POLITECNICO
DI TORINO

MASTER'S THESIS

Master of Science in
NANOTECHNOLOGIES FOR ICTs

Study of low power and radiation hard Design For Testability solutions for High Energy Physics applications

Supervisors
Dr. K. KLOUKINAS
Prof. A. KOUKAB
Prof. G. MASERA
Prof. C. RICCIARDI

Author
Gianmario BERGAMIN

Advisor
Dr. D. CERESA

Academic Year 2018/19

Acknowledgements

I would like to thank the CERN microelectronics section for the opportunity given for my master project, and in particular the Outer Tracker ASIC design team (Kostas, Alessandro and Simone) for the help. A special thank to Davide, an exceptional supervisor who introduced me to the art of IC design with precious advices, and supported always the innovative idea of scanchain insertion.

The master thesis represents the end of a long route, and I would like to express my gratitude to all the people who supported me during these years. Thanks to the "We don't care about this" folks for the semesters in Grenoble and Lausanne, where we faced many challenges in a pleasant atmosphere. A particular thank to Gianmarco, for his friendly help and suggestions. Thanks to the "MIP" group for the amazing time spent together, which makes me feel like I had never left. Thanks also to all the other people taking part to my life, to the alpine adventures and to all the enjoyable moments which allow me to rest from the engineering world.

And first, thanks to my family who strongly supported me along all this time. Thanks to my grandparent and brother who taught me from my early age how to ponder each problem to do my best, and thanks to my parents, who allowed me to choose always the best. Without them, I would have never reached this goal. An extra thank to Federica, a patient support point for long time.

Abstract

In Particle Physics, electronic systems are widely exploited to track and record collisions results: dedicated ASICs are used to read and analyse silicon sensors. The High Luminosity LHC phase requires the development of novel detectors for CMS Outer Tracker, able to perform real-time computations and to stand harsher conditions (-40°C, 100 Mrad of ionizing dose). The development stage of the three Front-End ASICs (MPA, SSA and CIC) is finalized and a successful verification phase will imply the large scale production. The limited manufacturing yield requires testing strategies to identify defective units, guaranteeing the correct functionality of the tracker modules. The aim of this thesis is to study and develop innovative procedures to replace the currently used functional tests, which present limited performances in terms of testing time, completeness and costs. The concept of structural test is introduced for the first time in the field of High Energy Physics and Design For Testability (DFT) techniques are studied for a novel testing approach. The latter include circuit configurations and netlist modifications to ease testing procedures, and must be developed at the same time of the development of the chip. Two solutions have been studied and implemented, a Built-In Self-Test for memory blocks (mainly SRAMs) applying a March algorithm and the Scan Design, which allows to exploit the powerful structural algorithms. Considering the stringent constraints for the ASICs, particular attention was paid to radiation tolerance and power overhead, and studies have been carried out to assess the best trade-off between fault coverage and negative impacts on area and timing.

Contents

List of Figures

1 Introduction

1.1 CERN, LHC and CMS upgrade

Particle physics investigates the basic constituents of matters through the interaction of particles at high energy: beams of charged particles are accelerated and products of collisions are observed through detectors capable of identifying, counting and tracing the emitted particles moving outwards from the collision point. The European Organization for Nuclear Research (CERN), created in 1954 and counting nowadays 22 member states, provides one of the most advanced scientific facility for High Energy Physics (HEP) experiments. In the 27km long LHC (Large Hadron Collider) two counter-rotating protons beams are accelerated to 7 TeV and brought into collision with a 40 MHz frequency (named bunch-crossing rate, BX) in four different crossing points, called experiments: ATLAS, CMS, LHCb and ALICE. The discovery of the Higgs Boson in 2012 declared by both CMS and ATLAS represents the most celebrated achievement of the first runs of LHC. An exhaustive presentation of CERN and LHC can be found in [1]. Since many other phenomena still need to be investigated (for example the neutrino oscillations, the matter-antimatter asymmetry and the super-symmetry extensions to the Standard Model), a significant upgrade (named High Luminosity) of LHC is foreseen within 2024. The instantaneous luminosity, defined as the number of interactions per second, proportional to the number of particles in each colliding bunch, and the revolution frequency are expected to quintuple. It will enlarge the data collected by the experiments by one order of magnitude, reducing the measure uncertainty and providing extensive studies to solve open questions in the field of HEP. Consequently, also the experiments will face important upgrades to

cope with the new requirements.

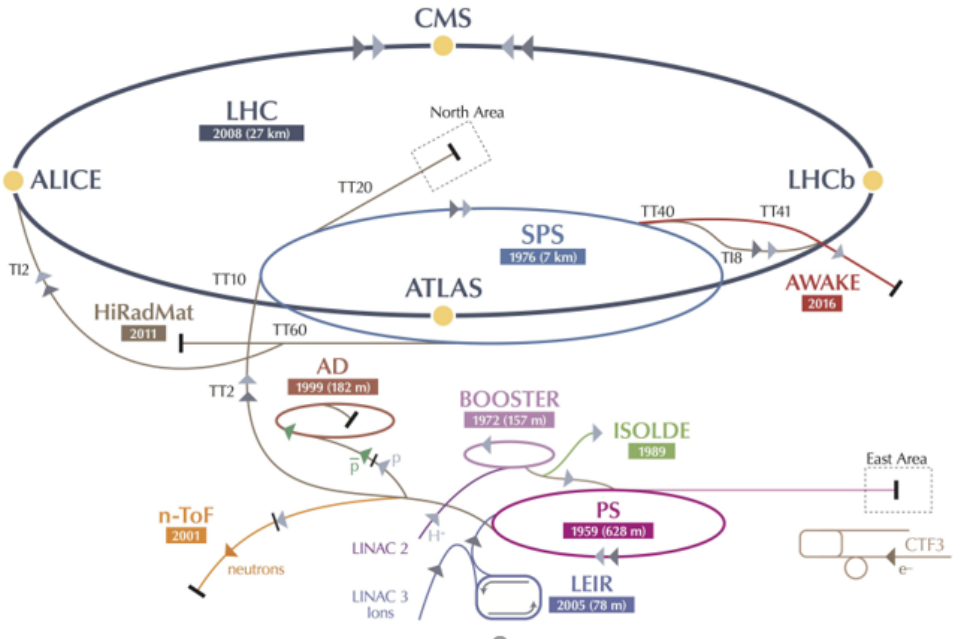


Figure 1.1. LHC accelerators and experiments

The Compact Muon Solenoid (CMS) is one of the four experiments located along the LHC orbit, it's a general purpose detector and acts as an high speed camera recording the stable particles produced during each collision. Measuring their momenta, direction and energy and assembling all the former information it's able to recreate the image of the collision. Figure 1.2 reports its characteristics. Composed by different concentric layers of detectors, CMS determines the particles properties performing the following operations:

- Bending particles: a strong solenoid magnet produces a 4T magnetic field to bend charged particles moving outwards from the collision point. It allows to identify their charge and momentum.
- Tracking the direction: CMS identifies the paths taken by charged particles deviated by the magnetic field thanks to concentric layers of silicon sensors composing the silicon Tracker. Interaction between silicon and charged particles leads to current spikes (hits), which are correlated to reproduce the full trajectory.
- Measuring energy: two calorimeters are present to filter emitted particles depending on their energy. The ECAL measures electrons and photons energy and

stop them completely while the HCAL measures the hadrons energies.

- Muons detection: since muons are not stopped by the calorimeters, they can be directly detected by the muon chamber present in the external layers.

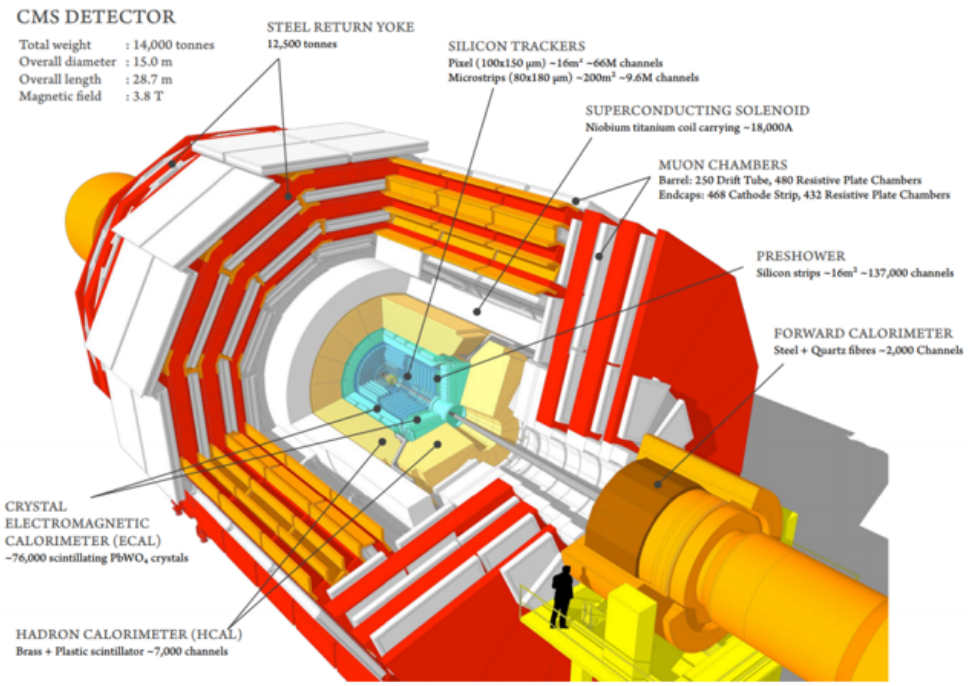


Figure 1.2. CMS structure

The inner part of CMS detector is the silicon Tracker, which is able to evaluate the position and the path of a traversing particle through a series of silicon layers exploiting reversely biased p-n junctions or MOS structures. The crossing coordinates of the different layers are sent to the back-end units which compute the event reconstruction and the particle momenta (higher the momentum, lower the bending under an applied magnetic field). Ten concentric cylindrical layers centered around the interaction point (see Figure 1.3) and 12 end-cap disks for each side compose the installed tracker, whose silicon strips cover an area of $\sim 200\,m^2$. A triggered readout mode manages data recording and transmission to reduce data rates and storage: the front-end ASICs store in memory the event information (i.e. the hits read on the strips) and transfer them to the back-end only in case the event is of physics interest. The triggering signal is called Level-1 trigger, is generated entirely from the muon chambers and calorimeter information and in case it's not received within a certain time latency the stored event is discarded. The intrinsic limitations of the current

CMS tracker in the HL-LHC environment make necessary its replacement with new detectors characterized by higher radiation tolerance, trigger rates and computation capabilities. The increase in luminosity and particle density brought by the LHC upgrade requires the following features for the new CMS tracker:

- Higher granularity, pixel/strip density and resolution to the z-axis allowing a better precision.
- Reduced material budget to limit shielding to particles (the tracker must be transparent to crossing particles).
- Improved trigger rates and bandwidth to stand higher number of collisions, together with data analysis and compression procedures to make a pre-selection of interesting events directly on chip.

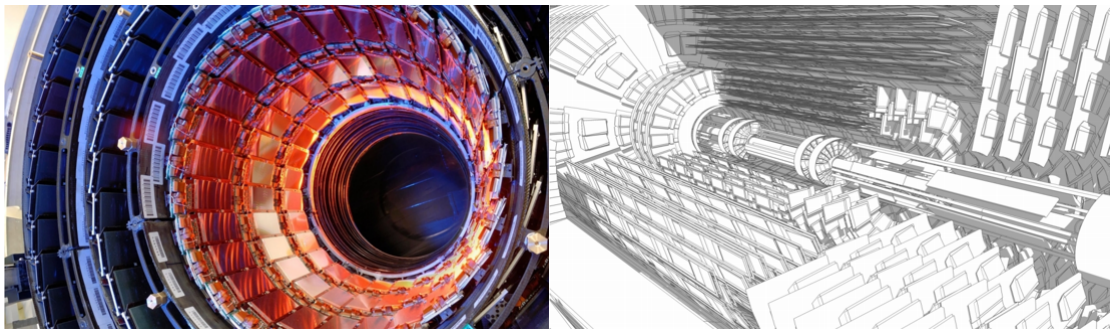


Figure 1.3. CMS tracker

Differently from the previous tracker which merely records the generated particles, the developing of an active sensor was necessary to evaluate in real time significant information, discarding already on chip the misleading combinations of random patterns. Providing the back-end with the most relevant hits for every event, the main feature of the upgraded tracker is the capability of participating in the Level-1 trigger signal generation, enabling the events reconstruction in the high luminosity context. The most effective way to discriminate useful data is to exploit the magnetic field to measure the momentum of the incoming particle, studying already locally its trajectory and transmitting direct information about high transverse momentum (P_T) particles, called stubs.

1.2 The Pixel-Strip (PS) module

As proposed in [2], the desired requirements can be achieved by measuring not only the position but also the direction of the impinging particle through two planar silicon sensors spaced apart, which are hence able to measure the particle curvature and perform data discrimination. The PS module, i.e. the detection unit implementing these concepts, presents a pixel and a strip silicon sensor to combine the high spatial resolution of the former with lower power density of the latter and thanks to the 4T magnetic field is able to successfully identify high P_T particles with a few mm spacing among the layers.

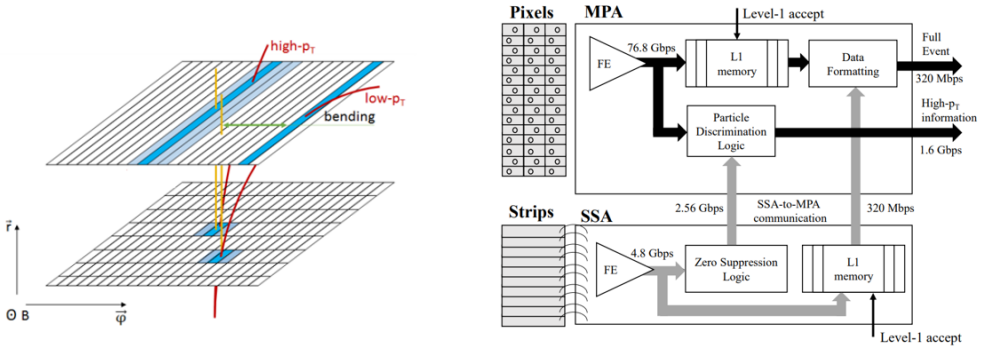


Figure 1.4. The PS module approach (left) and MPA-SSA communicaton (right)

Due to the mechanical setup, the mix of two types of sensors requires two different ASICs for their readout: the Macro Pixel ASIC (MPA) and the Short Strip ASIC (SSA). Designed in the same technology, the former reads the pixel matrix and computes real-time particle discrimination while the latter decodes the strip sensor hits and transfers the information to the MPA for processing. Since 16 MPAs and 16 SSAs are required in each module for the full sensor readout (presenting an active area of $\sim 5x10\,\mathrm{mm}^2$), two additional chips are needed for data aggregation and transmission: the CIC (Concentrator, one for each group of 8 Front-End chip pairs) and the LpGBT respectively. A Front-End (FE) block instantiated inside each ASIC digitalizes the deposited charge into an electrical pulse which is then sampled by the input stages of the chips with the 40 MHz bunch crossing clock. Figure 1.4 summarizes the communication between each SSA-MPA pair: SSA reads strip signals, stores the event (L1 data) into radiation resistant SRAMs and transmits Trigger-Data (the hits centroids) to the MPA. The latter interprets the pixel signals storing L1 data and process the

combined trigger data correlating the pixel and strip information to reject low P_T particle hits. In this way, the above mentioned system can provide both a continuous transmission of stubs for every event rejecting useless information and a triggered readout, transmitting the complete events under a trigger request (L1 level). Figure 1.5 represents a tridimensional view of the PS module.

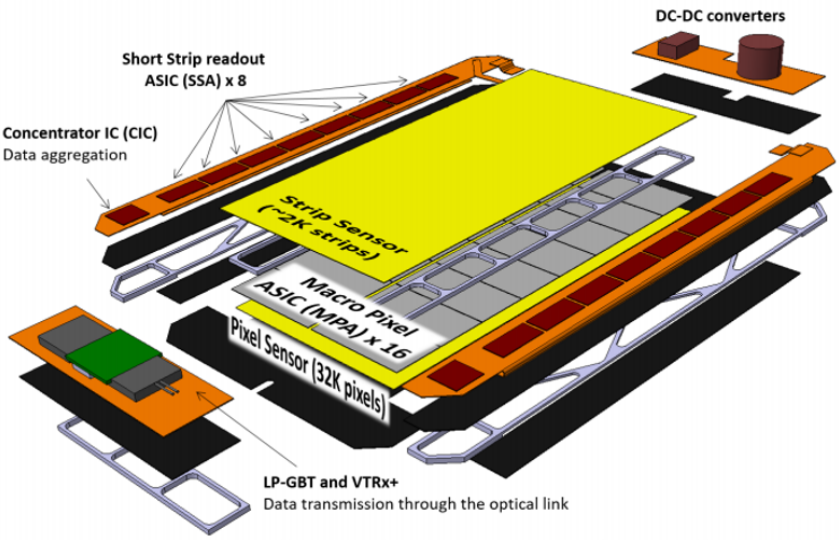


Figure 1.5. 3D structure of the PS module

Figure 1.6 shows one quarter of the tracker section with the modules dispositions: the blue lines represent the Pixel-Strip module while the red ones, called 2S modules, are units composed by two strips layer to limit power consumption where the high pixel granularity is not necessary. In the latter case MPA and SSA are replaced by a different ASIC, called CBC.

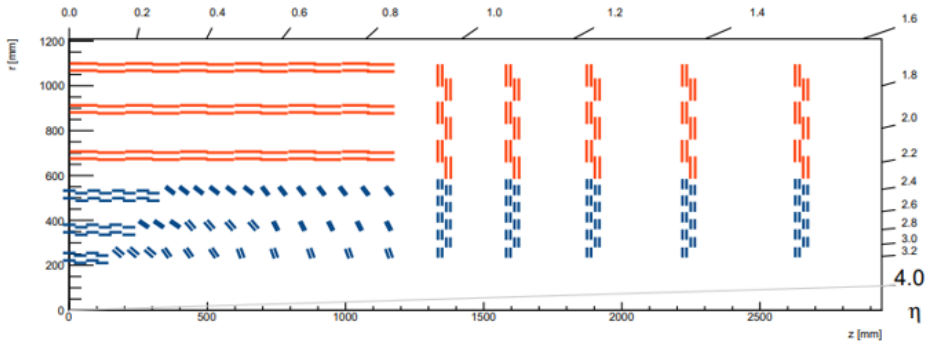


Figure 1.6. Modules disposition in the tracker

1.3 Radiation effects on CMOS electronics

The maximum instantaneous luminosity expected for the High Luminosity LHC is $5*10^{34}$ cm^{-1} s^{-1}, with a maximum total dose after 10 years of 57 Mrad [3], hundreds times more than spatial and satellite orbital applications [4]. Radiation effects, which comprise any alteration of the expected functions of the circuit due to the interaction of impinging particles, are for this reason one of the main challenges in the design of the front end ASICs.

The main source of the radiation effects are represented by beam losses and primary and secondary particles generated after collisions. In addition, the structures exposed to stray radiations become radioactive, emitting beta and gamma rays [5]. Consequences of radiations are mainly determined by the energy, mass and charge of incident particles. Charged particles such as electrons and protons interact with silicon through coulomb force, inducing atomic excitation until ionization. When decelerated, electrons can also produce X-ray radiations. Protons and neutrons, being massive particles, can also interact with the nuclei of the target material, leading to displacement damage effects. Photons have no mass neither charge and interact with silicon mainly by photoelectric effect, indirectly ionizing the material and generating an electron-hole pairs. Depending on the resulting impact on the device, radiation effects are divided in two categories, cumulative and non-cumulative. The former collects the consequences of accumulation over time of single particle interactions and includes Total Ionizing Dose and Displacement Damage. The latter category includes instead the functional failures (called Single Event Effects, SEE, and Single Event Upset, SEU) due to energy deposition of a single particle, usually localized in space and time.

