POLITECNICO DI TORINO

DOCTORAL THESIS

in Electronic Devices - XXVIII cycle -

DISAT Department

Electronic systems for intelligent particle tracking in the High Energy Physics field

DAVIDE CERESA

Supervisors:
Prof. Fabrizio Pirri
Prof. Felice Iazzi

Ph.D coordinator:
Prof. Giovanni Ghione

March 2016

POLITECNICO DI TORINO

Abstract

DISAT department

Ph.D in Electronic Device

Electronic systems for intelligent particle tracking in the High Energy Physics field

by Davide Ceresa

English version

This Ph.D thesis describes the development of a novel readout ASIC for hybrid pixel detector with intelligent particle tracking capabilities in High Energy Physics (HEP) application, called Macro Pixel ASIC (MPA). The concept of intelligent tracking is introduced for the upgrade of the particle tracking system of the Compact Muon Solenoid (CMS) experiment of the Large Hadron Collider (LHC) at CERN: this detector must be capable of selecting at front-end level the interesting particle and of providing them continuously to the back-end. This new functionality is required to cope with the improved performances of the LHC when, in about ten years' time, a major upgrade will lead to the High Luminosity scenario (HL-LHC).

The high complexity of the digital logic for particle selection and the very low power requirement of $< 100\,\mathrm{mW/cm^2}$ drive the choice of a 65 nm CMOS technology. The harsh environment, characterized by a high ionizing radiation dose of 100 Mrad and low temperature around -30°C, requires additional studies and technology characterization. Several architecture for intelligent particle tracking has been studied and evaluated with physics events from Monte-Carlo simulation. The chosen one reaches an efficiency $> 95\%$ in particle selection and a data reduction from $\sim 200\,\mathrm{Tb/s/cm^2}$ to $\sim 1\ \mathrm{Tb/s/cm^2}$.

A prototype, called MPA-Light, has been designed, produced and tested. According to the measurements, the prototype respects all the specifications. The same device has been used for multi-chip assembly with a pixelated sensor. The assembly characterization with radioactive sources confirms the result obtained on the bare chip.

Versione italiana

La tesi di seguito riportata descrive lo sviluppo di un nuovo ASIC studiato per la lettura di rivelatori ibridi constituiti da pixel con un sistema di tracciamento intelligente di particelle per applicazioni nel campo della fisica delle particelle, chiamato Macro Pixel ASIC (MPA). Il concetto di tracciamento intelligente è stato introdotto per l'aggiornamento dell'esperiemento Compact Muon Solenoid (CMS), uno dei due grandi rivelatori per la fisica delle particelle generale, costruito sull'acceleratore Large Hadron Collider (LHC) al CERN. Il sistema di tracciamento deve essere in grado di selezionare a livello front-end le particelle interessanti, e di fornirle ininterrottamente al back-end. Questa nuova funzionalità è stata richiesta per far fronte al miglioramento delle prestazioni del acceleratore quando, in circa dieci anni, un vasto aggiornamento condurrà allo scenario chiamato High Luminosity (HL-LHC).

La complessità della logica digitale necessaria per la selezione di particelle e la richiesta di una densità di potenza totale $< 100\,\mathrm{mW/cm^2}$ conducono alla scelta della tecnologia CMOS a 65 nm. L'ambiente ostile, caratterizzato da un'alta dose di radiazioni ionizzanti fino a 100 Mrad e le basse temperature intorno ai -30°C, richiedono studi aggiuntivi e la caratterizzazioni della tecnologia. Differenti architetture per il tracciamento intelligente di particelle sono state studiate e valutate con eventi fisici ottenuti grazie a simulazioni Monte-Carlo. L'architettura scelta permette di raggiungere un'efficienza $> 95\%$ nella selezione di particelle e una riduzione di dati da $\sim 200\,\mathrm{Tb/s/cm^2}$ a $\sim 1\ \mathrm{Tb/s/cm^2}$.

Un prototipo, chiamato MPA-Light, è stato disegnato, prodotto e testato. Sulla base delle misure effettuate, il prototipo rispetta tutte le specifiche. Lo stesso dispositivo è stato usato per l'assemblaggio di moduli multichip con un sensore pixellato. L'illuminazione del modulo con fonti radioattive ha inoltre confermato i risultati ottenuti sul chip senza il sensore connesso.

Dedicated to Stefania,

and to Our Dreams.

Contents

List of Figures

Abbreviations

MOS	**M**etal **O**xide **S**emiconductor
CMOS	**C**omplementary **M**etal **O**xide **S**emiconductor
PMOS	**P**-type **M**etal **O**xide **S**emiconductor
NMOS	**N**-type **M**etal **O**xide **S**emiconductor
LHC	**L**arge **H**adron **C**ollider
CMS	**C**ompact **M**uon **S**olenoid
MPA	**M**acro **P**ixel **A**SIC
MIP	**M**inimum **I**onizing **P**article
HPD	**H**ybrid **P**ixel **D**etector
FE	**F**ront **E**nd
CSA	**C**harge **S**ensitive **A**mplifier
TOT	**T**ime **O**ver **T**hreshold
DAQ	**D**ata **A**c**Q**uisition
RTL	**R**egister **T**ransfer **L**evel
VCD	**V**alue **C**hange **D**ump
TID	**T**otal **I**onizing **D**ose
SEE	**S**ingle **E**vent **E**ffect
MOS	**M**etal **O**xide **S**emiconductor
STI	**S**hallow **T**rench **I**solation
LDD	**L**ight **D**oped **D**rain

SEU	**S**ingle **E**vent **U**pset
LET	**L**inear **E**nergy **T**ransfer
ELT	**E**nclosed **L**ayout **T**ransistor
TMR	**T**riple **M**odular **R**edundancy
ECC	**E**rror **C**orrection **C**ode
SST	**S**ilicon **S**trip **T**racker
L1	**L**evel **1**
IP	**I**nteraction **P**oint
LHC	**L**arge **H**adron **C**ollider
HLT	**H**igh **L**evel **T**rigger
BX	**B**unch **X**crossing
MaPSA	**Ma**cro **P**ixel **S**ub **A**ssembly
CIC	**C**oncentrator **IC**
LV	**L**ow **V**oltage powering
HV	**H**igh **V**oltage biasing
DTC	**D**ata **T**rigger **C**ontrol
SSA	**S**hort **S**trip **A**SIC
PS	**P**ower **S**upply
FEA	**F**inite **E**lement **A**nalysis
STI	**S**hallow **T**rench **I**solation
DACs	**D**igital to **A**nalog **C**onverters
DLL	**D**elay **L**ocked **L**oops
PSP	**P**ower **S**kew **P**roduct
ToA	**T**ime **o**f **A**rrival
BX-ID	**B**unch **X**crossing **I**dentification **D**ata
SNM	**S**ignal to **N**oise **M**argin
SET	**S**inge **E**vent **T**ransient
ESD	**E**lectro**S**tatic **D**ischarge
SLVS	**S**calable **L**ow **V**oltage **S**ignaling
FIFO	**F**irst **I**n **F**irst **O**ut
RTL	**R**egister **T**ransfer **L**evel
MC	**M**onte **C**arlo
FPGA	**F**ield **P**rogrammable **G**ate **A**rray

CML	**C**urrent **M**ode **L**ogic
LSB	**L**east **S**ignificant **B**it
ENC	**E**quivalent **N**oise **C**harge
UBM	**U**nder **B**ump **M**etalization
MPW	**M**ulti **P**roject **W**afer

Chapter No. 1

Introduction

In 1910 Wilson built his masterpiece and realised his dream: a cloud chamber that could visualise particle tracks – resembling something like the vapour trails left in the wake of an airplane. The next year he took his first photographs of tracks, exclaiming excitedly:

"they are as fine as little hairs"

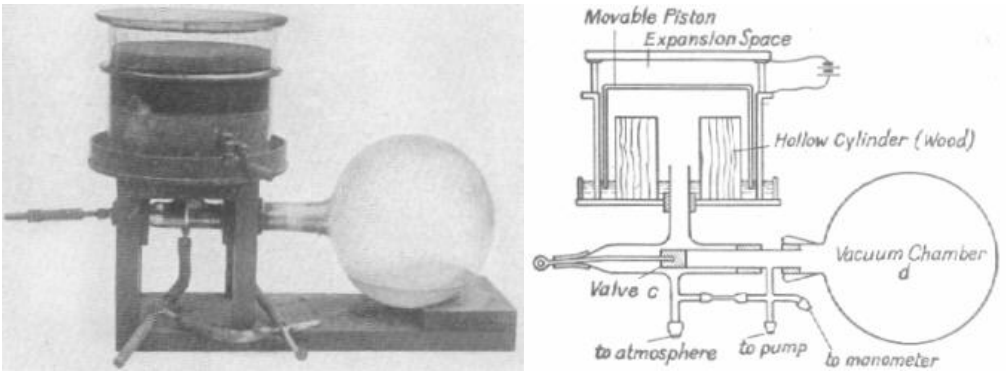
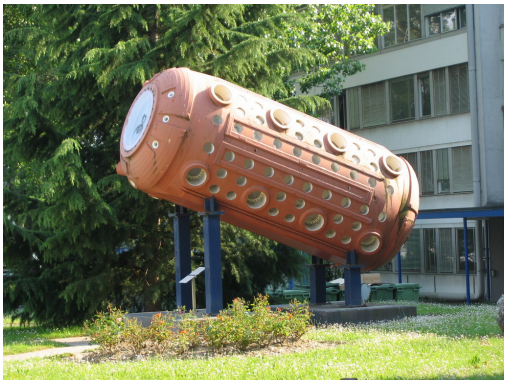


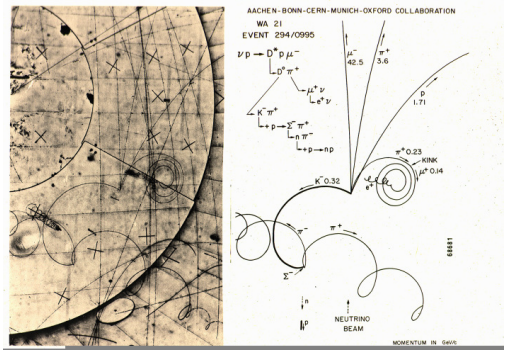
FIGURE 1.1: Wilson's 1910 Cloud Chamber. The diameter of the chamber is 16.5 cm, depth 3 cm. The movable piston is suddenly lowered by opening the valve c and so connecting the vacuum chamber d with the part of the apparatus beneath the piston

For the first time, physicists could see the activity of the subatomic world. Wilson's cloud chamber (in figure 1.1) allowed visualization of the particle tracks coming from the natural radioactivity and from the cosmic rays. The particle could be easily identified by a visual analysis of the tracks left in real time. During this exciting time of exploring the unknown domain of the reality, many new particles like for example muon, pion and positron have been discovered by analyzing the "pictures" recorded from the cloud chambers. Wilson received the Nobel Prize in Physics in 1927 for his invention.

In 1973 , a modified version of the cloud chamber, a bubble chamber called "Gargamelle" (see Figure 1.2(a)) allowed one of the greatest discoveries at European Organization for Nuclear Research (CERN): the neutral current. The bubble chamber is filled with a superheated liquid and contains a piston which allows to make a fast change of the pressure inside the chamber, thus bringing the liquid into a superheated state. As the particles traverse the superheated medium, they vaporize the liquid along their paths. These trails of bubbles were photographed and the films were analyzed manually (see Figure 1.2(b))



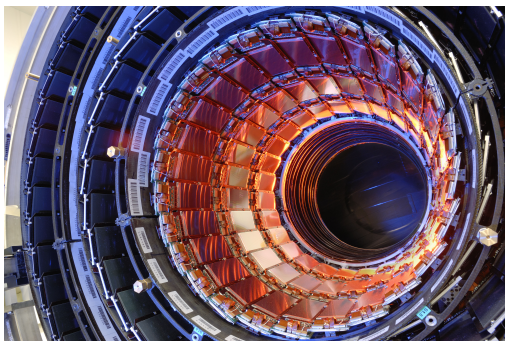
(A) The Gargamelle bubble chamber.



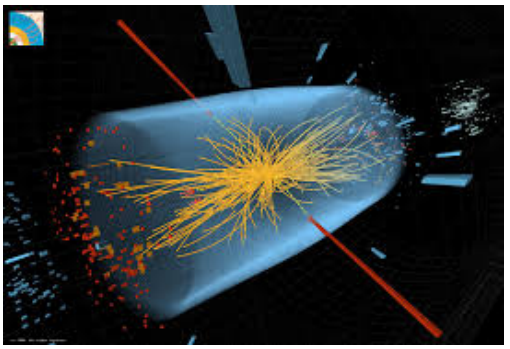
(B) Track reconstruction and analysis.

FIGURE 1.2

Towards the end of the 1970s, also semiconductor detectors began to be explored for their tracking capabilities. The spatial resolution in the 5-10 μm range, the introduction of planar technology and the possibility to develop a dedicated readout for integration of detector and electronics provided the definitive boost to the use of this technology. Nowadays, all particle physics spectrometers have inbuilt vertex detectors, which deliver excellent results. The application of silicon tracking detectors has expanded to nuclear physics, solid-state physics, astrophysics, biology and medicine.



(A) The CMS Silicon Strip Tracker.



(B) CMS event from Run 1.

FIGURE 1.3

The four experiments on the Large Hadron Collider (LHC) at CERN use silicon detectors in their central vertexing/tracking systems. The largest silicon detector is the Silicon Strip Tracker (see Figure 1.3(a)) of the Compact Muon Solenoid (CMS) experiment: about $210\,\mathrm{m}^2$ of active silicon subdivided into many thousands of detector modules. The readout system is based on radiation hard electronics, fabricated in submicron commercial processes. It samples, amplifies, buffers and processes signals from the silicon detector. Samples are stored for a latency of a few μs, and are transmitted if a trigger is received.

As explained in the paper from the CMS collaboration [1], the CMS silicon strip tracker worked throughout LHC Run 1 from 2009 to 2013 (an event is shown in figure 1.3(b)), providing fundamental data to the discovery of a particle at about 125 GeV "consistent with the Higgs boson". Currently, the tracker is acquiring data during LHC Run 2 with a center of mass energy of 13 TeV.

The path of evolution of the LHC is now oriented towards the High Luminosity (HL-LHC) scenario, described in the preliminary design report [2], where event reconstruction will become impossible with the current detectors due to the increased luminosity and pile-up. A major upgrade is foreseen after 2020 which comprises the complete substitution of the CMS tracking system. This upgrade will introduce the concept of "intelligent" particle tracking which requires a front-end electronics capable of selecting the interesting physics events. Thanks to this capability the detector will provide selected information to the experiment back-end for every collision event, making possible the event reconstruction in the High Luminosity environment.

This work is dedicated to the development of the electronic system for intelligent particle tracking. In particular, it is focused on the desing of a dedicated readout electronics for hybrid pixel detector with particle discrimination capabilities, called Macro Pixel ASIC (MPA). This represents one of the key components and one of the main engineering challenges in the construction of the new CMS tracking system.

1.1 Main challenge

The capability of reconstructing physics event in the High Luminosity scenario together with the requirement of reducing the material makes the design of this new particle tracking system an hard engineering challenge. On the one hand, introducing the "intelligence" to select interesting information at front-end level requires very complex and power angry readout electronics. On the other hand, minimizing the amount of interaction with the incoming particle makes necessary the reduction of cooling material and consequently of power consumption.

Moreover, in order to keep the occupancy at a few percent and allow the track reconstruction at higher luminosity, the objective is to design a particle tracking system with an higher granularity (x5 respect current one) introducing pixelated sensors. Since the amount of material in the tracker must be sensitively lower than the current system, the use of pixel detector in some region requires an important effort in power and data readout optimization.

In particular, bandwidth and power limitations require a front-end module able to reduce "on-line" the amount of data to be transmitted to the back-end. The data selection is made by a novel concept of silicon detector modules, the so-called p_T-modules. The main novelty is the capability of tracking charged particles with a transverse momentum (p_T) higher than a certain threshold (nominally 2 GeV/c) at every Bunch Crossing (BX) i.e every 25 ns. This value is provided by the intrinsic collision frequency of the collider. Thanks to the technology improvements, the front-end electronics can include complex digital circuits providing the "intelligence" to select and transmit the interesting events.

Furthermore, the requirement for radiation hardness increases of $\sim$ one order of magnitude, and requires a full characterization of the technologies used as well as radiation hardening technique to limit the performance degradation. All these features cost in

term of power, and require the introduction of power reduction technique as clock and memory gating, low power supply, and dedicated design for many components in the system.

In conclusion, this Ph.D thesis describes the development of an electronic device with particle discrimination capabilities for the readout of hybrid pixel detector. The main challenge is represented by the very low power density of < 100 mW/cm^2 available for the complex data processing required. Further studies address the problem related with high radiation level up to 100 Mrad and low temperature operation around -30°C.

1.2 Thesis organization

This document is organized as follow:

- The second chapter (the first is this introduction) provides a theoretical background on the silicon detectors for High Energy Physics application with emphasis on the pixel electronics and readout system;
- The third chapter reports about the various aspects of the particle tracking system of the CMS experiment for the High Luminosity LHC. It provides an overview of the module architecture, power distribution, electronics components and readout system;
- The fourth chapter describes the author's work on the development of an electronic device with particle discrimination capabilities for the readout of hybrid pixel detector;
- The fifth chapter provides the description of the first prototype produced with a 65 nm CMOS technology. Electrical characterization as well as radiation test and module assembly results are reported.

Chapter No. 2

Silicon Detectors for High Energy Physics

This chapter describes the hybrid pixel detectors, which is the technology used for the electronic device developed in this thesis. The principles about silicon detector, readout electronics and their application in the silicon tracking systems for High Energy Physics experiments are introduced. The effect of radiation on the electronics and the technique for radiation hardness conclude the chapter.

2.1 Silicon for particle detection

Silicon presents several features which make its use favourable for particle detection [3]. The small energy band gap (1.12 eV at room temperature) produces a large number of charge carriers per unit of energy loss by the ionizing particles to be detected. Besides, the high material density (2.33 g/cm^3) leads to a large energy loss per traversed length of the ionizing particle (3.8 MeV/cm for a minimum ionizing particle). It is then possible to build thin detectors that still produce large enough signals to be measured. The mobility of electrons and holes is high at room temperature, and only moderately influenced by doping. The charge can thus be rapidly collected, with collection times in the order of ns, and detectors can be used in high-rate environments. On the other hand, the silicon band gap is large enough to have a sufficiently low leakage current due to electron-hole pair generation.

As far as detector fabrication is concerned, the major advantage of silicon resides in the availability of a developed technology, which also allows the integration of detector and electronics on the same substrate. Moreover, its excellent mechanical rigidity allows the construction of self-supporting structures. An overview of silicon detector and of other particle detectors types can be found in Grupen and Swartz [4].

2.2 Detector structure

The basic element of silicon detectors, shown in figure 2.1, is a highly doped area of silicon on a resistive substrate of the opposite polarity acting as diode, which is then reverse-biased. Usually, the applied voltage is above the full depletion in order to use the whole volume for charge collection. Aluminium contacts connects the doped area to the readout electronics, while the back-side metallization provides the bias voltage.

If an ionizing particle is traversing the detector sensitive volume, electron-hole pairs are created along its path. The electric field in the depleted volume separates electrons and holes, which drift to the positive, respectively negative electrode inducing a current in the readout circuit. This current can be amplified and integrated by a charge sensitive amplifier resulting in an output voltage which is proportional to the collected charge.

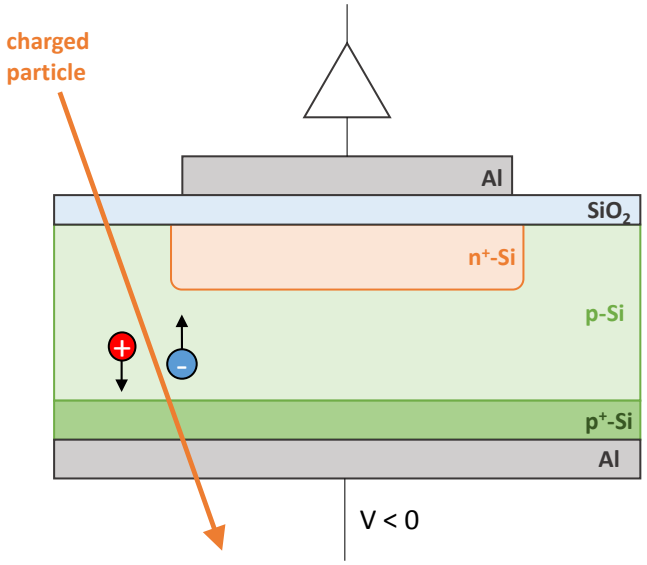


FIGURE 2.1: p^+-on-n silicon detector structure.

If the particle is stopped inside the detector, the measured charge is proportional to the energy of the particle, otherwise the particle will traverse the detector and the measured signal will be proportional to the energy loss of the particle. The energy loss is due to Coulomb interaction, Bremsstrahlung and scattering with the electrons and the core of the silicon atoms. The mean energy needed for the creation of an electron-hole pair in silicon is of 3.6 eV. In a silicon detector with a thickness of 200 μm the most probable value of electron-hole pairs generated by a Minimum Ionizing Particle (MIP) is of 16000.

2.2.1 Microstrip detector

Microstrip detectors are obtained by segmenting the doped side into strips over the full length of the detector. The strips are usually from few tens to few hundreds of μm apart, with the detector position resolution increasing with the decreasing strip pitch. The segmented side is usually covered by a few μm layers of SiO_2 or Si_3N_4, which protect the wafer during fabrication but also the detector itself. The Aluminium contacts can be placed directly on the doped strips (DC coupled detectors) or on a thin oxide or nitride layer, in which case the doped strips are capacitively connected to the readout electronics (AC coupled detectors). The latter solution is more expensive, due to the additional steps needed in the production, however capacitive coupling prevents leakage currents to flow through the electronics. The electrical connection between the strips and the readout electronics is usually realized via thin wires (wire-bonding).

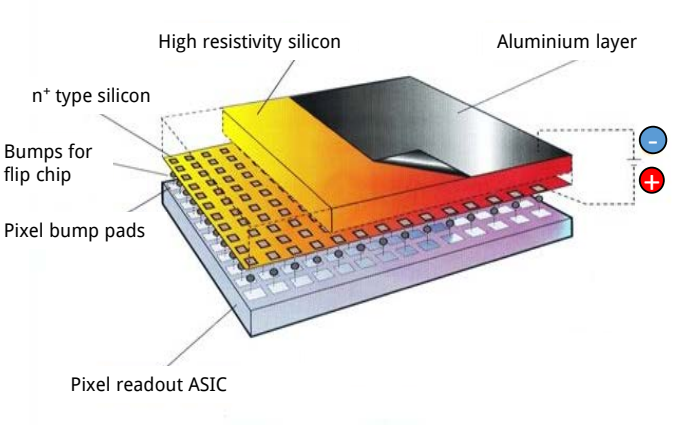


FIGURE 2.2: p^+-on-n silicon sensor bump-bonded with the pixel readout ASIC.

2.2.2 Hybrid pixel detector

The planar process allows also the segmentation of one of the detector sides into a two-dimensional array of pixels. In this case an unambiguous 2-dimensional information about the position of the hit is achieved. The lateral size of pixels usually ranges between a few tens of μm and a few mm. The number of pixels and by that the number of readout channels increases linearly with the active area of the detector, while for silicon strip detectors the number of readout channels increases with the square root of the active detector area. A higher cost of pixel detectors results from the complexity of the readout electronics and of the mounting techniques, especially when the pixel dimensions are small. The use of pixel detectors is nevertheless inevitable in environments in which the detector occupancy is high, i.e. the sensor is traversed by many close-by particles. The use of strip detectors is in this case impossible due to ambiguities in the determination of the hit positions.

Several categories of pixel detector are available and they can be divided according to the technology used for charge collection. As mentioned in Rossi et Al. [5], Hybrid Pixel Detector (HPD) is the most used technology for HEP application. It uses high-resistivity silicon substrates like in the case of microstrip detectors. The sensor is divided in pixels with the same pitch as the readout chip and the two are connected using flip-chip technology. This technique, also known as controlled collapse chip connection or its acronym, C4, is a method for interconnecting chips to external circuitry with solder bumps that have been deposited onto the chip pads.

2.3 Readout ASIC

In HPD, since the two parts are produced separately, they can be optimized and designed independently from each other. Also, any standard CMOS technology can be used to design the readout electronics, so the advances in the lithographic process can be exploited to build more advanced systems, with smaller features and/or more features. The main disadvantage of this architecture is the cost of the flip chip process, especially for detectors with very small pixels.

2.3.1 Front-End Electronics

Silicon sensors provide a typical signal in the range of tens of thousands of electrons within the collection time of a few nanoseconds. Front-end (FE) electronics process the signal from the sensor. Signal processing starts with the conversion of the signal charge into voltage which is performed by a Charge Sensitive Amplifier (CSA). This voltage is then discriminated and digitized. The resulting data are then coded into a comprehensive data format so that information about pixel address, time and amplitude can be disentangled from the data pattern. The common circuit blocks, shown in figure 2.3, are:

The Charge Sensitive Amplifier is an electronic circuit converting the electric charge Q to the voltage V_{out}. The core of the CSA is a voltage amplifier (core amplifier) providing high open loop gain A. The voltage amplifier has a capacitive feedback C_f .

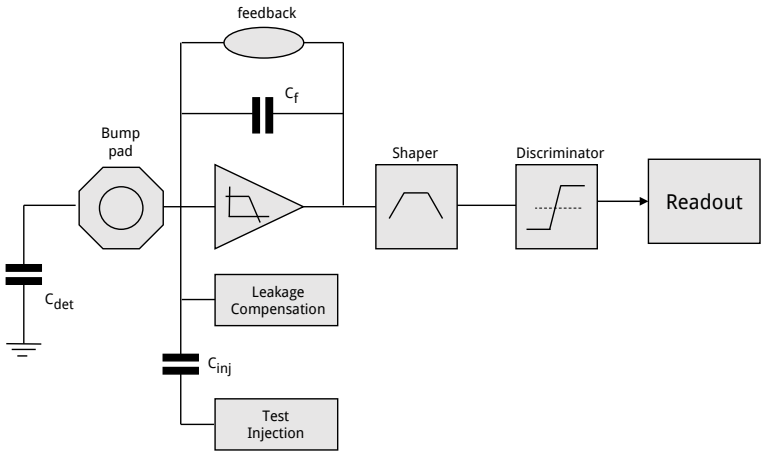


FIGURE 2.3: Component of a generic front-end circuit.[5]

In the ideal case, the CSA behaves as an integrator with a closed loop gain inversely proportional to the feedback capacitance. In the real life, the finite open loop gain A of the core amplifier and the capacitance of the sensor C_d affects the gain:

$$V_{out} = \frac{Q}{C_f} \qquad g_{ideal} = \frac{1}{C_f} \qquad g = g_{ideal} \cdot \frac{1}{1 + \frac{1}{A} + \frac{C_d}{C_f A}} \tag{2.1}$$

The Feedback Circuit is required to define the DC-operation point of the charge sensitive preamplifier and to remove signal charges from the input node (or from C_f after the dynamic response of the amplifier) so that the preamplifier output voltage returns to its initial value.

The Leakage Compensation Circuit is implemented to sink all or a significant fraction of the leakage current. The sensor pixels are usually DC-coupled to the preamplifier inputs (this eliminates the need for biasing structures and AC-coupling capacitors on every sensor pixel), thus a leakage current I_{leak} of up to a few hundreds of nA after irradiation must be sunk/sourced by the pixel circuit.

The Shaper is an electronic filter, usually a band-pass, defining the bandwidth of the signal output of the CSA. The shaper helps to suppress the low and high frequency noise component and therefore improves the signal to noise ratio of the pixel detector.

The Discriminator compares the voltage at the output of the shaper (or CSA) with a reference value and sets its output either to HIGH level or LOW level depending on whether the input voltage is above or below the detection threshold, thus providing digitization with a single bit resolution. If the discriminator processes a signal from the CSA with a constant current source feedback, the length of the pulse at the output of the discriminator is directly proportional to the signal charge. The length of the pulse can be used for amplitude measurement called Time Over Threshold (TOT).

Test Charge Injection allows to verify the correct operation of the Front-End electronics. The controlled injection of known charges into the preamplifier is accomplished by applying a known voltage step to a well-defined calibration capacitor C_{inj}.

2.3.2 Readout architecture

The digital hit signals of the discriminator must be further processed by circuitry in the pixel and at the chip periphery. The architecture of this readout depends strongly on the target application. In some application like medical device, the number of hits during a given time interval in every pixel can be sufficient. This requires simple counters in the pixels and a mechanism to transfer the counter values to I/Os. More detailed information is required in HEP applications. The positions, often also the times and possibly the corresponding pulse amplitudes, of all hits belonging to an interaction must be provided. This requires a timing precision of 25 ns (the bunch crossing interval) for the detectors at LHC. Concerning the readout of events, it can be started immediately after the interaction. Very often, however, a trigger system selects only a fraction of the events for readout in order to reduce the data volume sent to the Data Acquisition (DAQ). Since the trigger signal arrives with a significant delay, all hits must be identified and buffered for some time. At the LHC experiments, the current trigger latency is in the order of 2–3 μs which corresponds to $\sim$100 interactions. Almost all architectures perform an immediate zero suppression (i.e. process only pixels with amplitudes above a threshold) to reduce the size of the required buffers. Anyway, the limited buffer space available can lead to a loss of hits in high-rate events.

The choice of a suited architecture mainly depends on the available chip technology, on the required information and on the acceptable hit losses which can have very different characteristics for different readout concepts. Detailed simulations of the hit losses are therefore required before a choice can be made. Important parameters for analyzing the performance of a readout ASIC are the occupancy, the hit rate and the efficiency.

Occupancy is an indication of the utilization of the front-ends or of the data traffic inside a chip within a certain period. It indicates which fraction of its resources a readout chip is using at a given moment and is defined as:

$$O = \frac{N_{pixel_{hit}}}{TN_{pixels_{chip}}} \cdot 100\,\% \tag{2.2}$$

where O is the occupancy, $N_{pixel_{hit}}$ the number of pixels containing hit information, $N_{pixels_{chip}}$ the number of pixels in a full chip and T is a period. In many applications, occupancy is a time-averaged quantity and the period is not explicitly mentioned. In the case of HEP experiments on LHC, the occupancy is expressed as occupancy per Bunch Crossing.

Hit rate indicates the flux of particle on a certain area, which in the case of a readout ASIC is the active area of the sensor:

$$R = \frac{N_{pixel_{hit}}}{TA_{chip}} \tag{2.3}$$

where R_{hit} is the hit rate, T the acquisition time and A_{chip} area of a chip. Typically, hit rate is expressed as $\frac{hits}{s \cdot cm^2}$.

Efficiency. Occupancy and hit rate provide the quantity of data to be processed by a digital readout architecture. Instead, the efficiency of the digital readout architecture gives the ratio of the correct work performed by the architecture and is defined as:

$$E = \frac{N_{output}}{N_{input}} \tag{2.4}$$

where E is the readout efficiency, N_{output} the number of correct hits received at output of the chip and N_{input} the number of hits received from the output of the front-end stage.

2.3.3 Power estimation technique

Another important parameter to estimate the performance of a readout electronics is its power consumption. Consequently, the contributions from logic circuits, interconnections, clock distribution, on chip memories and I/Os must be estimated during the design. Yet, power consumption of a system cannot be solely determined from high-level models, but the models can be used as tools to estimate the power consumption of the full architecture.

Once an architecture with sufficient performance in terms of efficiency and latency has been found, a pixel region or even a pixel block can be designed at Register Transfer Level (RTL). This block can then be synthesized and a prototype of the physical design completed. A back-annotated netlist of this prototype can be used in simulation to obtain toggling rates for all nets in the design in the form of Value Change Dump (VCD) information. Again back-annotating this information into a physical design tool, an accurate estimate for power consumption of this block is obtained.

In a large design, the modelling of the interconnections becomes an important contribution to the total power consumption. Three wire categories are defined: local wires connect gates, intermediate wires connect subsystem and global wires, which includes data, control and address buses, connect different regions of the design. The wire width is assumed to be minimum, excluding the clock distribution, as the wire RC constant does not change with wire width. Knowing the data activity and the bus dimensions, the power contribution is calculated as:

$$P_{dynamic} = \alpha C_{total} {V_{DD}}^2 f \tag{2.5}$$

where α is the switching factor, C_{total} the total capacitance, V_{DD} the power supply and f the operating frequency. The total capacitance includes also the repeaters input capacitances and the wire parasitics.

The I/O power is consumed in two parts. One is the power used to drive off-chip capacitance, bonding wires and the pad capacitance. The other is the power consumed by the driver itself. The value strongly depends on the architecture chosen for the driver: CMOS driver power depends on the activity and on the load capacitance, while differential driver uses a contant current. In the latter case, the power it does not change significantly with activity and load capacitance.

