Development of the Readout Electronics for the High Luminosity Upgrade of the CMS Outer Strip Tracker

Davide Braga

High Energy Physics Group
Department of Physics
Blackett Laboratory
Imperial College London

A thesis submitted to Imperial College for the degree of Doctor of Philosophy and the Diploma of Imperial College.

January 2016
The copyright of this thesis rests with the author and is made available under a Creative Commons Attribution Non-Commercial No Derivatives licence. Researchers are free to copy, distribute or transmit the thesis on the condition that they attribute it, that they do not use it for commercial purposes and that they do not alter, transform or build upon it. For any reuse or redistribution, researchers must make clear to others the licence terms of this work.
Declaration

The work presented in this thesis is my own work, and the work of others is acknowledged or referenced.

Chapter 1 Through the Ph.D I have actively engaged in working meetings of the CMS Tracker Upgrade group, and I have had the opportunity to contribute to the evolution of the 2S-module, for example by liaising closely with the hybrid design team at CERN to identify the optimal floorplan for the integration with the chips, or by producing early proposals for the module data formatting.

Chapter 2 The CMS Binary Chip 2 inherited many of the circuits of the previous prototype, predating my involvement in the project, such as the pipeline control logic or the analogue channel chain, the latter designed by Mark Raymond. As the principal designer of the CBC2 I was responsible for floorplanning, circuit design and layout, top level integration and verification; the only exceptions being the test-pulse generator and the DC-DC converter. Some of the main tasks included adopting a bump bonding layout, doubling the number of channels and improving power and bias distribution, modifying the comparator hysteresis and implementing Mark’s improvements for the analogue chain. I was also responsible for every aspect of the stub-finding logic: cluster width discrimination, offset correction and correlation logic, inter-chip communication and test features (fast ORs, stub slow readout). With respect to the pipeline memory, I reduced the footprint of the SRAM cells and doubled the number of cells with minimal increase to the circuit area.

Chapter 3 I participated in the installation and data-taking shifts for the CERN H8 beam test of CBC1 modules. The electrical characterization of the CBC2 was performed by Mark Raymond at IC; my contribution was to verify the results through circuit simulation, and the thermal imaging of the CBC2.

Chapter 4 I was responsible for the coordination of the TID measurements, and together with Mark Raymond for setup and the data taking. I performed the dosimetry and calibration described in Appendix A, and ran all the different tests (room-temperature, low-temperature, masked and high-dose). I also analysed the data, interpreted the results and investigated them through simulations. While I was not directly involved with the SEU testing, I analysed the layout and the relevant parameters for the memory elements, and contributed to the interpretation of the results.

Chapter 5 The beam test of the CBC2 at DESY and the corresponding data analysis are the work of a large team. My individual contributions included assisting with the beam test installation and setup; daily data-taking shifts; organizing the database and playing a role in the coordination of the analysis; performing analysis of the cluster width distribution, hit efficiency and stub bend distribution. I also adapted the CMSSW Digitizer for the 2S module readout and performed the initial parametrization.

Chapter 6 I worked on the specification and architecture of the CBC3; designed and laid out a new stub-finding logic to include half-strip spatial resolution, stub-bend information, layer swapping logic and a new inter-chip communication scheme. I also proposed a new layout for the SRAM cell, to limit TID-induced leakage and to accommodate longer latency without expanding the chip dimensions.

Some of the work presented in this thesis resulted in the following publications: [1, 2, 3, 4, 5, 6].
Abstract

The High-luminosity upgrade of the LHC, scheduled for 2024, will deliver the dramatic increase in luminosity required for precision measurements and to probe Beyond the Standard Model theories. At the same time, it will present unprecedented challenges in terms of pileup and radiation degradation.

The CMS experiment is set for an extensive upgrade campaign, which includes the replacement of the current Tracker with another all-silicon detector with improved performance and reduced mass. One of the most ambitious aspects of the future Tracker will be the ability to identify high transverse momentum track candidates at every bunch crossing and with very low latency, in order to include tracking information at the L1 hardware trigger stage, a critical and effective step to achieve triggers with high purity and low threshold.

This thesis presents the development and the testing of the CMS Binary Chip 2 (CBC2), a prototype Application Specific Integrated Circuit (ASIC) for the binary front-end readout of silicon strip detectors modules in the Outer Tracker, which also integrates the logic necessary to identify high-$p_T$ candidates by correlating hits from two silicon strip detectors, separated by a few millimetres. The design exploits the relation between the transverse momentum and the curvature in the trajectory of charged particles subject to the large magnetic field of CMS.

The logic which follows the analogue amplification and binary conversion rejects clusters wider than a programmable maximum number of adjacent strips, compensates for the geometrical offset in the alignment of the module, and correlates the hits between the two sensor layers. Data are stored in a memory buffer before being transferred to an additional buffer stage and being serially read-out upon receipt of a Level 1 trigger.

The CBC2 has been subject to extensive testing since its production in January 2013: this work reports the results of electrical characterization, of the total ionizing dose irradiation tests, and the performance of a prototype module instrumented with CBC2 in realistic conditions in a beam test. The latter is the first experimental demonstration of the $p_T$-selection principle central to the future of CMS.

Several total-ionizing-dose tests highlighted no functional issue, but observed significant excess static current for doses <1 Mrad. The source of the excess was traced to static leakage current in the memory pipeline, and is believed to be a consequence of the high instantaneous dose delivered by the x-ray setup. Nevertheless, a new SRAM layout aimed at removing the leakage path was proposed for the CBC3. The results of single event upset testing of the chip are also reported, two of the three distinct memory circuits used in the chip were proven to meet the expected robustness, while the third will be replaced in the next iteration of the chip.

Finally, the next version of the ASIC is presented, highlighting the additional features of the final prototype, such as half-strip resolution, additional trigger logic functionality, longer trigger latency and higher rate, and fully synchronous stub readout.
Acknowledgements

First and foremost, I would like to thank my supervisor Prof. Geoff Hall for the opportunity to work on such a stimulating project, for the advice and guidance, and for his constant support and encouragement.

I would also like to express my gratitude to Kirika, Andrew, and the High Energy Physics group at large for the good time at Imperial College. A special mention goes to Mark Raymond, great designer and mentor, from whom I have learnt a great deal. I am also extremely grateful to Mark Pesaresi for taking me under his wing and for the invaluable help in navigating the CMSSW waters, and for the good company.

My gratitude also goes to Andrei Nomerotski, my supervisor at Oxford University before starting at IC, and to Todd Hoffman for the help and advice.

This enterprise would have been impossible without the substantial support of my managers at RAL, Mark Prydderch and Marcus French, to whom I am extremely grateful for being so accommodating, and for endorsing my ambitious plans. The ASIC and CMOS Sensors groups at RAL are a professional, friendly and collaborative work environment of wonderful people, too numerous to name, who have been like second family for many years. I am thankful to the other designers who contributed to the CBC2: Mark Prydderch, Mark Raymond and Peter Murray, and to Lawrence Jones who designed the CBC1. I am also grateful to the Diamond Light Source Detector Group for providing equipment and assistance for the TID test, and to the Engineering Instrument Pool group for the thermal camera.

I would like to acknowledge the rest of the Phase II Tracker Upgrade collaboration for the friendly and prolific meetings and discussions; as well as the beam test and data analysis teams.

A big thank you also goes to all my housemates and friends.

Above all, I am most grateful to my family: to my parents for always being there for me; to the kids, for putting up with my \textit{maybe-once-the-PhD-is-over} mantra; to my little baby, for charging my batteries with every smile, and to my wife and muse, for her unwavering support, for believing in me and sacrificing so much for me.

Dearest F, this is for you
Contents

Declaration 2

Abstract 3

Acknowledgements 4

Contents 5

List of Figures 11

List of Tables 17

Chapter 1. Introduction to the HL-LHC and CMS 19

1.1 The High Luminosity Large Hadron Collider 19

1.1.1 Physics Opportunities 20

1.1.2 LHC and CMS Upgrades 21

1.1.3 Challenges and Requirements 23

1.2 Track Trigger 27

1.2.1 $p_T$-Selection Principle 29
## Contents

1.3  Tracker Geometry 31

1.4  Module Design 33
  1.4.1  Module Readout 36

1.5  Performance Estimate 38

1.6  Summary 42

## Chapter 2.  Front-End ASIC Development 43

2.1  Choice of Technology: 130 nm CMOS Process 43

2.2  Choice of Architecture 48
  2.2.1  APV “Digital” 48
  2.2.2  CMS Binary Chip 49
  2.2.3  CBC1 51

2.3  CBC2 52
  2.3.1  Architecture 54
  2.3.2  Front-End 54
  2.3.3  Stub-finding Logic 65
  2.3.4  Data Readout 73
  2.3.5  On-chip Powering Circuit 75
  2.3.6  Bump Bonding 76

2.4  Summary 78
Chapter 3. CBC2 Measurements and Performance 80

3.1 Test Activities 80

3.1.1 Wafer Probing 81

3.1.2 Single Die Testing 83

3.1.3 Dual-CBC2 Hybrid 83

3.1.4 Mini 2S-Module Beam Test 84

3.1.5 Eight-CBC2 Hybrid 85

3.2 Analogue Test 86

3.2.1 Test Pulse Circuit 86

3.2.2 Noise and Power 86

3.2.3 Gain 87

3.2.4 Offset Dispersion and Threshold Tuning 90

3.2.5 On-Chip Power Elements 91

3.3 Stub-Finding Logic Test 94

3.3.1 Beta-Source Test 95

3.3.2 Test with Cosmic Rays 95

3.4 Summary 96

Chapter 4. Radiation Effects and Testing 97

4.1 Radiation Damage at the HL-LHC 97

4.2 Radiation Effects in CMOS Electronics 98

4.2.1 Radiation-Induced Effects and Mitigation Techniques 100
<table>
<thead>
<tr>
<th>Contents</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.2.2 130 nm CMOS Radiation Hardness</td>
<td>107</td>
</tr>
<tr>
<td>4.3 Total Ionizing Dose Irradiation Test</td>
<td>108</td>
</tr>
<tr>
<td>4.3.1 X-ray Testing</td>
<td>108</td>
</tr>
<tr>
<td>4.3.2 Experimental Setup</td>
<td>109</td>
</tr>
<tr>
<td>4.3.3 Dosimetry</td>
<td>110</td>
</tr>
<tr>
<td>4.3.4 Biasing</td>
<td>110</td>
</tr>
<tr>
<td>4.3.5 Annealing</td>
<td>112</td>
</tr>
<tr>
<td>4.4 Room-Temperature Irradiation</td>
<td>113</td>
</tr>
<tr>
<td>4.5 Low-Temperature Irradiation</td>
<td>116</td>
</tr>
<tr>
<td>4.6 Masked Irradiation</td>
<td>118</td>
</tr>
<tr>
<td>4.7 High-Dose Irradiation</td>
<td>121</td>
</tr>
<tr>
<td>4.8 TID Irradiation Results</td>
<td>122</td>
</tr>
<tr>
<td>4.8.1 Origin of the Leakage Current</td>
<td>125</td>
</tr>
<tr>
<td>4.8.2 Effect of Analogue Biasing</td>
<td>130</td>
</tr>
<tr>
<td>4.8.3 Static Noise Margins</td>
<td>132</td>
</tr>
<tr>
<td>4.8.4 Difference for Test-Pulse and Pedestal S-Curves</td>
<td>134</td>
</tr>
<tr>
<td>4.9 Single Event Upsets</td>
<td>137</td>
</tr>
<tr>
<td>4.9.1 Memory Elements in the CBC2</td>
<td>142</td>
</tr>
<tr>
<td>4.9.2 SEU Testing</td>
<td>146</td>
</tr>
<tr>
<td>4.10 Conclusions on Irradiation</td>
<td>151</td>
</tr>
</tbody>
</table>
Chapter 5. Performance Studies 153

5.1 CBC2 Beam Test 153
   5.1.1 Experimental Setup 153
   5.1.2 Analysis and Results 159

5.2 Digitizer Studies 169
   5.2.1 Conclusions 171

Chapter 6. Design Modifications for the Final CBC 172

6.1 CBC3 172
   6.1.1 CBC3 Overview 172
   6.1.2 Front-End 174
   6.1.3 CBC3 Logic 175
   6.1.4 Stub Gathering Logic and Readout 182
   6.1.5 Control Interface 183
   6.1.6 L1 Trigger Memory 183
   6.1.7 L1 Trigger Memory Buffer 185

6.2 Summary 185

Conclusions 187

Appendix A. TID Dosimetry 188
   A.0.1 Dose Calculation 188
   A.0.2 Dosimetry Diode 194
   A.0.3 Uncertainty 195
List of Figures

1.1 Schematic representation of the CMS experiment. 20
1.2 LHC run and upgrade plan. 21
1.3 Quarter section view of the CMS Phase II Tracker. 23
1.4 $p_T$-cut working principle. 29
1.5 Relation between the transverse momentum $p_T$ and the crossing angle $\phi$. 30
1.6 Sensor spacings adopted in the Tracker. 31
1.7 Comparison between the Phase II upgrade and the current Tracker: number of hits and material budget. 33
1.8 Exploded view of the 2S module. 34
1.9 Prototype of the 2S module Front-End Hybrid, with 8 CBC2s. 36
1.10 2S module electronic system and data readout. 37
1.11 Cluster-to-stub ratio and stub multiplicity for barrel and endcap modules. 39
1.12 Efficiency and rates for single muon and electron triggers with track trigger information. 40

2.1 Schematic and SEM section of the 130 nm CMOS process used by the CBC. 46
List of Figures

2.2 Relative size of the two CMS Binary Chip prototypes. 50
2.3 CBC2 block diagram, arranged to reflect the layout of the chip. 55
2.4 CBC2 architecture. 56
2.5 CBC front-end readout channel. 56
2.6 Charge sensitive preamplifier. 57
2.7 Effect of the input transistor polarity on noise injection from the detector biasing. 60
2.8 CBC gain amplifier. 61
2.9 Schematic of the CBC2 comparator. 63
2.10 Effect of the Hit-Detect logic at the output of the comparator. 65
2.11 CBC2 stub-finding logic. 66
2.12 CBC2 Cluster Width Discriminator block. 68
2.13 Coincidence window width settings across the Tracker volume. 68
2.14 CBC2 Φ-shift correction. 69
2.15 Detail of the layout of front-end channels and coincidence logic. 72
2.16 Detail of the stub-finding logic. 72
2.17 Block diagram of the CBC2 readout control logic. 73
2.18 CBC2 L1 data format. 74
2.19 Power supply distribution nets inside the CBC2. 76
2.20 C4 bump metallisation and parasitic capacitance. 78
2.21 Photograph of C4 bumps on a CBC2 wafer. 79

3.1 Wafer probe station and custom probe card. 81
List of Figures

3.2 Results of wafer screening and yield. 82
3.3 CBC2 test setups. 83
3.4 Mini-2S test module. 84
3.5 Eight-CBC2 hybrid. 85
3.6 Readout noise vs. strip capacitance for electron and hole readout. 87
3.7 S-curve mid-points for all the 254 channels, as a function of the test pulse charge. 88
3.8 Test pulse amplitudes and pedestals for the mini 2S-modules CBC2s. 88
3.9 CBC2 channel threshold tuning. 89
3.10 S-curve for a binary system, and for the CBC2. 90
3.11 Threshold dispersion for all the 254 channels (h+ readout), before and after calibration. 91
3.12 Effect of the DC-DC converter switching activity on the chip noise. 93
3.13 Thermal images of a CBC2. 94
3.14 \( \beta \) source test setup and hit profile. 95
3.15 Arrangement of the test-pulse groups of channels and output data waveforms. 96
4.1 Monte Carlo estimation of radiation dose for the Tracker, at 3000 fb\(^{-1}\). 99
4.2 Parasitic lateral transistors caused by positive charge trapped in the STI. 103
4.3 Elements used in the calculation of equivalent width and length of an enclosed-layout transistor. 104
4.4 Transistor cross-section and parasitic Silicon Controlled Rectifier (SCR) structure responsible for latchup. 107
List of Figures

4.5 Interior of the irradiation cabinet and positioning of the spectrometer. 111
4.6 Dosimetry diode and silicon drift spectrometer. 111
4.7 Accelerated annealing run and DAQ. 111
4.8 Results for the room-temperature irradiation of the CBC2. 114
4.9 Results for the room-temperature irradiation of the CBC2: accelerated annealing. 116
4.10 Analogue biases during room temperature irradiation. 117
4.11 CBC2 setup for TID irradiation. 118
4.12 Results of two cold irradiation test of the CBC2. 119
4.13 S-curves for cold irradiation of the CBC2. 120
4.14 Results of the high-dose irradiation test of the CBC2. 121
4.15 Digital current for different irradiation conditions. 124
4.16 Digital current for different irradiation conditions. 126
4.17 SRAM layout highlighting possible inter-device leakage paths, the two internal nodes A and B, as well as the internal symmetry. 127
4.18 Effect of biasing on the I-V curves of nMOS devices. 131
4.19 SRAM cell: effect of the stored value on the TID-induced leakage current. 132
4.20 Simulation of SRAM static noise margins. 133
4.21 Equivalent circuits highlighting the leakage paths for read operation, and the effect on the output data line. 135
4.22 Simulation of the degradation of the pipeline readout lines during irradiation. 138
4.23 Critical charge simulation for the majority-voting register. 140
List of Figures

4.24 Critical charge extracted from the pulses of Fig.4.23a. 140
4.25 Circuit schematic of one of the two flavours of majority voting registers used for the configuration data (MVR-Set). 145
4.26 Circuit schematic of the SEU-hardened flip-flop used for the pipeline control logic. 146
4.27 CBC2 current, beam flux and calculated TID dose during SEU testing. 148
4.28 Layout of the TMV-SRAM register. Highlighted are the three couples of storage nodes and their separations. 150

5.1 Top view of the detector planes. 154
5.2 Generation of electron and positron beams at DESY. 155
5.3 Stopping power for electrons in silicon. 156
5.4 Hit profiles for the four sensors. 156
5.5 Beam test module and DAQ. 157
5.6 Distribution of cluster size as a function of the angle of incidence. 160
5.7 Cluster width: data and simulation. 161
5.8 Geometrical displacement of tracks within the thickness of the sensor. 161
5.9 Beam divergence for three angles of incidence. 163
5.10 Efficiency of one of the CNM sensors as a function of threshold (VCTH) for different angles of incidence. 164
5.11 Stub bend as a function of module rotation. 167
5.12 Stub-finding efficiency and derived $p_T$ cut efficiency. 168
5.13 CMSSW Digitizer drift principle. 170
<table>
<thead>
<tr>
<th>Figure Number</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.1</td>
<td>Block diagram of the CBC3 ASIC.</td>
<td>173</td>
</tr>
<tr>
<td>6.2</td>
<td>Effect of the module flipping in one stave on the stub coincidence window.</td>
<td>176</td>
</tr>
<tr>
<td>6.3</td>
<td>CBC3 Cluster Width Discriminator block.</td>
<td>177</td>
</tr>
<tr>
<td>6.4</td>
<td>CBC3 programmable look-up table for stub bend mapping.</td>
<td>179</td>
</tr>
<tr>
<td>6.5</td>
<td>Detail of the layout of the CBC3 logic.</td>
<td>180</td>
</tr>
<tr>
<td>6.6</td>
<td>CBC3 output data format.</td>
<td>182</td>
</tr>
<tr>
<td>6.7</td>
<td>Comparison between an 8x2 SRAM block for the CBC2 and CBC3.</td>
<td>184</td>
</tr>
<tr>
<td>6.8</td>
<td>Trigger losses vs. L1 rate, for different buffer depths.</td>
<td>186</td>
</tr>
<tr>
<td>A.1</td>
<td>Calculation of the dose delivered to the sensitive volume of the CBC2 during TID irradiation.</td>
<td>189</td>
</tr>
<tr>
<td>A.2</td>
<td>Acquired spectrum.</td>
<td>192</td>
</tr>
<tr>
<td>A.3</td>
<td>X-ray spectra from the molybdenum source, as corrected for the dosimetry analysis.</td>
<td>192</td>
</tr>
<tr>
<td>A.4</td>
<td>Linear attenuation coefficients for Si and SiO$_2$ and effect of BEOL adsorption.</td>
<td>193</td>
</tr>
<tr>
<td>A.5</td>
<td>Profile of the current in the dosimetry diode scanned across the irradiation area. The CBC2 is superimposed.</td>
<td>195</td>
</tr>
<tr>
<td>A.6</td>
<td>BEOL cross-section and effect on total transmission coefficient.</td>
<td>197</td>
</tr>
<tr>
<td>B.1</td>
<td>Comparison between the ESD protection devices of the CBC and of the APV25.</td>
<td>200</td>
</tr>
<tr>
<td>B.2</td>
<td>Layout comparison of the input protection diodes for the APV25 and the CBC2.</td>
<td>200</td>
</tr>
</tbody>
</table>
List of Tables

1.1 Comparison between the main parameters of the current Outer Tracker and its Phase II upgrade. 32

2.1 Comparison between the electrical properties of typical wire and bump bonds (from [68]) 77

4.1 Increase in static current consumption resulting from the resistors simulating the TID-induced leakage path across the nMOS devices in the SRAM. 130

4.2 Effect of resistors, simulating the TID-induced leakage path across the nMOS devices in the SRAM, on the correct behaviour of the pipeline. 135

4.3 SEU strategy and circuits used for the CBC2 registers. 143

4.4 SRAM cell: SEU-relevant parameters. 144

4.5 Majority-voting SRAM (Reset): SEU-relevant parameters. 145

4.6 Number of SEU events and respective cross-section for the Inter-Integrated Circuit bus (I²C) majority-voting registers (from [128]). 149

5.1 Major characteristics of the sensors employed for the beam test. 158

5.2 Hit efficiency for the CNM sensor, for a threshold of 6440 $e^-$ (VCTH=120) and for several angles of incidence. 165
A.1  TID for diode and BEOL thicknesses at $\pm 6\sigma$ (normalized against nominal value).
Chapter 1

Introduction to the HL-LHC and CMS

1.1 The High Luminosity Large Hadron Collider

Having recently restarted operations at the record centre-of-mass energy of 13 TeV, the Large Hadron Collider (LHC) and its detectors are exploring uncharted territory in the quest for a deeper understanding of Nature.

The first run of the LHC, from 2009 to 2013, was extremely successful and saw, amongst a wealth of new results, the discovery of the long-sought Higgs boson by the two general purpose detectors, the Compact Muon Solenoid (Fig.1.1) and A Toroidal LHC ApparatuS (ATLAS) [7, 8]. The LHC is now poised for a series of upgrades that will extend its discovery potential by increasing the luminosity to $5 \times 10^{34}\text{cm}^{-2}\text{s}^{-1}$ (baseline), and possibly as high as $7.5 \times 10^{34}\text{cm}^{-2}\text{s}^{-1}$ (“ultimate” luminosity), a phase known as High-Luminosity LHC (HL-LHC).

To cope with the resulting unprecedented operating conditions, the experiments will also require substantial maintenance and improvements. This chapter includes a brief overview of the main upgrade activities for CMS, focusing on the development of a new Tracker detector and on the requirements for the Strip Tracker front-end readout ASIC which is the subject of this thesis.
Figure 1.1: Schematic representation of CMS experiment in its current Run 1 configuration, highlighting the main subdetectors.

1.1.1 Physics Opportunities

The exploitation of the rich discovery potential of the HL-LHC has been identified in both Europe and the U.S. as the highest strategic priority for particle physics [9, 10]. The new data will enable both precision measurements and direct Beyond the Standard Model (BSM) searches.

Precision measurements improve the Standard Model (SM) of particle physics, while at the same time providing a window into possible BSM phenomena. The study of the newly discovered Higgs boson will of course be of paramount importance: CMS has already established that its spin and $0^+$ parity are compatible with the current theoretical expectations [11]; a much larger dataset of 3000 fb$^{-1}$ will however allow to measure precisely nearly all its couplings, including rare production and decay processes such as self-coupling and couplings to $\tau$ and $\mu$, the latter with a branching ratio of only $10^{-4}$. 
Any deviation from the expected values of rare decays which are well described within the SM, would in fact indicate enhancing or suppression by new physics processes. An example of such suppressed channel already investigated during Run 1 is the decay $B_s \to \mu^+\mu^-$ [12]; the HL-LHC will give access to the branching ratio of the even rarer $B_d \to \mu^+\mu^-$. 

Examples of direct searches for exotic processes are the detection of states with large missing transverse energy, which could indicate the production of Weakly Interacting Massive Particles (WIMPS) in the final state, as well as the signature of the lightest supersymmetric particles. The discovery reach for new heavy gauge bosons will also extend to 6 TeV or more; while the search for extra dimensions and other exotica will probe regions up to a few TeV [13].

### 1.1.2 LHC and CMS Upgrades

The current plan for the operation and upgrade of the LHC is presented in Fig.1.2. It includes a series of long data-taking periods ($Runs$), and long shut-downs ($LS$) for maintenance and upgrade.

Run 1 completed in 2012, and the following LS1 allowed the machine to reach a centre-of-mass energy of 13 TeV and to reduce the bunch spacing from 50 to 25 ns. The current Run 2 will see CMS reach its original pileup design figure of 25
inelastic interactions per Bunch Crossing (BX), and should yield approximately 150 \( fb^{-1} \). Further upgrades of the LHC injection chain will allow the luminosity to reach \( 2 \times 10^{34} \text{cm}^{-2}\text{s}^{-1} \) during Run 3, for a projected \( 300 \, fb^{-1} \) by 2024.

After this, the experiment focusing magnets will need to be replaced during LS3, and the new HL-LHC configuration will provide a substantial increase in instantaneous luminosity, with a potential peak value of \( 2 \times 10^{35} \text{cm}^{-2}\text{s}^{-1} \). Depending on the filling and levelling scheme of the machine, the integrated luminosity at the end of operations will range between 3000 and 4000 \( fb^{-1} \).

Two major upgrades, known as Phase I and Phase II, will allow CMS to operate during Run 3 and the HL-LHC stages respectively.

**Phase I Upgrade:** an extensive program which, between LS1 and LS2 will see the upgrade of the Muon, Calorimeters and Pixel systems among others. The addition of a fourth layer of Cathode Strip Chambers and endcap Resistive Plate Chambers will increase coverage and reduce the accidental trigger rate for the Muon system. Hybrid photo diodes will be replaced with superior silicon photomultipliers in the Electromagnetic Calorimeter (ECAL), which will have higher segmentation, and the front-end electronics of the Hadronic Calorimeter (HC) will be upgraded. The Pixel Detector will also be replaced during LS2 with a CO2-cooled, 4-layer barrel and 3-disk endcap system with improved tracking, vertexing and b-tagging capabilities. A detailed description of the program can be found in the Phase I Technical Proposal [15].

**Phase II Upgrade:** at the end of Run 3, radiation damage and more challenging operating conditions will require a total or substantial replacement of many of CMS subdetectors.

The Strip Tracker, which was designed to withstand \( 500 \, fb^{-1} \), will have reached the end of its lifetime, and will need replacing with a new system able to operate under much higher event pileup and radiation. The Phase II Outer Tracker (Fig.1.3) is the focus of the following sections, where it will be discussed in more details, together with the related Trigger system.
1.1 The High Luminosity Large Hadron Collider

Figure 1.3: Quarter section view of the CMS Phase II Tracker. Three distinct regions are clearly visible: the inner Pixel Detector, which has extended coverage in $\eta$ and one additional layer; the Inner Strip Tracker with pixel-strip (PS) modules ($R>20$ cm, in blue), and the Outer Strip Tracker with strip-strip (2S) modules ($R>60$ cm, in red).

In the ECAL, while the barrel crystals will be able to continue operation, requiring only the replacement of the front-end electronics to provide higher readout bandwidth and contribute to the hardware trigger, a completely new High Granularity Calorimeter with hadronic and electromagnetic sections will be installed in the forward regions. Data acquisition and computing resources will also undergo major improvements to cope with the data and bandwidth of the HL-LHC era. A comprehensive description of the various aspects of the Phase II upgrade program is provided in the Technical Proposal and Upgrade Scope documents [13, 14].

1.1.3 Challenges and Requirements

To fully benefit from the increase in luminosity and centre of mass energy, it is essential that the future CMS continues to reconstruct all the standard physics analysis objects with high efficiency and background rejection, low fake rate and high resolution.

The primary goal of the upgrade is to maintain the Phase I performance of the detector through the HL-LHC operation, preserving efficiency, resolution and background...
rejection rates in spite of the drastic increase in irradiation and pileup brought by the increase in luminosity. Given the heterogeneous composition of the detector, these translate into specific requirements for each of the experiment’s highly-specialized detectors and systems; while only the Tracker is discussed in this work, a comprehensive description of the scope of the upgrades for the various sub-detectors can be found in Ref.[13].

Irradiation

It is certain that the current CMS Tracker will not be able to survive irradiation beyond Phase I and will have to be completely replaced before Phase II.

Aside from the contribution due to beam halo and beam gas interactions, the detector irradiation is in fact a linear function of the integrated luminosity, and with one year of HL-LHC operation delivering a dose comparable to the entire irradiation prior to the Phase II upgrade, it is clear how the future Tracker must be significantly more radiation hard.

The present Outer Tracker was designed to operate up to an integrated luminosity of 500 fb\(^{-1}\) without loss of efficiency, and with a maximum average pileup of 50 collisions/BX. Simulations of radiation-induced degradation and leakage current up to 1000 fb\(^{-1}\) at the minimum operating temperature of -20\(^\circ\)C predict the loss of almost all of the current stereo layer in both barrel and endcap regions [13].

While easy access to the Pixel detector will be maintained to allow for future repairs, or even replacement, the Strip Tracker must be radiation tolerant to 3000 fb\(^{-1}\).

For silicon detectors, radiation damage causes defects in the lattice which reduce the charge collection efficiency (thereby reducing the signal) and increase the leakage current. Aside from the direct degradation of the signal due to the leakage shot noise, the performance of the detector is degraded by the increase in material required to deliver the dissipated power and extract the consequent heat.
For the Outer Tracker, n-in-p strips have been selected, as they degrade gradually after heavy irradiation, without the type inversion observed in other options. The active thickness has been reduced from 300 to 200 µm to reduce leakage current and operating voltage while providing adequate signal charge; together with this preferred option, additional inactive material up to a total thickness of 320 µm is being considered to limit the fabrication cost.

Pileup

Pile-up (PU) refers to the overlap in space or time of signals coming from distinct collisions. Two forms of pileup can affect the event reconstruction. In-time pileup is caused by the fact that each BX typically results in multiple collisions. Most of the collisions are soft or low-energy collisions which do not contribute to the search for new physics and are considered background, as opposed to so-called hard collisions which result in high transverse-momentum ($p_T$) partons and decay products originating from high-mass primary particles.

In-time PU is already the dominant source of hits in the Tracker, and is poised to increase dramatically from the initial CMS design figure of 25 average PU, to the projected 140 average PU of the HL-LHC (possibly as high as 200, depending on the filling scheme of the machine).

Out-of-Time (OOT) pileup refers to the overlap of signals from different BXs. Although OOT pileup typically affects signal formation in the calorimeters, even if the Tracker has a signal collection and processing time sufficiently shorter that the BX interval, the strong magnetic field combined with the multitude of primary and secondary low energy particles creates so-called loopers, particles confined within the Tracker volume for more than one BX which are a significant source of OOT PU.

By increasing the complexity of the event signature, pileup increases the fake rate and makes it harder to reconstruct the primary interactions, while at the same time requiring larger bandwidth for data readout and longer latency for trigger processing.

One strategy to cope with high pileup is requiring good vertexing capability along the direction of the beam, given that in CMS the colliding bunches have
a spatial spread of 5 cm (RMS) in this dimension (z). To this end, the three innermost layers of the Outer Tracker (OT) will have a combination of shorter strips (2.5 cm long) and 1.5 mm long pixels.

The Tracker extension in $\eta$ is also designed to mitigate the effect of pileup in the calorimeter by greatly improving jet identification in the peak production region for vector boson fusion and scattering processes, as well as improving the measurements of total and missing energy [14].

To obtain occupancies similar to the current level, the segmentation of the Tracker will increase substantially from the current total channel count of 9.6 M strips, to 47.8 M strips and 218 M long pixels. This will also help by reducing the contribution of the leakage current to the noise after irradiation.

Overall, the main requirements for the high luminosity upgrade of the Tracker can be summarized as:

- increased granularity, to maintain the occupancy below 1% across all Tracker regions;
- radiation tolerance up to 3000 fb$^{-1}$ with no maintenance for the Outer Tracker;
- operation up to 200 PU;
- higher hardware trigger rate (750 kHz, compared to the current 100 kHz) and longer latency (12.5 $\mu$s, compared to the current 4 $\mu$s) (the Trigger is discussed in the following section);
- extended coverage to $\eta = 4$;
- reduced material budget;
- synchronous processing and readout of L1 trigger primitives from the Outer Tracker, compatible with the increased L1 rate and latency.
1.2 Track Trigger

The current event storage rate is limited to $\sim 1$ kHz. To achieve the required reduction of $\mathcal{O}(10^4)$ from the collision rate of 40 MHz, CMS relies on a two-level trigger architecture. The Level 1 Trigger (L1) ($40 \mathrm{MHz} \rightarrow 100 \mathrm{kHz}$) is implemented in hardware, and is limited to information from the muon and calorimeter systems. The following High Level Trigger (HLT) ($100 \mathrm{kHz} \rightarrow 1 \mathrm{kHz}$) performs a full-reconstruction similar to the offline analysis on a large computer grid.

