UNIVERSITÀ DEGLI STUDI DI PERUGIA

Dottorato di Ricerca in Ingegneria Industriale e dell'Informazione - XXX Ciclo

DIPARTIMENTO DI INGEGNERIA

VIA G. DURANTI 93 - 06125 - PERUGIA (I)

TEL. 075-5853653 • FAX 075-5853654

DESIGN AND OPTIMISATION OF LOW POWER HYBRID PIXEL ARRAY LOGIC FOR THE EXTREME HIT AND TRIGGER RATES OF THE LARGE HADRON COLLIDER UPGRADE

Ph.D. candidate: Sara Marconi

Supervisor:
Ph.D. Pisana Placidi

Ph.D. Coordinator:
Prof. Ermanno Cardelli

Co-supervisor:
Eng. Jørgen Christiansen

A dissertation submitted for the degree of
Doctor of Philosophy in
Industrial and Information Engineering

A.A. 2017/2018

"Engineers like to solve problems.
If there are no problems handily available,
they will create their own problems."
Scott Adams

Preface

Digital design of integrated circuits in nanometer technology requires to address several design challenges. Among those, system complexity has to be handled with modern techniques and tools, power density needs to be considered as a major player in design choices (trade-off versus performance), clock distribution and timing closure require special attention due to large chip size (impact of interconnections) and variability issues demanding additional timing safety margins. These issues are common to multiple research and industry applications, among which the design of the readout electronics of next generation hybrid pixel detectors for the High-Luminosity Large Hadron Collider (HL-LHC) at CERN. In addition, in this application circuits operate in harsh radiation environments, experiencing performance degradation and various classes of hard/soft errors. Pixel detectors are devices capable of detecting different forms of radiation with high resolution (up to a few micrometers), thanks to the small size of the sensing element. In High Energy Physics (HEP) applications particles are detected based on their ionizing interaction with the sensor and collected information has to be readout through dedicated high density electronics. For applications demanding fast detection, tolerance of high radiation levels (above tens of Mrads), reliability with high input rates (in the order of few GHz/cm^2), the sensor and the electronics are normally fabricated in separate substrates. Such a systems is referred to as Hybrid Pixel Detector (HPD).

This work is part of the effort to design the digital readout electronics of next generation hybrid pixel detectors. The most relevant examples are the pixel detectors for the HEP experiments *A Toroidal LHC ApparatuS* (AT-

LAS) [1], *Compact Muon Solenoid* (CMS) [2] and *A Large Ion Collider Experiment* (ALICE) [3] at the LHC. New generation pixel detector systems and ASICs for High Energy Physics (HEP) applications will be a big step forward and will have to meet specifications in terms of smaller pixels to improve tracking resolution, much higher hit rates ($3\,\mathrm{GHz/cm^2}$), much higher output bandwidth and large integrated circuits with low power consumption and low power fluctuations. Their electronics will also have to work reliably for years under the hostile radiation conditions, requiring unprecedented radiation tolerance (up to 1 Grad). The PhD activity is part of the design effort to develop the digital readout electronics of such complex systems using commercial high-scaled technologies and requires to face challenges which are of common interest in the technological and scientific context, i.e. system complexity, low power consumption, reliability in hostile environments (shared also with space applications).

The design of next generation pixel detector systems has driven the creation of multiple collaborations and projects, to which the PhD activity has been an active contribution:

- RD53 [4], an international collaboration of universities and research institutes, targeted to design the next generation of hybrid pixel readout chips to enable the phase 2 pixel upgrades of the ATLAS (A Thoroidal LHC ApparatuS) and CMS (Compact Muon Solenoid) experiments. The readout chip, named RD53A, has been prototyped in August 2017;
- CHIPIX65 [5], an italian project born with the primary goal of developing an innovative CHIP for a PIXel detector, using a CMOS 65nm technology, for experiments with extreme particle rates and radiation (effort shared with the RD53 Collaboration). Such a chip has been prototyped in June 2016;
- AIDA-2020 (Advanced European Infrastructures for Detectors at Accelerators) [6], an European project aiming at pushing detector technologies beyond the state-of-the-art and offering highly equipped infrastructures for testing.

In the remainder the contents and the organization of the report will be described. In Chapter 1, an overview on the state of the art of hybrid pixel detectors is provided and the motivations and requirements of this work are introduced; Chapter 2 describes the contribution to the development and optimization of a simulation framework for complex integrated circuits, based on advanced verification methodology. The framework is aimed to handle system complexity, allow architectural studies for design optimisation and achieve extensive verification. In Chapter 3: first, the pixel readout chip architecture and floorplan choices are introduced; second, the contribution to the architecture optimization and comparison by means of the developed framework is summarized and results are reported for multiple stages of the design optimisation (behavioural level, initial architectures for small-scale prototypes, digital architectures for the RD53A chip). In Chapter 4, the focus moves to the power methodology defined for the optimization of the chip design and its critical serial powering scheme. Results from detailed power analysis and optimizations of the digital pixel array architecture are reported. Finally, Chapter 5 is centred on issues regarding radiation tolerance in the harsh environment and timing-related optimisations. Radiation effects and chosen design techniques are initially presented, in order to introduce the reader to the results of top-level timing optimisation (including radiation degradation), clock distribution and tolerance to bit upsets.

Contents

List of Figures

List of Tables

List of acronyms

ACB Analog Chip Bottom 70
ASIC Application-Specific Integrated Circuit 24
BCR Bunch Counter Reset 73
CBA Centralised Buffering Architecture 87
CDR Clock Data Recovery 71
CDC Clock Domain Crossing 74
CMD CoMmand Decoder 73
CS Channel Synchroniser 73
CTS Clock Tree Synthesis 119
DAQ Data Acquisition System 30
DBA Distributed Buffering Architecture 87
DCB Digital Chip Bottom 70
DFE Differential Front End 71
DICE Dual Interlocked storage CEll 140
DPI Direct Programming Interface 59
DUT Design Under Test 109
ECC Error Correction Coding 141
ECR Event Counter Reset 73
ELT Enclosed Layout Transistor 131
EOC End of Column 30
FIFO First-In First-Out 74
FE Front End 71

FSM Finite State Machine 77
HDVL Hardware Description and Verification Language 45
HEP High Energy Physics 3
HLS High Level Synthesis 47
HPD Hybrid Pixel Detector 3
IBL Insertable b-Layer 102
ICG Integrated Clock Gating 117
LDO Low-DropOut 103
LFE Linear Front End 71
MBU Multi Bit Upset 134
MMMC Multi Mode Multi Corner 138
MIP Minimum Ionising Particle 63
OOP Object-Oriented Programming 45
OVM Open Verification Methodology 47
P&R Place&Route 91
PLL Phase Locked Loop 71
PR Pixel Region 39
RTL Register Transfer Level 46
SAIF Switching Activity Interchange Format 110
SEB Single Event Burnout 133
SEE Single Event Effect 133
SEGR Single Event Gate Rupture 133
SEL Single Event Latch-up 133
SET Single Event Transient 133
SEU Single Event Upset 41
SFE Synchronous Front End 71
SLDO Shunt and Low Drop Output 103
SPEF Standard Parasitic Exchange Format 110
SDC Synopsis Design Constraints 146

SOI Silicon on Insulator 130

STA Static Timing Analysis 138

STI Shallow Trench Isolation 131

SV SystemVerilog 44

TCF Toggle Count Format 110

TID Total Ionising Dose 41

TL Transaction Level 45

TMR Triple Modular Redundancy 141

ToT Time over Threshold 27

UVM Universal Verification Methodology 44

VCD Value Change Dump 110

VCO Voltage Controlled Oscillator 71

Chapter 1

State of the art and requirements for next generation hybrid integrated chips in harsh environments

The sensing and readout components of state of the art hybrid detectors are introduced in Section 1.1 and their applications are described. Moreover, Section 1.2 presents the requirements of next generation hybrid pixel detectors, highlighting challenges which are addressed in this work.

1.1 Hybrid pixel detectors and applications

The notion of pixel comes from image processing applications and it describes the smallest discernible element in a given device. A pixel detector is therefore able to detect an image and the size of the pixel corresponds to its granularity. Devices used in everyday life such as photo cameras, video cameras and X-ray films are basic examples of such systems composed of a sensing element (pixel) which interacts with photons of different energies and generates an intensity distribution i.e. the image. For HEP applications, images or patterns are not

generated by visible light, but by charged particles or photons in the keV to MeV energy range, which experience an ionizing interaction with the detector.

HEP experiments demand the use of the so-called HPD, since they are particularly fast and able to detect high-energy particles and electromagnetic radiation [7]. Detection is performed through different devices with specific functions: *i)* a sensor converts part of the energy of the radiation into an electric signal, *ii)* the signal is pre-processed by the front-end electronics and further treated by a digital readout circuitry, *iii)* eventual processing and storage allow for later inspection and data analysis. The peculiarity of "hybrid" pixel detectors comes from the fact that the sensor and the readout ASIC are fabricated separately and are then joined together through a process called bump bonding, as shown in Figure 1.1 for one single pixel. Such a process is characterised by a rather high cost, but they are also capable of standing high radiation levels and suitable for high resolution and high rate applications. The main characteristic of a hybrid pixel detector is the high density connectivity between the sensing elements and the readout electronics. For this reason it is required that the connectivity is vertical, that there is exact match between the size of the pixel and the size of the front-end electronics channel and that the electronic chip is very close (tens of μm) to the sensor [7].

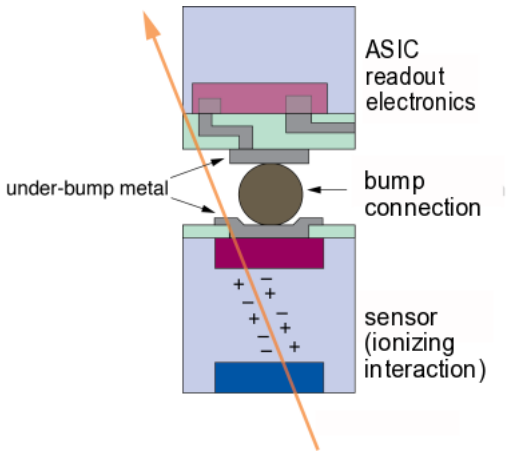

Figure 1.1: Basic building block (i.e. pixel) of a hybrid pixel detector: sensor and the readout electronics are separate and feature a bump connection [8].

Section 1.1.1 summarises the sensor functionality, while the Application-Specific Integrated Circuit (ASIC) and applications are presented in Section 1.1.2 and 1.1.3.

1.1.1 Silicon radiation sensor

At the state of the art, many different kinds of radiation sensors based on different materials have been developed (e.g. gas electron multipliers, silicon strips, pixels and drift detectors, CCDs, active pixel sensors, vacuum tube photomultipliers, avalanche photodiodes, etc.) [9]. In particular, planar silicon sensors are considered as they have been adopted for previous generation experiments (e.g. [10]) and constitute a valuable option for future detector upgrades. Nevertheless, it can be at the same time highlighted that other materials (e.g. diamond) and technologies (e.g 3D) are also being evaluated within the HEP community [11]. The geometry of the cross-section of a single-sided p-in-n silicon sensor, is shown in Figure 1.2 as a basic example: a large

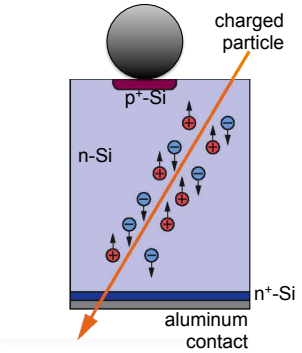

Figure 1.2: Cross section of a single-sided p-in-n silicon sensor, with n-bulk and p+ implant.

area p^+ implantation is placed in a n-bulk and a positive bias is applied to the back side through a ohmic n^+ contact and metal layer. The electric field in the generated depletion zone allows the collection of the signal charge (electron and holes) liberated by ionizing particles. The sensor acts therefore as a reversed-biased p-n junction. The collected charge is fed to the analog front-end in the ASIC readout chip through the bump-bond connection, i.e. by DC-coupling. P^+-in-n sensors have been extensively used for their simplicity,

above all for applications where radiation damage is not too significant [7]. Other planar silicon sensors topologies have been also implemented and optimised for parameters such as maximum charge collection, spatial resolution, radiation harness. To provide an example, the current CMS barrel pixel detector features a so-called double-sided n-in-n approach [10] and implements special layout inter-pixel isolation techniques, which assure better radiation tolerance and spatial resolution. It is anyway not the purpose of this work to describe in details all possible sensor topologies and layouts, but it can be mentioned that for CMS upgrade a different topology (i.e. a single-sided n-in-p approach [11]) is currently identified as the silicon planar sensor candidate.

1.1.2 ASIC Readout chip

Pixel detectors readout chips feature different geometries, readout approaches and analog devices, but the main properties and the hierarchy are common to most of them. They are indeed composed of an active area which contains a repetitive matrix of elementary pixels directly interfacing the sensor and of a chip periphery, in charge of global control, data buffering and readout and global configuration. The described hierarchy is shown in Figure 1.3.

The active area, also referred to as pixel matrix or pixel array, is composed of elementary electronic units called Pixel Unit Cells (PUCs). In the small pixel size, a PUC integrates an analog front-end, required to perform analog-to-digital conversion of the charge collected in the sensor, and digital processing, possibly including data storage. A basic block diagram is reported in Figure 1.4, where the interface between analog and digital logic is highlighted. Its components are described more in detail in Section 1.1.2.1 and 1.1.2.2.

1.1.2.1 The analog front-end

An analog front end is usually implemented with a cascade of a few amplifying stages. The first stage is the preamplifier, while the following ones are band-limited and determine the frequency spectrum of the output pulse and its shape, forming the filter or pulse shaper. This filtering is required as detector signals are very fast and their shape cannot be preserved with limited band-

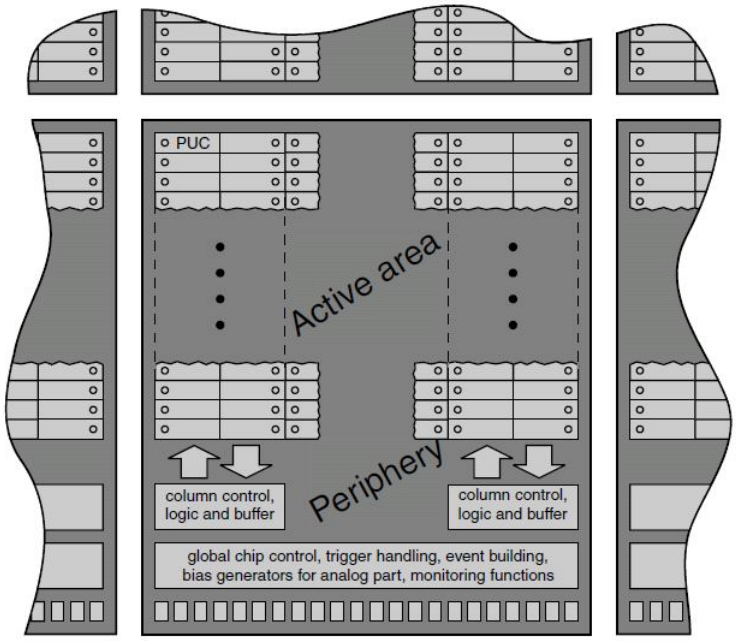

Figure 1.3: Generic pixel detector: active area and periphery circuitry [7].

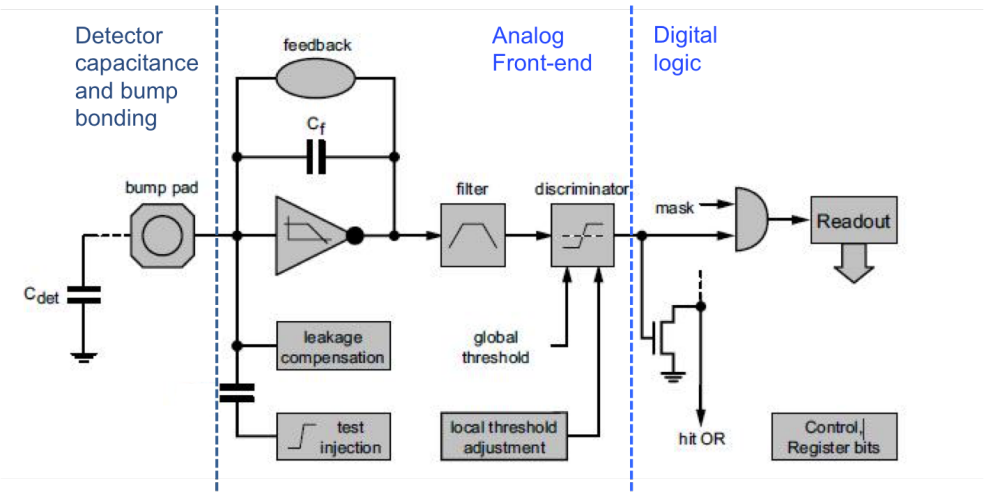

Figure 1.4: Block diagram of a generic PUC [7].

width and power consumption. Different front-end architectures are available in literature and they have been used for different applications and requirements [9]. The most relevant front-end architecture is the one whose generic scheme, without implementation details, is showed in Figure 1.4. The output of the analog circuitry is fed to a discriminator, which outputs a digital signal. This can be either considered a binary hit, as done in a few cases in literature (e.g. [12], [13]), or a further amplitude measurement can be performed. In the analog front end, an inverting amplifier with feedback capacitance converts the input charge to a voltage. The preamplifier is a crucial part of the circuit and it is designed taking into account many metrics (e.g gain, bandwidth, power, noise, etc.). In circuits where the preamplifier output is directly interfaced with the discriminator (without a separate filter), the discharge must be completed before the next signal arrives, to avoid overlap. On the other hand, a fast discharge can lead to a reduction of the peak amplitude if it starts before the signal has reached its peak. This concept can be noticed in Figure 1.5 (a), where discharges with different feedback time constants are shown. Analog to digital conversion of the collected charge is performed through a Time over Threshold (ToT) measurement, where the ToT is the number of clock cycles during which the signal is higher than the discriminator threshold. Ideally the pulse width is supposed to be proportional to the input charge. Digitization into a defined number of bits can be done with simple approaches, i.e. using a clock signal and a digital counter for each channel. Alternatively a clock counter can be centralised and its output is latched into local registers when the leading and trailing edge transitions are detected by the single channels and the ToT is obtained by difference. Either way, the time constant of the discharge is defined based on the digital ToT counting speed capabilities (dependent on the clock period). Aiming to obtain a linear behaviour, a constant current discharge is normally adopted to extend the preamplifier output pulse. The concept is shown in Figure 1.5 (b) for multiple charge amplitudes.

Dead-time is introduced in the measurement due to the limited ToT counting speed. Dead-time is an important metric, since it is source of hit losses that can become severe when high hit rates are encountered. In the field of radiation detection [14], dead-time is usually modelled either by a paralyzable

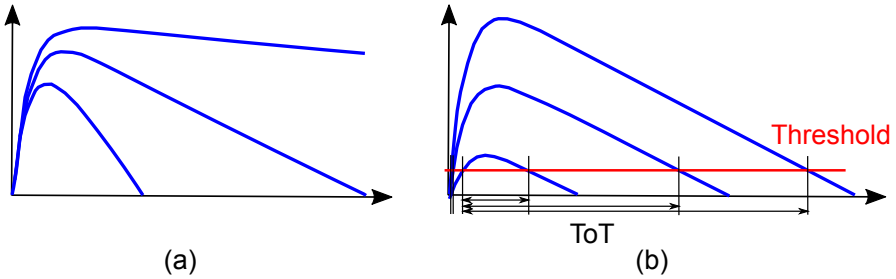

Figure 1.5: Preamplifier signals (amplitude vs time) obtained with constant current feedback. A variation in the feedback time constant is shown (a). ToT measurments for different input charges with fixed time constant are displayed (b).

or non-paralyzable system. In the first case, events occurring during the dead period are not recorded, but still extend the dead-time duration. In the second case, events received during the dead-time are lost and do not have an effect on the detector behaviour. The first model resembles more the behaviour of the system of interest, since additional charge extends the ToT pulse width. For a paralyzable system the distribution of intervals between random events occurring at an average rate n is $P_1(T)dT = ne^{-nT}dT$ [14]. The probability of intervals larger than τ is obtained by integration:

$$P_2(T) = \int_{\tau}^{\infty} P_1(T)dT = e^{-n\tau}. \tag{1.1}$$

The recorded input rate m corresponds to the true rate n multiplied by this factor. In terms of first-order losses the two dead-time models are equivalent and differ only for very high input rates. For $n << 1/\tau$, both non-paralyzable and paralyzable models can be approximated to:

$$m \cong n(1 - n\tau). \tag{1.2}$$

Even if τ is not a fixed value for the system of interest (as it varies with the input charge) the average dead-time can be used as a reference to estimate correspondent losses. An example related to the target application is provided, considering a pixel as the paralyzable system, a 75 kHz input rate and a bunch crossing period of 25 ns. If an average ToT=4 is assumed, this leads to dead-time losses of

$$\frac{n-m}{n} = n\tau = 75\,kHz \cdot 100\,ns = 0.75\%. \tag{1.3}$$

In general, the amount of tolerable dead-time depends on the particular application. Requirements related to this work will be discussed in Section 1.2.1.

Besides dead-time, other key parameters of analog front-end designs are [9]:

- peaking time, the time required for the signal to swing from the baseline to the peak: it has to be fast enough for the signal to go above threshold in the right cycle;
- gain, ratio between the peak of the output voltage and the input charge;
- noise, due to intrinsic disturbances generated within the sensor and the front-end amplifier;
- time resolution, with an accuracy that strongly depends on the application;
- power consumption is also a central metric and constitutes a trade-off with speed and analog performance.

1.1.2.2 Main concepts on digital readout architectures

The readout architecture of the ASICs of HPDs depends very much on the target application. Position, time and possibly the corresponding pulse amplitude, of all hits belonging to an interaction must usually be provided in HEP. The choice of a suited architecture depends on a number of factors, e.g. on the available chip technology and on the acceptable hit losses.

General aspects of digital readout architectures will be introduced in the following paragraphs, while a detailed evaluation of architectures suitable for this work will be presented in Chapter 2.

Architectures with zero suppression In hybrid pixel detector ASICs, digital architectures typically process only pixels with amplitudes above a threshold, in order to reduce the size of buffers which are often required to store data for a certain amount of time. This readout approach, where only a reduced number of pixels of the full pixel matrix are processed, is referred to as zero suppression [7]. Ideally, the aim of any architecture is to read out exclusively non-zero hits to optimise the use of local buffering and readout bandwidth.

Trigger-less and triggered architectures Experiments with low input rates can afford to read out all the impinging pixel hits immediately after the interaction. If higher rates are involved, on-detector data reduction is needed in order to obtain a feasible data rate towards the Data Acquisition System (DAQ). For this reason, usually a trigger signal is used for selection of hits of interest. The generation of the trigger signal is based on the analysis of many sub-detectors of the experiment. This analysis has to happen within a fixed latency after the particle has been detected, for the trigger to be correctly produced. Storage logic is required to maintain the data until the trigger latency has expired. Trigger latency is currently in the order of around 100 cycles (correspondent to bunch crossing interactions) and will be incremented in the future detectors, see Section 1.2). Moreover, depending on the trigger rate and possible bursts of consecutive triggers, chips must be capable of accepting new triggers before data of the previous ones are fully read out. Some relevant references of trigger-less architectures, each of one supporting different input rates, are the following: Timepix [15], CLICpix demonstrator [16], ToPix [17], Timepix3 [18] and Velopix [19]. The details of these architectures will not be reported, since the focus of this thesis is on triggered readout architectures for very high hit rates.

As far as triggered architectures are concerned, data buffering can be implemented with different approaches and storage elements can be located in different parts of the pixel chip (End of Column (EOC), single PUC, region of certain number of PUCs), implementing different readout schemes. Moreover, limited buffering constrained by the area available, can be a substantial source of hit loss unless this issue is properly addressed at design time. For the target application this point will be analysed in Chapter 2. An overview of state of the art triggered readout architectures [20] and their evolution over the years is herein summarised. In the so-called “timer architecture" [21], an analog timer delay is used to perform a coincidence with the trigger signal, identifying the pixels to be read out. Readout architectures have evolved to a digital implementation of the the trigger matching mechanism. In particular a “digital delay”, in form of a timestamp counter has been used with different implementation approaches. For many relevant architectures, such counters are located

in the EOC. In particular, the initial development of the Front End-A chip of the ATLAS group features counters in the EOC and implements a “conveyor belt architecture", since pixel hits are transported uniformly to the periphery: each clock cycle hit addresses are moved from the pixels to the EOC where they are assigned a timestamp counter, counting the remaining clock cycles to reach the latency. In order to avoid hit data loss during the trigger latency without increasing pixel area with local storage a “column drain architecture" has been similarly used by the CMS pixel chip [22] and the ATLAS FEI-3 [23]. The peculiarity of this approach is that buffering during the trigger latency is performed in the column periphery, whereas in the PUC only one hit at the time is stored. To this end, pixel data are moved to the EOC as quickly as possible. This reduces the pixel dead-time and makes the latency loss only dependent on the EOC buffering. The main difference between the CMS and the ATLAS approach is in the way the timestamp is made available to the pixels. Association of the timestamp to the pixels is performed with a pointer mechanism or with distribution of the bus to the whole chip. An alternative approach, i.e. with a 2-deep buffer and trigger matching logic in the pixel, was used by the ALICE chip [24]. In this case both the timestamp bus and the trigger are distributed to all the pixels: the timestamp is expressed as a particular 8-bit up-down counting time pattern which is stored locally and later compared to the time pattern itself to assess the expiration of the trigger latency.

The second generation of ATLAS [25] and CMS [26] pixel chips also implement a triggered readout, but with different schemes. The CMS PSI46DIG pixel chip keeps the architecture concept almost unaltered, but changes the readout implementation from analog to digital and increases the EOC buffering in order to cope with increased input and trigger rates. ATLAS FE-I4 has successfully explored the possibility of introducing a regional readout architecture, which combines digital processing and triggerring logic from every group of 4 pixels into one synthesized logic block. Placing most digital processing within the pixel matrix makes it possible to sustain higher hit rates while reducing digital power, because most hits are held within their respective region until the trigger latency expires, and then erased, with no need for high data

bandwidth between pixels and periphery. Moreover, this chip has profited from a more scaled technology (130 nm vs the previous generation 250 nm), which has allowed the required storage resources to fit in the pixel array. The drawback of local digital processing is the need to distribute clock and trigger signals throughout the pixel matrix, with potential digital noise injection into the front ends [27].

1.1.3 Applications

The main applications of pixel detectors can be found in particle physics but their use has spread in a variety of other fields related to imaging. An brief description of both fields is provided in this section.

HEP applications can be found in the context of the Physics Program of the experiment at the LHC, which is aimed at answering fundamental questions in particle physics (e.g. the origin of elementary particle masses, nature of the dark matter, fundamental forces, difference between matter and antimatter, etc.). To this purpose, protons are accelerated up to 7 TeV (design value) and circulate in a 27 km-long accelerator vacuum pipe 100 m underground: one beam of protons rotates clockwise, and the other beam counterclockwise in separate but close orbits and they can be forced to collide in specific regions around which the experiments are located [7]. Collisions cause the so called events, i.e. fundamental interaction between subatomic particles, occurring in a very short time span, at a well-localized region of space. Therefore, individual charged particles, usually triggered by other subdetectors, have to be identified with high demands on spatial resolution and timing. Most of the mentioned LHC-collider-detectors at CERN, i.e. ALICE, ATLAS, CMS, LHCb, as well as fixed target experiments (e.g. NA62 [28]) employ the hybrid pixel technique to build pixel detectors covering large scale surfaces ($\sim$ few m^2). The detectors are normally arranged in cylindrical barrels layers and disks, as shown in the example from the CMS detector in Figure 1.6.

The main purpose that these detectors must serve is particle tracking in order to allow *i)* identification of short lived particles, *ii)* pattern recognition and event reconstruction, *iii)* momentum measurement [30]. An example of

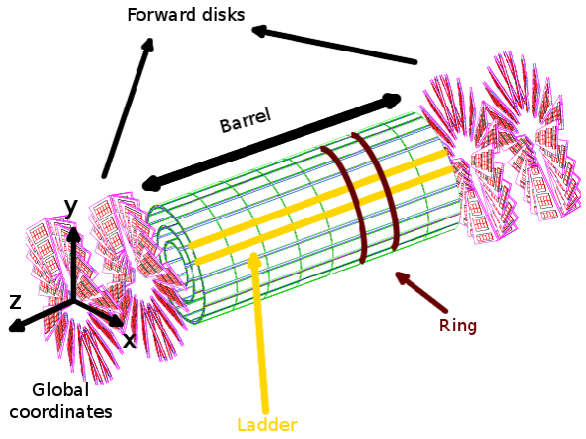

Figure 1.6: Three-dimensional view of the CMS pixel layout [29].

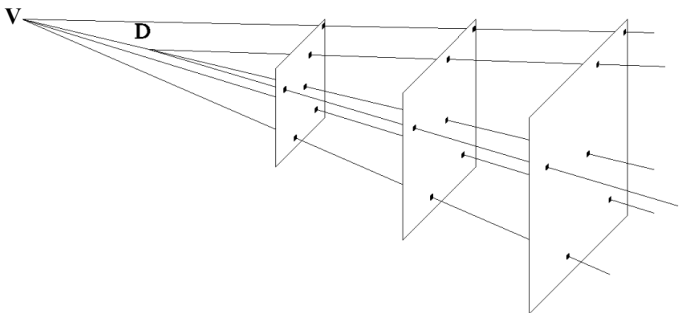

Figure 1.7: Tracking example of a decay topology with collision vertex V and decay vertex D. Tracks are measured by three pixel detectors and detected hit pixels are highlighted [7].

a decay of a short-lived particle is shown in Figure 1.7. It is required that tracks emerging from the fast decay are measured as close as possible to the interaction point. Time and spatial resolution are therefore important for such an application, as well as good granularity, which helps to distinguish the track of interest from many others which may confuse the picture (above all in a high rate context). Figure 1.7 shows this concept by using three pixel detectors to reconstruct the desired particle track, by finding positions of charged particles at a number of key points and therefore recording their paths. Pixel detectors for particle tracking are very demanding since they need to reconstruct a huge number of charge particle trajectories in three dimensions, whereas the time

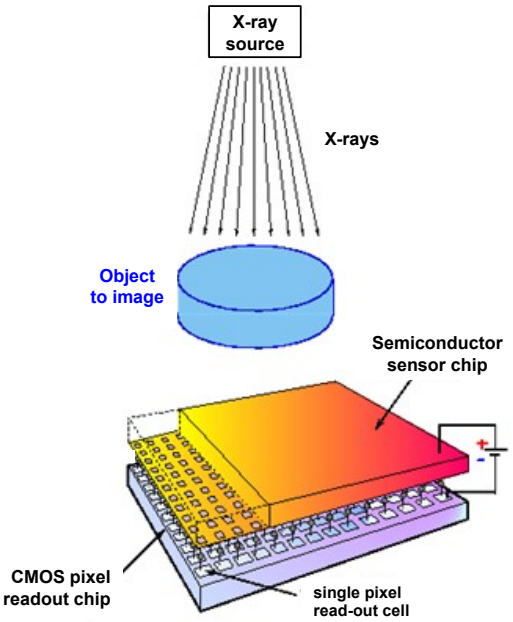

Figure 1.8: Hybrid pixel detector application for X-ray radiation imaging [31].

between two beam crossings is only a fraction of a microsecond. The most critical part of these detectors consists in the inner layers that are usually referred to as the “vertex detector" or “Inner tracker". In this context, which is related to the subject of this thesis, the use of hybrid “pixelated" detectors, is necessary.

Although these devices were developed for high-energy ionizing particles and radiation beyond visible light, they have been also adopted in many other areas. In particular, radiation imaging has become one fundamental target for hybrid pixel detectors. The basic detection mechanism is shown in Figure 1.8: the object to be imaged is placed between the X-ray source and the detector and a certain amount of X-ray is absorbed by the object (depending on its density and composition). The X-rays that pass through the object are captured by the detector, which is capable of reconstructing the image and possibly also determine properties of the material. X-ray imaging performed by hybrid pixel X-ray cameras has represented an advancement with respect to

usual CCD or CMOS cameras and numerous applications have been opened in material sciences (crystallography), non-destructive control, biomedical imaging and clinical imaging leading to a growing industrialization [32]. Moreover, neutron transmission radiography has also shown to be a valuable application for structures which are hardly distinguishable with X-ray radiography, thanks to the different attenuation factors in the two cases [33]. The interest shown from these communities has pushed experts to develop hybrid pixel circuits dedicated imaging applications, such as X-ray detection. The Medipix collaboration at CERN has been one outstanding example of this effort and it has delivered a whole family of chips [15], [34], [35]. These ASICs will not be described in further details, as the main application of this thesis is particle tracking for next generation high energy physics experiments.

1.2 Phase 2 upgrade and requirements

A description of the pixel detector upgrade conditions and quantitative requirements is herein provided [36], as it defines the specifications for the readout chip subject of this work [37]. Within the LHC, protons acceleration and control of their trajectories is achieved by grouping them into bunches which cross each other with constant frequency. An important parameter in accelerator experiments is the number of events that one can expect for a particular reaction. For fixed-target experiments the interaction rate ϕ depends on the rate of beam particles n hitting the target, the cross section for the reaction under study, σ, and the target thickness d (in cm) according to where σ is the cross section per nucleon, N_A Avogadro's number and ϱ the density of the target material (in g/cm^3) [38]:

$$\phi = \sigma \cdot N_A \ [mol^{-1}]/g \cdot \varrho \cdot n \cdot d \ . \tag{1.4}$$

Equation 1.4 can be written by defining a reference quantity L called luminosity

$$\phi = \sigma \cdot L \, , \tag{1.5}$$

which can be seen as the interaction rate for unitary cross section. In collider experiments luminosity definition adds a level of complexity as it is a combination of two particle beams which are one the target of the other. It is out

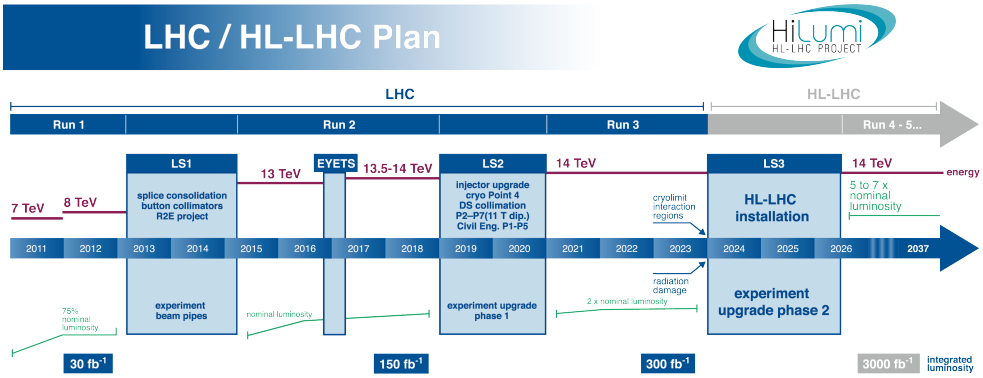

Figure 1.9: Plan for the LHC in the next 10 years [42].

of the scope of this work to give a detailed physics explanation, which can be found in [38] and [39].

In the first major physics run (Run 1) in 2011 and 2012, the collider reached a peak luminosity of 7.7 x 10^{33} cm $^{-2}$ s $^{-1}$. The pile-up, i.e the number of interactions per bunch crossing, has reached a peak of almost 40 in 2012 [40]. Each of its two general purpose experiments ATLAS and CMS, have acquired and processed a huge amount of data which has yielded a vast quantity of physics results [41]. Nevertheless, many physics studies and research are needed to expand the physics potential of the LHC, in particular for rare and statistically limited standard model (SM) and beyond standard model (BSM) processes. Major revisions to the machine and the experiments are therefore necessary and a series of long periods of data-taking (referred to as Run 1, Run 2, etc.) interleaved with Long Shutdowns, designated LS1 (2013-2014), LS2 (2019-2020), LS3 (2024-2025), have been planned, as shown in Figure 1.9.

Run 2 is ongoing at the time of writing, reaching pile-up peaks of 70-80 [43], whereas this thesis is part of the effort for developing electronics systems for the LS3, also referred to as High Luminosity LHC or Phase 2 upgrade. The proposed operating scenario is to level the instantaneous luminosity at 5 x 10^{34} cm $^{-2}$ s $^{-1}$ for a further 10 years of operation, with potential peaks of 2 x 10^{35} cm^{-2} s^{-1} [41]. The foreseen pile-up for the ATLAS and CMS experiments is much higher than previously, i.e. 200. The HL-LHC is expected to run at

a centre-of-mass energy of 14 TeV and with a bunch spacing of 25 ns i.e. at 40 MHz frequency. As a general remark concerning the upgrade of hybrid pixel detectors for the Inner Tracker, closest to the interaction point, the increase in radiation levels requires improved radiation hardness, while the larger particle density requires higher detector granularity, increased bandwidth to accommodate higher data rates, and improved trigger capability to keep the trigger rate at an acceptable level while not compromising physics potential. ATLAS and CMS are carrying out a common development in the framework of RD53 collaboration [4] to develop a pixel readout integrated chip in 65 nm CMOS technology for the for extreme rate and radiation. Aforementioned operating conditions have an impact on the requirements of the sensor and readout electronics: the latter, subject of this work, will be described in detail in 1.2.1.

1.2.1 Pixel chip requirements

Phase 2 upgrade operating conditions, with high instantaneous luminosity and consequently high pile-up, contribute to define a set of requirements for the ASIC readout chip. The specifications of the pixel chip demonstrator object of this work are summarised in Table 1.1 for completeness. Requirements addressed by this work are described more in detail in Section 1.2.1.1.

1.2.1.1 Requirements addressed in this work

Hit rate and efficiency One of the most challenging design requirements for the readout electronics is to be capable of withstanding a hit rate of $3\,\mathrm{GHz}/cm^2$ with negligible losses (<1%). For a readout ASIC, the hit rate (R_H) indicates the flux of particle on a certain area on the active area of the sensor:

$$R_H = \frac{N_{hit\ pixels}}{T \cdot A_{chip}}\ , \tag{1.6}$$

where $N_{hit\ pixels}$ indicates the number of pixels hit in the area A_{chip} over the time period T. The high hit rate requirements poses challenges on guaranteeing the target hit efficiency for the overall pixel chip E_H, defined as the ratio

$$E_H = \frac{N_{readout\ pixels}}{N_{hit\ pixels}} \tag{1.7}$$

Table 1.1: Demonstrator pixel chip specifications

Technology	65 nm CMOS
Chip size	20x11.8 mm^2 ($\sim$half size of final chips)
Pixel size	50x50 μm^2, 25x100 μm^2
Detector capacitance	< 100 fF (200fF for edge pixels)
Detector leakage current	< 10 nA (20 nA for edge pixels)
Detection threshold	$< 600\, e^-$
In-time threshold	$< 1200\, e^-$
Hit rate	< 3 GHz/cm^2
Noise hit occupancy	$< 10^{-6}$
Charge resolution	4 bit ToT
Hit loss	$< 1\%$ at 3 GHz/cm^2
Trigger rate	≤ 1 MHz
Readout data rate	< 5.12 Gb/s
Radiation tolerance	500 Mrad (TID) 10^{16} n_{eq}/cm^2 at –15°C
SEU affecting whole chip	< 0.05/hr/chip at 1.5 GHz/cm^2 particle flux
Power consumption at max. hit/trigger rate	< 1 W/cm^2
Temperature range	–40°C to +40°C

between the number of pixel correctly read out at the output of the chip $N_{readout\ pixels}$ and the number of hit pixels $N_{hit\ pixels}$. Losses are composed of a combination of dead-time of the analog front-end and digital losses due to limited buffering for hit storage.

Small pixel size and large IC format One of the fundamental requirements for the design of the next generation Inner Tracker is the use of a smaller pitch compared to the present pixel detector, which featured a pixel size of 100×150 μm^2, for better resolution. As far as the sensor is concerned, thin silicon sensors (of thickness 100-150 μm), segmented into pixel sizes of 25×100 μm^2 or 50×50 μm^2, are expected to exhibit the required radiation tolerance and to deliver the desired performance in terms of detector resolution, occupancy, and separation of multiple tracks. Consequently the design of a readout chip with a small PUC size is required, which poses area density challenges. This requires to minimise the logic per pixel, which has to be accommodated in 2500 μm^2 including also the analog front-end. In addition, a large IC format is demanded in order to maximise the fill factor (i.e. maximise the active area and minimise edge effects when building the detector with those chips), which significanlty increases the design complexity (number of devices). Moreover, non-linear scaling of inteconnect parasitics complicate in-time distribution of global signals and power distribution.