1.3.1 Cumulative effects

Displacement damage consists in the dislocation of atoms from their position in the lattice of a crystal. It is mainly caused by hadrons (pions, neutrons, etc) and heavy ions but CMOS technology is almost insensitive to it.
On the other hand, due to the oxides present in the structures, MOSFETs are very sensitive to cumulative ionizing doses. First of all, radiation-induced charge trapped in the oxides gives raise to threshold voltage variations (lowering it in a n-channel

transistor and increasing its absolute value in case of p-channel transistor). TID produces also interface traps with energy in the energy gap of silicon, leading again to an increase (in absolute value) of the threshold voltage, which worsens the behavior of p-MOSFETs while compensates the one of n-MOSFETs. In addition, in advanced CMOS technologies the influence of lateral oxides used to insulate adjacent devices (Shallow Trench Isolation) and spacers became significant. These two oxides provoke RISCE (radiation induced short channel effect) and the RINCE (radiation induced narrow channel effect), making the effect of TID more evident in small size devices. Moreover, parasitic parallel channels are created in nMOS devices, leading to an increase of the leakage current. More information on radiation effects can be found in [6].

1.3.2 Non cumulative effects

As an energetic particle hits the die, the energy loss generates by ionization electron hole pairs which can be separated under an electric field: collected charges can reversibly modify logic values stored in the affected nodes giving origin to soft and non-destructive errors called Single Event Effects (SEEs). The parameter used to assess their impact it's called cross section, and corresponds to the number of functional failures generated divided by the flux of incoming particles. SEEs have different impact depending on the type of logic affected. In case of combinational logic (buffers, delay cells, logic gates, etc) they lead to Single Event Transients (SETs), namely current spikes generating glitches on affected signals whose duration equals the charge injection time. In general, SETs do not represent a problem if the glitch duration is not long enough to be sampled by a memory element. On the contrary, SETs on clocks or asynchronous set/reset paths lead to wrong duty cycles or data losses. SEEs striking sequential logic are named Single Event Upsets (SEUs) and happen whenever data stored in memory elements is overwritten following the ionization event.
Since the impact of particle hits on the circuit functionality cannot be predicted deterministically, the only way to protect the circuit is to store values in multiple bits, so that it is possible to recover and rewrite the information lost.

1.4 ASIC implementation and Rad-Hard technology

1.4.1 Technology and timing corners

The technology chosen to implement MPA, SSA and CIC is the 65nm. Among the ones present in CERN design kits and already characterized with radiation tests, namely 250nm, 130nm and 65nm, is the one which presents the best trade off. First, thermal and mechanical constraints limit the maximum power density for the whole tracker to 100 mW/cm^2 [7] (note that it refers to the average dissipation of the die where the cooler is joined) and a scaled technology reduces the average switching power (thanks to a reduction of the capacitances, P= $C * VDD^2 * f$). Since the gate thickness decreases with the technology, smaller nodes MOSFETs reduce also radiation induced charges trapped in the gate oxide (which cause a threshold voltage shift). SOI technology has been excluded due to the buried oxide which creates parasitic channels in NMOS transistors, increasing source to drain leakage current. On the other hand, going towards more scaled technologies like the 28nm one presents additional issues, especially for what concern SEU, since the critical charge needed to change the logic value is proportional to the node capacitance, which decreases with technology scaling. The radiation characterization has not been done yet but studies on the radiation effects in 65nm technology demonstrate short-channel transistors are more prone to TID damages with respect to longer ones [8].
Based on the 65nm characterization [9], the 7-tracks foundry standard cells present a huge degradation with impinging radiations already at 200Mrad, which is significantly reduced in the 9 and 12 tracks ones. Since the 12-tracks cells have larger minimum width, resulting in higher power consumption and area occupancy, the 9 tracks library was chosen in the design accepting a lower speed. Minimum width buffers and delay elements were removed during synthesis to reduce TID effects due to STI oxides: the former suffer of radiation induced narrow channel effects while the latter, adopting a non-minimum length, degrade differently from all the other standard cells under radiations. The main disadvantage related to adopting larger width buffers is an increase in power spent.
Since MPA and SSA will operate in the range between -40°C and 0°C, with possible different supply voltages (1.1 – 0.9 V) and under different levels of radiation, the timing analysis corners must be carefully defined for all these cases. The typical case represents a temperature of 25°C and a voltage supply of 1V. In the fast-fast corner

used for the hold check (for which the data has to arrive later than the capture clock), the temperature is set to -40° and the supply voltage to 1.1V: at large VGS, lower the temperature (i.e. higher mobility) and higher the power supply result in lower delay. In advanced technology nodes, and in the 65 nm one, when V_{GS} approaches V_T the variation of the current with respect to temperature is mainly given by threshold voltage shift (temperature inversion effect, [10]), which dominates over mobility and leads to smaller delay with higher temperature (higher temperature, lower the threshold voltage and faster the device). For this reason, the setup check (for which the data launched by one flip-flop must be stable before the capture edge) is instead performed in case of -40°C and 0.9V of power supply.

1.4.2 Single event effects hardening techniques

SEUs and SETs hardening techniques are based on redundancy: if data are stored in different circuit nodes (or several bits), it is possible to reconstruct the information even in case of SEUs. To reach a tradeoff between area and power overhead and SEE tolerance the maximum error rate acceptable for the chips has to be carefully addressed, considering separately the control logic and the data path. Any upset in the former brings in fact the system to work in an unknown state and only a reset signal can recover its functionality, making the ASIC not working for an extended period of time. On the contrary, single events on the data path (both stubs and L1 data) makes the module transmitting a wrong information, without affecting however the tracker operations or synchronicity. For this reason, and mainly due to power budget constraints (the area doesn't represent a real issue, being the final density of these chips determined by the large sensor area), hardening based on redundancy applies only to the control state machines and configuration registers. The approach used for all the ICs in the CMS outer tracker is the Triple Module Redundancy (TMR): registers and their logic are instantiated three times and the corresponding outputs are voted using majority voters: the resulting output is the one present with more frequency at its inputs. In Figure 1.7 the triplication of a state machine is reported as example. Concerning the control part, three different states machines clocked by three different clocks (A, B and C) and triplicated primary inputs are instantiated. Before feeding the following stage, outputs from the first column of FFs are voted, so that an upset on the register value is not propagated to the following logic. A feedback loop is instantiated as example in the second column of flip-flops to highlight the

importance of a periodically refresh of their values with the voted ones: in case two SEUs occur after a certain amount of time one from the other, without this technique the functionality would be lost. In this case instead, a possible error in one of the FFs of the second stage is corrected the following clock cycle. Voters are triplicated as well to avoid propagating wrong values due to SETs on them. On the other hand, the non-triplicated datapath requires single signals: a voted clock is distributed to those FFs and outputs of the control logic are voted to become single ended (in this case only a voter is employed). The triplication step is managed by a tool developed at CERN called TMRG [11]. To make triplication technique effective, the distance among triplicated instances must be sufficient to guarantee that an impinging ionizing particle doesn't affect more than one node at the same time; the minimum distance was set to 15 μm based on the measurements performed in [12].

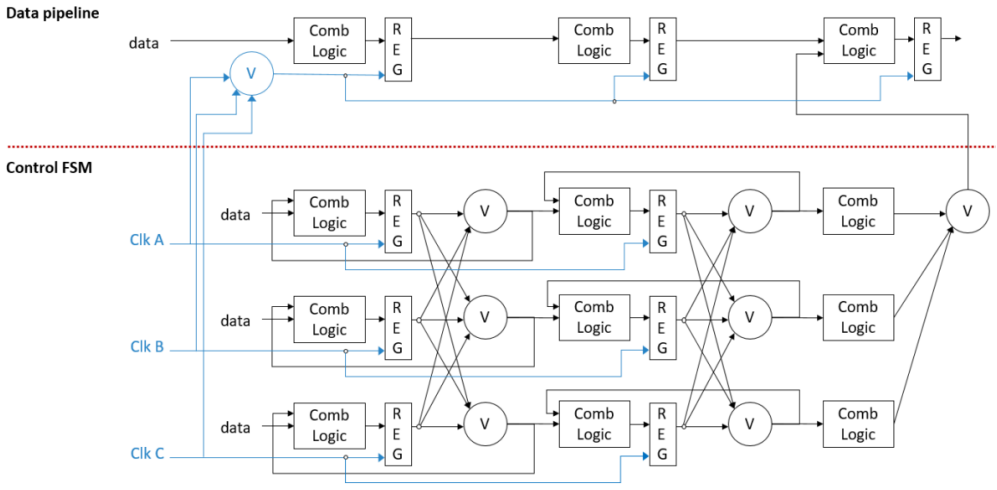


Figure 1.7. Example of Triple Module Redundancy implementation.

Dual Interlocked Cell (DICE) is another technique which prevents SEUs on sequential elements: for MPA and SSA it has been used only for FFs or latches inside the SRAM memories. Instead of storing the information in two nodes as in classical memory elements based on cross coupled inverters, the DICE latch presents a symmetric structure in which 4 nodes keep the value. The cell layout (Figure 1.8) allows data propagation from each node in two directions, different for each logic level (a 0 can only propagate rightwards while a 1 can only propagate leftwards) so that a certain logic value cannot propagate more than one stage. For example, a low logic value on node B can pull up node C without affecting node D and, on the contrary, a high level

on B can pull down node A without changing node D. The A, B, C and D nodes have two stable configurations (0101, 1010) and an SEU on one of the memory nodes can affect only one of the closest nodes. As soon as the upset extinguishes, the untouched nodes can restore the value of the whole cell. A DICE structures occupies double the area with respect to a normal latch and spends almost twice more power, but it is an effective solution for the SRAMs input stages.

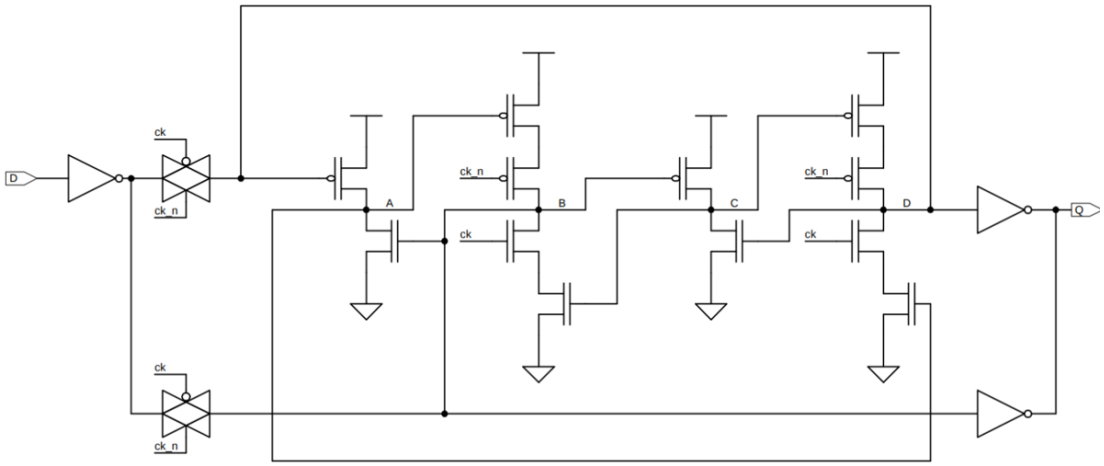


Figure 1.8. DICE flip-flop layout.

2 Functional and Structural tests

Each step reported in the simplified IC design flow of Figure 2.1 involves interaction of humans with humans, computer softwares and machines: misunderstandings and errors occurring in any of these steps, together with limited fabrication yield and vulnerability to process variation make necessary the setup of testing procedures. VLSI testing is usually performed in different stages to detect different types of failures: a chip can be defective both due to development errors (the circuit doesn't comply with the project specifications) or because of physical defects, in case the manufactured circuit doesn't match with the developed one.

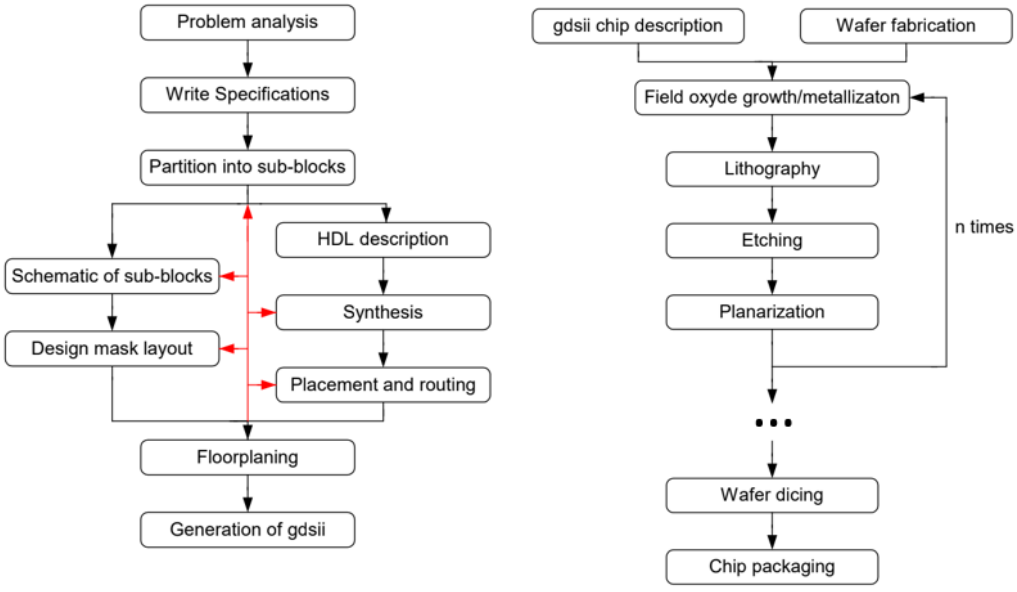


Figure 2.1. Simplified IC design flow.

In the first case, a functional test ensures the circuit will perform the given input to output function: this step, called verification, is performed both before and after prototype production, by verification tools and functional tests respectively. Successful results implies the large scale production. Manufactured ICs undergo to a final test to assess the integrity of all their parts. The functional test, which completely exercises the circuit behaviour, is replaced in this phase by a structural one, which verifies instead the physical design and the ability of each signal line to carry logic value 0 and 1. Differently from a functional procedure, which presents drawbacks in terms of testing time and fault coverage, structural test allows both to detect whether a device under test has any faults and to locate their exact positions, allowing further failure mode analysis. An explanation on structural testing techniques can be found in Section 2.2.
If the behavioral level is not used anymore in the structural test, it is still adopted to create functional tests for semiconductor memories. Thanks to their simple model, exhaustive functional tests are in fact still possible and normally used in practice, allowing checking memory attributes such as address uniqueness, decoding, cell crosstalk, column and row coupling and data retention. Further details on memory functional test are provided in Section 2.1.

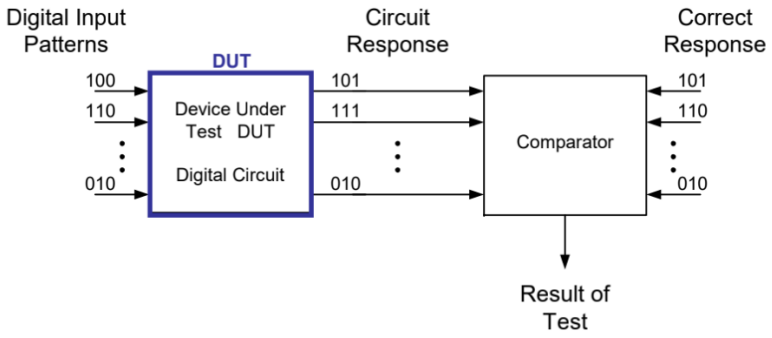


Figure 2.2. Principle of VLSI tests.

The basic principle of digital testing is represented in Figure 2.2: binary patterns, called test vectors, are applied as inputs and circuit responses are compared with the expected ones. ICs scales and complexity, together with the amount of test data and fault coverage (defined as the number of error detected over the total number present in the design), represent the main issues for nowadays testing procedures. Since most of the time the internal logic is unreachable only with the limited number

of input and output ports, testing strategies must include circuit configurations and netlist modifications (named Design For Testability techniques) allowing to control and observe internal nodes.

2.1 Memory test

2.1.1 Memory faults

In order to build a simple but rather complete fault model for the memory blocks in the design, needed to choose the correct testing algorithm, the reduced functional scheme of Figure 2.3 is considered. It comprises three main blocks: the memory cell array, the address decoder and the peripheral circuitry required to perform the operations. The main faults which are prone to occur within this model are reported in Figure 2.4.

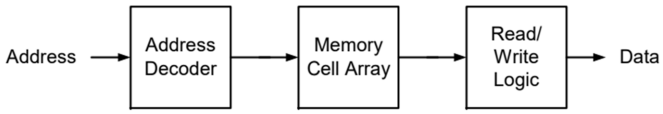


Figure 2.3. Simplified memory block

Among single-port defects, which are sensitized by one port access, simple faults do not interact with any other fault present in the array, whereas linked ones can influence each other's behavior and cause masking (i.e. the combination of two or more failures hide the faulty behavior, leading the test to fail). Simple faults (highlighted in figure by the dashed line) can be further classified into single-cell, two-cell, k-cell and address decoder faults. Depending on the number of operations required to activate them (just one or more), they are further classified into static and dynamic faults respectively.
Each error is characterized by a fault model, composed by a list of sensitizing memory operations (Figure 2.5) and the result of the faulty behavior. The notation employed to represent fault models is <S/F>, where S is the sensitizing operation and F the faulty cell response.

Single cell faults

This category includes faults occurring inside the same cell where the sensitizing operation is performed (i.e. the exact cell that is being read or written). The most

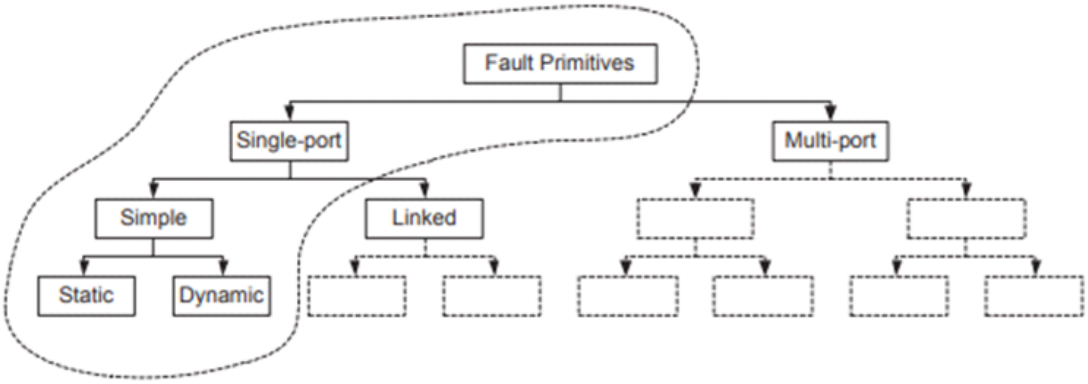


Figure 2.4. Memory faults [13]

general fault primitives, which represent many physical failure behaviors, are reported below.

- Stuck-at-fault (SAF), in which the logic value of the cell is permanently 0 (stuck-at-0, *sa0*) or 1 (stuck-at-1, *sa1*) . The condition to detect and locate them is to read a 0 and a 1 from each cell. The fault is described using the following notation: < ∀/0>, <∀/1>.
- Transition faults (TF) characterize cells that fail to make a transition when written. To be able to detect an up transition fault (<↑/0>) or a down transition fault (<↓/1>) the cell must undergo a switching operation and be read just after.
- Data retention fault (DRF) happens when the cell fails to maintain its state after some period without any access: the cell can unexpectedly flips from 1 to 0 ($<1_T/\downarrow>$) or viceversa after a delay time ($<0_T/\uparrow>$).

Two cells faults

The following fault primitives describe the influence the cell state (or an operation performed on it) has on the neighboring ones. The former, accessed during the sensitizing operation, is called aggressor (a-cell) while the latter victim (v-cell).