In Phase II, even considering the Phase I trigger upgrades, the rates for isolated muon, electron and jet triggers at L1 will greatly exceed the current acceptance rate of 100 kHz, and cannot be reduced sufficiently by raising the $p_T$ thresholds within reasonable values. The muon trigger is a prime example of how increasing the $p_T$ threshold alone would not be sufficient for a high purity and efficient trigger. Due to tails in the momentum resolution of the current system, low $p_T$ muons are occasionally reconstructed as high momentum tracks (Fig. 1.12b). Given the high rate of muon production, this effect becomes considerable, and results in a flattening of the trigger curve for thresholds above 20 GeV. As discussed in Section 1.5, the addition of track information at L1 completely resolves this issue, and also sharpens the momentum resolution.

CMS already relies on information from the Tracker for the so-called Particle Flow reconstruction algorithm implemented in the HLT, in which the precise position and momentum resolution of the Tracker are combined with informations from the calorimeters and muon chambers to reconstruct individually each decay product, improving jet isolation, missing transverse energy measurement and lepton identification.

One of the major, and arguably the most innovative goal of the upgrade is to introduce the use of tracking information already at L1, thus enabling more powerful trigger algorithms which will benefit form the excellent momentum and position resolution of the Tracker [16, 17, 18].
This ambitious goal has profound implications on the design of the Tracker, and comes with considerable challenges. The most obvious is that it will not be possible to read out all the hits at 40 MHz, and therefore data reduction for the L1 trigger primitives will be essential. Unlike ATLAS, which plans to implement a region-of-interest readout following a first L0 hardware trigger [19], in CMS the reduction will be performed by the Tracker front-end readout. The goal is for the Strip Tracker modules to reject hits associated with tracks with $p_T < 2$ GeV, considered low energy background, and to provide L1 trigger primitives at every BX (40 MHz) with minimum latency.

The target $p_T$ threshold was chosen to provide sufficient background rejection, without affecting the physics reach of the detector. For example in the case of the Higgs studies, given its relatively low mass, it is essential for the hardware trigger to maintain a low threshold to achieve reasonable selection efficiency for all the decay products. Simulations show that about 90% of the tracks in minimum bias events have $p_T < 1$ GeV/c, with 97% below 2 GeV/c [16, 17]. A value of 2 GeV therefore translates into a data reduction of approximately one order of magnitude, by rejecting most of the background particles which fail to reach the outer part of the detector in the strong magnetic field of CMS and are therefore not useful for triggering.

To allow sufficient time for the additional on-detector hit processing, as well as for track reconstruction and matching with the muon and calorimeter objects, the L1 latency will be extended from the current 3.4 $\mu$s to 12.5 $\mu$s. This requires the upgrade of the readout electronics, which must store the full hit data until a L1 trigger requests readout.

The trigger rate will also be increased: assuming an online event selection of $10^{-2}$ for the HLT and a pileup of 140, a L1 trigger acceptance rate of 500 kHz is considered sufficient to maintain thresholds comparable to those used in Phase I across the current trigger menu. Given the uncertainty on the final pileup, the Tracker will be designed for a maximum L1 rate of 750 kHz, which should allow comparable performance up to 200 PU.
1.2 Track Trigger

Figure 1.4: $p_T$-cut working principle (from [13]). a) Sketch of the module’s two strip sensors, with a separation between 1 and 4 mm and with strips aligned along $z$. A programmable correlation window on the outer sensor identifies when two hits form a valid trigger primitive candidate, called stub. For a given $p_T$, the relative offset between hits in the two sensors $\Delta R$ increases with radius $R$. In the barrel, the offset in the transverse plane is given directly by the sensors spacing (b), while in the endcaps it depends on the position of the module ($\tan \theta$) (c).

1.2.1 $p_T$-Selection Principle

$p_T$ selection, or $p_T$-cut, exploits the correlation between a charged particle’s transverse momentum and the curvature of its trajectory within the strong magnetic field of the superconducting solenoid of CMS.

The $p_T$ selection and the required data processing will be performed at the front-end of each module. The working principle is presented in Fig.1.4: each module comprises of two silicon strip sensors closely spaced, with strips aligned along the direction of the beam ($z$). The hits of the two sensors are correlated in the front-end readout ASICs: if the hits on the two sensors fall within a programmable correlation window, and are therefore compatible with the trajectory of a high-momentum particle, they are identified as a candidate trigger primitive called a stub [20, 16, 17].

In the plane transverse to the magnetic field of the solenoid ($r\phi$), the radius of curvature $\rho$ of a charged particle with charge $q$, transverse velocity $v_T$ and mass $m$ can be derived by equating Lorentz force and centripetal force:

$$qv_TB = \frac{mv_T^2}{\rho}$$
which yields:

\[ \rho = \frac{m v T}{q B} = \frac{p T}{q B} \]

The transverse momentum can therefore be expressed as a function of the radius of curvature \( \rho \) by the expression (in natural units):

\[ p T \left[ \text{GeV} \right] = 0.3 \frac{q}{e} B [\text{T}] \rho [\text{m}] \]

with \( e \) the unit charge.

At radius \( R \), the relation between the transverse momentum \( p_T \) and the crossing angle \( \phi \) is (Fig.1.5):

\[ R = \rho \sin(\phi) \quad \text{and} \quad \rho = \frac{p_T}{0.3 B} \quad \Rightarrow \quad p_T = \frac{R}{2} \frac{0.3 B}{\sin(\phi)} \quad (1.1) \]

which, in the case of small crossing angles (\( \sin(\phi) \approx \phi \)) and a magnetic field of 4 T, approximates to:

\[ p_T \approx \frac{0.6 R}{\phi} \]

For a pair of closely separated layers at radii \( R_1 \) and \( R_2 \), the previous expression becomes:

\[ p_T = \frac{0.3 B}{2} \sqrt{R_1^2 + R_2^2 - 2R_1 R_2 \cos(\Delta \phi)} \]

From the above formulae it is clear how the \( p_T \)-cut is a function not only of the angle of incidence, and therefore of the coincidence window applied, but also of the
radius of the module in the Tracker: at larger radii the bending for a give $p_T$ is more pronounced and results in larger $\Delta \phi$ between the two sensors (Fig.1.4). For this reason, for the chosen strip pitch of 100 $\mu$m and for a 4 T magnetic field, this concept is limited to radii above 25 cm approximately.

To simplify the production of the modules, only two different values of $\Delta R$ will be implemented (Fig.1.6): the vast majority of the modules will have a 1.8 mm spacing, with the exception of a forward region in the endcaps, where a larger spacing of 4 mm will be necessary. A uniform $p_T$ threshold will be obtained by programming the coincidence window independently for every module, as described in Chapter 2.

1.3 Tracker Geometry

The tkLayout software package [22], specifically designed for quickly evaluating the consequences of design choices on the performance of the CMS Tracker, has been instrumental in guiding the design of the Tracker and its modules [23]. Several options for the geometry of the Tracker have been considered: the current baseline consists of a classical barrel and endcaps design (Fig.1.3). Thanks to the introduction of a fourth pixel layer, the number of strip layers could be reduced from 10 to
1.3 Tracker Geometry

<table>
<thead>
<tr>
<th></th>
<th>Existing Tracker</th>
<th>Phase II Tracker</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of modules</td>
<td>15148</td>
<td>15504</td>
</tr>
<tr>
<td></td>
<td></td>
<td>7084 PS</td>
</tr>
<tr>
<td></td>
<td></td>
<td>8424 2S</td>
</tr>
<tr>
<td>Total active surface</td>
<td>210 m²</td>
<td>218 m²</td>
</tr>
<tr>
<td></td>
<td></td>
<td>155 m² strips (2S)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>31 m² strips (PS)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>31 m² macro-pixels (PS)</td>
</tr>
<tr>
<td>Number of channels</td>
<td>9.3 M strips</td>
<td>47.8 M strips</td>
</tr>
<tr>
<td></td>
<td></td>
<td>218 M macro-pixels</td>
</tr>
<tr>
<td>Power in the tracking volume</td>
<td>( \sim 30 ) kW</td>
<td>( \sim 80 ) kW</td>
</tr>
</tbody>
</table>

Table 1.1: Comparison between the main parameters of the current Outer Tracker and its Phase II upgrade.

6 in the barrel, and from 9 to 5 in the endcap disks. Table 1.1 highlights the main parameters of the planned upgrade with respect to the existing detector.

As shown in Fig.1.7, the radiation length along the entire volume of the Tracker will be significantly reduced, mainly thanks to the use of a two-phase CO\(_2\) cooling, which will maintain the system at \(-20^\circ\text{C}\) or below. In particular, the present peak at the boundary between barrel and endcaps will be avoided with the removal of dedicated boards for module control, now integrated with the module itself, and with the reduction of the power cables section thanks to the use of on-module DC-DCs (as described later in Section1.4). A smaller radiation length will be highly beneficial and will greatly reduce multiple scattering, bremsstrahlung and photon conversion in the Tracker, which is the main limitation of the current system and a limiting factor for the resolution of the ECAL [24].

As a further improvement, an alternative option with tilted modules in the inner Strip Tracker is also being considered as an attractive solution to significantly reduce the material budget and the cost of the sensors, while at the same time improve the performance of the Tracker [13, 14]. Apart for the different mechanical structure, such Tilted Barrel Tracker configuration utilizes the same modules as the baseline design, and is therefore not discussed further.
Figure 1.7: tkLayout comparison between the Phase II upgrade and the current Tracker: number of hits and material budget. Left: number of hits as a function of pseudorapidity, indicating the redundancy and robustness against possible module or section failures. Right: the upgrade will reduce significantly the present Tracker radiation length, especially in the peak pseudorapidity region where $x/X_0 \approx 1.8$ (the Phase I Pixel expected contribution, in grey, is provisionally used in this plot for Phase II).

### 1.4 Module Design

The existing CMS Strip Tracker comprises 27 different flavours of modules for optimal coverage [25]. By adopting only two types of modules, the new Tracker design aims instead at reducing complexity, with obvious benefits in terms of design resources, assembly, operation and possibly maintenance.

Designed as independent units individually connected to the back-end with power lines and optical fibres for data readout and control, the modules are intended to minimize single point failures and facilitate assembly.

While the basic concept and many of the components are common, the two designs are optimized for different regions of the Tracker:

**Strip-Strip module (2S):** optimized for the outer region of the Strip Tracker ($R > 60$ cm), the 2S module consists of two AC-coupled sensors, each about $10 \times 10$ cm$^2$, with two rows of 5 cm long strips with 90 µm pitch, for a total strip count of 4064 (Fig.1.8). The expected power budget is approximately 5 W.
1.4 Module Design

Figure 1.8: 2S (Strip-Strip) module for the Outer Strip Tracker. Left: exploded view and main components. Right: sketch of the folding of the high-density flex-hybrid to allow the bonding of both sensors to the same front-end readout CBC (from [13]).

Pixel-Strip module (PS): in the inner region (R > 20 cm), higher granularity is required to obtain the desired occupancy and vertexing capability. The PS module correlates 2.5 cm long strips with columns of 16 1.6 mm long pixels, both with 100 µm pitch, for a total of 30720 pixels and 1920 strips [26]. For roughly half the sensor area of the 2S, at 5 × 10 cm² this module requires anyway a higher power budget (∼ 7 W), due to the higher granularity and irradiation levels.

Compared to the existing Tracker, both the PS and 2S modules aim to exploit five main technical choices and advances [13]:

- the use of binary data, instead of the present analogue readout, where the digitization is performed already on the front-end ASICs. This is justified by the vast increase in channel count (see Table 1.1), and is further discussed in Chapter 2.

- replacement of the separate readout and control systems with a common one for data, clock and control;

- use of DC-DC powering to minimize ohmic losses in the power cables by increasing the distributed voltage and thereby reducing the associate current;
• use of high density and low mass hybrids, for the dense routing required between the sensor strips and the front-end ASICs [4];

• use of flip-chip bump-bonding interconnection for all the ASICs on the module, to target assembly of the module hybrids in commercial high-volume production lines. As an example, each strip read out ASIC is bonded to the front-end hybrid by approximately 800 bumps on a 250 $\mu$m pitch. A significant improvement in assembly time is expected from this step, which will halve the number of wire bonds required for the connections between the sensors and the populated hybrids.

Naturally, the upgrade will also exploits other technological advances that followed the design of the original Tracker, in particular the scaling of integrated circuit technologies. While this is discussed in more detail in Chapter 2 in the context of the ASIC design, all the front-end and data processing circuits on the module will employ either of the radiation-proven and now mature 130 nm or 65 nm node CMOS technologies.

The availability of high density and low mass hybrids has been essential in the design of the module. Both front-end and service hybrids consist of a flexible, four layer polyimide substrate, with a total thickness of $\sim 130$ $\mu$m. The high-density routing achievable with this technology (minimum line pitch of 63 $\mu$m) removes the need for pitch adapters from the strips to the readout ASICs (CBCs), and helps with the dense routing of the CBCs output lines to the Concentrator IC (CIC) chip. Crucially, the flexibility of the front-end hybrid and its small bending radius allow the wire-bonding of both silicon sensors to the same face of the hybrid (bottom of Fig.1.8 and Fig.1.9): this greatly simplifies the signal routing from both sensors to the same CBC chip, where hit correlation for the two layers is performed, since it removes the need for a large number of relatively low density vias through the substrate.

The total estimate for the 2S mass currently amounts to 25 g, with the highest contribution from the aluminium spacers and carbon fibre support (12.6 g), followed by the sensors (9 g).
1.4 Module Design

High density flex hybrids

- 25 µm double-sided polyimide core layer, plus two single-sided 12.5 µm polyimide layers on either side.
- 25 µm coverlay on the bottom, solder mask on the top. Total thickness ~130 µm.
- 2S hybrid. Wirebonding pitch to sensor 90 µm × 2 sensors. Bump bonding pitch of CBC 250 µm, 800 bumps × 8 chips. High-density routing: thinnest line 30 µm with spacing 33 µm.

Prototype with Flat Flexible Connector (CIC not yet available). Eventually will be wire bonded to the Service Hybrid. Key element for a lightweight module design!

![Prototype of the 2S module Front-End Hybrid](image)

**Figure 1.9:** Prototype of the 2S module Front-End Hybrid, with 8 CBC2s bump-bonded, but without CIC chip and with flat flexible connector (FFC), to be replaced in the module by wire bonds to the Service Hybrid. The inset clearly shows how the hybrid is folded on one edge to wire bond sensors from both sides.

### 1.4.1 Module Readout

2S and PS modules are based on a similar electronic scheme, illustrated in Fig.1.10 for the 2S case. The main difference being that, while in the case of the 2S module strips from both sensors are read out by the same CMS Binary Chip (CBC), where the correlation is performed, in the case of the PS module strips are read out by a Short Strip ASIC, which sends the hit information to a different front-end chip for the pixel sensor (Macro Pixel ASICs [27]), where the correlation is performed.

In the 2S module, (Fig.1.10a) the sensor strips are read out by 16 CBCs, eight on each front-end hybrid. Each CBC reads out 127 strips from the inner sensor, and 127 from the outer one. Both the stub data and the binary unsparsified DAQ output data from the eight CBCs on one Front-End (FE) Hybrid are passed to a Concentrator IC, which buffers, aggregates and formats the data to interface with the Service Hybrid (cf. Fig.1.8). The symmetry between the two front-end hybrids is reflected in the even allocation of the total bandwidth (Fig.1.10b).

At the target transmission frequency of 320 Mbps, each CBC will output 8b of DAQ data and up to 40b of trigger data per BX to the CIC, compatible with sending up to 3 stubs per CBC per BX (as described in Chapter 6). This amounts to a potential maximum of 768 b/BX from the front-end ASICs, far in excess of the available module output bandwidth of 80 b/BX. The vast majority of these data are however
1.4 Module Design

(a) 2S module connection to the back-end power supply through low and high voltage lines (12 V and \( \approx 800 \) V), and to the DTC board through bidirectional optical fibres for readout and control data.

(b) Data flow from the CBCs front-end to the Low Power GigaBit Transceiver (GBT) (LP-GBT)

**Figure 1.10:** 2S module electronic system and data readout (from [13]). The readout flow follows the two front-end hybrids (Left and Right), with the LP-GBT total readout bandwidth evenly allocated between the two CIC chips.
trigger data (stubs), and at the expected average occupancy of less than one stub per module (Fig.1.11), sparsification in the CIC chip will ensure negligible trigger data loss by exploiting the statistical distribution of high-$p_T$ hits in space and time.

To limit the latency introduced by sparsification, the CIC adopts a so-called block-synchronous scheme, whereby all the CICs in the Tracker operate synchronously on a period of 8 BX, transmitting a maximum of 24 stubs/module from all the combined CBCs. Simulations show that such a system, which has a fixed latency of 8 BX, is expected to be at least 99% efficient for the nominal condition of 140 PU [28]. A fully synchronous, unsparsified mode for the CIC, limited to a maximum L1 rate of 100 kHz but nevertheless deemed useful for commissioning, will also be available.

The Service Hybrid hosts all the functionality for data transfer to and from the back-end electronics and for power management. Data transfer is handled by the LP-GBT, a low-power version of the existing GBT Integrated Circuit (IC) [29] currently being developed: a serializer/deserializer chip with an expected bidirectional bandwidth of 4.8 Gbps (effective 4.48 Gbps after error correction). The LP-GBT, which also provides I$^2$C control data, clock and trigger distribution, as well as local monitoring capabilities, interfaces with the back-end Data, Trigger and Control (DTC) board through bidirectional multi-mode optical link and an optoelectronics transreceiver (VTRx+) [30] on each module.

A DC-DC converter, based on a common development for several LHC experiments [31], will provide the voltages required for the front-end electronics (1.2 V) and optoelectronics (2.5 V).

### 1.5 Performance Estimate

The $p_T$-cut concept has been demonstrated experimentally in a test beam with 2S prototype modules; both test and outcome are the subject of Chapter5. Fig.5.12, for example, shows the excellent agreement between the reconstructed stub efficiency from experimental data, and the expected one from Monte Carlo simulation; for all
1.5 Performance Estimate

Figure 1.11: Simulation results demonstrating the efficacy of the proposed $p_T$ cut for data reduction, for 140 PU. Top: the cluster-to-stub ratio for barrel (left) and endcap modules (right) is presented as a function of $z$ and radius $\rho$ respectively. As expected, selecting stubs over clusters allows a reduction of one order of magnitude in readout objects. Bottom: stub multiplicity per module per BX. Because of the 2S module’s larger area, the last PS and first 2S barrel layers (51 and 69 cm) experience the same rate. Also, due to the excessive stub rate at low radii, endcap PS modules with $\eta < 2.4$ will be excluded from the L1 trigger. (From [13]).

The maximum efficiency is reached above 2 GeV/c, with a sharp turn-on behaviour.

On the detector scale, the efficacy of selecting stubs over hits or clusters for data reduction has been confirmed with Monte Carlo simulations. As shown in Fig.1.11, stubs provide a one order of magnitude reduction in readout data with respect to clusters, and the average stub rates across all trigger layers ($\sim 1/$module$\times$BX) are wholly compatible with the block-synchronous readout scheme described in Section 1.4.1.

Below are some of the case studies presented in the Phase II Technical Proposal: the results of an extensive simulation campaign to evaluate the benefit of L1 track availability to the performance of the trigger.

**Single Muon Trigger**

As briefly anticipated in Sec.1.2, the current stand-alone muon system occa-
(a) Trigger efficiencies for single muon with (black) and without (red) L1 track information.

(b) Trigger rates for single muon with 20 GeV threshold, with (black) and without (red) L1 track information.

(c) Trigger efficiencies for single electron with 20 GeV threshold, with and without L1 track information. The open squares pass an additional isolation cut.

**Figure 1.12:** Efficiency and rates for single muon and electron triggers with track trigger information (from [13]).
sionally promotes low-energy muons to high-$p_T$ particles. As evident from Fig.1.12b, this results in a flattening of the muon trigger rate vs. threshold; even increasing the trigger threshold beyond values that would compromise the physics performance of CMS would not bring the trigger rates to acceptable levels. The installation of additional muon stations during the Phase I upgrade will lessen this issue, but will not suffice for the muon rates expected at the HL-LHC.

Monte Carlo simulations of dedicated matching algorithms which exploit the trigger information from the Tracker have been shown to perform extremely well by providing a rate reduction of one order of magnitude. As visible in Fig.1.12a, they perform with high efficiency and, thanks to the excellent resolution of the Tracker, much sharper turn-on.

**Single Electron Trigger**

For electrons, $e^-/\gamma$ trigger tower objects in the ECAL are matched to L1 tracks. Because of bremsstrahlung or multiple scattering, the electron triggers have a lower efficiency than muons or pions, and the efficiency tends to degrade with $\eta$ as the material increases. Despite this, for a threshold of 20 GeV the inclusion of track information reduces the trigger rate by approximately five times (Fig.1.12c right).

**Single Photon Trigger**

Similarly to the electron case, single photon identification is seeded by calorimeter objects, with the important difference that the position of the primary vertex is not known, since the calorimeter trigger object cannot be matched to L1 tracks. In this case, tracker-based isolation of the $e/\gamma$ candidates performs considerably better than the calorimeter-only isolation requirement, lowering trigger rates by a factor of three, while retaining 90% efficiency.

**Track vertexing, Jets and Multi-Object Triggers**

Tracks at L1 can be used to reconstruct the primary vertex for an event. This is a highly effective way to increase trigger efficiency in high-pileup conditions, in particular for multi-object triggers such as those estimating the total or
missing transverse energy (HT), and multi-jet or hadronic triggers in general, for which the rate is significantly reduced by requiring that the jets originate from the same vertex. As an example of the possibilities opened by the new Track Trigger, the same constraint could easily enhance any other future multi-object trigger in CMS.

1.6 Summary

The plans for the High-Luminosity upgrade of the LHC which will extend the life of the experiment and its discovery potential were briefly outlined, together with the case for a new Tracker for CMS with data reduction and trigger capabilities.

The new Tracker will replace the current largest silicon detector with an even larger and more complex system, capable of providing trigger primitives for track reconstruction at the hardware trigger level, while also having superior performance in terms of radiation length and hit and momentum resolution.

By correlating hits between two closely separated sensors in each module, it was shown that it is possible to reject low transverse momentum tracks from suitable trigger primitives. This $p_T$-cut concept is central to the design of the new CMS Tracker and drove the design of its layout and of the individual modules, as well as the design of the CMS Binary Chip (CBC) which performs both front-end signal readout and stub identification in the Outer Tracker, and which is described in detail in the next chapter.

Finally, a selection of case studies demonstrating the benefits of the $p_T$-cut concept was presented. Thanks to the availability of the CBC2, the concept has also been proven experimentally in the test beam described in Chapter 5.
Chapter 2

Front-End ASIC Development

2.1 Choice of Technology: 130 nm CMOS Process

CMOS processes are broadly categorized on the basis of the smallest feature-size offered, typically referring to the gate length of a transistor; a 130 nm process for example offers a minimum gate length of approximately 130 nm. While these families, also known as technology nodes, often share the same underlying lithographic technology, the complexity of the manufacturing stages and the wide spectrum of target applications and markets make every process unique.

With the successful development of the APV25 CMS Tracker readout in the 250 nm IBM CMOS process in 1999 [32], it became apparent that by employing a commercial standard technology the particle physics community could benefit from the exceptional standards of process control and reliability which are driven by the mass-production electronic market [33, 34]. Other benefits include the scaling of devices, which allows more functionality per area, and the reduction in power supply voltage and hence lower power consumption [35]. Deep-submicron technologies are also inherently more resistant to radiation exposure, an important requirement for High Energy Physics (HEP) experiments. Gate oxide thicknesses of only a few nm (~3 nm for 130 nm CMOS) allow the removal of positive charges trapped in the dielectric through tunnelling of channel electrons [36]. This beneficial trend has
continued with technology scaling, as suggested by the recent characterization of a 65 nm process by CERN [37, 38], even though it is yet to be established whether the extreme levels required by the inner Pixel Phase II upgrade (∼1 Grad) can be sustained, as suggested by recent results from the RD53 collaboration∗ [39].

There are, however, several practical obstacles to the seemingly unstoppable scaling of CMOS technologies known as Moore’s Law. The most pressing is cost: with investments currently of the order of 10 billion dollars for a 22nm fabrication plant [40], and the exponential increase in complexity in the number of processing steps required, the cost of manufacturing an ASIC in such advanced technologies has become prohibitive for all but a few companies with volumes for mass production. Research institutions benefit from Multi Project Wafers (MPWs): special services offered by silicon brokers such as MOSIS† and EUROPRACTICE‡, which allow a number of designs to share the same lithography reticle, thereby sharing most of the cost of manufacturing. Although MPWs are effective for low-volume production and prototyping, they are not adequate for the volumes required to instrument large areas such as for the LHC silicon trackers, for which so-called engineering runs are necessary.

Cost is therefore one of the reasons why the gap between the leading-edge processes used by the HEP community and those used for commercial devices has widened in recent years. When the APV25 was first delivered in 1999, the next generation 180 nm process node was just becoming available. For comparison, at the time of writing, the most advanced designs for the HEP community are targeting the 65 nm node, which became commercially available in 2006.

Another complication is the exponential increase in design complexity for deep-submicron processes; the number of rules that must be observed for a successful design is in excess of a thousand for a 130 nm process [41]. While the latest generation of computer-aided design tools can help to manage such complexity, in order to obtain a robust design with high yield it is still important for the designer to

∗http://rd53.web.cern.ch/RD53/
†https://www.mosis.com/
‡http://www.europractice-ic.com/
be familiar with all the limitations of the process and with the available degrees of freedom. This has a direct impact on the time necessary to adopt a new technology and to complete a design. Yield and robustness have been a constant focus in the design of the CBC (cf. Sections 2.3.3 and 6.1.6); a few examples include the use of non-minimum length devices where possible, which are less prone to hot-electron degradation, and metal tracks with non-minimum width and separation to reduce the risk of opens and shorts respectively. Particularly useful in this respect is a set of optional rules and guidelines provided by the foundry, known as design for manufacturability (DFM), which contain recommendations and yield-enhancing techniques [41].

It should also be stressed that the pressure in industry for smaller transistors is primarily driven by purely digital circuits, which benefit the most from small feature sizes and are less sensitive to the limitations and non-linear effects that are prominent in analogue circuits. Reduced power supply voltage, sub-threshold effects and tight restrictions on the dimensions and layout of transistors are only a few of the difficulties affecting analogue design at 65 nm and below. Use of small technology nodes is most applicable for pixelated sensor readout ASICs, such as those instrumenting vertex detectors, to fit an increasing number of logic gates into each pixel and enable new capabilities.

Strip-readout ASICs have less stringent constraints for circuit density; the front-end electronics is typically arranged into parallel channels with a pitch to approximately match the input pads, and are therefore free to extend in the other dimension. Long and narrow channels are possible (cf. for example [42], with 128 channels about 9 mm long), although at the expense of increased routing density and potential for channel-to-channel crosstalk. In the case of the CBC2, each channel measures 40 µm × ∼700 µm and, although a pitch smaller than 40 µm would be feasible, the maximum routing density of the hybrid limits the number of input signals that can be fanned into the ASIC. For a strip read-out ASIC a more mature technology offers a better compromise in terms of digital and analogue performance [43].

On the basis of these and other practical considerations, such as present and long-term availability of the vendor [44], the IBM 130 nm CMOS8RF process, optimized
Figure 2.1: Schematic and scanning electron microscope section view of the 130 nm CMOS process used by the CBC. Highlighted are the small feature size of the low-level metals, as well as the thickness of the low-resistance top routing layers (adapted from [45]).

for radio-frequency (RF) analogue and mixed-signal applications, was chosen for the design of the CBC.

A cross section of the process is shown in Fig.2.1. The process offers eight routing metals: three fine-feature layers (with sub-micron minimum pitch) for local interconnection, four intermediate copper routing layers for low-resistance, long-distance interconnection and power distribution, and an extra-thick top level aluminium layer designed for inductors, with resistance down to the mΩ/square level. Since the ASIC does not require any inductors, this last layer is used for power distribution in order to reduce voltage drops across the chip.

Other features offered by this process include:

- Nominal power supply voltage of 1.2 V and minimum lithographic image of 120 nm.

- Twin-well CMOS technology on non-epi p-substrate: allows the placement of nMOS transistors within p-type well isolated from the bulk silicon by means of a buried n-type well, increasing isolation and allowing independent biasing of the body.
2.1 Choice of Technology: 130 nm CMOS Process

- Shallow trench isolation (STI): a processing step that enhances the isolation between adjacent devices.

- Single and dual gate oxide for 1.2 V and 3.3 V transistors respectively; thick-oxide transistors are typically used on I/O devices to interface with standard link protocols. To avoid the poor performance of thick gate oxides after irradiation, all transistors in the CBC are thin-oxide, except for the DC-DC step-down converter which must be able to sustain the 2.5V input voltage. Low and zero-threshold transistors, also available at additional cost, were not used in the CBC.

- Wire-bond pads or Controlled Collapse Chip Connections (C4s); both these options have been exploited in the production of the CBC, as described in Section 2.3.6.

- Metal-to-metal capacitor; also called metal-insulator-metal (MIM) capacitor, this parallel-plate device makes use of one or two thin, high-dielectric special layers and provides high capacitance per unit area (up to $\sim 4 \text{ fF/} \mu \text{m}^2$). The thin layers require additional masks and are susceptible to breakdown from high-voltage transients and from the build-up of static charge that can occur during the fabrication process, but specific layout rules are in place to prevent this failure. MIM capacitors are used in the CBC in the front-end channels, where the area constraint makes high-density capacitors desirable, and for on-chip decoupling where large capacitors (hundreds of pF) are required.

- Vertical natural capacitor (VNC); also called metal-oxide-metal (MOM) capacitor, it is created by interdigitating multiple levels of routing metals (optionally connected by contact vias) and separated by the normal inter-metal dielectric. Thanks to the small feature sizes of modern processes, this device achieves a capacitance of $\sim 1 \text{ fF/} \mu \text{m}^2$ and does not require additional fabrication steps. VNC capacitors were used for the channel test-pulse capacitors (Section 3.2.1), since the desired value of 20 fF was below the minimum design size for MIM capacitances.
Several types of resistor devices are available, with different and typically conflicting performance in terms of resistance per unit area and precision. Other aspects to consider are temperature and voltage sensitivity, absolute tolerance, parasitic capacitance to the ground plane and potential for latch-up (cf. Sec. 4.2.1) created by parasitic diode junctions.

- BFMOAT isolation trench: this technique is used to create an area of resistive substrate to reduce the coupling of substrate noise between the analogue and digital sections of the CBC.

- Electrostatic Discharge (ESD) protection macros are available to suit several ESD models and strategies, as discussed in Appendix B.

## 2.2 Choice of Architecture

### 2.2.1 APV “Digital”

The APV25 was an extremely successful chip and the natural starting point for the front-end electronics for the HL-LHC. However it soon became clear that there were good reasons to drop the all-analogue approach in the high data-rate environment of the high-luminosity LHC, as the custom analogue off-detector links required for transmitting the analogue pulse height information would be replaced by high-speed digital links [46, 33].

To retain the pulse-height information, a digitization stage would have to be introduced either before or after the memory buffer. Digitization before the pipeline allows a digital memory, which should minimize the area required by this circuit, with the disadvantage of requiring an Analog-to-Digital Converter (ADC) on every channel. An estimate based on the International Technology Roadmap for Semiconductors yielded 1.6 mW for a 130 nm CMOS 6-bit ADC operating at 20 MHz (50 ns bunch crossing frequency was being considered for the HL-LHC) [46]. The APV25 achieved a power consumption of 2.7 mW/channel, but a much lower figure, close
to 0.5 mW/channel, is needed for the HL-LHC given the increase in the number of channels and the available cooling budget per module.

Digitization after an analogue pipeline, on the other hand, could achieve 50 $\mu$W per channel since a relatively high-power ADC could be shared between many front-end channels. With the analogue pipeline still in place, it would also be possible to retain the slow shaping and analogue deconvolution approach implemented in the APV25. These allow a slower, more power-efficient analogue front-end, while the correct bunch crossing of the signal is reconstructed from a weighted sum of three consecutive samples of the signal [32]. The main disadvantage of this architecture is complexity, since it required all the circuitry of the analogue APV25 chip with the added complexity of a fast ADC. Readout latency would also be affected by the conversion speed of the ADC and the hit occupancy. Moreover, other features currently performed on the off-detector Front-End Drivers (FEDs) would have to be integrated into the front-end: in particular the sparsification necessary to keep the data volume at a manageable level and possibly a common-mode subtraction mechanism. It was also not clear whether an analogue pipeline using gate capacitances would still be feasible with the increased gate leakage of the 130 nm process ($\sim 20$ pA/$\mu$m$^2$), especially in the case of a possible increase in pipeline length. Data sparsification in particular was seen as adding excessive complexity and making the system less robust and less easy to operate and debug [46].