Trigger rate and latency The motivation of the need for a trigger signal to select events of interest, instead of a full readout, has been introduced in 1.1.2.2. As far as the phase 2 upgrade is concerned, two different solutions are being considered: *i)* a higher trigger rate and longer latency, or *ii)* two different levels of triggers, with a level-0 trigger that will feature additional data reduction techniques. In the context of the RD53 collaboration the requirement on the trigger rate has been set to 1 MHz, whereas the target trigger latency is 12.5 μs. Due to the very high hit rate, hits need to be stored during the 12.5 μs trigger latency in the pixel array within a Pixel Region (PR) (made of multiple pixels, e.g. 2×2 or 4×4), in order to not saturate the bandwidth along the columns and to avoid the high power consumption which would be

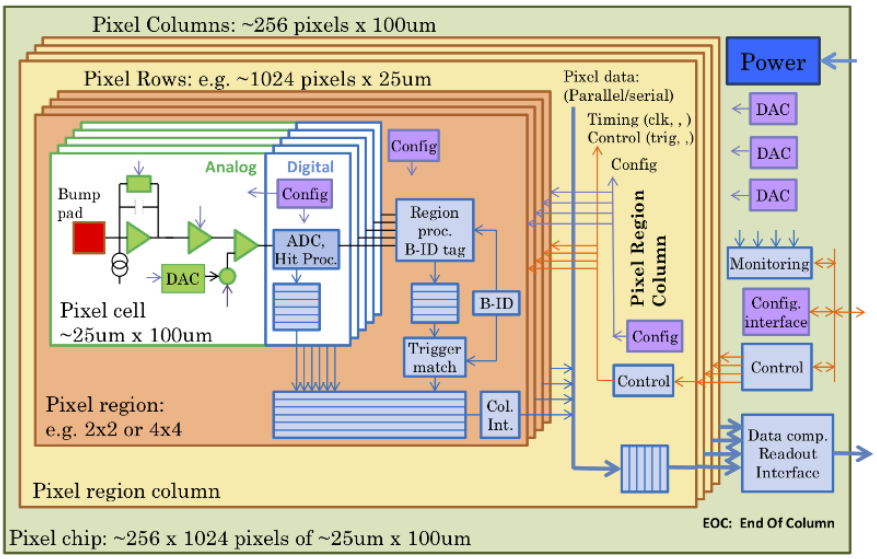

Figure 1.10: Diagram of the hierarchical organization in a 3rd generation pixel chip, showing how pixels are grouped in regions, regions in columns, and column pairs in a full matrix [4].

demanded by the continuous data transfer. The buffering required is direcly related to the specification on the hit efficiency and on the pixel size: the goal is to achieve sufficient efficiency by optimising buffering resources to be arranged in the limited pixel area. Triggered event data are then collected from the pixel array and buffered at different stages for derandomization. Finally appropriate data formatting is applied on-chip before sending readout data on a configurable number (1-4) of differential Electrical links (E-links) at 1.28 Gb/s. A block diagram of the buffering stages and final readout is shown in Figure 1.10.

Radiation hardness and technology choice A 65 nm CMOS technology has been chosen by CERN as an appropriate technology platform for high rate and high density applications for experiments, based on a technology evaluation among multiple foundries. In the the innermost regions of next generation Inner Trackers in CMS and ATLAS, the fluence, i.e. the number

dN of particles incident on a sphere of cross-sectional area da:

$$\phi = \frac{dN}{da}, \tag{1.8}$$

is foreseen to reach 1 MeV neutron equivalent fluence of the order of $10^{16}\, n_{eq}/cm^2$. The maximum value is mainly depending on the radius distance from the particle beam, while the variation along the z direction is very moderate. Cartesian detector coordinates have been shown in Figure 1.6 and are often equivalently described with a cylindrical coordinate system (r,Φ,z).

The Total Ionising Dose (TID) absorbed by the medium (the pixel chip, mostly made of silicon) is measured in Gray (Gy, i.e. $J \cdot kg^{-1}$) in the international system, while it is often expressed in rad ($1\,\text{rad} = 0.01\,\text{Gy}$) within the HEP community. This quantity is a measure of accumulating ionising effects which cause performance degradation as a device is exposed to ionising radiation. The chosen technology had been studied and was seen to have excellent radiation tolerance up to 100 Mrad [44] and RD53 has been evaluating the feasibility of its use for radiation levels of up to 1 Grad (corresponding to 10 years of operation for the inner part of the detector). The use of a high density 65 nm CMOS technology is also critical for the HL-LHC pixel detectors in order to have the required circuit density to implement the small pixels and to buffer hit information during the trigger latency. The studies performed by the RD53 collaboration have led to the conclusion that by dedicated design functionality up to 500Mrad can be guaranteed and this is therefore the target specification. Moreover, the specification in terms of Single Event Upset (SEU) is set to a maximum of 0.05 upsets (affecting the whole chip) per hour per chip. Although important for final phase 2 chips, low priority has been given to it for the development of RD53 prototype, since it is not considered a critical design aspect to be demonstrated.

Power consumption and serial powering scheme Phase 2 upgrade pixel chip will have to be in a large IC format featuring low power consumption, in order to instrument large areas of the detector keeping material interfering with the particles as low as possible. Current generation pixel chips consume power in the order of $0.3\,\text{W/cm}^2$, which already demands challenging cooling

and power distribution. For next generation detectors, a CO_2 cooling system with better cooling performance is assumed. This allows a target power density of $1\,\mathrm{W/cm^2}$, important to allow the increased specifications in terms of rates and radiation hardness. In particular, in the context of RD53, the power consumption specification is given as $4\,\mu\mathrm{A/pixel}$ for the analog and $< 4\,\mu\mathrm{A/pixel}$ for digital [45].

The optimisation of the analog pixel front-end, targeted to achieve lowest possible power consumption with acceptable noise and discriminator thresholds, it is not covered by this work. The focus will instead be on digital power consumption, which must be optimised for both static consumption (e.g. leakage currents) and dynamic consumption (e.g. switching nodes and parasitics) and requires an appropriate design strategy and methodology.

As far as the powering scheme is concerned, delivering power to a detector with multiple modules (composed of a certain number of chips) is increasingly challenging due to the low supply voltages (1.2 V) of the modern technology adopted, requiring problematic high currents to deliver a given power. The use of a classical passive parallel powering system and local DC-DC power conversion are both excluded. The first because of the high currents and the second due to the radiation environment, magnetic field and tight constraints on space and material budget. A serial power distribution system is considered to be the only viable solution to supply the Inner Tracker with the required power, within an acceptable material budget and power cable losses. In a serial powering scheme, the current consumption is fixed and is required to provide sufficient power and some additional headroom current for fluctuations, as it will be discussed more in detail is Chapter 4. In this context, the overall ASIC power density specification is set to $1\,\mathrm{W/cm^2}$ including losses of the on-chip power regulators.

Chapter 2

Development and optimisation of a SystemVerilog framework for the architectural study, simulation and verification of the readout electronics

A system-level design framework based on SystemVerilog (SV) and the standard Universal Verification Methodology (UVM) [46] is a valuable tool to handle system complexity, evaluate multiple system architectures and achieve design optimisation through the concurrent contribution of multiple designers. A first version of a SV-UVM simulation and verification framework has been implemented and described in [47], where an initial study on buffering architectures modelled at behavioural level was presented. In this thesis, such a platform has been optimised and extensively used in order to meet fundamental requirements for the specific application, i.e. the design and verification of the large scale RD53A prototype, as well as for two small scale demonstrator chips, CHIPIX65 and FE65-P2. In order to allow these different projects to share the same framework, the work has been aimed to achieve high modularity

and re-usability, to handle complexity and better support integration of architectures being investigated by different designers at various abstraction levels. This has also allowed the framework to be partially re-used for the simulation and verification environment presented in [48]. The state of the art of system-level simulation methodologies and motivations are discussed in Section 2.1, whereas Section 2.2 describes the structure of the developed framework and its optimisation for modularity, flexibility and re-usability. Section 2.2.2 also provides details on which hardware description levels are supported and which have been used in this work.

2.1 State of the art and motivations

The persistent push to shrinking process node continues in industry and in scientific research, with the primary goals of reducing area and thus the cost and improving performances. For the last forty years, the decreased transistor and wire sizes also brought increased speed and reduced power consumption, but those benefits have declined as the devices approach the atomic limits [49]. The target gain is accompanied by challenges of current leakage, power management, timing predictability and production yield. Therefore, designing chips in 65 nm processes requires more planning, extra analysis and complex trade-offs: all aspects contributing to success need to be incorporated into the design from the start. Such a trend raises new design challenges, therefore the need for faster integration of complex systems, high-level system design and extensive functional verification are becoming mandatory [50]. Since verification is the most time-consuming part of the design, requiring a significant engineering effort, industry CAD tools, languages and methodologies are evolving towards higher-level and class-based testbenches capable of addressing design complexity and meeting time-to-market needs [51]. SV is a combined Hardware Description and Verification Language (HDVL), based on extensions to Verilog, created with the aim of fully supporting system-level design and verification [52]. It offers both enhancements for the description of the Device Under Test (DUT) at multiple levels of abstraction, from gate-level to Transaction Level (TL), and advanced verification features, based on Object-Oriented

Programming (OOP) techniques and high-level communication through transactions. An extensive description of SV features for design and verification can be found in [53] and [54], respectively. SV is being used extensively in the industry and also in the academics for different purposes. In particular, some examples available in literature are related to early TL system description [55], mixed-signal system validation [56], [57], [58], SoC verification for a variety of applications (e.g. image signal processing [59], memory controllers [60], [61]). Furthermore, interest is shown in using UVM verification frameworks to simulate SystemC IP models [62], also in mixed-signal automotive use cases [63]. Such a widespread adoption of UVM has also been possible thanks to the availability of a rich class library provided by a reference verification methodology, known as the UVM [46], [64].

Such advanced industry design tools are also being considered in research wherever complexity is a relevant issue. The High Energy Physics (HEP) community has also turned to high-level design, simulation and verification techniques: description language such as Simulink [65], C++ [66], SystemC [67] and SV are being used for different applications in addition to standard Register Transfer Level (RTL) simulation with classical testbenches. In [68], Simulink has been adopted, but it is considered more suited for algorithmic design since it does not guarantee a high level of granularity in modelling and clock cycle accuracy in simulation. In order to evaluate system functionality and estimate data losses, C++ architecture modelling and simulation has been adopted in [69] and [70]. Since C++ only provides with cycle-based communication and un-timed computation, in both cases, the model had to be at a second stage translated into an RTL description (i.e. Verilog and VHDL) in order to study further details of the architectures. SystemC is an expansion of the ANSI Standard C++ with a C++ class library needed for system-level and HDL modelling. A SystemC simulation environment has been defined in [71] to guide the evaluation of the performance of a new protocol at TL with clock-cycle accuracy. With respect to [71], where a separate VHDL testbench was needed to verify the synthesizable RTL description of the design, SV and verification methodologies allow the designer to re-use the same environment as the design progresses from TL to detailed gate level description.

As summarised in Table 2.1 , SystemC and SV languages address the needs of specific constituents in the system-design process.

The first supports software compatibility and it is an ideal candidate to improve the design process of the software/hardware interface at TL [72]. On the other hand, transactions can also be implemented at a cycle-true and bit-accurate signal level and High Level Synthesis (HLS) tools are capable of synthesizing a subset of SystemC constructs. The second can be considered a bottom-up oriented approach: it has roots in HDLs, it provides full compatibility with Verilog (containing all the features necessary for a complete path to implementation including synthesis and simulation with back-annotation), but it also offers hardware description abstraction and additional verification capabilities. In a research environment with a limited number of experts, who are taking care of all the steps of the design process, the use of a unique language can be considered an added value to reduce complexity. In this context, SystemVerilog was considered a more solid solution for the community since the RD53 collaboration was started [4], whereas the performance of HLS tools was not sufficiently explored and studied to rely on their use for fine logic optimisation.

Table 2.1: SystemC and SystemVerilog complementary design capabilities and support of emerging methods including TLM and assertion-based verification (ABV) [73].

	SystemC	**SystemVerilog**
Core abstraction level	*Events and messages*	*HW implementation view*
Architectural design	System-level hardware view and SW programmer's view	Logic states and transitions DPI link to C/C++/SystemC
Architectural verification and HW/SW co-verification	Cycle accurate TLM@ > 10,000 cps	Timing accurate RTL @ 1-10 cps; TLM capability; C-like externsions
RTL-to-gates design	High Level Synthesis	Logic synthesis
RTL-to-gates verification	TLM/RTL co-simulation	Implementation testbench including ABV and functional coverage

Concerning SV, it has already been adopted in HEP for the design of integrated circuits for the readout of hybrid pixel detectors such as FE-I4 [74], where Open Verification Methodology (OVM) methodology was used for chip final verification. Moreover, for the Timepix3 and Velopix [19] designs, ar-

chitecture modelling was both performed with TL modelling, achieving higher simulation speed, and with synthesizable RTL, closer to the details of the hardware implementation. In addition, SV and UVM have been also used in the community for pure verification goals of chips in their final design stage, as presented in [75]. It should be underlined that such examples in literature were mainly targeted to be used by small groups for very specific simulation or verification goals, whereas the proposed approach is aiming to a higher level of flexibility, generality and modularity of the environment. It is indeed meant to perform extensive architecture evaluation and verification in a world-wide spread community of designers working on the RD53A prototype chip, as it has been highlighted in [76]. Furthermore, the herein presented framework and methodology has been also made available to the community and has provided a starting point for the development of the system-level simulation framework for multiple front-end readout ASICs verification with performance evaluation described in [48].

2.2 The VEPIX53 environment

In this section the VEPIX53 environment will be described. VEPIX53 stands for Verification Environment for PIXel chips developed in the framework of the RD53 Collaboration. The fundamental goal of this SV-UVM framework is to guide the design of next generation pixel chips at different steps of the design flow, from initial global architectural studies to extensive simulation and verification of the final design. The main requirements of such a platform are therefore the following:

- flexible generation of input stimulus data, coming both from external full detector or sensor simulations and generated within the framework itself (with given constrained random distributions which enable designers to test alternative and extreme cases);
- simulation of DUTs or sub-blocks described at different abstraction levels;
- automated verification i.e. capability of predicting expected chip outputs

(depending on randomly generated inputs), verifying conformity with actual outputs, report messages on matches, mismatches, errors/warnings, and collect statistics on performance (addressing also needs of post-production testing).

IC designers from various experiments and institutes, also building different Intellectual Properties (IPs), will contribute to the evaluation and simulating alternative pixel chip architectures. Therefore, it is important to allow them to benefit from a single simulation framework for performing global architecture optimisation. The use of advanced UVM OOP features has been essential to achieve high-customisability of the framework and to enable the simulation of the DUT at various description levels as clarified in Section 2.2.1. A dedicated project organisation structure, described in Section 2.2.2, has also been defined in order to assure the stability of the core of the environment but still provide a flexibility for the specific needs of different DUT architectures.

2.2.1 Universal verification methodology components, testbench and tests

A characterising feature of UVM environments is the presence of a testbench class (derived from the standard *uvm_env* class), which can be seen as a container object that instantiates all the reusable verification components and defines their configuration. A recommended approach in UVM is to explore alternative possible scenarios by using different test classes derived from the standard *uvm_test* class, without changing the testbench, which remains unique. Multiple tests can instantiate the testbench and determine the nature of traffic to generate and send to the DUT [46] as shown in Figure 2.1. A more challenging goal needed for this particular application is enabling designers to perform several tests of alternative DUTs re-using the same top level environment (i.e. the testbench). Therefore, the identification of common interfaces and verification components is mandatory. The basic features of the VEPIX53 framework are below summarised:

- four generic input and output interfaces to the pixel chips have been

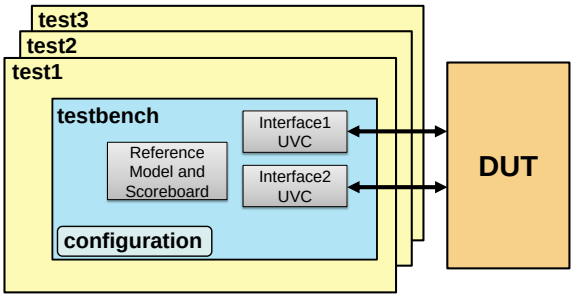

Figure 2.1: Hierarchical Layers of a UVM testbench: reuse of the same testbench for different tests.

identified (i.e. hit, trigger, output data and pixel array analysis), but also additional project-specific interfaces could be added;

- each of the previously listed interfaces communicates with a specific Universal Verification Component (UVC), or so-called environment, that wraps all classes devoted to it;
- each UVC uses a transaction object to represent data coming from or going to the corresponding interface, the format of which has been defined;
- UVCs connected to input interfaces not only provide stimuli to the DUT, but also monitor them and send corresponding transactions to the reference model;
- output interface monitors DUT outputs and the corresponding UVC converts them in an according transaction format;
- additional UVCs are defined to predict the expected behaviour of the chip and perform automated verification.

The main UVCs of the framework are herein described in further details and shown in the block diagram in Figure 2.2:

- the *hit* UVC, associated to the hit interface, has the main function of generating the charge signals associated to particles crossing the detector,

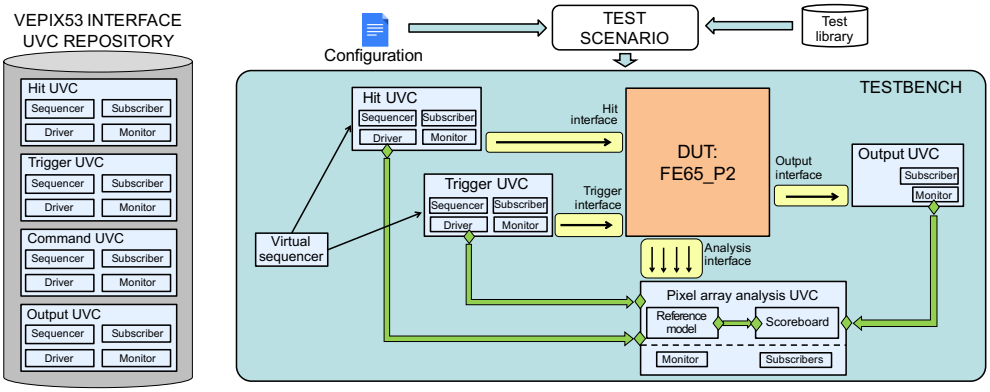

Figure 2.2: Block diagram of the VEPIX53 simulation and verification environment, highlighting a set of the developed UVCs [77].

and injecting them into the pixel matrix. More details on this component will be given in Section 2.2.3;

- the *trigger* UVC, associated to the trigger interface, is in charge of generating the external trigger signal of the pixel array according to configurable trigger rate and latency;

- the *virtual sequencer* controls the coordinated generation of hit and trigger transactions;

- the *output* UVC, associated to the pixel array output interface, takes care of producing data transactions by monitoring the data at the output of the pixel array;

- the *pixel array analysis* UVC is conveniently defined for containing different components. The *reference model* predicts the pixel array output according to the monitored hit and trigger transactions (it is, in practice, a transaction level description of the pixel array used as a golden reference for the DUT). The *scoreboard* checks for the conformity between predicted and actual output. Monitor and subscribers are associated to an analysis interface, which contains pixel array internal signals, and monitor the status in order to collect statistics on performance;

- for the functional verification of the pixel chip, two more UVCs are defined: the *command* UVC, which is in charge of generating the commands of the chip (e.g. calibration pulses, read and write registers) in agreement with a dedicated serial input protocol, and the *Aurora* UVC, which monitors data transactions at the actual pixel chip output, encoded based on the Xilinx Aurora protocol [78].

Most UVCs, made of a set of modular classes, are essential to guarantee the functionality of the verification environment. Nevertheless, different designers could need to modify some of them to make the environment more compliant with specific simulation and verification goals. For example, a designer could need to use a more detailed description of the reference model, taking into account well-known and accepted sources of error or to model custom functionality. Furthermore, it is not excluded for a user of the framework to incorporate completely new verification classes in the existent testbench. Such issues can be addressed thanks to the use of UVM OOP features like the configuration database, factory registration and class overrides. Configuration is easily achieved by defining a class that contains all the parameters for a given component. The configuration object parameters for a specific test are defined by calling the *uvm_config_db #(T)::set* method. Then UVM components that need to use a certain configuration object can access it by calling the *uvm_config_db #(T)::get* method and store its parameters to a local configuration object of the same type. This approach has extensively been used in the design framework to configure stimuli generation, monitoring and automated verification. As regards the factory, the recommended UVM methodology dictates that engineers should never construct components and transactions directly using the basic *new()* class constructor, but should make calls to a special look-up table (i.e. the factory) to create and register components and transaction types. With the factory registration it is possible to control overrides in top-level tests in order to substitute component or transaction types at run-time, before building the entire testbench environment. This is possible thanks to the so-called factory overrides. In the adopted approach, due to the high level of flexibility required, we define first basic classes for all

```
factory.set_type_override_by_type
(analysis_env::get_type(),
analysis_custom_env::get_type(),"*");
factory.set_type_override_by_type
(analysis_ref_model::get_type(),
analysis_custom_ref_model::get_type(),"*");
```

Figure 2.3: Example code: factory override of the basic reference model and analysis environment with custom ones [76].

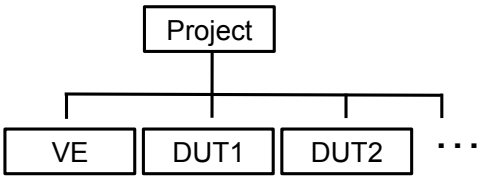

Figure 2.4: Top level project directory organisation [76].

the required components, then more detailed classes when needed for specific DUTs and at the end we override such classes in the DUT-specific UVM related tests. An example of such a class override is reported in the code in Figure 2.3.

2.2.2 Project organisation for reusability and modularity

The definition of the directory structure has been driven by the need to keep all classes, which are shared among different users, separate from the ones eventually overridden for specific needs. For this reason, the top-level of the project has been organised as displayed in Figure 2.4, with a unique folder (VE, Verification Environment) for the common files and separate ones for each DUT/architecture, which can be progressively added. In the first, one can find the subdirectories shown in Figure 2.5: it can be noticed how one folder has been dedicated to each SV interface while another one is used for top level classes. Each DUT-specific folder is organised in five subdirectories, as reported in Figure 2.6.

The define subdirectory contains files used to group Verilog defines whereas harness gathers all the DUT source files. Work is the folder where simulations are actually run and output files are generated. Custom classes contain DUT-specific UVM classes which can be used to override the base classes using the

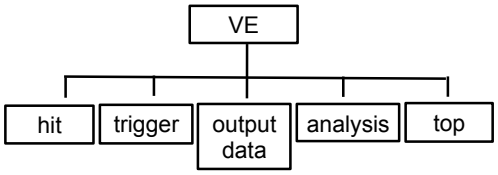

Figure 2.5: Verification Environment directory organisation [76].

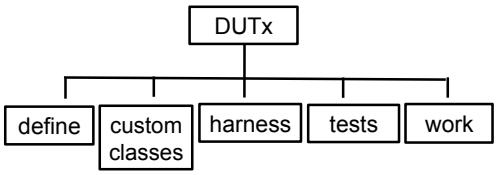

Figure 2.6: Specific DUT directory organisation [76].

mechanism described in Section 2.2.1. In the same way, also test files have been kept separate and a test folder has been dedicated to each DUT. It has also been mentioned that the verification environment is capable of simulating DUTs at different description levels. In order to guarantee such a support for them and also for alternative DUTs (possibly using different protocols) it has been essential to use a modular approach combined with the factory override mechanism. While at TL description the concept of interface is a port, i.e. a channel where transactions are passed to transmit information, working at behavioural/RTL/gate level, interfaces are made of physical analog or digital signals and constitute the boundary between the high-level environment and the chip. In UVM, normally drivers and monitors are respectively responsible for the conversion of signals into input or monitored transactions and vice versa. This is at the same time in conflict with simulation of chips at TL. For this reason, all driver and monitors have only been described at TL and they use TL ports. Such classes are part of the common verification environment directory. A DUT described at TL can directly interface to the UVCs through TL ports. Instead, separate blocks are required to simulate more detailed DUT implementations (behavioural, RTL, gate-level). In particular, transactions coming from drivers need to be converted into physical inputs (*TLM2SIG*) and information coming from DUT outputs have to be packed into protocol-

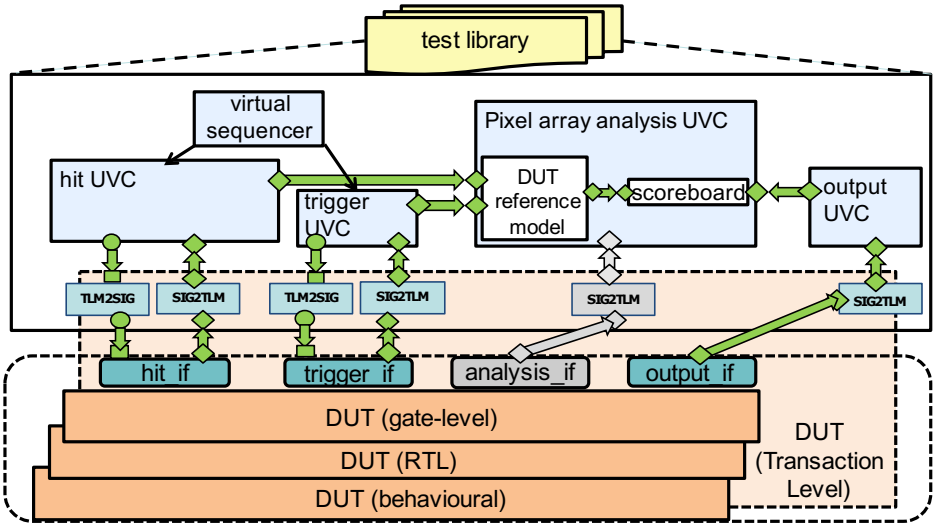

Figure 2.7: VEPIX53 block diagram emphasising its support for DUTs described at TL, behavioural, RTL and gate-level.

independent transactions (*SIG2TLM*), as highlighted in the block diagram in Figure 2.7. These classes can vary depending on the specific signals contained in the different interfaces and are classified as custom ones. Respecting such a structure and conveniently overriding classes, designers are able to simulate not only alternative architectures and DUTs but also at different description levels. Architectures described at behavioural, RTL and gate-level have been mainly simulated, whereas the feature allowing support for TLM description has not been used in this work.

2.2.3 Pixel hit stimuli generation

The stimulus to the DUT is generated by running sequences, extended from the *uvm_sequence* class, which control transactions that are passed to the drivers, translated and subsequently injected to the DUT. Sequences are therefore defined to control the input interfaces to the chip (i.e. hit and trigger) and they can be set and configured in the tests defined in the test scenario, through the UVM configuration database facility. At the top-level, the *virtual sequencer* is responsible of coordinating the launch of concurrent lower level sequences.

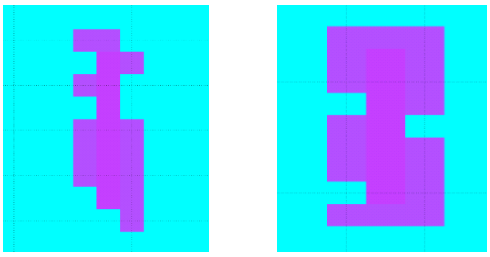

Figure 2.8: Example of the signal generated by a single particle on a group of pixels.

VEPIX53 supports multiple schemes for hit generation (and corresponding *uvm_sequence* extended classes), i.e. :

1. constrained random fashion, according to a set of pre-defined classes of clustered hits;
2. read from physics data in ROOT format produced by Monte Carlo pixel detector simulations;
3. a combination of the externally sourced hit data with the constraint random ones generated by the framework.

As far as 1. is concerned, a SV-UVM stimuli generator for pixel hits emulation has been presented in [47], and various hit scenario examples were described in [79]. In these works, multiple classes of realistic (and extreme) hits have been identified on the base of expected pixel hits at the HL-LHC e.g. single or grouped particles, background effects and noise hits. Moreover, the hit UVC provides flexible input monitoring at different levels of detail for debug, graphical representation and statistics collection of generated data [47]. A graphical example of the signal generated on a group of pixels is shown in Figure 2.8. The main goal of the hit generator is to be capable of generating patterns of hit pixels within the framework itself that emulate as well as possible the physics ones, still providing high level of flexibility. In order to clarify the core functionalities of the hit UVC, it can be said that interactions are modelled by taking into account the shape of the cluster of fired pixels. The generation of these clusters is based on several parameters which

can be set through the UVM configuration database: e.g. sensor parameters, a rate for each class of hit, expressed in Hz/cm^2, a list of specific parameters for each class of hits (e.g angle between the charged particle and the sensor, amount of fired pixels surrounding the core of the cluster in the case of charged particles), the range of possible amplitudes for the charge deposited by each hit. As regards the latter, a constrained randomisation of the amplitude value is performed by the framework and assigned to the pixels in the core of the cluster. Both Verilog and SV offer a set of embedded random constrained generators for several distributions (e.g. uniform, Gaussian, Poisson,..). In [47] amplitude values had been randomised with a uniform distribution in a given range. Actually, the expected distribution of the charged deposited by a particle crossing a sensor have peculiar distributions, depending on multiple sensor parameters. For this reason, a more detailed model of the random constrained stimuli has been developed, implementing a non-uniform random generator in the SV framework where the chosen distribution can be read from a file. Such a feature could also be useful for different purposes, allowing to choose any desired amplitude distribution. The specific distribution, which has been provided within the RD53 collaboration from more detailed sensor simulations is shown in Figure 2.9. The format used already takes into account the conversion from an analog charge to a digital amplitude, represented in the form of a ToT value. The hit configuration switches make it possible to also choose between the uniform and non-uniform distributions through the test.

2.2.3.1 SystemVerilog interface to externally provided Monte Carlo data

With respect to [47] and [79], the possibility of importing hit patterns from physics data, has been additionally included in VEPIX53. Such data have been provided by both the CMS and ATLAS experiments, featuring HL-LHC operating conditions and the specifications related to the Phase 2 upgrade. Physics simulations emulate collision happening in the LHC beam-pipes. Protons are circulated in several very closed packed bunches, in order to maximise the probability of protons colliding with each other. Every time these bunches

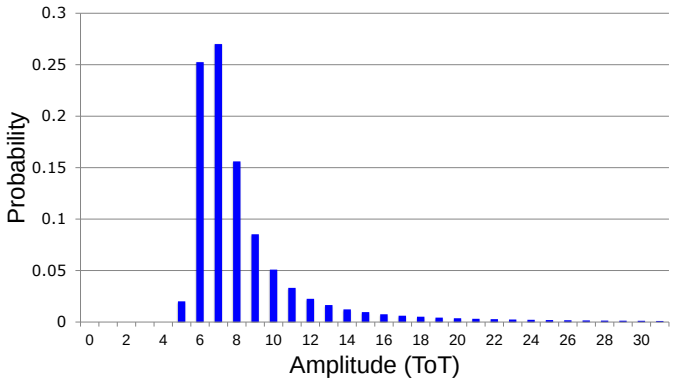

Figure 2.9: Distribution of amplitude imposed to fired pixels in the SV environment based on a non-uniform distribution provided through a file (example provided from detailed sensor simulations).

cross one another, more than one proton-proton collision takes place. As introduced in Section 1.2, the number of these collisions is the so-called pile-up, a quantity directly influencing the hit rates seen by the pixel readout chips in the experiments. The CMS data, produced by a workflow based on the CMS data analysis framework (CMSSW), contain events related to layer 0 of the pixel detector with different pixel sizes (50×50 or $25\times100\,\mu m^2$), sensor thickness of 150 μm, a digitizer threshold of 1500 e^- and a pile-up of 140. The ATLAS data, on the other hand, were extracted from Analysis Object Data (xAOD) generated with the ATLAS simulation chain and are related to all the four layers of the detector, with a pixel size of 50×50 μm^2, sensor thickness of 150 μm and digitizer threshold of 500 e^-. No pile-up was initially provided for these set of data. The structure of the barrel in the case of the CMS pixel layout has been shown in Figure 1.6. For both the CMS and ATLAS data, subsets have been extracted related to modules at the center and edges of the barrel, i.e. with particles hitting the sensor at different angles (more perpendicular at the center, in an oblique fashion at the edges). The common characteristic of the CMS and ATLAS data analysis simulation chains is that they use a common framework, so-called ROOT, to flexibly obtain statistics over the huge amount of events. It provides all the functionalities needed to deal with big data processing, statistical analysis, visualisation and storage [80] and it is developed

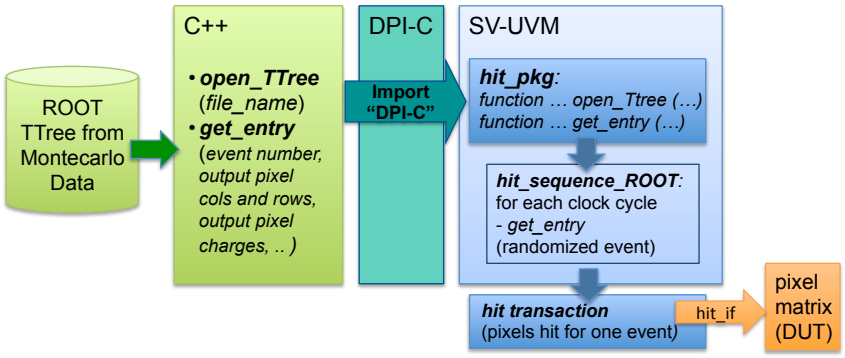

Figure 2.10: Implemented DPI C++/SV interface for the generation of hit transactions. C++ functions calls are imported and used to get entries of the ROOT TTree provided from Monte Carlo simulations, picking a randomised entry for each iteration.

in C++. The format used by ROOT to represent big data sets, in an optimized structure, is the so-called TTree. Thanks to the SystemVerilog Direct Programming Interface (DPI), it is possible to set up transparently interfaces with other languagues, such as C++. This allows the definition of functions in the SV hit generator which directly call ROOT routines. In particular, a hit sequence has been implemented, which iteratively chooses one event (for the whole pixel chip) from a ROOT TTree generated by the ATLAS/CMS simulation chain, as shown in Figure 2.10. The number of events simulated is configurable and the entry of the TTree to be selected is randomised. The ROOT powerful tool can be re-used to analyze data imported from ROOT into the VEPIX53 framework, producing statistics on the pixel chip hit stimuli. In order to provide a few examples, the cluster size distribution of the CMS pixel data mentioned is shown in Figure 2.11 and 2.12 for different locations in the barrel. The cluster size distribution for a sensor with $50\times50\,\mu\text{m}^2$ pixel size (i.e. 1:1 to the pixel size of the readout chip) is shown. Statistics are shown for the size of the cluster in the two perpendicular directions, z (along the barrel) and ϕ (cylindrical coordinate). It can be noticed that in the center of the barrel clusters have in average a similar size on the two directions (since particles are often hitting the sensor perpendicularly), while they are strongly elongated along Z in the case of the edges of the barrel. This is due to the fact that there

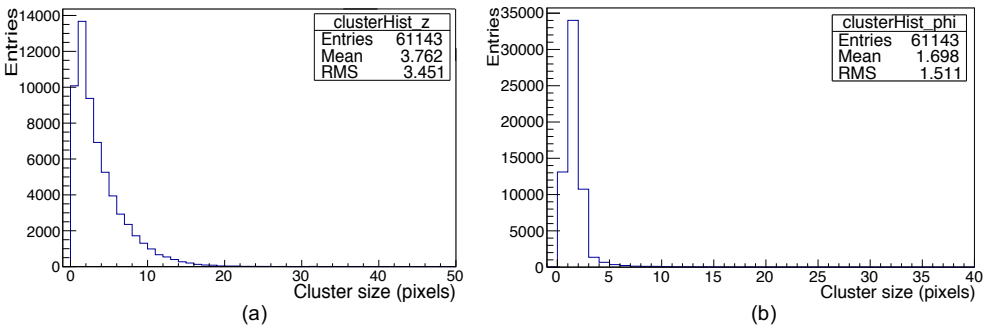

Figure 2.11: Cluster size histograms for modules in the center of the barrel (obtained from CMS ROOT TTrees). Sizes both along z direction (a) and ϕ direction (b).

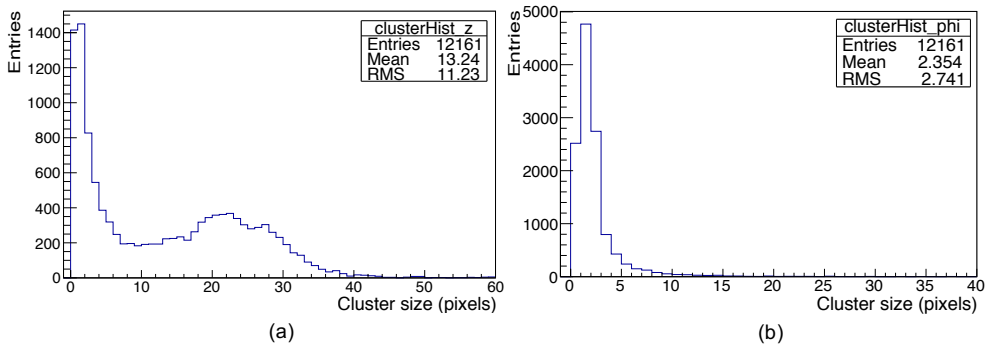

Figure 2.12: Cluster size histograms for modules in the edges of the barrel (obtained from CMS ROOT TTrees). Sizes both along z direction (a) and ϕ direction (b).

is a significant population of particles hitting the sensor in a oblique fashion. In addition, slow particles being bended by the magnetic field in the detector (and moving with helicoidal trajectory) also determine a significant population of clusters with smaller size along Z. Extracting such an information separately in the framework would require the implementation of clustering algorithms (not necessarly respecting the characteristics of the physics event), while profiting from the ROOT features gives additional control on the provided data.

In addition to the statistical analysis performed on the whole ROOT TTree, it is also possible to extract basic statistical information on the simulated input

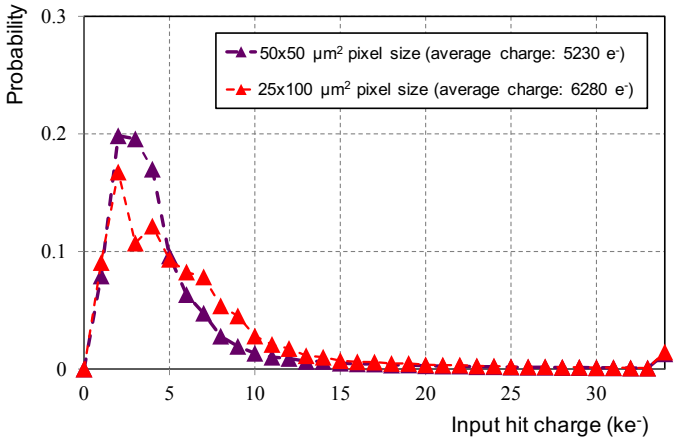

Figure 2.13: Monitored pixel charge amplitude distribution for CMS Monte Carlo data with different pixel sizes [77].

data sets within the VEPIX53 framework, such as the monitored hit rate on the full matrix and the charge amplitude distribution per pixel, an example of which is shown in Figure 2.13.

A final remark can be done concerning Monte Carlo hit data and constrained-random ones generated in the framework. The specific subsets provided from the experiments, will be in the future refined as the definition of the characteristics of the sensors and pixel detector layout for the Phase 2 upgrade progress. Obtaining data from heavy Monte Carlo simulations, which are themselves under study in the physics community, can be a time-taking process. This motivates the need for further flexibility on the input pattern generation, such as the capability of mixing the externally sourced hit data with constraint random ones. Even if the available Monte Carlo data are not simulated using the expected pile-up, it is still possible to stimulate the DUT with the expected hit rate and with meaningful cluster distribution. To this end, the capability of imposing a non-uniform distribution on charge amplitudes, is a valuable feature for the constraint random stimuli to resemble Monte Carlo hits. In addition, the flexibility on input stimuli generation is of particular importance to provoke extreme simulation conditions not covered by Monte Carlo data sets. This is a vital aspect for the verification of final chips for the experiments.

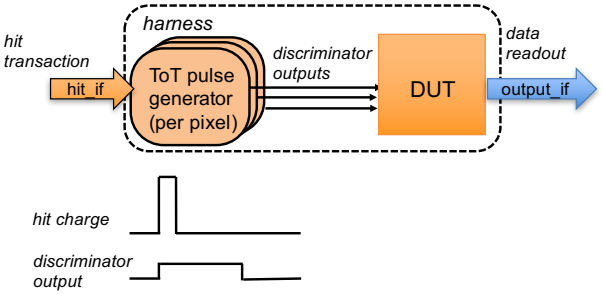

Figure 2.14: Block diagram of the chip harness containing multiple ToT pulse generators.

Indeed, this stage has been overlooked in some cases during the design of past pixel chips, causing readout bugs to be observed only during operation in the experiments. Similarly to the hit interface, also the trigger interface (*trigger_if*) is used to provide constrained-random stimuli to the DUT in terms of trigger accepts, which is a dedicated input to the chip for performing selection of events of interest. Monitoring and statistics collection are also available in the trigger UVC and different levels of detail can be configured through the test depending on specific needs.

2.2.3.2 Behavioural modelling of the analog front end

The verilog model of the various analog front ends provided by the designers does not completely describe their behaviour, e.g. it is already expected to receive a binary signal (discriminator output) as a hit input. For this reason, a behavioural model has been implemented for the charge-to-discriminator-output conversion and used across multiple front ends (also allowing to maintain only one generic model). Such ToT pulse generator modules are instantiated within the harness interfacing to the UVM environment and provide the binary signal to the DUT, as shown in Figure 2.14. It can be mentioned that for one specific front end the designer added a dedicated model to emulate a fast oscillator for operation in fast ToT mode (within the FE itself). The ToT pulse is still produced inside the ToT pulse generator, just referring to a faster clock. Dead-time is taken into account to measure hit losses in the UVM environment, by monitoring internal flags of the module. In order to simplify the

reference model, when a incoming hit overlaps with a precedent one, the second is neglected. The same approach is used for the time needed to the digital pixel logic to be capable of receiving a new incoming hit. Since it is deterministic for a given architecture, a DUT busy flag is directly used in the ToT generator and monitored for output prediction and inefficiency monitoring.