- A State coupling fault (CFst) rises whenever a v-cell is forced into a certain value by the state of an a-cell. Considering all possibilities, there are four CFsts: <0/0> <0/1> <1/0> <1/1> (the notation <A/B> means that A state in the aggressor will force B state in the victim).
- Inversion coupling fault (CFin) indicates that a transition in an a-cell inverts the

Notation	Action
r	read operation
w	write operation
r0	read a 0 from the memory location
r1	read a 1 from the memory location
w0	write a 0 to the memory location
w1	write a 1 to the memory location
↑	write a 1 to a cell containing a 0: the cell has a rising transition
↓	write a 0 to a cell containing a 1: the cell has a falling transition
↕	complement the cell contents
∀	any memory write operation
⇑	increase memory address
⇓	decrease memory address
⇳	addressing order can be either increasing or decreasing

Figure 2.5. List of sensitizing operations and their notation

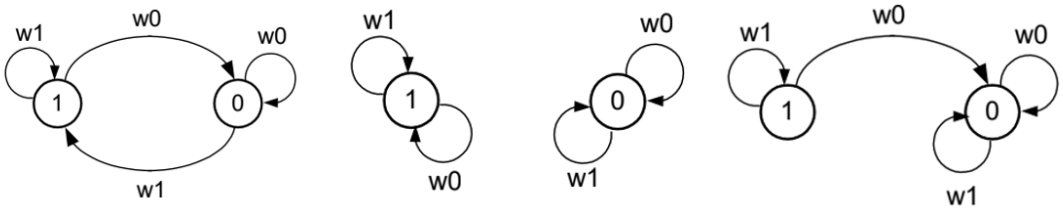


Figure 2.6. From left to right, state diagram of a good cell, *sa1* fault, *sa0* fault and <↑/0> fault.

content of a v-cell. The two possible types are denoted as <↑/⇳> and <↓/ ⇳ >. To detect them, each v-cell should be read after a write operation is performed in the aggressor, avoiding odd number of transitions to prevent fault masking.

- Idempotent coupling fault (CFid): an up or down transition in an aggressor cell forces the victim into a certain state, 0 or 1. The four possible CFids are annotated as <↑/↑>, <↑/↓>, <↓/↑>, <↓/↓>. For example, the notation <↑/↓> detects an up transition in the aggressor causing a down transition in the v-cell).

K-CELL faults: neighbourhood pattern sensitive fault (NPSF)

More generally, the content of a victim cell can be influenced by the content of its neighbouring aggressor cells (horizontally, vertically and diagonally), which may be

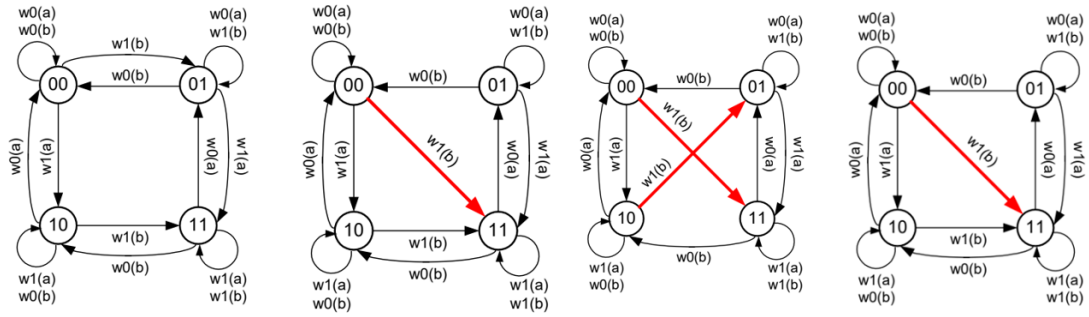


Figure 2.7. From left to right: state diagram of a good cell operation a and b, < 1;1 > CFst (victim a), <↑/⇕> CFin, <↑/↑> CFid.

patterns of states (0 and 1) or transitions. Within this group are categorized active NPSF, in which the base cell changes due to a change in the pattern of the neighborhood, passive NPSF in which the base cell cannot change when a specific pattern is present in the neighborhood and static NPSF, when the base cell is forced into a particular state when a specific pattern is present in the aggressor cells.

Address decoder faults (AF)

AFs are defects striking the address decoder and are assumed to be the same during both read and write cycles. Four possible cases are identified: no cell is accessed for a certain address, no address can access a certain cell, multiple cells are simultaneously accessed with same address and multiple addresses can access a certain cell.

2.1.2 March test algorithms

To test whether the memory is failing or not, typical memory operations must be applied to activate the identified wrong behaviours and the resulting outputs must be compared with expected responses. Before the fault primitives of the previous paragraphs were introduced, traditional tests consisted in simple writing and reading operations of a base and a victim cell respectively. Examples are the zero-one test (the memory is written with all 1s, read, written with 0s and finally read again), the check-board test (in which a chessboard pattern of 1 and 0 is applied and reversed in the memory matrix) and the GALPAT test, where a single bit is made walking through the array. Contrary to the first methods, which present unacceptable low coverage, the latter test exhibits good performances but results impracticable for nowadays array

dimensions.

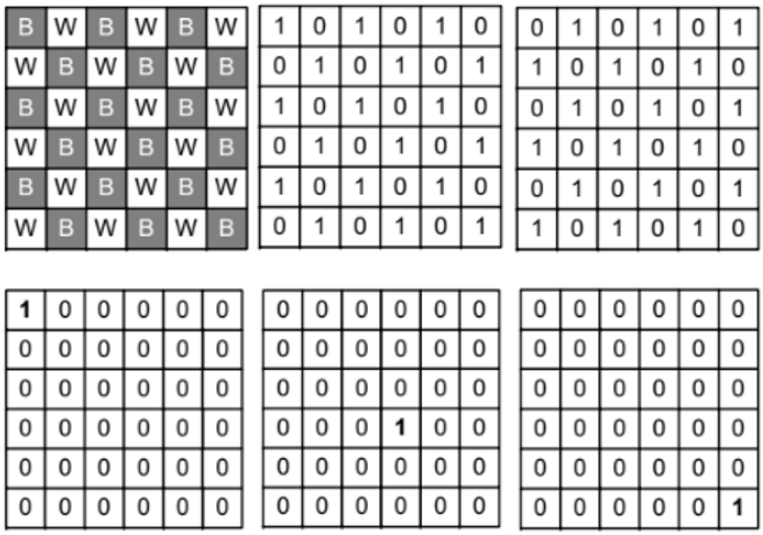


Figure 2.8. Check-board (top) and GALPAT (bottom) test patterns.

Currently used memory tests are called March algorithms and consist of a finite sequence of memory operations, named March elements, which are applied to each memory cell. The different operations are writing or reading 0 or 1 (w0, w1, r0 and r1 respectively) and are performed in a certain order ($\Uparrow$ monotonic increase of the address, $\Downarrow$ monotonic decrease of the address or $\Updownarrow$ irrelevant address order). The March test notation groups into curly brackets the March elements, which must be entirely applied to the cell before proceeding to the next one. The sequence is of crucial importance in order to detect dynamic faults in a single cell or coupling faults in the memory matrix. Figure 2.9 reports some of the March test currently used in IC market.

Several algorithms are able to detect different faults and involve different time and complexity. For example:

- March A is a complete test for linked idempotent coupling faults, it's the shortest test for AFs, SAFs, and TFs not linked. Its test time is 15.2^N, where N is the number of bits in the memory matrix.
- March B is obtained by the modification of the first elements in March A to detect AFs, SAFs and TFs linked with CFs. The time required for a March B test is in the order of 17.2^N.

Algorithm	Description
MATS	$\{ \Updownarrow (w0); \Updownarrow (r0, w1); \Updownarrow (r1) \}$
MATS+	$\{ \Updownarrow (w0); \Uparrow (r0, w1); \Downarrow (r1, w0) \}$
MATS++	$\{ \Updownarrow (w0); \Uparrow (r0, w1); \Downarrow (r1, w0, r0) \}$
MARCH X	$\{ \Updownarrow (w0); \Uparrow (r0, w1); \Downarrow (r1, w0); \Updownarrow (r0) \}$
MARCH C−	$\{ \Updownarrow (w0); \Uparrow (r0, w1); \Uparrow (r1, w0); \Downarrow (r0, w1); \Downarrow (r1, w0); \Updownarrow (r0) \}$
MARCH A	$\{ \Updownarrow (w0); \Uparrow (r0, w1, w0, w1); \Uparrow (r1, w0, w1); \Downarrow (r1, w0, w1, w0); \Downarrow (r0, w1, w0) \}$
MARCH Y	$\{ \Updownarrow (w0); \Uparrow (r0, w1, r1); \Downarrow (r1, w0, r0); \Updownarrow (r0) \}$
MARCH B	$\{ \Updownarrow (w0); \Uparrow (r0, w1, r1, w0, r0, w1); \Uparrow (r1, w0, w1); \Downarrow (r1, w0, w1, w0); \Downarrow (r0, w1, w0) \}$

Algorithm	Fault coverage								Operation count
	SAF	AF	TF	CF in	CF id	CF dyn	SCF	Linked faults	
MATS	All	Some							$4 \cdot n$
MATS+	All	All							$5 \cdot n$
MATS++	All	All	All						$6 \cdot n$
MARCH X	All	All	All	All					$6 \cdot n$
MARCH C−	All	All	All	All	All	All	All		$10 \cdot n$
MARCH A	All	All	All	All				All linked CFids, Some CFins linked with CFids	$15 \cdot n$
MARCH Y	All	All	All	All				All TFs linked with CFins	$8 \cdot n$
MARCH B	All	All	All	All				All linked CFids, All TFs linked with CFids or CFins, Some CFins linked with CFids	$17 \cdot n$

Figure 2.9. Memory test algorithms fault coverage and description.

- March X is designed to detect unlinked inversion coupling faults. Its test time is proportional to 6.2^N
- March Y detects TFs linked to CFins in addition to all faults detected by March X. This test requires 8.2^N time to be completed.
- March C-, which is obtained by removing a redundant element of March C test, detects all kinds of unlinked faults with an acceptable testing time (10.2^N). Considering the frequency of appearance, P.K Veenstra demonstrated in [14] its ability to reach higher fault coverage with respect to other common tests (Figure 2.10).

For the previous reasons, and thanks to the limited number of operations and complexity, it was decided to apply a March C- algorithm in the memory test, trying to

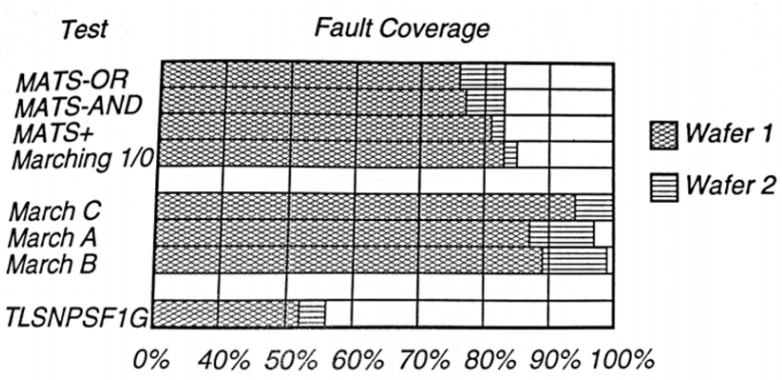


Figure 2.10. Fault coverage for different March tests, from [14]

implement a reconfigurable method which can be adapted to new algorithms. The sequence of operations to be performed is reported below. The description of its implementation is reported in Chapter 3.

- No matter the writing order, initialize all memory matrix to 0.
- Starting from the first bit of the first word read the previously written value (0) and subsequently, for the same cell, write 1. All stuck-at-0 faults and some coupling faults can be detected at this stage: if, for example, cell i+1 changes value (from 0 to 1) when cell i is written, the read-0 operation will fail.
- Read 1 and write 0 for all cells starting from the beginning. In addition to stuck-at-1 and transition faults (from zero to one), it recognizes the inversion and coupling faults (similarly to the previous step).
- Read 0 and write 1 starting from the last cell (similar to the second step, but in the opposite order).
- Read 1 and write 0 for every cell starting from the last one (to test coupling and inversion faults between cell i+1 and i).
- Read 0 in all the cells in whatever order.

2.2 Structural test

Differently from functional tests, structural checks do not focus anymore on the function implemented by the circuit but rather on its structure of gates and interconnects, assuming the design as functionally correct. In this way it's possible to generate universal algorithms able to detect errors regardless the circuit functionality. The introduction of fault models, which are representation of defects at the abstracted behavioural level, fills the gap between the real defects in the physical domain (which are too numerous and diverse to be controlled) and the manifestation in the digital behavior. The introduction of a fault model makes the analysis independent also of technology.

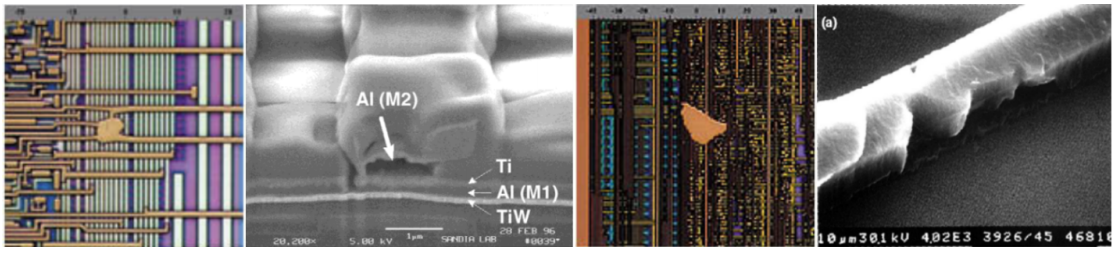


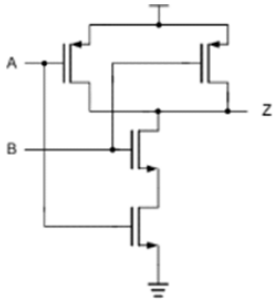
Figure 2.11. Examples of physical defects in ICs, from left to right: bridge formation, defective via, masking particle, electromigration notch.

2.2.1 Fault models

In the Register Transfer Level, where the whole circuit is modeled as a netlist of Boolean gates, stuck-at fault model is adopted to represent whatever issue occurs at the interconnects, reflecting defects affecting the connected gates. Each line is characterized by two kinds of stuck-at faults: stuck-at-0 (*sa0*), for which the line is blocked to the 0 value no matter the effort to drive it to 1 and stuck-at-1 (*sa1*), which maintains the logic state equal to 1 irrespective of the correct value of the gate driving it.

Figure 2.12 reports the fault model for a NAND gate. In the abstract level of the fault model, whatever physical defect affecting the cell (missing or bridged metals, contaminations by particles, oxide breakdown, over etching, mechanical damages, defective masks and photoresists, crystal imperfections and many other phenomena) is represented by six possible faults on the input and output lines (one *sa1* and one *sa0*

for each of them). For example, a short circuit leading the output Z to be grounded is represented as Z stuck-at 0: whatever combination is present on the input, the output will be always 0 (third column of the faulty response). Similarly, if the A line is not able to carry a high value (and for this reason is *sa0*), the pull down path is never activated while the pull up leads the output always to an high value (in this case the result is the same as if Z was stuck-at 1). Note that the presence of certain stuck-at fault does not hinder the output to be correct under certain input combinations.



Inputs		Fault-free response	Faulty response					
A	B		A_0	B_0	Z_0	A_1	B_1	Z_1
0	0	1	1	1	*0*	1	1	1
0	1	1	1	1	*0*	*0*	1	1
1	0	1	1	1	*0*	1	*0*	1
1	1	0	*1*	*1*	0	0	0	*1*

Figure 2.12. Fault model for a NAND gate. The faulty response is shown in red and italic.

Figure 2.13 represents a more complex netlist involving more gates and interconnects: the red spots indicate the possible fault sites, where both *sa1* and *sa0* must be addressed one by one in a structural test.
Therefore, fault detection requires first of all an input test vector that activates that precise defect, driving the corresponding line to the opposite value of the fault (0 in case of stuck-at-1 and one in case of stuck-at-0). In case the error is present, the given line will assume the faulty value (1 for a *sa1* and 0 for a *sa0*). Once the error has been activated, it must be transmitted to an observation point, usually a primary output or a memory element accessible from the external. For this purpose, in order not to hinder the propagation, all off-path gate inputs have to be set to non-controlling values. For example, if a stuck-at-0 has to be detected for line E in the circuit represented in the figure, it must be activated setting line B to 1, and propagated through the G-Z path. For this purpose, lines D and H must be set both to 1 (the non-controlling value for NAND and AND gates), to allow the *sa0* propagating to output Z. Input binary patterns allowing to discover faults are called test vectors. They must ensure the controllability and the observability of each fault, namely being able to drive a certain node to a certain value and allow the transmission of the faulty behaviour to an observable node for comparison.

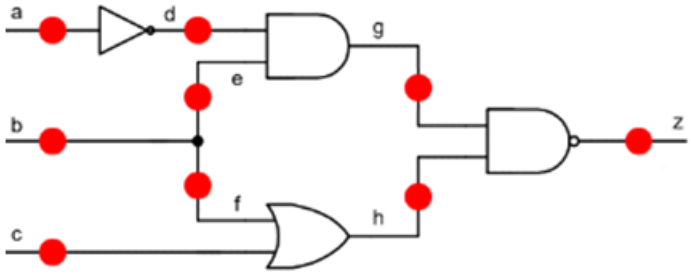


Figure 2.13. Fault sites in a digital circuit

The total number of fault sites in a Boolean circuit is given as the sum of primary inputs, gates and fanout branches and can grow to a significantly large value in modern circuits. Since test vectors need to be generated for each of them, reducing the number of faults in a circuit would be beneficial for reducing the number of test vectors necessary to test the whole circuit: the fault equivalence approach is used for this purpose. Fault equivalence states that two faults *f1* and *f2* are equivalent if all test vectors detecting *f1* also detect *f2*: being the two faulty functions indistinguishable, a single test vector is sufficient to detect both. For example, in case of a single NAND gate, each *sa0* in input leads the gate output to a constant logic high value, regardless the input test vector: for this reason, A *sa0*, B *sa0* and Z *sa1* are equivalent. Figure 2.14 reports the fault equivalence rules applied to Boolean gates and interconnects. The previous example demonstrates the difference between functional and structural tests: the former generate a full set of test patterns to completely check the circuit functionality, resulting in long and expensive procedures; structural test exercises instead only the minimal set of stuck-at faults after discarding equivalent ones.

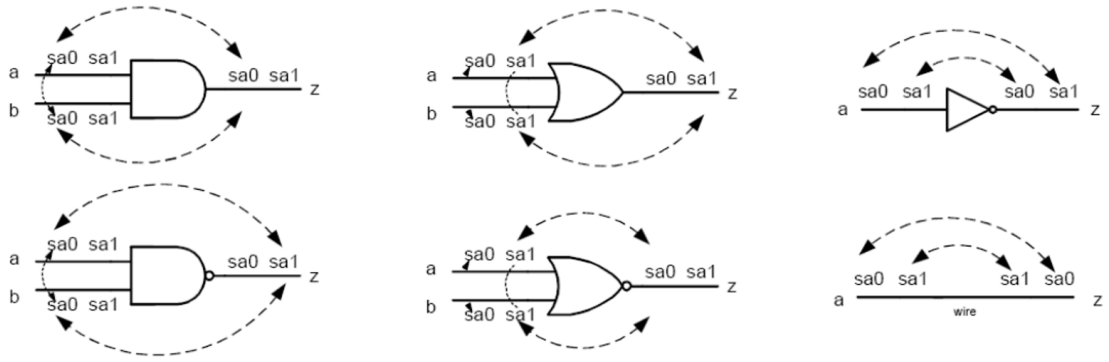


Figure 2.14. Fault equivalence method for Boolean gates. No fault equivalence is allowed for fanout branches and XOR/XNOR gates.

2.2.2 ATPG algorithms

The process of generating test vectors, along with the expected outputs for faults in the netlist is called Automatic Test-Pattern Generation (ATPG). ATPG injects a fault into the circuit and uses a variety of mathematical algorithms to activate it and to make it propagating through the logic, to be manifested at the circuit outputs.

2.2.3 Combinational ATPG

The very first ATPG algorithm for combinational circuits was developed by J. P. Roth which created an higher order algebra (called D-Algebra) able to represent both the good and faulty behaviour of each signal line, allowing one single ATPG calculus to determine both the values [15]. A detailed explanation on ATPG algorithms and rules can be found in [16]. Considering each fault one by one, the ATPG approach consists in three steps:

- Fault sensitization, where the stuck-at fault is activated forcing it to the opposite value of the fault value. Without any internal controllable point, the fault must be activated starting directly from a primary input.
- Fault propagation, in which the fault consequence is propagated through different paths to a primary output of the circuit. This step, called also forward trace, requires the inputs of the gates along the fault propagation path to be set to non-controlling values, in order not to block the fault effect. For instance, fault effects are propagated from an AND/NAND gate by setting the other inputs to 1, while from an OR/NOR gate by setting other inputs to 0.
- Line Justification, in which internal signals previously assigned to sensitize or propagate a given fault (step 1 and 2) are brought back (justified) to the primary inputs of the circuit.