For these reasons, it was decided to retain the un-sparsified nature of the APV25, at the expense of the pulse height information.

2.2.2 CMS Binary Chip

What emerged from this preliminary analysis was a recommendation for a binary, unsparsified architecture. The CMS Binary Chip (CBC) [5, 6, 47, 48], without analogue deconvolution, needs a faster front-end and an additional comparator in each channel, but this architecture guarantees low overall power consumption, as there is no ADC, making the readout and pipeline purely digital. By having a synchronous
system, it is relatively straightforward to recognise upsets in the operation of the chips by checking the consistency of the pipeline address data.\footnote{Individual APV25 upsets have not actually been detected in CMS, because of the resets required by the pixel readout to cope with beam-gas interactions \cite{49}.} It is also possible to emulate the behaviour of the chip off-detector, in the same way as is currently done with the APV Emulator (APVE) and protect the system against buffer overflow with trigger back-pressure \cite{16, 50}. This is essential since the transit time of $\sim 0.5 \mu s$ from the counting room to the Tracker and the time needed to process the front-end data frames do not allow a veto system based on counting triggers off-detector \cite{33}. Other added advantages of the binary approach include the possibility to port the design to more advanced processing nodes, where the use of analogue pipelines could be impossible due to high leakage currents. Importantly, dealing with binary information would also facilitate on-chip implementation of the trigger functionality. The disadvantages, as already mentioned, are loss of $dE/dx$ information for particle identification and likely degradation of position resolution, together with a more difficult evaluation of noise and common mode effects.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure2.2}
\caption{Relative size of the two CMS Binary Chip prototypes.}
\end{figure}
2.2 Choice of Architecture

2.2.3 CBC1

Measuring 7 mm × 4 mm, the first prototype of the CBC was a full-size ASIC, meaning that it did not comprise only a few channels or a small subset of the final functionality, but included 128 channels, internal biasing and all the additional functionality required to correctly read out a silicon strip sensor at 40 MHz. The decision stemmed partly from the success of the APV25 and the similarities to its architecture, but also from the fact that only a full-size ASIC can highlight system issues related to power distribution, signal crosstalk and common-mode noise, all of which could be critical for the performance of the Tracker.

Fig.2.2 shows a photograph of the chip, and outlines its main functions. The 128 pads are arranged in two staggered columns to obtain an effective pitch of 50 µm, while the other three sides use pads on 150 µm pitch. Each front-end channel includes a charge integrating preamplifier, followed by a shaping amplifier and a comparator for the binary digitization. The outputs of the comparators are stored into a 256-deep memory block, which at the 40 MHz operating frequency of the LHC provides an effective 6.4 µs latency. Upon receipt of an external L1 trigger, the data are passed into an additional 32-deep memory buffer, where they are stored before being transmitted off-chip on a 40 Mbps serial line. Additional features, such as an on-chip DC-DC converter to power the chip from a 2.5 V external supply and a Low Dropout Linear Regulator (LDO) to power the analogue sections, were also included, together with I2C interface, bias circuitry and other ancillary circuits. Since the CBC1 was very successful, most of its circuits were included in the next prototype ASIC, and will be discussed in more detail in the following section on the CBC2. The few issues discovered during testing and later fixed in the CBC2 will also be described in Chapter 3.

The CBC1 chips were fabricated in a 130 nm CMOS process and were delivered in February 2011. They underwent thorough evaluations, culminating with the instrumentation of a telescope plane at the Super Proton Synchrotron (SPS) at CERN [51].

*This was the envisaged latency at the start of the CBC development, owing to the limit imposed by the existing ECAL read-out pipeline.*
2.3 CBC2

Despite the success of the CBC1, after its completion a number of significant new requirements had emerged from the parallel development of the 2S module. The initial assumption that only a couple of dedicated layers in the Tracker would suffice for trigger primitives [52], was dismissed in favour of several distributed layers extending to the outer Tracker, as described in Chapter 1. The trigger functionality had to be provided by the strip readout modules, and therefore be integrated into the CBC development. To address this, a second, full-scale prototype ASIC was required: the CMS Binary Chip v.2 (CBC2) [1, 2, 3, 53].

By the time design started on the CBC2, the development of the 2S module concept had progressed significantly and a number of additional features, besides the trigger capability, were incorporated in the new prototype. The most prominent changes and new features introduced are summarized here:

- The number of channels doubled to 254, divided equally between the top and bottom sensors of the 2S module. The reason this number is not a canonical power of two is to have a true zero which can be recognized at the output as an absence of hits, distinguishable from the zero-th channel address. The number of channels is even to maintain the symmetry between the two sensors. The loss of two symbols out of the 256 available compares very favourably with the overhead otherwise introduced by an additional output bit to flag the presence of hits. This is especially important given the tight constraints on the available output bandwidth, as discussed in Section 1.4.1.

- The mounting of the front-end chips on the hybrid was proposed to change from wire-bonding to flip-chip, or bump-bonding, assembly. The aim was to target industrial assembly processes, since, from experience in building the present CMS Tracker, it was recognised that assembly of the more than 8000 modules required in the Outer Tracker is a significant task in terms of time and resources. As described in Section 1.4, the aim was to retain wire-bonding
only for the sensors, while the hybrid components, chips included, would be assembled by external companies.

The choice of pad pitch is a compromise since it affects the hybrid and the chip in opposite ways; a larger pitch increases the size of the chip, especially in the case of the CBC2 where it was decided not to place any active circuit below the input pads to avoid cross-talk and noise injection. Conversely, a small pitch puts severe constraints on the manufacture of the hybrid, since it dictates the number of routing layers required, the minimum track width and separation (∼30 µm), and the technology and minimum features required for via drilling to reach the high number of pads on the chip. Since hybrid manufacture is already challenging due to the large area of the 2S module, it was decided to ease the constraints on the hybrid at the cost of some additional chip area. Flip-chip bonding (Section 2.3.6) is a well-established technology and a conservative pitch of 250 µm was chosen for the pads. However, unlike ball-grid array (BGA) packages, such coarse bump-bonding had not been previously used for front-end chips or for hybrid-ASIC assembly for such applications, so proof from a full-size ASIC was required before adopting it.

- The CBC2 includes logic functionality to correlate hits between the two sensors and to identify stubs. The logic is described in detail in Section 2.3.3. Synchronous readout of stubs was not implemented, since at the time the data format and the requirements were still fluid. However a number of test features for the readout of stubs have been implemented and are described in Section 2.3.4.

- A test-pulse circuit is included to test the response of individual channels (Section 3.2.1).

- New versions of the DC-DC step-down converter and linear regulator, to improve the noise performance of the chip, are included (Section 3.2.5).

- New biasing circuits address the few issues uncovered in the testing of the CBC1 [54].
2.3 CBC2

2.3.1 Architecture

A functional view of the architecture of the CBC2 is presented in Fig.2.3. In this picture, signal processing flows from left to right, from the input pads to the output pads on the right-hand side. This is a natural arrangement for strip readout ASICs since it minimizes the risk of injecting signals into previous readout stages (through parasitic coupling between adjacent nets, for example).

The signal path includes amplification, analogue signal processing for noise optimization, followed by digitization. The binary output of the comparators is then sent to the stub-finding logic and at the same time it is stored, waiting for the external L1 trigger before readout. These main functionalities require several additional ancillary circuits, described in this chapter; among these are programmable bias generator, bandgap reference, power-on reset circuit, I2C clock and data serial link interface, powering features (DC-DC step-down converter and LDO) and a test-pulse generator.

The 254 channels are identical and are repeated with a pitch of 40 µm, to match the pitch of the input pads while allowing extra space at the two extremities to include the test pulse generator circuit. This configuration also allows good segmentation and matching, since every stage of the readout channel is enclosed by similar circuits, and allows effective bias distribution through the array.

The binary data from the channels are sent to a memory pipeline to provide the required latency for the external L1 trigger, and at the same time to the stub-finding logic which identifies high-$p_T$ candidates. This is composed of several functional blocks, to mask noisy channels, reject wide-clusters and correct for geometrical offsets within the detector (Fig.2.4).

2.3.2 Front-End

Despite being split between two different sensors, up to the digital processing stage for stub identification, all of the 254 input channels are identical. Each channel
Figure 2.3: CBC2 block diagram, arranged to reflect the layout of the chip (signal processing flow from left to right).
Figure 2.4: CBC2 architecture, highlighting the two parallel data paths through the stub-finding logic (trigger data, in green) and directly to the memory pipeline (L1 data, in red).

Figure 2.5: CBC front-end readout channel.

consists of a charge sensitive preamplifier, gain amplifier and comparator as shown in Fig.2.5, plus additional local biasing and programmable registers.

Charge generated by an ionising event in the strip is read out by a preamplifier and integrated onto a feedback capacitor. The feedback capacitor is discharged by a resistive network, selectable between hole and electron readout, to allow flexibility in the choice of sensor, which at the time was still open.

The resulting voltage pulse from the preamplifier is further amplified by a capacitive gain second stage, also with selectable readout polarity. A large value programmable feedback resistor stabilises the amplifier. To compensate for any channel-to-channel
offset and mismatch in both the gain amplifier and the comparator, an 8-bit programmable offset is implemented at the output of the gain amplifier.

The comparator detects signals over a programmable global threshold and produces a digital-high pulse until the signal returns below threshold. A select circuit is used to keep the same pulse polarity irrespective of whether electrons or holes are integrated. A 4-bit programmable hysteresis level is also implemented to reduce switching noise.

The individual stages are described in detail in the following sections.

**Preamplifier**

The charge-sensitive preamplifier presents a high-value dynamic input capacitance on which the charge is integrated [55]. The CBC preamplifier (Fig.2.6) consists of a single-ended cascode amplifier, a source-follower at the output, and a feedback capacitor and resistor network.

The total power consumption for this stage is 150 $\mu$W. The currents in the input transistor, cascode and source follower are programmable through 8-bit on-chip Digital-to-Analogue Converters (DACs), so that noise performance can be tuned to the final sensor capacitance.
The open loop voltage gain ($A_V$) must be high for most of the signal charge to be removed from the detector. The design detector capacitance of $\sim\!5$ pF plus any parasitic capacitance at the input, must be small compared to the product $A_V C_f$ so that most of the charge is integrated onto the feedback capacitor $C_f$. In addition, the feedback capacitor connects to the input of the source follower and therefore helps compensate the amplifier. The original specification for 2.5 cm long strips resulted in a target sensor capacitance of 5 pF, including safety margin (typical values of strip capacitance for the pitch and thickness under consideration are around 1.5 pF/cm). The 2S module design has since converged on 5 cm strips, so the next version of the ASIC will be optimized for higher capacitance. The ability to program most front-end currents independently, however, allows a high degree of flexibility which has proven valuable during CBC2 commissioning, with a variety of sensors used according to availability.

The integrated charge is discharged through the feedback network. With reference to Fig.2.6, if electrons are read out, the switches $e$ are closed and a single resistor of 200 k$\Omega$ forms the feedback resistance. If holes are read out then switches $h$ are closed and a T-network forms the resistive feedback.

The circuit is able to cope with leakage current of both polarities up to 1 $\mu$A, since DC-coupling of the sensors to the input pads was an option to reduce the cost of the fabrication of the sensors\footnote{The design later converged on AC-coupled strips}. The preamplifier does not include a dedicated leakage compensation circuit, such as that described in Ref.[56]; the DC shift caused by the sinking or sourcing of the leakage current through the feedback circuit is instead accommodated within the signal voltage range, and later removed by AC-coupling the following gain stage. Without leakage current, for n-in-p sensors ($e^-$ readout mode) the input and output nodes of the preamplifier are biased around a nominal voltage of $\sim\!200$ mV. A maximum leakage current of 1 $\mu$A through the 200 k$\Omega$ feedback resistor will introduce an additional 200 mV shift at the output node, still well within the linear dynamic range of the preamplifier. In hole readout mode, a negative offset would bring the preamplifier outside its operating range, hence the T
feedback network is designed to introduce a positive baseline shift at the output to allow extra headroom to accommodate the leakage-induced negative shift. For both electron and hole readout modes, the feedback network has an equivalent resistance of 200 kΩ, designed for a decay time of 20 ns to reduce pileup effects. The choice of feedback resistance is often a trade-off. The resistance must be large enough so that the time constant $R_fC_f$ is not small compared to the rise time of the signal, otherwise the signal is not completely integrated and causes a loss in pulse height known as ballistic deficit. However, the leakage current through too large a resistance would introduce excessive offset for the dynamic range of the amplifier, and a small $R_f$ would also become a significant noise source.

The second stage is a source follower, a common-drain configuration with near-unity voltage gain and low output resistance, well suited to drive low impedance or capacitive loads. This configuration is often used as an output stage for its current drive capability and for its high input impedance, which helps the first stage achieve high voltage gain. It also introduces a level shift, which ensures that the cascode transistor remains in saturation. The cascode bias ($V_{pc}$) is a global voltage generated on-chip by an 8-bit programmable DAC.

**Choice of input transistor** Historically, pMOS transistors have been favoured over nMOS as input devices since they have lower $1/f$ noise [55]. However, in recent processes the frequency at which the $1/f$ spectral density equals the white noise (known as corner frequency) has been pushed to lower frequencies so the difference is not as pronounced as in the past [57], and there are advantages in using an nMOS instead in terms of power supply rejection, as presented in Fig.2.7. For a pMOS device, any noise on the power supply is in series with the signal and is therefore amplified. This could be avoided by grounding the source of the pMOS and having a negative second supply, but requires an unconventional powering scheme which would complicate the assembly of the module and was therefore not the preferred option. In a nMOS input device, the source and the sensor are coupled to the same ground, which provides some supply immunity with a conventional powering scheme [58, 59].
The most important parameter for a transistor is the transconductance \( g_m = dI/dV \), which changes significantly depending on the operating region of the device. The maximum value of transconductance, and therefore the minimum theoretical noise, is achieved when the transistor operates in strong inversion. Noise and gain must however be balanced against minimizing power consumption, especially in a system with millions of channels, such as the CMS Tracker. A better design figure for sizing the input transistor is the transconductance efficiency, defined as the ratio of the transconductance \( g_m \) to the drain current \( I_D \). The transconductance efficiency is maximum in weak inversion, so input transistors for front-end readout usually operate in this region [60], where \( g_m \) is independent of the particular geometry of the gate [57]. There are good reasons not to use minimum-length input devices: transistors with very short channel exhibit excessive thermal noise [57], which is dominant at the frequencies of interest for fast pulsing systems such as the 40 MHz synchronous CBC, which has a peaking time of 20ns.

**Post Amplifier**

The second gain stage of the front-end channel (Fig.2.8) amplifies the signal before the comparator and removes any leakage current-induced offset by being AC-coupled to the preamplifier. The circuit includes a differential amplifier with a tail current of 20 \( \mu \)A. The ratio of the input capacitor \( C_{in} \) and the feedback capacitor \( C_f \) set the voltage gain of 12.5. DC stability is provided by a feedback network, selectable.
depending on the polarity of the signal. Two diode-connected pFETs are used in the first stage of the amplifier to limit the differential swing in order to speed recovery from highly ionizing particles (HIPs), which generate signals up to three orders of magnitude larger than minimum ionizing particles (MIPs) [61]. A capacitor $C_c$ is used to compensate the stability of the amplifier.

A resistor in the output branch allows the level shifting of the gain amplifier output voltage. This is done by adjusting the differential currents $I_{pao1}$ and $I_{pao2}$. The currents are controlled by a 8-bit control register in each channel. The nominal current of this stage is 32.7 $\mu$A and the total power consumption is therefore about 40 $\mu$W.

The quiescent output voltage is controlled globally by $V_{plus}$. The feedback network includes a long pFET biased with a small current to have equivalent high resistance and, as for the preamplifier, there are two distinct selectable configurations for hole or electron readout. Figure 2.8 shows the configuration for electrons. The three

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure2_8}
\caption{Left: CBC gain amplifier. Right: feedback network for electron (top) and hole readout (bottom).}
\end{figure}
switches $e$ connect a current mirror in the configuration shown. Since the output signals are negative going, the sources of the current mirror transistors must be connected to the higher potential ($V_{\text{plus}}$). A further circuit in the bias generator (common to all channels) can be used to adjust the current in the feedback circuit. Also in Fig. 2.8 is the configuration for reading out holes; in this case the output signal is positive-going so the sources of the current mirror transistors must connect to the output. A 1 pF capacitor is used to ensure that the gates of the transistors in the current mirror follow the output to maintain linearity.

One of the problems identified with the first version of the CBC was related to common-mode sensitivity of the feedback network of the gain amplifier in holes mode. When many channels were active at the same time, for example during the capture of S-curves when the threshold is lowered below the noise floor, their 1 MΩ resistance was effectively in parallel, resulting in a much lower equivalent resistance, lower than the impedance driving that node ($V_{\text{PAFB}}$ in Fig.2.8). It was therefore possible for self-coupling to occur between the output signal and $V_{\text{PAFB}}$. During the testing of the CBC1, the problem was circumvented by loading the $V_{\text{PAFB}}$ net with an external 100 nF capacitor to ensure stability. In the CBC2, however, the issue was resolved by buffering $V_{\text{PAFB}}$ locally with a source follower in every channel to avoid coupling between channels. The fix was successful and the instability has not been observed during the testing of the CBC2.

**Comparator**

The CBC uses binary readout and the digitizer stage consists of a comparator in each channel, which separates the analogue and digital domains of the chip. The combination of high-frequency limiting bandwidth in both the preamplifier and the gain amplifier results in a peaking time of approximately 20 ns at the input of the comparator.

The comparator is a two-stage differential input architecture, as shown in Fig.2.9, followed by additional standard-logic inverters. The differential stage is powered from the analogue supply, the logic gates from the digital supply. The polarity of
the output is selectable depending on whether electrons or holes are being read out, so that a hit in the channel always corresponds to a positive transition at the output of the comparator.

The global threshold voltage (VCTH) is generated on chip in 256 steps of $\sim 2.3$ mV, from 0.2 V to 0.8 V, programmable via I2C.

The magnitude of signal charge affects the response time of comparators. For a system such as the CBC with a peaking time still significantly longer than the charge collection time in the sensor, this so called time-walk is mostly dependent on the rise time of the input signal and the gain of the comparator [55]. The specification in this case was set to a limit of 16 ns for the maximum time-walk between a 1.25 fC and a 10 fC signal, for a 1 fC threshold [6, 62].

The comparator also has programmable hysteresis, so that the input threshold changes as a function of the output level. A small amount of hysteresis is useful to suppress switching noise, especially when the amplitude of the signal is comparable to the threshold setting, so when the switching of the comparator is at its slowest and noise on the input signal could cause multiple consecutive threshold crossings. The switching activity is highly undesirable since it is not only noise in the supposedly clean digital domain, but also involves high current spikes that can potentially propagate to adjacent channels, therefore resulting in threshold and occupancy-dependent common-mode noise [63].
In the first version of the CBC1, the hysteresis was implemented with a resistive feedback network. This solution suffered from a sharp non-linearity in the high threshold range, due to the limited current sinking capability of the bias circuit used to generate the global threshold voltage $V_{CTH}$. The combined current created by the resistive hysteresis feedback of all the 128 channels would exceed this limit and cause the circuit to fail at high thresholds. By over-driving this voltage with an external DAC, it was possible to effectively work around this problem and successfully test the functionality of the comparator. A different solution was adopted for the CBC2 to resolve this issue: an internal hysteresis circuit consisting of an additional differential pair biased with a small, programmable current (Fig.2.9). This current controls the amount of hysteresis and is programmed globally by a 4-bit register. The default setting during operation is for minimum hysteresis.

**Hit-Detect Logic**

The output of the comparator returns to zero when the pulse from the gain amplifier falls below threshold. The duration of the comparator output depends therefore on the magnitude of the signal, and could result in a large signal being recorded as two, or more, consecutive hits in the same channel. This would not only lead to out-of-time data in the pipeline but would also affect the triggering logic in a high-occupancy scenario. For these reasons, a logic block called *Hit-detect* is inserted after the comparator to synchronise the output of the comparator with the bunch crossing clock which defines the sampling of the data into the pipeline. Two modes of operation are possible and are described in Fig.2.10; in the first mode the logic produces a synchronised version of the comparator output, which is sent to the pipeline memory. In the second mode, the output is a pulse which is one clock cycle long, regardless of the duration of the comparator output pulse. The Hit-Detect can be disabled by programming the relevant register through the I^2C interface, in which case it acts as a global mask for all the channels.

The Hit-Detect circuit was present on the CBC1, but dead-time associated with this particular circuit meant that in the case of two consecutive hits on the same channels,
2.3 CBC2

Figure 2.10: Effect of the Hit-Detect logic at the output of the comparator: (a) and (b) the circuit is enabled (EN=1) and SEL=1 which means a synchronised version of the input should pass through to the output. This is true of (a), however (b) is not high for enough time to be synchronised to the clock. In case (c), the circuit is disabled (EN=0) and nothing passes to the output. For (d) and (e), the circuit is enabled and SEL=0.

the second would not be captured. Although the probability of this occurring is low, given the low channel occupancy (∼ 1%), the circuit was modified in the CBC2 to remove this limitation, and the chip can now accommodate consecutive hits.

2.3.3 Stub-finding Logic

The operation of the stub-finding logic is based on a simple procedure: the first stage rejects wide clusters of hits on both inner and outer sensors; subsequently for every valid cluster on the inner sensor, the logic looks for a corresponding 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 [1].

The logic circuitry responsible for these tasks is structured into combinatorial blocks, which repeat every two adjacent channels (corresponding to one inner strip and the outer strips directly above it) (Fig. 2.11). This structure is therefore highly uniform across the chip, which is beneficial in terms of power consumption and distribution, timing and design verification and test coverage.

Channel Mask

Given the low occupancy of high-$p_T$ tracks (Sec.1.2), noisy or faulty channels should not be processed by the combinatorial logic since they would generate fake stubs and significantly reduce the bandwidth available for genuine stubs, therefore decreasing the trigger efficiency. Every channel can therefore be individually masked.
Figure 2.11: CBC2 stub-finding logic: channels from both inner and outer strips have a corresponding Cluster Width Discrimination block, followed by a Offset Correction and Correlation Logic block for every inner strip. Also highlighted are the inter-chip signals and the two fast-OR signals which can be selected to latch the stubs into the readout shift register.
by programming 32 8-bit registers via I²C. The masks do not affect the data stored into the memory pipeline, since the full data frame is read out from the CBCs without zero-suppression and therefore there is no bandwidth gain in suppressing noisy channels.

As described in Sec.1.4.1, the increase in L1 trigger rate to 750 kHz decided after the design of the CBC2 required data sparsification at the module level, with zero-suppression performed by the CIC ASIC. To avoid including memory in the Concentrator chip to recognize noisy strips, the channel mask will be extended to the pipeline data in the next version of the CBC.

Cluster Width Discriminator

The Cluster Width Discrimination logic (CWD) is the first stage of the combinatorial logic and analyses adjacent strips from the same sensor to reject wide clusters of hits, which are typically associated with low-momentum tracks. Specifically, a CWD block is associated with each channel and its output goes high whenever this strip is found to be the centre of a valid cluster. A global 2-bit register sets the value of the maximum cluster width to one, two or three adjacent strips; clusters wider than this are suppressed (Fig. 2.12). The same register also allows the CWD to be disabled and to pass every single hit to the subsequent stage.

This is a first step in reducing the low energy and combinatorial background for the following correlation stages; from the equation of Sec.1.2.1, clusters of four strips or above correspond to a maximum $p_T < 0.71$ GeV/c, calculated for the outermost layer ($R = 1$ m), $B = 3.8$ T and sensor active thickness of 200 μm (see also Fig.5.8). The low threshold cut, less than half the target threshold of 2 GeV, ensures that this coarse selection does not affect the efficiency of the following $p_T$ selection.

At the output of the CWD only the central strip of a cluster is active; in the case of a two-strip cluster, the centre is assigned to the strip with lower address (Fig. 2.12). Despite the obvious loss of resolution, this choice was made to reduce the number of bits per stub to a minimum, to accommodate more stubs in the output bandwidth
Figure 2.12: Every channel has a Cluster Width Discriminator block which indicates whether the strip (indicated by the red arrow) is at the centre of a cluster. Clusters wider than a programmable width (1, 2 or 3 adjacent strips) are suppressed. A loss of resolution occurs for two-strip clusters where the centre is assigned to the strip with the lowest address.

Figure 2.13: Study of the coincidence window width settings necessary to achieve a $p_T$ threshold cut of 2 GeV across the Tracker volume, based on the results from the tkLayout simulation tool. The maximum window width is 15 strips in the barrel and 12 in the endcaps (from [64]).

available. The readout scheme has since been revised and as a result the future version of the ASIC will implement half-strip resolution for the stub position to avoid this loss of resolution, as described in Sec. 6.1.3.

Φ-shift Correction and Correlation Logic

For every central strip of a valid cluster in the inner layer, the Offset Correction and Correlation Logic (OCC) block following the CWDs looks for valid clusters in a coincidence window in the outer layer. If there are any, then the output is high, which indicates that the inner strip corresponds to a valid stub.

The width of the correlation window in the outer layer strips is what defines the $p_T$ cut, once the modules have been assembled. For a given $p_T$, the displacement
(a) *Left:* the combination of the offset correction (±3 strips) and of the coincidence windows (±8) result in a total search window of ±11 strips. *Right:* the ability to shift the search window by a maximum of 3 strips (with 1 strip resolution) allows one to correct for the parallax offset across the module.

**Figure 2.14:** CBC2 Φ-shift correction.

between the hits in the two sensors depends on the position of the module in the tracker, increasing at larger radii in the case of the barrel. In order to obtain a uniform $p_T$ cut throughout the Tracker volume it is possible to adjust the sensor spacing during assembly, and, more effectively, the width of the coincidence window, which for a single chip is programmable in the range ±8 strips, symmetrical around the central strip. This range is sufficient to cover the maximum coincidence window, which for the outermost region of the Strip Tracker extends to 15 strips (Fig.2.13).

Also depending on the position of the strip along the module in the $r$–$\phi$ plane, the coincidence window must be adjusted to account for the geometrical lateral displacement across the same module, as illustrated in Fig. 2.14. This offset is independently programmable in the range ±3 strips in the two halves of the ASIC. Every module is therefore divided into 16 regions of programmable offset. The maximum offset expected in the Tracker is $\sim$ 1.5 strips [65], well within the adjustable range.

It should be stressed that while the readout channels are identical for the two strip sensors, the OCC block is ‘seeded’ by the inner layer clusters and the correlation
window is applied to the outer layer. The distinction between inner and outer is a natural one when referring to the interaction point and to the hermetic design of the Tracker. However, as discussed in more detail in Section 6.1.3, this asymmetry imposes a constraint on the orientation of the modules during assembly and is therefore undesirable. The CBC3 will address this by allowing one to assign the role of seeding layer to either of the sensors.

**Inter-chip Boundaries**

To be able to resolve clusters or stubs crossing two adjacent chip domains in the silicon sensors, up to 15 signals must be transmitted by each CBC2 to each neighbouring chip. Readout modules are designed to be independent units so tracks which extend across module boundaries will be resolved by overlapping the modules along $\eta$.

Both the CWD and the offset correction and correlation logic rely on inputs from neighbouring channels. To guarantee continuity of the stub coverage across the module, many of these signals must be transmitted across the chip boundaries. In fact, in the extreme case of the maximum search window and offset, the correct operation of more than 15% of all the stub logic in a module would depend on such inter-chip signals.

Specifically, the numbers of inputs required by each chip at each of its two boundaries are:

- CWD inner sensor channels: 2 adjacent inner channels;
- CWD outer sensor channels: 2 adjacent outer channels;
- Offset correction: 3 adjacent coincidence logic channels;
- Coincidence logic: 8 adjacent coincidence logic channels;
for a total of 15 inputs. The same chip also needs to output the same number of signals to its neighbour. The CBC2 therefore needs a total of 60 pads dedicated to inter-chip signals: 15 inputs and 15 outputs on each side.

An alternative scheme involving direct transfer of the comparator outputs would have been possible, reducing the number of inter-chip signals from 15 to 13. This solution requires a duplication of CWD blocks on each chip to process the additional strip outputs. Given the marginal reduction in the number of links at the cost of additional logic, the more straightforward solution described above was favoured for the CBC2. This scheme is discussed in more detail in Sec.6.1.3, however, where the introduction of half-strip resolution makes this alternative more attractive.

At the edges of the module, for the first and last chip the signals are grounded on the hybrid so that they do not cause spurious signals; for the same reason, the inter-chip pads are designed to drift low when not driven or left floating, to guarantee intrinsic robustness of the chip regardless of the hybrid (for example during testing when the chip is not bonded to the full hybrid).

As discussed in Chapter 1, modules are designed as independent units and there is no data transmission between adjacent modules. Overlap of modules in $\eta$ and $\phi$ will ensure that high-$p_T$ tracks straddling adjacent modules are not lost. The acceptance loss due to stubs straddling the two halves of the same module in $\phi$ is compensated by the multiple layers in the Tracker.

**Logic Layout**

The CWD logic differs slightly in drive strength for signals from inner and outer sensors, since the latter are propagated for a longer distance, across 11 inner channels in each direction (for a total length of about 1.8mm). The vertical tracks which distribute the output of the comparators and of the CWD blocks across adjacent channels were carefully laid out to have uniform length, be well-separated (well above the allowed minimum spacing) and with the least number of crossings to minimize crosstalk and the extra load due to stray capacitance between adjacent lines (Fig. 2.16).
2.3 CBC2

Figure 2.15: Detail of the layout of front-end channels and coincidence logic: (a) ESD protection; (b) test capacitor; (c) front-end channel; (d) 8-bit comparator offset registers; (e) channel mask registers and Hit-Detect circuit; (f) channels OR (254-OR); (g) CWD; (h) Offset Correction and Correlation Logic, including stub readout shift register; (i) stubs OR (127-OR).

Figure 2.16: Detail of the stub-finding logic: (A) CWD for inner strip; (B) CWD for outer strip; (C) Offset Correction and Correlation Logic; (D) flip-flop for stub readout shift register; (E) Hit-Detect output lines to/from adjacent channels; (F) CWD output lines to/from adjacent channels.
2.3.4 Data Readout

The CBC2 provides two separate and independent output data streams: the first consists of L1 data, which includes the output of all channels for a given bunch crossing, and is requested by an external L1 trigger. The second consists of stubs, which in the final version of the ASIC will be output synchronously at every bunch crossing, but in the CBC2 can be read out at 40 Mbps for test purposes.

L1 Data Readout

Block diagrams of the data readout and of the relative control logic are shown in Fig. 2.17. L1 data are continuously written, synchronously with the BX clock, in a 256-deep pipeline memory. Once a L1 trigger is issued to the chip, the control logic copies the data for the BX of interest, together with the pipeline column address (time slice), into the following 32-deep memory buffer, where the data are held for readout before being serialized to the output pad. Data which are not triggered are overwritten once the pipeline cycles at the end of the latency period.

The pipeline memory is 254-bit wide to match the channel count; the latency of the L1 trigger is accommodated by its depth of 256, which at the LHC nominal frequency of 40 MHz corresponds to a maximum latency of 6.4 \( \mu s \) (to be increased...
to 12.5 $\mu$s in the CBC3). The memory cells are dual port Static Random-Access Memory (SRAM), which allow read and write access at the same time, and are discussed in detail in Chapter 4 in the context of the chip irradiation performance.

The writing operation is controlled by two pointers, write and trigger pointer, separated by a programmable delay. The readout time for an event (266 bits at 40 Mbps $= 6.65$ $\mu$s, see Fig. 2.18) is shorter than the average trigger interval (10 $\mu$s at 100 kHz), however due to the random nature of the L1 triggers, losses due to vetoing of the trigger to avoid buffer overflow are both inevitable and easily quantifiable. Losses can be controlled by ensuring that the memory buffer can hold the data until previous data are read out and the output line becomes available again; in so doing increasing its depth, at the cost of added maximum latency.

Simulations show that a depth of 32 is sufficient to achieve an average trigger rate of 135 kHz for a trigger vetoing of $10^{-5}$ (see Fig. 6.8). This figure could be easily improved by increasing the readout speed or the number of output pads. In Chapter 6, however, it will be shown that a buffer depth of 32 is sufficient even for the future CBC3, which for the same level of data loss targets a much more aggressive 1 MHz trigger rate.

Two error flags are output with the header of each data packet: the circuit continuously monitors the latency between write and trigger counters, against the programmed value in the relevant I$^2$C register. It also issues a flag in case of a full memory buffer.

The data packet is shown in Fig. 2.18: it includes a short header, the two error flags, the pipeline address and the channel data, for a total of 266 bits.
2.3 CBC2

Stub Data Readout

As a test feature, the output of the combinatorial logic for each channel is latched into a parallel-input, serial-output shift register. The resulting 131 bits (127 bits for the stub data, 3-bit header and 1-bit trailer) can be streamed out at 40 Mbps to verify the correct operation of the logic.

