MANCHESTER 1824

The University of Manchester

DOCTORAL THESIS

Developing a Silicon Pixel Detector for the Next Generation LHCb Experiment

Peter Švihra

Supervisor: Prof. Chris Parkes, PhD.

A thesis submitted to The University of Manchester for the degree of Doctor of Philosophy

in the

Department of Physics and Astronomy
Faculty of Science and Engineering

2021

Blank page

Contents

Contents 3

Figures 7

Tables 11

Acronyms 13

Abstract 15

Declaration of Authorship 17

Copyright Statement 19

Acknowledgements 21

Preface 25

Part I
LHCb Upgrade

1 LHCb experiment and its Upgrade 31

1.1 CERN and LHC 31
1.1.1 LHC experiments 32
1.2 LHCb and LHCb Upgrade 34
1.2.1 VELO 36
1.2.1.1 Upgrade overview 37
1.2.2 Tracking stations 37
1.2.2.1 Upgrade overview 38
1.2.3 RICH 39
1.2.3.1 Upgrade overview 40
1.2.4 Calorimeters 40
1.2.4.1 Upgrade overview 41
1.2.5 Muon Chambers 42
1.2.5.1 Upgrade overview 42
1.2.6 WinCC/OA software 42
1.2.6.1 JCOP framework 44
1.3 VELO Upgrade 45
1.3.1 Mechanics 45
1.3.1.1 RF foil 46
1.3.1.2 Vacuum-Feedthrough Board 47
1.3.2 Cooling 48
1.3.2.1 Performance 48
1.3.2.2 Production 49
1.3.2.3 Safety 50
1.3.3 Electronics 50
1.3.3.1 Sensors 51
1.3.3.2 VeloPix ASICs 52
1.3.3.3 On detector 54
1.3.3.4 Off detector 55
1.4 Summary 56

2 Semiconductor detectors 57

2.1 Interaction of radiation with matter 57
2.1.1 Energy loss of charged particles 57
2.1.2 Carrier generation 59
2.1.3 Multiple Coulomb scattering 60
2.2 Semiconductors 61
2.2.1 Silicon 62
2.2.2 Carrier motion 63
2.2.3 Carrier density 64
2.2.3.1 Doping 66
2.2.4 p-n junction 66
2.2.4.1 Reverse biasing 70
2.2.5 Detectors 71
2.2.5.1 Readout 72
2.2.5.2 Segmentation and resolution 72
2.3 Radiation damage 74
2.3.1 Surface damage 74
2.3.2 Bulk damage 75
2.3.2.1 Leakage current 76
2.3.2.2 Depletion voltage 78
2.3.2.3 Charge trapping 78
2.3.2.4 Annealing 79
2.4 Summary 80

3 VELO Upgrade modules 83

3.1 Assembly 83
3.1.1 Bare module 83
3.1.2 Tiles attachment 85
3.1.2.1 Alignment 85
3.1.2.2 Gluing 85
3.1.3 Hybrid attachment and wire-bonding 88
3.1.3.1 Alignment 88

3.1.3.2 Gluing 88
3.1.3.3 Wire-bonding and HV cabling 89
3.1.4 Cabling 89
3.1.4.1 Interconnect and data cables 89
3.1.4.2 LV cables 90
3.2 Metrology 90
3.2.0.1 SmartScope 90
3.2.1 Substrate flatness 93
3.2.2 Glue layer thickness and tile flatness 93
3.2.3 Tile position 93
3.2.4 Hybrid position 94
3.2.5 Wire-bond pull testing 94
3.2.6 Photo inspection 94
3.3 Module installation 94
3.3.1 Test setup 94
3.3.1.1 Vacuum 94
3.3.1.2 Cooling 95
3.3.2 Installation 96
3.3.3 Displacement 97
3.3.4 Cooling connection and leak-check 98
3.3.5 Thermal cycling 98
3.4 Module performance testing 99
3.4.0.1 Software 99
3.4.1 Communication 100
3.4.2 TFC 100
3.4.3 PRBS 101
3.4.4 Noise and Equalisation 102
3.4.5 Cooling performance 103
3.4.6 High voltage 104
3.4.6.1 Software 104
3.4.6.2 IV curves 106
3.4.6.3 Ramp-up 106
3.5 Summary 106

4 Database and quality assurance 107
4.1 Database 107
4.1.1 Modules project 108
4.1.1.1 Grading 112
4.2 Assembly 113
4.3 Metrology 113
4.3.1 Substrate flatness 113
4.3.2 Glue layer thickness and tile flatness 117
4.3.3 Tile position 119
4.3.4 Wire-bond pull testing 122
4.3.5 Module displacement and thermal cycling 123
4.4 Electrical testing 125
4.4.1 Communication 125
4.4.2 TFC 127
4.4.3 PRBS 127
4.4.4 Noise and Equalisation 129
4.4.5 Cooling 131
4.4.5.1 Bandgap 132
4.4.5.2 Thermal degradation . . . 136
4.4.6 High voltage 136
4.4.6.1 IV 137
4.4.6.2 Ramp-up 137
4.5 Evaluation 138
4.5.1 Grading 139
4.5.2 Quality 140
4.6 Summary 141

5 Charge collection efficiency scans 143
5.1 Charge collection efficiency 143
5.1.1 Scan definition 144
5.1.1.1 Threshold dependency . . 145
5.1.1.2 Beam dependency 145
5.1.2 Simulation 145
5.1.2.1 Angle and position dependency 146
5.1.2.2 Landau reconstruction . . 149
5.1.2.3 Radiation damage 150
5.1.3 Scan duration estimates . . . 151
5.1.4 Verification and required work 154
5.2 Summary 155

Part II

LHCb Upgrade II

6 Future upgrades **159**
6.1 LHCb Upgrade II 159
6.1.1 VELO Upgrade II 161
6.2 Ultra fast silicon detectors 162
6.2.1 Impact ionisation 163
6.2.2 Timing performance 164
6.2.3 Fill Factor 165
6.2.4 Gain 166
6.3 Sensor simulations 167
6.3.1 Sentaurus TCAD 167
6.3.2 SRIM 169
6.4 iLGAD simulation 169
6.4.1 Doping profile 169
6.4.2 IV curves 171
6.4.3 Charge collection 172
6.4.3.1 Thickness 250 µm 175
6.4.3.2 Thickness 50 µm 175
6.4.4 Test beam simulation 178
6.4.4.1 Energy deposition 179
6.4.4.2 Angle-voltage scan 180
6.5 Summary 181

7 Summary and Conclusions **183**
7.1 LHCb Upgrade I 183
7.2 LHCb Upgrade II 185

Bibliography **187**

Appendices

High Energy particle physics Ventilator **199**

Word count: 39,513

Blank page

Figures

1.1 Schematic overview of the CERN accelerator complex. 32
1.2 The four main LHC experiments. 34
1.3 The angular distribution of $b\bar{b}$ pair production. 35
1.4 Schematic view of VELO R and Phi sensors. 36
1.5 Schematic view of VELO modules with the RF box. 36
1.6 Intensity of the magnetic field for trackers. 38
1.7 Geometry of the UT detector. . . 39
1.8 Reconstructed Cherenkov angle for isolated tracks. 40
1.9 Schematic view of RICH1 detector. 40
1.10 Segmentation of the ECAL and HCAL. 41
1.11 Simplified WinCC/OA manager schematic diagram. 42
1.12 VELO Upgrade detector visualisation. 45
1.13 Material budget of the VELO Upgrade. 46
1.14 Simulated IP_x resolution of VELO and VELO Upgrade. 46
1.15 Steps of the RF foil production. 47
1.16 Final design of the silicon cooling substrate. 49
1.17 Steps of the silicon microchannel substrates production. Silicon (gray), mask (blue), metallisation (gold). . 50
1.18 Fully assembled VELO module. 51
1.19 Simulated radiation map 52
1.20 Schematic of the pixel architecture 53
1.21 Schematic of module readout. . 54
1.22 Simulation of the expected data rates in the VeloPix ASICs 56
2.1 Mean energy loss of a positive muon in copper as a function of momentum. 58
2.2 Probability of energy loss interactions for 500 MeV pions in silicon. 60
2.3 Schematic visualisation of the Coulomb scattering. 61
2.4 Comparison of the band structure for metal, semiconductor and insulator materials. 62
2.5 visualisation of the silicon valence band with four covalent bonds. . . . 63
2.6 Schematic semiconductor band diagram, density of states $N(E)$, Fermi-Dirac distribution $f(E)$ and $n_{n(p)}$ electron (hole) carrier concentration. 67
2.7 Schematic visualisation of the p-n junction. 68
2.8 Spatial profiles of charge density Q, intensity of the electric field E and electric potential ϕ across the p-n junction. 68
2.9 Schematic cross-section of a hybrid pixel detector. 73
2.10 Comparison of the weighting field for a planar and segmented device. 74
2.11 Visualisation defects in the crystal lattice. 76
2.12 Dependency of damage cross-section on particle energy . . . 77
2.13 Dependency of leakage current on the neutron fluence equivalent for different silicon resistivity and doping. 77
2.14 Dependency of depletion voltage and effective doping concentration on the neutron fluence equivalent. . . . 77

2.15 Velocity profiles of electron-hole pairs for different bias voltages and irradiation. 79
2.16 Annealing behaviour of selected detector properties. 80

3.1 Assembly steps of the VELO upgrade modules. 84
3.2 Gluing robot Fisnar F5200N with a heat gun. 86
3.3 Comparison of glue patterns deposited on tiles. 87
3.4 CT scan of tiles attached to the cooling substrate using simple star pattern. 87
3.5 Zoomed in image of a glued FE hybrid next to a tile 89
3.6 SmartScope device with laser and optical sensors used for mechanical measurements. 91
3.7 SmartScope marker alignment. . 92
3.8 SmartScope routine to measure glue thickness and tile flatness. . . . 93
3.9 CAD model of the module test setup in UoM. 95
3.10 Pfeiffer Vacuum D-35614 Asslar Turbo Pump. 96
3.11 Liquid CO_2 cooling machine used for cooling the modules. 97
3.12 GasCheck G3 leak detector used to check for CO_2 leaks. 98
3.13 Main VELO FSM WinCC/OA panel. 100
3.14 Module monitoring panel. 100
3.15 Module TFC panel. 101
3.16 Module PRBS panel. 102
3.17 Module noise and equalisation panel. 103
3.18 Screenshot of the software developed for controlling the PSU 105

4.1 Database module overview page. 109
4.2 Database single module page. . 110
4.3 Comparison of substrate flatness measurements performed by CERN and UOM on SUB79 116
4.4 Trend of substrate flatness results, showing the quartiles of the Δz distribution. 117
4.5 Measurement of glue layer thickness and tile flatness of CLI, MOD_120. 118
4.6 Trend of CLI (VP0) tile mean glue thickness and mean tile flatness results. 119
4.7 Average glue thickness of ASICs under CLI and NLO tiles. 119
4.8 Visualised tile position calculated variables x, y and angle. 120
4.9 Visualised closest distance between NLO (VP1) and NSI (VP2) tiles. 120
4.10 Trend of the tile position metrics. 121
4.11 Distribution of pull test strengths before and after TC. 122
4.12 Trend of all pull test forces. . . 123
4.13 Trend of the maximal displacement during the module cool-down. 124
4.14 Displacement and temperature during TC of MOD_98. 124
4.15 Trend of analogue and digital voltage during communication tests of NSI (VP2) tile. 128
4.16 Trend of TFC response before TC. 128
4.17 Example of PRBS test data . . 129
4.18 Trend of PRBS test BER from before TC. 129
4.19 Example of equalisation results for NSI (VP2). 130

4.20 Comparison of noise results in NSI (VP2) for different tests. The tile is used in MOD_103. 131
4.21 Trend of mean noise and percentage of masked pixels. 131
4.22 Example of the automatically produced bandgap plots for MOD_119. 134
4.23 Position of tiles and NTC sensors on a module. 135
4.24 Comparison of the cooling performance based on the glue layer thickness. 135
4.25 Trend of cooling performance for CLI (VP0) before TC. 135
4.26 Dependency of thermal degradation on glue layer thickness after TC. 136
4.27 All IV curves for the NSI (VP2) tile. 137
4.28 Example of voltage ramp with current monitoring for MOD_110. 138
4.29 Trend of ramp-up results for NLO (VP1) tile. 138
4.30 Final module production grades. 139
4.31 Resulting quality of the modules. 140
4.32 Final grade and quality of the modules as of 31 August 2021. . . . 141

5.1 Dependency of effective depletion voltage on fluence for the old VELO sensors. 144
5.2 Example of the CCE scan for the Run 1 and 2 VELO. 144
5.3 Raw data output from simulated dependency of charge on voltage for non-irradiated VELO. 147
5.4 Schematic view of a MIP traversing two pixels under an angle θ. 147
5.5 Angular dependency of the pixel collected charge, obtained at 100 V, full depletion of the simulated sensors. 148
5.6 Dependency of the hit reconstruction efficiency on the applied threshold for different angular cuts. 149
5.7 Simulation results of a threshold scan for non-irradiated VELO sensors, made for different bias voltages. 149
5.8 Simulation results of a threshold scan for irradiated VELO sensors, made for different bias voltages. . 151
5.9 Dependency of the MPV from the Landau fit on voltage for normal and irradiated sensors. 151
5.10 Schematic view of the CCE scan where one of the five modules is tested. 152
5.11 The CCE scan duration estimates, applicable to a subset (fifth) of the modules. 153

6.1 Schematic view of the LHCb Upgrade II detector. 160
6.2 Visualisation of the interactions from a single bunch crossing in VELO. 161
6.3 Schematic view of different types of the timing plane solutions and corresponding performance of the primary vertex resolution. 162
6.4 Comparison of LGAD types of devices. 166
6.5 Q-effective representation from doping profile of LGAD 167
6.6 Simulation results of dependency of LGAD gain on Q-effective 167

6.7 Sentaurus workbench used to control the Sentaurus TCAD simulation. 168
6.8 Implant masks used for the implantation in the recent iLGAD sensors run. 170
6.9 Approximation of the process simulation data of n-implant (phosphorus) and p-multiplication (boron). 171
6.10 Meshing map of the simulated diode 171
6.11 Absolute electric field profile of the simulated LGAD 171
6.12 IV curve of no gain and iLGAD devices, varying thickness 172
6.13 2D current profiles of 250 µm thick sensor without gain. 173
6.14 2D current profiles of 250 µm thick iLGAD sensor. 174
6.15 Current profiles registered by the central electrode of 250 µm thick sensors. 176
6.16 Collected charge and gain as a function of the injected charge for the 250 µm thick sensors. 176
6.17 Induced current of the neighbouring channels for a different distance from the central point. . . 177
6.18 Current profiles registered by the central electrode of 50 µm thick sensors. 177
6.19 Collected charge and gain as a function of the injected charge for the 50 µm thick sensors. 177
6.20 LET dependency of 1.6 MeV in silicon, simulated in SRIM. 178
6.21 Example of the TCAD implemented single proton energy deposition. 179
6.22 Comparison of the collected charge Q for sensors without gain and iLGADs, based on the bias voltage and incidence angle. 180
6.23 Collected charge and gain as a function of incidence angle for protons. 181

7.1 Duration of the module assembly and testing done at the UoM. . . . 184

Tables

1.1 Comparison of the VeloPix and Timepix3 ASICs 53

2.1 Bethe-Bloch formula variables. . 59

4.1 Grading definition for metrology tests. 114
4.2 Grading definition for electrical performance tests. 126
4.3 Module grade counts. 139
4.4 Weights used for quality calculation. 140

5.1 Boole simulation parameters. . . 146
5.2 List of 15 thresholds used for the simulated CCE scan. 150
5.3 Summary and comparison of CCE scan duration parameters. 154

6.1 Parameters defining pair creation from 1.6 MeV proton in silicon. . . 179

Blank page

Acronyms

This document is incomplete. The external file associated with the glossary 'acronym' (which should be called `thesis.acr`) hasn't been created.

Check the contents of the file `thesis.acn`. If it's empty, that means you haven't indexed any of your entries in this glossary (using commands like `\gls` or `\glsadd`) so this list can't be generated. If the file isn't empty, the document build process hasn't been completed.

You may need to rerun LaTeX. If you already have, it may be that TeX's shell escape doesn't allow you to run makeindex. Check the transcript file `thesis.log`. If the shell escape is disabled, try one of the following:

- Run the external (Lua) application:

  `makeglossaries-lite "thesis"`

- Run the external (Perl) application:

  `makeglossaries "thesis"`

Then rerun LaTeX on this document.

This message will be removed once the problem has been fixed.

Blank page

Abstract

Developing a Silicon Pixel Detector for the Next Generation LHCb Experiment

by Peter Švihra

University of Manchester
Faculty of Science and Engineering
Department of Physics and Astronomy

The second long shutdown of the LHC presents an opportunity for the LHCb experiment to upgrade its detector systems and switch to a fully software triggered readout. Its first tracking layer, the VELO detector, is no exception to this and is undergoing an upgrade increasing the number of sensitive channels from 180 thousand silicon microstrips to about 41 million pixels. The new system will operate with zero-suppressed readout at 40 MHz, while cooled down using evaporative liquid CO_2 in silicon microchannel plates.

The VELO Upgrade will consists of 52 modules, placed around the beam-pipe, built at the University of Manchester and Nikhef. The construction of the modules is a complex process that consists of a number of tight tolerance steps, their results verified both in metrology and in the electrical and thermal performance testing. In order to store data and track the performance a database has been developed, used to automatically analyse the uploaded values as well as compute the grades and quality of the individual steps and final modules. By the end of August 2021, 42 modules have been produced in Manchester, 37 of them with high quality and no issues present.

Due to the nature of the harsh radiation environment, the sensors have to withstand a fluence up to $1 \cdot 10^{16}\,1\,\mathrm{MeV}\,\mathrm{n_{eq}}\,\mathrm{cm}^{-2}$ and still provide a good signal to noise ratio. A new method of a charge collection scan has been proposed, linking the commonly used voltage scan with a threshold scan and using the extrapolated tracking information to estimate the amount of collected charge. The simulation indicates that the scan of a subset of modules will take about 8 min, a feasible duration despite the impact on the physics data taking.

A further upgrade of the LHCb is planned for Long Shutdown four of the LHC. This will operate at higher luminosities leading to a significant increase in the pile-up of the collisions from a single proton-proton bunch crossing. For this reason a precise time stamping $\mathcal{O}(50\,\mathrm{ps})$ is to to be added. This could be achieved in silicon detectors by using $\mathcal{O}(10)$ internal gain in the sensor. Simulations of the expected performance of a recently produced batch of sensors are presented. These characterise the anticipated performance of these $\mathcal{O}(50\,\mu\mathrm{m})$ segmented devices in a test beam, providing the impact of charge sharing and device response to an angular scan.

Declaration of Authorship

Candidate Name: Peter Švihra

Faculty: Faculty of Science and Engineering

Department: Department of Physics and Astronomy

Thesis Title: Developing a Silicon Pixel Detector for the Next Generation LHCb Experiment

This work represents the combined efforts of the author and his colleagues in the LHCb collaboration. Some of the content has been published elsewhere and/or presented to several audiences. No portion of the work referred to in the thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

Signed:

Date:

Blank page

Copyright Statement

Blank page

Acknowledgements

I would like to thank everyone who had helped me on my PhD journey, and those whom I had met during my studies. Your support, trust, and encouragement were vital.

First of all, I would like to give a huge thank you to my supervisors Chris and Marco. Both of you supported me, trusted in my decisions, and always managed to guide me in the right direction. I know you always stood behind me and did all you could to help, even if it was not obvious. I would also like to thank Mark and Adam, both of whom had the infinite patience with me when I was focused on everything else, rather than what they needed from me. I am glad I could work with you on such amazing projects.

Stefano, thank you. You are the perfect combination of a demanding and understanding project leader one could wish for. I knew I could ask you anything and you would always found time and had fought for me. I am very grateful that you managed to push me further and was not afraid to give me responsibility. I would also like to extend my gratitude to Alex for all his help and patience. I will, surely, never forget the CO_2 enthalpy diagram and the endless module testing sessions we had to endure. Oh, and the occasional Friday Sandbar with both of you when you joked how neither Dónal nor I can talk about work unrelated topics. Dónal, you are my brother-in-arms. Together we suffered through it all – be it the module production or the HEV. I think we were constantly trying to be better than each other and I am glad we did so, because the motivation of not slacking behind made me perform at my best capacity. To the rest of the VELO assembly team and all of the fifth floor personnel, I am really happy I had the chance to work with you. I knew I could rely on you for anything I needed.

While being in Manchester, I got quickly used to it being my second home. Regardless of your experiment, everyone in the HEP group deserves my thanks. I will miss the painstakingly long Christmas meetings when everyone got together after a difficult year, the weekly free lunch (what an excellent motivation to attend a seminar!) and the Fridays at Sandbar which on occasions ended up on Saturday morning. In particular, I would like to thank Ivan, Matteo, and Gianluca, I am glad I could spend so much time with you, talking about everything and nothing and sharing the daily struggle over a cup of proper espresso. Kiran, for a theory guy, you are cool. Thank you, Ste, for the late night lectures when I wanted to learn what you do and for the trivia that you always had on hand. And finally, Nico, you coming back from the LTA caused a golden age of social events and I will not forget the one-too-many beers (or Bloody Marys) after the long (working) days. Thank you for it all, including the lovely slides for the Christmas meeting.

I also would like to give huge thanks for the whole VELO team based at CERN and all other institutes. Thank you, Paula, I will always remember that no plot is ever final and that it is always possible to "add just one more thing". I am grateful for all the times you stood behind me and helped me, regardless of how late I asked. Special thanks to the database team as well – Victor, Kurt and Karlis. I learned a lot from all of you, thank you for trusting me with the divine privileges, I hope I left it in a better state. Karol, it was a pleasure working with you even if it was mostly the HEV and not the commissioning. I know I can be a pain sometimes, but I think you managed me along with the whole HEV software team well and I am glad that you trusted me with the chance to help.

My thanks are also in place for all the people that taught me and that I collaborated with over the years, both from the FNSPE CTU in Prague and BNL. Naming a few, Václav, you showed me the importance of knowing people and taught me that one should not worry about things one cannot change. I am really grateful that you trusted in me all these years ago and did what you could to support me. I know you would have been happy for me. Andrei, you accelerated my whole career by leaps and I cannot begin to imagine where I would have been without your help. You brought me into a fast-paced research environment and called me an expert even when I did not feel like one. Thank you both for forcing me out of my comfort zone and trusting in my skills. Thank you, Maria, for all of the long-distance support you provided me with. A special thanks goes to Michal. You taught me that I can endure anything and that all bad things will eventually pass. Thank you for trying to pass down all your experiences and wanting only the best for me, it means a lot.

Even though I am not great at keeping in touch, I am happy I have friends that are always here for me, regardless of where we are in the world. I would like to thank all occupants of the Hospoda Discord channel, the random talks, games and memes we shared were a nice distraction. Mito and Honza, thank you both for the jokes, profound discussions and all the good food we shared. I knew I could rely on you regardless of the problem. Thank you Tomáš and Ondra, the endless talks we had were always relaxing. Thank you, Andrej, our random get together gaming sessions always brought back kind memories. Jakub, thanks for the unending chat and all of the random shares. And Maty, I know you will always have my back and that I can rely on you for everything. Thank you for being there whenever I needed it.

Lastly, I want to thank my family. My parents, who had always supported me regardless of the choices I had made and for always wanting the best for me. My brother Janko with his wife Katka, who were cheerful on so many occasions and were happy to help. And finally, Lucka, my other half, thank you for the unending support, I would not have made it without you. Regardless whether I felt like I could not do it anymore or when I naively thought I perfected something, you managed to push me even further and help me with it.

I would like to dedicate this work to Malá babka, for she made sure I am who I am. I made it further than you ever anticipated and I know you would have been proud.

Thank you.

Vi Veri Veniversum Vivus Vici

Preface

While the goals of various research programmes are focused on very specific topics, the general motivation towards a better precision and larger statistical samples drives the development of new technologies. This is also true for High Energy Physics (HEP) experiments, where the systems are designed years in advance having to consider increased data rates, cooling requirements, radiation hardness and aiming towards improved resolution of the measured events.

Over the past few decades, the advancements in the silicon detector development have resulted in an excellent spatial resolution of $\mathcal{O}(10\,\mu\text{m})$, while improvements in the processing power and Application Specific Integrated Circuit (ASIC) design provide access to much higher readout speeds. Because of the planned increase of the instantaneous luminosity delivered by the Large Hadron Collider (LHC), the detector development has to further improve to keep up with the pile-up. For this reason a drive remains to add a precise time-stamping information of $\mathcal{O}(50\,\text{ps})$, that can be used to distinguish between the tracks from multiple proton-proton interactions during a bunch crossing in the interaction point.

This thesis is focused on the present and future upgrades of the tracking heart of the Large Hadron Collider beauty (LHCb) experiment – the VErtex LOcator (VELO) detector; and is structured as follows:

The first chapter gives an overview of the European Organization for Nuclear Research (CERN) with its accelerator complex, focused on the LHC and experiments built at its interaction points. One of them, the LHCb detector system, is described in more depth and its ongoing upgrade towards a trigger-less readout is outlined. The focus is then dedicated on the VELO Upgrade and its main components – mechanics, cooling and electronics.

The second chapter provides an overview of the interaction of ionizing radiation with matter along with a detailed insight into the physics of semiconductor detectors. This provides a backbone for the rest of the work as aspects ranging from diode reverse biasing and energy deposition to the radiation damage are described.

The third and the **fourth chapter** present the bulk of the work of the author as the description of the testing and analysis of the VELO Upgrade module assembly done at the University of Manchester (UoM) is provided. While the individual construction and testing processes are described in the **third chapter**, the **fourth** is dedicated towards the database development, automatic data analysis and final Quality Assurance (QA) of each test as well as of the finalised modules.

The fifth chapter establishes a starting point for the commissioning of the detector, namely the Charge Collection Efficiency (CCE) scans that are used for the monitoring of the sensor full depletion voltage which will change with the increasing adsorbed radiation dose during data taking.

The sixth chapter is focused on the future of the LHCb experiment. In order to better distinguish the information coming from the increased number of tracks, the detector system will once again undergo an upgrade where a timing information of the hits/tracks will be added. This requires new technologies, the use of the Ultra Fast Silicon Detectors (UFSD) by the tracking stations is investigated in more detail. The chapter also describes simulations of such devices, laying a ground work for the recently produced inverse Low-Gain Avalanche Diode (LGAD).

The design and upgrade of the VELO detector required large amounts of work which could not have been achieved by a single person. For this reason, most of the work performed has been done within the LHCb collaboration where the author's contribution has been dedicated to the individual aspects of the module assembly and their characterisation. This includes establishment of the metrology processes performed at the UoM, realization of the glue deposition studies, and execution and debugging of the module testing. For this purpose a number of software tools have been developed by the author, ranging from scripts simplifying the glue robot control through user interface controlling High Voltage (HV) Power Supply (PSU) up to some aspects of the WinCC/OA based control. The author has also been a major contributor to the development of the online database, containing all information from the module QA, where he developed algorithms used for data analysis and established automatized processes of the module grading and quality estimates. Parts of the work were presented by the author during multiple international conferences (IEEE2019, LHCC2020, ICHEP2020) and published in proceedings [1]. Besides this, the author designed a procedure to perform CCE scan of the VELO detector based on the binary readout of the VeloPix ASICs. The proposed method is supported by the results of minimum bias Monte-Carlo simulations performed using the LHCb developed framework. Finally, the author also contributed to the development of LGADs by means of simulating the device characteristics and their particle detection performance.

As most of the work presented was based on the production of the VELO Upgrade modules, the COVID-19 lockdowns prevented institute access and interrupted all of the relevant processes. The author of this theses therefore decided to participate in the CERN coordinated development of High Energy particle physics Ventilator (HEV) with the aim to create a high quality device for a fraction of the cost of commercially available solutions. While this work is not described in the thesis in detail, a short summary and a pre-print of a paper is provided in the **appendix**.

Apart from this, the author continued in his collaboration with groups in Brookhaven National Laboratory (BNL) and Czech Technical University in Prague (CTU), working on applications and data analysis of the Tpx3Cam – a single photon sensitive fast time stamping device. The author contributed to a number of research topics ranging from molecular dynamics to quantum information science and co-authored a wide range of papers (to list a few, see [2–7]). These works are not described here.

Blank page

Part I

LHCb Upgrade

Blank page

Chapter 1

LHCb experiment and its Upgrade

1.1 CERN and LHC

European Organization for Nuclear Research (CERN) is an international organization focused on the nuclear and HEP research. It is a significant meeting point for the scientists from the whole world and it provides bases for a number of collaborations. The organization is located on the Swiss-French border near Geneva, containing individual research facilities as well as supporting structures for assembly, testing and Research & Development (R&D). The most notable facility is the accelerator complex, shown in Fig. 1.1.

Large Hadron Collider (LHC) [9] is currently the largest particle accelerator in the world, with a length of 27 km located about 100 m below the ground. It is designed to collide both protons and heavy ions at four different intersections. The accelerator consists of two parallel vacuum tubes, liquid helium cooled superconducting magnets (1232 dipoles for keeping the circular path and 392 quadrupoles used for beam focusing) and 16 Radio Frequency (RF) cavities used for the acceleration. The operation of the complex is split to different so-called Runs, separated by Long Shutdown (LS) used both for upgrades and maintenance.

The process of the particle acceleration is done in multiple stages. After stripping the hydrogen atoms from their electrons, the protons are accelerated to 50 MeV in the Linear Accelerator (LINAC)2 (replaced by LINAC4 during LS2). They are then accelerated in a set of circular accelerators Proton Synchrotron Booster (PSB) (1.4 GeV), Proton Synchrotron (PS) (25 GeV) and finally Super Proton Synchrotron (SPS) (450 GeV) before being injected into the LHC. Both beams in the accelerator are filled according to a predefined filling scheme, usually containing 2808 bunches 25 ns apart with $1 \cdot 10^{11}$ protons each. The instantaneous luminosity [1] delivered by the accelerator is up to about $\mathcal{L}_{\text{inst}} = 1 \cdot 10^{34}\,\text{cm}^{-2}\,\text{s}^{-2}$, dependent on the actual beam cross-section and profiles in the interaction point. While the LHC is designed to operate at the maximal beam energy

[1] represents a measure of the accelerator's ability to produce inelastic collisions and serves as proportionality factor between the ratio of the detected events per second and their cross-section σ

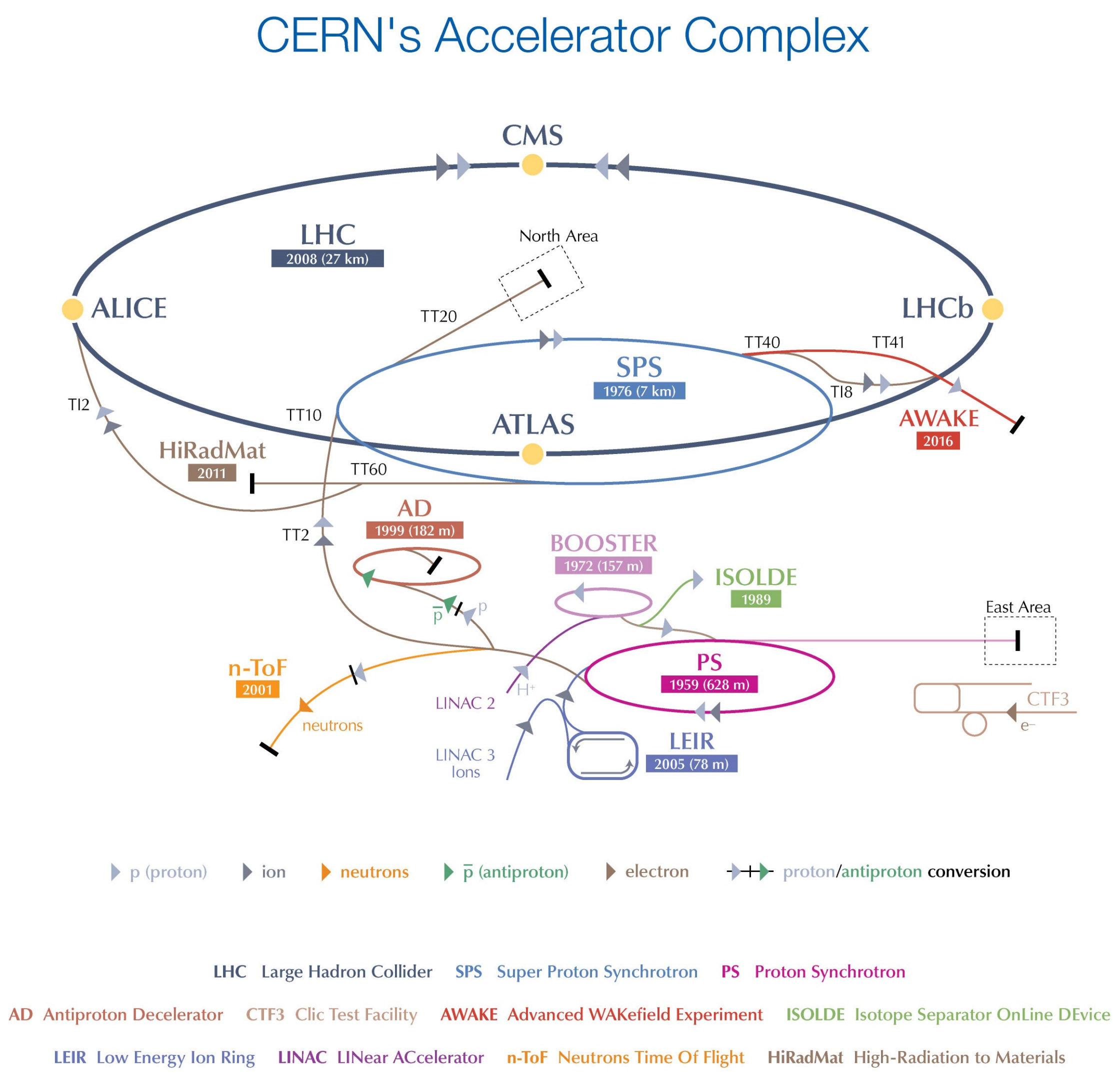


Figure 1.1: Schematic overview of the CERN accelerator complex. [8]

7 TeV, the centre of mass collision energy [2] has been set to only $\sqrt{s} = 8$ TeV during the Run1 operation (2011-2012; $\sqrt{s} = 7$ TeV during the first year). After the LS1 it has been increased to $\sqrt{s} = 13$ TeV for the Run2 data taking period (2015-2018) and is expected to further increase closer to $\sqrt{s} = 14$ TeV for Run3 (2022-2024). An upgrade to a High Luminosity LHC (HL-LHC) is expected at LS3 [10].

1.1.1 LHC experiments

While there are many experiments performing physics studies at CERN, the LHC contains only a few. The four biggest experiments located at the beam intersections are described in more detail, their schematic overviews are given in Fig. 1.2. All of them contain similar

[2] represented by the square root of the Mandelstam variable s, defined as the Lorentz-invariant squared sum of the colliding protons 4-momenta

components, varying in the actual technological implementation and design. This is both to have independent collaborations whose results may be used to verify the results of the others and to allow the experiment to have designs focused on particular physics goals. This underlying structure consists of tracking stations (reconstructing the vertices and tracks from the interaction point), Particle Identification (PID), calorimeters (to measure full energy deposition) and muon detectors. The experiments also heavily rely on the use a magnetic field to allow particle charge and momentum determination.

A Large Ion Collider Experiment (ALICE) [11] is dedicated to studying interactions of the heavy ions, primarily with lead ions but also with other ion and protons. The heavy ion experiments are usually performed only for a few weeks during the operational year with a smaller number of particle bunches in the LHC. The main physics interests are focused on improving the understanding of the strong interaction and formation of the quark-gluon plasma.

It is a 26 m long, 16 m tall and wide detector, weighing ≈10 000 t, constructed as a main barrel with a forward arm trigger and tracking modules. This provides full coverage of the collision point without saturating the trigger. The experiment uses a central 0.5 T solenoid enclosing the barrel (called L3, reused from the previous experiment), providing a 3 T m bending force on the particles. Another dipole magnet is used in the forward tracking stations.

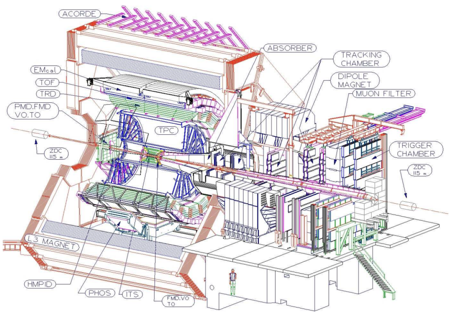
A Toroidal LHC ApparatuS (ATLAS) [12] represents a general purpose detector, meaning that it can be used to study most of the high transverse momentum LHC relevant physics. The collaboration (with CMS) made the discovery of the Higgs Boson [13] and has been researching its properties. It also performs many searches for direct production of new physics particles. It performs other studies of the strong and electroweak interactions as well as heavy ion physics and flavour physics studies.

The experiment is the largest one, 44 m long and 25 m in diameter, weighing ≈7000 t. It is constructed as a set of barrel detectors along with endcaps, providing excellent acceptance around the whole collision point. Magnetic field is provided using a 2 T solenoid for the inner detector and large toroidal ones for both barrel and endcap muon detectors, each with 0.5 T and 1 T, respectively.

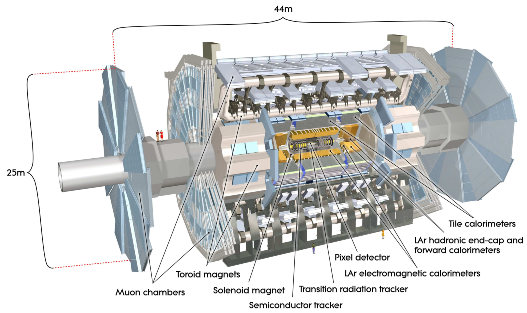
Compact Muon Solenoid (CMS) [14] is also a general purpose detector, its physics program being similar to that of ATLAS, the collaboration has an extensive flavour physics programme.

CMS is one of the smaller experiments, 22 m long and 15 m in diameter, however, weighing ≈12 000 t. The detector is again constructed in a form of a barrel with endcaps, its magnetic field generated in a compact 4 T superconducting solenoid.

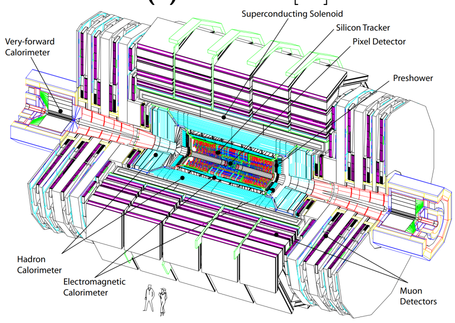
Large Hadron Collider beauty (LHCb) [15] was primarily designed with a focus on precision studies of Charge Parity (CP) violation and rare decays, however,



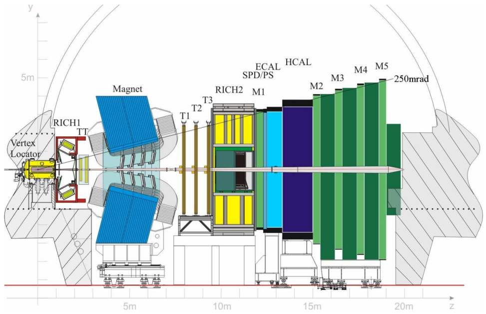
(a) : ALICE [11].



(b) : ATLAS [12].



(c) : CMS [14].



(d) : LHCb [15].

Figure 1.2: The four main LHC experiments.

expanding to other fields in the recent years. It is specialised in studies of the heavy quarks, namely based on the b and c quarks in their corresponding meson and baryons. It has discovered CP violation in the $B^{\pm}$ [16], B_s^0 [17] and charm systems. It has discovered 55 of the 62 hadrons at the LHC, including 16 exotic hadrons (tetraquarks and pentaquarks).

It is a 21 m long, 10 m wide and 13 m tall detector, weighing ≈5600 t, constructed as a forward arm from the collision point. The magnetic field is created with a dipole magnet, providing up to 1 T field intensity localised between the trackers. The polarity of the magnet can be changed, providing a tool to minimize the error when studying CP violation effects.

1.2 LHCb and LHCb Upgrade

As mentioned above, the LHCb detector [15] is a single arm forward spectrometer, used to study CP violation, focusing on b- and c- hadron decays. The experiment has to have excellent PID and tracking capabilities in order to select the interesting events from the large background.

Due to the limits imposed by the trigger rate and radiation damage, the detector operated at the maximal luminosity $\mathcal{L}_{\text{inst}} = 4 \cdot 10^{32}\,\text{cm}^{-2}\,\text{s}^{-2}$ which is 50 times lower than the LHC is capable of delivering. This is achieved by focusing of the beam, keeping the luminosity mostly constant throughout the fill. One advantage of the lowered rate is

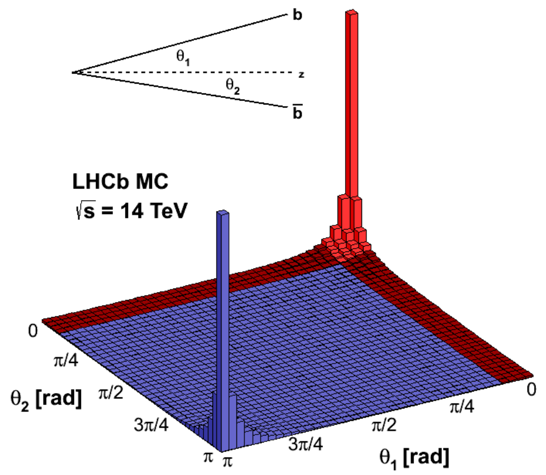


Figure 1.3: The angular distribution of $b\bar{b}$ pair production. The red region represents the LHCb acceptance. [18]

in having a simpler analysis due to typically only having one, or a small number of, pp collisions per event, the rates produce about $1 \cdot 10^{12}$ $b\bar{b}$ pairs in $1 \cdot 10^{7}$ s.

The detector coordinates are defined using the Cartesian system, where the z axis points downstream as the particles travel from the collision through the detector. The x axis with ± 300 mrad acceptance and y axis with ± 250 mrad are defining the horizontal and vertical directions, respectively. The minimal acceptance around the beam-pipe starts at 15 mrad. The coverage can be best defined using pseudorapidity – a relation of the angle of secondary particles with respect to the beam. It's value is $\eta = 0$ for the $\Theta = 90°$ angle and infinite for the parallel particles. In the LHCb about 24 % of $b\bar{b}$ quark pairs are contained within the whole acceptance $1.6 < \eta < 4.9$, even though it covers only about 2.3 % of the solid angle. [18] The angular distribution of the pair production is plotted in Fig. 1.3.

After a successful 10 year-long operation and delivered integrated luminosity of $L = 10\,\mathrm{fb}^{-1}$, the system is being upgraded with novel sub-detectors. [19] The upgrade is occurring during LS2 over a period of three years 2019-2021 with expected commissioning during the year 2022.

The most important aspect of the upgrade is in the increase of the data rates to the full 40 MHz readout along with the removal of the hardware level trigger. This way the maximal luminosity increases to $\mathcal{L}_{\mathrm{inst}} = 2 \cdot 10^{33}\,\mathrm{cm}^{-2}\,\mathrm{s}^{-1}$, the expected integrated luminosity by the end of Run4 is $L = 50\,\mathrm{fb}^{-1}$. Some detector systems also have to be upgraded in order to withstand the increased radiation damage from the higher fluences and provide better measurement resolution.

As mentioned before, all of the large LHC experiments consist of similar systems, providing information about tracking, PID, calorimetry and muon detection. The LHCb

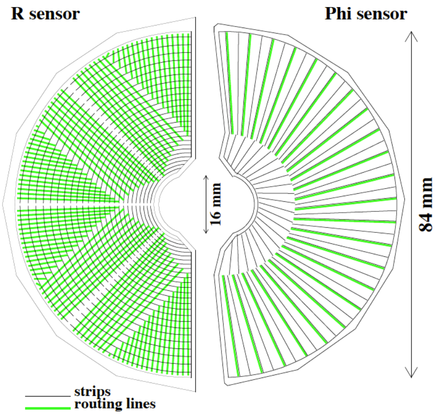


Figure 1.4: Schematic view of VELO R and Phi sensors. [22]

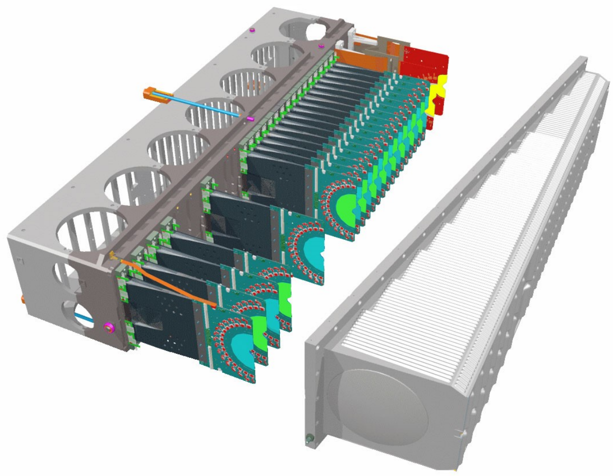


Figure 1.5: Schematic view of VELO modules with the RF box. [15]

specific implementation is discussed below.

1.2.1 VELO

The VELO [20] is a crucial sub-detector used for tracking and locating vertices, the original system consists of silicon strip detectors, arranged to determine r and ϕ coordinates of the passing particles. It is the system closest to, and surrounding, the interaction point, constructed as 21 sensitive planes on each side of the beam-pipe. The main purpose of the VELO is not only to identify tracks of the particles but also to separate the primary vertices of the beam collision from the secondary ones, originating from the short-lived decay of particles (namely b- and c- mesons). At least four stations have to register hit for the track reconstruction.

The majority of the silicon detector is made of 300 µm thick n^+-on-n sensors attached on the modules in a half-moon shape. The strips on one side of the module are arranged as radial concentric semicircles, while on the other side as angular stripes (see Fig. 1.4), both readout by a number of Beetle [21] ASICs placed on the periphery of the module. It is cooled down to $-10\,°C$ using evaporative CO_2 in small aluminium blocks close to the edge of the Printed Circuit Board (PCB) with the sensors.

The whole system is enclosed in a corrugated RF aluminium box, positioned only 5 mm from the beam. Its main purpose is to separate the VELO and beam vacua as well as protect the sensitive electronics from the beam-induced charge. The schematic of the assembled half detector with the RF foil is in Fig. 1.5. For the safety of the operation, the detector is kept in a retracted position from the beam and only moved in once the stable beam conditions have been declared.

Important metrics of the VELO performance are different resolutions characterising the tracking – Impact Parameter (IP), primary vertex and decay time resolutions. First of all, the IP is defined as the smallest distance of the track to the primary vertex. This is useful for the measurements of the long-lived particles as their decay occurs offset from the primary vertex, providing an excellent selection variable in data for their identification. The resolution is affected by the multiple scattering in the material, the detector design, notably the distance to the first measured point. The precision achieved by the detector is 35 µm for a transverse momentum higher than 1 GeV/c [22]. Secondly, the resolution of the primary vertex achievable by the system depends on the track multiplicity (number of tracks from a vertex) and is 13 µm in x, y and 75 µm in z for 25 tracks. Finally, an important factor to consider is resolution of the decay time, used in the particle lifetime measurements, the achieved precision being ≈50 fs. [23].

1.2.1.1 Upgrade overview

Contrary to the first generation system, the Upgrade [24] consists of silicon pixel detectors, increasing the effective number of channels from 180k up to almost 41M. In addition, whereas the previous VELO was triggered with 1 MHz readout, the new system is readout untriggered at the beam-crossing frequency of 40 MHz. However, while the old detector provided Analog-Digital Converter (ADC) counts (translatable to the amount of deposited energy), the Upgrade will produce only binary pixel hits during the nominal operation conditions. The upgrade will be further described in Sect. 1.3.

1.2.2 Tracking stations

Apart from the VELO, the LHCb contains other dedicated tracking stations. They are located upstream (close to the VELO) and downstream of the magnet as the particles travel.

The stations are assembled as pairs of 1D detection modules, each module consisting of two sensitive planes rotated 5° with respect to each other. The structure of all detectors therefore follows the x-u-v-x pattern (vertical, 5° anti-clockwise, 5° clock-wise, vertical). The use of such pattern minimizes the amount of fake tracks, known as ghosts, typical for a set of 1D detectors.

The average track finding efficiency in the whole acceptance of the system is above 96 % for the momentum range $5\,\mathrm{GeV/c} < p < 200\,\mathrm{GeV/c}$. The momentum resolution being 5‰ to 10‰. [23]

Tracker Turicensis (TT) [25] The detector is located before the magnet, covering the whole acceptance of the experiment. It is based on a silicon strip technology, made from 500 µm thick sensor with 512 strips and a pitch of 183 µm.

Magnet [26] The magnet separates the upstream (VELO, TT) and downstream (IT, OT) tracking stations and provides an integrated magnetic field 4 T m for 10 m long tracks. Is has a saddle shaped design weighing 1500 t, its polarity may be changed in order to minimize systematic effects of the overall detector construction. The field is required to reduce rapidly outside the z-extent of the dipole in order not to affect

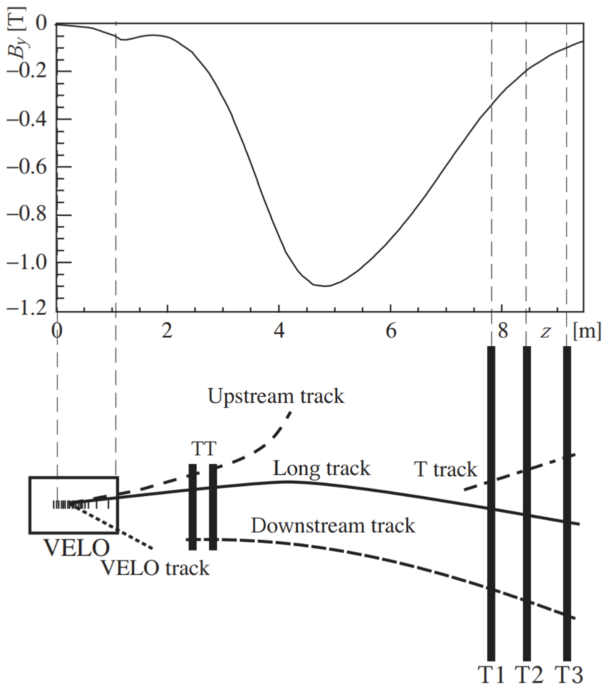


Figure 1.6: Intensity of the magnetic field for trackers. [25]

any sensitive electronics of the detector stations (notably the HPDs of the RICH), the visualisation of the field intensity is given in Fig. 1.6.

Inner Tracker (IT) [27] The IT is in the inner parts, around the beam-pipe, of the T1-T3 tracking stations downstream from the magnet. The detectors thus cover only part of the overall acceptance but the region with the highest particle fluence. It is also based on a silicon strip technology, however, using only 320 µm thick sensors with 384 strips and 198 µm pitch.

Outer Tracker (OT) [28] The last system is also located in the T1-T3 tracking stations, filling the rest of the experiment acceptance not covered by the IT. It is made of drift tubes with 4.9 mm diameter, filled with a mixture of Ar and CO_2 gas in a 7:3 ratio.

1.2.2.1 Upgrade overview

The upgrade [29] of the tracking system is comparable in extent to that of the VELO. In order to match the increased rate both due to radiation hardness and data throughput, while changing into a fully software triggered experiment, most of the hardware has to be renewed. Even though the technology and construction is completely new, the tracking modules are built in the same x-u-v-x pattern with 5° separations.

Upstream Tracker (UT) The detector stands as a replacement of the TT, utilizing the silicon strip technology. The silicon sensors are 250 µm thick with a nominal

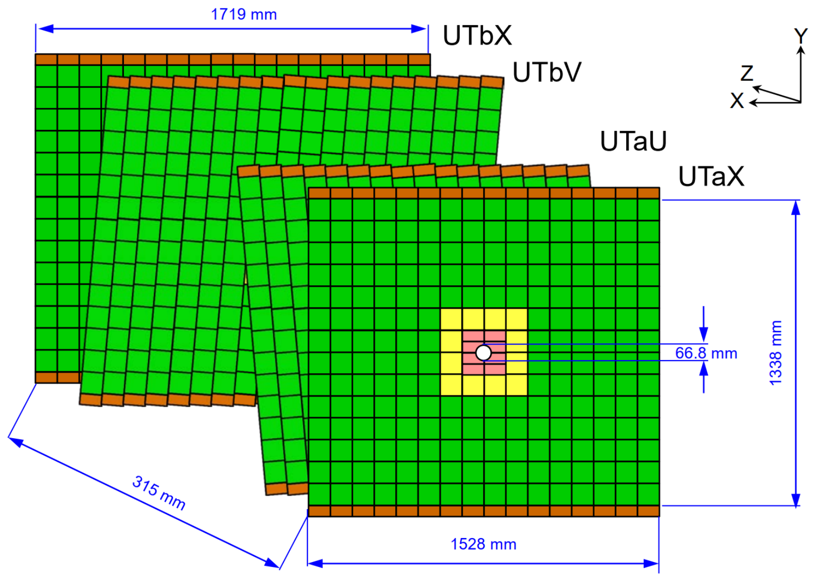


Figure 1.7: Geometry of the Upstream Tracker (UT) detector, showing the x-u-v-x pattern. The colour coding represent modification of the strip pitch and number of channels. [29]

pitch of 190 µm (halved to 95 µm in the region around the beam). The geometrical disposition is visualised in Fig. 1.7.

Scintillating Fibre Tracker (SciFi) It is built as a replacement of both IT and OT in the same geometry of 3 stations made of 4 layers, fully assembled from layered scintillating fibres with 270 µm diameter. The stacks are then read out using Silicon Photomultiplier (SiPM) with pixel pitch ≈60 µm, assembled into 250 µm thick channels. The multiple fired pixels along with the overlap of the channels and fibres provide for better position resolution via signal weighting.

1.2.3 RICH

The Ring Imaging CHerenkov (RICH) [30] system is used for the PID in the experiment, its primary goal is to separate pions and kaons and measure their momenta. The physics behind the measurement is Cherenkov radiation, where charged particles travelling in a dielectric medium faster than light cause an emission of radiation in a cone around the particle trajectory. The opening angle of the cone is proportional to the refractive index and the speed of the particle, thus the measurement directly provides information about the particle momentum. The dependency of the angle on the particle momentum for different radiators is in Fig. 1.8.