Two different implementation were developed for the ToT pulse generators block: *i)* a simple conversion from a given ToT value to a binary pulse with a duration of ToT clock cycles (in case constraint randomization or external hits already are provided as ToT values), *ii)* the conversion from an actual charge, either provided from constraint randomization or Monte Carlo data. The definition of a conversion function is required for the second implementation. An ideally linear relation has been considered between the input charge and correspondent pulse duration. The digitization to a finite number of bits is then performed by a floor function, which associates the pulse duration to the number of clock cycles covered. In [77] a series of linear approximations were also used to emulate extreme and intermediate FE settings, with saturation at a certain full scale charge (e.g ToT=15 for all charges higher than full scale), as also reported in Section 3.2.2. For final results presented in this work, an alternative function has been adopted. Indeed, the slope of the conversion function can be controlled based on the FE bias and it is normally defined as a trade-off between efficiency, charge resolution and dynamic range (for high energy particles) [81]. The linear conversion function has been defined based on such considerations: the first point in the line associates the threshold to a unitary ToT, while the second allows dead-time inefficiency to stay in the order of 1% (hit loss requirement). In particular, the second point is defined such that a Minimum Ionising Particle (MIP) traversing perpendicularly a 150 μm-thick sensor with $50\times50\,\mu m^2$ size (i.e. 12 ke^-) corresponds to a "high" ToT value (i.e. 10), in order to improve resolution without excessively compromising losses and dynamic range. The conversion function is reported in Figure 2.15. It can be noticed that no saturation of the conversion function is present, emulating the actual behaviour of FEs designed so far (which do not feature any discharge mechanism in case of overflow). This clearly refers to the dead-time cycles needed for the ToT conversion (even if the value stored

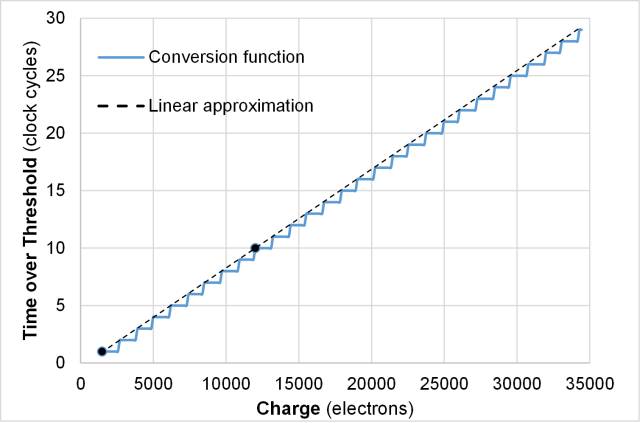

Figure 2.15: Charge to ToT conversion function for the analog front-ends: a linear relation between charge and discriminator pulse duration is defined. The duration is then digitized to the number of clock cycles (ToT value). The y-axis corresponds to the dead-time cycles for ToT conversion (even if the digital ToT counted saturates at ToT=15).

by the digital counting logic saturates at ToT=15, for a 4-bit measurement). It should be highlighted that the conversion function's choice is not meant to be absolute and it will depend on FE and digital architecture optimisations as well as on updated simulations, including position on the detector and/or experiments' needs. For example, not all the FE implemented so far feature a linear characteristic. However, based on the current knowledge, the model herein shown is considered a valuable example and has been used for final simulation results presented in Chapter 3.

Chapter 3

RD53A prototype for the phase-2 pixel upgrades: digital array architectural study

The RD53 collaboration was started in order to design the next generation of hybrid pixel readout chips to enable the ATLAS and CMS Phase 2 pixel upgrades [4]. In particular, the so-called RD53A integrated circuit is meant to demonstrate in a large format IC the suitability of the selected 65nm CMOS technology (e.g. radiation tolerance), stable low threshold operation, and high hit and trigger rate capabilities. The main characteristics of the adopted technology, provided through Europractice, are summarised in Table 3.1. RD53A is intended to be a prototype chip and not a final production chip for the experiments and for this reason it contains design variations for testing purposes. After RD53A implementation in silicon and testing, final production chips will be developed as designs revisions of RD53A, with possible modifications targeted to experiment specifications (e.g. pixel chip size, different functionalities and features, etc.). Section 3.1 introduces the pixel chip top level organisation, while Section 3.2 reports on the digital pixel array architecture study and optimisation at multiple design stages, which is the focus of this work.

Table 3.1: Main characteristics of the 65 nm technology [82].

TECHNOLOGY	MS/RF
Geometry	65 nm
Device Application	Low Power
Core Voltage (V)	1.2V
I/O Voltage (V)	2.5V
Poly Layers	1
Metal Layers (Min)	4
Metal Layers (Max)	9
Back end of line Dielectric	Low-K
Back end of line Metal	Cu

3.1 RD53A pixel readout chip floorplan and architecture

The RD53A chip is composed of two main parts (see Figure 3.1): the active area, a matrix made of 192×400 pixels with a pixel size of 50×50 μm^2 and the chip periphery, located at the bottom of the chip. As far as the chip dimensions are concerned, the width of RD53A is 20 mm, similar to what is expected for final production chips, whereas the height is constrained to 11.8 mm by the space available on the reticle, since the chip submission is shared with other projects in order to reduce the cost [83]. The peripheral circuitry is placed at the bottom of the chip and contains all global analog and digital circuitry needed to bias, configure, monitor and read out the chip. The wire bonding pads are organised as a single row at the bottom chip edge and are separated from the first row of bumps by 1.7 mm in order to allow for wire bonding after sensor flip-chip. In addition to those, a row of test pads has been included at the top of the floorplan for debugging purposes in such a prototype chip. It can be highlighted that in the pixel array analog front ends are placed in so-called analog islands composed of 4 fronts ends each, which are embedded in a digital synthesized "sea". The basic layout block composing the pixel matrix is a 8×8 pixel core containing 16 analog islands. Three flavours of pixel core have been integrated in RD53A, as described more in detail in Section 3.1.1.1 for the analog part and in Section 3.2.3 for the digital architecture.

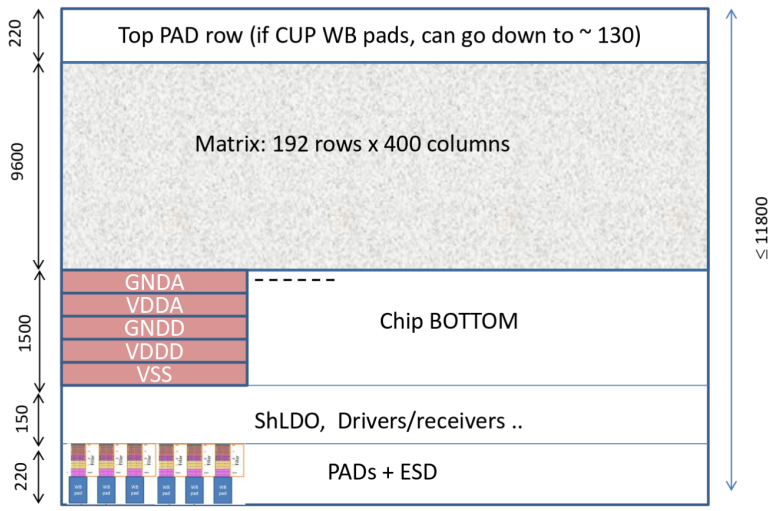

Figure 3.1: RD53A floorplan organisation showing the pixel matrix, the chip bottom including power regulators (ShLDO), drivers/receivers, chip PADs and ESD protection as well as a row of top pads [83].

Pixel matrix floorplan With reference to Table 3.1, the full metal stack (9 metals with an additional redistribution layer) has been used in the project. They are referred to as M1-M9 and AP for the top layer. The first 7 metals (M1-M7) are thin metals, while M8 and M9 are thick and ultra-thick metals, respectively. The selection of the metal stack is motivated by project choices on power and bias distribution, as well as analog-digital isolation. In particular, within the pixel matrix the following approaches have been adopted:

- the three top metals (M8-M9-AP) with maximum wire width are used for best possible power distribution as shown in Figure 3.2. The distribution is performed for double pixel columns (100 μm pitch). With this implementation, analog simulations have shown a 10.5 mV static power drop on both ΔVDDA=ΔGNDA for a full size chip, which have been seen to be sustainable for analog performance. The same conservative approach has been adopted for the digital power, more critical in terms of power fluctuations and peaks;
- M6 is used for analog bias distribution and shielding of bias lines is

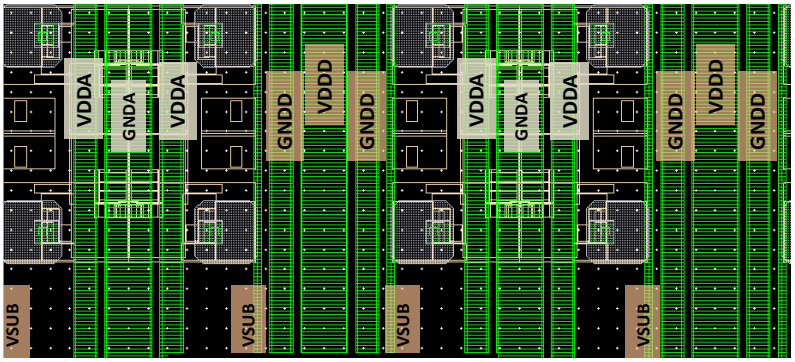

Figure 3.2: Power distribution scheme for the analog (VDDA, GNDA) and digital (VDDD, GNDD) power within the pixel matrix.

performed on top (M7) and bottom (M5), as highlighted in Figure 3.3. This means that on the top and on the bottom of analog islands these metals are not available for digital routing;

- M1-M2-M3-M4 are fully available for digital routing in the array.

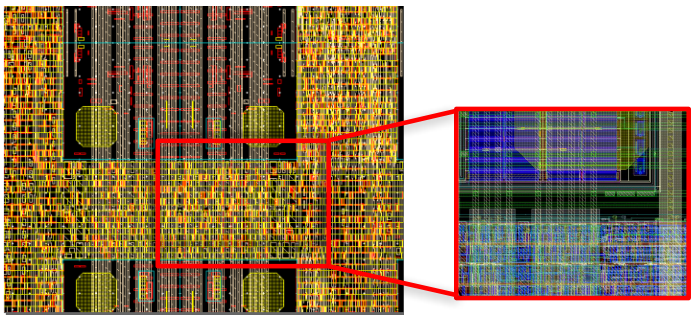

Figure 3.3: Zoom on analog bias distribution along the matrix, using M6 for bias and M5/M7 for shielding.

3.1.1 Architecture of main building blocks

The block diagram in Figure 3.4 shows in a functional view the main building blocks of the chip, starting from the bottom: the distribution of power regulators for the analog ($ShLDO\_An$) and digital ($ShLDO\_Dig$) power in the

IO frame; the Analog Chip Bottom (ACB) and the synthesized Digital Chip Bottom (DCB) at the bottom; the analog buses to the array for bias distribution and the digital signal lines from the DCB to the digital array. In the chip

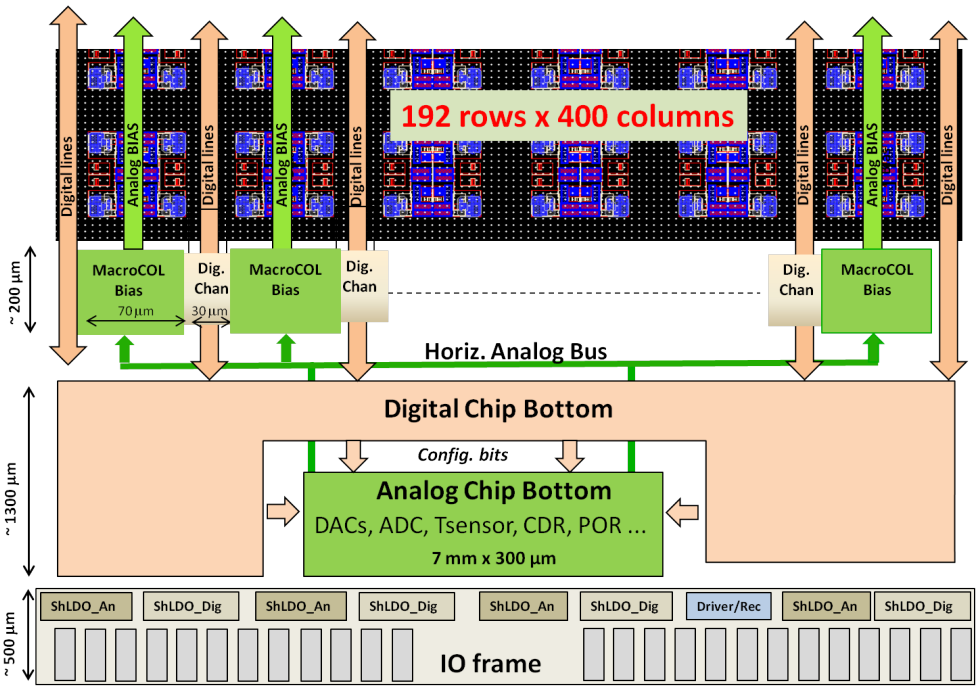

Figure 3.4: RD53A floorplan functional view [83].

periphery, all the analog building blocks are grouped in the ACB macro-block, which is fully assembled and characterised in an analog environment. The main functionalities of this block are: *i)* provide different references to current DACs., *ii)* monitor different signals of the chip (current references, temperature and radiation sensors, etc.) and digitise them through a 12 bit ADC, *iii)* provide two voltage levels for the calibration circuit. It also includes:

- power on reset block, whose main function is to ensure that the chip has a reasonable configuration immediately after startup and stored logic states are well defined. An asynchronous signal resets the global configuration memory to default values, whereas the pixel configuration is switched to use hard-wired default configuration instead of the values stored in their

registers;

- the Clock Data Recovery (CDR) is made of an internal Voltage Controlled Oscillator (VCO) and a Phase Locked Loop (PLL) to lock to the incoming 160 Mbps control serial stream. It is in charge of generating three of the clocks used within RD53A: 160 MHz, also referred to as "command clock" and used mostly in the digital periphery; the 1.28 GHz, which is the maximum frequency for data output from the serializer; the 640 MHz clock used to fine delay the command clock, in order to synchronise every chip's internal clock to the LHC bunch crossing (40 MHz);

- the output serializers, using the 1.28 GHz clock (or a 2-to-8 fraction of it, based on configuration) from the CDR to serialize the encoded chip output data on 4 lanes.

All the building blocks have been previously prototyped, tested and characterised in foreseen radiation environment. The ACB block is surrounded by a synthesized block, the DCB, which implements the Input, Output and Configuration digital logic. In Section 3.1.1.2 the functionality implemented in the DCB is summarised in order to provide some insight in the digital architecture of the whole chip.

3.1.1.1 Analog front ends

RD53A contains three different Front End (FE) designs, developed by different groups within the RD53 collaboration, to allow detailed performance comparisons. They are identified as Synchronous Front End (SFE), Linear Front End (LFE) and Differential Front End (DFE), with the last two being considered asynchronous front end designs. The three designs have common constraints and features which ease their integration in a unique pixel matrix: layout area for a 4-pixel analog island is limited to $70\,\mu\text{m}\times70\,\mu\text{m}$ for all variants and it contains also the bump bond pads (same for all designs) in a $50\,\mu\text{m}\times50\,\mu\text{m}$ grid. The calibration injection circuit, which allows front end testing through the injection of a defined amount of charge in pixels defined by configuration, is also common among the three. This choice guarantees

direct performance comparison. The bias distribution also follows the same organisation for all 3 analog designs. In order to cover a 400 pixel width with a 8×8-pixel building block, a 16-core width has been assigned to the SFE, whereas a 17-core width is used for the asynchronous ones. The latter are also placed next to each other as shown in Figure 3.5, as they have the most similar functionality and a large area with uniform response can be desired for sensor characterisation in test beams.

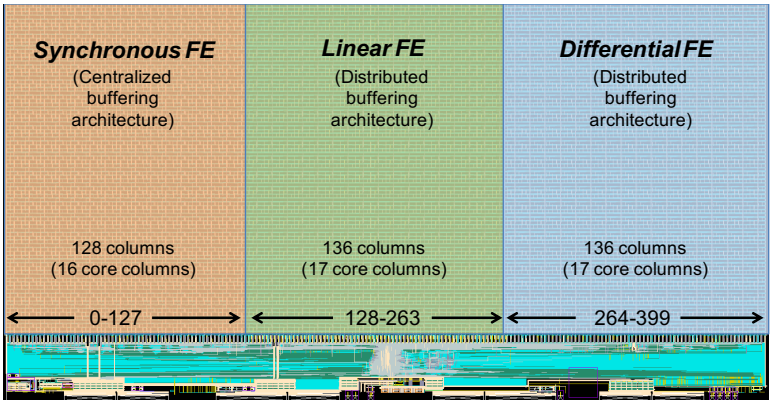

Figure 3.5: Arrangement of front end flavours in RD53A. The pixel column number range of each flavour is shown along the bottom. The type of digital architecture used in each flavour is also written in parenthesis.

The main characteristics which distinguish the 3 front end flavours are the following:

- the SFE features a synchronous discriminator composed of a differential amplifier and a positive feedback latch, which can be turned into a local oscillator up to 800 MHz using an asynchronous logic feedback loop in order to perform a fast Time-over-Threshold (ToT) counting. In addition, this front end does not use a pixel-by-pixel threshold trimming with local threshold adjustment and instead adopts an auto-zeroing scheme that requires periodic acquisition of a baseline;

- the LFE is a fully analog circuit and implements a linear pulse amplification in front of the discriminator, which compares the pulse to a threshold

voltage; locally the threshold is trimmed in each pixel using 4-bit resistor ladders;

- the DFE is also a fully analog circuit which uses a differential gain stage in front of the discriminator and implements a threshold by unbalancing the two branches. This FE features local circuitry for threshold adjustment, based on a 4-bit binary weighted DAC.

The interested reader can find a more detailed description of the FEs in the following references for the SFE [84], LFE [85] and DFE [86].

3.1.1.2 Digital chip bottom

The DCB includes all the digital chip periphery (with the only exception of the pixel array column readout) and its block diagram is shown in Figure 3.6. Its main sub-blocks are summarised in the following:

- the Channel Synchroniser (CS) block is used to generate from the 160 MHz clock the 40 MHz clock, i.e. the only clock distributed to the pixel matrix, which can be phase-aligned to the symbols (Sync) in the command stream when they are sent. Its synchronisation to the LHC bunch crossing cycle is achieved thanks to the fine delaying of its source clock (160 MHz clock);

- the CoMmand Decoder (CMD), which is in charge of decoding commands incoming from a single differential serial input. The custom protocol used transmits encoded clock and commands on a single link, is DC-balanced with short run length for A/C coupling and reliable transmission, and has built in framing and error detection. The main commands are defined to send triggers, read and write global and pixel configuration, perform full data path reset through an Event Counter Reset (ECR) or only reset of the bunch crossing counter through a Bunch Counter Reset (BCR). A generic global pulse is also used for multiple purposes, including generating calibration injection pulses. Details on the command protocol and signals are available in [83];

- the data builder is composed of a tree of First-In First-Out (FIFO) which receive data from each the columns composing the pixel matrix and progressively aggregate them in packets to be sent to the output Clock Domain Crossing (CDC) FIFO. This block is a boundary between the 40 MHz and the 160 MHz clock domain, data are first received with the pixel array clock and further processed/aggregated with the 160 MHz. Timing constraints are defined so that the two clocks are treated as synchronous to each other by timing verification;
- the output CDC FIFO used to transfer data across the 160 MHz clock domain to the data clock used by the Aurora transmitter (i.e. the serializer 1.28 GHz divided by 20 = 64 MHz);
- the Aurora frame transmitter implements data framing and 64b/66b encoding according to the Aurora Xilinx documentation [78], and sends pixel and user data with proper format to the serializer. Multi-lane frame transmission (2 or 4 lanes) is supported as well as single-lane transmission.

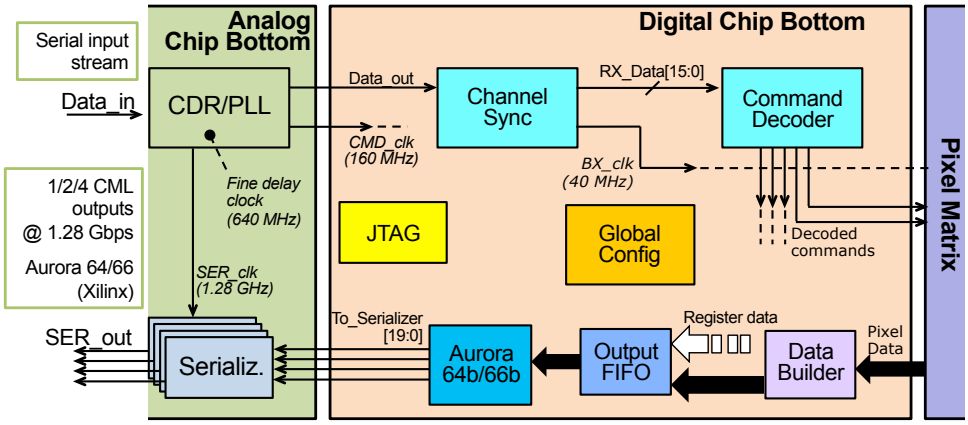

Figure 3.6: Block diagram of the digital chip bottom and its interface to the pixel matrix and ACB [87].

With RD53A being a prototype chip, multiple features have been added to allow debug and testing. A JTAG interface is included to control the chip

(bypassing the command decoder) and to run scan chain tests on global configuration. In addition to the CMD, also most of other critical blocks (e.g. the CDR, the power on reset, the serializers, power regulators, etc.) can be bypassed thanks to dedicated pads and backup structures included in the prototype. A detail description of the testing modes is found in [83].

3.2 Digital array architectural study and choice

The requirements on pixel size, hit efficiency (with the defined hit rates and trigger latency) and low power discussed in Section 1.2.1.1 demand dedicated optimisation of the logic of the digital array with the chosen CMOS 65nm technology. Optimising buffering resources to be arranged in the limited pixel area can be addressed by storing together information from multiple hits from the same physical cluster. Sharing of trigger latency buffers can lead to more compact circuitry and lower power. Such an investigation is herein addressed: first, an architectural exploration is performed in Section 3.2.1 with pixel region architectures described at behavioural level; second, more detailed RTL descriptions are optimised and compared in Section 3.2.2, as implemented in small-scale prototypes; finally, further optimisation of the chosen architectures for RD53A and final simulation results are shown in Section 3.2.3.

3.2.1 Architectural exploration at behavioural level

The first questions to be answered are how many pixels should share storage logic within Pixel Region (PR), in what pattern, with what internal organisation, and how are region boundaries handled. The optimisation depends on cluster size distributions, which in turn depend on sensor type and location in the detector, and on physics input. An initial study of shared buffering performance was performed analytically in [88] and results identified square regions with size from 2×2 to 4×4 pixels as the most suitable ones. The developed VEPIX53 framework has been adopted to simulate the pixel chip architectures with more detail and with more elaborated cluster models for input hits. The candidate architectures, both implementing triggered readout

and regional storage during the trigger latency, differ on the organisation of this buffering and its control logic. In particular, the following architectures have been evaluated at behavioural level:

- distributed latency counters, where only the timestamp information is handled in a centralised fashion, whereas independent pixel buffers are used to record the hit charge;
- centralised FIFO, where the complete hit information is stored in a unique shared buffer.

The behaviour of the analog front end was abstracted with a charge converter module, which converts the hit charge into a discriminator output. Both architecture feature a counter per pixel to measure the ToT. The hit timestamp, on the other hand, is provided to the PR as a 9-bit bus coming from a counter module defined at the end of column sector of the pixel chip. Additional details on the implementation of the selected architectures are provided below. In both architectures a 40 MHz bunch crossing clock is provided to the PR.

Distributed latency counters architecture In this architecture, based on the ATLAS FE-I4 [89], the arrival of the hit enables latency down counters, defined inside each cell of the PR latency memory, and trigger matching is checked when such counters reach zero (i.e. after the latency); a memory management unit links the read and write pointers to the memory cells among the local pixel ToT buffers and the latency memory and assembles the triggered hit packets containing the timestamp value and the ToT values from each PUC. The correspondent block diagram is reported in 3.7.

Centralised FIFO architecture In the centralised FIFO architecture the regional buffer stores hit packets containing both ToT and timestamp information; trigger matching is checked by comparing the external counter signal, subtracted by the trigger latency, with the stored timestamp. The block diagram of the architecture is shown in Figure 3.8: since different pixels, possibly recording multiple ToTs, need to access a unique memory, the control Write Logic is also shared within the PR. The same approach is used for the trigger

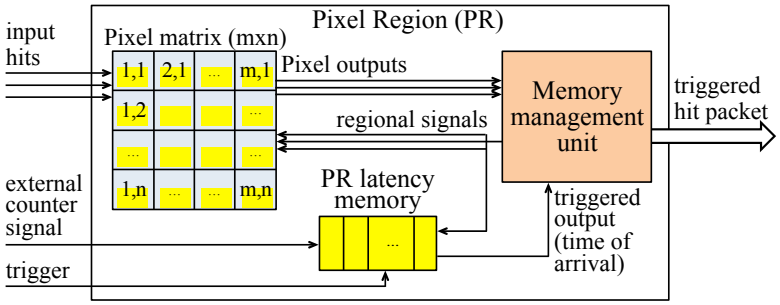

Figure 3.7: Block diagram of the distributed counters buffering architecture [90]. Memory elements are highlighted in yellow. ToT information is stored in the correspondent pixels, whereas the latency counters are centralised in the PR.

matching logic. The implementation of the pixel Finite State Machine (FSM),

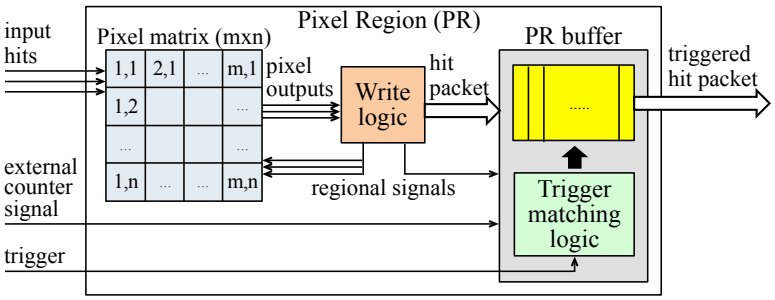

Figure 3.8: Block diagram of the centralised FIFO buffering architecture [90]. The stored information is contained in a centralised PR buffer (yellow).

when a pixel is hit moves from a idle state to counting one, to then be readout and saved in the centralised buffer. If instead the pixel is not hit but another one the same region is, the common write logic will make the first pixel blind until the information from the second is stored.

The architectures have been described at behavioural (not synthesizable) level as part of a parameterised pixel chip model which can alternatively implement the two buffering architectures and a PR of parameterised size and shape [90]. On one hand, the description level chosen allows to profit from SV high-level structures, for faster development. On the other hand, when compared with a TL description, it is closer to the physical implementation,

making it possible to describe also details of the logic which could be missed otherwise. The lack of connection between a TL description and the physical one was reported to be a disadvantage of a TL implementation in [91], whereas a behavioural description eases the translation to a synthesizable one (only requiring to replace a set of un-synthesizable constructs with synthesizable logic). A key factor of this choice is that the optimisation of the latency buffering has to be achieved at local pixel region level, as low level details are critical in terms of area and power performance. Since the expected hit rate is also rather uniform across the matrix, for this application it is not considered mandatory to simulate very big structures. Therefore, no elaboration or simulation time bottleneck has forced to move to higher description levels than the one adopted.

Both architectures were simulated for relevant pixel region configurations 1×1, 2×2 and 4×4 pixels. In order to evaluate the worst case conditions available at the time of the work, the presented simulations have been run using Monte Carlo data sets related to the innermost layer of the detector at the edges of the barrel, featuring a pixel size of 50×50 μm^2 and a pile-up of 140. For these data the corresponding monitored hit rate is 2.7 GHz/cm^2. Simulations were run with 10 μs trigger latency for ∼12 ms (average simulation time: 2 hours), in order to collect sufficient statistics on the pixel region performance using the available Monte Carlo data. The architecture performance for pixel regions at behavioural level is evaluated by monitoring *i)* hit loss and *ii)* buffer occupancy through the VEPIX53 analysis UVC. As introduced in Section 1.2.1.1, the hit loss is due to two main sources: dead-time of the PUC/PR and latency buffer overflow. The latter, on the other hand, is used for building the occupancy distribution, from which it is possible to carry out the corresponding buffer overflow probability. The hit loss rate due to dead-time for each architecture and configuration is reported in Figure 3.9 (a). These results are compatible with those produced using hits generated within the SystemVerilog framework [47] and show an increasing dead-time for the centralised FIFO architecture as the region gets bigger. 40 MHz ToT counting is assumed for both architectures. In the distributed latency counters architecture, on the other hand, the hit loss rate is constant with respect to the

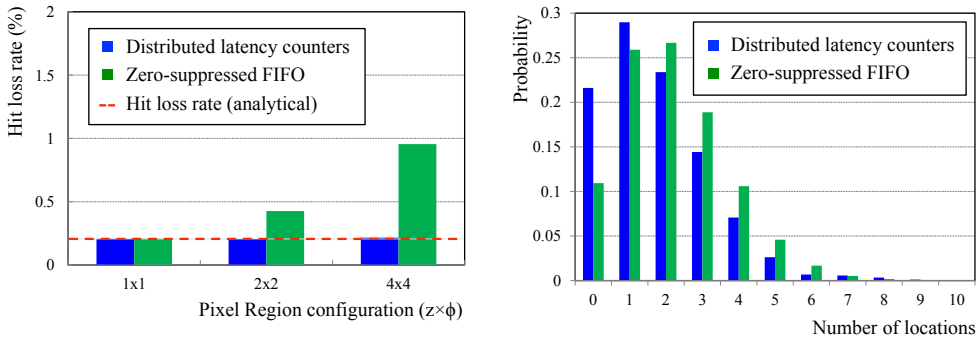

Figure 3.9: (a) Hit loss rate in pixel region due to dead-time; (b) Occupancy histograms of trigger latency buffers for a 2×2 pixel region [90].

Table 3.2: Hit loss rate due to buffer overflow [90].

Pixel region ($z \times \phi$)	Buffer locations	Hit loss rate	
		Centralised FIFO	Distributed latency counters
2×2	8	0.030%	0.129%
4×4	12	0.002%	0.032%

PR size and it has also been proven that it is comparable with the hit loss rate that is calculated analytically using the average ToT of the pixel hits [69]. The latency buffer occupancy was monitored and examples of histograms are shown in Figure 3.9 (b). DUTs were simulated with oversized latency buffers, in order to carry out the buffer overflow probability as a function of the number of locations. From these it is possible to determine the required number of locations that keep such a probability below a certain design value (e.g. 1% or 0.1%). Using the suggested number of locations related to an overflow probability below 0.1%, further double check simulations have been run with fixed size buffers: as reported in Table 3.2, the monitored hit loss due to buffer overflow is in most cases below ∼0.1%. At this stage of the architecture evaluation, the two architectures have show a comparable behaviour in the number of buffering locations, whereas the distributed one has been seen to be preferable in terms of dead-time losses.

3.2.2 Optimisation and comparison of selected architectures implemented in small-scale prototypes

Initial architectures studied at behavioural level have been replaced by synthesizable RTL descriptions and design improvements and some choices (with respect to the generic behavioural chip) have been addressed, before their implementation into two small-scale chip prototypes. The VEPIX53 framework has been used:

- to simulate a 2×2 PR distributed buffering architecture produced in the FE65-P2 prototype [92] in order assess its compliance to the RD53A specifications;
- to optimise and verify of an alternative 4×4 PR centralised architecture, then implemented in the CHIPIX65 prototype chip [93].

Performance assessment of the FE65-P2 distributed architecture
The architecture implemented in the FE65-P2 prototype features local ToT storage and centralised time information storage. With respect to the distributed latency counters architecture (based on FE-I4), no more downcount counters are used to perform the trigger matching, with the aim of saving the power they consume during the trigger latency. Instead, two separate timestamps (shifted of the latency time) are distributed as global signals: at any incoming hit, the timestamp with the higher count is stored and when its value gets equal to the second timestamp (i.e. after the trigger latency), if a trigger pulse is detected, trigger matching takes place. As far as the buffering is concerned, in the prototype chip all the memories are composed of 7 locations: for four pixels in a region, this means 7×4×4-bit for the ToT memories plus the 7×10-bit for the timestamp memories. From the implementation point of view, no dedicated memory structure has been used, instead storage has been performed with banks of registers, made of edge-sensitive flip flops.
Simulations have been run of a 4×64 pixel multicolumn containing the RTL description of the described architecture and taking into account several different parameters, such as: *i)* different analog FE models, with multiple charge-ToT relations, *ii)* different numbers of memory locations, *iii)* input hits featuring

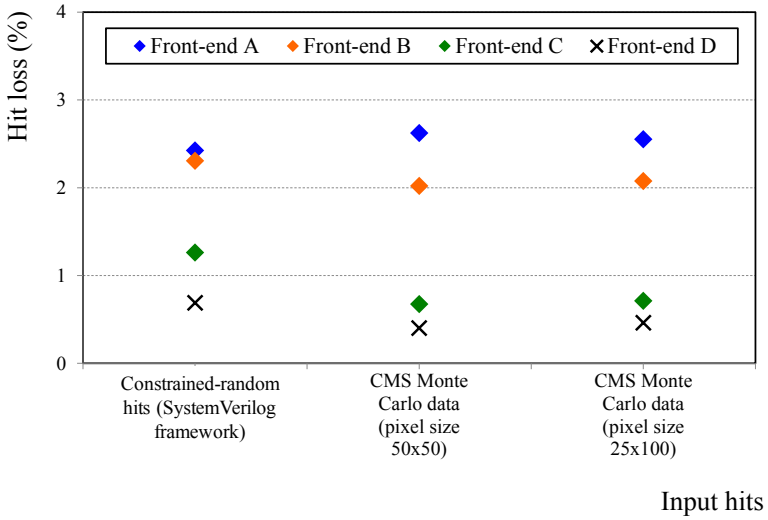

Figure 3.10: Monitored hit loss due to dead-time for different analog front ends and input hit charge distributions [77].

a 3 GHz/cm^2 rate with different charge distributions, also taken from CMS Monte Carlo data. Since data available were featuring a pile-up of 140 (not enough for producing the desired 3 GHz/cm^2 hit rate), the imported data were mixed with constraint-random hits generated by the SV framework, contributing with an additional 1 GHz/cm^2 rate and featuring the same hit charge distribution as that of the Monte Carlo data (shown in Chapter 2, Figure 2.13).

In [77] it was concluded that a total hit loss rate smaller than 1% can be achieved with the evaluated architecture. In particular, as shown in Figure 3.10, losses due to dead-time can be obtained by tuning the analog front end in such a way that the corresponding output ToT distribution will feature a low average value. In Figure 3.10, four analog front ends were modelled at behavioural level with different charge-ToT relations and different ToT clock periods. The front-ends A, B and C operate at the bunch crossing clock frequency of 40 MHz, featuring a full scale charge of 4500, 7500 and 35000 electrons, respectively. The front-end D, instead, operates at 128 MHz with a full scale charge of 10500 electrons. The latter models a front-end including a fast oscillator, designed in CHIPIX65 and RD53. For simulation purposes,

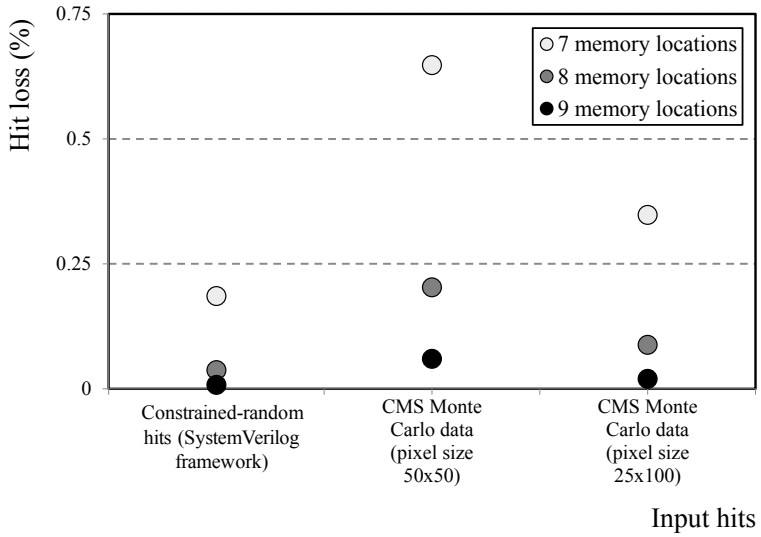

Figure 3.11: Monitored hit loss due to buffer overflow for different numbers of locations and input hit charge distributions [77].

this capability has been added at behavioural level combined with the digital architecture of the prototyped FE65-P2. As discussed in Section 2.2.3.2, the actual choice in terms of charge-ToT relation will depend on the trade-off between efficiency, charge resolution and dynamic range. The ToT clock period depends on the chosen analog front end (i.e. only the SFE features fast clock operation). From the buffer overflow point of view, the adoption of 8 buffer locations instead of 7 is already sufficient for considerably reducing hit loss (Figure 3.11). The increase in area occupation associated to a higher number of memory locations was estimated as well: this was done by breaking down the area of the FE65-P2 digital pixel region logic (4046 μm^2 in the synthesized prototype) into its different components and recalculating it for 8 and 9 locations. As summarised in Table 3.3, pixel regions featuring 8 or 9 locations occupy 7% or 15% more area, respectively.

Simulation and optimisation of the CHIPIX65 centralised architecture Performance limitations described in Section 3.2.1, due to the common management of the single FIFO inducing pixel region dead-time, have been ad-

Table 3.3: Occupation of area for different pixel memory sizes.

Number of buffer locations	Digital pixel region logic area (μm^2/pixel)
7	1012
8	1080
9	1165

dressed within the CHIPIX65 project, while translating the architecture from behavioural to RTL description. The pixel region freezing problem has been overcome thanks to a fixed buffer writing-time. The block diagram of the proposed 4×4 pixel region architecture is reported in Figure 3.12. The system is composed of the following building blocks [94]: *i)* 16 independent pixels which consist of both a front end and a digital interface, which computes the ToT information within a fixed dead-time, and flags the end of processing to the shared digital logic; *ii)* a shared region digital writer synchronously checks ready pixels flags and, if any hit is detected in the region, saves into the region buffer a reduced (up to 6 pixels) information packet, selected by a priority encoder; *iii)* a shared region digital buffer, whose depth is 16, which saves packets consisting of a timestamp, a binary hit map of every pixel in the region (more efficient than the pixel addresses), and up to 6 5-bit ToTs from the pixels; *iv)* a shared region digital trigger matcher, where comparators are multiplexed to either perform trigger matching against the buffer rows or to mark (through anti-aliasing logic) the buffer locations as invalid once the corresponding valid trigger window has elapsed; *v)* the region digital output finally selects the triggered entries and sends them down the macro column: column arbitration is based on a busy signal in a fast-or configuration. Simulations have been run for 20 ms constraint random input hits featuring a 3 GHz/cm^2 rate and with both the analog front end flavours integrated in the CHIPIX65 project prototype. The adoption of VEPIX53 in the project from early stages has allowed to guide implementation choices. The hit loss results shown in Figure 3.13, are particularly interesting in the case of the fast front end, where the total analog and digital dead-time equals only 5 clock cycles. Detailed simulations under the same conditions have been performed with parametrised buffer depths of

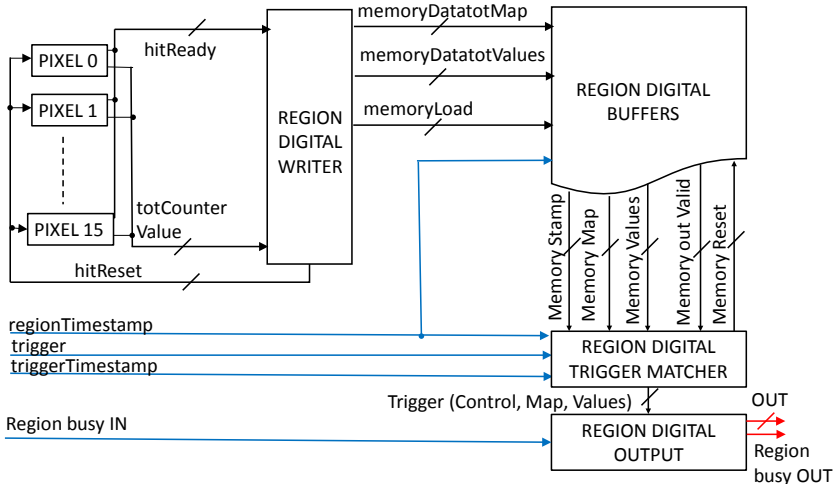

Figure 3.12: Centralized 4×4 pixel region architecture of the CHIPIX65 small-scale prototype [94].