Two additional test features were included in the CBC2: the OR of all the stub outputs ($\text{OR}127$: to flag the presence of at least one stub candidate), and the OR of all the channels ($\text{OR}254$: to flag any signal above threshold). Either can be selected to be available on a dedicated output line, and provides a prompt indication of the chip activity (Fig.2.11). The stub shift register was designed to be controlled by either of these OR signals, or by their latched versions which are also made available to control the readout. Unfortunately, an error in the polarity of a signal meant that an additional clock cycle is required for the correct latching of the stub data. This could be considered as the only shortcoming of the CBC2; however it should be stressed that the OR functionality was not part of the specifications for the chip and was inserted as an additional test feature. Despite this limitation, which for example prevented the use of the OR output as an external trigger to control a telescope module such as the one described in Chapter 5, the OR can still be used effectively to control the stub readout when injecting a fixed pattern in the channels through the test-pulse circuit.

The logic blocks for the two ORs result in a very long ($\sim10$ mm) and thin ($\sim10 \mu$m) design; they were designed as a tree structure, to minimize propagation delays. A counter is also included to veto further latching of the shift-register and avoid overwriting of the data, until the readout is complete.

2.3.5 On-chip Powering Circuit

The CBC2 incorporates a DC-DC switched capacitor converter, designed by CERN [66], which can provide the required 1.2 V from a 2.5 V supply. This voltage is filtered off-chip and used to power the digital circuits on the ASIC. An on-chip low-dropout
2.3 CBC2

Figure 2.19: Features of the power supply distribution nets inside the CBC2. Left: the gnd net is common for both analogue and digital circuits; the main functional blocks, where the density of circuits is higher and the gnd is distributed to a capillary level, can be clearly discerned from the picture. In the top-right corner lies the DC-DC converter, which has dedicated gnd and power supplies; center: the digital power supply does not propagate to the analogue front-end; right: the analogue power supply feeds mainly the analogue front-end and the bias generator.

linear regulator is used to provide a clean power supply to the analogue front-end and filter out switching noise from the DC-DC converter. Both these elements were functional in the first iteration of the CBC, but some noise injection from the switched capacitor converter was observed in the front-end. The CBC2 incorporates an improved version of the DC-DC converter optimized to reduce switching noise; the effect of the power converter on the performance of the chip is reported in Sec.3.2.5. All other measurements refer to the chip powered by an external power supply through the on-chip linear regulator.

The digital and analogue power domains are well separated on chip by BFMOAT isolation trenches (Section 2.1), while the ground net is common to both analogue and digital circuits. The domains are highlighted in Fig. 2.19.

2.3.6 Bump Bonding

Bump-bonding, or flip chip assembly, is a process first introduced in 1964 [67], in which solder bumps are deposited on the chip and the chip is later flipped and
bonded to the package or hybrid (through solder reflow or mechanical pressure).

The process used with the CBC2 is known as Controlled Collapse Chip Connection (C4) and requires the placement of special under-bump metallisation on top of the chip pads to accommodate and provide adhesion for the solder bump. A consequence is that the top-level metallisations are different for bump and wire-bonded pads. Being the only chip of its run designed for bump-bonding, a number of wafers had to undergo a manufacturing split and be processed for C4 finishing. To take advantage of the chips manufactured on the other wafers, additional wire-bondable pads were included in the design for all the back-end output and control input signals, so that the chip could be powered, controlled and read-out when wire-bonded. The sheer number of input pads would not allow the same duplication, so the chip cannot be connected to a sensor; however the wire-bonded version was useful for the quick development of test procedures for wafer-probing (based on the CBC1 setup), for probing the bump-bonded dice without damaging the solder bump, and for x-ray irradiation of the CBC2 (Chapter 4).

The CBC2 has 768 5-on-10 bump-bond pads, meaning a ball diameter of 5 mil\(^{\ast\ast}\) (127 \(\mu\)m) with a pitch of 10 mil (254 \(\mu\)m). The large number of pads available is one of the main advantages of bump-bonding, together with the much lower inductance of the bond compared to wires [67]. A comparison of the electrical properties of the two methods is presented in Table 2.1; the much lower inductance is especially beneficial for input pads, where it would degrade the sensor signal, and for high speed data I/O. The large number of pads, coupled with the fact that the pads are not confined to the periphery of the chip, reduces the voltage IR drop for grounding and power distribution.

\(\ast\ast\)1 mil = \(\frac{1}{1000}\) inch = 25.4 \(\mu\)m

### Table 2.1: Comparison between the electrical properties of typical wire and bump bonds (from [68])

<table>
<thead>
<tr>
<th></th>
<th>R (m(\Omega))</th>
<th>L (nH)</th>
<th>C (pF)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wire</td>
<td>90</td>
<td>2.58</td>
<td>0.02</td>
</tr>
<tr>
<td>C4</td>
<td>30</td>
<td>0.06</td>
<td>0.17</td>
</tr>
</tbody>
</table>
The effect of the higher parasitic capacitance on the input pads was investigated (Fig. 2.20), and was one of the reasons for separating the input pads from the inter-chip signal pads with the insertion of three columns of grounded bonds (b in Fig. 2.21). Such bonds shield the sensitive inputs from the digital activity of the back-end pads. With reference to Fig. 2.20a, a simple parallel plate capacitor model gives:

$$C = \frac{A \epsilon \epsilon_0}{d}$$

where \( A \) is the area (assuming a bump height after bonding of 85 µm, \( 46 \times 85 \) µm\(^2\)), \( d \) the separation (880 µm), \( \epsilon \) the relative permittivity of the underfill (\( \sim 4 \) for polyimide), and \( \epsilon_0 \) the permittivity of air (8.85 x 10\(^{-12}\) F/m). The resulting parasitic capacitance between the two bumps is limited to \( C = 157 \) aF, therefore negligible.

2.4 Summary

The development of the front-end readout ASIC for the CMS Strip Tracker was presented, starting with an introduction to the CMOS technology of choice, followed by an account of the initial design considerations behind the CMS Binary Chip program.

Building on a very successful first full-scale prototype, the CBC2 introduced several major improvements and new functions, such as the ability to correlate inputs.
2.4 Summary

Figure 2.21: Photograph of C4 bumps on a CBC2 wafer (detail of lower die): (a) input pads; (b) dummy pads (grounded on the hybrid); (c) inter-chip pads for stub-finding logic; (d) ground pads; (e) I/O pads; (f) wire-bonded test pads.

from two different sensors and to identify candidate hits above a programmable $p_T$ threshold.

Since covering every aspect of the design of the chip would be excessive and outside the scope of this work, preference was given to the new areas or those with substantial upgrades with respect to the CBC1. The analogue front-end chain, consisting of preamplifier, gain amplifier and comparator, was presented in detail, together with the subsequent data storage and readout. The stub-finding logic and its building blocks are implemented to provide flexibility in the design and operation of the $p_T$ modules, for example by programming the width of the correlation window, the offset correction or the maximum cluster width. Finally, the implications for inter-chip communication and practical aspects such as chip layout and bonding were also presented (ESD protection forms Appendix B).

The chip was received in January 2013, and has since demonstrated excellent performance, while also being the backbone of several prototype modules for the CMS Phase II Tracker. The results of the chip testing, both in isolation and instrumenting a 2S module prototype, is the topic of the following chapters.
3.1 Test Activities

The testing of the first iteration of the CBC culminated with two beam tests at CERN in the H8 line [51, 69], in which two modules with a CBC1 bonded to a 5 cm strip sensor were used as parasitic planes for a telescope equipped with APV25 analog readout ASICs. Thanks to the fine spatial resolution of the telescope (compared to the sensor strip pitch), and to the small scattering of the 400 GeV proton beam, it was possible to perform precision measurements such as the module spatial resolution and track residuals, which would otherwise be difficult to extract from a binary system.

The results from the electrical testing of the chip are presented in Ref.[54]. Many of the findings are also valid for the CBC2, which inherited many of the circuits from the previous version. At the same time the tests highlighted a few issues, later resolved in the CBC2: the post-amplifier feedback network for example was modified to eliminate distortion introduced when many channels were active at once; and a similar problem was also addressed by redesigning the distribution of the global comparator threshold VCTH.
3.1 Test Activities

Testing of the CBC2 started immediately after the delivery of the wafers with yield screening, and progressed from single-die characterization, to the integration with sensors and the assembly of a mini 2S-module.

This chapter will present the main results from the electrical characterization of the chip [2], while irradiation performance and the results from a beam test will be covered in Chapters 4 and 5 respectively.

3.1.1 Wafer Probing

Four of the eight bump-bonded wafers have been tested so far to provide known-good chips for the assembly of the CBC2 hybrids.

The wafers were screened at Imperial College (IC) (Fig.3.1), with a test setup consisting of:

- semi-automatic 8-inch probe station*;

3.1 Test Activities

- two custom-made probe cards, with conventional blades and epoxy-mounted needles respectively;

- VME-based Data Acquisition (DAQ), including 8-bit ADC, clock and I\(^2\)C cards;

- LabVIEW\(^\dagger\) software to analyse supply current, bandgap and LDO output voltages, bias generator current and voltages, comparator offsets and S-curves. All the configuration registers and pipeline cells are also accessed, to identify stuck bits and addresses.

The probes make contact with the right-most column of redundant wire-bond pads, which were specifically included to avoid damaging the I/O bump-bond pads. The fully-automated test takes about 5 minutes per chip. This is due to the relatively high number of I\(^2\)C operations required to obtain the S-curves, and to the limited speed of the I\(^2\)C for the current setup (100 kHz). Even when operating the I\(^2\)C at its nominal frequency of 1 MHz, however, wafer testing of the \(~150k\) chips required for the Outer Tracker will have to be carefully planned to be feasible in a reasonable time.

The test procedures were refined for the last two wafers, with the offset tuning as described in Ref.[70].

For the first wafer, four die were rejected because of accidental damage. In the other cases, except for one failing visual inspection, the causes for rejection included: too

\(^\dagger\)http://www.ni.com/labview/
3.1 Test Activities

(a) Wire-bonded CBC2 test board, including the I\textsuperscript{2}C Scalable Low-Voltage Signalling (SLVS) interface card for level translation. (b) Dual-CBC2 test board bonded to a pitch adapter with calibrated external charge injection and precision capacitors to simulate interstrip capacitance.

Figure 3.3: CBC2 test setups.

low supply current and average gain, biases not working, one pipeline bit stuck, one channel output stuck, and one unresponsive chip (Fig.3.2).

Overall, excluding the mishandled die, the yield is relatively high at \(~97\%\).

3.1.2 Single Die Testing

The testing of the CBC2 started in February 2013 with the wire-bonded version; because of its similarities with the first CBC, this version was useful to quickly develop test procedures for the bump-bonded wafers (Fig. 3.3a). It was also used in the total ionizing dose irradiation tests described in Section 4.3.

3.1.3 Dual-CBC2 Hybrid

A high-density hybrid for the CBC2 has been designed by CERN as a prototype for the 2S-module hybrids [4, 71, 72]. By accommodating two readout chips side-by-side, it allows one to fully exercise the inter-chip links and the coincidence logic, as well as to test the performance of the bump-bonded version of the CBC2, thereby benefiting from the full connectivity of the high number of power and ground connections.

The hybrid is the so-called “rigid” design which was developed at the beginning of the project, and later abandoned in favour of the flexible version described in
Chapter 1. Despite resulting in a successful and fully functional assembly, the rigid substrate exhibited excessive warpage due to thermal expansion mismatch with the ASICs, and the rigidity was insufficient for reliable wire bonding without the aid of auxiliary stiffeners, as employed by the current design [72].

Initially, this setup was bonded, via a pitch adapter, to a set of precision capacitors to inject input charge and extract the gain of a subset of channels (Fig. 3.3b). Two 5 cm silicon sensors were later mounted and wire-bonded to the hybrid to create a mini 2S readout module (Fig. 3.4).

3.1.4 Mini 2S-Module Beam Test

The sensors used for this assembly were as close as possible to the final specification of 90 $\mu$m strip pitch and 5 cm length. Both polarities were tested, with p-on-n sensors from Infineon Technologies AG (Infineon) and n-on-p from Centro Nacional de Microelectrónica, Barcelona (CNM), which are discussed in more detail in Chapter 5.
3.1 Test Activities

The availability of the front-end ASICs and of the hybrids was important to progress with the module development and integration; in particular the difficulties in bonding both sides of the module helped in moving towards the flexible hybrid design.

So far, the mini 2S-modules have been tested with radioactive source, cosmic rays, and used in several test beams. Several of these mini 2S-modules have been distributed to collaborating institutes, where they have been used to test and develop the DAQ firmware and software. System issues such as common-mode noise and electromagnetic pick-up by the bond wires, for example, are being investigated at a dedicated setup at CERN [73].

3.1.5 Eight-CBC2 Hybrid

A further hybrid prototype has been designed by CERN, which accommodates eight CBC2 chips (Fig.3.5) [72]. This is intended to assess the production of large area substrates, as well as allow further progress with the integration of a module which will be very similar to the final one. Together with the fine pitch of the input pads, the high-density of SLVS differential tracks required for the trigger data readout to the Concentrator (96 per hybrid) has proven to be one the most difficult design constraints. Testing of the first prototype of this assembly carried out at IC [74] has shown very good assembly, with no disconnected channel out of the total 2032 inputs.
3.2 Analogue Test

3.2.1 Test Pulse Circuit

The test-pulse circuit integrated into the CBC2 front-end provides an essential means to test individual chips, without the need for external components. It allows injecting charge into the input nodes, and is used to generate S-curves, evaluate individual channel connectivity, gain and noise. By masking the channels appropriately, it is also possible to verify the functionality of the stub-finding logic.

The circuit consists of a programmable delay line with 1 ns resolution, regulated by a delay-locked loop, coupled to a charge step generator able to provide a maximum charge of 22 fC in 256 steps of 0.086 fC (±10%). The magnitude, polarity and delay with respect to the master clock, are programmable via I²C. The small input capacitance value of 20 fF limited the design choice to the vertical natural capacitors described in Section 2.1, which in this instance have a 3σ absolute precision of ±25%.

3.2.2 Noise and Power

As mentioned in Section 2.3.2, the front-end was originally designed for a capacitance of about 5 pF, corresponding to a strip length of 2.5 cm. Fortunately, sufficient headroom was allowed in the design to tolerate larger capacitances by increasing the input device bias current to meet the original noise target of less than 1000 $e^-$. The noise performance is a function of the input capacitance and the Equivalent Noise Charge (ENC) has been measured to be:

$$ENC[e_{\text{RMS}}] = 500 + 64 \ C_{\text{in}}[\text{pF}]$$

for both electron and hole readout\(^\dagger\). For a 5 cm strip sensor with typical capacitance of 7.5 pF, the CBC2 meets the target readout noise of 1000 $e^-$ for a total power consumption of 350 $\mu$W/channel. Noise and analogue power consumption are shown

\(^\dagger\)This expression is a simplified fit of the data and does not correspond to the series and parallel noise components, which add in quadrature.
3.2 Analogue Test

Figure 3.6: Readout noise vs. strip capacitance for electron and hole readout: full points are measurements, empty points simulations. Simulations included only the front-end, and therefore the power consumption was consistently slightly underestimated.

in figure 3.6 as a function of the external capacitance. The measurements are in good agreement with the expected results from simulation. The increase in power consumption is due to the fact that to maintain the peaking time of the preamplifier, the biasing current in the input transistors is increased according to the input capacitance.

The total digital power consumption was measured at 4.4 mA (Fig.4.8 and 4.9), corresponding to $\sim 20 \, \mu\text{W/channel}$.

3.2.3 Gain

The internal test pulse allows the extraction of the gain of the front-end, which is about 45 mV/fC for both readout polarities. The slightly undulating characteristic of the gain plots in Fig.3.7 is due to a non-linearity of the DAC used to set the threshold. This problem was present but not visible in the CBC1, where the global threshold was supplied externally, therefore masking the performance of the bias DAC.

This is not considered to be a problem, though, as the gain only needs to be linear in the region where the threshold will be set, between 1 and 2 fC.

The test pulse response also allows extraction of the pedestal and the comparator threshold (I²C register setting) VCTH in terms of charge. Fig.3.8 presents the
3.2 Analogue Test

Figure 3.7: S-curve mid-points for all the 254 channels, as a function of the test pulse charge.

<table>
<thead>
<tr>
<th>test pulse</th>
<th>Threshold INF_A</th>
<th>Threshold INF_B</th>
<th>VCTH units CNM_A</th>
<th>VCTH units CNM_B</th>
</tr>
</thead>
<tbody>
<tr>
<td>12</td>
<td>149.8</td>
<td>149.7</td>
<td>119.6</td>
<td>119.8</td>
</tr>
<tr>
<td>18</td>
<td>157.8</td>
<td>156.4</td>
<td>111.0</td>
<td>111.9</td>
</tr>
<tr>
<td>24</td>
<td>165.1</td>
<td>162.3</td>
<td>101.8</td>
<td>103.8</td>
</tr>
</tbody>
</table>

Figure 3.8: Average of test pulse amplitudes, derived from the S-curve mid-points for all channels, for all the four chips in the two mini 2S-modules developed for the beam test described in Chapter 5. The pedestals are extrapolated as intercepts from three different test pulse amplitudes (12, 18, 24). The sensors have opposite readout polarities: as illustrated in the top left figure, representing the output of the post-amplifier and the comparator threshold VCTH; larger settings of the latter correspond to larger signal threshold for hole readout, while the opposite is true for electron readout.
3.2 Analogue Test

**Figure 3.9:** In the CBC2, three registers are available for tuning the channel thresholds (from [70]). **VPLUS:** global setting for DC baseline voltage setting at the post-amplifier output; **VCTH:** global setting for comparator threshold voltage; **OFFSET:** local (per channel) fine tuning of the comparator DC input voltage, by varying a dc current through the resistor R.

Average test pulse amplitudes, derived from the S-curve mid-points for all channels, for the two chips (labelled CBC2_A and CBC2_B) in each of the two mini 2S-modules developed for the beam test described in Chapter 5. Two different sensors with different readout polarities were used, hence while for electron readout larger test pulse amplitude corresponds to smaller VCTH values, the opposite is true for the hole readout. Focusing on the electron readout module (labelled CNM), all the channel thresholds were tuned to 120 VCTH using the test pulse setting of 12/256 (maximum output range of 1.1 V), which for a nominal test pulse capacitance of 20 fF ±10% corresponds to a threshold of:

\[
12/256 \times 1.1 \, \text{V} \times (20 \pm 0.2 \, \text{fF}) = 1.031 \pm 0.1 \, \text{fC} = 6440 \pm 644 \, e^{-}
\]

Similarly, the second reference point at 24/256 for the CBC2_A yields 12880 ± 1288 e⁻.

From Fig.3.8: \( \text{VCTH} \big|_{12/256} - \text{VCTH} \big|_{24/256} = 119.6 - 101.8 = 17.8 \, \text{VCTH} \) which corresponds to \( 12880 - 6440 \pm 10\% \, e^{-} \).

For the CBC2_A, one VCTH unit therefore corresponds to \( (6440 \pm 10\%)/17.8 = 361.8 \pm 36.2 \, e^{-} \). For the CBC2_B, this value amounts to \( 403.8 \pm 40.4 \, e^{-} \).
3.2 Analogue Test

Figure 3.10: Left: S-curve for a binary system, plotting hit count rate vs. threshold scan (from [55]). This should be compared to the ideal step function, in absence of noise. The width of the distribution gives the Gaussian noise. Right: example of S-curve for one channel of the CBC2 (from [70]).

3.2.4 Offset Dispersion and Threshold Tuning

Three different I²C registers affect the comparator threshold setting in the CBC2, as shown in Fig.3.9. Of these, two are global settings (VPLUS and VCTH) which affect all channels at once, and are responsible for setting the DC value at the output of the post-amplifier and at the input of the comparator respectively. Being global, such registers cannot correct for channel-to-channel mismatch, therefore an additional means to calibrate every channel to reduce threshold dispersion is required.

Rather than locally adjusting the threshold voltage VCTH itself, the CBC2 includes a current 8-bit DAC in each channel to introduce a programmable voltage offset in series with the input signal, thereby changing the effective differential threshold VCTH - (VPLUS + OFFSET) (Fig.3.9).

Offset and noise are extracted from the channel S-curves, obtained by plotting the hit count rate versus a sweep of the global threshold voltage VCTH, over a number of repeated measurements (e.g. 100). The mid-point of the resulting S-curve represents the crossing of the threshold, which includes the offset due to channel mismatch, while the width of the distribution is directly proportional to the noise (Fig.3.10).

The offset tuning is performed by adjusting the value of the OFFSET current in every channel until the S-curve midpoints overlap at the desired VCTH threshold.
While it is possible to obtain S-curves by triggering on noise only, without input signal, for the best calibration the chip is triggered with a fixed test pulse from the on-chip input pulse circuit, so that the threshold is calibrated for the target operating threshold. This also avoids having all channels firing at once for low threshold levels, a non-realistic operational condition which introduces artefacts such as common-mode noise and distorts the calibration.

As shown in Fig. 3.11, after calibration the threshold dispersion across the 254 channels is reduced to sub-millivolt level ($\sigma = 70\text{e}^-$), which is the resolution of the calibration DAC.

In the CBC2, a non-monotonicity of the VCTH DAC means that a linear sweep of VCTH results in a discontinuous S-curve. While the software for the tuning procedure can recognize such discontinuity and operate the VCTH register accordingly, this is clearly undesirable and will be addressed in the CBC3, in which the binary-weighted current steering architecture of the current DAC will be replaced by an inherently monotonic resistor architecture, with a finer resolution for even smaller threshold dispersion.

### 3.2.5 On-Chip Power Elements

The CBC2 can be powered by a 2.5 V supply through an on-chip DC-DC converter. The output of the DC-DC converter is filtered off-chip by external capacitors, and
feeds the digital part of the ASIC directly. A low-dropout linear regulator filters the voltage ripples caused by the DC-DC converter, and provides a clean, regulated supply at $\sim 1.1$ V for the analogue circuitry.

**Low-dropout Regulator**

Besides filtering the output switching noise of the DC-DC converter, the main reason for the LDO was to improve the power-supply rejection (PSR) of the front-end. Not being differential, the preamplifier dominates the power supply noise response of the system.

Although the active LDO does not regulate above the bandwidth of its amplifier, filtering is still provided at high frequencies by the resistance of the pass transistor and by the external capacitance.

The efficiency has been measured at $\sim 88\%$, with a standing current of 750 $\mu$A [75].

**DC-DC Converter**

The on-chip DC-DC switched capacitor converter, designed by CERN [66], can provide the required 1.2 V from a 2.5 V supply. The output voltage is filtered off-chip and used to power the digital circuits on the ASIC. Although functional with high efficiency (\(\sim 90\%\)), the circuit was found to affect the front-end performance. The structure of the additional noise was examined by sweeping the timing of the external trigger with respect to the phase of the DC-DC converter, in 25 ns steps [76].

The results for the CBC1 are presented in Fig.3.12a: the shift in pedestals indicates the coupling of the switching transients through the internal chip ground. Excess noise also appears with a frequency of 1 MHz, matching the converter operating frequency.

To alleviate this issue, a new version of the DC-DC converter with slower switching edges, as well as improved radiation hardness, was included in CBC2. The measurements presented in Fig.3.12b confirm the effectiveness of these improvements;
3.2 Analogue Test

(a) CBC1

(b) CBC2 (1 VCTH unit \(\sim \) 2.5 mV)

Figure 3.12: Effect of the DC-DC converter switching activity on the chip noise (courtesy of M. Raymond). The phase between the converter clock and the front-end trigger is found to have an effect on the mid points of the front-end S-curves (common-mode shift) and noise. Although reduced, the correlation is still visible in the CBC2.

the pedestal shift was reduced from \(\sim 100 \text{ mV} \) to \(\sim 25 \text{ mV} \) peak-to-peak, whilst the worst-case noise remained similar at \(\sim 2.5 \text{ mV} \).

Alternatives for powering the module ICs are currently being considered. All the other measurements presented in this work were obtained with an external power supply powering the on-chip linear regulator.

Temperature Measurement

Even though the power consumption of the chip is modest, a wire-bonded CBC2 was imaged with a ThermaCAM SC 3000 camera to identify possible localized failure points (Fig. 3.13). During imaging, the chip was biased in its nominal conditions and acquiring S-curves.

Although the camera offers automatic calibration of the emissivity of standard materials, such as silicon, the calibration could not be performed for the top-most material of the die. An artefact is also evident in the minimum temperature measured: given that the camera sensor is cooled to 70 K, reflection off the highly-reflective ground plane on the board can alter the measurement when the sample is placed very close
3.3 Stub-Finding Logic Test

(a) Thermal image of the wire-bonded CBC2.

(b) Thermal image of the CBC2 ASIC, including scan lines.

c) Temperature profile along the scan lines of Fig.3.13b. The thermal camera was not calibrated so the absolute value of the temperature is not correct.

Figure 3.13: Thermal images of a CBC2.

to the optics. The absolute reading of the temperature is therefore unreliable for figures 3.13b and 3.13c; in the case of Fig. 3.13a however no reflection occurs and the temperature is correctly measured with respect to the ambient.

In Fig.3.13c features such as power buses and shielding layers are clearly visible in the temperature profile along the vertical scan, although the effect could be due to the change in reflectivity of the metal layers distributing the power across the chip. Overall, the measurement confirmed good uniformity across the chip and the absence of hot-spots.

3.3 Stub-Finding Logic Test

The 32-channel test groups are such that channels corresponding to strips aligned above each other on the two sensors layers can be injected at the same time (Fig.3.15): this feature coupled with the possibility to mask individual channels and to offset the coincidence window allows thorough testing of the offset correction and correla-
3.3 Stub-Finding Logic Test

3.3.1 Beta-Source Test

$\beta$ electrons from a Sr$^{90}$ source positioned on top of the microstrip sensors on the dual-CBC2 module were used to qualitatively test the functionality of the CBC2 logic (Fig. 3.14-left). Two distinct profiles are clearly identifiable in Fig. 3.14, corresponding to the hit counts for channels belonging to the two different sensors but read out by the same chip.

3.3.2 Test with Cosmic Rays

The same dual-CBC2 module with its self-triggering capabilities can also be used to effectively trigger on cosmic rays, although the rate is very low ($\ll 1$ Hz) even when the coincidence window is set to its maximum width. An example of a measurement from cosmic ray is given in Fig. 3.15-right: a scintillator is used to trigger the oscilloscope, but the fast-OR of the stubs on the chip also provides a low latency signal that can be used for triggering, after which the stubs stored in the shift register can be read out to compare them with the data stored in the memory buffer for L1 readout.
3.4 Summary

The main results from the extensive characterization of the CBC2 were presented.

The chip meets the design specification of 1000 $e^{-}$ noise for a total power consumption of 350 $\mu W$/channel. It also allows the correction of offsets and mismatch across channels, with an offset spread reduced to 70 $e^{-}_{\text{RMS}}$ after calibration. Finally, the correct operation of the correlation logic and of the data readout could be verified, thanks to the inclusion of an-on chip test pulse circuit (missing in the CBC1), able to inject simultaneously into groups of correlated channels, and later by testing a mini 2S-module prototype with a radioactive source and cosmic rays.

Apart from a minor issue involving non-monotonicity of the global threshold setting, to be corrected in the CBC3, the CBC2 performs extremely satisfactorily and meets all its specifications.
Chapter 4

Radiation Effects and Testing

4.1 Radiation Damage at the HL-LHC

The increase in luminosity and centre-of-mass energy at the HL-LHC will exacerbate the already stringent requirements for radiation hardness for both sensors and electronics. While the degradation of silicon sensors under irradiation greatly affects signal readout through excess leakage current, excess noise, reduced charge collection due to recombination, etc., the scope of this chapter is limited to the readout ASIC and to the direct effects of irradiation on the performance of the CBC2. The results of a comprehensive R&D program to identify silicon sensors for the Phase II upgrade can be found in Refs.[77, 78] and in the Phase II Technical Proposal [13].

Although the highest levels of irradiation affect the pixel detector and the forward regions, the volume of the strip tracker to be read out by the CBC will also be exposed to extreme dose levels, and crucially must be sustained without any maintenance for the duration of the experiment. The projected particle fluence at 3000 fb$^{-1}$ is plotted in Fig.4.1b: it shows the expected strong dependence on radius, while being almost constant in $\eta$. For the outer strip Tracker instrumented with the CBC ($R > 60\text{cm}$), the fluence does not exceed $6 \times 10^{14} \text{ cm}^{-2} 1 \text{ MeV n}_{\text{eq}}$. The corresponding absorbed dose (Fig.4.1a) does not exceed 300 kGy$.\quad\text{\textsuperscript{*}}$ Despite the increased

$\text{\textsuperscript{*}}$The old unit for absorbed dose, the $\text{rad}$, is still commonly used beside the SI standard unit $\text{gray}$ ($1\text{Gy} = 100\text{rad}$)
radius (from $\sim 30$ cm to $\sim 60$ cm), this is twice the maximum dose for the current operation of the APV25 ASICs (150 kGy).

### 4.2 Radiation Effects in CMOS Electronics

Extensive reviews on the effects of radiation in deep-submicron CMOS technologies have been produced within the HEP and Space Science communities [81, 82, 83, 84, 85]. This section will briefly review the dominant effects affecting MOS transistors, since bipolar devices are not present in the CBC.

Radiation effects in electronics can be divided according to their development over time.

**Cumulative Effects**

These are gradual effects due to the accumulation of microscopic defects, which eventually cause a measurable shift in performance and lead to failure once the cumulative radiation exceeds the tolerance limits of the device.

Cumulative effects are subdivided into:

**Total Ionizing Dose (TID)** Ionizing radiation, such as charged particles and high energy photons, lose some of their energy in the form of electron-hole pairs when interacting with the material in the sensor as well as in the ASIC. While a fraction of these recombine, others drift or diffuse according to the local electric field, and can result in the formation of trapped charge via hole-trapping in defect precursors in the oxide, or in the formation of trapping centres at the interface between silicon and silicon dioxide. In CMOS devices, these defects translate into a shift in threshold voltage, reduced mobility and transconductance, increase in noise and in the creation of leakage paths. The total amount of energy deposited by a particle through ionization is referred to as Total Ionizing Dose (TID) [85].
4.2 Radiation Effects in CMOS Electronics

(a) Absorbed dose with the present Tracker layout.

(b) 1MeV neutron equivalent flux in silicon at CMS using FLUKA. The baseline layout of the Phase2 Tracker is superimposed.

**Figure 4.1:** Monte Carlo estimation of radiation dose at 3000 fb$^{-1}$ using FLUKA [79, 80]. The projected dose and 1 MeV n$_{eq}$ fluence for the outer strip Tracker instrumented with the CBC ($R > 60$cm) do not exceed 300kGy and $6 \times 10^{14}$cm$^{-2}$ respectively.
4.2 Radiation Effects in CMOS Electronics

Displacement damage Non-ionizing energy loss can cause atomic displacement in the silicon lattice, which degrades the electrical properties of the device. While bipolar and optoelectronic devices can be particularly susceptible to non-ionizing radiation, CMOS technologies below 250 nm are relatively insensitive at the doses present in the Tracker.

Single Event Effects (SEEs)

These events are due to the energy released by a single particle, and can therefore happen at any moment. Their probability is expressed in terms of cross-section (area), which gives the expected rate when combined with the particle flux. Of particular concern for SEE are secondary ionization events from the recoil of lattice atoms, which can deposit energies three order of magnitude larger than the primary particle, and can be caused by neutral as well as charged particles. These events are referred to as Highly Ionizing Particles (HIPs) [61].

SEEs are further classified into:

Transient SEEs (SETs) are spurious signals or glitches created by an ionizing event, which can propagate through a circuit without necessarily being captured by a latch or register.

Static SEEs are permanent but not destructive, the most common being a Single Event Upset (SEU), which occurs when a transient SEE is captured by a storage node, and one or several bits of information are altered. SEUs and their effect on the CBC2 are discussed in detail in Section 4.9.

Permanent SEEs can be destructive or at best require a power cycle to restore functionality. While power MOS devices are prone to a variety of destructive events, low-power CMOS devices can be affected by Single Event Latchup (SEL), discussed in more detail in Section 4.2.1.

4.2.1 Radiation-Induced Effects and Mitigation Techniques

Several strategies can be adopted when designing radiation-hard integrated circuits. Many of the processes and techniques used stemmed from applications in the
aerospace and nuclear industries; however the requirements for high-energy physics instrumentation and tracking at the LHC experiments in particular exceed the levels of radiation encountered in space environments by many orders of magnitude, and required ad-hoc development. Radiation-hardening can target the fabrication process of an IC, Radiation Hardening By Process (RHBP), or the use of circuit design techniques known to mitigate radiation effects, Radiation Hardening By Design (RHBD).

**Hardening by Process**

RHBP involves the modification of doping profiles, the use of special materials or processing steps to minimize the effects of radiation on performance.