The system is constructed as RICH1 and RICH2 detectors placed upstream and downstream from the magnet, respectively. Both detectors are similar in construction, consisting of a radiative medium, combination of spherical and flat mirrors and photodetectors. Hybrid Photon Detectors (HPDs) are used as single photon sensitive detector, containing 1024 pixels each 500 µm^2 large and biased to 20 kV. They are sensitive to

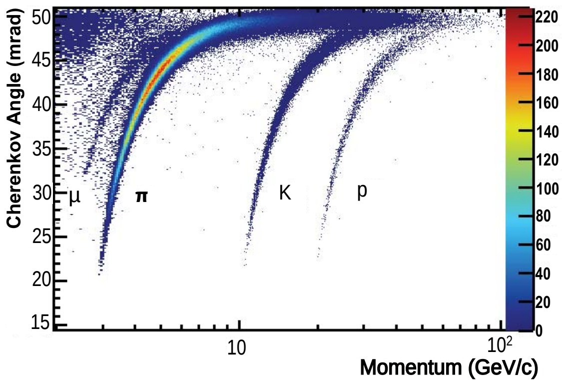


Figure 1.8: Reconstructed Cherenkov angle for isolated tracks based on the particle momentum in the C_4F_{10} radiator. [23]

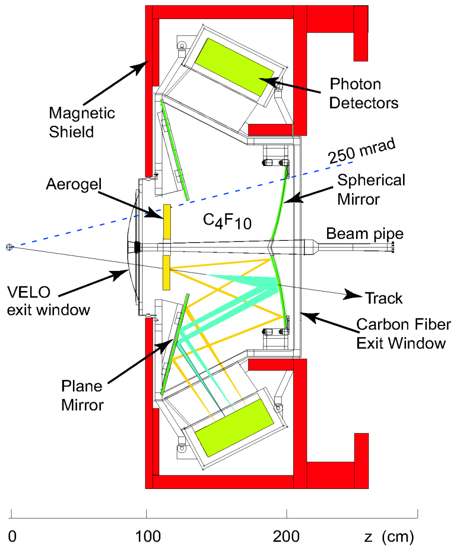


Figure 1.9: Schematic view of RICH1 detector. [15]

the wavelengths 200 nm to 600 nm. The whole system is placed in a MuMetal enclosure, minimizing the magnetic field near the sensitive components (see Fig. 1.9).

RICH1 The first station uses C_4F_{10} radiator, sensitive up to 60 GeV/c, respectively. The resolution of the detector is ≈1.6 mrad.

RICH2 The second station is filled with CF_4 gas, measuring particle momenta from 50 GeV/c to 100 GeV/c, complementary to the first station. The lower measurement limit is also higher since the low-momentum particles are swept out by the magnet; the detector achieves an overall resolution ≈0.68 mrad.

1.2.3.1 Upgrade overview

As part of the system upgrade [31], the general layout of the detector will not be changed, however, the RICH1 will have an increase to the optical length. This will increase the area of the sensitive detector plane and help with lowering the increased occupancy caused by the increased luminosity. Furthermore, the photodetectors will be replaced with Multi-anode Photo Multiplier Tubes (MaPMTs) with higher granularity than the HPDs and equipped with 40 MHz readout.

1.2.4 Calorimeters

The calorimeters [32] provide essential information for the PID of the whole system. The particles travelling through the detector deposit energy in the absorber generating showers of secondary particles, which in turn interact with the scintillators. Segmenting the volume provides information about the shape and the trajectory of the particle shower, the example is shown in Fig. 1.10. In the LHCb the light from the scintillators is gathered

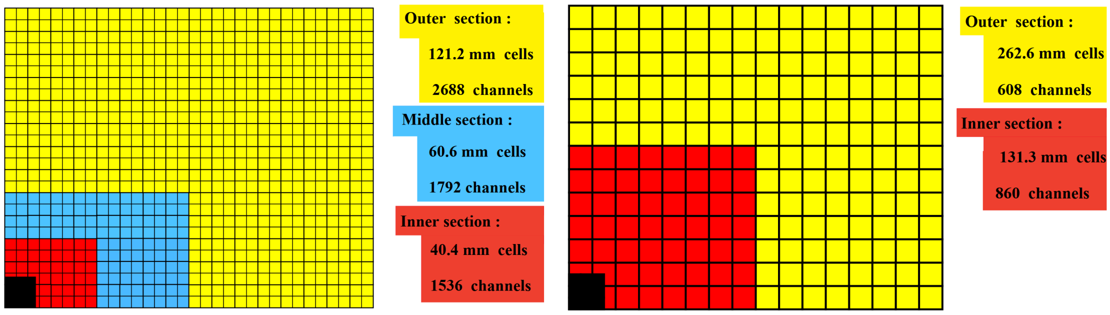


Figure 1.10: Segmentation of the quarter of the Electromagnetic Calorimeter (ECAL) and Hadron Calorimeter (HCAL) to readout channels. The black square represents clearance around the beam-pipe. [15]

using wavelength-shifting fibres which transfer the light to the Photo Multiplier Tubes (PMTs). Its gain is set based on the distance from the beam-pipe as well as on the type of the interaction (hadronic calorimeters have purposefully lower light yield).

Scintillator Pad Detector (SPD) & Pre-Shower (PS) The first layer of the system is built in order to better distinguish high momentum neutral and charged pions from electrons. The system is built with the SPD in front (detecting only charged particles), a 15 mm thin lead converter and then the PS detector. The sub-detector has 12032 channels, providing great trigger capabilities.

Electromagnetic Calorimeter (ECAL) This is made from stacked layers of 2 mm lead, 120 µm Tyvek (white paper-like reflective material) and 4 mm scintillating tiles. In total the calorimeter has 6016 channels, the overall thickness is 0.835 m (corresponding to 25 radiation lengths), designed to fully contain electromagnetic showers.

Hadron Calorimeter (HCAL) The hadronic calorimeter is built with detection planes parallel to the beam-pipe, made from alternating tiles of iron and scintillator every 20 cm, the total thickness being 1.65 m (corresponding to 5.6 interaction lengths). The iron layers are 16 mm thick in the lateral direction while the scintillators only 3 mm with a 120 µm layer of Tyvek, similarly to the ECAL. For the purpose of the event triggering, the resolution of the hadronic showers is not required to be as high as for the electromagnetic ones, the system has only 1468 channels and does not cover full interaction length.

1.2.4.1 Upgrade overview

During the LS2 the electronic readout is being upgraded to enable for the higher data throughput. Due to the removal of the hardware level trigger, the SPD/PS detector is removed. [31] In the end this provides the full event information with much greater flexibility due to software triggering needed namely in the fully hadronic decays.

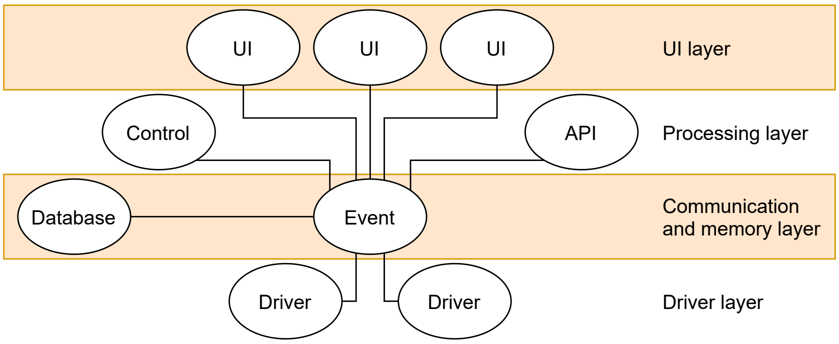


Figure 1.11: Simplified WinCC/OA manager schematic diagram.

1.2.5 Muon Chambers

The last detector in the LHCb is made for muon detection only [33]. It consists of 5 stations (M1-M5), where the first station is placed before the calorimeters and the rest completes the experiments. In order to provide an energy cut on the muons, the M2-M5 stations are separated by 80 cm of iron. The minimum momentum to pass the whole system is 6 GeV/c.

The system consists of 1380 chambers in total and has been designed in a way that each chamber has similar occupancy (the channel spacing increases according to the distance from the beam). The detection in the first layer is made by Gas Electron Multiplier (GEM) detectors, while the rest is done using Multiple Wire Proportional Chambers (MWPCs). A combination of Ar, CO_2 and CF_4 gases is used.

1.2.5.1 Upgrade overview

As part of the upgrade process, the first layer of the detector is removed and additional shielding is installed along the beam-pipe in order to reduce the occupancy. Apart from this the electronic components are upgraded similarly to the rest of the LHCb upgrade. [31]

1.2.6 WinCC/OA software

The Experiment Control System (ECS) and Data Acquisition (DAQ) for the whole LHCb is implemented using SIMATIC WinCC/OA [34], the same software as is used for the operation of many of the CERN experiments. A more detailed introduction to the software is given here as these tools were maintained and used during the module production testing described in Sect. 3.4.

WinCC/OA is developed by ETM (later acquired by Siemens) and is a type of a Supervisory Control And Data Acquisition (SCADA) building system which can be easily deployed for large projects. The main advantage of the system is in its reliability, scalability, openness and simplicity of automatized operations.

The architecture of the software is based on a set of managers, each taking care of different aspects of the operation. The communication is established using Transmission Control Protocol (TCP) where the individual managers (or whole systems) can span multiple machines. The system architecture is simplified in Fig. 1.11 and further described.

Event manager Fulfills the job of linking all managers (evaluation and distribution of events between them) thus can be only one per the system.

Database manager Provides options to archive and access historical system values in Raima or Oracle database.

UI manager Serves a purpose of both accessing and building the User Interface (UI) elements. Development is achieved from a large set of graphical widgets with modifiable scripts for all of their functionality (e.g. initialize, onClick, ...). The objects are then linked to the internal system values and provide both control and readback options.

Control manager Processes scripts defined outside of the UI scope, the functions may be user defined in C++.

API manager Enables linking of an externally compiled code into the whole system. All managers are built using the same Application Programming Interface (API) thus providing option for any customization of the WinCC/OA.

Driver manager Works as a middle layer between WinCC/OA and physical hardware. Responsible for any data smoothing, transformation and communication.

The system accessible values are defined as DataPoint Type (DPT), presenting a description of a complex type similar to structures in C/C++. The container may include any of the basic types (bool, float, string, ...) as well as more complex containers. The DataPoints (DPs) are then the actual instances of the DPT stored in the memory. The lowest level values are DataPoint Element (DPE) – apart from storing the value of the basic type, they can specify alarm ranges and other processing flags used for smoothing or storage. All of the definitions and creations of the values may be accessed through a *para* module.

WinCC/OA contains a *gedi* module, a tool for the graphical development and testing of the UI. The basic graphical unit is a panel – a program window that may be opened stand-alone or initialized within another panel. The DPE may be linked using their specified names (along with the names of the parent DP) to the UI widgets within the panels, both for visualisation as well as control of the values. In order to simplify the development of a large and complex system where the same objects are used repetitively (e.g. PSU channels, temperature probes, ...), so-called delta-parameters may be used by individual panels. The UI elements are built for the single low-level object with defined control or read-back functionality of the DPEs and the parent DP accessed as a variable delta-parameter. The final system then may combine multiple instances of the same object into a single panel, assigning the name of the DP to the delta-parameter. Furthermore, the variable parameters may be used to specify any other DPs and widgets in the panels.

In order to provide better monitoring, WinCC/OA contains alarm handling features. As mentioned above, the alarm limits are defined as part of every single DPE. The reason why it is not part of the DPT definition is to provide individual alarm settings for the same object (the alarm range on the same PSU channel might depend on its use). The ranges may be defined to trigger upper/lower limit for both warnings and errors. Since some values may oscillate around the alarm, a hysteresis may be defined to reduce the amount of triggers. The general handling also provides an option to acknowledge the alarms so that the occurrence is always noted by the operator.

Additional frameworks have been developed for the usage by the CERN groups, simplifying the operation of the frequently performed tasks. One such framework is being developed by the CERN Joint Controls Project (JCOP) team, containing number of tools in a form of installable components.

1.2.6.1 JCOP framework

JCOP framework serves as a template for easier development of the ECS for all experiments as their needs are almost identical. Some examples include quick generation and modification of large amount of devices of the same type, alarm handling panels, trending plots of monitored values or even panels for commonly used devices (such as CAEN boards). A noteworthy component is Finite State Machine (FSM), providing control and monitoring tools for a complex system operated by shifters.

FSM The main goal of the FSM it to provide an abstraction layer for the ECS. The objects are structured in a tree-like scheme, propagating commands downstream and states or alarms upstream. This ensures simple configuration of the whole system from the top while being able to quickly go down and find the error of a single device or channel. For safety reasons users and groups may be defined in order to restrict control of some settings. Furthermore, the system may be partitioned to enable expert intervention or detach the command/state propagation in either direction.

The FSM is built stand-alone, linked via the API to the WinCC/OA managers. It is using a custom State Machine Interface (SMI++) configuration language for both communication between the layers as well as internal decision making. The structure of the FSM is similar to the WinCC/OA objects using DPT and DPE – Control Units (CUs) or Logical Units (LUs) define the object form and set of Device Units (DUs) represents the physical hardware.

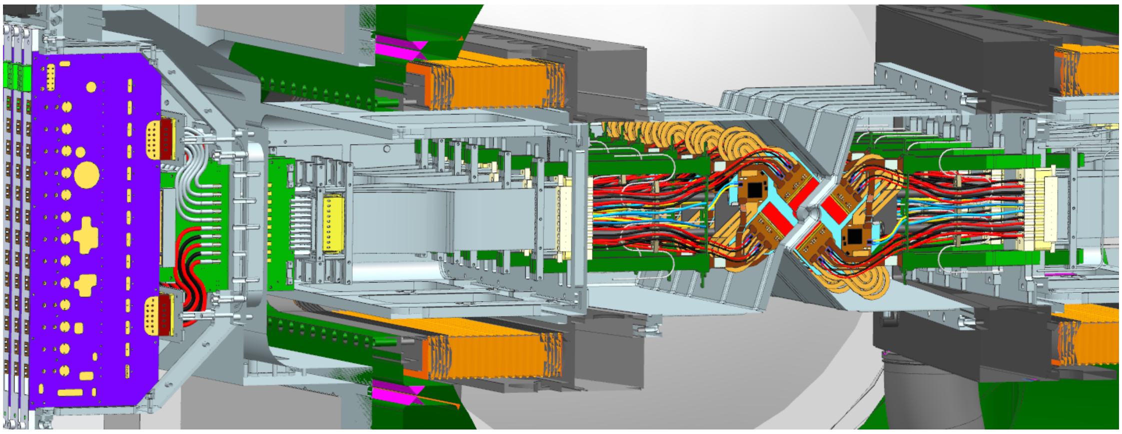


Figure 1.12: VELO Upgrade detector visualisation. The main components being: sensors (red), microchannels (blue), RF foil (grey), OPB (purple). [35]

1.3 VELO Upgrade

As main body of this work represents various aspects of the construction of the VELO Upgrade subsystem [24], a more in depth description of its components is given. In total, the subsystem consists of 52 modules positioned in a secondary vacuum inside the LHC beam-pipe, surrounding the interaction point. For the best tracking performance the smallest possible amount of material has to be used in constructing all components.

The visualisation of the whole detector is in Fig. 1.12. The subsystem can be split into three main parts discussed in more detail - mechanics, cooling and electronics.

1.3.1 Mechanics

Since the VELO upgrade is similar in construction to the old VELO some of the old structural parts may be reused, however, most are refurbished or replaced.

The detector is built in a large vacuum tank, containing two bases with modules, two RF foils, a bellows system, wake-field suppressor and all required cables as well as cooling pipes. On the downstream side, the tank is closed with a thin aluminium window, minimising the amount of material used. Two hoods are used to seal the access ports with the halves inserted, containing Vacuum Feed-through Board (VFB) for each of the modules as well as the Opto-and-Power Boards (OPBs) for the rest of the data and power transmission. [24] The total material budget of the upgraded system is in Fig. 1.13. Again as a precaution during the operation, the VELO halves will start retracted from the beam with the maximal distance ≈20 mm. This requires the system to be able to move in the x direction on a daily basis as the halves are inserted only during the stable beams conditions. Adjustments in the y can also be made, needed if the beam position has moved.

The major changes with respect to the former VELO include a reduced distance from the silicon sensor to the beam (from 8.2 mm to 5.1 mm). Furthermore the readout

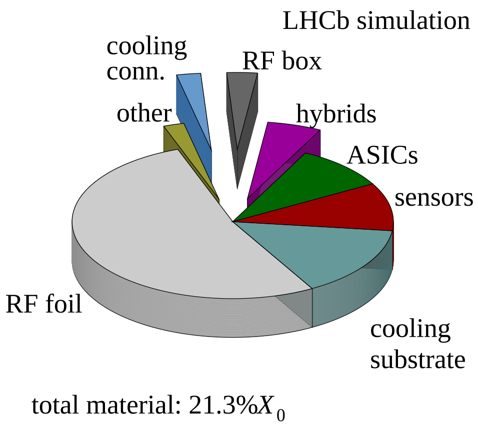


Figure 1.13: Material budget of the VELO Upgrade. [24]

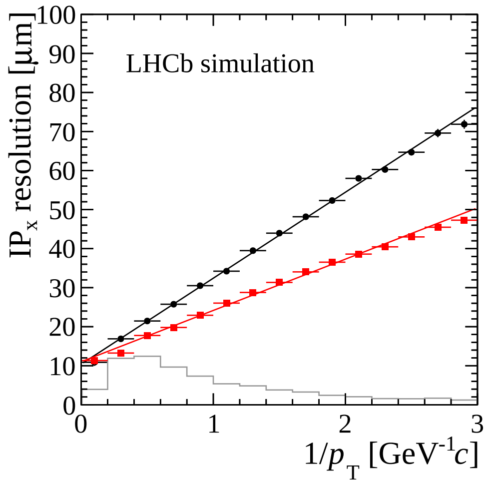


Figure 1.14: Simulated IP $_x$ resolution of VELO and VELO Upgrade as a function of inverse momentum. The black and red points correspond to the old and upgrade detector, respectively. [24]

geometry change from strip to pixel detector requires modifications of the RF foil and a new way to get all connections out of the vacuum.

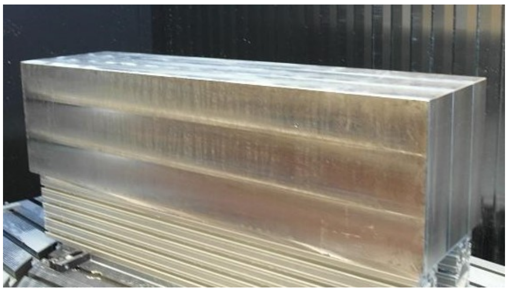
1.3.1.1 RF foil

In order to operate the detector in a vacuum close to the interaction point without any possible contamination of the primary beam-pipe vacuum, a thin aluminium alloy RF foil is used for separation as well as protection against beam-induced currents. The requirements on the foil due to higher luminosity and smaller distance to the beam are following: [24]:

- achieve minimal thickness (less than 250 µm),
- radiation hard for up to 1000 MRad,
- withstand a pressure difference between the two volumes of less than 10 mbar,
- tight tolerances to enable movement of the halves and provide for module insertion as close to the beam as possible,
- dimensional stability – less than 200 µm variation in 1 m of material with a temperature difference of 10 °C.

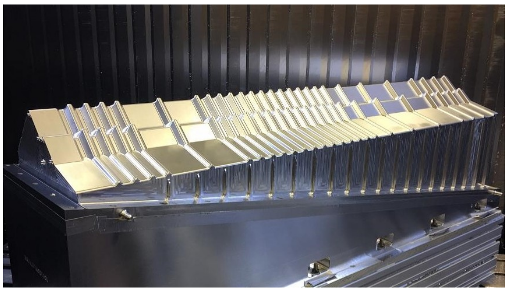
Furthermore, the foil has to not only withstand the evacuation but also filling of the beam-pipe with nitrogen at the atmospheric pressure.

The drive to achieve the minimal thickness comes from the IP resolution. The comparison of the old and upgrade system performance from the simulation is in Fig. 1.14 where the slope is driven by the material in the RF foil.

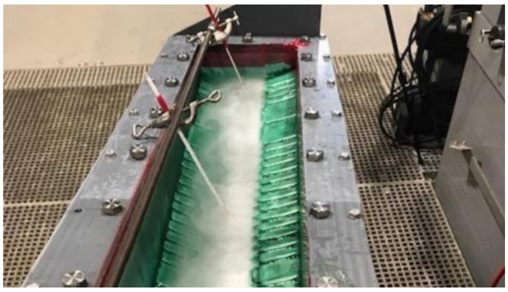


(a) : Starting aluminium alloy block.

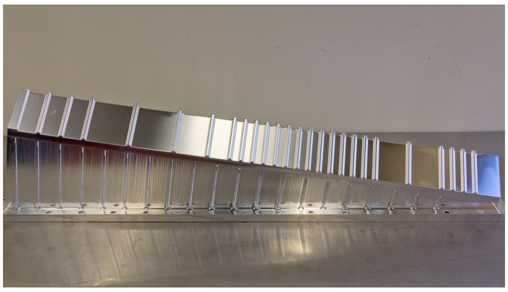
(b) : Milling of the inner part.



(c) : Support mold for milling the outer part.



(d) : Etching.



(e) : Final foil.

Figure 1.15: Steps of the RF foil production at Nikhef. [35]

The RF foil is machined in Nikhef from a single block of aluminium alloy ($AlMg_3$), milled down to 500 µm thickness. In order to minimize the amount of material between the beam and the sensitive parts of the detector, the final design is then further milled to 250 µm near the beam and etched to 150 µm in the innermost parts. Finally the beam-facing part of the foil is Non-Evaporable Getter (NEG) coated, activated at the temperature 160 °C. The production steps are visualised in Fig. 1.15.

1.3.1.2 Vacuum-Feedthrough Board

The transition of the data, Low Voltage (LV) and HV is achieved with a custom-made VFB made from a 12 layered 2 mm thick PCB. The feed-through has connectors soldered both on the in air and vacuum sides, enabling a quick connection of the OPB and module, respectively.

The VFB is built by inserting the PCB into two half-moon holders, making a connection between the PCB and the rest of the vacuum flange. Araldite 2011 is used to glue everything in place and left for curing. Afterwards both sides are one by one potted using Araldite 2020 making the whole object leak-tight.

1.3.2 Cooling

Much of the upgraded cooling is based on the concept of CO_2 cooling, which has been widely adopted at the LHC in upgrades, thus relying on an already tested and performant system, however, the mechanical and thermal cooling support of the modules is an innovative development. The main requirements on the cooling of new modules come from the increased heat generation of the readout electronics and smaller distance to the beam, defined as follows: [24]

- reliably dissipate up to 43 W,
- provide a safe margin on thermal run-away (caused by the linked increase of the sensor leakage current and its temperature), achieved by maintaining −20 °C for the silicon sensor using coolant of around −30 °C,
- contain the minimal amount of material,
- avoid any leaks, in case of rupture in one of the modules keep the differential pressure below 10 mbar across the RF foil,
- minimize mechanical distortions due to thermal expansion of the components.

Following these requirements, a number of options were evaluated. The solutions considered were silicon microchannel substrates, 3D printed titanium cooling channels and ceramic substrates with embedded channels. [36]

The silicon solution was adopted as it was the most advanced in development both in R&D and LHCb and building on work for the NA62 GigaTracker [37]. The final design is shown in Fig. 1.16. The high thermal conductivity of silicon provides very good heat transfer between the coolant and the rest of the system, while matching the Coefficient of Thermal Expansion (CTE) of the readout electronics. This solution is radiation hard and also minimizes the amount of material present close to the beam.

The utilised cooling method is based on evaporative CO_2 , which is an isothermal process keeping mostly constant temperature across the module regardless of the heat-load (unless all of the CO_2 evaporates before reaching the outlet – so called dry-out). The flow can be easily regulated by adjusting the pressure. [39] An important aspect that had to be studied extensively was pressure tolerance, since the standard operation expects to reach 20 bar to 30 bar at −30 °C.

1.3.2.1 Performance

To validate the performance of the silicon substrates, a number of simulations, and extensive prototype and development work had to be done. [24] This included tests using heat sources mimicking beam conditions as well as simulations of heat transfer from the readout electronics used for the final channel size and spacing. The most important for

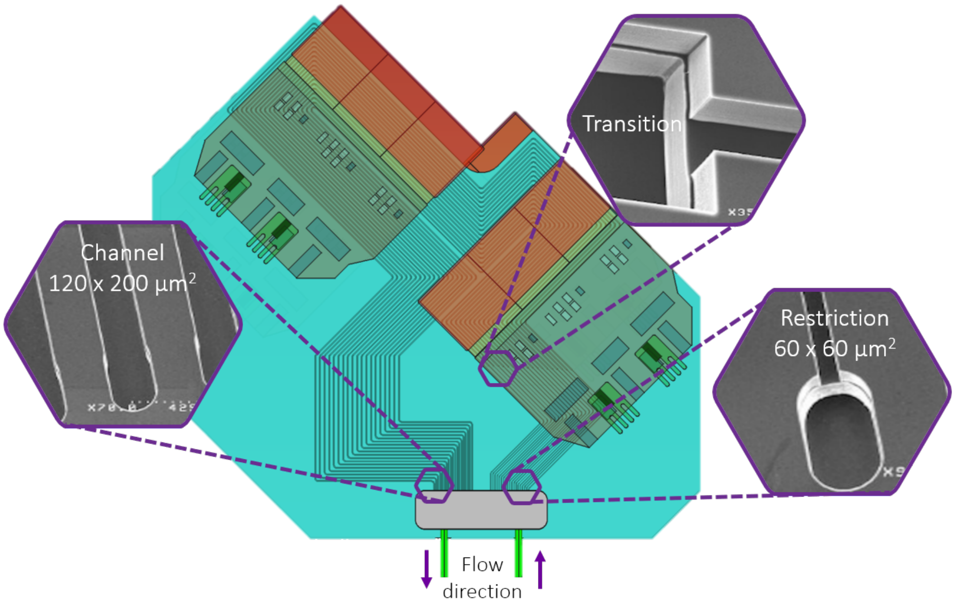


Figure 1.16: Final design of the silicon cooling substrate with visualised position of the channels and front-end readout electronics. [38] The embedded images show detailed photographs of various channel regions provided by ©CEA.

the production were the pressure testing and development of the connector soldering process.

In the nominal operation conditions, the VeloPix ASIC generate up to 3 W of heat (26 W per module), with additional 5 W per GigaBit Transceiver ASIC (GBTx). Since some of the sensitive regions are overhanging the cooling substrate, an important aspect that had to be considered is the maximal ΔT between the coolant and the hottest tip of the chip. Provided that the cooling plant can achieve CO_2 temperature of $-30\,°C$, the expected ΔT is about $7\,°C$, thus keeping the readout above $-20\,°C$. [24]

1.3.2.2 Production

The silicon substrates have total thickness of 500 µm, containing 19 channels positioned based on the location of the readout electronics. The CO_2 boiling condition is achieved by firstly having the channels restricted to a cross section of 60 µm · 60 µm, later expanded to 120 µm · 200 µm in order to cause a pressure drop and trigger the evaporative condition. [39]

The production process is based on a well known technology of Micro-Electro-Mechanical Systems (MEMS), relying on the standard processes of lithography and etching. However, the process is complex and the 69 % yield of the delivered wafers was lower than originally foreseen, mostly if comparing with different production runs with smaller components (reaching 85 % yields) [35].

The production follows a number of steps, starting by Deep Reactive Ion Etching (DRIE) 120 µm deep trenches in a 260 µm thick silicon wafer, which is then covered by another wafer. The connection of the two wafers is achieved by hydrophilic bonding. The

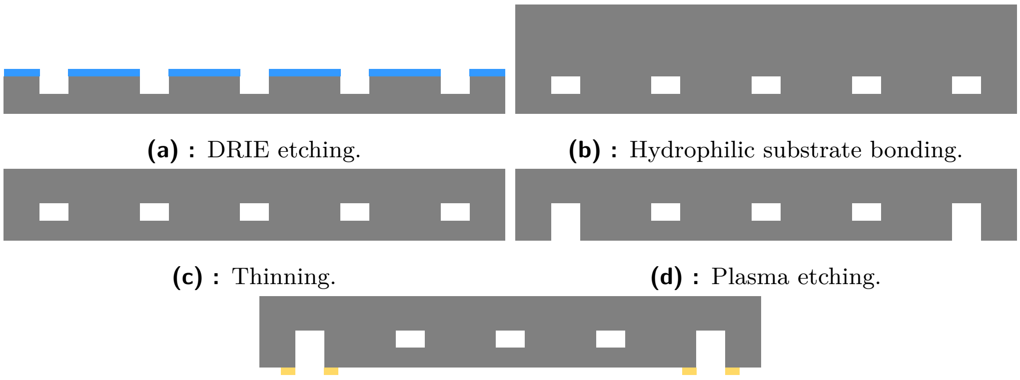


Figure 1.17: Steps of the silicon microchannel substrates production. Silicon (gray), mask (blue), metallisation (gold).

whole substrate is then thinned down to the final 500 µm.

Afterwards the holes for the fluidic connector are plasma etched into the substrate. Both sides (silicon and the connector), are metallized with layers of nickel and gold with an additional titanium layer on the substrate (in order to prevent diffusion of metals). The two parts are then aligned and soldered together using a 55 µm PbSn solder foil. [39] The production steps of the substrate are visualised in Fig. 1.17.

1.3.2.3 Safety

As mentioned above, the nominal operational pressure of the microchannels is 20 bar to 30 bar, however, at 22 °C it might reach 60 bar. Applying a standard safety factor of 2.5 yields 150 bar, the final safety requirement matching that of the old VELO at 170 bar. [39] The high ratings are not only important due to the safety of the modules themselves but also to prevent damage to the whole VELO and LHC.

Pressure testing of the used substrate design resulted in a maximum pressure at 400 bar, well above the defined limits. This, however, does not include the soldered fluidic connection, which should withstand more than 700 bar. The prototype samples were tested over a thousands of endurance cycles (each lasting 800 s at 12 bar) and several thousand pressure cycles (from 1 bar to 180 bar) without any visible degradation. [24]

All produced microchannels are pressure tested to 180 bar before being used for the module assembly. During the operation issues could arise if any particulates were present in the system (channel restrictions having the smallest cross-section), which will be mitigated by filtering of the used CO_2 and pressure monitoring. Furthermore the final cooling system will contain pressure-safety points set to 100 bar [24] in order to prevent any damage.

1.3.3 Electronics

The increase of the data rates coming both from the increased luminosity as well as from positioning the active channels 3 mm closer to the beam required a new approach to the

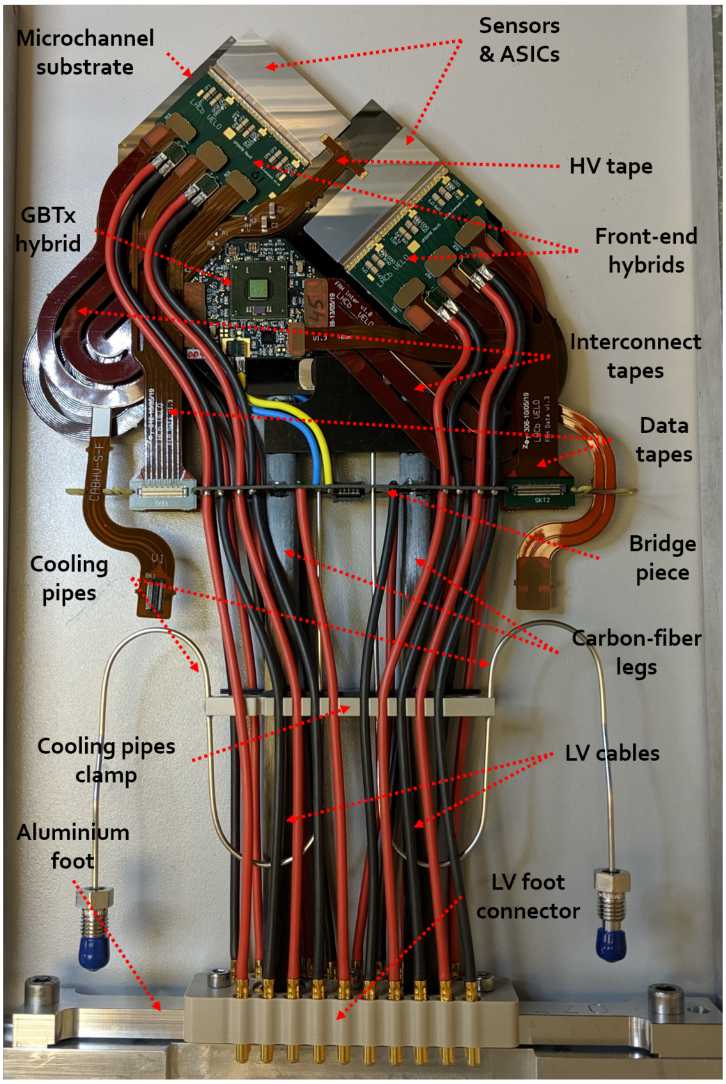


Figure 1.18: Fully assembled VELO module.

readout. In order to achieve this a technology has been changed – relying on hybrid pixel detectors instead of strip based ones. To keep the same coverage, the individual modules are populated with 4 such detectors, configured in an L-shape. In total the whole VELO upgrade consists of ≈41 Mpixel (compared to ≈0.18 Mstrips used in the old system). A VeloPix ASIC has been developed, providing zero-suppressed trigger-less readout with much higher data throughput [40].

The individual components of the readout chain will be further discussed, the basic building block being a module (shown in Fig. 1.18), its assembly and testing later described in details in Chapter 3.

1.3.3.1 Sensors

The sensors used have to match the pixel size of the read-out electronics, in this case $55\,\mu m \cdot 55\,\mu m$. Since the individual sensors are attached to 3 ASICs in line, the pixels at the inner sides have to be elongated to $55\,\mu m \cdot 137\,\mu m$ in order to minimize the dead

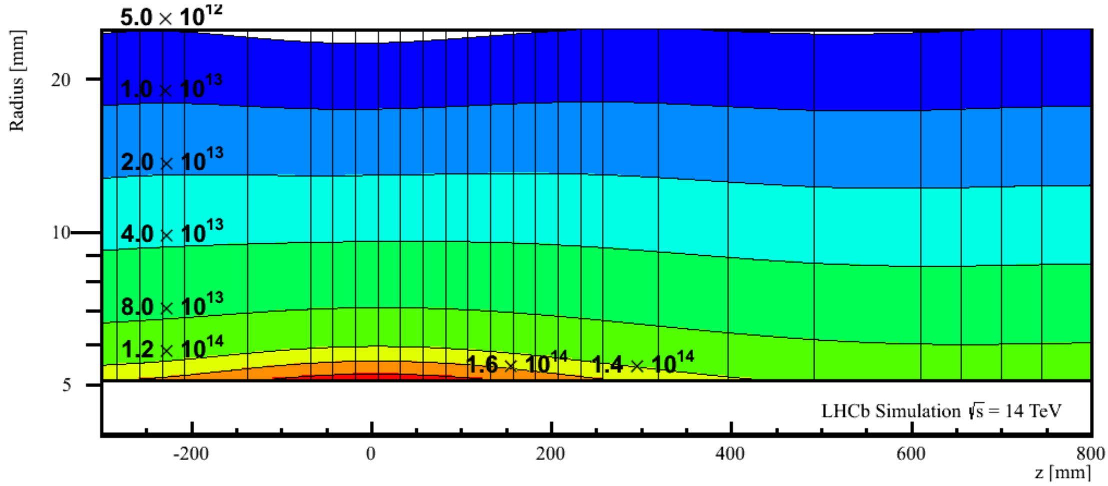


Figure 1.19: Simulated radiation map as a function of position and radial distance from the beam. The colours represent units of $n_{eq}cm^{-2}$ for integrated luminosity $L = 1\,fb^{-1}$ [24]. The expected value by the end of Run4 is $L = 50\,fb^{-1}$.

area. The ASICs are attached to the sensors using a flip-chip bonding technique, each pixel on both the sensor and the ASIC is linked using a solder bump-bond.

Furthermore, the sensors have to withstand the fluence expected at the end of the LHC Run4 ($8 \cdot 10^{15}$ 1, the radiation map along the beam axis shown in Fig. 1.19) while maintaining the best performance.

A number of studies have been made to develop and test such sensors, the chosen ones being 200 µm thick n-on-p silicon planar sensors made by Hamamatsu. They provide detection efficiency higher than 99 % and can be operated up to 1000 V after the irradiation. [41] In order to mitigate the time-walk effect (the time-of-arrival depends on the deposited signal) the sensors utilize e^-collection.

1.3.3.2 VeloPix ASICs

The utilized VeloPix ASIC is based on a Timepix3 design, the comparison of the two summarized in Table 1.1. Triple modular redundancy has been used to protect the digital components from the single event effects [42].

The main advantage of the VeloPix is in its high data throughput, sending out only the signal from the pixels which registered a particle interaction. This is a so so-called zero-suppressed information, where only the active pixels are read-out. The analogue signal is at first amplified, compared to a global threshold level and digitized in the pixel architecture. In order to achieve the high throughput, the $256 \cdot 256$ pixel matrix is then read out as a Super Pixel Packet (SPP). An SPP is a $2 \cdot 4$ pixel large cluster from where the data are transmitted to the End of Column block (EoC) forming a 30 bits large packets which are sent out of the ASIC. The schematic of a single pixel architecture is in Fig. 1.20.

The data leaving the VeloPix at 40 MHz are combined into a 128 bits packets consisting of a group of four SPP along with a header and parity bits, scrambled to achieve better

Table 1.1: Comparison of the VeloPix and Timepix3 ASICs, based on [40]

ASIC	Timepix3	VeloPix
Technology	130 nm CMOS	130 nm CMOS
Power consumption	<1 W cm^{-2}	<1.5 W cm^{-2}
Charge collection	e^{-}, h^{+}	e^{-}
Pixels	256x256	256x256
Pixel size	55 µm · 55 µm	55 µm · 55 µm
Timing resolution	1.5625 ns	25 ns
Readout	Triggerless, ToT	Triggerless, binary
Peak hit rate ASIC	80 Mhits/s	900 Mhits/s
Peak hit rate pixel	1.2 khits/s	50 khits/s
Total bit rate	5.12 Gbit s^{-1}	20.48 Gbit s^{-1}
Radiation hardness	-	400 Mrad

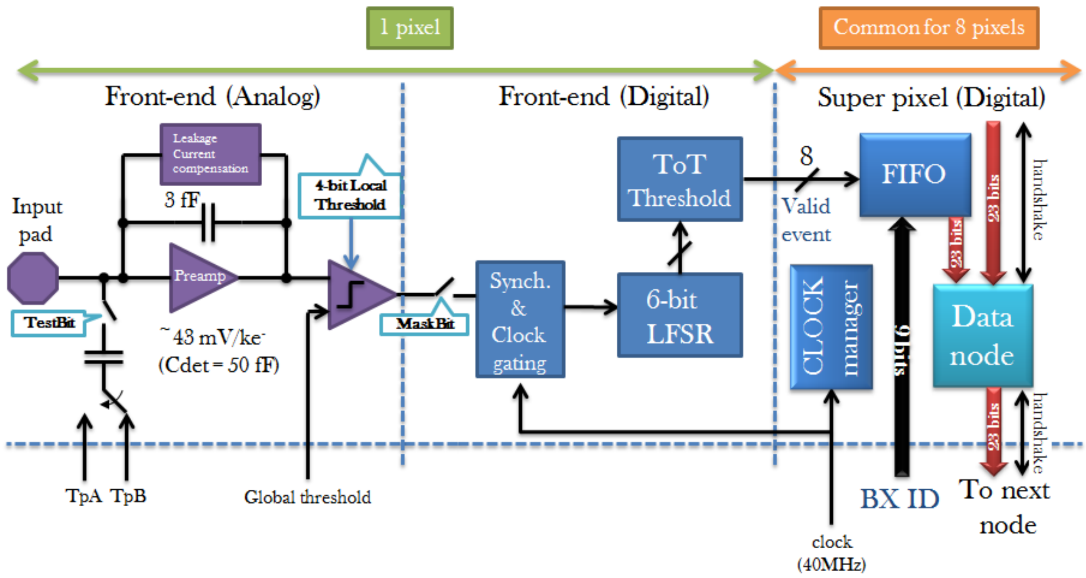


Figure 1.20: Schematic of the pixel architecture. [40]

data transmission [40, 43]. For the best performance, a dedicated Gigabit Wireline Transmitter (GWT) protocol has been designed [44] instead of using the standard GigaBit Transceiver (GBT) [45]. Apart from this, the ASICs are also connected to the ECS slow control interface, providing monitoring and configuration options.

The VeloPix frames are not only used for the pixel data readout but also for other control:

Data Each of the four SPP contain 9 bits Bunch Crossing ID (BXID), 13 bits Super Pixel (SP) address and 8 bits hitmap.

Since the collisions are at 40 MHz rate, the hits that are registered at the opposite end of the EoC may arrive later than the ones near it. For this reason the ASICs have to keep track of the BXID which is later used to assign the hits to the correct events. In total 12 bits would be needed contain the full information (2808 bunch crossings of the LHC), however, in order to optimize the data bandwidth only 9 bits

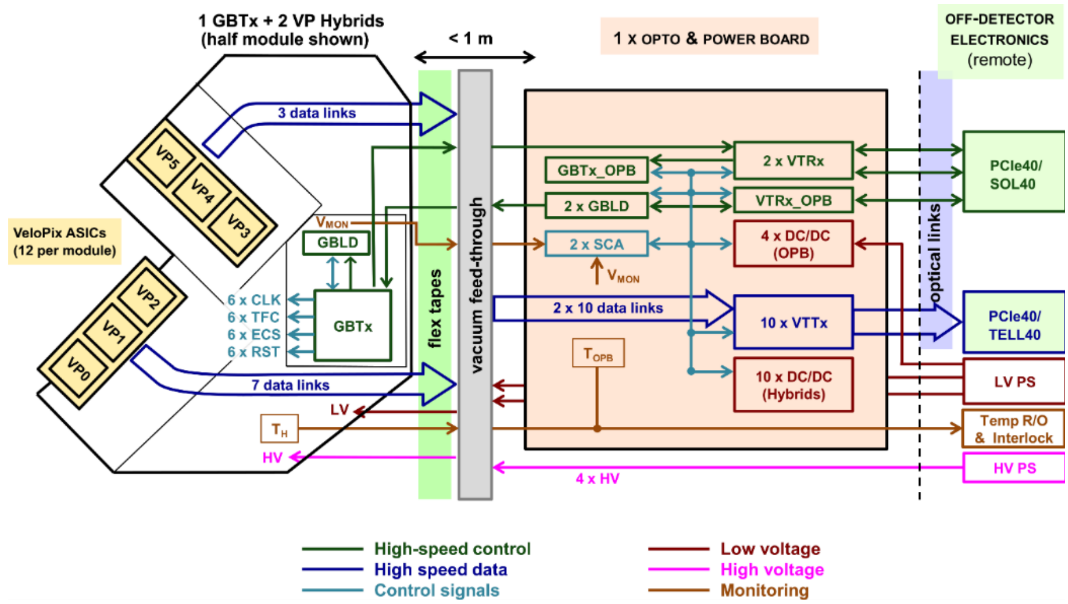


Figure 1.21: Schematic of module readout. The visualisation is made only for one side of the module (effectively halving the number of links and ASICs). [46]

are used, potentially grey-coded [43].

The SP address is split into 7 bits and 6 bits, representing the columm (128 options) and row (64 options) positions, respectively.

The hitmap information provides the precise mask of the hits for each individual SP. The whole approach effectively reduces the amount of data transferred since clusters (hits in neighbouring pixels) are expected from the tracks registered by the VELO.

Special frames The purpose of additional special frames is to provide synchronization and identification capabilities. They are easily distinguished from the data packets, as the individual SP contain a dedicated 4 bits header and the hitmap information is always empty.

There are 5 special frames in total: BXID, Timing and Fast Control (TFC) Sync, Chip ID, TFC Align Mode and Invalid (Idle); all distinguished by the header. The commands are described in more detail in the module testing Sect. 3.4.2.

1.3.3.3 On detector

The VELO Upgrade modules are based on a microchannel cooling plate with four tiles (made of a sensor bump-bonded to three VeloPix ASICs) attached in an L shape, two on each side. The naming of the tiles is a three-letter acronym: first letter represents the microchannel side (**C**ooling/**N**on-cooling); second letter corresponds to the microchannel edge (**L**ong/**S**hort); the last letter defines the distance from the beam (**I**nner/**O**uter). The four used combinations are: CLI, CSO, NSI, NLO.

In order to protect sensitive electronics, dissipate heat and minimize the amount of material, only the DAQ and control ASICs are attached to the module and operated in the vacuum. The signals are then transmitted using sets of data tapes, converted to optical signal and forwarded to off-detector electronics. The electronics overview is in Fig. 1.21.

Control ASICs In order to configure and synchronize the VeloPix ASICs, each side of the module contains a GBTx and GigaBit Laser Driver (GBLD).

The GBTx decodes and distributes the clock and other control signals for the VeloPixes. It also monitors the configuration and voltages of the ASICs (read out by GBT Slow Control Adapter (GBT-SCA) [47] on the Front-End (FE) hybrids). The monitoring data are sent upstream via the GBLD in order to minimize losses over the distance.

Data tapes The signal transmission is done using four thin flexible PCB tapes 50 cm to 70 cm long, each containing up to seven differential links. The dielectric material was chosen as Pyralux AP-Plus due to its excellent radiation hardness and low dissipation of high frequency signals [24, 48].

Since the expected data rate is higher in the ASICs closer to the beam (estimated from the simulation in Fig. 1.22), four GWT links are assigned to the inner ASICs, two to the middle ones and only one on the outer side. Each link should provide a data rate of 5 Gbit/s, in total 20 links are used for the read out of the full double-sided module.

The tapes are connected using MOLEX Slimstack connectors to the FE hybrid PCBs (wire-bonded to the ASICs) and to the VFB where the signal gets transferred from the module to the OPB. HV for biasing is delivered similarly through thin HV tapes, bonded directly on top of the sensors, operated up to a maximum of 1000 V.

OPB The OPB is responsible for the following: electrical to optical signal conversion, power delivery and monitoring. There is an additional GBTx with GBT-SCA on the board itself.

The 20 data links coming from the VeloPix ASICs are converted in 10 Versatile Twin Transmitter (VTTx) modules (each with two sockets) and sent via optical fibres to the off detector electronics. The GBTx data are converted in a similar manner, however, in total they use 3 Versatile Transceiver (VTRx) modules which provide communication both up- and down-stream. The control data sent to the GBTx in the module are again sent via the GBLD.

The LV power delivery is achieved using radiation hard DC/DC converters [49]. In total ten converters at 1.3 V are used for the module and four at 2.5 V for the board.

1.3.3.4 Off detector

The LHCb upgrade Back-End (BE) is built as an interface between the FE modules, DAQ, TFC and ECS systems. The ≈300 m long optical fibres transfer the zero-suppressed

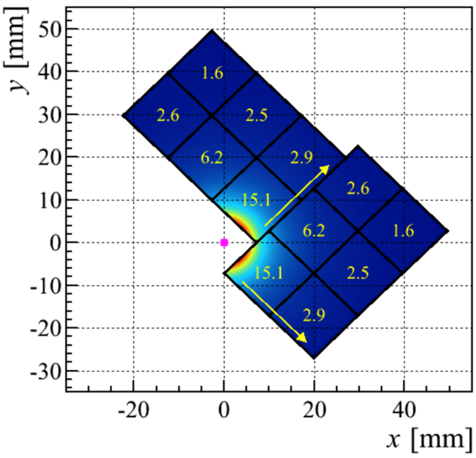


Figure 1.22: Simulation of the expected data rates (Gbit/s) in the VeloPix ASICs. [43]

data from the readout modules to the off detector electronics in the counting room for further processing. [50]

The whole system is managed by readout supervisor S-ODIN boards which control interface (SOL40) and readout (TELL40) boards. The S-ODIN boards also provide an interface for the TFC synchronization with the LHC. [43]

Both SOL40 and TELL40 are based on an PCIe40 board, namely an Intel Arria 10 Field Programmable Gate Array (FPGA). [51] Each board contains 4 transceiver and 4 receiver Multi-fibre Push On (MPO) optical connections, in total providing 48 links in both directions. The optical signal is converted back to an electrical one where each link is connected to an FPGA high speed input/output. The firmware of the SOL40 handles the TFC communication and general slow control system, while the TELL40 is responsible for the DAQ.

1.4 Summary

CERN with its accelerator complex provides excellent HEP research conditions for a number of international collaborations. Its largest particle accelerator, the LHC, primarily provides proton-proton collisions at the frequency 40 MHz, the products of which are detected by four main experiments. LHCb is one of them with a research focused on the CP violation studies and b and c hadron decays, and is currently undertaking a planned upgrade. This required extensive R&D of different components, the presented work focused on the upgrade of VELO detector subsystem. The new tracker will have a vastly increased number of channels, all with a trigger-less readout, while will still be operated in a vacuum cooled down using evaporative liquid CO_2 .

Chapter 2

Semiconductor detectors

Based on their type and properties, the particles traversing matter interact and lose energy. This can be either via electromagnetic or nuclear interactions, however, the principle of the tracking detectors for HEP experiments is based on the former. The result is generation of charge carriers which are collected on the attached electrodes and produce an electrical signal. First the basic principles of the interaction will be described, followed by an in depth description of physics of the semiconductors, their usage as detectors and damaging effects of the radiation.

2.1 Interaction of radiation with matter

As mentioned above, the radiation can interact either with the nuclei or electromagnetically, mostly with the electrons in the medium. Based on this, the particles may be split to three groups

- charged particles,
- photons,
- neutral hadrons.

The energies of the particles coming from the LHC collision point are very high compared with their mass. Because of this, the hadrons and leptons have a very low cross-section for the nuclear interaction while for the high-energy photons the dominant process is an electron-positron pair production. This means that the main means of energy deposition is via Coulomb interactions between the charged particles and the atomic electrons.

2.1.1 Energy loss of charged particles

As particles transverse the material a portion of their energy is lost until they are completely stopped. The mean energy loss (also called stopping power) of charged particles is defined as energy lost per surface density and depends on the relative momentum of the

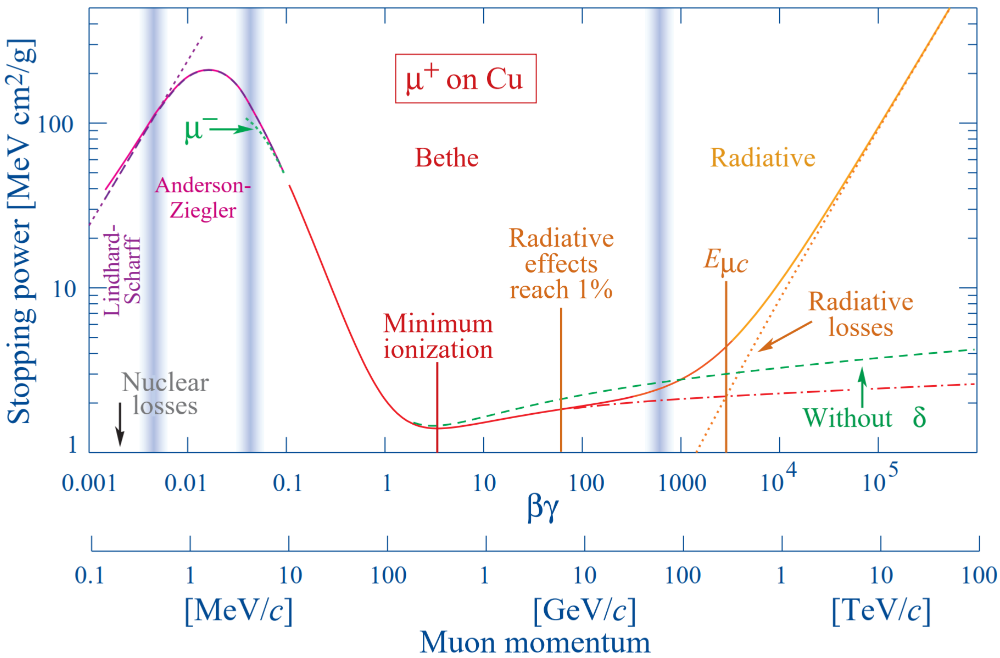


Figure 2.1: Mean energy loss of a positive muon in copper as a function of momentum. The graph is horizontally split to regions based on the approximations of the interactions. [52]

particles to the absorber medium. An example of such dependency is shown in Fig. 2.1 for positive muons in copper.

The dependency of the mean energy loss is best described in a form of $\beta\gamma$ factor, where β represents the particle velocity relative to the speed of light c and γ is a Lorentz factor defined as $1/\sqrt{(1-\beta^2)}$. Based on this, the interactions may be split according to the increasing energy of the particle to low energy region, Bethe region and radiative region. It is important to note that the dependency is valid for both electrons/positrons and heavier charged particles, however, additional effects have to be considered in the former case due to their matching masses with the atomic electrons.

The mean energy loss $\left\langle \frac{\mathrm{d}E}{\mathrm{d}x} \right\rangle$ in the range of $1 \cdot 10^{-1} < \beta\gamma < 1 \cdot 10^{3}$ is described by the Bethe-Bloch formula [53] as

$$-\left\langle \frac{\mathrm{d}E}{\mathrm{d}x} \right\rangle = \frac{K}{\beta^2} z^2 \frac{Z}{A} \left[\frac{1}{2} \ln \left(\frac{2 m_{\mathrm{e}} c^2 \beta^2 \gamma^2 T_{\mathrm{max}}}{I^2} \right) - \beta^2 - \frac{\delta(\beta\gamma)}{2} \right], \tag{2.1}$$

with the individual variables described in Table 2.1.

The maximal energy transfer for a particle of mass M from a head-on collision is calculated as [54]

$$T_{\mathrm{max}} = \frac{2 m_{\mathrm{e}} c^2 \beta^2 \gamma^2}{1 + 2\gamma \frac{m_{\mathrm{e}}}{M} + \left(\frac{m_{\mathrm{e}}}{M} \right)^2}. \tag{2.2}$$

For small energies the particles follow a decrease of the stopping power with β^{-2} until reaching the minimum at around $\beta\gamma \approx 3$ (corresponding to about $0.94c$). Afterwards, the logarithmic part $\ln(\beta\gamma)$ slowly increases the value. Particles in this valley with the

Table 2.1: Bethe-Bloch formula variables.

K	$4\pi N_\mathrm{A} r_\mathrm{e}^2 m_\mathrm{e} c^2$
N_A	Avogadro's number
r_e	electron radius
$m_\mathrm{e}c^2$	electron rest mass
c	speed of light
z	particle charge
A	atomic mass of absorber
Z	atomic number of absorber
β	particle velocity (as v/c)
γ	Lorentz factor $1/\sqrt{(1-\beta^2)}$
T_max	maximal energy transfer in a single collision
I	mean excitation energy
$\delta(\beta\gamma)$	density correction (for high energy particles)

minimal energy loss are called Minimum Ionizing Particles (MIPs) and deposit the least amount of the energy in the detector. While the Bethe region described kinematic losses which caused excitation or ionization in the atomic electrons, the high energy region is then mostly dominated by radiative losses such as Bremsstrahlung. The stopping power caused by the collisions only may be also called Linear Energy Transfer (LET).

2.1.2 Carrier generation

The energy deposited in the material may result in an excitation of atoms which then conduct electric current. The average number of such carriers, discussed in depth for semiconductors in Sect. 2.2, corresponds to the ionization energy I needed by the absorber material, simplified as

$$N_\mathrm{pairs} = \frac{E}{I}, \tag{2.3}$$

where E is the total deposited energy.

Given thick enough material, the mean energy deposition of a MIP will follow the Bethe-Bloch formula, however, the individual interactions are statistical in nature. The shape of the probability distribution is affected by the generation of δ-electrons, a type of radiation created by large energy transfers in the interaction. These electrons then follow the standard rules of the energy deposition and may cause spatial broadening of the tracks.