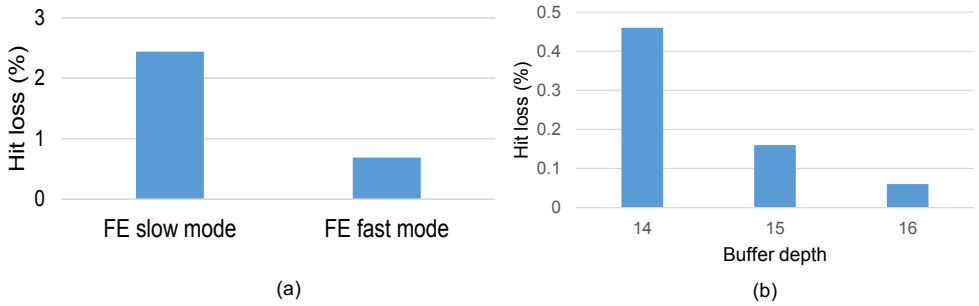

Figure 3.13: Centralised architecture performance results [94]: hit loss due to pixel deadtime for both the slow (40 MHz) and fast (128 MHz) front end modes (a), hit loss due to buffer overflow for increasing values of buffer depth (b)

.

14, 15 and 16 locations. The choice of the highest buffer depth, which has also been proven to fit in the digital area available, makes digital losses negligible. A significant optimisation of the area available has been achieved thanks to the reduction of the number of fired pixels stored in each region packet. It has been indeed seen that even in the extreme case of the detectors sitting at the edges of the barrel, where elongated clusters cause bigger cluster sizes,

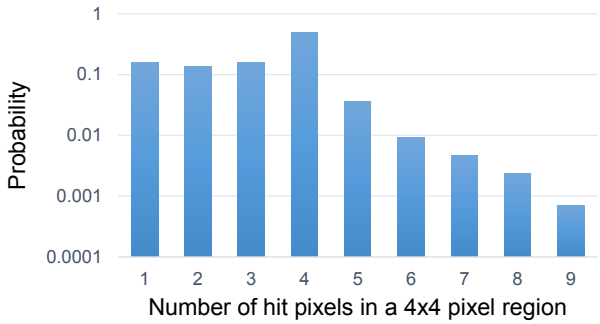

Figure 3.14: Histogram of number of hit pixels per pixel region (4x4) simulated with external Monte Carlo data in the extreme scenario at the edges of the barrel [94].

the average number of pixels fired per region is lower than 4. The detailed probability distribution is reported in Figure 3.14. These results justify the choice of limiting the number of stored hits with ToT to 6, instead of 16. For the remaining 10 pixels the binary information is anyway available, i.e. the hit information is not lost. On the other hand, this brings an effective area gain in terms of storage resources.

3.2.2.1 Architecture comparison: summary of results

The results presented separately for the two architectures are summarised and extended (with area and power metrics) in Table 3.4 for a comprehensive comparison. The comparison metrics defined are:

- inefficiency i.e. hit loss, achieved through RTL simulation within VEPIX53;
- cell area occupation (comparison of synthesized architectures);
- power consumption (comparison of place-and-routed architectures in the typical corner).

Also it has been necessary to make some common choices to make the comparison meaningful:

- same number of pixels (4x64). This corresponds to a single column for the centralised architecture (PR = 4×4 pixels; 1×16 PRs = 4×64 pixels)

and to a double column for the distributed architecture (PR = 2×2 pixels; 2×32 PRs = 4×64 pixels);

- number of buffer locations that keeps inefficiency in the order of 0.1%;
- no use of SEU tolerant design techniques (refer to Chapter 5 for radiation-hard design techniques).

Table 3.4: Comparative table between centralised and distributed buffer architecture. Simulation conditions: pile-up 200 (∼3 GHz/cm^2 hit rate), constraint random hits generated within VEPIX53, simulation run for 800,000 bunch crossing cycles, trigger latency: 12.5 μs, trigger rate: 1 MHz. *The centralised architecture features a higher number of ToT bits (5 instead of 4).

Metrics		4x4 centralised buffer architecture*		2x2 distributed buffer architecture	
Inefficiency (%)	Dead-time (single pixel loss)	Slow FE	Fast FE	Slow FE	Fast FE
		2.44	0.69	0.58 - 2.67	0.33 - 0.71
	Buffer (PR cluster loss)	14 locations: 0.46 15 locations: 0.16 16 locations: 0.06		7 locations: 0.57 - 0.81 8 locations: 0.16 - 0.21 9 locations: 0.04 - 0.07	
	Limit on number of ToTs (only ToT info loss)	0.29 (6 TOTs saved out of 16)		- (all ToTs stored)	
Area	μm^2/pixel (only digital logic)	761 (14 loc.) 786 (16 loc.)		1039(7 loc.) 1165 (9 loc.) (EST)	
Average Power	μW/pixel (only digital logic)	∼7.5		4.8	

Some conclusion can be drawn based on obtained results for the FE65-P2 and CHIPIX65 architectures:

- in terms of dead-time, the centralised architecture shows significant losses due to the implemented fixed buffer writing-time (equal to the time needed to compute the longest possible TOT, i.e. 5 clock cycles for the fast FE and 15 with a standard "slow" FE), when no fast FE is used. On the contrary the digital logic of the distributed architecture does not introduce any dead-time in addition to the analog FE contribution (the

range of results is due to multiple charge-to-ToT conversion functions used);

- as concerns buffering resources, the centralised architecture shows lower buffer losses with the number of buffer locations (16) which could be actually fit. This has been possible thanks to an efficient sharing between more pixels and thanks to the data reduction performed (limited number of stored ToTs);

- in terms of area, the centralised architecture has been seen to have advantages which could potentially allow the use of more buffer locations or additional features; a remark is related to the fact that to keep the comparison independent from implementation details, the area calculation is only based on the gate cell area (not including full clock-tree, routing, timing optimisation);

- as regards the power consumption, the centralised buffer architecture shows significantly higher values per pixel. Due to the limited time available, power optimisation was not extensively performed in the CHIPIX65 digital architecture. Therefore, at this stage it is not obvious to discriminate whether the higher consumption is also due to slightly higher complexity of the logic (due to data reduction, etc.) when compared to the distributed one.

3.2.3 Optimisation for the RD53A chip

Given the complementary advantages and disadvantages of the two architectures, the conclusion for RD53A has been to integrate both of them: the Centralised Buffering Architecture (CBA) together with the SFE (since capable of fast ToT measurement with lower pixel dead-time), while the Distributed Buffering Architecture (DBA) has been adopted for the asynchronous FEs. This decision has allowed the design team to address limitations of both architectures: limited area (and therefore constrained buffer size with increased buffer losses) for the DBA; power consumption and dead-time losses for the

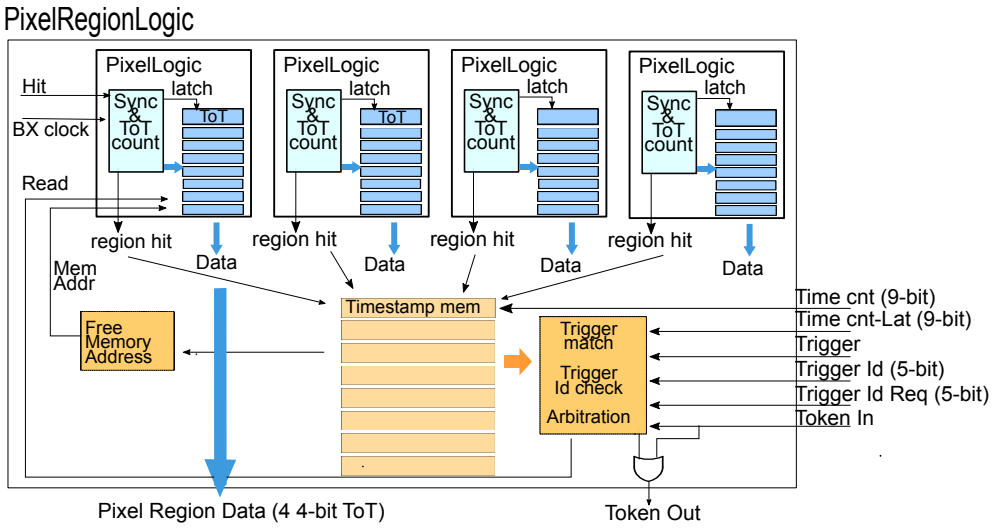

Figure 3.15: Block diagram of the PR logic of the DBA architecture.

CBA. It can be mentioned that the architectures adopted in previous prototypes have been re-organised, both from the point of view of the hierarchy and of the code cleanness. The goal has been to achieve better modularity and minimise error-proneness in the integration of multiple FEs and architectures. Based on the RD53A design team organisation, the main focus of this thesis in terms of design optimisation has been the DBA. Nevertheless, many of the techniques and results achieved and described in the next chapters can be adopted across both architectures (e.g. timing, power optimisation, etc.). Moreover, in this work a comparison of design performances of both architectures has been performed by means of VEPIX53 and digital design tools. A summary of the main optimisations affecting simulation performance are reported in this section for both architectures (with a higher level of detail for the DBA), whereas the next chapters cover implementation-related aspects.

3.2.3.1 Distributed Buffering Architecture

The block diagram of the pixel region logic of the DBA architecture implemented for RD53A is shown in Figure 3.15. It does not include the logic directly interfacing to the analog FE, which mainly contains pixel configura-

tion and logic for both analog and digital calibration pulse injection. Since this part is FE-dependent, it has been implemented separately from the rest of the digital pixel logic, in order to keep the latter identical for the two FE flavours. The *PixelLogic* is responsible of synchronising incoming hits, performing ToT counting and saving the resulted value in the local memories. The leading edge also instructs the common logic in the PR to save the timestamp in a free memory. In case some pixels out of the four are not hit, the leading edge from a hit pixel does still trigger writing of a none-hit (to complete the packet). This means that no pixel address needs to be used for the pixel in the region. The expiration of the trigger latency is implemented as in FE65-P2, since it reduces logic and power in the region. After the latency has expired, hits are either selected for readout or discarded according to the trigger signal. Readout is based on token arbitration, with priority given to PRs on the top of the array. A further *TriggerId* check is performed to match the specific event which is being read from the periphery while also subsequent events may have been triggered.

Optimisation of ToT storage elements Apart from the hierarchical organisation, the logic optimisation has been aimed to reduce area and buffer losses. The design was initially featuring 7 memories (both for each pixel ToTs and for the timestamp) and after place and route the area utilisation was close to 90%. In this design, this value has been seen to be approximatively the maximum allowed to close design at the final stages, i.e. no additional logic could be fit without causing implementation issues. As shown in Table 3.4, a buffer-depth of 7 is not sufficient to make the buffer losses negligible with respect to the analog ones and to keep the overall hit losses below 1%. The first optimisation addressed has been modifying the implementation of the ToT memories ($4\times4\times7$ bits): previously edge-sensitive flip flops have been replaced with level-sensitive latches, achieving a significant area utilisation decrease ($\sim$10%). In this case, the replacement did not cost additional logic or timing complication, whereas for the timestamp memories (also involved in the trigger matching and data readout) this approach was discouraged and not further investigated as it could complicate the timing and readout. It should be underlined that the

Table 3.5: Area utilisation reduction achieved with a latch-based implementation of the ToT memories.

ToT memories implementation	Single cell area reduction	Overall digital area utilisation (DFE flavour)
flip-flop (7 mem)	-	89%
latch (7 mem)	26%	80%
latch (8 mem)	26%	82%

area gain has allowed fitting an additional memory and obtaining improved buffering performance, still keeping area utilisation of ∼82%. A summary of the area utilisation results for the discussed design variants is reported in Table 3.5. Even if the area increase caused by a memory location is limited, no other location was added to keep the same number of memories on both FE integrated with the DBA. Indeed, additional area margin is needed for the second front end flavour (LFE), whose size is the maximum allowed from specifications ($35\,\mu$m×$35\,\mu$m). The DFE, initially used during the design optimisation, is instead slightly smaller ($34.71\,\mu$m×$32.44\,\mu$m) and this leaves more space for the digital logic. Even few micrometers on the FE size, then grouped in analog islands, have a non-negligible effect on the digital part. Moreover, at this stage of the design no extremely slow logic corner was considered for meeting timing during implementation and this is further addressed in Chapter 5. These reasons have motivated to put more effort into reducing area utilisation. The actual results of the final designs will be summarised in Section 3.2.3.3.

Evaluation of a compact 4-bit latch for ToT storage One of the design techniques which has been considered further on to reduce digital area has been the integration of a full-custom 4-bit multi bit latch, implementing one complete ToT memory. The block was obtained by gathering together in a unique layout 4 latches from the technology library and by merging common logic and removing single latch output inverters. In order to allow routing on the bigger cell, the designer has used routing resources in addition to M1 (i.e. up to M3). The size of the compact multi-bit latch made of 4 of the negative-level sensitive latches used for the ToT storage are summarised in Table 3.6,

Table 3.6: Area reduction achieved with the 4-bit latch full-custom block.

Cell	Cell area reduction	Overall digital area utilisation (LFE flavour)
1-bit latch (8 mem)	-	87%
4-bit latch (8 mem)	52%	80%

together with the area gain obtained after Place&Route (P&R). Although area utilisation has improved significantly, in the final implementation for RD53A multi-bit latches have not been included. The design has indeed shown to be not area but routing limited. The compact layout using higher level metals than standard cells introduces routing congestion issues which complicate the final design stage. Even if not evaluated during the RD53A development, this suggests that a 2-bit latch or a more compact 1-bit latch, possibly requiring only metal 1, may be a more efficient trade-off between area utilisation and routing congestion. Moreover, floorplan reconsiderations for more efficient use of available routing layers could be also very effective (if more metal layers available for digital routing in the array).

Pixel region size optimisation As far as the buffering performance are concerned, an additional design optimisation has been implemented concerning the pixel region shape. In particular, simulations performed by means of VEPIX53 with Monte Carlo data, have shown that an elongated pixel region shape is to be preferred with respect to a square one. This has been studied in the case of the DBA architecture, i.e. for a region made of 4 pixels. Simulation results for multiple pixel sizes and positions in the barrel of the detector are compared in Table 3.7 for a 2×2 and 4×1 pixel region (where the size is expressed as $z \times \phi$). The simulated matrix is made of 4×64 pixels and the trigger latency buffer size is fixed to 8 locations for monitoring overflow hit loss. It should be highlighted that pixel dead-time was neglected during the simulations in order to maximise buffer overflow probability. The reported results are therefore pessimistic and represent a worst-case scenario in terms of buffer hit losses, whereas dead-time losses have not been included in the study, as independent from the PR shape. It can be concluded that 4×1 PR shape brings a

Table 3.7: Buffer performance improvements thanks to a 4×1 PR pixel region shape.

Pixel size μm^2	Portion of barrel	2×2 PR hit loss	4×1 PR hit loss	Hit loss Δ
50×50	center	0.76	0.57	-0.18
	edges	1.87	0.60	-1.27
25×100	center	0.78	0.44	-0.33
	edges	2.20	0.54	-1.66

relevant gain in terms of buffering performance, especially at the edges of the barrel, both with square pixels (approach foreseen by ATLAS) and elongated ones (approach foreseen by CMS). Therefore, in the RD53A implementation of the DBA architecture, the pixel region shape has been changed to 4×1. Thanks to the flexibility of the adopted analog island concept, the change of pixel region shape has only required a few modifications to the RTL code and P&R scripts.

3.2.3.2 Centralised Buffering Architecture

The limits of the architecture implemented in CHIPIX65 in terms of simulation metrics have been addressed for the design of RD53A. In particular, the main optimisations implemented by the design team have allowed:

- the overcoming of the pixel fixed dead-time issue (source of high dead-time losses for standard asynchronous FE designs);
- an increase on the number of ToT values saved per event for each 16-pixel region from 6 to 8.

The resulting pixel region architecture is shown in Figure 3.16. It can be noticed that the main difference with respect to the DBA lays on the pixel region size and on the approach used for the ToT buffering, whereas the trigger matching logic, arbitration and memories are using the same scheme of the previous architecture. The fixed-pixel dead-time problem has been solved by introducing additional levels of buffering. A waiting time is still required before the hit information can be stored in the centralised buffer, as all pixel need to have finished processing. The waiting time is different depending on the FE

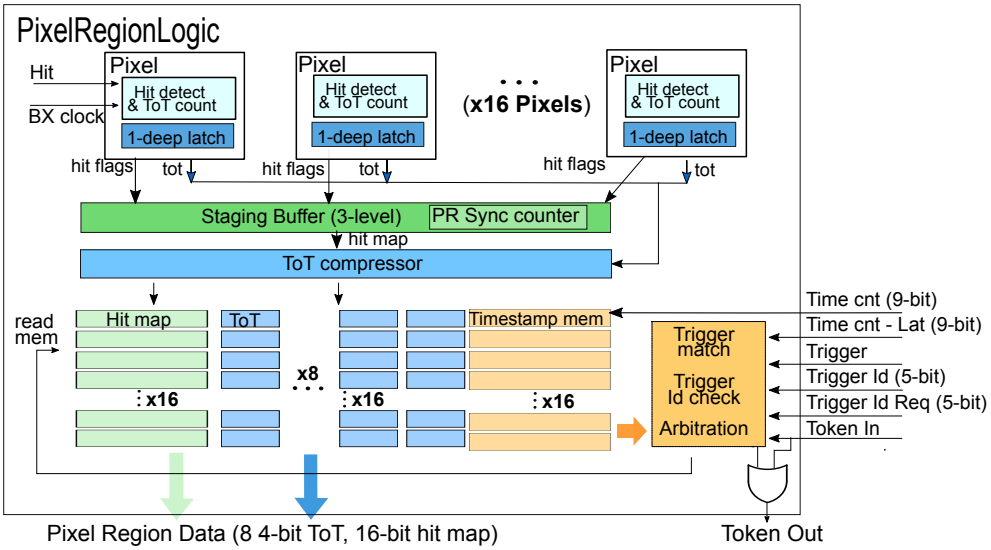

Figure 3.16: Block diagram of the PR logic of the CBA architecture.

mode (40 MHz clock or faster ToT measurement) and it is settable by global configuration. The pixel ToT information is latched in a 1-deep ToT buffer in the pixel logic. A staging buffer has been instead implemented to store the region hit map during the waiting time. This block also includes a region synchronisation counter which replaces the previous pixel dead-time counter and it is used to synchronise the access of each pixel to the shared memory: when this counter reaches the waiting time, the hit map is propagated to the ToT compressor together with the ToT values.

As far as the pixel region shape is concerned, the 4×4 implementation has been adopted as in CHIPIX65, as no further investigations on a different approaches (e.g. elongated 8×2) has been investigated. As regards the implementation of the PR latency memories: *i)* timestamp memories are flip-flop based (the logic was adopted from the DBA and it is very similar), *ii)* the hit map and ToT memories are also flip-flop based. The second choice has been done by the designer at late stages, due to difficulties encountered in properly constraining the design to meet timing at top level with a latch-based implementation. The simulation results obtained with the architecture will be reported in Section 3.2.3.3 together with the ones of the DBA.

3.2.3.3 Simulation performance results

The simulation performance of the optimised RD53A architectures for the digital matrix are reported in Table 3.8. In addition, implementation metrics are reported for a more complete comparison, even if treated in the following chapters. The metrics defined are similar to the ones previously shown in Section 3.2.2.1 for the FE65-P2 and CHIPIX65 architectures:

- inefficiency i.e. hit loss, achieved through simulation of non-synthesized RTL (within VEPIX53);
- area utilisation (comparison of place-and-routed architectures);
- power consumption (comparison of place-and-routed architectures in the typical corner, i.e. process: TT, voltage: 1.2 V, temperature 25°C).

The hierarchical block simulated for the DBA and CBA comparison is a 8-pixel wide column with full chip height, i.e. a total of 8×192 pixels. Each of them corresponds to a pixel core column in the RD53A chip. The common simulation conditions have been:

- mixed Monte Carlo and constraint random hits to achieve the target specification of 3 GHz/cm^2;
  - Monte Carlo data characteristics: provided by CMS, with 140 pile-up, 50×50 μm^2 pixel size for the centre of the barrel and 25×100 μm^2 for the edges (pixel size choice foreseen by CMS);
  - constraint random hits characteristics: featuring the same pixel charge distribution and similar cluster shapes as extracted from the Monte Carlo data;
- charge-to-ToT conversion function described in Section 2.2.3.2;
- simulation run for 500,000 bunch crossing cycles, i.e. 12.5 ms (simulation time ∼3h30m);
- trigger latency: 12.5 μs;

- trigger rate: 1 MHz.

It should be underlined that the fast-mode of the SFE has not been taken into account, as it does not provide further insight on the digital architectures and it has been only integrated with one of them. Therefore, simulation results are obtained using the SFE in its non-fast mode. It is anyway obvious that a faster ToT counting (as also seen in Section 3.2.2.1), can allow a reduction on dead-time losses.

Table 3.8: Comparative table between centralised and distributed buffering architectures.

Metrics		DBA		CBA	
Inefficiency (%)	Sources	Center	Edges	Center	Edges
	Dead-time (analog)	0.73	0.99	0.73	0.99
	Dead-time (digital)	0.18	0.19	0.36	0.37
	Latency buffer	0.10	0.10	0.09	0.13
	Total hit loss (single pixel)	1.01	1.28	1.18	1.49
	Limit on number of ToTs (only ToT info)	-		0.13	0.22
Area Utilisation		LFE	DFE	SFE	
%		88	82	87	
Average Power (μW/pixel)		4.7		6.19	

In terms of inefficiency results, the DBA and CBA architectures integrated in RD53A feature similar losses, which are close to specs if the charge-to-ToT curve is chosen to limit dead-time losses. The latter are higher at the edges of the barrel unless a different conversion function is used, since Monte Carlo data show in average a higher charge per pixel (for the pixel size: 25×100). The digital logic has also an impact on dead-time, in particular for the case of the CBA (2 clock cycles versus 1 clock cycle of the DBA). Latency buffer overflow achieves an order of magnitude lower losses than the analog part, as required. Moreover, it can be noticed that the DBA architecture profits from the implemented elongated pixel region shape, as latency losses are the same in the

different portions of the detector. It is evident that inefficiency is dominated by dead-time at the high hit rates of operation and that a certain charge-ToT distribution may need to be chosen, possibly penalising physics considerations (e.g position and charge resolution). To this end, the adoption of a faster ToT measurement (i.e. SFE) is a valuable solution. On the other hand, the latch able to turn itself into a fast oscillator needs to be continuously clocked with the 40 MHz clock. This custom digital cell in the analog macro is powered digitally for noise reasons and therefore causes a non-negligible power overhead in the digital domain (higher than 1.5 μA@1.2 V/pixel, similar in standard/fast mode and with/without hit activity). Instead, the analog power consumption is below specs (3.3μA/pixel with a requirement of 4). This contribution to the consumption in the digital domain needs to be added to the values on the table (for the CBA, since the DBA is not integrated with the SFE and the result is not straightforward). Therefore, the trade-off between the power budget and the dead-time losses should be studied for future implementations. In alternative, an interesting and simple solution to be investigated is the use of a 80 MHz ToT counting scheme, implemented in the digital logic with a double-edge ToT counter, which would not require any analog fast oscillator. Even if this was not implemented in the context of the RD53A prototype, it will certainly be studied for final chips. As far as implementation metrics are concerned, the DBA and CBA architectures integrated in the RD53A have rather similar area utilisation, with the main difference being the FE size (LFE: 35 μm$\times$35 μm, DFE: 34.71 μm$\times$32.44 μm, SFE: 35 μm$\times$33.2 μm). Nevertheless, the area of the CBA can be reduced by using a latch-implementation of ToT memories, which was not implemented in RD53A. Instead, the digital power consumption of the pixel array is $\sim$30% higher in the case of the CBA (further details are discussed in Chapter 4).

Validity checks and simulation warm-up In order to cross-check the validity of simulation results, whenever possible it is appropriate to compare them with statistically expected ones. While an analytical calculation is not trivial in the case of latency buffer inefficiency, it can be carried out for dead-time losses, with the assumptions of pixels being independent (hit rate per

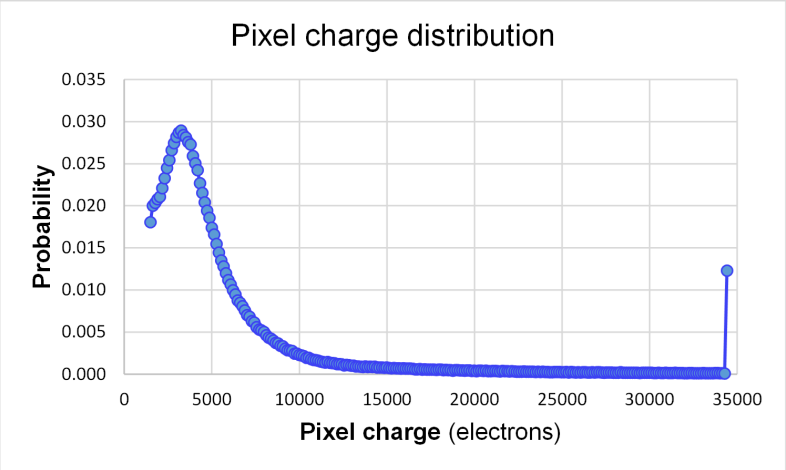

Figure 3.17: Pixel charge probability distribution of CMS Monte Carlo data in the center of the barrel (pixel size $50{\times}50\,\mu\text{m}^2$).

pixel: 75 kHz, clock period: 25 ns). An example is herein reported, in the case of the simulations with Monte Carlo at the centre of the barrel. Assuming a Poisson distribution for arrivals and a paralyzable system model for the pixel dead-time (as described in Section 1.1.2.1), the probability of a hit being received within a N-clock-cycle dead-time is: *Pr(2nd hit before or at the Nth clock cycle) = (1-e$^{-N\cdot cycle\,period\cdot rate}$)* [14], [81]. The pixel charge probability distribution from the data is known and shown in Figure 3.17. Combining it with the defined charge-to-ToT conversion function (i.e. ToT correspondent to each charge), it is possible to calculate the probability of any hit being received within the combinations of dead-time, i.e. the dead-time inefficiency:

$$\sum_{N=1}^{\infty}[Pr(N)\cdot(1-e^{-N\cdot 25\,ns\cdot 75\,kHz})] = 0.77\%. \tag{3.1}$$

The analytical result is in accordance with the simulation one, taking also into account that the analytical description does not model clusters and considers each pixel independently. This basic validation phase is reported to highlight its importance, especially in the case of externally provided data, imported in the framework with a defined strategy (therefore not directly randomised). By performing such cross-checks, correlations effects were found, which caused

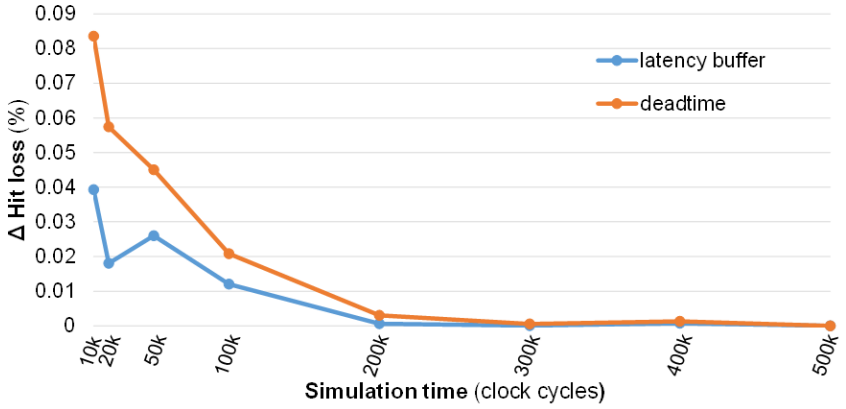

Figure 3.18: Absolute difference (Δ) of hit loss percentage results with respect to value measured at the end of the simulation, both in the case of dead-time and latency buffer overflow. On the x-axis, k stands for a factor of 1000.

systematically higher dead-time losses ($\sim$+30% for the case shown). This was due to the fact that the set of events imported was too small. The problem was solved for the reported results, without asking for additional Monte Carlo events, instead by randomly shifting the simulated sub-matrix (8×192 pixels) with respect to the full-matrix provided with Monte Carlo data.

Moreover, as far as the simulation time is concerned, it has also been cross-checked that the chosen duration is sufficient for statistics collection. As it can be seen in Figure 3.18, both hit loss results have converged (with extremely low, i.e. 0.01%, difference with respect to the final % hit loss value). This shows that the impact of the simulation warm-up is not any more visible and simulation results are meaningful. It can be noticed that a lower number of simulated bunch crossing cycles (e.g. with $\Delta < 0.1$-0.05%) would most likely represent an acceptable compromise between uncertainty (5-10%) on hit loss estimation and simulation time.

Chapter 4

Low-power methodology and optimisation for operation with serial powering

In the last decades, low power circuit design has been a vital issue in VLSI design since the system feature size has shrank gradually and clock frequency has increased rapidly. Power density has become a prime concern for system designers for multiple reasons (e.g. heat removal and challenging power delivery networks). Nowadays, with scaling of voltage and current that is reaching its limits, power density is determining an interruption of the increasing clock frequency trend in high performance electronics (e.g. microprocessors). The demand for power-sensitive design has also grown significantly in recent years due to tremendous growth in portable applications [95]. Consequently, the need for power-efficient design techniques has increased considerably in the past decades and has been targeted to the different applications, from high-performance complex systems to portable systems.

In the context of the phase 2 pixel upgrade at HL-LHC, a serial power scheme is foreseen to power thousands of modules (each consisting of multiple chips) and this requires project-specific considerations with respect to the state of the art low power design techniques. Proving the feasibility of serial powering is itself one of the main goals of the RD53A prototype.For these reasons,

achieving low power is considered a critical challenge and it needs to be tied to the simulation framework and the design methodology. Due to the very high number of pixels integrating large amount of logic, the pixel matrix plays the major role in the overall power consumption of the chip. Moreover, the digital logic can also introduce power fluctuations which could possibly couple into the analog power supply or cause variations on the effective threshold of the pixels. These aspects are herein addressed: the concept of serial powering and its motivations are introduced in Section 4.1; a critical review of state of the art low power techniques for the purpose of this work is performed in Section 4.2; the defined power methodology and initial results obtained for prototyped pixel chips are shown in Section 4.3, whereas Section 4.4 reports the final results of the low-power optimisation of the pixel array logic for RD53A and further considerations.

4.1 Serial powering concept and motivations

A classical passive parallel powering scheme with a constant voltage, as present in the current LHC pixel systems, has numerous problems which exclude its use for the phase 2 pixel detector upgrade. This can be explained from a combination of multiple factors: *i)* the electronic circuits must be powered with high currents (approx. 2 A), to cope with the very large number of pixels and the high hit and trigger rates, and at the low voltages (1.2 V) used by modern CMOS technologies to meet the density requirements and fast operation of the detector; *ii)* the power is transmitted over a long distance since power supplies are located outside the active detector volume; *iii)* the power cables must be low mass to minimise interactions of particles with the material, which compromises physics analysis. The combinations of these factors would cause power losses in the cables to exceed the actual power consumption of the electronics. The further the granularity increases for new generation detectors, the more power cable losses become critical and not sustainable. A second possible powering option, i.e. the use of local DC-DC power conversion within the pixel volume, has also been excluded because of the extremely high radiation and magnetic field levels, combined with very tight constraints on space and mate-

rial budget. Therefore, from a certain granularity onwards serial powering has been found to be the only viable solution to supply the inner tracker with the required power, within an acceptable material budget and power cable losses. In support of this approach, the ATLAS pixel community has already experimentally proven the feasibility of the scheme with previous generation pixel chips as FE-I3 [96] and FE-I4 [97], even though it has not been installed in the experiment during the Insertable b-Layer (IBL) upgrade due to the limited time available. Additional testing with FE-I4 modules have been performed to validate the powering scheme for the ATLAS and CMS pixel detectors at the HL-LHC [98].

A serial powering scheme consists in powering a chain of pixel chip modules in series with a constant current. In this way, the power supply current is "re-used" among multiple loads connected in series, in contrast to a classical parallel powering scheme where the supplied voltage is shared among loads and the injected current is only used once. The fact that the injected power supply current is re-used multiple times (n times) in a power chain reduces significantly the total power losses in the power supply cables (factor n less cables). A basic comparison of serial powering across modules with a full parallel scheme is shown in Figure 4.1:

- within a module, chips are powered in parallel: this solution both profits from the serial powering advantages and allows chips in the same module (connected to the same sensor) to be at the same potential;
- even lower material can be achieved by using lighter power cables [36], since higher voltage losses can be tolerated;
- the peculiarity of the scheme is that it is designed for a certain constant current consumption: the pixel modules cannot consume more (digital power variations could otherwise cause chip failure);
- if the consumption is lower, a dedicated circuit is needed to dissipate the "surplus" power, in order to maintain the serial power current constant.

A specialised local shunt regulator is required to convert the injected current into a stable local voltage supply for the pixel chip. This local regulator must

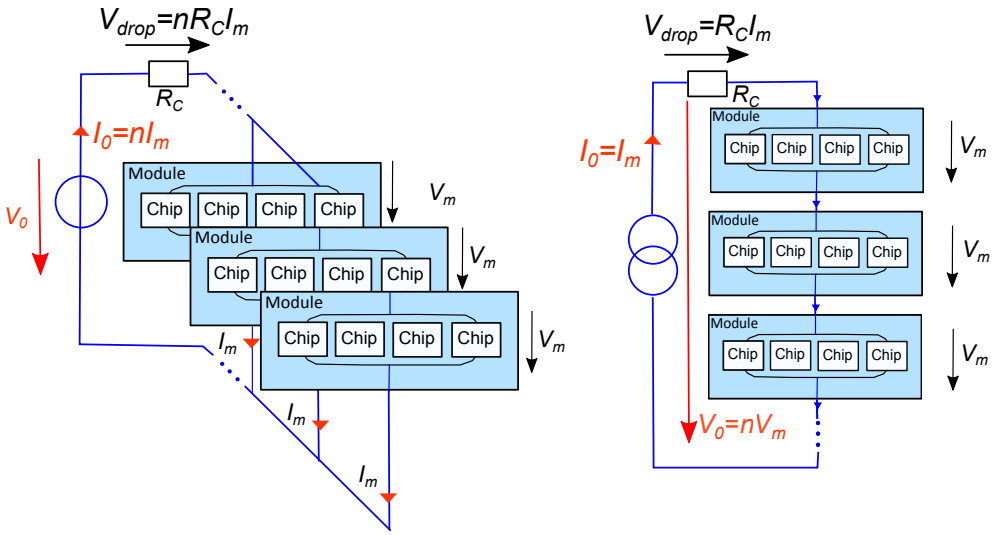

Figure 4.1: Power cable losses in parallel and serial powering. V_0 and I_0 are the supply voltages and currents, while V_m refers to the voltage across a module and I_m to the module's current. The number of modules is indicated by n [36].

also assure that any dynamic load variation is not visible from the outside of the shunt regulator. A dedicated combined Shunt and Low Drop Output (SLDO) voltage regulator has therefore been developed [99] for integration on the RD53A chip. It is composed of a Low-DropOut (LDO) regulator generating the low supply voltage and a shunt regulator consuming the current not drawn by the load. In particular, two SLDOs are integrated in the RD53A chip to power the digital and the analog domains of the chip separately (see Figure 4.2), in order to minimise noise coupling from the digital logic to the noise sensitive analog front-ends. Even if they are independent, the two SLDO are both powered in parallel from a single common power loop (to assure that they are at the same potential level).

4.1.1 Design challenges for low-power

While the analog domain normally features a rather constant power consumption, the digital power can have large fluctuations within the clock cycle and across clock cycles depending on the logic activity. Such digital power varia-

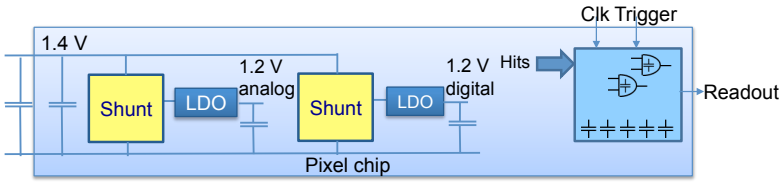

Figure 4.2: Block diagram of a serial powered chip with integrated regulators for analog and digital domains.

tions constitute a main worry as they could cause chip failure, if higher than the current provided to the serial power chain, as sketched in Figure 4.3. More-

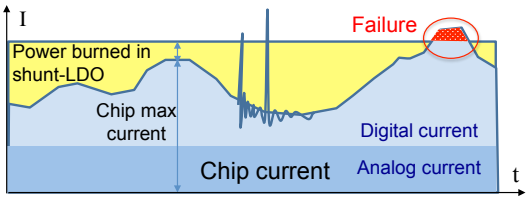

Figure 4.3: Sketch showing the effect of power variations in a serial powering scheme.

over, dynamic current variations in modern CMOS circuits are extremely fast and it is in practice not possible to dynamically adjust at system level injected power supply currents to match dynamic load changes. A serial power loop has indeed significant inductance and cannot support fast current changes [36]. Therefore, it is vital to feed enough current to the loads to comply with the maximum current needs, including dynamic current variations due to the activity of the digital logic. Given that the highest possible load current is injected, the shunt regulator is then in charge of maintaining the total current constant independent of the actual current consumed by the load.

These considerations need to be taken into account for the low power optimisation of the digital logic and a dedicated analysis is required. The power metric which needs to be defined is the maximum current consumed by the chip during operation. For this purpose, it is fundamental to consider that current variations will be filtered by on-chip and off-chip decoupling and will be not visible to the power delivery circuits. Moreover, the choice of low power design techniques is not straightforward: state of the art low power techniques are

mainly meant to minimise average power consumption and do not necessarily help in minimising maximum current. In this work, the optimisation goal is not only to reduce average energy consumption but also to quantify and limit digital logic power fluctuations as much as possible, taking into account the impact of local decoupling on power variations.

4.2 State of the art and selection of low power design techniques

Transistor scaling in the last decades has not only enabled performance improvement, but also caused significantly increased power density due to higher integration. Therefore, at the state of the art of VLSI circuits several design techniques have been proposed to reduce different sources of power consumption [100]. Total power in a CMOS technology is given by [101]:

$$P_{total} = P_{switching} + P_{short-circuit} + P_{leakage} \tag{4.1}$$

where $P_{switching}$ is due to charge and discharge capacitances; $P_{short-circuit}$ is dissipated by the instantaneous current between the supply and ground during a switch of state; $P_{leakage}$ is a combination of parasitics currents of the CMOS device, which are present whenever it is on, regardless of its activity. The first two are also referred to as dynamic power, whereas the last as static. Identifying the main source of power consumption for the considered application is essential to guide power optimisation and choose the most appropriate design techniques. For this reason, power consumption has been broken down into its different components: internal power (43%), switching power (56%) and leakage power (1%). Compared to power factors summarised in (4.1), internal power is reported from implementation design tools based on multi-variable models of standard cells which include both $P_{short-circuit}$ and $P_{switching}$ of only internal nodes. This first report is based on the FE65-P2 digital pixel array consumption, for which details will be given in the Section 4.3.2. It is evident that power consumption is dominated by the dynamic component, due to the high rates and continuous operation in nominal conditions. Therefore,

power reduction techniques discussed are mainly targeted to dynamic power optimisation.

Potential highly effective design techniques for dynamic (and at the same time leakage) power reduction, such as multiple supply voltages and dynamic scaling of voltages and frequency [102], cannot be used for the RD53 chip. Because of system considerations, it is planned to have only one digital supply (1.2 V) and frequency is also fixed to 40 MHz in the pixel matrix, to stay synchronised to the LHC system frequency. It should also be highlighted that such a chip will be used in a very hostile radiation environment, causing considerable performance degradation and bit upsets. For this reason the voltage supply has been chosen to limit performance degradation after radiation and simplicity of the design has been preferred. In case radiation effects will be proven to be less critical than what it is currently expected, lower voltage for the digital logic (1-0.8 V) may be tested for potential use in future versions. It should anyway be highlighted that with the adopted serial powering scheme, the power gain which can be obtained by reduced supply voltage is not quadratic as for standard powering schemes. Indeed, even if the power consumption of the digital chip logic scales quadratically, the power losses taking place in the LDO increase due to the higher voltage drop across it, as shown by the example in Figure 4.4. The overall power budget savings scale with an approximatively linear behaviour with respect to the reduced supply voltage. It is also not foreseen to mix high-V_t cells (featuring lower leakage power) with standard-V_t ones, since they experience more degradation due to radiation [103] and leakage power is negligible in this application. Moreover, power gating has not been adopted in the design for multiple reasons: separate handling of each pixel would be required, as the activity is randomly spread over the matrix (for such a goal not enough area, routing and logic resources are available); system requirements do not allow for sleep times which would cause data losses above specifications.