In sub-micron CMOS processes, with a gate oxide thickness of less than 3 nm, the charge trapped in the oxide can be removed through direct tunneling of electrons, and therefore the focus has become the charge trapped in the Shallow-Trench Isolation (STI) and in the buried oxide (BOX) in the case of a silicon-on-insulator (SOI) process. An additional advantage of device scaling is that the doping concentration has increased to maintain relatively high values of threshold voltage, which in turn requires higher level of trapped charge in the surrounding oxide to influence the native field in the transistors [86]. This inherent radiation hardness of deep-submicron technology nodes, the scarcity of industrial providers of radiation-specific processes and the fact that these imply higher cost and lower yield than standard CMOS processes, were some of the reasons for choosing a standard 130 nm CMOS process for the design of the CBC. The higher level of integration and smaller minimum feature size offered by standard CMOS processes can also contribute to reducing the cross-section for single-event upsets.

**Hardening by Design**

ASICs in standard fabrication processes can be made radiation-hard by applying circuit techniques aimed at limiting the negative consequences of radiation on voltage
threshold, leakage current and SEE. This section includes a brief description of some of the most common and effective techniques, and their application to the design of the CBC2.

- **Triple wells**: the main advantage of triple wells is the possibility to bias the well potential to a different voltage with respect to the bulk substrate; this is typically exploited by biasing the body of the nMOS to the same voltage as the source and therefore avoiding the body-effect in the stacked transistors often used in analogue circuits. Thanks to the isolation effect provided by the buried n-type layer, triple wells have been shown to reduce alpha-particle and neutron-induced soft error rate cross-sections in 130 nm and 90 nm SRAMs and latches. However the same argument calling for good contacting of the substrate applies to the buried n-well and care must be taken in reducing the resistance by means of frequent well contacts [86].

In the design of the CBC2, triple well nMOS devices were used for all analogue circuits, and frequent contacts were placed to reduce the resistance of the well. Since they require more area than standard nMOS, they were not used for the digital logic or memory of the CBC2.

- **Enclosed Layout Transistor (ELT)**: several transistor layouts have been proposed to avoid or reduce the parasitic conduction channel induced by TID (Fig.4.2), such as annular, edgeless, H-gate and Dog Bone transistors [87]. The Enclosed Layout Transistor is the most effective, and has been extensively used in HEP applications [88]. The gate is annular and separates the two diffusions so that any leakage current must flow beneath the gate, rather than along the STI oxide. Since the drain is often the most sensitive area to SET, choosing the inner diffusion as drain node can also help in reducing the area and therefore the cross-section for SETs [87]. Poor TID behaviour has been reported for the configuration with inner source [89], which was therefore avoided. A major limitation of this design, other than a precision of ±10% in the extraction of the effective width and length of the device from the physical dimension [81] (Fig.4.3), is that the minimum aspect ratio (width/length) must
be greater than $\sim 4$, and even larger for short channel devices. Transistors with smaller form factor cannot be laid out as ELT.

In the CBC2, all the nMOS devices in analogue circuits compatible with this limitation are implemented as ELTs; the logic and the memory instead employ nMOS with traditional layout. pMOS transistors, being in this instance more radiation hard, are laid out as linear devices to avoid the additional area penalty.

- **Guard rings**: these implants, also called guard bands, bias the substrate and n-wells, and are effective against all radiation-induced effects, requiring only an increase in area. Findings on Field-Oxide FETs (FOXFETs) and logic test structures reported in Ref.[81] demonstrated how even partial guard rings are effective in reducing inter-device leakage current, between diffusions belonging to adjacent transistors, by two orders of magnitude. By providing a low-resistance collection path for charge created by ionizing radiation, guard rings also limit the magnitude and the duration of single-event transients [87]. Guard rings and n-well contacts are profusely used in the layout of the CBC2 to provide isolation between devices, as well as to avoid latch-up, as discussed later in this section.

- **N-well contacts**: the area and the location of contacts in a n-well has been proven to affect the duration of SET pulses generated in a pMOS device [90].
The SET pulse is in fact proportional to the vertical and lateral resistance offered by the well. This is due to the so-called parasitic bipolar amplification effect, where the ionization electrons in the well temporarily change the electric field and activate the parasitic bipolar by forward biasing the source-body junction, which injects additional electrons in the cell. Large n-wells and additional contacts can limit this effect.

- Shortening the SET duration: if the length of a transient is too short, the circuit or the memory element do not have time to respond, in which case the SET is equivalent to a glitch and will not propagate or be captured. Similarly, limiting the speed of a circuit by adding capacitance or resistance is equivalent to rejecting longer pulses. This technique obviously limits the maximum operating frequency of the circuit, and might require additional dynamic power as well as area. Designing larger transistors can also increase the node capacitance, although this might be offset by the increase in collecting area. In this respect, the use of larger ELT addresses both SET and TID resistance, but requires higher dynamic power, customized design libraries and additional area.
4.2 Radiation Effects in CMOS Electronics

- Spatial redundancy: in this case the information is stored in multiple locations and, provided a hit affects only one node, the upset will not be able to propagate, or the correct data can be reconstructed from the remaining uncorrupted data. Several memory elements with multiple storage nodes have been developed [91, 92]: the Dual Interlocked Cell (DICE) [93, 94] for example uses four nodes instead of the two of a SRAM, and relies on adjacent nodes to restore the upset one.

Triplication, or Triple Mode Redundancy (TMR), with majority voting is also a classic example of spatial redundancy, in this case three elements process the same information, and an arbitration circuit outputs the majority output [95]. This of course is only effective against a single upset: accumulation can be prevented by a periodic or autonomous refresh when an upset is detected, but multiple-node upsets can only be limited by interleaving the nodes and spacing them apart. For critical components, TMR can also be applied to functional blocks or more complex circuits, and can be combined with the use of special layout techniques and temporal redundancy.

As described in Section 4.9.1, TMR was used in the configuration registers for the CBC2.

- Temporal redundancy: in this case a delayed version of the input is used in conjunction with the original one, to filter SET shorter than the duration of the delay. The penalty is an increase in the propagation time for the logic, and a limit on the maximum frequency attainable.

- Encoding: another form of redundancy is the addition of supplementary information to the data stored in memories or transmitted off-chip, which can identify or even correct for soft (temporary) and hard (persistent) errors. Several Error Correction Codes (ECCs) are possible [96], differing in the overhead of the additional bits, the complexity of the encoding/decoding operations and the error correcting and detecting performance. One of the most common ECC is the Hamming Encoding (HE). For single-bit correction of a \( m \)-bit data packet, HE requires \( k \) additional parity bits, where [97]:

\[
2^k \geq m + k + 1
\]
Single-bit error correction requires therefore $\log_2 m + 1$ additional bits, so for longer data packets the relevant overhead decreases substantially.

In the CBC, single-bit correction with Hamming encoding is implemented in the data frame header to be stored in the buffer RAM, and the decoding takes place before the loading of the data in the output shift register. An error in the pipeline header would in fact invalidate the whole data from the chip for a particular L1 trigger, and is therefore more critical than other readout data. The pipeline 8-bit address is encoded into 12-bit with the addition of 4 parity bits ($m=8$, $k=4$), which are decoded back to the original 8-bits before writing the address into the output shift register.

- Single Event Latchup protection: when complementary transistors are placed in close proximity, the $pnpn$ parasitic structure highlighted in Fig.4.4 formed by the substrate, $n^+$ and $p^+$ diffusions and the n-well of the pMOS can be activated and initiate a self-sustained, high-current state known as latchup [98]. This configuration with coupled complementary bipolar devices is also known as SCR when used as a power device. Conduction is triggered by overshoot or undershoot of VDD and GND respectively, or by the collection of free charges in the bulk due to hot electron injection or ionizing particles, and leads to a high-current state that will ultimately destroy the device unless it is powered off.

The substrate and well resistances (Fig.4.4) determine the minimum current necessary to turn on the devices; the susceptibility of a circuit to latchup can therefore be reduced by minimizing these. While the vertical resistance is a characteristic of the process, the lateral resistance $R_H$ is proportional to the number of contacts and their spacing from active diffusions. The separation between n-well and nMOS devices is also an important parameter since it determines the current gain of the bipolar devices.

Heavy doping of the n-well, shallow-trench isolation oxide and low power supply contribute to the lack of sensitivity of the 130 nm process to SEL [81, 99]. SEL sensitivity has however been shown to be extremely dependent on the layout: the design kit used for the CBC includes rules to prevent the occurrence
of internal and external latchup; however in a extreme radiation environment such as the HL-LHC, stricter layout practices and RHBD techniques must be followed. Guard rings, with high numbers of contact vias and metal strapping for low resistance, are an effective way to collect minority and majority carriers from the substrate, and to maintain a minimum distance between nMOS and n-well devices.

During the testing of the CBC2, monitoring of the current consumption did not highlight any event compatible with latchup; however no heavy-ion irradiation of the chip has yet been performed.

### 4.2.2 130 nm CMOS Radiation Hardness

Starting with the commercial 250 nm CMOS process currently employed at the LHC, HEP experiments have benefited from the inherent radiation hardness associated with technology scaling [100]. This trend is set to continue, as the 130 nm and 65 nm nodes have already demonstrated higher levels of radiation hardness, as described in Refs.[37, 38, 99, 101], even though the feasibility of the ∼1 Grad goal for the Phase II upgrade of the pixel tracker has yet to be established [39].

The CMOS8RF process used for the CBC has been characterized extensively at CERN. In terms of cumulative dose effects, provided some of the RHBD techniques aforementioned are adopted, it is considered a mature and well-established technology, rad-hard to more than 100 Mrad, well above the requirement for the CBC [99].
One significant TID effect is the threshold voltage shift, which is more pronounced for nMOS, while affecting only narrow core pMOS devices above 100 Mrad. nMOS and pMOS also exhibit a different long-term shift, with the former rebounding after peaking between 1 and 6 Mrad ($\Delta V_{th} < 160$ mV), and the latter increasing steadily ($\Delta V_{th} < 80$ mV). This is due to the different accumulation dynamics and opposite effects of trapped holes and interface states for nMOS, while for pMOS both of these contribute to increasing the threshold.

Higher-voltage devices with thicker oxide are also offered in this process, but being severely affected by charge trapping in the gate oxide ($\Delta V_{th} < 400$ mV), they have been avoided in the design of the CBC2.

The other severe effect for low-power application is the creation of parasitic conduction paths around the edges of nMOS transistors, due to charge trapping in the isolation oxide between devices (Fig. 4.2). As described in Ref. [102], this leads to an increase in static current and for short devices also influences the electric field of the main transistor (Radiation-Induced Narrow Channel Effect - RINCE).

4.3 Total Ionizing Dose Irradiation Test

4.3.1 X-ray Testing

X-ray testers have been extensively used for qualifying electronic components for radiation environments [103, 104]. They offer competitive advantages over the other common source, cobalt-60. Since they can be switched off and the photons emitted have much lower energy and are therefore less penetrating than the $\sim 1.3$ MeV $\gamma$ emission of $^{60}$Co, they have less stringent safety requirements. For the same reason, however, precise dosimetry of the X-ray sources is more critical and must take account of the absorption and attenuation in the Device Under Test (DUT). Other advantages over radioactive sources include more precise collimation (since they produce fewer energetic delta rays in the DUT), high flux rates and availability.

The output spectrum is a superposition of the bremsstrahlung continuum from the primary electrons and of the characteristic lines emitted by the fluorescence of the
target material. The number of photons produced varies linearly with the current in the tube filament, whereas the spectrum depends on the voltage applied. Both are constrained by the maximum cooling power, with typical values around 3.5 kW.

While tungsten is the most commonly used anode material for TID irradiation studies [103, 102, 105], a molybdenum-tube X-ray tester was used for the TID irradiation of the CBC2, because of its availability at the STFC Diamond Light Source Detector Group†.

The nominal wafer thickness of 737 µm, and the fact that the CBC2 is designed to be bump-bonded to the support hybrid, would result in excessive attenuation of the incident flux reaching the layer which houses the active circuitry, either through the bulk silicon of the device itself or through the hybrid material, depending on the direction of irradiation. For this reason the wire-bonded version of the chip, for which the active layer lies only ~25 µm beneath the surface of the ASIC, was irradiated instead.

### 4.3.2 Experimental Setup

- **X-ray irradiation system**: assembled by Hiltonbrooks‡, with a molybdenum X-ray tube (Thales MC 61-04x12 MO LFF - Long Fine Focus): a water-cooled, sealed tube with a maximum supply voltage of 50 kV and maximum tube current of 60 mA (maximum power dissipation of 3000 W) (Fig.4.5).

- **Filter**: 150 µm-thick aluminium foil, positioned in front of the source to remove the low-energy radiation.

- **XY Stepper**: Newport ESP301 Motion Controller§ with 16.7 µm maximum absolute precision.

- **Fluorescent screen**: for initial alignment to the beam.

†[http://www.diamond.ac.uk/Science/Research/Detector.html](http://www.diamond.ac.uk/Science/Research/Detector.html)
‡[http://hiltonbrooks.co.uk/](http://hiltonbrooks.co.uk/)
4.3 Total Ionizing Dose Irradiation Test

- Dosimeter: a 263 µm-thick silicon diode was used for calibration. A Keithley 6485 Picoammeter was used to read the diode current and Keithley 2410 1100V SourceMeter to supply the diode bias voltage.

- Radiant Detector Technologies LLC Vortex™ silicon drift detector [106] to characterize the source spectrum (Fig.4.6).

- DAQ: Custom-made DAQ hardware based on the USB FPGA module DLP-HS-FPGA† developed by Mark Raymond and NI LabVIEW software to control the operation of the chip. Three different setups were developed for room-temperature and cold irradiation (Fig.4.7).

4.3.3 Dosimetry

Details of the dosimetry, from the experimental procedures to the sources of error, are presented in Appendix A. In particular, by measuring the X-ray spectrum and calculating the energy-dependent absorption in the top layers of the ASIC, it was found that the TID in the sensitive volume is attenuated to approximately two thirds of its surface value. This should indicate that the values quoted for the radiation hardness of the CBC2 are inherently conservative with respect to what is found in the literature for other integrated circuits, since a similar analysis, which relies on both the incident spectrum and the metal composition of the ASIC being known, has not been presented elsewhere.

4.3.4 Biasing

Biasing conditions during both irradiation and annealing have been shown to affect drastically the response of a device [36, 89, 102, 103, 107].

In the case of single devices, for example when characterizing a technology for radiation hardness, it is recommended to operate the device in the worst case bias

†http://www.dlpdesign.com/fpga/hsfpga.shtml
4.3 Total Ionizing Dose Irradiation Test

Figure 4.5: Left: interior of the irradiation cabinet, the CBC is positioned on top of the x-y stepper in front of the X-ray source, while the biasing electronics is shielded at the back of the cabinet. Right: positioning of the spectrometer.

Figure 4.6: Left: dosimetry diode, biased and mounted on the shielded CBC2 mounting card. Right: silicon drift spectrometer (beryllium window protected by red cap).

Figure 4.7: Left: accelerated annealing run. The flexible interface bus allows continuous operation of the chip inside the oven. Right: custom-designed DAQ: LabVIEW software, mini-FPGA control module and CBC2 card (shielded).
4.3 Total Ionizing Dose Irradiation Test

conditions: all terminals grounded for pMOS, and gate at VDD and drain and source grounded for nMOS [102]. This however, as discussed in Section 4.8.2, can result in a gross overestimate of the degradation occurring in normal operation [101]. Worst-case biasing is also not possible when characterizing a complex ASIC, with hundreds of thousands of transistors all operating in different conditions. To obtain a realistic picture of the behaviour of complex ASICs under irradiation, the chip should be biased in its normal conditions, possibly allowing for adjustment of control biases to compensate for performance degradations [105].

One advantage of the bespoke DAQ system used for the CBC2 is that it allows continuous operation of the chip under realistic conditions. During the TID testing, with the chip biased to normal conditions and constantly triggered, a one-hour sequence of operations was run continuously, specifically:

- every hour: auto-tune of the channel offsets (for about 10 minutes);
- every 10 minutes: perform a bias scan of one of 10 front-end analogue biases (each recorded every ∼100 minutes);
- every minute: temperature, bandgap voltage, and power supply current and voltage reading, for the chip biased in normal conditions and when all analogue settings are set to minimum value.

4.3.5 Annealing

Some form of thermal annealing is also recommended after irradiation, to capture possible failure mechanisms such as the rebound of nMOS thresholds. One of the standards used for ionizing radiation testing of electronic components, ESA 22900 [108], foresees 168 hours (7 days) of room temperature annealing, followed by 168 hours of accelerated ageing at 100°C. The other common standard, US-MIL-883 [109], prescribes 168±12 hours at 100±5°C. One week of accelerated annealing at 100°C has become common practice for HEP tests [37, 94]; the “ATLAS Policy on
4.4 Room-Temperature Irradiation

Radiation Tolerant Electronics” [110], for example, recommends 24 hours of room-temperature annealing should be followed by 168 hours at 100°C, with electrical measurements after 24 and 168 hours.

The first room-temperature irradiation of the CBC2, presented in the following section, underwent a 136 hours accelerated annealing at \(\sim100^\circ\mathrm{C}\), plus an additional \(\sim12\) days of room temperature annealing. This demonstrated the expected complete recovery of the power consumption to pre-irradiation level.

4.4 Room-Temperature Irradiation

During its first TID test, the CBC2 was irradiated to 10.3 Mrad at room temperature.

The DAQ constantly monitored the currents in the bias/readout board for the 1.2 V and 3.3 V supplies (by measuring the voltage drop across a 1 Ω resistor), the temperature in the irradiation cabinet, bandgap and LDO output voltages. The chip was always biased and operated as described in Section 4.3.4.

The initial irradiation, the duration of which was limited by the availability of the X-ray set, was followed by a period of room temperature and accelerated annealing at 100°C during which the chip was run continuously. The results are presented in Fig. 4.8.

Since the analogue circuits are designed to be radiation hard, one of the aims of the test was to monitor the increase in leakage current from the digital circuits. To be able to discern the component of the 1.2 V current due to digital activity, the analogue settings were periodically programmed to 0 to inhibit the analogue activity. Fig.4.8a shows both the total chip current and this “digital” current.

The “see-saw” structure superimposed on the waveform of Fig.4.8a is due to the periodic activity-related current of the chip, as described in Section 4.3.4. Although more pronounced when the absolute value of the current is higher, the structure is not a result of irradiation.
4.4 Room-Temperature Irradiation

(a) Total chip current with and without analogue biasing.

(b) Instantaneous and integrated dose.

(c) Temperature in the irradiation cabinet.

Figure 4.8: Results for the room-temperature irradiation of the CBC2.
With reference to Fig. 4.8c, the temperature in the irradiation cabinet is clearly correlated with the operation of the X-ray tube, and increases by \( \sim 6^\circ \text{C} \) during irradiation. Because the X-ray power supply would occasionally trip at high current (for example at \( t \approx 4000 \) minutes in Fig. 4.8b), an occurrence particularly inconvenient during overnight operation, the current in the tube was maintained mostly below the initial setting of 50 mA.

A clear feature of Fig. 4.8a is the initial current spike which was observed after an initial delay, once the dose reached approximately 120 krad. The online analysis of S-curve tuning indicated a failure for digital currents above \( \sim 14 \) mA, and the source was subsequently turned off. A swift recovery followed, after which the irradiation was resumed. The consequent current rise was less pronounced and peaked at \( \sim 14.5 \) mA. Further increases in the dose rate (Fig. 4.8b), even to the level that caused the initial current spike, did not result in appreciable increase in current.

This delayed start, suggesting initial accumulation and the presence of a threshold level, the exponential-like rise and fall and the dose-dependent rate, are all features compatible with the current understanding of TID-induced defect development.

As evident from Fig. 4.8a, an almost complete return to pre-irradiation values of the current is attained even during room-temperature irradiation, which is only slightly enhanced by the subsequent room temperature and accelerated annealing (Fig. 4.9).

Along with the current consumption, the bias voltages, which are accessible via an analogue multiplexer pad, were also monitored during irradiation (Fig. 4.10), and exhibit a similar trend, diverging progressively from the nominal value, to recover to the pre-irradiation values after peaking at about 1 Mrad. It should be stressed that at no point during the irradiation were the bias voltages adjusted to compensate for this drift, unlike what would happen during operation in the Tracker.

Apart from the initial current increase during the early stages of irradiation up to 1 Mrad, which promptly decayed, no significant change in behaviour was noticeable, even at very high dose rates. As such, the performance of the CBC2 was considered satisfactory. The initial current spike was attributed to the high dose rate, vastly...
4.5 Low-temperature Irradiation

In excess to what is expected at the HL-LHC, and was therefore not considered a concern for the operation in the Tracker. Nevertheless, to investigate the effect of low operating temperature and to understand the nature and the source of the transient excess current, additional tests and studies, described in the following sections, were performed.

4.5 Low-temperature irradiation

To investigate the effects of the low operating temperature (down to \(\sim-20^\circ\text{C}\)) on the performance of the CBC2 under irradiation, a new setup was created (Fig. 4.11a). The CBC2 board was mounted on stacked Peltier heat pumps and heatsink, and placed in a dry air enclosure. The chip temperature was regulated to \(-15^\circ\text{C}\) for the duration of the test. Crucially, the analogue and digital supplies were provided independently, in order to understand the origin of the transient excess current. The
4.5 Low-Temperature Irradiation

LabVIEW interface was also modified to allow independent monitoring of the chip DAQ and control unit.

As evident from Fig.4.12a, the TID-induced current spike was clearly confined to the digital current, which excluded degradation in the front-end or the bias circuit. Surprisingly, however, the magnitude of the digital current varied substantially depending on the biasing of the analogue part of the chip.

A second irradiation at T=-15°C was performed to try to limit the digital current below 15 mA, to demonstrate that at the dose level expected in operation the CBC2 would not lose functionality (Fig.4.12c). Unfortunately, the minimum continuous
dose rate allowed by the test setup was too high at 85 rad/min, so the X-rays were pulsed to provide an average dose of $\sim 40$ rad/min. This is still higher than the $\sim 10$ rad/min maximum expected dose in the Outer Tracker.

The cooled irradiation therefore showed that the damage is limited to the digital portion of the chip, and that rapid annealing of the damage occurs even at low temperature.

Together with the effect of the analogue biasing on the magnitude of the digital current, one further puzzling result was the observation that in the high-current region, while no effect was visible on the S-curves obtained with the pedestal tuning, the S-curves obtained with the on-chip test pulse circuit were clearly affected (Fig. 4.13). Some, eventually all, channels would stop responding to the test-pulse once the digital current went above 14 mA, to recover completely once the current returned to be below this value. An explanation for this effect will be given in Section 4.8.4.

## 4.6 Masked Irradiation

To pinpoint the origin of the leakage, a further test setup was devised to selectively irradiate only a portion of the chip, through the use of adjustable lead masks.

---

\[ 30 \text{ Mrad} / (10 \text{ years} \times 200 \text{ days} \times 24 \text{ hours} \times 60 \text{ mins}) \]
4.6 Masked Irradiation

(a) First cold irradiation: power supply currents, for analogue biased or unbiased.

(b) First cold irradiation: instantaneous and cumulative dose.

(c) Second cold irradiation: power supply currents, for analogue biased or unbiased.

(d) Second cold irradiation: instantaneous and cumulative dose.

Figure 4.12: Results of two cold irradiation test of the CBC2.
4.6 Masked Irradiation

(a) Pedestal S-curves are not affected by the TID-induced digital current.

(b) S-curves obtained through the test pulse circuit show some (eventually all) channels not responding. All channels recover functionality once the current returns below 14mA.

**Figure 4.13:** S-curves for cold irradiation of the CBC2.

(Fig.4.11b). This excluded the bias generator, the front-end channels, the Hit-Detect logic, the Test Pulse circuit and other subcircuits. The pipeline memory was eventually identified as the source of the leakage. However while unmasked channels failed to respond to the test-pulse, they could still change state when the comparator threshold was brought to pedestal level. The explanation for such behaviour is presented in Section 4.8, following the results of the last high-dose irradiation of the CBC2.
4.7 High-Dose Irradiation

One of the CBC2 chips was irradiated to \( \sim 41 \text{ Mrad} \) to investigate the behaviour at high dose (Fig. 4.14). Having proven the annealing of TID-defects at cold temperature, the chip was irradiated at room temperature, with a dose rate of 5050 rad/min.

The digital current exhibits the expected rise and decay, peaking at \( \sim 30 \text{ mA} \). After the initial peak, reached at \( \sim 600 \text{ krad} \), the power consumption returns to be static.

The output of the bandgap reference circuit is also plotted in Fig. 4.14a: a shift of 23 mV is observed across the whole range, which is modest and not cause for concern. The analogue current, which is referenced to the bandgap voltage, increases accordingly.
4.8 TID Irradiation Results

The results of the different TID tests are presented together in Fig.4.15, which shows the rise and fall of the digital current against the cumulative dose deposited in the CBC2, for various dose rates and temperature conditions.

The current peaks between 500 krad and 1 Mrad, in good agreement with what is reported in the literature for the same technology [99, 102, 111, 101, 36]. Refs.[94] and [37] report a similar behaviour for the digital current, with an increase of $\times 80$ and $\times 300$ at $\sim 1$ Mrad with respect to pre-irradiation values, for 130 nm and 65 nm technologies respectively.

- **Rise**: the waves of Fig.4.15a refer to different dose rates and operating temperatures, which both contribute to the creation and annealing of defect states in the oxide.

TID damage has been shown to be strongly dependent on the dose rate [102]: in Ref.[99], a reduction in dose rate from 24.4 krad/min to 65 rad/min, resulted in a reduction in leakage current of almost 3 orders of magnitude. In the case of the CBC2, although the exposure for the low-dose irradiation performed by “pulsing” the X-ray source (equivalent to $\sim 40$ rad/min) was not sufficient for the current to peak, it is clear how it results in a smaller magnitude for a given TID with respect to the high-dose irradiation at the same temperature.

This dependence is however due uniquely to the prolonged annealing at low doses, rather than to a true dependence on the dose [112]. Reaction rates such as those governing the evolution of trapped charges in the oxide can be modelled by Arrhenius’ Law [102, 113]:

$$k = Ae^{-E_a/RT}$$

where $k$ is the rate constant, $A$ the frequency factor, $E_a$ the activation energy, $R$ the gas constant and $T$ the absolute temperature. Higher temperatures result in higher recombination, and therefore lower TID-induced leakage current. This can be seen in the curves in Fig.4.15a relating to $T=15^\circ$C at 850
rad/min and $T=+25^\circ C$ at 5050 rad/min; despite the higher dose rate of the latter, the higher operating temperature results in a overall lower maximum current.

Overall, results from irradiations at the doses allowed by the X-ray setup used for TID testing of the CBC2 should be considered overly pessimistic.

Simple models, such as those describing the threshold voltage shift or drain-source leakage for a single transistor, are inadequate for a complex phenomenon such as the total leakage current of an ASIC with more than a million individual devices. Leakage arises from a combination of layout, circuit state and operation; the effect of the analogue biasing on the current is attributed in Section 4.8.2 to the operating state of the pipeline.

- Fall: Fig.4.15b presents the annealing rate as a function of time, for different dose rates and temperatures.

The impulse response function of a device to a short irradiation pulse has been successfully used to model the response of a device to irradiations with arbitrary exposure and dose, provided the response of the system is linear [114, 115, 84]. Long-term annealing behaviour for such a system can be described by a linear-with-log($t$) function of the form:

$$-\Delta V_0(t) = -\frac{At}{\ln(t/t_0)} + C$$

where the coefficients $A$, $C$ and $\gamma_0$ can be extracted from a series of steady state irradiations at different doses [114]. However, besides being beyond the scope of the CBC2 testing, this model is inadequate when the system is not linear (for example in the case of hole trapping saturation [84]) or when the effect under study is determined by a number of factors, such as in the case of the chip total digital current.

A double-exponential decay curve was used to fit the data in Fig.4.15b, since the rate of annealing is proportional to the concentration of the defects, and dominated by a fast and slow processes with different decay constants:

$$Y(x) = y_0 + A_1 \exp \left(\frac{x-x_0}{\tau_1}\right) + A_2 \exp \left(\frac{x-x_0}{\tau_2}\right)$$
4.8 TID Irradiation Results

Figure 4.15: Digital current for different irradiation conditions.

The effect of the temperature is correctly reflected in the longer time constants of the -15°C irradiation, with respect to the +25°C.

The evolution of the analogue biases, voltages and S-curves with TID are presented in Fig.4.16. No significant degradation or change are visible.

Interestingly, the S-curve mid-points obtained through the test-pulse circuit (Fig.4.16c) show a monotonic increase of the mid-point across the chip. The effect is too pronounced to be caused by on-chip mismatch between the test capacitances, and is most likely due to a voltage drop along the lines carrying the test-pulse, which originates from one side of the chip (above channel 254). In this case, low-number
channels would see a reduced voltage input across the test capacitor, and be tuned to a lower VCTH value, as observed in Fig.4.16c.

To summarize the results of the TID testing: the CBC2 was found to be rad-hard in excess of 41 Mrad, with no significant shift in performance other than a short-lived initial excess in current consumption, peaking at about 1 Mrad and rapidly annealing back to baseline even at the operating temperature of -15°C. The magnitude of the excess current was demonstrated to be a strong function of the dose-rate, and is expected to therefore be negligible at the rates expected at the HL-LHC.

Despite the excellent performance of the chip under irradiation, significant effort went into understanding the origin of the transient excess current, the effect of biasing on the magnitude of this current, and into explaining the observed disparity between the s-curves obtained with pedestals and those observed with the test-pulse. The findings are presented in depth in the following sections.

4.8.1 Origin of the Leakage Current

Once the temporary increase in static current was found to originate in the memory pipeline, it was possible to investigate the cause of the leakage at the circuit and transistor levels.

The vast majority of the pipeline area is taken by the ∼74k static random-access memory (SRAM) cells. These cells, described in Section 4.9.1, contain minimum size access nMOS transistors and small pull-down nMOS transistors, which are particularly susceptible to the lateral parasitic conduction channels activated by ionizing radiation, as described in Section 4.2.2.

Two possible sources of leakage were investigated: inter-device and drain-to-source leakage.

Inter-Device Leakage

Parasitic leakage paths can form between diffusions biased at different potentials. These paths are studied through the characterization of FOXFETs, in which the
4.8 TID Irradiation Results

(a) Analogue bias currents.

(b) Analogue bias voltages.

(c) S-curves mid-points and noise.

Figure 4.16: Digital current for different irradiation conditions (courtesy of M.Raymond).
Measurements of FOXFET devices under irradiation were performed in Ref.[99], where a device with $W/L=200\mu m/0.92\mu m$ exhibited a leakage current of $\sim 0.3 \mu A$, with $V_{gs}=1 V$. This value was used to estimate the scale of the inter-device leakage in the CBC2 memory, under the simplistic assumption that $I_{\text{leak}} \propto W/L$ and without discriminating between diffusion-to-diffusion and diffusion-to-nwell paths.

Fig.4.17 shows the layout of the SRAM cell used in the memory pipeline. Adjacent cells are well separated, with abundance of substrate and n-well contacts, so inter-cell leakage is excluded.

With reference to the paths highlighted in Fig.4.17, and exploiting the symmetry of the cell and the fact that the two internal nodes store opposite levels, the leakage current of individual paths would be proportional to the $W/L$ of the equivalent FOXFET:

- path $A$ is present in only one of the two nodes during read, not present otherwise (when both Dout and DoutB are precharged high): $I_{\text{leak}A} \propto 0.41/0.3$

- paths $B$ and $E$ are always present in only one of the two sides of the cell:

$$I_{\text{leak}B} \propto 0.41/0.3$$
\( I_{\text{leakE}} \propto 0.34/0.64 \)

- path \( C \) is present on both sides if \( D_{\text{in}} \neq D_{\text{out}} \). The probability of this happening during normal operation is \( P(D_{\text{in}} \neq D_{\text{out}}) = 2\Omega(1 - \Omega) \), where \( \Omega \) is the channel occupancy.

\( I_{\text{leakC}} \propto 2 \cdot 0.28/0.64 \cdot 2\Omega(1 - \Omega) \)

- path \( D \) is present on both sides if \( D_{\text{in}} \neq D'_{\text{in}} \).

\( I_{\text{leakD}} \propto 2 \cdot 0.28/0.61 \cdot 2\Omega(1 - \Omega) \)

- path \( F \), between the n-well biased at VDD and a grounded diffusion, is always present:

\( I_{\text{leakF}} \propto 0.44/0.64 \)

The total inter-device leakage of the memory is therefore:

\[
I_{\text{leakTOT}} \propto \#\text{cells} \left( \frac{0.41}{0.3} + \frac{0.41}{0.3} + \frac{0.34}{0.64} + \frac{0.44}{0.64} + \left( \frac{0.28}{0.64} + \frac{0.28}{0.61} \right) 4\Omega(1 - \Omega) \right)
\]

\[\triangleq \#\text{cells} \left| \frac{W}{L} \right|_{eq}\]

which, for a 2% occupancy, gives:

\[
I_{\text{leakTOT}} \simeq 0.3 \mu\text{A} \left( \frac{200}{0.92} \right)^{-1} 73408 \left| \frac{W}{L} \right|_{eq} = 0.4 \text{mA}
\]

The inter-device leakage is therefore responsible only for a small fraction of the increase in digital current observed under irradiation. This conclusion is in line with what is reported in the literature for a 130 nm CMOS process node [99, 36], for which inter-device leakage was found to be more than two orders of magnitude lower than single device drain-to-source leakage.
4.8 TID Irradiation Results

Drain-to-Source Leakage

Leakage paths between the drain and source diffusions of a single device (Fig.4.2) are the dominant contributors to TID-induced static leakage current for the technology node used for the CBC2 [99, 36].