The estimation of on-chip memory power requires the characterization of the cell. The latter provides models which give the consumption for write and read operations, and consequently, the power consumption can be estimated knowing operating frequency and switching factors.

2.4 Partilce tracking system in HEP experiments

Silicon detectors are largely used in HEP experiments. This is also the case of the LHC which is the world's largest and most powerful particle collider, the largest, most complex experimental facility ever built, and the largest single machine in the world. It was built by CERN between 1998 and 2008 in collaboration with over 10,000 scientists and engineers from over 100 countries, as well as hundreds of universities and laboratories. As shown in figure 2.4, it lies in a tunnel 27 kilometres in circumference, as deep as 175 metres beneath the France–Switzerland border near Geneva, Switzerland.

The LHC accelerator is designed to collide protons at a centre of mass energy of 14 TeV and luminosity of $10^{34}\text{cm}^{-2}\text{s}^{-1}$. Bunches of 10^{11} protons collides every 25 ns (Bunch Crossing period). A detailed description of LHC can be found in its design report [6]. At the collision point, where about one billion proton-proton collisions are produced every seconds, are installed four large experiment. ATLAS [7] and CMS [8] are general purpose experiments, while ALICE [9] is dedicated to the study of heavy ion collision

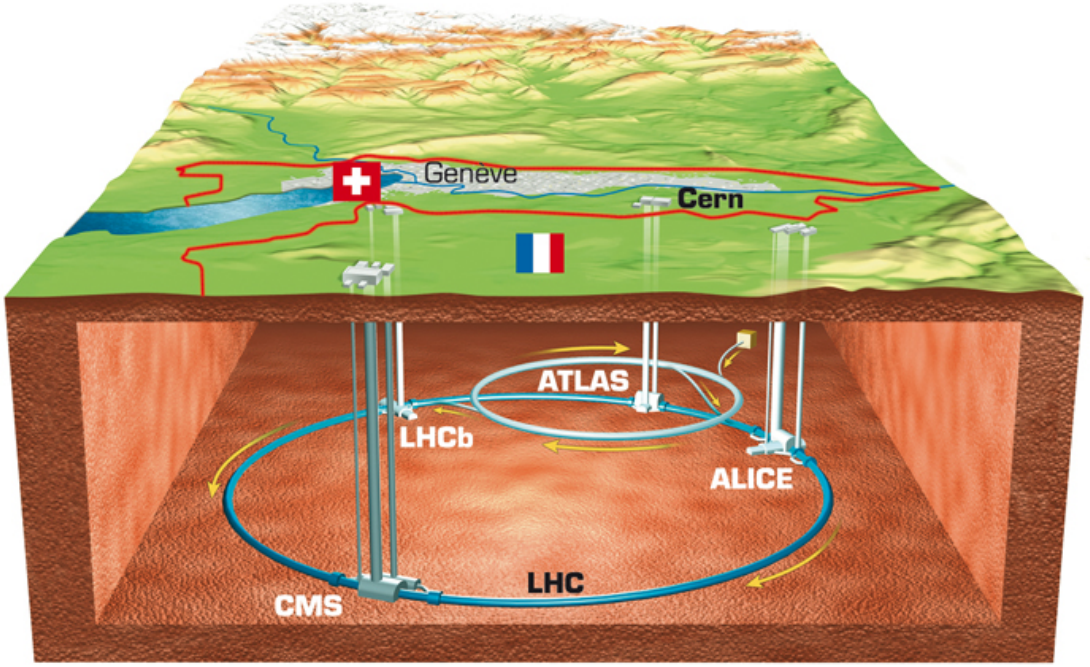


FIGURE 2.4: Section of the LHC accelerator with the four experiment.

and LHCb [10] to the study of the b-physics. These experiments measure parameters of secondary particles produced in the collisions of the high energetic primary particles. In order to precisely measure trajectories and origins (vertices) of these secondary particles, the tracking detectors are situated very close to the collision point. Tracking and vertexing detectors are finely segmented silicon pixel or strip detector systems with short recovery time and high data throughput.

The tracking system, called tracker, is an essential component of a large HEP experiment. It performs a precise measurement of particle trajectories. Electrically charged particles are detected in sensitive layers of the detector.

The coordinates system used throughout this thesis is based on r, z, ϕ, η where:

- r is the radial distance from the nominal beam line;
- z axis coincides with the nominal beam line;
- ϕ is azimuthal angle and is measured in the plane perpendicular to the beam line;
- η, called pseudorapidity, is the angle of a particle relative to the beam axis.

The sensitive layers of a tracker are usually arranged in barrel layers, coaxial with respect to z axis, and several end-cap disks to provide a broad angular coverage. Trackers operate in a strong magnetic oriented along z axis. The magnetic field curves the particle trajectories proportionally to their transverse momentum (p_T) as shown in figure 2.5. By measuring the curvature of the trajectory, the momentum of the particle is determined.

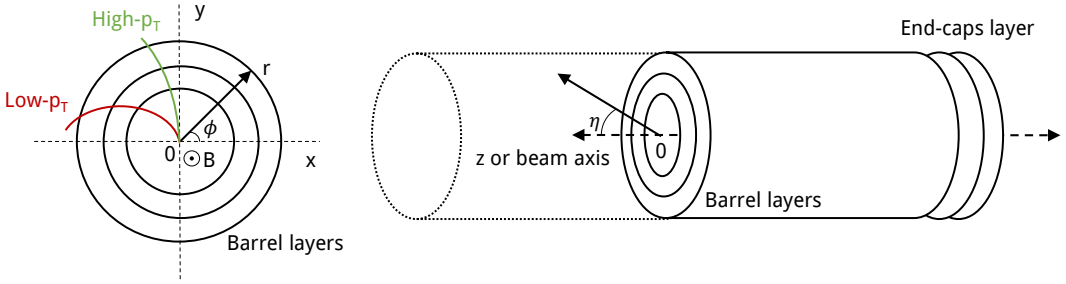


FIGURE 2.5: Structure and coordinate system of a generic tracking system. On the left it is shown the r-ϕ plane with an example of low and high p_T tracks. On the right it is shown the r-Z plane with the end caps layer at the end.

An additional challenge for the silicon detectors in HEP application is the resistance to radiation. Moreover in the case of the tracker, the detectors work in close proximity of the collision point and undergo extremely high radiation levels. For the application described in this thesis, the expected radiation level are more than three order of magnitude higher than the one reached in space application. This is the reason why, it is very important to study the radiation effects and to use radiation hardening technique which can prevent failures.

2.5 Radiation induced effect on CMOS technologies

Ionizing radiation can cause parametric degradation and ultimately functional failures in electronic devices. High energy electromagnetic radiation or particle radiation are a common hazard in environments such as outer space, high-altitude flight or near particle accelerators. These effects are particularly important for High Energy Physics detectors, as their role is to perform measurements in close proximity of particle collisions where they are exposed to higher radiation fluxes than in all the other applications. Thus, it is important to understand the different types of radiation effects and how to properly design electronics to minimize their impact on the performances of the systems.

Cumulative effects are gradual effects taking place during the whole lifetime of the electronics exposed in a radiation environment. A device sensitive to Total Ionizing Dose (TID) or displacement damage will exhibit failure in a radiation environment when the accumulated TID (or particle fluence) has reached its tolerance limits. It is therefore in principle possible to foresee when the failure will happen for a given, well known and characterized component. On the contrary, Single Event Effects (SEE) are due to the energy deposited by one single particle in the electronic device. Therefore, they can happen in any moment since the beginning of its operation in a radiation environment, and their probability is expressed in terms of cross-section.

2.5.1 Total Ionizing Dose effects

TID is the measurement of the dose, that is the energy, deposited in the material of interest by radiation in the form of ionization energy. The unit to measure it in the

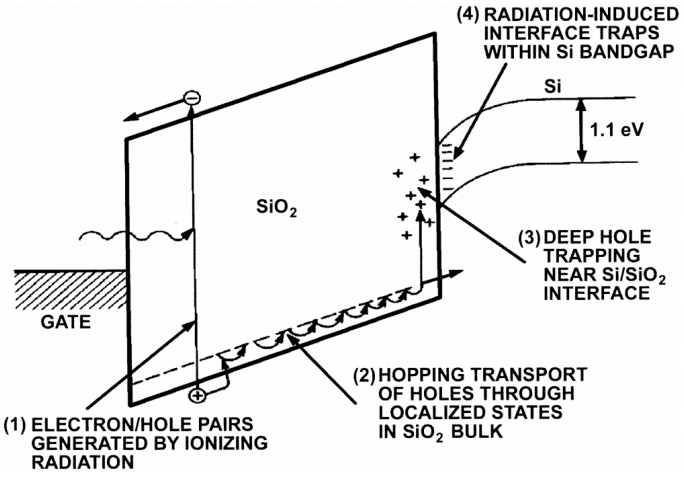


FIGURE 2.6: Schematic energy band diagram for MOS structure, indicating major physical processes underlying radiation response[12].

International System (SI) is the Gray, but the radiation effects community still uses most often the old unit, the rad. The equivalence between the two is:

$$1\,Gray\,(Gy) = 100\,rad. \tag{2.6}$$

TID effects are a typical case of cumulative effects. The ionization dose is deposited by particles passing through the materials constituting the electronic devices as described in Oldham and McLean [11] and shown in figure 2.6. These physical processes lead from the initial deposition of energy by ionizing radiation to the creation of ionization defects in the dielectric of a Metal-Oxide-Semiconductor (MOS) structure and can be summarized in: 1) the generation of electron hole pairs, 2) the prompt recombination of a fraction of the generated electron hole pairs, 3) the transport of free carriers remaining in the oxide, and either 4a) the formation of trapped charge via hole trapping in defect precursor sites or 4b) the formation of interface traps via reactions mostly involving hydrogen.

Using as reference the MOS structure shown in figure 2.7, the charges at the gate oxide of the transistor will screen or enhance (depending on the polarity of the transistor) the gate electric field. This will led to a threshold voltage shift. In lateral oxide as the Shallow Trench Isolation (STI) oxide used to isolate transistors from each other, they might attract an image charge in the semiconductor which can invert the interface and open leakage path [12]. This effect is typical of nMOS transistor. The defects formed at

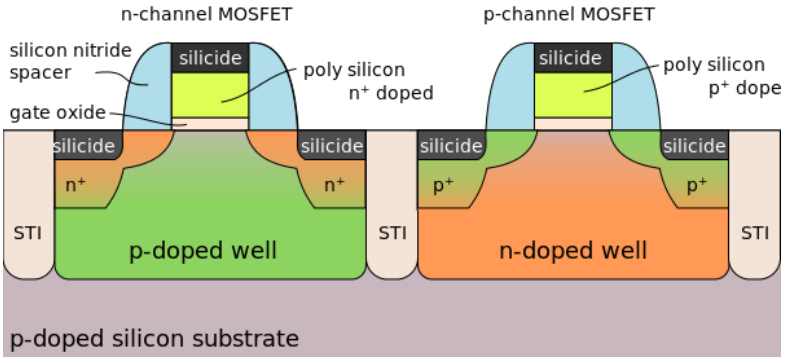


FIGURE 2.7: n-channel and p-channel MOSFET's. Individual transistors are electricaly separated by STI trenches.

the interface between silicon and silicon dioxide (this is the region where the conductive channel forms in a MOS transistor) are called interface states. They trap charge from the channel, which leads to both a threshold voltage shift and also affects the mobility of carriers in the channel. The two types of effects, the trapping of holes and the creation of interface states, have a very different dynamic. Holes are trapped very quickly, and can be detrapped by thermal energy (this is called annealing). Therefore, increasing the temperature is a good method to anneal the trapped charge. Interface states instead exhibit a slow formation, and they do not anneal at temperature below about 400°C.

2.5.2 Single event effects

Single Event Effects are caused by very localized event induced by a single particle. The incoming ionization particle loses energy in the semiconductor through Rutherford scattering (Coulomb interaction) with the lattice structure. The energy is transferred to the lattice as an ionization tail of free electron-hole pairs. In the bulk of the semiconductor, these will recombine with no effect. In a p-n junction or in its proximity, the pairs will be separated and collected, giving rise to a current spike. The collection of charge at a circuit node might give origin to:

- Transient errors which are frequent in analog circuits, or in combinational logic. The generated signals are asynchronous, they can propagate through the circuit during one clock cycle and also sometimes propagate to a latch and become static.

- Static errors, called Single Event Upset (SEU), that can be corrected by outside control. They overwrite information stored in the circuit, but a rewrite or power cycle can correct the error with no permanent damage.
- Hard errors which are those leading to a permanent error, possibly causing the failure of the whole circuit. They cannot be recovered unless detected at their very beginning in some cases (as for Latchup). In that case, it is possible to interrupt the destructive mechanism and bring back the circuit to functionality.

Concerning SEU, the probability of a particle causing one (or more) bit upset is dependent mostly part on two parameter: how much energy the particle is able to transfer to the silicon and the minimum charge needed for a storage element to flip state. The energy transferred is defined as Linear Energy Transfer (LET), a measure of the energy transferred to the device per unit length as an ionizing particle travels through a material. The common unit is MeV cm^2/mg of material. The minimum LET to cause a detectable effect in a node is called LET threshold. Experimental tests can be conducted to calculate the "cross section" of a device, which is a measure of the response of the device to the radiation. For a given LET, the cross section is the number of errors divided by the incoming particle fluence (#particles/cm2).

2.5.3 Radiation-hardening techniques

Techniques for radiation hardness can be divided in physical and logical. Physical techniques act at layout level and a very well known technique is the use of Enclosed Layout Transistor (ELT) [13]. As shown in figure 2.8, one of the diffusions (drain or source) is surrounded with the gate oxide, while the other diffusion is all around the gate. Such a layout avoids the STI oxide to touch both the ends of the channel of a transistor, forming a parasitic channel where leakage current could flow.

Single event upsets need a different approach. On a circuit level, cells that are more robust to injected charge in sensitive nodes can be designed. The simplest way of achieving this is to increase the capacitance of the sensitive nodes, in order to increase the minimum charge needed to upset the stored value. By accepting an area and power consumption penalty, the error rate can be decreased by more than one order of magnitude. Another solution is to create structures that have multiple nodes that must be

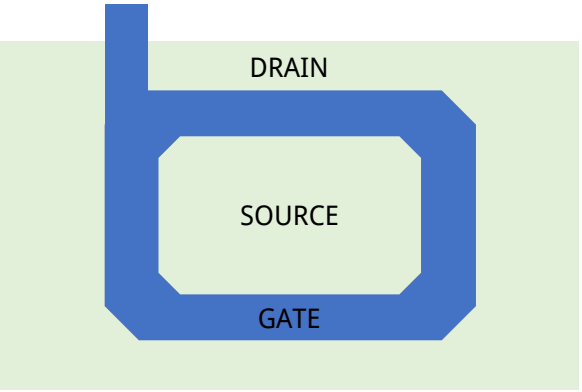


FIGURE 2.8: Layout of an Enclosed Layout Transistor.

upset at once to change the stored value: if the two nodes are spaced out in the layout, the probability of both of them being hit by a particle at the same time is drastically reduced. An example of this is the DICE cell, as described in Naaser[14].

Other techiques, as Triple Modular Redundancy (TMR) and Error Correction Coding (ECC), can be used to make circuits very tolerant to SEEs. TMR is based on triplicated logic in which the correct result is a vote of the three outputs. If only one device has been upset, the output of the voting is still correct. ECC can also be used to correct single-event upsets or even detect multiple bit upsets. These techniques, however, introduce area, power and timing penalties which for a fully triplicated design can be easily estimated: the area overhead is always more than 200% as voting logic is required in addition to the triplication overhead.

In conclusion, there are effective techniques to reduce the radiation effect on ASIC devices, but the area and power penalty must be considered from the begin of a project. So, in power and area constrained design the expected radiation level and the acceptable error rate must be well defined in order to implement the correct radiation hardness structures.

Chapter No. 3

A particle tracking system for future HEP experiments

In order to maintain or improve the physics performance of the CMS detector in the high pileup conditions of the upgraded LHC performance, the entire tracking system will be replaced with new detectors featuring higher radiation tolerance and enhanced functionality. The particle tracking system will contribute to the Level-1 trigger decision, in order to maintain and improve the ability to select interesting physics events at high pileup. A new concept of "intelligent"silicon detector modules with local data selection and data reduction integrated in the front-end readout electronics will enable the implementation of the trigger functionality. The first part of this chapter provides an overview of the current tracker architecture and performance, explaining which part must be substituted or improved. This introduction is followed by the description of the different component of the upgrade, focusing mainly on the front-end modules.

3.1 The CMS experiment

This thesis is dedicated to the development of a new particle tracking system for one of the experiment on the LHC accelerator, the CMS experiment. This detector measures 21.6 meters in length, 15 meters in diameter and weighs about 14000 tonnes. Approximately 3,800 people, representing 199 scientific institutes and 43 countries, form the CMS collaboration who built and now operate the experiment [8].

The overall layout of CMS is shown in figure 3.1. The core of the detector is a huge 4-T superconducting solenoid magnet which measures 13 m and has a inner diameter of 6 m. The magnetic field is confined by a steel yoke that forms the bulk of the detector's weight of 12500 tonnes.

The detector contains subsystems which are designed to measure the energy and momentum of photons, electrons, muons, and other products of the collisions. The innermost layer is a silicon-based tracker. Surrounding it is a scintillating crystal electromagnetic calorimeter, which is itself surrounded with a sampling calorimeter for hadrons. The tracker and the calorimetry are compact enough to fit inside the solenoid. Outside the

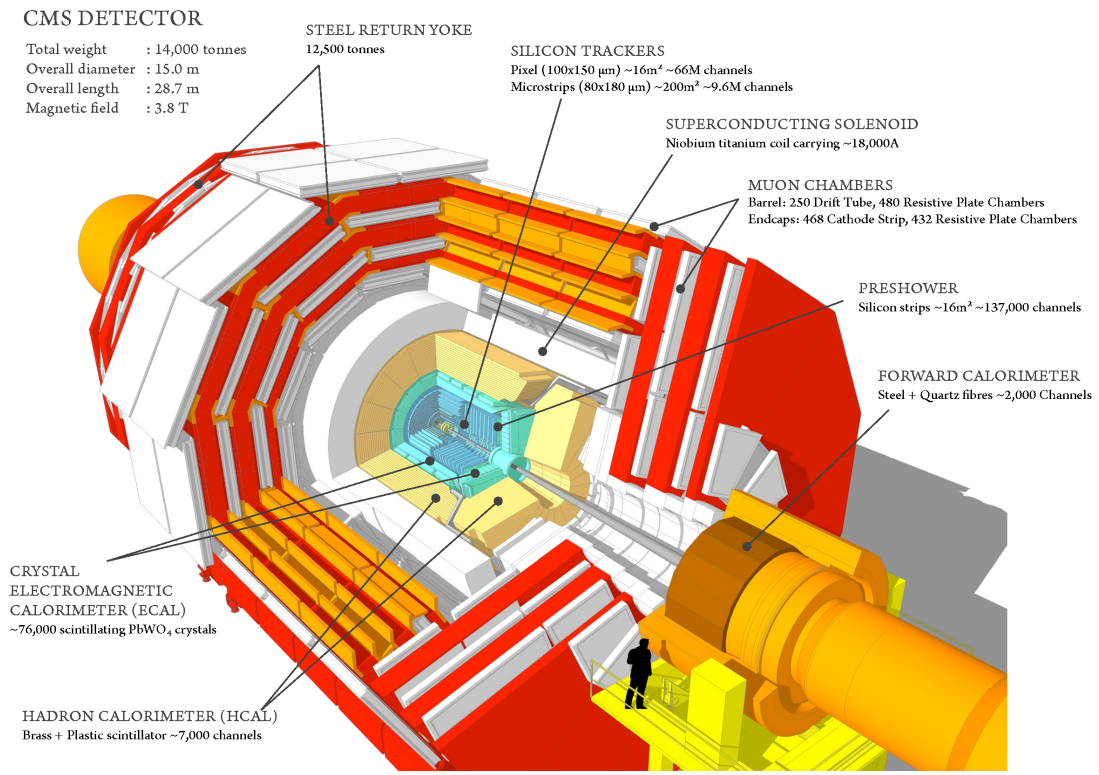


FIGURE 3.1: A perspective view of the CMS detector.

magnet are the large muon detectors, which are inside the return yoke of the magnet. A detailed description of the CMS detectors can be found in the collaboration's article [15].

3.1.1 The current CMS Silicon Strip Tracker

The current version of the particle tracking system of the CMS detector is completely based on silicon strip technology. It is called Silicon Strip Tracker (SST) and is the largest detector of its kind ever built and operated. It is roughly 5.6 m long in the z direction, with a diameter of 2.2 m. With its 10 barrel layers and 12 endcap disks per side, it features about 200 m^2 of sensitive surface in 15148 modules with 9.3 x 109 channels read out through 36392 analogue optical links. This sub-detector was designed to operate at luminosities up to about $10^{34}cm^{-2}s^{-1}$. Significant robustness and redundancy in the tracking capability was implemented in the detector layout, to ensure optimal performance for several years of operation (up to an integrated luminosity of about 500 fb^{-1}), with basically no maintenance or repairs.

The Silicon Strip Tracker is based on a triggered readout. The trigger, called Level-1 (L1) trigger, is generated by custom electronics that process data from the calorimeters and muon detectors in order to select the most interesting events from LHC collisions. During the trigger generation, the readout data of the tracker are stored in front-end pipelines. The latency time between the event and the L1 trigger arrival time is fixed to a value of 3.2 μs and is called L1 latency. When the trigger reaches the front-end, the modules send to the experiment back-end the requested event.

The SST extends from 20 to 120 cm in radial distance from the Interaction Point (IP) and up to 280 cm in length. Tracker modules vary in shapes and dimensions among the different regions of the detector. These modules host four or six readout chips, called APV25. This chip has 128 amplifying channels and is designed in 0.25 μm CMOS technology. The signal shaping with a de-convolution filter has a shaping time of 25 ns. Further a pipeline buffer of 192 columns can store LHC bunch crossings over 4.8 μs, to allow a decision from the CMS first level trigger system. A full description of this component can be found in the article by Raymond et Al.[16].

3.2 The High Luminosity LHC

The Large Hadron Collider (LHC) project concluded its first phase of physics exploitation with a center-of-mass collision energy of 7 TeV (half of the nominal value) and a record peak instantaneous luminosity of $\sim 3 \times 10^{33} \mathrm{cm}^{-2}\mathrm{s}^{-1}$.

The performance of the LHC in delivering luminosity to the experiments is continuously growing. According to the machine plans [17] in figure 3.2, the nominal luminosity of $\sim 1 \times 10^{34} \mathrm{cm}^{-2}\mathrm{s}^{-1}$ and centre-of-mass energy of 14TeV should be reached during the current run, after a machine upgrade which was performed during the first long shutdown of the LHC carried out during 2013-2015. Two more long shut-downs are planned towards the end of 2010s and mid 2020s and the instantaneous luminosity of the machine should exceed the design goal. Before the third long shutdown the silicon pixel vertex detector of CMS will be replaced to cope with the increased particle density. After the last upgrade of LHC the instantaneous luminosity is expected to reach $\sim 5 \times 10^{34} \mathrm{cm}^{-2}\mathrm{s}^{-1}$. This scenario is known as High-Luminosity LHC

The HL-LHC scenario represents a challenge for the detectors due to the high radiation dose (up to $10^{15} n_{eq} cm^{-2}$ at 20 cm) and the high number of pile-up events (up to 200). For the CMS detector, these high levels of pile-up are a problem for event reconstruction (separating the signal of different particles and reconstructing primary events). Under these conditions the Silicon Strip Tracker must be replaced during the third long shutdown in preparation of HL-LHC. The whole upgrade of the experiment is called "Phase-2" upgrade.

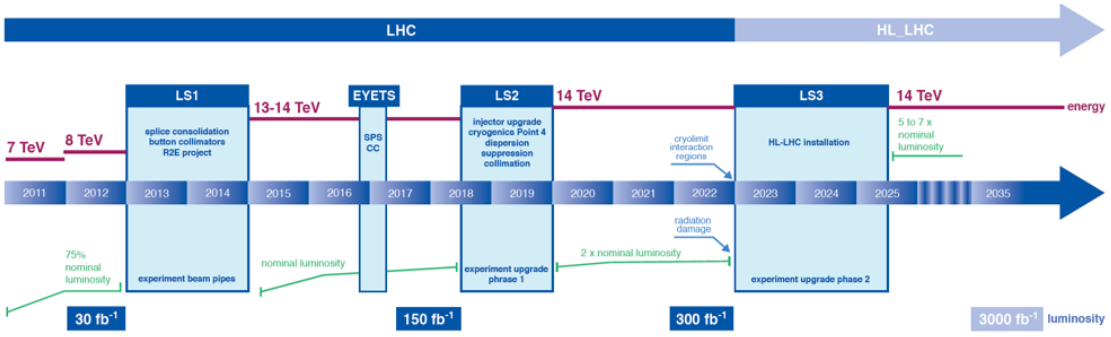


FIGURE 3.2: High Luminosity LHC plan [17].

3.3 The CMS Phase-2 upgrade

The phase 2 upgrade gives the CMS collaboration the opportunity to improve the performance of the detector: increasing the resolution of the track reconstruction and possibly reducing the amount of particle interactions in the tracking volume.

The quantity of interaction is expressed respect the radiation length (X0), which is the characteristic length that describes the energy decay of a beam of electrons. As shown

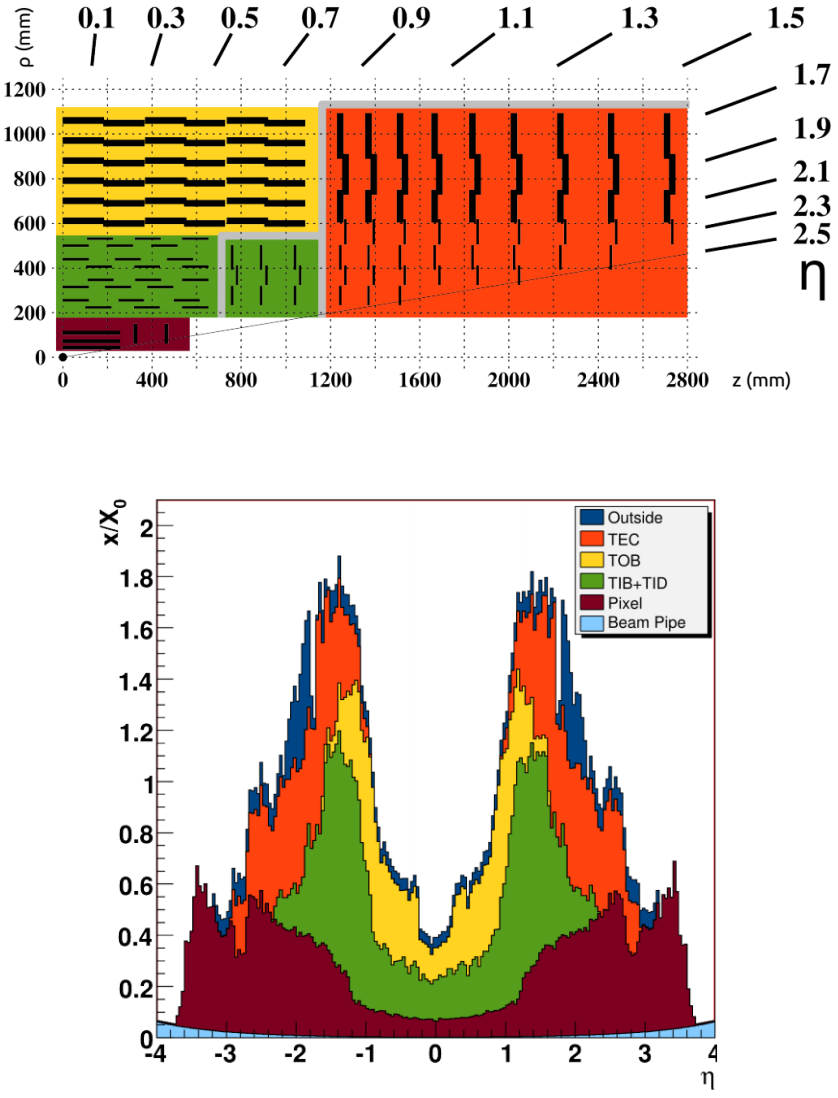


FIGURE 3.3: Sketch of the Silicon Strip Tracker layout (top), and distribution of material in radiation lengths as a function of pseudorapidity (bottom). The peak in the region $1< \eta <2$ contains an important contribution from the services routed between barrel and end-cap.

in figure 3.3, the material in the current CMS detector ranges from approximately 50% of a radiation length in the barrel section to 180% at $\eta = 1.5$. This is detrimental to the resolution of the electromagnetic calorimeter because of the loss of energy, and affects also the tracking resolution of charged particles through multiple scattering effects which modify the particle trajectory.

The granularity of the detector should increase approximately by a factor 5 in order to cope with a maximum foreseen pile-up of 200 collisions per bunch crossing and keep the occupancy to a few % level as in the current detector. The radiation hardness of the sensor material should increase to move from the current LHC design integrated luminosity of 500 fb^{-1} to an expected one of 3000 fb^{-1} for HL-LHC.

To this date the only sub-detector to be fully replaced in CMS for HL-LHC is expected to be the Tracker. On the other hand the services for the tracker (readout, power supply, cooling, etc.) are not accessible without partially dismantling those for other sub-detectors. For this reason the upgraded Tracker is constrained to use the same services as the present one. This fact together with the need of an increased granularity requires the development of new solution for data readout, powering and cooling.

In a collision lots of particle are produced. The event reconstruction often looks for a particle not directly but through its decay products. Looking for a specific decay pair have a certain probability of finding a random combination of other products which look similar. This combinatorial background poses problems to the Level-1 trigger to the point where it is not possible to reconstruct the event only with the information from the calorimeter. A possible solution would be to increase the L1 rate which is defined as the average frequency of L1 trigger. Currently, this rate is 100 kHz but to cope with the number of collision of the HL-LHC it should be increased above the possibility of the readout system.

In the High Level Trigger (HLT), a streamlined version of the CMS offline reconstruction software running on a computer farm, a large trigger rate reduction factor is currently achieved by using information from the Tracker. The study summarized in Pesaresi [18] shows as including this information in the Level-1 trigger generation would allow a much better rejection of combinatorial background. The collaboration decided therefore to include the option of sending tracking information in real time to the Level-1 trigger

in the new tracking system for the HL-LHC, launching the development of the so-called p_T modules [19].

3.3.1 The p_T modules concept

The higher granularity and the requirement of sending tracking information at each bunch crossing correspond to a very large increase of the needed bandwidth from the tracker module to the experiment back-end. Even exploiting newer technologies which provides higher communication speed, the available bandwidth is not enough to deliver all the data. This limitation excludes the measurement of the particle charge, and drives the choice of a binary readout. The front-end electronics will provide 1-bit information per channel per BX which indicates if the signal from the sensor was above or below threshold.

Nevertheless, the amount of data generated is still above the bandwidth availability. Indeed, in order to provide data in real time, the system must send the full event with a frequency of 40 MHz, which corresponds to a flow of 10^4 Tb/s for the whole tracker. A possible solution consists in making a front-end electronics capable of selecting only the interesting events. The discrimination among particles is based on their transverse momentum (p_T). The requirement for the module design is therefore to include the "intelligence" to reject locally the signals from particles with low transverse momentum. These particles are not interesting for the Level-1 reconstruction and rejecting them with p_T threshold of 1–2 GeV/c reduces the bandwidth requirements by at least one order of magnitude. This information, with a reasonable increase in the L1 rate, would be enough to make possible the event reconstruction in the High Luminosity scenario.

The front-end electronics thus must be able to recognize the particles with low-p_T. This feature is achieved by measuring, not simply the position of the particle, but also its direction. The curvature of a charged particle in the magnetic field of 4 T, provided by the superconducting solenoid surrounding the tracking system, is directly proportional to its p_T. Therefore, if two planar sensors are placed one on top of the other with a spacing of a few millimiters, the curvature of a particle can be estimated by measuring the distance between the incidence positions of the particle in the two layers.