Conflicts among signal assignments may arise among the three steps, and the algorithm has to backtrack, that is finding a new combination of compatible values which allow both to sensitize the fault and to propagate it. A fault for which no test patterns can be found after all possible backtrackings is called untestable, and limits the fault coverage of the test. Within this class, redundant faults are the ones that do not modify the input to output function, and for this reason cannot be tested.

2.2.4 Sequential ATPG

Combinational ATPG algorithm is not sufficient for nowadays circuits, presenting many flip-flops and latches: sequential ATPG has to be applied on top of it. It allows fault propagation even through memory elements, making possible to test faults affecting sequential logic. In this case flip-flops and latches are considered as ideal memory elements, without any faults neither on the clock path nor inside: their output and input faults are reported to the combinational logic they feed. ATPG for a sequential circuit consists in a sequence of vectors which initializes the circuit to a certain known state, activates the faults and propagates the effects to primary outputs and for this reason is much more time consuming. The sequential netlist is unrolled into a larger combinational one by the so called time-frame expansion method, where temporal copies of the combinational gates are linked with ideal memory elements (see Figure 2.15). Within this structure, the single stuck-at fault present in the real sequential circuit appears as multiple faults, existing in each unfolded time-frame. Combinational ATPG is then used to generate a test vector for the multiple stuck-at fault.

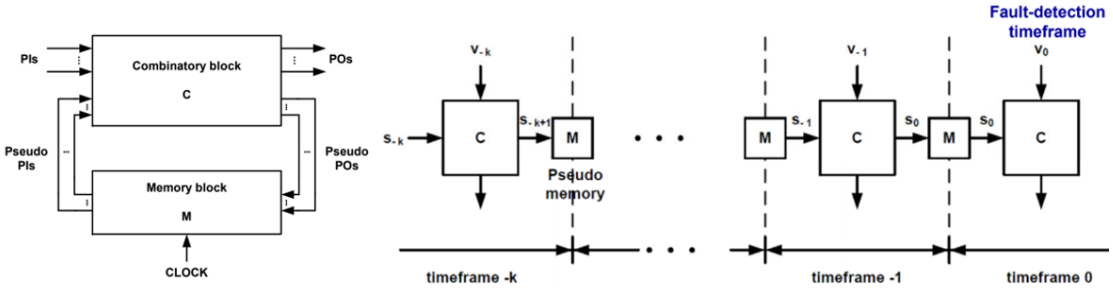


Figure 2.15. Time-frame expansion method

Any modern ASIC has a complex and deep logic which is hardly accessible with the limited number of input and output pin, making most of the time the ATPG algorithms to fail if merely applied to input and output pins. DFT techniques are then necessary to provide additional controllable and observable points, to ease fault sensitization, propagation and justification. An example of a structured DFT method is reported in Chapter 4.

3 Built-In Self-Test for SRAM testing

3.1 The SRAM block

The triggered readout implemented by SSA and MPA consists in temporarily buffering event hits in radiation tolerant memories and allows reducing the quantity of data sent to the back-end, saving power and cabling resources. Embedded storage elements located in the chips store the full sensor image for the time needed for the trigger computation: raw data are then transmitted only in case a L1 trigger is received within a certain latency time, otherwise they are discarded. The memory elements act as circular buffers, storing in consecutive lines events for consecutive bunch crossings at 40 MHz, up to 12.5 μs in total (which is for this reason the maximum time for the trigger computation). Depending on their activity, both static random access memories and latch based memory are employed for power consumption optimization: MPA presents sixteen SRAMs (one for each sensor row) while SSA features a latch memory for the L1 data packets and an additional SRAM to keep track of high ionizing particles (HIP flags, see [17]).

A memory gating technique has been implemented to reduce power dissipation in SRAMs, taking advantage of the low number of incident particles on the sensor area (hit occupancy) of the CMS outer tracker: the write operation is performed only if a hit is detected, otherwise the memory clock is gated. Since the SRAM lines are not overwritten or refreshed at every cycle, not updated values could be successively read from the memory: to avoid this error, a label is saved together with the data, to record the time they have been written. When the trigger signal is asserted, a control logic

reads out the memory lines and compare the label: data are sent in output only if the label corresponds to the expected one. Figure 3.1 displays the memory block together with the control circuit needed for its gating, tag generation and comparison. Its description can be found in [18].

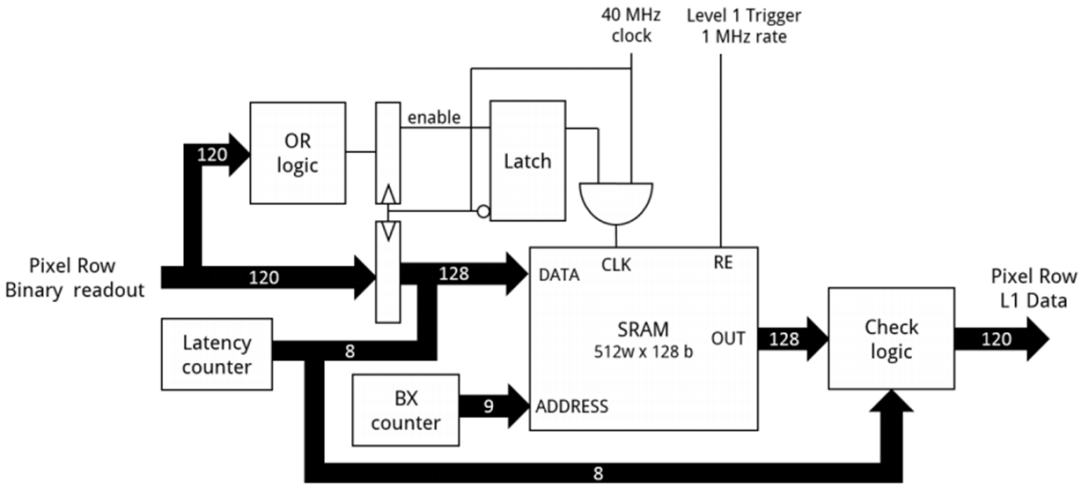


Figure 3.1. SRAM block with gating logic.

The SRAM block counts 512 lines per 128 bits, where 120 bits are occupied by the sensor image (which counts 120 pixels for each row) and the remaining 8 by the tag. A rad-hard SRAM compiler built in 65nm technology is provided by IMEC institute [19] and the cell is integrated in the flow as a macro block. The main requirements for the memory are the capability to stand radiation doses up to 200 Mrad and to perform a simultaneous read and write operation at 80 MHz, behaving like a pseudo dual port SRAM. For this purpose, a single clock is provided in input and an internal self-timing clock generator creates the clock signals needed for the two operations. Disregarding the internal architecture details, the SRAM timing diagram and functionality relevant for testing purposes are reported in Figure 3.2. Three are the possible operations within a single clock cycle: read, write or both read and write (in this case the read operation is the first to be performed). At the rising edge of the clock, read address, write address and write data have to be stable at inputs, and output data is provided after the read latency time.

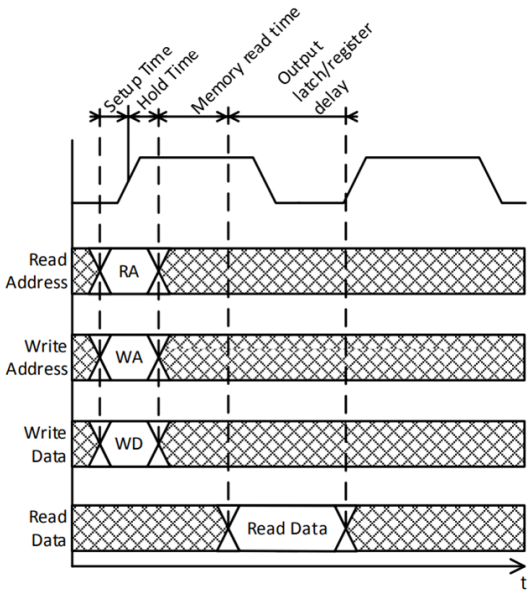


Figure 3.2. SRAM timing diagram, from [19].

During the preliminary tests of the first prototypes of MPA and SSA, some L1 data memories showed unexpected failures at low voltage, leading to a significant decrease in the chips yield. Atypically high input capacitance, direct connection of clock input to internal latches without rebuffering and an optimistic timing model revealed to be its leading causes. In view of the development of the second prototypes, the memory compiler was eventually corrected, introducing the needed buffers and balancing the clock tree.
The currently used test can be performed only stimulating the functional triggered readout paths and takes 12 seconds to perform a single write and read operation on one line. An ad-hoc testing strategy able to selectively target memory faults and to provide more accurate and precise characterization is needed for the next designs. The introduced method will allow also faster checks for the former failures during production testing.

3.2 BIST development

The embedded memory inside SSA and MPA chip does not allow its inputs controllability or outputs observability, making impossible to perform the chosen March C- test only through an external software. The difficulty and the cost of a scan approach

employing an external ATE made necessary the implementation of a Built-in Self-Test (BIST) design. BISTs are techniques and circuit configurations that enable a circuit to test itself and simply consist in the addition of extra circuitry to ease the testing procedures. In other words, the March algorithm is fully implemented in hardware inside the IC boundaries. Many are the benefits of this technique:

- No need of an external test equipment, resulting in a cheaper test (and maintenance is not needed).
- Tests are performed at chip speed, which is increasingly difficult using ATEs and results in a lower testing time.
- Contrary to ATEs technology, BIST hardware follows VLSI technology improvements.
- Low development cost can be further reduced reusing it for different chips.

In addition to developing time, the main drawbacks come from the additional hardware placed inside the module, which conducts to yield and reliability decrease. In case of MPA and SSA SRAMs, particular attention must be paid to timing issues, power overhead and radiation impact. For these reasons, the BIST structures was thought to be as simple as possible.

3.2.1 Implementation

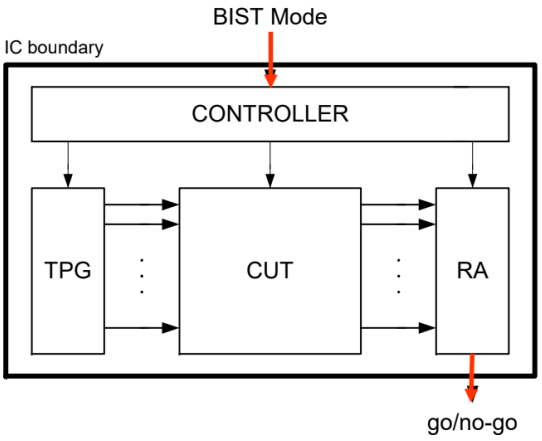


Figure 3.3. General BIST structure

Figure 3.4 reports the structures generally needed in a BIST: a test pattern generator TPG (i.e. the inputs to be sent to the circuit under test to sensitize the faults), an output

response analyzer RA, which is in charge of comparing the responses to assess the good or faulty behavior and a controller, to manage the operations. All this additional hardware has to interact with the memory during the testing time, in which the CUT stops the normal activity, and be transparent during the functional operation of the chip: input multiplexing is needed for this purpose. This last requirement has to be carefully addressed in ICs working with high radiation levels: a SET on the control signal of the multiplexer (which decides whether the system is in normal or test mode) makes the memory to get a wrong input.
Two different designs were implemented for the SRAM block alone and for the SRAM block with gating circuitry respectively. Depending on the requirements and on the testing strategies employed at top level, one of the two cells can be instantiated.
Three considerations were taken into account for the BIST integration:

- The working frequency for the BIST structures to correctly test the memory operations must be equal to the functional clock frequency, 40 MHz.
- The structure must be fully embedded in the memory block, and no additional external signals and pads are allowed. Configuration signals can be integrated in the already existing I^2C protocol. The latter is present in the PS module to configure all the other internal parameters, and a radiation hard I^2C slave handles the operations through Wishbone protocol [20].
- A maximum of 3 mW is assigned for the test of each SRAM, to be able to test simultaneously all the MPA memories within the given power budget. Considering the SRAM power consumption (1.9 mW in the typical corner), 1.1 mW is left for the BIST hardware.

Test pattern generation

Test inputs required by March algorithms can be created using a 128 bits programmable shift register. Contrary to ROM-based and LFSR approaches, which are the standard techniques employed in Built-In Self-Test designs, it allows to generate the desired patterns with a limited area overhead and a simple controller. The controller provides the shift direction (left/right) and the bit to be scanned in (0 or 1). For example, in case the line is initialized to zero and the second operation of the second march element has to be performed (that is writing 1 to all cells starting from the LSB), the shift register will assume for each writing cycle the following status: 100000, 110000,

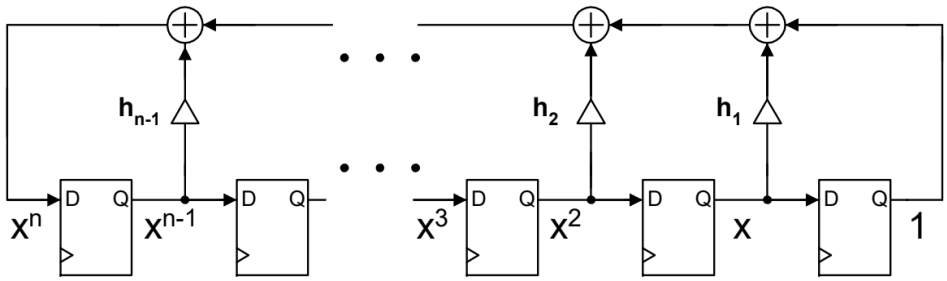


Figure 3.4. Linear Feedback Shift Register architecture. LFSRs produce psudo random patterns depending on the number of flip flops and XOR gates.

111000 until 111111 to write the MSB. Oppositely, as required in the third step, to write 0 starting from the MSB in a line containing all 1s the shift register has to be programmed as follow: 111110, 111100, 111000, etc.

Output response analyzer

The output read from the memory has to be compared with the expected response: a comparator is instantiated for this purpose. To simplify the control logic the good machine response comes from a delayed version of the input shift register, which takes into account the latency of the memory operations and the controller timing (it simply consists in a pipeline stage which delays the data presented at the memory data input). In correspondence of each reading operation the comparator assesses the good or faulty behaviour for each cell and saves the result (1 for a failing cell and 0 for a good cell) in an internal register, which can be accessed through I^2C protocol. Theoretically, a 512 bits register is required to store the result of the whole March test (at least 1 bit for each line) but the limited space available for addressing within the protocol makes necessary a compression of the comparator outputs. Being 128 the maximum number accessible bits, each of them will contain the result of the March test performed on 4 consecutive lines. At the beginning of the test the register is reset to 0 and it's written only in case a 1 is provided in input, to make sure not to lose the fault information. Starting from the information contained in the latter register, an additional functional test can be successively performed to detect the exact faulty lines. Eventually, they can be also replaced by other redundant lines.

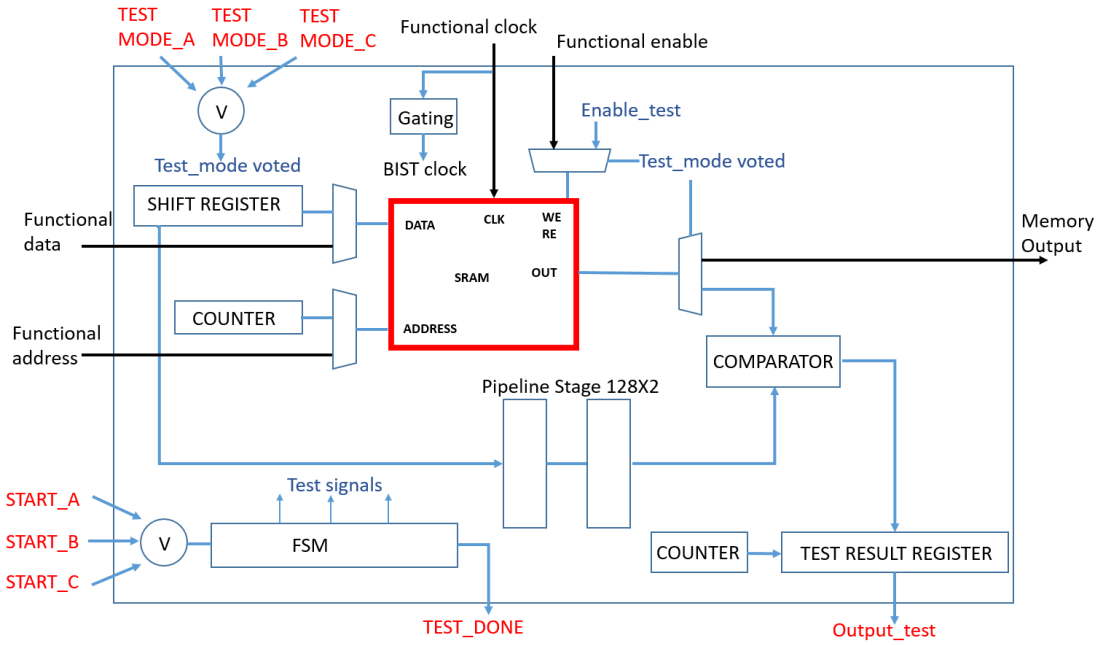


Figure 3.5. SRAM memory block with BIST. The functional paths are reported in black and the test structures in blue.

Architecture

Figure 3.5 reports the modified SRAM memory block (in black) embedding the BIST structures (highlighted in blue). Figure 3.6 displays the changes needed to include them in the MPA memory block with gating circuitry. In both cases, the additional area comprises:

- A 128 bit programmable shift register to create the test patterns according to the desired March test.
- A 9 bit counter to address the line (0 to 511) where the March element has to be performed.
- A pipeline stage to feed the second input of the comparator with the good output response.
- A 7 bit counter to be incremented each 4 memory lines to access the 128 bits register used to store the test result, written in case a 1 is provided at its inputs.
- A comparator (composed by 128 XOR ports and a final OR one) asserting its output signal to 1 in case of miscomparings.
- A multiplexer collar to separate the normal operation from the tests inputs. The multiplexer selection (TEST_MODE) is an additional input of the mem-

ory block. The functional output is disconnected during test phase to limit power consumption of external logic, which may be triggered during reading operations.

- A clock gating cell to disable the BIST structures clock pins during the normal operation of the ASIC, to limit power overhead (which eventually consists only in leakage). A latch is instantiated in the cell to protect the gated clock from glitches in the enable signal.
- A voter to protect multiplexers from single event transients on TEST_MODE signal buffers. TEST_MODE_A, B and C count 4 bits each and only in case all the voted bits are 1 the system enters in test mode.
- A Finite State Machine, which provides the control signals for registers, counters and memory.

In the second design, a modified clock gating cell to activate the 40 Mhz clock of the memory in test mode together with an additional pipeline stage to take into account the read latency is needed.

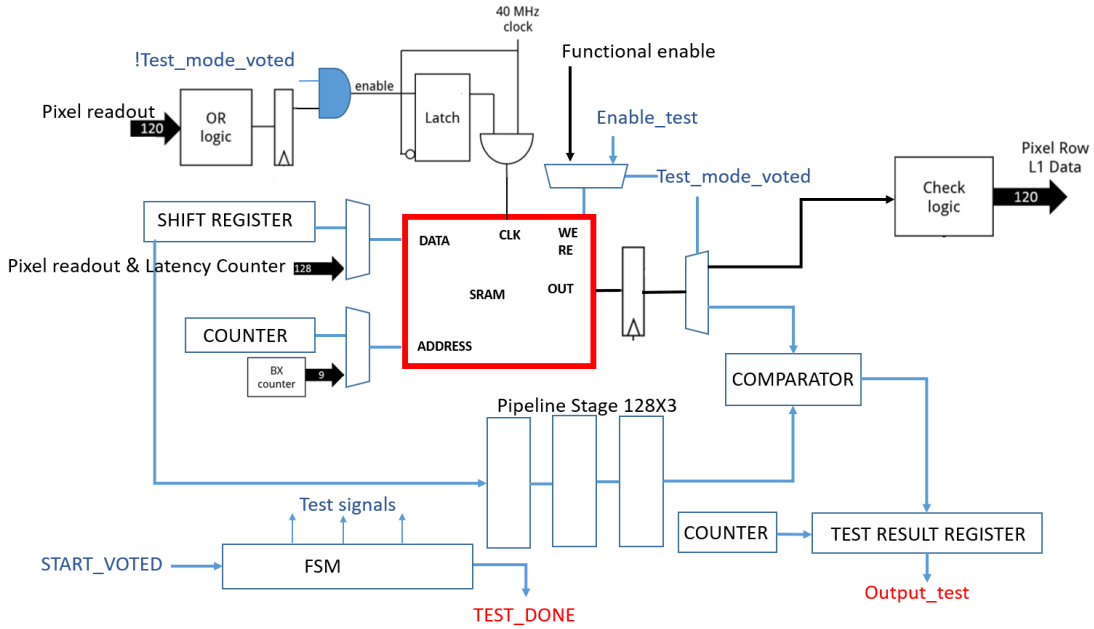


Figure 3.6. MPA and SSA memory block with gating circuitry and embedded BIST structures.