The SRAM cell in particular contains minimum-size access nMOS and small pull-down nMOS transistors, which are especially susceptible to the effect of lateral parasitic conduction channels. While it was not possible to measure directly the off-resistance of these devices in the CBC2, the leakage current measurements reported in Refs.[99, 101] for the same technology indicates a post-irradiation off-resistance of the order of $\sim 10 \, \text{M}\Omega$ for minimum-size nMOS transistors (100 nA at $\sim 1\, \text{V} \, V_{ds}$).

To verify whether such degradation of off-resistance could explain the excess current observed during testing, the SRAM circuit model was modified to account for the TID-induced off-state leakage; resistors were placed in parallel to the nMOS devices, and their resistance changed until the simulation matched the results of the tests. The pMOS off-resistance was not altered, since they are intrinsically more resistant to radiation damage. The resistances of the access transistor ($R_{\text{access}}$) and of the pull-down transistor ($R_{\text{pull-down}}$) were changed independently in the range 1:10 MΩ.

The digital current resulting from this simulation is presented in table 4.1. Levels compatible with the test results ($\sim +30\, \text{mA}$) are obtained in several cases. By comparing the cases where either of the resistors is 1 GΩ (which corresponds to no degradation), it is possible to see that the resistance of the pull-down transistors is more critical than the one of the access devices. This is likely due to the fact that the former is in the path of any leakage to ground.

The degradation of the nMOS off-resistance under irradiation can therefore, by itself, account for the increase in static current observed in the CBC2.

These simulations also ruled out problems with the memory data line drivers, confirming that, even after severe degradation, the write operation remains effective and the read operation non-destructive.
4.8 TID Irradiation Results

<table>
<thead>
<tr>
<th>$R_{\text{access}}$ nMOS</th>
<th>$R_{\text{pull-down}}$ nMOS</th>
</tr>
</thead>
<tbody>
<tr>
<td>1MΩ</td>
<td>+128mA</td>
</tr>
<tr>
<td>2MΩ</td>
<td>+110mA</td>
</tr>
<tr>
<td>5MΩ</td>
<td>+97mA</td>
</tr>
<tr>
<td>10MΩ</td>
<td>+89mA</td>
</tr>
<tr>
<td>1GΩ</td>
<td>+79mA</td>
</tr>
</tbody>
</table>

| 1MΩ | +89mA | +35mA | +27mA | +19mA |
| 2MΩ | +71mA | +31mA | +23mA | +15mA |
| 5MΩ | +58mA | +34mA | +18mA | +10mA |
| 10MΩ| +50mA | +27mA | +14mA | +6mA  |
| 1GΩ | +39mA | +16mA | +8mA  | 0mA   |

Table 4.1: Increase in static current consumption resulting from the resistors simulating the TID-induced leakage path across the nMOS devices in the SRAM. Values in excess of 30mA are highlighted in red.

4.8.2 Effect of Analogue Biasing

The effect of the state of the analogue biasing registers on the magnitude of the digital current can be explained by the fact that the outputs of the front-end comparators flip when unbiased. Because of the effect of the biasing in the front-end channel, when the analogue circuitry is programmed to be disabled, the output of the comparators are all high, while in normal operation only channels which detect a hit are high.

The biasing conditions of a device during irradiation have already been shown to greatly affect the resulting performance degradation (Section 4.3.4, [99, 101, 107]). Ref.[101] reports that the threshold shift is strongly reduced when all the terminals are grounded, even for minimum-size transistors (Fig.4.18c). Moreover, in this configuration no degradation is measurable in terms of leakage current. A lower gate voltage in fact allows for more recombination of radiation-induced charges under the gate oxide, which reduces the trapping in the oxide and the consequent threshold voltage shift. At the same time, it also favours charge trapping away from the STI-bulk junction, thus reducing the effect on the main and the lateral parasitic transistors. Analogous measurements showing negligible degradation after irradiation at low gate voltage bias are presented in Ref.[107] (Fig.4.18a, 4.18b).

In the case of the CBC2 SRAM, with reference to Fig.4.19, if the comparators are biased during irradiation transistor N2 will degrade faster than N1 because of the higher gate voltage; however the leakage current through N2 will be limited since
the drain node is at 0. When the circuit is unbiased, the comparators flip state and so do the storage nodes of the SRAM. In this configuration, N1 will start degrading more significantly than N2, and at the same time leakage through N2 will increase significantly since the drain node voltage is high. The leakage currents in the biased and unbiased states are the same only if the duration of the irradiation is the same in both states, otherwise, from the above analysis:

\[ I_{\text{biased}} < I_{\text{unbiased}} \quad \text{if} \quad t_{\text{biased}} > t_{\text{unbiased}} \]

\[ I_{\text{biased}} > I_{\text{unbiased}} \quad \text{if} \quad t_{\text{biased}} < t_{\text{unbiased}} \]

which agrees with the results obtained for the CBC2 and presented in Section 4.5, for which the analogue registers were turned off only for a fraction of the irradiation test.
4.8 TID Irradiation Results

4.8.3 Static Noise Margins

The static noise between the two cross-coupled inverters of a SRAM cell is a DC perturbation caused by offsets and process mismatch. It is therefore a well-suited parameter to monitor the behaviour of radiation-induced transistor threshold shifts when investigating the operation of the SRAM under irradiation.

The Static Noise Margin (SNM) is defined as the maximum possible value of DC-offset between the two nodes in the SRAM cell before the cell flips status. The standard way to visualize the SNM is the so-called “butterfly plot”, which shows the DC transfer function of the two inverters in the SRAM cell (Fig.4.20a). The SNM is equivalent to the length of the side of the biggest square that fits in the “eyes” of the butterfly plot. Write, read and hold modes have different SNM: the content of the cell however is more easily perturbed during the read operation, as evident in Fig.4.20a, therefore only the read SNM was considered as the worst case for the following failure analysis.

The calculations of the SNM for the CBC2 SRAM followed the method described in Ref.[116]. Monte Carlo simulations were performed to account for process variation and transistors mismatch.
4.8 TID Irradiation Results

(a) SRAM: "butterfly plot" for hold and read modes. The size of the squares inside the "eyes" represent the maximum SNM.

(b) Simulation of SNM as a function of TID.

Figure 4.20: Simulation of SRAM static noise margins.

From the analytical expression for the SNM, as derived in Ref.[116], it emerges that the SNM decreases with the shift in voltage threshold caused by irradiation. Such offsets depend on the particular geometry of the transistor, and crucially on the fabrication process, which can vary greatly across different manufactures. Although no such data on single transistors were available from the irradiation of the CBC2, the same 130 nm CMOS process has been qualified by CERN, and results for devices of the same flavour and size of those included in the SRAM are published in Ref.[102]. The data refer to 2005, and as manufacturers continually alter the fabrication process in an effort to improve yield and performance, the performance under irradiation might have changed. This particular process is however a mature one, and extensive use in the HEP community has not highlighted any deviation: very similar results for a subsequent batch of the same process are presented in Refs.[99, 101].

For this analysis, the transistor models used in the simulation were modified to account for the TID-induced shift in threshold voltage according to the values from [102]. Data for nMOS of the exact size of both pull-down and access devices were available. For the pMOS, the threshold shift is not significant but for the minimum size device. In the absence of data for the SRAM pull-up transistor (W/L=320/120)
the overly-pessimistic values for a minimum-size device (W/L=160/120) were used instead.

The resulting degradation of the read SNM with irradiation is plotted in Fig.4.20b. As expected, the evolution of the SNM closely resembles the change in $V_{th}$ for nMOS devices, reaching a minimum around 6 Mrad, to recover completely after high-temperature annealing.

Despite a 25% reduction from $\sim$200 mV to $\sim$150 mV, given the number of conservative and worst-case assumptions made, this analysis confirms that the CBC2 pipeline should not suffer from destructive read operations (i.e. the upset of the node state following read access), and suggests that the failure of the pipeline under high-dose irradiation must lie elsewhere.

### 4.8.4 Difference for Test-Pulse and Pedestal S-Curves

Having understood the origin of the TID-induced excess current (traced to drain-to-source leakage in the SRAM nMOS pull-down devices) and the effect of biasing on its magnitude (due to the flipping of the output of the comparators when the chip is unbiased, and to the data-dependent degradation of the SRAM cell), the remaining open question is the failure of the S-curves obtained through the test-pulse circuit in high-leakage state, despite the chip responding correctly to the to S-curves obtained through the pedestal threshold.

To investigate this, simulations similar to those employed in Section 4.8.1 were performed, with resistive elements mimicking the degradation in off-resistance of the nMOS transistors in the SRAM cells. The results are presented in table 4.2. Although the resistance of the pull-down devices was shown in Section 4.8.1 to play the main role in the magnitude of the leakage current, in this case it is the resistance of the access transistors which seems to determine the behaviour of the pipeline.

The simulation highlighted a problem on one of the complementary output bit lines (DoutB) during reading. Figures 4.21a and 4.21b represent the equivalent circuits
Table 4.2: Effect of resistors, simulating the TID-induced leakage path across the nMOS devices in the SRAM, on the correct behaviour of the pipeline. As the colouring highlights, the equivalent resistance of the access transistor seems to be the critical factor.

<table>
<thead>
<tr>
<th>Resistance nMOS access</th>
<th>Resistance nMOS pullDown</th>
<th>Pipeline output</th>
</tr>
</thead>
<tbody>
<tr>
<td>1GΩ</td>
<td>1GΩ</td>
<td>Correct</td>
</tr>
<tr>
<td>10MΩ</td>
<td>1GΩ</td>
<td>Failure</td>
</tr>
<tr>
<td>1GΩ</td>
<td>10MΩ</td>
<td>Correct</td>
</tr>
<tr>
<td>10MΩ</td>
<td>10MΩ</td>
<td>Failure</td>
</tr>
</tbody>
</table>

Figure 4.21: Equivalent circuits highlighting the leakage paths for read operation, and the effect on the output data line.
connected to the DoutB line when a 0 and a 1 are read, respectively (only the half of the SRAM cell connecting to DoutB is included).

In total 256 SRAM cells are connected to a data line, to provide the required trigger latency. Of these, only the one corresponding to the triggered address is read at any given time. The branch to the left of the circuit represents the cells storing a 0 \((N\) cells), the central branch the cells storing a 1 \((255 - N\) cells), and the right branch is the SRAM being read out. The output lines Dout and DoutB are both pre-charged to 1 before a read operation. The on-resistance of the devices (typically in the kΩ range) is much lower than the off-resistance due to the TID-induced conduction channels (in the MΩ range), and can be ignored in the analysis of the leakage paths.

When a 0 is read out \((\text{Dout}=0, \text{DoutB}=1)\) (Fig.4.21a), DoutB remains high. A leakage current through the \(255 - N\) cells storing 1 subtracts from this, but the simulations indicated that even low values of parasitic off-resistance are in this case not sufficient to overcome the drive strength of the pull-up pMOS driving the node. A separate leakage current through the pull-down transistors in the cells storing 0 is also highlighted; this path does not affect the reading operation, even though it contributes to the total static current.

When a 1 is read instead \((\text{DoutB}=0)\) (Fig.4.21b), a leakage through the access transistors of the cells storing 0 prevents DoutB from falling to GND. A separate leakage path exists (central branch), but this is only active during the initial transient on DoutB, and in fact contributes to DoutB going low.

The results from the simulation of the equivalent circuit (Fig.4.21c) are plotted in Fig.4.21d for several values of TID-induced off-resistance, as a function of the number of cells storing a 0. It is clear how, for values of \(R_{\text{access}}\) in the MΩ range, the minimum voltage on the DoutB node is proportional to the number of cells storing 0. Above \(\sim 520\) mV, the digital gate sensing the data output line fails to trigger and the data in the pipeline are therefore incorrectly read as a 0.

The fact that the read operation is sensitive to the number of cells storing 0 can explain why the pipeline fails when channels are triggered through the test pulse.
4.9 Single Event Upsets

SEUs have been introduced in Section 4.2; however in order to analyse the susceptibility of the CBC2 to SEU, the charge collection mechanism is here revisited in greater detail.

The ionization charge deposited by a particle is collected by the sensitive nodes of the device, usually reverse-biased junctions such as drain-bulk junctions. The charge collection is due to three components: drift, diffusion and parasitic bipolar amplification in the case of a well-implanted device. The first gives rise to a fast component (in the order of tens of picoseconds), while diffusion contributes to a slower, prolonged current (hundreds of picoseconds).

One way to harden the cell against SEU is to increase its time constant by inserting resistors or extra capacitance. In the case of a SRAM cell, since a write operation
Figure 4.22: The degradation of the readout lines of the CBC2 pipeline, as observed during testing under heavy irradiation, was replicated in simulation for different data configurations, by modelling leakage paths with parasitic resistors as described in Section 4.8.1 ($R_{\text{access}} = 10\,\text{M}\Omega$, $R_{\text{pullDown}} = 1\,\text{G}\Omega$). When the readout line (Rd) is high, if the bit being read differs from most of the other data in the pipeline (cases b and c) a clear degradation of the data output lines is noticeable.
also relies on overcoming the restoring feedback, any hardening aimed at slowing its
response will also reduce the maximum operating frequency of the memory. Other
drawbacks of this solution are the negative temperature coefficient and poor match-
ing of the polysilicon resistors used, and the area and power penalty of both resistors
and capacitors [117].

Critical Charge

A figure of merit commonly used in defining the radiation performance of a circuit,
the critical charge $Q_{\text{crit}}$ is defined as the minimum charge collected at a sensitive
node required to create an upset [118]. Even for a static circuit, not only the
amount of charge deposited but also the temporal evolution of the pulse determine
the outcome of an SEE. As already mentioned, whether a voltage spike results in an
upset or not depends on the dominant pole of the feedback mechanism for SRAM.
The duration of the pulse can vary from few to hundreds of picoseconds [119], and
depends on the layout of the cell, the relative position of the ionization charge with
respect to the sensitive node, the local electric field (and therefore on the doping
and fabrication process) and the particle species (Linear Energy Transfer (LET) and
charge density).

Among the several current models used to simulate a SEE event, the double expo-
nential is the most commonly used [120, 121, 122]:

$$I(t) = I_{\text{max}}(e^{-t/\tau_f} - e^{-t/\tau_r})$$

where $\tau_r$ and $\tau_f$ are the rise and fall time constants of the pulse, and $I_{\text{max}}$ is the
maximum current which is swept during simulation until the circuit under study
is upset. The corresponding critical charge is obtained by integrating the current
pulse:

$$Q_{\text{crit}} = \int_0^{T_F} I(t) \, dt$$

where $T_F$, called the flipping time, defines the point at which the internal restor-
ing feedback takes over to irreversibly flip the state of the circuit [123]. Simpler,
piecewise linear pulses have also been used [124], but are not suitable when the time
4.9 Single Event Upsets

Figure 4.23: *Left:* double-exponential current pulses used for the simulation of the critical charge for node A of the majority-voting register. The rise time constant is increased from 8ps to 32ps, with $\tau_f = 10\tau_r$. *Right:* corresponding node voltage.

Figure 4.24: Critical charge extracted from the pulses of Fig.4.23a. $Q_{\text{crit}}$ is the total integrated charge, while $Q_{\text{crit}2}$ is the charge up to the flipping time.

constant of the circuit becomes comparable to the pulse duration. 3D device simulations of SRAM cells [122] have highlighted how, among other models proposed in the literature, the double-exponential curve best approximates the current pulses generated by particles with relatively low LET, whereas in the case of higher-LET particles or ions the longer collection tail from the diffusion of deep carriers is not correctly modelled.

For dynamic circuits, the critical charge also varies with respect to the timing of the clock and other input signals.

Figures 4.23 and 4.24 show the results of a simulation of the critical charge for one of the majority-voting SRAM registers used in the CBC2. The pulses in 4.23a all deliver the minimum charge required to upset the cell, but differ in time constant,
although the structure of the pulse is maintained. Longer pulses are associated with overall higher charge (Fig.4.24), although from the temporal evolution of the voltage on the upset node (Fig. 4.23b) it is clear how a significant fraction of this charge is collected after the node voltage has already tipped. In general, shorter pulses, which model ionizing particles directly striking a sensitive node, represent the worst-case condition and give the lowest critical charge. The dependence of $Q_{\text{crit}}$ on the precise evolution of the pulse was already identified in Ref.[118], in which the authors recommend that only the charge collected before the flipping of the node is included in the value of $Q_{\text{crit}}$. Results similar to Fig.4.24, highlighting the linear dependence between $Q_{\text{crit}}$ and the pulse width, are presented in Ref.[119]. If only the charge collected up to the flipping time is considered, the dependence of $Q_{\text{crit}}$ is clearly reduced, but still evident. This linear relation is captured by the classical equation for the critical charge:

$$Q_{\text{crit}} = C_N V_{DD} + I_{DP} T_F$$

where $C_N$ is the total node capacitance, $V_{DD}$ the power supply voltage, $I_{DP}$ the maximum restoring current (in the case of a SRAM cell, the pMOS drain current), and $T_F$ is the flipping time [123]. The change in $T_F$ due to the pulse time-constant (Fig.4.23b) is thus responsible for the increase of $Q_{\text{crit}}$.

While the absolute value of $Q_{\text{crit}}$ obtained from the simulations depends strongly on the assumptions made about the input pulse, the fact that the double exponential curve has been verified against realistic 3D models, and the linearity of the relation between $Q_{\text{crit}}$ and $\tau_r$ evident in Fig.4.24, suggest a linear relation between the values extracted through circuit simulation and the real $Q_{\text{crit}}$. When a fixed current pulse is used to compare different circuits on the same device, $Q_{\text{crit}}$ therefore gives a useful indication of the relative SEU-hardness of the architecture, irrespective of the missing linear factor.

The sensitive nodes for the collection of ionization charges are the highly-doped drain and source $n^+$ and $p^+$ implants. Being reversed biased junctions, $n^+$ collect ionization electrons, so that such a node can only be upset when storing a 1. The
opposite is true for $p^+$ diffusions. Just like the critical charge, which accounts for the strength of the restoring current and for the capacitance of a particular node, the area of the diffusion also determines the SEU cross-section for a particular cell. Complex memory comprising several transistors can be upset in multiple ways: by defining an upset due to an $n^+$ node as $n$-mode, and one due to a $p^+$ as a $p$-mode, the total SEU cross-section can be expressed as [124, 125]:

$$\sigma = K_n \sum_{i=0}^{N} \frac{\text{area}_i}{Q_{\text{crit}_i}} + K_p \sum_{i=0}^{P} \frac{\text{area}_i}{Q_{\text{crit}_i}}$$

where $N$ and $P$ are the number of $n$ and $p$-modes, while $K_n$ and $K_p$ account for the collection efficiency of electrons and holes ($K_p < K_n$ because of the reduced sensitive volume due to the limited depth of the n-well).

By measuring the SEU cross-sections for particles with different LETs, it is possible to extrapolate the values of $K_n$ and $K_p$, as performed in Refs.[124, 125]. While such measurements have not yet been performed for the CBC2, the area of the sensitive regions, as extracted from the layout, is provided together with the critical charge in Tables 4.4 and 4.5.

### 4.9.1 Memory Elements in the CBC2

Since SEU-hard storage elements require additional area, their use must be balanced against the expected upset rates and severity.

While low-rate corruption of readout data can be tolerated, thanks to the fact that multiple points from the tracker are needed to identify a track, an upset in the data stored in the pipeline control logic (such as read and write pointers) would cause a loss of synchronization which could only be restored through an external reset command. An upset in the configuration registers would also change the behaviour of the chip. Given the number of front-end ASICs in the Tracker, the downtime associated with an external reset could be much more critical than strip data corruption, and different strategies must be considered for the protection of memory elements against SEU (Table 4.3).

In the case of the CBC2, three different memory elements were used:
### 4.9 Single Event Upsets

<table>
<thead>
<tr>
<th>SEU effect</th>
<th>Readout Control</th>
<th>Configuration Registers</th>
<th>Readout Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Critical</td>
<td>SEU-hardened FF + ECC</td>
<td>Triplicated SRAM</td>
<td>-</td>
</tr>
<tr>
<td>Tolerable</td>
<td>-</td>
<td>-</td>
<td>SRAM</td>
</tr>
<tr>
<td># cells</td>
<td>1037</td>
<td>2464</td>
<td>73536</td>
</tr>
</tbody>
</table>

**Table 4.3**: SEU strategy and circuits used for the CBC2 registers.

- **SRAM**: the pipeline and buffer memories amount to the vast majority of the storage elements (~74k), so small-area registers are required. Provided it does not add significantly to the percentage of spurious hits due to the noise in the system (specified to be < 1%), corruption of readout data can be tolerated. Therefore a standard eight-transistor (8T), non rad-hard and custom-designed dual-port SRAM was used.

- **Majority-voting registers**: these registers employ triplication and majority voting output to reduce the SEU cross-section. The three basic cells are SRAM cells. These registers are used to store all the configuration bits programmed through the I2C interface. The version used for the CBC1 was modified in the CBC2 to include an external refresh command, to avoid accumulation of upsets.

- **Rad-hard flip-flop**: the chosen architecture is based on the cell proposed by Liu and Whitaker [126, 127]; a similar version designed in the same IBM 130 nm CMOS process used for the CBC was reported in Ref.[92] to have an extremely low cross-section ($2.69 \times 10^{-16}$cm$^2$/bit).

**SRAM**

The SRAM is a dual-port design, which allows simultaneous read and write access to the memory (except for the same cell). Differential inputs and output lines are used to improve the noise margins in read and write operations. The two storage nodes form the outputs/inputs of two cross-coupled inverters; two nMOS access transistors on each side are controlled by the read and write lines and connect to the input and output lines. The transistors are sized so that the write operation can
overwrite the value stored in the cell, while a read access to the storage nodes does not affect its value.

The cell size in the CBC2 was reduced by 50% with respect to the CBC1 SRAM through the sharing of substrate contacts and power lines between adjacent cells: the area of the CBC2 pipeline is the same as the one in the CBC1, despite having twice as many channels. To compensate for the increased load of the address lines, the write and read pointer drive strength was increased. The net result was a saving in die area without affecting the performance of the pipeline.

**Triple Majority-Voting SRAM**

This register consists of three SRAM storage cells, followed by majority voting logic (Fig. 4.25). Over time, subsequent upsets can accumulate and upset the cell, so the version in the CBC2 was modified to include an external refresh signal, which overwrites the content of all cells with the output of the majority logic. The refresh command is a global signal with a dedicated input pad, and its frequency can be chosen according to the rate of upsets experienced. A self-correcting version, able to detect an upset and initiate its refresh, was considered for the CBC3, although in light of the SEU test results described in Section 4.9.2, an implementation based on the rad-hard flip-flop is now favoured.

Two flavours of this register exist, here named Majority Voting Register (MVR) Set and Reset, which differ solely in the power-up default value (1 and 0 respectively).

**SEU-Hardened Flip-Flop**

The SEU-hardened flip-flop used in the pipeline control logic is based on the cell first proposed in Ref.[126] and later improved in Ref.[127].

<table>
<thead>
<tr>
<th>Node</th>
<th>Capacitance [fC]</th>
<th>$p^+$ diffusion area [$\mu$m$^2$]</th>
<th>$n^+$ diffusion area [$\mu$m$^2$]</th>
<th>$Q_{\text{crit}}$ (0 → 1) [fC]</th>
<th>$Q_{\text{crit}}$ (1 → 0) [fC]</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.88</td>
<td>0.102</td>
<td>0.314</td>
<td>32.0</td>
<td>13.7</td>
</tr>
<tr>
<td>B</td>
<td>0.91</td>
<td>0.102</td>
<td>0.307</td>
<td>13.9</td>
<td>32.0</td>
</tr>
</tbody>
</table>

Table 4.4: SRAM cell: SEU-relevant parameters.
Figure 4.25: Circuit schematic of one of the two flavours of majority voting registers used for the configuration data (MVR-Set).

Table 4.5: Majority-voting SRAM (Reset): SEU-relevant parameters.
4.9 Single Event Upsets

Figure 4.26: Circuit schematic of the SEU-hardened flip-flop used for the pipeline control logic.

The basic memory element of this circuit is, like a SRAM cell, based on cross-coupled and self-restoring storage nodes (Fig.4.26). As explained in Ref.[127], the resistance against SEU is attained through redundancy, restoring feedback and by exploiting the polarity of charge collection. N-junctions, which are biased positively with respect to the substrate, only collect ionization-induced electrons, therefore a 0 stored on a node consisting exclusively of nMOS junctions is inherently incorruptible. A symmetric argument holds for 1’s stored on p-junctions. Both the master and the slave sections are split into such complementary storage nodes.

Aside for the pipeline control logic, these registers are also used for the trigger data and stubs shift-registers.

4.9.2 SEU Testing

A first empirical characterization of the Single Event Upset rates for the CBC2 was conducted in September 2014 using the 62 MeV proton beam at the UCLouvain Light Ion Irradiation Facility (LIF)**, the findings of which are summarized in Ref.[128].

The SEU contribution from low-energy neutrons (<20 MeV) is roughly one order of magnitude below that of higher energy hadrons. Moreover, above 20 MeV different

**http://www.cyc.ucl.ac.be/LIF/LIF.php
hadrons have similar cross-sections [129]. Without differentiating between hadron species, a first approximation of the expected SEU rates at the HL-LHC, \( R_{\text{TOT}}^{\text{HL-LHC}} \), can therefore be derived by scaling the measured proton SEU cross-section according to the relative fluences:

\[
R_{\text{TOT}}^{\text{HL-LHC}} \simeq \frac{\Phi_{\text{HL-LHC}}^{>20\text{MeV}}}{\Phi_{\text{LIF}}^{p}} R_{\text{LIF}}^{p} \simeq \frac{1 \times 10^{7} \text{cm}^{-2} \text{s}^{-1}}{2.5 \times 10^{8} \text{cm}^{-2} \text{s}^{-1}} R_{\text{LIF}}^{p} = 0.04 R_{\text{LIF}}^{p}
\]

where the total expected hadron fluence in the outer strip Tracker \( (R = 60 \text{ cm}) \) was derived from [130].

A single wire-bonded CBC2 module, analogous to the one used for TID testing (Fig.4.11b, without lead mask), was irradiated. The CBC2 current, beam fluence and TID dose are plotted in Fig.4.27. The equivalent TID can be calculated as the product of integrated fluence and linear energy transfer, or to a first approximation from the integrated fluence only knowing that \( 10^{11} \text{p/cm}^2 \) at 60 MeV correspond to 13.8 krad in silicon [81]. The slow increase in standing current can be attributed to the effect of TID on the SRAM memory described in Section 4.8. The peaking at about 1 Mrad is also consistent with what was presented in the same chapter. Sudden changes in current are instead due to the programming of the CBC2 in different configuration states.

**SEU Tests and Results**

Different tests were conducted to investigate the different memory elements in the CBC2:

**SRAM** In the case of the pipeline, the test looked at the data stored in one of the 256 frame columns. After an initial external reset command and a short data acquisition phase, the clock was disabled to avoid overwriting the data before the issuing of a trigger, in order to store the data in the pipeline for longer than the 6.4 \( \mu \text{s} \) maximum latency.
No data flips were observed during the $3.1 \times 10^3$ s duration of the test, corresponding to 22 hours of operation at the HL-LHC. This allows the setting of an upper limit for the SEU rate per chip (90% CL):

$$R_{\text{chip}}^{\text{SRAM}} < 7.1 \times 10^{-3}\text{s}^{-1}$$

which corresponds to a cross-section of:

$$\sigma_{\text{bit}}^{\text{SRAM}} < 1.1 \times 10^{-14}\text{cm}^{-2}$$

The performance of the SRAM is therefore more than adequate and well below the specification for noise hits reported in Section 4.9.1.

**Triple Majority-Voting SRAM**  The I²C registers were periodically read and compared against the written configuration. The direction of the upset was also recorded (from 0 to 1 or vice versa). The external refresh signal which instantiates
### Table 4.6: Number of SEU events and respective cross-section for the I\(^2\)C majority-voting registers (from [128]). The flux error is primarily due to the stability of the run. * denotes 1 ms long refresh pulse sent from software; † CBC2 rotated at 45° angle of incidence.

<table>
<thead>
<tr>
<th>I(^2)C Refresh</th>
<th>Time [s]</th>
<th>(\Phi[10^8\text{cm}^{-2}])</th>
<th>SEUs</th>
<th>Bit rate [s(^{-1})]</th>
<th>(\sigma[\text{cm}^{-2}])</th>
</tr>
</thead>
<tbody>
<tr>
<td>disabled</td>
<td>4200</td>
<td>2.44 ± 0.09</td>
<td>5</td>
<td>(9.9 ± 4.4) (\times 10^{-7})</td>
<td>(4.0 ± 1.8) (\times 10^{-15})</td>
</tr>
<tr>
<td>1*</td>
<td>4200</td>
<td>2.42 ± 0.03</td>
<td>1</td>
<td>(2.0 ± 2.0) (\times 10^{-7})</td>
<td>(7.9 ± 7.9) (\times 10^{-16})</td>
</tr>
<tr>
<td>10*</td>
<td>3600</td>
<td>2.39 ± 0.14</td>
<td>6</td>
<td>(1.4 ± 0.6) (\times 10^{-6})</td>
<td>(5.6 ± 2.3) (\times 10^{-15})</td>
</tr>
<tr>
<td>10</td>
<td>2400</td>
<td>2.64 ± 0.01</td>
<td>5</td>
<td>(1.7 ± 0.8) (\times 10^{-6})</td>
<td>(6.9 ± 3.1) (\times 10^{-15})</td>
</tr>
<tr>
<td>disabled†</td>
<td>3600</td>
<td>2.57 ± 0.61</td>
<td>4</td>
<td>(9.3 ± 4.6) (\times 10^{-7})</td>
<td>(3.7 ± 1.9) (\times 10^{-15})</td>
</tr>
<tr>
<td>disabled</td>
<td>4200</td>
<td>2.44 ± 0.09</td>
<td>11</td>
<td>(2.1 ± 0.6) (\times 10^{-6})</td>
<td>(8.4 ± 2.5) (\times 10^{-15})</td>
</tr>
<tr>
<td>1*</td>
<td>4200</td>
<td>2.42 ± 0.03</td>
<td>5</td>
<td>(9.5 ± 4.2) (\times 10^{-7})</td>
<td>(3.8 ± 1.7) (\times 10^{-15})</td>
</tr>
<tr>
<td>10*</td>
<td>3600</td>
<td>2.39 ± 0.14</td>
<td>4</td>
<td>(8.9 ± 4.5) (\times 10^{-7})</td>
<td>(3.6 ± 1.8) (\times 10^{-15})</td>
</tr>
<tr>
<td>10</td>
<td>2400</td>
<td>2.64 ± 0.01</td>
<td>21</td>
<td>(7.0 ± 1.5) (\times 10^{-6})</td>
<td>(2.8 ± 0.6) (\times 10^{-14})</td>
</tr>
<tr>
<td>disabled†</td>
<td>3600</td>
<td>2.57 ± 0.61</td>
<td>16</td>
<td>(3.6 ± 0.9) (\times 10^{-6})</td>
<td>(1.4 ± 0.4) (\times 10^{-14})</td>
</tr>
</tbody>
</table>

the majority voting was either disabled or operated at different frequencies to evaluate the effect of rate on the SEU cross-section. The results are summarized in Table 4.6.

A surprising observation is the absence of any significant improvement when the refresh was active. Increasing the refresh rate or pulse length (from the standard 25 ns to a 1 ms long, software-controlled refresh) also failed to affect substantially the upset rates. This should not be the case if the hits in the three SRAM nodes were uncorrelated, and therefore points to an interaction that can be explained by examining the cell layout.

Fig. 4.28 shows the physical placement and highlights the separation between the storage nodes of the cell: the three bits are separated by only 2.4 \(\mu\text{m}\), which is below the range in silicon of secondary fragments such as heavy ions or delta electrons, and are therefore susceptible to simultaneous upset, against which the majority voting is ineffective. This hypothesis is substantiated by the observation of multiple-bit upsets in the 8-bit I\(^2\)C configuration registers: up to three bits from the same register were observed in one instance. The proximity between the individual elements in a 8-bit register could account for such an effect.
Configuration errors are critical and can only be corrected by an external write operation. The measured cross-section does not meet the requirement for CMS, where the configuration is written before a one-day data-taking run. This issue must therefore be addressed in the next iteration of the chip. While it would be possible to maintain the majority voting scheme by spacing the storage nodes further apart (possibly through automatic placing of the cells [95]), this would require a significant change in the design and potentially a departure from the current channel-based register layout (Fig. 2.15). The results would also still need to be qualified against SEU, and presents the risk of introducing additional pitfalls. The preferred strategy for the CBC3 is therefore to replace the existing I²C registers with the SEU-hardened flip-flop used in the pipeline control logic, which demonstrated excellent SEU performance.