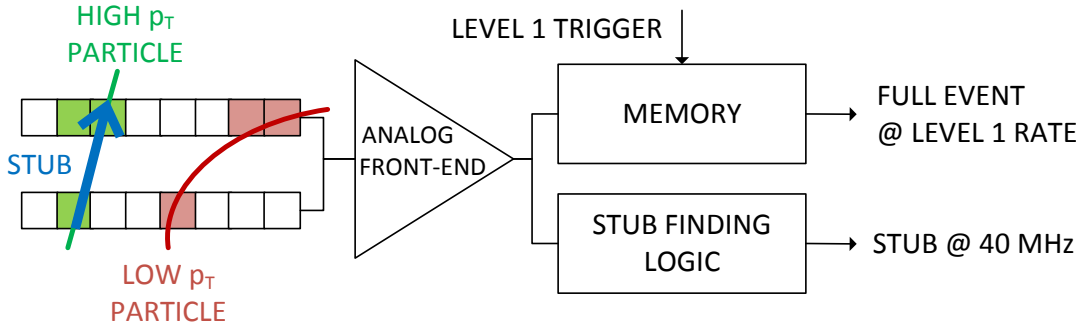


FIGURE 3.4: A simplified view of the readout electronics including p_T discrimination.

For this reason, as shown in figure 3.4, the modules are composed of two closely-spaced silicon sensors read out by a common electronics. The front-end correlates the signals collected in the two sensors, and select pairs compatible with particles above the chosen pT threshold. These pairs are called "stubs". The strong magnetic field of CMS provides sufficient sensitivity to measure pT over the small sensor separation, enabling the use of pT modules in the entire radial range above ∼20 cm from the interaction point.

Summarizing, the new tracking system combines the triggered readout of the full event present in the SST with a trigger-less readout of selected information. Indeed, stub data are sent out at every bunch crossing, while all other signals are stored in the front-end memories for reading out when a L1 trigger is received.

3.4 A particle tracking system for the HL-LHC

Table 3.1 summarizes the general requirements of the tracking system for the High Luminosity upgrade compared with the SST specifications:

Parameter	LHC specs	HL-LHC specs
Bunch Crossing frequency (BX)	40 MHz	40 MHz
L1 rate	100 KHz	750 KHz
L1 latency	3.2 μs	12.8 μs
Radiation tolerance	500 fb^{-1}	3000 fb^{-1}
Participation to L1 trigger	No	Stubs at 40 MHz
Temperature	-20 °C	-30 °C

TABLE 3.1: Comparison between the specs for the tracking system of CMS for LHC and HL-LHC

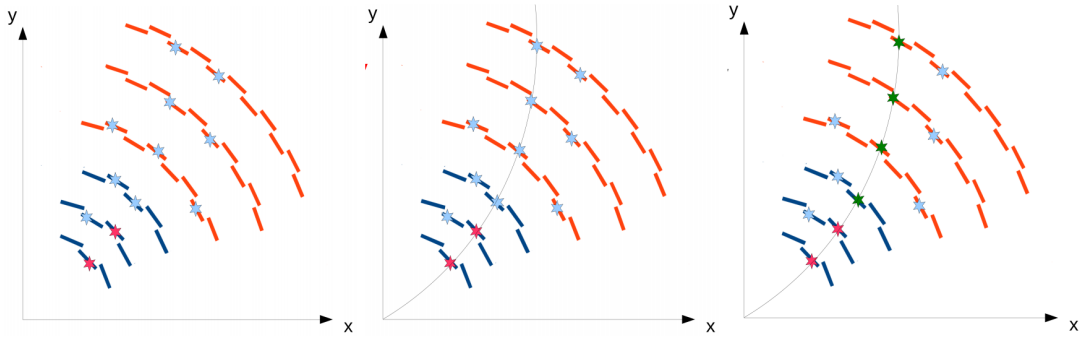


FIGURE 3.5: A three steps track reconstruction algorithm is illustrated. Step 1: pairs of stubs in neighboring layers are combined to form the seeds. Step 2: the seeds are projected to the other layers and matching stubs are found. Step 3: the matched hits are included in the final track fit.

The L1 rate needs to be increased to 750 KHz although the introduction of p_T modules to allow event reconstruction in the high rate environment. Also the L1 latency needs to be increased in order to process the additional information from the front-end electronics for the L1 trigger generation. The processing of these p_T data in the experiment back-end is called "Level-1 Track finding". As explained by Pesaresi [20], it consists of a track trigger system which would identify tracks with transverse momentum above 2 GeV/c. A track is generated combining the stubs received by the front-end modules of the different layers. An example of track reconstruction is given in figure 3.5. The role of the tracking trigger is to deliver track objects to the L1 trigger within about 5 μs, in order to allow this information to be merged with that from other sub-detectors. The whole L1 latency will be $< 12.8\,\mu$s.

A complete and detailed description of the new tracking system for the CMS experiment can be found in the technical proposal for the CMS upgrade [21]. In the following paragraphs only an overview of the full system with particular attention for the electronics will be given.

3.4.1 CMS Tracker structure

The CMS tracking system extends about 1.2 m radially from the beam and cover a length of approximately 5 m. Hit rate and radiation level vary strongly with the distance from the interaction point. It is thus impossible to fulfill the requirements in the different area of the tracker with the same module design and front-end electronics. For this reason, the tracker has been divided in two part:

- **Pixel Detector** will extend between 3 cm and 20 cm from the beam line. HL-LHC is expected to deliver, over 10 years of operation, about 1 Grad of radiation dose in silicon, at about 3 cm from the interaction region where the first layer of the pixel detector could be located. Hence, radiation hardness represents the main challenge for this sub-detector. Moreover, so close to the interaction point the curvature of particles due to the magnetic field is too small to measure and discriminate the particle p_T. Therefore, this sub-detector will not participate to the Level-1 trigger. More information can be found in the technical proposal [22].

- **Outer Tracker** will extend from 20 cm to 120 cm from the beam line. The expected dose is limited to 100 Mrad. The distance from the interaction point makes possible the implementation of p_T modules. The main challenge is the integration of pixels together with the development of a readout architecture for intelligent particle tracking in a module that fulfills the material requirement.

3.4.2 Outer Tracker module design

The module design for the Outer Tracker is based on hybrid circuit board technology. The main components of a module are the two sensor layers, the hybrid circuit board which host the readout electronics, called Front-End hybrid, the one which host the service electronics, called Service Hybrid, and the cooling structure. The Service Hybrid contains the DC/DC converters for powering and the electronics for optical fiber communication. Power will be delivered at a higher voltage to reduce ohmic losses in the services and then converted to the required low voltage inside the tracking volume. The use of one optical link per module provides the bandwidth needed for the trigger functionality, and at the same time offers significant advantages in the overall system design by avoiding additional electrical interconnectivity in the tracking volume. In order to optimize the amount of material, two types of p_T modules are under development for the Outer Tracker:

- **"2S" modules** are composed by two superimposed strip sensors of approximately $10 \times 10\,cm^2$, mounted with the strips parallel to one another. They populate the outer regions, above r ∼60 cm (in red in the sketch of figure 3.6), which entails approximately $150\,m^2$ of sensing area. Wire bonds at opposite ends of the sensor

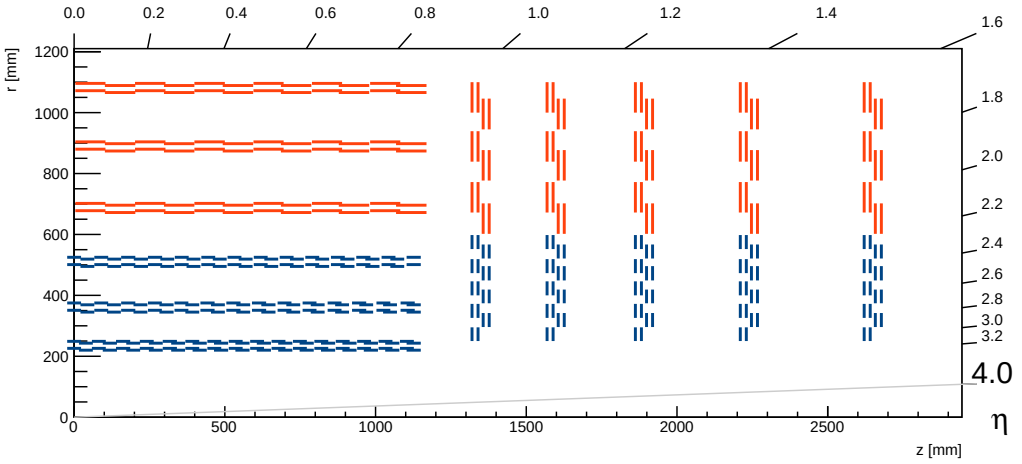


FIGURE 3.6: Sketch of one quarter of the Tracker Layout. Outer Tracker: blue lines correspond to PS modules, red lines to 2S modules. The Pixel detector, with forward extension, is shown in green.

provide the connectivity of both sensors to the Front-End Hybrid. A single Service Hybrid carries a 5 Gb/s data link, an optical converter, and the DC/DC converter that provides power to the module electronics.

- **"PS" modules** are composed of two sensors of approximately 5 x 10 cm^2, one segmented in strips, and the other segmented in "macro-pixel" of size 100 μm x 1.5 mm. The chosen pixel size permits the use of the flip-chip bump-bonding technology, an industrial process that is expected to be affordable for a large-scale production. As for the 2S module, wire bonds provide the connections from the strip sensor and from the macro-pixel readout chip to the front-end hybrid, and, in turn, to the auxiliary electronics for powering and readout, all of which is integrated in the module assembly. PS modules are deployed in the radial range between r ~20 cm and r ~60 cm (blue in the sketch of figure 3.6), resulting in a sensor surface area of about 60 m^2 (30 m^2 short strip sensors and 30 m^2 macro-pixel sensors).

On the one hand, the pixelated sensors provide sufficiently precise measurements of the z coordinate for tracking to enable primary vertex discrimination at Level 1 trigger generation. At the same time, three additional layers of unambiguous 3D coordinates each with associated an estimate of the particle pT, are of particular use for track finding, offering enhanced robustness for the pattern recognition in a more cost effective way than, for instance, an extension of the Pixel detector to include additional layers at larger radii.

On the other hand, despite the higher resolution, the PS module can not be used in all the tracker because of the higher power consumption and the consequent higher material needed for cooling. To remove heat from electronics and sensors, CO_2 two-phase cooling will be used. Such choice of cooling technology helps to reduce the amount of passive material in the tracking volume. The design of the modules should provide for efficient removal of the heat generated by the electronics and sensors, accurate geometrical positioning, minimal mass, as well as a simple and reproducible assembly procedure. For the thermal performance, the tracking system cannot be operated at room temperatures since damage induced by traversing particles would render it inoperable after only a fraction of its foreseen lifetime: electrical currents through the sensors increase linearly with radiation damage. Fortunately, these currents are also exponentially dependent on the temperature, and can be largely reduced by running at low temperatures. The design requirement is to achieve a sensor temperature of -20 °C or lower with a coolant temperature of -30 °C for modules irradiated with the full HL-LHC integrated luminosity.

The layout of the Outer Tracker has been the subject of extensive studies and detailed modeling, exploring several variants, including geometries with barrels only, and geometries with different numbers of barrel layers, and/or different numbers and size of end-cap disks. The version shown in figure 3.6 has been adopted as baseline design, as it provides efficient use of the silicon sensors while providing good tracking performance and minimizing both cost and material in the tracking volume. Further optimization of the inner barrel region may be possible, and is being explored. It should be noted that all end-cap disks are equipped down to the lowest radius, to be compatible with an extension of the tracking acceptance up to $\eta = 4$, while in the present tracker rings located beyond $\eta \leq 2.5$ are not equipped.

3.4.3 Silicon sensors choice

The particle fluence is the main constraint in the choice of the silicon sensors. They will be exposed to fluences up to $1.5 \times 10^{15} \mathrm{n}_{eq}\mathrm{cm}^{-2}$, a factor of ten larger than the design requirement for the present Tracker. In this hard environment, the performance of p-in-n float zone sensors degrades too much [23]. Sensors with electron read-out are more robust in terms of high field effect after irradiation and provides higher charge collection

than p-in-n sensors. The choice of thin sensors (200 μm) could offer advantages in terms of reduced leakage current and less material in the tracking volume.

The Tracker modules requires three types of sensors:

- 2S sensor: strip sensor for the 2S module. AC coupled sensor of approximately $10\,\mathrm{x}\,10\,\mathrm{cm}^2$, with two rows of 5 cm long strips with 90 μm pitch.
- PS-s sensor: strip sensor for PS module. AC coupled sensor of approximately $5\,\mathrm{x}\,10\,\mathrm{cm}^2$, with two rows of 2.5 cm long strips with 100 μm pitch.
- PS-p sensor: "macro-pixel" sensor for PS module. DC coupled sensor of approximately $5\,\mathrm{x}\,10\,\mathrm{cm}^2$, with 32 rows of macro-pixels 1.5 mm long with 100 μm pitch.

In summary, in the baseline layout the Outer Tracker consists of 15508 detector modules (8424 2S and 7084 PS), with a total active surface of 218 m^2 , 47.8 million strips and 218 million macro-pixels.

3.5 Outer Tracker electronics

After the previous general introduction, the electronic system for the Outer Tracker will be discussed in details. Its block diagram is shown in 3.7. The system is designed to deliver trigger and L1 readout data with high efficiency up to a L1 accept rate of 750 kHz, and to cope with latencies up to 12.5 μs.

At the front end, the electronic system is built around the sensor modules (2S or PS). Electrically, it consists of two FE Hybrids, detailed by Blanchot et Al. in [24], interfacing to the two rows of strips of the silicon sensor(s) (2S and PS modules), plus, in the case of the PS module, of a Macro-Pixel-Sub-Assembly block (MaPSA) integrating the pixelated sensor with its readout chips. Data generated by eight FE chips are buffered, aggregated and formatted by the Concentrator IC (CIC) that acts as a data hub for the Service Hybrid. The latter hosts all services to/from the counting room: data transfer, low voltage powering (LV) and high voltage biasing (HV). It connects to the FE Hybrid through wire bonds made at the same time as those that connect the FE hybrid to the sensors. The LP-GBT is a communication and monitoring ASIC which serializes (deserializes) data sent to (received from) the optoelectronic transceiver, which is called

VTRx$^+$. It also acts as I2C master of the module, controlling the FE ASICs and contains monitoring functions that will be used to check environment and functionality. Both FE and Service Hybrids are multilayer high density flexible kapton circuits that will be tightly folded during module assembly to provide connectivity to both top and bottom sensor and/or to sensor backsides, while limiting impact on material budget.

The front-end electronics generates two separate data path: a trigger-less data path, called Trigger path, which contains the high-p$_T$ information for the Level-1 event reconstruction, and a triggered data path, called DAQ or L1 Data path, which provides the full event (raw data or L1 data) when requested by the Level-1 trigger. At the back-end, the Data, Trigger and Control board (DTC) sends and receives data to/from multiple modules (typically 50-70 modules per card). This board is a custom development based on off-the-shelf commercial FPGAs and multi-channel optoelectronic transceivers, which is located outside the cavern wall in a radiation free environment. It processes three data streams to/from the detector: DAQ, Trigger and Timing and Control.

As explained by Abbaneo in [25], the DTC board sends the Trigger data, a stream of stubs selected by the front-end electronics, to the L1 Track Finding system. The latter combines the stubs from the whole tracker to perform pattern recognition to reconstruct the tracks of primary particles with pT $>2\,$GeV, and discard as many as possible of all

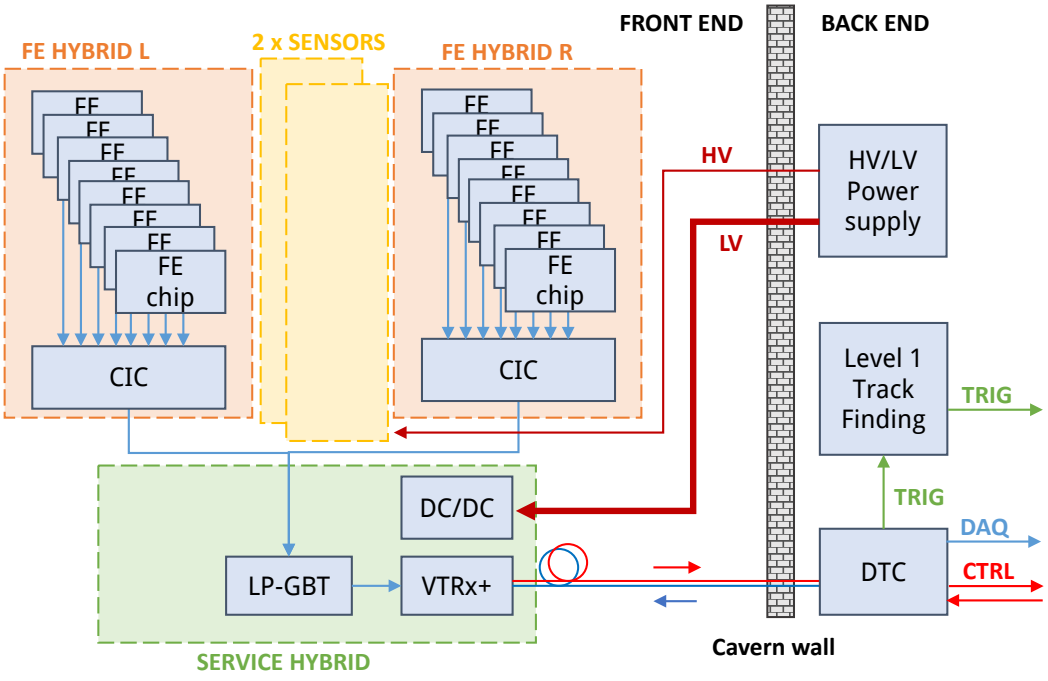


FIGURE 3.7: Electronic system block diagram.

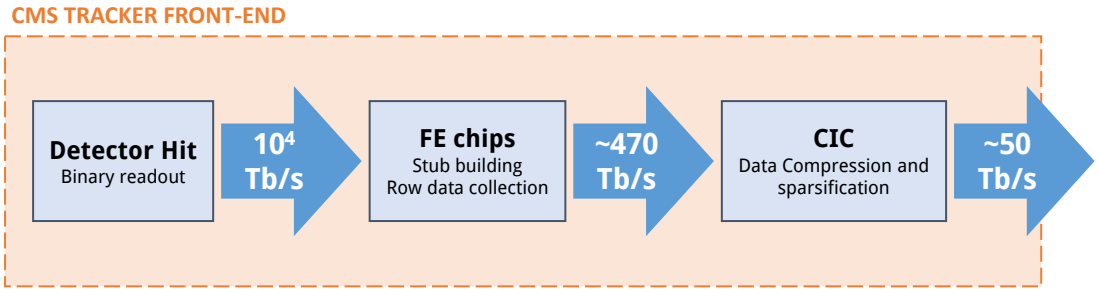


FIGURE 3.8: CMS Tracker front-end data reduction scheme.

the other stubs. The found tracks are sent to the CMS Level-1 trigger and will be used for the trigger generation with the information from the other sub-detectors.

The back-end contains also the power supply for high and low voltage: power will be supplied to the FE at a voltage of 10-12 V while the HV supply will be designed to reach a voltage of -800 V. The High Voltage for the sensor is provided with cables from the cavern and requires a power level of 1 W at the end of life of the innermost sensors.

3.5.1 Data flow reduction

The objective for the introduced system is to extract simultaneously the trigger data at 40 MHz and the L1 data from event passing the first trigger level up to 750 kHz. The different front-end compression steps are represented in figure 3.8. Everything starts with the binary signals of the 250 M channels of the future tracker. At 40 MHz, it represents roughly a total amount of 10^4 Tbps.

The data is extracted from the module by 5 or 10 Gbps bi-directional optical link. Excluding error correction and upstream, the available bandwidth for downstream data transfer is 3.7 Gbps. Considering the entire system, the available bandwidth is ∼50 Tbps, consequently the front-end electronics must provide a compression factor of 200. The first compression step is carried out in the FE ASICs which combine the zero-suppression technique, position encoding and rejection of low p_T particles to achieve a compression factor of around 20. Clearly, another data compression step is mandatory. The second compression step is performed by the CIC and strongly relies on the low tracker occupancy ensured by the high granularity of the tracker, both in space and time. In the CIC chip, the signal of every module will be gathered over time (8 BX) and space (8 chips) as sketched on figure 3.10, mixing high rate events with low rate ones. The two

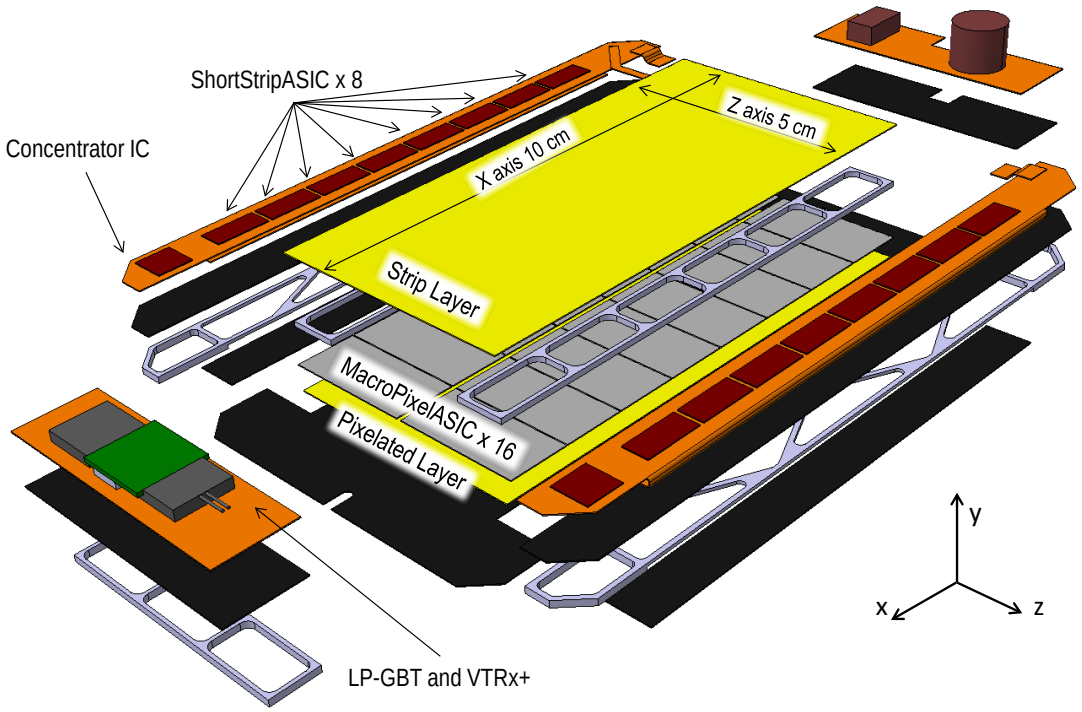


FIGURE 3.9: PS module 3D model.

CIC chips per module ensure thus another factor 10 in data compression, fulfilling the requirement.

3.5.2 Pixel-Strip module

Between the two modules for the Outer Tracker, the Pixel-Strip one is the more challenging from the points of view of, both, mechanics and electronics. As shown in figure 3.9, since the PS module is constructed using one pixelated and one strip sensor, two front-end chips must be developed: the Short Strip ASIC (SSA) and the Macro-Pixel ASIC (MPA).

The SSA sits on the FE Hybrid and it is connected with flip-chip technology. Strip sensor is wire-bonded to the same hybrid, which provides the strip signals to the SSA. The MPA is bump-bonded to the macro-pixel sensor. Due to the large pixel size, a standard bump pitch of 200 μm can be used, relaxing the assembly requirements. Two rows of eight MPAs are bumped to each macro-pixel sensor, resulting in an assembly of approximately 32000 pixels per module (2000 bumps per MPA). The MPA periphery is wire-bonded to the Front-End Hybrid and the connectivity between SSA and MPA is made with folded kapton connections.

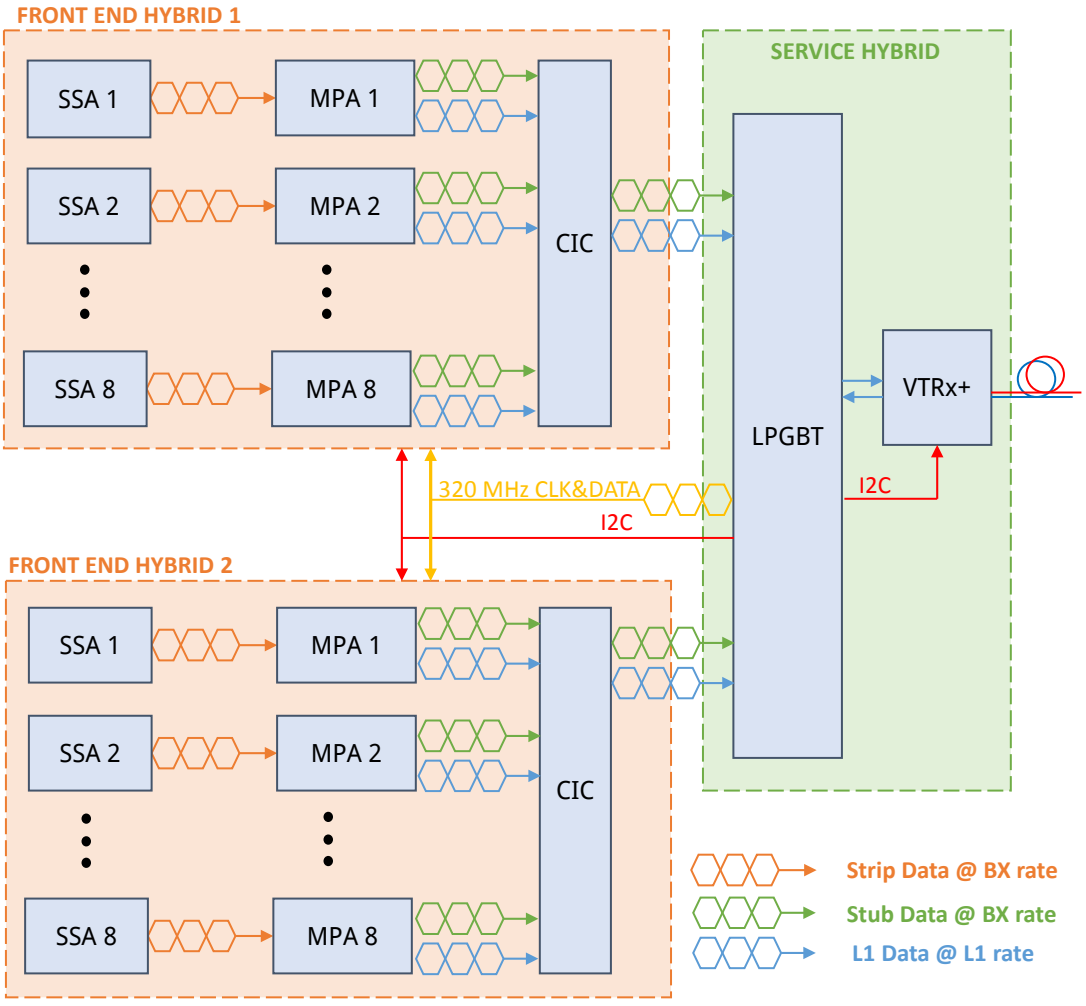


FIGURE 3.10: PS module block diagram.

The data flow is shown schematically in Figure 3.10. The SSA processes the strip sensor signals, and sends strip data without any compression through high speed links to the corresponding MPA. On the other side of the module, the MPA processes and sparsifies the signals from the pixel sensor. Moreover, it correlates the bottom macro-pixel sensor hits with the data received from the SSA strips in order to generate stubs and stores the full pixel and strip events for the L1 latency.

L1 Data and Trigger paths are divided in the transmission to the CIC. The latter, located on the FE Hybrid, aggregates the data received from the MPAs, and sends them to the Service Hybrid. However, the limited bandwidth available from the LP-GBT constrains the amount of data that the CIC can pass on to the Service Hybrid and requires detailed simulation in order to optimize the readout efficiency. The LP-GBT distributes the clock to the two front-hybrid, acts as I^2C master and send/receive data to/from the optical

module.

Differently from the 2S module, the PS one shows a second Service Hybrid dedicated to the power distribution. The Service Hybrid on the PS module measures only 5 cm respects the 10 cm of the 2S version, thus a single one can not host the DC/DC converters for powering and the other services. The power distribution is one of the key elements for the success of the PS module since the power consumed in the conversion is estimated as 1/3 of the total one. For this reason the next paragraph describe the architecture chosen for the power distribution on the module.

3.6 DC/DC converter for the Outer Tracker

All the electronics that operate inside the detectors have to work in an extremely hard environment because of the high magnetic flux density and high level of radiation within the experiments. Because of that, the solution currently implemented (top of figure 3.11) for the distribution of low-voltage power to the front-end electronics inside the LHC trackers is based on remote power supplies that are located in areas safe from radiation hazards, that feed the front-end systems via cables about 100 meters long. To cope with the large currents at low voltages and long distances, thick copper sections are used. Despite this, the cables develop large voltage drops across them and an increase of temperature due to the losses. The CMS tracker for the HL-LHC will contain an increased number of sensor channels, that is associated with an increase of power to be delivered, without increasing the amount of cables neither their cross sections. For these reasons, and in view of the upgrade of the LHC, it is necessary to evaluate the possibility of an alternative power distribution scheme.

Several powering topologies are being explored to power the Outer Tracker. The chosen baseline is based on the use of on-board DC-DC converters to efficiently distribute power to the on-detector electronics of tracker module. The use of DC to DC converters provides a more efficient power distribution, allowing for smaller and fewer input cables to be brought in the front-end area. Such a scheme allows to keep the current per module lower than 1 A and consequently limits the voltage drop along the cables. It also provides the option to move the power supply (PS) from the balcony to the counting room where it can be accessed also when LHC is running.

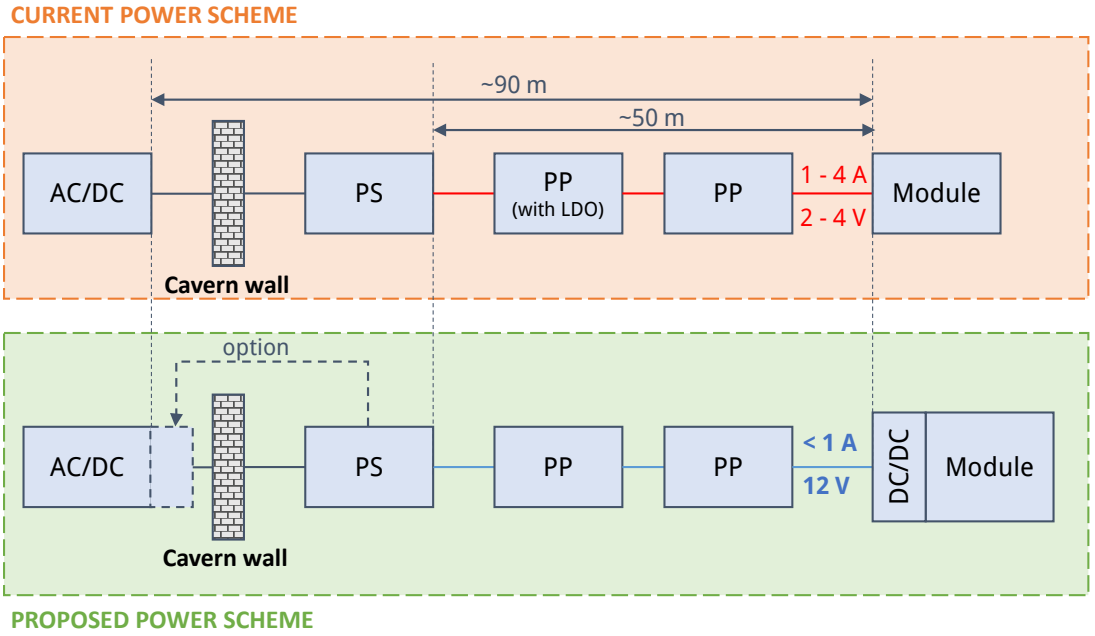


FIGURE 3.11: Current and proposed configuration for tracker power scheme.

3.6.1 On-module power distribution

An on-module DC/DC converter must fit the following requirements:

- **Radiation tolerance**: there is no commercial DC/DC chip able to tolerate the TID levels of the Outer Tracker (∼100 Mrad). This is the reason why an ASIC must be designed using layout techniques that increase the tolerance of the components of the chip.
- **Tolerance to magnetic flux density**: The tracker is surrounded by a strong magnetic field (up to 4 Tesla) parallel to the beam axis which would saturate any ferromagnetic material, therefore an air-core inductor is used.
- **Reduced size**: The limited space on the module requires an optimization of the ASIC and of the components on the hybrid in order to make the integration possible.