3.2.2 Operation and Timing

Since the BIST is supposed to be used only in the preliminary verification phases before MPA and SSA installation in the tracker and consequently without any incident radiation, differently from the triplicated control signals of the rest of the circuit, the BIST controller does not need to be triplicated.
The FSM encoding style chosen has registered outputs allowing to avoid glitches and uses two sequential blocks (with assignments coded using non-blocking statements) and one combinational block (with blocking assignments). Figure 3.8 reports the FSM chart where each column represents the operations needed to complete a certain March element of the March C- test chosen (see Chapter 2). Apart from the first and last column, in which the memory is initialized and read respectively, the structure is similar for all the others. At the beginning, the memory address counter, the 7bit counter (which controls the addressing of the test result register) and the shift register are initialized (SET/RESET WB_COUNTER and SHIFT_REGISTER), then the shift register is enabled (states 3, 8, 13, 18) and finally the writing and reading operations on the memory are performed (WRITE & READ states). Two checks are present: the first one (SHIFT_REG) ensures every bit of the line is both read and written with the new value, while the latter (ADDRESS) stops the current March element at the end of the memory lines. Both read and write enable are set active in the same cycle to stress the dual port functionality. To provide the correct input for the writing operation the shift enable signal must be advanced of one cycle, making necessary a pipeline stage before the comparator input. It allows to take into account the reading and control signal latencies. Since the TEST_MODE input (which controls the multiplexers) is used in the clock gating cell (which disables also the clock for the FSM flip flops), an additional START input is allocated to let the FSM working. To respect the I^2C command protocol it is triplicated and voted inside the block. The only output of the BIST is TEST_DONE, asserted when the test is completed and the result is ready to be read in the dedicated register.
Figure 3.7 displays the BIST functionality: at the beginning TEST_MODE and START signals are set to their active value, a fault (for example a stuck-at-0) is injected in line 49 and at the end of the test a 1 is present on the 12^{th} bit (equal to 49/4) of the TEST_RESULT register. The done signal indicates the test ends in 6.7 ms.

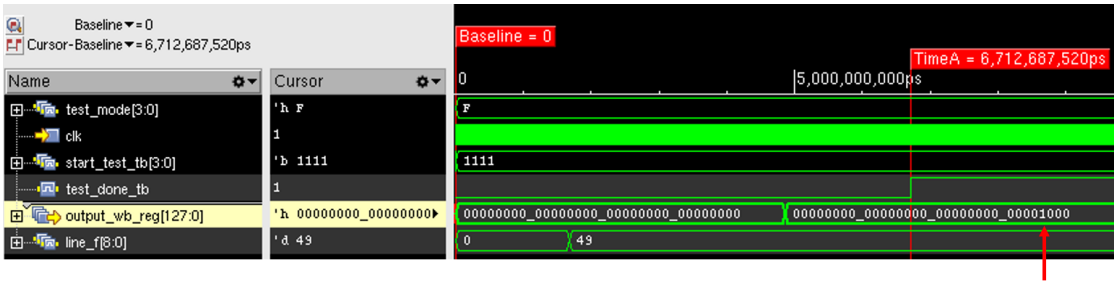


Figure 3.7. Testing interface

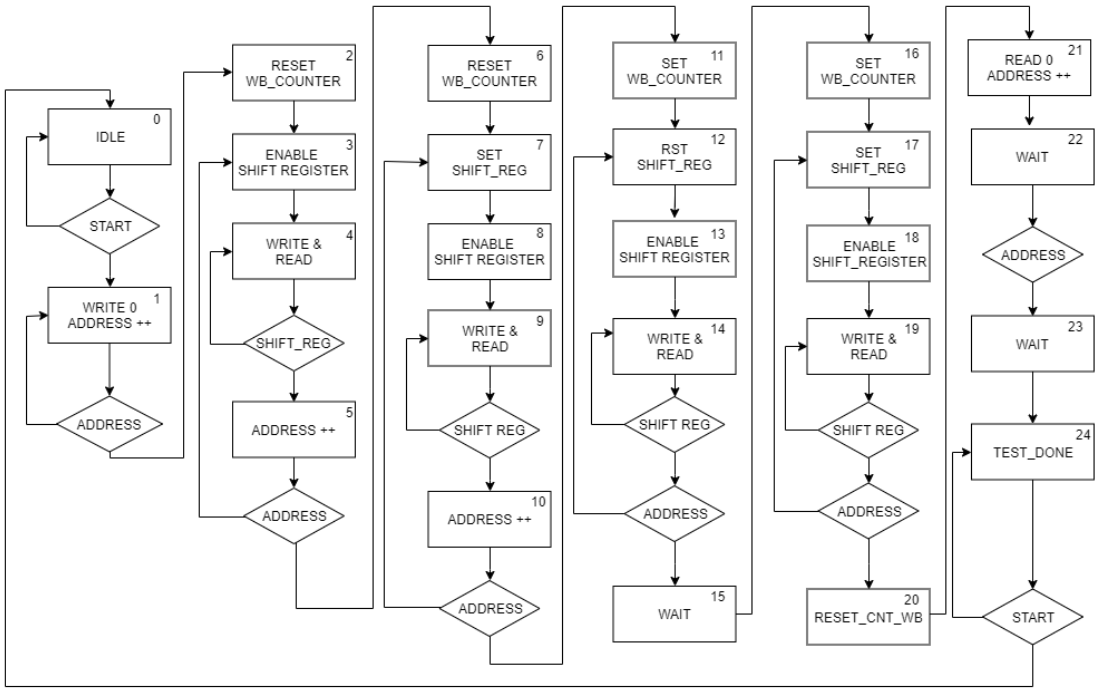


Figure 3.8. FSM state diagram

3.2.3 Place&Route and Power results

The routed memory blocks, with and without gating circuitry, are reported in Figure 3.9 and 3.10 respectively. The technology constraints and timing corners employed are the ones used for the full chip flow, reported in Chapter 1. MPA and SSA floorplans limit the block height to 550 μm, implying some routing congestion due to SRAM dimensions. The width was kept around 1000 μm in both cases. Only the first five layers of metals were used (metal 1, 2 and 3 for routing and metal 4 and 5 for the power grid), to leave other metals for the upper level connections.

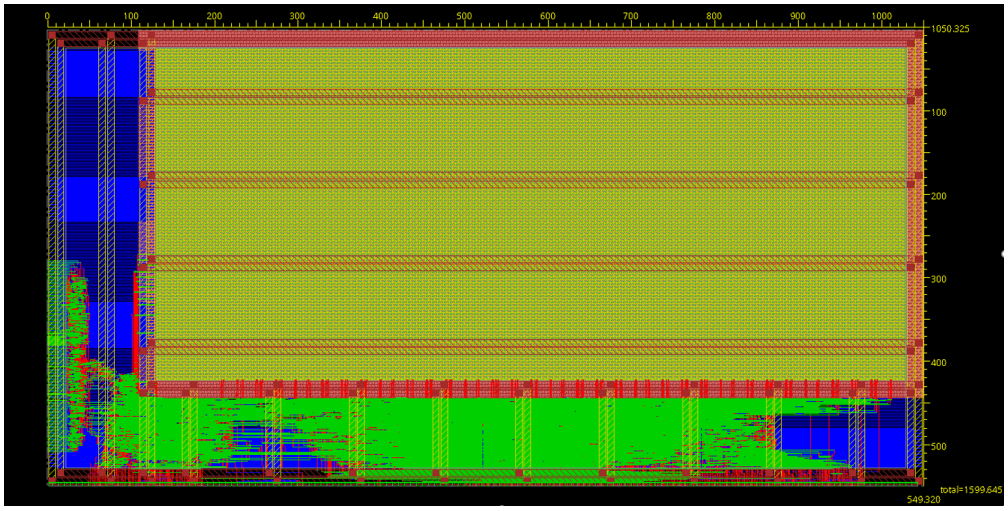


Figure 3.9. Routed memory block with BIST hardware

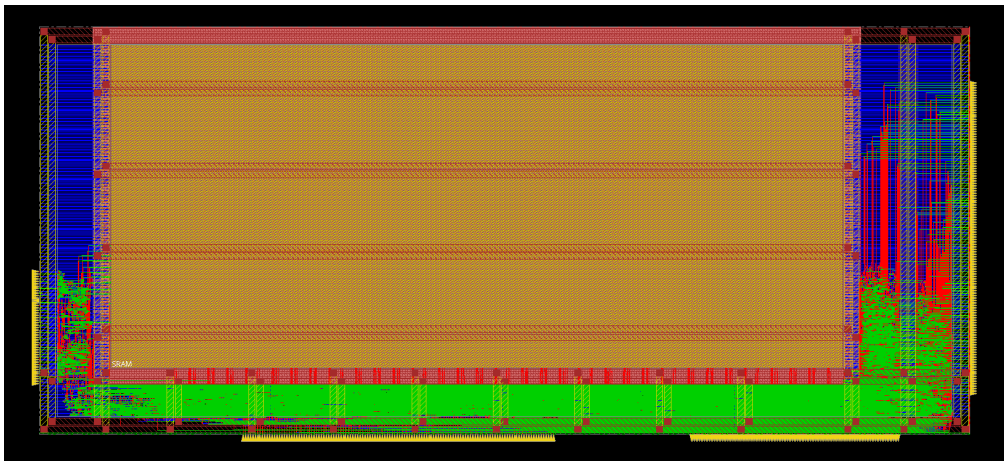


Figure 3.10. Routed SRAM block with BIST hardware

Vector-based power analysis was performed during the test mode operation of the block, where excessive switching activity might cause peaks in power consumption. The total dynamic power spent for the test is equal to 2.24 mW in the typical corner, which allows the parallel testing of all MPA memories to be still within the budget.

CORNER	TOTAL	SRAM	BIST
Typical 1.0 V	2.13	1.89	0.24
Maximum 0.9 V	0.42	0.24	0.18
Minimum 1.1 V	3.62	3.33	0.29

Figure 3.11. Power analysis results during March C- test. From top to bottom: typical, maximum and minimum corners. The results are expresssed in mW.

3.3 BIST performances

The designed BIST represents an efficient, low cost and self-contained testing method and will replace the currently used functional test, which takes 12 seconds to perform a single write and read operation accessing only one line. In less than 7ms (a thousand times faster), without any external software, it allows to detect the most frequent failures in SRAMs (stuck-at, address, transition and coupling faults) and to test its operations at working speed (40MHz), which results to be almost impossible with other approaches. The dynamic power spent during test allows the parallel testing of all 16 lines of MPA and of the whole module at the same time. The additional hardware inserted doesn't represent a problem neither in terms of area overhead nor in power consumption (being limited to a negligible leakage power thanks to the clock gating cells); in addition it's suitable for whatever type of March test and only the control FSM has to be revised. The already present Slow Control block, based on the I2C protocol, manages the BIST operations and makes it transparent and clock-gated during the normal operation. A negligible increase in the cross section may be caused by possible SETs on the multiplexers, which have however to be sampled to become effective SEUs.

4 Scanchain implementation in the CIC

4.1 The Concentrator IC

The Concentrator Integrated Circuit is used in both the 2S and PS modules and aims at buffering, collecting and formatting data coming from each group of 8 Front-End ASICs of the hybrids and sending them to the lp-GBT, which represents the connection with the beck-end. It performs both spatial (i.e. collecting information coming from 8 different ICs) and temporal compression, grouping together data for 8 consecutive bunch crossings. Two CICs are employed for each module, and around 30000 of them will be produced for the total outer tracker. Being a fully digital chip gathering results from all the other ICs on the hybrid its functioning is of fundamental importance: any faults in the CIC causes the whole module to fail. The block level architecture is reported in Figure 4.1 and its technical specifications can be found in [21].

The main structures inside the chip are four: the front-end blocks which treat data coming from 8 front-end chips (MPA or CBC), the output block which groups and formats them in an output packet, the System Manager in charge of generating clock and reset signals and the I^2C control block which interprets the configuration signals. Two fully independent paths are present to manage the two data types coming from the FE chips; the trigger data produced at 40 MHz are processed with minimum delay by the Trigger_Path block while L1 inputs follow a different flow in the L1_Path and are sent out only in case a trigger is received. L1 data frame coming from MPA and CBC are stored into eight FIFOs (one for each FE block) holding them for a maximum latency of 12.5us (the maximum time allowed to receive a trigger signal) and an output

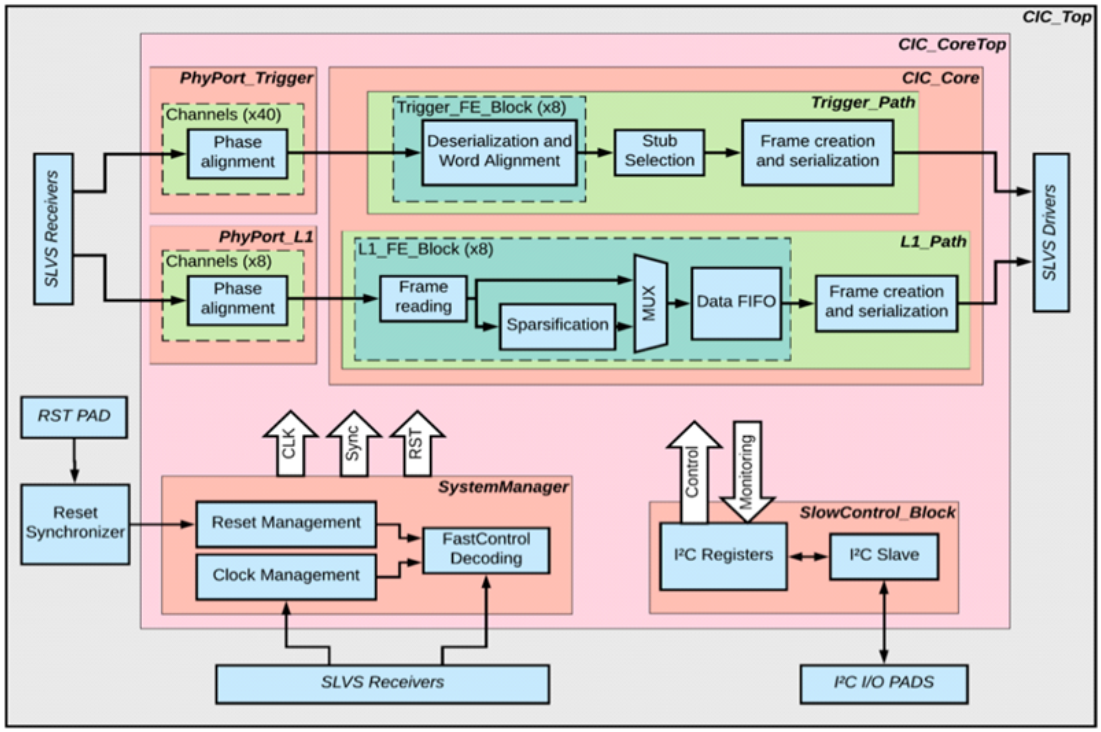


Figure 4.1. CIC block scheme

formatter block merge together data stored in the FIFOs that correspond to the same L1 event. Raw L1 data coming from CBCs are processed to extract information on hits (sparsification process) before being saved in the FIFO. Due to the limited number of output channels and their bendwidth, the Trigger_Path block processes received stubs according to their bending before sending them in output. Each channel presents a phase alignment system to align the input data (which is sent out according to the MPA or CBC clock) with the internal CIC clock and a word alignment block, which ensures each input line is aligned with the others. The CIC receives only one differential clock at 320 MHz as input and all the other clock signals (20 MHz, 40 MHz, 160 MHz and 640 Mhz) are generated internally by the SystemManager block according to the SlowControl configuration (saved into the SlowControl_Block). The same approach holds for the reset signal: apart from an hard reset (which reset the state configuration of the SlowControl_block), all the reset signals for the FIFOs, internal state machines and registers are generated internally.

The CIC is a fully digital chip comprising around 50K flip-flops and 130K latches (mainly composing the FIFO memories). Its architecture adopts the same radiation hardening techniques of MPA and SSA: the control part (System_Manager, SlowCon-

trol_block, control state machines) presents triplicated logic and signals (including triplicated clocks and reset for each copy instantiated) while the data path (FIFOs, pipeline stages and data processing) is kept not triplicated with single ended signals. Majority voters are instantiated to restore the good value in case of SETs and to transmit the signals from the triplicated logic to the not triplicated one, including clock and reset signals.

4.2 The Scan Design approach

The lack of internal controllability and observability points makes the ATPG algorithms to fail for most of faults in the design, leading to an unacceptable low fault coverage. Differently from the customized BIST implemented for the memory block, the high complexity of a circuit like the CIC requires an already structured DFT method, which can benefit from design automation and can exploit the powerful ATPG procedures. The DFT technique introduced in this case is the Scan Design approach, which aims at obtaining controllability and observability for the flip-flops in the design, so that they can be directly controlled as they were primary inputs (for this reason they are called Pseudo Primary Inputs, PPI). In addition to the normal functional mode, Scan Design approach introduces a mode of operation of the chip (named test mode) where memory elements in the design are serially connected to create one or more configurable shift registers called scanchains. These structures can be used to set and read flip-flops values, and test pattern generation has to be done only for the combinational part. The flip-flop states are set by shifting values inside the scanchain from an input pin and can be read by an output one. In order to embed the scan capability and to separate the two functionalities, D flip-flops must be replaced with scan flip-flops (Figure 4.2), which can select two different inputs (the output of the previous element in the scanchain or the functional signal coming from the logic) according to a control signal (usually called shift enable). Figure 4.3 represents the modified circuit with the insertion of a scanchain: the normal functionality is preserved (through the D input of the multiplexer) and a scan path is created, so that the functional paths can be bypassed to reach controllability and observability of internal nodes. The path highlighted in red represents the data flow during the two modes.

The scanchain insertion procedure is automatically performed during the synthesis and Place and Route flow, but the circuit must comply with the following rules to be

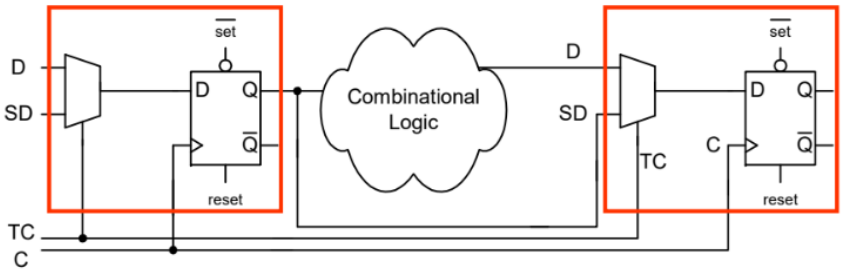


Figure 4.2. Scan flip-flop

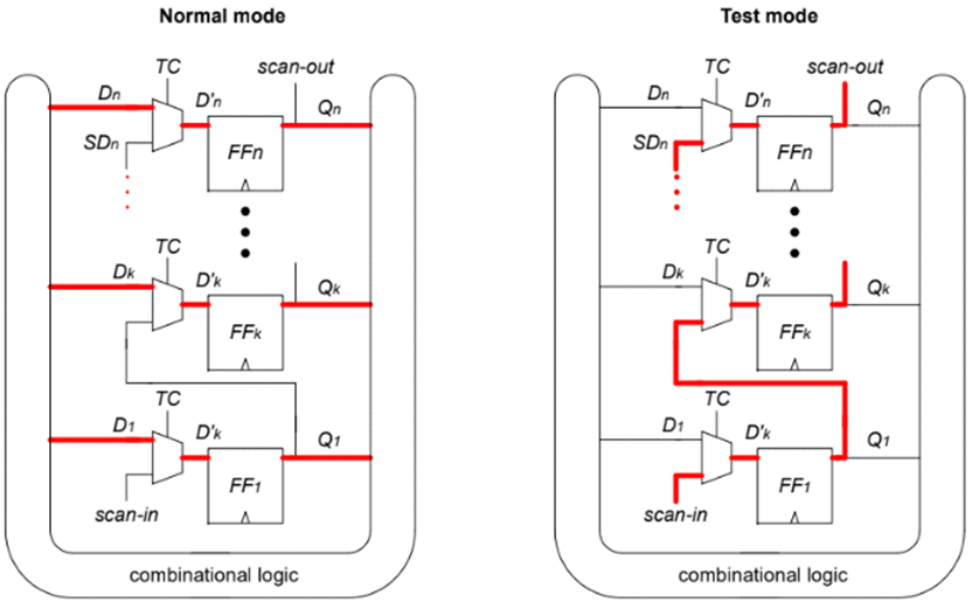


Figure 4.3. Scan circuit modes

scan-testable.