For a thinner material, the statistical fluctuations result in a longer tail of the distribution as is shown in Fig. 2.2, in general described by a Landau-Vavilov distribution [56, 57]. The Most Probable Value (MPV) of the energy loss is then [55]

$$\Delta_\mathrm{p} = \xi \left[\ln\left(\frac{2m_\mathrm{e}c^2\beta^2\gamma^2}{I}\right) + \ln\left(\frac{\xi}{I}\right) + 0.200 - \beta^2 - \delta(\beta\gamma) \right], \tag{2.4}$$

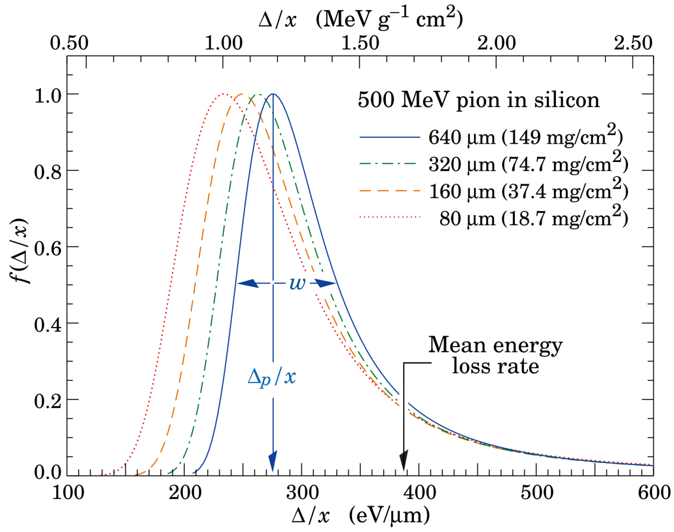


Figure 2.2: Probability of energy loss interactions for 500 MeV pions in silicon, Δ_p represents the most probable value. A broadening of the distribution following the decrease of the silicon thickness is visible, best described by Straggling functions. [52, 55]

where

$$\xi = \frac{K}{2} \left\langle \frac{Z}{A} \right\rangle \frac{x}{\beta^2}, \tag{2.5}$$

for the material of thickness x. The description of other constants follows from Table 2.1. The average energy loss per thickness then corresponds to the Bethe-Bloch formula, lowered based on the thickness.

For very thin materials (such as silicon sensors $\mathcal{O}(100\,\mu\mathrm{m})$) the Landau-Vavilov distribution does not describe the deposition properly and Straggling functions have to be used. The Eq. (2.4) mostly corresponds to the distribution, however, the profile gets significantly wider (shown for varied thickness of silicon in Fig. 2.2). [55] For this reason the Landau function is often convoluted with a Gaussian to better represent the measurements.

2.1.3 Multiple Coulomb scattering

As a result of the interaction, the trajectory of the particles may change. Apart from the inelastic scattering where the traversing particle loses energy, the Coulomb interaction may also result in an elastic Rutherford scattering, affecting only the direction of the flight. Due to a large number of interactions, the overall effect can be approximated by a Gaussian distribution with a standard deviation θ representing the scattering angle as [58]

$$\theta = \frac{13.6\,\mathrm{MeV}}{\beta c p} z \sqrt{\frac{x}{X_0}} \left[1 + 0.038 \ln \left(\frac{x}{X_0} \right) \right], \tag{2.6}$$

with particle momentum p and material thickness x relative to its radiation length X_0 (defined as a distance after which the particle contains only $1/e$ of its energy).

The scattering is visualised in Fig. 2.3, it can result both in the offset of the track as well as change of its direction. For this reason the tracking detectors aim to minimize the

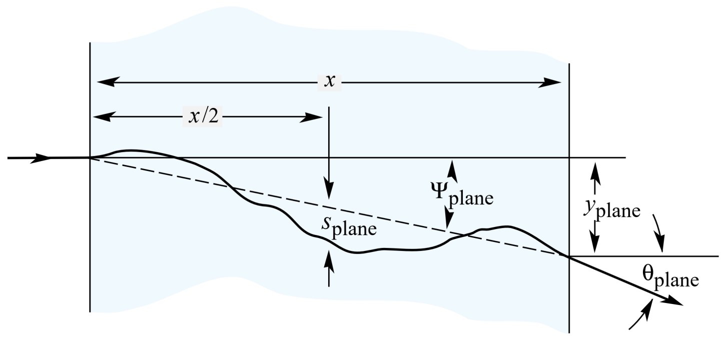


Figure 2.3: Schematic visualisation of the Coulomb scattering. [52]

amount of material present in the path of the particles.

2.2 Semiconductors

Due to their properties, semiconductors are commonly used sensor media in the tracking detectors. In order to describe them, one must look at the quantum mechanical description of electrons and their discrete energy levels. In crystalline structures the orientation and size of the lattice influences the disposition of these levels, effectively creating different bands separated by forbidden energy gaps. Two such bands are conduction – the electrons are not bound to the atom and therefore may conduct electric current; and valence – the electrons are bound.

Semiconductors (e.g. silicon, germanium,…) are materials where the separation between the conduction and valence bands is small enough to permit for a thermal excitation of electrons. The size of this forbidden gap is called a bandgap energy E_{g} and for semiconductors is $\mathcal{O}(1\,\mathrm{eV})$. [59] For comparison, insulators also have these two bands separated, however, the energy required for an electron to jump from one band to another is much larger ($E_{\mathrm{g}} > 5\,\mathrm{eV}$). Finally, conductors do not experience any separation and outermost electrons freely conduct current. A graphical representation of the difference is in Fig. 2.4. An important aspect of the electrons jumping to the higher energy state is creation of vacancies (called holes) in their original place. These holes behave as positively charged particles moving in the material and contribute to the overall conductivity.

The excitation of electrons from one band to another does not only depend on the energy of the forbidden gap, but also on the lattice momentum of the states at the bottom and top of the conduction and valence bands, respectively. In the direct semiconductors momenta of both states match, however, they are different in the indirect ones. This increases the energy requirements for the electron to change states. [60]

Semiconductors exist in two forms – intrinsic or extrinsic. In the first case the semiconductor is made of one element only and the free charge carriers are a direct consequence of the thermal excitations. In an ideal case at the absolute zero temperature, the intrinsic semiconductor would behave as an insulator. Since it is almost impossible to produce materials with just one element, the semiconductors are usually extrinsic with

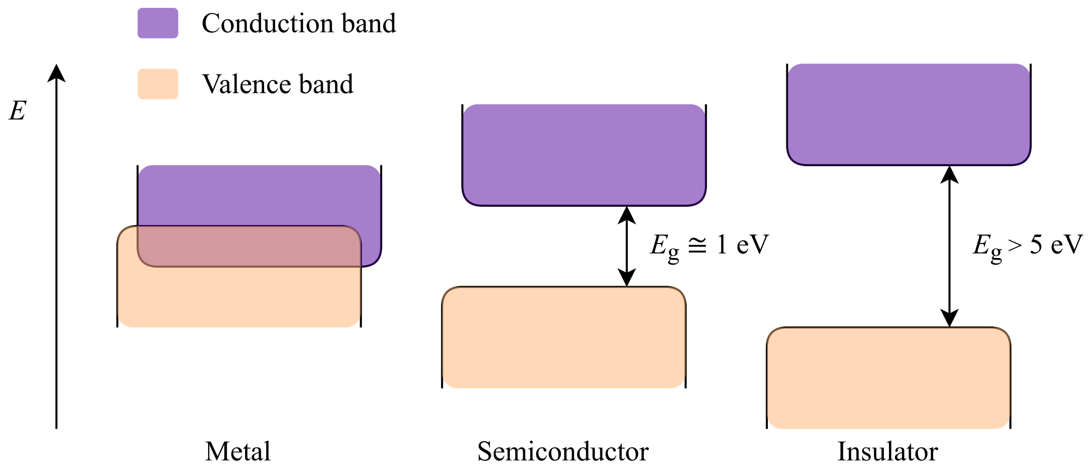


Figure 2.4: Comparison of the band structure for metal, semiconductor and insulator materials.

impurities purposefully introduced into them. One example is replacing the atoms in the crystal lattice with elements with a different number of valence electrons, thus changing the balance of the charge carriers. These properties will be discussed later in the text, using silicon as a commonly used building block.

2.2.1 Silicon

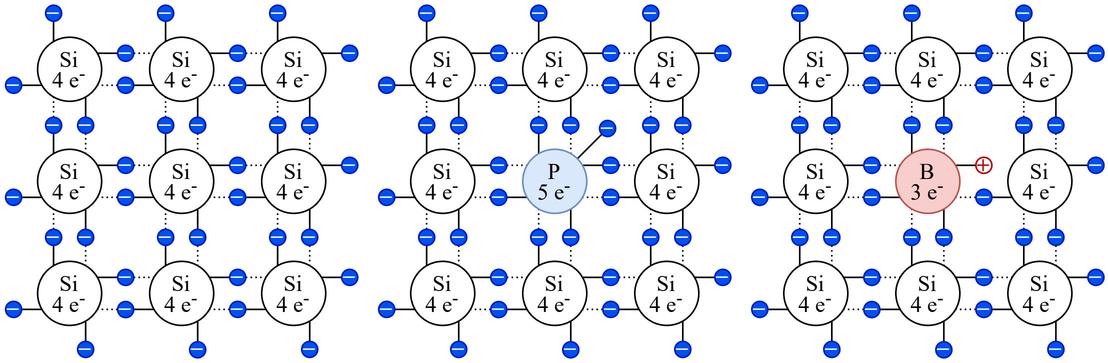
Silicon is an element with an atomic number $Z = 14$ and four valence electrons forming covalent bonds with neighbouring atoms, commonly present in the same crystal structure as a diamond. It is a naturally occurring indirect semiconductor with a bandgap energy $E_g = 1.12\,\text{eV}$, however, the excitation energy needed to produce an electron-hole pair is about $3.6\,\text{eV}$.

Due to its abundance on the Earth, silicon is one of the most widely used semiconductors in the industry. Among other applications, silicon is the main element used to develop transistors, the latest technology achieving size $\mathcal{O}(5\,\text{nm})$. These are in turn used to create ASICs and other electronic components.

Its usefulness and processing availability is also heavily utilized in research, most commonly as an interaction medium for particles to deposit their energy in. [61, 62] The excitation energy of silicon is an order of magnitude lower than gas-based detectors, effectively reducing the size of the used sensors while providing options for an excellent resolution. [63]

As mentioned before, impurities are commonly introduced into the material lattice in order to achieve the desired performance. Silicon semiconductors are then used in a form of n- or p- types, based on the number of valence electrons in the impurity. The difference of the covalent bonds is visualised in Fig. 2.5 and further described here.

n-type The n-type semiconductors have 5 valence electrons, effectively providing an extra electron carrier to the conduction band. They are also referred to as donor



(a) : Intrinsic semiconductor. **(b) :** n-type semiconductor. **(c) :** p-type semiconductor.

Figure 2.5: visualisation of the silicon valence band with four covalent bonds.

impurities, commonly used are phosphorus or arsenic.

p-type Opposing to the previous, p-type semiconductors have only 3 valence electrons thus inserting a hole into the valence band. Because of this they are called acceptor impurities, dopants are boron or aluminium.

2.2.2 Carrier motion

The electrons and holes in the semiconductors are freely moving particles in the lattice until they recombine. Such macroscopic carrier motion can be defined as a current density $\boldsymbol{J}$, split into its drift and diffusion components

$$\boldsymbol{J} = \boldsymbol{J}_{\mathrm{drift}} + \boldsymbol{J}_{\mathrm{diff}}. \tag{2.7}$$

The drift current density is caused directly by the applied electric field written as

$$\boldsymbol{J}_{i,\mathrm{drift}} = q_i n_i \boldsymbol{\nu}_i = q_i n_i \mu_i \boldsymbol{E}, \tag{2.8}$$

where n_i is the density of charge q_i with the velocity $\boldsymbol{\nu}_i$, defined using charge mobility μ_i and intensity of the electric field $\boldsymbol{E}$.

The mobility is dependant on both temperature and intensity of the electric field. It is mostly constant for the intensities less than $1 \cdot 10^4\,\mathrm{V\,cm^{-1}}$ after which it starts to scale as $\mu_i \sim 1/E$. This is mostly caused by the increase of random collisions of the charge carriers with the material. The velocity increases linearly with the electric field and reaches a saturation value $\mathcal{O}(1 \cdot 10^7\,\mathrm{cm\,s^{-1}})$. The temperature dependency follows as $\mu_i \sim T^{-2.3}$. [64] The values for electrons and holes at 300 K are estimated as [65]

$$\mu_{\mathrm{e}} = (1415 \pm 46)\,\mathrm{cm^2\,V^{-1}\,s^{-1}},$$
$$\mu_{\mathrm{h}} = (480 \pm 17)\,\mathrm{cm^2\,V^{-1}\,s^{-1}}.$$

The difference between the two is caused by the fact that while the electrons move in the conduction band, the holes rely on the movement of the electrons within the valence band, where they are more strongly bound to the atom.

For the combination of electron-hole pairs, the Eq. (2.8) is then rewritten to

$$\boldsymbol{J}_{\text{drift}} = \sum_i \boldsymbol{J}_{i,\text{drift}} = q\left(n_{\text{e}}\mu_{\text{e}} + n_{\text{h}}\mu_{\text{h}}\right)\boldsymbol{E} = \sigma\boldsymbol{E}, \tag{2.9}$$

where q is the elementary charge and subscripts e,h denote electrons and holes, respectively. The constant σ represents conductivity of the material, its inverse value ρ resistivity

$$\rho = \frac{1}{\sigma} = \frac{1}{q\left(n_{\text{e}}\mu_{\text{e}} + n_{\text{h}}\mu_{\text{h}}\right)}. \tag{2.10}$$

The diffusion component is governed by the motion of the charges due to their gradient of concentration $\boldsymbol{\nabla n_i}$

$$\boldsymbol{J}_{i,\text{diff}} = -q_i D_i \boldsymbol{\nabla n_i} = -\mu_i k_{\text{B}} T \boldsymbol{\nabla n_i}, \tag{2.11}$$

D_i presenting diffusion coefficient for the given charge, substituted for using the Einstein equation

$$D_i = \frac{k_{\text{B}}T}{q}\mu_i \tag{2.12}$$

where k_{B} is Boltzmann constant and T is absolute temperature.

Substituting for the current densities in the Eq. (2.7) from the Eq. (2.9) and Eq. (2.11), the total current density based on electron-hole pairs is

$$\boldsymbol{J} = \boldsymbol{J}_{\text{drift}} + \boldsymbol{J}_{\text{diff}} = \sigma\boldsymbol{E} + k_{\text{B}}T(\mu_{\text{e}}\boldsymbol{\nabla n_{\text{e}}} - \mu_{\text{h}}\boldsymbol{\nabla n_{\text{h}}}). \tag{2.13}$$

The combination of these effects results in a Gaussian spread $\sigma_{i,\text{charge}}$, modifying the spatial resolution of the charge collection [63]

$$\sigma_{i,\text{charge}} = \sqrt{2D_i t_{i,\text{drift}}} = \sqrt{\frac{2k_{\text{B}}Tx_{i,\text{drift}}}{q|\boldsymbol{E}|}}, \tag{2.14}$$

where $t_{i,\text{drift}}$ and $x_{i,\text{drift}}$ represent the drift time and distance, respectively.

2.2.3 Carrier density

Based on the [59], the concentration of any free carriers n_i at the energy level E for an intrinsic semiconductor is estimated as

$$n_i = \int_{E_b}^{\infty} N(E)f(E)\,\mathrm{d}E, \tag{2.15}$$

where E_b is the lowest energy level in the band, $N(E)$ is the corresponding density of states and $f(E)$ the Fermi probability distribution. The density of states can be calculated as

$$N(E) = 4\pi\left(\frac{2m_i}{h^2}\right)^{3/2} E^{1/2}, \tag{2.16}$$

using the effective mass of the carriers m_i in the medium and Planck constant h. The probability distribution from the Fermi-Dirac statistics is

$$f(E) = \frac{1}{\mathrm{e}^{\frac{(E-E_\mathrm{F})}{k_\mathrm{B}T}} + 1}, \tag{2.17}$$

where k_B is Boltzmann constant, T is absolute temperature and E_F is the Fermi level. The level effectively defines a point at which the probability of the carrier occupancy is $f(E) = 0.5$. Approximating the distribution for $(E - E_\mathrm{F}) \gg k_\mathrm{B}T$ yields

$$f(E) \approx \mathrm{e}^{-\frac{(E-E_\mathrm{F})}{k_\mathrm{B}T}}. \tag{2.18}$$

Substituting back into the Eq. (2.15) and solving using gamma functions we get

$$n_\mathrm{n} = 2\left(\frac{2\pi m_\mathrm{n} k_\mathrm{B}T}{h^2}\right)^{3/2} \mathrm{e}^{-\frac{E_\mathrm{C}-E_\mathrm{F}}{k_\mathrm{B}T}} = N_\mathrm{C}\mathrm{e}^{-\frac{E_\mathrm{C}-E_\mathrm{F}}{k_\mathrm{B}T}}, \tag{2.19}$$

$$n_\mathrm{p} = 2\left(\frac{2\pi m_\mathrm{p} k_\mathrm{B}T}{h^2}\right)^{3/2} \mathrm{e}^{-\frac{E_\mathrm{F}-E_\mathrm{V}}{k_\mathrm{B}T}} = N_\mathrm{V}\mathrm{e}^{-\frac{E_\mathrm{F}-E_\mathrm{V}}{k_\mathrm{B}T}}, \tag{2.20}$$

concentrations of electrons n_n and holes n_p, their effective masses m_n, m_p and energies of the corresponding conduction E_C and valence E_V bands. The factor $N_{\mathrm{C(V)}}$ in the front represents effective density of states in the conduction (valence) band.

Since the derivation was done for an intrinsic semiconductor having the same amount of electrons and holes, we can estimate the intrinsic density n_I corresponding to the amount of thermally generated pairs as

$$n_\mathrm{I} = \sqrt{n_\mathrm{n} n_\mathrm{p}} = \sqrt{N_\mathrm{C} N_\mathrm{V}}\mathrm{e}^{-\frac{E_\mathrm{g}}{2k_\mathrm{B}T}}. \tag{2.21}$$

The E_g in this case is the bandgap energy, defined as difference between the energy levels of conduction and valence bands. It is important to note that the intrinsic density is not dependent on the Fermi level. The factor $\sqrt{N_\mathrm{C} N_\mathrm{V}}$ scales with temperature as $\sim T^{3/2}$ thus the amount of thermally generated pairs for silicon at room temperature $T = 300\,\mathrm{K}$ is about $1.45 \cdot 10^{10}\,\mathrm{cm}^{-3}$.

Rewriting the Eq. (2.19) to solve for the E_F

$$\begin{aligned} E_\mathrm{F} &= E_\mathrm{C} + k_\mathrm{B}T \ln\left(\frac{n_\mathrm{n}}{N_\mathrm{C}}\right), \\ E_\mathrm{F} &= -E_\mathrm{V} - k_\mathrm{B}T \ln\left(\frac{n_\mathrm{p}}{N_\mathrm{V}}\right), \end{aligned} \tag{2.22}$$

the Fermi level E_I for an intrinsic semiconductor is then derived as

$$E_\mathrm{I} \equiv E_\mathrm{F} = \frac{E_\mathrm{C} - E_\mathrm{V}}{2} + \frac{3k_\mathrm{B}T}{4} \ln\left(\frac{m_\mathrm{p}}{m_\mathrm{n}}\right). \tag{2.23}$$

It is clear that at the absolute zero temperature the Fermi level is in the middle of the bandgap, at higher temperatures the deviation is caused due to different effective masses of electrons and holes.

2.2.3.1 Doping

The commonly used dopants have a shallow energy level in the forbidden bandgap region, either close to the conduction (donor impurity) or valence (acceptor impurity) bands. The comparison between intrinsic, n- and p-type semiconductors is visualised in Fig. 2.6.

Since the energy difference of impurities is very close to their respective bands (only about 0.05 eV), almost all of the atoms are in an ionized state at the room temperature. Given that the concentration of dopants is much larger than the intrinsic density and the silicon is heavily doped only with one element, the carrier density $n_{\mathrm{n(p)}} \approx N_{\mathrm{D(A)}}$ is equal to the concentration of the donors (acceptors). For $N_{\mathrm{D}} \gg N_{\mathrm{A}}$ the shift of the Fermi level can be estimated as

$$E_{\mathrm{F}} = E_{\mathrm{I}} + k_{\mathrm{B}} T \ln\left(\frac{N_{\mathrm{D}}}{n_{\mathrm{I}}}\right), \tag{2.24}$$

analogically for $N_{\mathrm{A}} \gg N_{\mathrm{D}}$

$$E_{\mathrm{F}} = E_{\mathrm{I}} - k_{\mathrm{B}} T \ln\left(\frac{N_{\mathrm{A}}}{n_{\mathrm{I}}}\right). \tag{2.25}$$

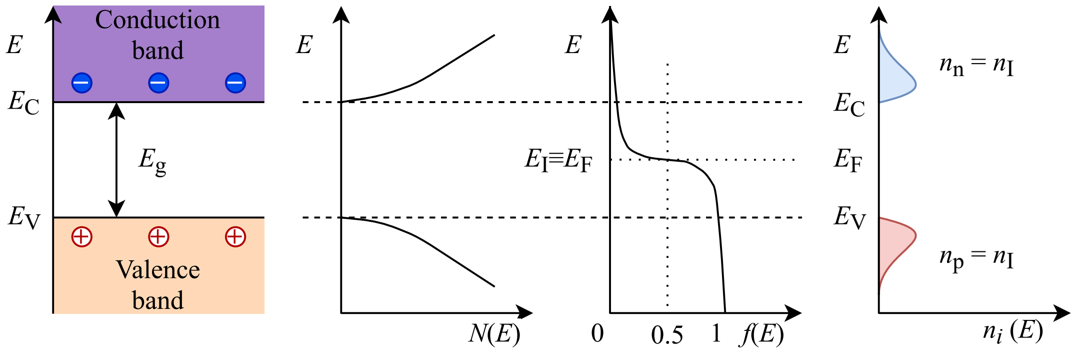
Using $n_{\mathrm{I}}^2 = \mathrm{const} \approx 1 \cdot 10^{20}\,\mathrm{cm}^{-6}$ from the Eq. (2.21) in a thermal equilibrium means that in both doping cases the concentration increase of one charge carrier is at the expense of the other. This effectively increases the total amount of charge carriers and thus also the conductivity. As an example, doping with donor impurities $n_{\mathrm{n}} \approx N_{\mathrm{D}} \equiv 1 \cdot 10^{17}\,\mathrm{cm}^{-3}$ satisfies the condition for concentration of holes $n_{\mathrm{p}} = 1 \cdot 10^{3}\,\mathrm{cm}^{-3}$. The density of charge carriers $n = n_{\mathrm{n}} + n_{\mathrm{p}}$ is then $n \approx 2 \cdot 10^{10}\,\mathrm{cm}^{-3}$ and $n \approx 1 \cdot 10^{17}\,\mathrm{cm}^{-3}$ for intrinsic and doped semiconductor, respectively.

In reality, semiconductors are usually doped with both n- and p- types, in some cases even different concentrations of the same type (n^+-on-n semiconductors). In such case the effective doping concentration is used, defined as an absolute value of the difference of the two.

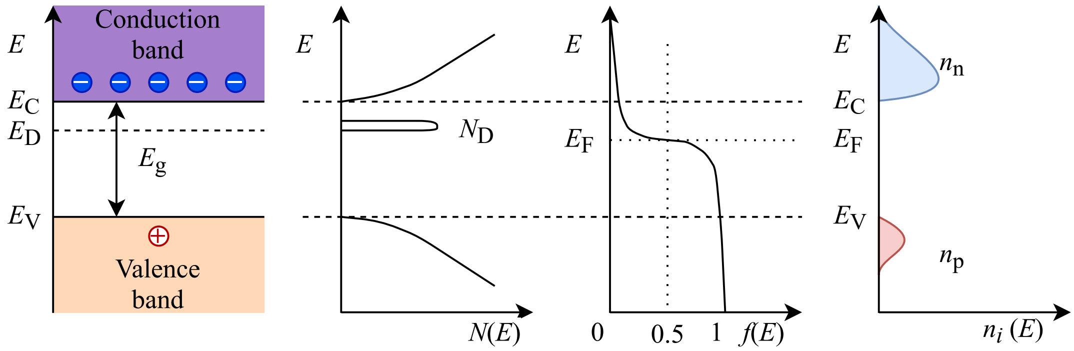
2.2.4 p-n junction

Probably the most important property of extrinsic semiconductors is the existence of p-n junction, created on the contact border of the n- and p- type devices. [59, 61, 62] They are used in diodes or transistors, controlling the direction and flow of the current.

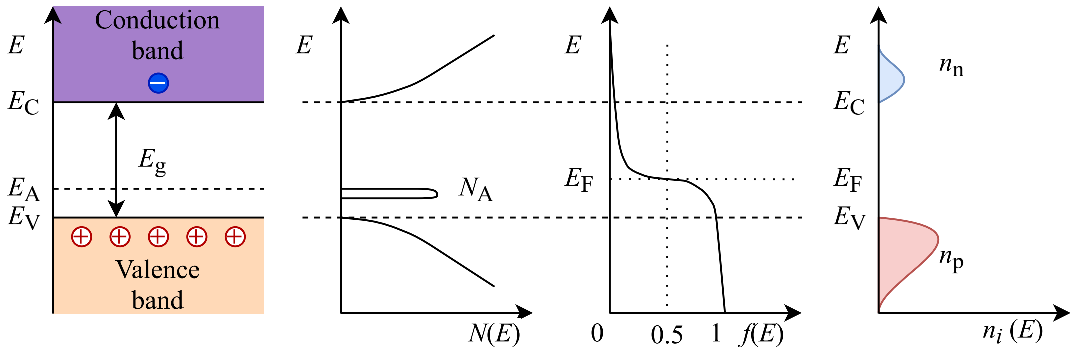
As a result of different Fermi levels, a gradient of the concentration of the charge carriers occurs in the junction (see Fig. 2.7). Electrons from the donor impurity flow in the direction of the p-type region, the holes experience the exact opposite flow towards the n-type region. The charge carriers in the centre recombine and create a depletion region with an electric field opposing further diffusion. The system eventually reaches an



(a) : Intrinsic semiconductor.



(b) : n-type semiconductor. E_{D} and N_{D} represent the donor impurity energy level and concentration, respectively.



(c) : p-type semiconductor. E_{A} and N_{A} represent the acceptor impurity energy level and concentration, respectively.

Figure 2.6: Schematic semiconductor band diagram, density of states $N(E)$, Fermi-Dirac distribution $f(E)$ and $n_{\mathrm{n(p)}}$ electron (hole) carrier concentration. E_{C} and E_{V} represent energy limits of the conduction and valence bands, respectively. Based on [59].

equilibrium, also resulting in the matching of the Fermi levels across the junction.

The equilibrium state of the junction is best described using an abrupt junction (large jump in concentration) in Fig. 2.8. The charge carriers in the depletion region are recombined, its width d written as

$$d = d_{\mathrm{n}} + d_{\mathrm{p}}, \tag{2.26}$$

where d_{n} and d_{p} correspond to the width of the depletion region in the donor and acceptor

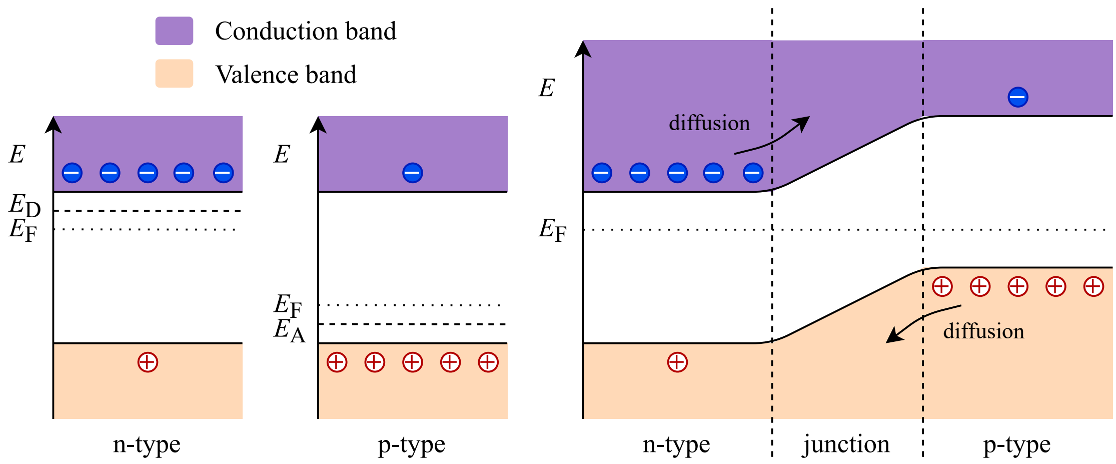


Figure 2.7: Schematic visualisation of the p-n junction.

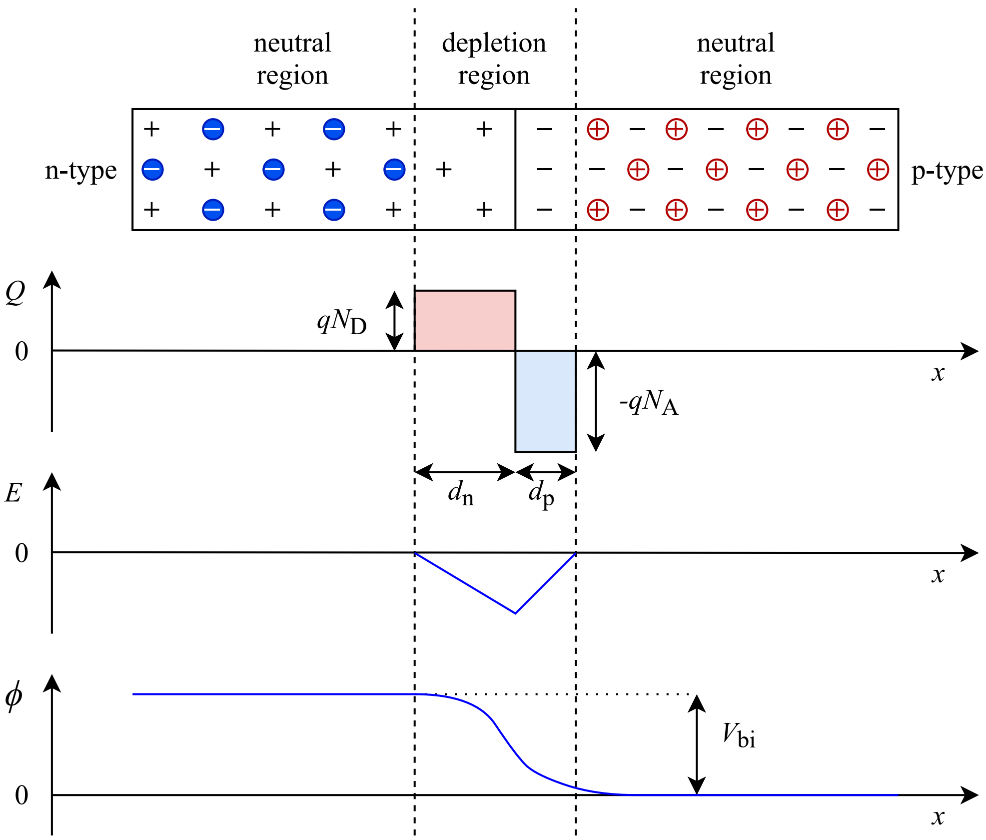


Figure 2.8: Spatial profiles of charge density Q, intensity of the electric field E and electric potential ϕ across the p-n junction.

parts, respectively. The charge density $Q(x)$ can be written as a function of depth

$$Q(x) = \begin{cases} 0, & x \leq -d_n \\ qN_D, & -d_n < x \leq 0 \\ -qN_A, & 0 < x \leq d_p \\ 0, & d_p < x \end{cases}, \tag{2.27}$$

the electric field $E(x)$ and potential $\phi(x)$ can be then derived from the Poisson's equation

$$\frac{\mathrm{d}^2\phi}{\mathrm{d}x^2} = -\frac{Q(x)}{\varepsilon}. \tag{2.28}$$

The $\varepsilon = \varepsilon_0\varepsilon_\mathrm{r}$ is a product of permittivity of vacuum ε_0 and relative permittivity of the material ε_r.

After integrating Eq. (2.28) and applying the boundary conditions $E(x) = 0$ for both d_p and d_n, the electric field is

$$E(x) = \frac{\mathrm{d}\phi(x)}{\mathrm{d}x} = \begin{cases} -\frac{qN_\mathrm{D}}{\varepsilon}(x + d_\mathrm{n}), & -d_\mathrm{n} < x \leq 0 \\ \frac{qN_\mathrm{A}}{\varepsilon}(x - d_\mathrm{p}), & 0 < x \leq d_\mathrm{p} \end{cases}. \tag{2.29}$$

Since the two values have to match at the contact position $x = 0$, a condition fulfilling charge neutrality in the semiconductor is obtained

$$N_\mathrm{D}d_\mathrm{n} = N_\mathrm{A}d_\mathrm{p}. \tag{2.30}$$

Depletion depth is therefore proportional to the doping concentration of the two parts. In the case where one side is heavily doped with respect to the other, the region with larger concentrations is narrow and vice-versa for the other side. This is frequently used in the particle detectors as one side of the sensor tends to have much larger doping concentration than the other.

The further integration of Eq. (2.29) results in the electric potential

$$\phi(x) = \begin{cases} -\frac{qN_\mathrm{D}}{2\varepsilon}(x + d_\mathrm{n})^2 + V, & -d_\mathrm{n} < x \leq 0 \\ \frac{qN_\mathrm{A}}{2\varepsilon}(x - d_\mathrm{p})^2, & 0 < x \leq d_\mathrm{p} \end{cases}, \tag{2.31}$$

the integration constants were selected based on the boundary conditions $\phi(d_\mathrm{p}) = 0$ and $\phi(-d_\mathrm{n}) = V$, where V is a potential difference present between the two sides. Matching the equations at $\phi(0)$, the voltage can be estimated as

$$V = \frac{q}{2\varepsilon}(N_\mathrm{D}d_\mathrm{n}^2 + N_\mathrm{A}d_\mathrm{p}^2). \tag{2.32}$$

In the case where no external field is applied, the difference represents a built-in voltage $V_\mathrm{bi} \equiv V$ in the diode. This also corresponds to the difference of the Fermi levels between the two semiconductor types and can be obtained in a different form from Eq. (2.22)

$$V_\mathrm{bi} = \frac{1}{q}(E_{F_\mathrm{p}} - E_{F_\mathrm{n}}) = \frac{k_\mathrm{B}T}{q}\ln\left(\frac{N_\mathrm{A}N_\mathrm{D}}{n_\mathrm{I}^2}\right). \tag{2.33}$$

Since the value of the built-in voltage is very small, the particles that interacted in the depleted volume would be missed compared to the large amount of thermally produced e-h pairs.

2.2.4.1 Reverse biasing

Connecting the opposing ends of the p-n junction to an external PSU affects the properties of the depletion region. The voltage to the electrodes can be applied in two directions:

Forward , with the positive terminal connected to the p-type semiconductor. The applied voltage effectively cancels the V_{bi} and forces electrons from the n-type semiconductor to drift to the depletion region and recombine. With the voltage $V > V_{\mathrm{bi}}$ the depletion region eventually disappears and the current may freely flow through the diode.

Reverse , has the positive terminal connected to the n-type semiconductor. This is used to further expand the depletion width, used in particle detectors.

Focusing on the reverse biasing case, the external voltage V_{bias} may be added to the Eq. (2.32) as

$$V_{\mathrm{bias}} + V_{\mathrm{bi}} = \frac{q}{2\varepsilon}(N_{\mathrm{D}}d_{\mathrm{n}}^2 + N_{\mathrm{A}}d_{\mathrm{p}}^2). \tag{2.34}$$

Substituting from Eq. (2.30), and solving for depletion width d from Eq. (2.26) we get

$$d = \sqrt{\frac{2\varepsilon(V_{\mathrm{bi}} + V)}{q}\left(\frac{1}{N_{\mathrm{A}}} + \frac{1}{N_{\mathrm{D}}}\right)}. \tag{2.35}$$

The built in voltage may often be ignored due to its small value (less than 1V).

Since particle detectors usually utilize much larger doping on one side, the equation can be simplified using the lower doping concentration value N

$$d \approx \sqrt{\frac{2\varepsilon}{qN}V_{\mathrm{bias}}}. \tag{2.36}$$

After connecting the voltage to the electrode, the depletion region starts expanding in the direction from p- towards n- type. This is mainly due to the higher mobility of electrons.

Solving the Eq. (2.36) for the voltage in case where the depletion corresponds to the thickness of the device $d \equiv d_{\mathrm{th}}$, the estimate of the full depletion voltage is

$$V_{\mathrm{dep}} = \frac{qNd_{\mathrm{th}}^2}{2\varepsilon}. \tag{2.37}$$

Rewriting the equation using the resistivity ρ and the carrier mobility μ

$$V_{\mathrm{dep}} = \frac{d_{\mathrm{th}}^2}{2\varepsilon\mu\rho}. \tag{2.38}$$

where $\rho = 1/(q\mu N)$. This means that the voltage needed to achieve a full depletion is lower for a material with higher resistivity. The best performance is therefore achieved with the silicon of the highest purity.

The fact that the depletion region holds an electrical potential difference and does not contain any free charge carriers means it can be approximated as a charged capacitor.

[61,63] Using the Eq. (2.36), the capacitance per unit area C_{A} is estimated similar to a simple parallel plate capacitor

$$C_{\mathrm{A}} = \varepsilon \frac{1}{d} = \sqrt{\left(\frac{\varepsilon q N}{2V_{\mathrm{bias}}}\right)}. \tag{2.39}$$

Even though the values for a standard thickness ≈300 µm of a silicon sensor are very small $\mathcal{O}(50\,\mathrm{pF\,cm^{-2}})$ [63], the capacitance has an important consequence on the signal formation as it limits the frequency response of the detector. This is effectively minimized with an increase of the depletion width and the applied bias voltage. Even though the capacitance is not a desired effect, it can be utilized in a form of a $C - V$ measurement which provides an excellent way of estimating the depletion width of a device.

The electric field across the depleted region affects all e-h pairs created within the volume. This is desired for detecting particles interacting in the medium, however, it also includes any charge carriers created from the thermal excitations or from diffusion from the underdepleted regions. The extra component is called leakage or dark current and is mostly dominated by the temperature

$$I(T) \propto T^2 \mathrm{e}^{-\frac{E_{\mathrm{g}}}{2k_{\mathrm{B}}T}}. \tag{2.40}$$

Semiconductors are therefore often cooled down, a rough estimation is that this volume current doubles every 8 K [62].

2.2.5 Detectors

The base structure in the detector is a sensor which provides an interaction volume for the particles. It is based on a combination of p- and n- types implants, their size and shape varying according to the desired application. The sensitive parts of the sensors are made from the depleted region with a metallic layer on the opposing sides, representing electrodes. The rest of the sensor surface is then covered with an insulating layer of silicon oxide SiO_2. [61,62]

The signal from the semiconductor sensor is not caused by the physical collection of the charge carriers at the electrodes, rather by the current induced by the movement of the charges. This is described by a Shockley-Ramo theorem [66,67], the induced current i is calculated as

$$i = q\boldsymbol{E}_{\mathrm{w}}\boldsymbol{\nu}, \tag{2.41}$$

where $\boldsymbol{E}_{\mathrm{w}}$ is a weighting field based on a dimensionless potential and $\boldsymbol{\nu}$ is the velocity of the moving charge q. Using the mobility and intensity of electric field instead of the charge velocity, the equation can be rewritten to

$$i = q\mu\boldsymbol{E}_{\mathrm{w}}\boldsymbol{E}. \tag{2.42}$$

The weighting field only affects the strength of the induced signal and not the movement of the carriers itself. It is defined by applying zero potential to all electrodes apart from the observed one, where the potential is set to one. The total collected charge is then a time integral of the induced current.

As an example using a simple planar sensor (two parallel electrodes on each end), the weighting potential would change linearly with the depth, therefore an e-h pair created in the middle of the sensor would result in the same induced current both from the electron as well as from the hole. However, the instantaneous profile of the electron and hole currents would not match as the the drift of the carriers is affected differently by the electric field.

2.2.5.1 Readout

The signal from the sensors is usually amplified and digitized, the actual implementation based on the specific needs of the application. [62] This can be done using a simple oscilloscope or in particle physics experiments is usually performed with a specially designed ASIC. Regardless of the device used, there are three main types of the measurement of the collected charge.

Binary The simplest way to record data is a binary readout. The particle interaction is registered based on a comparison of the voltage level with a predefined threshold. This is often useful in applications that do not require energy resolution and need to record data with a very high rate.

Analog-Digital Converter (ADC) counts The ADC counts are obtained by directly sampling the voltage/current levels from the amplifier. The number of samples can be tuned to provide peak values or a more detailed representation of the current profile induced on the electrodes. Depending on the signal this method might require a very high data throughput which is often difficult to achieve.

Time over Threshold (ToT) The measurement of the time that the amplifier signal exceeds a threshold, the ToT is an indirect representation of the total deposited charge. It provides a timing information about the duration of a shaped signal representing charging and discharging of a capacitor.

The Charge Collection Efficiency (CCE) and the corresponding scans used to characterise the device performance are described in detail in Chapter 5.

2.2.5.2 Segmentation and resolution

In order to achieve better spatial resolution, the sensor is often segmented. [61, 62] For the particle tracking purposes, the most commonly used configuration are strips, pixel matrix or even a complex 3D column structure [68].

Strip sensors provide information only in one direction, however, they are cost-effective and may be created in larger sizes. The multi-dimensional tracking can be achieved

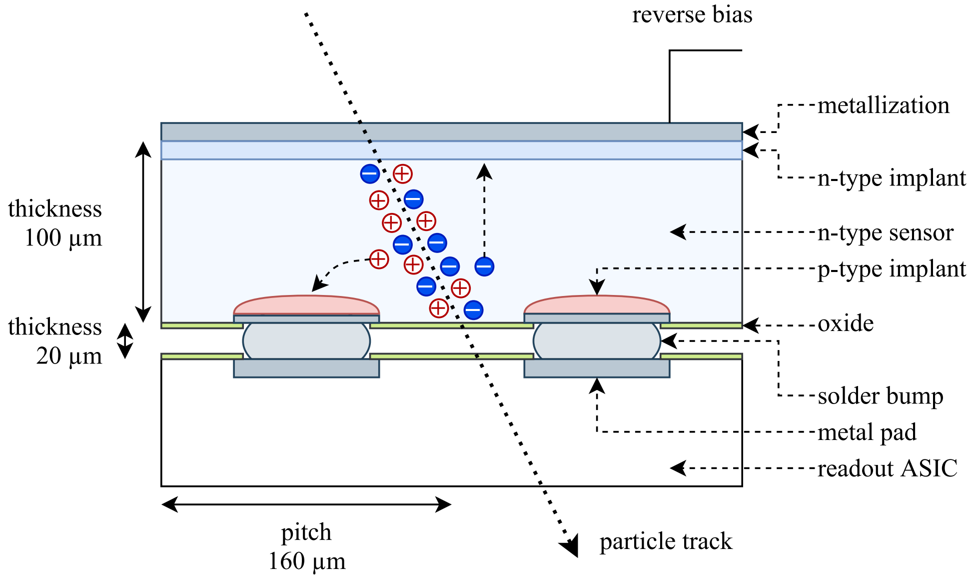


Figure 2.9: Schematic cross-section of a hybrid pixel detector.

by using at least two layers rotated with respect to each other. A small opening angle $\mathcal{O}(10°)$ between the strips can be used in order to minimize the probability of ghost hits (a dis-ambiguity of the interaction location). The signal readout and processing may be done on the periphery of the sensor using wire-bonds to connect the ASIC to the individual channels.

Another commonly used option is to do a full 2D readout with sensors split to pixels. These can be then read out either as a hybrid detector, where the ASIC is flip-chip bonded to the sensor; or as a Monolithic Active Pixel Sensors (MAPS) [69], where the whole sensor and readout part is implemented into the same piece of the silicon substrate. A cross-section of a hybrid pixel detector is in Fig. 2.9.

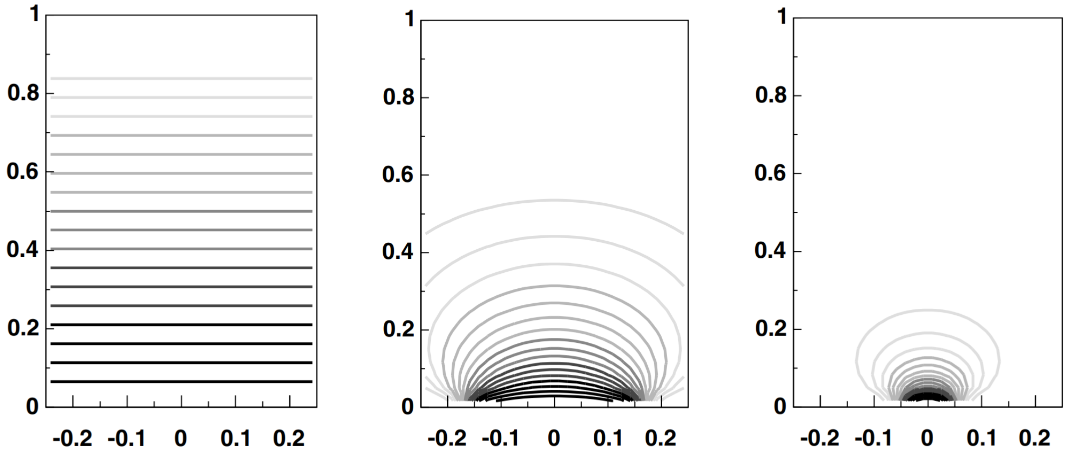
It is clear from the Eq. (2.41) that the amount of the induced charge is affected as the weighting potential is modified based on the neighbouring channels. Comparison of the field lines for a planar and strip sensor is in Fig. 2.10. The main outcome is that while in a simple diode the induced current is registered from the whole depth of the sensor, in a segmented device only the areas close to the segmentation will contribute far more.

The decrease of the channel size is driven by a desire to improve spatial resolution, which is in turn affected by the type of the readout. In addition, particles do not travel perfectly aligned through the middle of the channel. When the interaction occurs on the border of two channels, both of them may register the signal.

In a case with a binary strip readout without any charge sharing, the resolution is estimated as [62]

$$\sigma_{\mathrm{pos}} = \frac{p}{\sqrt{12}}, \tag{2.43}$$

where p is the strip pitch. This simple scenario considers the particle hits assigned to the centre of the channel, regardless of their actual position of interaction within the



(a) : Infinite parallel electrodes.

(b) : Collection electrode 1/3 of the sensor thickness.

(c) : Collection electrode 1/10 of the sensor thickness.

Figure 2.10: Comparison of the weighting field for a planar and segmented device based on the electrode size. [62]

channel volume. Including the fact that the neighbouring channels may register the same interaction on the border improves the resolution, however, at the risk of mis-tagging two independent particles occuring at the same time. The resolution could be further improved by a readout measuring values proportional to the amount of the collected charge or even by tilting the sensors and forcing multi-channel interactions.

2.3 Radiation damage

While the interaction of radiation in the sensor volume is desired to record the information about the particles, very high rates result in radiation induced effects to the individual components. The damage caused to the sensors will be discussed in this section, however, it is important to note that similar effects also occur in the semiconductor-based readout components. [61]

These radiation effects are triggered both via ionization and direct interaction of particles with the atoms in the crystal lattice. This can result in charge trapping, increase of leakage current or change of the space-charge in the oxide or depleted region [62]. It also results in the lowered CCE.

2.3.1 Surface damage

Since the surface is mostly made of a dielectric material, the lattice structure is highly irregular. For this reason the main damage is caused due to ionization and subsequent trapping of holes in the crystal structure. The scaling of the effects may be done using Total Ionizing Dose (TID) – representing the total amount of deposited energy which results in the e-h pair creation.

Similarly to the depleted region of a semiconductor, the pairs created in the oxide follow the present electric field and most of them recombine immediately after generation. However, some electrons escape the recombination process and due to a large difference in the carrier mobility

$$\mu_{\mathrm{e,oxide}} = 2 \cdot 10^{1}\,\mathrm{cm^2\,V^{-1}\,s^{-1}},$$
$$\mu_{\mathrm{h,oxide}} = 2 \cdot 10^{-5}\,\mathrm{cm^2\,V^{-1}\,s^{-1}},$$

drift out of the volume while the much slower holes get trapped. [70] This is often a case for the transition region between the semiconductor and the oxide regions, caused by the interstitial oxygen atoms. The positively charged states may also be introduced by ionizing interactions with the trivalent silicon atoms, effectively removing the valence electron. [71]

Using the TID metric, the oxide charge gets saturated after $\mathcal{O}(100\,\mathrm{kRad})$, the resulting value $N_{\mathrm{ox,sat}} \approx 3 \cdot 10^{12}\,\mathrm{cm^{-2}}$. [72] The extra charge affects the electric field near the transition region, causing a creation of electron/hole accumulation layer in n- and p-types, respectively.

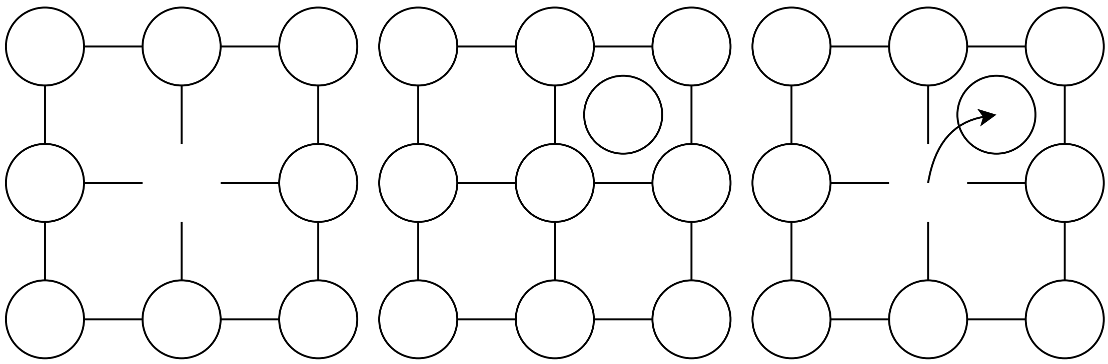
In order to remove most of the damage, the device can be annealed at a high temperature 150 °C, though in particle detectors this is usually impractical due both to mechanical constraints and the effect on the more important bulk damage discussed below. Annealing of the surface damage is also achieved by electrons to tunnel from the bulk to the oxide where they recombine with the holes. [73]

2.3.2 Bulk damage

Contrary to the surface, the ionization in the bulk does not present a major issue since the carriers have large mobility and do not get trapped easily. However, the structure of the semiconductor is a highly regular crystal and given sufficient energy, the interacting particles may knock-off an atom from the lattice. These defects are called vacancies (missing atom), interstitial (extra atom) or Frenkel pairs (vacancy next to an interstitial). A visualisation is in Fig. 2.11.

In silicon 25 eV is needed to break a bond and create a point-like defect. [74] Due to the energy and momentum conservation, the corresponding kinetic energies are about 190 eV and 260 keV for protons (neutrons) and electrons, respectively. In case the transferred kinetic energy reaches values from 2 keV to 12 keV, the silicon atoms will continue interacting with the rest of the lattice and generate a localized cluster – several point-like defects. Higher recoil energies result in multiple such clusters.

Since the interactions depend heavily on the energy and type of particle, a normalization is needed to simplify the description of the damage effects. Non-Ionizing Energy Loss (NIEL) has been introduced for this reason, describing the amount of energy deposited per area density that does not result in ionization (an opposite of TID). It corresponds to the displacement damage cross-section D summing up all possible interactions i for a



(a) : Vacancy. **(b) :** Interstitial atom. **(c) :** Frenkel pair.

Figure 2.11: Visualisation defects in the crystal lattice.

given energy E as [75]

$$D(E) = \sum_i \sigma_i(E) \int_0^{E_{\mathrm{R,max}}} f_i(E, E_{\mathrm{R}}) P(E_{\mathrm{R}})\, \mathrm{d}E_{\mathrm{R}}, \tag{2.44}$$

where σ_i is an interaction cross-section, $f_i(E, E_{\mathrm{R}})$ represents a probability of a collision resulting in a recoil energy E_{R}. $P(E_{\mathrm{R}})$ is a Lindhard partition function [76] describing the energy fraction needed for the displacement ($P(E_{\mathrm{R}} < E_{\mathrm{D}}) = 0$ for a displacement energy E_{D}). For the normalization purposes, the displacement damage cross-section of 1 MeV neutrons is used as $D_{\mathrm{n}}(1\,\mathrm{MeV}) = 95\,\mathrm{MeV\,mb}$.

The fluence of an arbitrary type of particles Φ can be therefore scaled as

$$\Phi_{\mathrm{eq}} = \kappa\Phi, \tag{2.45}$$

where Φ_{eq} is the 1 MeV neutron equivalent fluence (units written as $1\,\mathrm{MeV}\,\mathrm{n_{eq}}\,\mathrm{cm}^{-2}$) and κ is a hardness factor specific for a given particle. This provides a straightforward conversion between different types of particles and energies to the same unit, the dependency of NIEL on the particle type and energy is plotted in Fig. 2.12.

Even though the defects are generally not stable and move through the lattice, they can further interact with impurities in the crystal forming permanent defects. The damage therefore causes a modification in the semiconductor band structure, resulting in additional generation of e-h pairs or formation of traps in the band-gap region. This in turn affects the overall space-charge present in the depletion zone, worsening the performance of the detector.

2.3.2.1 Leakage current

The extra thermal generation and recombination of the e-h pairs is caused by the added energy levels between the valence and conduction band. The existence of additional charge carriers results in a larger leakage current, effectively increasing the total power consumption and heat output of the sensor. The temperature then increases the resistivity,

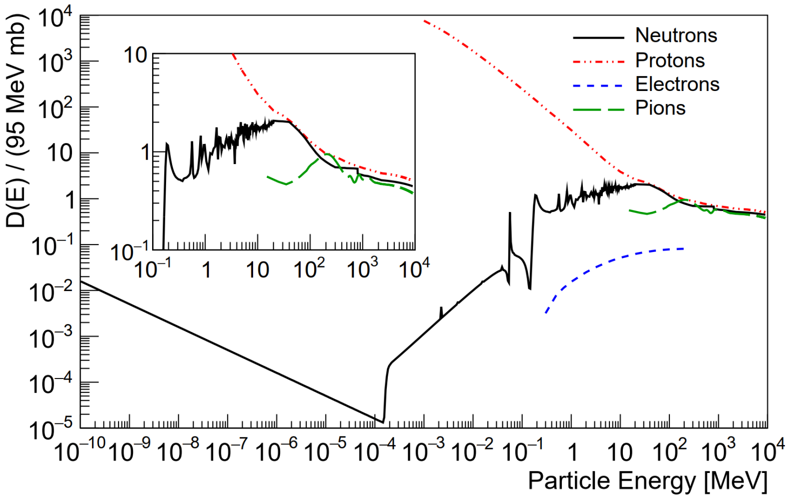


Figure 2.12: Dependency of damage cross-section on particle energy, plotted for different particles. [77]

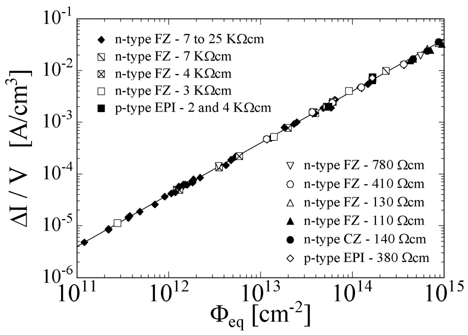


Figure 2.13: Dependency of leakage current on the neutron fluence equivalent for different silicon resistivity and doping. Measured right after the irradiation. [75]

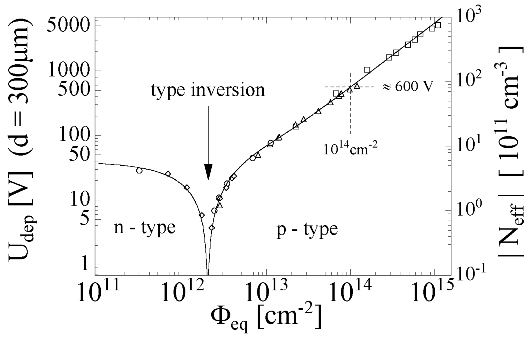


Figure 2.14: Dependency of depletion voltage and effective doping concentration on the neutron fluence equivalent. Measured right after the irradiation. [73]

which in turn further increases the current causing a thermal runaway of the device. For this reason the devices in a strong radiation environment have to be actively cooled.

