The development of a radiation tolerant low power SRAM compiler in 65nm technology.

R. Brouns[a], S. Bonacini[b], L. Berti[a], S. Verhaegen[a], K. Kloukinas[b], G. Thys[a], S. Redant[a], A. Marchioro[b]

[a]*imec, 3001 Leuven, Belgium*
[b]CERN, 1211 Geneva 23, Switzerland

brouns@imec.be, Laurent.Berti.ext@imec.be, verhaegs@imec.be, thys@imec.be, redant@imec.be
Sandro.Bonacini@cern.ch, Kostas.Kloukinas@cern.ch, Alessandro.Marchioro@cern.ch mailto:amicsa2014@cern.ch

Abstract

With the upcoming upgrades of the LHC experiments, it will be necessary to improve the performance and reduce the power consumption of the detector readout electronics. CERN has chosen to use a 65nm technology for part of the new generation of ASICs targeted to these upgrades. For this technology the SRAM memories within the readout circuitries need special attention as the commercially available IP blocks don't give the necessary radiation tolerance.

This paper will describe the design of a SRAM compiler design platform with a custom SRAM design underneath. The generated SRAMs have clock synchronous write/read operations and pseudo dual-port addressing.

They are implemented in the LP (Low Power) version of the technology and are designed to be radiation tolerant to reduce excessive power leakage due to TID (Total Ionizing Dose) and to minimize the impact of SEE (Single Event Effects) in the memory address decoding circuitry. The generated SRAMs can handle a TID >200Mrad and Linear Energy Transfer (LET) of 15 MeV.cm^2/mg. The max. frequency is at least 80MHz. This is also verified post-layout in all PVT corners. An additional challenge for these SRAMs is to keep the dynamic power consumption to a minimum whilst maintaining the radiation tolerance.

I. INTRODUCTION

The Large Hadron Collider (LHC) upgrade forces higher requirements to the electronics of many sub-detectors to be installed in the experimental chambers. With respect to the previous generation of front-end systems, the upgrade will need lower power consumption, smaller mass, volume and faster data channels, in addition to the robustness to a harsher radiation environment. A step towards more modern integrated circuit CMOS technologies was necessary to fulfil these requirements and 65nm is currently the chosen node to be used in the design of near-future particle detectors for LHC applications.

SRAM memories are used extensively in front-end chips in particle physics instrumentations. These memories are used for several functions, the main one being the temporary buffering of data waiting to be shipped off-detector after a triggering event. This trigger system allows to reduce greatly the quantity of data to be transmitted from inside to outside the experimental chamber, saving power and volume occupied by cabling resources.

This paper describes the design of an SRAM compiler, a software capable of generating SRAM blocks of different sizes starting from a custom design in the chosen technology.

As the target application of these memories will be in an intense radiation field, the design is such to be robust to both Total Ionizing Dose (TID) effects and Single-Event Effects (SEE). The techniques used to assure this radiation robustness are discussed in this article.

II. SPECIFICATIONS

A. General specifications

The generated SRAMs are designed in a 65mn technology using only standard-Vt devices, and occupy only the 4 lowest levels in the metal stack so that the upper layers can be used for routing or for the power grid.

The generated SRAMs can do simultaneous read and write operations, even though in the primary application will do more frequently write operations.

Description	Value	Unit
Supply	1.2 ±10%	V
Frequency	> 80	MHz
TID hardening	>200	Mrad
LET threshold	>15	MeV.cm^2/mg

Table 1: General specifications

B. Timing diagram

The SRAM has 3 possible working states within the same clock cycle : either a read, a write or both read and write operations. In the latter, first a read operation will be executed and then a write.

At the rising edge of the clock, the read & write address and the data to be written in the memory need to be available at the input terminals. See figure 1.

After the rising edge, the data read from the memory will be available at the output port.