**SEU-Hardened Flip-Flop** To evaluate the pipeline control logic memory, the CBC2 was triggered after a fixed number of clocks following an initial reset. The chip monitors the difference between the trigger and write pointers, and flags one of the error bits in the header if they do not match the programmed latency. An error can also be detected if the read out pipeline address does not match the triggered one.
A total of three SEU events was recorded during a 3.3 hour test, corresponding to a SEU rate per chip:

\[ R_{FF} = (3.6 \pm 2.0) \times 10^{-2}\text{hour}^{-1} \]

Specifically, two of the events consisted of the wrong pipeline header address, while in one case all the channels were set to 1 instead of 0, which could be the result of a SEU on the clock line. Such an event could easily be identified as spurious and discarded.

The SEU rate for the pipeline logic is therefore very low, and only a low frequency fast reset will be required in the tracker to recover upset chips.

4.10 Conclusions on Irradiation

The CBC2 has been proven to exceed the required level of radiation hardness against total ionizing dose [131].

A series of tests up to 41 Mrad, performed under realistic conditions (at cold temperature and with the ASIC biased and acquiring data or telemetry), highlighted no functional issue other than a transient excess current, the nature of which was presented at length. The current was traced to static leakage in the memory pipeline, and was attributed to the very intense dose rate used for the test, much higher than that which will be present in the Tracker: the combination of lower dose rate and longer annealing (even at -15°C) should make this effect vanish during operation at the HL-LHC. Despite this, several improvements emerging from these studies for the next iteration of the chip will be presented in Section 6.1.4. It should also be stressed that no adjustment to the bias registers was performed to compensate for the effects of irradiation.

More recent tests performed at the Karlsruhe Institute of Technology with X-ray irradiation up to 120 Mrad (~0.7 Mrad/h) confirmed no additional degradation of the analogue bias signals or loss of channels [132].
Finally, the results from the SEU testing of the memory elements in the CBC2 were presented. While two out of the three registers types used show adequately low cross-sections, the triple-voting circuit used for the I²C registers will be replaced in the CBC3 with the flavour currently used for the pipeline control registers.
Chapter 5

Performance Studies

5.1 CBC2 Beam Test

In December 2013, two mini 2S-modules were tested at the Deutsches Elektronen-Synchrotron (DESY) in Hamburg. The beam test itself as well as the main results from the extensive collaborative data analysis that followed (cf. [133, 134]) are presented in this chapter.

5.1.1 Experimental Setup

Two mini 2S-modules (described later) were inserted between the telescope planes of the DESY DATURA telescope* (Fig. 5.1). The telescope is composed of two arms, with three planes each equipped with MIMOSA26 Monolithic Active Pixel Sensors (MAPSs) [135].

The aim was to exploit the 4 μm resolution of the telescope to study the spatial resolution of the strip module; however a timestamp issue prevented the synchronization between the DUT and telescope data sets. During the test, an issue was discovered with spurious, random double triggers from the Trigger Logic Unit (TLU), which

*https://twiki.cern.ch/twiki/bin/view/MimosaTelescope/WebHome
Figure 5.1: Top view of the detector planes (strips perpendicular to the page): two mini 2S-modules are inserted between the telescope planes, with the upstream one mounted on a rotational stage. Scintillators at the two extremities provide trigger information.

were later suppressed through a modification of the DAQ firmware. Since the telescope DAQ was not affected by the fake triggers, large discrepancies in the number of events existed between the two early data sets. Moreover, even after fixing the trigger, a difference in the definition of timestamp between the two DAQs meant that a direct match was not possible. Later attempts to synchronize the data by looking at the correlation in hit patterns were ultimately unsuccessful [134].

A module with p-type sensors was mounted on a high-precision rotational stage between the two arms of the telescope. By controlling the angle of incidence of the monoenergetic beam, it was possible to replicate the effect of charged particles with tuned $p_T$ in the magnetic field of the CMS Tracker, as described in Sec. 5.1.2. The second module, instrumented with n-type sensors and positioned downstream with respect to the beam, served as a fixed reference for the event selection.

Two scintillators at both the entrance and exit windows were used to provide a trigger for the data acquisition.

The data-taking included several runs required to commission the software, the TLU and to calibrate the system as described in Section 3.2.4. Subsequent data included:

- threshold scans: the threshold of the CBC2 in one of the two modules was swept to investigate related performance such as noise, occupancy, efficiency, etc.;
Figure 5.2: Generation of electron and positron beams at DESY (from http://testbeam.desy.de).

- trigger latency scan: to identify the optimum delay between the external trigger and the pipeline readout address;

- angular scan: the CNM module was positioned at fixed angles to evaluate the effect on cluster width and stub performance;

- cluster width scan: the maximum cluster width setting in the CBC2 (which defines the maximum width of a cluster to be considered as valid by the on-chip stub-finding logic, as described in section 2.3.3) was changed from 3 adjacent strips (default) down to 2, 1 and bypassed.

In total, more than 1.2 M events were saved to disk.

**Beam Characteristics**

The T21 beam line used for the test provides electrons or positrons with energies from 1 GeV to 6 GeV. A bremsstrahlung beam of photons is extracted from the synchrotron electron beam by a carbon fibre target; the photons are subsequently converted into $e^-e^+$ pairs by a secondary metal target. A dipole magnet and collimator select the energy of the final beam (Fig.5.2). The resulting beam has an energy spread of $\sim 5\%$, a divergence of $\sim 1$ mrad [136] and an average rate of $O(1kHz)$. 
In the case of the CBC2 testing, a 4 GeV positron beam was selected. Fig. 5.3 shows that at this energy radiative energy loss is the dominant process. Most of the radiated energy, however, is deposited outside the fiducial volume of the strips, and the beam can be considered as minimum ionizing particles [137].

The four sensor planes were centred on the beam, as clearly visible from the hit counts of Fig. 5.4.
(a) Mini 2S-module in the temperature-controlled, light-tight aluminium box, which also acts as the heat sink for the Peltier. Windows above and below the sensor, protected by aluminium foil, are wide enough to give large beam acceptance even at large tilt angles (courtesy of Mark Raymond).

(b) Data acquisition system (from [139]).

**Figure 5.5:** Beam test module and DAQ.
5.1 CBC2 Beam Test

<table>
<thead>
<tr>
<th>Sensor</th>
<th>Type</th>
<th>Pitch</th>
<th>Strip width</th>
<th>Thickness</th>
<th>Length</th>
<th># strips</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Infineon</td>
<td>n-type</td>
<td>80</td>
<td>20</td>
<td>300</td>
<td>50</td>
<td>256</td>
<td>region of disconnected channels</td>
</tr>
<tr>
<td>CNM</td>
<td>p-type</td>
<td>90</td>
<td>25</td>
<td>270</td>
<td>54</td>
<td>254</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 5.1: Major characteristics of the sensors employed for the beam test.

Modules

The two mini 2S-modules were equipped with different sensors, described in Table 5.1. Both have strip pitch and length very close to the final outer Tracker design, and enough strips to bond to the top or bottom channels of two adjacent CBC2s. They mainly differ in manufacturer and in readout polarity: the fixed module being an Infineon n-type, and the one mounted on the moving stage a CNM p-type. At a time when the choice of the sensor polarity for the Tracker had not yet converged on electron readout, testing the CBC2 in both configurations was one of the aims of the tests. Availability of the sensors was also a decisive factor in the choice of the sensors. Unfortunately, the Infineon sensor suffered from a limited region of non-responsive strips in the central region (Fig.5.4), which were removed from the subsequent data analysis. The electrical characterization of the sensors can be found in Ref.[140]. A third module, also with Infineon sensors, was available as backup but was not used.

Modules and interface cards were enclosed in the light-tight, temperature-controlled aluminium boxes of Fig.5.5a, actively cooled to a constant 20°C via an external Arduino-based controller. An interface card inside the casing provides power to the module, as well as SLVS level translation for the CBC2 slow ($I^2C$ data and clock) and fast control signals (Trigger, Fast Reset, $I^2C$ Refresh and Test Pulse). Entry and exit windows, protected by aluminium foil, provide wide beam acceptance even when the module is rotated, and a 34-way connector carrying SLVS signals connects to the DAQ.
Data Acquisition

The beam test was as much a test for the DAQ as for the front-end, and as such data acquisition and control systems (Fig. 5.5b) were designed to resemble as closely as possible the final system, in spite of the fact that several key components were not available at the time. The role of the GBT, for example, was emulated by the CERN Giga Bit Link Interface Board (GLIB) [141]. A front-end GLIB with two FPGA Mezzanine Cards (FMCs), close to the modules and the telescope, emulated the module readout of a 2S module, while a back-end GLIB connected through optical links controlled triggers and clock reference. In particular, a Time to Digital Converter (TDC) implemented in the back-end GLIB conditioned the data with the phase difference between the 40 MHz clock of the DAQ (designed to mimic operation at the LHC) and the DESY beam clock, so that it was possible to discard events with an arrival time incompatible with the acquisition phase of the system.

The software was based on the standard CMS XDAQ framework for data acquisition.

5.1.2 Analysis and Results

In most of the following analysis, the Infineon module is used as a reference for the event selection cuts, while the focus is on data from the CNM module. Apart from its superior assembly with no ill-connected strips, it was the module equipped with the n-in-p sensor which was mounted on the rotational stage, a requirement for the test of $p_T$-cut performance.

The comparator threshold is expressed both in terms of the value set in the $I^2C$ register on the CBC2 (VCTH), and electrons. The conversion between the two, as derived in section 3.2.3 and averaging between the two CBC2s in the CNM module, is:

$$\text{Threshold} \ [e^-] = 380 \ (120 - \text{VCTH} \ [I^2C]) + 6440$$

In electron readout configuration, the output of the CBC2 front-end is a negative pulse, therefore higher VCTH values correspond to a lower threshold (cf. Fig.3.8). The channels were calibrated for a threshold of VCTH=120 ($6440 \ e^-$), which, unless otherwise stated, was the nominal threshold for data taking.
Figure 5.6: Normalized distribution of cluster size from 1 to 10 adjacent strips as a function of the angle of incidence: one and two strip clusters dominate, with three strip clusters becoming significant only at large angles. Larger clusters are not appreciably affected by the angle of incidence, suggesting they are caused by delta electrons.

Noise Estimate

An estimate of the noise of the modules (excluding the aforementioned defective strips) was performed by looking at one million randomly triggered events in the absence of beam. At nominal threshold, the noise rate per strips for all four CBC2s was found to be extremely low at $O(10^{-7})$ [133].

Cluster Width Distribution

The analysis of the cluster width is important to evaluate the design choice of rejecting clusters larger than three strips in the CBC2 CWD logic.

The normalized distribution of individual cluster widths as a function of the angle of incidence is presented in Fig. 5.6, for clusters from one to ten adjacent strips. At low angles, the vast majority are single-strip hits, with an increase in two-strip clusters as the angle of incidence increases. A marked increase in three-strip clusters is noticeable between $22^\circ$ and $32^\circ$. This can be attributed to the tracks spanning several strips: for reference the track displacement within the sensor thickness ($t \ast$...
5.1 CBC2 Beam Test

Figure 5.7: Cluster width as a function angle of incidence for the top CNM sensor (S0): a parametrization of the Digitizer (Section 5.2) applied to data (SimHits) from a Monte Carlo simulation yields results in good agreement with the experimental data, even though it underestimates large clusters (courtesy of Suchandra Dutta).

Figure 5.8: Geometrical displacement of tracks within the thickness of the sensor, in the direction perpendicular to the strip length, due to the angle of incidence of the beam ($t \times \tan(\theta)$, with $t = 270\mu m$ the detector thickness, and $\theta$ the angle of incidence). The data points match the rotation angles used for the measurements of the DUT. The offset is also expressed in terms of strips, for a pitch of $90\mu m$. 
$\tan(\theta)$, with $t = 270 \mu\text{m}$ the detector thickness, and $\theta$ the angle of incidence) is plotted in Fig. 5.8. The offset is also expressed in terms of strips, given the sensor strip pitch of 90 $\mu\text{m}$: for large angles the displacement approaches two adjacent strips, which means a tracks impinging on the centre of a strip will cause a three-strip clusters.

From Fig. 5.6, it is also evident how the distribution of wide clusters ($> 5$) is not affected by the track angle. Although the noise figure quoted in the previous section was performed without beam, and slightly higher noise could be expected otherwise due to higher occupancy and possible common mode effects, the strip noise is too low to account for these large clusters. Such clusters are therefore attributed to secondary delta electrons within the silicon sensor, the propagation of which is broadly independent from that of the primary particles.

Although limited to a single particle species, this study confirms how the upper acceptance limit of three adjacent strips for the cluster width discrimination logic of the CBC2 does not significantly affect the stub detection efficiency, as more than 99% of hits fall in this category for even very large angles of incidence.

As described later in section 5.2, the test beam provided the first realistic data on the performance of the CBC2, and the data on cluster width distribution in particular were used to tune the parameters in the CMS Software (CMSSW) simulation. Figure 5.7 demonstrates how the results of Monte Carlo simulation using the cluster parametrisation closely match the experimental data, even though they slightly underestimate the fraction of large clusters.

**Hit Efficiency**

High hit efficiency is a necessary prerequisite for the following studies of stub-finding efficiency.

To allow the selection of candidate track events and to account for the rotation of the CNM stage, the analysis started with the calculation of the beam divergence for every angle of rotation of the CNM module. Clean, high-$p_T$ events were selected
Figure 5.9: Beam divergence for three angles of incidence. The RMS value of the gaussian fit corresponds to the beam divergence used in the selection cut for the CNM efficiency.

by requiring only one cluster in all four sensor planes, and limiting the distance between the hits in the two CNM sensors to less than four strips. The difference between the centroid of the stubs in the INF and in the CNM modules was then histogrammed (Fig. 5.9): the standard deviation of the resulting Gaussian profile is taken as the beam divergence, and used in the selection cuts for the calculation of the sensor efficiency.

Since the telescope data were not available, a subset of clean events was used to evaluate the hit efficiency, which was studied for different values of threshold setting. After masking the noisy signals from the INF module, the candidate events for the estimation of the sensor hit efficiency were selected by requiring only one cluster per sensor in three of the four sensors, and looking at the presence of a hit in the remaining plane when:

- the offset between the clusters in the other module is less than four strips (high-$p_T$ stub), and
- the offset between the centroid of this stub and the hit on the other sensor of the module under consideration falls within the mean of the fit ±1σ as calculated above for the beam divergence (Fig.5.9).
Figure 5.10: Efficiency of one of the CNM sensors as a function of threshold (VCTH) for different angles of incidence. The data are fitted with error functions (Erf). The horizontal error bars represent the bin width and not true errors.

The sensor hit efficiency is defined as the ratio:

\[ \text{Hit Efficiency} \quad \epsilon = \frac{\text{HITS}}{\text{TRACKS}} \]

where the denominator is the total number of selected clean tracks, and the numerator the events where the expected hit was observed.

The hit efficiency for one of the CNM sensors is plotted in Fig. 5.10 as a function of the comparator threshold VCTH, for all the angles of incidence measured during the beam test.

A fit using the integral of the Landau distribution expected for the energy absorption was performed in Ref.[133], including second order effects in the fit to match the slope and the tail of the Landau. The non-idealities considered included non-monoenergetic beam, the variation in energy deposition per strip due to charge sharing, and uncertainty in the effective threshold across all channels.

For the purpose of evaluating the efficiency setting, however, a simpler fit is sufficient, where the Landau is approximated by a Gaussian distribution. The data are hence
<table>
<thead>
<tr>
<th>Angle [degrees]</th>
<th>Efficiency $\epsilon$</th>
<th>Inefficiency $1 - \epsilon$</th>
<th>Uncertainty</th>
</tr>
</thead>
<tbody>
<tr>
<td>$2^\circ$</td>
<td>1.00E+00</td>
<td>6.86E-05</td>
<td>-2.61E-05 + 1.95E-05</td>
</tr>
<tr>
<td>$12^\circ$</td>
<td>9.97E-01</td>
<td>2.54E-03</td>
<td>-3.25E-04 + 2.90E-04</td>
</tr>
<tr>
<td>$15^\circ$</td>
<td>9.97E-01</td>
<td>3.46E-03</td>
<td>-1.87E-04 + 1.77E-04</td>
</tr>
<tr>
<td>$17^\circ$</td>
<td>9.94E-01</td>
<td>6.10E-03</td>
<td>-2.60E-04 + 2.50E-04</td>
</tr>
<tr>
<td>$22^\circ$</td>
<td>9.81E-01</td>
<td>1.86E-02</td>
<td>-3.32E-04 + 3.26E-04</td>
</tr>
<tr>
<td>$32^\circ$</td>
<td>9.91E-01</td>
<td>9.15E-03</td>
<td>-3.68E-04 + 3.55E-04</td>
</tr>
</tbody>
</table>

Table 5.2: Hit efficiency for the CNM sensor, for a threshold of 6440 $e^-$ (VCTH=120) and for several angles of incidence.

fitted with an error function (erf) of the form:

$$\text{fit}(x) = \frac{1}{2} \text{erf}\left(\frac{x - a}{b} + 0.5\right) + c$$

with $a$ the centre of the fit, $b$ the half width, $c$ the maximum efficiency of the fit function, and the error function erf is:

$$\text{erf}(x) = \frac{2}{\sqrt{\pi}} \int_{-\infty}^{x} e^{-t^2} dt$$

The hit efficiencies for a threshold setting of 6440 $e^-$ and for different angles of incidence are reported in Table 5.2. The results confirm high efficiency ($> 99\%$) for the chosen threshold in all cases.

**Stub Bend Distribution**

The CBC2 does not incorporate synchronous stub readout, except for the fast-OR test output which aggregates all 127 correlation channels. Since this output was not saved during the beam test, a study of the performance of the coincidence logic must rely on reconstructed data. The coincidence logic had already been comprehensively tested on the bench as described in section 3.3. The beam test however provided the first realistic hit data patterns to evaluate the efficacy of the $p_T$ cut concept.

A first analysis of the stub distribution was performed by correlating hits in the CNM module, after selecting only clean events with only one hit per sensor. The stub bend, defined as the offset between the top and bottom clusters in the DUT sensors, is
plotted against the angle of rotation of the CNM module in Fig.5.11a. As expected, the mean of the distribution shifts with the angle of rotation, up to +20 strips for an angle of incidence of 32°. It should be noted that on the CBC2, despite the window of coincidence being limited to ±8 strips, the Φ-shift can be accommodated by setting the offset parameter for the coincidence logic. The maximum shift for the target $p_T$ is also controlled by the separation of the two sensors, and the offset range of ±8 strips is therefore sufficient in the Tracker.

The mean and standard deviation of the stub distributions of Fig.5.11a have been used to extrapolate the non-zero initial angle of rotation of the DUT module. A correction of +2° to the recorded angle was applied beforehand, based on early interpretation of data, to account for a non-zero initial angle. Assuming perfect alignment between the strips of the two sensors, the data were then fitted with:

$$\text{stub bend} = p d \tan(\Theta_{\text{recorded}} + \Theta_0)$$

with $\Theta_0 = 2° + \Theta_{\text{FIT}}$

with $p = 90\mu m$ the strip pitch, and $d = 2.5$mm the distance between the sensors (Fig.5.11b). The small deviation of the fit value of $\Theta_{\text{FIT}} = -0.07°$ confirmed the validity of the initial assumption of $\Theta_0 = +2°$.

$p_T$-Cut Efficiency

Having confirmed a hit efficiency in excess of 99%, it is possible to investigate with confidence the performance of the $p_T$-selection of the CBC2.

To avoid biasing the selection of events in the DUT, the selection cut was performed only on the fixed reference module, by requiring one cluster on both top and bottom sensors, clusters smaller than four adjacent strips, and a difference in the centroid of the two clusters of less than two strips. These cuts are intended to reject particles scattered upstream of the DUT, traversing the sensor at oblique angles and therefore underestimating the efficiency. The last criterion for example constrains the angle of incidence of the beam to less than 70 mrad, well above the measured beam divergence of 1 mrad but enough to effectively reject this background.
Figure 5.11: Stub bend (offset between the top and bottom sensor clusters) for the CNM module, as a function of the angle of rotation.

The stub-finding efficiency as a function of the angle of incidence for the CNM module, for a coincidence window of ±7 strips (VCTH=120), is presented in Fig.5.12a. The efficiency drops sharply from 99.5% for angles below 10°, to approximately 0 for angles above 12°. This is the desired outcome, as it demonstrates how straight tracks are detected with very high efficiency, while the system rejects tracks impinging with an angle greater than \( \sim 12° \), for which the stub bend becomes greater than the correlation window width setting of 7 strips (cf. Fig.5.11b).

To visualize the effect on the \( p_T \) selection, this result can be transposed to the \( p_T \)
(a) *Left:* stub-finding efficiency as a function of the angle of incidence. *Right:* equivalent \( p_T \) cut efficiency for a module at \( R=75 \) cm, as derived from the graph on the left (from [142]).

(b) Simulation of stub efficiency for muons (c) Simulation of stub efficiency for muons, in all barrel layers (from [13]). The lower efficiency for electrons and pions is due to the interaction with the pixel detector.

**Figure 5.12:** Stub-finding efficiency and derived \( p_T \) cut efficiency from test beam data (*top*), and from simulation in all barrel layers and for different particles (*bottom*).
5.2 Digitizer Studies

Cut of a layer in the Tracker via the relation (Eq.1.1):

\[ p_T = \frac{0.3 RB}{2 \sin \phi} \approx \frac{0.3 RB}{2 \phi} \]

For a module at R=75 cm in the barrel, and a magnetic field of 4 T, this corresponds to a \( p_T \) threshold of 2.17 GeV, with a resolution of \( \sigma_{pT-cut} = 0.1 \) GeV. Such modules would therefore be able to suppress most low-\( p_T \) tracks and provide the expected trigger primitives for the required data reduction in HL-LHC conditions.

The data agree well with the results of simulations, presented in Fig. 5.12b, where the expected reduced efficiency for electrons due to scattering in the pixel detector is also observed (Fig.5.12c).

5.2 Digitizer Studies

As part of this thesis, the readout performance of the CBC2 was integrated into the algorithm responsible for the digitization of the Tracker hits within the CMSSW simulation environment [143]. The Digitizer class loops over the SimHits and, based on the entry and exit points and on the energy loss of a track, calculates the position of the detectors hits (Digis) [144]. The SiPixelDigitizer, as the name suggest, was initially borrowed from the equivalent class in the pixel detector (already a digital readout system) and adapted to reflect the strip pitch, module geometry and stub-finding algorithm of the 2S modules.

The Digitizer algorithm proceeds as follows (Fig.5.13) [144]:

- the track segments within the silicon sensor volume are subdivided into 100 segments, and each is allocated a corresponding fraction of the total energy loss in the sensor calculated by the GEANT4 simulator;

- the point charge from each segment is drifted to the detector strips under the influence of the magnetic field, then diffused along the strips with a Gaussian smearing;
Figure 5.13: CMSSW Digitizer drift principle to create hits (Digis) from the output of the GEANT4 simulation (SimHits) (from [144]).

- all the charges falling within a strip boundary are summed to obtain the strip signal;
- detector and readout noise (RMS) are applied before determining which pixels are above threshold;
- finally a threshold (smeared with a Gaussian, the standard deviation of which is one of the noise parameters) is applied before the digitization of the signal.

To adapt the pixel Digitizer to the strip CBC2 module, the following modifications were required:

- the geometry of the sensor was updated to match 90$\mu$m, 200$\mu$m thick and 5 cm long strips;
- the charge drift parameter was modified to account for hole readout, before reverting to electrons once the choice of sensor polarity converged onto n-on-p;
- the electronic noise was set to the CBC2 noise figure (1000 $e^-$);
- a crosstalk contribution between adjacent strips was added, which from simulations was estimated to be 8% for the nearest neighbours;
- hard-coded values, such as sensor thickness, strip pitch and electronic noise were parametrized and moved into configuration files;
- the digitization step was changed to binary.
The threshold and inter-strip coupling parameters were tuned by comparing the results of the simulations to the test beam data, in particular the cluster width distributions as a function of the angle of incidence. As shown in Fig.5.7, the agreement is within 5%, although more particle species at different energies should be considered to validate the result.

5.2.1 Conclusions

The results of the beam test described in this chapter prove the correct functioning of the stub-finding logic of the CBC2, with the chip performing as desired without any unexpected behaviour and under realistic conditions. The $p_T$-cut efficiency study of section 5.1.2 in particular demonstrated that the 2S module instrumented with the CBC2 can discriminate between low and high $p_T$ tracks, confirming the validity of the stub reconstruction concept at the heart of the future CMS Tracker design.

Having demonstrated the functionality of the CBC2 and the effectiveness of the design and of the settings available on the CBC2, it was possible to proceed confidently to inserting the remaining features of the chip, as discussed in the following chapter.

The testing of the CBC2 and of the 2S module continued with a second test beam, performed in June 2015 at the CERN SPS, to verify the performance after irradiation of the sensors and of the CBC2s to the expected HL-LHC doses [145]. An additional one in November 2015 to test a full-size 2S prototype module, with a focus on the performance at the centre of the module where the two rows of strips meet, was also carried out successfully.
Chapter 6

Design Modifications for the Final CBC

6.1 CBC3

Following the successful development of the CBC2, design started on the CBC3, which is intended as the final pre-production prototype (although provision for one additional iteration exists). As such, the CBC3 needs to expand the functionality of the CBC2 by including features such as a synchronous stub readout. At the same time, the ASIC specifications reflect the updated requirements for the Tracker readout, capturing many aspects not yet crystallized at the time of the CBC2, or later revised.

While including a general overview of the CBC3, this chapter focuses on the main contributions as part of this thesis, namely the stub-finding logic and the memory design.

6.1.1 CBC3 Overview

The block diagram of the CBC3 is presented in Fig.6.1. The architecture is that of the CBC2: 254 channels with analogue front-end and binary conversion, followed by digital signal processing blocks for stub formation, and by a memory pipeline. The main differences can be summarized as:
Figure 6.1: Block diagram of the CBC3 ASIC (from [146]).
6.1 CBC3

- updated front-end;
- updated coincidence logic, including additional functions and a new inter-chip signalling scheme;
- synchronous stub readout, with additional clock domain (320 MHz) and relevant synchronization interface;
- new memory pipeline with extended latency and SRAM design for improved TID performance;
- extended control interface.

Each of these features will be discussed in more detail in the following sections.

6.1.2 Front-End

The final choice of n-in-p silicon as the sensor material for the 2S module (electron readout) removes the need for the dual-polarity readout described in section 2.3.2. A similar simplification of the front-end derives from the choice of AC-coupled sensors, which makes the CBC2’s ability to cope with leakage current superfluous.

The resulting relaxed constraints on the operating range of the front-end, and the resources freed by removing one polarity and therefore simplifying the feedback network of both pre and post-amplifier, can be used towards optimizing the design in terms of powering, pulse shaping and overload recovery. The target is to shorten the peaking time slightly to less than 20 ns, to obtain a return to baseline within 50 ns, while maintaining the current ENC figure of 1000 $e^-$. 

The integration of the 2S module with other ICs based on a 65 nm CMOS process makes it desirable to have a common powering scheme. The CIC and GBT ASICs will in fact tolerate a maximum supply voltage of 1.2 V, including a safety margin of +10%, to be provided by two cascaded DC-DC converter chips on the module (11 V → 2.5 V → 1.2 V). To comply with this lower voltage tolerance, the CBC3
should be able to operate with an input voltage as low as 1 V after the voltage drop in the on-chip regulator.

The design and optimization of the front-end is being performed by Mark Raymond at Imperial College.

### 6.1.3 CBC3 Logic

#### Channel Mask and Hit-Detect Logic

The $I^2C$-programmable channel mask designed to prevent noisy or faulty channels from generating fake stubs (Section 2.3.3) will now also prevent masked channels from generating stored hits in the pipeline. The block-synchronous readout of the 2S module after zero suppression of the hits in the CIC [147] would in fact rapidly become inefficient otherwise.

A similar issue is caused by HIPs, which saturate the front-end and can give rise to several out-of-time spurious hits in the pipeline. To prevent this the Hit-Detect Logic will be modified to include a HIP-suppress mode of operation, whereby the 40 MHz-sampled output of the channel comparator is forced back to zero if the output stays high for longer than a number of consecutive clock cycles, to be programmed via $I^2C$. The inclusion of a Delay-Locked Loop (DLL) will also allows for a fixed 25 ns pulse mode.

#### Layer Swapping Logic

In the Tracker assembly, modules of the same layer overlap slightly to achieve complete coverage. To keep the clearance to a minimum, adjacent modules are flipped with respect to each other so that the surface-mounted components are outward-facing. Because of the asymmetric nature of the CBC2 stub-finding logic, this causes a problem if the seed layer is hard-wired to the bottom sensor of the 2S module. The offset for the coincidence window, as well as the stub bend information (discussed later in this chapter), would then have to be adjusted for the module orientation,
Figure 6.2: Rationale behind the CBC3 Layer Swapping Logic: in the assembly of the Tracker adjacent modules are flipped (as in the TOB stave pictured here), so that the seed layers alternate between inner and outer layer. The Layer Swapping Logic in the CBC3 makes sure that the stub seeding is consistently associated with the inner sensor.

as presented in Fig.6.2. A degradation in performance due to extra scattering and conversion could also be expected from seeding from the outer layer.

To avoid this limitation and have uniform performance across the Tracker, the CBC3 incorporates a Layer Swapping Logic block, programmable via $I^2C$. The logic works as a simple multiplexer at the output of the Channel Mask, effectively allowing the swapping of the upper and lower sensor channels at the input of the stub-finding logic. Inter-chip communication is not affected since all the CBCs on the same module share the same setting.

**Stub-Finding Logic**

The most significant change affecting the stub-finding logic in the CBC3 is the ability to output not only the address but also the bending of stubs, defined as the offset between the cluster centroids of the seed and the correlation layers, after removal of the programmable Φ-shift offset. Referred to as *stub bend information*, this value is output synchronously with the stub address to enhance track recognition for the L1 trigger, since it provides information about the charge of particles, as well as a rough transverse momentum estimate.

Another improvement is the finer resolution of half-strip pitch for both stub address and bend.
Cluster Width Discrimination Logic  As mentioned in section 2.3.3, in the case of even-strip clusters, the CBC2 assigns the centre of a cluster to the strip with the lower address, causing a loss of resolution.

The Cluster Width Discrimination logic in the CBC3 has one additional output to be able to identify such clusters (Fig.6.3), providing effectively half-strip resolution for the following correlation stage.

The maximum programmable cluster width, above which clusters from either sensor are rejected for the purpose of stub formation, was also increased from three to four adjacent strips.

Offset Correction, Correlation and Stub Bend Logic  The Offset Correction, Correlation and Stub Bend Logic (OCCB), in addition to the OCC functions of correlating clusters from top and bottom sensor and correcting for the geometrical $\Phi$-shift offset, also extracts the stub bend information.

Both the stub address and bend are provided with half-strip resolution. The maximum correlation window width, symmetrical around the central strip, is now $\pm 7$ strips ($\pm 14$ half-strips) and is programmed globally via a 4-bit $I^2C$ register. The maximum offset is up to $\pm 3$ strips ($\pm 6$ half strips), and is independently programmable across four groups of 32 channels per chip (32 groups per 2S module).
To be able to cover the maximum window width and offset, the OCCB block must look at one seed strip from the bottom sensor, plus $\pm 7 \pm 3$ strips around the strip immediately above it. Accounting for the half strip resolution, this implies a total of 44 inputs for every OCCB block: a significant increase from the 24 inputs of the CBC2 OCC, which required a rethinking of the inter-chip signalling scheme, as discussed later in section 6.1.3.

It should be reiterated how, just as in the CBC2, when more than one valid cluster is found in the correlation window, only the one closest to the centre of the window (with the least bend, and therefore associated with a larger $p_T$) is selected for the stub formation.

The offset correction is applied before the stub bend extraction, so 5 bits are used for the stub bend to cover a range of $\pm 14$ half strips. Simulation studies on the efficacy of the stub bend information for L1 trigger formation are still under way (e.g. [148, 149]); however a resolution of 4 bits is considered sufficient. Since a 4-bit scheme does not provide enough granularity to cover the full correlation window with uniform resolution, a bend code mapping scheme needs to be introduced.

Instead of a hard-wired mapping, the CBC3 includes an $I^2C$ programmable look-up table: by retaining full 5-bit resolution at the output of the OCCB and providing a programmable mapping scheme, the CBC3 allows full flexibility so that the optimal mapping can be tuned during operation. Fig.6.4 represents the look-up table and one possible mapping scheme [150, 151].