An effective design technique to reduce dynamic power in digital circuits is clock gating [104], i.e. masking the clock to synchronous circuits during idle state, in order to avoid unnecessary switching. Since the hit rate per pixel (75 kHz) is significantly lower than the clock frequency, this design technique

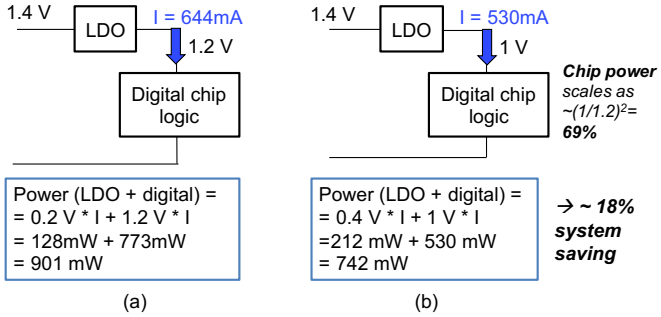

Figure 4.4: Power estimations of the power budget improvement obtainable with a reduced power supply for the digital domain. For the overall power gain both the digital chip and LDO power consumption are considered. Estimation are based on the FE65-P2 architecture scaled to a full-size chip with 160,000 pixels. For both cases a TT process at 25 °C is assumed, with technology libraries at different voltages 1.2 V (a) and 1 V (b).

is of particular interest in the context of this work. It was also adopted in FE-I4, i.e. for pixel chip detectors being currently used in LHC. It also has the "side effect" of causing large power variation as when there are no hits the power is very low, while as the hit occupancy increases so does the power. Thus it may be preferable for the chip to consume digital power all the time, even when there is no occupancy, in particular for a serial powering scheme. These sets of considerations/questions are studied in this thesis in order to optimise the design. Clock gating choices are one of the aspects concerning the optimisation of the local clock distribution, as clock gating cells propagate the clock to the sequential logic. Since the clock tree cells switch at maximum rate of the application and have the largest capacitive loads, placement or clock tree constraints also play a role in the power budget and must be considered.

4.3 Power analysis methodology and results of a prototyped digital architecture

Digital design tools have been used to generate power models starting at RTL and moving to detailed post-layout power estimation with parasitic annotation. The design flow shown on the left hand side of Figure 4.5 and the software

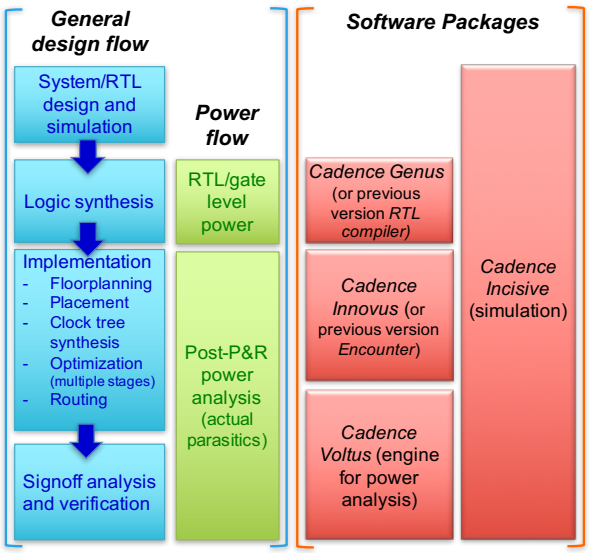

Figure 4.5: Digital design flow and Cadence software packages used for the design, power analysis and optimisation.

packages on the right hand side have been extensively used for the design and parallel power analysis and optimization. The main steps of the related power methodology are herein summarised:

1. power analysis after synthesis to gates (without full layout information), is used to drive substantial architectural choices before going into a full detailed design;

2. detailed and accurate post P&R power analysis and optimisation including:

   (a) average power estimations under different activity conditions to assess power impact of different factors, important to understand variations in different operation modes and guide design choices;

   (b) dynamic power variations analysis under the variety of operating conditions and at different time constants (1 ns, 25 ns, 100 ns, 1 μs, 10 μs), to emulate the impact of filtering on power variations performed by local decoupling;

3. application of the analysis methodology to drive detailed design choices.

For the application of interest, it was essential to perform simulations under realistic operation conditions in order to accurately estimate power and its variations (to be capable of proving the reliability of the serial powering scheme). With the high hit and trigger rates the dynamic power consumption due to high activity has indeed been seen to be the dominant contribution. To this purpose, the power analysis has been integrated with the high level SystemVerilog-UVM simulation framework VEPIX53, capable not only of generating the proper input stimuli but also to simulate the Design Under Test (DUT) up to detailed post-layout netlist. The resulting full activity can be provided to the power analysis tools for accurate power predictions.

4.3.1 Power estimation for architectural choices

Power analysis at gate-level, i.e. after synthesis to gates without layout parasitics information, is useful to drive substantial architectural choices before developing the complete and detailed design. In the context of RD53, a key design choice is related to the use of clock gating technique, since it is a source of power variations. As anticipated in Section 4.2, its use was initially discouraged to keep power as constant as possible. A 4×64 pixel array based on the FE65-P2 prototype, has been synthesized by means of the *Cadence RTL Compiler* tool. Simulations of the obtained netlists (i.e. with and without the implementation of clock gating) have been run within VEPIX53 under $3\,\mathrm{GHz/cm^2}$ hit rate, 1 MHz trigger rate and $12.5\,\mu$s trigger latency. A power profile showing power variations averaged over a $1\,\mu$s time scale is shown in Figure 4.6. It has been obtained by means of a defined iterative algorithm which instructs power reports in sequence over small time windows. As it will be described in more detail in the following sections, such a time scale ($1\,\mu$s) emulates the effect of the on-chip decoupling and required SLDO decoupling. A significant increase (∼x5) in power consumption is seen when excluding any form of clock gating in the architecture and cannot be tolerated in pixel detectors. Moreover, power variations observed at this initial stage of the design flow (excluding accurate parasitics information) are not particularly critical.

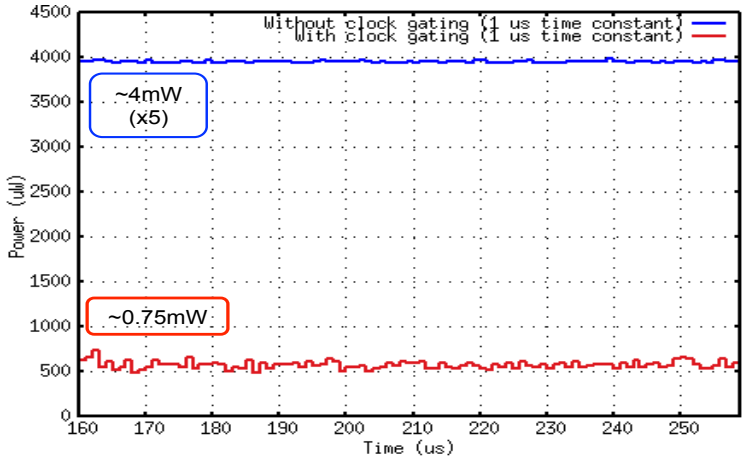

Figure 4.6: Gate-level power profiles of small 4×64 pixel matrix for clock gating evaluation (in red for the design featuring clock gating, in blue for the same without it) [105].

4.3.2 Post-layout power analysis

A more detailed power analysis is necessary to provide initial specifications to the powering system, whereas gate-level analysis has shown around 50% underestimation due to limited modelling of parasitics and clock tree. For this reason, the implementation flow with the *Cadence Encounter/Innovus* tool has been advanced to the post P&R stage. Parasitics have been extracted in the form of the Standard Parasitic Exchange Format (SPEF) file and the more detailed post P&R netlist has been simulated by means of the VEPIX53 framework to annotate activity in a Value Change Dump (VCD) file. The latter not only contains average switching activities over the whole simulation, as for lighter Switching Activity Interchange Format (SAIF) and Toggle Count Format (TCF) files, but also its detailed evolution over time, important to study power variations.

At first, average power estimations have been obtained under multiple hit and trigger rate conditions in order to assess the power impact of different factors and guide design choices. A summary, where results are given per pixel and also scaled to the full pixel matrix (assuming 400×400 pixels initially foreseen for both ATLAS and CMS experiments) is reported in Table 4.1. Pre-

sented results are based on the technology typical corner (TT, 1.2 V, 25 °C). The activity conditions included are: extreme hit and trigger rate as described in Section 4.3.1, high hit rate and trigger absence (to decouple hit and trigger effect), without hits and without triggers (i.e. just clocking the logic). It can be highlighted that the power consumed by the clock tree, including both global and local clock delivery, is dominant. This is mainly due to a combination of high switching activity of the clock and high total load of clock buffers, with many registers as well as interconnects. Power variations of the reference

Table 4.1: Average power results for the typical corner at 1.2 V under a variety of activity conditions.

Activity conditions	Single pixel power (μW)	Full chip power (W)
with hits (3 GHz/cm^2) and triggers (1 MHz)	4.84	0.774
with hits (3 GHz/cm^2) and without triggers	4.7	0.752
without hits and without triggers	3.81	0.61

4×64 pixel array have also been studied at post P&R stage by extensive power profiling under a variety of operating conditions. Two relevant and opposite examples, i.e. power profiles produced with extreme hit and trigger rate conditions and with only the clock feeding the logic are reported in Figure 4.7. In these plots, power peaks are evaluated at different time scales (1 ns, 25 ns, 100 ns, 1 μs, 10 μs): absolute peak value and percentage increase with respect to average power are highlighted. This study allows to investigate the impact of decoupling seen from the chip to the serial power network, which acts as a low-pass filter to current variations. Digital power variations are very high within the clock time period (25 ns), but they get much smoother after averaging over 1 μs. Even in the case of high hit and trigger rates, variations at this time scale are limited within 20%. In the plot at the bottom of Figure 4.7, variations are already filtered at short time scales, since digital activity is stable. These results have been used as an input to SLDO simulations to verify its functionality and demonstrate the reliability of the serial powering scheme.

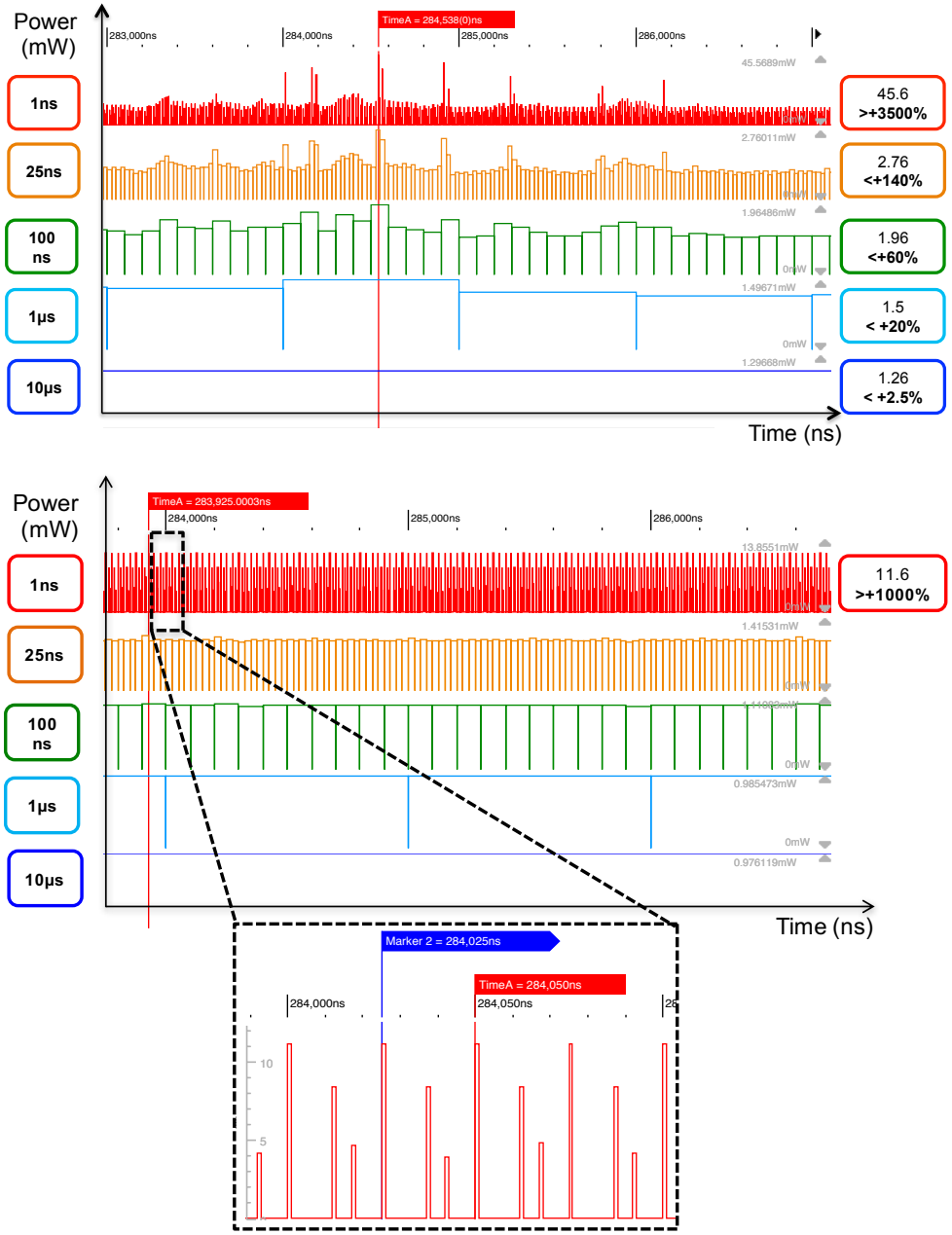

Figure 4.7: Power profiles of a 4×64 pixel array at different time scales: at the top with high activity (3 GHz/cm^2 hit rate and 1 MHz trigger rate) and at the bottom with low activity (only clocking digital logic) [105]. A zoom shows the highest peaks in correspondence of the clock edges, above all the rising one.

4.3.3 Validation of serial powering approach with digital power profiles

The topology shown in Figure 4.8, made of two serially powered modules, each composed of four chips, was simulated based on the detailed SLDO design. The chip in red represents the one with simulated digital activity, the green coloured chips are its neighbouring chips within the same module and the light blue the neighbouring module in the serial power chain. Each chip was simulated as a pair of SLDOs for analog and digital operated in parallel. In the case of digital active chip (red), the load was simulated as a current sink based on VCD files extracted from the power profiles shown in Section 4.3.2 (with scaling to a full-size chip of 160,000 pixels). In the other cases, the load was simulated as a constant current sink of 800 mA (assuming for simplicity 5 μW per pixel for a voltage of 1 V). As shown in Figure 4.8, local decoupling capacitances (chip, power grid, input/output SLDO capacitors with equivalent series resistance), parasitic inductances (wire-bonds, cabling), resistances and capacitances (pads) were also included in the simulation. The impact of the digital activity of a chip on the regulated output voltages was studied. A maximum limit of 10% and 1% for the digital and analog domain, respectively, were considered to be acceptable without compromising functionality and performance. The digital activity of a chip was simulated for the extreme case with maximum peaks (1 ns resolution) in Figure 4.7, in order to confirm the effect of decoupling capacitance. As shown in the top plot of Figure 4.9:

- in the digital domain, a variation of less than 100 mV is noticed for the active chip itself and less than 10 mV for the voltages of the rest of the chips on the chain;
- in the analog domain, the digital activity of one chip causes a variation of less than 1 mV in the rest of the module, while the impact on the rest of the serial power chain is negligible.

Overall, the performance of the SLDO regulator with a digitally active load is demonstrated to be within acceptable limits. The presence of local decoupling is proven to filter short power fluctuation which get averaged over the μs

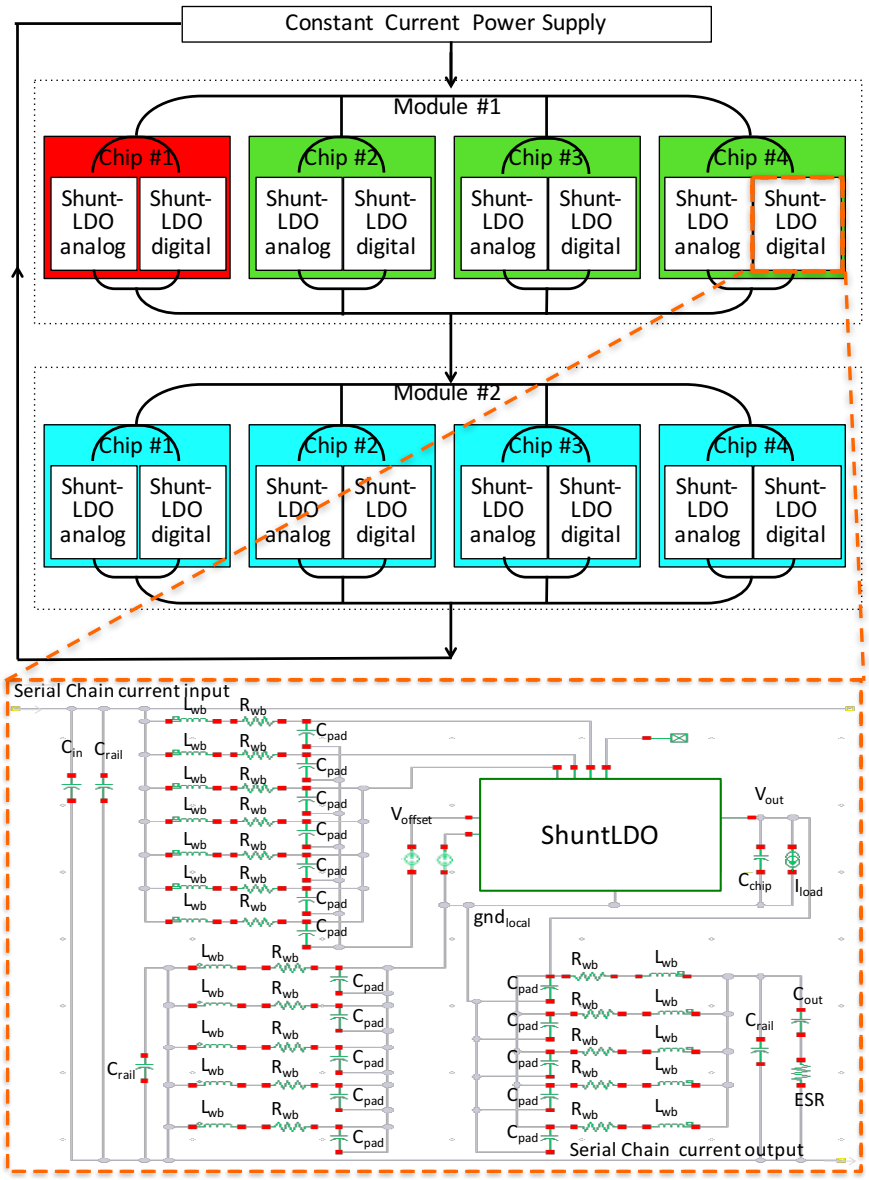

Figure 4.8: Serial powering topology: two modules powered in series with the four chips within a module and the two SLDOs per chip powered in parallel. Detailed schematic of the basic unit is also shown.

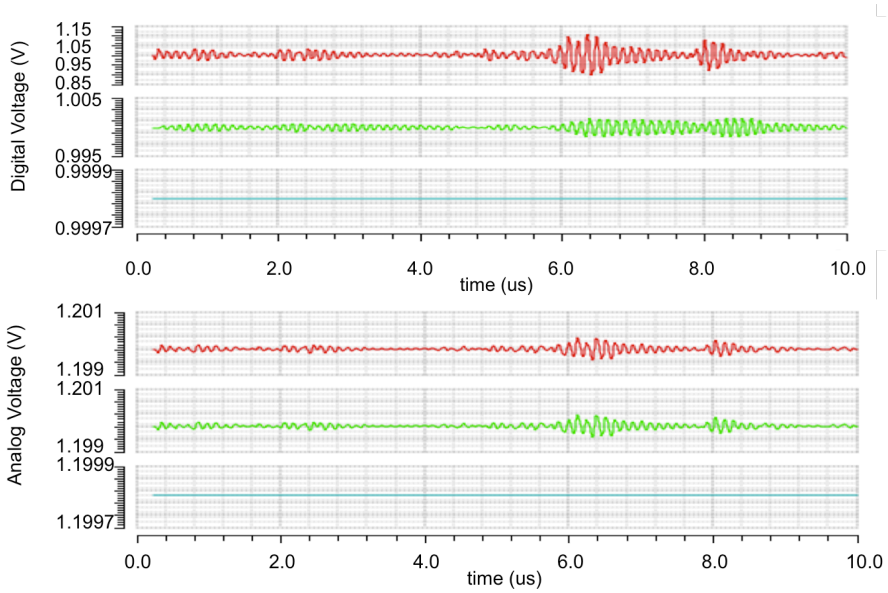

Figure 4.9: Impact of the digital activity of a chip to the digital power domain of the chips in a serial power chain [105].

timescale. This effect assures stable operation of serially powered modules.

4.4 Low-power optimisation of the pixel array logic

The described methodology has been adopted to assess power performance of the digital array logic throughout the design process. As already discussed in Section 3.2.3, the focus of the work has been the DBA array logic. Some of the techniques herein reported have been adopted also by the CBA, as it will be mentioned. The design optimization has clearly not only addressed low power, but a trade-off among design metrics such as area, hit efficiency and power. Therefore, in the following, design choices are always taking into account a combination of the three.

4.4.1 Evaluation of architecture variations

With respect to the initial FE65-P2 architecture, a set of changes have been evaluated at early stages of the RD53A implementation in order to improve

the trade-off between design metrics. The building block under study is a RD53A 8x8 pixel core, integrated with the DFE. A summary is reported in Table 4.2, where area utilisation, average power and peak power per pixel, with averaging over 1 μs time scale, are shown. It can be highlighted that the peak power design metric has been defined in order to address requirements of the serial powering scheme, as described in the methodology in Section 4.3. In

Table 4.2: Results for the typical corner at 1.2 V on average power consumption, peak power (averaging at 1 μs time scale) and digital area utilisation for different pixel architectures.

Architecture implementation (identified by #number)	Average power per pixel (μW)	Peak power per pixel (μW)	Area Utilization
#1 ToT storage: flip-flops, 7 mem	4.8	6.26	89%
#2 ToT storage: latches, 7 mem	4.98	5.8	80%
#3 no ToT counters	5.6	6.54	85%
#4 ToT counters, 7 mem, synch readout	4.84	5.6	80%
#5 additional memory (8 mem)	5.2	6.1	82%

the initial implementation (case #1), the ToT information was calculated with 4-bit counters, stored in flip-flops and read through asynchronous readout. A reduction in peak power per pixel has been successfully obtained by storing the ToT data in latches (case #2), which has at the same time improved area utilisation, as also reported in Section 3.2.3. With the aim of reducing power fluctuations, a different approach for ToT calculation has been evaluated in case #3, with the aim of limiting power variations. In this case, ToT counting was not implemented with per-pixel counters but with local subtraction of the timestamp value at the trailing and leading edge of the incoming hit. An increase in average and peak power has been actually observed. At the rates of interest for the application, gated ToT counters have been seen to demand for less power than the additional logic needed in case #3 for the gray-to-binary conversion and subtraction of the timestamps. The solution studied has been therefore not adopted. In case #4, slightly improved results have been achieved (with respect to case #2) with fully synchronous memory readout, which is also preferable for timing constraint reasons. As it can be seen in case #5, the

area gain has allowed an additional memory to fit, which significantly reduces hit loss of the digital logic. The additional memory has implied a small power increase. For comparison with results in the following, it can been mentioned that at this stage of the implementation the technology corners adopted during the design flow were the "standard" typical, fast and slow corner (i.e. no timing pessimism for radiation degradation was considered). Nevertheless, power analysis was performed with power models including total ionizing dose effects (500 MRad) to study impact on power consumptions. Results showed less than 5% power increase and no dominant impact of leakage power induced by radiation, as expected for this technology.

4.4.2 Custom clock gating and local clock distribution choices

Clock gating has been implemented from the beginning in the RTL since significant power savings, in the order of 5x, can be obtained even after synthesis, as discussed in Section 4.3.1. The initial implementation is an RTL description of a glitch-free clock gating cell like the one shown in Figure 4.10, which the synthesiser translates into two separate cells from the library. It should be highlighted that a detailed choice on which parts of the logic to gate, based on its architecture, was required to achieve power optimisation. In particular, with respect to Figure 3.15 in Section 3.2.3, clock gating is manually performed in:

- the PixelLogic, just after the first synchronising stage for the analog input;
- the ToT counter inside the *PixelLogic*, to prevent the clock from being propagated to the counter when it is not enabled;
- the common pixel region logic for trigger matching and timestamp storage (in this case a clock gating cell is used for each memory cell and its write/read control logic, with 2%-3% lower area utilisation seen compared to automated clock gating insertion).

The use of special Integrated Clock Gating (ICG) cells, which integrate in a unique standard cell (with a predefined and optimised layout) the logic

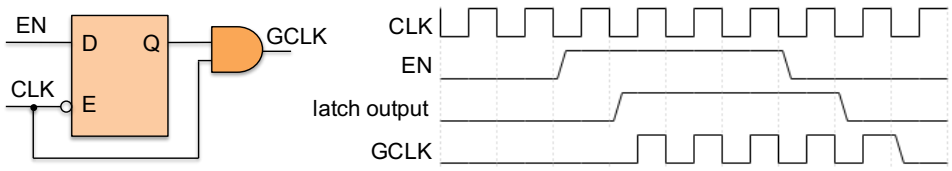

Figure 4.10: Clock gating cell including an AND cell and a negative-level sensitive latch to prevent glitches.

shown in Figure 4.10, has been evaluated. It can be noticed that they profit from more efficient placement and timing, which is reflected in improved area and power with respect to Table 4.2. Moreover, the use of automated clock gating has been considered in order to explore further possible optimisations on clock gating choices. The synthesis tool is capable of implementing clock gating by recognising sequential logic which features enabling logic. The trade-off between power saved from the reduced activity and power consumed by the additional cells has been studied. Results are summarised in Table 4.3 and the optimal one has been adopted as the baseline. In the first case the synthesiser is left free to add clock gating wherever possible as the minimum number of gatable flip flops is set to 1. Additional gating can be identified in the netlist and it is in particular located downstream of the manual ones in correspondence of: the free memory address, one bit of the state machine in the triggering logic, five bits of the timestamp memories which are not reused for the *TriggerId* check. It can be noticed that it has a positive impact on power, while it causes an expected area increase. Since using a clock gating cell to gate one single flip flip is intuitively not very efficient, the minimum number of gateable flip flops has been set to 3. As a consequence, only the memory address and the timestamp memories feature additional gating, which achieves improved power. The additional gating logic is also limited enough to not have a visible impact on area utilisation, after the optimisation stages of the implementation flow. It can be highlighted that the area utilisation increase with respect to the results presented in the previous section is mainly due to the use of the bigger LFE (so less area is available for the digital logic). Indeed, at later stages of the design for the RD53A chip, when all of them

Table 4.3: Results of the clock gating optimisation with adoption of ICG cells and additional automated clock gating with variable number of flip flops indicated in parenthesis (FF).

Clock Gating (CG) approach	Average power per pixel (μW)	Peak power per pixel (μW)	Area Utilization
Manual CG with ICG cells	4.7	5.5	85.8 %
Additional automated CG (1 FF)	4.55	5.2	87 %
Additional automated CG (3 FFs)	4.5	5.1	84.7 %

had been integrated, the one with the biggest size has been used for evaluating design variations, in order to consider the "worst case" conditions for the area available to the digital logic.

Some additional considerations can be drawn, regarding more generically local clock distribution in the RD53A pixel core building block. In modern digital design tools, Clock Tree Synthesis (CTS) is part of the design flow which comes directly after placement of standard cells and before P&R and fine timing optimisation. This "automated" step requires designers to provide constrains based on the needs of the application. For example, at first CTS is performed taking into account only one primary design corner. The choice of the latter has been seen to have non-negligible impact on the clock tree power. Indeed, if a slow corner is used, the tool can over-buffer the clock tree at first and never get rid of superfluous buffers during the optimisations, since they are timing driven. In particular, for this application, the adoption of the slowest corner (SS, 0.9 V, -40°C) as primary corner has shown to give $\sim$15% digital power increase with respect to the use of the fastest corner (FF, 1.32 V, -40°C), which has been instead chosen. Indeed, even if at the beginning clock distribution can be timing-wise "weak", additional optimisations steps, taking into account all corners defined in the design flow, are free to add and resize buffer to optimise timing only where necessary. This still allows proper timing closure and at the same time clock tree power minimisation. A similar consideration is related to the driving strength of clock buffers which the CTS is allowed to use. Since within the core timing requirements are not particularly tight for a 65nm technology (i.e. frequency of operation = 40 MHz), local clock tree cells can be constrained to a maximum driving strength lower than the

highest available, in order to reduce power consumption. It has been seen that only allowing the use of clock buffers with driving strengths lower than 8 can have a non-negligible impact on digital consumption: ∼10% increase has been observed otherwise. Another interesting aspect is related to the RC characteristics of the clock routing. Using low-resistive high metal layers for clock tree routing can potentially help reducing the RC and therefore the buffer count. This should not constitute a worry for local clock distribution (where the resistance of wires on short distances is not strongly affecting delays). Anyway, for the purpose of the project only standard thin metals are made available and have therefore been used for local clock distribution (since thick metals have been dedicated to the critical global power distribution). As far as the clock nets capacitance is concerned, higher spacing between clock routing nets is a factor that could reduce net capacitances and therefore power consumption. Due to the congested design routing for some of the architecture combinations, this has not been explored during RD53A design, but could be considered for future developments.

4.4.3 Summary of results for RD53A architectures

Table 4.4 reports a summary of the power consumption of the different pixel cores flavours as they have been finally integrated in the RD53A chip. Following the defined power methodology, not only average power consumption, but also peak power (averaging at 1 μs time scale) and digital area utilisation are shown. It can be noticed that the DBA architecture achieves lower power consumption and it is at the limit for acceptable area utilisation (to close the design) when integrated with the LFE flavour, whereas some margin is still available with the smaller DFE. The higher density, also affecting routing congestion and therefore net loads, has a slight impact on power consumption, even if the two digital architectures are identical in RTL. As far as the CBA is concerned, power estimations have shown a higher power consumption, still reduced with respect to the CHIPIX65 prototype chip. Some of the solutions presented previously have also been adopted in the CBA, e.g. clock gating (with a similar structure) by means of ICG cells, clock tree constraints. Even

if the designer had initially shown the possibility of achieving a power consumption comparable to the DBA one, timing closure with all design corners both locally and at top level has caused the power budget to increase.

Table 4.4: Results for the typical corner at 1.2 V on average power consumption, peak power (averaging at 1 μs time scale) and digital area utilisation for the final RD53A architectures.

Architecture implementation	Average power per pixel (μW)	Peak power per pixel (μW)	Area Utilisation
DBA integrated with LFE	4.7	5.49	88%
DBA integrated with DFE	4.6	5.36	82%
CBA integrated with SFE	6.18	7.18	87%

4.4.3.1 Studies on further power optimisation

After RD53A final verifications and submission, further investigations have been performed concerning power consumption, as it constitutes one of the critical issues for future chips which will be designed for detectors in the HL-LHC upgrade. With the serial powering scheme, current headroom has to be considered to take into account digital power fluctuations and avoid system failures due to power peaks. Achieving the lowest possible power consumption is essential to minimise the system power budget of CMS and ATLAS pixel detectors. Additional studies were performed in the digital array pixel core for the DBA architecture (possibly extendible to the CBA), mainly concerning the local clock tree structure and hierarchy. The clock distribution to the regions was broken into its components, analysing in detail their contribution to the power budget, in order to address further optimisations. The structure of the clock distribution to the sinks for a pixel region is shown in Figure 4.11.

The local clock distribution deserves special attention, as it has a high impact on power with respect to the global clock distribution (including the skew compensation mechanism described in Section 5.3), as summerized in Table 4.5. As in the previous FE65-P2 chip, a “hard” clock disabling (AND cell at the root of tree) is implemented for the whole region when all pixels are disabled. This can be seen as a “reliability” feature to make sure that regions

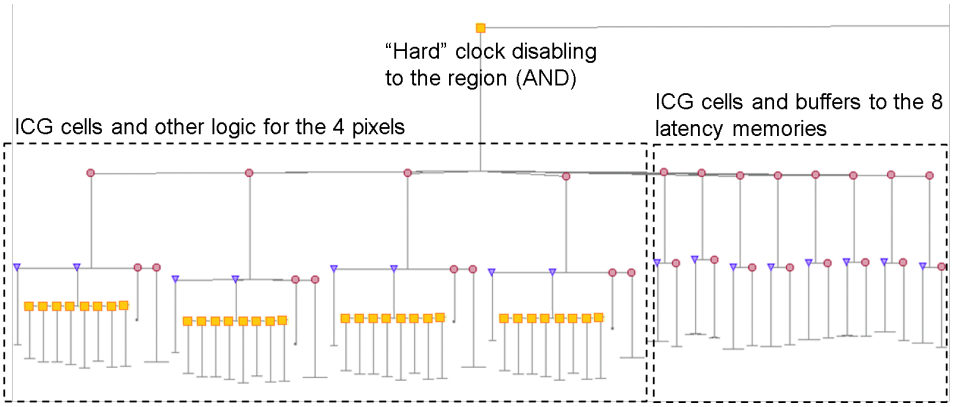

Figure 4.11: Local clock distribution down to the sinks for one pixel region made of 4 pixels. Clock gating cells are shown in red, buffers in purple and other combinatorial logic along the clock tree in orange.

Table 4.5: Percentage contribution of global and local clock distribution to power consumption. The values shown apply to the DBA architecture integrated in RD53A with the LFE.

	Contribution to clock tree power	Contribution to total power
Global clock distribution	13%	8%
Local clock distribution	87%	54%

can be fully disabled and cannot affect the data readout in case of major problems. A summary of the average contribution to the power consumption of each class of cells along the clock distribution within a PR for is reported in Table 4.6, in order to motivate further developments. As expected, the main

Table 4.6: Average power consumption of each class of cells along the clock distribution for a PR, excluding buffers down in the tree.

Clock tree cells	Contribution to average power in a PR (μW)	Percentage
Hard disabling of the clock	4.8	25.5%
1^{st} stage of ICG cells in the pixels	1.33	7%
2^{nd} stage of ICG cells in the pixels	$\sim$0.12	$< 1\%$
1^{st} stage of ICG cells in the latency memories	2.2	1.5%
2^{nd} stage of ICG cells in the latency memories	$\sim$0.003	$< 0.1\%$

source of power consumption are cells which are always active or constantly receiving the clock, i.e. the AND and the ICG cells up in the hierarchy, gating each of the pixels and each of the memory cells. The AND cells are always active and loaded by 12 clock gating cells, which make its power consumption 25% of the overall region power. The power consumption of these cells is also highlighted in the instance power map in Figure 4.12 for the full core. Moreover, these 12 ICG are constantly receiving the clock, which also gives a non-negligible contribution. On the contrary, 2^{nd} stage gating cells only have a minor impact on the power budget. Few design modifications have been evaluated in order to address such limitations and are summarised in Table 4.7. The trade-off between average power, peak power and area utilisation has been studied with the DBA architecture integrated with the LFE, i.e. the most critical from area density point of view. First (#1), a common clock gating has been manually implemented in the RTL in order to gather the 8 cells in the 1^{st} stage to latency memories, which have been removed. This has required a slight modification to the minimal control logic in each latency memory (since the clock is received also when other memory cells in the region are "active"). For this reason, a small % area increase has been observed, enough to complicate the design closure with the LFE. As previously, the tool has been

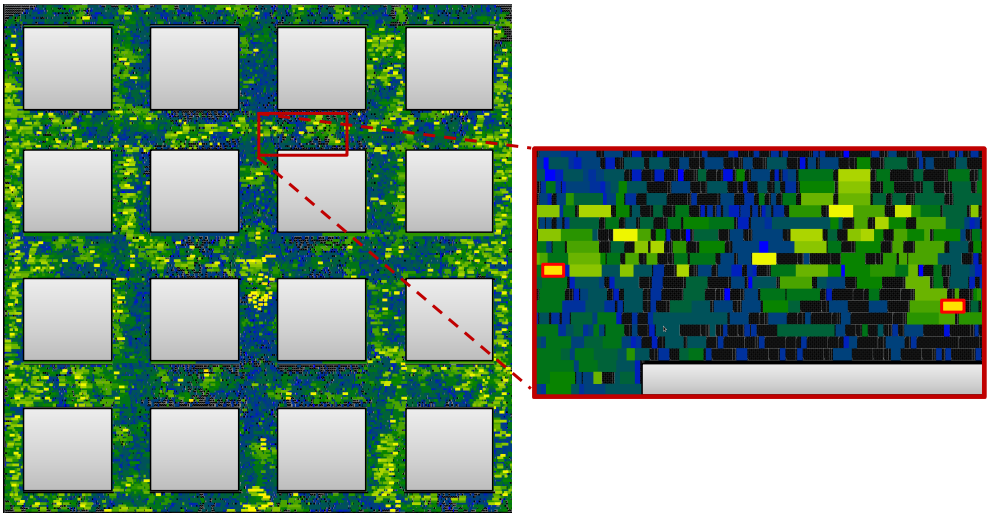

Figure 4.12: Instance power map of the pixel core: the AND cells hard disabling the clock (highlighted) are among the few cells with a dark yellow colour.

left free to automatically add 2^{nd} stage of ICG cells in the latency memories. The reduced number of 1^{st} stage CG for latency memories and reduced load on the AND gate has achieved a 20% average power reduction and improved power peaks. Secondly (#2), the same approach of removing per-pixel clock

Table 4.7: Results for the typical corner at 1.2 V on average power consumption, peak power (averaging at 1 μs time scale) and digital area utilisation for different clock gating implementations.

Architecture implementation (identified by #number)	Average power per pixel (μW)	Peak power per pixel (μW)	Area Utilisation
#0 DBA integrated with LFE (RD53A)	4.7	5.49	88%
#1 Common 1^{st} stage CG for lat. memories	3.8	4.7	91%
#2 Common 1^{st} stage CG also for pixels	3.5	4.7	91%
#3 Removal of hard clock disabling	3.29	4.53	90%
#4 Manual 2^{nd} stage CG for lat. memories	3.1	4.3	88%

gating and using a unique ICG cell has been adopted for the pixels. In this case, the effect of the lower number of gating cells has allowed a small average

power reduction (with very similar peak fluctuations) without affecting area. Indeed, the common gating also causes higher activity for the clock logic down in the tree of the pixels. At this stage, the consumption of the AND cells for hard clock disabling is significantly lower thanks to the reduced load, i.e. from 4.8 to 1.7 μW/PR, but still not negligible. For this reason, it can be worth discussing whether the feature is required or if simply forcing the disabling in the data path (data output and arbitration signals) could be sufficient. In this case AND clock cells are replaced by a buffer tree, with a lower number of cells, which obviously allows a power reduction. The power gain which could be obtained has been studied in case #3 and it is ~5%. Finally, starting from case #3, an additional design variation has been considered to solve the observed area increase. In case #4, the 2^{nd} stage of clock gating per latency memory has been brought back to be manual in the RTL as it was initially. Area-wise, gating the clock to the memories (manually minimising the control logic required) has been seen to be slightly more efficient than an automated synthesised implementation. The local clock distribution in case #4 is shown in Figure 4.13 for comparison with the initial implementation. To conclude

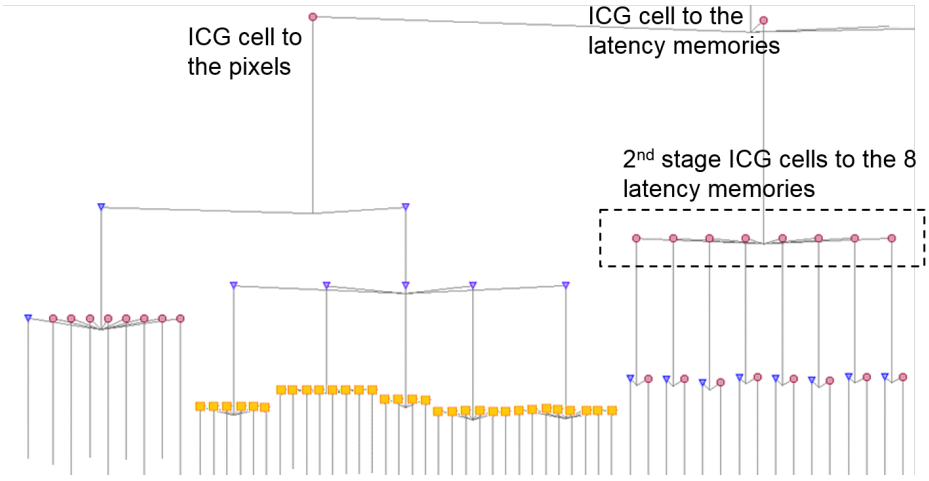

Figure 4.13: Local clock distribution for one pixel region in case #4 from Table 4.7, after the first stage of clock buffers in the core. Clock gating cells are shown in red, buffers in purple and other combinatorial logic along the clock tree in orange.

this Section, a general remark regarding clock tree power optimisation should be made. As the clock tree power reduces, both average and power peaks have been seen to reduce. Nevertheless, since the constant contribution to power consumption is reducing, the ratio between peak and average power is increasing. This is in conflict with the goal of having almost "constant" digital power consumption. Besides serial powering issues, achieving constant consumption is meant to limit fluctuations of the effective threshold of the pixels and coupling between analog and digital domains. These factors have to be taken into account to identify the best trade-off between power budget minimisation and digital power fluctuations' reduction.