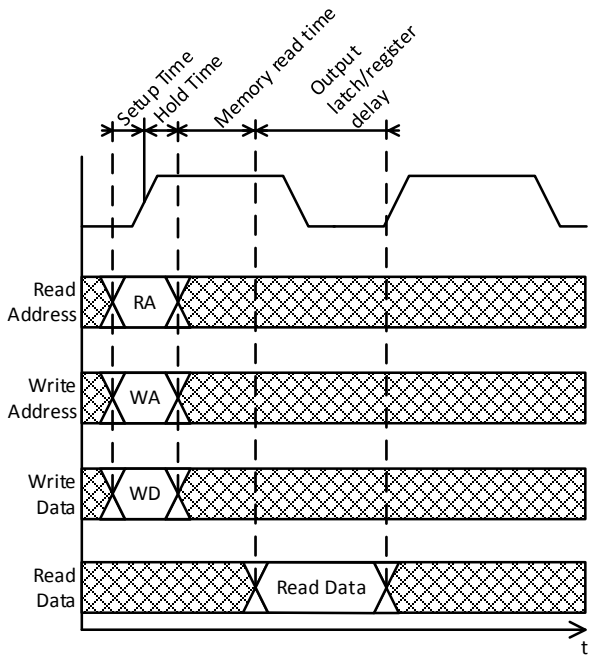

Figure 1: Timing diagram

III. SRAM ARCHITECTURE

A. Radiation hardening

The applied strategy for making the generated SRAMs radiation hardened attacks the issue on 2 levels, TID and SEE. The hardening for SEE of the addressing circuit is done by drive strength. This prevents reading or writing the wrong word. Protection against latch-up is foreseen by placing p+ guard bands between all n- regions. It has been chosen to not protect the data within the memory cells in a hard way: an error-correction code can be used for such purpose as an additional external layer. In order to minimize TID effects (drive loss, Vt shift) no minimal width transistor are used in the design but a larger width was chosen for the nMOS and pMOS transistors based on TID measurement data [1]: all nMOS are larger than 200nm and all pMOS are larger than 500nm.

B. Architecture

The SRAM is composed out of a set of memory blocks either in series (increasing word size), in parallel (increasing # words) or a combination of both. A bank of flip-flops is foreseen to store the target address of the word. A self-timing clock generator will generate all the clock signals at the correct time. 2 sets of decoders will convert the address to select the correct word. A buffer to insure signal integrity. Figure 3 illustrates the architecture more clearly. To achieve the required pseudo dual-port behaviour, there are 2 banks of flip-flops that will clock in the write address in one bank, and read address in another. The address is decoded in 3 pieces, the 3 LSB will be used for selecting the interleaved word (more info in subsection E. Periphery), the MSB (1 or 2 bits) will be used to select the memory block row, and the intermediate bits will select a specific memory block.

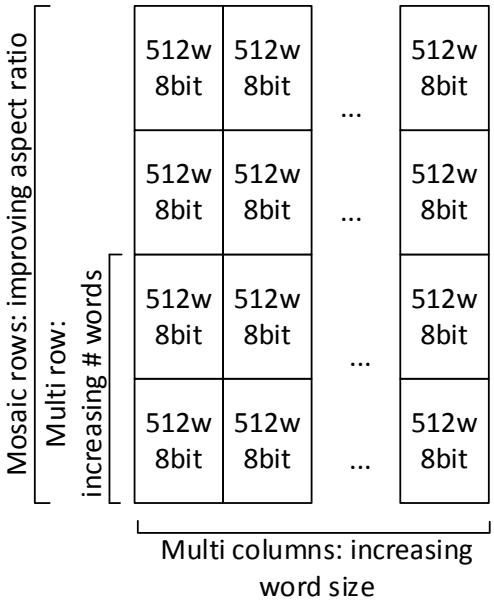

Figure 2: Scalability of the SRAMs

To improve the aspect ratio in case of big word sizes, the SRAM can be cut into X pieces and stacked on top of each other. To prevent confusion with multiple memory rows for increasing # words, these stacked rows are called mosaic rows. To calculate how many mosaic rows are needed to get the aspect ratio below a specific amount; the formula below is used to define numbers of rows needed to get a targeted aspect ratio:

$$\#mosaic\ rows = \sqrt{\frac{current\ aspect\ ratio}{target\ aspect\ ratio}}$$

C. The Memory cell

For the memory cell the classic 6-transistor design was used. The schematic was changed to fulfil the requirements of the minimal transistors widths for TID mitigation. The design was verified across process/voltage/ temperature (PVT) variations to confirm that the specifications for speed and stability are met. The static noise margin is larger than 18% of the supply voltage over the full PVT range.

D. The DICE flip-flop

The DICE flip-flop [3] is the cornerstone of the SEU protection strategy for the periphery and addresses storage. They have been hardened against SEU thanks to the well-known DICE latch structure (see figure 4). Internal SETs in the flip-flops are mitigated thanks to the redundancy of the DICE memory and the addition of cCell. In the DICE memory there are 2 times 2 identical outputs (see figure 4: A=C & B=D). This redundancy may be used to filter internal SET in the flip-flop by using A and C or B and D as input of the cCell. In case of SET, only one of the inputs is affected/toggled, it means that the cCell is in high impedance (i.e. the SET is not propagated).

For the hardening of the flip-flops used to store the write and read address, the output cCell has been sized such that the

cCell itself is insensitive to SET (hardening by drive strength).