The on-module DC/DC converters for the Tracker are derived from common developments for LHC experiment upgrades. They are based on Buck converters with air core inductors. The voltage conversion from the input voltage (12 V) to the low voltage input of the ASICs exploits a two stages architecture as shown in Figure 3.12. The first stage

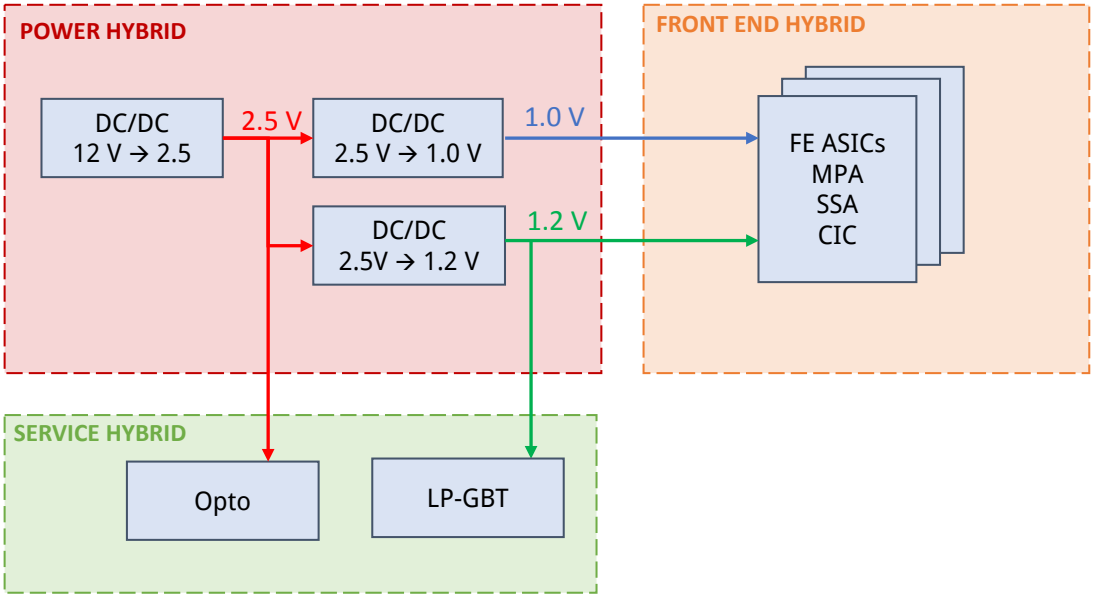


FIGURE 3.12: PS-module power distribution scheme.

converts from 12 V to 2.5 V and powers the optical device for fibre optics transmission (VTRX+) and the second stage. The second stage comprises two DC/DC converters which convert from 2.5 V to 1.2 V and ∼1 V. The 1.2 V powers all the remaining ASICs on the module while the lower voltage powers the digital core of the front-end electronics. The required power levels are between 6 W and 8 W assuming a 75% DC/DC conversion efficiency in the first stage and 85% in the second.

3.7 Chapter Summary

The High Luminosity scenario of the LHC machine represents an engineering challenge from several points of view. The increased number of collision drives the development of a new concept of tracking system which includes high complexity data processing at front-end level. The need for lower interaction with particles forces to reduce as much as possible the material in the tracker, leading to a change in power distribution architecture and cooling system. The higher radiation level requires different approaches for different parts of the tracker. One of the key element for the new tracker is the Pixel-Strip module which integrates all the requirements for the upgrade: pixelated sensor and intelligent particle tracking. The core of this module is the Macro Pixel ASIC, which is the main subject of this Ph.D thesis and is described in details in the next chapter.

Chapter No. 4

A readout chip with momentum discrimination capabilities for pixel detector

This chapter describes the design of a readout ASIC for pixel detectors. This design is dedicated to the requirement of the Pixel-Strip module, described in chapter 3, for the upgrade of the CMS tracking system. Besides the triggered readout architecture used in the current tracker, it includes an event-driven readout path with particle recognition capabilities. The integration in a tracker module which must operate down to -30 °C provides a very limited power budget. This constraint with the high radiation level demand the use of advanced technology, the design of a dedicated readout architecture and the development of low-power techniques.

4.1 Readout electronics requirements

As explained in chapter 3, the upgrade of the CMS Tracker for the High Luminosity LHC requires new readout electronics for the Front-End modules. Focusing on the Pixel-Strip module design, the concept of readout electronics for particle tracking strongly evolved from the present tracker. The combination of a strip and a pixel sensors requires the design of two different chips. The Macro Pixel ASIC reads out the pixel sensor, while the Short Strip ASIC reads out the strip one. The main challenge for this system is the requirement of participating to the Level 1 event reconstruction. It entails the definition of an algorithm for Stub Finding which correlates the strip and pixel signals, and the development of a readout architecture to provide the found stubs at every event to the experiment back-end. This trigger-less readout path, called Trigger path, is a dominant part of the design in terms of power consumption due to the continuous data processing and transmission.

Nevertheless, the front-end electronics preserve the triggered readout for the entire event as the current tracker, which is called L1 data path, but its performances must be improved:

- **Level-1 Latency:** The latency between the data acquisition and the trigger is fixed. In the SST, it was 4 μs, while, in the upgrade it can increase up to 12.8 μs. This change impacts strongly the storage capability available on the front-end modules, which is already larger due to the pixelated sensor. For these reasons, Level-1 memories become another dominant contribution to the power consumption of the design.
- **Level-1 Rate:** The rate of Level 1 trigger should increase from 100 KHz to 750 kHz. This specification together with the higher granularity impact the size of the bandwidth needed for the on-module communication and between front-end modules and CMS back-end, forcing the use of zero suppression technique and the choice of a binary readout.

The work presented in this chapter concerns mainly the Macro Pixel ASIC, which is the processing core of the module. This chip is responsible for the storage and encoding of L1 data as well as for the correlation of the detector signals to discriminate particles

and generate high-p_T information. Besides these digital functionalities, it contains the front-end electronics to readout the pixel sensor.

4.1.1 Pixel Sensor specifications

The specifications about the pixel sensor are summarized in table 4.1.

Parameter	Value
Sensor type	Si n-in-p
Sensor thickness	$200\,\mu$m
Nominal Signal	16000 e-
Active area	$\sim 5\,\mathrm{x}\,10\,\mathrm{cm}^2$

TABLE 4.1: Sensor and ASIC dimensions specifications

The choice of the sensor polarity has been defined in the CMS technical proposal [21]. The electron read-out sensor (n-in-p) shows higher charge collection and robustness in terms of high field effect after irradiation than p-in-n sensor. The choice of thin sensors, namely $200\,\mu$m, reduces leakage current and material in the tracking volume. Such a sensor provides a nominal signal of $16000\,\mathrm{e}^-$ to the front-end electronics. The requirement for the minimum threshold is set around $3000\,\mathrm{e}^-$, which provides the constraint of an intrinsic noise lower than $500\,\mathrm{e}^-$. However, for the front-end electronics as for the entire design, the constrain which makes it an engineering challenge is the limited power density.

4.1.2 Power requirements

The design of the Pixel-Strip module needs information about thermal performance and mechanical deformation while being cooled to low temperature which are obtained with the Finite Element Analysis (FEA) [26]. FEA models include the module geometry, its support and cooling structure, as well as a realistic description of the power dissipation of the front-end electronics and of the sensors. The power density considered for the readout chips is $100\,\mathrm{mW/cm}^2$. Considering only the MPA, the power density requirement is $< 70\,\mathrm{mw/cm}^2$, which provides a power budget per chip of $200\,\mathrm{mW}$.

The readout electronics are the main contribution to the power consumption of the full module (5 W of the 8 W estimated), hence they represents the main load for the power converters and for the cooling system, which are optimized for the estimated load. Any excess in power consumption of the electronics would overload the power converters and the cooling structure, increasing the temperature of the module. Such a problem must be strictly avoided because the sensor must be kept below a temperature around - 20 °C to avoid breakdown or thermal runaway of the sensor. These are the reasons why it is very important to fulfil the consumption requirements. In order to address all the open questions, the first step is the study about which technology is more suitable for the project.

4.2 The 65 nm CMOS technology

The complex logic and the large storage capabilities required by the design would suggest to move to a very down-scaled technology in order to reduce the power consumption. Commercially available CMOS technologies include deep sub-micron nodes down to 28 nm, but the decision on the technology concerns also the project organization. The use of a technology supported by CERN is fundamental to access software and tool support, design courses, legal and administrative issues.

Many projects at CERN uses a commercial 130 nm CMOS technology. This technology has already been fully characterized also for radiation effects. Furthermore, the cost for prototyping can be shared with other projects. The main problem is the power consumption. The large memories and the complex logic would requires a too large amount of current in such a technology respect the given power density constraint.

The other available choice is a commercial 65 nm with low-power feature. The down-scaling of the technology provides a factor 2 gain in the power consumption of the memories and digital logic. The radiation tests show promising results for the CMS Outer Tracker dose. The technology is fully supported by CERN and many project are now starting to work with this node. Consequently, also in this technology the prototyping cost can be reduced sharing the cost among projects.

Going to deep downscaled technologies would ensure an even lower power consumption, but it would strongly increase the prototyping cost, require a new campaign for radiation

induced effect studies and would make the organization of the project more difficult due to the absence of support inside CERN. This is the reason why the chosen technology is the 65 nm node.

This 65 nm low-power CMOS technology was developed for logic and mixed-signal/RF circuits, and allows multiple supply voltages for core and I/O. Its nominal supply voltage is 1.2 V and it features a high-resistivity epitaxial substrate process, shallow trench isolation (STI), two gate oxide options (1.2 and 2.5 V), nickel-silicided low-resistance n+ and p+ polysilicon and diffusion areas. Its device options contain nMOS and pMOS with several different threshold values. High-voltage 5V-drain-tolerant devices are optional. The backend offers 3 to 9 copper metal layers for interconnection plus 1 top aluminum layer for wire-bond/flip-chip pad, pad redistribution layer and laser fuses. Low-k dielectric is used as inter-metal insulator in thin metal layers.

4.3 Macro Pixel ASIC Architecture

Once the technology has been chosen, the design starts from the definition of a chip architecture which can fit the specification summarized in table 4.2.

Parameter	Value
Acquisition frequency (BX)	40 MHz
Acquisition type	Continuous
Acquisition mode	Binary
Readout mode	Event-driven and triggered
Data type	Sparsified coordinates and parameter
Triggered data latency	$\leq 12.8\,\mu$s
Trigger rate	750 KHz
Transmission frequency	320 MHz
Power budget	$\sim$70 mW/cm^2

TABLE 4.2: Data Readout specifications

The block diagram explaining the architecture proposed in the following pages is shown in figure 4.1. The front-end electronics read out the silicon detectors. It implements a binary readout where at each clock cycle the output goes to 1 if the signal from the

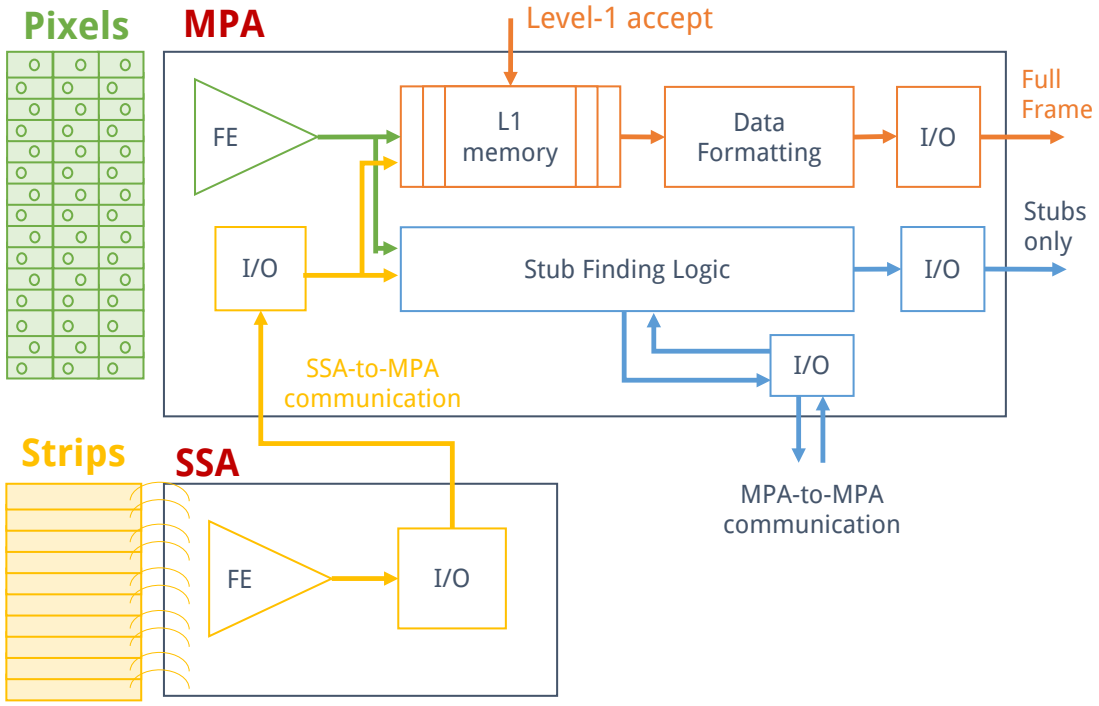


FIGURE 4.1: Pixel and Strip Data Block diagram.

sensor is above threshold and to 0 if not. The Short Strip ASIC (SSA) contains the front-end electronics for the strip sensor, while the Macro Pixel ASIC (MPA) the one for the pixel sensor. The data from the two front-end electronics are processed in the MPA through two readout paths:

- **L1 data path:** It stores the full event information and sends them out if receives a L1 trigger. The hit information from the two detectors (Pixel and Strip) is stored without any data reduction in the L1 memories for the duration of the L1 latency. Upon arrival of a L1 trigger, the event is processed, the hits are encoded with position and width of the hit clusters, and sent in output.
- **Trigger path:** It receives the same input of the L1 data block and, synchronously with the 40 MHz bunch crossing frequency, correlates the data. Its digital logic looks for coincidences within a narrow geometrical angle between pixel and strip clusters in order to generate and encode the high-pT information needed for the event reconstruction in the L1 trigger generation.

Concerning communication, the continuous operation requires a chip-to-chip data transmission synchronous with the BX frequency. Each Bunch Crossing the SSA provides the

strip data to the MPA and the MPA needs also some information from the neighbour MPAs in order to reconstruct the particles passing between two MPAs.

4.3.1 Dimensions and connectivity

Besides functionalities, the other specifications which drive the design are the physical constraints. The pixel segmentation is constrained by the granularity requirements of the upgrade, while the ASIC size is limited by the requirement of planarity for the module assembly. Very large ASICs show some bending which makes very difficult the assembly of multi-chip modules. Consequently, given a pixel of 100 μm x 1446 μm, the pixel array per MPA contains 16 rows and 120 columns. With these dimensions, the large size of the pixel sensor requires the use of 16 MPAs for reading out a single sensor.

Parameter	Value
Pixel Array	120 x 16
Pixel Size	100 μm x 1446 μm
ASIC length	25 mm
ASIC width	11.9 mm
ASIC thickness	$\sim$250 μm

TABLE 4.3: Sensor and ASIC dimensions specifications

The dicing precision is usually in the order of tens of μm, so a spacing of 100 μm divides two adjacent MPAs. Since the pixel sensor does not have gaps, two channels are not connected to the readout electronics due to the spacing between the chips and the margin at the edge of them. In order to not lose any channel the unconnected pixels are shorted to the neighbour in the sensor layout, so at the edge of each pixel row there is a pixel with double width. A simple representation is shown in figure 4.2.

The readout chip connects with the sensor through standard bump bonding technique. In order to keep bump bonding pitch large enough to avoid the use of high density technique, the bonding pad are staggered with a vertical spacing of 200 μm and a resulting horizontal spacing of 200 μm. Such a choice limits also the cost of the bonding process. A line of bump bonding pad at the bottom of the pixel array provides the common ground connection between the pixel sensor and the ASIC.

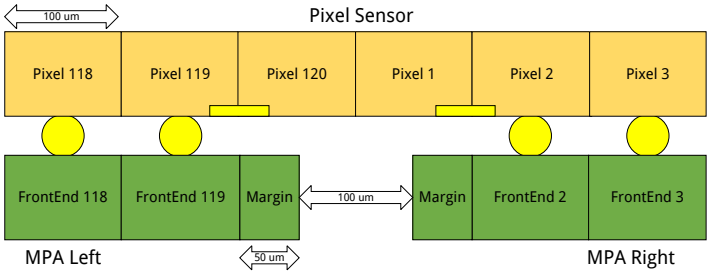


FIGURE 4.2: The figure shows the spacing between two adjacent MPA. The yellow circles represent the bump bonding connections, while the small yellow rectangles between pixel 119 and pixel 120 represents the connection on the sensor layout.

The chip-to-chip communication exploits differential lines on the front-end hybrid. For the SSA-to-MPA links, the SSA is wire-bonded to the hybrid as well as the MPA for the MPA-to-CIC links. The MPA-to-MPA links could exploit a chip-to-chip direct wire-bond connection. The choice of single-in-line wire-bond pads allows a simple and cheap wedge bonding technique. The pitch is 100 μm, but it could be lowered down to 70 μm, if needed, without impact on the bonding technique.

4.3.2 Preliminary power estimation

After the general functionalities and the dimensions have been defined, the first step of the development consists in a preliminary estimation of the power consumption. The power budget for the MPA, about 200 mW, has been equally divided in the three main parts of the design:

- **Analog front-end:** The power allocated for the front-end electronics corresponds to a consumption of 30μA/channel including also the bias structures, which have a minor contribution to the power consumption.
- **L1 memory:** The MPA stores the full event, $\sim$2000 bits, for 12.8 μs. Considering the bunch crossing frequency of 40 MHz, the total storage capability is of 512 words of $\sim$2000 bits. The writing operation dominates the power consumption since it runs at 40 Mhz respect the 750 KHz of the reading operation. Consequently, its consumption must be limited to less than 1.5 mW/MHz.
- **Readout paths, I/Os, clock distribution:** The remaining power is allocated for the data processing, the clock distribution and the I/Os. Several studies have

been carried out to optimize these three components making possible the integration of all the required functionalities. Scalable Low-Voltage Signaling (SLVS) has been investigate to limit the power consumption of the I/Os, while several strategies for the clock distribution have been investigated since, due to the large size of the pixel array, a typical distribution can consumes up to 50 mW.

Due to the large chip area of $\sim 3\,\mathrm{cm}^2$, another non-negligible contribution to the power consumption is the data transport on-chip, i.e. the power needed to move the data from the pixel array to the periphery and from the pixel cells to the memories. This is the reason why the floorplan of the ASIC becomes of fundamental importance, and in the following section not only algorithms will be described, but also the spatial placement of the different blocks will be covered.

4.3.3 Floorplan

As shown in figure 4.3, the area of the chip is divided in pixel array and periphery. The pixel array connects to 16 rows and 118 columns of pixels on the sensor through bump bonding connection. The pixel width is $100\,\mu$m, and at the edges of the matrix the die extends of only $50\,\mu$m in order to leave $100\,\mu$m among MPAs in multi-chip assemblies. Consequently, the width of the chip is 11.9 mm. The pixel length is 1.446 mm, while the full chip has a length of 25 mm, so that the periphery length is 1.864 mm.

Pixel Matrix. In a general hybrid pixel detector ASIC, the analog front-end electronics is integrated in a single pixel, while the synchronization logic and the digital logic is shared among the pixels in the same region. Common structures are usually called pixel region and extend in the two directions. The pixels or pixel regions are grouped into pixel columns and the number of horizontal connections is very limited between the columns.

In the MPA the single pixel integrates the analog front-end and share the synchronization logic, the digital logic and buffers. But, unlike general HPD, the pixel regions extend only along the x-direction due to the high aspect ratio. These pixel regions are grouped together into a pixel row. This architecture is peculiar for pixel readout ASIC and requires an important effort to determine the best architecture for data and signal distribution.

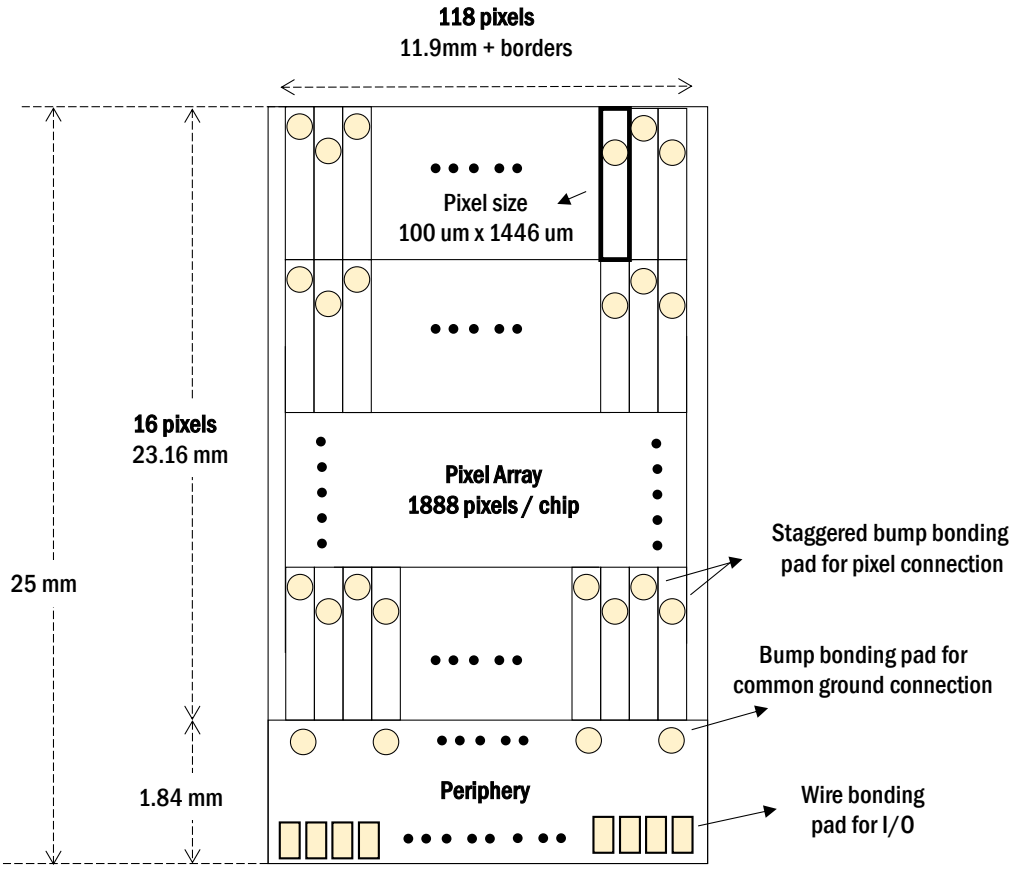


FIGURE 4.3: Structure and the dimensions of the MPA.

The architecture of a pixel row is schematically shown in figure 4.4. The analog circuitry is integrated in the single pixel. Due to the staggered distribution of the bumps, the layout of the front-end is different between left and right pixels, which can provide slightly different noise performance. Analog front-end receives signals from the sensor and, after the dead-time associated with analog processing of the signal, the output is synchronized with the clock and is transmitted to the pixel row block through row digital bus. The output data from the row block are sent to the periphery block through column digital bus which cross the analog part in reserved lanes.

Periphery. The periphery area extends outside the sensor area for the placement of input and output pads. It contains biasing circuitry, global configuration, and digital logic for data processing. Digital-to-Analog Converters (DACs) are used for the global biasing of the analog front-end circutry. Bandgap reference circuits provide a almost temperature and radiation independent voltage reference. Clock divider and Delay Locked Loops

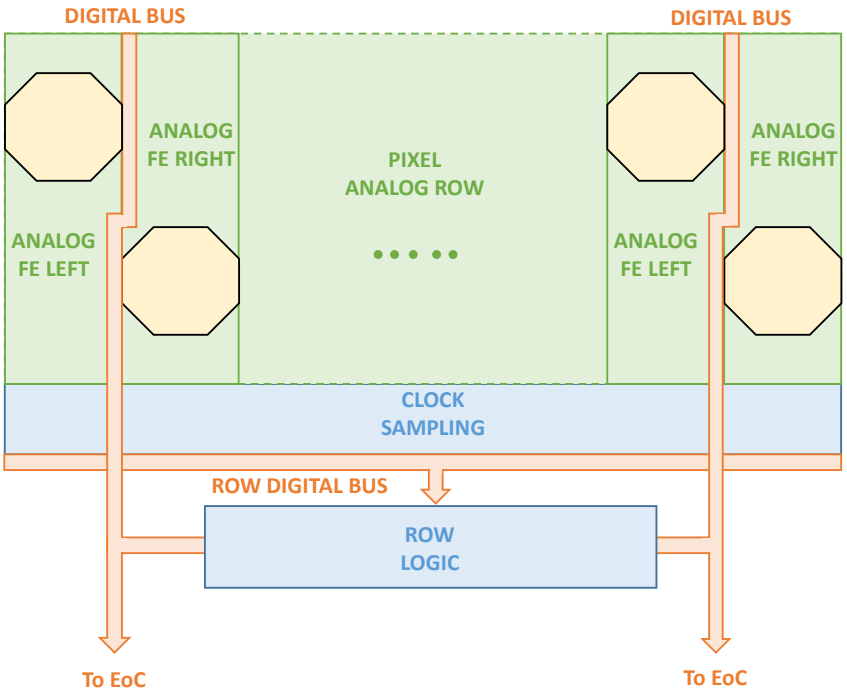


FIGURE 4.4: Schematic representation of the Pixel Row architecture.

(DLL) allows slow clock generation and de-skewing. At the bottom of the periphery, the wirebond pads provide the I/O connectivity.

The architecture just presented is based on a binary readout which necessitates of a timing reference distributed to the entire pixel matrix. This reference is the 40 MHz clock which is synchrnous with the frequency of the collider. The large size of the pixel matrix and the requirement of a skew lower than < 1 ns make compulsory an optimization of the distribution.

4.4 Clock Distribution

A simple architecture based on column clock distribution, where each pixel column includes a clock line, would consume a large amount of power (more than 50 mW, to be compared with the total power budget, of the order of 200 mW), due to the large number of columns. The proposed solution is instead based on a row distribution scheme where a central buffer column distributes the clock along the 24 mm pixel matrix and one clock line per row distributes the clock to the 120 pixel cells in the row, as shown in figure 4.5.

A study of the optimum number and dimension of repeaters to be placed on the central column has been carried out by Gaioni et Al. [27]. The implementation based on CMOS buffers supplied with a reduced power supply voltage has been investigated and evaluated in terms of total power consumption (including the contribution from receivers and column repeaters) and the maximum skew between pixels (i.e. skew between the so-called first pixel and last pixel in figure 4.5). A comparison with conventional clock distribution circuits based on full swing buffers, supplied with 1.2 V, has been carried out as well. Figure 4.6 shows the solution where both the transmitters and the receivers are implemented by means of standard CMOS buffers supplied with a reduced supply voltage VDDL (equal to 800 mV in the circuit simulations). The figure reports the nMOS and pMOS transistor size as a multiple of a unit transistor. The best solution

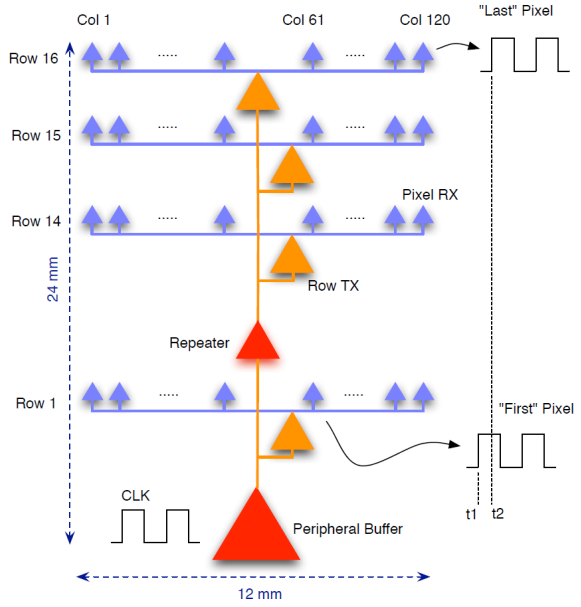


FIGURE 4.5: Row-based clock distribution architecture

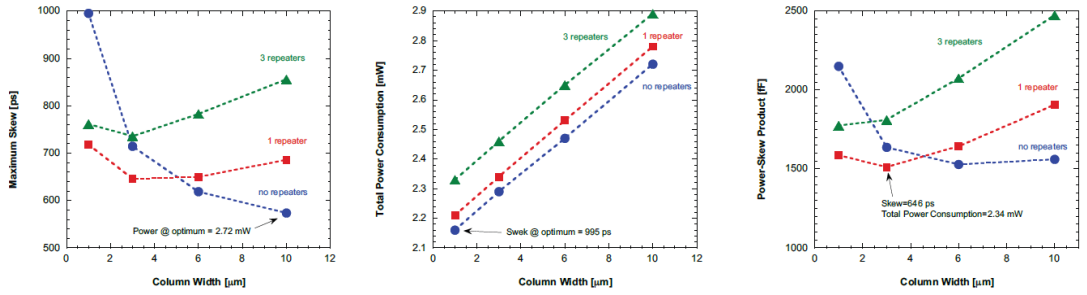


FIGURE 4.6: Maximum skew (left), total power consumption (centre) and power-skew product (right) as a function of the column width.

	Average slew [ps]	Total Powerv[mW]	PSP [fJ]
0.8 V CMOS	646	2.34	1512
1.2 V CMOS	441	5.41	2386

TABLE 4.4: Simulation results for different solutions for the clock distribution circuits of the MPA.

in terms of skew is obtained when no repeaters are added in the column, and with a column width equal to 10 μm. In this condition, the total power consumption is equal to 2.72 mW. On the other hand, the best solution in terms of power is obtained, once again, in the case of no repeaters added in the column, with a width equal to 1 μm. In this condition, the skew is close to 1 ns. Considering the power-skew product (PSP) as a figure of merit to be minimized, an optimum value is obtained in the case of 1 repeater, with a column width equal to 3 μm. In this optimum condition, the skew is equal to 646 ps with a total power consumption of 2.34 mW. Table 4.4 gathers such data and provides a comparison with a solution based on standard CMOS driver supplied with 1.2 V.

These results proves the advantages of the proposed architecture in terms of power consumption and the feasibility in terms of skew. Indeed, also in the case of a reduced power supply, the skew is < 1 ns which is the maximum allowed skew for the system. Once the feasibility of a clock distribution for the binary readout within the power and skew constraints has been demonstrated, the next step is the design of a front-end electronics which amplifies and sample the signal from the sensor.

4.5 Front-end electronics

The front-end circuitry is common for Trigger and L1 data path and includes the analog signal processing, the sampling stage which acts as interface with the digital logic and some testing features.

4.5.1 Analog front-end

The full analog chain, shown in figure 4.7, is composed by a preamplifier, a shaper and a discriminator. The preamplifier is built with a buffered cascode loaded with a

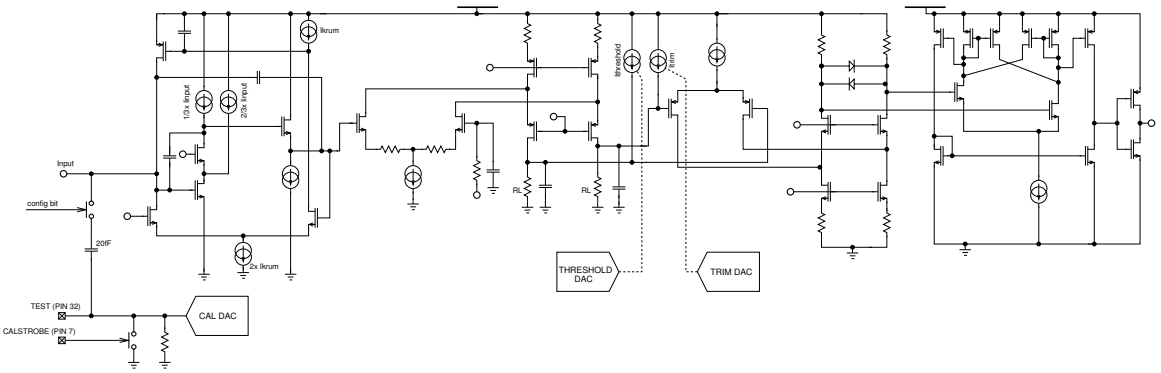


FIGURE 4.7: MPA Analog schematic.

degenerated PMOS cascode current source and enclosed with a Krummenacher feedback [28] providing leakage compensation for the n-on-p silicon sensors up to 200 nA. An extra current source, directly supplying the input transistor, provides the extra boosting of the bandwidth and the minimization of the noise contribution from the active loads.

The second stage, working as an amplifier/integrator, and the threshold interface are built with a differential folded cascode loaded with resistors. The common threshold for the discriminator is provided by high impedance current source mirroring the output current from an 8-bit mutual DAC and sourcing it to one of the load resistors which produces a DC voltage imbalance. The local per-pixel 5-bit DAC is connected to the second load resistor which provides the equalization of the discriminators offset spread.