Chapter 5

Design optimisation of the RD53A large format IC for timing and reliability in harsh radiation environments

The design of the RD53A large scale integrated circuit involves facing challenges common to nowadays deep-submicron semiconductor technologies. While CMOS device dimension has shrank (allowing higher speed and lower power consumption), the production of larger die sizes has become economically feasible, resulting in increased design complexity and average length of interconnects. Parasitic effects of interconnects display a scaling behaviour which differs from the active devices and which has gained in importance, starting to dominate relevant metrics such as design speed [106]. In this technological context, for the purpose of this work it is fundamental to define a design strategy to achieve low-skew clock distribution (Section 5.3) and timing closure (Section 5.4). In addition to these common design issues, assuring reliability in unprecedented levels of radiation is a major challenge for this application. Timing closure and cumulative radiation effects are strongly related, as it will be introduced in Section 5.1, and are therefore treated concurrently. To this

end, the chosen design approach will be discussed in Section 5.2, expanding the discussion to all relevant classes of radiation effects.

5.1 Radiation effects on CMOS technologies

The presence of ionizing radiation is in general a significant threat to the correct operation of electronic devices, both in the terrestrial environment (due to atmospheric neutrons and radioactive contaminants inside chip materials) or in space (particles emitted by the Sun and galactic cosmic rays). Artificial radiation are also generated for biomedical devices, nuclear power plants as well as for HEP experiments [107]. Radiation-hard design is of transversal interest and important for this work, since the target level of radiation can compromise the functionality of the chip if no measures are taken against it. In this Section an introduction to two main classes of effects on the logic in radiation environment will be given and dedicated approaches for a radiation-tolerant design will be discussed.

5.1.1 Cumulative effects: Total Ionizing Dose

The fundamental interactions between an energetic particle and a semiconductor device can be *i)* ionizing, i.e. creating free electron-hole pairs by disrupting electronic bonds, *ii)* displacement damage i.e. causing atoms to be displaced from their lattice site and leaving a vacancy [107].. For CMOS technologies, displacement damage is known to be not as critical as ionizing effects and this has been also confirmed for 65 nm technology [108]. For this reason, it will not be addressed in this work. On the contrary, Total Ionizing Dose (TID) has a significant influence on CMOS technologies, including 65 nm. It is an accumulating effect which gets worse and worse as a device is exposed to ionizing radiation. A radiation-induced charging of the oxide is caused and involves several different physical mechanisms, which take place on very different time scales, with different field and temperature dependencies [109]. High-energy electrons (secondary electrons generated by photon interactions or electrons present in the environment) and protons can ionize atoms, gener-

ating electron-hole pairs, also in a sequence until energies are sufficiently high (thousands of electron-hole pairs can be created). When a CMOS transistor is exposed to high-energy ionizing irradiation, electron-hole pairs are created in the oxide and cause oxide-trapped holes and interface-trapped charges, as shown in Figure 5.1. Electron-hole pair generation in the oxide leads to almost

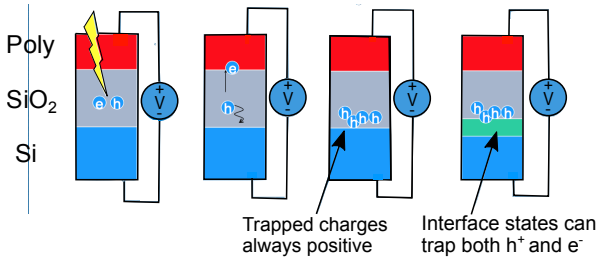

Figure 5.1: Electron-hole pair generation in the silicon oxide, induced by radiation, leads to oxide-trapped holes and interface-trapped charges (holes for PMOS and electrons for NMOS) [110].

all total dose effects [111]. In addition to oxide-trapped charge and interface-trap charge buildup in gate oxides, charge buildup occurs also in other oxides including field oxides, Silicon on Insulator (SOI) buried oxides, and alternate dielectrics. The accumulation of charge in the oxides and at their interface influences the electrical parameters of transistors (for the gate oxide) and of the parasitic structures unavoidable in CMOS. This can have multiple effects at transistor level (e.g. threshold voltage shift, leakage current increase, transconductance degradation), causing secondary effects at circuit level such as timing degradation and power increase.

For old technologies, the charge accumulated in the gate oxide had a major role in TID degradation, due to the thickness of the gate oxide (the total charge accumulated in the oxide is proportional to thickness). A sharp decrease of TID effects has been seen in commercial CMOS processes with lithographic dimensions as small as 250 nm, using gate oxides 5.2 nm thick [112]. One of the relevant effects induced by radiation in 250 nm technology was NMOS transistor leakage, caused by the formation of an inversion layer in the p-type

substrate or p-well underneath the field oxide or at the edge of the active area. This inversion layer is formed due to the radiation-induced accumulation of positive charge in the silicon oxide and leads to source-to-drain leakage and inter-transistor leakage between neighbouring implants. The same effect is not seen for PMOS transistors, since the positive charge accumulated in the oxide pushes the n-substrate or n-well more into accumulation without creating any inversion layer. In the past, Enclosed Layout Transistor (ELT) were used to prevent radiation-induced leakage at the edge of NMOS transistors. The basic concept of a ELT is shown in Figure 5.2. P+ guardrings were also added to

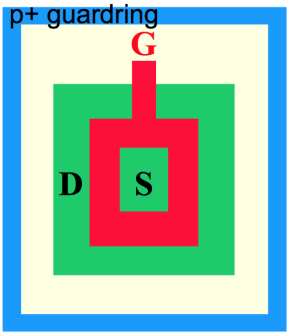

Figure 5.2: NMOS transistor laid out in enclosed geometry to prevent transistor leakage. The implementation of p-guardring prevents leakage between adiacent transistors [110].

cut leakage paths between adjacent n+ junctions at different potential [112].

Although in more modern technologies the gate oxide becomes thinner and hence less sensitive to TID, the Shallow Trench Isolation (STI) oxide does not scale down correspondently. As a consequence, radiation-induced charge trapping in the STI oxide still leads to macroscopic effects limiting the radiation tolerance of conventional CMOS circuits. The TID response of transistors and isolation test structures for a 130 nm technology was studied up to 100 Mrad in [113]. Contributions from oxide-trapped charge and interface states to radiation-induced edge effects were found and seen to have a significant influence on the transistor characteristics. The channel length of the transistor was seen to have significant impact on degradation (mainly threshold

voltage shift), since edge effects are more visible for shorter channel lengths. NMOS leakage was also seen in 130 nm technology, but with values up to 2-3 orders of magnitude smaller, with respect to older CMOS technologies. This conclusion made the need for ELT design less and less critical [113].

The complexity of the phenomena, dependent also from the details of the complex technology process (which cannot be revealed completely by the foundry), has motivated the HEP community to carry out a radiation campaign both at transistor and circuit level on different technologies and multiple vendors to study reliability. In synergy with RD53 collaboration, the commercial 65 nm technology chosen for the project has been extensively studied, up to an unprecedented TID of 1 Grad (= 10 MGy). Alternative high density (65 nm) CMOS technologies have also been evaluated and have not shown better radiation tolerance. A very positive characteristic of the chosen technology is that it does not feature a significant increase of leakage current, which was instead observed in other technologies (both with bigger and same technology node). In particular, the leakage current increases maximum of 2 order of magnitudes, with values below the nA even at 1 Grad [114]. The high dose tolerance of the thin gate oxide was confirmed, but defects in the spacer and lateral STI oxides have shown a strong effect on the performance of the transistors. Observed radiation damage of transistors depends on a large set of parameters and conditions: radiation level and dose rate, temperature during irradiation, type of transistor, width and length of the transistor, biasing of the transistor during irradiation, annealing temperature, and time including biasing during annealing. Transistor performance degradations are mostly due to "short-channel" and "narrow-channel" effects and some guidelines are know on minimum length and width, treated in details in [115]. It can be highlighted that during the extensive radiation campaigns an unexpected behaviour of irradiated PMOS transistors was encountered at very high temperature (100°C), with detrimental effects on the device performance. This behaviour has been studied and seen to be highly temperature and bias driven (i.e. it needs high temperature and bias to be activated): it will therefore not cause significant performance degradation over long time periods when the chip is operated at temperatures between −10°C and −20°C. Moreover, if the chip will not op-

erate for some weeks/months and will therefore be at at room temperature and without bias/power, the detrimental annealing effect will be significantly smaller: in fact mainly recovery annealing is seen.

5.1.2 Single Event Effects

A Single Event Effect (SEE) is the result of an instantaneous impact of radiation affecting the state of the electronics. SEE can be in general classified in destructive (hard-error) and non-destructive (soft-error) [116]. Soft errors are temporary and recoverable by applying power shut down, reset or rewriting the corrupted data, but are clearly undesirable at too high rates since they would not allow the system to operate with proper functionality for a long time, required for data acquisition. In CMOS-based circuits, possible hard errors are Single Event Burnout (SEB) that can occur in power MOS devices, Single Event Gate Rupture (SEGR) or gate-rupture. The CMOS p-n-p-n parasitic structures can also be vulnerable: a Single Event Latch-up (SEL) can cause a strong current which can lead to overheating of the device. If it is not stopped by a power cycle, it can have destructing impacts on the transistor [116]. These hard errors are not discussed in this work as there is no strong evidence of their impact on the chosen technology based on the radiation campaigns available [117]. Most relevant soft errors are Single Event Transient (SET) and Single Event Upset (SEU), of which some examples are shown in Figure 5.3. The former causes a transient change of voltage in one of the capacitive nodes of a logic gate. The likelihood of an SET decreases with increasing node capacitance. If this change is captured by a memory device, it becomes a persistent effect. Instead a SEU occurs when deposited charge directly causes a change in the state of a sequential element such as a flip-flop, latch or memory [116]. If no other measures are taken, the corrupted state will persist until a new value is written into the memory device and will propagate to the logic connected to the fan-out of the memory. The most important figure for SEEs is the rate of occurrence (i.e. how many events take place per hour/day/year) in a particular environment. A generic way of characterising SEEs is the cross section σ, defined as the number of observed upset events di-

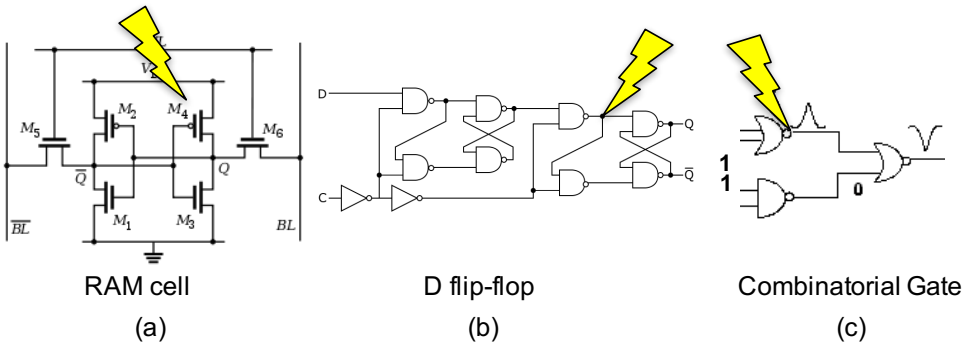

Figure 5.3: Examples of SEE: SEUs on a RAM cell and on a flip-flop and a SET causing a glitch on combinatorial logic are shown respectively in (a), (b) and (c).

vided by the incoming particle fluence (#particles/cm^2). In literature studies are available about SEUs in 65 nm CMOS technologies [117]. Digital prototypes were irradiated in a heavy-ion beam facility and it was concluded that the probability of an SEU in a single device decreases as transistor size is decreased. On one hand, the critical energy needed to cause a upset diminishes (due to the reduction in supply voltage and node capacitance), but on the other hand physical dimensions also reduce. With an area ratio between the standard library cells in 130 nm and 65 nm of about 4×, the cross-section has been seen to scale almost proportionally by a factor 3.4×. Even though the cross-section of the cells in 65 nm is lower with respect to previous technologies, this is not sufficient to consider the technology to be SEU robust. The higher density of the technology node also allows the integration of much more logic in the circuits, causing many more nodes to be exposed to SEUs. Moreover, the probability of a Multi Bit Upset (MBU) to take place is higher and therefore separation between redundant storing cells is required when designing for SEU tolerance. Therefore, SEUs need to be taken into account in the design process with the chosen technology.

5.2 Design approach for reliability in the radiation environment

In this Section, the design approach adopted to assure the reliability of the digital pixel array in the harsh radiation environment will be discussed. In particular, the strategy used to model TID effects is described in Section 5.2.1, whereas Section 5.2.2 discusses which measures against SEEs are required for the target application.

5.2.1 Performance degradation of the digital logic

Based on the TID effects introduced in Section 5.1.1, in the context of the RD53 collaboration the following main choices and developments have been carried out:

- the design has been targeted to remain functional up to 500 Mrad. This translates into the need to replace electronics of the inner layer (a small fraction of the total area) of the CMS and ATLAS pixel detectors after five years of operations, unless chips are proven to remain functional after higher dose);

- SPICE simulation models of transistors at 200 and 500 Mrad have been extrapolated from the results of the radiation campaign (with worst case bias and room temperature), in order for the designers to take them into account [114];

- analog circuits have been designed to simulate correctly with such radiation models and the produced test chips have demonstrated the required radiation tolerance;

- developed SPICE simulation models have been used in the community to generate digital design library files with re-characterized timing and power information (to account for radiation effects both at 200 and 500 Mrad);

- a dedicated Digital RADiation (DRAD) test chip [118] was designed to study experimentally the impact of TID on digital logic gates.

The DRAD chip includes nine different versions of standard cell libraries (differing in the device dimensions, threshold flavour and layout of the device) and each library has test structures designed to characterize delay degradation of the standard cells. In particular, four different sized digital libraries have been integrated (7, 9, 12, 18 track), whose height is shown in Figure 5.4. Measurements of time delays for gates of different size and type have been per-

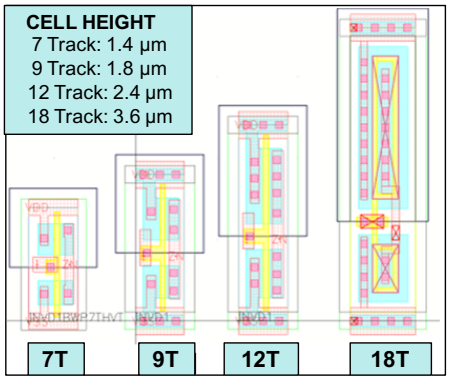

Figure 5.4: Cell height for different sized digital libraries integrated in the DRAD chip: 7, 9, 12 and 18 track [119].

formed and compared with circuit simulations of worst case (worst case bias) radiation models of single transistors. In Figure 5.5 the observed delay degradation of differently sized digital libraries (7, 9, 12, 18 tracks) with different types of transistors (normal threshold voltage, V_t , high V_t , and low V_t) is summarised and compared to delay degradation obtained with the simulation model. As anticipated, it is evident that delay degradation is increasing with smaller size of the transistors. Measured speed degradation of differently sized digital cells after 500 Mrad radiation is significantly less than predicted by the correspondent simulation model, as transistors in digital circuits are only under worst case bias conditions during short signal transitions. Moreover, when the chip is operated cold (and never kept biased if not cooled), as planned for LHC experiments, modest delay degradation within ∼20-50% is observed. The

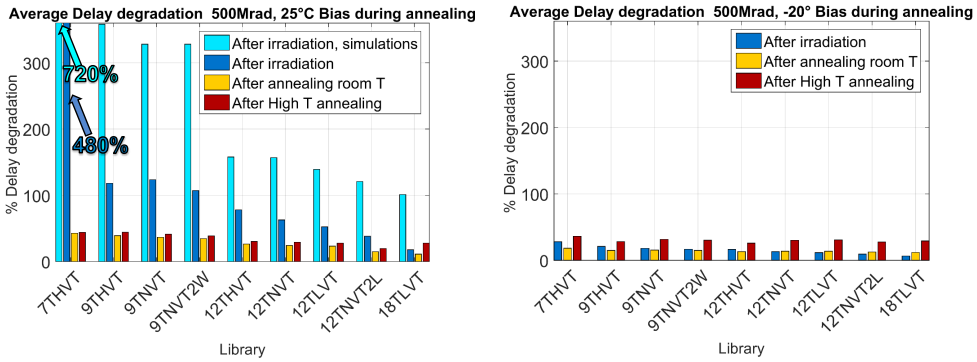

Figure 5.5: Average delay degradation of standard cells from different libraries integrated in the DRAD test chip [120]. For the multiple libraries first the number of tracks is indicated (e.g. 7T), followed by the V_t flavour (e.g. HVT, high V_t) and eventual indication of double width (2W) or length (2L) transistors. Measurements results are shown: after irradiation up to 500 Mrad performed at room temperature (left) and cold (right). Room and high temperature annealing are also displayed. Results from 500 Mrad simulation models (derived at worst case bias and room temperature) are reported on the left plot for comparison.

9-track library is of particular interest for the purpose of this work, since it is used for the implementation of the pixel array logic. Indeed, the area density on the digital array is too high to consider any library with bigger cells. At the same time, the use of even smaller devices (7-track library) is not optimal considering their significant performance degradation after irradiation. Even if at cold temperature the behaviour it is not detrimental, it is still important to maintain margin for operating the chip in test setups at room temperature. Figure 5.6 shows relative time delay degradation after irradiation and annealing with bias (both at room and high temperature) for various types of gates of the 9-track library, tested in the DRAD chip. The largest degradation after annealing, is observed in gates with small PMOS transistors (latch cell delay: LH_DEL) and having multiple PMOS transistors in series (NOR gates). The delay degradation when irradiated to 500 Mrad at room temperature reaches ∼160% for the worst case gate type. When irradiated cold the worst case delay degradation is only at the level of ∼20%. A conservative approach considered for the design of the digital array logic is to use worst case timing models (obtained from 500 Mrad SPICE models in worst case bias

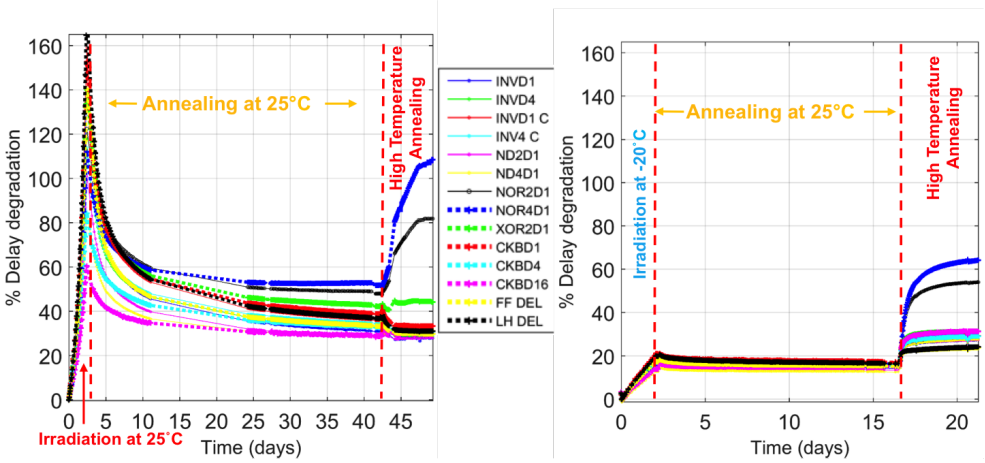

Figure 5.6: Measurements of delay degradation for standard cells from 9-track normal V_t library after irradiation and with annealing with bias [120]. The name of the cells in the legend describes in order: cell type, number of inputs of the cells (where applicable), driving strength. *FF_DEL* and *LH_DEL* stand respectively for the delay of a flip flop and a high-level sensitive latch.

and room temperature), throughout the design flow. The logic synthesis tools can indeed take such models into account when synthesizing the detailed gate level design, by either avoiding less performant gates or only using such gates for un-critical timing paths. Moreover, the same timing models can also be used during the Static Timing Analysis (STA) and optimisation stages of the P&R flow, until final sign-off. Whereas synthesis is performed with a single (worst case) corner, for P&R a Multi Mode Multi Corner (MMMC) digital flow is adopted. The tools perform STA in parallel for multiple functional modes and library corners: radiation models can be included as additional corners. This approach was used during the design of the RD53A pixel array. However, the adoption of 500 Mrad models (in worst case bias and room temperature) as worst case corner, has been seen to be at the limit for achieving timing closure, most of all for signal propagation along the column. Moreover, using over-pessimistic corners to close the design has a negative impact on design metrics, e.g. power, hit efficiency (depending on the memories with can be accomodated in the area needed to close timing). In Figure 5.7 it can been seen

that experimental results at 500 Mrad are way better than simulation models, which feature worst case biasing on all transistors. Instead, experimental re-

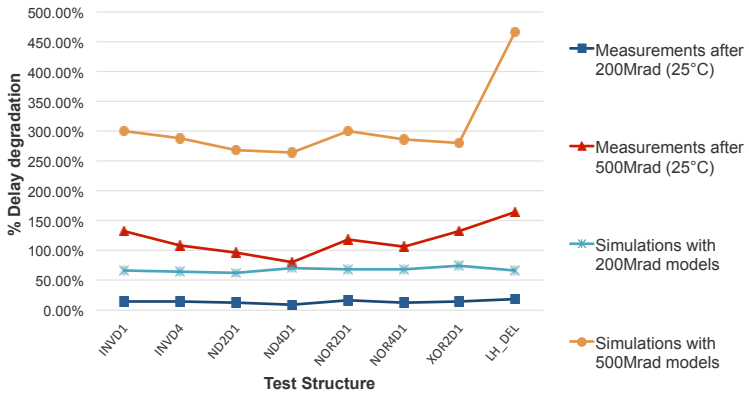

Figure 5.7: Percentage delay degradation of standard cells from 9-track normal V_t library after irradiation with respect to the ones before radiation. Measurements results of the DRAD chip at different temperatures are compared with results from correspondent simulation models (worst case). The name of the cells in the x-axis describes in order: cell type, number of inputs of the cells (where applicable), driving strength.

sults at 500 Mrad (at 25°C) have a degradation similar to the simulation results at 200 Mrad (worst bias, 25°C), whereas experimental results at 500 Mrad at −20°C (operation temperature) are well below the 200 Mrad simulation models. A trade-off design choice has been done for RD53A digital design: the 200 Mrad timing library was included in the multi-mode multi-corner analysis, while the over-pessimistic 500 Mrad was excluded. In order to have a conservative margin on TID tolerance, the very worst case technology library offered by the foundry (SS, 0.9 V, −40°C) was added as worst case corner. Such a corner is at cold temperature since at the 0.9 V supply the technology experiences the so called "temperature inversion": below a certain supply voltage, the transistor V_t increase caused at low temperature dominates over the carrier mobility increase, causing cells delay to be higher at low temperature. This technology corner shows more pessimistic delays than the 200 Mrad library characterized from simulation models, as shown in Figure 5.8. The plot has been gener-

ated with Cadence Liberate, by comparing the timing files of the two libraries. Approximatively a +50% pessimism on delays can be observed, which is con-

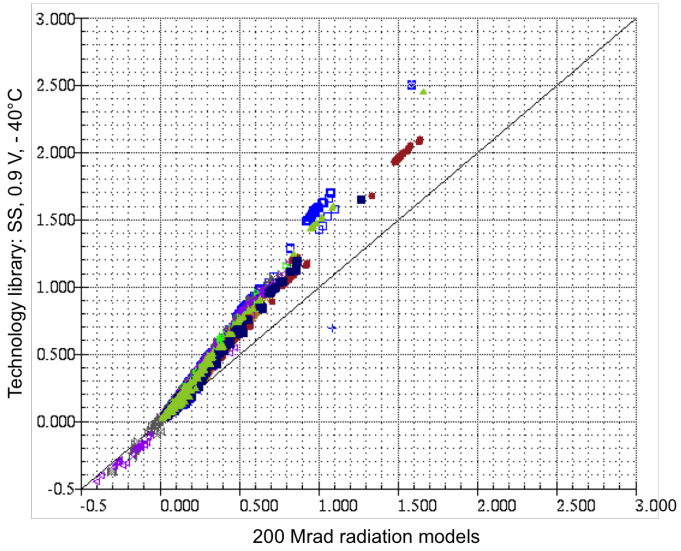

Figure 5.8: Graphical library comparison between 200 Mrad radiation models and the SS, 0.9V, −40°C technology corner. Values from the first are given on the X-axis and the values from the second library are given on the Y-axis (i.e. when the values in the two libraries match, the plotted data points fall on the 45-degree axis). The scattered-plot is based on the same library cells integred in the DRAD chip.

sidered enough to take into account experimental results at 500 Mrad at room temperature, without the need for excessive over-design.

5.2.2 Single Event Effects

SEE mitigation techniques are well known in the context of HEP, space, aeronautic and terrestrial applications. They can be classified as: *i)* hardening by technology, where the technology process is modified to minimize sensitivity to soft errors (SOI has been for example seen to be more resistant to SEE [121]); *ii)* hardening by cell design, where memory circuits are made to store the information in multiple nodes in order to automatically correct flipped nodes e.g. Dual Interlocked storage CEll (DICE) [122]; *iii)* hardening at system level, using techniques capable of correcting bit flip either by means of redundancy and

voting mechanisms as in the case of Triple Modular Redundancy (TMR) or by actual Error Correction Coding (ECC) as in the case Hamming or advanced Reed-Solomon encoding schemes. These techniques, however, have also an impact on area, power and timing. Generally, for a fully triplicated design, the area overhead is always more than 200% as voting logic is required in addition to the triplication overhead. Error correction scheme can achieve better performance only when the number of bits is sufficiently high to compensate for the additional logic required (e.g. [123]). The latter clearly causes also power increase and possibly timing complication.

As regards SET tolerance, effects on signals such as clock, reset, etc., are the most critical. At the same time, global distribution of aforementioned signal features high capacitance, implying that small current injections may not be sufficient to provoke visible transients. On the other hand, trees of buffers built to distribute such signals (in particular for the clock), reduce the load of the nets, possibly posing problems for SET. The SET sensitivity of high loaded nets needs to be verified with SPICE simulations emulating current injection. Mitigation techniques used rely on full triplication, including clock and reset signals. Therefore, comprehensive SET tolerance is hard to be achieved in the pixel array logic without a major impact on area and power. For the design of RD53A pixel array, a low priority has been given to design against SET and its verification, since it is a prototype chip needing to be produced in a limited time with the resources available. On the other hand, experimental tests will be perform to spot any potential SET issue and it will be possible to reproduce such effects on simulations after the chip has been implemented. In the digital chip bottom (i.e. DCB, Section 3.1.1.2), a minimum approach adopted against SETs has been using deglitchers for critical global signals (e.g. reset).

As far as SEU tolerance is concerned, some considerations about its cross section are necessary to decide which sequential elements need to be protected. Based on [117] and [124] a conservative cross section with the order of magnitude of 0.5-5 $\cdot 10^{-14}\,\mathrm{cm}^2$ can be assumed for latches and flip-flops of the chosen technology. Moreover, considering the accepted detector inefficiencies and noise hits, the criteria for corrupted hits from the pixel chip itself have been set to 10^{-4}. For the pixel array, this implies that:

- no protection is needed for hit data during trigger latency. Indeed, assuming 24 bits for pixel region data stored in memory during 12.5 μs of latency and given the conservative hadron rate and cross section (i.e. 500MHz/cm^2 and $5 \cdot 10^{-14}$ cm^2) this brings to a corruption probability of:

$$\begin{aligned} P_{hit\,corr.} &= 500 \cdot 10^6\, Hz/cm^2 \cdot 12.5 \cdot 10^{-6}\, s \cdot 5.0 \cdot 10^{-14}\, cm^2 \cdot 24 \\ &< 10^{-8}, \end{aligned} \tag{5.1}$$

  well below the defined criteria;

- protection is needed for pixel configuration latches (assuming that it is written once and not refreshed during operation), which can affect front end and pixel region logic functionality, as well as data readout. Depending on the front end design, in RD53A each pixel has a maximum of 8 configuration bits, bringing to a failure rate of (for a 2 cm$\times$2 cm pixel chip with 50 μm pitch):

$$\begin{aligned} R_{failure} &= 500 \cdot 10^6\, Hz/cm^2 \cdot 160 \cdot 10^3\, pixels \cdot 8 \cdot 5.0 \cdot 10^{-14}\, cm^2 \\ &\sim 30\, upsets/second/chip, \end{aligned} \tag{5.2}$$

  i.e. $\sim$1% pixels affected after 40 s of operation;

- protection is not required for FSM and control logic in the pixel array, as long as it is proven that they are capable of recovering after a certain transient of non-functionality and they are within defined criteria for efficiency and noise hits.

SEU injection needs to be performed during simulation in the digital array under operating conditions in order to prove the aforementioned capabilities of the FSMs and digital logic. SEU simulation is a known problem both in HEP community, space applications and recently also in industry, due to the growing presence of electronics in the environment which needs high robustness to faults (e.g. automotive applications) [125]. In particular, approaches available in the technological context and in the HEP environment have been analysed and adopted for the target application. Further details are reported in Section 5.5.2, together with preliminary simulation results. As regards protection

of configuration registers in pixel array, design evaluations including the use of DICE latches are reported in Section 5.5.1.

As far as the DCB is concerned, SEU may lead to loss of event synchronisation or corrupted pixel chip configuration, and should therefore be as low as possible since it requires global system actions to recover correct functionality. A high number of global configuration bits and synchronisation bits (state machines, counters,..) are controlling the global functionality of the chip and are therefore a critical SEU target. For this reason, SEU protection is mandatory for operation in the experiments. As in the case of SETs, SEU-tolerant design has not been a priority for the RD53A prototype, whereas it does require a more careful study for following chips that will have to operate in the HL-LHC experiments. Nevertheless, in the DCB a TMR approach was adopted by the designers for the global configuration, by mapping selected registers with triplicated cell during synthesis. Moreover, additional constraints during place and route were used to guarantee minimum distance between memory elements, as presented in [126].

5.3 Hierarchical low-skew clock distribution along the column

In a synchronous digital system, the clock signal defines a time reference for the movement of data and its function is therefore vital to the operation of a the system. Clock signals have special characteristics: they are loaded with the greatest fanout, travel over the longest distances and operate at the highest speeds of any signal within the entire system. They are also particularly affected by technology scaling, since long global interconnect lines become more and more resistive with decreasing line dimensions [127]. Moreover, choices on clock tree distribution have a main impact on the performance trade-off among system speed, physical area, and power. In this work, the clock is not only a reference for the sequential logic, but also represents a global timing reference for the full matrix which is meant to perform in-time sampling of incoming particle hits. Therefore, it is not sufficient for the clock distribution

to achieve timing closure, instead, it is also required that its propagation along the column (to the hit-synchronisation stage) features a skew lower than 2 ns. This is fundamental to assure that each event is correctly sampled across the whole matrix and needs to be guaranteed across technology corners and including radiation damage. Finally, the chip hierarchy is another aspect which influences clock network choices. Indeed, clock tree synthesis is part of the digital design flow and needs to be performed in the building blocks of the system which are used for synthesis and physical implementation as a layout block. For large chips, maintaining hierarchy is fundamental for design tools to handle complexity.

Related works on pixel readout chips have developed different solutions to address similar design specifications, depending on the technology adopted and other system specifications. In the ATLAS FE-I4, clock skew was controlled by balancing the clock along the column with different delays from the top (no-delay) to the bottom (maximum delay) [89], as shown in Figure 5.9. This was achieved by manually placing delays in multiple points along the column, i.e. partially breaking the design hierarchy with final layout adjustments. This approach allows to control the skew along the columns (up to some extent): preserving a clock skew of about 1-2 ns has also the advantage of reducing the risk for sharp power spikes, which may be caused by a perfectly un-skewed clock distribution if local decoupling is not sufficient to absorb them. In LHCb

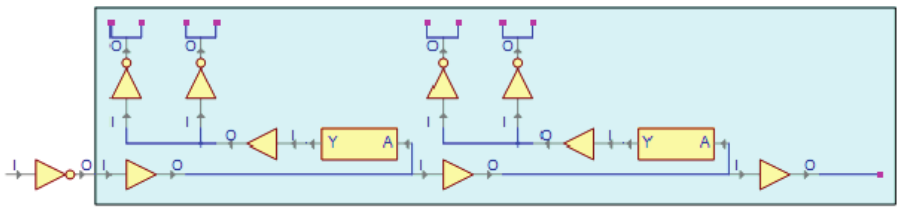

Figure 5.9: FE-I4 clock distribution along a double column. Rectangular cells represent different delays used to compensate for the clock skew [89].

Velopix [128], the use of clock repeaters along the clock trunk in combination with second local level buffering, was sufficient to achieve a skew lower than 2 ns. This was obtained by determining the optimal distance between repeaters

(880 μm) and the clock wire width (0.4 μm) for the adopted 130 nm technology. Within the hierarchy of the design, not all basic building blocks (so-called 2×4 super-pixels) contain them, since the optimal distance is found with one repeater every 4 super-pixels. For the MPA project (submitted together with RD53A in the same technology), a low power clock distribution strategy was presented in [129]. In contrast to the approaches previously described, the clock distribution is row-based, i.e. with only one column buffer and one buffer per each pixel row. This solution was shown to achieve substantial power savings for the chip, specially since it features a high aspect ratio between the number of pixel columns and rows (118/16). Moreover, the use of a reduced power supply was investigated to further reduce power consumption. In order to achieve a clock skew lower than 1 ns, thick metals were used in the MPA design to minimise the resistivity of interconnect lines for the chosen 65 nm technology.

5.3.1 Preliminary clock distribution study

A set of constraints/recommendations concerning the implemented clock distribution along the column in RD53A are discussed before treating it in details. First, the high power budget and the large scale of the chip makes power distribution a critical issue and it has encouraged the use of low-resistivity top level metals for power distribution, as reported in Section 3.1. Due to the conservative power distribution approach, no thick metal was available for clock routing. An extensive evaluation of alternative power distribution schemes (allowing to use thick metals for clock distribution) was not addressed during the RD53A design, but it is not excluded for future developments. Another relevant aspect is the hierarchical structure of the chip and its design density. RD53A is integrating multiple frond-end and digital architectures, already introducing three different layout building blocks in the pixel array. Any additional hierarchical variation is not ideal, since it can increase complexity and design effort. At the same time, the area density makes it undesirable to allocate any empty space in the array area for manual clock delaying/routing and this also introduces some complications to the automated design flow.

Given these conditions, a preliminary study on clock distribution was performed in order to assess which results could be obtained from the technology and the routing metals available. To this purpose, the synthesis tool Cadence RTL compiler was used to quickly evaluate the clock skew of a column-based clock distribution with clock buffers of different sizes and "placed" at multiple distance from each other (e.g. used as repeaters along the column). The basic clock unit being synthesized is a clock buffer for column propagation loaded by a buffer for local clock distribution, as shown in Figure 5.10. This block is replicated for a certain number of times needed to cover a 400-pixel column (aiming to final pixel chip size, ∼double than RD53A). Instead of using the

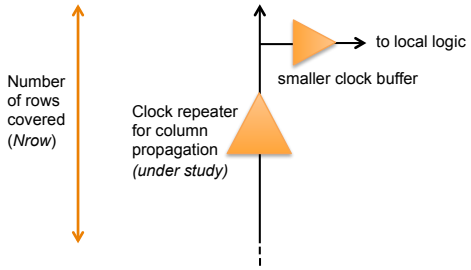

Figure 5.10: Basic clock unit with one clock repeater every *Nrow* pixel rows.

synthesizer wire load models, the net load from one stage to the following has been set manually through Synopsis Design Constraints (SDC). The load of a minimum-width thin metal wire was estimated by simulations from analog designers and a load range was provided. For a 100 μm, the range 6.6-21 fF was considered to be representative of load extremes (an ideal fully isolated wire and a with considerable routing in the surrounding). While evaluating schemes with repeaters at multiple distances, a linear scaling of the capacitance was assumed. The resistive behaviour of the net was instead not modelled in detail (underestimated wire models were used by the synthesizer). It should be highlighted that the goal of the study was not to obtain accurate clock skew estimation, but rather to guide the design choice between an approach with or without skew compensation.

In Table 5.1 a set of different configuration for the total number of clock unit is reported as function of the distance between repeaters and the selected

Table 5.1: Total propagation delay as function of different distance between repeaters, load net capacitances and buffers. Results are based on the technology corner: SS, 1.08 V, 125°C.

Totol Number Clock Unit	**Nrow**	**Distance between repeaters (μm)**	**Net load @ 100 μm**	**Net load (fF)**	**Repeater cell**	**Total Delay (ns)**
200	2	100	6.6	6.6	CKBD2	21.28
100	4	200	6.6	13.2	CKBD2	14.33
50	8	400	6.6	26.4	CKBD2	10.68
25	16	800	6.6	52.8	CKBD2	8.39
20	20	1000	6.6	66	CKBD3	6.09
200	2	100	21	21	CKBD2	37.12
100	4	200	21	42	CKBD2	28.95
50	8	400	21	84	CKBD3	18.42
25	16	800	21	168	CKBD6	9.08
20	20	1000	21	210	CKBD8	6.67

buffer. The two extreme net load estimation values are used and the size of the clock buffers progressively increased. This table shows that the total bottom-up delay decreases with increased *Nrow* and also decreases by considering more powerful buffers. Concerning the first point, it should be highlighted that this approach is underestimating quadratic RC delay effects along the column, but it can still be seen as an optimistic case to assess the feasibility of such a clock distribution. At this point, *Nrow* has been fixed to a ∼high value (20) and all clock buffers offered by the technology have been studied in the range of net load considered (Figure 5.11). It is reminded that the skew specification needs to be met with the slowest corner, accounting for TID degradation, as discussed in 5.2.1. During this study the choice of the RD53A design corners was not finalised, mainly since the understanding of radiation effects on digital logic was under investigation. Therefore, the timing analysis was performed with the "standard" worst case corner from the technology (SS, 1.08 V, 125°C). At the same time, it was known that radiation effects would have most likely implied a more severe timing degradation. Even if approximate, the outcome of the study suggested that achieving the required skew with available clock buffers and standard routing with thin metals is difficult (if at all possible). Anyway, a complete evaluation at layout level of metal-width alternatives or use of thick

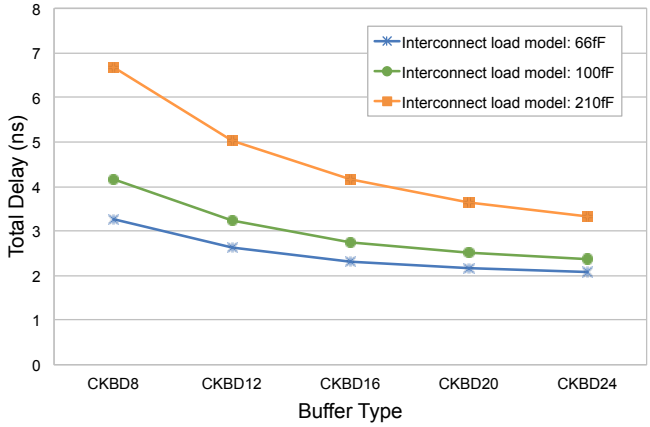

Figure 5.11: Propagation delay of different clock repeaters "placed" every *Nrow*=20 pixel rows, assuming 3 wire load scenarios.

metals could not performed within the time available for the RD53A design. Based on the studies performed, the alternative solution, i.e. the adoption of a (fixed) skew compensation scheme, was chosen for implementation. This decision has also allowed to early define the design hierarchy with some level of flexibility with respect to the clock distribution structure, as it will become more clear in the following.

5.3.2 Implemented clock distribution scheme and results

Before describing in details the clock distribution scheme adopted in RD53A, it should be reminded that the basic building block of the array has been defined to be a 8×8 pixel core, as shown in Figure 5.13. This choice was motivated by multiple factors: *i)* feasibility to simulate in an analog environment (small circuit to limit complexity), *ii)* sufficiently fast propagation of signals along the column (details in Section 5.4), *iv)* data readout "cluster-oriented", with a pixel core capable of containing long horizontal clusters of hits, *iv)* low layout size aspect ratio (square), supposed to better profit from P&R algorithms. The latter is mentioned as another extreme approach was initially evaluated within RD53, i.e. using an entire column (twice 2×2 PR wide) as the basic layout

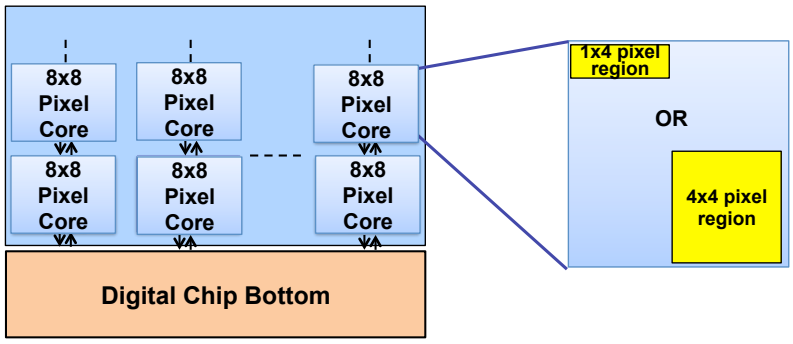

Figure 5.12: Pixel array hierarchy, with a pixel core as building block. The sizes of the different pixel region architectures integrated in the chip are also shown.

block. This was successful for the FE65-P2 small-scale prototype (4×64), but up-scaling it to a 400-pixel height was found to be not feasible with the adopted design tools. With the aim of defining a "as simple as possible" design hierarchy and reducing error-proneness across flavours, it is advisable to have cores identical to each other (per FE-flavour). In RD53A this has been achieved by:

- statical pixel core address calculation, i.e. the bottom core address is set at the end of column and each core contains arithmetical logic to calculate its own address and propagate it to the next one;
- a "programmable" clock delay block is used in each core, which statically multiplexes the amount of skew compensation based on the pixel core address (i.e. position along the column).