**Inter-Chip Signals**

The stub-finding logic layout (Fig.6.5a) closely resembles its earlier version in the CBC2: one logic block for every couple of seed and correlation channels, to guarantee identical performance across all channels. Besides the additional area required by the added functionality, the main difference is the number of inputs to each logic block.
Figure 6.4: CBC3 programmable look-up table for converting the 5-bit resolution stub bend from the CBC3 logic to the 4-bit output data (from [146]).
(a) Detail of the CBC3 coincidence logic. The notation follows that of Fig.2.16 to highlight the similarities with the CBC2: (A) CWD for seeding strip; (B) CWD for outer strip; (C) Offset correction, correlation and stub bend logic; (E) Hit-Detect output lines to/from adjacent channels; (F) CWD output lines to/from adjacent channels.

(b) Detail of the layout of the CBC3 logic. The additional CWD blocks required to reduce the number of inter-chip signals have been accommodated in the existing space at the top and bottom extremities of the chip, thereby maintaining the channel pitch mapping of the CBC2 and avoiding expanding the chip dimensions.

Figure 6.5: CBC3 coincidence logic and inter-chip signal blocks.
Adopting the CBC2 inter-chip communication scheme described in section 2.3.3, where the outputs of the CWDs are propagated between adjacent chips, would in fact result in an excessive number of signals to be transmitted (52 at each chip boundary).

A better alternative is to propagate the output of the Layer Swapping Logic, and reconstruct the cluster selection by adding additional CWD blocks at each boundary, dedicated to processing the signals coming from adjacent ASICs.

By doing so, the only signals required are:

- 2 channels from the seed layer, from channels with lower strip address, for the seed CWD;
- 3 channels from the seed layer, from channels with higher strip address, for the seed CWD;
- 12 channels from the correlation layer, from channels with lower strip address, for the OCCB (2 of which are also for the correlation CWD);
- 13 channels from the correlation layer, from channels with higher strip address, for the OCCB (3 of which are also for the correlation CWD);

for a total of 30 signals at each chip boundary: the same number as the CBC2, and a significant reduction from the 52 pads of the original scheme described above.

The downside is the requirement for an additional ten CWD blocks per chip, per boundary, to reconstruct the clusters in the upper layer. The additional power consumption associated with this is however negligible, especially when compared to the power saving deriving from the fewer pad drivers. Also, as visible in Fig.6.5, careful layout ensured that the additional logic did not increase the chip area.

As for the CBC2, the inter-chip pads are designed to float to zero when not connected externally, for example in the case of the first and last of the eight CBCs on a 2S hybrid.
6.1 CBC3

Figure 6.6: CBC3 output data format (courtesy of Mark Raymond). Left: synchronous trigger data, with each column representing one SLVS line. ($S =$ stub address (8b); $B =$ stub bend information (4b); $SoF =$ stub overflow; $OR254 =$ any unmasked channel above threshold; $Error =$ latency OR FIFO overflow error; $Sync =$ for synchronization with the CIC; $L1 =$ L1 trigger data, unsparsified.) Right: L1 trigger data frame.

6.1.4 Stub Gathering Logic and Readout

The Stub Gathering Logic of the CBC3 takes the stub addresses (8-bit) and bend information (5-bit) at the output of the stub-finding logic, and propagates them with one bunch crossing latency to the bend lookup table, and ultimately to the Data Packet Assembly and Transmission Logic.

The available bandwidth for the L1 trigger allows only three stubs to be output synchronously by every CBC to the Concentrator IC. Should more be present, the ones associated with the lowest strip addresses will be selected, together with an extra bit to flag stub overflow.

With full stub readout capabilities, the CBC3 has no need for the $OR127$ test signal, which in the CBC2 indicated whether at least one stub was detected for the whole chip. The $OR254$ test output however will be retained as a useful low-latency indication of signal activity, and will be included in the output data packet.

The output of the CBC3 consists of six differential SLVS lines operating at 320 Mbps. Of these, one is dedicated solely to L1 data, while the remaining five transmit trigger data in the form of stubs, error and synchronization bits. Fig.6.6 illustrates the proposed data packets; more details can be found in [146, 152].
6.1.5 Control Interface

The $I^2C$ serial interface of the CBC3 will operate at 1 MHz. As concluded in section 4.9.2, to improve the SEU immunity of the $I^2C$, the current triple majority-voting cells will be replaced by the more effective Whitaker-style registers.

Additionally, a new fast control interface via one SLVS input will be able to issue eight bit commands synchronously. These will be shared with other chips on the module; CBC-specific commands include Fast Reset, Trigger, Test Pulse Trigger and L1 Count Reset.

6.1.6 L1 Trigger Memory

One of the major changes to the specifications of the CBC3 was the doubling of the maximum L1 trigger latency from 6.4 $\mu$s to 12.8 $\mu$s. While desirable for trigger processing, a longer latency directly translates into additional front-end memory. The 12.8 $\mu$s limit was mainly dictated by the readout chip on the PS module, which employs in-pixel memory and has stricter power and area constraints. In the case of the CBC, an extension of the pipeline to 512-deep SRAM cells based on the CBC2 design would have increased the width of the chip by approximately 1.2 mm. Because of the dense routing on the 2S side hybrids, there would be no margin to accommodate such an increase without extending the hybrid itself.

More importantly, although the conclusion from the TID irradiation of the CBC2 presented in Chapter 4 was that the observed failure and excess current should not happen under operating conditions, it was decided to improve the radiation hardness of the SRAM to rule out any possible failure scenarios.

A new SRAM cell was proposed which focuses on radiation hardness, and at the same time is optimized for the layout of the longer CBC3 pipeline (Fig.6.7). The read and write nMOS access transistors have been replaced by pMOS devices, which do not experience the same level of degradation and do not suffer from the parasitic lateral leakage current effect presented in section 4.2.1. The nMOS pull-down transistors
of the SRAM cross-coupled inverters have also been replaced by enclosed layout transistors.

These two measures combined should completely remove the main leakage current path under irradiation. The distance between adjacent devices was also maintained at the same level as the CBC2 SRAM, so that field oxide parasitic paths should yield negligible currents, as calculated in section 4.8.1.

As evident from Fig.6.7, the new design achieves a significant reduction in area. To exploit the area available on chip in the vertical direction, and given that the length of the pipeline increases the width of the ASIC, every transistor in the SRAM cell was stacked in line, to achieve minimum cell width. While the increase in height is limited to a modest 30% (in the CBC2 about 50% of the area was used for inactive fill structures), the overall cell area is reduced by 44% and, crucially, the width scales by more than 55%.

The new design is also optimized for manufacturing yield: in accordance with a
well-established industry trend for so-called thin cell SRAM design, the number of notches and corners in the active pattern was minimized to aid lithography and therefore reduce mismatch [153].

Although difficult to speculate on without experimental data, the new SRAM might also have a smaller SEU cross-section. Even if the cell pitch is considerably reduced, the two cell storage nodes are now separated by more than 10 μm. And while a reduced pitch increases the probability of multi-bit SEUs, the fact that the ionization charge will be collected by more than one cell should increase their individual SEU threshold.

### 6.1.7 L1 Trigger Memory Buffer

As well as longer latency, the average sustainable L1 trigger rate for the Phase II upgrade has been increased substantially from 300 kHz to 1 MHz. Just as the latency affects the depth of the pipeline, a higher trigger rate has implications for the buffer memory that follows it. The role of this block in fact is to store data awaiting readout while previous triggers are being transmitted off-chip. If the trigger rate is too high, the buffer will progressively fill up, until additional triggers have to be vetoed to avoid data losses.

Because of the stochastic nature of the trigger, losses are inevitable; however a sufficiently deep buffer can limit these to negligible values. As visible in Fig.6.8, the CBC2’s existing 32-deep buffer is sufficient since losses are limited to < $10^{-5}$ at 1 MHz trigger rate, and only a marginal gain would result from doubling the size to 64 samples.

### 6.2 Summary

The design of the last prototype of the CMS Binary Chip was presented. The CBC3 incorporates all the functions required by the final production ASIC, including optimized front-end, extended memory for 12.8 μs L1 trigger latency, and full stub
Figure 6.8: Percentage of triggers rejected to avoid buffer overflow vs. trigger rates. A buffer depth of 32 is sufficient to maintain losses to less than $<10^{-5}$ for the proposed maximum L1 trigger rate of 1 MHz (courtesy of Mark Raymond).

readout. The stub-finding logic extends the functionality of the CBC2 to include stub bend extraction, half-strip stub resolution and provision for layer swapping. The inter-chip communication scheme was also revisited to avoid the increase in the number of extra pads otherwise dictated by the introduction of half strip resolution. Finally, based on the results of the TID irradiation, a new rad-hard SRAM cell was proposed to eliminate the source of leakage current and to accommodate the longer latency without growing the chip dimensions.
Conclusions

CMS faces the enormous task to rebuild the central silicon tracking detector to cope with the challenges of operation and survival at the HL-LHC. The electronics to read out the sensors will take years to develop fully. This process began some years ago for the outer strip modules and novel ideas have been implemented to allow the data to be used in the L1 trigger.

The CMS Binary Chip is the first ASIC to be developed for the new detector and has pioneered several developments, such as the design of a flexible front-end hybrid, the design of a data Concentrator IC, and the prototyping of a smaller but fully-instrumented version of the new strip module.

The CBC2, designed as part of this work, is the first chip to include the capability to correlate hits between pairs of sensors to identify high-$p_T$ track candidates. The logic includes a number of programmable features designed to tune the desired $p_T$ threshold. The results of extensive electrical tests confirm that the ASIC performs extremely well. The radiation hardness of the CBC2 was also investigated with a series of x-ray tests, and found to be satisfactory, with recommendations to be adopted in the next iteration.

This thesis has demonstrated how close the design is now to the final requirements, and included the first experimental demonstration of the functionality of suitable track-trigger modules in beam test, a concept central to the High-Luminosity upgrade strategy of CMS.

An overview of the final pre-production prototype, the CBC3, concluded the exposition.
Appendix A

TID Dosimetry

A.0.1 Dose Calculation

As described in Section 4.2, radiation mainly affects the gate and field oxide of a device. To have an accurate estimate of the dose delivered it is essential to be able to extrapolate the reading of a dosimeter to the dose deposited into the sensitive layer of the DUT. The usual assumption [103] is to consider the incident radiation to be monochromatic, thereby requiring the calculation of only one absorption coefficient. However given the scarcity of results in the literature for TID testing performed with a molybdenum tube, and the different characteristics of molybdenum and tungsten spectra, to be confident in the calculation of the dose for the CBC2 test, the x-ray spectrum was measured and the absorption coefficients calculated for many energy bins.

The dose calculation proceeded as follows (Fig. A.1):

- A dosimetry diode is placed in the same position as the DUT and fully depleted. The power deposited in the diode under different machine settings is calculated from the current measured in the diode, minus the dark current. After this initial acquisition, the diode is replaced by the CBC2 until the end of the irradiation.
Figure A.1: Calculation of the dose delivered to the sensitive volume of the CBC2 during TID irradiation.  

a) The spectrum measured in the Vortex detector is corrected for absorption in the silicon sensor and in the beryllium window to obtain the incident spectrum.  

b) The spectral power is obtained from the absorption and the resulting current in the dosimetry diode.  

c) Dose delivered to the CBC2 sensitive volume (in red).  

d) The absorption in the outer layers of the CBC2, which amounts to a significant fraction of the incident radiation, is accounted for.
The spectral distribution of the incident radiation is measured with a Vortex silicon-drift detector [106] (Fig. A.2). Even for minimum current setting in the x-ray tube, the beam intensity was too high for the Vortex, so to avoid saturation this had to be placed slightly off-axis of the beam, but at the same distance and, importantly, at the same tube bias of 50 kV. The measured spectrum is binned in 4096 bins of 13 eV, from 0 eV to 53235 eV (Fig. A.3).

Absorption and transmission coefficients are obtained from the NIST Standard Reference Database [154] and the CXRO Filter Transmission Coefficient database [155] for silicon, silicon oxide and beryllium. The database generates absorption and transmission coefficients for elements and compounds for coherent and incoherent scattering, photoelectric absorption, and pair production in the fields of the atomic nucleus and electrons. These subsets of values (typically 200 points) are interpolated to obtain a value for each of the 4096 photon energy bins.

The spectrum read out by the Vortex is corrected to account for the absorption of its 350µm-thick silicon sensor (nominal thickness). With reference to Fig.A.1, if $I_D$ is the intensity in the Vortex detector for a particular spectral component, the intensity of mono-energetic radiation at depth $x$ in an absorber is:

$$I(x) = I_0 e^{-ux}$$

where $I_0$ is the intensity of the radiation at the surface ($t=0$) and $u$ is the total linear attenuation coefficient for the material of interest (in this case Si). The linear attenuation coefficient describes the fraction of the beam which is not transmitted per unit thickness of the absorber, and is derived from the mass attenuation coefficient found in the literature simply by multiplying the latter by the nominal density of the material (2.32 g/cm³ for Si, 2.65 g/cm³ for SiO₂) (Fig. A.4a). A different coefficient is used for every photon energy bin of 13 eV, so that the assumption of mono-energetic components is satisfied to good approximation. To extract the incident intensity $I_0$ from the intensity $I_D$ measured in the Vortex sensor:

$$I_0 = I(x)e^{ux} \quad \text{and} \quad I_D = I_0 - I(t) = I_0 - I_0e^{-ut}$$
with \( t \) the thickness of the Vortex detector.

The intensity incident on the Vortex silicon sensor is therefore:

\[
I_0 = \frac{I_D}{1 - e^{-ut}}
\]

- The resulting spectral components are further corrected to account for absorption through the 25 \( \mu \text{m} \)-thick beryllium window, to obtain the spectrum incident on the detector, which is the same spectrum incident on the diode and DUT. In this case, instead of the absorption coefficient, the transmission coefficients \( T_{x_{Be}} \) for thin layers available from [155] were used. The intensity incident on the Vortex beryllium entry window is:

\[
I_1 = \frac{I_0}{T_{x_{Be}}}
\]

As shown in Fig. A.3, the thin window of beryllium, with its low atomic mass and low density, does not significantly affect the incident spectrum, except for the low energy components, which would anyway be absorbed in the first few microns of the DUT.

- A new set of attenuation coefficients \( u_{t_{\text{diode}}} \) is then used to obtain the spectral component absorbed in the 263 \( \mu \text{m} \)-thick silicon dosimetry diode:

\[
I_{\text{diode}} = I_1(1 - e^{-u_{t_{\text{diode}}}})
\]

Having extracted the relation between the radiation stopping in the diode and the incident flux, and by calculating the energy deposited in the diode from the value of the induced current, it is possible to obtain the power of the spectrum and therefore the absolute values for the spectral components of the incident radiation.

- The topmost layer of the chip is a stack of a polyimide protective layer and the Back-End Of Line (BEOL) part of the IC fabrication process, which includes metal wiring layers, contacts and insulating dielectric. The composition and the nominal, minimum and maximum thicknesses of the layers are provided in the design manual, so that is possible to obtain the transmission coefficient in
Figure A.2: Acquired spectrum. The main emission lines of molybdenum are clearly visible; additional lines have also been marked.

Figure A.3: X-ray spectra from the molybdenum source, as corrected for the dosimetry analysis.
A TID Dosimetry

Figure A.4: Left: Linear attenuation coefficients for Si and SiO2 (data from [154]). Right: Difference in dose rate when the dose is calculated for the whole chip volume, as opposed to a thin layer below the BEOL surface.

a way analogous to the case of the beryllium thin window. Given the number of interconnecting metal layers in the 130 nm CMOS process used (eight) and their complex, non-uniform pattern (Fig.A.6a), it is not possible to obtain an accurate value applicable to the entire chip surface. Multiple layers of the same material were combined into a single layer of equivalent total thickness, with the assumption of 50% metal coverage of the chip.

• The combined total transmission is plotted in A.6b. It results in significant absorption, more pronounced for the lower part of the spectrum but still substantial at the Kα emission line of the molybdenum tube.

• The calculation of the contribution of the topmost layer of the DUT is usually neglected and considered unnecessary for most practical purposes [103]. However in this case it was found to account for almost a third of the dose delivered to the DUT (Fig. A.4b). This is due to the thick copper redistribution layers designed to allow low-resistivity power routing and inductors in the top layer of the RF CMOS process used for CBC2 fabrication; their relatively high atomic number and density, and substantial thickness cause non-negligible absorption of the incident beam.

• At this point the spectral components transmitted through the superficial layer of the CBC2 can be considered as the incident spectrum on the gate and field
oxide. The intensity stopping in the DUT gate and field oxide is:

\[ I_{DUT\text{gate}} = I_1 T x(1 - e^{-\mu t_{DUT\text{oxide}}}) \]

where \( t_{DUT\text{oxide}} \) is the thickness of the volume considered for dose calculation. Since most of the low-energy photons stop in the \( \sim 30 \mu m \) above this layer, the exact value of the thickness considered here does not affect the resulting dose significantly, which confirms that once the absorption of the soft part of the spectrum in the outermost layer of the ASIC is accounted for, the energy can be thought to be deposited uniformly in the sensitive volume. In calculating the dose for the CBC2, a thickness of 5 \( \mu m \) was chosen for this volume.

### A.0.2 Dosimetry Diode

Silicon diodes are commonly used to estimate the radiation dose in the DUT [156]. From the current in the diode it is possible to extract the energy deposited in the silicon layer and therefore the incident photon flux. Thin dosimetry diodes are preferred for this purpose, since the energy can be considered as deposited uniformly across the thickness of the device, which simplifies the dose calculation. For silicon, the depth at which the intensity of the beam is reduced to 1/e of its surface value, the *attenuation length*, is 699.5 \( \mu m \) for photons in the molybdenum \( K_\alpha \) principal emission line (17.479 keV) [155]. This value, however, drops sharply for lower energy photons, which represent a significant fraction of the spectral power emitted by the molybdenum tube; for 10 keV x-rays for example the attenuation length is just 134.3 \( \mu m \). The diode used for this irradiation was 263 \( \mu m \pm 10 \mu m \) thick, with an active area of 5 \( \times \) 5 mm\(^2\). The same diode had been used in the TID testing of the APV25 chip, and to account for attenuation of the x-ray beam in the non-negligible thickness of the diode a correction factor was calculated with a simulation program [105]. However this was obtained for a tungsten (W) x-ray tube, and so could not be compared directly with this analysis.

The beam profile of x-ray testers is usually asymmetric [105, 157], so the position of the DUT in the beam becomes very important to ensure uniform irradiation. Excessive dose variation across the chip area would in fact affect not only the dose delivered, but also the behaviour of distributed circuits which rely on accurate matching,
Figure A.5: Profile of the current in the dosimetry diode scanned across the irradiation area. The CBC2 is superimposed.

such as the bias distribution. This also prevents the positioning of the DUT too close to the source, and a compromise is required between the maximum flux deliverable and its uniformity. The beam profile measured during the irradiation of the CBC2 is plotted in Fig.A.5.

A.0.3 Uncertainty

Total ionizing dose levels are typically quoted in the literature with an uncertainty of ±10% [157]. In the case of the CBC2 TID test, the sources of uncertainties and assumptions considered are:

- Thickness of the spectrometer silicon sensor: the specifications for the instrument do not quote uncertainty, however wafer thickness is a very well controlled parameter (usually in the order of 1 μmRMS); such modest variance and the effect of the 0.1 μm inactive surface layer can be considered negligible in the calculation of the incident spectrum and of the corresponding dose. It was verified that an increase of 6 μm (≈6σ) yields a dose increase of <0.5%.

- Energy resolution of the spectrometer (150 $e^{-}_{FWHM}$): as shown in Fig. A.4a, at energies corresponding to the molybdenum emission lines, the absorption
coefficients for silicon and silicon oxide are not a strong function of the photon energy, so the broadening of the spectrum peaks on the dose calculation is considered negligible.

- The thickness of the dosimetry diode used is quoted as 263 µm ±10 µm. A more exact value could be derived by rotating the diode and measuring the resulting current under constant irradiation [156]; however this was not possible in this setup. The effect on the dose was calculated to be < ±3%. The resolution of the picoammeter used for the diode current (20 fA) is negligible.

- The effect of the uncertainty in the thickness of the top layers of the CBC2, quoted in the design manual as ±6 σ values, was estimated by performing the same dose reconstruction described in Section A.0.1 for the extreme cases of maximum and minimum layer thickness. The total BEOL transmission for the different scenarios considered is plotted in Fig.A.6b. The difference in the dose for the two cases accounts for less than 5% of the dose delivered to the active layer of the chip A.1. Much more influential is the exact composition of the layers: full-metal coverage for all layers yields a reduction in dose of ~14% with respect to the assumption of 50% coverage used in the above calculation. 100% metal coverage across the whole chip is however an unrealistic scenario, also prevented by specific design rules, and any realistic deviation from the assumed 50% coverage would translate into a much lower dose variation.

- The spectrum acquisition was conducted once the detector became available half-way through irradiation, which required the CBC2 to be moved out of the beam to insert the Vortex. The translation was done through the motorized stepper, but to account for possible changes in the absolute position of the CBC2 a second dosimetry was performed at the end of the irradiation, confirming the pre-irradiation readings.

- This analysis does not account for the so-called Absorbed-Dose Enhancement Effect (ADEE) [103, 121], which can increase the effective delivered dose when the target layer is surrounded within hundreds of nanometres by other layers.
Figure A.6: Left: cross-section of the BEOL layers of the CMOS process [41]. Right: total transmission coefficient for several BEOL thickness and coverage scenarios.

<table>
<thead>
<tr>
<th>BEOL thickness</th>
<th>min.</th>
<th>nom.</th>
<th>max.</th>
</tr>
</thead>
<tbody>
<tr>
<td>diode thickness [µm]</td>
<td>253</td>
<td>1.08</td>
<td>1.03</td>
</tr>
<tr>
<td></td>
<td>263</td>
<td>1.05</td>
<td><strong>1.00</strong></td>
</tr>
<tr>
<td></td>
<td>273</td>
<td>1.02</td>
<td>0.97</td>
</tr>
</tbody>
</table>

Table A.1: TID for diode and BEOL thicknesses at ±6σ (normalized against nominal value).

with different atomic number and composition. The effect is due to Compton and photoelectrons generated in one layer diffusing and depositing their energy onto nearby layers: it is therefore extremely difficult to quantify since it depends on layer composition, local electric field and geometry of individual circuits. For this reason, ADEE has only been studied for simple devices, typically a silox structure (SiO₂/polysilicon gate/SiO₂/Si), for which dose enhancements of ~ 10% were observed [103]. Although in the CMOS process used for the CBC2 the first metal lies some 350 nm above the gate oxide, the stacked, inhomogeneous nature of the BEOL could cause dose enhancement; however this is never considered during TID qualification of complex ASICs and only results in an additional safety margin.
Appendix B

CBC2 ESD Protection

Electrostatic discharge (ESD) is a failure mechanism caused by discharge of static charge through the chip. With shrinking gate oxide thickness and interconnect features, modern IC technologies have become increasingly susceptible to destructive damage due to ESD [158].

Two factors must be considered when designing against ESD: the failure current and the series resistance for both wires and devices [158]. I/O wires and ESD protection devices must be sized to withstand discharge current without fusing, and the resistance to the nearest discharge path through a protection device must be low enough so that the discharge current does not translate into a voltage sufficient to break down the dielectric inside the chip. The gate oxide of CMOS devices, only a couple of nm thick, is a weak point and oxide punch-through is one of the basic failure mechanisms. Others include junction burnout and metallization burnout. ESD damage can be catastrophic or latent: when an ESD pulse is not sufficiently strong to destroy a device, it nevertheless damages it and degrades performance and lifetime, which of course would be detrimental in an environment with no access for maintenance such as the CMS Tracker. In industry such devices are referred to as walking-wounded, and they are particularly difficult to screen for since still functional [158].

Three different models exist which describe the standard ESD failure mechanisms [159]:
Human Body Model (HBM) simulates the human discharge through the chip; because of the relative high-impedance of the human body, this model accounts for the longest duration of a current pulse, with the lowest values of current and voltage. While most commercial devices, destined for consumer markets, should be resilient to ESD due to human contact, this particular mechanism can be avoided in a well-controlled environment where simple but strict precautions, such as use of wrist-straps, are observed whenever ICs are handled. For this reason, the CBC was not designed to be HBM-immune.

Machine Model (MM) represents a metal tool discharging through the chip, and has medium voltage, current and duration. This should also be easy to avoid in a controlled environment by appropriate grounding of all equipment. This model is now considered redundant to the HBM [160].

Charged Device Model (CDM): this model simulates the effects of a charged chip discharging through any two pads, due to either production equipment or manufacturing and assembly processes. This mechanism has the highest voltage and current and shortest pulse duration. For the CBC this is probably the most relevant ESD failure mechanism, hence the protection devices have been suitably sized.

Each of these models requires different protection devices with different size active and passive components. The simplest ESD protection is to use reversed-biased junctions in every I/O pad, to provide an alternative path for the discharge current and to clamp the voltage. Designed for high-current capabilities and fast turn-on, these diodes can have a non-negligible parasitic capacitance that, when added in parallel to the input pads of the ASIC, can degrade the noise performance of the preamplifier. Hence a trade-off exists between an optimum low-noise system, with no input protection, and a noisy but ESD-robust one.

In the case of the CBC, the input pads of the ASIC are protected by a standard double-diode configuration between supply and ground (Fig.B.1). The diodes are placed as close as possible to the input transistor, to minimize the resistance of the
Figs. B.1 and B.2 shows a benchmark comparison between the protection diodes for the CBC and the APV25. Both these front-end designs have the same low-noise constraints and similar production volumes and module assemblies. In terms of area and perimeter the protection diodes on the CBC are rather small, however important information on the depth and profile of the implant is missing. The CBC also makes use of a p-n-p up-diode with low on-resistance due to the low resistance of its own n-well, which delivers fast clamping performance with reduced area and reduced capacitive loading.

Although not enough CBC chips have been bonded and tested to make a conclusive assessment at this stage, the yield of bonded chips appears to be extremely good, with no dead channels observed for underfilled devices, so the ESD protection seems
adequate. The feasibility of larger ESD devices was investigated in response to one of the early CBC2 dual modules showing some channel failing after bonding: such devices could ultimately be accommodated into a future version of the ASIC without altering significantly the existing layout. However, the cause of the dead channels is thought to be due to mechanical stress in the non-underfilled version of the hybrid resulting from the bowing of the thin substrate during bonding.
Glossary

2S Strip-Strip module. 33, 51, 52, 57, 79, 80

ADC Analog-to-Digital Converter. 48

ADEE Absorbed-Dose Enhancement Effect. 196

ATLAS A Toroidal LHC ApparatuS. 19, 27, 112

BEOL Back-End Of Line. 191, 196

BSM Beyond the Standard Model. 19, 20

BX Bunch Crossing. 21, 25, 27, 38, 39, 73

C4 Controlled Collapse Chip Connection. 47, 77

CIC Concentrator IC. 35, 36, 35, 36, 38, 67, 174, 175, 182

CMS Compact Muon Solenoid. 19

CMSSW CMS Software. 162, 169

CNM Centro Nacional de Microelectrónica, Barcelona. 84, 87, 156, 159, 162, 164

CWD Cluster Width Discrimination logic. 67, 71, 94, 160

DAC Digital-to-Analogue Converter. 57, 59, 63, 87, 90, 91

DAQ Data Acquisition. 82, 85, 112, 113, 116, 153, 158
**DESY** Deutsches Elektronen-Synchrotron. 153, 155, 158

**Digis** detectors hits. 169

**DLL** Delay-Locked Loop. 175

**DTC** Data, Trigger and Control. 35, 38

**DUT** Device Under Test. 108, 153, 156, 166, 188, 191, 193, 194

**ECAL** Electromagnetic Calorimeter. 22, 32, 41, 51

**ECC** Error Correction Code. 105

**ELT** Enclosed Layout Transistor. 102, 103, 104, 183

**ENC** Equivalent Noise Charge. 86, 174

**ESD** Electrostatic Discharge. 48, 198

**EYETS** Extended Year End Technical Stop. 21

**FMC** FPGA Mezzanine Card. 158

**FOXFET** Field-Oxide FET. 103, 125, 127

**GBT** GigaBit Transceiver. 35, 38, 158, 174, 204

**GLIB** Giga Bit Link Interface Board. 158

**HC** Hadronic Calorimeter. 22

**HE** Hamming Encoding. 105

**HEP** High Energy Physics. 43, 44, 98, 102, 107, 112, 132

**HIP** Highly Ionizing Particle. 100, 175

**HL-LHC** High-Luminosity LHC. 19, 21, 22, 97, 106, 146, 147, 151

**HLT** High Level Trigger. 26, 27, 28
**Glossary**

**I²C** Inter-Integrated Circuit bus. 17, 51, 54, 64, 65, 74, 83, 86, 87, 89, 143, 148, 149, 151

**IC** Integrated Circuit. 38, 100, 191, 198

**IC** Imperial College. 81, 85

**Infineon** Infineon Technologies AG. 84, 156, 159

**L1** Level 1 Trigger. 26, 27, 54, 67, 71, 74

**LDO** Low Dropout Linear Regulator. 51, 54, 82, 92

**LET** Linear Energy Transfer. 139, 142

**LHC** Large Hadron Collider. 19

**LIF** Light Ion Irradiation Facility. 146

**LP-GBT** Low Power GBT. 35, 38

**LS** Long Shutdown. 21, 22

**MAPS** Monolithic Active Pixel Sensor. 153

**MPW** Multi Project Wafer. 44

**MVR** Majority Voting Register. 144

**OCC** Offset Correction and Correlation Logic. 68, 69, 177

**OCCB** Offset Correction, Correlation and Stub Bend Logic. 177, 178, 181

**OOT** Out-of-Time. 25

**OT** Outer Tracker. 25

**PS** Pixel-Strip module. 33

**PSR** Power Supply Rejection. 92
| **PU** | Pile-up. 25, 26, 38 |
| **RHBD** | Radiation Hardening By Design. 100, 106, 107 |
| **RHBP** | Radiation Hardening By Process. 100, 101 |
| **SCR** | Silicon Controlled Rectifier. 13, 106, 107 |
| **SEE** | Single Event Effect. 100, 101, 139, 205 |
| **SEL** | Single Event Latchup. 100, 106 |
| **SET** | Transient SEE. 100, 102, 103, 104, 105 |
| **SEU** | Single Event Upset. 100, 137, 141, 142, 144, 146, 147, 148, 149, 150, 151, 182 |
| **SimHits** | hits generated by the GEANT-4 simulation. 160, 169 |
| **SLVS** | Scalable Low-Voltage Signalling. 83, 85, 158, 182, 183 |
| **SM** | Standard Model. 20 |
| **SNM** | Static Noise Margin. 132, 134 |
| **SPS** | Super Proton Synchrotron. 51, 171 |
| **SRAM** | Static Random-Access Memory. 73, 104, 127, 128, 130, 131, 132, 136, 137, 142, 143, 144, 147, 148, 149 |
| **STI** | Shallow-Trench Isolation. 101, 102, 130 |
| **TDC** | Time to Digital Converter. 158 |
| **TID** | Total Ionizing Dose. 98, 102, 104, 107, 109, 110, 112, 113, 122, 147 |
| **tkLayout** | Software package developed to quickly evaluate the design choices for the development of the CMS Tracker. 31, 32, 68 |
| **TLU** | Trigger Logic Unit. 153, 154 |
| **TMR** | Triple Mode Redundancy. 105 |
VCTH Comparator Threshold ($I^2C$ register setting). 63, 87, 124, 159, 164, 166

VME Asynchronous master-slave architecture (from: VERSA-Module Euro card). 82

WIMPS Weakly Interacting Massive Particles. 21

XDAQ CMS Online Software. 159
References


http://stacks.iop.org/1748-0221/7/i=09/a=C09001.


http://stacks.iop.org/1748-0221/7/i=01/a=C01075.


http://stacks.iop.org/1748-0221/8/i=01/a=C01041.

http://stacks.iop.org/1748-0221/10/i=01/a=C01046.


[138] M. Berger, J. S. Coursey, M. A. Zucker, and J. Chang, “ESTAR, PSTAR, and ASTAR: computer programs for calculating stopping-power and range tables for electrons, protons, and helium ions (version 1.2.3),”.


https://indico.cern.ch/event/348920/session/2/contribution/18/
attachments/688987/946200/EGallo_141106.pdf.

https://indico.cern.ch/event/350910/session/4/contribution/12/
attachments/693561/952319/Electronics_111114.pdf.

https://indico.cern.ch/event/375536/session/4/contribution/21/
attachments/748512/1026858/Electronics_240215.pdf.

[152] D. Braga, L. Caponetto, D. Ceresa, M. Raymond, F. Vasey, S. Viret, and
Y. Zoccarato, “I/O data formats for the Concentrator Integrated Circuit,”
04/05/2015. https://espace.cern.ch/Tracker-Upgrade/Electronics/
CIC/SharedDocuments/Dataformats/CIC_I0_Formats_v2.pdf.

SRAM cell design for the 32 nm node and beyond,” in VLSI Technology,
2005.

[154] M. J. Berger, J. Hubbell, S. Seltzer, J. Chang, J. Coursey, R. Sukumar,

Photoabsorption, Scattering, Transmission, and Reflection at E= 50-30,000

“Determination of X-ray flux using silicon pin diodes,” Journal of