4.5.2 Digital front-end

Figure 4.8 shows the block acting as interface between the analog front-end and the digital world. It synchronises the pulse from the discriminator with the acquisition clock, identifying the Time of Arrival (ToA) with the associated Bunch Crossing Identification Data (BX-ID).

When the discriminator in the front-end senses a signal above threshold, it propagates a square pulse which is detected by a flip-flop, acting as edge detector. A second flip-flop samples the edge detector output with the rising edge of the 40 MHz sampling clock. The result is a synchronous pulse with one clock cycle duration for each pulse from the analog front-end. The same signal is also used to reset the edge detector on the falling edge of the 40 MHz sampling clock.

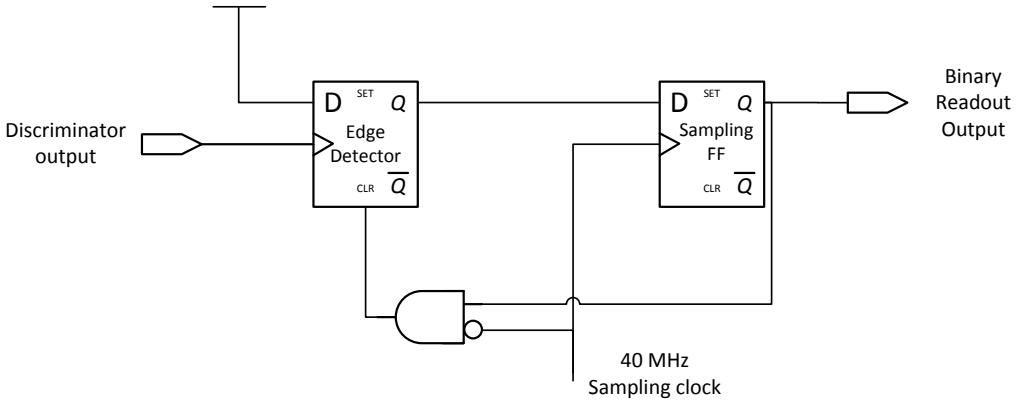


FIGURE 4.8: Binary readout schematic.

The described front-end was prototyped in the first MPA prototype described in chapter 5.

4.6 Trigger Path

The signal from the front-end electronics are transmitted to the Pixel Row region which contains circuitry for both readout path. First, the Trigger path is considered, and, in the next section, the L1 data path will be discussed. The main challenge for the Trigger path is the discrimination of particle based on their transverse momentum. The algorithm implementing this functionality is called Stub Finding.

4.6.1 Stub Finding algorithm

The Trigger path provides the high p_T information to the Level 1 Tracking by sending out the position of the stubs which have been found with an estimation of their transverse momentum. The position is summarized with the coordinates of the point of incidence on the pixel sensor, while the momentum with the bending angle in the r-phi plane of the particle (see figure 4.9). The Stub Finding algorithm discriminates the particles based on the bending value: if it is lower than a given threshold the particle is accepted, and the information about the corresponding stub is sent to the L1 Tracking.

In order to build a stub, the MPA must process the hit informations from both pixel and strip sensors. Since the front-end block implements a binary readout, the trigger

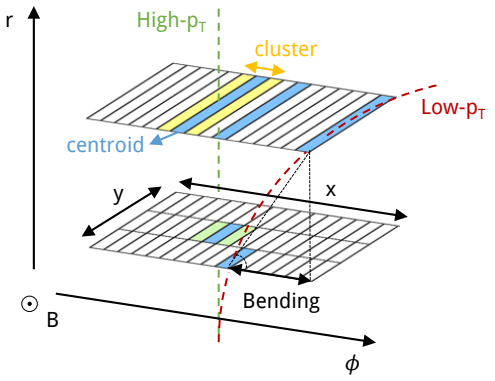


FIGURE 4.9: Bending calculation in a Pixel-Strip module

path receives a binary matrix of 120 x 16 pixels and a binary vector of 120 x 1 strips. In output it provides the encoded position and bending of the stubs found. The following paragraphs describe the main steps of the data processing:

Clustering and centroid extraction (Clustering): A cluster is a group of adjacent hits. It can be generated due to cross-coupling between adjacent pixels, by an high-p_T particle passing through two adjacent pixels, by a not interesting particle as a very low-p_T or a secondary particle (see figure 4.10). In the first two cases, the algorithm must reduce the cluster to his geometric centre, called centroid, which should correspond with the real incidence point of the particle. In the other cases, the cluster must be discarded because it is generated by a not interesting particle. For this reason, on the r-phi plane (pitch = 100 μm), clusters wider than a programmable threshold are rejected, while the centroids of accepted clusters are calculated. The width of the accepted cluster depends on the coupling between pixels. Accepting larger clusters will reduce the data reduction capability of the clustering step, so the cross coupling among pixels must be reduced as much as possible.

On the y-axis (pitch $\sim$ 1.5 mm), the cross-coupling is excluded. Hence, clusters wider than 2 pixels are rejected, because they comes from not-interesting particles. For 2-pixels wide clusters, the coordinate of the pixel closer to the periphery is chosen as centroid, since they could be generated from a high-p_T particle passing in between two pixels.

Offset correction: As shown in figure 4.11, approximating the ideally cylindrical geometry of the tracker barrel with sensors that are actually planar provides an offset at

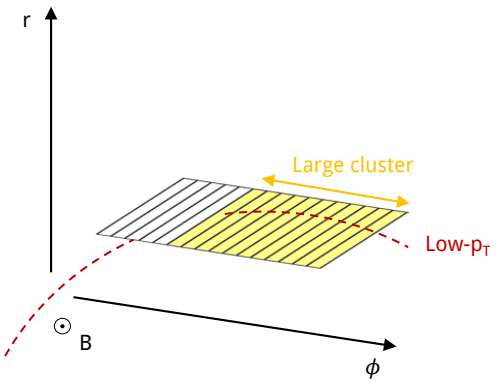


FIGURE 4.10: Large cluster caused by very low-p$_T$ particle.

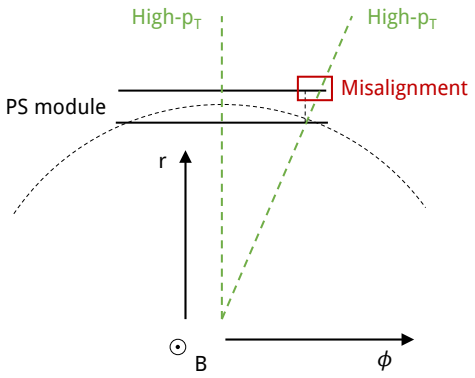


FIGURE 4.11: Sketch of the misaligment caused by approximating the cylindrical geometry of the tracker with planar sensor.

the edges of each module. The value of the offset depends on the distance from the vertex and from the distance between the two sensors. So, a programmable shift must be applied to one of the two layers depending on their module coordinates. The Δ_x offset for a generic position x can be easily computed through triangles similarity:

$$\Delta_x = x(1 - \frac{R}{R+\epsilon}) \tag{4.1}$$

where R is the radius of the innermost layer and ϵ the distance between the 2 silicon sensors. The position x is defined as the distance from the centre of the module along the ϕ axis.

Correlation: The core of the stub finding algorithm is the correlation between the two sensor layers. For each pair of pixel and strip centroid, the distance between their

x-coordinates is computed. If this value is lower than a given threshold a stub is found. If the stub is induced by a primary particle, this distance is related to the transverse momentum by the following relation:

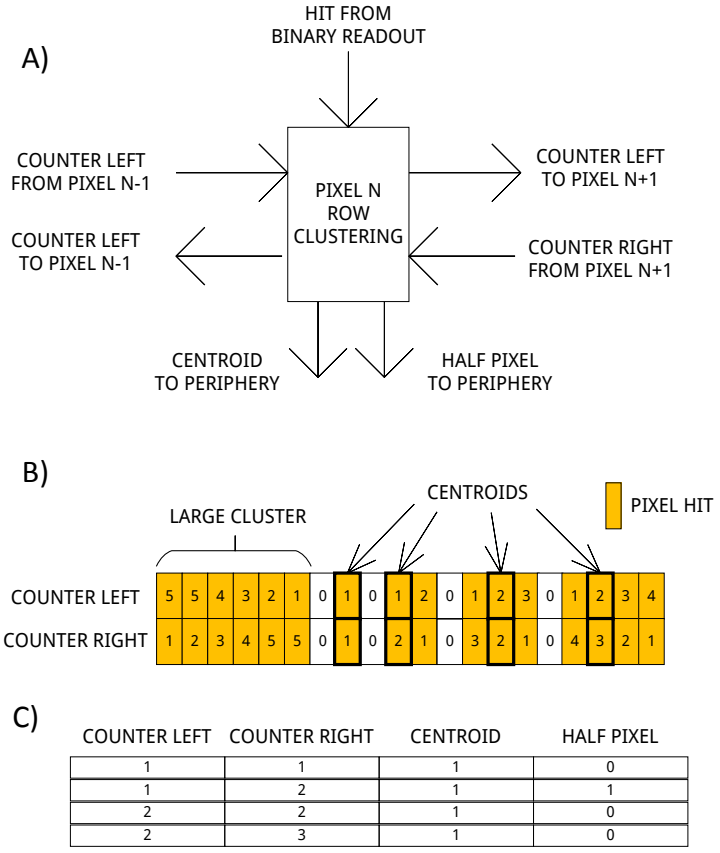
$$p_T = \frac{\epsilon B_C R}{2\delta_\theta}\sqrt{1+\frac{\delta_\theta^2}{\epsilon^2}} \tag{4.2}$$

where B is the magnetic field, c the speed of light, R is the radius of the innermost cluster, ϵ the distance between the 2 silicon sensors, and δ_θ the angular distance between the two cluster centroids. In pratice, δ_θ is the distance computed between the two centroids and it is called Stub Bend. The maximum measurable bend value, corresponding to a stub p_T limit of 2GeV/c, depends on the layer radius.

4.6.2 Clustering and centroid extraction implementation

In the following discussion, the technique studied, and published in [29], for the implementation of the clustering and centroid extraction algorithm will be presented. The main parameters for the analysis are the data reduction efficiency, data losses and power consumption.

The Clustering implementation requires an horizontal (x-axis) communication between pixels as shown in figure 4.12. Each pixel block receives the output of the front-end electronics and, considering pixel *n*, it also receives the left counter from pixel *n-1* and the right counter from pixel *n+1*. The counters contain the number of adjacent hits on the two sides and saturate at the maximum allowed size of the cluster + 1. If pixel *n* received an hit from the front-end circuitry, it increases the counters and it propagates them on the two directions. If not, it sets to zero the counters on the two directions. In order to generate the centroid, the pixel computes his position inside the cluster from the value of the two counters as shown in the example of figure 4.12. If it is in the geometric centre it sends a centroid signal in output. When the number of adjacent hits is odd, the centroid coordinate is the geometrical centre of the cluster. While, when the number is even, the geometrical centre is between two pixels. In this case, as convention, the pixel on the left of the geometrical centre sends the centroid signal to the periphery. The same pixel sends also a second signal, called "Half Pixel", in output which indicates that the coordinate of the centroid is between that pixel and the next one.



COUNTER LEFT	COUNTER RIGHT	CENTROID	HALF PIXEL
1	1	1	0
1	2	1	1
2	2	1	0
2	3	1	0

FIGURE 4.12: A) Connectivity of Row Pixel Clustering module. B) Examples of centroid extraction and cluster elimination with counters method (max.cluster size = 4). C) Table for centroid and half pixel values calculation from the value of counters.

Pixel Column OR-ing The first approach proposed for the implementation of Clustering and Centroid Extraction by Marchioro in [30] included also a block called Column OR-ing, shown in figure 4.13. It carries out the logic OR of the hits from the front-end electronics in every pixel column. In the case where one column contains more than one hit, the priority goes to the hit closer to the periphery. The pixel matrix is reduced to a strip vector where every hit has an associated row coordinate. This technique reduces the data size to be processed of a factor 16, but it also reduces the efficiency of the Stub Finding algorithm. Indeed, large cluster generated by low p_T particles, secondaries or loopers can hide good cluster generated by high p_T particle. The output of this block is then processed by the Clustering and Centroid extraction module.

Row Pixel Clustering A possible method to mitigate the efficiency loss provided by the pixel column OR-ing consists in moving the pixel Clustering block in the Pixel Row region before or without OR-ing the pixel columns. This solution avoids the transmission

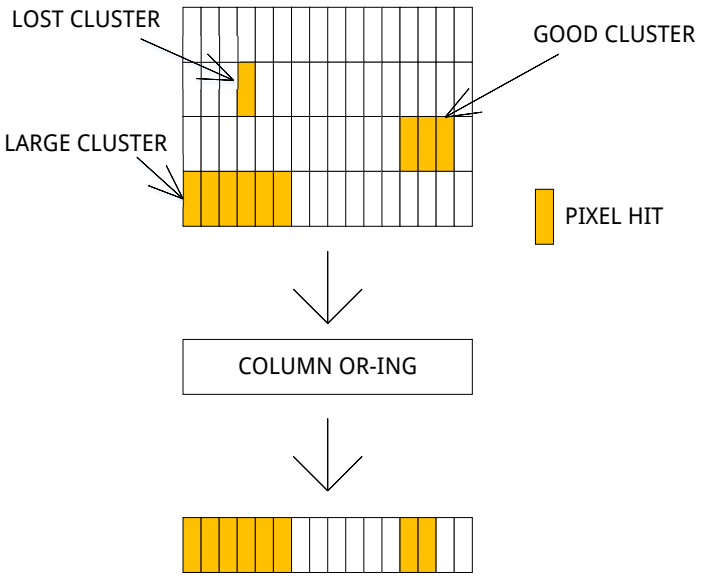


FIGURE 4.13: Column OR-ing example. The figure shows the hiding problem caused by large cluster during the Column OR-ing. A large cluster on the first row of the pixel frame hides the good cluster above it.

of large clusters along the 24 mm long pixel matrix decreasing the amount of data moved to the periphery. Considering an average line length of 1.8 cm from the pixel to the periphery and a line capacitance of 2 pf/cm, the power consumed per hit is:

$$P = 40\,MHz \cdot 2\,pf/cm \cdot 1.8\,cm \cdot 1.2\,V^2 \sim 200\,\mu W \tag{4.3}$$

Therefore, the transport of large cluster to the periphery has a not negligible cost in terms of power, which can be avoided thanks to the Row Pixel Clustering. On the other hand, the pixel clustering logic is placed in every pixel row, increasing the amount of static power. Anyway, this contribution is strongly limited by the use of the 65 nm technology which shows very low leakage currents.

In conclusion, using the Row Pixel Clustering shows advantages in terms of algorithm efficiency and power consumption. The drawback of an increased pixel logic does not impact this design since the pixel density is not a concern. Avoiding the pixel column OR-ing maximizes the efficiency of the algorithm, because it completely excludes the problem of cluster masking. Using it after the Row Pixel Clustering strongly reduces the losses due to cluster masking. Indeed, large clusters have already been discarded by the clustering, therefore, only in the case of two centroids on the same column there is a data loss. The choice about Pixel Column OR-ing is also related with the following

processing step and from their power consumption: if the digital block carrying out the offset correction and correlation operation are able to fit the power budget also without the data reduction from the OR-ing, the latter will be excluded. On the contrary, the Pixel Column OR-ing must be used.

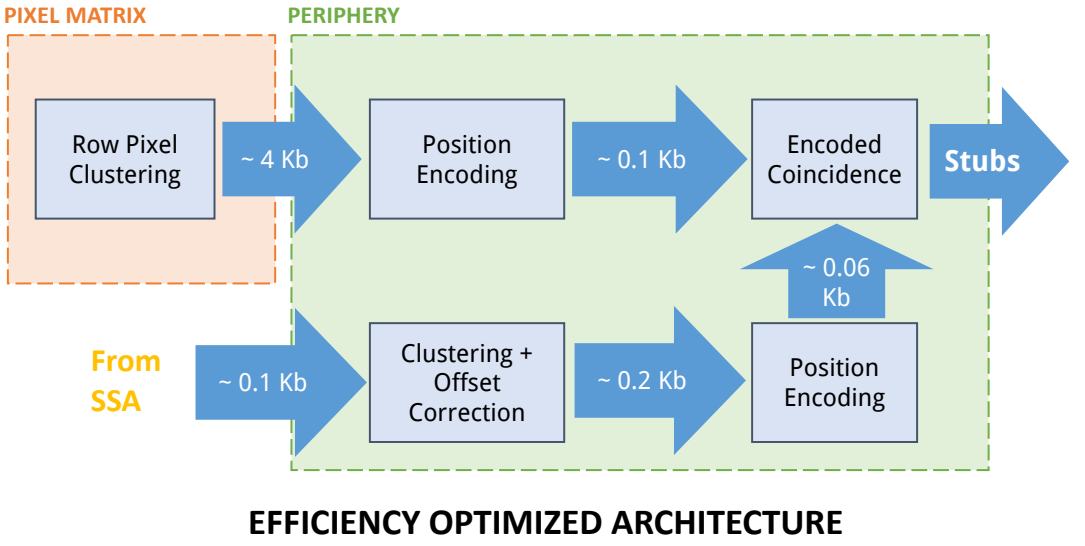
4.6.3 Offset correction and correlation implementation

The step after the clustering and centroid extraction in the Stub Finding algorithm is the offset correction on the strip data followed by the correlation between pixel and strip centroid. The offset correction shifts of a programmable amount the strip data. The implemented shift is in the range of $+/-400\,\mu$m with a precision of $+/-50\,\mu$m. Four offsets can be set, one every 40 strips. These parameters are enough to fit the requirement of the whole tracker.

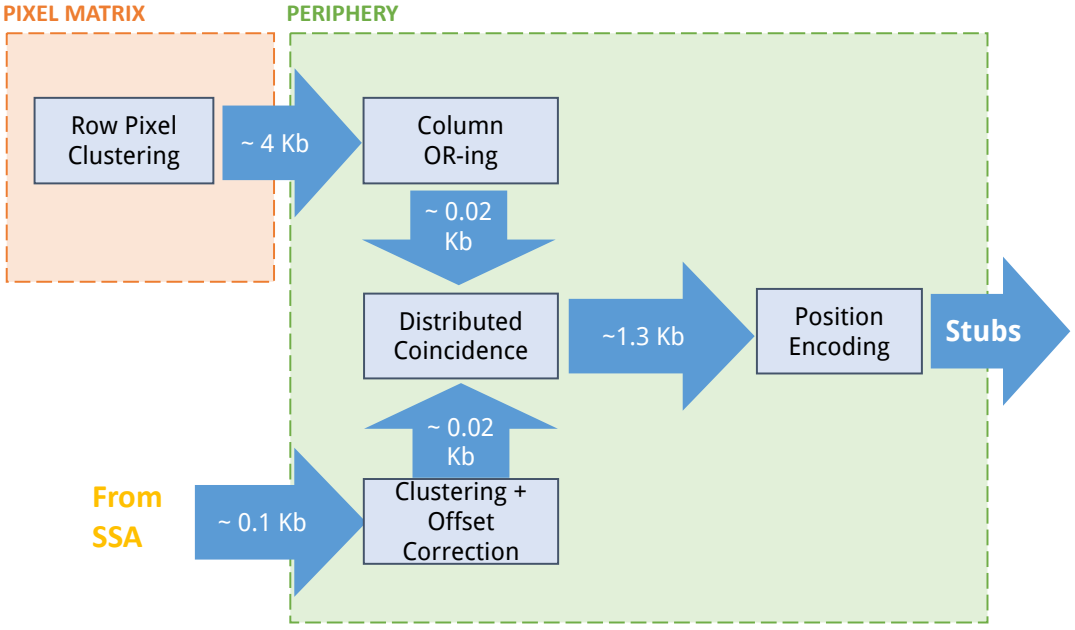
Once the offset is applied, the correlation logic receives pixel and strip centroids and it generates the stub information. The implementation of this logic is influenced by the architecture of the entire Stub Finding logic. Two different architectures, schematically represented in figure 4.14, have been designed:

Distributed Correlation Logic. If the Clustering and Centroid Extraction includes also the Pixel Column OR-ing, the Correlation logic for every valid cluster on the inner sensor looks for a hit in a coincidence window on the outer sensor. If a hit is present within this window, the inner strip is considered a valid stub. This coincidence logic is structured into combinatorial blocks which have as input one channel n from the inner sensor and the channels from n - k to n + k, where k is the maximum allowed bending, from the outer sensor.

Encoded Correlation Logic If the Pixel Column OR-ing is not used, the Correlation logic should process about ∼4000 bits. For this reason, a data encoding step before the Correlation logic has been considered. Including the half pixel resolution and 4 bits for the row number, the pixel position can be encoded on 12 bits, while for the strip 8 bits are sufficient. Limiting the number of centroids that can be encoded to 8 (8 pixel and 8 strip centroids), the amount of data is reduced to 160 bits, obtaining a data compression factor ~ 25. Following this approach, the correlation logic computes the x-position difference between pixel and strip centroids, and if it is within the maximum



EFFICIENCY OPTIMIZED ARCHITECTURE



POWER OPTIMIZED ARCHITECTURE

FIGURE 4.14: Top: Trigger path architecture optimized for efficiency. Bottom: Trigger path architecture optimized for power consumption.

allowed bending it generates a stub. The stub position is defined as the pixel centroid xy-position, while the stub bending is defined as the difference between the centroid x-coordinates. Since the encoders accept up to 8 pixel and strip centroids, 64 cells process each possible combination of pixel and strip centroids. Anyway, the correlation logic limits the output to 2 stubs per pixel centroid. The last step of the Correlation logic orders the stubs giving priority to the lower row and transfers them to the output interface.

Summarizing, two architectures have been developed for the Trigger path. Both includes the Row Pixel clustering, but the one optimized for efficiency does not include the Pixel Column OR-ing and uses the position encoding step before the Correlation logic. On the contrary, the one optimized for power consumption includes the Pixel Column OR-ing and uses the position encoding step after the Correlation logic. The baseline is the version optimized for efficiency, the other solution remains a back-up if the power budget is not fitted with the first architecture.

4.7 Position encoding technique

Before opening the discussion about the L1 data path, it is important to introduce the technique used for position encoding. As explained in the previous paragraph, data encoding is one of the main steps in the Trigger path, since encoder modules are needed for x and y position of the stubs.

For L1 data, since the binary readout provides one bit per pixel, a full MPA+SSA system at each BX generates an amount of data $\sim$2 Kb. Taking into consideration the L1 rate of 750 KHz, the bandwidth needed for an unsparsified L1 Data path to CIC communication is $\sim$1.5 Gbps. Estimating the bandwidth between MPA-CIC around 2 Gbps and considering it should be mainly used for the Trigger path, the unsparsified approach strongly exceed the available bandwidth. For this reason, also the L1 Data path requires a data encoding step.

The encoder designed for the MPA is inspired by the MEPHISTO architecture, described by Fischer in [31] and shown in figure 4.15. A linear scan through 120 channels in 25 ns would require a scan rate of 5 GHz which is difficult to achieve in a low power CMOS

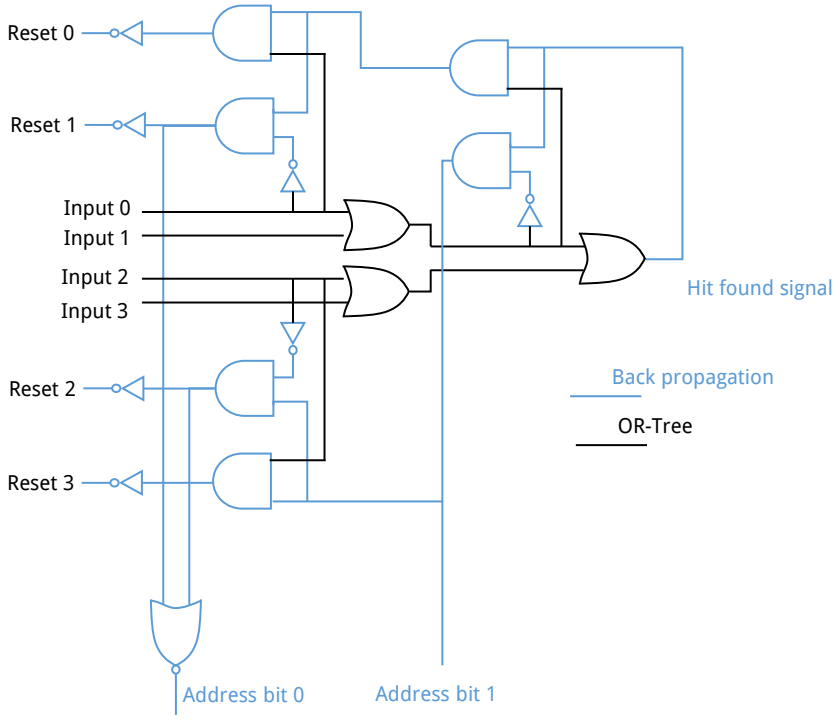


FIGURE 4.15: Logic diagram for a 4 bit MEPHISTO priority encoder. Black lines show the OR-tree, while blue lines indicate the backpropagation blocks. For simplicity, only the back-propagation with uppermost priority is shown.

design. The proposed architecture is instead based on a binary tree of OR gates (OR-tree), with 120 flip flops in input. After the propagation time, the OR-tree provides a hit found signal. On top of the OR-tree, two back-propagation blocks with opposite priority are associated to each OR gate. One back-propagation block sets the address line to 1 if the hit is on the upper input of the OR gate, while the other does the opposite. Consequently, the hit address is automatically coded in binary format in the tree so that no additional address memories are needed. In case of more than two hits, the uppermost and the downmost hits are encoded, and the corresponding input flip-flop is set to 0, so that, in the next clock cycle, the encoder will process the remaining hits.

This encoder has been used in both the readout paths as building block of more complicate architectures, which allow dual clock rate operation or pipelined operations to improve its performance.

4.8 L1 Data path

As it was shown in figure 4.1, the L1 Data path is divided in two functions: event storing and event processing. The L1 memory stores one event per bunch crossing and after the L1 latency, if it receives a L1 trigger signal, it sends the event to the processing, if not, it discards the event. The processing extracts the cluster width of the strip and pixel clusters and encodes the position of the first pixel or strip in the clusters. From the power consumption point of view, the most critical part is the event storing because of the continues memory write cycles at 40 MHz, while the probabilistic memory read operations and the consecutive event processing works at an average frequency of 750 kHz (L1 rate).

The area of almost $3\,cm^2$ makes the floorplan of the chip fundamental to reduce the power consumed in data transport, as in the Trigger path. In the L1 data path less than one event out of 40 is processed, while the remaining events are discarded. Therefore, placing the memories in the pixel region instead of placing them in the periphery would reduce the data transport of more than 40 times. This choice makes the data transport contribution negligible for this readout path.

L1 memories are circular memories which store every event for the L1 latency; in particular one memory per pixel row is foreseen plus one for the strip data, resulting in 17 L1 memories in total. The word size is 128 bits and the depth is defined from the L1 latency: up to a L1 latency of 12.8 μs, a 512 words memory is sufficient.

The two main issues about the memory are radiation tolerance and power consumption which are analysed in the next paragraphs.

4.8.1 A radiation tolerant low power SRAM compiler

The development of a radiation tolerant low power SRAM compiler in 65 nm technology, described by Brouns in [32], is the result of a collaboration between CERN and IMEC institute. The main parameters are summarized in table 4.5.

The design uses only standard-Vt devices, occupies only the 4 lowest levels in the metal stack, and supports simultaneous read and write operations (dual rate operation). TID

Description	Value
Supply	$1.2 \pm 10\%$ V
Frequency	> 80 MHz
Operation type	Dual rate
TID Hardening	> 200 Mrad
LET threshold	> 15 MeV cm^2/mg

TABLE 4.5: SRAM specifications

effects as drive loss and Vt shift are limited avoiding the use of minimal width transistor. NMOS are larger than 200 nm and PMOS than 500 nm. The hardening for SEE of the addressing circuit is done by drive strength, while protection against latch-up is reached by placing p^+ guard bands between n^- regions. These strategies impact the area and the power consumption of the SRAM.

The penalty in area is not a problem for the MPA design, while the power consumption must be carefully evaluated. The characterization of the memory provides the cost per operation and considering write frequency of 40 MHz and a read frequency of 750 KHz the total power consumption for 17 memories in the typical case is ∼110 mW.

Such a value represents more than half of the power budget for the MPA and it is clearly exceeding the requirement of the design. In order to decrease the power consumption one solution is to limit the activity of the memory by using a gating technique as shown in the next paragraph.

4.8.2 Memory gating technique

At each BX the binary readout from the front-end electronics is stored in a memory in the pixel row region. Considering a maximum particle occupancy ∼1 ‰ , a pixel row shows an hit every 8 BX in average, while in the remaining 7 cycles the memory is written with all zeros. The studied memory gating technique, which is shown in figure 4.16, avoids this useless operation by adding a minimum amount of logic. Thanks to the fixed latency (nominally 12.8 μs), the memory can be used as a circular memory, where a bunch crossing counter provides the address. A latency cycle is defined as the time needed to go through all the memory. Another counter, called Latency counter, counts the number of latency cycles. This counter starts from 0 and it is increased every latency

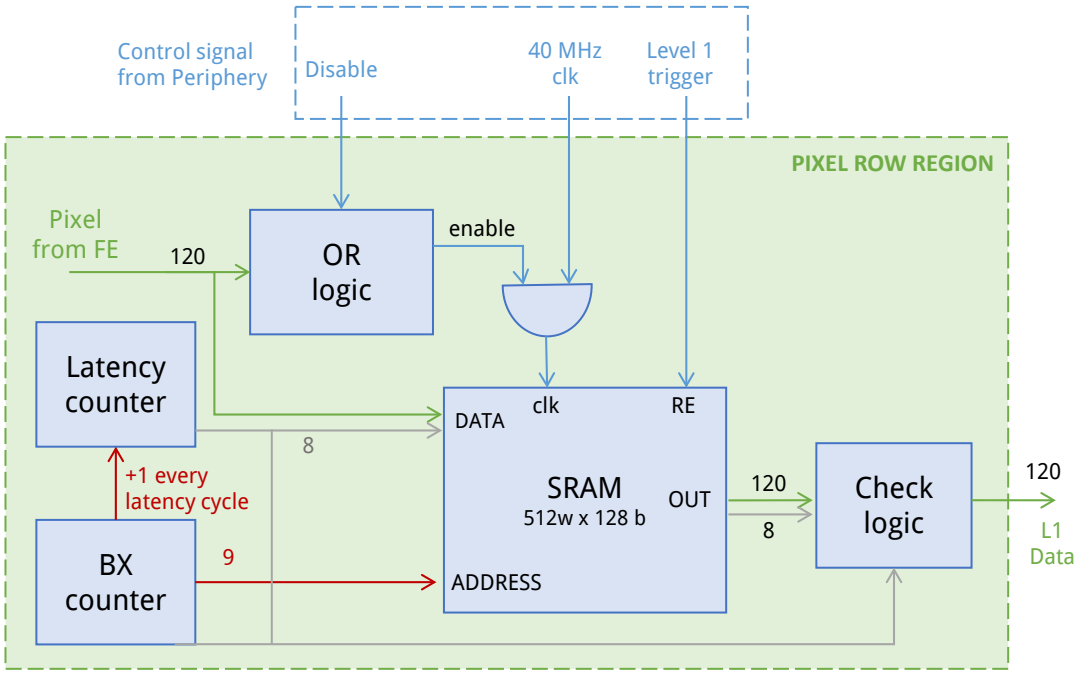


FIGURE 4.16: Memory gating schematic.

cycle i.e when the bunch crossing counter return to 0. Every address in the memory should be over-written every latency cycle if no gating is applied.

Considering the address n during the latency cycle k, the presence of at least one hit in the 120 bits from the front-end electronics is checked with an OR-tree which provides an enable signal. If this signal is high, the event is stored, if the signal is low, the memory is not clocked. With the pixel data, the logic also stores the latency counter value which tags the event. After the latency time, the same address n is read only if a L1 trigger is received. If this is the case, the system must check that the read event is the correct one. Indeed, if in the previous latency cycle the memory was gated, the event stored in address n does not come from the correct latency cycle k but from a previous one. This can be easily checked comparing the current latency counter $k+1$ with the latency counter stored in the memory. If its value is $< k$, the event is discarded and all zeros are sent to the L1 data processing because the event is not the requested one. If it is $= k$, the event is sent to the processing stage.

A latency counter of 8 bits saturates after 256 cycles, which means every $\sim$3.2 ms the counter will restart. Before the counter restarting, the memory must be clean i.e. every address must be written with the last latency counter in order to avoid wrong checks.

For this reason, the OR-logic is disabled with a control signal from the periphery for a full cycle and the memory is written every bunch crossing.