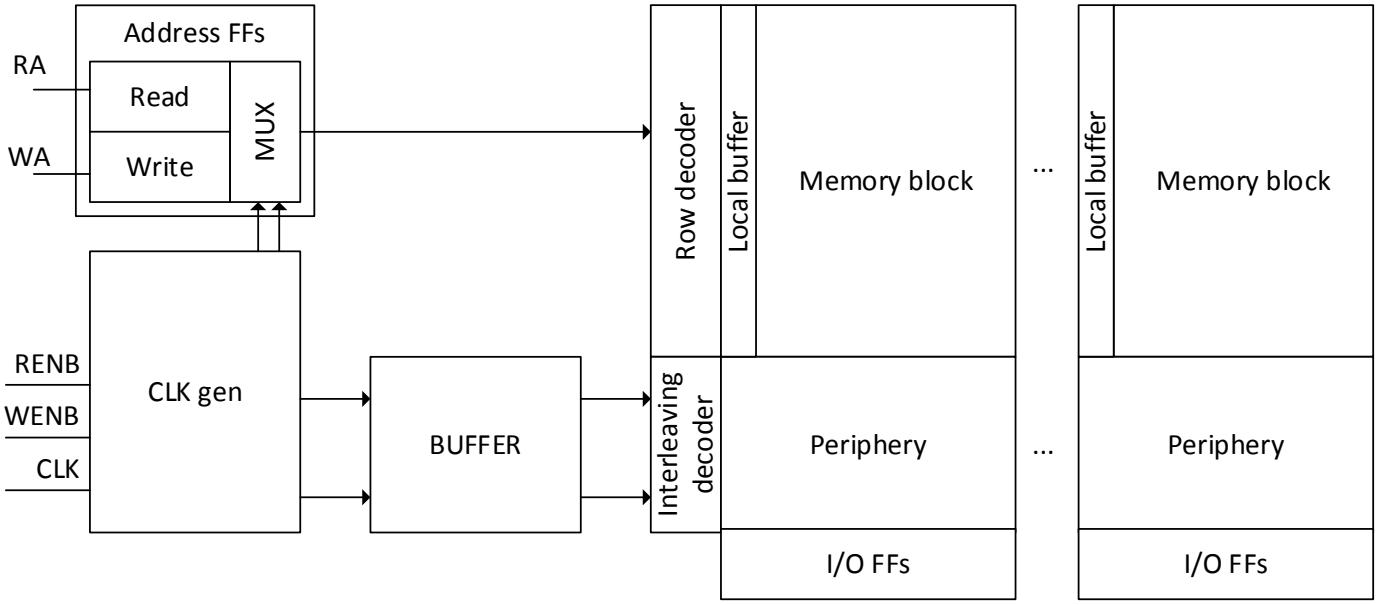

Figure 3: Block diagram of the architecture

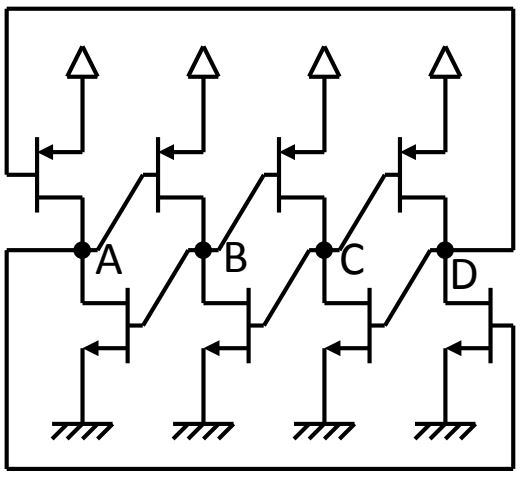

Figure 4: The Dice latch structure

E. Periphery

The memory cells in the memory block are interleaved by 8 words. This has a dual benefit. First of all it improves the aspect ratio to have 8 words on 1 row. Second of all it provides extra radiation hardening. When a charged particle hits the memory block under an acute angle, the particle can hit 2 neighbouring cells and alter both contents. A multi bit upset (MBU) would occur if these 2 bits belong to the same word.

To correct this, multiple parity bits would need to be added to the word. Alternatively by interleaving (figure 5), the particle will not hit 2 bits in the same word, causing less corruption within the data word. To decode this interleaving during write and/or read operations MUXs have been foreseen between the BIT-lines and the sense amplifier.

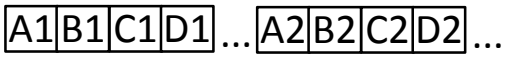

Figure 5: Interleaving

F. Memory block

Memory blocks have been made in 2 flavours, 128 words by 8 bits and 512 words by 8 bits, respectively 16 word lines and 64 word lines high. As mentioned before 8 words are interleaved per word line.

The word line itself is split up into multiple sections with one backbone, a global word line that is buffered inside every memory block. The advantage is that the local buffers inside the memory blocks can be better optimized for the load than one big central buffer.

All memory blocks have multiplexing gates on the outputs, so that multiple rows can be used to increase the number of words without shorting any outputs directly.