- At least one additional pin must be reserved for test purpose (i.e. the control signal to separate the two modes of operations). If possible, another input pin and an output pin must be assigned for shift-in and shift-out purpose (usually called SCAN-IN and SCAN-OUT respectively); if they are not available they can be multiplexed with a functional primary input through the shift enable signal.
- Only D-type flip-flops can be used in the design.
- All reset and clock signals of flip-flops composing the scanchains must be controllable from primary inputs. Internally generated or gated clock and reset signals are not valid testing signals.
- Data inputs of flip-flops must not be fed by clocks signals to avoid race condi-

tions in normal mode.

EDA software tools include scan insertion procedures, which involve the following steps:

- The needed DFT pins and signals are allocated in the RTL design.
- Scan configuration (scanchain properties and test signals) is defined during the synthesis step and flip-flops are replaced with the corresponding scan flip-flops during the mapping phase. Netlist modifications to ease scanchain insertion are carried out at this stage. The Synthesis step is managed by Cadence Genus solution.
- Scan reordering (i.e. reorder the position of each flip-flop inside the chain to minimize interconnection) and scan stitching (physically routing the interconnections among memory elements and control signals, from SCAN-IN to SCAN-OUT pins) is performed during Place and Route phase, managed by Cadence Innovus software.
- Test vectors are generated through ATPG on the final netlist by Cadence Modus tool.

The production test will exploit an external Automatic Test Equipment, which interprets the generated patterns and provides the correct timing for the circuit under test.

4.2.1 Scanchain timing

Scan-based testing is performed in two different phases. First of all, the scan register is tested by a shift pattern aiming at verifying the stuck-at faults in the flip-flops, the shifting operations and the connectivity. A toggle sequence of length N_{ffs}+4, where $N_{ffs}$is the length of the chain, is used to test all the possible transitions (0->0, 0->1, 1->1, 1->0). In this first step the shift enable signal is kept active, the reset signals are de-asserted and the sequence is clocked through the test clock signal. Single stuck-at faults in the combinational logic (and in the sequential part not connected to the scanchain) are targeted in the second phase of testing, where ATPG vectors are shifted inside the scanchain and successively applied to the logic. Assuming the shift enable signal (TC) active low (that it, when TC=0 the circuit is in shift mode), the following procedures are applied consecutively for each test vector:

- LOAD PHASE: TC is set to 0 and clock is activated for a number of cycle equal to the length of the scanchain to shift the test vector inside. The outputs generated and the SCAN-OUT bit are irrelevant at this step.
- CAPTURE PHASE: primary inputs (part of the test vectors) are presented and the circuit is forced in normal mode (TC is set to 1) for one clock cycle. The logic response is captured both in the scanchain elements and in output.
- UNLOAD PHASE: the sampled response is shifted out of the scanchain. This latter step can be merged with the scanchain load (step 1) to save time.

Figure 4.5 displays the timing in a scan design approach. Each test vector consists of S_n bits to be shifted serially in the scanchain concatenated with I_n bits to be applied as primary inputs in the capture cycle. Correspondingly, each output response consists of O_n bits sampled by the scan segment and N_n primary output to be compared with the expected response. The scan-in and scan-out phases overlap to save testing time.

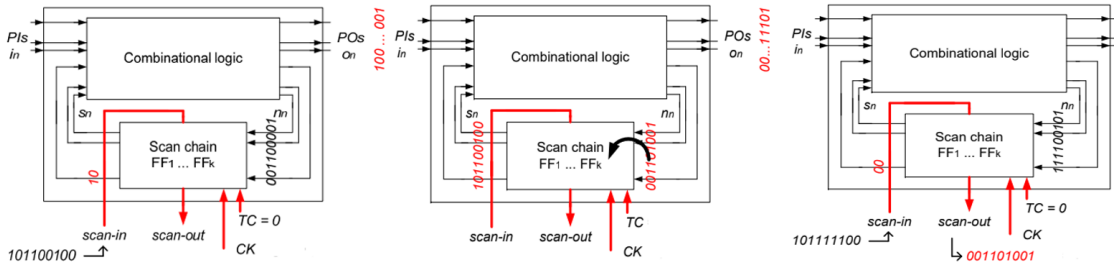


Figure 4.4. Scan-path architecture. From left to right: scan-in, capture and scan-out phases.

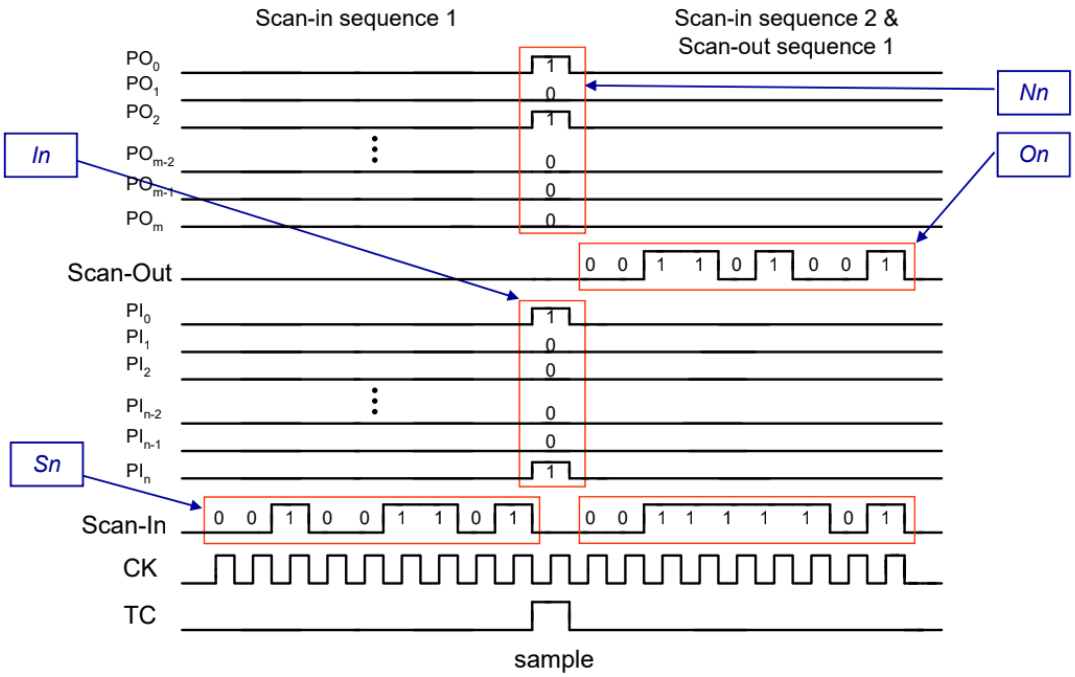


Figure 4.5. Scanchain timings.

The full-scan approach is considered so far one of the best DFT techniques for IC testing: it can be completely automated using commercial EDA tools and presents multiple advantages. It permits to access many internal nodes without requiring a single separate connection for each node accessed: being the bistable elements accessed and tested directly, test pattern generation for sequential circuit is not required. Serialization of test data allows also to reduce the number of additional input and output pins. The change from normal to test mode can be controlled simply by a test mode signal and the clock can be separated to create different timing modes to relax constraints. The main drawbacks concerning scan insertion are represented by the additional DFT circuitry which leads to area and power overhead and performance penalties in terms of timing closure and routing congestion. In addition, control signals and scan flip-flops conversion must be carefully handled in highly radiation environment.

4.3 Block-Level scanchain insertion

Apart from the submodule SlowControl_block, which is synthesized and routed separately to carefully control the correct spacing among triplicated registers, the synthesis

of the CIC is carried out at top level, without hierarchical distinctions among modules. However, a preliminary study of scanchain insertion performed only on a submodule (the TriggerPath block) highlighted the main issues of the scan design approach applied to this type of architecture.

4.3.1 Implementation

Three different clocks at 40 MHz, 320 MHz and 640 MHz are present in the block (they become nine after triplication), and a voted clock is distributed to the non-triplicated sequential logic. The reset operation is managed by three different reset signals (9 after triplication). Synthesizing only this block allows both the reset and clock signals to be present at top level as primary inputs; however, in the flat synthesis of the full chip they are derived internally as output of the System Manager, leading to violations of the DFT rules of Chapter 2. Even if clock and reset pins are present as input ports, only 30% of the whole number of FFs present in the design respect the DFT rules, namely can be scan-testable. In fact, being the block part of the data path, very few instances are triplicated and receive the corresponding clock and reset from input; the majority of the logic is instead non-triplicated and a voter is present on its clock and reset path. The limited number of flip-flops which can be connected into a scanchain with the actual configuration doesn't allow the MODUS automatic test pattern generation to converge to a good fault coverage (only 10%): a restricted number of control and observation point makes the ATPG process to fail for the most part of the stuck-at faults present in the design. Modifications to the netlist are necessary to be able to connect much more memory elements into a controllable and observable scanchain.

Testpoints insertion

Genus DFT synthesis software allows to detect and solve the clock and reset DFT violations, letting the automatic insertion of test points, consisting in blocks that bypass the source of violations to make the affected nets controllable again from primary input during test mode.
Figure 4.6 reports the two violations present in the Trigger_Path block which hinder the most part of flip-flops from being connected into a scanchain (they would cause a wrong shift and capture operations) and the methods employed by Genus Synthesis tool to fix them. In case of gating of the clock (typically due to clock gating cells or

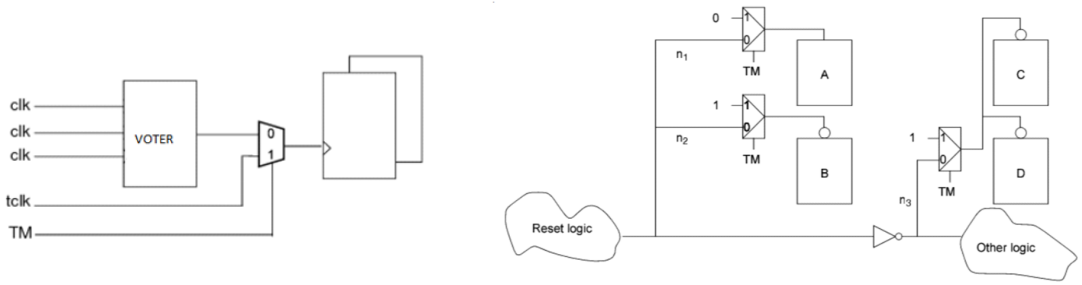


Figure 4.6. Fixing of DFT rules violations for clock (left) and reset (right)

voters), a multiplexer is inserted to select a certain test clock (which can be also a functional clock controllable from PI) while in case of reset violations a non-resetting value is selected during test phase. In addition to SCAN_IN, SCAN_OUT and SCAN_ENABLE (shift enable) needed for the scan design, the insertion of test points requires another dedicated input signal TEST_MODE, to control the multiplexer selection: it is asserted to its active value all along the testing time, both in shift and capture operations. TEST_MODE signal must not be confused with SCAN_ENABLE one and they cannot be used for the same purpose: scanchain operation corruption may arise in case of toggles of the TEST_MODE (the functional path would be selected in that case, leading to unknown values of clock and reset signals). On the contrary, SCAN_ENABLE determines the difference between shift and capture operations and switches during testing phase.
91 testpoints were automatically added by Genus DFT during the Synthesis step, allowing 100% controllability of clock and reset flip-flops pins.

Scanchain configuration in the Trigger_Path block

Being the clock signals available for testing as primary inputs, the 7433 scannable flip-flops (corresponding to 100% of the sequential elements in the design) were divided into three different scanchains according to their clock domain (40, 320 and 640 MHz). Each scanchain has a separate input and output shift port with common SCAN_ENABLE and TEST_MODE signals and joins together flip-flops within the same clock domain (including the three clocks and their voted ones) to avoid the insertion of lockup latches. In fact, in case different test clock domains are present in the same scan chain, lockup cells (mainly latches) are inserted on the scan path to synchronize the scan data between adjacent scan cells: they simply hold the previous scan data

and delay it so that it can be correctly captured by the following element.

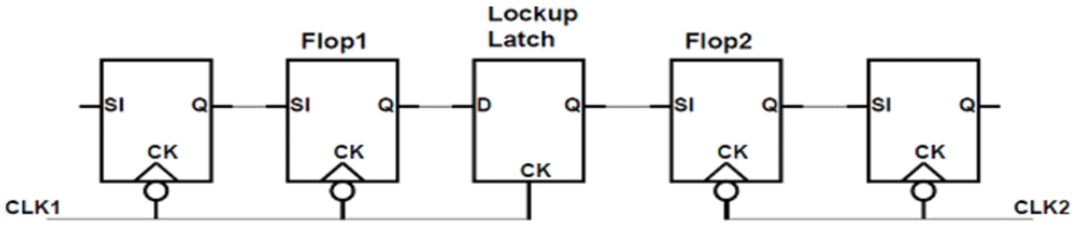


Figure 4.7. Lockup element insertion

The increased controllability and observability of internal nodes obtained through the scanchain insertion allows Modus ATPG to reach more than 97% of fault coverage: considering that in this testmode (i.e. adopting the scan design approach) 97.59% of the total logic is active, a final fault coverage of 95% is obtained for the whole block. The inactive logic reported by Modus includes faults which cannot be observed due to structural characteristics (dangling logic or TIE blocks) or to testmode constraints, and for this reason are not processed by the ATPG algorithms. In presence of a TEST_MODE signal which constrains fixed values for test control signals, the functional path of the multiplexer inserted as test point becomes for example part of the inactive logic.

Without any precise constraints on the area occupied by the instances of the Trigger_Path block in the total CIC, it was not possible to assess the impact of scanchain insertion on the routability of the design. In this case, for the Place and Route (PnR) phase a standard density of 60% was chosen to define the dimensions of the floorplan.

4.3.2 Power study and optimization

The final routed block was integrated at top level in the CIC and a vector-based power analysis was performed to check the normal mode of the chip, excluding in this way the shift operation. Power reports displayed in Figure 4.8 show a significant increase in the power consumption (more than 50%, from 20.7 mW to 31.3 mW) due to the additional hardware inserted. In fact, in addition to the multiplexers used to fix the DFT rules violations, around 20000 buffers where added along the scan-path connections (namely on the nets from the Q output of one flip-flop to the SCAN_IN input of the following one in the chain). The test points inserted on the clock path represent alone 18% of the total power overhead (they switch at each clock cycle)

while the remaining increase is due to the buffers inserted along the scan path, which toggle together with the output of the flip-flops they are attached to. Using non-minimum width buffers (which are more tolerant to radiation effects) worsens the power consumption.

Hold racing conditions

The large number of buffers inserted along the scan path to meet scanchain timing constraints are a direct consequence of the structure of the design, where multiple flip-flops receive a voted clock. The different latency among triplicated clocks, which are then merged into a single voted signal, makes necessary the insertion of buffers to respect hold and setup constraints. The latter (especially the hold timing checks) are worsen during the shift phase, in which the flip-flops behave like a simple shift register, where no logic delays the data launched. Figure 4.9 shows the analysis performed by Innovus tool which considers all the possible combinations of changes in the three clocks before the voter and reports the worst case slack calculation: the launch and capture latency of different clocks are different due to their distinct routing distributions. The same figure reports also an hold timing analysis performed after an optimization of the hold racing condition on the Trigger_Path.

Starting from three separate clock signals, a voter and a multiplexer are present on the clock path: the former to distribute a single voted clock to the datapath and the latter acts as testpoint (inserted during synthesis) to control the clock during the test phase. The top part presents an example of passing path, while in the bottom part a failing path it's reported. In the former case the inserted buffers delay the data enough to be correctly sampled. As the output of the first flip-flop switches during normal operation, all the buffers switch correspondingly, increasing the combinational power spent.

Solution: a new clock distribution

A way to balance the different launch and capture clock latencies is to remove voters on clock and distribute a single clock signals also to the data path instances that previously received the voted one. Figure 4.10 displays the modifications to the clock structure in the Trigger_Path block. On the left, the older version of the design presents

Group	Internal Power	Switching Power	Leakage Power	Total Power	Percentage (%)
Sequential	5.829	0.5458	0.001293	6.376	30.84
Macro	0	0	0	0	0
IO	0	0	4.917e-08	4.917e-08	2.379e-07
Combinational	7.115	5.083	0.006515	12.2	59.04
Clock (Combinational)	0.6289	1.462	0.0001147	2.091	10.12
Clock (Sequential)	0	0	0	0	0
Total	13.57	7.091	0.007922	20.67	100

Group	Internal Power	Switching Power	Leakage Power	Total Power	Percentage (%)
Sequential	6.352	0.4969	0.001678	6.851	21.89
Macro	0	0	0	0	0
IO	0	0	5.985e-08	5.985e-08	1.912e-07
Combinational	12.66	9.004	0.01014	21.68	69.26
Clock (Combinational)	1.071	1.698	0.0001509	2.769	8.849
Clock (Sequential)	0	0	0	0	0
Total	20.08	11.2	0.01197	31.3	100

Group	Internal Power	Switching Power	Leakage Power	Total Power	Percentage (%)
Sequential	6.359	0.7107	0.001696	7.071	29.06
Macro	0	0	0	0	0
IO	0	0	1.471e-07	1.471e-07	6.046e-07
Combinational	8.5	6.16	0.008499	14.67	60.29
Clock (Combinational)	0.913	1.678	0.000145	2.591	10.65
Clock (Sequential)	0	0	0	0	0
Total	15.77	8.549	0.01034	24.33	100

Figure 4.8. Power analysis on Trigger_Path block: starting from the the top, results on the design without scanchain, with scanchain (middle section) and with scanchain without buffers (lower part).

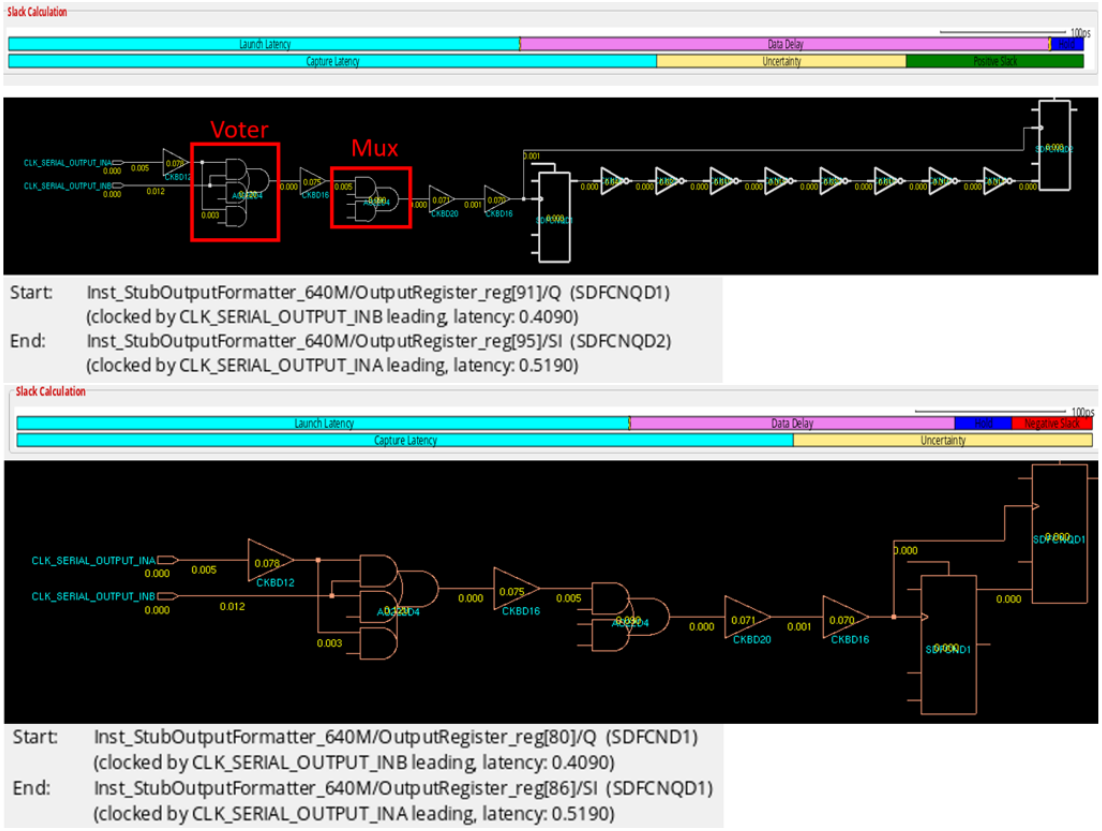


Figure 4.9. Slack calculation Innovus

control logic triplicated (with three different clock and reset signals) and datapath instances not triplicated. Voters (V1, V2) on clock and reset are present to create single signals for those instances. On the right, the new Trigger_Path structures without voters on clock: a new clock is distributed (in this case it's provided as additional input pin) and voters are removed. Triplicated instances still maintain their clock trees while clock pins of datapath instances are fed directly with the new clock. As demonstrated during the place and route flow of the total chip, all the design benefits from this different approach, which reduces the overall turnaround time and eases timing closure also on the functional paths.