The leakage current I per volume V can be written as

$$\frac{I}{V} = \frac{I_{\phi_{\text{eq}}=0}}{V} + \frac{\Delta I(\Phi_{\text{eq}})}{V}, \tag{2.46}$$

where the change is specified as

$$\frac{\Delta I(\Phi_{\text{eq}})}{V} = \alpha\Phi_{\text{eq}}. \tag{2.47}$$

Φ_{eq} represents the NIEL equivalent fluence and α is an experimentally measured factor representing current-related damage rate. The increase of the current linearly scales based on the fluence. The factor α does not depend on the silicon resistivity and its doping profiles, the dependency of the leakage current of various devices on the fluence is in Fig. 2.13. [75]

2.3.2.2 Depletion voltage

The defects in the material may act both as donors, acceptors or neutral atoms, altering the concentration of the implanted impurities. From the Eq. (2.37) the depletion voltage depends on the effective doping concentration N_{eff} (difference of the concentrations of donors and acceptors of the bulk). With the increasing fluence, the concentration (and subsequently the depletion voltage) changes as shown in Fig. 2.14.

The effect can be described as creation of acceptor states close to the valence band, effectively inducing negative space charge in the volume. This mostly results in an increase of the voltage needed to reach the full depletion for a p-type bulk. However, for an n-type bulk the process at first cancels out the positive charge present in the depletion region and lowers the overall effective concentration of the device. Afterwards the negative charge starts to dominate and the effective concentration gets increased. The point where the change happens is called a type inversion. [75] The impact on the concentration can be written as

$$N_{\text{eff}} = N_{\text{eff},\Phi_{\text{eq}}=0} - \Delta N_{\text{eff}}(\Phi_{\text{eq}}, T, t), \tag{2.48}$$

where

$$\Delta N_{\text{eff}}(\Phi_{\text{eq}}, T, t) = N_{\text{c}}(\Phi_{\text{eq}}) + N_{\text{a}}(\Phi_{\text{eq}}, T, t) + N_{\text{y}}(\Phi_{\text{eq}}, T, t). \tag{2.49}$$

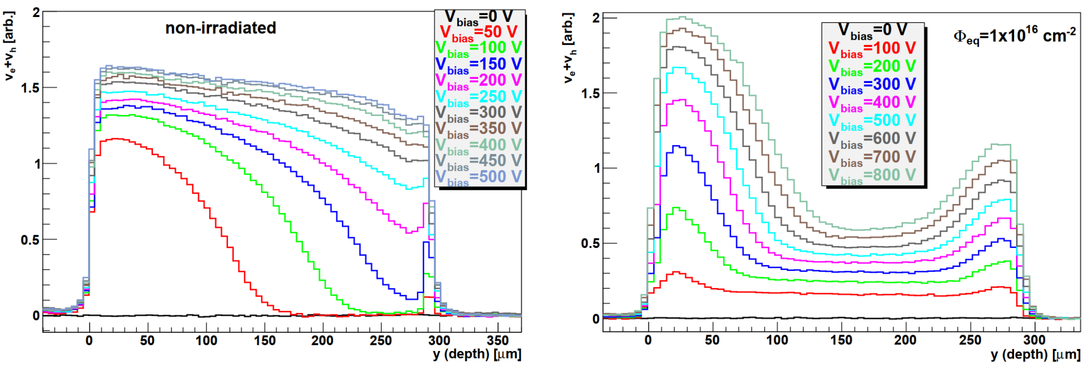
The theory describing the fluence dependency is called Hamburg model [75]. The individual components correspond to the stable damage $N_{\text{c}}(\Phi_{\text{eq}})$ from fixed displacement in the lattice, $N_{\text{a}}(\Phi_{\text{eq}}, T, t)$ and $N_{\text{y}}(\Phi_{\text{eq}}, T, t)$ represent short-term benefical and long-term reverse annealing, respectively. The components are dependent on the time t passed from the irradiation and temperature T of the material, described in more detail in Sect. 2.3.2.4.

The Hamburg model is valid up to the fluence $\approx 1 \cdot 10^{15}\,1\,\text{MeV}\,\text{n}_{\text{eq}}\,\text{cm}^{-2}$ after which the dependency of N_{eff} starts to saturate. [78] This effect is crucial for the VELO Upgrade sensors, as the estimated fluence at the end of life is up to $1 \cdot 10^{16}\,1\,\text{MeV}\,\text{n}_{\text{eq}}\,\text{cm}^{-2}$ [24]. However, it has been shown that even for such high radiation fields the silicon sensors can operate below 1000 V while having the MPV of the collected charge above $6\,\text{ke}^{-}$ [41].

The effect of the lowered effective charge N_{eff} comes from the formation of a double peak [79], caused by the trapping of the free carriers near their respective collection electrodes. This can be shown by the change of the electric field profile, plotted using the carrier velocity profiles (see Fig. 2.15). In the non-irradiated sensor the electric field grows from one side, while for a strongly irradiated sensor it grows from both front and back. [78]

2.3.2.3 Charge trapping

The recombination of the charge carriers caused by the created traps decreases their generation lifetime τ_{g}. The measured signal is therefore lowered in case of a long trapping lifetime τ_{t} with respect to the duration of the readout. This results in a decrease of the



(a) : Not irradiated. **(b) :** $\Phi_{\rm eq} = 1 \cdot 10^{16}\,1\,\rm MeV\,n_{eq}\,cm^{-2}$

Figure 2.15: Velocity profiles of electron-hole pairs for different bias voltages and irradiation. [78].

CCE of the detector.

The overall impact on the rate can be estimated as

$$\frac{1}{\tau_{\rm g}} = \frac{1}{\tau_{\rm g,\Phi_{eq}=0}} + \frac{1}{\tau_{\rm t}}, \tag{2.50}$$

where

$$\frac{1}{\tau_{\rm t}} = \beta\Phi_{\rm eq}. \tag{2.51}$$

Similarly to the Eq. (2.47), the β is an experimentally measured constant, the trapping rate scales linearly with the fluence $\Phi_{\rm eq}$. It is important to note, that the trapping time depends on the type of the charge carrier. [80]

2.3.2.4 Annealing

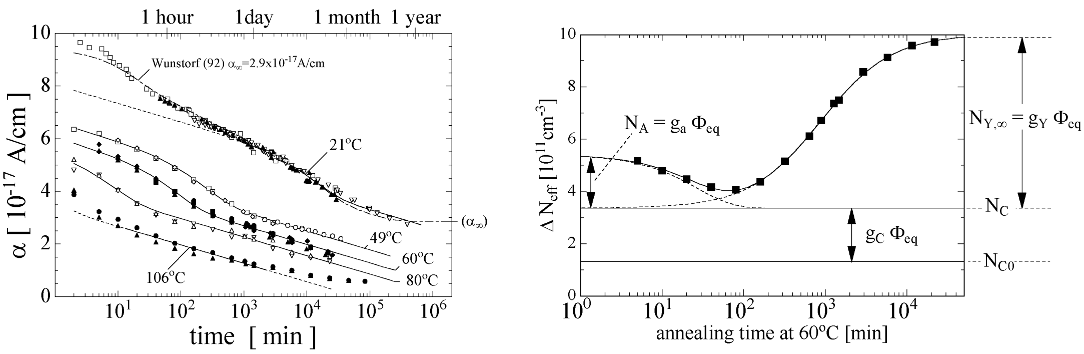
Since some radiation induced damage is not fixed, the resulting effects may change over time. This process is called annealing, an example of change of the current-related damage rate based on the sample temperature and time from the irradiation is in Fig. 2.16a.

The time-dependent changes may be described using the Hamburg model in Eq. (2.49), describing change of the concentration of the carrier density after annealing time t. Contribution of the individual components is in Fig. 2.16b with the following definitions [75]:

Stable damage The first part does not depend on the annealing and can therefore have the biggest effect on the long-term performance of the silicon sensors in strong radiation fields. Its dependency on fluence is written as

$$N_{\rm c}(\Phi_{\rm eq}) = N_{\rm c,0}(1 - {\rm e}^{-c\Phi_{\rm eq}}) + g_{\rm c}\Phi_{\rm eq}, \tag{2.52}$$

where the first term represents an incomplete donor removal resulting in a final concentration $N_{\rm c,0}$, the second term is an addition of stable acceptors.



(a) : Current-related damage rate as a function of sample temperature and time from the irradiation.

(b) : Dependency of change of the effective doping concentrations ΔN_{eff} on time from irradiation. Individual contributions based on the Hamburg model are highlighted.

Figure 2.16: Annealing behaviour of selected detector properties [75].

Short-term annealing The short term process is caused by removal of acceptors in the semiconductor and can be described as a sum of exponential terms. Only the part with the largest decay constant $\tau_{\mathrm{a}}(T)$ is usually considered, resulting in

$$N_{\mathrm{a}}(\Phi_{\mathrm{eq}}, T, t) \approx g_{\mathrm{a}} \Phi_{\mathrm{eq}} \mathrm{e}^{-\frac{t}{\tau_{\mathrm{a}}(T)}}. \tag{2.53}$$

Reverse annealing The last component is responsible for the change of the depletion voltage after longer time-periods $\mathcal{O}(1\,\mathrm{week})$. It is caused by the build-up of acceptors and can be written as

$$N_{\mathrm{y}}(\Phi_{\mathrm{eq}}, T, t) = g_{\mathrm{y}} \Phi_{\mathrm{eq}} \left(1 - \frac{1}{1 + \frac{t}{\tau_{\mathrm{y}}(T)}}\right), \tag{2.54}$$

with g_{y} proportionality and $\tau_{\mathrm{y}}(T)$ decay constants.

The decay constants from the previous description follow the Arrhenius relation

$$\frac{1}{\tau_x} = k_x = k_{x,0} \mathrm{e}^{-\frac{E_x}{k_{\mathrm{B}} T}}, \tag{2.55}$$

where k_x is a frequency factor and E_x represents the activation energy.

2.4 Summary

Semiconductor detectors present an excellent way to measure the interaction of ionising radiation. This is namely due to the properties of a reverse-biased p-n junction where a depletion region which is empty of free charge carriers gets created. Based on the amount of the deposited energy, a number of electron hole pairs is created along the particle track in this region from where they drift towards their respective electrodes. The HEP experiments primarily utilise such devices for tracking, segmenting the sensors and reading out each individual channel with dedicated ASICs. The detectors are usually operated in strong radiation environments, therefore the radiation induced effects have to be understood in order to achieve the best performance throughout the device lifetime.

Blank page

Chapter 3

VELO Upgrade modules

As mentioned in the Sect. 1.3, the basic building block of the VELO Upgrade detector is a module. These are assembled at both the UoM (Manchester, UK) and Nikhef (Amsterdam, NL), requiring high precision in many aspects. For this purpose extensive procedures for their construction and QA have been deployed. This includes visual inspections, metrology, electrical communication tests and thermal performance assessment of the module, the checks are done at every stage of the production process. Final assembly of the whole VELO system then takes place at the University of Liverpool (UoL) (Liverpool, UK) from where it gets transported to the CERN (Geneva, CH).

The assembly, metrology and testing processes are all performed in a Clean Room (CR). To ensure repeatability, every step has to follow a pre-agreed set of instructions and the results are recorded. The information is uploaded to the online database and automatically analysed, providing instantaneous information about the quality of the components and the performed tasks. The database is discussed in depth in the Chapter 4. It also provides a safe off-site backup of data about each of the modules, which can be stored for the long term.

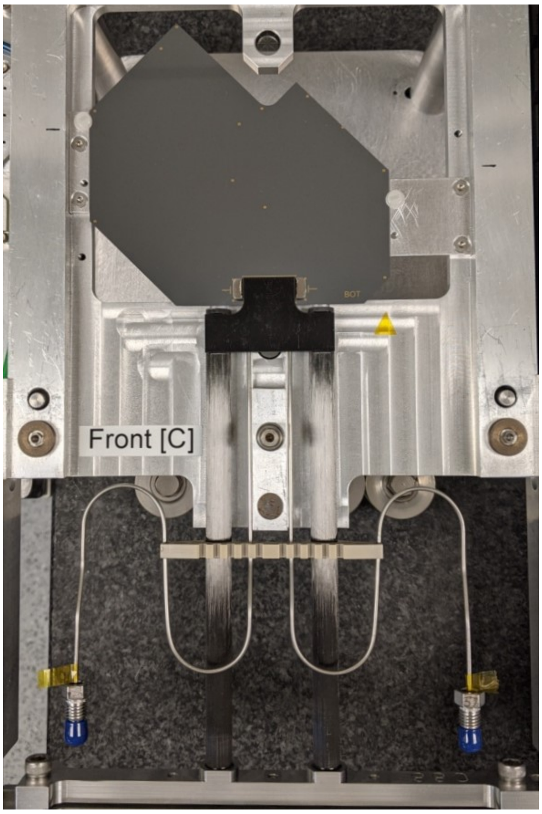
3.1 Assembly

The module assembly consists of four major steps shown in Fig. 3.1 and described in more detail below. The description is based on the construction processes performed at the UoM, those at Nikhef are similar but vary in some details.

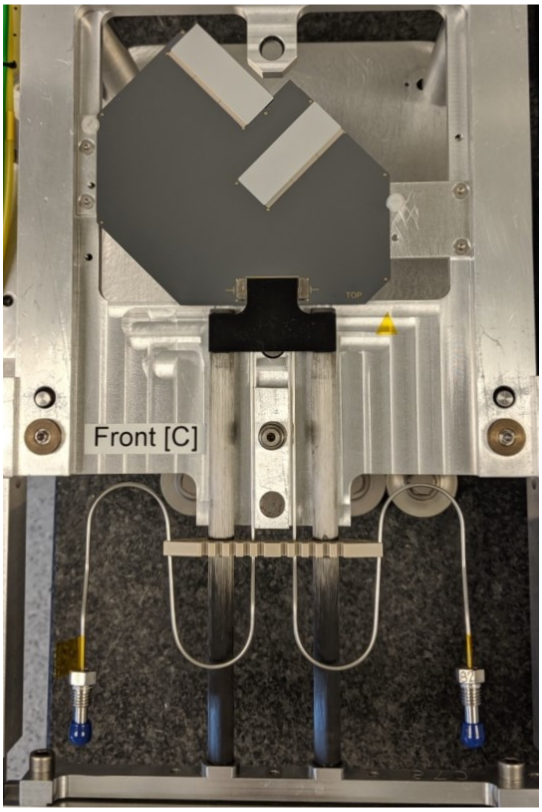
3.1.1 Bare module

The assembly starts with the making of a bare module, based on a silicon microchannel cooling substrate. The substrate is responsible for both heat removal and mechanical support for the readout components. More details about the design and performance of the cooling design are in Sect. 1.3.2. The assembled bare module is shown in Fig. 3.1a

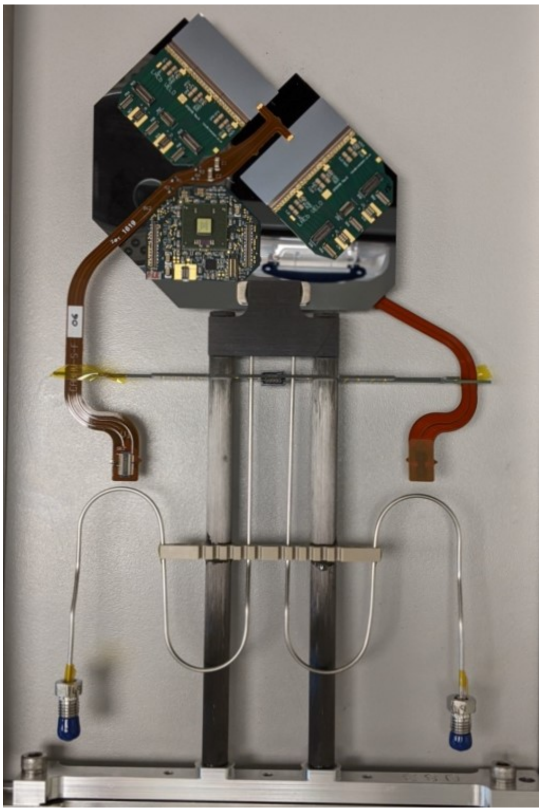
The base of the module is made from an aluminium foot with a steel dowel pin on one side, used for alignment of the reference systems. Two carbon fibre legs are then glued to the foot and everything is inserted into a custom-made building frame. Afterwards the



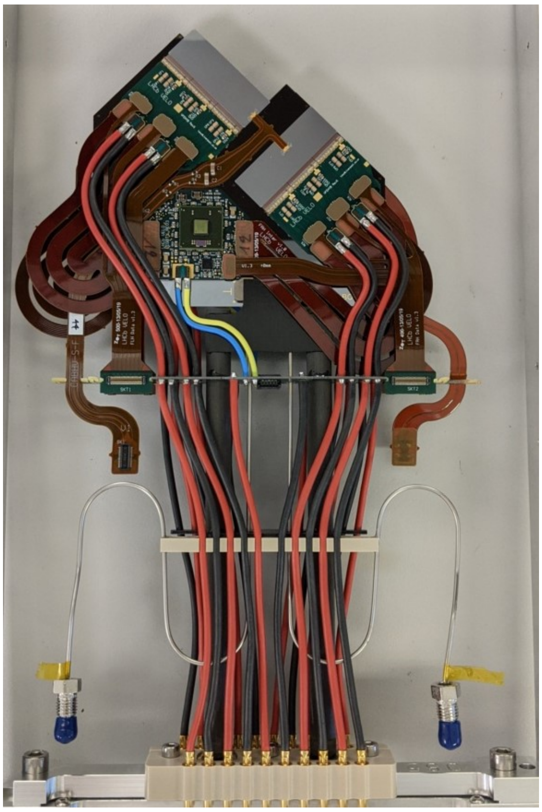
(a) : Bare module construction.



(b) : Tile attachment.



(c) : Hybrid attachment and wire-bonding.



(d) : Final cabling.

Figure 3.1: Assembly steps of the VELO upgrade modules.

cooling substrate along with the cooling pipe clamp is added. The clamp is responsible for fixing the pipes to the legs and minimizing the propagation and thermal stress onto the substrate from the rest of the cooling system. Finally the substrate is aligned using pins on the building frame and glued, using a piece of carbon fibre – mid-plate – to hold on the foot.

After curing, the module is placed into another custom-made turn-plate frame, which contains two glass precision markers visible from both sides of the module. Their position is known with respect to the steel dowel pin, providing a transformation between the LHCb and the frame reference systems described in Sect. 3.2.0.1. In order to spot any defects of the substrate which could have occurred during the assembly, it's flatness is measured as described in Sect. 3.2.1 and data analysed according to Sect. 4.3.1.

3.1.2 Tiles attachment

Placement of the tiles on the modules is one of the tasks requiring the highest accuracy during the module construction, aiming at achieving a precision better than 30 µm. Furthermore, the thickness of the glue layer used for the tile attachment affects the final cooling performance of the whole module. The temperature of the tip of the overhanging tile is the most critical element, the targeted glue layer thickness underneath each tile is 80 µm. While the production steps are described further in the text, the metrology measurements are discussed in detail in Sect. 3.2.2 and 3.2.3 and their results in Sect. 4.3.2 and 4.3.3.

3.1.2.1 Alignment

The tiles are first placed in pair on a dedicated alignment system, held in place using vacuum. The system consists of two translation and one rotation stage per tile and uses a camera for pattern recognition. The alignment process then locates the markers etched on the ASICs and uses the information to automatically move and rotate the individual stages. Afterwards the two tiles are moved to a dedicated transfer-plate frame by first turning on the vacuum on it and then disabling the vacuum on the alignment system, the success of the transfer verified visually as the transfer-plate contains holes near the locations of the ASIC markers. The whole process is repeated for the second set of tiles.

3.1.2.2 Gluing

For obtaining the best cooling performance of all VeloPix ASICs, a thin uniform layer of glue has to be deposited. The glue deposition is performed using a programmable robot, a Fisnar F5200N [81], with a syringe and a needle and fluid dispenser running on compressed air. The robot is shown in Fig. 3.2. Due to the system size constraints, the process is done in two parts using the pairs of tiles held by transfer-plates. For repeatability, the transfer-plates are held in place using precision dowel pins on a mechanical breadboard.

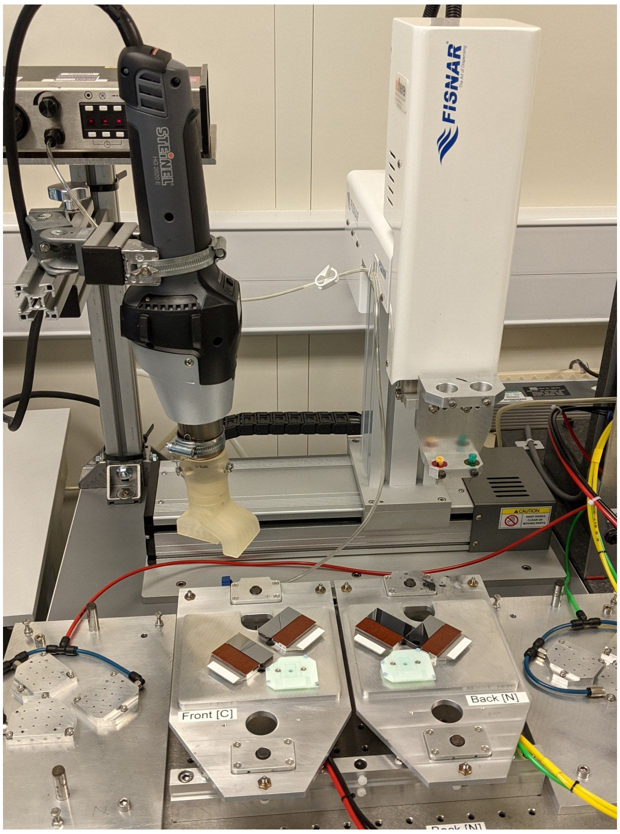


Figure 3.2: Gluing robot Fisnar F5200N with a heat gun.

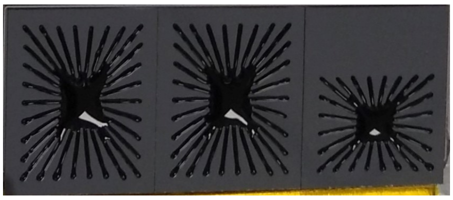
Different types of glue as well as deposition patterns had to be investigated in order to achieve the desired performance. This required an extensive R&D process.

Pattern Since the detector is operated in vacuum, air pockets trapped in the glue layer could cause damage and need to be avoided.

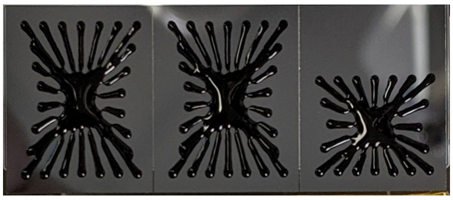
A snake deposition pattern (see Fig. 3.3a) was proposed at first, enabling the air to escape through the openings while keeping full coverage. However, if the points near the openings closed before all air could escape, an air pocket could have been created. For this reason a star shaped pattern (see Fig. 3.3b) was tested and has been selected as it is inherently better at preventing occurrence of the air bubbles. Another advantage of the star pattern is a smaller chance of glue spillage, as there is less glue deposited in the outer regions of the ASIC, while still providing the full coverage needed for the heat transfer. Small amounts of glue were still occasionally spilled or got very close to the inner gaps between the ASICs, as is visible in Fig. 3.4. The pattern was therefore further enhanced by concavely shaping the inner points (see Fig. 3.3c).

Since the gluing utilizes a programmable robot, the 3D cartesian coordinates as well as line speeds have to be defined in advance. The robot moves the syringe in lines connecting each two consecutive coordinates, depositing glue as programmed. To ease the usage of the robot, a simple software for generating patterns in the correct

(a) : Snake pattern.



(b) : Star pattern.



(c) : Star pattern with concave inner sides.

Figure 3.3: Comparison of glue patterns deposited on tiles.

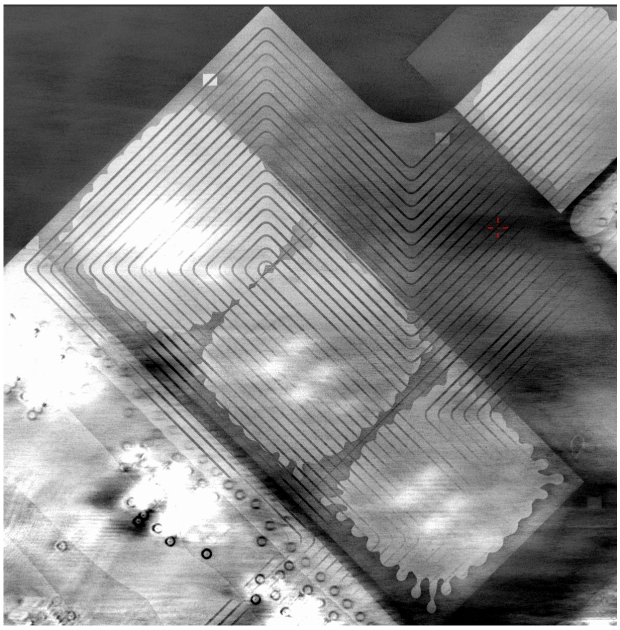


Figure 3.4: CT scan of tiles attached to the cooling substrate using simple star pattern. The small lines are individual channels in the silicon substrate. [35]

format has been developed. It was mainly used during the optimizations steps of the star shape pattern as it provided options to quickly define line speeds and the number of lines.

Glue Stycast 2850FT with catalyst 9 was selected as the preferred adhesive based on the irradiation to the same fluence as is expected in the innermost VELO region and shear-force testing [82]. However, it was later discovered that the mechanical connection of the glue was not ideal, leading to a failure of adhesion on an early module. For this reason an extensive study, in collaboration with the CERN and Nikhef groups, was performed, testing new types of glue with respect to their mechanical strength, radiation hardness and thermal performance. The glue agreed upon for the tile attachment is Stycast 2850FT, the same as originally proposed, but with catalyst 23LV. The main improvement being the flexibility of the adhesive, better withstanding the bending forces.

In order to normalize the glue properties before every use, the container with epoxy is first heated up to 50 °C for one hour and left to cool down to a room temperature. This also ensures proper mixing of the aluminium oxide filler in it.

A transfer-plate with tiles is put in place and the robot is manually calibrated so that the tip of the needle is set at 200 µm distance from the surface of the tiles. Afterwards the glue is mixed in a planetary mixer, inserted into the syringe and its viscosity is monitored (based on the mass of glue dispensed in a given time). The robot is then started, depositing glue in a predefined pattern with the line speed manually set based on the viscosity, such that a consistent amount of glue is deposited in every module. The desired glue thickness is about 80 µm, both to absorb the $\mathcal{O}(10\,\mu\text{m})$ non-uniformity of the cooling substrate, to provide enough flexibility due to the different CTE between the epoxy and the silicon, and to minimise the chance of damaging the silicon due to the size of the filler particle (max 5 % of the particles are bigger than 45 µm of which max 1 % are bigger than 63 µm [83]). When a whole side is finished, a heat gun pointing towards the glue is turned on to 60 °C in order to remove any adsorbed water as it was found to cause the adhesion failure during the glue R&D. Within a minute from the heating, the transfer-plate is placed on the turn-plate, held in place with retention springs.

The whole process is then repeated for the second transfer-plate, adjusting the glue deposition speed using current viscosity, the vacuum on the plates is turned off only after the glue is finished curing. The module after the tiles are attached is shown in Fig. 3.1b.

3.1.3 Hybrid attachment and wire-bonding

The process of attaching hybrids is very similar to that of the tiles. However, this does not require such high placement precision. After gluing, the FE hybrids and HV tapes are connected using wire-bonds to the tiles. As part of the metrology, the attachment is visually inspected (see Sect. 3.2.4), the wire-bonds are pull tested and analysed following the description in Sect. 3.2.5 and Sect. 4.3.4, respectively.

3.1.3.1 Alignment

The FE hybrids as well as the GBTx hybrid are put on a dedicated alignment frame where they are mechanically aligned using set of pins. This achieves precision of about 100 µm. They are then held in place using vacuum and transferred to the transfer-plates, similarly to the process of tile attachment.

3.1.3.2 Gluing

The main requirements on the gluing process were good radiation hardness and flexibility. The flexibility of the adhesive is due to the different CTE of the silicon substrate and the hybrid PCB, namely its copper layers. Since the cooling of these components is not as critical as of the tiles (less produced heat), Loctite 5145 has been chosen. Deposition is done using the same Fisnar F5200N robot but with different types of syringe and needle.

Pattern As it is not necessary to cover the whole area of the hybrids by glue, a smaller number of lines has been used for the attachment. The FE hybrids have 5 evenly spaced lines, the most important being beneath the bond-pads in order to provide mechanical stability for the wire-bonding process. The GBTx hybrid is attached using 9 lines, closer together compared to the FE, in order to provide some heat transfer capabilities.

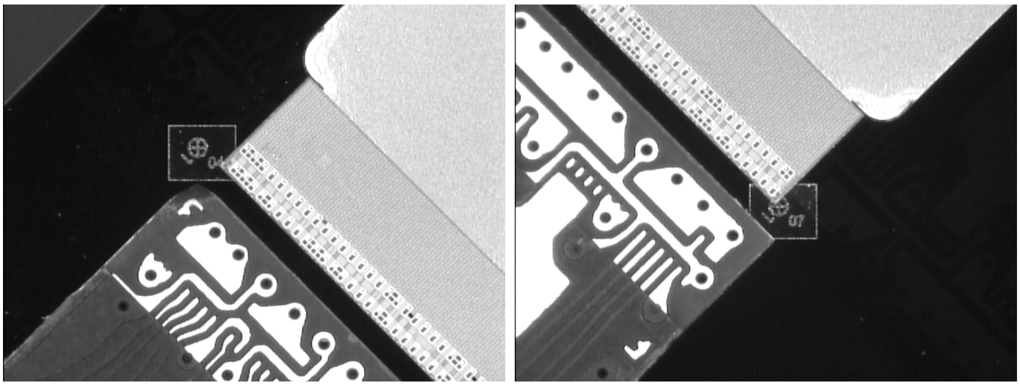


Figure 3.5: Zoomed in image of a glued FE hybrid next to a CLI tile, showing both corners.

The glue is put inside the syringe and the robot program execution is started. When the deposition is finished for all hybrids, the transfer-plates holding them are again fixed to the turn-plate with the module and left to cure. Afterwards the vacuum on the transfer-plates is switched off. A detailed view of the FE hybrid and a tile is shown in Fig. 3.5. The module at the end of these steps is shown in Fig. 3.1c.

3.1.3.3 Wire-bonding and HV cabling

The module is put into a dedicated holder and the HV cable is glued in using drops of Araldite 2012. The glue is left to cure and the module is fixed in place using a vacuum. The bonder is aligned using image recognition and checked, then the automatic program is executed bonding a total of 420 bonds per tile. Any lifted bonds are manually re-bonded when the program finishes.

Afterwards the HV tape is manually bonded to the sensor surface. This process is repeated for the other side of the module.

3.1.4 Cabling

The final step in the module assembly is attaching all of the remaining cables. These provide both LV power to the module components and bi-directional communication from the VeloPix ASICs and GBTx hybrids out to the OPB.

3.1.4.1 Interconnect and data cables

The cables used are thin tape-like copper cables with connectors at their ends. Matching connectors are placed at both FE and GBTx hybrid PCBs. Since each connection establishes a large number of data links, the plugging in requires much attention.

Before attachment, the cables are plasma cleaned and the module is held in the same mechanical jig as is used for the wire-bonding. To support the module during the process, a set of 3D printed parts is positioned on a base plate to which the frame with the module is attached. Starting on the C-side, the interconnect cables are attached, linking the FE with GBTx hybrids. The process is repeated for the other side, followed by the attachment of the data cables before returning to the C-side and finishing the data cable connection.

3.1.4.2 LV cables

To simplify the connection of all of the needed LV cables, an externally manufactured harness is used. It is made of silicon coated copper cables, a PCB transition piece and connectors.

The harness is attached to the module in the same frame as the interconnect and data cables. The transition piece is glued to the carbon fibre legs where it establishes a mechanical link while changing from thinner gauge cables close to the readout, to thicker ones going back to the module foot. The LV cables pins are inserted to the foot connector before establishing the rest of the connections on the hybrid PCBs. The finished module is shown in Fig. 3.1d.

3.2 Metrology

Every component is graded when produced and tested, however, the final module grade and quality might be affected by either an imprecision during some of the assembly steps, non-ideal contact or other unforeseen effects. To counter this, every assembly step has a corresponding metrology check which helps with detection of the issues. All of these tests are done at both assembly sites, differing only in the tools used for the measurements. The description will focus on the ones performed at UoM, the data analysis and grading described in Sect. 4.3.

3.2.0.1 SmartScope

The SmartScope [84] (see Fig. 3.6) is one of the most widely used devices for the metrology tests in UoM. It utilizes an optical system for precise 3D measurements. The device can be controlled both by a controller and Measure-X software [85]. An automatic pre-programmed routine can be defined in the device, requiring only a minor supervision with a rough alignment of the components done by the user.

The device contains a flat plane with a backlight moving in the x and y coordinates and a camera with a laser, attached on a z positioning mechanism. Two additional light sources around the camera, and below the translucent plane, help in setting the ideal lighting conditions for the measurement. Any measured object has to be fixed on the surface of the xy plane, which is firmly attached using screws on the sides of the backlit panel.

For the optical measurements, the SmartScope can record data with the precision of 2 µm in the xy plane and 5 µm in z. It also contains feature-finding methods to automatically locate shapes according to their different contrasts or shades. This enables

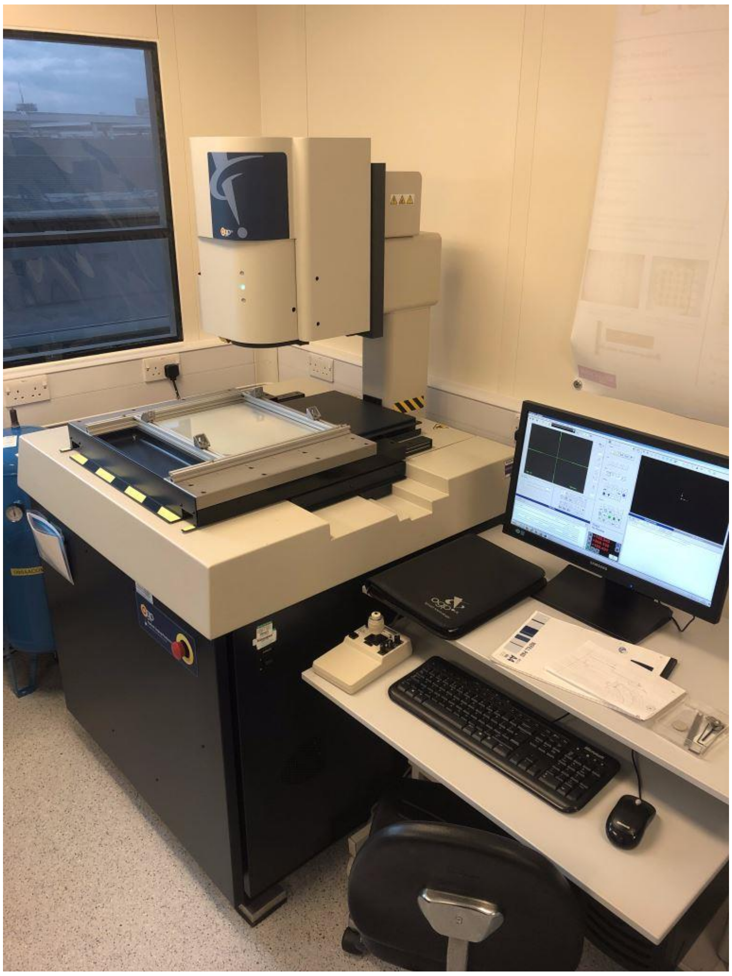


Figure 3.6: SmartScope device with laser and optical sensors used for mechanical measurements.

the precise location of lines or circles which can be used as an input to calculate more complex geometrical shapes and intersects.

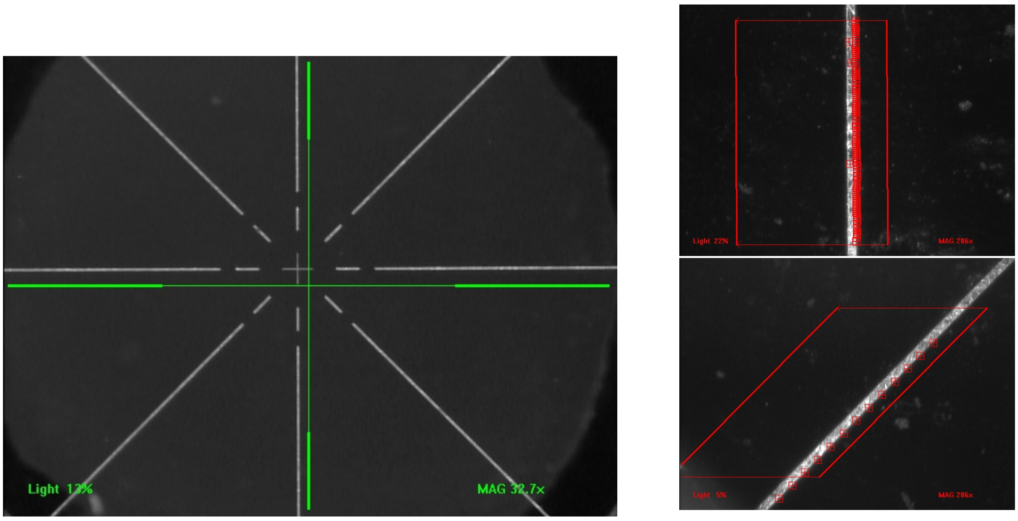
The laser is best suited to scan over larger areas with many measurements points along its path. It is an independent measurement module, its offset automatically accounted for by the software both during the execution of the routine as well as in the data output.

The SmartScope is used for the following measurements at the UoM:

- substrate flatness,
- glue layer thickness / tile flatness,
- tile position,
- hybrid position.

It has also been used for investigation of issues such as glue spillage or visual artifacts on the surface of the used objects.

For the measurement purposes the module is kept in the turn-plate, using its two glass markers to establish a reference system. The frame is fixed by a set of dowel pins in a



(a) : Manual alignment of the marker.

(b) : Automatic fitting of the vertical and diagonal lines.

Figure 3.7: SmartScope marker alignment.

dedicated holder which is screwed into the SmartScope. To better relate the measurement to the LHCb coordinate system, a transformation had to be identified.

Coordinate transformation The common point of the two reference systems is the dowel pin on the module foot. To find the transformation, first a measurement of the glass markers on the empty turn-plate is performed. The marker is at first manually aligned (Fig. 3.7a) then sides of both diagonal and vertical line are automatically fitted (Fig. 3.7b), their intersections constructed and the precise point is set as the centre of the circle containing all four intersects. The same process is repeated for the other marker.

Afterwards, the dowel pin and the bottom part of the frame are measured in a similar manner. Intersections between the line and two perpendicular pin lines are constructed, the final position is defined as their midpoint. Using the Computer-Aided Design (CAD) model of the frame, a reference system between the pin and the frame markers is defined. Since the dowel pin's position is matched to the module foot in the LHCb coordinate system, the final transformation is obtained.

Due to small imperfections in positioning the frames, locating the lines and optical effects on the markers, the measurement has to be repeated for each side.

In order to simplify the measurements for the operator, the transformation is automatically applied to the SmartScope. The metrology tests start by measuring the glass markers on the turn-plate as described above, this time with a module inserted. Two virtual points are then constructed based on the measured transformation, representing the LHCb reference point and one point along the vertical axis. These are then used to match the SmartScope reference system with the LHCb one, ensuring all subsequent measurements are recorded directly in the desired reference system.

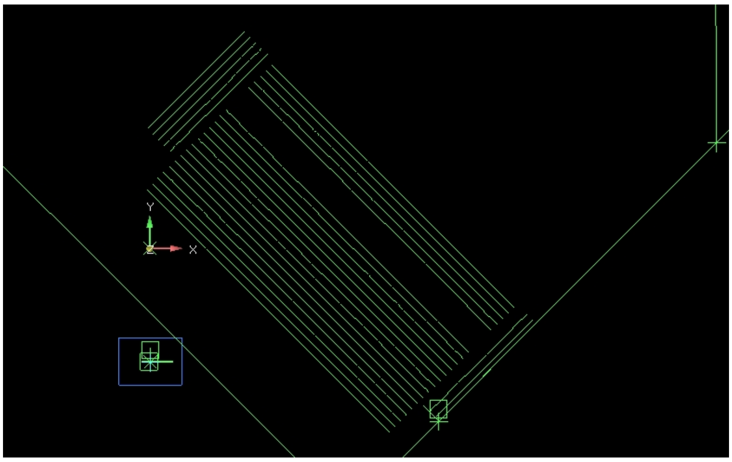


Figure 3.8: SmartScope routine to measure glue thickness and tile flatness. Groups of lines represent measurements of the tile and surrounding cooling substrate.

3.2.1 Substrate flatness

Flatness measurement of the cooling substrate is performed in order to verify the quality and spot any damage that could have happened during transport or storage of the substrates. It is done right after the bare module assembly step.

The data are taken using the laser system of the SmartScope. The routine at first automatically locates the outline of the substrate, using the information to level the plane to provide a better starting point for the scan. Subsequently, the laser scan is performed in vertical lines, saving hundreds of x, y and z data points to a file. The analysis of the data from the scan is described in Sect. 4.3.1.

3.2.2 Glue layer thickness and tile flatness

Measurement of the glue layer thickness underneath the tiles is performed in order to check the amount of deposited glue, an important factor for the cooling performance and heat transfer from the tiles. As a by-product of this test, the flatness of the tiles is also measured. The step is done between the attachment of the tiles and hybrids.

The test is done in a similar manner to the substrate flatness, with the one difference that the scan is repeated individually for each tile. The laser-scanned area covers the tile as well as its surrounding cooling substrate (see Fig. 3.8), saving x, y and z data to a file, results are in Sect. 4.3.2.

3.2.3 Tile position

The exact position of all tiles is crucial for a better starting point of the detector alignment from data in the LHCb experiment.

The SmartScope measurement is based on automatically locating the centroids of a number of fiducials etched into both ASICs and sensor. Since this test relies only on the contrast differences of the etching, the operator has to pay close attention to the routine execution. The selected markers are in the bottom left and right corners of the tiles as well as on the back side of the sensors (where visible due to overhang). The data-stream again saves the x, y and z of the points to a file, the analysis then performed in the online database with the process described in Sect. 4.3.3.

3.2.4 Hybrid position

After the FE and GBTx hybrids are attached, the alignment of the FE hybrids and tiles has to be checked. This step helps with locating any potential glue spillage or offset, in order to provide as much information for the wire-bonding as possible. The check is done by taking zoomed-in pictures using the SmartScope, see Fig. 3.5.

3.2.5 Wire-bond pull testing

In order to verify the strength of the attachment of the wire-bonds, each ASIC contains 4 sacrificial ones which do not carry any signal or LV. These bonds are pulled using a dedicated device, which uses a small hook to catch and pull on the bonds and records the breaking force. The parameters of the break are then uploaded to the database and analysed as written in Sect. 4.3.4.

3.2.6 Photo inspection

Every assembly step is followed by an extensive photo documentation, where detailed pictures of the module are taken. This information is useful for spotting macroscopic artifacts and verifying the quality of connectors, cables or even glue deposition patterns.

3.3 Module installation

Given that the assembly and metrology of the module succeeded, it still does not guarantee the final performance will be ideal. The module has to be electrically tested in vacuum while being cooled down to $-30\,°\mathrm{C}$. A number of QA tests are done to make sure the connections are good and that the module works well under the nominal conditions.

3.3.1 Test setup

For the testing purpose a small vacuum tank connected to a cooling system has been built, a CAD model with a module attached is shown in Fig. 3.9. The electrical part of the test setup will be discussed later in the Sect. 3.4.

3.3.1.1 Vacuum

The vacuum tank is split in to two parts, sealed with a large rubber o-ring and screwed into place. The evacuation and pressure measurement is done using a Pfeiffer Vacuum

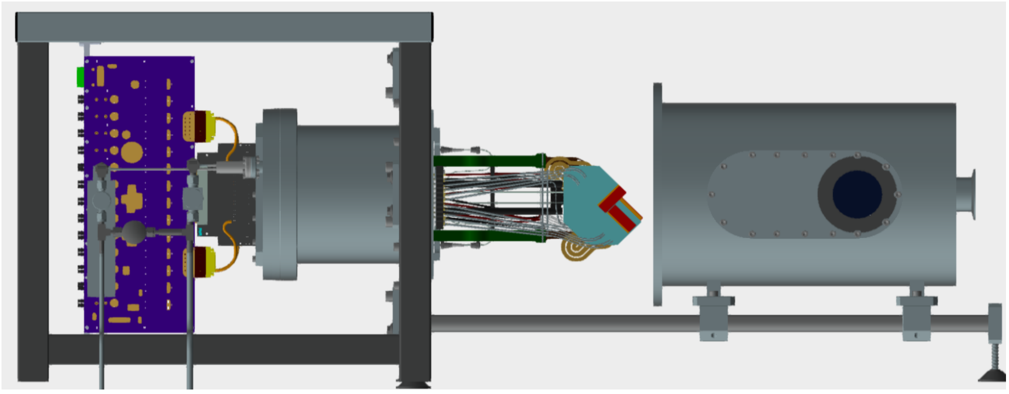


Figure 3.9: CAD model of the module test setup in UoM.

D-35614 Asslar Turbo Pump (see Fig. 3.10). The pump is toggled and read-out manually with the desired pressure lower than $1 \cdot 10^{-3}$ bar.

The first half is connected to the electronics system via a number of dedicated vacuum-feed-throughs, enabling transfer of data from both displacement and thermal probes, as well as all of the signals, LV and HV connections coming to the module. It also contains a frame where the module is securely attached to.

The second half is an empty cylinder movable on two rails, attached to the vacuum pump via flexible pipe. Each side of the cylinder contains a small germanium window, providing visual access for a thermal camera.

3.3.1.2 Cooling

The cooling of the module is achieved by a manually operated CO_2 machine, shown with its schematic view in Fig. 3.11. The temperatures at various points of the system as well as pressure and flow are monitored by a PC. For safety reasons, the pressure and flow in the whole system is also monitored using gauges on the machine itself.

For cooling purposes a bottle with liquid CO_2 is used, the desired conditions are set by adjusting valves. The system works in a blow-through mode, therefore the used CO_2 is heated up and transitioned into a gas phase before being released into the exhaust tube leading to the atmosphere. Since CO_2 in larger concentrations might be lethal, the machine has to always have at least two operators closely following the defined precautions and checking the CO_2 monitor placed in the CR.

The standard process of running the cooling machine begins with evacuation of the cooling pipes using a separate vacuum pump. This prevents blockages of the small pipes with particles or frozen water after the CO_2 is let in. The attached bottle is then opened and CO_2 fills the whole system. Due to the restrictions in the microchannel cooling plates, the CO_2 begins to expand and transit into a gas phase. The back pressure after the module outlet is regulated to a defined value of around 12 bar. The CO_2 that leaves the module is then used to pre-cool the inlet and is afterwards fully converted to gas in a warm water bath, before being let out of the machine.

Figure 3.10: Pfeiffer Vacuum D-35614 Asslar Turbo Pump.

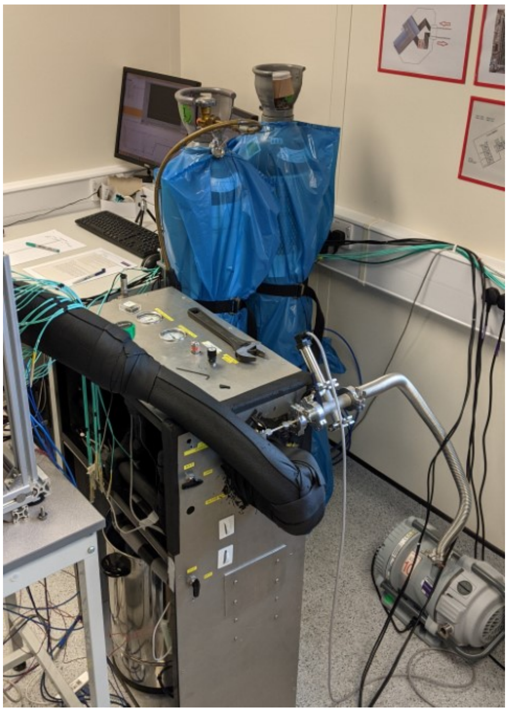
At first the CO_2 is flowing with restrictions setting the flow to $0.8\,\mathrm{g\,s^{-1}}$ at which point the system takes about 20 min to reach stable conditions. Afterwards a needle valve is used to restrict the flow to about $0.4\,\mathrm{g\,s^{-1}}$, defining the nominal conditions. Small adjustments to both the back-pressure regulator and the needle valve have to be made in order to trigger the ideal boiling conditions underneath all of the tiles.

When the testing is finished, the CO_2 bottle is fully closed and the system is slowly emptied of the residual CO_2 by fully opening the valves. Once the flow reaches $0.06\,\mathrm{g\,s^{-1}}$, the machine is considered fully shut down. With no heat generation inside the tank, the module might take about 25 min to fully warm up.

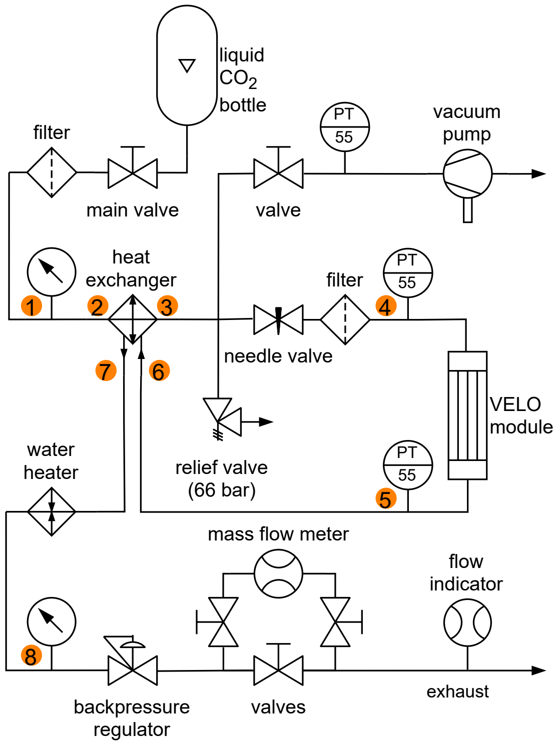
As a way of estimating amount of CO_2 used, a script is executed at the end which integrates the flow over time from the run. A full bottle has about 25 kg of liquid CO_2 , however, due to the lowered pressure the system becomes unstable at around 8 kg and the bottle has to be replaced.

3.3.2 Installation

Before the installation, the module is already stored in the Survey frame covered with plexiglass from both sides. The whole frame with the module is aligned using the LV connector with its matching part on the vacuum tank and then inserted. While held in hands, the frame is screwed into the tank and fixed in place. Finally the plexiglass covers are carefully removed.



(a) : Picture. The probes and the heat exchanger are inside the large box.



(b) : Schematic view. Numbers in orange represent temperature probes.

Figure 3.11: Liquid CO_2 cooling machine used for cooling the modules.

The module installed in the tank has at first only the LV connection made, both data and HV tapes still need to be attached. The data connections are carefully made using 3D printed bars guided with screws in order to minimize the chance of damaging either of the connectors.

3.3.3 Displacement

Due to various CTE of the module components and many points of attachment, it is expected that the module will thermally distort. Since the final data based alignment process of the whole VELO works well only for the xy plane, an estimate of the distortion in the z would help the process. As means of monitoring such distortions, two Micro-Epsilon capacitive displacement sensors are used, inserted in a dedicated holder which is directly attached to the Survey frame close to the tiles on the N-side. The sensors are placed to be about 500 µm away from the sensor surface. The displacement are read out and recorded in 1 s intervals using a LabView software.

The measurements are done during the following parts of the test:

- cooling connection,
- evacuation of the tank,
- cool down of the module.

Figure 3.12: GasCheck G3 leak detector used to check for CO_2 leaks.

The analysis is described in Sect. 4.3.5. The final z position in the assembled detector halves at room temperature still has to be measured in the UoL, once the modules are assembled into the VELO.

3.3.4 Cooling connection and leak-check

Due to the operation in vacuum and usage of liquid CO_2 for the cooling, all of the connections have to be leak tight. The module is connected to the rest of the system using Swagelok Vacuum Coupling Radiation (VCR) fittings.

In order to check for any CO_2 leaks, the whole process of cooling down the module is started, however, the CO_2 bottle is opened only for a short duration to reach the back pressure 20 bar. The valve on the machine is then adjusted to keep the pressure at a constant value. At this point the GasCheck G3 leak detector , with a minimum sensitivity to CO_2 of $4 \cdot 10^{-5}\,\mathrm{cs}^3$ [86] (see Fig. 3.12), is used to verify that none of the connections leak, otherwise the VCRs are tightened. The remaining CO_2 is then slowly released.

In order to protect the tiles against a thermal runaway, the Negative Temperature Coefficient (NTC) thermistor sensors placed on the FE hybrid are monitored. The NTCs are connected to an interlock box developed for the ATLAS experiment [87] which is defined to operate up to 40 °C (the range is extended from the default 30 °C by resistors soldered in parallel). In case of reaching the maximal value, the interlock cuts the power to the LV PSU. The temperature values are read-out and saved using Universal Serial Bus (USB) connection to the same LabView software as is used by the displacement sensors.

3.3.5 Thermal cycling

During the operation of the LHCb experiment, the VELO will normally be kept cold to reduce radiation damage, but has to withstand a number of cycles of heating and cool-downs from room temperature for maintenance or in the case of an incident. To simulate the possible effects and monitor any impact on the performance, a set of 10 thermal cycles ranging from 20 °C to −30 °C is performed. The process consists of the full manual cool-down at 6.5 °C min^{-1} and warm-up at 2.5 °C min^{-1}, each cycle takes about 25 min. The only difference with respect to the standard cooling machine operation is in not restricting the flow to 0.4 g s^{-1} but starting the warm-up process straight away.

3.4 Module performance testing

In order to test the electrical performance of the modules in circumstances as closest as possible to the final VELO system, a copy of the full readout chain has been built in CERN, Nikhef and UoM. The setups were cross-validated using modules produced at both construction sites.

The module tests mentioned in this chapter are done both before and after Thermal Cycling (TC). This helps with spotting any issues that could be triggered by the large temperature variation. The results and analysis of the individual tests is described in Sect. 4.4.