This gating architecture has been designed, placed and rooted to check functionalities and power consumption. The results show the memory gating strongly reduces the power consumption. Assuming this technique is not applied to the strip SRAM and one pixel SRAM is always in cleaning state, the total power consumption simulated in the fast corner is 50 mW. This value fits the initial requirement of < 1.5 mW/MHz and corresponds to a saving $> 50\%$.

4.9 Supply Voltage scaling

The L1 data path with the memories and the Trigger path with the fast data processing consume around half of the power budget in the MPA. One common technique for reducing power is to reduce the supply voltage. For CMOS circuits, the cost of lower supply is lower performance. Scaling the voltage supply increases the delay of a gate, but, if the timing of a circuit is not marginal, the scaling will save power without corrupting the design at a given frequency.

The two main sources of power dissipation in CMOS circuit are static current, which results from resistive paths between power supply and ground, and dynamic power, which results from switching capacitive loads between different voltage levels. For a CMOS gate, the dynamic power is:

$$P = \alpha C V^2 f \tag{4.4}$$

where α is the activity factor, C is the load capacitance, V is the supply voltage and f is the operating frequency. This equation shows as, if the supply is scaled by a factor x, the dynamic power consumption scales by x^2. Consequently, scaling the voltage to 0.8 V provides a gain in power consumption about 55%.

On the other hand, the voltage scaling has a negative effect on the performance of the transistor, which must be studied to prove the feasibility of the design at a lower voltage supply.

4.9.1 Temperature inversion effect

In order to study the delay of a cell is necessary to study the behaviour of its output current. The drain current in a transistor is given by the equation:

$$I_d = \mu_n C_{ox} \frac{W}{L} (V_{GS} - V_t)^2 \tag{4.5}$$

where μ_n is the mobility, C_{ox} the oxide capacitance, W the transistor width, L the transistor length, V_{GS} the gate voltage and V_t the threshold voltage. From equation 4.5, the current is proportional to the mobility of the semiconductor. Lattice vibrations cause the mobility to decrease with increasing temperature. Therefore, the cell delay increases with temperature due to the mobility. The other parameter which changes with temperature is the threshold voltage:

$$V_t(T) = V_{t0} + \alpha_{Vt}(T - T_0) \tag{4.6}$$

where α_{Vt} is a constant variable which reduces the threshold voltage when the temperature increases ($\sim$ -3 mV/°C).

The final drain current depends on the dominating effect at a certain V_{GS}. In advanced technology node as the 65 nm one, the mobility effect is dominating only when the V_{GS} is large. In this case, the difference $(V_{GS} - V_t)$ is almost constant respect the temperature variation of V_t. Instead, if V_{GS} approaches V_t, the variation with temperature of the difference is larger. When the variation of the threshold dominates over the mobility, the cell delay decreases with high temperature and viceversa. This condition is called "temperature inversion effect".

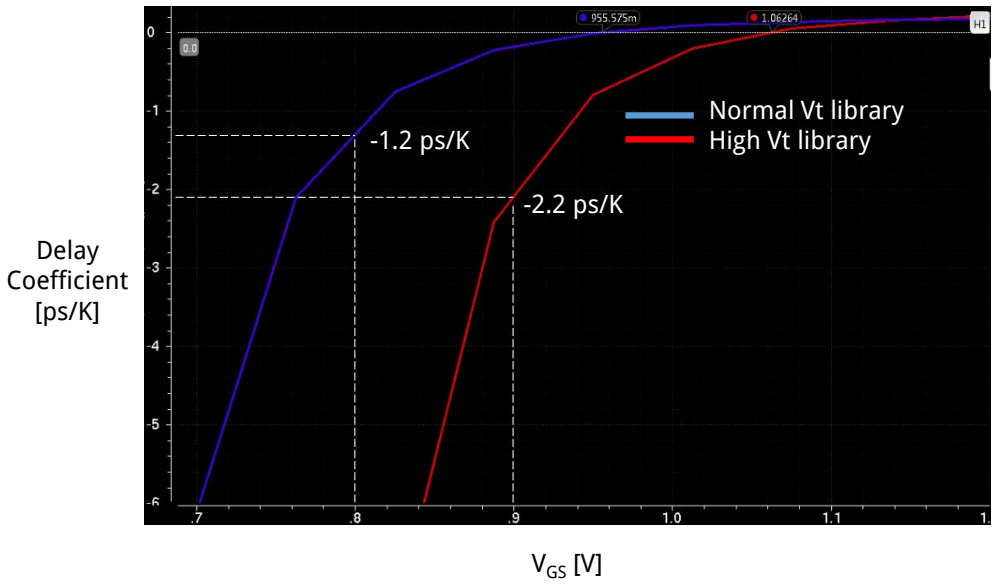


FIGURE 4.17: Delay temperature coefficient variation respect V_{GS}.

Figure 4.17 shows the delay variation coefficient for different supply voltage in the case of high-Vt and normal-Vt transistors. A negative coefficient indicates the temperature inversion effect. As expected, high-Vt transistor degrades more than normal-Vt. In the latter the inversion point is around 0.95 V. Moreover, the coefficient amplitude is higher when the device is in temperature inversion respect the normal effect i.e the variation of the delay with temperature is larger in inversion than in the typical condition.

In the case of the MPA, where the cooling temperature is -30 °C, the temperature inversion effect plays a negative role. For this reason, the 0.8 V option was discarded as well as the use of high-Vt cells at least in the low supply voltage domain. The choice for the supply voltage is orientated towards 1 V, where the normal-Vt transistor are not in temperature inversion. The power saving will be anyway ∼30% respect the nominal voltage of 1.2 V.

One of the main reason to exclude the 0.8 V supply is the performance degradation of the L1 memories when simulated in this condition. SEU resistance of DICE Flip-flop and Signal-to-Noise Margin (SNM) of memory cell are are not affected. On the contrary, the read and write operation timing at 40 MHz is marginal in typical condition (TT corner and 25°C), while it does not fit the specification in the worst case condition (SS corner and -30°C), where the memory is no more functional. Furthermore, due to the long transition, the design will be 4 times more sensitive to Singe Event Transient (SET).

A redesign to resize the buffers would be needed to solve these problems, but it would also increase the power consumption. Consequently, also in the case of the memory, the optimal trade-off between power and performance is found with a supply voltage of 1 V.

4.10 I/O interfaces

Once the architecture for the two readout paths have been developed and different power saving techniques have been investigated, the next step is defining the interfaces of the readout chip with the other components of the front-end module. The periphery includes three different I/Os:

- Analog signals connect to the internal circuits with wire bond pads which include electrostatic discharge (ESD) protection structure.
- Slow digital signals use CERN IP pads, developed by Kremastiotis [33] with I/O buffer powered at 1.2 V. The CERN IP avoids the use of the foundry IP pad which uses 2.5 V transistor with thick oxide. This kind of transistor shows a worse performance with radiation and would require an additional supply.
- High frequency digital signals use Scalable Low-Voltage Signaling (SLVS) pad. This is a well-know and widely used technique to provide a low-power, high-speed I/O interface for point-to-point transmission. The large number of connection requires an optimized design which is described by Traversi [34].

The large bandwidth, the high frequency and the continuous transmission make the differential I/Os an important contribution to the power consumption and require an optimization, not only of the component, but also of the transmission scheme.

4.10.1 High speed communication

Differential I/Os are used for the SSA-MPA, the MPA-MPA and MPA-CIC data transmissions. The connectivity scheme is shown in figure 4.18. The highest available frequency on the PS module is 320 MHz which allows data rates of 320 Mbps or 640 Mbps with a double data rate operation. SSA-MPA and MPA-MPA links are short connections

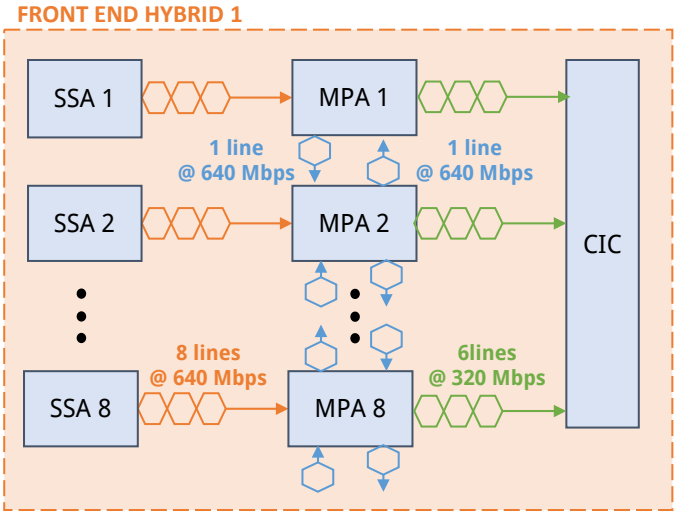


FIGURE 4.18: FE ASICs connectivity scheme.

(< 2 cm), while the MPA-CIC link is longer (< 10 cm). Since the SSA sends unsparsified strip hits to the MPA every BX, the amount of transmitted data is ~ 5 Gbps. In order to minimize the number of links, the choice of 640 Mbps is preferred since allows the use of 8 lines instead than 16. The same choice is made in the MPA-MPA communication where one link per side is sufficient. Concerning the MPA-CIC communication, the 320 Mbps is instead selected to ensure the signal integrity along the longer lines.

Reducing the number of links is an advantage from the point of view of hybrid design and also power consumption. The hybrid design can accommodate six lines from each MPA to the CIC corresponding to a total input bandwidth ~ 15 Gbps per CIC. The proposed SLVS driver shows a power consumption of 2.8 mW with nominal current setting and can operate up to 1.2 Gbps. Thus, choosing the double rate approach in the SSA-MPA and MPA-MPA provides a power saving of 25 mW which corresponds to 15 % of the total power budget. Another method to save power is decrease the current output of the driver. Considering a termination resistance of 100 Ω, the nominal current is 2 mA which provides a differential mode voltage (V_{DM}) of ± 200 mV around the common mode (V_{CM}) of 200 mV. The low-power receiver designed proved to work at 640 Mbps also with a $V_{DM} = \pm 30$ mV, which allows to reduce the current in the driver to the minimum value of 0.5 mA. In conclusion, the power consumption for data transmission on the Macro Pixel ASIC is estimated around 30 mW.

4.11 Simulation Studies

Besides the power estimation, another very important step in the design flow is the verification and the performance evaluation of the proposed architectures. The first check is a verification obtained by injecting randomly generated events in the MPA ASIC functional model at the Register Transfer Level (RTL) and at the synthesised gate level. The results of the verification with random input help designers to improve the model and to find bugs, but does not provide any information about the performance or about the switching activity (directly related to the power consumption) of the chosen architecture. Consequently, also Monte Carlo generated events [35] have been used to evaluate the performance of the Macro Pixel ASIC architecture described in the previous sections.

4.11.1 Simulation with Monte-Carlo events

Monte-Carlo (MC) programs for computer simulation of complex interactions in high-energy particle collisions provide event samples for the entire CMS Tracker. These MC events contain information of all the particles with a p_T larger than 0.1GeV/c like p_T, impact parameter and hit positions. A script translates the format of the MC generated events into input files adapted for the MPA Verilog model and into files which contain the expected output stubs. Running simulations with the MC input files and comparing the obtained output with the expected one, the designer can evaluate the performances of the MPA architecture and also the limitations introduced from the limited bandwidth in the module or from the module to the CMS back-end.

4.11.2 Efficiency analysis

By using the tracker geometry generated by Bianchi et Al.[36], which is shown in figure 3.6, the MPA simulation with MC generated events provides an estimation of the MPA performance in terms of stub finding efficiency in the HL-LHC environment. This efficiency is defined as the percentage of particles with p_T larger than 2 GeV/c which the model detects.

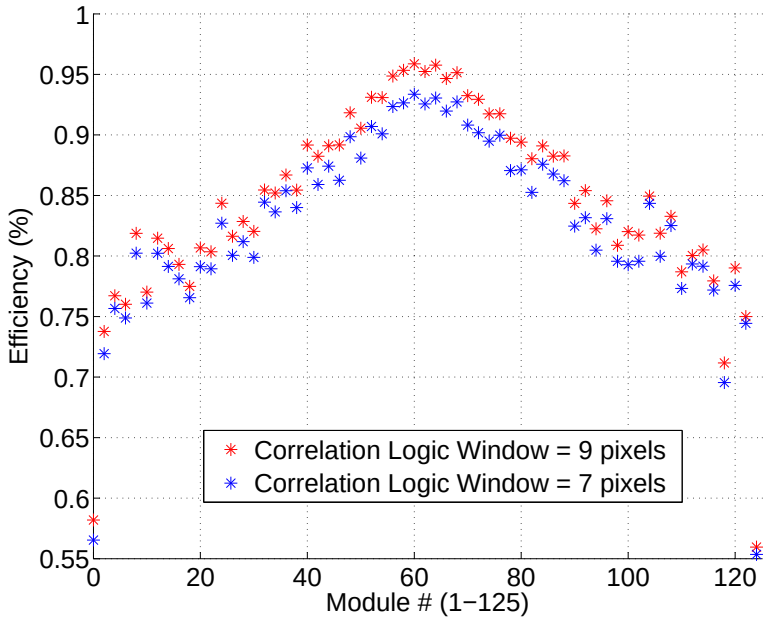


FIGURE 4.19: Stub finding logic efficiency for different module number in Layer 1. Red points represents the stub finding Logic efficiency with a correlation logic window of 9 pixels, while blue points represents the same efficiency with a correlation logic window of 7 pixels.

This analysis allows to study the parameters space of the MPA model. The results for the first barrel layer are shown in figure 4.19: the worst case performance corresponds to the densest environment, which is found in the Layer 1 located at approximately 23 cm from the beam line, where the total stub finding efficiency is around 88.5%. Several simulations with different parameters as correlation window and cluster width provide the comparisons among readout architectures and chip configurations. In figure 4.19, the simulation is repeated with different correlation window dimensions and, decreasing the window size, the total layer efficiency decreases by $\sim 2\%$.

In general, the stub finding efficiency increases to $\sim 95\%$ in the center of the tracker while it decreases at large Z values due to geometrical inefficiencies stemming from the absence of z-communication between the two chip rows in the PS module. This limitation makes impossible to detect the particles crossing two different chip rows in the bottom and top sensors and is the cause of the largest inefficiency. Several solutions are being evaluated to solve this problem, because by solving it the efficiency distribution in Z becomes almost flat, providing an overall efficiency $\sim 96\%$ for the first barrel layer as shown in figure 4.20 (green points).

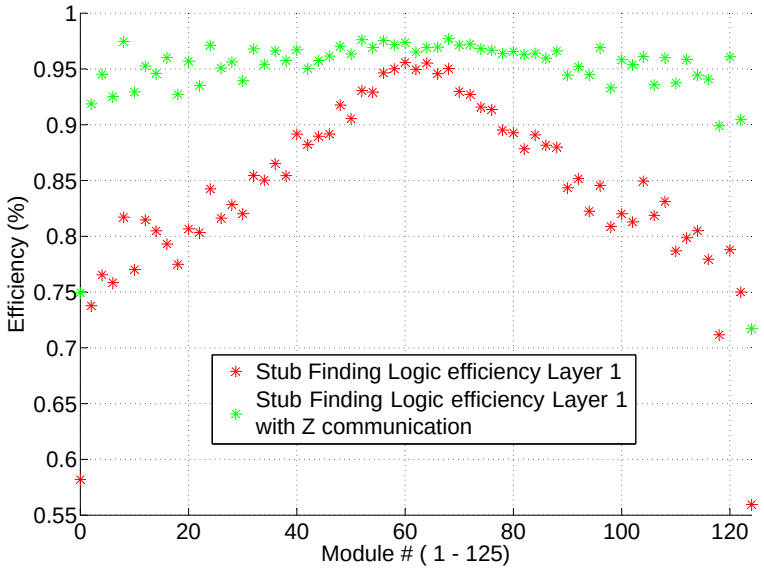
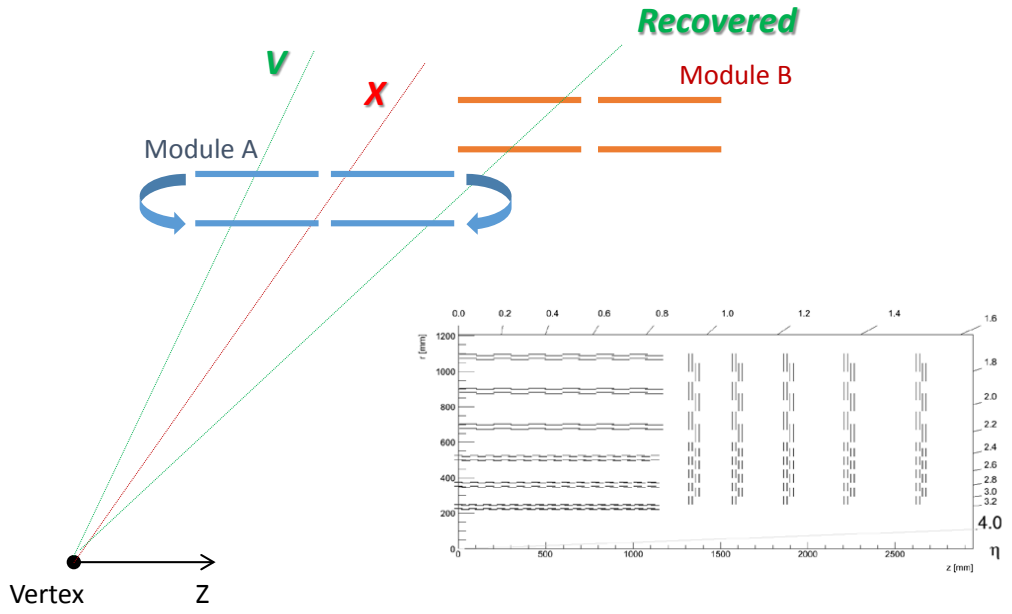


FIGURE 4.20: Stub finding logic efficiency respect to the module number in Layer 1. Module 63 is located at z = 0. Higher and lower module numbers correspond to positions with larger absolute z values, with a maximum z of +/- 1100 mm. The lower efficiencies of module 1 and 125 are artifacts due to the absence of the end-caps in the simulation.

A promising alternative to the introduction of the z-communication is the tilted tracker layout, introduced by Mersi et Al. [37], which allows the PS module in the barrel layer to be always almost perpendicular to the particles. Consequently, the number of particles crossing two different chip rows in the bottom and top sensors will be minimized and the expected efficiency will be the same as the module located at Z=0. The two alternatives are represented in figure 4.21.

Another interesting result verified in the simulation is the amount of fake stubs. Only between 5 and 10 % of the stubs generated by the MPA model correspond actually to a high-p_T particle. The remaining ones are caused by non-interesting particles such as low p_T particles, secondaries or combinatorials. This large amount of fake stubs is filtered out by L1 tracking system in the CMS back-end which excludes them correlating the stubs from different tracker layers. However, all the generated stubs need to be transmitted from the MPA to the CMS back-end before the filtering. Reminding the diagram in figure 3.7, the bandwidth bottlenecks are the communication between MPA and Concentrator IC and the Optical Link between the PS module and the CMS back-end. Consequently, in addition to the MPA inefficiencies shown before, also the limited bandwidth for stub transmission can introduce inefficiencies in the stub finding process



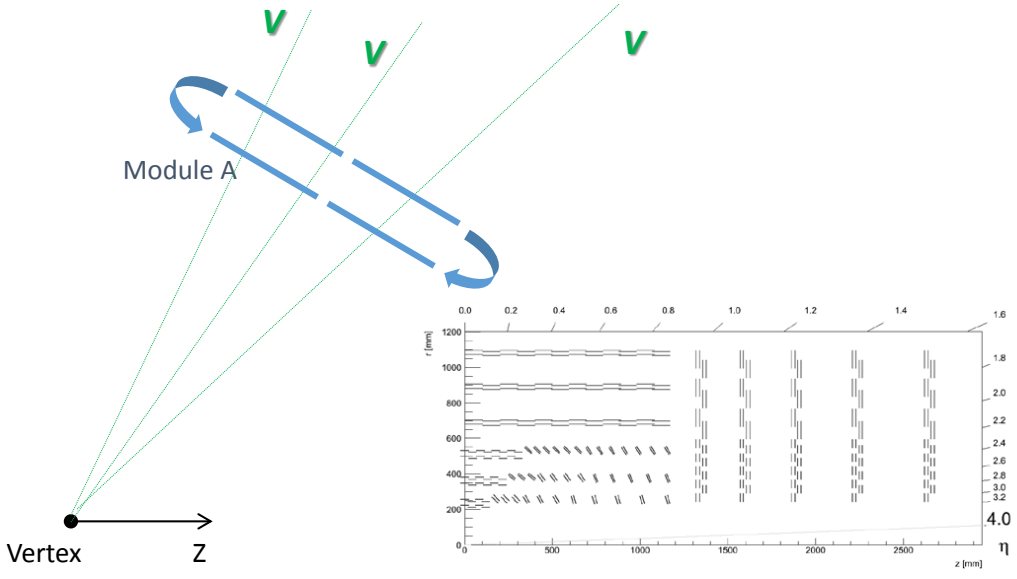


FIGURE 4.21: Top: The baseline tracker layout shows inefficiencies due to the absence of z-communication between the two part of a module. The super-position of modules compensate for the inefficiencies at the edge of the module. Bottom: The tilted layout strongly reduces the inefficiencies due to the absence of z-communication, solving the problem.

and must be studied. The next paragraph discusses the optimizations carried out on the data format and the estimated inefficiencies.

4.12 Data Format

The details of the data format on the PS module are summarized in the specification document [38], while the studies on the efficiencies losses in the CMS report by Viret [39]. Several architectures were studied for the communication protocol between MPA-/CIC and CIC/LP-GBT:

- **Fixed Blocks:** The bandwidth is divided in two asymmetrical channels. Most of the bandwidth is reserved for the Trigger path, while the remaining is given to the L1 data path. This approach is the simpler from the point of view of the design, but it can lead to higher inefficiencies.
- **Priority to trigger:** The bandwidth is shared between the two readout path. The Trigger path has the priority and the L1 data path uses the bandwidth only when it is free from Trigger data. Such a solution minimizes the losses, but it adds complexity from the point of view of the design.
- **Trigger/Raw separation:** This solution exploits the LHC bunch crossing structure. Particle collision does not take place every bunch crossing, therefore these empty bunches can be used to send L1 data while the remaining bunch crossing are used to send only Trigger data. The limitation of this approach is the possibility of a change in the LHC bunch crossing structure, which could affect the system. For this reason, this solution was discarded.

4.12.1 Trigger path transmission

In order to take a decision, simulations with MC events were carried out to estimate the loss with a Fixed Blocks architecture, where 5 lines are reserved for the trigger path and 1 line for the L1 data path. Considering the half pixel resolution and the bending value encoded on 5 bits, the maximum number of stubs per BX is 2. Using MC events with 140 pile up, the losses gets to $\sim$5% in the inner layers of the tracker, which is too high to ensure a correct event reconstruction.

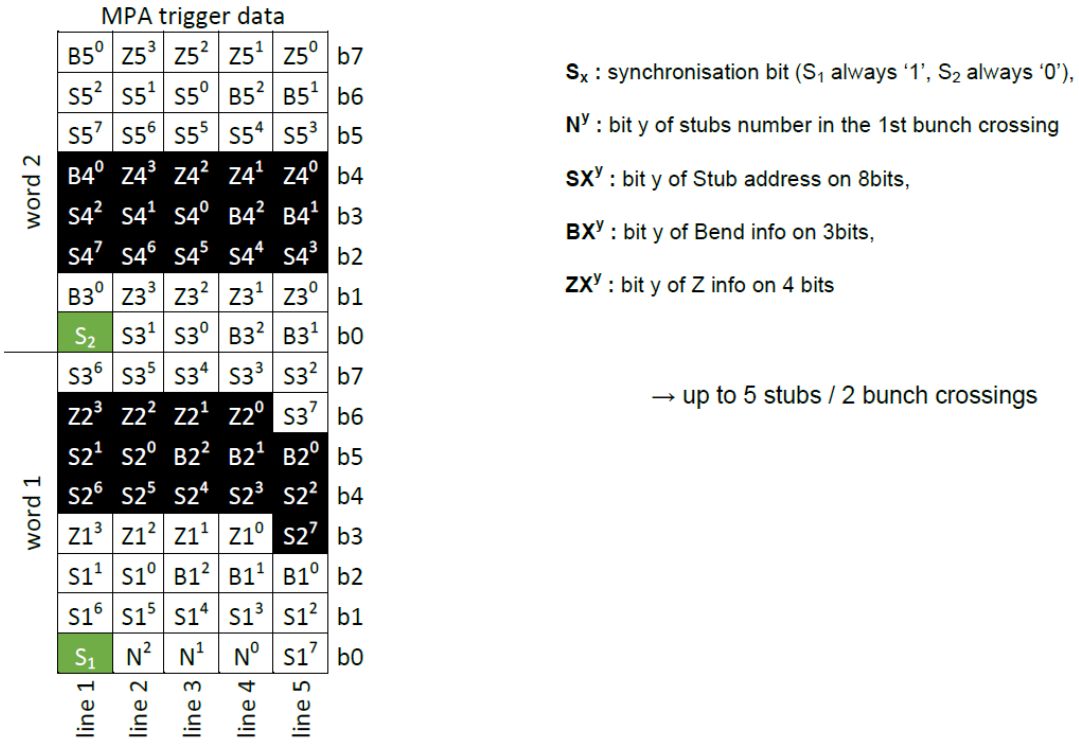


FIGURE 4.22: Trigger path data format.

A possible solution to handle the higher occupancy in the tracker inner layer is sending the trigger data from the MPA in synchronous blocks of 2 clock cycles. This technique allows to average the stub rates between two consecutive bunch crossings. Trigger data format is shown in figure 4.22. As the packet is spanning over two clock cycles, every word begin with a synchronisation bit, always '1' in the first word and always '0' in the second one. Then, the first word contains the number of stubs in the first BX encoded over 3 bits, in order to determine the BX to which each stub belongs. All stubs in the payload beyond this value belong to the second BX.

Further improvements are obtained with the bending encoded on 3 bits. Including this option in the MPA allows to pass 5 stubs per 2 BX limiting the losses only where needed and keeping the maximum resolution for the bending in the rest of the tracker. Table 4.6 summarize the average front-end losses measured in the different pile-up scenarii 140 and 200.

The simulation results show losses under 1% at PU 140. While at PU 200, the losses get to 2% but it is sufficient to use a tighter stub building cuts to mitigate the problem. Such reconfigurability is kept in the MPA and the final setting can be adjusted after the design. However, it is interesting to see how the efficiency is affected by a tighter cut.

Layer:	TIB1	TIB2	TIB3
PU 140	0.9 ± 0.4 %	0.2 ± 0.5 %	0.1 ± 0.6 %
PU 200	1.7 ± 0.3 %	0.3 ± 0.4 %	0.1 ± 0.4 %

TABLE 4.6: Average values of losses in the MPA-CIC communication for the Tracker Inner Barrel (TIB) layers

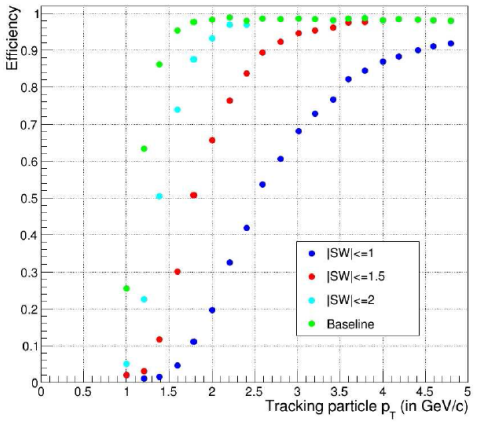


FIGURE 4.23: Stub efficiency for different SW (Stub width) cuts in TIB1.

Figure 4.23 shows, for different stub width cuts, the stub reconstruction efficiency in the barrel innermost layer. Tighter cuts shows large reduction of the low p_T stubs rate at the cost of a poorer resolution in the stub pT resolution, in fact the turn-on curve being less sharp.

The low losses reached with the Fixed Block architecture avoids the additional complexity of the Priority to Trigger architecture.

4.12.2 L1 data path transmission

The choice of a Fixed Block architecture leaves one line for the L1 data which are sparsified in the MPA. The data format, shown in figure 4.24, consists of two basic parts:

- The header have fixed length and five data blocks: a start sequence (19 bits), an error field (2 bits), a L1ID (9 bits), the strip cluster multiplicity (5 bits) and finally the pixel cluster multiplicity (5 bits).

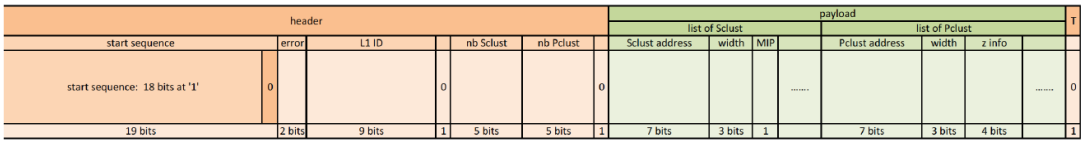


FIGURE 4.24: MPA L1 data format where Sclust list contains the strip clusters while the Pclust list the pixel clusters.

- The payload has a variable length and is divided into two parts. The first part is the strip clusters list, each strip cluster being encoded on 11 bits (7 bits address, 3 bits width, and 1 bit for MIP flag). The second part is the pixel cluster list, each pixel cluster being encoded on 14 bits (7 bits address, between 1 and 120, 3 bits width and 4 bits for the Z position).

The goal of the start sequence is to create a unique and well identified sequence of 19 bits that cannot be found anywhere else in the data format. In case of error in the L1 data format, at worst the erroneous L1 data frame will be lost, but the concentrator will be able to handle easily the next frame. In order to guarantee the uniqueness of the start sequence, a fixed '0' bit is inserted in the header part between the L1 ID field and the number of strip cluster, and also between the number of pixel cluster and the payload part.

Once a data format is defined also for the L1 data path, the simulations with MC events provide a size estimation for the L1 data path FIFO (First In First Out). After the L1 accept trigger is received by the MPA, the L1 processing prepares the data for the transmission. The L1 accept trigger can arrive at random time, thus also when another event is being transmitted. In this case, the L1 FIFO stores the events which are waiting to be transmitted. The FIFO width is defined by the maximum payload, while the length by the L1 accept rate. These two values are obtained from the MC events at PU200, which provide the maximum cluster multiplicities corresponding to the maximum size of an event, as shown in figure 4.25 and 4.26.

The cluster multiplicity shown in the figure confirms the choice in the L1 data format. Concerning the size of the sparsified L1 word, only one word out of a 10 milions is larger than 500 bits. Consequently, an MPA FIFO size of 512 bits would be sufficient to ensure an almost lossless data transmission. Concerning the depth of the FIFO, the value 5

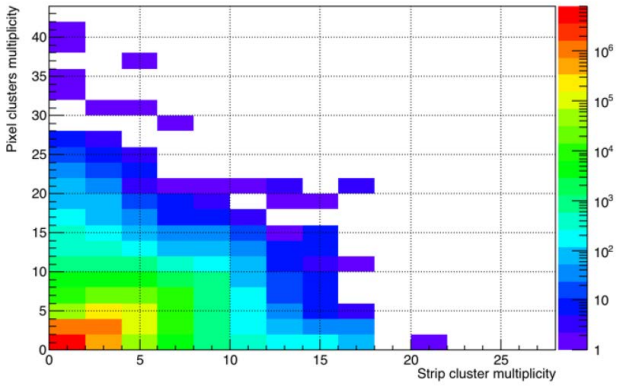


FIGURE 4.25: MPA L1 data cluster multiplicities, for PU200 events.

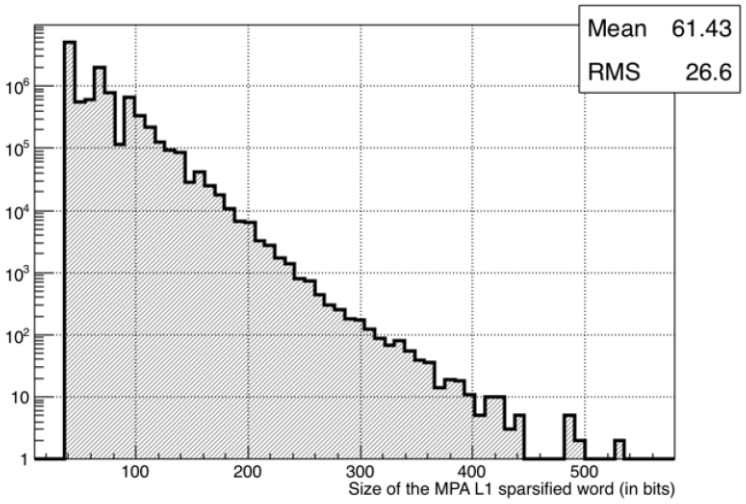


FIGURE 4.26: Size of the MPA L1 sparsified words, for PU200 events.

seems appropriate for the whole detector, since during the simulations never more than 4 events are stored in one MPA.