4.3.3 Results

Together with a more balanced clock distribution, which reduces the number of buffers inserted along the scan path (and clock buffers too), the new clock simplifies the scanchain approach: testpoints insertion is not needed anymore (all the flip-

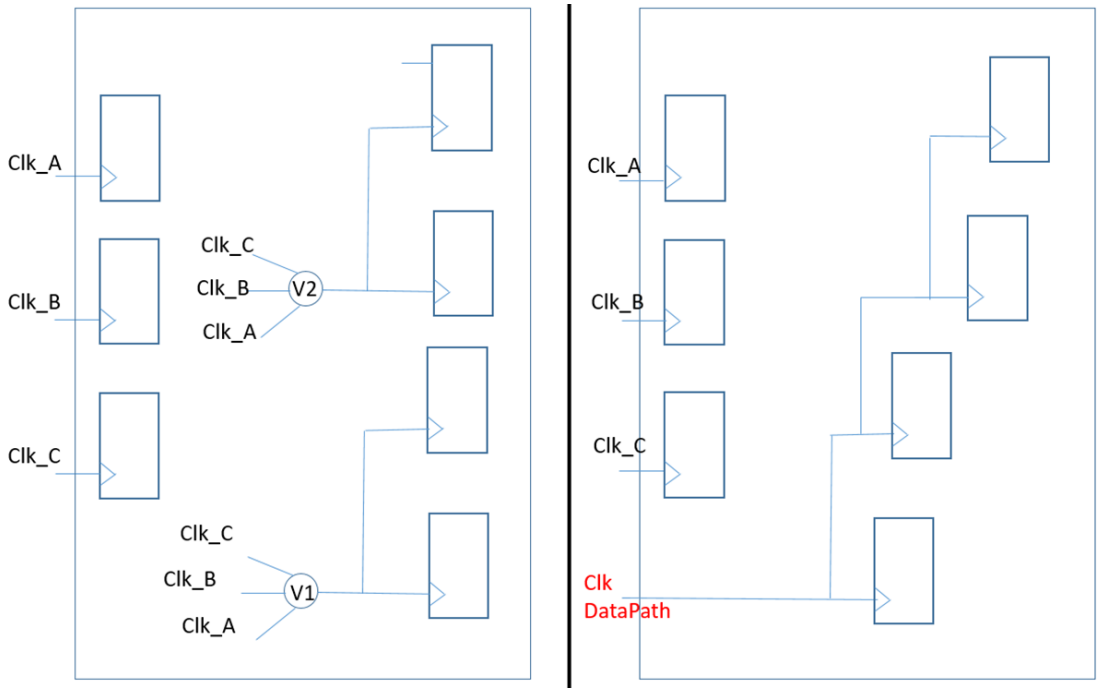


Figure 4.10. Clock tree modification

flops receive the clock signal as PI, according to the DFT rules). The results of the power analysis performed on the latest version of the Trigger_Path block are reported in Figure 4.11. The modified clock structure allowed to save 10% of power in the design without scanchain (from 20.67 mW to 18.46 mW) and the new Trigger_Path with scanchain resulted to be less power hungry even than the older block without scanchain (5% of power is still saved). The scanchain insertion in this latest version implies a power overhead of 6%.

Radiation tolerance represents the only drawback of this new technique: possible SETs on the clock buffers are not anymore filtered out by the voting system. However, since the modifications affects only the registers of the data path which were not triplicated (and for this reason already not protected against SEUs), the increase in the cross section should be negligible. In addition, a single voter was instantiated for a large number of flip-flops in the former design, and a re-buffering of the clock signal line was necessary before clock pins of sequential elements: SETs on those elements would not have been filtered as well.

Group	Internal Power	Switching Power	Leakage Power	Total Power	Percentage (%)
Sequential	6.337	0.3671	0.001681	6.705	34.24
Macro	0	0	0	0	0
IO	0	0	6.698e-08	6.698e-08	3.42e-07
Combinational	5.716	5.131	0.007215	10.85	55.43
Clock (Combinational)	0.5522	1.469	0.0001341	2.022	10.32
Clock (Sequential)	0	0	0	0	0
Total	12.6	6.967	0.00903	19.58	100

Group	Internal Power	Switching Power	Leakage Power	Total Power	Percentage (%)
Sequential	5.844	0.3325	0.001287	6.178	33.47
Macro	0	0	0	0	0
IO	0	0	4.823e-08	4.823e-08	2.613e-07
Combinational	5.83	4.552	0.006002	10.39	56.28
Clock (Combinational)	0.5158	1.376	0.0001113	1.892	10.25
Clock (Sequential)	0	0	0	0	0
Total	12.19	6.26	0.0074	18.46	100

Figure 4.11. Power analysis on the modified Trigger_Path with and without scanchain (top and bottom part respectively)

4.4 Full chip scanchain insertion

4.4.1 Scanchain setup

Some modifications were necessary to make the CIC compatible with the scan design approach. First, it was decided to instantiate all the scan related pins (SCAN_IN, SCAN_OUT, SCAN_ENA, TEST_MODE, SCAN_CLK), without multiplexing none of them with the functional inputs, to save as much as possible the radiation tolerance. In fact, in case of SETs on the multiplexers and their control signals, the whole corresponding data flow would be lost. With respect to the Trigger_Path block, in which the clocks were already present as primary inputs, an additional input test clock (SCAN_CLK) was needed in this case: being the only clock used for all the scannable elements, no lockup elements were needed along the scan path. The corresponding physical ports have been taken from the already present ground pads of the IO ring; for this reason the SCAN_ENA and TEST_MODE signals are defined as active high, namely the chip is in test mode when they are assigned a logic 1, so that no modifications

in the PS and 2S module PCB are needed. The only resulting drawback could be a possible IR drop in the power distribution. With the current IO ring layout (i.e. a single SCAN_IN and SCAN_OUT pad present), and excluding any form of compression logic (which allows to build different scanchains fed by the same SCAN_IN pin and merging in the same SCAN_OUT pin), the only possible configuration consists in a single scanchain connecting together all the flip-flops. Testing time represents the drawback of a single, thus long, scanchain which requires many shift cycles for each test pattern. However, different compression logic techniques introduce many disadvantages in terms of fault coverage reduction and hardware complexity increase and instantiating more scan-related pads (to create shorter scanchains with its own input and output) would entail a significant change in the ground distribution. For all these reasons, it was decided to build a single scanchain to limit negative impacts of the scan design implementation.
Secondly, following the example of the Trigger_Path Block, the clock distribution was modified to provide single clock signals also to the datapath instances (to replace the voted ones). In this case it was internally generated from the System_Manager, connecting it to the A version of each clock (see Figure 4.12). This artifice allowed also improving the timing closure steps, resulting in a decrease in the Synthesis and Place and Route flow turnaround time and lower number of failing paths.

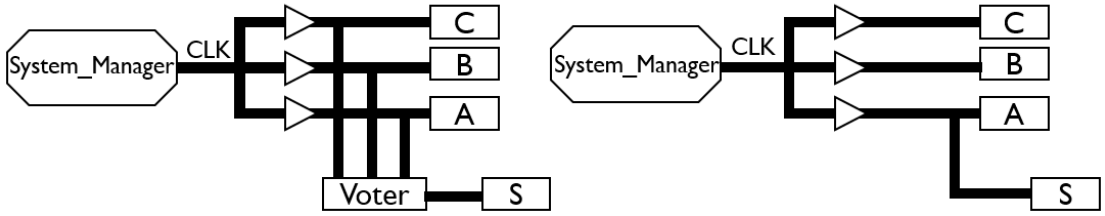


Figure 4.12. On the left, the old clock distribution with voters on clock, on the right the modified one with a dedicated clock for the datapath instances.

4.4.2 Rad-Hard solutions

Testpoints insertion

Due to the internal structure of the CIC, in which the System Manager manages all the timing signals, none of the flip-flops clock or reset pins are directly controllable

from outside its boundaries, leading to the violations of the DFT rules. Clock gating cells and combinational logic along the clock path contribute to those violations. As already presented in the case of the Trigger_Path block, the fixing procedures of Genus DFT software allows to automatically solve those violations with the help of control logic and TEST_MODE signal. Letting the automatic insertion of tespoints (as did for the hierarchical block) presented problems in the routability of the design: more than 1400 sources of DFT rules violations were found, implying a large number of multiplexers and gates added and nets routed (TEST_MODE and SCAN_CLK). The designed chip presented an high fault coverage (100% of flip-flops were connected in the scanchain) but timing and routability (DRV) issues. In addition, the automatic testpoints inserted (multiplexers and simple NAND gates) don't present any radiation hardening techniques thus increasing the cross section: for example, a SET on one of the TEST_MODE signal buffers may reset a large number of flip-flops. Even though the testpoints inserted are a negligible number with respect to the total instances in the design, they may also have an impact in the power consumption.
To solve the routability issue and to obtain a better control on the netlist modifications it was decided to fix those violations manually, looking for their sources and placing customized testpoints during the synthesis step. The latter were designed so that possible SETs on TEST_MODE buffers do not affect the multiplexer output: for this purpose a glitch filter was added to the usual multiplexer on its selection pin (Figure 4.13). As displayed in Figure 4.14, any pulse shorter than 14 ns can be filtered out, making the testpoint robust against single event effects along the control signal (the large number of buffers may have a significant impact on the cross section). Two different testpoints were built for the reset and clock violations respectively. In the former, the test input is internally tied to 0 (i.e. de-asserted during test time), while it's connected to SCAN_CLK in the clock testpoint.

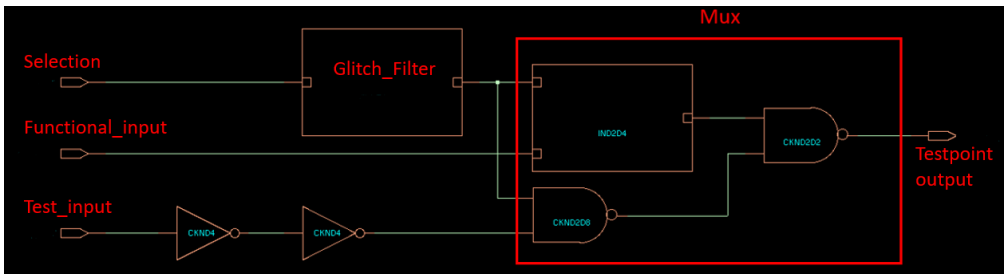


Figure 4.13. Manual testpoint

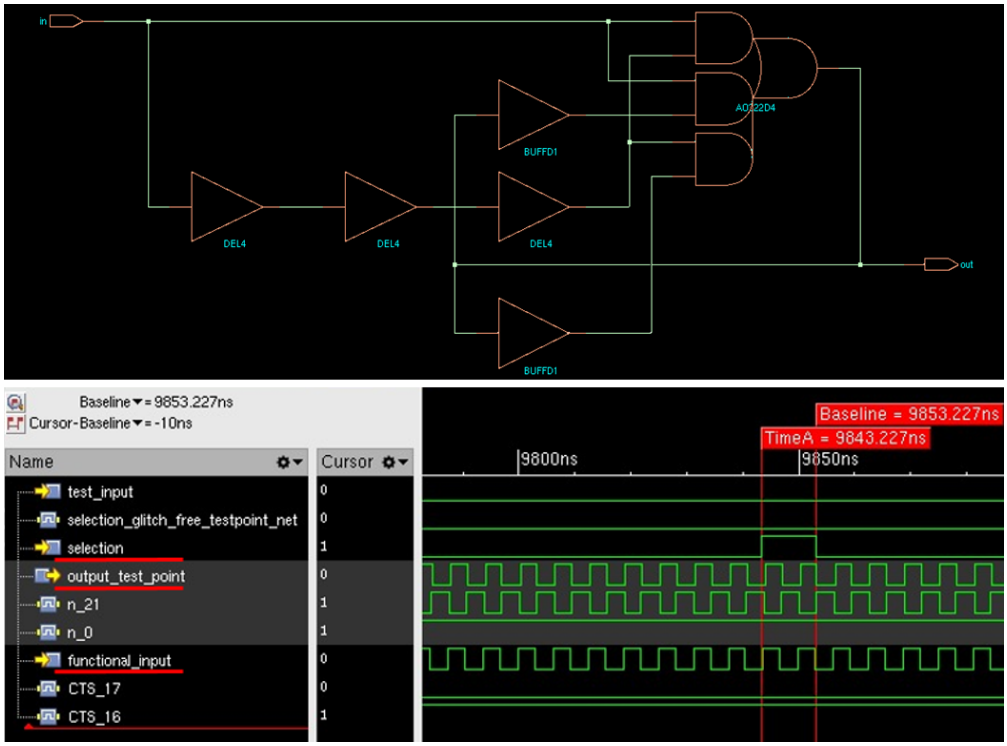


Figure 4.14. Glitch filter structure and simulation. It consists in a majority voter taking as inputs the glitch filter input, its delayed version and the output.

Manual modification of the netlist allows also choosing the exact nodes where test-points are added. Contrary to the automatic procedure it was at first decided to confine them only to the output of the System_Manager block, on the clock and reset signals sources: with a limited number of instances added (less than 30) around 85% of the flip-flops were made scannable (corresponding to a fault coverage of 70%). For each DFT violation (which prevents a certain number of flip-flops from being part of the scanchain), the corresponding testpoints must be inserted and the whole flow (from synthesis to signoff) has to be run to assess the best trade-off between routability, timing closure, scanchain length, fault coverage, SCAN_CLK frequency and power consumption. Keeping the test clock frequency equal to 20 MHz (which resulted to be the highest one to avoid setup timing issues) many iterations and attempts are needed to reach a good compromise.

Shift enable signal

SCAN_ENA, namely the test signal which activates the shift operation of the scanchain, represents another weakness of the Scan Design approach: single event transients along its line induce the flip-flops in the design to behave like a unique shift register (the test input of the scan flip-flops would be selected instead of the functional one). Even though the wrong signal needs to be sampled to become a real issue, thus decreasing the error probability, a radiation hardening technique must be implemented also in this case. Since Genus DFT does not allow triplication and voting of the SCAN_ENA, and in addition voters on scan flip-flops would imply routing congestion, the signal was internally triplicated at the output of the IO ring and each version (*ibuf_SCANena*, *ibuf_SCANenaB* and *ibuf_SCANenaC*) distributed to the corresponding triplicated flips flops of the scanchain (Figure 4.15). The SCAN_ENA pins of the remaining non-triplicated flops were tied to the three triplicated signals evenly.

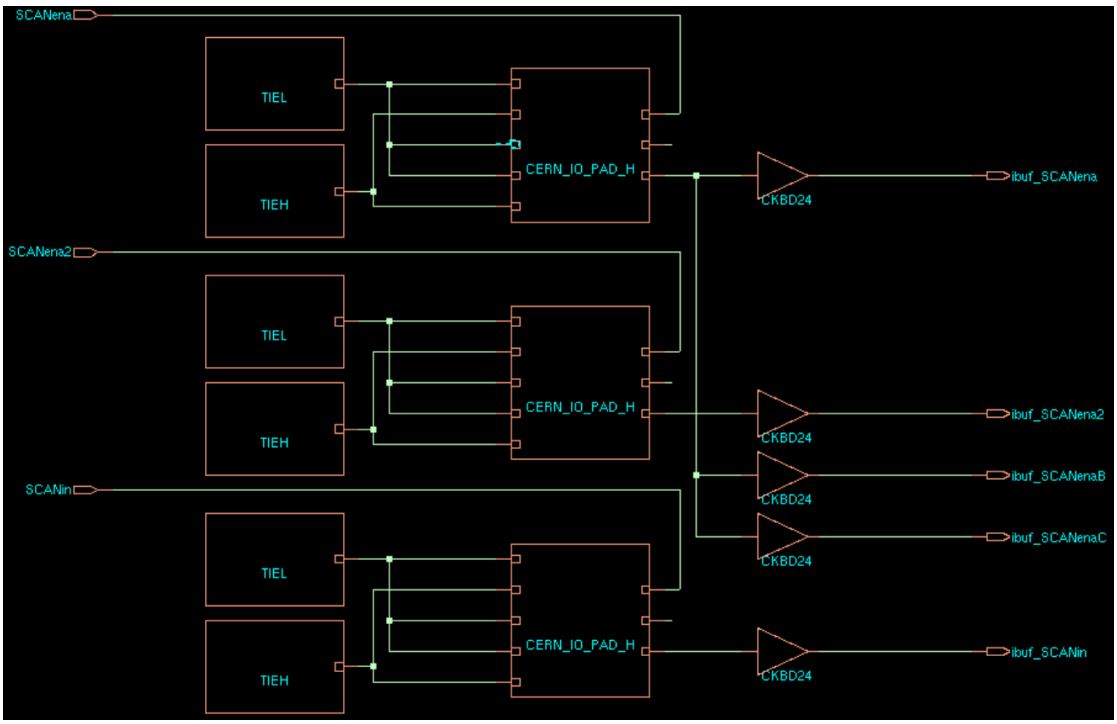


Figure 4.15. Triplication of SCAN_ENA signal at the output of the PADs (CERN_IO_PAD_H).

The I^2C Slave block

Since the I^2C Slave block groups together all the static configuration registers responsible for all the chip settings, it was decided to process it separately from the remaining flattened modules to carefully control its routing. Conversely, during the final routing optimization stages which move the instances, the distance (15 μm) among triplicated registers would not be respected and an impinging particle may cause multiple SEUs, creating severe errors on the chip functionality. For this reason, it was synthesised and routed separately and integrated in the top level flow as a macro (it's treated as a blackbox in synthesis and successively included as a library cell). Following the same techniques previously described (triplicated SCAN_ENA for triplicated registers and glitch filters inserted on testpoints) a single scanchain connecting together 747 flip-flops was created inside the I^2C block and four replica were instantiated in the CIC. The four scan segments were eventually connected at the head of the top scanchain, paying attention to correct also the DFT violations along their clock and reset signal paths (clock and reset pins are not anymore primary inputs once instantiated inside the total chip). Due to its internal structure presenting almost entirely triplicated and voted logic, the Scan Design approach results in this case the only way to successfully test all the possible faults thus guaranteeing the proper radiation robustness. In effect, the presence of majority voters may mask a faulty response coming from one of its three input branches when stimulated with a pure functional test. More than 95% of fault coverage was reached performing Modus ATPG algorithms on the I^2C block.