3.4.0.1 Software

The testing of the modules is done using SIMATIC WinCC/OA, described in more details in Sect. 1.2.6. The control components specific for the VELO have been developed mostly by the CERN team, however, some development and modifications were done in UoM as well.

Most of the testing is done from the main FSM VELO module panel (see Fig. 3.13), providing tools to configure and monitor all of the components. The following modes are used, each serving a different purpose:

1 – Low power Module operation with reduced analogue currents (minimizes the corresponding Digital-Analog Converters (DACs) settings).

2 – Minimal/QA Default startup mode enabling module monitoring and control.

3 – PRBS Configuration of the VeloPix ASICs periphery to generate the Pseudo-Random Bit Sequence (PRBS) pattern and opens the shutter. The latter increases the power consumption of the tile due to the activity of the digital part of the FE.

4 – GWT (Test pulse) Similar to the PRBS mode, however, the periphery is configured to transmit SPPs and the pixels matrix is configured as well. Mode used for test-pulse generation in the VeloPix ASICs.

5 – GWT (Physics) Nominal operation mode of the module for the data taking, almost identical to the test pulses (apart from their generation).

Individual tests are described below in more detail.

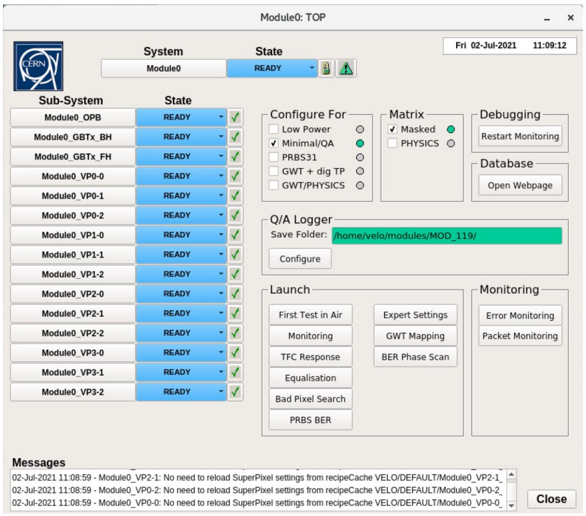


Figure 3.13: Main VELO WinCC/OA panel used for FSM control.

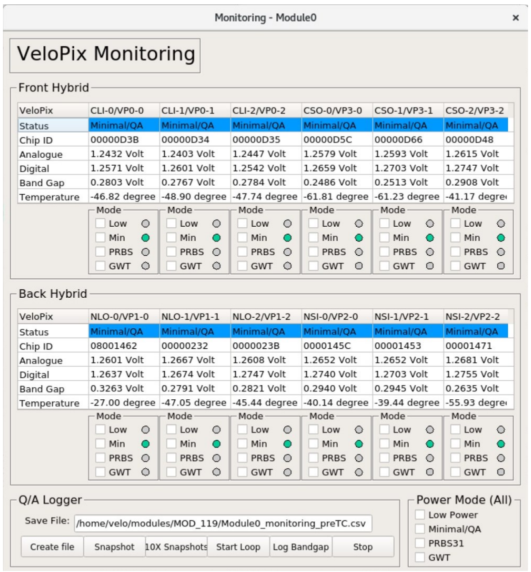


Figure 3.14: Module monitoring panel. Used for communication and cooling performance tests.

3.4.1 Communication

In order to make sure the readout chain is properly established (no broken ASICs, wire-bonds or other connections), a simple read-back of the values is tested. Only the information from the ECS data stream is verified thus testing the path to the tiles via the GBTx. It is always the first performance test performed after the module has been installed in the tank.

To check the communication, the module is fully configured into the Minimal/QA mode via the FSM. Afterwards a module monitoring window is opened (see Fig. 3.14), showing detailed information about the ASICs – id, analogue, digital and bandgap voltages and temperatures. The operator has to make sure that all of the ASICs are present with the analogue and digital voltages above 1.2 V and record a few snapshots of the values. The results are logged and evaluated in the database (see Sect. 4.4.1). During the experiment operation, the status and voltage are automatically monitored in real-time using the WinCC/OA FSM software, warnings and errors are raised accordingly.

Since the test may be run very quickly (configuration of the module and testing takes $\mathcal{O}(10\,\mathrm{s})$), this is the only test that can be run without the cooling and the vacuum. This also presents an option to quickly spot and fix issues (bad data tape connection, broken GBTx or lifted wire-bonds), without having to evacuate the tank and cool down / warm up the module (taking $\approx$1 h).

3.4.2 TFC

Another part of the communication chain ensures that the whole experiment is properly synchronized and can be controlled. This is especially important since the hardware level

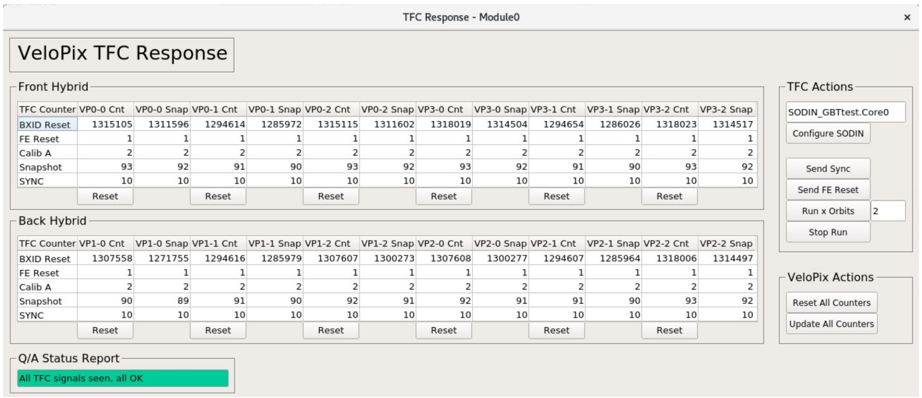


Figure 3.15: Module TFC panel.

trigger has been removed and all data are passed through for the software to process.

To test the performance a set of TFC commands are sent to the VeloPix ASICs and their response is monitored in panel shown in Fig. 3.15. Apart from the module being configured in Minimal/QA, the SODIN (the readout supervisor responsible for timing and command distribution) has to be configured as well.

BXID reset The reset is triggered once per LHC orbit, resetting the 12 bit bunch ID counter.

FE reset Triggers a reset of the VeloPix ASICs.

Calibration The VeloPix injects a test-pulse with a predefined pattern.

Snapshot Saves the current state of the TFC counters into dedicated registers, enabling their readout via ECS.

Synchronization A special frame is inserted into the outgoing packet stream, containing pattern 0x2BABE. The reception of the packet ensures proper synchronization between the FE and BE, this is especially needed due to fact that only the 9 Least Significant Bits (LSB) bits (512 unique numbers) are read from the bunch ID counter.

The BXID reset and Snapshot counters should increase over time, while the rest has to be manually triggered and the response recorded and evaluated (see Sect. 4.4.2).

3.4.3 PRBS

To speed-up the process during nominal DAQ, the data from the tiles bypass the GBTx and use dedicated links. Since any errors in the stream might corrupt the tracking and performance of the final experiment, the integrity is verified using a PRBS generated in the VeloPix ASIC.

The tiles are configured in the PRBS mode via the FSM, triggering generation of 31 bit long sequence (PRBS31). The stream is registered and decoded directly by the FPGA using the provided transceiver toolkit. A dedicated panel (see Fig. 3.16) is used to

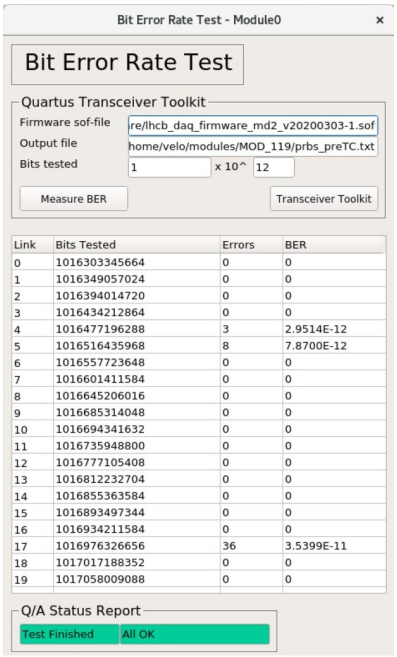


Figure 3.16: Module PRBS panel.

monitor and configure the settings showing the number of bits tested as well as Bit Error Rate (BER). The results are then uploaded to the database and analysed, see Sect. 4.4.3. As the desired BER is below $1 \cdot 10^{-12}$, only $1 \cdot 10^{12}$ bits are tested in order to shorten the test duration. Such scan takes about 5 min, longer duration would have a significant impact on the testing setup stability. Furthermore, the idea behind the scan is mostly to spot very noisy links as the proper phase-matching will be done with the fully assembled VELO.

The information from the PRBS test is not always sufficient in order to understand the performance issues. For this reason a dedicated tool can be used, combining the BER information with phase variation and thus producing eye diagrams.

3.4.4 Noise and Equalisation

The noise level in the ASIC is a key operational parameter, lowering the temperature reduces somewhat the electronic noise of the readout circuit. This affects the overall performance of the defined threshold and thus has to be characterized. Furthermore, the response of each pixel slightly varies due to small imperfections of the silicon used for both sensor and ASIC, causing different signal response given the same deposited energy and threshold setting. To counter the effect, a threshold trim value specific for each pixel may be set thus equalising the response of the whole matrix.

The test is performed automatically in Minimal/QA by running a global threshold scan covering the range of the electronic noise, while biasing the sensor to the nominal -140 V. Such a scan is repeated twice, both for the minimal (trim-0) and the maximal (trim-F) configuration of all of the pixels. The hit count response of the whole pixel matrix is

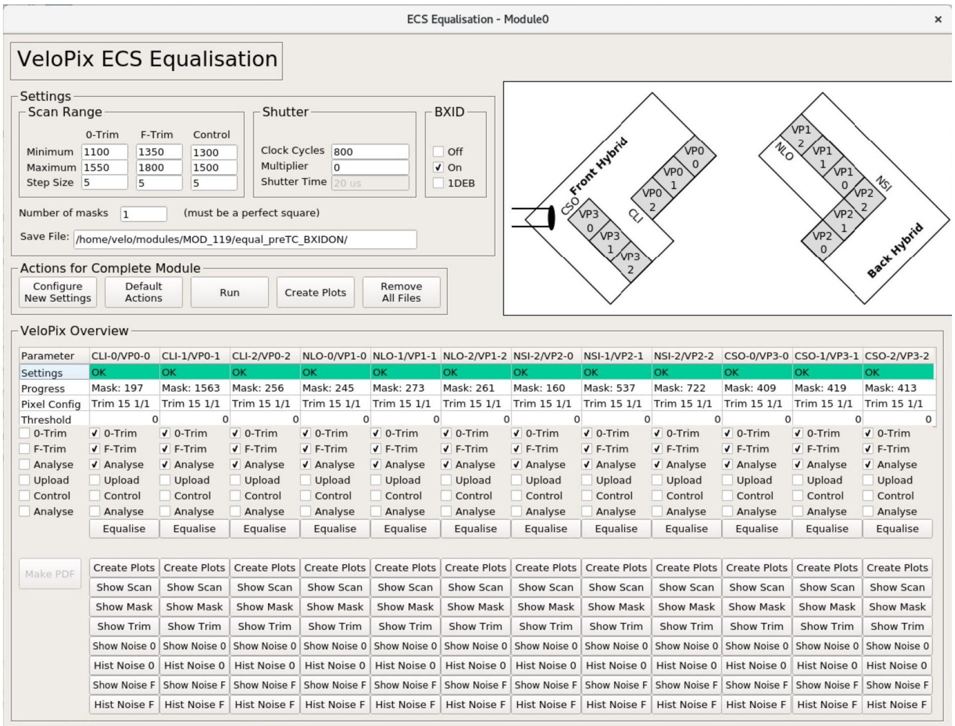


Figure 3.17: Module noise and equalisation panel.

saved from every step of the scan. The results, described in detail in Sect. 4.4.4, are then used to calculate the VeloPix ASIC noise and the ideal pixel trim setting which ensures equal response of the detector. The panel used for the test and real-time analysis is in Fig. 3.17.

3.4.5 Cooling performance

Testing the cooling of the module is important in order to verify the heat transfer capabilities from the tiles. This is mostly important for the long-term operation of the modules as it is expected that the heat generation will increase with the radiation damage.

Since the performance relies on the ideal conditions of the CO_2 boiling in the microchannels, the cooling machine has to be precisely set to a flow of $0.3\,\mathrm{g\,s^{-1}}$ with the back pressure at 12 bar. The test is then done by configuring the tiles in different modes (Low power, Minimal/QA and GWT), effectively changing their power consumption. After each configuration the system takes ≈1 min to reach stable temperature before the data are taken. The whole process has been made semi-automatic using WinCC/OA monitoring panel (Fig. 3.14) in order to simplify the operator's job requiring only configuration and data upload. The logged values are bandgap temperature from each VeloPix ASIC, analogue and digital voltages and currents, as well as temperatures from the NTC sensors in the FE hybrids. Values are then automatically processed and maximal ΔT estimated, the methodology described in Sect. 4.4.5.

3.4.6 High voltage

Since the used silicon sensors are designed to operate with HV applied, the performance when biasing the sensor has to be verified. Each tile has been tested in a probe station as part of its QA, whereas, the module testing includes the full readout chain (from the tiles, through the HV tapes down to the VFB).

The tests are performed using the Keithley 2410 1100V Sourcemeter PSU [88], one tile at a time. The PSU provides a range up to 1100 V with precision $\mathcal{O}(1\,\mathrm{nA})$.

3.4.6.1 Software

The testing of the HV required the development of a dedicated software to perform DAQ and safely ramp the voltages. The tool has been developed as a multi-platform and stand-alone Graphical User Interface (GUI) in Python3, with the use of additional libraries (pyvisa, Numpy, PyQt5 and Matplotlib). The communication between the PC and the PSU is done via a General Purpose Interface Bus (GPIB) to USB interface.

The software is initialized based on a settings file. This includes the number of connected devices with their GPIB addresses, but also widely used Current-Voltage (IV) parameters such as ramp speeds, starting and ending voltages as well as current limits. In addition to this a waiting time and number of averaging points for each step of the IV curve is added, which reduces the measured noise in the data. All of these settings can be changed from within the GUI and are directly applied to the selected devices.

The design of the software is done in a modular way, consisting of multiple files/objects:

main.py Initialization of the software and its execution.

gui.py Initialization of the GUI, device control, DAQ and plotting. The object provides direct access to the settings file and ensures upload of the new values to the devices. All data and commands are propagated using signal-slot mechanism, enabling thread safe operation of the code.

design.py Design of the GUI, made in the QtDesigner software [89]. The code sets the locations of the buttons and plots on the screen, defined in layouts in order to enable resizeability of each object based on the size of the main window.

utils.py Utility functions taking care of the logging, providing a simple way to verify the operation and debug any present issues.

plotter.py Definition of a simplified real-time plotting. The object can easily enable or disable parts of the plots or rescale the axes.

readout.py Execution of the voltage setting, device initialization and DAQ. This part is run in a dedicated thread and linked directly to the main window to preserve the correct timings of the defined voltage scan. The object automatically controls all attached and selected PSUs, thus providing the option of parallel testing.

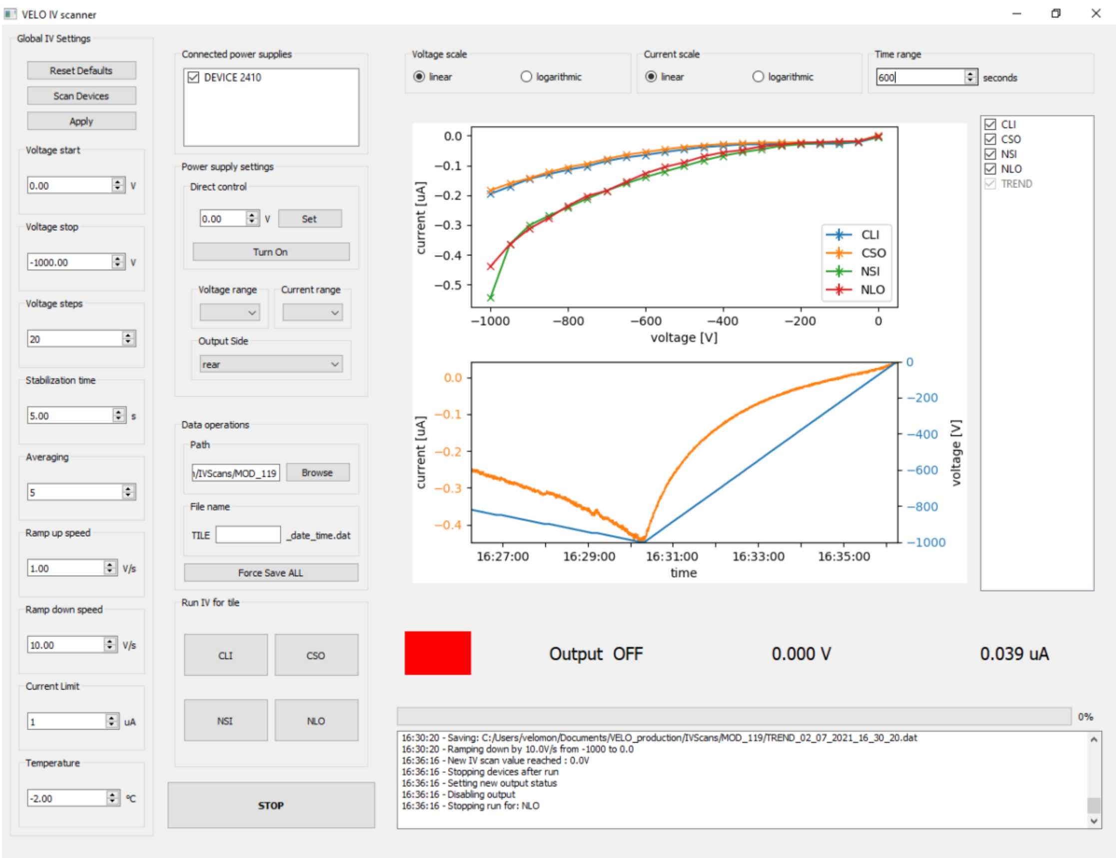


Figure 3.18: Screenshot of the software developed for controlling the PSU.

In order to safely operate the devices, this object may receive a terminating signal at any point of the scan, ensuring the safe ramp-down of the PSU.

PyKeithleyMonitor.py Implementation of the link between different PSU devices. The functions link the python layer with the GPIB commands defined in the manual of a given device.

The object includes the definition for Keithley versions 2410, 487 and a dummy device.

The operation was made in an user-friendly way (a screenshot of the GUI is in Fig. 3.18), ensuring that the users could use it as simply as possible while having access to all of the information about the scan performed.

Since the main purpose of the software was to test VELO modules or directly control the PSU, the whole interface was accommodated to it. For this reason a set of dedicated buttons with names corresponding to the four module tiles is added (CLI, CSO, NSI, NLO). The IV plot shows the individual measurements, providing a tool for a simple comparison of the biasing performance. Apart from this, the PSU can be directly operated by setting the desired voltage, the software will automatically perform ramping based on the settings.

The output is in the form of two files, one providing current and voltage from the IV scan points in the standardized format used during the tile testing. The other file contains a more detailed information with more measurement points along with time stamps.

For testing purposes, the software will automatically provide a dummy set of devices (based on the definition in settings) in case no matching device is connected.

3.4.6.2 IV curves

For the purposes of setting the proper ground of the sensors, the LV has to be applied to the tiles under test by configuring them in the Minimal/QA mode. This establishes a common ground between the sensors, readout, and the rest of the setup – including that of the HV PSU. In this configuration, the module requires active cooling. In order to match the conditions of the probe station testing, the IV curves of the tiles should be taken at room temperature. However, due to the CO_2 cooling machine constraints, the highest temperature that can be stably achieved is ≈0 °C.

The scan is done in the range from 0 V to −1000 V every 50 V, with a current limit at 1 µA. The ramp-up and ramp-down speeds are defined as $1\,\mathrm{V\,s^{-1}}$ and $10\,\mathrm{V\,s^{-1}}$, respectively. A waiting time of 10 s is set to stabilize the reading, the final value is an averaged measurement of five values. Data are then uploaded to the database, the analysis described in Sect. 4.4.6.1.

3.4.6.3 Ramp-up

The ramp-up test is done with the same settings as the IV curve scan, the only difference being that the test is performed only after TC at the nominal cooling conditions (−30 °C). The purpose of this test is to verify the capabilities of the tiles when cold, therefore only the maximum current and voltage are recorded. The test is stopped either after reaching a current of 1 µA or at the maximum voltage 1000 V. Results of the scan are summarised in Sect. 4.4.6.2.

3.5 Summary

The VELO module assembly and testing consists of a number of procedures requiring high precision where every step is documented using dedicated software or hardware tools. The processes can be split to three parts – assembly, metrology and module performance. While the metrology is primarily used to verify the outcome of the individual assembly steps (ranging from the tile placement and gluing all the way to the cabling), the module performance checks the communication, control and cooling capabilities of the system. These tests provide the information about every aspect of the module with the results later used to characterise the quality and grade of the individual steps.

Chapter 4

Database and quality assurance

The main role of the database is to provide a simplified online tracking of the VELO assembly progress, saving all module relevant information and assigning grades. Data are used to locate any defects on the fly but also for any future references and comparisons.

The database is built to be used by all of the assembly and testing sites, it includes information from the base structural components all the way to the individual electrical tests of the finished modules and VELO Upgrade halves. Since at the time of writing a majority of the modules has been tested at UoM, the data presented in this chapter are only considering this subset of the production.

4.1 Database

The database is a PostgreSQL, accessed via Django framework based on Python. The used structure primarily consists of models.py and admin.py files which define links between the Python and database objects and create a simplified web UI. The code base is split to a number of projects, with information between them easily shared.

Common Shared objects accessible by all other projects are defined here, ranging from institutes through operators all the way to issues.

Pixel tiles The project tracks the VeloPix ASIC testing, bonding to the silicon sensors and all probe-station measurements.

Cooling Construction and testing of the microchannel cooling substrates is defined in the this project.

Electronics components Information about all of the remaining electronics parts that are made and tested before use are stored here, including cables, VFBs and OPBs.

Modules The project tracks assembly and testing of the modules, based on the items defined in the previous projects. This will be described in more detail.

Installation This is the final step, where individual modules and other electronics components are installed into the VELO halves and tested.

Since the database serves as a common place for file upload and simple data analyses, online plotting has been added to provide a visual representation of data. This is done using Bokeh and Holoviews projects, enabling interactive control of the plots.

4.1.1 Modules project

The modules project links all of the database-tracked components for the purpose of the module assembly and testing. An overview of the UoM produced modules is in Fig. 4.1, the lines represent individual modules while the columns correspond to the assembly/testing steps.

The columns are arranged based on their order of the production which is almost identical for both UoM and Nikhef. The letters in the table match the status of the test, A-F defined as grades, P as test pending, M as multiple QA tagged tests uploaded and Z for test not graded. When adding a new object, the corresponding column-row matched letter may be clicked, which will open a pre-filled window that is to be modified with the object specific information. The already existing objects are opened in the same way, providing a quick access to their data and plots.

In order to better track the problems of these complex objects, a simple issue listing capability has been added. While this functionality has been developed namely for the modules project (shown in 4th column from the right in the Fig. 4.1), it is part of the common components and is therefore easily accessible in other parts of the database as well.

Issue listing The module assembly related issues may be of two kinds – general problems or processes/tests specific ones. Since all of these objects are defined independently, the issues would be visible only from the given object itself. Because of this an extra linking method had to be defined, which would enable tagging the module as a parent. The result is an accessible list of all relevant issues within the module page, while the test specific ones are also accessible from the test itself.

All of the objects are defined using parent classes corresponding to construction and testing, the individual groups in the module project are following:

items defining the materials used,

processes including the assembly steps as well as TC,

metrology representing the mechanical qualifications,

electrical testing containing the communication and electrical operation.

This approach simplifies the creation of new methods, where all of the common fields (grades, operators) are automatically defined without the need to duplicate the code.

The individual module is accessed from its dedicated web page, shown in Fig. 4.2. This contains more detailed information compared to the module overview page, where the assembly or testing object may be filled in directly using inline methods.

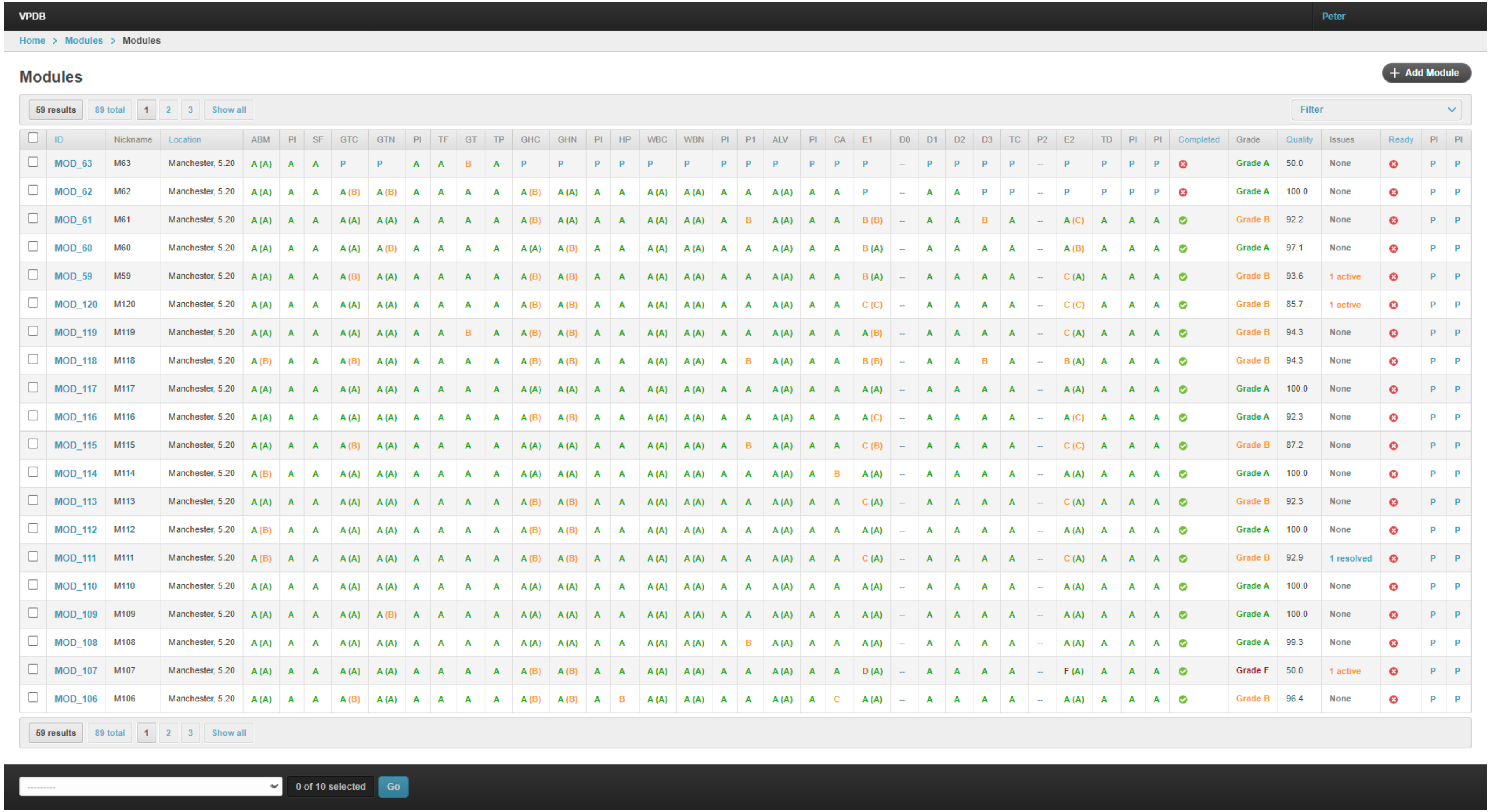


Figure 4.1: Database module overview page. Showing modules (lines) with assembly and testing steps (columns) in the same order in which they are performed, along with their status and grades.

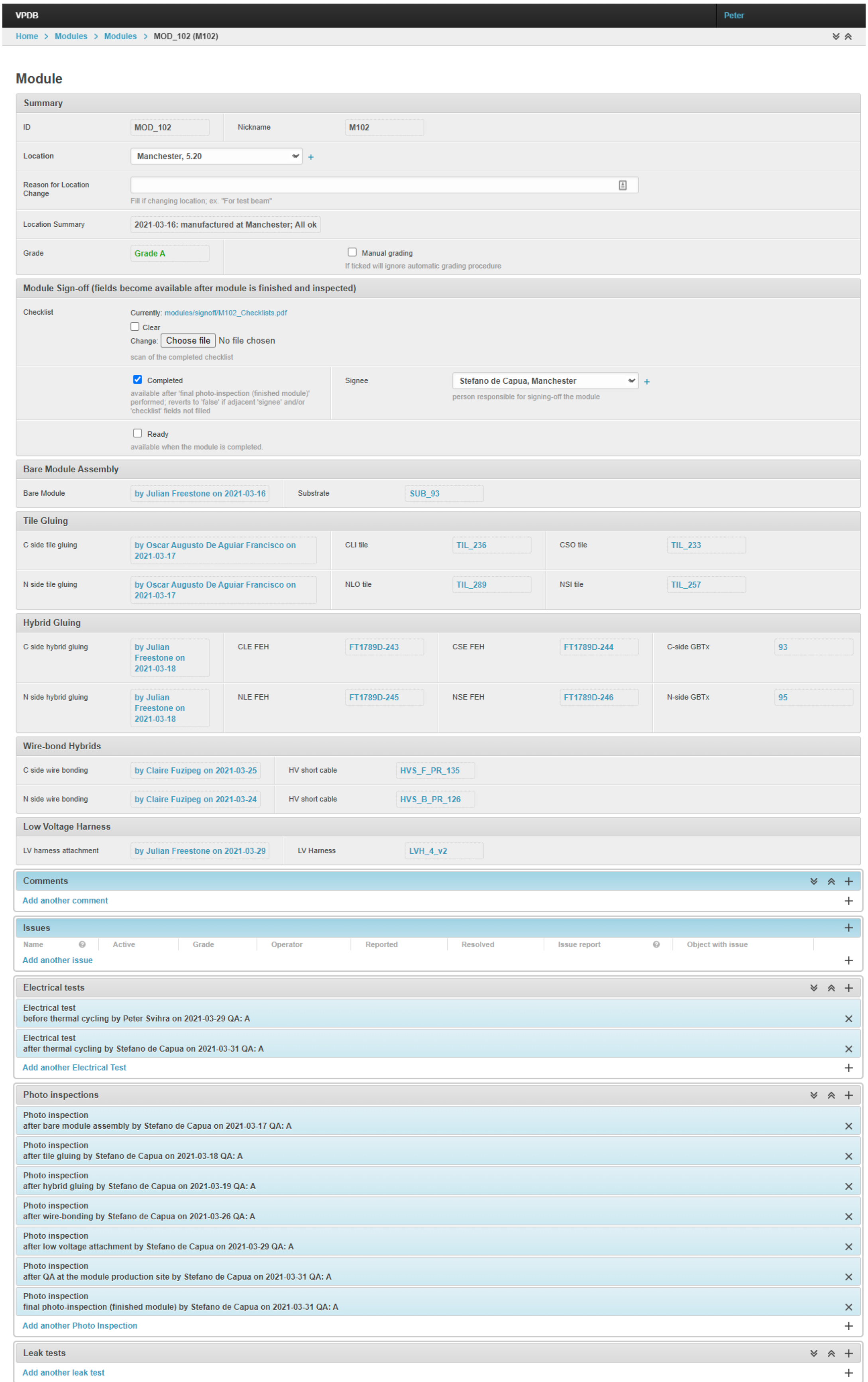


Figure 4.2: Database single module page with list of individual components and tests.

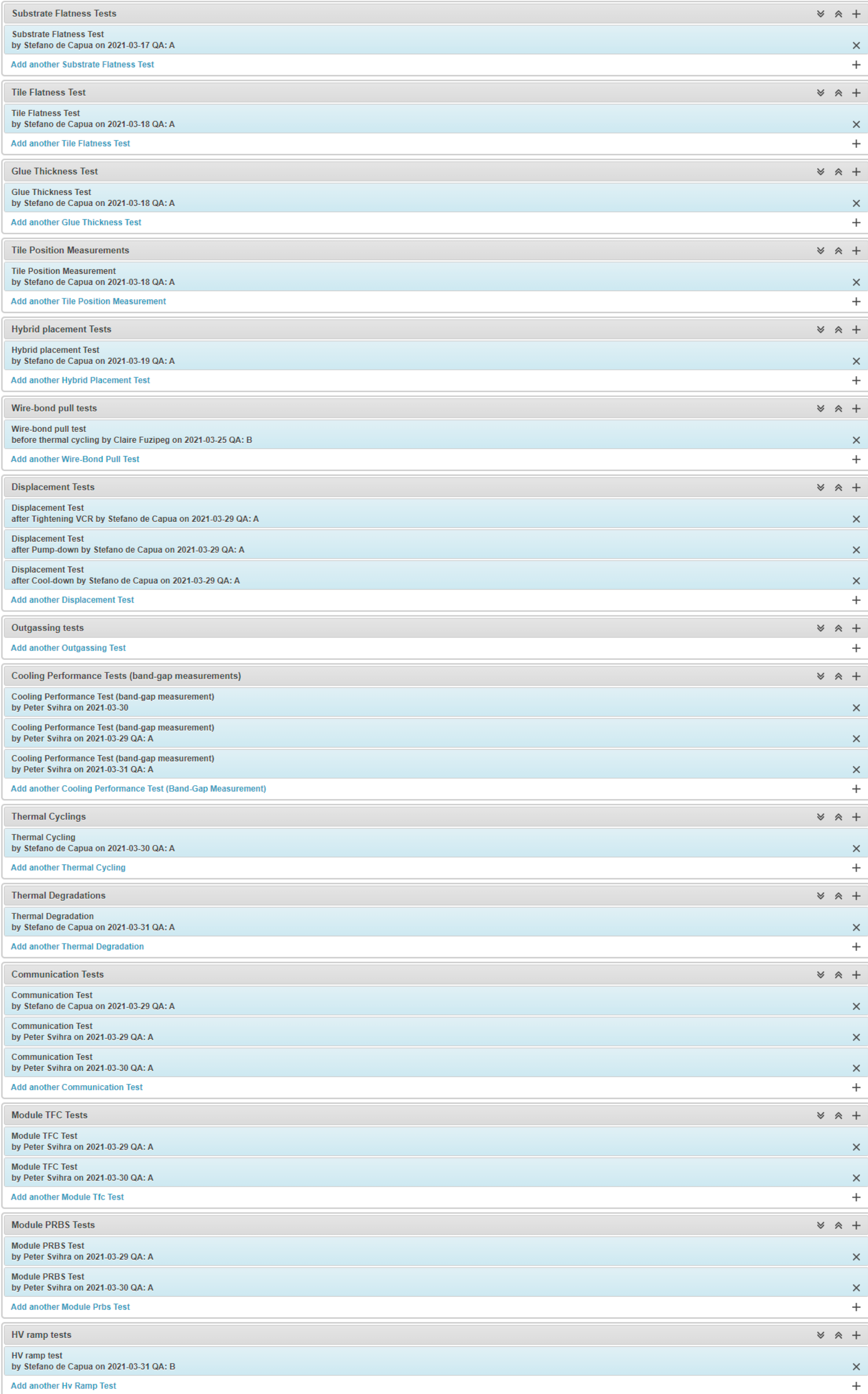


Figure 4.2: Database single module page with list of individual components and tests.

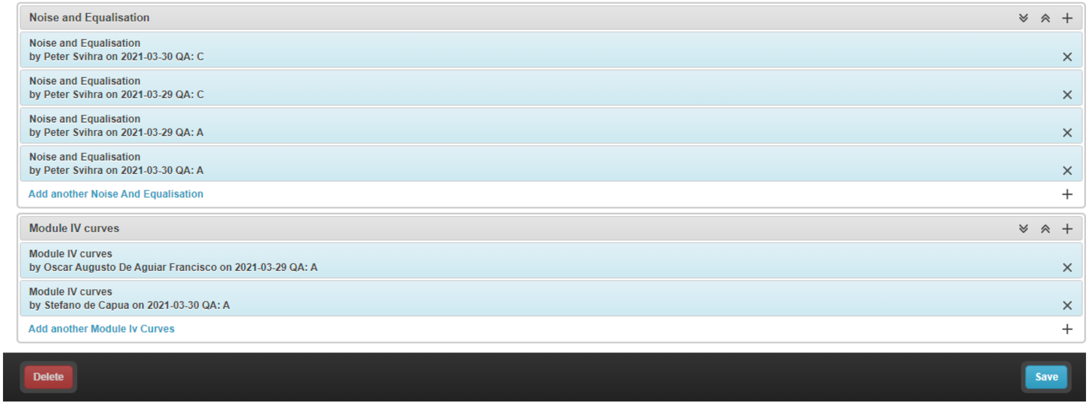


Figure 4.2: Database single module page with list of individual components and tests.

4.1.1.1 Grading

Most of the tests have automatic grading procedures implemented using the uploaded values to mark their performance. In a case when the object should not be considered a part of the module QA, a dedicated checkbox can be unticked. This will still provide the database processing of the results without affecting the overall grade and can be useful especially during the R&D phase and performance testing.

The grades used are defined as follows, a more detailed definition later in the text:

A excellent,

B good,

C acceptable,

D irresolute,

F failed.

These grades can be always manually changed, however, the change should contain an explanation in the comment box. The manually graded tests (namely the assembly steps or visual inspections) should match the same definitions.

The individual grades are based on the expected performance limits of the specific test as well as an approximate distribution of results. In case the test did not fail, most of the results should be grade A and only a bare minimum grade C.

An important change has been the addition of the grade D, used to characterize results which need more investigation or decision before the module can be used. In general, such modules are not suited to be installed as they may have unresponsive ASIC or a potential issue with the cooling, however, they might be used in a lower intensity regions if really needed.

Finally, the grade F represents a failure of the test, either due to a test setup issue or a bad component/assembly. These problems are first investigated in detail with an effort to fix them before deciding whether such modules can be used as they could affect a system-wide performance of the whole VELO.

The individual grade limits as well as the overall module grading and quality is described further in the text.

4.2 Assembly

The whole process of the module assembly (as described in Sect. 3.1) is tracked in the database. The individual objects representing the steps are simply linking the used components with user comments. This provides both storage tracking (number of pixel tiles available, ...) as well as an unified way of keeping the information.

The final process grade is based on the lowest grade of the used components and of the operator assigned grade for the given step. The only caveat is that grade B components do not propagate and behave as grade A, mainly so that the component does not affect the overall module grade in the case when everything works well. Since the component grade would then be hidden in the module overview in Fig. 4.1, the assembly steps are shown with the process grade and the worst grade of the used components in brackets.

4.3 Metrology

The metrology tests performed in between the module assembly steps require quick access to the results as any potential issues may stop the production. For this purpose almost all of the tests have automatic processing of the uploaded data. The tests are summarised in Sect. 3.2 and 3.3, with the grading limits in the Table 4.1. The individual data processing algorithms and grading choices are described later in the text. Apart from the automatically obtained data, module pictures are taken and uploaded to the database after every step, providing a tracking of the module progress as well as a visual scan for any defects.

4.3.1 Substrate flatness

After performing the measurement described in Sect. 3.2.1, the data are uploaded to the database for an automated processing. This is also done for the test from other institutes measuring the substrate flatness, the only difference being the format and number of points provided.

CERN Data provided by CERN are from the final check of the substrate before shipping, performed as a high resolution optical scan of the whole surface. This results in a a large data size with lot of background, the processing is therefore done by roughly fitting a plane and rejecting distant points, saving x, y and z coordinates in a Comma-Separated Values (CSV) format.

Table 4.1: Grading definition for metrology tests.

(a) : Substrate flatness.

grade	absolute value of Q_1 and Q_3
A	$< 50\,\mu m$
B	$< 75\,\mu m$
C	$< 100\,\mu m$
D	$> 100\,\mu m$
F	–

(b) : Tile flatness.

grade	mean±2·stdev
A	$< 30\,\mu m$
B	$< 60\,\mu m$
C	$< 120\,\mu m$
D	$> 120\,\mu m$
F	–

(c) : Glue thickness.

grade	mean±2·stdev
A	$(40\,\mu m, 120\,\mu m)$
B	$(20\,\mu m, 160\,\mu m)$
C	$(0\,\mu m, 200\,\mu m)$
D	outside
F	–

(d) : Tile position.

grade	position	angle	distance
A	$< 55\,\mu m$	$< 1000\,\mu rad$	$> 130\,\mu m$
B	$< 110\,\mu m$	$< 2000\,\mu rad$	$> 100\,\mu m$
C	$> 110\,\mu m$	$> 2000\,\mu rad$	$> 50\,\mu m$
D	–	–	$< 50\,\mu m$
F	–	–	$< 10\,\mu m$

(e) : Wire-bond pull force.

grade	weight
A	$> 5\,g$; or at most 2 between $4\,g$ to $5\,g$
B	$< 4\,g$; or more than 2 less than $5\,g$
C	–
D	$< 2\,g$
F	–

(f) : Displacement VCR tightening.

grade	before-after
A	$< 50\,\mu m$
B	$< 80\,\mu m$
C	$< 100\,\mu m$
D	$> 100\,\mu m$
F	–

(g) : Displacement pump-down.

grade	maximum
A	$< 50\,\mu m$
B	$< 80\,\mu m$
C	$< 100\,\mu m$
D	$> 100\,\mu m$
F	–

(h) : Displacement cool-down.

grade	maximum
A	$< 100\,\mu m$
B	$< 150\,\mu m$
C	$< 200\,\mu m$
D	$> 200\,\mu m$
F	–

UoM As the measurement is a laser scan, the input is directly in a required CSV format and does not have to be modified in any way.

Nikhef Since a different optical technique is used for the measurement, only up to 20 different points are saved. Because of this, the data are fitted during the upload using the second degree polynomial in 2D, generating the points to fill the expected substrate shape. The result is saved in the same format as above.

The final analysis of the flatness is then automatically performed by a least-square plane fitting of z. The resulting data describing flatness Δz are then saved in the form of

$$\Delta z \equiv z - z_{\text{plane}}, \tag{4.1}$$

together with the nominal x and y coordinates. In order to minimize the effect of outlier points, in total 5 % of the most distant values from both sides are disregarded.

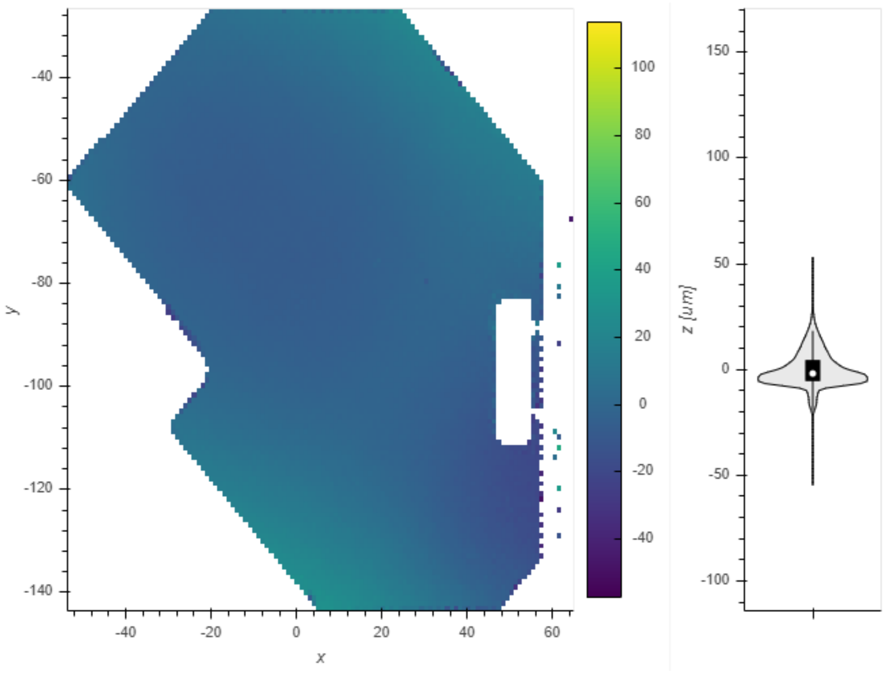
The flatness results are an important step of the construction process as spotting any large substrate defects at an early stage of the module assembly can prevent its failure at a later step. An interactive map of Δz is plotted in the database web page along with a corresponding violin plot of the Δz distribution with highlighted quartiles. This provides for a simple visual comparison between results of different sites or substrates.

Results of the flatness measurements of the microchannel substrate used for assembling the MOD_100 module are in figure 4.3. The measurements performed by both CERN and UoM do not match, probably due to change in the fixation caused either by the carbon fibre legs or by the turnplate holder used in UoM (in the CERN measurement the substrate is freely placed on a flat surface). This, however, does not present any issues for the module assembly as the variation along the whole substrate is negligible with no cracks present.

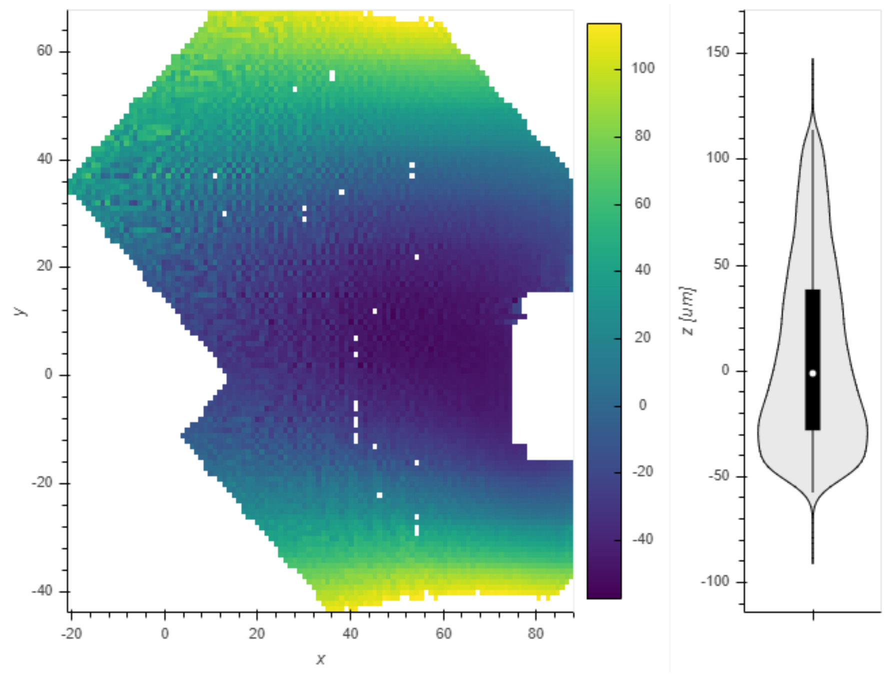
It is important to note, that due to the different procedures of obtaining data, some extra points (clearly distant from the substrate) are plotted for the CERN measurements, whereas regions with missing/scattered data entries are present in the the UoM procedure. The extra data-entries occupy the same plane as the substrate, and are thus indistinguishable by the processing script. On the other hand, the missing entries are caused by the SmartScope being unable to properly focus the laser in some regions. Neither should impact the data quality of the results.

In order to provide a simple tracking tool where any drift in the UoM measurement could be spotted, quartiles of the Δz distribution are plotted for each module in Fig. 4.4. It can be seen that some modules have a very large Q_4 (corresponding to the most distant measured point), caused by a wrongly focused or scattered laser in the few points. The least-square minimization of the plane results with the median of the flatness to lie at around 0 while quartiles Q_2 and Q_3 are consistent across the produced modules.

grading Since the measurement covers a very large area with aforementioned outliers in the distribution (difficult to remove by automatic processing), instead of mean and



(a) : Measured by CERN



(b) : Measured by UoM

Figure 4.3: Comparison of substrate flatness measurements performed by CERN and UOM on SUB79, used in MOD_100. Left plots are x, y maps of Δz and the right violin plots show Δz distribution, highlighting median and quantiles.

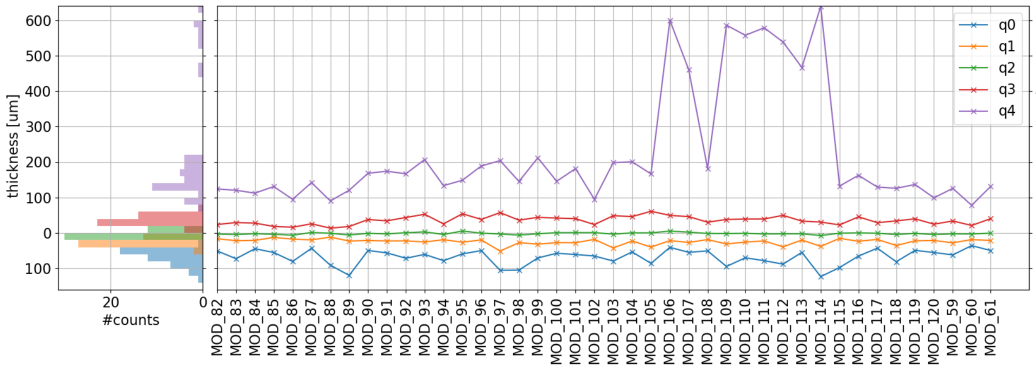


Figure 4.4: Trend of substrate flatness results, showing the quartiles of the Δz distribution.

standard deviation the quartiles were selected for the grading purposes. The individual levels based on quantiles at 25 % and 75 % of data are specified in Table 4.1a.

4.3.2 Glue layer thickness and tile flatness

This measurement is done for both UoM (described in Sect. 3.2.2) and Nikhef, once again the data-input from the two sites differ. The differences are bigger than for a simple substrate flatness measurement, therefore the calculation has to be performed differently as well. The results are not only needed to verify the uniformity but also used later in the cooling performance described in Sect. 4.4.5.

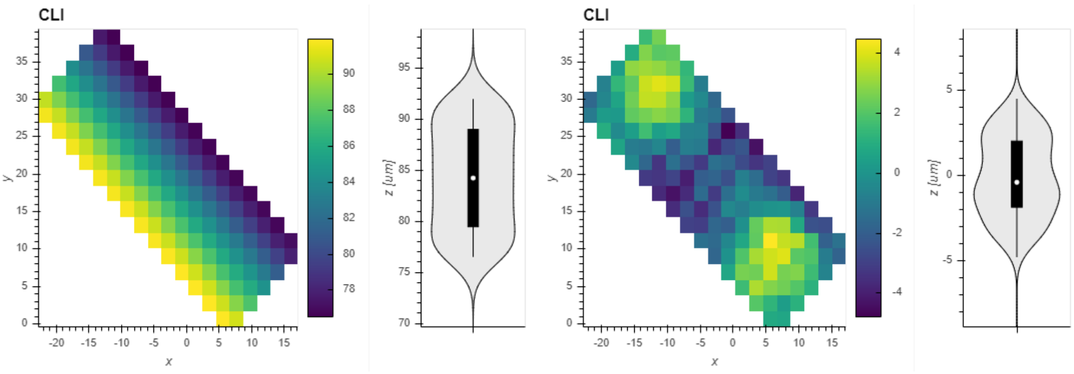
UoM The final analysis of the glue thickness and tile flatness is performed at the same time by the least-square plane fitting of both tile z_{t} and of surrounding substrate z_{s}. As the data file contains an uninterrupted stream of x, y, and z coordinates, the points corresponding to the tile and substrate are located automatically using a simple peak-finding algorithm, relying on the fact that the distance between the top of the tiles and substrate should be at least $d = 445\,\mu\mathrm{m}$ (representing the thickness of tiles, bump-bonds and VeloPix ASICs). The data are then saved in the form of

$$\Delta z_{\mathrm{t}} \equiv z_{\mathrm{t}} - z_{\mathrm{t}_{\mathrm{plane}}}, \tag{4.2}$$

$$\Delta z_{\mathrm{g}} \equiv z_{\mathrm{t}} - \Delta z_{\mathrm{t}} - z_{\mathrm{s}_{\mathrm{plane}}} - d \tag{4.3}$$

together with the x and y coordinates, where Δz_{g} is the final glue layer thickness and Δz_{t} the corresponding tile flatness.

Nikhef Since the number of measured points is not large and does not contain any points from the substrate, the analysis relies on the already obtained results of its flatness. The previously mentioned second order 2D polynomial fit of the substrate is used as the underlying reference system $z_{\mathrm{s}_{\mathrm{fit}}}$. While the tile flatness is simply obtained from the least-square plane fit of the data, the result has to be used to estimate the glue



(a) : Glue thickness Δz_{g}. **(b) :** Tile flatness Δz_{t}.

Figure 4.5: Measurement of glue layer thickness and tile flatness of CLI, MOD_120. Heatmaps show the spatial profile of $\Delta z_{\mathrm{g(t)}}$, the violin plots show the distribution, highlighting quartiles. Overhanging regions are plotted in both cases, visually worsening the glue thickness profiles.

thickness as

$$\Delta z_{\mathrm{t}} \equiv z - z_{\mathrm{t_{plane}}}, \tag{4.4}$$

$$\Delta z_{\mathrm{g}} \equiv z - \Delta z_{\mathrm{t}} - z_{\mathrm{S_{fit}}} - d. \tag{4.5}$$

Since two separate results $\Delta z_{\mathrm{t(g)}}$ are created from a single measurement, data can be uploaded to either the glue thickness or the tile flatness object and the other database entry is automatically populated.

Similarly to the substrate flatness, an interactive map of both Δz_{g} and Δz_{t}, as well as their violin plots showing distribution are plotted on the database web page. An example for the CLI (VP0) tile from MOD_120 is in Fig. 4.5. The glue thickness varies due to the non-uniformity of the substrate and, in the case of the CLI and NSI, is plotted in the overhanging regions as well.

Using the LHCb coordinate system, the regions of the individual ASICs are easily selected from the data (this, however, relies on their precise placement). Mean and standard deviation of the thickness are then calculated for each ASIC along with the same values for the flatness of the whole tile. This provides a simple tool to compare the results of many tiles and modules, the trend for the CLI tiles is in Fig. 4.6.

The glue layer thickness is an important factor of the thermal performance of the whole module, the target thickness being 60 μm to 100 μm. While most of the modules have a uniform layer with negligible slope, in a few cases a larger variation is present among the individual ASICs of a tile. Since the glue deposition on a single tile is a reasonably stable process (its speed defined based on the actually estimated glue viscosity), any large variation could mean impurities in the layer, potentially damaging the readout. However, the glue also performs a function of a filler absorbing non-uniformities between the substrate and tiles. Plotting the average glue thickness for the back-to-back ASICs (see Fig. 4.7 for CLI and NLO tiles), the variation is within the standard deviation. This means that any increase or decrease of the average thickness is caused by an issue in the

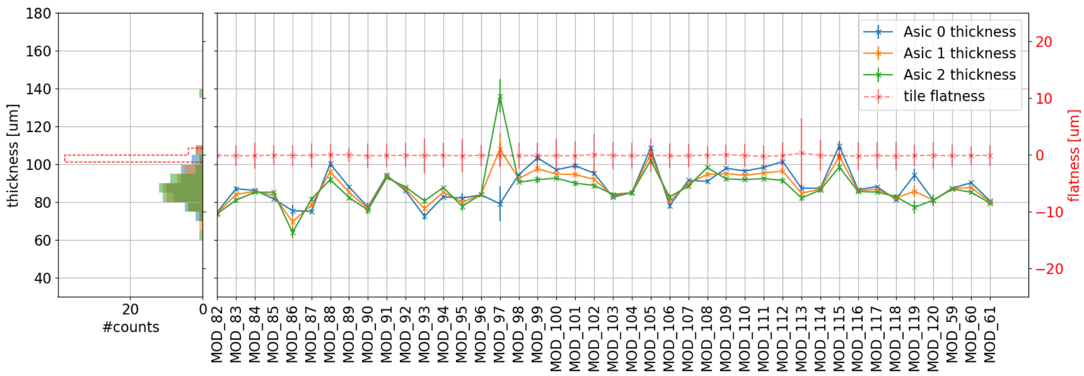


Figure 4.6: Trend of CLI (VP0) tile mean glue thickness and mean tile flatness results.