4.13 Chapter Summary

This chapter presented the Macro Pixel ASIC, a readout ASIC for hybrid pixel detector. Pixel and array sizes are defined by the particle tracking application. The choice of a 65 nm CMOS technology is forced by the requirement of a power density lower than 100 mW/cm^2, of an high complexity logic capable of intelligent particle tracking and of storing the full event for more than 10 μs. A peculiar floorplan and clock distribution have been studied in order to reduce power consumption of signal distribution and

data transport. A front-end electronics with current consumption lower than $30\,\mu$A per channel and a noise lower than $200\,e^-$ have been designed. The algorithm for intelligent particle tracking, called Stub Finding, has been implemented at gate level, and verified with randomly generated events as well as with Montecarlo generated physics events, which allowed also the estimation of the performance. Supply voltage scaling and memory gating techniques have been used to fulfil the power requirements. Data formats for chip-to-chip communication have been studied with the objective of optimizing the bandwidth. In conclusion, the complete architecture of a readout ASIC for intelligent particle tracking has been defined in details, and it will be implemented in the coming years.

Chapter No. 5

A prototype in 65 nm technology

A readout ASIC for hybrid pixel detector with the capability of performing quick recognition of particles with high transverse momentum has been designed for the requirements of the CMS Outer Tracker at the High Luminosity LHC. The choice of a 65 nm CMOS technology has made it possible to satisfy this power requirement despite the fairly large amount of logic necessary to perform the momentum discrimination and the continuous operation at 40 MHz. This chapter describes a first prototype chip in 65 nm including a large part of the final functionality and the full front-end electronics. Measurements of the analog front-end characteristics closely match the simulations and confirm the consumption of $< 30\,\mu$A per pixel. Radiation tests provide important knowledge about the technology. Prototype module production allows further studies on the readout chip, on the sensor and on the assembly techniques.

5.1 Description of the MPA-Light

The high complexity of the digital logic and the very low power density requirement make the MPA a very challenging project. This prototype, called MPA-Light, addresses the challenge related to the design of a 65 nm analog front-end with a power consumption $< 30\,\mu$A per pixel and explores several techniques for low power digital design. The front-end implements binary readout, while a fully synthesized logic in the chip periphery includes the Stub Finding logic for quick recognition of particles with high p_T. Input pins using a 160 MHz data rate provide the strip data used to generate stubs. A summary of the main MPA-Light specifications can be found in table 5.1.

Parameter	Value
Active area	1.7 x 4.5 mm
Periphery area	1.7 x 2 mm
Technology	65 nm CMOS
Pixel Size	100 μm x 1446 μm
Detector type	n+ on p-
Input Capacitance	≤ 500 fF
Measurement type	Binary readout
Clock input	160 MHz
Data type	Raw or encoded coordinate

TABLE 5.1: MPA-Light specifications.

5.1.1 ASIC architecture

The MPA-Light is a 48 channels readout ASIC for hybrid pixel detectors designed in 65 nm CMOS technology to readout n+ on p- silicon detectors with an estimated capacitance < 500 fF per channel. As shown in Figure 5.1, the die size is ~ 1.7 mm x 6.5 mm and includes pads for bump-bonding to the detector and wire-bonding for read-out and powering signals. The pixel matrix is composed by 16 x 3 pixels, where every pixel measures 100 μm x 1446 μm. The sensitive area is $\sim 7.65\,\text{mm}^2$. The side edges of the ASIC extend only by 50 μm outside the pixel matrix to allow multi-chip assembly with a single sensor and therefore a minimal dead-area. At the bottom edge of the pixel matrix a row of dedicated bumps provides ground connection to the pixel sensor from the chip.

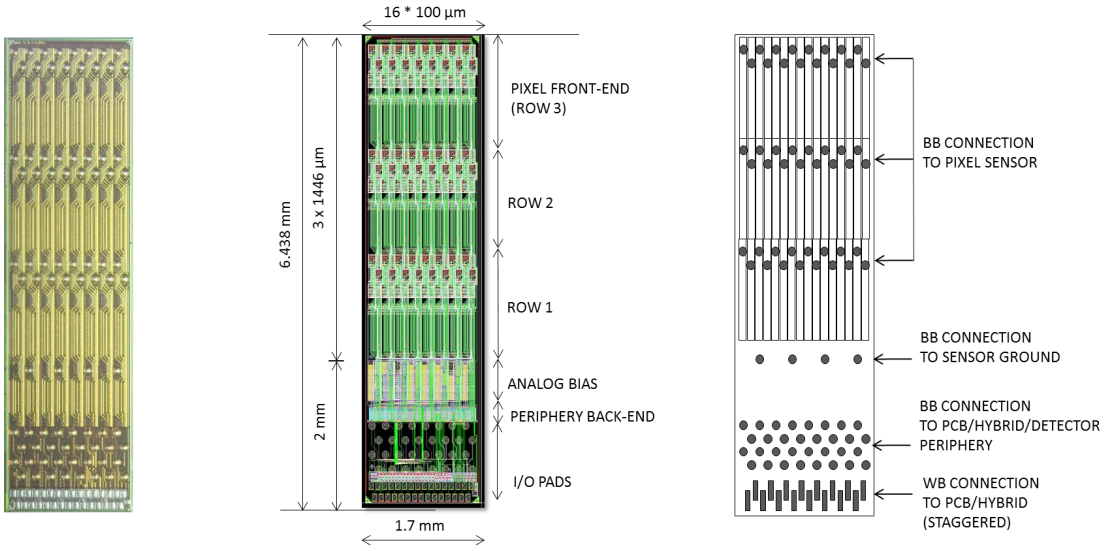


FIGURE 5.1: Left: picture of the MPA-Light. Centre: layout view of the MPA-Light with dimensions and components. Right: connectivity view of the MPA-Light (WB = wire-bond, BB = bump-bond).

Two staggered rows of wire-bond pads at the bottom edge of the periphery connect to the module substrate system. The wire-bonding pitch is 50 μm. The same signals are also provided by four rows of bump-bonds which allow a fully bump-bonded assembly with a bump-bonding pitch of 200 μm. The chip can be operated in several modes and can be set in acquisition or configuration/readout modes by an external signal. A serial shift register is used to load the configuration, while another is used to read out the data. When acquisition starts, an internally generated signal clears the data from the previous acquisition. During acquisition, the chip needs an external 160 MHz clock through a Current Mode Logic (CML) receiver designed by Felici [40]. An internally derived 40 MHz clock is used for acquisition.

5.1.2 Pixel front-end

The pixel architecture is shown in figure 5.2. The analog front-end consists of a pre-amplifier, a shaper and a two stage discriminator with hysteresis. The details about the analog front-end can be found in paragraph 4.5.1. A global 8-bits DAC (threshold DAC) sets the threshold for the pixel matrix and a 5-bits DAC per pixel allows to compensate the pixel-to-pixel threshold variations. Another global 8-bits DAC (calibration DAC) sets the amplitude of a pulse that can be injected with a test capacitance (Cc) of 20 fF to the front-end.

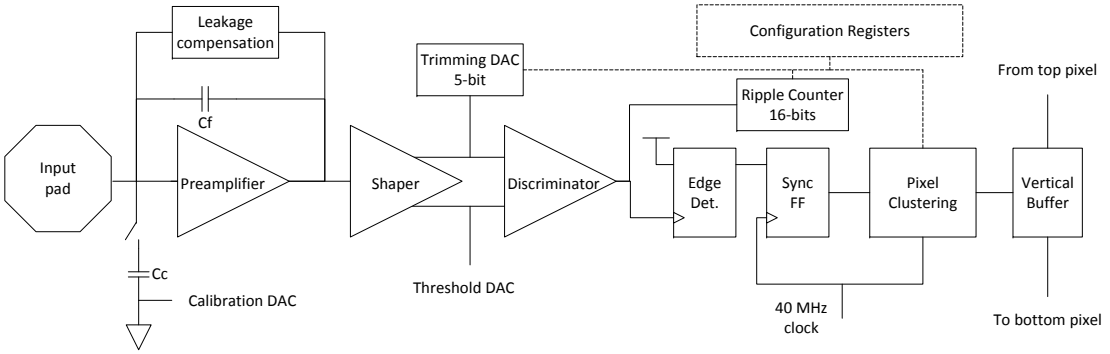


FIGURE 5.2: Pixel front-end. Dashed lines represent configuration signals.

The output of the discriminator connects to an edge detector followed by a flip-flop which synchronizes the pulses from the front-end with the 40 MHz clock. Before being transmitted to the periphery, the pixel data is reduced by the pixel clustering logic, described in details in 4.6.2: pixel clusters larger than a certain value depending on the position in the tracker cannot be generated by high transverse momentum particles and consequently such clusters are immediately discarded; the clusters within the defined range are instead reduced to the centre of the cluster, called centroid. The centroids are transmitted to the periphery with a fixed latency of two clock cycles.

For testing purposes, a 16-bits ripple counter connects to the discriminator output and is read out at the end of each acquisition. Configuration registers allow to disable this counter, the binary readout and the pixel clustering logic, as well as to define the width of clusters accepted by the pixel clustering logic and the values for threshold equalization.

5.1.3 Periphery back-end

The chip periphery includes the logic shown in Figure 5.3. It performs the quick recognition of high p_T particles using the Stub Finding logic described in details in paragraph 4.6.3. The logic includes strip cluster reduction/elimination, offset correction, centroid position encoding and correlation logic. The block receives the pixel data from the front-end electronics and the strip data from an external input. The strip interface runs at a frequency of 160 MHz with a strobe signal for synchronization. A deserializer prepares this data for the Stub Finding logic.

The data processing elements are pipelined. Registers store one step of the data path according to the processing mode chosen during configuration. The available modes

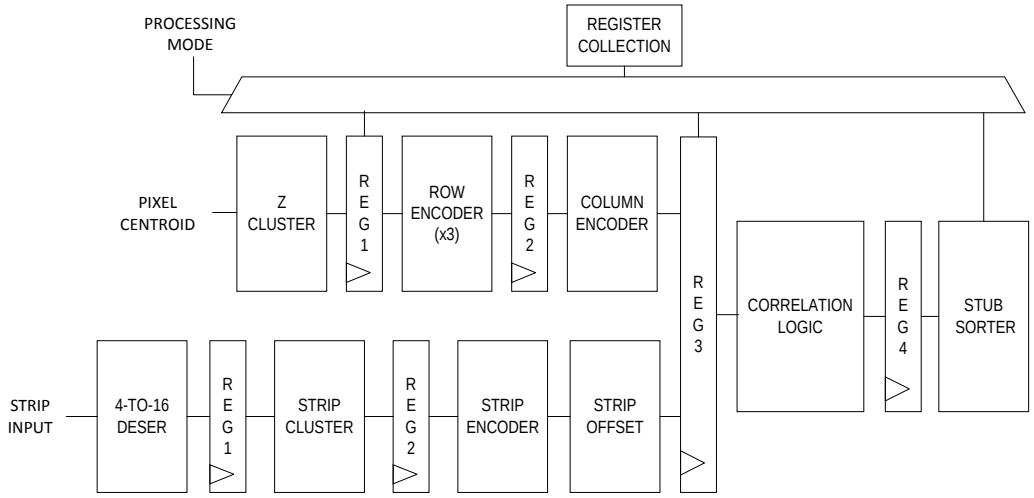


FIGURE 5.3: Periphery schematic.

allow to store the data from the front-end, the coordinate from the pixel encoders or the stubs from the stub sorter. A time-stamp, representing the cycle counter from the acquisition start, identifies the saved data.

The MPA-Light can also emulate the functionalities of the SSA, i.e it works as a readout chip for strip detectors, by OR-ing the pixel columns. When it is set in this mode, a serializer block sends the strip data at 160 MHz to the output. A strobe signal is generated for data synchronization.

The Stub Finding logic allows evaluating on silicon the timing and power performances obtained in simulation. Furthermore, the SSA emulator mode can be used to build a reduced size module with the capability of performing quick recognition of high p_T particles using only MPA-Light ASICs.

5.2 Electrical characterization

The MPA-Light was submitted for production and the first batch was available at the beginning of 2015. Tens of samples were tested so far and proved functional. A custom test set-up, shown in figure 5.4, was developed to test and characterize the MPA-Light. It includes a carrier board holding the chip, an interface board for voltage and current generation and monitoring, an FPGA development board and a command line interface program running on a Linux PC.

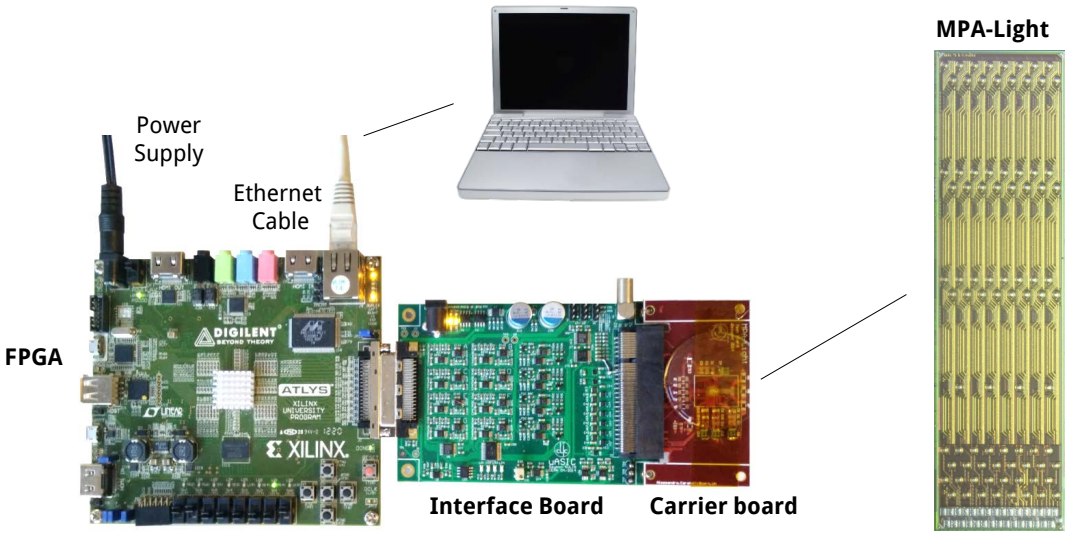


FIGURE 5.4: MPA-Light test system.

The characterization was performed using the internal test capacitor in a bare chip, without connection to a sensor. The quantity of charge injected is determined by the relationship $Q_i = C_c V$, where V is the voltage determined by the calibration DAC in the periphery. A wire-bond connection allows the measurement of the calibration DAC voltage which provides a Least Significant Bit (LSB) value of 0.035 fC, when an ideal value of 20 fF is assumed for C_c, and an Integral Non-Linearity (INL) $< \pm 0.5$ LSB. Power consumption confirms the simulation values, in particular the analog front-end provides a consumption of $\sim 25\,\mu$A per channel.

5.2.1 Front-end characterization

A 5-bits threshold-equalization DAC per pixel allows to correct for the pixel-to-pixel threshold mismatch. Figure 5.5 shows the number of events counted while scanning the threshold voltage for each possible value of the 5-bits DAC. The measured dynamic range is 90 LSB ($\sim 9.5\,\mathrm{ke}^-$) and the INL is ± 0.2 LSB. The r.m.s threshold variation before equalization is 16 LSB ($\sim 1.7\,\mathrm{ke}^-$), while afterwards the achieved noise free variation is 0.8 LSB ($\sim 95\,\mathrm{e}^-$) as shown in figure 5.6.

The σ of the gaussians in figure 5.5 corresponds to the Equivalent Noise Charge (ENC). Its distribution is shown in figure 5.7 and provides an r.m.s value of 1.5 LSB ($\sim 162\,\mathrm{e}^-$). The minimum detectable charge can be calculated by quadratically adding the measured

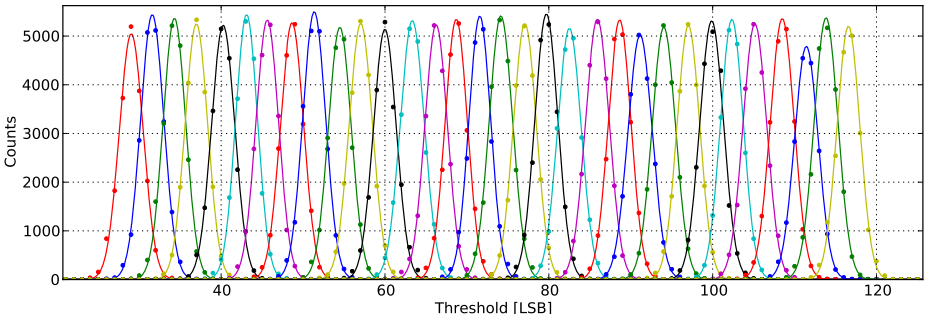


FIGURE 5.5: Baseline scan of a pixel for the 32 different DAC codes of the trimming DAC.

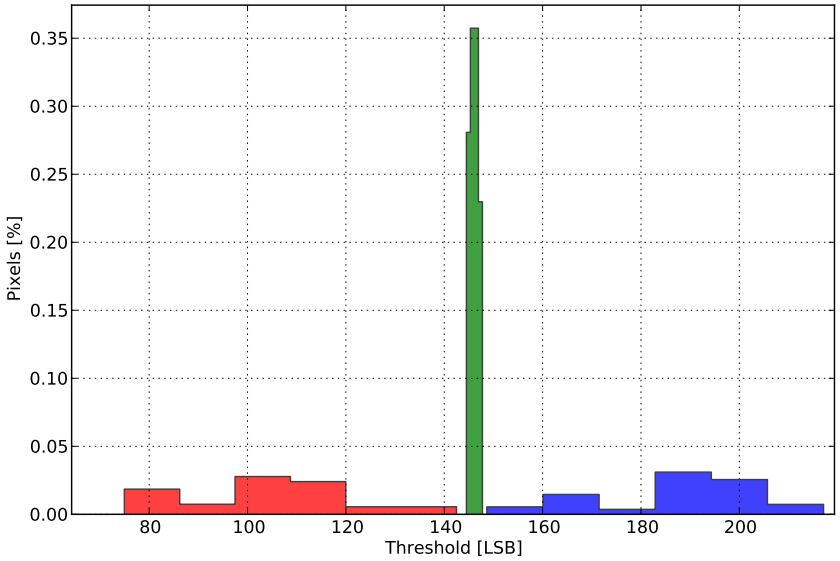


FIGURE 5.6: Red and blue histograms show the threshold distribution of all pixels with the minimum and maximum DAC codes; the green histogram shows the same distribution for calibrated matrix.

electronic noise and the threshold variation because both measurements are uncorrelated. The quadratic sum of mismatch and noise is 1.7 LSB ($\sim$175 e^-). After threshold equalization, the 6σ minimum threshold could be set at 10.2 LSB ($\sim$1.1 ke^-).

Figure 5.8 plots the S-curve of the whole pixel matrix for different pulse amplitudes. A thousand pulses are injected for each value of threshold. Threshold equalization is optimized for values between 0.5 fC and 1.5 fC since this is the range expected during operation. Complementary error function fitting allows the extraction of the S-curve mid-point which is the effective threshold. The front-end gain can be calculated from the distance among the effective threshold for different input charges. The measured gain is 85 ± 5 mV/fC.

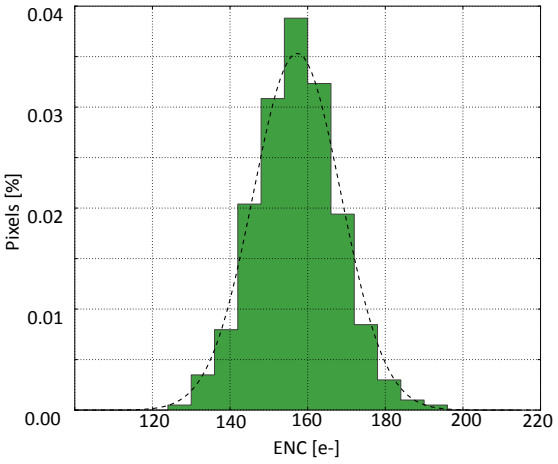


FIGURE 5.7: Noise distribution

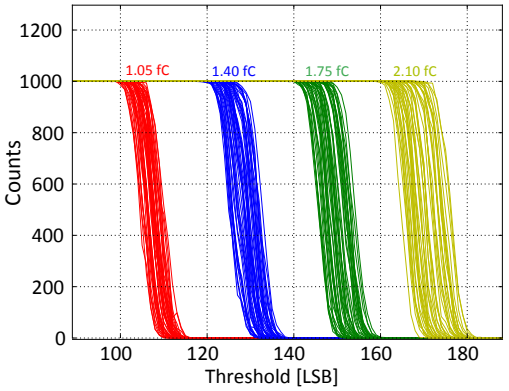


FIGURE 5.8: S-curves for different pulse amplitudes

The binary readout allows timing characterization of the analog front-end. Charge injection time scans for different threshold values provide the shaper output as shown in figure 5.9. Peaking time is 24 ± 1.6 ns as expected from simulation. An important parameter to ensure correct operation at 40 MHz is the front-end time walk. It is measured as the difference between charge injection time and detection by the edge detector in the front-end. Figure 5.10 shows the measurements with thresholds at 0.5 fC and 1 fC which give a walk time $< 15 \pm 1.6$ ns over a charge range from 0.5 to 9 fC. The characterization of the front-end provides results very close to simulations. Consequently, the front-end respects all the requirements and can be used without further optimization in the MPA.

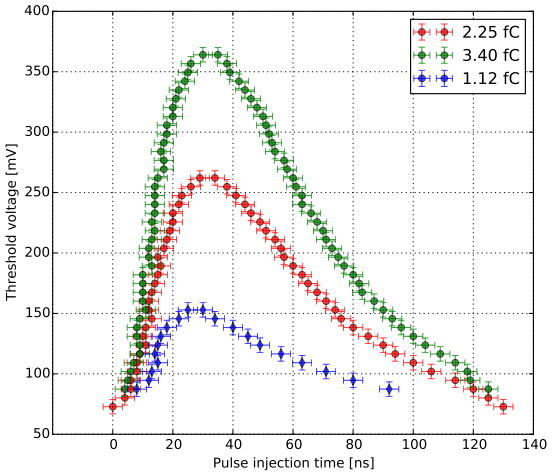


FIGURE 5.9: Shaper output for different input charges

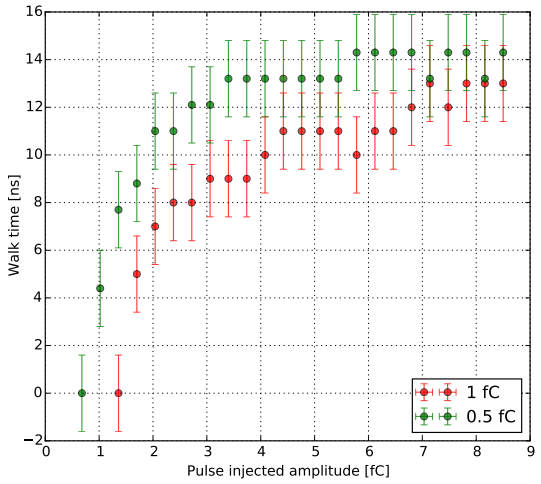


FIGURE 5.10: Time walk for different thresholds

5.3 Multi-chip module prototype

The MPA prototype was produced not only for the development of the ASIC itself, but also to help the prototyping of the sensor and of the module. In the PS module, the assembly of pixelated sensor with its readout ASIC is called "MaPSA" i.e Macro-Pixel-Sub-Assembly. The first MaPSA prototype, called MaPSA-Light includes six MPA-Lights distributed on two rows which are connected to one pixelated sensor as shown in figure 5.11. The objective of this prototype development is:

- Evaluate the MPA chip functionality on a reduced set of pixel channels;
- evaluate the flip chip bump bonding process between the MPA-Light and the Sensor-Light objects;

- evaluate potential handling and testing difficulties with the MaPSA subassembly.

As explained by Bergauer [41], the pixelated sensor for the MaPSA-Light has been produced on p-type Float-Zone wafer of 4 inches. The thickness of the sensor is 200 μm, the resistivity is between 4 kΩcm and 8 kΩcm and the oxygen concentration is $< 2\text{x}10^{16}\text{cm}^{-3}$. The pixels are DC coupled with punch through bias. Different variant for the pixel isolation has been implemented. The baseline consists in p-stop implant with peak concentration $\sim 1\text{x}10^{16}\text{cm}^{-3}$ and full depth $> 1.5\mu$m. Peak concentration is defined as the maximum concentration slightly below the bulk surface, while full depth is defined as the depth below the bulk surface where the doping concentration has almost reached the bulk concentration (within 10%). The full depletion voltage is $\sim$90 V while the breakdown voltage is > 500 V.

The wafer sensor includes also a different geometric design of the detector for a possible variant of the MaPSA assembly: the flipped-MaPSA described in the next paragraph.

5.3.1 Assembly architecture

Figure 5.12 shows the two alternative architectures for the PS-module. The baseline foresees a pixel readout chip bump-bonded to the sensor and wire-bonded to the hybrid. This configuration requires two different I/O processes on the same chip which can not be provided by the ASIC foundry. Consequently, the bump-bonding processing is carried out by the module assembler with additional cost. Moreover, the pixel readout ASIC

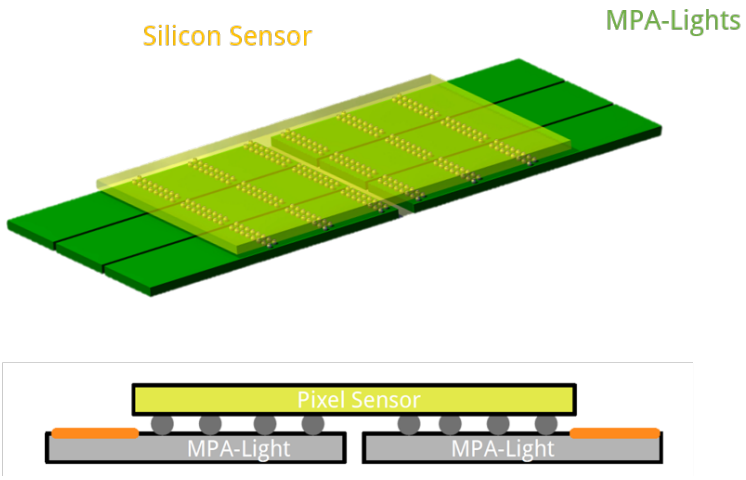


FIGURE 5.11: Top: MaPSA-Light 3-D view. Bottom: MaPSA-Ligth side view.

in between the two sensors can interact with the passing particles, and disturb the p_T measurement. Last, the MPA, which is the most power consuming component of the PS module, is cooled through the bump-bonds and the sensor which does not provide the best cooling efficiency.

A possible solution to these problems is the Flipped-MaPSA. Inverting the position of MPA and pixel sensor, the main advantage is the absence of material between the two sensors. Furthermore, the MPA design would contains only bump-bonding I/O and the cooling would reach the maximum efficiency. On the other hand, this configuration reduces the active area in the pixel sensor which can be critical in the final design production. Also the cooling of the sensor can become critical due to temperature gradient across the sensor surface. However, the main concern are the digital signals which cross the periphery of the sensor from the MPA to reach the Front-End hybrid through wire-bonding connections.

In order to explore the feasibility of this new approach, the MPA-Ligth is equipped with wire-bonding and bump-bonding pads. The latter allows to build a flipped MaPSA-Light which can test the impact of the digital signal and of the cooling. Some dedicated sensors have been designed with an isolated periphery where the links between the signals from the MPA and the Front-End Hybrid is implemented.

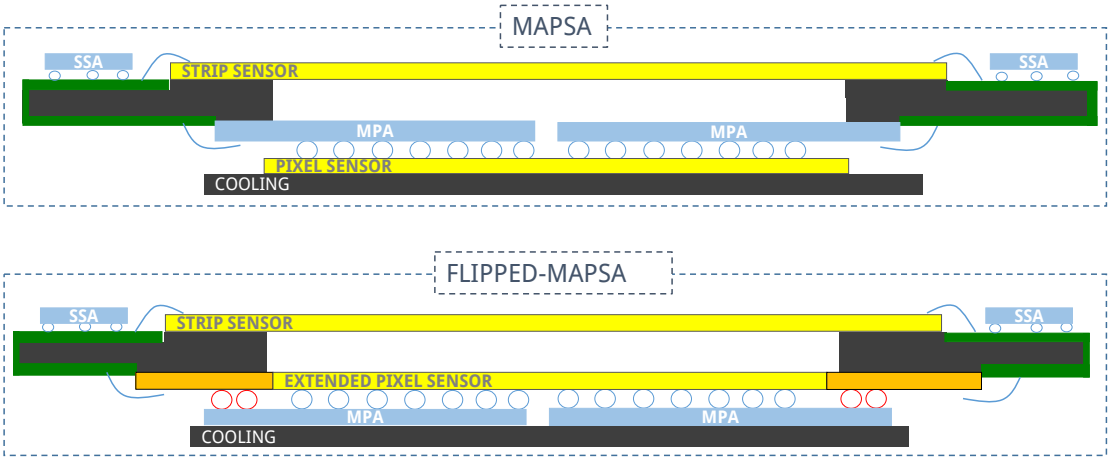


FIGURE 5.12: Top: Schematic of the PS-module with MaPSA. Bottom: Schematic of the PS-module with Flipped-MaPSA.

5.3.2 MaPSA-Ligth assembly process

The assembly process for the MaPSA-Light consists in Under Bump Metalization (UBM) deposition on both sensor and readout ASIC, bumps fabrication on one of the two components and flipping of one component on the other. The readout ASIC are received already diced by the foundry because the production is carried out on Multi Project Wafer (MPW). Thus, the deposition of UBM and of the bumps require a special effort. On the contrary, the sensors are received as wafers and the dicing step is part of the assembly process.

The MaPSA-Ligth assembly has been outsourced to external companies which provides bump-bonding and assembly services. Two different approaches have been proposed:

- The first strategy consists in Indium bumping on the sensor wafer. The first step prior assembly is the indium bumps fabrication on the sensor wafers. Afterwards, an Under Bump Metalization (UBM) must also be deposited on the MPA. This step requires wafer rebuild on temporary mask since they are provided already diced by the foundry. At this point the assembly step is done by flipping the MPA-Ligth on the sensor wafer with the bumps. After the bonding of the MPAs, an under-filling step is necessary for reliability issues. The result is shown in figure 5.13
- The second approach is based on solder sphere jetting on the MPA-Light. This technique is very specific for single die prototyping. The bumping is selective, so the extra bumps reserved for flipped-MaPSA will be skipped. The balling material is $Sn_{98}Ag_2$. The steps followed for the assembly are the same of the first described approach.

The first production batch has provided 15 MaPSA-Light assemblies. The preliminary testing, whose results are shown in the next paragraph, has been carried on three samples from different manufacturers.

5.3.3 MaPSA-Light results

The testing of the MaPSA-Light was focused mainly on the MPA-Light ASIC perfomances when a pixel sensor is bonded. However, the tests also provided a verification

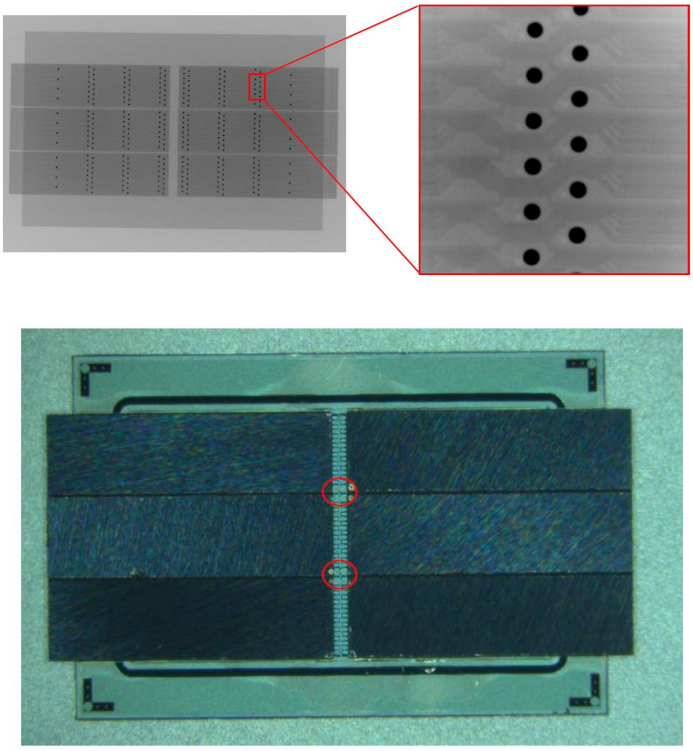


FIGURE 5.13: Top: X-ray image of MaPSA-Light. Zoomed image allows to see the alignment of the bumps. Bottom: Image of the MaPSA-Light after under-filling. Red circles shows the application points.

of the basic functionality of the MaPSA-Light and made a preliminary investigation of the assembly quality. The main advantage of this test campaign is the possibility to exploit the existing setup for the MPA-Light testing. The same test carried out on the single ASIC can be repeated on the full assembly. This method provides a comparison between the ASIC perfomances without and with sensor, besides it allows to use an already debugged system. The only requirement is a larger carrier board to include the space and connectivity for the MaPSA-Light.