4.4.3 Final configuration and Fault coverage

Following the procedures explained in the previous paragraphs, 32 test points were added to provide control and observation nodes in crucial parts of the design, where a limited hardware overhead could result in a significant increase in fault coverage. The final version of the CIC presents 91% of flip-flops connected into a single scanchain counting 47407 bits, allowing Modus ATPG to reach a global fault coverage of 79.81%. The fullscan fault coverage, which takes into account only the testable logic within the built testmode, resulted to be 85%. The former values have been obtained performing ATPG algorithms with only scan-related pins (i.e. SCAN_IN and SCAN_OUT), without controlling any functional primary inputs or observing any primary outputs. Compared to the traditional approach which generates patterns involving all inputs and

outputs, this technique (called Internal Fullscan) requires a simpler tester and probe card, presenting nevertheless lower fault coverage.
Since the conversion to scan flip-flops and the additional signal routing may damage some critical functional timing paths, the affected flip-flops have to be excluded from the scanchain: in the CIC some configuration registers were excluded for this reason. The shift registers at 640 MHz of the frame creation and serialization block were excluded during scanchain stitching phase due to their timing violations in shift and capture mode. Since the structure was optimized for a working frequency of 640 MHz, the scan and capture phases at 20 MHz presented negative slacks, which could not be fixed without worsening the timing of the functional paths.

4.4.4 Place&Route and Power results

Additional timing modes generated by the insertion of DFT features have to be taken into account to correctly constraint the design under the three different modes of operations (SHIFT, CAPTURE and NORMAL). The SDC constraint generation is managed during the synthesis step: the default mode constraint file is loaded as input and the two additional files for SHIFT and CAPTURE modes are created. In the shift mode the scan chains need to be timed to make sure the shift registers can successfully shift test data from scan input to scan output; in the capture mode the design is timed for the capture operation according to the test clock characteristics. The typical constraints include creation of test clocks, setting test signals to certain fixed values and disabling timing checks on certain paths. During the routing step performed in Cadence Innovus software different timing analysis are performed simultaneously for the three modes and three different SDF files are generated for simulation of the functional and test mode (for shift and capture phases). Since the scanchain hardware is intended to be used during the testing phases before the installation of the chips in the tracker, the temperature corners for the CIC were weakened to ease timing closure: the temperature in the hold check was set at 0°C instead of -40°C and the corner for the setup check was moved from -40°C to 125°C. Three signoff optimizations were however needed after the routing stage to reach timing closure.

Figure 4.17 reports a comparison of the power consumption during the normal operation of two routed versions of the CIC, with and without scanchain respectively. The four corners represent the possible modes of operations of the CIC, which is

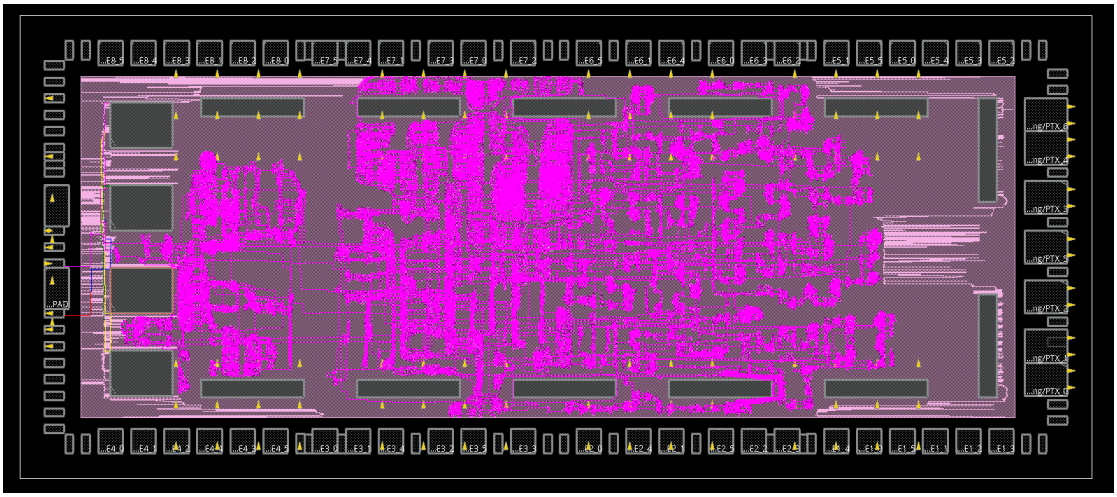


Figure 4.16. CIC layout with scanchain connections.

employed in the PS and 2S module, in connection with the MPA and CBC respectively. WORD ALIGNMENT and DATA SENDING refer to two different activities of the ASIC, namely the synchronization of the incoming data and the formatting of output data. Apart from the MAX corner at 1.32 V, which is not representative of the real working conditions, a maximum power overhead of 28% is reached in the worst case. Though significant, it is still acceptable and within the power budget.

corner	version	WORD ALIGNMENT	DATA SENDING
CBC 1.32V	NO SCANCHAIN	315	249
	SCANCHAIN	398	303
CBC 1.2V	NO SCANCHAIN	235	167
	SCANCHAIN	301	211
MPA 1V	NO SCANCHAIN	207	186
	SCANCHAIN	261	184
MPA 1.1V	NO SCANCHAIN	264	192
	SCANCHAIN	332	236

Figure 4.17. Summary of power analysis performed on the two versions of the CIC (with and without scanchain) in different corners. It takes into account the digital, analog and I/O parts. The values are expressed in mW.

Figure 4.18 displays a detailed comparison of the power spent by the digital part of the chip for the two versions of the CIC, without and with scanchain respectively. Two

are the main contributions to the power overhead, namely the increased switching activity of combinational gates and clock buffers. The former is mainly due to scan-path buffers, which are a significant number and toggle following the Q output of the related flip-flops. The latter component comes instead from the additional buffers placed along the clock paths, to compensate timing violations due to scanchain insertion. In our particular case, as a result of tespoints inserted at the System Manager output, the test and functional clocks share the same distribution, and hold racing conditions of scan-paths are fixed adding delay elements along the same clock tree. In this sense, the automatic insertion procedure of testpoints would have probably spread the two clock trees, allowing the optimization tool to work on different paths.

Group	Internal Power	Switching Power	Leakage Power	Total Power	Percentage (%)
Sequential	37.25	0.6141	0.02341	37.89	18.34
Macro	57.37	22.04	28.13	107.5	52.04
IO	0	0	6.312e-07	6.312e-07	3.055e-07
Combinational	34.62	11.5	0.08766	46.21	22.37
Clock (Combinational)	5.463	9.333	0.003888	14.8	7.163
Clock (Sequential)	0.1516	0.04204	1.155e-05	0.1937	0.09373
Total	134.9	43.53	28.24	206.6	100

Group	Internal Power	Switching Power	Leakage Power	Total Power	Percentage (%)
Sequential	37.32	1.783	0.02682	39.13	14.99
Macro	57.37	22.64	28.12	108.1	41.43
IO	0	0	4.343e-07	4.343e-07	1.664e-07
Combinational	40.62	39.63	0.09961	80.35	30.79
Clock (Combinational)	5.108	27.98	0.003666	33.09	12.68
Clock (Sequential)	0.1643	0.118	1.369e-05	0.2823	0.1082
Total	140.6	92.16	28.25	261	100

Figure 4.18. Comparison of digital power consumption of the CIC without (top) and with (bottom) scanchain.

In case the power overhead exceeds the power budget, a possible technique to avoid the drawback could be the scan path gating, namely separate the functional output with respect to the scan path. The result can be achieved both instantiating library elements (flip-flops) with a dedicated scan output pin or gating the scan path directly on the netlist, adding simple gates controlled by TEST_MODE signal before the long

sequences of buffers.

Patterns generated by ATPG algorithms have been simulated to assess the power consumption during the testing phase, where excessive switching (especially during the shift operation) may result in very high peak activities. Two possible test-benches can be built following the timing diagram presented in Figure 4.5: the vectors can be either serially shifted through the SCAN_IN pin emulating the true timing of the ATE or simultaneously loaded in parallel in the interested memory elements, allowing a significant reduction of simulation time. Similarly, the capture and scan unload phases can be verified by sequentially checking SCAN_OUT pin or comparing all the scan flip-flops outputs concurrently. In both cases, two different SDF files (shift or capture timings) have to be annotated at runtime according to SCAN_ENA signal. By means of MODUS software it was possible to classify sequences according to their switching activity and perform vector-based simulations for the most demanding vectors. Figure 4.20 displays toggle reports for SCAN and CAPTURE cycles, while Figure 4.19 reports the vector-based power analysis for the 208^{th} pattern performed in the worst corner (1.32V), which is still inside the power budget. In case power consumption in test mode exceeds the specifications, low power ATPG algorithms can be exploited to generate test vectors with limited switching activities.

```
Total Power
-----------------------------------------------------------------------------------------
Total Internal Power:       29.62472059           30.6747%
Total Switching Power:      25.86048318           26.7770%
Total Leakage Power:        41.09198279           42.5483%
Total Power:                96.57718720
-----------------------------------------------------------------------------------------
```

Figure 4.19. Vector-based power analysis on the 208^{th} sequence.The values are reported in mW.

4.5 Scan Design performances

The introduction of the Scan Design technique in the Concentrator ASIC represents an efficient probe for internal nodes, allowing to test faults in the design through the ATPG procedures. Exploiting only five I/O pins, the structural test reaches 80% of fault coverage in less than 6 seconds thanks to 1567 test vectors. The standard implementation approach has been modified to include radiation-hard (glitch filters,

```
INFO (TBM-054): Scan Shift Toggle report categorized by SCAN CYCLE [end TBM_054]

+=============+==========+============+=======+==========+==========+==========+================+================+
|  Odometer   | Relative | Tester     | Scan  | Total    | Scan Load | ScanUnload|  Unload        |  Load          |
|             | Sequence | Cycle      | Cycle | Switching | Switching | Switching |  Event         |  Event         |
|             |          | Number     | Number| Percentage| Percentage| Percentage|  Odometer      |  Odometer      |
+=============+==========+============+=======+==========+==========+==========+================+================+
|1.2.1.14.12  |    208   |   9904494  |43423  |    50.93  |    46.52  |     4.40  |1.2.1.14.12.6.1 |1.2.1.14.13.1.1 |
|1.2.1.44.12  |    688   |  32664757  |47366  |    50.72  |    50.67  |     0.04  |1.2.1.44.12.6.1 |1.2.1.44.13.1.1 |
|1.2.1.94.12  |   1483   |  70353692  |46146  |    50.70  |    49.32  |     1.37  |1.2.1.94.12.6.1 |1.2.1.94.13.1.1 |
|1.2.1.48.14  |    754   |  35791641  |45256  |    50.69  |    48.31  |     2.37  |1.2.1.48.14.6.1 |1.2.1.48.15.1.1 |
|1.2.1.46.2   |    710   |  33704833  |44444  |    50.69  |    47.43  |     3.26  |1.2.1.46.2.6.1  |1.2.1.46.3.1.1  |
|1.2.1.92.4   |   1444   |  68503698  |45103  |    50.65  |    48.14  |     2.50  |1.2.1.92.4.6.1  |1.2.1.92.5.1.1  |
|1.2.1.30.7   |    459   |  21808132  |47402  |    50.65  |    50.65  |     0.00  |1.2.1.30.7.6.1  |1.2.1.30.8.1.1  |
|1.2.1.87.1   |   1362   |  65325639  |43447  |    50.64  |    46.38  |     4.25  |1.2.1.87.1.6.1  |1.2.1.87.2.1.1  |
|1.2.1.74.5   |   1159   |  54993625  |46595  |    50.63  |    49.75  |     0.87  |1.2.1.74.5.6.1  |1.2.1.74.6.1.1  |
|1.2.1.64.13  |   1009   |  47883039  |47359  |    50.63  |    50.56  |     0.06  |1.2.1.64.13.6.1 |1.2.1.64.14.1.1 |
+=============+==========+============+=======+==========+==========+==========+================+================+

INFO (TBM-054): Capture Toggle report categorized by CAPTURE CYCLE [end TBM_054]

+=============+==========+================+==========+=================================================+
|  Odometer   | Relative | Event          | Total    |  Clock Name(s)                                  |
|             | Sequence | Odometer       | Switching |                                                 |
|             |          |                | Percentage|                                                 |
+=============+==========+================+==========+=================================================+
|1.2.1.73.10  |    1148  |1.2.1.73.10.4.1 |   48.65  | SCANclk                                         |
|1.2.1.66.5   |    1033  |1.2.1.66.5.4.1  |   48.20  | SCANclk                                         |
|1.2.1.2.6    |      10  |1.2.1.2.6.4.1   |   42.31  | SCANclk                                         |
|1.2.1.65.1   |    1013  |1.2.1.65.1.4.1  |   41.35  | SCANclk                                         |
|1.2.1.51.14  |     802  |1.2.1.51.14.4.1 |   40.28  | SCANclk                                         |
|1.2.1.22.11  |     335  |1.2.1.22.11.4.1 |   39.83  | SCANclk                                         |
|1.2.1.60.10  |     942  |1.2.1.60.10.4.1 |   39.71  | SCANclk                                         |
|1.2.1.90.15  |    1423  |1.2.1.90.15.4.1 |   39.49  | SCANclk                                         |
|1.2.1.17.7   |     251  |1.2.1.17.7.4.1  |   39.42  | SCANclk                                         |
|1.2.1.57.6   |     890  |1.2.1.57.6.4.1  |   39.39  | SCANclk                                         |
+=============+==========+================+==========+=================================================+
```

Figure 4.20. Switching activities report. The odometer values are the hierarchical number assigned to each test vector within the tester cycles. Sequence 208 shows the highest average switching activity.

shift enable signal triplication) and low power solutions (precise control of test points, clock tree modifications), to make scanchain insertion compatible with the ASIC constraints. The discovered drawbacks (routing congestion, timing closure) have been overcome and a first prototype has been included in the design and verified. Further optimizations can still be achieved in terms of fault coverage and testing time, applying recursively test point insertion. Power overhead and radiation tolerance revealed to be the main challenges for this DFT technique; the former can be addressed by the scan-path gating method while the real cross section increase can be estimated only by irradiation tests.

5 Conclusions and perspectives

The development of the final Front-End ASICs prototypes for CMS upgrade made necessary the setup of testing strategies for production phase. The currently used functional test, which stimulates the circuit simulating its real working activity, presents several limitations in terms of fault coverage, testing time and costs. A structural approach has been introduced to separate the functionality from the physical failures which may arise during manufacturing phase, allowing to build fault models and to exploit different testing algorithms. Logic complexity and limited number of I/O pins of modern ASICs represent the main issues for the latter algorithms to converge, requiring circuit modifications which ease fault sensitization and observation. Two different DFT techniques have been studied, for MPA/SSA SRAMs and CIC respectively. In the former case, a Built-In Self-Test has been developed and integrated in the memory block, so that the functionality is preserved and the March C- test is realized in hardware inside the IC boundaries. The architecture was thought to be as simple as possible to avoid power overhead and radiation issues. Furthermore, it can be easily customizable to adapt to new March algorithms.
In the second case, the Scan Design approach was implemented for the first time in the Concentrator ASIC. Additional constraints imposed in the field of High Energy Physics in terms of radiation tolerance and power increase have been considered and a simple scanchain has been built. The digital flow has been modified accordingly to automatize the procedure, and was run from RTL changes to verification phase and power analysis. The first results are encouraging, showing 80% of fault coverage reached in less than 10 seconds, one order of magnitude less than the current functional test. Compression logic, additional test points insertion and scan-path

gating techniques might further optimize the scan design in terms of testing time, fault coverage and power overhead respectively. Considering the developed methods, a new testing strategy is now available for the next prototypes of the PS module ASICs. The memory blocks (both SRAMs and flip-flop/latch based) will be tested separately from the rest of the circuit thanks to the BIST structure, while the remaining combinational and sequential logic can be tested through structural tests within a scan design approach.

Bibliography

References

[1] O. Bruning, H. Burkhardt, and S. Myers, "The Large Hadron Collider", *Prog. Part. Nucl. Phys.*, vol. 67, pp. 705–734, 2012. DOI: 10.1016/j.ppnp.2012.03.001.

[2] A. Marchioro, "A hybrid module architecture for a prompt momentum discriminating tracker at SLHC", *PoS*, p. 037, 2011. DOI: 10.1088/1748-0221/7/09/C09001.

[3] D. Contardo, M. Klute, J. Mans, L. Silvestris, and J. Butler, "Technical proposal for the phase-ii upgrade of the cms detector", Geneva, Tech. Rep. CERN-LHCC-2015-010. LHCC-P-008. CMS-TDR-15-02, Jun. 2015. [Online]. Available: https://cds.cern.ch/record/2020886.

[4] E. Benton and E. Benton, "Space radiation dosimetry in low-earth orbit and beyond", *Nuclear Instruments and Methods in Physics Research Section B: Beam Interactions with Materials and Atoms*, vol. 184, no. 1-2, pp. 255–294, 2001. DOI: 10.1016/S0168-583X(01)00748-0.

[5] F. P. La Torre, "Study of induced radioactivity in proton accelerator facilities", Tech. Rep., 2014.

[6] G. Borghello, "Ionizing radiation effects in nanoscale CMOS technologies exposed to ultra-high doses", PhD thesis, University of Udine, Dipartimento Politecnico di Ingegneria e Architettura, Udine, Italy, 2018.

[7] A. Mussgiller, "Detector Modules for the CMS Phase II Outer Tracker", no. DESY-2014-03109, 2014. [Online]. Available: https://bib-pubdb1.desy.de/record/170990.

[8] F. Faccio, G. Borghello, E. Lerario, D. M. Fleetwood, R. D. Schrimpf, H. Gong, E. X. Zhang, P. Wang, S. Michelis, S. Gerardin, *et al.*, "Influence of ldd spacers and h+ transport on the total-ionizing-dose response of 65-nm mosfets irradiated to ultrahigh doses", *IEEE Transactions on Nuclear Science*, vol. 65, no. 1, pp. 164–174, 2018. DOI: 10.1109/TNS.2017.2760629.

[9] L. J. Casas, D. Ceresa, S. Kulis, S. Miryala, J. Christiansen, R. Francisco, and D. Gnani, "Characterization of radiation effects in 65 nm digital circuits with the drad digital radiation test chip", *Journal of Instrumentation*, vol. 12, no. 02, p. C02039, 2017. [Online]. Available: http://stacks.iop.org/1748-0221/12/i=02/a=C02039.

[10] D. Ceresa, “Electronic systems for intelligent particle tracking in the high energy physics field”, PhD thesis, Politecnico di Torino, 2016.

[11] S. Kulis, “Single Event Effects mitigation with TMRG tool”, *Journal of Instrumentation*, vol. 12, no. 01, p. C01082, 2017. DOI: 10.1088/1748-0221/12/01/c01082.

[12] S. Miryala1, T. Hemperek, and M. Menouni, “Characterization of Soft Error Rate Against Memory Elements Spacing and Clock Skew in a Logic with Triple Modular Redundancy in a 65nm Process”, *PoS*, vol. Twepp-17, 2018. [Online]. Available: https://pos.sissa.it/343/029/pdf.

[13] S. Hamdioui, “Testing multi-port memories: theory and practice”, 2001.

[14] P. Veenstra, F. Beenker, and J. Koomen, “Testing of random access memories: theory and practice”, in *IEE Proceedings G (Electronic Circuits and Systems)*, IET, vol. 135, 1988, pp. 24–28.

[15] J. P. Roth, “Diagnosis of automata failures: a calculus and a method”, *IBM journal of Research and Development*, vol. 10, no. 4, pp. 278–291, 1966.

[16] M. L. Bushnell and V. D. Agrawal, *Essentials of Electronic Testing*. KLUWER ACADEMIC Publisher, 2002.

[17] A. Caratelli, D. Ceresa, J. Kaplon, K. Kloukinas, Y. Leblebici, J. Murdzek, and S. Scarfi, “SSA: design and implementation of a strip sensor readout ASIC for on-detector particle discrimination”, *Submitted for Journal of Instrumentation*, 2019.

[18] A. Caratelli, D. Ceresa, J. Kaplon, K. Kloukinas, Y. Leblebici, J. Murdzek, and S. Scarfi, “MPA: Design and implementation of a pixel readout ASIC, on a 65 nm technology, featuring on-chip real time particle discrimination.”, *Submitted for Journal of Instrumentation*, 2019.

[19] R. Brouns *et al.*, “Development of a radiation tolerant low power sram compiler”, 2014.

[20] A. Caratelli, “Research and development of an intelligent particle tracker detector electronic system”, PhD thesis, CERN, 2019.

[21] L. Caponetto, G. Galbit, B.Nodari, S.Scarfi, and S. Viret, “Cic1 technical specification”, Institut de Physique Nucléaire de Lyon, Tech. Rep., 2018. [Online]. Available: https://espace.cern.ch/Tracker-Upgrade/Electronics/CIC/Shared%20Documents/Specifications/CIC_specs_v2p1.pdf.