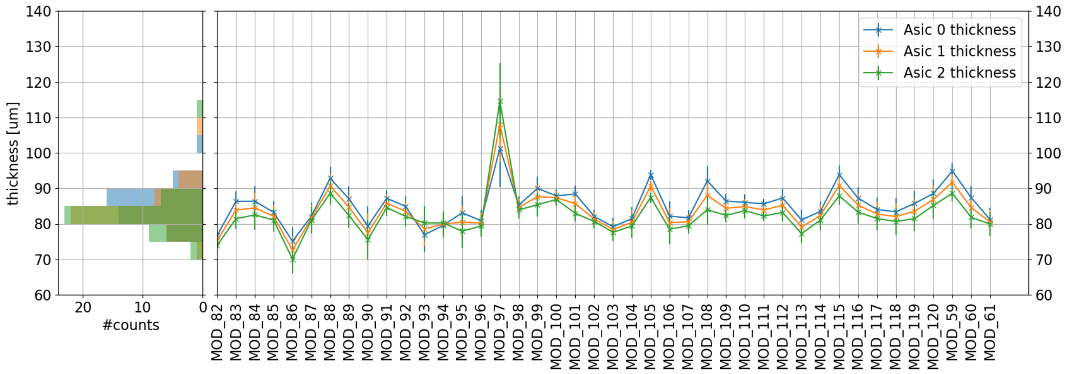


Figure 4.7: Average glue thickness of ASICs under CLI and NLO tiles.

gluing process.

Even though the tile flatness is gained as a by-product of the glue measurement, the results are important to spot any physical damage of the tiles. The results also serve as a check of the performance of the calculation algorithm, since a poorly located tile surface would result in a wrong estimate of the glue thickness.

grading Even though the test is similar in nature to the substrate flatness, the measurement points are better defined and mean with standard deviation may be used. In order to match the largest area of the tile, the grading levels are based as mean $\pm$ 2stdev, covering about 95 % of the area. The concrete grading limits used for the glue thickness and tile flatness are in Table 4.1c and Table 4.1b, respectively.

4.3.3 Tile position

Even though the UoM and Nikhef use different testing setups, the values are in a very similar format and only minor parsing on the database side is needed. After uploading the measured position of the ASIC and sensor fiducials, the points are directly saved in the database.

At first, a simple comparison of the x and y values has been used in order to characterise the tile placement. Even though functional, the approach did not consider the rigidity of

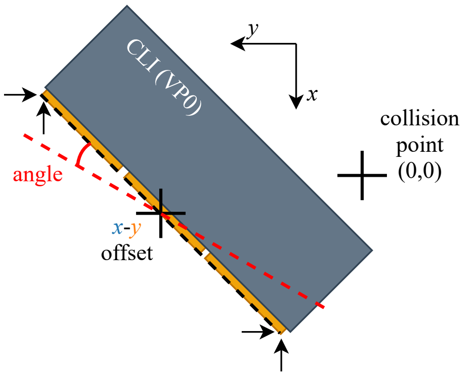


Figure 4.8: Visualised tile position calculated variables x, y and angle.

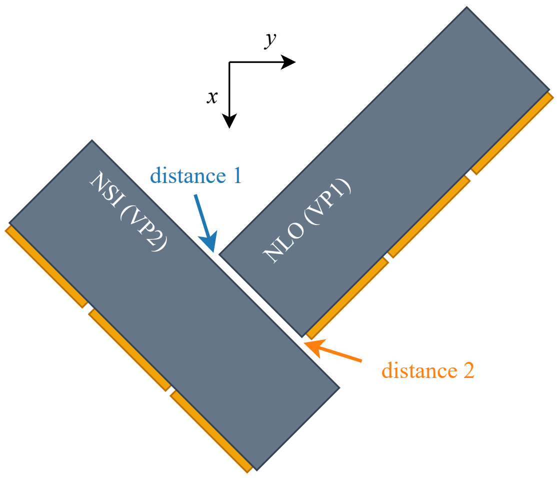


Figure 4.9: Visualised closest distance between NLO (VP1) and NSI (VP2) tiles.

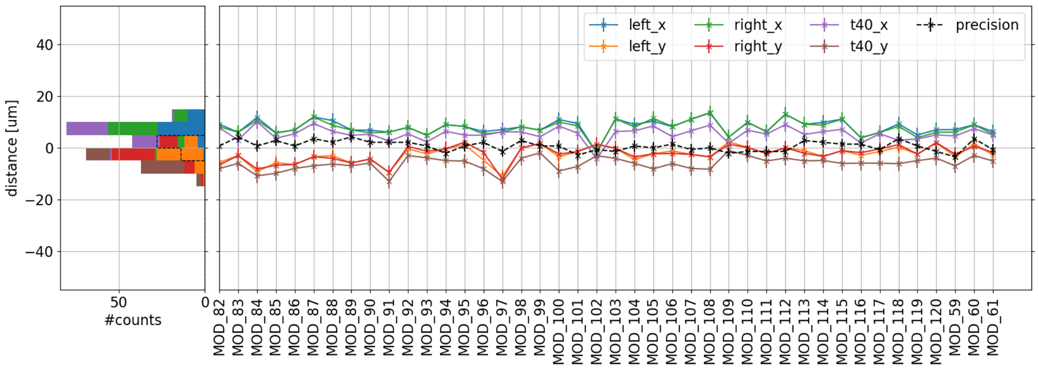
the tiles and the results could be wrongly graded based on a single shifted measurement. In order to better analyse the placement, an algorithm that calculates the middle point and rotation of the tile has been developed.

Transformation The algorithm locates a mid-point between the corner fiducial positions on the tile, its position simply compared as x_{off} and y_{off} to the ideal middle defined using the reference values. Afterwards, using the lines constructed for both measured and target corners, the opening angle θ is found. These final values are visualised in Fig. 4.8.

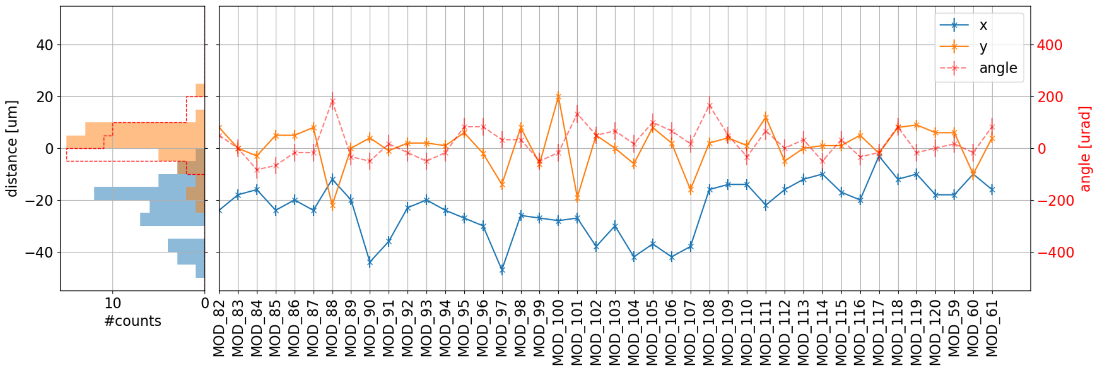
It is then possible to describe the whole surface of the tile using the offset and the rotation. The distance between the N-side tiles (see Fig. 4.9) can be simply estimated as well as compare the measured and expected position of other markers. The calculation is done by applying a 2D rotation matrix on the coordinates x, y relative to the ideal middle point as

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} \cos(\theta) & \sin(\theta) \\ -\sin(\theta) & \cos(\theta \end{bmatrix} \begin{bmatrix} x - x_{\mathrm{off}} \\ y - y_{\mathrm{off}} \end{bmatrix} + \begin{bmatrix} x_{\mathrm{off}} \\ y_{\mathrm{off}} \end{bmatrix}. \tag{4.6}$$

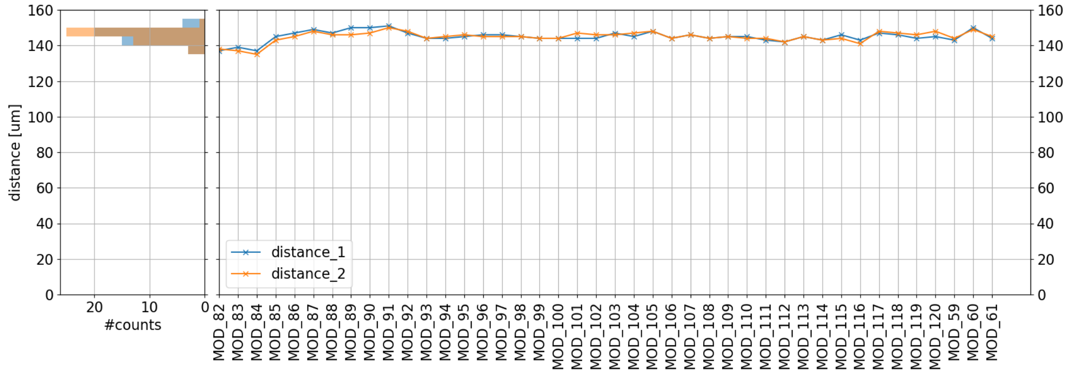
The estimated positions x', y' are then compared to the measurement, plotted for CLI in Fig. 4.10a. Apart from the distance between the estimated and measured sensor markers, the precision of the measurement is also plotted, calculated as a difference between the ideal and measured distance of the ASIC corner fiducials. While the variation of the precision is only a few µm (similar to the SmartScope optical system precision of 2 µm), the sensor markers have higher variation with the x and y values offset in the positive and negative direction, respectively. The inconsistency could be caused by the fact that the C-side sensor markers are measured from the N-side with the module flipped. The measurement is therefore affected by both the SmartScope coordinate system calculation as well as gravity pulling on the module in different



(a) : Difference between the ideal and measured distance of the corner fiducials.



(b) : Offset of the middle point in x and y along with the rotation of the CLI (VP0) tile.



(c) : Distance between corners of NSI and NLO.

Figure 4.10: Trend of the tile position metrics.

directions. It is important to note that the measurement expects the points to be in the same z plane and that the rotation angles are small ($\mathcal{O}(50\,\mu\text{rad})$), potentially introducing a calculation error.

In order to track the placement of the tiles, the trend of x_{off}, y_{off} and angle is plotted (example for CLI tile is in Fig. 4.10b). While the offset of the middle point is consistent in y around 0 µm to 10 µm, a small drift of the position is visible for the x from MOD_93 onwards. Plotting the trend helped in spotting the issue and after assembling MOD_107, the tile alignment constants used for the positioning have been slightly adjusted, resulting in an improvement of the tile placement.

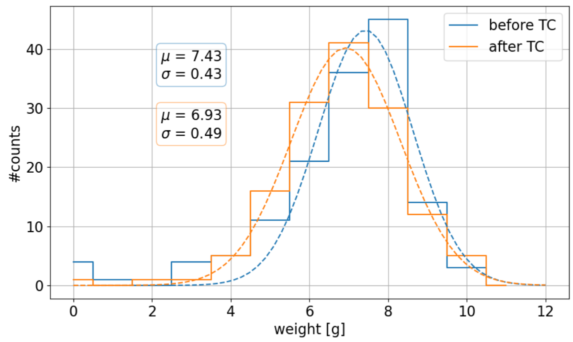


Figure 4.11: Distribution of pull test strengths before and after TC.

As mentioned above, the usage of a rotational and translational transformation provides option to calculate the distance between the N-side tiles. The ideal value is $d = 145\,\mu\mathrm{m}$, the trend for the modules plotted in Fig. 4.10c.

grading The placement and rotation of the tiles is important for the tracking alignment of the channels. Therefore the grading characterizes the translation and rotation of the individual tiles as well as the distance of the N-side tiles (due to safety concerns). The limits are defined in Table 4.1, the offset following a ratio of the pixel pitch ($55\,\mu\mathrm{m} \cdot 55\,\mu\mathrm{m}$). A similar condition is placed on the rotation, where $1000\,\mu\mathrm{rad}$ corresponds to roughly half of the pixel pitch positioned on the most distant place from the middle point.

4.3.4 Wire-bond pull testing

The metrology step described in Sect. 3.2.5 consists of many repeated pull tests, where breaking force of the individual wires is recorded. Both module assembly institutes provide values in the same format.

The information uploaded to the database contains the position of the wire under the test, specified by the tile, ASIC and a number assigned to each bond. Furthermore, the location of the break is recorded as well, defined by two letters:

first letter represents the location near ASIC (A) or hybrid (H),

second letter represents the type of the break as ankle (A) or foot lift (F).

At first, the pull testing has been done before and after TC in order to check the integrity of the bonds and compare the effects of the thermal shocks. This is compared in a histogram of the breaking forces shown in Fig. 4.11. Even though a minor worsening of the average forces may be observed, the change is very small and the overall strength of the bonds should not be affected as it is safely above 5 g (industry standard MIL-STD-883 for the used $25\,\mu\mathrm{m}$ thick wires being about 2 g). Furthermore, the potential danger of damaging the bonds when inserting a fully tested module inside the pull-testing machine

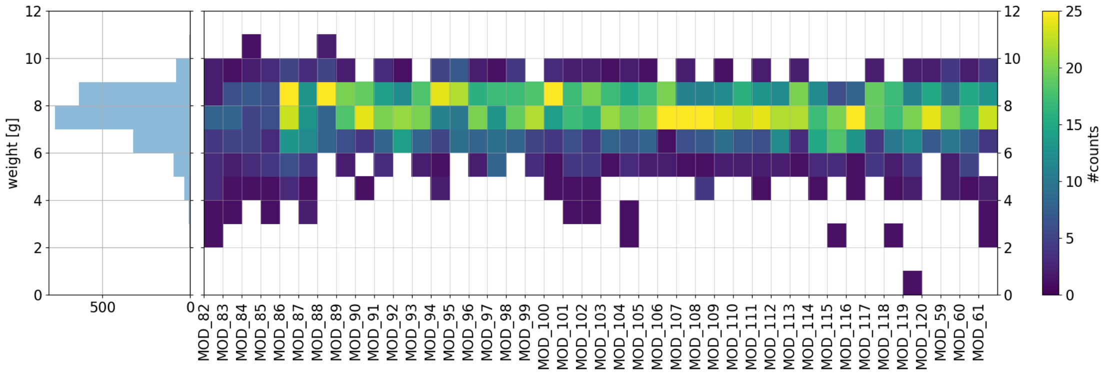


Figure 4.12: Trend of all pull test forces.

overweighted the need for the second round of testing. Because of this, all sacrificial wire-bonds are pulled in one go before TC.

The trend of the test is shown in Fig. 4.12, the first four modules having less entries due to being tested after TC as well. The UoM results improved over time, achieving the mean weight ≈8 g.

grading The grading is defined based on the values in Table 4.1e. Since there are 48 bonds pulled per a module in total, the probability of having some below the 5 g mark is increased. For this reason the criteria have been relaxed, enabling up to 2 bonds with force corresponding to at least 4 g.

4.3.5 Module displacement and thermal cycling

As described in Sect. 3.3.3, the displacement is a simple test monitoring the way any module bends at different stages of the operation. The file formats for UoM and Nikhef are once again slightly different, however, both contain timestamps, displacement values and temperature from the NTC sensors on all tiles.

As an example, the displacement trend of the maximal change during the initial cool-down to −30 °C is in Fig. 4.13. The values for the modules vary a lot, however, this is most probably caused by the way the cooling connectors are aligned, as even a small change may create a pull force on the whole module. These results are an important input for the uncertainty in the tracking alignment, as it is difficult to calibrate the z position of the installed modules.

grading The grading limits for the VCR connection, pump-down and cool-down are in Tables 4.1f to 4.1h, respectively.

Since the the first test monitors the connection of the cooling pipes, only the change of the displacement values before and after the connection are important. Contrary to that, the displacement during the other two tests depends on the vacuum or cooling of the module, therefore the maximal value difference defines the grade.

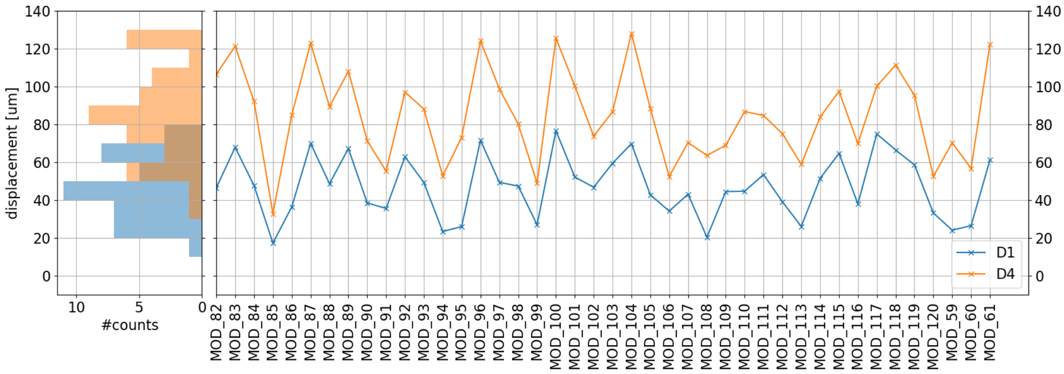
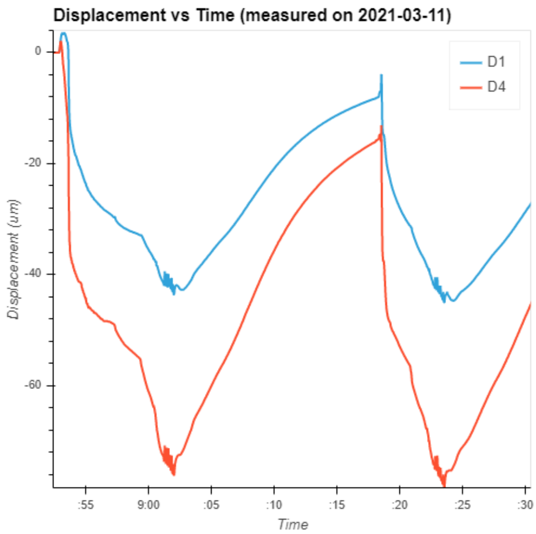
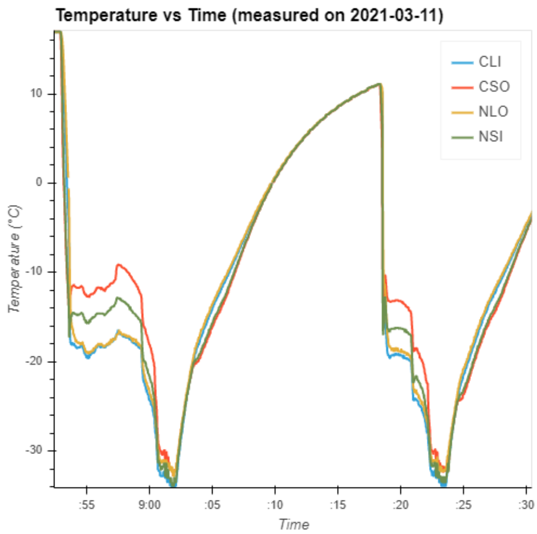


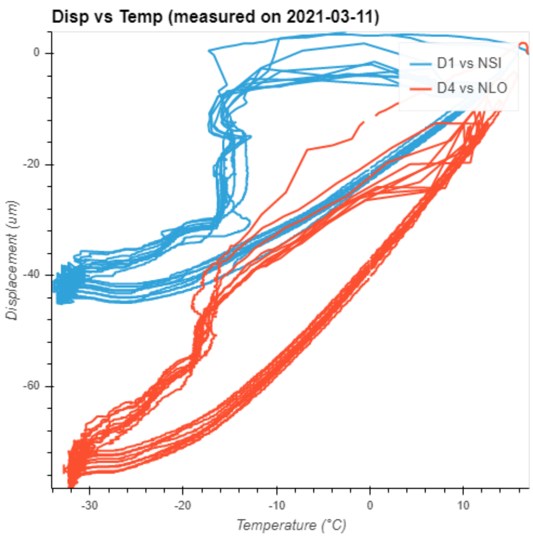
Figure 4.13: Trend of the maximal displacement during the module cool-down.



(a) : Time dependency of displacement.



(b) : Time dependency of NTC temperatures.



(c) : Correlation of displacement and temperature from 10 cycles.

Figure 4.14: Displacement and temperature during TC of MOD_98.

The information from the displacement and NTC sensors is also used during the TC, described in Sect. 3.3.5. While the process itself does not have any automatic grading, the values are uploaded to the database and performance plots are generated (see Fig. 4.14). These are useful as they provide a correlation of the displacement and temperature of the same set of tiles.

The dependency in Fig. 4.14c is very similar for all 10 cycles of the MOD_98, following a hysteresis. This is probably due to the quick cool-down of the module, where the displacement increases even during a stable temperature region (timestamp 8:55 to 9:00 in the Figs. 4.14a and 4.14b). When the module warms up, the displacement follows a much slower linear change.

4.4 Electrical testing

The electrical performance testing is done after the whole module has been assembled and installed in the vacuum tank. The tests are described in Sect. 3.4, each of them with automatic processing, analysis and grading implemented in the database. Since almost all of the tests are performed using WinCC/OA panels, the data formats are identical for all institutes.

Since all of the tests are done both before and after TC, an electrical test object has been created in the database which serves as a group containing all test types. This simplifies upload and tracking while distinguishing the test conditions without any extra overhead for the user. The tests can be still added directly to the module, either for debugging or when in need of uploading extra information.

Similarly to the metrology, the summary of the grading is in Table 4.2, the individual analysis of the data described further in the text. The grade for the whole suite of tests is summarised in the module overview in Fig. 4.1 as E1 or E2, corresponding to tests before and after TC, respectively.

4.4.1 Communication

The data obtained from the WinCC/OA monitoring panel (described in Sect. 3.4.1) are directly uploaded to the database. Since the ASICs need a stable voltage supply in order to properly function, any larger drop $\mathcal{O}(20\,\mathrm{mV})$ can hint at a potential issue. The mean values of the uploaded voltages are therefore automatically calculated and saved for every ASIC.

As the communication test is the only test done warm in air and cold both before TC and after TC, the comparison of the values may provide a better insight into the thermal dependency and performance worsening. Trends for all three cases for the CLI tile are plotted in Fig. 4.15, respectively. While the digital voltage for all ASICs is mostly stable across the produced modules, some drops are present in the analogue ones. It is visible that lowering the temperature improves the voltage performance (smaller resistivity of

Table 4.2: Grading definition for electrical performance tests.

(a) : Communication.

grade	analog and digital voltage
A	$> 1.20\,\mathrm{V}$
B	$> 1.18\,\mathrm{V}$
C	$< 1.18\,\mathrm{V}$
D	one missing
F	multiple missing

(b) : TFC.

grade	result
A	all OK
B	–
C	–
D	one missing
F	multiple missing

(c) : PRBS.

grade	BER of links	BER of more than half links
A	$< 1 \cdot 10^{-11}$	$< 1 \cdot 10^{-12}$
B	$< 1 \cdot 10^{-10}$	$< 1 \cdot 10^{-11}$
C	$> 1 \cdot 10^{-10}$	$> 1 \cdot 10^{-11}$
D	one missing	–
F	multiple missing	–

(d) : Equalisation and noise.

grade	tile masked	ASIC masked	noise	noise stdev
A	$< 1\,\%$	$< 5\,\%$	–	–
B	$< 5\,\%$	–	–	–
C	$> 5\,\%$	–	–	–
D	–	–	> 8	> 2
F	–	–	–	–

(e) : Bandgap.

grade	ΔT at 26 W
A	$< 8\,°\mathrm{C}$
B	$< 10\,°\mathrm{C}$
C	$> 10\,°\mathrm{C}$
D	missing data
F	–

(f) : Thermal degradation.

grade	mean abs	max
A	$< 1\,°\mathrm{C}$	–
B	$> 1\,°\mathrm{C}$	–
C	–	
D	–	$> 5\,°\mathrm{C}$
F	$> 5\,°\mathrm{C}$	–

(g) : IV curves.

grade	curve
A	$<$ high quality
B	$<$ good quality
C	$<$ good quality below 140 V
D	$>$ good quality below 140 V
F	–

(h) : HV-ramp.

grade	lowest voltage before 1 µA
A	$> 700\,\mathrm{V}$
B	$< 700\,\mathrm{V}$
C	–
D	–
F	$< 250\,\mathrm{V}$

the cooled metal results in a reduction of the voltage drop between the DC/DC converters and the FE) and that no major difference is visible after the TC. In a few cases the voltage readout during the cycling has been lost (for example CLI in MOD_101), however, all other tests remained working. This is a potential issue for the final VELO as the monitoring contains the information about the temperature as well, therefore, such a module should not be placed in a high fluence region.

grading The comparison for the voltage values is specified in Table 4.2a. The ASICs should work without any issues above 1.18 V, however, a higher voltage is safer for the operation. Any missing readings present a potential danger for the operation and have to be therefore individually assessed.

4.4.2 TFC

The results of the TFC test, described in the Sect. 3.4.2, are simple binary outputs telling whether the ASICs received/sent all of the required signals or not. The booleans are uploaded to the database by the operator in a form of checkboxes.

Any missing communication means that the given ASIC cannot be properly synchronised and controlled. The trend of the test results is in Fig. 4.16, all ASICs are working well apart from NLO ASIC-2 in MOD_101.

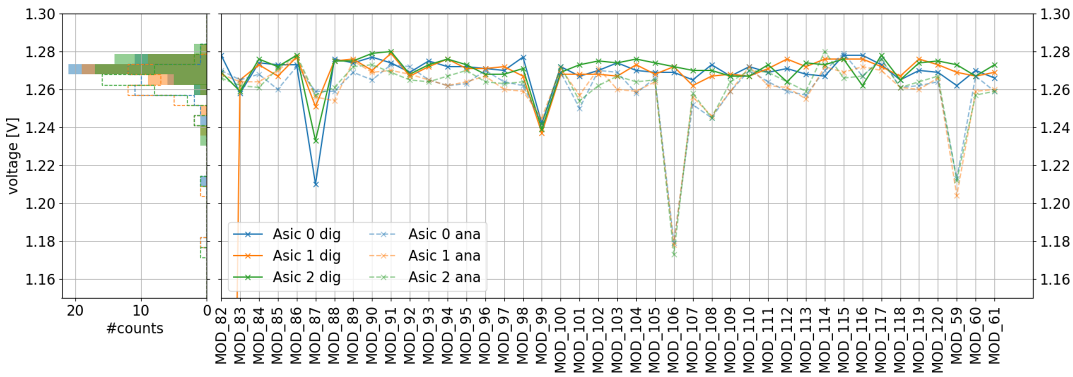
grading The grading is defined in Table 4.2b. Since any missing TFC responses mean bad communication, the grades are either A for a fully operational module or D/F based on the severity of the issue.

4.4.3 PRBS

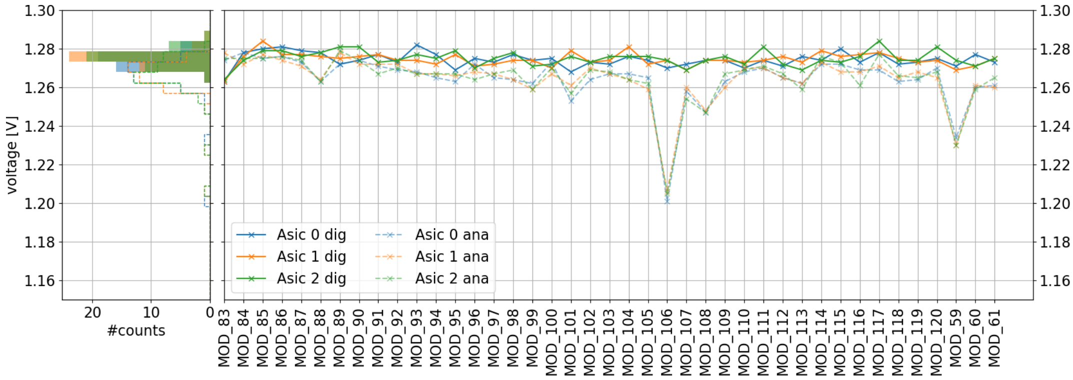
The data uploaded from the PRBS test described in the Sect. 3.4.3 consist of columns with number of bits tested and corresponding number of errors for each of the 20 links. The database then calculates the BER and produces a simple visualisation shown in Fig. 4.17.

Since the main metric followed is the BER, the trend for all UoM produced modules is in Fig. 4.18. While most of the links did not have any errors in the $1 \cdot 10^{12}$ bits tested, links 4-6 and 13 are in general more noisy, regardless of the module tested. This issue has not been fully understood yet, mostly due to the fact that reconfiguration of the MiniDAQ PC used for the testing randomly affects the result of the test. Furthermore, the performance improved between MOD_94 and MOD_114 even though no change in the assembly or the setup occurred. The eye diagram scans were performed as well, however, no visible worsening (closing of the eyes) has been observed.

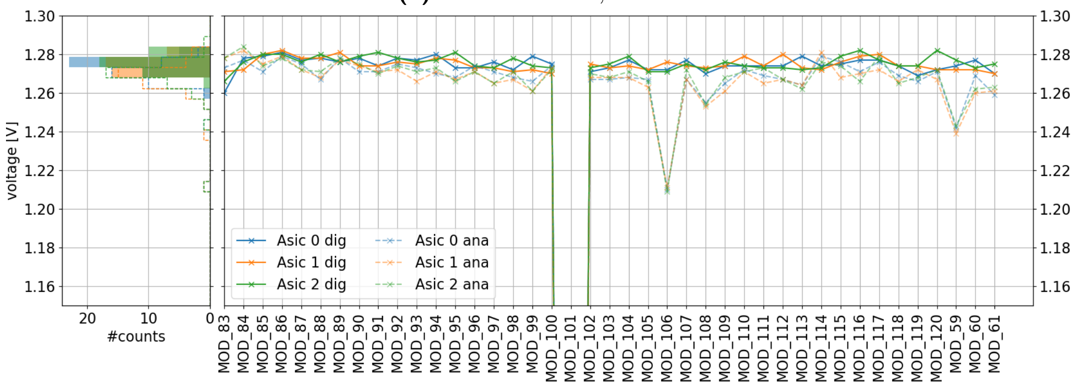
It is possible that the worsened results are caused by the change of the environmental conditions in the CR (humidity, temperature) as they cannot be controlled in UoM. This could be affecting the conditions of the VTTx and VTRx on the OPB. None of the



(a) : Warm, in air.



(b) : Cooled-down, before TC.



(c) : Cooled-down, after TC.

Figure 4.15: Trend of analogue and digital voltage during communication tests of NSI (VP2) tile.

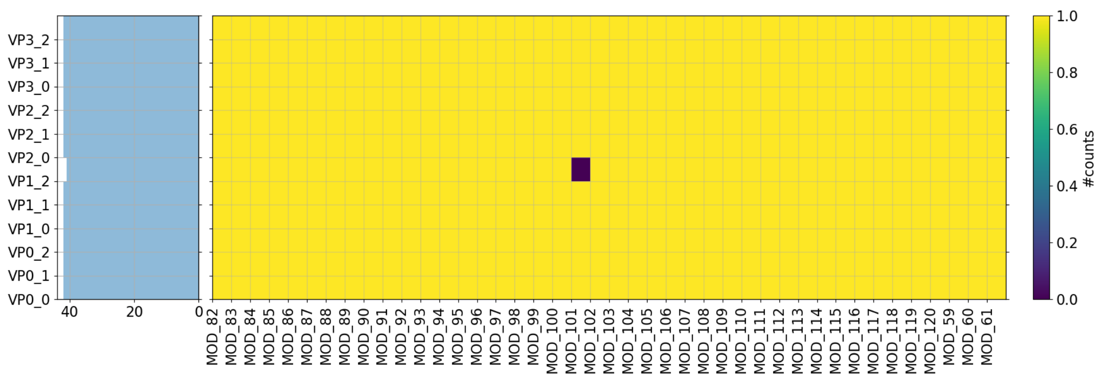


Figure 4.16: Trend of TFC response before TC.

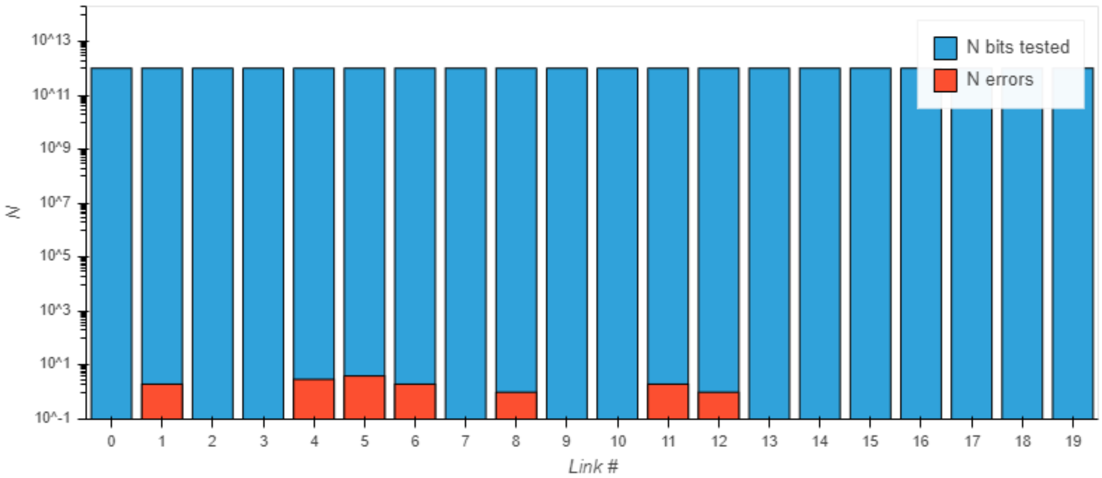


Figure 4.17: Example of PRBS test data showing $1 \cdot 10^{12}$ bits tested along with the number of errors. Test was done before TC on MOD_60.

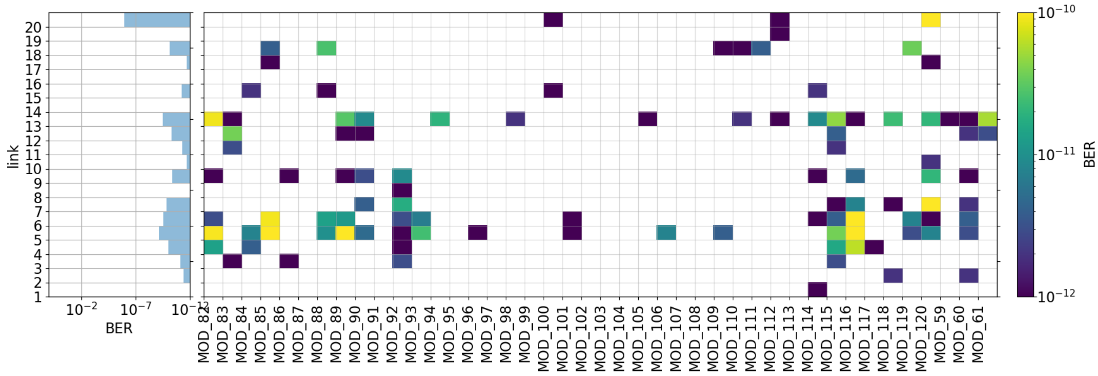


Figure 4.18: Trend of PRBS test BER from before TC. White means no errors detected in the $1 \cdot 10^{12}$ bits tested.

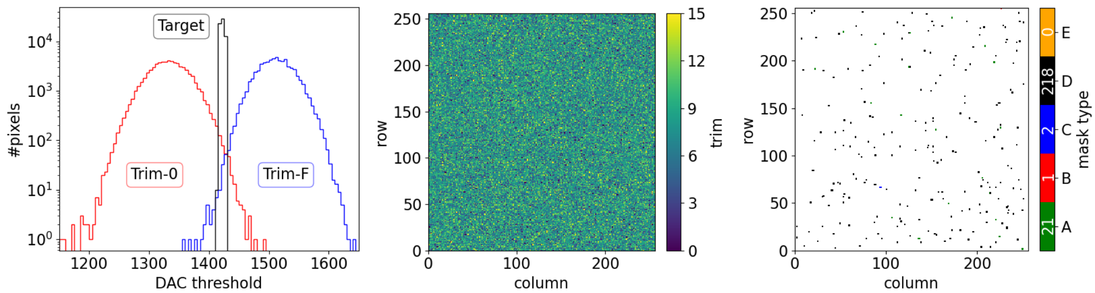
results would pose an enormous issue for the operation, the actual performance will be reconfirmed in the Liverpool during the VELO half assembly.

grading While the ideal performance would be with no errors at $1 \cdot 10^{12}$ bits tested, the limits had to be changed to better characterize the modules produced. For this reason two levels are used, describing the values of the worst link and of the half of the links, specified in Table 4.2c.

4.4.4 Noise and Equalisation

The noise and equalisation scan described in Sect. 3.4.4 produces large amount of data, specific for each ASIC and tile. While an automatic processing is a part of the WinCC/OA panel, the database compares results of different runs as well.

As a part of equalising the matrix response, the distribution of the threshold DAC values can be plotted in Fig. 4.19a, representing the threshold where the individual pixels started to register the noise signal. Furthermore, the distributions are based on the pixel Trim-0(F) setting, which offsets the global defined threshold. Following the design [40]



(a) : Distribution of pixel response during Trim-0(F) scans and resulting Target.

(b) : 2D trim DAC distribution.

(c) : 2D mask pattern.

Figure 4.19: Example of equalisation results for NSI (VP2). Scan done after TC in MOD_103.

a linear dependency of the 16 possible trim values may be assumed, the target settings are then calculated to lie in the middle of the two distributions. The corresponding 2D distribution of the trim settings is in Fig. 4.19b, no visible pattern should be present.

Since some pixels are noisy, not responsive or it might be impossible to equalise their response, a 2D mask pattern is also created in order to block their signal. This helps with reducing the unnecessary data coming from the ASICs as the mask is uploaded directly to the chip. While the uploaded one is a simple binary pixel switch, the masking reason can be defined based on the equalisation results as:

A dead pixel,

B missing response for Trim-0 ,

C missing response for Trim-F,

D cannot reach the Target,

E predicted noise too large.

An example of the mask pattern is in Fig. 4.19c.

The threshold DAC distribution based on the individual pixel hit counts is also obtained during the scan, its standard deviation then defines the corresponding noise response. The distribution of noise for all pixels is plotted in Fig. 4.20, comparing testing of the wafer and tile in the probe station and of the fully assembled module. While a small expected worsening is observed between wafer and tile testing, no significant effect is caused by attaching the tile to the module.

The noise is an important attribute as it defines the precision with which the pixels register signal. Using the default VeloPix settings, the signal has been measured as [42]

$$1\,\mathrm{DAC} \approx 15\,\text{e-h pairs}, \tag{4.7}$$

the mean noise of 6 DAC therefore corresponds to about 90 e-h pairs. For a comparison, in the 200 µm thick sensor the signal from a MIP creates about 16 000 e-h pairs, however,

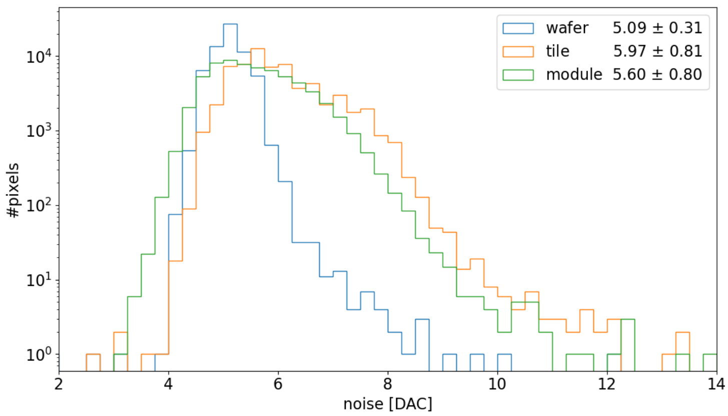


Figure 4.20: Comparison of noise results in NSI (VP2) for different tests. The tile is used in MOD_103.

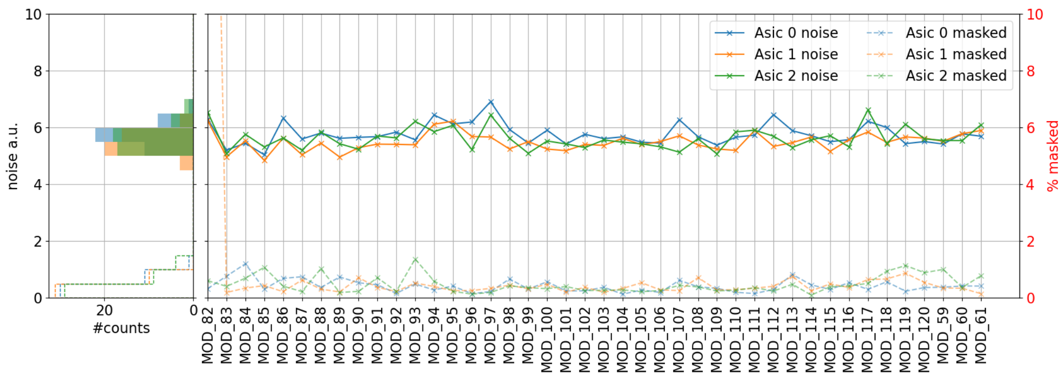


Figure 4.21: Trend of mean noise and percentage of masked pixels in VeloPix ASICs, shown for NSI (VP2) after TC.

the performance will worsen with the increased radiation damage. These values are important for design of CCE scan of the detector, further described in Chapter 5.

An example of the trend of the average noise and number of masked pixels is in Fig. 4.21. During the testing it has been found that the BXID synchronization frame increased the overall noise of the tiles. Since this has been always enabled during the R&D phase, the scan is performed four times in total, with BXID on and off both before and after TC.

grading The grades of the modules are based on the pixel noise performance as well number of the masked pixels. The concrete numbers are defined in the Table 4.2d.

4.4.5 Cooling

Cooling performance of the module is very important as there is an expected increase of the heat output with the radiation damage. While the general conditions are mostly dependant on the system itself (pressure, flow and temperature of the liquid CO_2), the glue thickness or TC may affect the results.

4.4.5.1 Bandgap

The bandgap test described in Sect. 3.4.5 has a fairly complex data analysis procedure as it relates information from many different subsystems.

The original method relied on the operator to visually estimate the NTC temperature, the currents supplied by the LV PSU and finally the FE hybrid voltage readings. Using these values the actual power output of the module had been calculated, using estimates of different power consumptions by the GBTx, GBLD and VeloPix ASICs, obtained from all power modes.

Power calculation The heat power output $P_{i,t}$ is defined as

$$P_{i,t} = I_{i,t} V_{i,t} \varepsilon(I_{i,t}) \tag{4.8}$$

where $V_{i,a}$ is voltage from the corresponding FE hybrid, $I_{i,t}$ is the current drawn by the appropriate ASIC and $\varepsilon(I_{i,t})$ efficiency of the DC/DC converters. The indices $i \in \{\mathrm{D}, \mathrm{A}\}$ distinguish between digital and analogue currents, while the indices t define the ASIC or tile. The efficiency is following an empirical function [35]

$$\varepsilon(I_{i,t}) = 1.017 - 0.3811 I_{i,t} + 0.0974 I_{i,t}^2. \tag{4.9}$$

Operating the module in the power modes 0a and 0b, the digital current drawn and the corresponding efficiency of the GBTx and GBLD ASICs has been estimated to

$$\begin{aligned} I_{\mathrm{D,GBTx}} &= 0.46\,\mathrm{A}, & \varepsilon(I_{\mathrm{D,GBTx}}) &= 0.86, \\ I_{\mathrm{D,GBLD}} &= 0.12\,\mathrm{A}, & \varepsilon(I_{\mathrm{D,GBLD}}) &= 0.97 \end{aligned}$$

and is expected to be consistent throughout the operation. The remaining current per tile t is then calculated as

$$4 I_{\mathrm{D},t} = I_{\mathrm{D}} - I_{\mathrm{D,const}} - I_{\mathrm{D,GBTx}} - I_{\mathrm{D,GBLD}}, \tag{4.10}$$

$$4 I_{\mathrm{A},t} = I_{\mathrm{A}} - I_{\mathrm{A,const}} \tag{4.11}$$

where the constant digital and analogue currents come from the power modes 0a, with module not configured and are $I_{\mathrm{d,const}} = 1.335\,\mathrm{A}$ and $I_{\mathrm{a,const}} = 0.052\,\mathrm{A}$.

The bandgap temperature recorded by the WinCC/OA panel was then uploaded to the database along with a file containing the manually calculated power consumption with the timestamps matching beginning and end of each power mode measurement. Because of this an effort has been made to simplify the whole process, minimizing the potential mistakes and speeding up the measurement.

The new method utilizes the already logged displacement data files which contain the NTC sensor readings. A modification has been made to the LV PSU software in order to log data in a simple time+current format. Finally, a functionality to the WinCC/OA panel has been added, linking the FE hybrid voltage logging and bandgap temperature readings. In total, three separate files are uploaded in the database, the timestamps separating the individual power modes are automatically determined from the WinCC/OA uploaded file (as every measurement also contains a mode-specific number).

The data uploaded produce two sets of plots – the time evolution of the temperatures and currents (see Fig. 4.22a) and linearly fitted dependencies of the temperature and power output (see Fig. 4.22b).

The first plot is an useful way to quickly verify the measurement conditions (no current spikes or visible increase of the temperature) with the automatically defined averaging regions for the three modes (red and black lines, corresponding to the beginning and the end of the averaging, respectively). Since the desired result is the cooling performance rather than the absolute temperatures, a temperature difference ΔT is defined. It is obtained by subtracting the offset of the linear fit from the data for all NTC and ASICs values. In order to better compare different modules and runs, the results are extrapolated to the 26 W mark, corresponding to the expected heat power output at the end of the VELO lifetime.

As an approach to minimize the systematic error from the manually operated cooling machine in the UoM, the final thermal performance values are obtained by subtracting the closest NTC sensor value from the corresponding VeloPix ASICs

$$\Delta T'_{\mathrm{VP}_{0,1}} = \Delta T_{\mathrm{VP}_{0,1}} - \Delta T_{\mathrm{NTC}_0}, \tag{4.12}$$

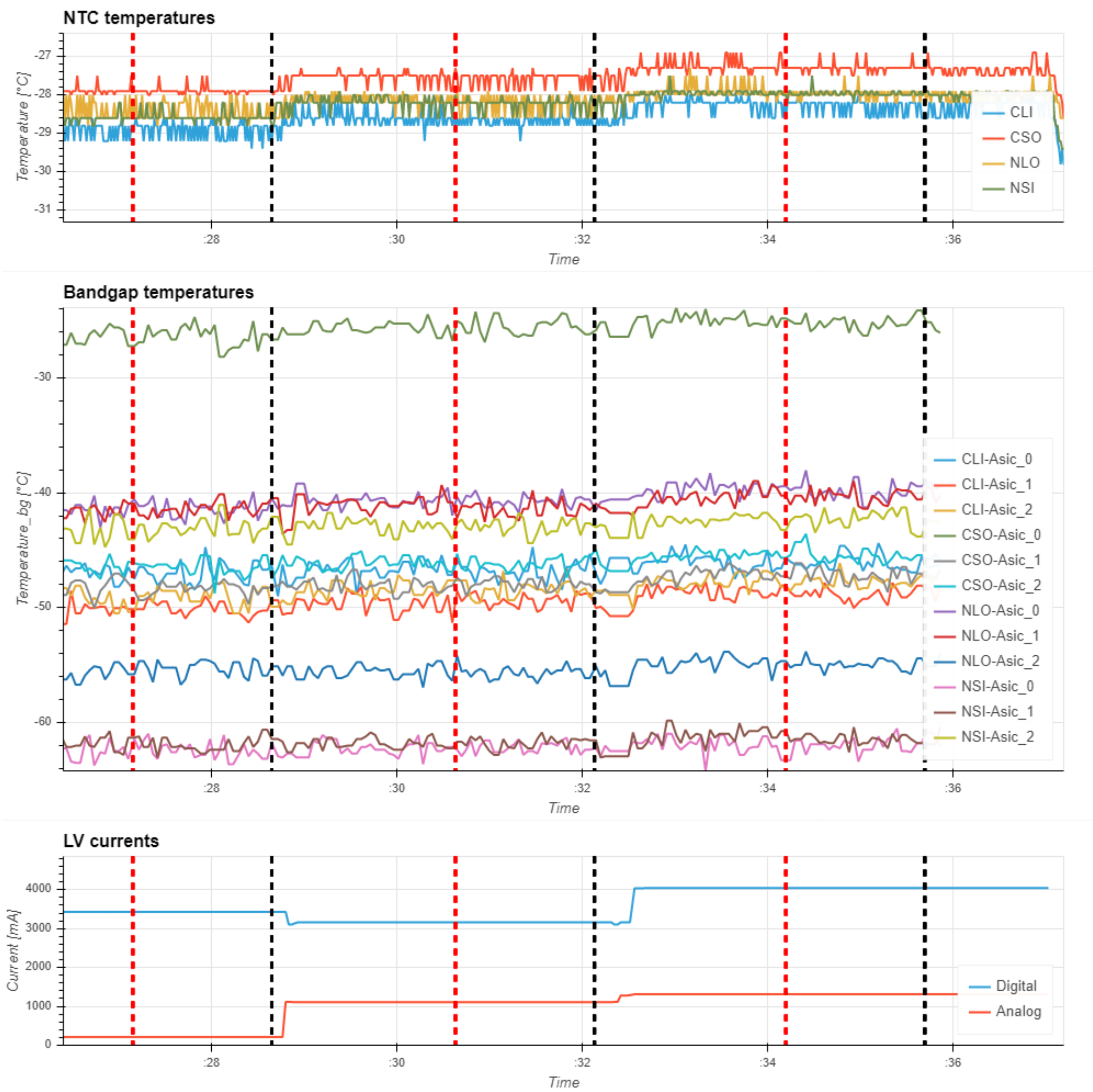
$$\Delta T'_{\mathrm{VP}_{2,3}} = \Delta T_{\mathrm{VP}_{2,3}} - \Delta T_{\mathrm{NTC}_2}. \tag{4.13}$$

The position of the individual tiles and NTC sensors is visualised in Fig. 4.23.

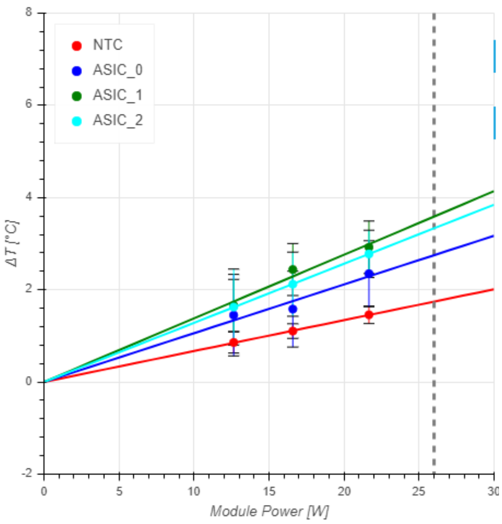
Since the cooling performance depends on the thermal conductivity and amount of the material present, a correlation between the glue layer thickness and the ΔT for each ASIC can be made. The dependency has been an important factor during the glue-selection studies before the start of the production discussed in Sect. 3.1.2.2, as it helped in ruling out Araldite as a possible replacement for Stycast with catalyst 9.

Glue studies The comparison of the three glue types tested is in Fig. 4.24, all with the NTC values subtracted. The measurements with Stycast catalyst 23LV are from 40 modules, while about 5 modules have been tested with catalyst 9 and only a single module with Araldite. The data points shown are from tests both before and after TC, as well as from any repeated module tests.

Due to the glue deposition issues, mostly during the R&D phase, some of the modules had a larger glue thickness underneath the ASICs. This turned out to be useful as it revealed an expected dependency of the cooling performance on the thickness. Even though the dependency seems linear, the values come from different tiles which may be differently cooled based on the amount of the overhang and the placement on the substrate. It is therefore only used for a visual estimate.



(a) : Measured quantities during the test.



(b) : Fitted cooling performance of the CLI (VP0) tile.

Figure 4.22: Example of the automatically produced bandgap plots for MOD_119.

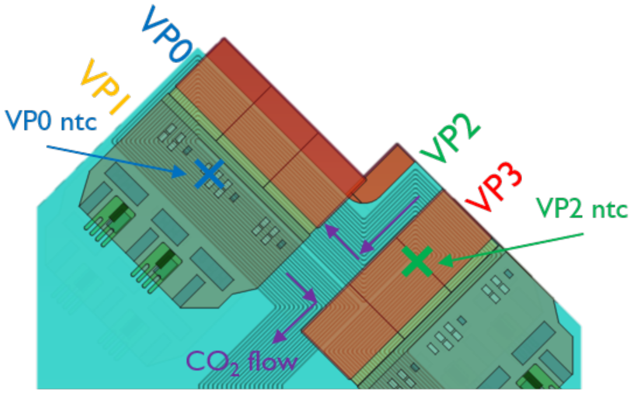
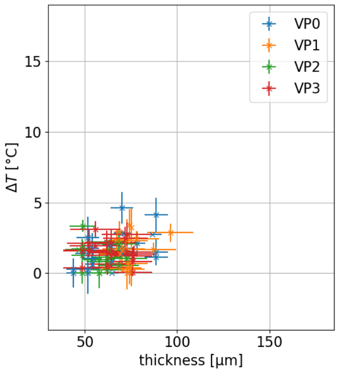
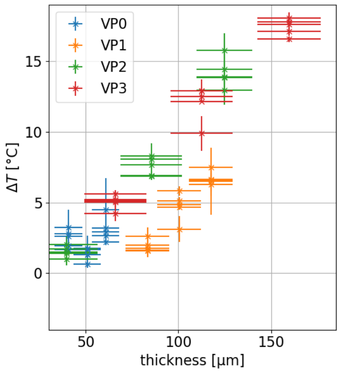


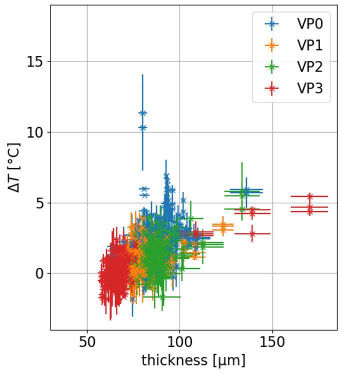
Figure 4.23: Position of tiles and NTC sensors on a module.



(a) : Stycast, cat9.



(b) : Araldite.



(c) : Stycast, cat23LV.