The first measurement was the IV characteristic of the sensor, which is shown in figure 5.14. Before the assembly process, sensor wafers have been measured and only the good sensors have been chosen for the assembly. Thus, any problem in the IV characteristic would be related to the assembly process. The measured sensors show a diode characteristic as expected, but with a current slightly higher than before assembly. This increase is related to the capacitance of the bonded ASIC which provides an additional leakage current.

Table 5.2 shows a comparison between the performance of the 3 MaPSA-Light measured.

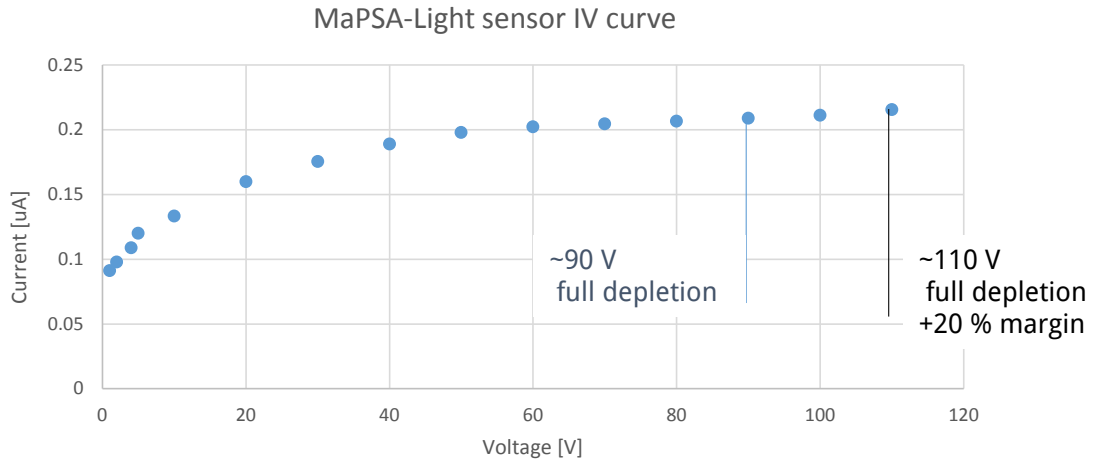


FIGURE 5.14: MaPSA-Light sensor IV characteristic.

The threshold spread is similar among the samples, and always below 1 LSB (∼108 e^-). Furthermore, it is smaller than the noise, which instead varies between the samples. The baseline noise in the table is the average value of the pixel matrix. In the case of the sample number 2, the noise value is very close to the value before the assembly process (see paragraph 5.2.1). Usually the noise value with the bumped sensor should be higher due to the additional capacitance of the sensor as in the case of the other two samples.

Sample	Baseline Noise	Threshold Spread	Minimum Threshold ($6\,\sigma$)
MaPSA-Light 1	185 e^-	80 e^-	∼1.2 ke^-
MaPSA-Light 2	165 e^-	88 e^-	∼1.1 ke^-
MaPSA-Light 3	200 e^-	67 e^-	∼1.3 ke^-

TABLE 5.2: MaPSA-Light performance results.

In order to investigate this difference, the noise distribution on the pixel matrix of the MPAs was analysed. In figure 5.15, the noise distribution of the sample 2 shows two different groups of pixels: one with a noise comparable to the MPA-Light before assembly and another with a noise close to the other two assemblies. The most probable reason for this effect is the presence of unconnected pixels. To prove it, a Sr-90 source has been placed on the second sample and the MPA-Light on the top has been placed in counting mode. As shown in figure 5.16, only a few pixels are connected to the sensor and are counting. Concerning the other two assemblies, both methods, noise distribution and source illumination, show all pixels connected. Since there are not connectivity problems,

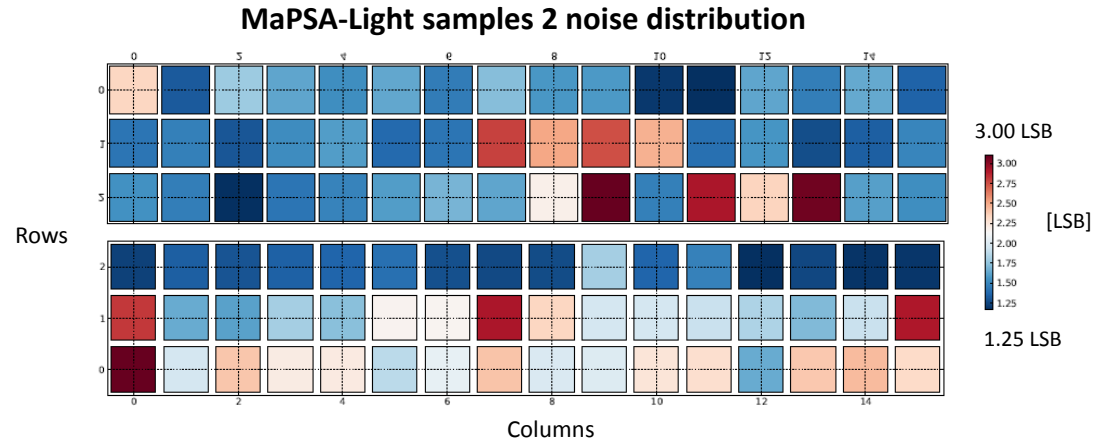


FIGURE 5.15: MaPSA-Light sample 2 noise distribution.

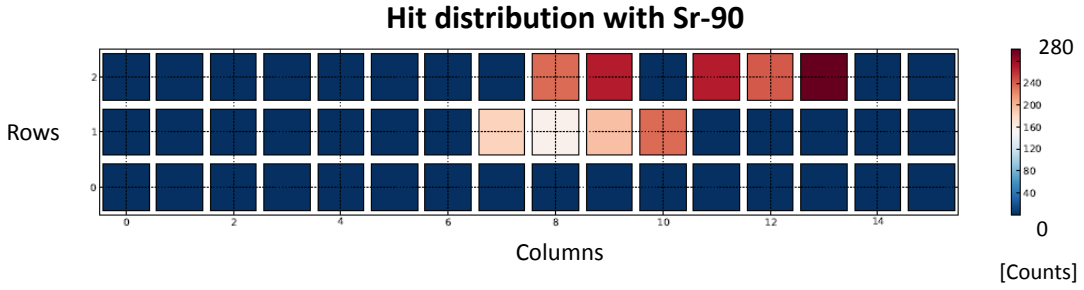


FIGURE 5.16: MaPSA-Light sample 2 hit distribution with a Sr90 source for the MPA-Light on the top of the assembly.

the 6σ minimum threshold for sample 1 could be set at 10.3 LSB ($\sim$1.2 ke$^-$), while for sample 3 it could be set at 10.4 LSB ($\sim$1.3 ke$^-$).

All the MPA-Light performances shown up to this point are measured injecting test pulses through a capacitance. This capacitance has a nominal value of 20 fF, but the actual value can vary of $\pm$20 % due to process variations. In order to measure the value of this capacitance, a threshold scan during an illumination with a known radioactive source can be used. The source is placed above the assembly, and for each value of the threshold the front-end will count how many particles are collected. For this measurement, the Cd-109 has been chosen since it shows a gamma peak in the dynamic range of the MPA-Light, around 6.2 ke$^-$. In reality, the Cd-109 has two gamma peaks around this value, but the resolution of the MPA-Light cannot distinguish them, because their difference is $<$ 1 ke$^-$. The threshold scan in figure 5.17 provides a Threshold LSB value $\sim$ 135 e$^-$, a Calibration LSB $\sim$ 255 e$^-$ and a test capacitance value $\sim$ 23.5 fF. This value is very close to the nominal and qualify all the results obtained with the test capacitance.

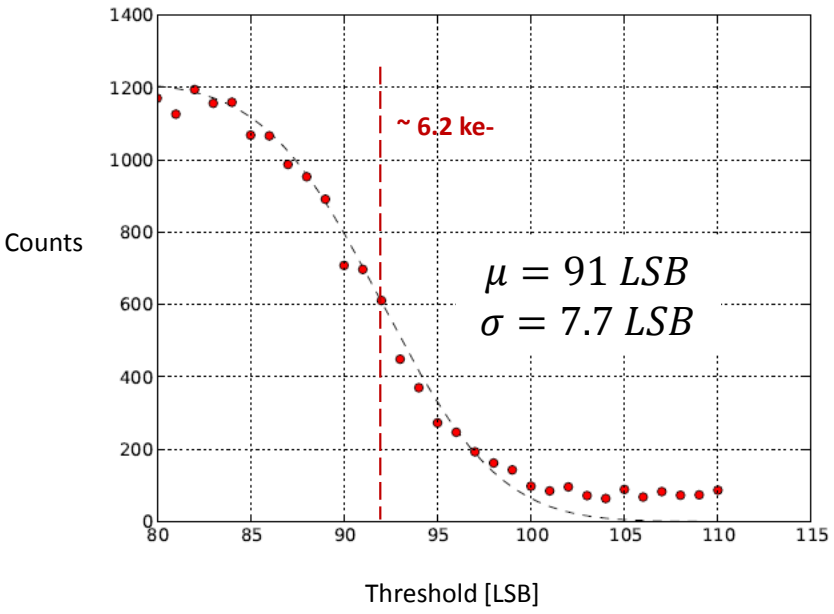


FIGURE 5.17: Threshold scan of a Cd-109 source.

Additional MPA-Lights were requested to the foundry for a total of 700 samples, which have been used and will be used for various studies, like additional modules, test beam, radiation tests, etc. A summary of the activities about MPA-Light and MaPSA-Light was given by Vasey in [42].

5.4 Total Ionization Dose characterization

An additional test needed for the MPA-Light is the TID characterization with X-rays. The expected dose is obtained with the FLUKA simulation package [43]. In figure 5.18 the Dose at 3000 fb^{-1} for the whole tracker is shown. The worst case for the PS module is found in the first barrel layer at 23 cm from the vertex and corresponds to a TID < 100 MRad.

Irradiation of the MPA-Light was performed using a calibrated 50-kV 3-kW X-ray generator (SEIFERT RP149). The characterization for TID was divided in two campaigns. The device was firstly irradiated focusing mainly on the analog blocks. Afterwards, several irradiation test were carried out to explore the effect of TID on the digital logic in the chip periphery. The purpose of this second part is not only extracting information concerning the MPA-Light, but also regarding the 65 nm technology itself. The results will be also compared with the transistor radiation effects reported by Faccio in [44].

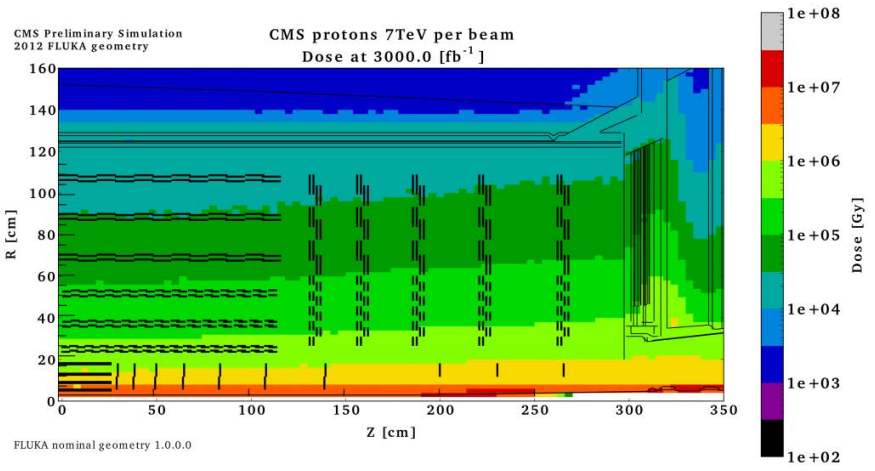


FIGURE 5.18: Total Ionizing Dose in Grey for the full tracker.

5.4.1 Analog blocks results

In order to evaluate the performances of the analog block in the MPA-Light ASIC, the chip was exposed to X-rays up to 150 MRad with a dose rate of 105 kRad/min. After the irradiation, it was annealed for 24 hours at 100 °C under bias. During the entire procedure, the test system monitored power consumption, front-end performance and digital logic. At 150 MRad the analog power consumption variation was -7 % while it was -4 % after annealing. The monitoring of the digital logic did not show and fail at nominal condition (1.2 V and 40 MHz). Minor variations were observed on the analog blocks: the calibration DAC LSB variation was -3 % at 150 MRad, but it increased up to -7 % after annealing (figure 5.19); the threshold LSB extracted from S-curves degraded up to -15 % at 150 MRad, while it partially recovered after annealing providing a final variation of -11 % (figure 5.20). However, the large dynamic range of the global DACs allows to correct easily the observed variations. Consequently, the analog blocks in the MPA-Light did not show any problem with X-ray TID irradiation up to 150 MRad.

5.4.2 Digital logic results

Faccio [44] reports severe radiation damage in short and narrow channel transistors in 65 nm technology. The performance degradation depends on the bias and temperature applied both during and after irradiation. The transistor study shows p-type transistors

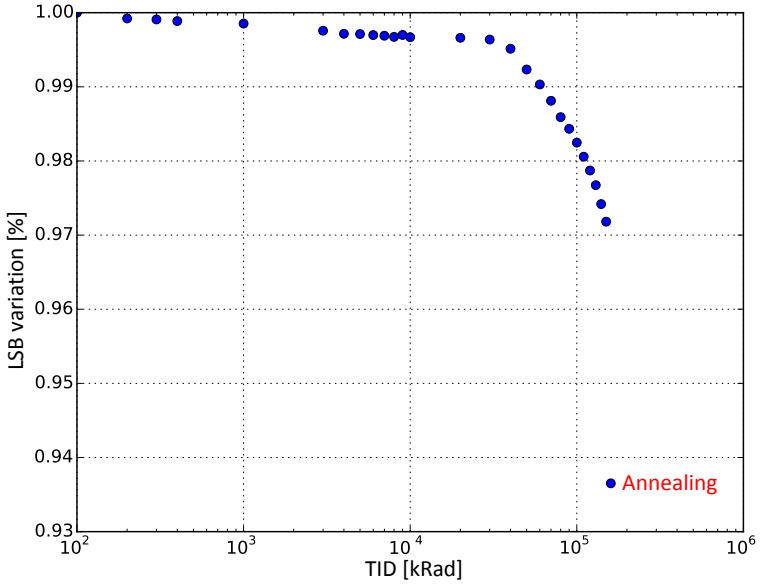


FIGURE 5.19: Calibration DAC voltage variation

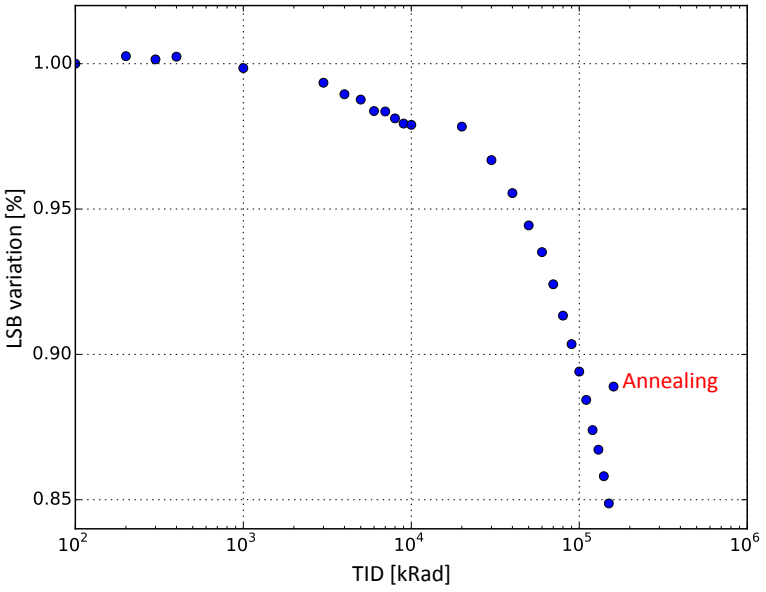


FIGURE 5.20: Threshold value variation extracted from s-curve

with minimal dimensions degrades in current of more than 40% already at 100 MRad. Since most of the standard cells in the digital libraries use minimum lenght transistor and a width < 200 nm, similar results could be expected also for the digital logic. However, during irradiation test transistors are kept in worst case condition of bias, while in a real application the transistors in a cell are not in the same condition. Therefore, several irradiations were carried out on the MPA-Light to explore the effect on the digital circuitry and compare it with the transistor results.

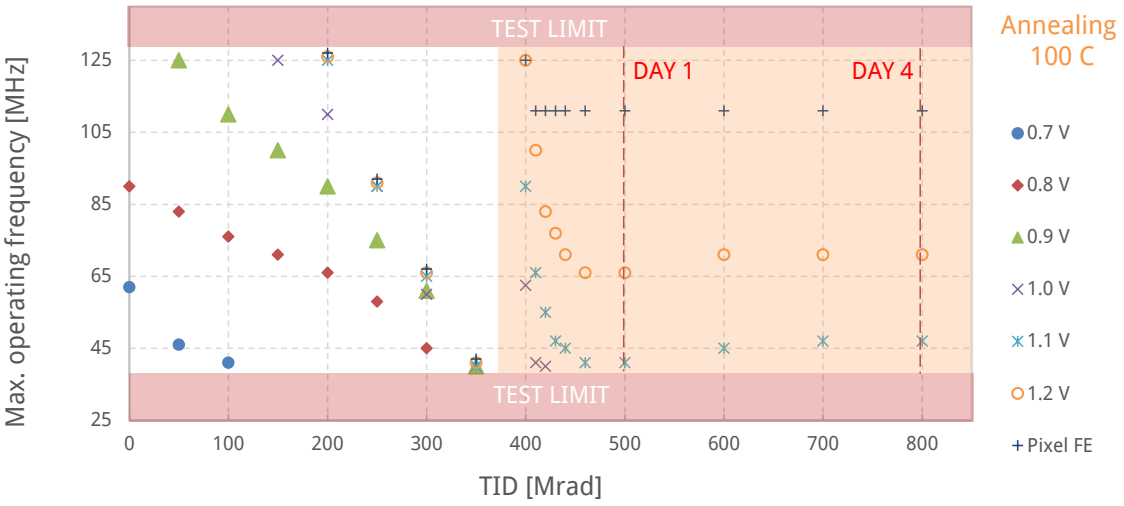


FIGURE 5.21: Degradation of the maximum operative frequency with TID. The orange coulored part shows the annealing at 100 °C.

In order to measure the degradation of the logic performance, the maximum operating frequency of a given data path is monitored. The part chosen for this test is the position encoder for the pixels because both the input and the output of this logic can be stored in the on-chip registers by changing the configuration. Due to a test system limitation, the maximum frequency provided to the MPA-Light is 500 MHz which is divided internally by 4 to get the operative frequency. Consequently, the maximum operating frequency is 125 MHz. Before the irradiation, the target logic does not fail at 125 MHz with nominal power supply, but it starts failing at 90 MHz when the supply is lowered down to 0.8 V.

In figure 5.21 the evolution of the maximum operating frequency is reported for different power supplies at room temperature. The measurement is stopped when the frequency goes below the nominal one, namely 40 MHz. Considering voltages lower than 1 V, the maximum operating frequency degrades linearly up to 300 MRad. This degradation is $\sim 15\,\%$ at 100 MRad. When the logic runs at higher voltage, the performance is limited by a second effect, probably caused from a different part of the design which has always the same power supply, since different supply curves show the same behaviour. Further investigations associate this degradation with the pixel front-end. At 400 MRad, the logic is not functional at 40 MHz, so the irradiation is stopped and the annealing started. The annealing temperature is 100 °C. The pixel front-end recovers part of the functionalities and shows stable performance during the annealing phase. Since the pixel front-end is working, the maximum frequency for supplies between 1 V and 1.2 V can be measured.

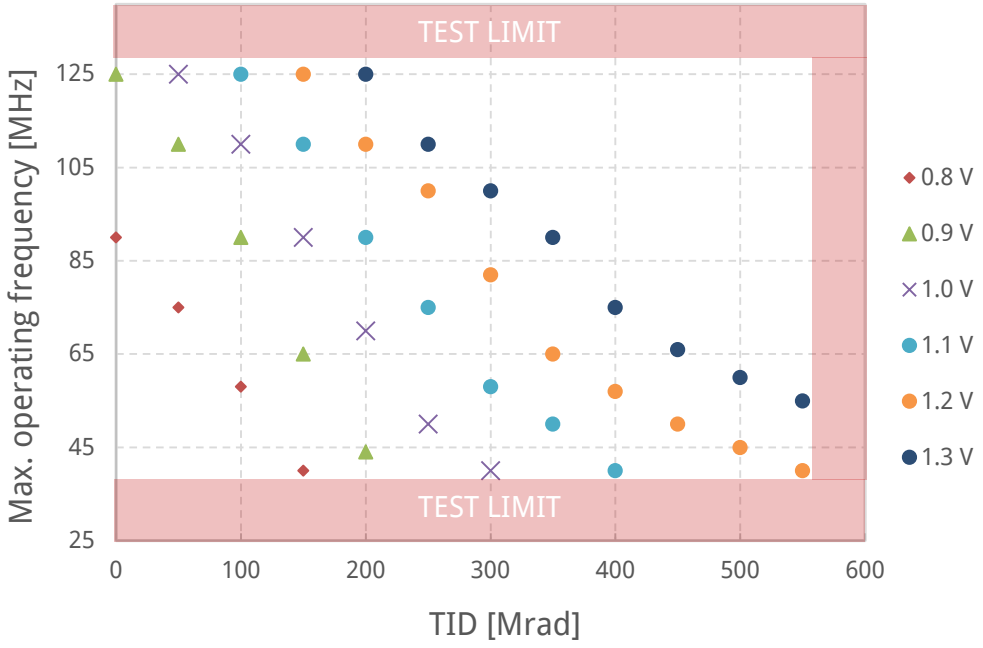


FIGURE 5.22: Degradation of the maximum operative frequency with TID at 100 °C.

These measurements show a still operating logic after 400 Mrad, but there is a further degradation during the first day of annealing.

In figure 5.22 the same measurements are reported for an irradiation at 80 °C. The choice of the temperature is limited by the test system which is not able to reach an higher temperature. The degradation at power supply lower than 1 V is double than the one measured at room temperature. The pixel front-end never fails at high temperature in the frequency range explored. Hence, the operation of the logic for supplies higher than 1 V can be observed up to 550 Mrad, and a linear degradation is also measured. The test stops at 550 Mrad due to I/O failures.

The results at different temperature provides important information for the comparison with the transistor results. The target logic shows higher degradation with irradiation at high temperature and further degradation in the first day of annealing, which is the behaviour observed in short channel transistors. On the contrary, the pixel front-end shows a lower degradation with irradiation at high temperature and a benefic effect of the annealing, which is the behaviour observed in narrow channel transistors [45]. Further investigation discovered the behaviour of the pixel front-end is related with delay cells for the reset of the edge detector in the pixel (see figure 5.2). The delays use long and

narrow channels which are not very representative of the standard logic cells. On the contrary, the effect observed on the target logic is representative of most of the standard logic cells where the short channel effect dominates.

One important achievement of these tests is the demonstration that the response of a digital circuit to TID can be correlated, at least qualitatively, with the degradation induced by radiation on individual transistors, both for the narrow and the short channels. Concerning the final application, the MPA-Light logic is fully functional also under low power supply at the expected dose. Furthermore, this test also provides an estimation of the margin to take during the design of the MPA: $\sim 15\,\%$ at 100 MRad.

5.5 Chapter summary

The MPA-Light chip has been designed using a commercial 65 nm CMOS technology with a macro pixel cell of 100 μm x 1446 μm. The Stub Finding logic for quick recognition of high transverse momentum particles was included and performed as expected. Front-end characterization with test pulses matched simulations closely, with a pixel-to-pixel threshold spread of 95 e^- r.m.s. after equalization, an ENC of 165 e^- r.m.s., a peaking time of 24 ns and a walk time < 15 ns.

This prototype has been also used to produce the MaPSA-Light assembly which consists in six readout ASICs with one pixelated sensor. This assembly has been characterized with test pulses and radiation sources, and it confirmed the performance of the MPA-Light with a bonded sensor. Furthermore, TID characterization proved the functionality of the MPA-Light at the expected dose of 100 Mrad and also provided important knowledge on the 65 nm technology.

In conclusion, the success of the MPA-Light is a very important step in the MPA project because it tests the complete front-end for the next iteration, and proves the feasibility of intelligent particle tracking in hybrid pixel detector with the given power budget.

Conclusions

This thesis was dedicated to the development of a readout chip with intelligent particle tracking capabilities for the future experiment on the High Luminosity LHC. This novel concept requires the front-end electronics to include the intelligence for discriminating particles based on their transverse momentum. The power density $< 100\,\mathrm{mW/cm^2}$ and the radiation level around 100 Mrad make this design a real engineering challenge.

The Stub Finding algorithm has been developed and simulated also with physics event from Monte Carlo simulations. The results have verified the functionalities and estimated an efficiency around 95% in the particle selection. Low power techniques as supply voltage scaling and memory gating have allowed to fit the power requirement for the digital logic. The use of rad-hard SRAM memory and dedicated design sLVS I/O have provided the storage capability and the high speed communication needed for the triggered and the event driven readout paths. Data format studies have optimized the available bandwidth minimizing the inefficiencies.

A first prototype in 65 nm have been designed, produced and tested. The electrical characterization has proven the functionalities of the full front-end analog chain and of a first version of the Stub Finding logic. Radiation tests have shown the functionalities of the desing up to doses higher than 300 Mrad. A first module production has provided important knowledge about the assembly procedures as well as a complete hybrid pixel detector for the sensor and readout chip characterizations with radioactive source and test beam.

Finally, the work carried out led to a complete model, fitting the all the requirements, which will be implemented in the coming years with the knowledge acquired from the successful prototyping iteration.

Bibliography

[1] CMS Collaboration, *Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC Physics Letters B*, Volume 716, Issue 1, 17 September 2012, Pages 30–61

[2] The HiLumi LHC Collaboration, *HL-LHC Preliminary Design Report*, CERN-ACC-2014-0300.

[3] G. Lutz. *Semiconductor Radiation Detectors.*, Springer Link, 2007.

[4] Claus Grupen and Boris Shwartz *Particle Detectors*, Cambridge monograph on particle physics.

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

[6] B. Oliver Sim, *LHC design report*, *CERN Document Server*, DOI: 10.5170/CERN-2004-003- V-1 (2004).

[7] The ATLAS experiment homepage, *www.atlas.ch.*

[8] The CMS experiment homepage, *cms.web.cern.ch.*

[9] The ALICE experiment homepage, *aliceinfo.cern.ch.*

[10] The LHCb experiment homepage, *lhcb.web.cern.ch.*

[11] 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, JUNE 2003

[12] F.Faccio et al., *Total Ionizing dose effects in shallow trench isolation oxides" Microelectronics Rliability*,48, (2008), 1000-1007.

[13] G. Anelli et Al. *Radiation Tolerant VLSI Circuits in Standard Deep Submicron CMOS Technologies for the LHC Experiments: Practical Design Aspects IEEE TRANSACTIONS ON NUCLEAR SCIENCE*, VOL. 46, NO 6, DECEMBER 1999

[14] R. Naseer. *The DF-DICE storage element for immunity to soft errors.* In Circuits and Systems. 2005. 48th Midwest Symposium on August 7-10, 2005 Page(s):303 - 306, 2005.

[15] CMS collaboration, *The CMS experiment at the CERN LHC*, 2008 *JINST*, 3, S08004.

[16] M. Raymond, G. Cervelli, M. French, J. Fulcher, G. Hall, L. Jones, L.-K. Lim, G. Marseguerra, P. Moreira, Q. Morrissey, A. Neviani, and E. Noah, *The CMS tracker APV25 0.25 μ*m *CMOS readout chip*, Proc. 6th Workshop Electronics for LHC Experiments , pp.130 -134 , 2000

[17] The High Luminosity project homepage, *hilumilhc.web.cern.ch.*

[18] M. Pesaresi, *Tracking trigger upgrade plans for CMS at SLHC*, PoS, vol. VERTEX2010, p. 047, 2010.

[19] D. Abbaneo and A. Marchioro, *A hybrid module architecture for a prompt momentum discriminating tracker at HL-LHC*, 2012 *JINST* 7 C09001.

[20] M. Pesaresi et Al, *Track Finding in CMS for the Level-1 Trigger at the HL-LHC*, CMS Conference Report, *CMS CR -2015/307*

[21] CMS collaboration, *Technical proposal for the Phase-II upgrade of the Compact Muon Solenoid*, CERN-LHCC-2015-010 / LHCC-P-008.

[22] Mauricio Garcia-Sciveres and Jorgen Christainsen, *RD Collaboration Proposal: Development of pixel readout integrated circuits for extreme rate and radiation*, CERN-LHCC-2013-008 ; LHCC-P-006

[23] Eichhorn, T for CMS Tracker collaboration. *Silicon strip sensor simulations for the CMS phase-II tracker upgrade*, Nuclear Science Symposium and Medical Imaging Conference (NSS/MIC), 2012 IEEE

[24] G. Blanchot et al., *Hybrid circuit prototypes for the CMS Tracker upgrade front-end electronics*, 2013 *JINST* 8 C12033.

[25] D. Abbaneo, *Upgrade of the CMS Tracker with tracking trigger*, 2011 *JINST* 6 C12065.

[26] Mussgiller et Al., *Detector Module R&D for the future CMS Tracker*, Forum on Tracking Detector Mechanics 2014.

[27] L. Gaioni et al., *Low-power clock distribution circuits for the Macro Pixel ASIC* 2015 *JINST* 10 C01051

[28] F. Krummenacher, *Pixel detectors with local intelligence: an IC designer point of view*, *Nucl. Instrum. Meth.*, A305 (1991) 527.

[29] D. Ceresa et al., *Macro Pixel ASIC (MPA): The readout ASIC for the pixel-strip (PS) module of the CMS outer tracker at HL-LHC*, 2014 *JINST* 9 C11012.

[30] A. Marchioro, *A hybrid module architecture for a prompt momentum discriminating tracker at SLHC*, PoS *(Vertex 2011)* 037.

[31] P. Fischer et al., *MEPHISTO – a 128-channel front end chip with real time data sparsification and multi-hit capability*, *Nucl. Instrum. Meth.*, A431 (1999) 134.

[32] R. Brouns et Al, *The development of a radiation tolerant low power SRAM compiler in 65nm technology.*, AMICSA 2014.

[33] I. Kremastiotis et al., *CERN IO Pad*, Internal IP documentation.

[34] G.Traversi et al., *Design of low-power, low-voltage, differential I/O links for High Energy Physics applications*, 2015 *JINST* 10 C01055.

[35] S.Viret, *https://sviret.web.cern.ch/*

[36] G.Bianchi, *tkLayout: a design tool for innovative silicon tracking detectors*, 2014 *JINST* 9 C093054.

[37] S. Mersi et al., *CMS Tracker Layout Studies for HL-LHC, TIPP 2011 - Technology and Instrumentation in Particle Physics*, Physics Procedia, 37 (2012) 1070-1078.

[38] D. Braga, D. Ceresa, M. Raymond, F. Vasey, S. Viret, Y. Zoccarato - *I/O data formats for the Concentrator Integrated Circuit* - 2014

[39] S. Viret - *Data transmission efficiency of the phase II tracker front-end system using new GBT transmission scheme - CMS Internal Note* IN-2015/XXX

[40] D. Felici et al., *A 20 mW, 4.8 Gbit/sec, SEU robust serializer in 65 nm for read-out of data from LHC experiments*, 2014 *JINST* 9 C01004.

[41] *Thomas Bergauer for the CMS Collaboration Silicon Sensor Prototypes for the Phase II Upgrade of the CMS Tracker* CMS Conference Report 2015/302

[42] F.Vasey, *CMS tracker electronics*, *ACES* '16

[43] *http://www.fluka.org/*

[44] F.Faccio, *Radiation hardness issues in 130nm and 65nm CMOS*, *ACES* '16

[45] F.Faccio and G.Cervelli, *Radiation induced edge effects in deep submicron CMOS transistors*, IEEE Trans Nucl Science, Vol.52, N.6, Dec2005, pp.2413-2420.