The implementation of this concepts is summarised in Figure 5.13 for a double-RD53A-size pixel core column. As concerns the skew compensation, a scheme with 6 delay options was considered sufficient to meet the timing requirement, while designing for a full-scale chip. Therefore, a 6-to-1 multiplexer has been used to select the amount of delay of the local clock distribution and only the three most significant bits of the core row address are needed to control the multiplexing (not all 8 combination are used). The block diagram of the *ProgrammableDelay* block is shown in Figure 5.14. It should be highlighted

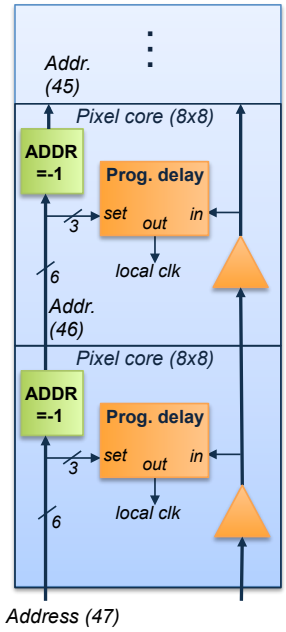

Figure 5.13: Block diagram showing the core row address calculation and clock skew adjustment schemes for the pixel cores.

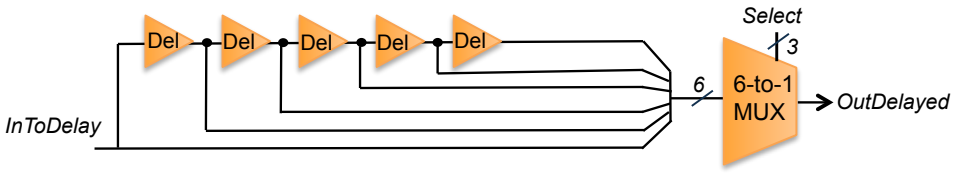

Figure 5.14: Block diagram of the clock skew adjustment for the pixel cores (*ProgrammableDelay*), with static dely selection based on the hierarchical core-row address calculation scheme.

that even if RD53A is a half-size chip, the distribution scheme has been done emulating the worst condition for a full-size case. Therefore, the maximum delay (5 delay units) has been used in the top core, whereas the bottom core contains 3 delay units. This allows to already face a situation with the highest clock insertion delays in the cores, as they would be seen for a full-scale chip. Few technical expedients were used to fine-control the placement, size and load of the clock repeaters along the column. During P&R optimisation, the tools were indeed not always making optimal choices, e.g. placing the clock repeater at the top or bottom of the core with the programmable delay far from it (in some cases increasing its load significantly), trying to fix it by adding multiple repeaters (only increasing the in-out propagation delay), etc. In order to deterministically allow tools to build a proper clock tree locally in the core, the placement of the clock repeater was forced to be in the center of the core, close to the programmable delay block (the repeater's only load). The correspondent clock input (output) pin at the bottom (top) of the core is also located centrally. Moreover, scripts were adopted to define the size of the repeater to be the biggest available in the technology being used. This was performed already at placement time, for all the following timing optimisations to already take it into account. Finally, it can be mentioned that, based on the results from the study in Section 5.3.1, it has been considered to use clock repeaters at longer distance than 400 μm (every core), in order to limit the skew to compensate for. Indeed, the propagation delay along the column is dominated by the buffer delay rather than by the interconnects. The side effect is that it partially breaks the design hierarchy by having twice the number of core variants (with and without the repeater). Placing the clock buffer at higher distance (e.g. every 2 or 4 cores) was seen to only give a small propagation delay reduction along a 48-core column ($\sim$1 ns). Indeed, the buffer delay is increasing significantly, not only due to the increased output interconnect load, but also because of the degraded input slew rate. This combined with the quadratic scaling of the interconnect propagation delay brings to the very limited gain, even if it was not evident in the preliminary study due to the poor resistive modelling of nets. The advantage of the skew compensation approach is that such a propagation delay can be sustained, still meeting timing requirements.

Therefore, the evaluated alternative approaches were abandoned in favour of a fully hierarchical scheme. It can be mentioned that the clock propagation along a 48-core column is accumulating a total of ∼10 ns in the worst case corner, while in the typical the delay is lower than 4 ns and below 3 ns in the fastest. Those are the delays that the skew compensation scheme is designed to compensate for. The results for the clock skew along the column for the submitted prototype are reported in Table 5.2 for multiple technology corners, including the most critical (i.e. slowest) ones used in the design. The skew

Table 5.2: Column clock skew results of the RD53A chip prototype, across multiple technology corners. The slowest corner is at cold temperature since at the low supply voltage (0.9 V) the adopted technology experiences temperature inversion.

Voltage	Process	Temperature	Column clock skew (ns)
1.32	FF	−40°C	0.5
1.2	TT	25°C	0.8
1.2	200 Mrad	25°C	1
0.9	SS	−40°C	2

results are obtained considering the first synchronising stage after the front-end discriminator output, i.e. where the requirement is critical to discriminate timing of incoming hits.

5.4 Optimisation for top-level system timing closure

Similarly to the clock distribution across the pixel matrix, many other signals need to be propagated to/from the whole matrix. Those also pose challenges in terms of propagation delay due to technology scaling, with increasing resistivity of long global interconnect lines affecting the skew along the large IC. This problem complicates for signals which are propagated after some logical computation (e.g. arbitration signals, address and readout data). The issue is initially discussed in Section 5.4.1 and possible solutions are proposed. The implementation and further optimisation stages performed to achieve timing closure in the RD53A chip are reported in Section 5.4.2.

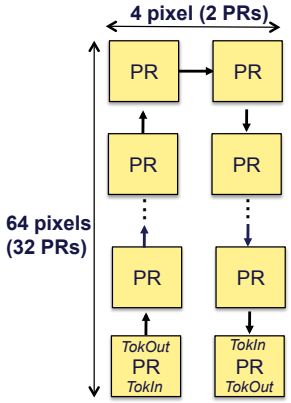

Figure 5.15: Propagation of the token signal across a double PR column (4×64 pixels) featuring the 2×2 PR distributed architecture from the FE65-P2 chip prototype.

5.4.1 Preliminary study on signal propagation across pixel regions

A preliminary study was carried out at early stages of the design to evaluate if the logic functionality in the digital array could be maintained after the significant delay degradation caused from TID effects. In particular, a double-PR column (4×64 pixels) from the FE65-P2 small-scale chip prototype (Figure 5.15), has been considered. Timing has been checked after synthesis using liberty files with modelled radiation (at this initial stage both 200 Mrad and 500 Mrad models were evaluated). The only timing paths seen to be critical are those crossing a chain of regions along the column (e.g. token path for arbitration and data bus). Indeed, such signals should ideally cross a whole chain of gates in one clock cycle. The token signal defines which pixel region is allowed to be read out first among those who have triggered data (priority is given to the first in the chain): in the FE65-P2 prototype the signal was propagated along 64 PRs by a daisy chain of OR gates, with a up-and-down path along the array. With such a propagation of signals across PRs, accumulated delays are not sustainable for a large IC: timing delay added per region exceeds by far the allowed limit for timing closure in the order of ∼100 ps (needed to propagate across 200 2×2 PRs in less than 25 ns). Even if at synthesis stage

the tool does not yet have full parasitic information and timing delays are underestimated, the investigation has allowed first conclusions on the design approach:

1. it is possible for the pixel region to operate functionally at 40 MHz using a highly integrated library (9-tracks) after radiation;

2. critical timing paths are those which propagate all along the columns and they would have been problematic when scaling to a full column of 400 pixels, independent of radiation;

3. a design strategy is required in order to meet timing for critical signals propagating along columns and needs to be conservative to include radiation degradation.

The proposed approach is shown in Figure 5.16 for the token signal, but can similarly be adopted for the data bus. The use of OR gates is conceptual, whereas they can be mapped into inverted logic with NAND gates, which suffer less from radiation effects. An additional hierarchical level, i.e. a pixel core made of multiple regions, is adopted. This allows faster token passing, by reducing the number of OR gates propagating the critical signals in the chain. Recalling the classical strategy used for adders, the chosen approach is referred to as "token-look-ahead". This solution is a compromise between reduced number of gates in the chain and limited increase in parasitics of long lines to connect them, in a similar way as it has been seen for clock propagation. The choice of a 8×8 core size has been also motivated by the other set of reasons already discussed in Section 5.3.2. With respect to Figure 5.16, the chosen N corresponds to eight pixels. It should be underlined that as far as the token and data propagation are concerned, it is not mandatory for them to be propagated in exactly one clock cycle. Indeed, a re-synchronisation of the signals from the array can be performed in the chip bottom and the FSMs can be made capable of waiting a higher number of clock cycle before processing the data. This is not ideal since it increases the latency readout of data packets, but it is at the same time a possible solution in case the "look-ahead" scheme is not sufficient to readout data in one clock cycle.

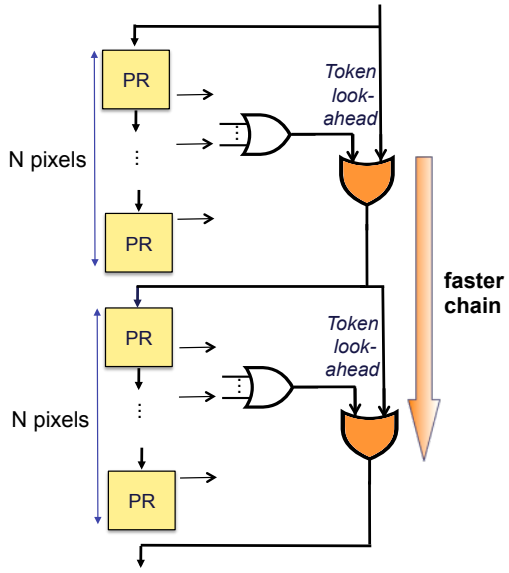

Figure 5.16: "Token-look-ahead" approach proposed to speed-up data propagation along columns (critical specially including radiation degradation).

An additional class of signals which require propagation along the column are global signals distributed to the whole pixel matrix (e.g. trigger, timestamp count, etc.). In this case, no major logic calculation is required and therefore the timing was considered to be less critical at early stages. On the other hand, in this case the propagation requirement is stricter (25 ns at most) for the whole column, since the signals are essential to properly readout triggered events. Further constraints related to such signals and the detailed approach followed are discussed in Section 5.4.2.1.

5.4.2 Optimised RD53A design and results

The detailed optimisation of the RD53A pixel array for timing is summarised in this Section. The focus is on timing-critical signals propagated along the array: inputs to the matrix are discussed in Section 5.4.2.1, wheareas token and readout data are treated in Section 5.4.2.2.

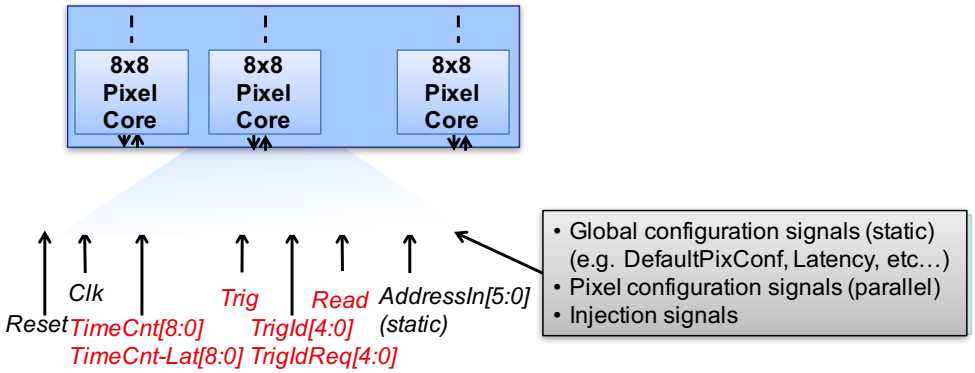

Figure 5.17: Pixel array inputs to each core column, with emphasis on signals whose timing is critical for correct data readout (highlighted in red).

5.4.2.1 Timing critical input signals to the array

In addition to the clock signal, featuring skew adjustment to compensate for the propagation delay along the column, other input signals to the array have strict timing requirements. The most relevant signals with this characteristic are shown in Figure 5.17. For example, the bunch crossing timestamp count (*TimeCnt*) and trigger (*Trig*) must be received in the correct clock cycle for a proper association of the stored/readout information and the event which generated it. The role of the additional signals, already described in Section 3.2.3 is herein reminded: the timestamp counter subtracted by the latency (*TimeCnt - Lat*) is used in the pixel regions to detect the expiration of the trigger latency for stored hit data; the trigger identifiers (*TrigId* and *TrigIdReq*) are used to match the specific event which is being read from the periphery (*TrigIdReq*) while also subsequent events may have been triggered (*TrigId* counts triggers as they are received); an acknowledge signal (*Read*) is used to allow cores with triggered data to put them into the data bus.

A known approach to reduce delays of a combinational logic chain in a synchronous design is to partition it into smaller sections, separating them with sequential elements (e.g. flip-flops), also known as pipeline design [106]. With negligible cost, a 1-stage pipeline can be added at the bottomo of each column (where power/area constraints are not as tight as in the pixel array)

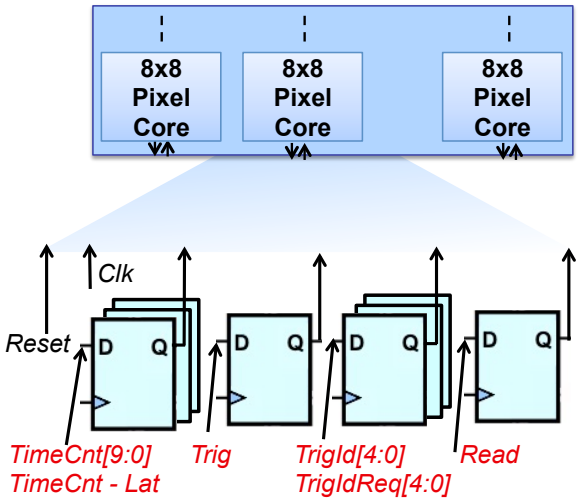

Figure 5.18: Pixel array timing critical inputs being re-synchronised in each column, to partition the timing paths from the chip bottom to the matrix.

for all the signals listed above. This allows to assure that the propagation delay is exclusively along the column, eliminating any logical elaboration and interconnect delay located in the chip bottom. This concept is sketched in Figure 5.18.

Independently from the presence of the pipeline stage, another crucial aspect related to the input signals to the pixel array is at which clock edge they are launched. In order to have a full 25 ns window available for signal propagation, the same edge used to receive signals (i.e. rising edge) was initially employed. This choice has a side-effect on the clock tree in the chip bottom controlling the launching flip-flops: their clock needs to be "almost" in phase with the clock of the cores. This means that an insertion delay similar to the skew compensation one is necessary and some additional margin is required to avoid hold violations in the cores at the bottom. Even if this could be initially achieved for a single core column, complexity arising from a big matrix and multiple (also extreme) timing corners, made clock tree synthesis hard to tune, computationally heavy and with unreliable results across subsequent iterations. For this reason, it was preferred to change the launching edge to the falling one; at the same time the window available for signal propagation was reduced

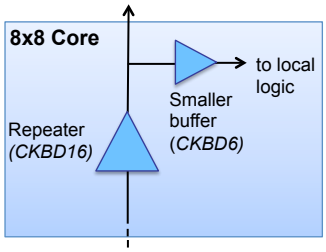

Figure 5.19: Signal propagation of timing critical inputs to the array, both from core to core and locally.

to 12.5 ns. At this point, a stricter requirement is forced on signal propagation. For this reason, in order to avoid too slow propagation, ineffective timing optimisation performed by the tools (e.g. addition of multiple in-out buffers) and assure more deterministic results across multiple P&R iterations, a well-defined approach has been adopted. In the RTL a repeater with high driving strength has been instantiated in each core for input-output propagation, followed by a buffer to drive local nets, as shown in Figure 5.19. Moreover, a 2-stage routing has been adopted in order for delays to be as independent as possible from specific flow iterations. In particular, the first routing stage involves only nets connecting input and output pins, whereas the rest of the local routing takes place only at the second stage. The goal is to use at best routing resources for critical signal propagation. It has a positive impact both for inputs and the outputs of the cores. In the layout in Figure 5.20, particularly straight vertical routing lines can be noticed after the first routing stage. The rest of the connectivity in other portions of the layout is coming from the previous design stage (CTS). In Table 5.3, propagation delays are reported across technology corners for the falling edge (i.e. launching edge) of timing critical input signals to the RD53A matrix. The trigger and bunch crossing count signals are shown as examples. With this approach, static timing analysis has succeeded for RD53A, as far as this category of signals are concerned. Thanks to the use of repeaters, the delays scale linearly for increasing column height: a chip with double size can still meet the constraint of total skew lower than 12.5 ns. Before targeting the RD53A design to a half-size column, this was actually

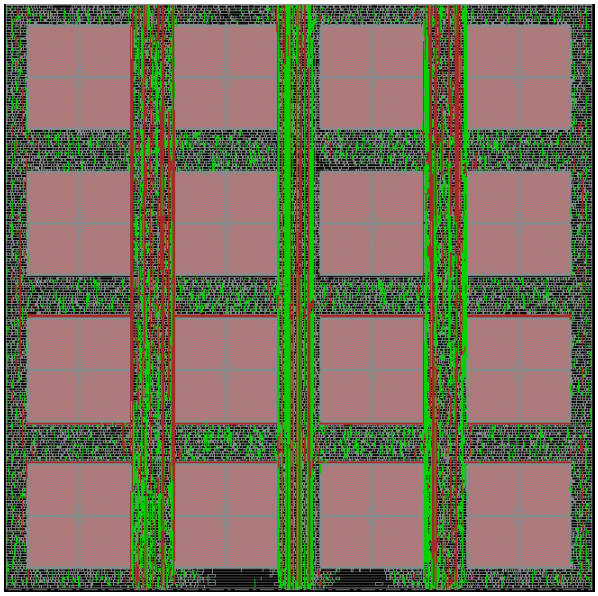

Figure 5.20: Routing of vertical nets connecting input and output pins for signals propagating from one core to the other along the column. Vertical metals M3 and M5 are shown respectively in green and red.

Table 5.3: Propagation delay of the trigger and of the bunch crossing count accumulated along the RD53A core column (192 pixels).

Voltage	Process	Temperature	*Trigger* column skew (fall)	*TimeCnt* column skew (fall)
1.32	FF	−40°C	1.7 ns	1.8 ns
1.2	TT	25°C	2.3 ns	2.2 ns
1.2	200 Mrad	25°C	3.3 ns	3.3 ns
0.9	SS	−40°C	6.1 ns	6 ns

verified with the timing analysis tools.

Power impact of bunch count distribution The distribution of many global signals across the full array has also an impact on power consumption which it is worth being quantified. As already discussed, it was chosen to distribute twice the bunch crossing timestamp count (*TimeCnt[8:0]* and *TimeCnt−Lat[8:0]*) to detect the trigger latency expiration in the array. This approach was implemented already in FE65-P2, in place of alternative solutions (e.g. latency memory counters as for the ATLAS FE-I4 [89] described in Section 3.2.1). The aim has been to minimise the logic in the array and reducing power. The power impact of the core-based distribution of the two bunch counts (including the buffers shown in Figure 5.19), is herein studied for the large scale RD53A chip. Power consumption of both buses is reported in Table 5.4 for one core, for one pixel and in percentage with respect to the total power. It can be concluded that the distribution of the global timestamps

Table 5.4: Power consumption of the core-based distribution of the two bunch counts, both as absolute value and in percentage with respect to the digital pixel array.

	Power per core (μW)	Power per pixel (μW)	Percentage of total pixel power
Bunch count (*TimeCnt[8:0]*)	10.6	0.17	3.6%
Second bunch count (*TimeCnt−Lat[8:0]*)	12.1	0.19	4%

with required timing constraints has not a very significant impact on power

(although not negligible), contributing for less than $\sim 8\%$ to the overall power consumption.

5.4.2.2 Arbitration scheme and data readout timing

As introduced in Section 5.4.1, arbitration and data signals are the most critical timing paths along the array column and travers multiple logical gates (not only repeaters as in the case of inputs). It has been already described as the "per-region" propagation has been changed to a "per-pixel-core" approach in order to speed-up timing paths. A simplified block diagram of digital core is shown in Figure 5.21, including the common logic for data readout shared between pixel regions at the core level. The *Token* signal is propagated across

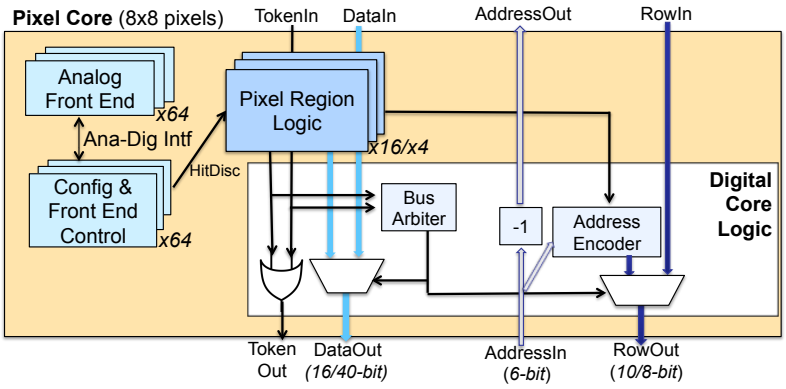

Figure 5.21: Block diagram of the readout of the pixel core. It includes 64 pixels, made of 64 AFEs and dedicated AFE control and pixel configuration logic. In the DBA (CBA) architecture 16 (4) PRs are instantiated. The region logic is identical for the Lin. and Diff. FE, whereas different for the Synch. FE, integrated with the CBA digital architecture. The core hierarchical level has been exploited to gather common digital core logic used for arbitration and data readout.

each region with a OR-chain as in the FE-I4 and FE65-P2, but a look-ahead approach is adopted from core to core. The *Token* can be seen as a combination of a readout request and priority encoding. It is indeed used from each region to sense whether another on top of it (i.e. with priority) has raised a request to read out data: if this is not the case, the specific PR can raise the request itself. The column control at the chip bottom keeps reading data

Table 5.5: Propagation delay of the token, data and address accumulated along the RD53A core column (192 pixels) for the DBA. For multi-bit signals, the worst case is reported.

Voltage	Process	Temp.	*TokOut* column skew (rise)	*DataOut* column skew (rise)	*RowOut* column skew (rise)
1.32	FF	−40°C	8.2 ns	6.6 ns	7.6 ns
1.2	TT	25°C	10.8 ns	8.5 ns	10.2 ns
1.2	200 Mrad	25°C	16.9 ns	13.5 ns	15.8 ns
0.9	SS	−40°C	31.7 ns	36.7 ns	30 ns

from a triggered event until the *Token* stays high. The *Token* signal of the regions in a core is also used to determine the address of the PR in the core (*Address Encoder*) and it is combined with the core address (*AddressIn*) to determine the full PR address. Based on this arbitration, if the core is granted access to the data bus, both ToT and address data are sent respectively on *DataOut* and *RowOut*. Otherwise, data from cores on top are simply propagated. Within a core, ToT data from multiple pixel regions are "forced" in RTL to be combined through a two OR stages with 4 inputs, to simplify the in-core *DataOut* path optimization performed by the synthesizer. The data packet propagated at the core column level is made of ToT data from a pixel region and its address, as shown in Figure 5.22 for the DBA. The core column address is added afterwards in the digital chip bottom during event assembly. At the core level, the packet size is digital-architecture dependent: an adapter is used for each CBA core column to make it to comply to the DBA packet, for seamless readout from the chip). The results of the propagation delay of the

Figure 5.22: Data packet propagated at the core column level for the DBA. ToT data are propagated by *DataOut* signals, whereas the address of the pixel region in the core and the core row are fed to the *RowOut* signals.

three data paths (token, data, address) are reported in Table 5.5 for a DBA column. It is evident that it was not possible to fit in one clock cycle for the slowest corner. For this reason, the FSM (in the chip bottom) controlling the core column has to wait 2 clock cycles for token and data to propagate, be-

fore processing them. Moreover, the possibility of configuring a longer waiting time is foreseen by global configuration for testing purposes, especially during radiation testing. As far as static timing analysis is concerned, this design choice has been taken into consideration with proper multi-cycle-path timing constraints (*set_multicycle_path*), i.e. the setup timing check on the flip flops receiving the token and data are performed at the 2^{nd} (or 3^{rd}) 40 MHz clock cycle. These timing exceptions have been verified by carrying out detailed gate-level simulation with delay back-annotationat top level, both of different core columns and full-chip.

It can be noticed that scaling up to a double-size chip such a readout latency (data waiting time) could double. On this aspect, simulations must be performed to study whether readout rates can be substained with the defined scheme. Otherwise, further approaches to speed-up the readout will have to be investigated e.g. adoption of faster low V_t standard cells, more detailed tuning of the netlist optimisation, pipeline stages along the column etc. The adoption of low V_t cells for the data bus was actually studied as a proof of concept and was seen to give ∼30% delay improvement.

5.5 Single event upset tolerance of RD53A digital pixel matrix

Design considerations regarding SEU tolerance of the digital pixel array are discussed in this Section, although not fully addressed during the design of the RD53A chip. In particular, the adoption of radiation-hard techniques for pixel configuration is evaluated in Section 5.5.1, whereas preliminary SEU simulations results of a 8×8 pixel core are reported in Section 5.5.2.

5.5.1 Pixel configuration

Pixel configuration registers have to cope with a too high failure rate due to the harsh radiation environment of the target application (as estimated in Section 5.2.2). The desired cross-section is in the order of 1000× lower than that offered by non-protected registers. For this reason, SEU-tolerant design

techniques need to be evaluated to address the issue. The adoption of TMR techniques has been initially discarded due to the strict area limitation in the pixel array. Given that triplication and voting logic cause more than 3× area increase, triplicated pixel configuration would occupy at least 10% of the area available. ECC schemes such as Hamming have a even higher overhead for the low number of bits in each pixel. Instead, handling the configuration of multiple pixels in common banks of register was seen to cause undesired placement and routing congestion complications. These initial considerations have discouraged a more in-depth investigation of techniques based on hardening at system-level. Instead, hardening by cell design seemed a more attractive solution due to the lower area overhead. The design of a 8-bit DICE cell in 65 nm technology has been carried out in the context of RD53, by the same team which implemented the rad-hard cell integrated in the ATLAS FE-I4 pixel configuration [130]. A DICE latch has redundant storage nodes and restores the original state when an SEU error is introduced in one node. Its basic building block and the functionality concept is shown in Figure 5.23. If both nodes storing the same value are upset at the same time (either X1 and X3 or X2 and X4, sensitive pairs), the SEU has effect and it is not corrected. In

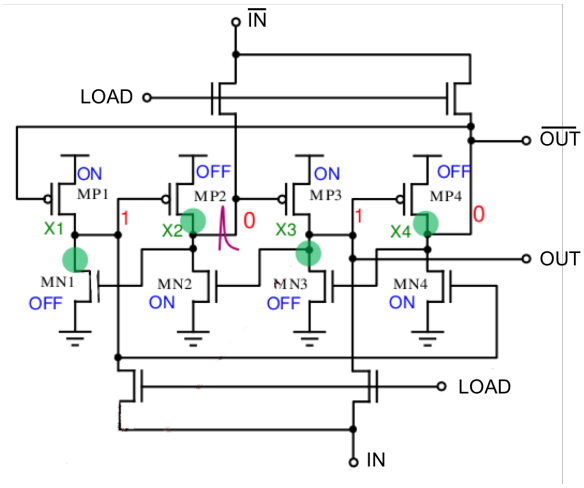

Figure 5.23: DICE latch structure and functionality. The case in which a 1-value is stored is shown: an upset of one of the nodes, e.g. X2, does not propagate to the following nodes and gets overridden by the previous one [130].

Table 5.6: 8-bit DICE latch area overhead versus 8 standard latches.

Cell	Cell area overhead	Overall digital area utilisation (Diff. FE flavour)
1-bit latch (8 cells)	-	82%
8-bit DICE latch	2.2×	86%

order to limit the probability of multiple critical nodes to be affected at once, an interleaved layout was implemented by the designers, separating as much as possible the sensitive pair nodes. Indeed, thanks to the implementation of a multi-bit latch, it is possible to separate nodes of the same latch with good utilisation of the area in the cell layout. The area overhead compared to a standard latch from the technology foundry is reported in Table 5.6, based on its integration with the Diff. FE (since this design variation offers area margin for additional logic). Although the design density is at an acceptable level for closing the design with digital design tools, the adoption of the standard flow was not sufficient to achieve a successful integration of the 8-bit DICE latch. Indeed, the cell has a more complex layout than standard cells: it occupies 4 rows, uses routing resources up to metal 4 and features 8×3 pins. By default, automated design tools place the cells very close to each other, since they have common inputs. Significant routing congestion can be already noticed after placement (based on initial trial routing), as shown in Figure 5.24 (left). The "big" DICE cells can be recognised as they are higher than the rest of the cells. If no dedicated approach is used, this leads in the following phases of the design flow to many violation (e.g. shorts, spacing rules, etc.) all around the custom cells, as highlighted in Figure 5.24 (right). For this reason, the placement of the cells has been guided to keep them separate from each other and close to the correspondent analog pixel. This has been achieved by defining floorplan regions for each DICE cell to be place in the assigned area. The defined regions are shown in Figure 5.25 for a portion of the pixel core (on top). This floorplan constraint (*createRegion*) does not prevent other modules to be placed within the region area. The reason for placing them vertically with respect to the front ends is related to the routing of power distribution. Indeed, digital power is distributed along the column with top level metals (metal 10, 9 and

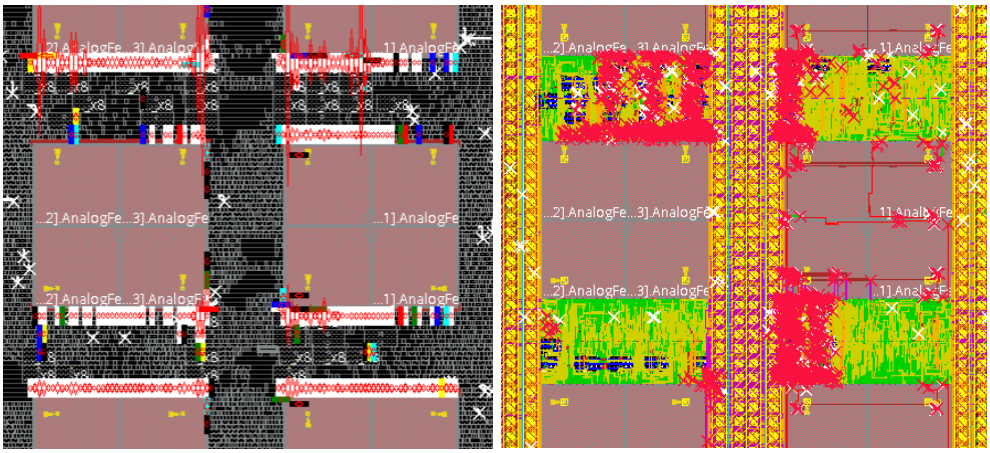

Figure 5.24: Zoom of the central area of the core where most of DICE latches are automatically placed by tools. High routing congestion issues are shown on the left, where big red "diamonds" and light colours point to limited routing resources. Congestion's effects on detailed routing are highlighted on the right (e.g. thousands of routing shorts are visible as red crosses).

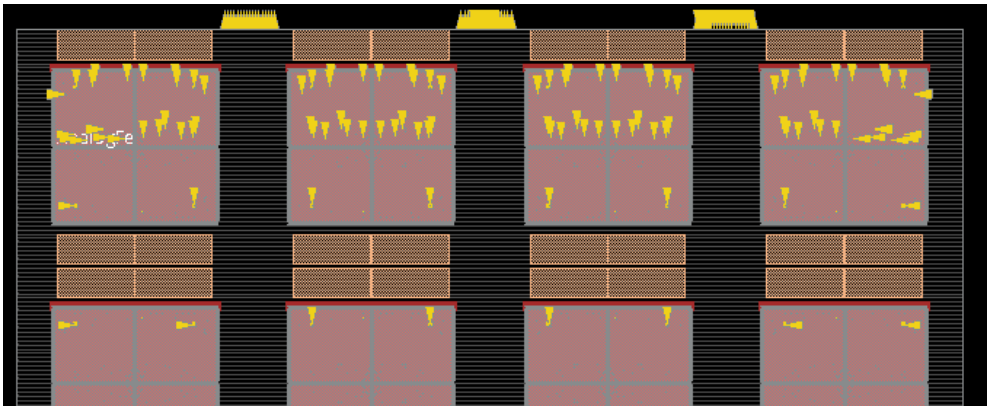

Figure 5.25: Floorplan regions assigned to each DICE latch, close to the correspondent analog front end and distant from each other.

8) in the space on the sides of the analog islands. Routing vias are then connecting metal 8 down to metal 1. This has been seen to complicate the routing of the DICE latches, e.g. easily causing shorts in intermediate metals: for this reason the alternative position was chosen. Moreover, once the placement is performed, it has been necessary to set the state of the cells to partially fixed, to avoid a complete rearrangement of their position during routing optimisation. Again, this has been only seen to give worst routing and congestion. It is clear that the use of custom cells bigger than usual standard cells (for which algorithms are optimised in the tools) brings additional complication to the design process. This has been also addressed by defining high congestion effort during placement and by setting a so-called "module padding" for the custom cells. It allows placement tools to consider the size of DICE cells a factor of 1.2 higher than their actual size, which helps to reduce congestion around them.

The integration of the DICE latch was successful with the Diff. FE, but could not be achieved with the Lin. FE due to area limitations. It was also not studied with the Sync. FE. The number of configuration bits of the latter is lower (only 3 in total), which makes the use of a 8-bit cell rather inefficient. Finally, it should be mentioned that the SEU testing of the cell (by the CPPM group who designed it) has shown an improved cross-section of 9× with respect to standard latches. This is unfortunately not sufficient to comply with specifications and it has reduced the interest on its use. For a combination of these factors, DICE latches were not integrated in the RD53A prototype. For future developments, it may actually be preferable to investigate a standard TMR approach, accepting a higher area overhead for a lower cross-section and possibly less routing congestion issues.

An alternative approach for SEU protection of pixel configuration is implemented in the RD53A prototype, which does not involve any hardening technique to be implemented in the digital pixel core. The so-called trickle configuration implies that pixel registers are refreshed continuously during operation without affecting data taking. Therefore, the chip is supposed to no longer require to hold configuration for more than a fraction of a second at a time. This feature has been implemented in the chip but not extensively simulated (in particular with SEU injection). It will be an interesting option

to be tested in the prototyped chip and to be further investigated for future developments.

5.5.2 SEU tolerance of the digital pixel array logic

As discussed in Section 5.2.2, the control, storage and readout logic in the pixel array is not strictly required to be protected. Such an assumption needs to be confirmed by proving that *i)* the logic is capable of self-recovering after upsets, *ii)* the inefficiency or noise hits caused are known and within acceptable criteria. Simulation with SEU injection is required and has to be integrated within the developed VEPIX53 framework. A feature of the *Cadence Incisive* simulator was evaluated, the so called *Functional Safety Simulator*, meant to check robustness of design in unplanned and unexpected events by causing SEUs (and/or SETs) during simulation. Such a tool has quickly shown to not be a proper solution for the application, since only capable of injecting errors at once in one simulation run, therefore not useful to emulate a certain SEU rate. Instead, approaches developed in the HEP context were found to be much closer to the needs of this work. In particular, the technical implementation of SEU injection has been mostly based on the one developed in the context of other hybrid pixel chips [128], which has been integrated in the VEPIX53 simulation framework. The main steps characterizing the adopted approach are herein described and graphically summarised in Figure 5.26:

1. the verilog gate-level design netlist is parsed to find all sequential elements, i.e. candidates for SEU injection;

2. hierarchical paths to the sequential elements are identified;

3. a dedicated SystemVerilog module is generated and defines functions for random SEU injection in the registers identified, with a settable rate and equal upset-probability for all the nodes;

4. SEU injection is performed on top of a simulation under operative conditions within VEPIX53, with automated checks to verify the absence of stuck conditions and to quantify losses and noise hits.

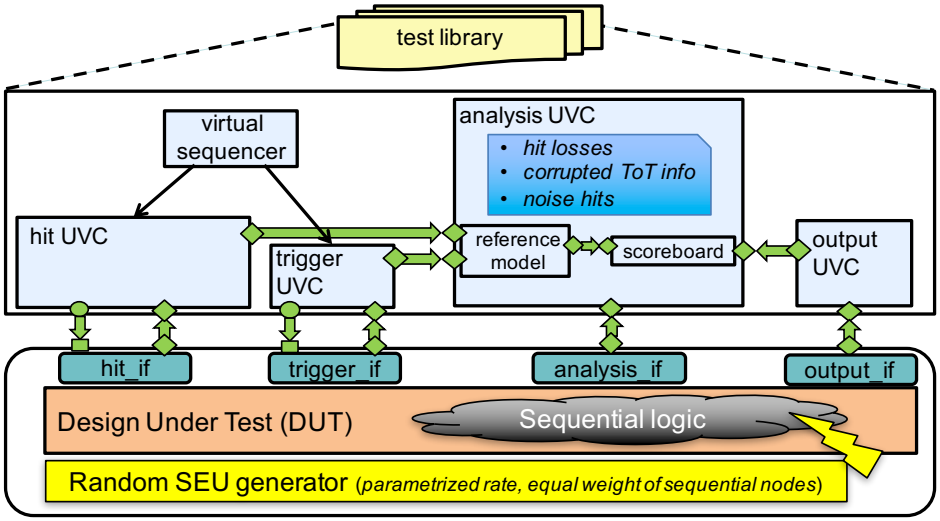

Figure 5.26: Block diagram of the simulation framework highlighting features for SEU injection.

Preliminary simulations have been carried out on the digital pixel array to assess the SEU tolerance of the DBA architecture integrated in the RD53A chip. The basic building block of the matrix has been studied, i.e. a 8×8 pixel core. Since no protection was implemented in the pixels, configuration registers have been excluded from SEU injection. Instead, verifying the functionality of the rest of the control, buffering and readout logic has been the simulation goal. In the pixel core, a total of 4864 flip-flops/latches are present, determining a worst-case failure rate of ∼1 upset every 8.2 s per pixel core (referring to Equation 5.2). For simulation purposes, a more efficient approach is the use of a much higher upset rate, in order to achieve coverage of most of the sequential elements with relatively short simulations. Mulfunctionalities can be observed and used to assess the recovering capabilities of the logic and to quantify the undesired effects. In particular, an average of 1 upset/μs/core has been set, i.e. $8.2 \cdot 10^6\times$ higher than the expected upset rate. For the results reported, a linear scaling of any SEU-induced error based on this factor is assumed.

Simulations, run for 900,000 BX cycles with the same hit and trigger conditions adopted in Section 3.2.3.3, have shown promising results. Although

data are partially corrupted by the injection, as expected, the logic is capable of recovering and correctly continuing to process subsequent incoming hits. No stuck condition and no unexpected malfunction were observed throughout the whole simulation. Thanks to the UVM scoreboard (automatically comparing expected and actual chip outputs), a first classification of errors caused by SEU injection has been performed. With respect to the expected chip output after the trigger selection (in absence of upsets), observed hit loss, corrupted ToTs and noise hits caused by SEUs are reported in Table 5.7. The coverage

Table 5.7: Observed hit losses, corrupted charge data, noise hits during simulation with SEU injection of the DBA architecture. Results are scaled down to emulate the worst case estimated rate of failure and quantified with respect to the total number of hits which are correctly readout in absence of SEU injection (i.e. 2438 hits).

	With increased SEU rate	Scaled to est. SEU rate	Ratio with respect to correctly readout hits
# Hits lost	48	$5.9 \cdot 10^{-6}$	$2.4 \cdot 10^{-9}$
# Hits with corrupted ToT	27	$3.3 \cdot 10^{-6}$	$1.3 \cdot 10^{-9}$
# Noise hits	1424	$1.7 \cdot 10^{-4}$	$7.1 \cdot 10^{-8}$

reached during simulation on the sequential elements of the pixel core is 98%. The preliminary results obtained confirm the assumption that the pixel array control and readout logic does not require SEU protection. Indeed, the logic is self-recovering after upsets and inefficiencies and noise hits are well below the acceptable criteria (i.e. 10^{-4}).

Conclusions

This thesis was dedicated to the development and optimisation of a low power integrated circuit for extreme rates and harsh radiation environments in nanometer technology. In particular, the work has focused on the optimisation and verification of the digital array logic of the RD53A readout chip prototype, targeted to the generation hybrid pixel detectors at the High-Luminosity Large Hadron Collider. The large scale prototype has been designed and submitted for production. This work has also contributed to the early development of a small-scale chip prototype, CHIPIX65, produced and tested before RD53A submission.