Figure 4.24: Comparison of the cooling performance based on the glue layer thickness.

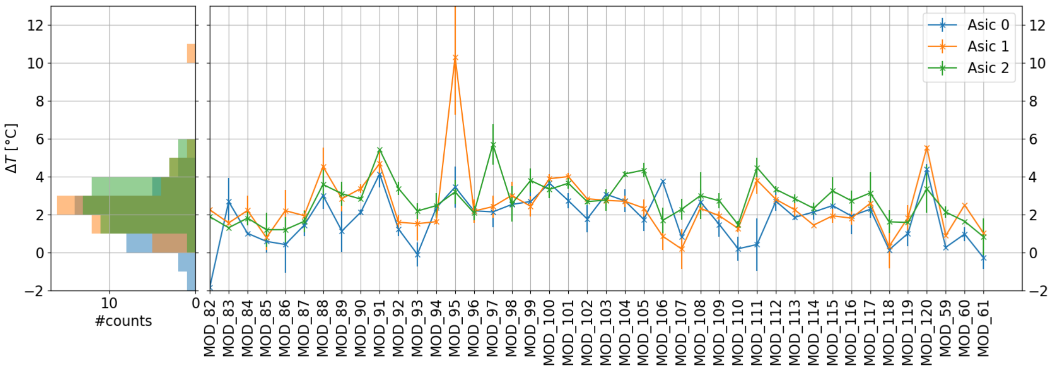


Figure 4.25: Trend of cooling performance for CLI (VP0) before TC.

From the Araldite plot it is clear that with the glue thickness above 100 µm, the end of life performance could easily be affected. While its adherence is excellent, it does not outweigh the need for a good cooling solution. For this reason, the Stycast with catalyst 23LV has been favoured as a replacement for the catalyst 9, since it provided very good cooling performance along with an improved adhesiveness.

grading The grading is based on the extrapolated cooling performance of all tiles at

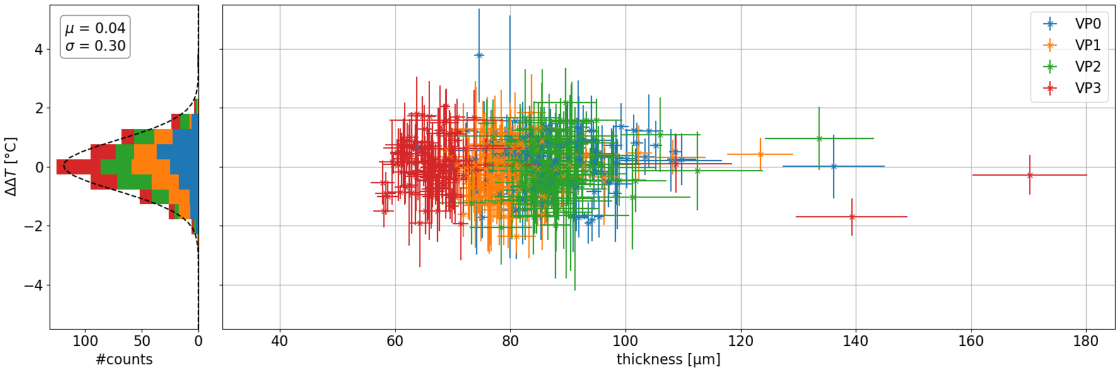


Figure 4.26: Dependency of thermal degradation on glue layer thickness after TC. The glue used was Stycast catalyst 23LV.

26 W with the NTC values subtracted. It accounts for the estimated worsening at the end of life of the VELO, the values specified in Table 4.2e.

4.4.5.2 Thermal degradation

The worsening in the cooling performance is an additional analysis based on the information from the bandgap test, comparing the results from before and after TC. No additional data have to be uploaded, the operator is only required to select two corresponding tests in the database they want to relate. The glue thickness test may be selected as an additional parameter, generating a thickness dependency plot of the cooling worsening.

The thermal worsening for all of the Stycast catalyst 23LV modules is plotted in Fig. 4.26. No visible change in the performance based on the TC is visible, regardless of the deposited glue thickness. This test works as an addition to the standard thermal performance which should spot bad cooling conditions.

grading Since the variation of the $\Delta\Delta T$ is quite large compared to the mean shift, the grading could be distorted with outlying values. For this reason, the mean and maximum of the absolute values of the degradation are used, better characterizing any overall effect of the TC (regardless if positive or negative). The corresponding grading limits are defined in Table 4.2f.

4.4.6 High voltage

The last component of the electrical performance tests verifies the HV biasing of the attached tiles. At first only the standard IV curves were processed, however, a simple ramp-up has been later added in order to verify the performance at the nominal cooling temperatures −30 °C. While the curves are done both before and after TC, the ramp-up is considered to be an extra safety check and can therefore be measured and uploaded only after the TC.

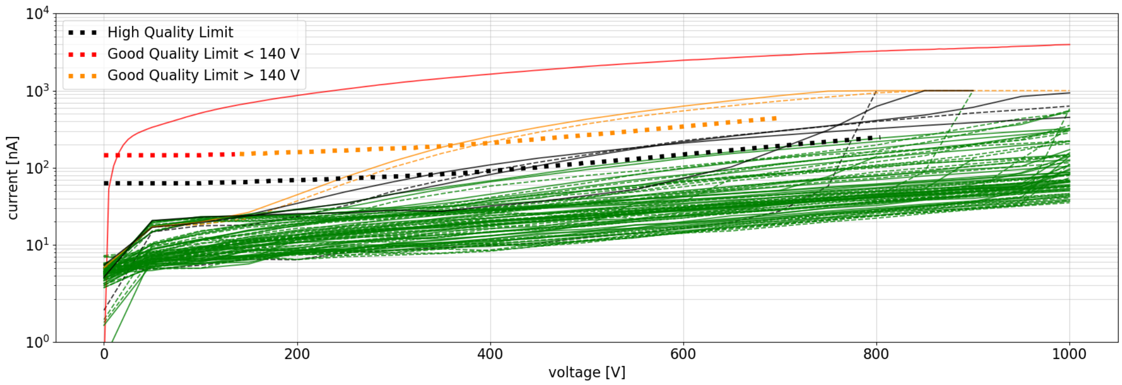


Figure 4.27: All IV curves for the NSI (VP2) tile. Plotted results before and after TC as solid and dashed dotted lines, respectively. The black, orange and red dots represent defined quality limits.

4.4.6.1 IV

The obtained IV curves from the scan (see Sect. 3.4.6.2) are saved in the same format in both UoM and Nikhef, matching the previous probe-station testing of the tiles done at CERN.

An example of all IV curves measured on tile NSI is plotted in Fig. 4.27, the before and after TC distinguished by solid and dotted lines, respectively. The black, orange and red lines represent the performance limits defined from the previous testing, the scan data are coloured based on crossing the individual lines (green in case they do not cross any).

Even though the tiles used in the test were mostly of high quality (not all were perfect due to the availability of the components), some of the scans indicate a worsening after the attachment. This could be caused by the large temperature difference at which the module is tested (0 °C vs the former 20 °C) as well as any system-wide effect (e.g. capacitor present on the HV tape or a VFB connection). In some cases the performance improved after the module remained in the vacuum, possibly due to the removal of any humidity adsorbed in the module or its components. In one of the measurement (see red line in the Fig. 4.27) the tile has been damaged during the assembly, causing an early breakdown.

grading As mentioned above, the grading defined in Table 4.2g is based on the limits from the probe-station measurements. This has been slightly modified in order to better characterize the standard operation performance at 140 V, as the module can still be used even if the current starts to rise after reaching the nominal bias voltage.

4.4.6.2 Ramp-up

Similarly to the IV test, the HV ramp-up contains data in the same format, the test is described in Sect. 3.4.6.3. Even though only the maximal voltage values are tracked by the database, the measurement records the whole progress of the data taking. An

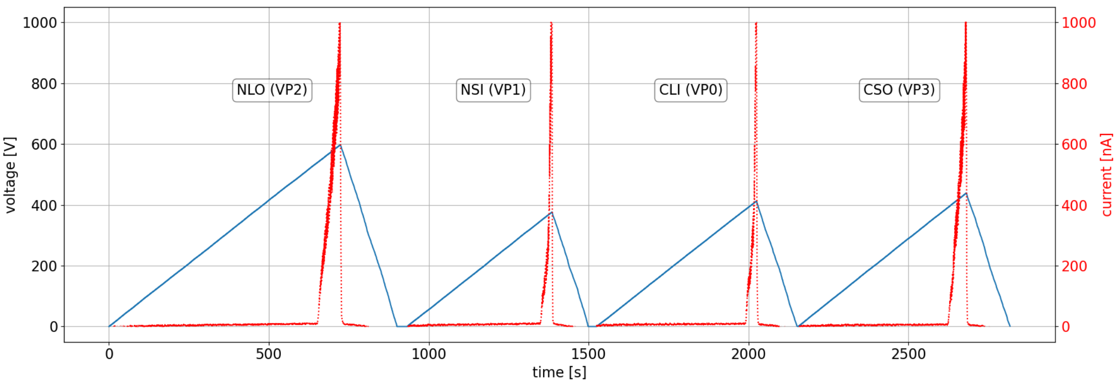


Figure 4.28: Example of voltage ramp with current monitoring for MOD_110.

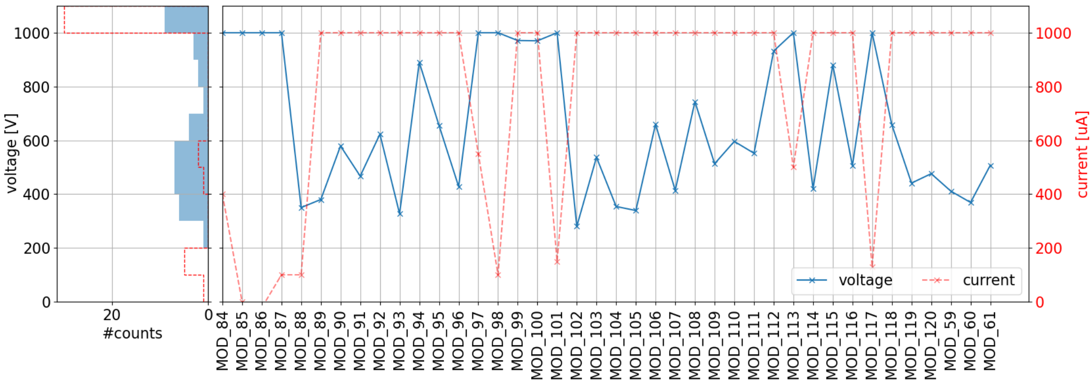


Figure 4.29: Trend of ramp-up results for NLO (VP1) tile.

example of such a scan for MOD_110 is in Fig. 4.28, with the trend of the maximal values reached by all modules in Fig. 4.29.

Since this test only serves as a quick verification of the HV operability at cold temperatures, the main outcome is whether the module can be biased well above the nominal voltage of 140 V or not. It is expected that the leakage current will change with the increase of the radiation damage, therefore the performance will have to be monitored during the VELO operation.

grading The grading levels defined in Table 4.2h check the minimal reached voltage of all tiles. This provides an extra information for the performance of the module, however, unless being graded F it is not considered an important factor.

4.5 Evaluation

While the previous sections described the individual analysis and grading of the tests, the information may be used to describe the overall grade and quality of the assembled module. This can be used to decide on the placement of the modules within the VELO halves, ensuring the best modules are in the most important regions.

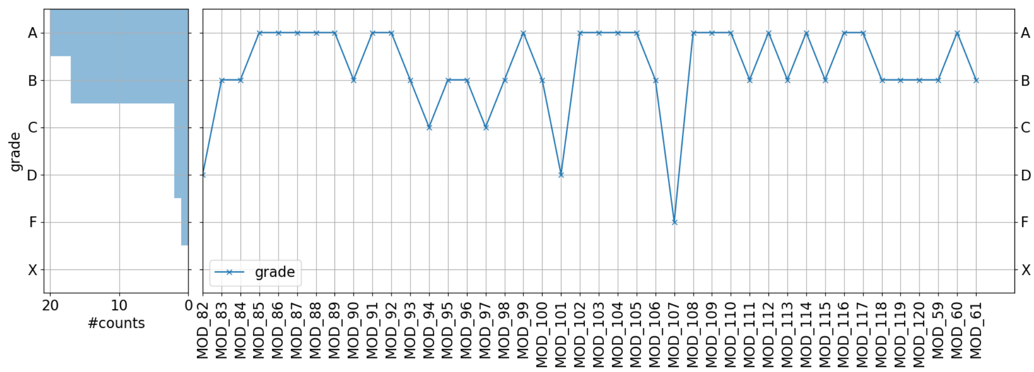


Figure 4.30: Final module production grades.

Table 4.3: Module grade counts.

grade	**A**	**B**	**C**	**D**	**F**
modules	20	17	2	2	1

4.5.1 Grading

The module grading is defined based on the number of the lowest graded assembly and testing steps. Single grades D and F are directly propagated to the module as such cases require further investigation which has to be easily visible in the database. On the other hand, the lowest grade from the A-C range is assigned only if there are more than two results with such grade, otherwise a grade one higher is selected. This provides a very quick metric to describe the performance of the module and its goodness, the results summarised in Table 4.3.

As was described before, only the components of grade C and worse affect the assembly grading steps, which in turn defines the module grade. It is important to note that instead of using the individual grades from the electrical tests before and after TC, only the final grades from E1 and E2 test suites are used. Because the PRBS test could not be reliably performed in the UoM (discussed in Sect. 4.4.3), a similar approach has been taken in propagating its value to the grade of the electrical tests. Therefore the results of the PRBS test are ignored unless graded D or worse, indicating a real problem in the system rather than only an increased value in BER at the level of $1 \cdot 10^{12}$ bits tested. The same applies to the HV-ramp test where only the failure propagates to the electrical test suite.

The main downside of this grading approach is that it makes it impossible to distinguish the perfect modules and the very good ones. For example, a module with all grades A would be graded the same as a module with cooling performance grade B, larger BER resulting in grade C and grade B wire-bond pull test results. For this reason a quality metric has been proposed to use in parallel to the simple grades.

Table 4.4: Weights used for quality calculation.

grade	**A**	**B**	**C**	**D**	**F**
weight	25	15	0	-50	-250

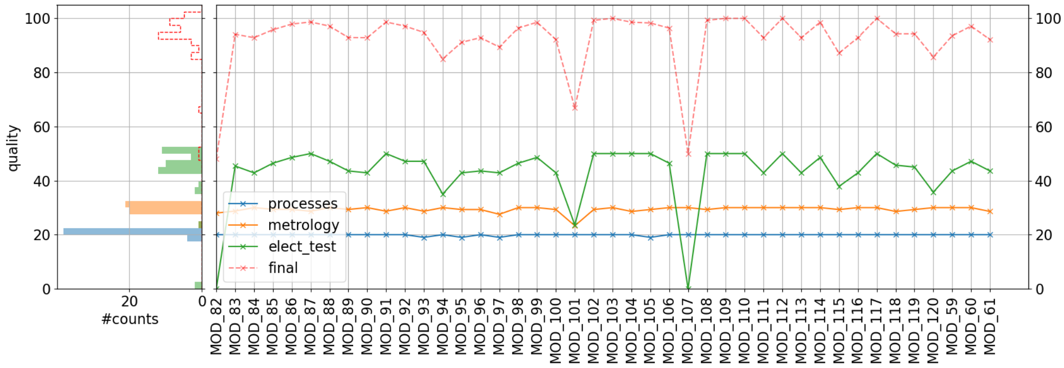


Figure 4.31: Resulting quality of the modules, defined from assembly processes, metrology and electrical tests in 20:30:50 ratios.

4.5.2 Quality

Similarly to the module grading, the quality is also based on the grades of tests, however, using the individual electrical tests instead of the test suites. This means that no grade is ignored, furthermore, the performance results contribute more to the final quality compared to the assembly itself.

The production steps are split to the three groups:

processes containing all assembly steps,

metrology consisting of metrology, TC and thermal degradation,

electrical representing the full suite of the electrical tests done before and after TC.

The quality number is defined in a range from 0 to 100, the ratio of the groups is 20:30:50, respectively.

The calculation of the corresponding quality values is then defined as weighed average of the grades within the group, the weights summarised in Table 4.4. Even though the weights of grades D and F have negative values, the minimal value for each group is fixed to be 0.

Using the definitions above, the trend of the quality values is plotted in Fig. 4.31, split based on the groups and the summed result. The plot clearly shows how the individual groups propagate to the final quality, its distribution mostly spanning 80-100 with minor variations.

While it is clear from the trend that MOD_107 has poor electrical performance with quality at 0, only using this information could mean that all of the tests were grade C or some other combination of grades A-D. Because of this the two QA metrics have to be used at the same time in order to properly qualify the modules. The combination can

Figure 4.32: Final grade and quality of the modules as of 31 August 2021.

be plotted together as shown in Fig. 4.32, visualizing both grade as well as its quality. The aforementioned MOD_107 is clearly graded F and cannot be therefore used for the installation in the VELO halves. On the other hand, there are some modules grade B with better quality than grade A ones. Ideally, A grade modules with the quality 100 are to be used in the most important positions in the VELO.

4.6 Summary

A database is an important tool used to store data and track progress of a range of objects. This is also true for the VELO Upgrade, where data from the various tests are uploaded and automatically processed. In order to minimise the error, simplify the analysis and track any issues originating during the module construction and testing, a number of algorithms has been developed that take care of decoding the inputs and produce simpler metrics and plots. This has been later improved upon by tracking the time-evolution of the individual tests and devising a grading scheme used to characterise the modules. These results will be used in the assembly of the VELO halves, aiming to have the highest quality modules closest to the collision point.

Blank page

Chapter 5

Charge collection efficiency scans

The VELO detector has to be prepared for its full operation in advance, making sure all DAQ and monitoring systems work as planned. This requires a large number of subsystems to be developed and tested. In addition, there are other tests which will need to be executed during the operation of the modules, making sure the operation of the individual detectors is the best achievable. For this reason various DAC and voltage scans have to be implemented, the work described here focuses on the Charge Collection Efficiency (CCE) scan.

5.1 Charge collection efficiency

As was described in the Chapter 2, the ionizing radiation traversing through the silicon sensor generates e-h pairs which induce current on the collection electrodes. The current is then amplified, shaped and read out by the circuitry.

In order to maximise the signal-to-noise ratio, it is beneficial to keep the sensors fully depleted throughout the operation. The depletion depth is mostly affected by the radiation-induced damage described in Sect. 2.3.2.2, where the voltage needed to fully deplete a sensor is proportionally increased based on the fluence. Since it is difficult to measure the CV curves in an installed tracker, a CCE scan characterising the ratio of collected and generated charge is commonly used. The scan results are translated to the depletion depth, effectively providing a dependency of the depletion voltage on a fluence (see Fig. 5.1).

The distribution of the deposited energy (and subsequently the number of generated e-h pairs) in a thin silicon sensor may be described by the Landau-Vavilov distribution convolved with a Gaussian distribution (see Sect. 2.1.2). Due to its elongated tail towards the large energy losses, the MPV is the best metric used to describe the CCE performance. With the increase of the bias voltage, the MPV shifts towards a higher mean energy loss up to a point of the full sensor depletion where it stays mostly constant.

The interaction of a MIP corresponds to about 80 e-h pairs generated per 1 µm of silicon. For a sensor of thickness 200 µm used in the VELO, the most probable amount of charge generated from a single particle is approximately 16 ke$^-$.

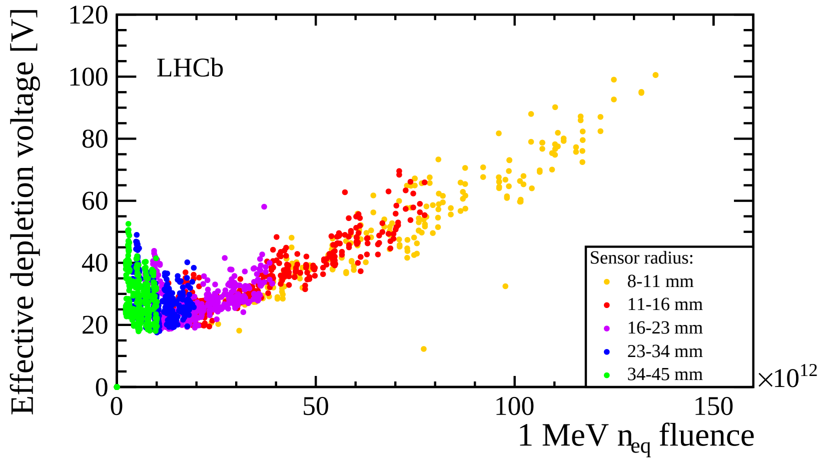
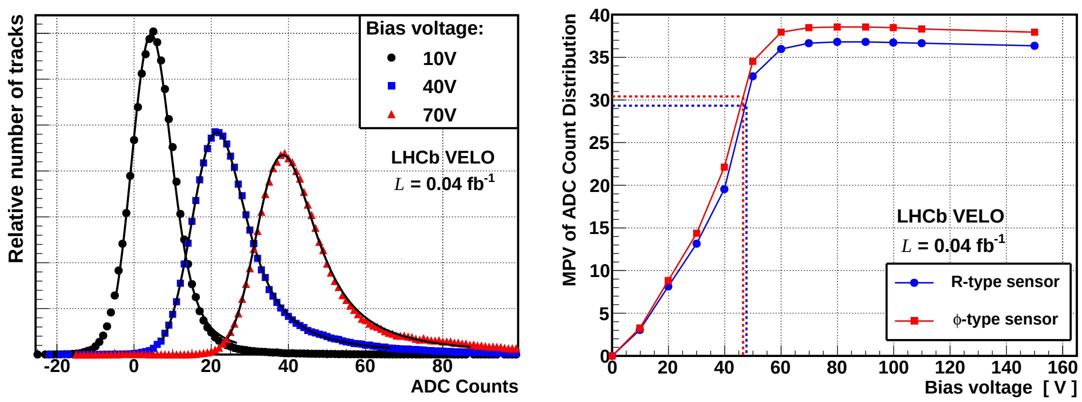


Figure 5.1: Dependency of effective depletion voltage on fluence for the Run 1 and 2 VELO sensors. Results after the end of Run 1 with total integrated luminosity $3.4\,\mathrm{fb}^{-1}$. [22]



(a) : Pedestal subtracted ADC distributions for R-type sensors.

(b) : Dependency of MPVs of the Landau distributions on voltage. Dashed line represents 80 % of the signal and corresponding effective depletion voltage.

Figure 5.2: Example of the CCE scan for the Run 1 and 2 VELO. [90]

5.1.1 Scan definition

The simplest way to obtain the depletion voltage from the CCE is by performing a voltage scan, ranging from an under-depleted to an over-depleted voltage values. At each voltage step, the MPV of the charge distribution has to be obtained. Relating the MPV to the voltage results in an S-curve dependency from where the full depletion is simply obtained. An example of such scan is plotted in Fig. 5.2.

These scans are normally performed using any form of a charge correlated channel response. This could be done using the ADC counts (representing the maximum of the induced signal) or ToT values (representing the duration of the shaped analogue signal). While the first method has been utilised in the Run 1 and 2 VELO [90] as it was measured by the the Beetle ASICs, the VeloPix in the upgraded detector uses either ToT or binary read out.

Even though the ToT values were previously used to characterise the tile performance [41], the complexity of linking the slow ECS data taking would be too large and time consuming as the full pixel matrix would have to be read out. For this reason a method using the standard readout of the detector with the GWT data has been proposed, utilising only the binary pixel hit information.

5.1.1.1 Threshold dependency

The main downside of the hit counting of a binary system such as the VELO upgrade is that it removes any discrimination of the collected charge. Since the scan is based on an estimate of the MPV, the signal information has to be somehow obtained. This can be achieved by performing a threshold scan for each voltage step, covering most of the expected signal range (from the noise limit of $1\,\mathrm{ke}^-$ up to about $30\,\mathrm{ke}^-$).

The obtained scan results represent an amount of all energy collections higher than the threshold. Given the number of both detected and undetected particle interactions per step is known, the threshold dependency will represent the cumulative sum of the charge collections. It can be either directly fitted with a Landau-Gaussian integral or reconstructed to obtain the original distribution and then fitted with the raw function. While both approaches yield the MPV, the latter is more robust as the initial fit parameters may be better defined.

5.1.1.2 Beam dependency

Performing such a 2D threshold and voltage scan would be simple for a uniformly irradiated device using a homogenous beam of MIP, however, the conditions in the real system strongly depend on the exact position of the individual pixels. The dependency on the distance from the beam results in a highly inhomogeneous irradiation and different number of particles traversing the detector. Because of this the pixel hits have to come from the reconstructed tracks which also helps in defining the ratios of the collected/missed interactions. Therefore the obtained data may be presented as hit finding efficiency and have to be obtained from the nominal LHCb operation conditions which imposes constraints on the scan duration.

Other geometrical effects also affect the performance of the scan, namely the track angle and the hit position, however, these will be discussed later in the text.

5.1.2 Simulation

In order to verify that the proposed scan works before testing it on a fully built VELO upgrade detector, parts of the scan have been simulated within the LHCb software framework. While the software tools range from the simulation of the particle collisions all the way to the reconstruction and data analysis, only the parts representing the simulation of the digitisation were used, relying on the previously simulated particle collisions and interactions. This is governed by the Boole project.

Table 5.1: Boole simulation parameters.

DDDBtag	upgrade/dddb-20201211
CondDBtag	upgrade/sim-20201218
Type	Monte-carlo Upgrade, minbias
Energy	7 TeV
Simulation	Sim10-Up08 – Digi15-Up04
Events	100
Pile-up	5.2

Boole [91] The project is based on the Gaudi framework, structured to separate individual subdetectors. It handles the digitisation of the Geant4 [92] produced detector hits and outputs the same format as comes from the experiment (with the additional truth information).

The simulation links to an existing file and loads all required conditions from databases. It then digitizes the interactions based on the options provided in a separate file.

The part that is relevant to the VELO takes care of generating the e-h pairs along the particle track so that they match the Landau distribution of charge deposition. Afterward it applies spatial smearing and sums up the charges collected by the individual pixels.

While the tool already provided a way to simulate the radiation damage and to change the depletion depth accordingly, modifications had to be implemented to mimic the dependency of the depletion on the applied bias voltage. This dependency follows the Eq. (2.36) and is simplified as

$$d = 20\sqrt{V_{\text{bias}}}, \tag{5.1}$$

corresponding to a full depletion depth $d = 200\,\mu\text{m}$ at $V_{\text{bias}} = 100\,\text{V}$.

The CCE simulation has been performed using the modified Boole project with tags and data specified in Table 5.1, including signal spillover from the previous and following events. The simulated dependency of the collected charge on the voltage is plotted in Fig. 5.3, the data match the ideal dependency described by Eq. (5.1). However, more background is present in the low-charge region, caused by the impact of the track angle and position, described in the next section.

5.1.2.1 Angle and position dependency

In the case where the particle is not in a perfectly perpendicular direction towards the detector surface and is not hitting the middle of the pixel, the e-h pairs may get created in multiple adjacent channels. Normally, the collected charge can be reconstructed by clustering these hits and summing the obtained signal. The resulting strength of the signal is then directly proportional to the depletion width, a small difference present due to the

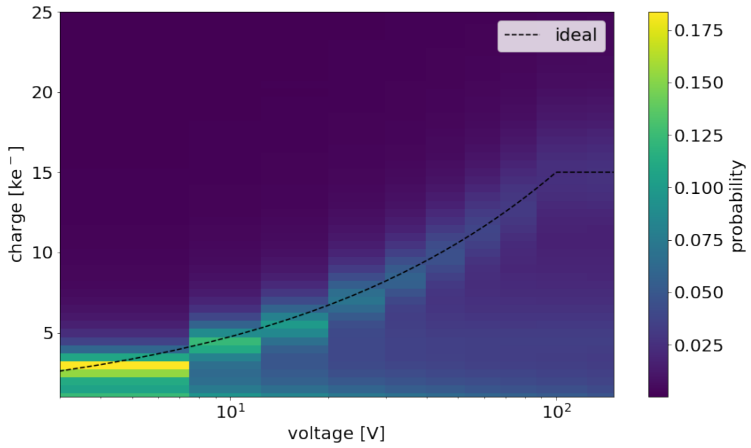


Figure 5.3: Raw data output from simulated dependency of charge on voltage for non-irradiated VELO. The Monte-Carlo results follow the ideal dependency defined by Eq. (5.1).

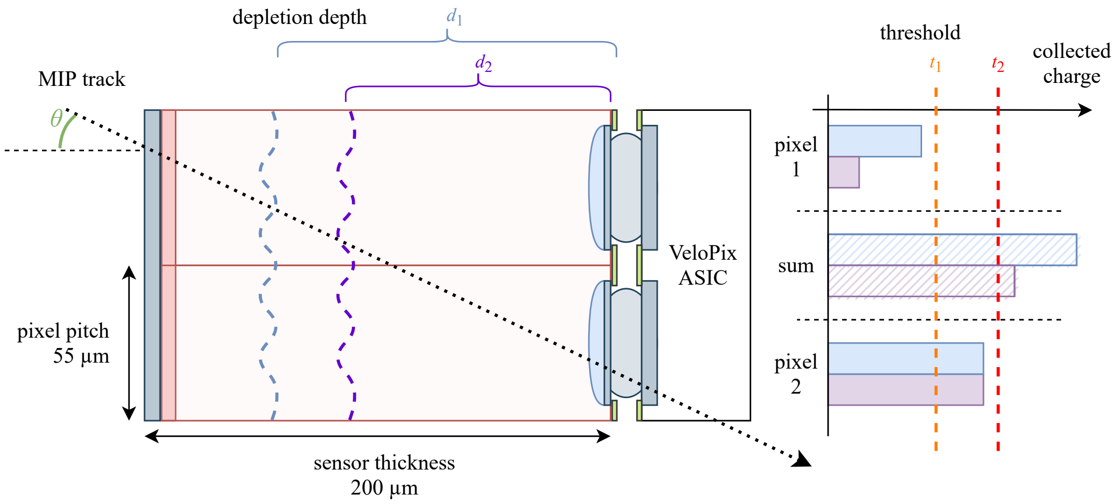


Figure 5.4: Schematic view of a MIP traversing two pixels under an angle θ. Dependency of the collected charge based on the depletion depth d is visualised on the right side. Since the signal is discriminated on a per-pixel basis, the threshold t affects the amount of reconstructed charge as it does not follow the sum of the two pixels.

impact angle θ which increases the track length in the sensor by a factor of $1/\cos(\theta)$. This causes a broadening of the Landau distribution (as described in Sect. 2.1.2), which is small at the LHCb due to its forward geometry (the acceptance being up to 17° at which point the effective depth changes at most by 10 %). A bigger effect may be caused by the particles traversing the sensor at larger impact angles and these need to be excluded by the tracking.

However, a binary read out relies on a threshold level and therefore has a much more complicated reconstruction process in which the collected charge strongly depends on the track (see Fig. 5.4). In this case the particle deposits energy in two channels with the sum related to the depletion depth d. In one pixel, the applied voltage results in a

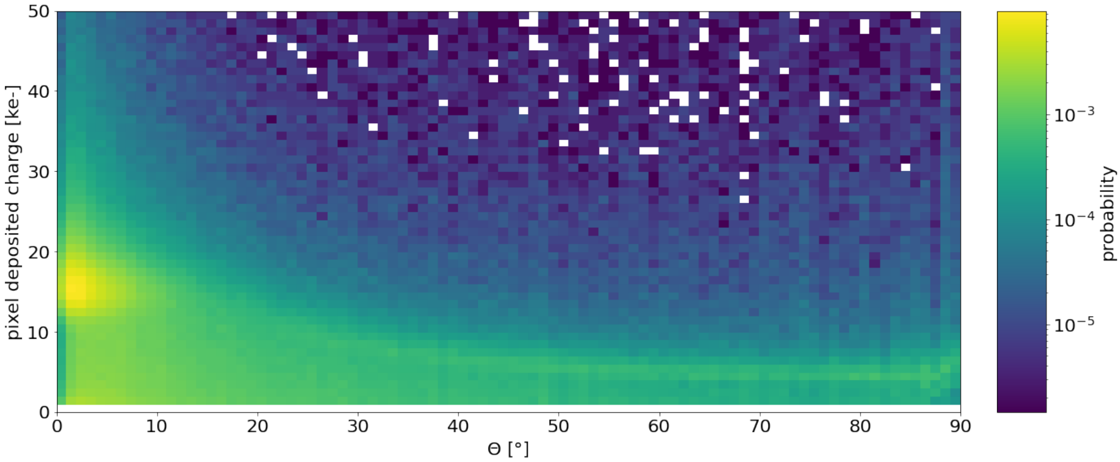


Figure 5.5: Angular dependency of the pixel collected charge, obtained at 100 V, full depletion of the simulated sensors. A visible decrease of the per-pixel deposited charge may be observed with an increasing impact angle.

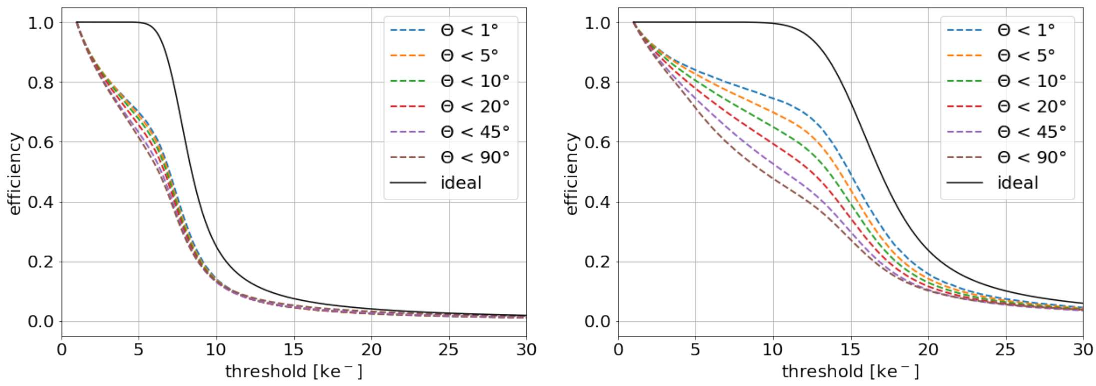
depletion larger than the corresponding track length, therefore the pixel will collect the same amount of charge in both cases. However, in the other pixel the signal strength is visibly affected as the depletion depth lowers the effective track length that can be read out.

Since the signal is discriminated on a per-channel basis and only binary yes/no information is recorded, it is impossible to obtain the sum of the collected charge from both pixels. The effect is best explained by applying a threshold t higher than the maximal charge collected by the pixels, however, lower than the sum. In this case no hits are detected, even though there was enough energy deposited to overcome the discrimination level for both depletion depths. Performing a threshold scan would therefore result in skewed data, reconstructing only a fraction of the charge and creating low-energy noise in the reconstruction.

This effect may be minimised by applying cuts on the extrapolated interaction positions, the simplest approach is done by ignoring all tracks with large impact angles θ with respect to the normal of the sensor plane. This can be visualised as a dependency of the pixel deposited charge in a fully depleted sensor on the angle, shown in Fig. 5.5.

The charge sharing effects may be investigated by plotting the scan results for different angular cuts, shown in Fig. 5.6. It is clear that with a tighter cut on the angle of incidence, the shape of the expected S-curve improves, however, even for a cut of $\Theta < 1°$ the resulting shape is skewed and does not reach the ideal efficiency. The reason for this is that while the angle significantly reduces the charge sharing, some tracks interact on the border of the pixels, therefore the deposited charge is split.

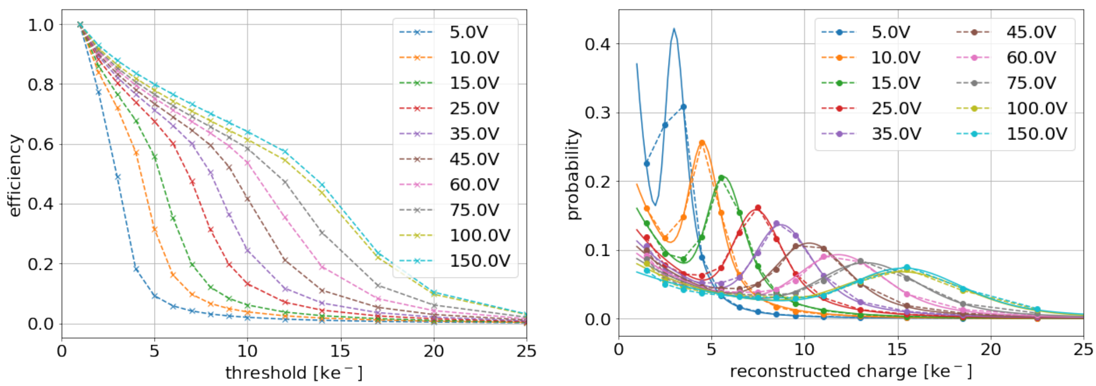
While the results are obtained from the simulation without any tracking and detector misalignment, it should be possible to apply the angular cut order of a few degrees. Applying a spatial cut on the track position within the pixel would be also beneficial, however, this is not possible due to multiple scattering and inadequate precision ($\mathcal{O}(10\,\mu\text{m})$). In the



(a) : $V_{\text{bias}} = 25\,\text{V}$.

(b) : $V_{\text{bias}} = 100\,\text{V}$

Figure 5.6: Dependency of the hit reconstruction efficiency on the applied threshold for different angular cuts.



(a) : Cumulative sum of hits obtained from the scans.

(b) : Reconstructed and fitted Landau distribution.

Figure 5.7: Simulation results of a threshold scan for non-irradiated VELO sensors, made for different bias voltages. Data are normalised based on the cumulative sum. The lines for 100 V and 150 V overlap due to reaching maximal depletion.

following simulations, a cut on the impact angle $\theta < 10°$ has been applied. This arbitrarily chosen value is proposed to improve the landau reconstruction while accounting for the multiple scattering and tilt of the silicon sensors. The cut may be later changed according to the more precise results from the track reconstruction, and detector operation and alignment.

5.1.2.2 Landau reconstruction

As was mentioned before, the proposed CCE scan relies on a 2D voltage and threshold scan which results in a cumulative sum of the collected charge. Considering the previously-discussed tracking inputs, the reconstruction of the MPV from the data is done by projecting the cumulative sum back into a Landau distribution for each of the voltage steps. This is achieved by subtracting the adjacent threshold bin counts and assigning their difference to a middle point. Such an approach can be easily used for data with a variable threshold step.

Table 5.2: List of 15 thresholds used for the simulated CCE scan.

threshold [ke$^-$]														
1,	2,	3,	4,	5,	6,	7,	8,	9,	10,	12,	14,	17,	20,	25

In order to apply the algorithm to the simulation data, at first the cumulative sum has to be created from the saved charge collection values. Since the numbers are obtained without any additional digitisation effect, a Gaussian smearing with $\sigma = 100\,\mathrm{e}^-$ (estimated from the VeloPix noise scans, see Sect. 4.4.4) is applied. The created cumulative sum along with the recreated and fitted Landau distribution from the simulation is plotted in Fig. 5.7. The background present in the low threshold bins has been mitigated using an exponential function.

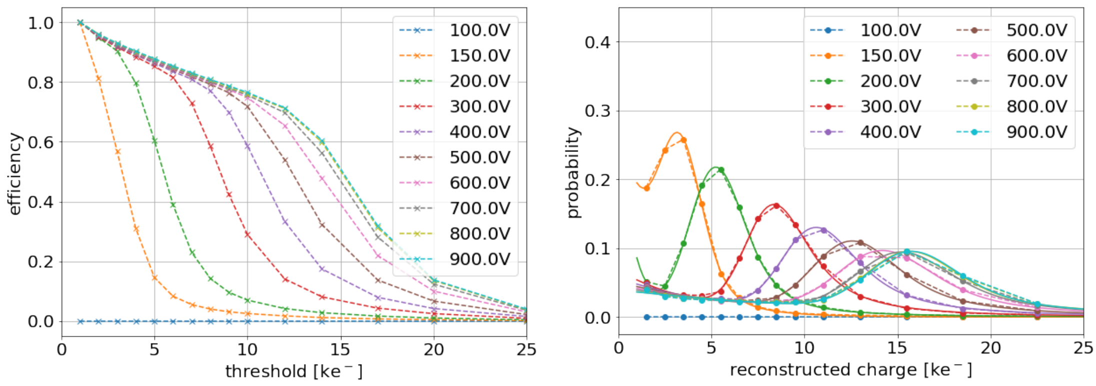
As it is important to properly match the variations in the region with a lower collected charge (defining the narrow Landau peak and noise), more scan points should be taken in this region. The threshold scan has been simulated using 15 different steps, summarised in Table 5.2. The precise DAC configuration controlling the threshold and the signal shaping circuitry required to obtain these settings in the physical VeloPix ASIC is yet to be investigated.

5.1.2.3 Radiation damage

Since the main reason to perform a CCE scan is to monitor the radiation damage of the individual modules such that the reverse-bias voltage can be increased accordingly, the performance of the proposed reconstruction has been tested using simulated irradiation of sensors. The effects of the fluence have been previously parametrised and implemented in Boole based on the exact radial distance from beam and module location along the z axis [93]. As a matter of simplification, the fluence in the simulation only scales the effective depletion depth.

The results of the simulation for integrated luminosity $L = 10\,\mathrm{fb}^{-1}$ are plotted in Fig. 5.8, produced with the same angular cut as before. Because of the strong positional dependency and missing tracking, another cut has been applied selecting the pixels based on the simulated fluence $0.5 \cdot 10^{15}\,1\,\mathrm{MeV}\,\mathrm{n_{eq}}\,\mathrm{cm}^{-2}$ to $1 \cdot 10^{15}\,1\,\mathrm{MeV}\,\mathrm{n_{eq}}\,\mathrm{cm}^{-2}$. This cut corresponds to pixels $(7 \pm 1)\,\mathrm{mm}$ distant from the collision point. Compared to the non-irradiated sensors, the total amount of the collected charge is lower as expected, obtainable only at much higher applied voltages.

The overall dependency of the MPV on the voltage can be therefore recreated using the fit parameters, the results plotted in the Fig. 5.9. The simulated results compare beginning of operation with the irradiated sensors, the outcome is that this scan approach should be feasible for the CCE scan.



(a) : Cumulative sum of hits obtained from the scans.

(b) : Reconstructed and fitted Landau distribution.

Figure 5.8: Simulation results of a threshold scan for irradiated VELO sensors, made for different bias voltages. Data are normalised based on the cumulative sum, the simulated fluence is in range $0.5 \cdot 10^{15}\,1\,\mathrm{MeV}\,\mathrm{n_{eq}}\,\mathrm{cm}^{-2}$ to $1 \cdot 10^{15}\,1\,\mathrm{MeV}\,\mathrm{n_{eq}}\,\mathrm{cm}^{-2}$.

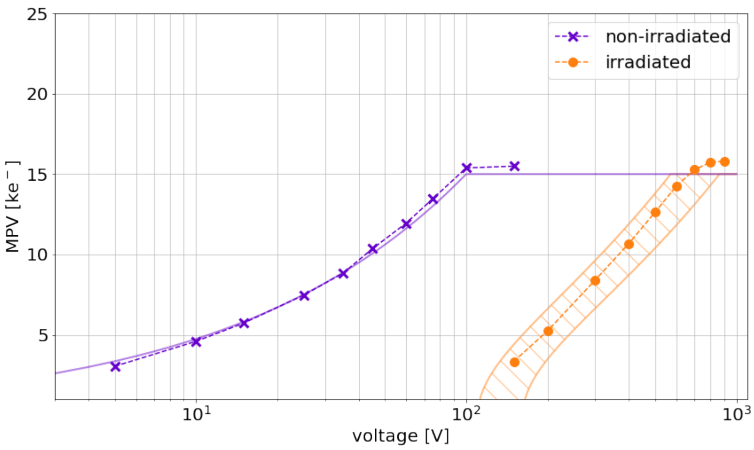


Figure 5.9: Dependency of the MPV from the Landau fit on voltage for normal and irradiated sensors. The solid line and filled region represents the underlying input of the simulation.

5.1.3 Scan duration estimates

Even though the simulation results indicate that it should be feasible to use the 2D voltage and threshold scan to reconstruct the dependency of the MPV on voltage, the duration of the scans is also an important factor to consider. The VELO upgrade consists of 52 modules in total where each of them should be regularly scanned in order to keep their performance at peak. The previously-used approach may be kept, where every fifth module is masked from tracking and the individual scan parameters are varied. The process is visualised in the Fig. 5.10, in order to scan the full VELO detector the same set of instructions has to be repeated 5 times, every time selecting a different subset of modules.

Using similar approach to the Run 1 and 2 detector, the scan should be performed approximately every $0.5\,\mathrm{fb}^{-1}$ to $1\,\mathrm{fb}^{-1}$ of delivered integrated luminosity. While for the

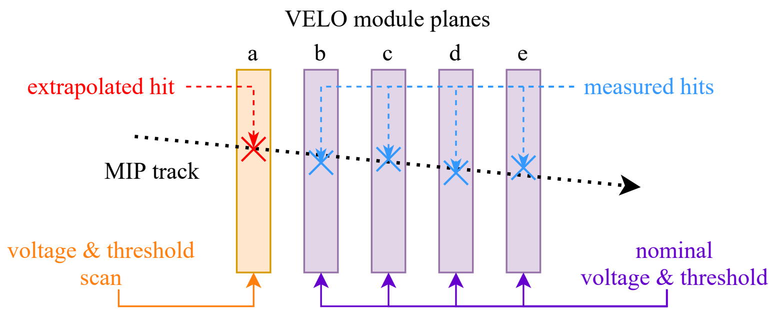


Figure 5.10: Schematic view of the CCE scan where one of the five modules is tested, the remaining four modules are operated in the nominal conditions.

old system this meant running the scan a few times throughout the year, due to the luminosity increase the VELO Upgrade would have to be scanned on a bi-weekly up to a monthly basis. Since the data taking has to come from the nominal LHC operation conditions, the scan duration needs to be as short as possible.

The estimate of the 2D scan duration t can be summarised as

$$t = t_{\mathrm{volt}}(V_{\mathrm{off}}) + t_{\mathrm{volt}}(V_{\mathrm{range}}) + \left\{ N_{\mathrm{volt}} \left[t_{\mathrm{stable}} + N_{\mathrm{DAQ}} \left(t_{\mathrm{config}} + t_{\mathrm{DAQ}} \right) \right] \right\} \tag{5.2}$$

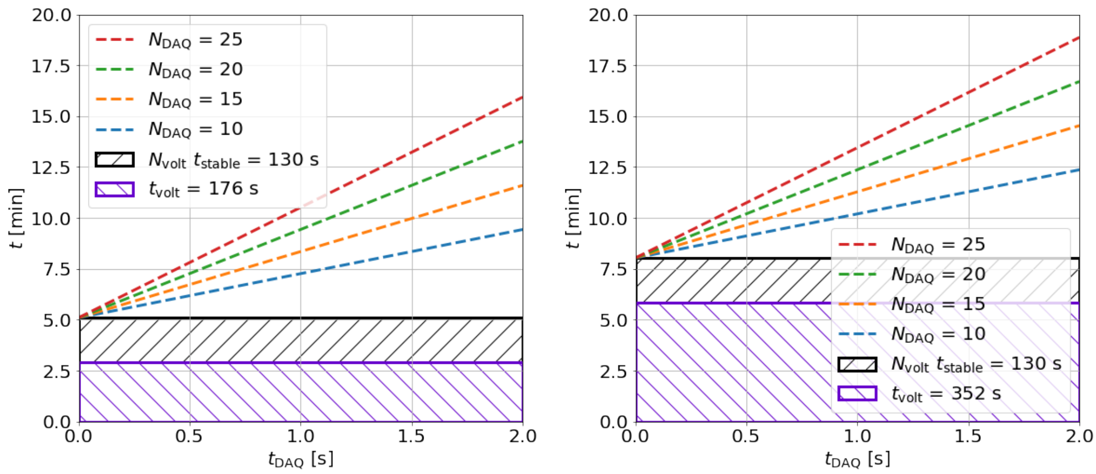
where $t_{\mathrm{volt}}(V)$, t_{stable}, t_{config} and t_{DAQ} represent the total time it takes to do a voltage ramp, stabilize current, configure VeloPix ASICs and take data, respectively. Similarly, the N_{voltage} and N_{DAQ} correspond to the number of voltage and DAQ steps. The voltages V_{off} and V_{range} define the offset of the first voltage point from 0 V and total voltage range covered from it.

In cases where the scan is performed after finishing the standard data taking, the sensors are already biased at the nominal voltage. Because of this the first factor $t_{\mathrm{volt}}(V_{\mathrm{off}}) = 0\,\mathrm{s}$ can be ignored. Using the same voltage ramp speed ν as was done during the IV-curve scanning ($1\,\mathrm{V\,s^{-1}}$ and $10\,\mathrm{V\,s^{-1}}$ for ramp-up and ramp-down, respectively) the duration depends only on the maximal range covered. This can be written as

$$t_{\mathrm{volt}}(V_{\mathrm{range}}) = \frac{V_{\mathrm{range}}}{\nu_{\mathrm{up}}} + \frac{V_{\mathrm{range}}}{\nu_{\mathrm{down}}} \tag{5.3}$$

and for a range of $V_{\mathrm{range}} = 160\,\mathrm{V}$ takes $t_{\mathrm{volt}}(V_{\mathrm{range}}) = 176\,\mathrm{s}$. Such duration estimate takes into the account the return to the starting voltage, needed for the next scan with a different subset of modules or for the return to physics data taking. Given it takes some time for the current to stabilise after changing the voltage values, the $t_{\mathrm{stable}} = 10\,\mathrm{s}$ is normally used in the module testing.

Another constant factor of the scan is the configuration time of the VeloPix ASICs, needed to change the threshold levels as well as covering any control software latency. Configuring the detectors via WinCC/OA script would represent a time penalty of $t_{\mathrm{config}} \approx 1\,\mathrm{s}$. Since the impact of this value quickly adds up based on the number of the



(a) : $V_{\text{range}} = 160\,\text{V}$.

(b) : $V_{\text{range}} = 320\,\text{V}$.

Figure 5.11: The CCE scan duration estimates, applicable to a subset (fifth) of the modules. Plotted for a different number of threshold scan points N_{DAQ} as a function of DAQ duration.

scan steps, a more beneficial approach is defining the threshold scan directly in the SOL40 firmware. A single DAC configuration performed this way takes only about $50 \cdot 10^{-6}\,\text{s}$.

The rest of the parameters covered by Eq. (5.2) are freely modifiable. Using the same configuration as was done for the former VELO [94], the number of voltage steps covering the range is $N_{\text{volt}} = 13$.

Applying the above defined values to the Eq. (5.2), the dependency of the total duration t can be then based on the number of threshold steps N_{DAQ} and corresponding data taking duration. The results for two different voltage ranges are plotted in Fig. 5.11.

In order to reproduce the Landau distribution using the proposed algorithm, the data should be taken for about 15 different thresholds, covering the lower energies in more detail. While the t_{config} should cover most of the system latencies, the t_{DAQ} needs to be found so that it produces enough tracks to perform a scan. This may be estimated using a full VELO simulation with track reconstruction, however, for the simplicity of this study a $t_{\text{DAQ}} = 1\,\text{s}$ will be considered.

The results for $V_{\text{range}} = 160\,\text{V}$ indicate that a single scan (performed on a fifth of the modules) should take about 8.3 min, the CCE scan of the whole VELO therefore done within 42 min. It is important to note that the minimal duration is governed by the voltage range which might have to be increased based on the radiation damage. Changing the range to $V_{\text{range}} = 320\,\text{V}$ while keeping the rest of the parameters the same, the full VELO scan would take up to 60 min. The duration could be lowered using a quicker voltage ramp-up, a change of ν_{up} from $1\,\text{V}\,\text{s}^{-1}$ to $2\,\text{V}\,\text{s}^{-1}$ would effectively halve the value of the t_{volt}. Similarly, a decrease of the number of the voltage steps or of the stabilisation time would lower the whole scan duration significantly. A comparison of some of these parameters is summarised in Table 5.3.

Considering the scan has not yet been tested on the detector itself, it would be beneficial to perform the first scans using more threshold points N_{DAQ} with a longer DAQ duration,

Table 5.3: Summary and comparison of CCE scan duration parameters. Colour highlights different values with respect to the first (default) configuration.

μ_{up} $\text{V}\,\text{s}^{-1}$	μ_{down} $\text{V}\,\text{s}^{-1}$	V_{range} V	**ramp** s	N_{volt} -	t_{stable} s	**stable** s	N_{DAQ} -	t_{DAQ} s	**DAQ** s	**total** min
1	10	160 320	**176** **352**	13	10	130	15	1	**195**	**8.35** **11.28**
2	10	160 320	**96** **192**	13	10	130	15	1	**195**	**7.02** **8.62**
2	10	160 320	**96** **192**	10	10	100	15	1	**150**	**5.77** **7.37**

potentially taking about 10 min per a fifth of the modules. In case when the time requirements turn out to be the upper estimate, the scan may be also weekly or bi-weekly on a single subset of modules only. This way the CCE would be monitored with a higher frequency while the total impact on the physics data taking would be the same as that of a single full scan performed a few times a year. A potential downside of this approach is correlating the results between the modules as the integrated luminosity will vary.

5.1.4 Verification and required work

While the presented work relies on the commonly-used Monte-Carlo simulations, the proposed method could be verified using similar type of detectors in a test beam environment. This could be achieved with any beam of MIPs (e.g. high energy electrons, pions, protons,...) as the e-h pair generation and signal collection should be the same regardless of the particle type.

An example could be using any telescope with a 200 µm thick sensor Device Under Test (DUT) and either a VeloPix or a Timepix3 based readout ASIC. The test would be repeated both before and after irradiation of the DUT, matching the expected fluence up to $8 \cdot 10^{15}\,1\,\text{MeV}\,\text{n}_{\text{eq}}\,\text{cm}^{-2}$ of the end-of-life VELO Upgrade. Similar CCE scans of the sensors have already been performed using CERN test beam facilities [41], however, always based on the ToT readout and charge summing. In the case of this newly proposed method, the 2D voltage-threshold scan would be performed, taking binary data for number of voltage and threshold combinations. In order to match the varied tilt and the impact angle of the particles in the experiment, the rotation of the DUT with respect to the beam normal could be varied as well. The results would be then obtained in the form of device tracking efficiency.

Regardless of the possible beam tests, the proposed method still needs to be implemented and tested in the full track reconstruction which is part of the LHCb software framework, with every fifth module scanned. This could be done using the same digitised data available from the Monte-Carlo simulation, however, with much more events in order to increase the statistics in regions further away from the beam-pipe. This would provide better estimates of the data-taking duration t_{DAQ} for a single threshold step as well as a better handle on the angular cut of the impact angle (as the tracking would already include the sensor tilt). Afterwards, the actual threshold scan and data-taking using the GWT data-stream needs to be implemented in the firmware.

5.2 Summary

The CCE provides an important insight into the performance of the silicon sensors and the effects of the radiation damage. The characterisation of this efficiency is done by relating a MPV of the interacting MIP for different depletion widths. The required values are obtained using a CCE scan where the deposited energy is correlated to the bias voltage values. A new 2D voltage and threshold scan for the VeloPix ASICs binary readout has been proposed, and an estimate of its duration for the whole VELO calculated. The performance should still be verified using simulated tracking and an optimal DAC configuration of the VeloPix ASICs should be found and the scan implemented in the firmware.

