in 934 ps. This is equivalent to 15 times FO4 delay in this technology. With Monte-

Carlo variation taken into simulation, the standard deviation of read and write times come out to be 35 ps and 65 ps, respectively. So after adding 3 sigma to nominal value (to cover the extreme case), total time (up to write done) is 1.126 ns. The default PVT is 1.2 V, 25C and both PMOS and NMOS at typical corners. Maximum delay up to write time is 2.65 ns (at 0.9V, 150C and both NMOS & PMOS at their 3 sigma slow corners)

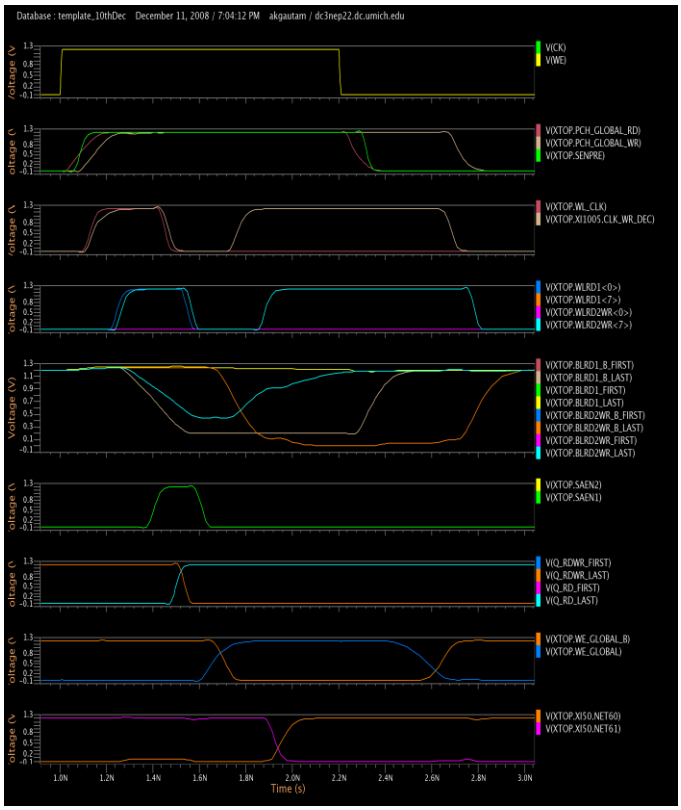


Figure 6:Performance Simulation showing Read and Write access times

5. ACKNOWLEDGMENTS

Our thanks to IBM for providing 130 nm CMOS technology SPICE models. Our sincere thanks to Prof. David Blaauw for his mentoring on the SRAM topics.

6. REFERENCES

[1] Digital Integrated Circuits by Jan M. Rabaey, Prantice-Hall publications

[2] Manoj Sinha, Steven Hsu, Atila Alvandpour, Wayne Burleson, Ram Krishnamurthy, Shekhar Borkar,LowVoltage Sensing Techniques and Secondary Design Issues for sub-90nm Caches