G. Clock generation

The timing of all signals is generated with continuous calibration on the background. This self-timing is achieved with a dummy memory block at the end of the memory. The clock generator will control the charge and discharge cycle of the BIT lines dummy block.

The Clock generator will also select the needed address from the flip-flop banks, either the read address or the write address.

The clock generator is build around 2 flip-flop. One controls the read and write phase, the other one read discharge duration.

The radiation hardening for the flip-flops is done by using DICE flip-flops. Other nodes in the clock generator are not hardened by drive strength, but the cross-section is low enough for the application.

IV. LAYOUT

The layouts are built upon 2 floorplans, one for 128w memory blocks and one for 512w memory blocks.

All block are optimized for compiler assembly and have been simulated post layout in all PVT corners.

The floorplan is always built up with the clock generator, address flip-flop banks, buffers and decoders located on the left side. Data flip-flops are located on the bottom of the flip-flop. The dummy memory is located the furthest form the clock generator to be sure that all parasitics are taken into account in the self-timing. Dimensions for a 1024w x 32b is approx. 450um x 375um.

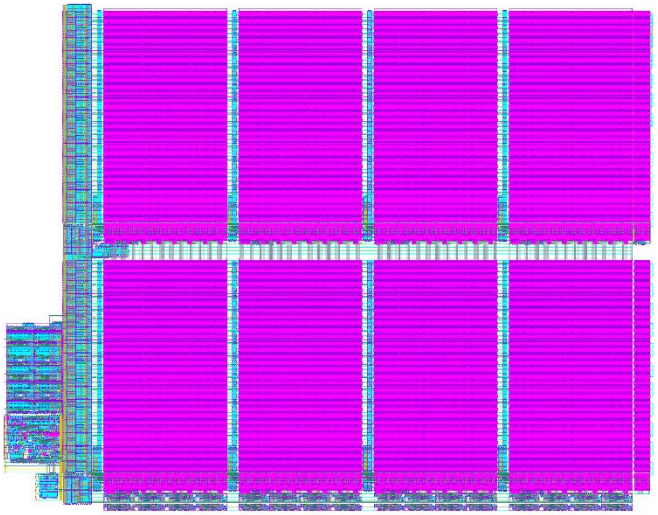

Figure 6: Layout of 1024word x 32bit SRAM

V. COMPILER

The compiler is build up out of 2 layers, a virtual design layer which is abstraction of SRAM design and an implementation layer that will generate the actual SRAM design database.

The compiler works in 2 stages. In the first stage the user can try out different possible designs and immediately see key parameters of that design e.g. physical dimensions, buffer configuration, etc. The name of this layer is “Virtual Design Layer”. In the second stage the Implementation Layer, will create the requested SRAM design.

The communication between the 2 layers happens via a TCP connection. This has the advantage that the implementation layer can run on a powerful server, while virtual design layer runs on a workstation.

Characterisation is done by interpolation between multiple pre-characterized instances. This has the major advantage that no expensive tools are needed for creating the liberty files at the user site.

The output from the compiler is an OA library containing the schematics, layouts, abstracts and symbol views. Additionally it will generate the needed liberty file, lef file and verilog model.

VI. FURTHER WORK

The first compiler version will be finished by the mid of Q3 2014. Two SRAM instances will be put on a test chip in July 2014 to prove the functionality in silicon.

VII. SUMMARY

In this paper, the development of a 65nm SRAM compiler was described. The SRAMs have clock-synchronous write/read operations and pseudo dual-port addressing, and are using the LP version of the technology with only standard Vt devices.

Synchronous write/read operations are controlled by the clock generation module. This module uses a dummy memory block with continuous calibration in the background.

The SRAMs are designed to be radiation hardened against cumulative and transient effect (TID and SEE). This is done by restricting the minimum dimensions of the devices, using DICE flip-flops and drawing p+ guard bands between the n-regions

The SRAM memories exist of combinations of memory blocks of either 128w x8b or 512w x8b. The word lines of these memory blocks are buffered locally to reduce power consumption. Additionally 8 words are interleaved per word line to mitigate MBU.

The compiler has the ability to first build a virtual design of the SRAM, so that the engineer, project leader or manager can use the tool to get an idea of the physical dimensions and the buffer sizing. When the virtual design satisfies the user, it can be fully generated and characterized.

VIII. REFERENCES

[1] S. Bonacini, P. Valerio et al. (2012, January). . “Characterization of a commercial 65nm CMOS technology for SLHC applications”. JINST, vol. 7, P01015

[2] Stefan Cosemans. (2010, Octobre). Variability-aware design of low power SRAM memories [PhD desertation]

[3] T.Calin, M. Nicolaidis, R. Velazco. (1996, December). Upset Hardened Memory Design for Submicron CMOS Technology. IEEE Transactions on Nuclear Science. Vol 43 (6), pp. 2874-2878