Modern advanced design and verification techniques have been adopted in order to address the challenging system specifications of the Phase 2 Upgrade of the CMS and ATLAS experiments, i.e. small pixel area ($50\times50\,\mu m^2$), extremely higher hit rates ($3\,GHz/cm^2$), high trigger rate (1 MHz) and long latency ($12.5\,\mu s$), low power consumption (less than $5\,\mu W$/pixel or lower) and serial powering scheme, radiation tolerance up to 500 Mrad, large scale chip format (20x11.8 mm^2, designed taking into account double the size) and system complexity (∼500M transistors), low-skew clock distribution (1-2 ns). The aforementioned requirements constitute a significant advancement with respect to the state of the art of CMS and ATLAS pixel detectors, currently featuring bigger pixel sizes (ATLAS FEI4: $50\times250\,\mu m^2$, CMS PSI46DIG: $100\times150\,\mu m^2$), an order of magnitude lower hit (100-400 MHz/cm^2) and trigger (100-200 MHz) rates, approximatively half the trigger latency and lower radiation tolerance (120-400 Mrad). The reduced pixel size and increased input rates and trigger latency have required the adoption of a higher density technology node (65 nm),

in order to accommodate all the required buffering and logic resources in the limited area and profit from technology scaling to limit power consumption.

The architecture optimisation of digital array logic treated in this work has addressed many of the required advancements. The original contribution of this thesis has covered both high-level architecture evaluation, simulation and verification aspects as well as detailed digital design including low power, timing and radiation tolerance optimisation. The main results achieved are herein summarised:

- a SystemVerilog-UVM simulation framework (VEPIX53) has been optimised with special focus on modularity and re-usability, to handle complexity and support integration of multiple DUTs at different description levels throughout the design flow, as presented in [131], [76]. Flexible stimuli generation [47] (including emulation of physics data and extreme cases), automated verification and performance assessment have been successfully implemented and extensively used. The framework is available in the HEP community for architecture optimisation, system simulation and verification;

- digital architectures for the digital matrix have been selected, simulated and compared in terms of hit loss and buffering performance thanks to the developed VEPIX53 framework. Behavioural-level studies of pixel region architectures have been reported in [90] taking into account high hit rates ($2.7\,\mathrm{GHz/cm^2}$) obtained from Monte Carlo data and long latency ($10\,\mu\mathrm{s}$); it was concluded that specifications can be met and that a distributed buffering approach is preferable in terms of dead-time losses in comparison to a scheme with centralized handling of a central FIFO (featuring region-level dead-time);

- selected architectures have been studied at RTL level as implemented in small-scale prototypes: in [77] the performance of a 2×2 pixel region architecture (already prototyped in the FE65-P2 chip) have been studied and possible optimisations to reduce losses have been proposed, in order to match with the demonstrator specifications; in [94], a novel 4×4 pixel region digital architecture with a centralised memory (with no shared

pixel region dead-time) and ToT data reduction has been presented. Results obtained based on pixel region simulations performed in VEPIX53 have guided design choices (number of ToT stored per region, buffer locations) and quantified hit losses. The main limitation found was related to the fixed pixel dead-time, causing high dead-time losses if operating with a classical ToT scheme;

- both selected architectures have been further optimised for the RD53A design. This work has focused on the distributed architecture: in Chapter 3, hit loss performance have been studied through simulations with Monte Carlo data at different positions in the detector. Losses in the order of 1% (with dead-time losses tunable through the analog front-end settings) have been achieved at the target hit (3 GHz/cm^2) and trigger rates (1 MHz) and maximum trigger latency (12.5 μs), thanks to the increased number of buffer locations and improved buffering performance obtained with an elongated pixel region shape (4×1);

- a power analysis methodology has been defined and adopted to guide the optimisation of the critical pixel array logic for the specific needs of the serial powering scheme, considering also power fluctuations. State of the art low power techniques have been analysed and selectively used depending on the applicability to the peculiar need in the context of the work. Average and peak consumption as well as area utilisation have been successfully reduced. System-level simulations including the produced digital activity have also proven the feasibility of the serial powering scheme. The aforementioned results have been partially summarised in [132], [133] and [134];

- design for the unprecedented levels of radiation (500 Mrad) has been addressed for the digital array logic in Chapter 5. Test results from irradiation campaigns at transistor level and on digital circuits have guided the selection of timing libraries adopted in the digital design. With such models including degradation effects, low-skew clock distribution (within 2 ns) has been achieved for the large IC by means of a hierar-

chical approach with skew compensation. Top level timing optimisation has been also performed for critical signals: a look-ahead approach has been defined for the token and readout bus, whereas fast propagation was obtained for inputs to the array;

- SEU tolerance of the array was also considered: the adoption of a 8-bit DICE cell has been evaluated for pixel configuration, showing a area overhead of 2.2× and routing complications to obtain a 9× cross-section reduction; SEU effects on control, storage and readout logic have been simulated for the designed distributed buffering architecture. Preliminary simulation results have shown good SEU tolerance of the logic, with losses and noise hits well below specifications;

- the coordinated effort of multiple institutes and universities part of the RD53 collaboration and CHIPIX65 projects, to which this PhD work has been an active contribution, has been reported in the following main publications [135], [136], [87] for RD53, and [93], [137], [138], [139], [140] for CHIPIX65.

Bibliography

[1] ATLAS Collaboration, "ATLAS pixel detector technical design report," CERN, Tech. Rep. CERN/LHCC/98-13, 1998.

[2] W. Erdmann, "The CMS pixel detector," *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 447, no. 1, pp. 178–183, 2000.

[3] F. Antinori, "A pixel detector system for ALICE," *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 395, no. 3, pp. 404–409, 1997.

[4] J. Christiansen and M. Garcia-Sciveres, "RD Collaboration proposal: Development of pixel readout integrated circuits for extreme rate and radiation," [Online]. Available: http://cds.cern.ch/record/1553467?ln=en, CERN, July 2013.

[5] "The CHIPIX65 Project," [Online]. Available: http://chipix65.to.infn.it/, 2014.

[6] "Advanced European Infrastructures for Detectors at Accelerators (AIDA)," [Online]. Available: http://aida2020.web.cern.ch/, 2015.

[7] L. Rossi, P. Fischer, T. Rohe, and N. Wermes, *Pixel detectors: From fundamentals to applications*. Springer, 2006.

[8] L. Rossi, "Pixel detectors hybridisation," *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detec-*

tors and Associated Equipment, vol. 501, no. 1, pp. 239 – 244, 2003, proceedings of the 10th International Workshop on Vertex Detectors.

[9] A. Rivetti, *CMOS: front-end electronics for radiation sensors*, ser. Devices, circuits, and systems. Boca Raton, FL: CRC Press, 2015.

[10] Y. Allkofer, C. Amsler, D. Bortoletto, V. Chiochia, L. Cremaldi, S. Cucciarelli *et al.*, "Design and performance of the silicon sensors for the cms barrel pixel detector," *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 584, no. 1, pp. 25 – 41, 2008.

[11] M. Dinardo, "The pixel detector for the CMS phase-II upgrade," *Journal of Instrumentation*, vol. 10, no. 04, p. C04019, 2015.

[12] R. Dinapoli, M. Campbell, E. Cantatore, V. Cencelli, E. Heijne, P. Jarron *et al.*, "A front-end for silicon pixel detectors in ALICE and LHCb," *Nuclear Instruments and Methods in Physics Research, Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 461, no. 1-3, pp. 492–495, April 2001.

[13] J. Kaplon and W. Dabrowski, "Fast CMOS binary front end for silicon strip detectors at LHC experiments," *IEEE Transactions on Nuclear Science*, vol. 52, no. 6, pp. 2713–2720, Dec 2005.

[14] G. F. Knoll, *Radiation detection and measurement.* Wiley, 2010.

[15] X. Llopart, R. Ballabriga, M. Campbell, L. Tlustos, and W. Wong, "Timepix, a 65k programmable pixel readout chip for arrival time, energy and/or photon counting measurements," *Nuclear Instruments and Methods in Physics Research A*, vol. 581, pp. 485–494, Oct. 2007.

[16] P. Valerio, R. Ballabriga, and M. Campbell, "Design of the 65 nm CLICpix demonstrator chip," Nov 2012. [Online]. Available: https://cds.cern.ch/record/1507691

[17] G. Mazza, D. Calvo, P. D. Remigis, M. Mignone, J. Olave, A. Rivetti *et al.*, "The ToPiX v4 prototype for the triggerless readout

of the PANDA silicon pixel detector," *Journal of Instrumentation*, vol. 10, no. 01, p. C01042, 2015. [Online]. Available: http://stacks.iop.org/1748-0221/10/i=01/a=C01042

[18] V. Gromov, M. van Beuzekom, R. Kluit, F. Zappon, V. Zivkovic, M. Campbell *et al.*, "Development and applications of the Timepix3 readout chip," *PoS (Vertex 2011)*, vol. 46, p. 1, 2011.

[19] T. Poikela, M. D. Gaspari, J. Plosila, T. Westerlund, R. Ballabriga, J. Buytaert *et al.*, "VeloPix: the pixel ASIC for the LHCb upgrade," *Journal of Instrumentation*, vol. 10, no. 01, p. C01057, 2015. [Online]. Available: http://stacks.iop.org/1748-0221/10/i=01/a=C01057

[20] R. Horisberger, "Readout architectures for pixel detectors," *Nucl. Instrum. Meth.*, vol. A465, pp. 148–152, 2000.

[21] "LHC1: A semiconductor pixel detector readout chip with internal, tunable delay providing a binary pattern of selected events," *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 383, no. 1, pp. 55 – 63, 1996, development and Application of Semiconductor Tracking Detectors.

[22] M. Barbero, W. Bertl, G. Dietrich, A. Dorokhov, W. Erdmann, K. Gabathuler *et al.*, "Design and test of the CMS pixel readout chip," *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 517, no. 1, pp. 349 – 359, 2004. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0168900203026391

[23] G. Aad, M. Ackers, F. A. Alberti, M. Aleppo, G. Alimonti, J. Alonso *et al.*, "ATLAS pixel detector electronics and sensors," *Journal of Instrumentation*, vol. 3, no. 07, p. P07007, 2008. [Online]. Available: http://stacks.iop.org/1748-0221/3/i=07/a=P07007

[24] K. H. Wyllie, M. Burns, M. Campbell, E. Cantatore, V. Cencelli, R. Dinapoli *et al.*, "A pixel readout chip for tracking at ALICE

and particle identification at LHCb," 1999. [Online]. Available: https://cds.cern.ch/record/431352

[25] T. Hemperek, D. Arutinov, M. Barbero, R. Beccherle, G. Darbo, S. Dube *et al.*, "Digital architecture of the new ATLAS pixel chip FE-I4," in *Nuclear Science Symposium Conference Record (NSS/MIC), 2009 IEEE*, Oct 2009, pp. 791–796.

[26] D. Hits and A. Starodumov, "The CMS Pixel Readout Chip for the Phase 1 Upgrade," *Journal of Instrumentation*, vol. 10, no. 05, p. C05029, 2015. [Online]. Available: http://stacks.iop.org/1748-0221/10/i=05/a=C05029

[27] M. Garcia-Sciveres, D. Arutinov, M. Barbero, R. Beccherle, S. Dube, D. Elledge *et al.*, "The FE-I4 pixel readout integrated circuit," *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 636, no. 1, pp. S155–S159, 2011.

[28] G. A. Rinella, E. C. Gil, M. Fiorini, J. Kaplon, A. Kluge, F. Marchetto *et al.*, "Test-beam results of a silicon pixel detector with Time-over-Threshold read-out having ultra-precise time resolution," *Journal of Instrumentation*, vol. 10, no. 12, p. P12016, 2015. [Online]. Available: http://stacks.iop.org/1748-0221/10/i=12/a=P12016

[29] M. Bartók, "Simulation of the dynamic inefficiency of the CMS pixel detector," *Journal of Instrumentation*, vol. 10, no. 05, p. C05006, 2015. [Online]. Available: http://stacks.iop.org/1748-0221/10/i=05/a=C05006

[30] N. Wermes, "Pixel detectors for particle physics and imaging applications," *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 512, no. 1, pp. 277 – 288, 2003, proceedings of the 9th European Symposium on Semiconductor Detectors:

New Developments on Radiation Detectors. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0168900203019053

[31] "Development and Applications of the Timepix3 chip," [Online]. Available: http://slideplayer.com/slide/7465783/, accessed: 2017-06-30.

[32] P. Delpierre, "A history of hybrid pixel detectors, from high energy physics to medical imaging," *Journal of Instrumentation*, vol. 9, no. 05, p. C05059, 2014. [Online]. Available: http://stacks.iop.org/1748-0221/9/i=05/a=C05059

[33] J. Jakůbek, "Semiconductor pixel detectors and their applications in life sciences," *Journal of Instrumentation*, vol. 4, no. 03, p. P03013, 2009. [Online]. Available: http://stacks.iop.org/1748-0221/4/i=03/a=P03013

[34] R. Ballabriga, J. Alozy, G. Blaj, M. Campbell, M. Fiederle, E. Frojdh *et al.*, "The Medipix3RX: a high resolution, zero dead-time pixel detector readout chip allowing spectroscopic imaging," *Journal of Instrumentation*, vol. 8, no. 02, p. C02016, 2013. [Online]. Available: http://stacks.iop.org/1748-0221/8/i=02/a=C02016

[35] X. Llopart, M. Campbell, R. Dinapoli, D. S. Segundo, and E. Pernigotti, "Medipix2: A 64-k pixel readout chip with 55- mu;m square elements working in single photon counting mode," *IEEE Transactions on Nuclear Science*, vol. 49, no. 5, pp. 2279–2283, Oct 2002.

[36] K. Klein, "The Phase-2 Upgrade of the CMS Tracker," CERN, Geneva, Tech. Rep. CERN-LHCC-2017-009. CMS-TDR-014, Jun 2017. [Online]. Available: https://cds.cern.ch/record/2272264

[37] M. Garcia-Sciveres, "RD53A Integrated Circuit Specifications," CERN, Geneva, Tech. Rep. CERN-RD53-PUB-15-001, Dec 2015. [Online]. Available: https://cds.cern.ch/record/2113263

[38] C. Grupen and B. Shwartz, *Particle detectors*. Cambridge university press, 2008.

[39] B. Muratori, "Luminosity and luminous region calculations for the LHC," CERN, Geneva, Tech. Rep. LHC-PROJECT-NOTE-301, Sep 2002. [Online]. Available: https://cds.cern.ch/record/691967

[40] "ATLAS experiment - Luminosity Public Results (Run 1)," [Online]. Available: https://twiki.cern.ch/twiki/bin/view/AtlasPublic/LuminosityPublicResults, accessed: 2017-10-26.

[41] B. Schmidt, "The High-Luminosity upgrade of the LHC: Physics and Technology Challenges for the Accelerator and the Experiments," *Journal of Physics: Conference Series*, vol. 706, no. 2, p. 022002, 2016. [Online]. Available: http://stacks.iop.org/1742-6596/706/i=2/a=022002

[42] "The HL-LHC project," [Online]. Available: http://hilumilhc.web.cern.ch/about/hl-lhc-project, accessed: 2017-07-04.

[43] "ATLAS experiment - Luminosity Public Results (Run 2)," [Online]. Available: https://twiki.cern.ch/twiki/bin/view/AtlasPublic/LuminosityPublicResultsRun2, accessed: 2017-10-26.

[44] S. Bonacini, P. Valerio, R. Avramidou, R. Ballabriga, F. Faccio, K. Kloukinas *et al.*, "Characterization of a commercial 65 nm CMOS technology for SLHC applications," *Journal of Instrumentation*, vol. 7, no. 01, p. P01015, 2012. [Online]. Available: http://stacks.iop.org/1748-0221/7/i=01/a=P01015

[45] M. Garcia-Sciveres, "RD53A Integrated Circuit Specifications," CERN, Geneva, Tech. Rep. CERN-RD53-PUB-15-001, Dec 2015. [Online]. Available: https://cds.cern.ch/record/2113263

[46] "IEEE Standard for Universal Verification Methodology Language Reference Manual," *IEEE Std 1800.2-2017*, pp. 1–472, May 2017.

[47] S. Marconi, E. Conti, P. Placidi, J. Christiansen, and T. Hemperek, "The RD53 collaboration's SystemVerilog-UVM simulation framework and its general applicability to design of advanced pixel readout chips,"

Journal of Instrumentation, vol. 9, no. 10, p. P10005, 2014. [Online]. Available: http://stacks.iop.org/1748-0221/9/i=10/a=P10005

[48] A. Caratelli, D. Ceresa, J. Kaplon, K. Kloukinas, and S. Scarfi, "Readout architecture for the Pixel-Strip module of the CMS Outer Tracker Phase-2 upgrade," CERN, Geneva, Tech. Rep. CMS-CR-2016-405, Nov 2016. [Online]. Available: http://cds.cern.ch/record/2235518

[49] S. Borkar, "Design challenges of technology scaling," *Micro, IEEE*, vol. 19, no. 4, pp. 23–29, Jul 1999.

[50] R. Jain, A. Chandrasekaran, G. Elias, and R. Cloutier, "Exploring the Impact of Systems Architecture and Systems Requirements on Systems Integration Complexity," *IEEE Systems Journal*, vol. 2, no. 2, pp. 209–223, June 2008.

[51] V. Berman, "System-level design language standard needed," *IEEE Design Test of Computers*, vol. 21, no. 6, pp. 592–593, Nov 2004.

[52] "IEEE Standard for SystemVerilog–Unified Hardware Design, Specification, and Verification Language," *IEEE Std 1800-2012 (Revision of IEEE Std 1800-2009)*, pp. 1–1315, Feb 2013.

[53] S. Sutherland, S. Davidmann, and P. Flake, *SystemVerilog for Design: A Guide to Using SystemVerilog for Hardware Design and Modeling*, 2nd ed. Springer Publishing Company, Incorporated, 2010.

[54] C. Spear, *SystemVerilog for Verification, Second Edition: A Guide to Learning the Testbench Language Features*, 2nd ed. Springer Publishing Company, Incorporated, 2008.

[55] A. Jain, H. Gupta, S. Jana, and K. Kumar, "Early development of UVM based verification environment of image signal processing designs using TLM reference model of RTL," *CoRR*, vol. abs/1408.1150, 2014. [Online]. Available: http://arxiv.org/abs/1408.1150

[56] B. C. Lim, J. E. Jang, J. Mao, J. Kim, and M. Horowitz, "Digital analog design: Enabling mixed-signal system validation," *IEEE Design Test*, vol. 32, no. 1, pp. 44–52, Feb 2015.

[57] C. Liang, G. Zhong, S. Huang, and B. Xia, "UVM-AMS based subsystem verification of wireless power receiver SoC," in *2014 12th IEEE International Conference on Solid-State and Integrated Circuit Technology (ICSICT)*, Oct 2014, pp. 1–3.

[58] S. Simon, D. Bhat, A. Rath, J. Kirscher, and L. Maurer, "Coverage-driven mixed-signal verification of smart power ICs in a UVM environment," in *2017 22nd IEEE European Test Symposium (ETS)*, May 2017, pp. 1–6.

[59] A. Jain, G. Bonanno, H. Gupta, and A. Goyal, "Generic SystemVerilog Universal Verification Methodology based reusable verification environment for efficient verification of image signal processing IPs/SOCs," *International Journal of VLSI design & Communication Systems (VLSICS)*, vol. 3, no. 6, 2012.

[60] A. El-Yamany, S. El-Ashry, and K. Salah, "Coverage Closure Efficient UVM Based Generic Verification Architecture for Flash Memory Controllers," in *2016 17th International Workshop on Microprocessor and SOC Test and Verification (MTV)*, Dec 2016, pp. 30–34.

[61] B. P. Biswal, A. Singh, and B. Singh, "Cache coherency controller verification IP using SystemVerilog Assertions (SVA) and Universal Verification Methodologies (UVM)," in *2017 11th International Conference on Intelligent Systems and Control (ISCO)*, Jan 2017, pp. 21–24.

[62] M. Mefenza, F. Yonga, and C. Bobda, "Automatic UVM Environment Generation for Assertion-Based and Functional Verification of SystemC Designs," in *2014 15th International Microprocessor Test and Verification Workshop*, Dec 2014, pp. 16–21.

[63] M. Barnasconi, M. Dietrich, K. Einwich, T. Vörtler, J. P. Chaput, M. M. Louërat *et al.*, "UVM-SystemC-AMS Framework for System-Level Ver-

ification and Validation of Automotive Use Cases," *IEEE Design Test*, vol. 32, no. 6, pp. 76–86, Dec 2015.

[64] R. Drechsler, C. Chevallaz, F. Fummi, A. J. Hu, R. Morad, F. Schirrmeister *et al.*, "Panel: Future SoC verification methodology: UVM evolution or revolution?" in *2014 Design, Automation Test in Europe Conference Exhibition (DATE)*, March 2014, pp. 1–5.

[65] T. M. Inc. (2017, Jan.) Simulink, MATLAB. [Online]. Available: https://www.mathworks.com/

[66] B. Stroustrup, *The C++ Programming Language*, 3rd ed. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 2000.

[67] "IEEE Standard for Standard SystemC Language Reference Manual," *IEEE Std 1666-2011 (Revision of IEEE Std 1666-2005)*, pp. 1–1163, Jan 2012.

[68] D. Markovic, C. Chang, B. Richards, H. So, B. Nikolic, and R. W. Brodersen, "ASIC Design and Verification in an FPGA Environment," in *2007 IEEE Custom Integrated Circuits Conference*, Sept 2007, pp. 737–740.

[69] D. Arutinov, M. Barbero, R. Beccherle, V. Buscher, G. Darbo, R. Ely *et al.*, "Digital architecture and interface of the new ATLAS Pixel Front-End IC for upgraded LHC luminosity," in *2008 IEEE Nuclear Science Symposium Conference Record*, Oct 2008, pp. 1923–1928.

[70] S. Heuvelmans and M. Boerrigter, "A pixel read-out architecture implementing a two-stage token ring, zero suppression and compression," *Journal of Instrumentation*, vol. 6, no. 01, p. C01093, 2011. [Online]. Available: http://stacks.iop.org/1748-0221/6/i=01/a=C01093

[71] N. Costantino, G. Borgese, S. Saponara, L. Fanucci, J. Incandela, and G. Magazzu, "Development, design and characterization of a novel protocol and interfaces for the control and readout of front-end electronics

in high energy physics experiments," *IEEE Transactions on Nuclear Science*, vol. 60, no. 1, pp. 352–364, Feb 2013.

[72] W. Rosenstiel, S. Swan, F. Ghenassia, P. Flake, and J. Srouji, "Systemc and systemverilog: Where do they fit? where are they going?" in *Proceedings Design, Automation and Test in Europe Conference and Exhibition*, vol. 1, Feb 2004, pp. 122–127 Vol.1.

[73] V. Berman. A Tale of two languages: SystemC AND SystemVerilog. [Online]. Available: http://chipdesignmag.com/display.php?articleId=116

[74] V. Zivkovic, J.-D. Schipper, M. Garcia-Sciveres, A. Mekkaoui, M. Barbero, G. Darbo *et al.*, "The FE-I4 pixel readout system-on-chip resubmission for the insertable B-Layer project," *Journal of Instrumentation*, vol. 7, no. 02, p. C02050, 2012. [Online]. Available: http://stacks.iop.org/1748-0221/7/i=02/a=C02050

[75] A. Fiergolski, "Simulation environment based on the Universal Verification Methodology," *Journal of Instrumentation*, vol. 12, no. 01, p. C01001, 2017. [Online]. Available: http://stacks.iop.org/1748-0221/12/i=01/a=C01001

[76] S. Marconi, E. Conti, J. Christiansen, and P. Placidi, "Reusable SystemVerilog-UVM design framework with constrained stimuli modeling for High Energy Physics applications," in *2015 IEEE International Symposium on Systems Engineering (ISSE)*, Sept 2015, pp. 391–397.

[77] E. Conti, S. Marconi, T. Hemperek, J. Christiansen, and P. Placidi, "Performance evaluation of digital pixel readout chip architecture operating at very high rate through a reusable UVM simulation framework," in *2016 IEEE Nuclear Science Symposium, Medical Imaging Conference and Room-Temperature Semiconductor Detector Workshop (NSS/MIC/RTSD)*, Oct 2016, pp. 1–4.

[78] "Aurora 64B/66B Protocol Specification," [Online]. Available: https://www.xilinx.com/support/documentation/ip_documentation/aurora_64b66b_protocol_spec_sp011.pdf, 2014.

[79] S. Marconi and P. Placidi, "A Flexible simulation and verification framework for next generation hybrid pixel readout chips in High Energy Physics," 2014, presented 2014. [Online]. Available: https://cds.cern.ch/record/1711786

[80] *ROOT - A Data Analysis Framework*, 2014. [Online]. Available: http://root.cern.ch/drupal/

[81] M. Garcia-Sciveres, B. Nachman, and F. Wang, "Optimal use of charge information for HL-LHC pixel readout."

[82] "TSMC 65 nm technology overview (MPW)," [Online]. Available: http://www.europractice-ic.com/technologies_TSMC.php?tech_id=65nm, 2017.

[83] M. Garcia-Sciveres, "The RD53A Integrated Circuit," CERN, Geneva, Tech. Rep. CERN-RD53-PUB-17-001, Oct 2017. [Online]. Available: https://cds.cern.ch/record/2287593

[84] L. Pacher, E. Monteil, A. Rivetti, N. Demaria, and M. D. R. Rolo, "A low-power low-noise synchronous pixel front-end chain in 65 nm CMOS technology with local fast ToT encoding and autozeroing for extreme rate and radiation at HL-LHC," in *2015 IEEE Nuclear Science Symposium and Medical Imaging Conference (NSS/MIC)*, Oct 2015, pp. 1–4.

[85] L. Gaioni, F. D. Canio, M. Manghisoni, L. Ratti, V. Re, and G. Traversi, "65 nm CMOS analog front-end for pixel detectors at the HL-LHC," *Journal of Instrumentation*, vol. 11, no. 02, p. C02049, 2016. [Online]. Available: http://stacks.iop.org/1748-0221/11/i=02/a=C02049

[86] A. Mekkaoui, M. Garcia-Sciveres, and D. Gnani, "Results of 65 nm pixel readout chip demonstrator array," *Journal of Instrumentation*,

vol. 8, no. 01, p. C01055, 2013. [Online]. Available: http://stacks.iop.org/1748-0221/8/i=01/a=C01055

[87] RD53 collaboration (including S. Marconi), "Development of a large pixel chip demonstrator in RD53 for ATLAS and CMS pixel upgrades," in *Topical Workshop on Electronics for Particle Physics TWEPP 2017*, 2017.

[88] E. Conti, J. Christiansen, S. Marconi, and P. Placidi, "Pixel chip architecture optimization based on a simplified statistical and analytical model," *Journal of Instrumentation*, vol. 9, no. 03, p. C03011, 2014.

[89] T. Hemperek, D. Arutinov, M. Barbero, R. Beccherle, G. Darbo, S. Dube *et al.*, "Digital architecture of the new ATLAS pixel chip FE-I4," in *2009 IEEE Nuclear Science Symposium Conference Record (NSS/MIC)*, Oct 2009, pp. 791–796.

[90] E. Conti, S. Marconi, J. Christiansen, P. Placidi, and T. Hemperek, "Simulation of digital pixel readout chip architectures with the RD53 SystemVerilog-UVM verification environment using Monte Carlo physics data," *Journal of Instrumentation*, vol. 11, no. 01, p. C01069, 2016. [Online]. Available: http://stacks.iop.org/1748-0221/11/i=01/a=C01069

[91] T. Poikela, J. Plosila, T. Westerlund, J. Buytaert, M. Campbell, X. Llopart *et al.*, "Architectural modeling of pixel readout chips Velopix and Timepix3," *Journal of Instrumentation*, vol. 7, no. 01, p. C01093, 2012. [Online]. Available: http://stacks.iop.org/1748-0221/7/i=01/a=C01093

[92] M. Garcia-Sciveres, "Results of FE65-P2 Pixel Readout Test Chip for High Luminosity LHC Upgrades," in *Proceedings, 38th International Conference on High Energy Physics (ICHEP 2016)*, vol. ICHEP2016, Nov 2016. [Online]. Available: http://cds.cern.ch/record/2231609

[93] A. Paternò et al. (including S. Marconi), "A prototype of pixel readout ASIC in 65 nm CMOS technology for extreme hit rate detectors at

HL-LHC," *Journal of Instrumentation*, vol. 12, no. 02, p. C02043, 2017. [Online]. Available: http://stacks.iop.org/1748-0221/12/i=02/a=C02043

[94] A. Paternò, L. Pacher, N. Demaria, A. Rivetti, G. Dellacasa, S. Marconi *et al.*, "New development on digital architecture for efficient pixel readout ASIC at extreme hit rate for HEP detectors at HL-LHC," in *2016 IEEE Nuclear Science Symposium, Medical Imaging Conference and Room-Temperature Semiconductor Detector Workshop (NSS/MIC/RTSD)*, Oct 2016, pp. 1–5.

[95] B. C. Paul, A. Agarwal, and K. Roy, "Low-power design techniques for scaled technologies," *Integration, the VLSI Journal*, vol. 39, no. 2, pp. 64 – 89, 2006, low-power design techniques. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167926005000477

[96] D. Ta, T. Stockmanns, F. Hï£¡gging, P. Fischer, J. Grosse-Knetter, ï. Runolfsson *et al.*, "Serial powering: Proof of principle demonstration of a scheme for the operation of a large pixel detector at the LHC," *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 557, no. 2, pp. 445 – 459, 2006. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S016890020502214X

[97] L. Gonella, D. Arutinov, M. Barbero, A. Eyring, F. Hï£¡gging, M. Karagounis *et al.*, "A serial powering scheme for the ATLAS pixel detector at sLHC," *Journal of Instrumentation*, vol. 5, no. 12, p. C12002, 2010. [Online]. Available: http://stacks.iop.org/1748-0221/5/i=12/a=C12002

[98] L. Gonella, V. Filimonov, F. Hï£¡gging, T. Hemperek, J. Janssen, H. Krï£¡ger *et al.*, "Performance evaluation of a serially powered pixel detector prototype for the HL-LHC," *Journal of Instrumentation*, vol. 12, no. 03, p. P03004, 2017. [Online]. Available: http://stacks.iop.org/1748-0221/12/i=03/a=P03004

[99] M. Karagounis, D. Arutinov, M. Barbero, F. Huegging, H. Krueger, and N. Wermes, "An integrated Shunt-LDO regulator for serial powered systems," in *2009 Proceedings of ESSCIRC*, Sept 2009, pp. 276–279.

[100] B. C. Paul, A. Agarwal, and K. Roy, "Low-power design techniques for scaled technologies," *Integration, the VLSI Journal*, vol. 39, no. 2, pp. 64 – 89, 2006, low-power design techniques. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167926005000477

[101] J. M. Rabaey, *Low Power Design Essentials*, ser. Integrated Circuits and Systems. Springer, 2009. [Online]. Available: https://link.springer.com/book/10.1007%2F978-0-387-71713-5

[102] Y. H. Chen, Y. L. Tang, Y. Y. Liu, A. C. H. Wu, and T. Hwang, "A novel cache-utilization-based dynamic voltage-frequency scaling mechanism for reliability enhancements," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 25, no. 3, pp. 820–832, March 2017.

[103] L. J. Casas, D. Ceresa, S. Kulis, S. Miryala, J. Christiansen, R. Francisco *et al.*, "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

[104] J. Shinde and S. S. Salankar, "Clock gating - a power optimizing technique for VLSI circuits," in *2011 Annual IEEE India Conference*, Dec 2011, pp. 1–4.

[105] S. Marconi, S. Orfanelli, M. Karagounis, T. Hemperek, J. Christiansen, and P. Placidi, "Advanced power analysis methodology targeted to the optimization of a digital pixel readout chip design and its critical serial powering system," *Journal of Instrumentation*, vol. 12, no. 02, p. C02017, 2017. [Online]. Available: http://stacks.iop.org/1748-0221/12/i=02/a=C02017

[106] J. M. Rabaey, *Digital Integrated Circuits: A Design Perspective*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 1996.

[107] M. Bagatin and S. Gerardin, *Ionizing Radiation Effects in Electronics: From Memories to Imagers*. CRC press, 2015.

[108] L. Ding, S. Gerardin, M. Bagatin, D. Bisello, S. Mattiazzo, and A. Paccagnella, "Investigation of total ionizing dose effect and displacement damage in 65 nm CMOS transistors exposed to 3 mev protons," *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 796, pp. 104 – 107, 2015, proceedings of the 10th International Conference on Radiation Effects on Semiconductor Materials Detectors and Devices. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0168900215003290

[109] T. R. Oldham and F. B. McLean, "Total ionizing dose effects in MOS oxides and devices," *IEEE Transactions on Nuclear Science*, vol. 50, no. 3, pp. 483–499, June 2003.

[110] F. Faccio, "Radiation hardness issues in 130nm and 65nm CMOS." Presented at the Fifth Common ATLAS CMS Electronics Workshop for LHC Upgrades, CERN, Switzerland, 2016.

[111] J. R. Schwank, M. R. Shaneyfelt, D. M. Fleetwood, J. A. Felix, P. E. Dodd, P. Paillet *et al.*, "Radiation effects in MOS oxides," *IEEE Transactions on Nuclear Science*, vol. 55, no. 4, pp. 1833–1853, Aug 2008.

[112] W. Snoeys, F. Faccio, M. Burns, M. Campbell, E. Cantatore, N. Carrer *et al.*, "Layout techniques to enhance the radiation tolerance of standard CMOS technologies demonstrated on a pixel detector readout chip," *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 439, no. 2, pp. 349 – 360, 2000. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0168900299008992

[113] F. Faccio and G. Cervelli, "Radiation-induced edge effects in deep submicron CMOS transistors," *IEEE Transactions on Nuclear Science*, vol. 52, no. 6, pp. 2413–2420, Dec 2005.

[114] M. Menouni on behalf of RD53 collaboration, "Main results of radiation characterization of CMOS 65nm technology and the impact on the design of RD53A, a large scale pixel chip prototype for HL-LHC," in *Presented at the International Workshop on Vertex Detectors, VERTEX 2017*, 2017.

[115] F. Faccio, S. Michelis, D. Cornale, A. Paccagnella, and S. Gerardin, "Radiation-Induced Short Channel (RISCE) and Narrow Channel (RINCE) Effects in 65 and 130 nm MOSFETs," *IEEE Transactions on Nuclear Science*, vol. 62, no. 6, pp. 2933–2940, Dec 2015.

[116] K. Iniewski, *Radiation Effects in Semiconductors*. CRC press, 2010.

[117] S. Bonacini, P. Valerio, R. Avramidou, R. Ballabriga, F. Faccio, K. Kloukinas *et al.*, "Characterization of a commercial 65 nm CMOS technology for SLHC applications," *Journal of Instrumentation*, vol. 7, no. 01, p. P01015, 2012. [Online]. Available: http://stacks.iop.org/1748-0221/7/i=01/a=P01015

[118] L. J. Casas, D. Ceresa, S. Kulis, S. Miryala, J. Christiansen, R. Francisco *et al.*, "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

[119] J. Christiansen, "RD53 Status Report," in *Presented at 131st LHCC Meeting - Agenda OPEN Session*, 2017.

[120] L. J. Casas, D. Ceresa, S. Miryala, S. Kulis, J. Christiansen, R. Francisco *et al.*, "Study of total ionizing dose effects in 65nm digital circuits with the DRAD, Digital RADiation Test Chip," in *Presented at the RADiations Effects on Components and Systems conference (RADECS 2017)*, 2017.

[121] IBM. (2002, Jan.) IBM. SOI Technology: IBM's next advance in chip design. [Online]. Available: https://www.ibm.com

[122] H. B. Wang, Y. Q. Li, L. Chen, L. X. Li, R. Liu, S. Baeg *et al.*, "An SEU-Tolerant DICE Latch Design With Feedback Transistors," *IEEE Transactions on Nuclear Science*, vol. 62, no. 2, pp. 548–554, April 2015.

[123] V. S. Veeravalli, "Fault tolerance for arithmetic and logic unit," in *IEEE Southeastcon 2009*, March 2009, pp. 329–334.

[124] M. Lemarenko, T. Hemperek, H. Krouger, M. Koch, F. Ltticke, C. Marinas *et al.*, "Test results of the data handling processor for the DEPFET pixel vertex detector," *Journal of Instrumentation*, vol. 8, no. 01, p. C01032, 2013. [Online]. Available: http://stacks.iop.org/1748-0221/8/i=01/a=C01032

[125] C. Lopez-Ongil, L. Entrena, M. Garcia-Valderas, M. Portela, M. A. Aguirre, J. Tombs *et al.*, "A unified environment for fault injection at any design level based on emulation," *IEEE Transactions on Nuclear Science*, vol. 54, no. 4, pp. 946–950, Aug 2007.

[126] S. Miryala, "SEE tolerant standard cell based design while guaranteeing specific distance between memory elements," in *Topical Workshop on Electronics for Particle Physics TWEPP 2017*, 2017.

[127] E. G. Friedman, "Clock distribution networks in synchronous digital integrated circuits," *Proceedings of the IEEE*, vol. 89, no. 5, pp. 665–692, May 2001.

[128] T. S. Poikela, J. Plosila, T. Westerlund, and K. Wyllie, "Readout Architecture for Hybrid Pixel Readout Chips," Apr 2015, presented 15 Jun 2015. [Online]. Available: https://cds.cern.ch/record/2042198

[129] L. Gaioni, F. D. Canio, M. Manghisoni, L. Ratti, V. Re, G. Traversi *et al.*, "Low-power clock distribution circuits for the Macro Pixel ASIC," *Journal of Instrumentation*, vol. 10, no. 01, p. C01051, 2015. [Online]. Available: http://stacks.iop.org/1748-0221/10/i=01/a=C01051

[130] M. Menouni, D. Arutinov, M. Backhaus, M. Barbero, R. Beccherle, P. Breugnon *et al.*, "Seu tolerant memory design for the atlas pixel readout chip," *Journal of Instrumentation*, vol. 8, no. 02, p. C02026, 2013. [Online]. Available: http://stacks.iop.org/1748-0221/8/i=02/a=C02026

[131] S. Marconi, E. Conti, P. Placidi, A. Scorzoni, J. Christiansen, and T. Hemperek, *A SystemVerilog-UVM Methodology for the Design, Simulation and Verification of Complex Readout Chips in High Energy Physics Applications*. Cham: Springer International Publishing, 2017, pp. 35–41. [Online]. Available: https://doi.org/10.1007/978-3-319-47913-2_5

[132] S. Marconi, S. Orfanelli, M. Karagounis, T. Hemperek, J. Christiansen, and P. Placidi, "Advanced power analysis methodology targeted to the optimization of a digital pixel readout chip design and its critical serial powering system," *Journal of Instrumentation*, vol. 12, no. 02, p. C02017, 2017. [Online]. Available: http://stacks.iop.org/1748-0221/12/i=02/a=C02017

[133] S. Marconi, T. Hemperek, P. Placidi, A. Scorzoni, E. Conti, and J. Christiansen, "Low-power optimisation of a pixel array architecture for next generation High Energy Physics detectors," in *2017 13th Conference on Ph.D. Research in Microelectronics and Electronics (PRIME)*, June 2017, pp. 201–204.

[134] S. Orfanelli et al. (including S. Marconi), "Serial powering optimization for CMS and ATLAS pixel detectors within RD53 collaboration for HL-LHC: system level simulations and testing," in *Topical Workshop on Electronics for Particle Physics TWEPP 2017*, 2017.

[135] N. Demaria et al. (including S. Marconi), "Recent progress of RD53 Collaboration towards next generation Pixel Read-Out Chip for HL-LHC," *Journal of Instrumentation*, vol. 11, no. 12, p. C12058, 2016. [Online]. Available: http://stacks.iop.org/1748-0221/11/i=12/a=C12058

[136] L. Gaioni et al. (including S. Marconi), "Design of analog front-ends for the RD53 demonstrator chip," *PoS*, vol. Vertex 2016, 2017.

[137] E. Monteil et al. (including S. Marconi), "A prototype of a new generation readout ASIC in 65nm CMOS for pixel detectors at HL-LHC, journal=Journal of Instrumentation," vol. 11, no. 12, p. C12044, 2016. [Online]. Available: http://stacks.iop.org/1748-0221/11/i=12/a=C12044

[138] S. Panati et al. (including S. Marconi), "First measurements of a prototype of a new generation pixel readout ASIC in 65 nm CMOS for extreme rate HEP detectors at HL-LHC," in *2016 IEEE Nuclear Science Symposium, Medical Imaging Conference and Room-Temperature Semiconductor Detector Workshop (NSS/MIC/RTSD)*, Oct 2016, pp. 1–7.

[139] L. Pacher et al. (including S. Marconi), "A Prototype of a new generation readout ASIC in 65 nm CMOS for pixel detectors at HL-LHC," in *International Workshop on Vertex Detectors, VERTEX 2016*, 2017. [Online]. Available: https://pos.sissa.it/287/054/pdf

[140] A. Paternò et al. (including S. Marconi), "Results from CHIPIX-FE0, a small scale prototype of a new generation pixel readout ASIC in 65nm CMOS for HL-LHC," in *Topical Workshop on Electronics for Particle Physics TWEPP 2017*, 2017.