Blank page

Part II

LHCb Upgrade II

Blank page

Chapter 6

Future upgrades

6.1 LHCb Upgrade II

Following the operation of the LHCb Upgrade, a next step is already in planning in order to improve the physics measurements done at the LHCb. The Upgrade II [95] (see Fig. 6.1) will be installed during LS4, ready for the operation during Run 5 and 6.

The central element of the physics programme will remain the flavour physics studies of CP Violation and rare decays, though the programme has become increasingly broad covering topics such as electroweak, QCD, dark sector searches and ion and fixed-target collisions. It is therefore beneficial to increase the recorded statistics while maintaining or improving the performance of the experiment to fully exploit the HL-LHC facility. [96]

Using the HL-LHC, the instantaneous luminosity will increase by nearly another order of magnitude to $\mathcal{L}_{\text{inst}} = 1.5 \cdot 10^{34}\,\text{cm}^{-2}\,\text{s}^{-1}$. By the end of Run 6 an estimate of the total integrated luminosity recorded by the LHCb is about $L = 300\,\text{fb}^{-1}$.

Another important change is the corresponding increase of the interactions per proton-proton bunch crossing. While the Upgrade I expects to have only an average of 5 interactions, the estimate for the mean number of interactions observed by the Upgrade II is 40. This poses a challenge for track reconstruction, namely with proper matching of the decays of heavy flavour hadrons to their primary vertex which is a key requirement of the LHCb physics programme. This is still planned with a full software trigger in mind.

Due to the previously mentioned reasons, most of the experiment will have to be upgraded in order to deliver the needed precision. A recurring theme for the upgrade is an addition of timing with precision $\mathcal{O}(20\,\text{ps})$, which provides the needed information used to reconstruct long-lived particles and correctly associate them to their origin primary vertex. The detector will also introduce the usage of radiation hard Depleted Monolithic Active Pixel Sensors (DMAPS) pixels. A brief description of the individual planned changes follows.

Tracking The main change for the VELO is the addition of the precise timing information and modification of the RF foil. The changes and approaches will be discussed in more detail in the following section.

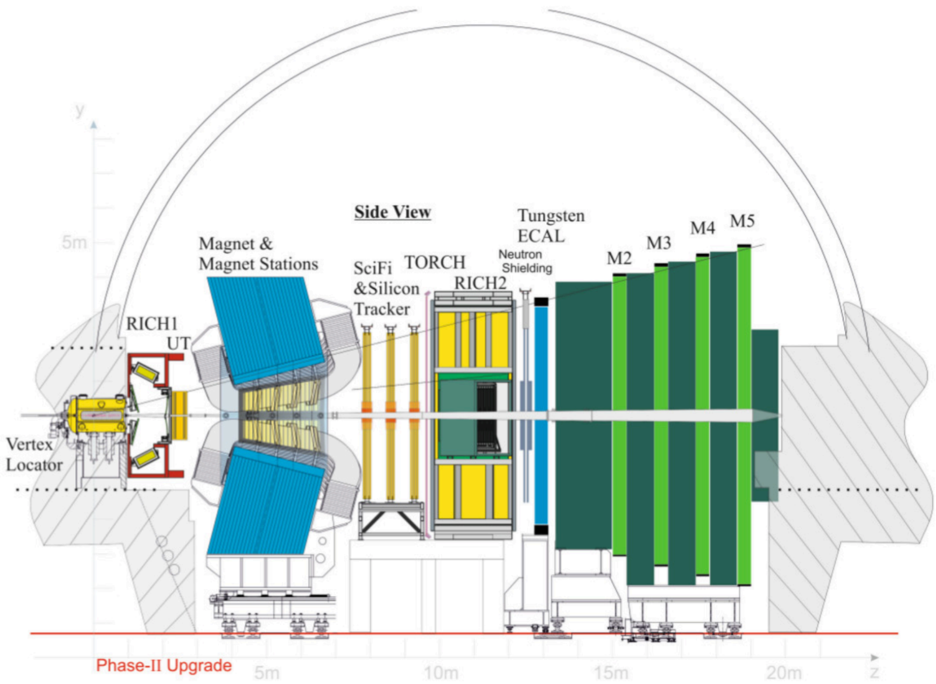


Figure 6.1: Schematic view of the LHCb Upgrade II detector. [95]

The rest of the tracking stations need an improvement of the granularity in order to cope with the increase of the tracks. The UT and the central part of the downstream tracking intend to use MAPS detectors. In the downstream tracking stations the outer region will still be covered with a SciFi, the whole sub-system called Mighty Tracker.

Since a number of low momentum particles ($p < 5\,\mathrm{GeV}/c$) get deflected towards the magnet, the information of their tracks is lost and worsens the reconstruction. A new scintillating fibre detector Magnet Station (MS) is proposed to be placed inside the magnet, tracking the particles interacting with its walls.

PID The RICH detectors will require photomultipliers with better spatial resolution and timing resolution of $\mathcal{O}(50\,\mathrm{ps})$. In order to identify low-momentum particles and thus provide benefits such as improving the b-hadron flavour tagging, a new system called Time of internally Reflected CHerenkov light (TORCH) is planned to be added as well. As the name hints, it utilises the time-of-flight of the Cherenkov light and provides a resolution of 15 ps per track.

The upgrade of the ECAL needs to fulfil a number of requirements given the demonstration of the strong physics programme with electrons, neutral pions and photons at LHCb. This requires higher granularity and $\mathcal{O}(10\,\mathrm{ps})$ timing resolution while meeting stringent radiation requirements in the inner region.

The muon system will be required to cope with increased rates and radiation damage, with a new system being developed for the highest fluence region.

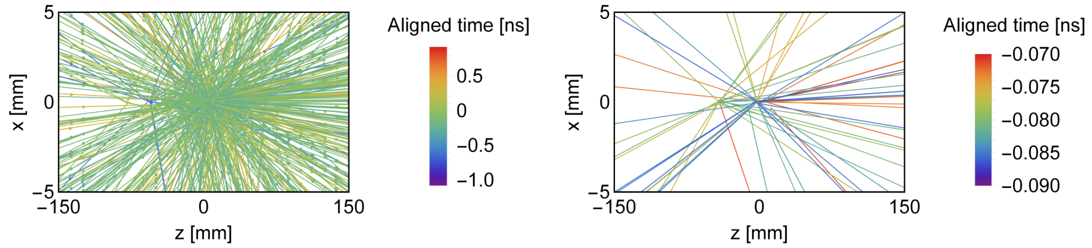


Figure 6.2: Visualisation of the interactions from a single bunch crossing in VELO (left plot). Using a time cut on the tracks of 20 ps (right plot) reduces the complexity of reconstructing the vertices. [95]

6.1.1 VELO Upgrade II

A more in depth description of the VELO Upgrade II is given, as the rest of the work follows discusses the ongoing R&D programme. There are many different aspects in designing a high performance system, ranging from new cooling techniques and data-readout up to designing new detector ASIC and sensors. [95]

Building on top of the VELO and VELO Upgrade designs, an RF foil is used in order to separate the beam and detector vacua, provide shielding against induced charge and minimising the wakefield excitations of the beam. This must be achieved at the smallest possible thickness in order to reduce scattering and improve the detector resolution. An extensive research programme is currently ongoing which is trying to solve the issues, potential candidates ranging from thinned down foils all the way to its complete removal.

Similarly, different options of the cooling are being investigated. This consists of researching coolants other than liquid CO_2 as well as substrates which provide mechanical support and heat transfer from the electronics. Ideally a lower material budget should be achieved while providing much lower temperatures.

Performed studies [95] have shown that the addition of timing information will be necessary at the luminosity of Upgrade II in order to properly reconstruct the primary and secondary vertices coming from the large amount of interactions. A simplified example is shown in Fig. 6.2 where the decrease of complexity of the reconstruction is clearly visible. The main goal is to provide a timing information for the tracks with a precision of $\mathcal{O}(20\,\mathrm{ps})$. This can be achieved either by adding a per hit timing information, or by adding fewer but more precise timing planes (see Fig. 6.3). The overall data rate increase will also require an updated read-out of the electronics.

Part of this work is dedicated to studying a specific type of silicon sensors with an internal gain. This has to be done with the radiation hardness as high as achievable, however, it is possible that the installed modules will have to be replaced after a few years of the operation.

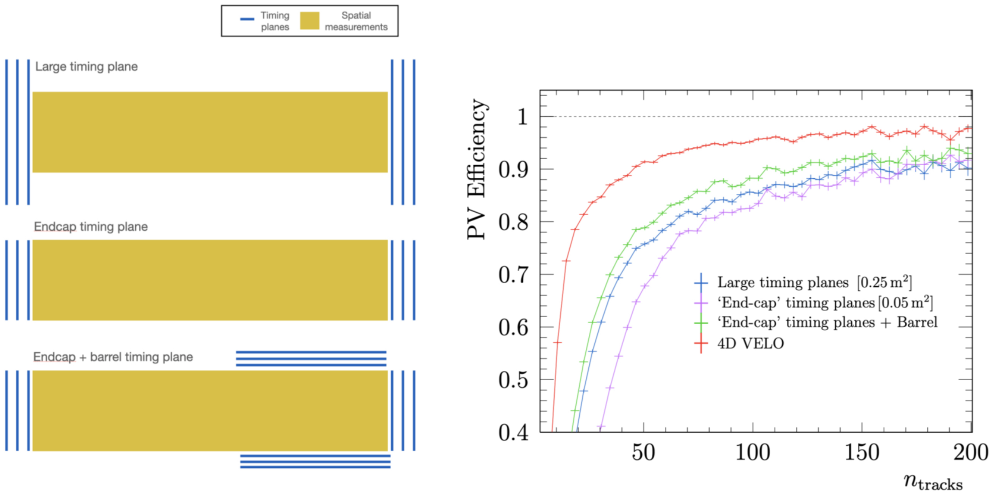


Figure 6.3: Schematic view of different types of the timing plane solutions (left) and corresponding performance of the primary vertex resolution (right), compared for a full 4D tracking VELO. [95]

6.2 Ultra fast silicon detectors

To achieve the exceptional spatial $\mathcal{O}(10\,\mu\text{m})$ and temporal $\mathcal{O}(10\,\text{ps})$ resolution required by the VELO Upgrade II, a large enough number of electron-hole pairs have to be created in every particle interaction and collected within a small time-window. This can be done by either having a very thin sensor with highly sensitive electronics, changing the topology of a sensor (using 3D technology [68]) or by multiplying number of generated charges (using multiplication layer [97]).

The gain G of these UFSD which rely on the charge multiplication can be simply defined as [98]

$$G = \frac{N_{\text{e,h}}}{N_{0;\text{e,h}}} \tag{6.1}$$

where $N_{\text{e,h}}$ represents the total number of the collected charges, compared to the number $N_{0;\text{e,h}}$ corresponding to a silicon without any multiplication layer. Based on the gain, the silicon devices may be split to three main groups:

Low-Gain Avalanche Diode (LGAD) [99] operating at low gain $\mathcal{O}(1 \cdot 10^1)$,

Avalanche Photodiodes (APD) [100] with gain of $\mathcal{O}(1 \cdot 10^2)$, close to a breakdown and working in a proportional regime,

Single Photon Avalanche Diode (SPAD) [101] achieving gain up to $\mathcal{O}(1 \cdot 10^5)$ while operating above breakdown in a Geiger mode.

The LGAD technology has been already demonstrated to give the required timing performance [102]. These devices are produced as $\mathcal{O}(1\,\text{mm})$ sized pads rather than desired pixels with pitch $\mathcal{O}(10\,\mu\text{m})$. Furthermore, the sensors built this way do not have the extreme radiation tolerance needed at the inner part of the VELO and their performance still has to be further researched.

6.2.1 Impact ionisation

The basis of the charge multiplication is in the continuous ionisation caused by the free charge carriers present in the sensor. Expanding upon the physics of semiconductors given in Sect. 2.2, for a strong enough electric field in the silicon ($\mathcal{E} = 300\,\mathrm{kV\,cm^{-1}}$ [103]), the free charge carriers accelerate and may produce a new e-h pair upon impact. The new set of carriers then starts accelerating once again, exponentially increasing the total amount of freely moving charges and resulting in an avalanche. The minimal required kinetic energy E_{kin} of the impeding particles to overcome the threshold energy E_{i} needed for the multiplication is [103]

$$E_{\mathrm{kin}} > E_{\mathrm{i}} = 1.5 E_{\mathrm{g}}, \tag{6.2}$$

where E_{g} represents the bandgap energy (discussed in Sect. 2.2). Due to the fact that the effective masses of electrons and holes vary, the actual thresholds for silicon are 3.6 eV and 5.0 eV, respectively. [59]

The dependency of the electron/hole concentrations $N_{\mathrm{e,h}}$ for a given depth d may be written as [98]

$$N_{\mathrm{e(h)}}(d) = N_{\mathrm{e(h)}}(0)\mathrm{e}^{\alpha_{\mathrm{n(p)}} d}, \tag{6.3}$$

where the ionisation rate $\alpha_{\mathrm{n(p)}}$ is defined as number of e-h pairs generated by the given carrier per unit of the distance travelled [103]. Its inverse corresponds to the mean free path between the interactions.

A number of approximations (Van Overstraeten-De Man [104], Okuto-Crowell [105] and Massey [106]) exist to describe the dependency of ionisation rate on the intensity of the electric field E, all based on [107]

$$\alpha_{\mathrm{n(p)}}(E) = A_{\mathrm{n(p)}}\mathrm{e}^{-\frac{B_{\mathrm{n(p)}}}{E}}. \tag{6.4}$$

The coefficients $A_{\mathrm{n(p)}}$ and $B_{\mathrm{n(p)}}$ are defined individually for the charge carriers and include dependency on temperature.

For the purpose of a better description of the multiplication, the rate of the generation and recombination is often followed. In general, when the device is not in an equilibrium a net rate $U_{\mathrm{n(p)}}$ is present, defined as

$$U_{\mathrm{n(p)}} = R_{\mathrm{n(p)}} - G_{\mathrm{n(p)}}, \tag{6.5}$$

where $G_{\mathrm{n(p)}}$ and $R_{\mathrm{n(p)}}$ correspond to the generation and recombination rates, respectively. The process results in the system restoring the initial equilibrium. In the indirect semiconductors, this rate is best described using Shockley-Read-Hall model [108] where the defects and impurities affect the recombination.

For the avalanche, the generation dominates and the net rate is [59]

$$U = \alpha_{\mathrm{n}} n \nu_{\mathrm{n}} + \alpha_{\mathrm{p}} p \nu_{\mathrm{p}}, \tag{6.6}$$

where $\nu_{\mathrm{n(p)}}$ is the drift velocity of the carriers. This is often implemented in the simulations of the multiplication region [98].

6.2.2 Timing performance

Since one of the main goals of the UFSD is to achieve an excellent timing resolution, the various processes affecting the performance have to be considered. The time resolution can be specified as [97]

$$\sigma_{\mathrm{t}}^2 = \sigma_{\mathrm{Time\text{-}walk}}^2 + \sigma_{\mathrm{Landau}}^2 + \sigma_{\mathrm{Distortion}}^2 + \sigma_{\mathrm{Jitter}}^2 + \sigma_{\mathrm{TDC}}^2, \tag{6.7}$$

where the individual components are discussed below.

$\sigma_{\mathrm{Time\text{-}walk}}$ The time-walk is an effect where a larger signal crosses the predefined threshold level sooner than a smaller signal. This can often be corrected for either by using a different discrimination technique or during the post-processing (in the case when the amplitude of the signal is measured). In order to minimize the effect for the MIP interactions, the multiplication should be uniform and create as similar signal as possible regardless of the particle track.

σ_{Landau} Since the generation of the free charge carriers is a statistical process, the energy deposition results in a signal variation. This is an inherent effect which can be only mitigated by using thinner sensors or longer signal integration times.

$\sigma_{\mathrm{Distortion}}$ The distortion is mainly caused by the weighting field and drift velocity of the carriers which depends on the electric field and thus may not be uniform. While the effect of the latter can be minimized by having a large enough bias voltage to reach the velocity saturation, the weighting field depends mostly on the size of the channel implant compared to the thickness of the sensor. Similarly to the time-walk, any non-uniformity of the multiplication region will have a strong effect on the total amount of the signal registered.

σ_{Jitter} Jitter is the effect of noise, which is always present, causing the signal to be registered earlier/later than the ideal case. This can be minimized by lowering the noise (e.g. by reducing the temperature or current) or increasing the slope of the signal leading edge. However, a balance of the two has to be found according to the desired application as amplifiers with a high resolution needed for the slope also have higher noise levels. [97]

σ_{TDC} The Time-to-Digital Converter (TDC) component represents the precision of the timing circuit used to measure the threshold crossing. The resolution is simply defined as $\Delta T/\sqrt{12}$, for a bin width ΔT that is defined based on the readout circuitry.

The mentioned effects result in the variation of the timing precision of the point in time when the signal crossed the predefined threshold value and the accuracy with which this is measured. Because of this, the UFSD should have a quick signal rise time along with a good signal-to-noise ratio. Furthermore, it is important to have a uniform response from the whole sensor volume as otherwise the threshold crossing point could depend on the interaction position. For the best results, a matching amplifier circuitry has to be selected [97].

6.2.3 Fill Factor

The next generation silicon detectors also need to achieve a very good spatial resolution, normally achieved by segmenting the sensor into small pixels. While this method is accessible and reliable (pixels with pitch 55 µm are used in the VELO Upgrade modules), the addition of the multiplication region has its own challenges.

Based on the technology used and separation between the individual channels, the gain layer may result in large non-uniformities for the individual channels. This can be characterised by a fill factor, defined as

$$\text{Fill Factor} = \frac{S_{\text{gain}}}{S_{\text{total}}}, \tag{6.8}$$

where S_{total} and S_{gain} represent the total area and the area with multiplication, respectively. In an ideal case, the fill factor should be equal to 1. Achieving a high fill factor is a particular concern for the small pitch devices required for the VELO.

In order to prevent a breakdown, the designs have to have a proper termination of the high intensity field in the multiplication region. A commonly used method for the p-type sensors is implanting an n-type Junction Termination (JTE) around the border of the multiplication region, however, this impacts the fill factor of the small pitch devices. Because of this, for a defined bulk substrate (considered a p-type due to better long-term performance in a high radiation fields), a number of options exist how to construct a low-gain multiplication region [109]:

LGAD , standard approach with pixels terminated using JTE, collecting electrons;

inverse LGAD (iLGAD) multiplication at the back side, collecting holes;

Trench Isolated LGAD (TI-LGAD) pixels terminated using trenches;

AC coupled LGAD (AC-LGAD) resistive AC-coupling of the insulated metal pads to the gain side.

The comparison is visualised in figure 6.4. The iLGADs need to have the JTE separation only around the outer sensor edges and the rest of the work in this chapter will be focused on them.

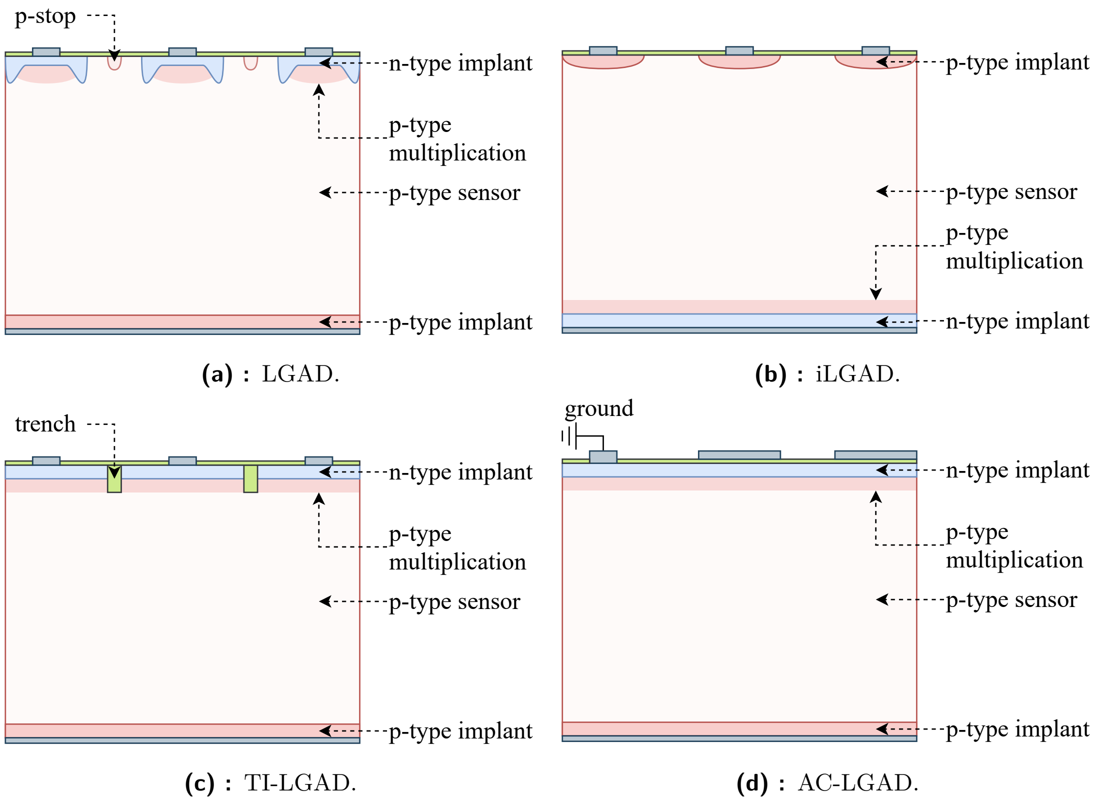


Figure 6.4: Comparison of LGAD types of devices, visualised as 100 µm thick sensor with pixel pitch 50 µm.

6.2.4 Gain

As was mentioned before, the gain characterises the increase of signal as compared to the same device without any charge multiplication. In order to achieve a high breakdown voltage along with the desired low gain, the doping profiles have to be closely controlled.

Since the multiplication depends on the intensity of the electric field (starts at about $\mathcal{O}(1 \cdot 10^5\,\mathrm{V\,cm^{-1}})$), it presents the main driving factor of the maximal device gain. Such values are achieved by the doping concentrations $\mathcal{O}(1 \cdot 10^{19}\,\mathrm{cm^{-3}})$ for the n-type electrode and $\mathcal{O}(1 \cdot 10^{16}\,\mathrm{cm^{-3}})$ for the multiplication region. The depth of the profiles is about 1 µm and 4 µm, respectively. [110].

A rough estimation of the gain may be done by following the Q-effective layer which corresponds to the net area of the multiplication region. This is computed as a difference of the integrals of doping concentration of the p- and n-type implants. The visualisation of this layer is in figure 6.5, with a relation between Q-effective and gain level in figure 6.6.

It is clear that the increase of the gain is achieved by a higher Q-effective, caused by the high doping levels of the multiplication layer. Because of the higher intensity of the field, this in turn lowers the breakdown voltage of the sensor. Both of these have to considered during the design of the device, along with the actual technological constraints of the production.

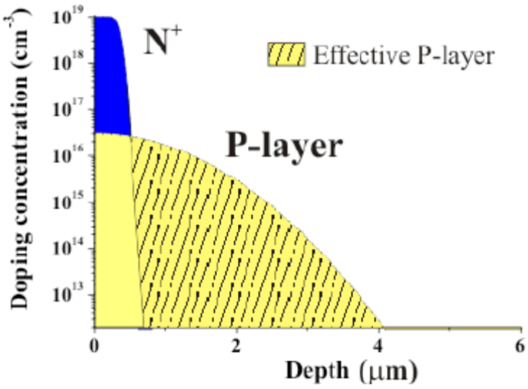


Figure 6.5: Q-effective representation from doping profile of LGAD. [111]

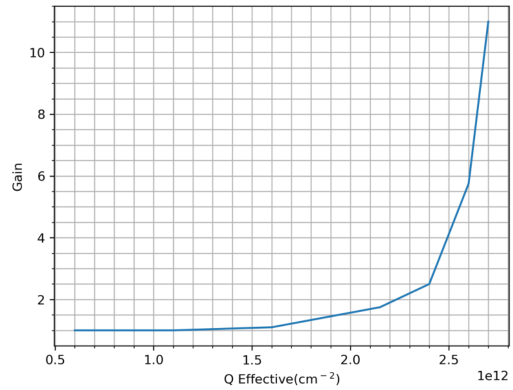


Figure 6.6: Simulation results of dependency of LGAD gain on Q-effective for 400 V reverse bias. [110]

6.3 Sensor simulations

Simulations are frequently utilised to better understand the performance and optimize the sensor design without the need to do a costly iterative process in a foundry. For this development a Technology Computer-Aided Design (TCAD) toolkit is used, simulating the doping profiles, calculating different electronic characteristics and even simulating charge injection and subsequent induced signal.

Since the TCAD contains only a limited amount of methods describing interaction of ionising radiation, other tools are often used for the purpose of detector development. One such widely used example is SRIM [112] as it can be quickly used to obtain energy depositions in complex material structures.

6.3.1 Sentaurus TCAD

Sentaurus [113] is a suite of TCAD tools developed by Synopsys, used for simulating fabrication and operation of the semiconductor based devices. The toolkit includes a number of GUI programs that simplify the definition and execution of the desired process.

The simulations are based on solving finite-element problems and can be performed in 2D or even 3D, each dimension giving an extra complexity and increasing the computing time and memory requirements. The results will slightly change when adding a dimension, however, smaller complexity gives a good starting point along with a quicker turnaround due to a much shorter time.

The two tools from the whole suite mainly used in this work are the structure editor – used to generate analytically defined doping profiles along with the finite elements (forming meshes) which describe the sensor; and sdevice – used for the electrical performance simulations.

structure editor The structure editor is used to initialise the device based on the desired

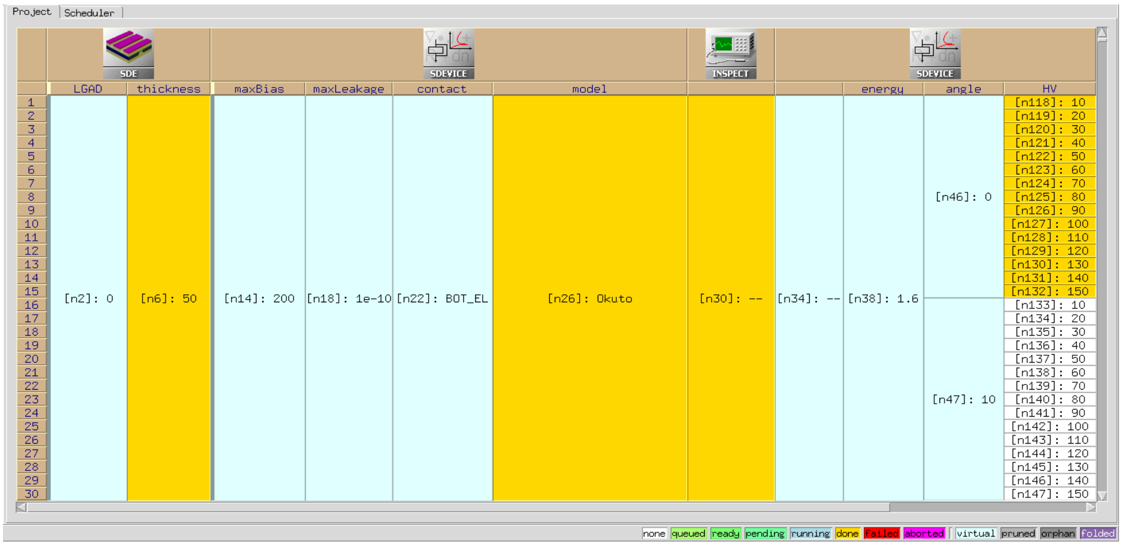


Figure 6.7: Sentaurus workbench used to control the Sentaurus TCAD simulation.

material compositions. While a much more precise approach would be simulating the actual processes as performed in the silicon foundry, this would require simulations of every single sensor modification and is computationally intensive. This is often not necessary as the profiles may be approximated using analytical functions.

The editor may be controlled either using a GUI or by a scripting language. In both cases, the software is based on a set of functions used to define the doping concentrations and distributions. Using the scripting approach, the values may be modified using global variables and complex algorithms with loops and conditions.

Finally, since the whole simulation is based on finite element methods, the whole structure has to be meshed. This is done by defining the maximal distance between the vertices as well as variations of the the profiles to be meshed.

sdevice Using the mesh result of the previous step, the device simulation applies the selected physics model on the device. This ranges from simulating the quasi-stationary conditions (IV or Capacitance-Voltage (CV) curves) all the way to a transient current evolution from a defined energy deposition.

For the simulation of LGADs, the model characterising the impact ionisation has to be selected (all models described in Sect. 6.2.1 are accessible). All of the physical constants may be modified in a parametric file.

The control of the whole simulation is done from the Sentaurus workbench which provides a way to monitor the simulation progress and define parameters accessed by the tools. The workbench is shown in Fig. 6.7, the individual cells correspond to nodes, representing an unique set of parameters for the simulation. These nodes are separately executed and contain all of the results. For a quick visualisation without any need to export the data, Sentaurus also contains an svisual tool which can plot 1D and 2D dependencies of the results.

6.3.2 SRIM

Stopping and Range of Ions in Matter (SRIM) [112] is a software package that provides tools to simulate the passage of particles (namely ions but neutron and electrons are included as well) in a medium. The interaction volume is defined using infinite parallel planes of specified material composition and corresponding densities. This helps to create complex silicon sensor structures which include oxide and metal layers.

The simulated interactions are of two types – ion-atom or ion-electron. While the latter is used only to define the energy loss and mean free path of the impeding ion, the interaction with the atom lattice may result in the change of the direction. The alternation of these two processes results in the final trajectory of the particle and its full energy deposition profile.

The data obtained from the simulation may be imported into the TCAD in a form of energy deposition profile, providing a simple way to simulate the response of the sensor for any particle.

6.4 iLGAD simulation

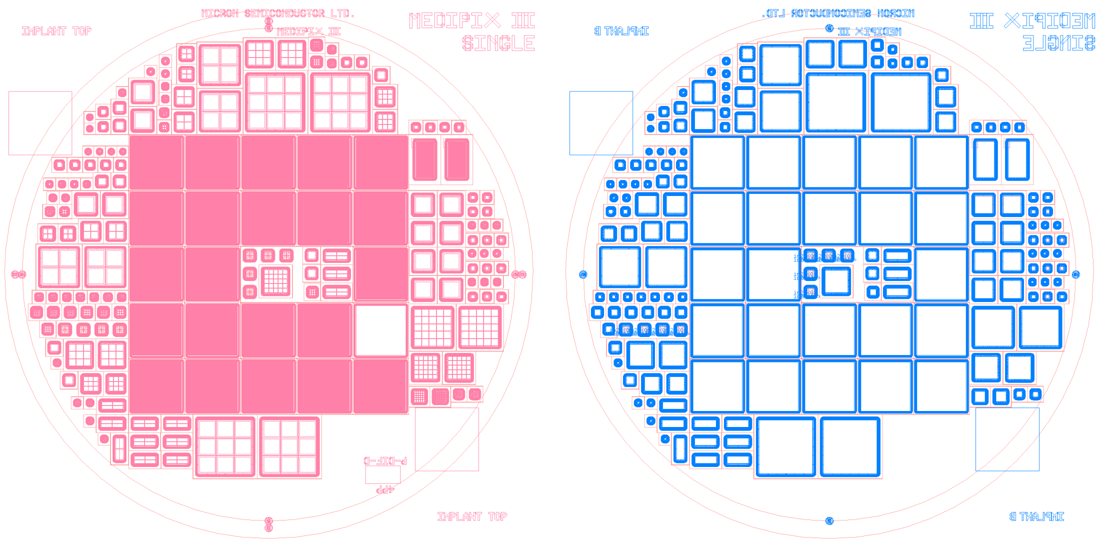
As part of the R&D for the future VELO upgrade, a collaboration between the UoM, the University of Glasgow (UoG), the University of Edinburgh (UoE) and Micron has been formed, developing and testing new sensors that could be eventually used. The simulation of the iLGAD devices was therefore made for the purpose of relating the results of a newly produced batch of sensors.

An example of the mask for the two doping layers used for the device production is in Fig. 6.8. The devices range from simple diodes of varying sizes ($0.04\,\text{mm}^2$ to $25\,\text{mm}^2$) to Medipix sized pixelated sensors with pitch 55 µm, 110 µm and 220 µm. At the time of writing the device testing and analysis is still ongoing, therefore the presented results do not contain any comparisons between the simulation and measurement.

6.4.1 Doping profile

Since the doping profile drives the performance of any gain-focused silicon device it is important to match the simulated profiles as close to the real ones as possible. This usually requires a detailed simulation of the implantation process for the specified topology where any change requires lengthy recalculation. For this purpose, an analytic model based on the data from the sensors previously produced by the UoG and Micron has been developed, using the provided Secondary Ion Mass Spectroscopy (SIMS) measurement [110]. The advantage of such an approach is its easy implementation into TCAD and a simple modification of the maximal values in order to modify the gain.

The match was obtained by fitting the results of the doping profiles with an arbitrary analytical function. The results are in Fig. 6.9, n-implant (phosphorus) fitted using three Gaussians, while p-multiplication (boron) by five Gaussian functions. The p-type bulk



(a) : Pixel side – boron.

(b) : Backside – phosphorus.

Figure 6.8: Implant masks used for the implantation in the recent iLGAD sensors run, produced by collaboration between the UoG, the UoM and Micron.

doping used in the simulations is $3 \cdot 10^{12}\,\mathrm{cm}^{-3}$, the p-type implant matching a simple Gaussian profile with the peak concentration $1 \cdot 10^{20}\,\mathrm{cm}^{-3}$ and $\sigma_{\mathrm{p}} = 3\,\mu\mathrm{m}$.

Even though the representation does not perfectly match the previously produced sensor, the overall shape should provide a good approximation for the new production run of the iLGADs, the Q-effective layer being $\approx 1 \cdot 10^{12}\,\mathrm{cm}^{-2}$. The main advantage of this approach is the possibility to quickly simulate different topologies of the devices.

The profiles were then defined in the TCAD structure editor, the simulation can be used to model a simple planar device or a strip one with implants covering the top part of the device. A set of externally accessible variables used to modify the devices is available, these include strip pitch, implant width, device thickness and avalanche parametrisation model. Finally, an option to specify a type of device (no gain, multiplication on top/bottom) has been programmed in as well.

As the newly produced devices match the pixel pitch of the Medipix-family sensors ($55\,\mu\mathrm{m}$), the width of the simulated sensor has been fixed to $440\,\mu\mathrm{m}$. The symmetry of the device is achieved by placing the mid point of one of the strips at the (0,0) location and automatically placing the rest of the strips around until they fill the device.

Since the whole simulation is a finite-element one, the physical quantities have to be assigned to a mesh. The meshing is automatically done by the toolkit, with higher-density regions enforced on both sides near the electrodes. An example of the meshed device is in Fig. 6.10. In order to quickly verify the modelled multiplication layer the profile of the electric field can be checked, shown in Fig. 6.11.

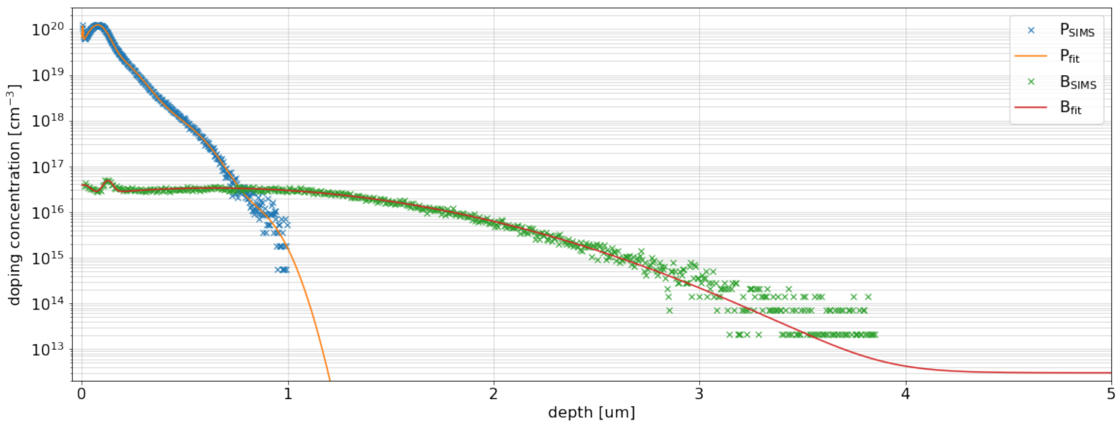


Figure 6.9: Approximation of the process simulation data of n-implant (phosphorus) and p-multiplication (boron). The data points compare the result of the SIMS process of an older sensor production run and the approximated analytical fits.

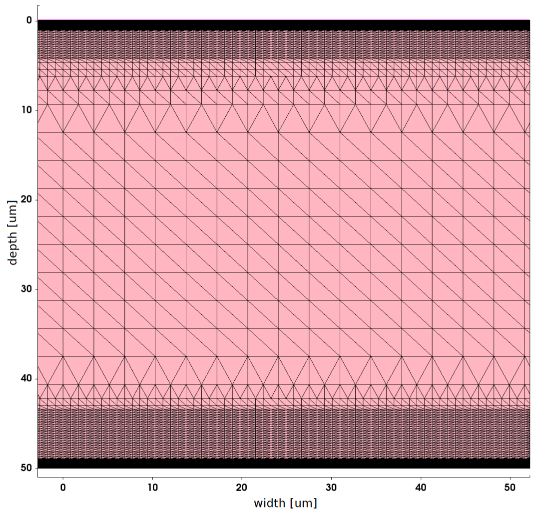


Figure 6.10: Meshing map of the simulated diode with thickness 50 µm. The meshing is more refined in the implant regions in order to better simulate the avalanche process in the multiplication layer.

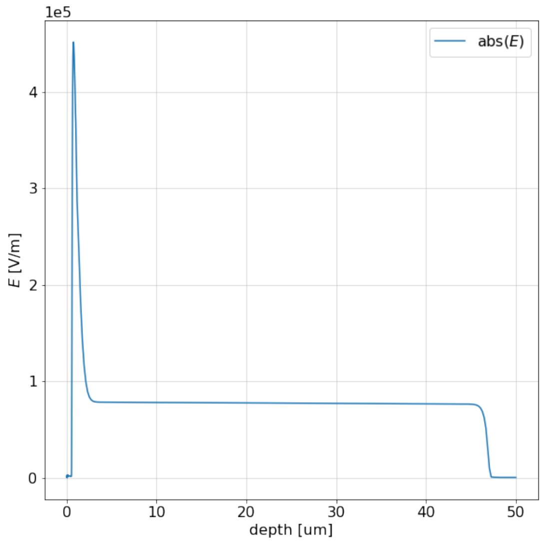


Figure 6.11: Absolute electric field profile of the simulated iLGAD with thickness 50 µm. A high field region is visible in the multiplication layer in the first few µm.

6.4.2 IV curves

The simulations were implemented for no gain and iLGAD devices, varying the thickness of the sensor. This was done both for planar sensors as well as segmented ones, the results are given in Fig. 6.12.

The IV curves of the simulated devices have the expected performance – with the increasing thickness, a larger value of the breakdown voltage is reached. Simple diodes without any gain are quickly depleted, breakdown occurring for a much higher values than for any diodes with internal gain (caused by the large electric field in the multiplication layer).

The shape of the IV curves for iLGAD contains an increase of the leakage current at

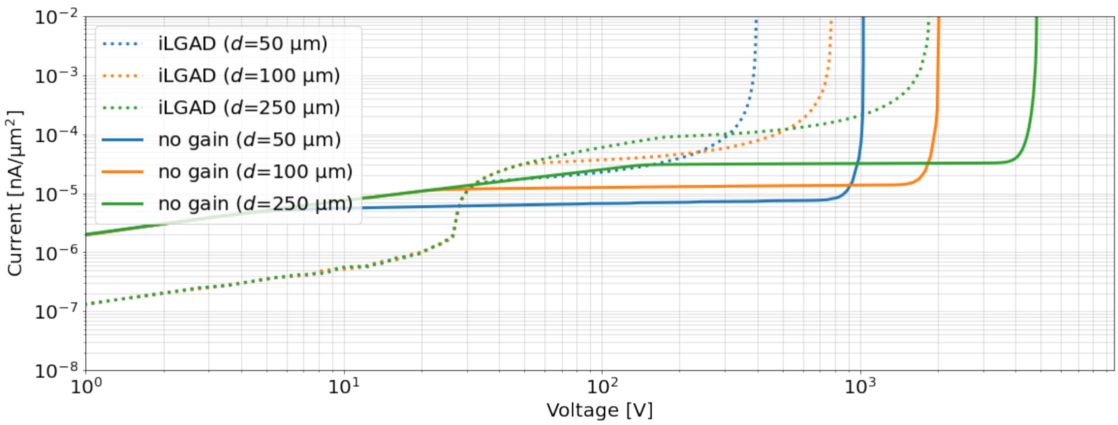


Figure 6.12: IV curve of no gain (solid lines) and iLGAD (dotted lines) devices. The dependency is shown for different thicknesses from 50 µm to 250 µm, thinner diodes have smaller breakdown voltage value.

around 30 V. This is caused by first depleting the multiplication region, later followed by the depletion of the bulk. A better insight into the precise depletion depths would be given by CV curves.

6.4.3 Charge collection

Since the main idea behind the sensors with the built-in multiplication is to provide a better signal to noise ratio, the formation and collection of the induced current is simulated to provide a better insight into the device performance. This is done via a transient simulation where the energy deposition profile is defined and the TCAD framework calculates the drift and diffusion of the e-h pairs.

The segmentation of the simulated devices was based on a newly produced iLGAD samples made by Micron, their thickness being 50 µm, 250 µm. Even though the production run was focused on pixelated sensors, the simulation was done only in 2D, representing a strip configuration. The three implemented pixel pitches are 55 µm, 110 µm and 220 µm, their corresponding implant widths 36 µm, 71 µm and 181 µm, respectively. As was mentioned before, the width of the simulated region is fixed to 440 µm with strips centred around 0.

A comparison of the time evolution of the 2D current profiles for devices without and with gain is in Fig. 6.13 and Fig. 6.14, respectively. The simulation was done for an injected charge 10 keV placed 10 µm below the surface. Both devices were 250 µm thick, the gain layer of the iLGAD located at the bottom part.

It is clear that at first the current in both cases flows from the top to the bottom of the device, representing the flow of the electrons from the injected charge. This takes about 4 ns at which point the electrons arrive at the multiplication layer in the iLGAD. The produced holes start flowing in the opposite direction, taking another ≈ 6 ns to arrive at the collection electrodes at the top. The time evolution profiles will be discussed in more depth in the following text.

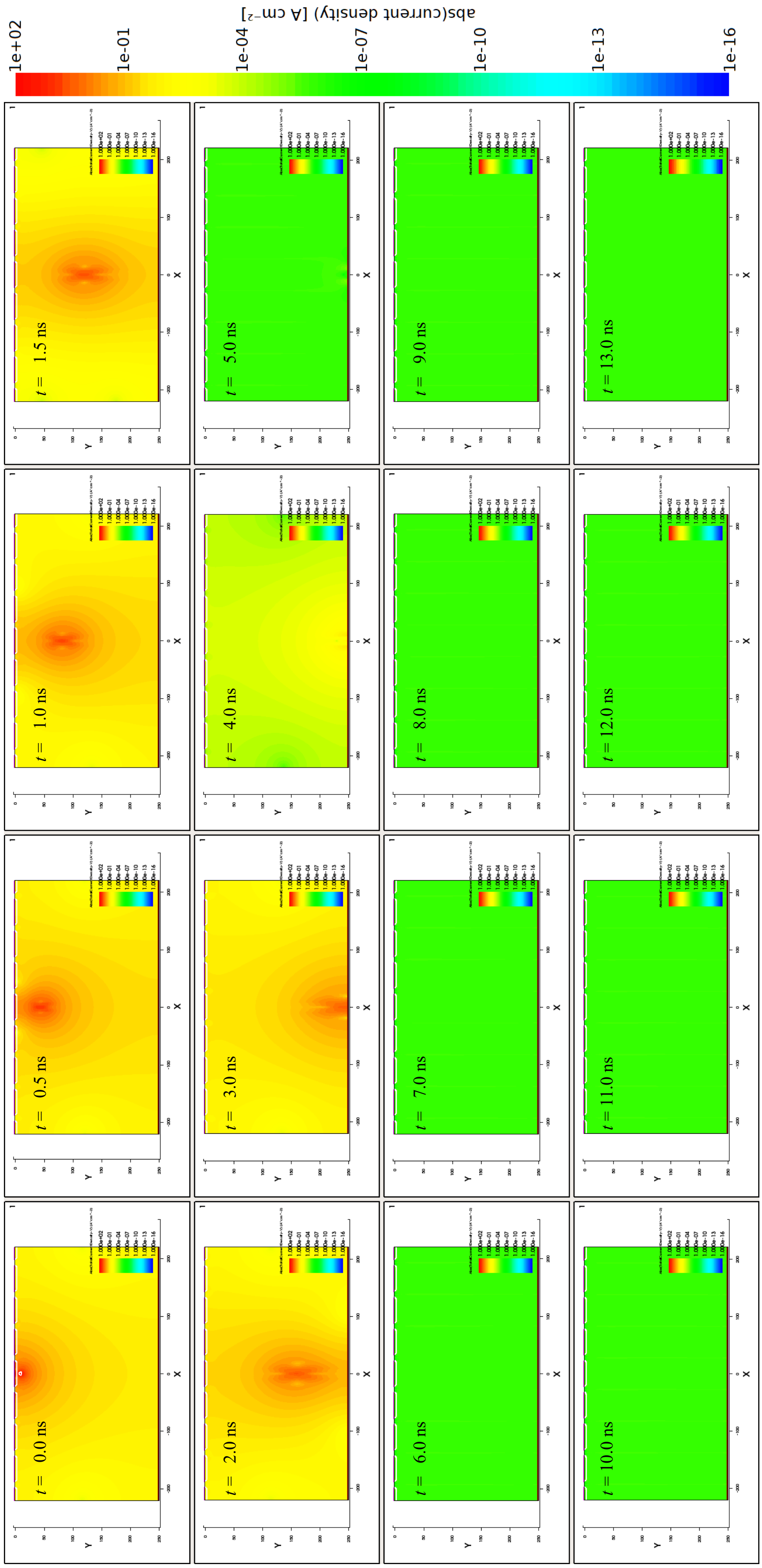


Figure 6.13: 2D current profiles of 250 µm thick sensor without gain, biased to 400 V. The injected charge is 10 keV.

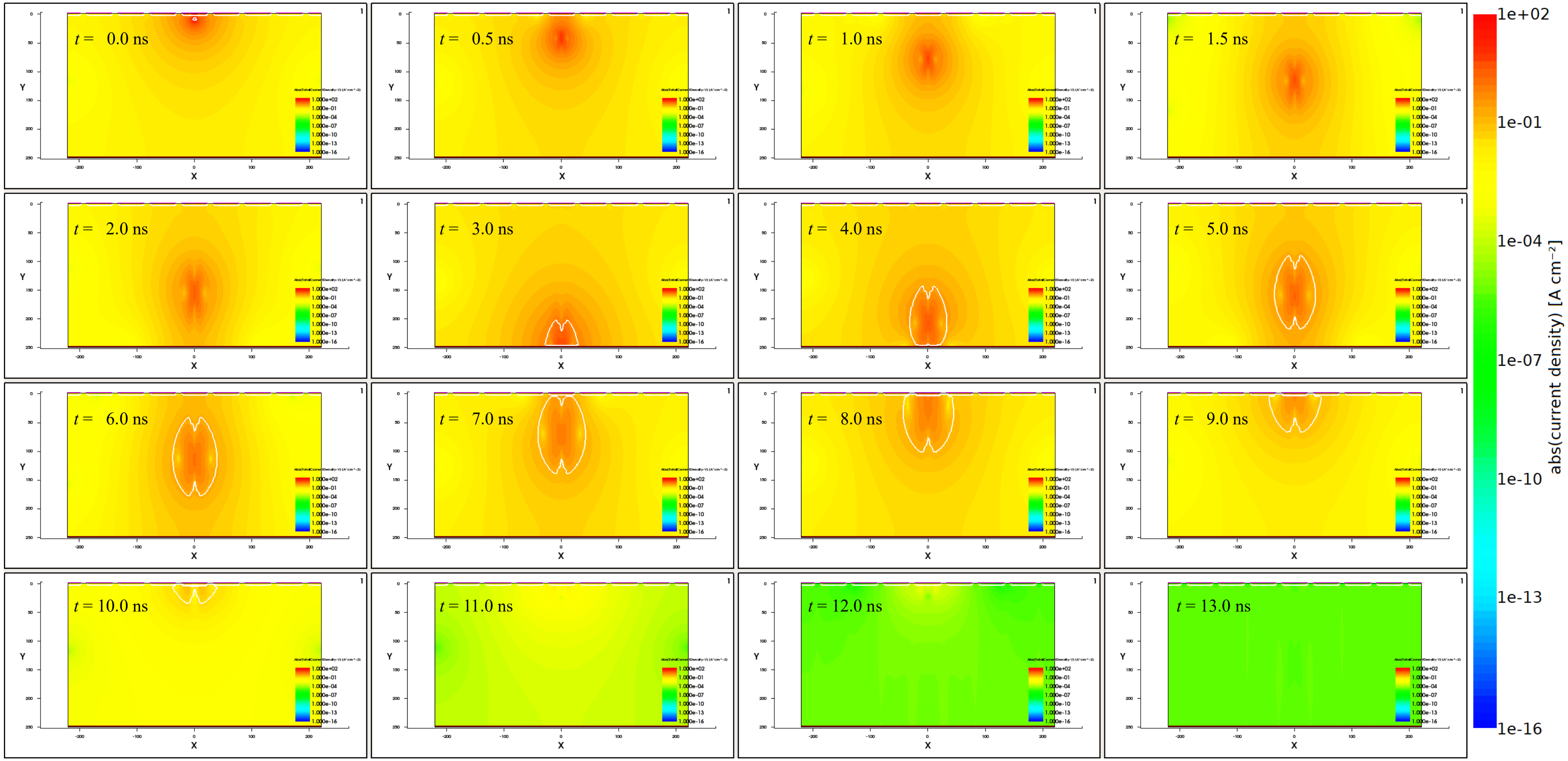


Figure 6.14: 2D current profiles of 250 μm thick iLGAD sensor (multiplication layer at the bottom), biased to 400 V. The injected charge is 10 keV. Multiplication of the electrons occurs at around 4 ns.

6.4.3.1 Thickness 250 µm

The more detailed view of the signal collected by the central electrode of a 250 µm thick sensor biased to 400 V (representing a flat region in the IV curve shown in Fig. 6.12) is in Fig. 6.15. The simulation was done for all pitches and four different energies, all injected at the same position $(x, y) = (0, 10)$ µm. The multiplication is visible at around the 3 ns mark, where the current registered by the iLGAD devices rises, whereas the devices without any gain already collected all of the charge.

The iLGAD current profiles may be split into three regions [114] – a quick rise time caused by electrons at the beginning (shared by both iLGAD and no gain diodes); a slower rise time from the onset of the hole multiplication; and a quick rise time from the hole collection. The readout may be tuned to discriminate on each of these regions achieving different timing precision [114], the importance of the low-gain devices is in an improved signal-to-noise ratio.

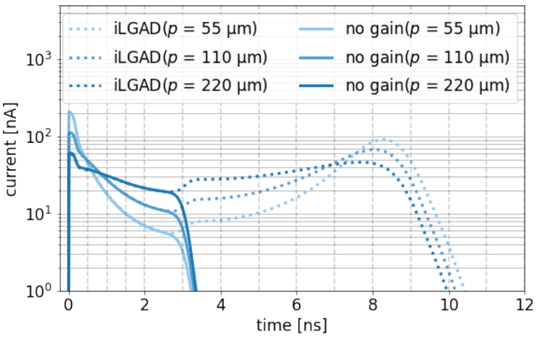
Integrating the current over the time domain provides information about the total charge Q collected by the electrodes. This is used to calculate the gain factor, defined as a ratio of the charge collected by the iLGAD and no gain devices. The results for the central electrode are plotted as a function of the injected charge in Fig. 6.16. While the devices with a larger pitch (comparable to the device thickness) have a constant gain of about 3.5, a decrease of both the collected charge and gain is visible for the iLGAD device with pitch 55 µm. This inconsistency is expected due to the charge sharing between neighbouring channels.

Charge sharing In order to investigate the charge sharing of the induced signal, the response of all the channels has been simulated for the 20 keV charge injection. Since the strips are symmetrically placed around zero, the current is summed for the ones with the same distance from the central strip. The results are plotted for all channels for the 55 µm and 110 µm pitches in Fig. 6.17.

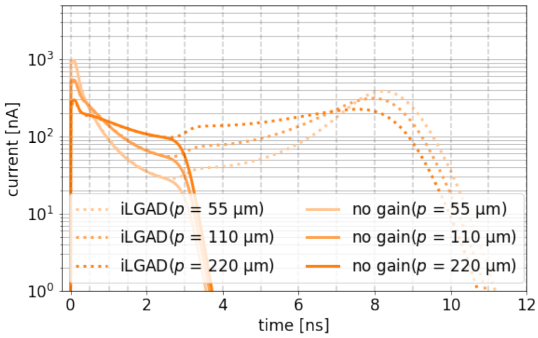
It is visible that the summed signal matches in the two cases, the charge sharing being a more dominant effect in the device with smaller pixel pitch. This is important as it presents a new constraint on the usage of small pitch devices with internal gain – the charge sharing has to be considered in the readout electronics, potentially increasing the data throughput or requiring a more complex internal logic of the readout ASIC. However, the effect could be minimised with thinner devices, where the pitch is comparable to the sensor thickness.

6.4.3.2 Thickness 50 µm

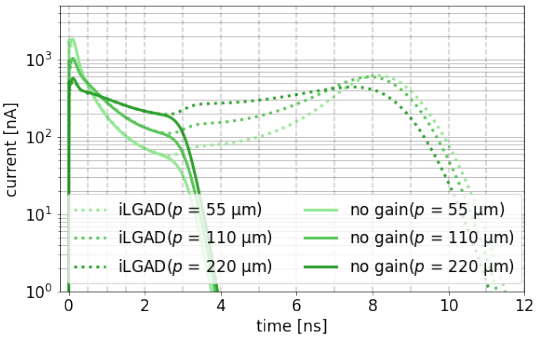
Since the internal gain provides an excess of signal compared to a standard device, it is possible to use thinner devices while keeping the signal to noise ratio high. This has been implemented in the simulation with a changed thickness of sensors to 50 µm. The devices were biased only to 200 V as the breakdown is much lower compared to the thicker ones.



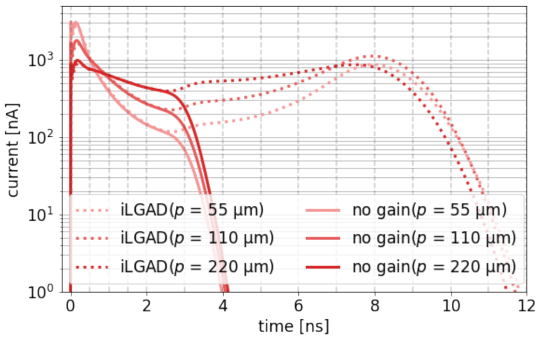
(a) : $E_{\mathrm{kin}} = 1\,\mathrm{keV}$.



(b) : $E_{\mathrm{kin}} = 5\,\mathrm{keV}$

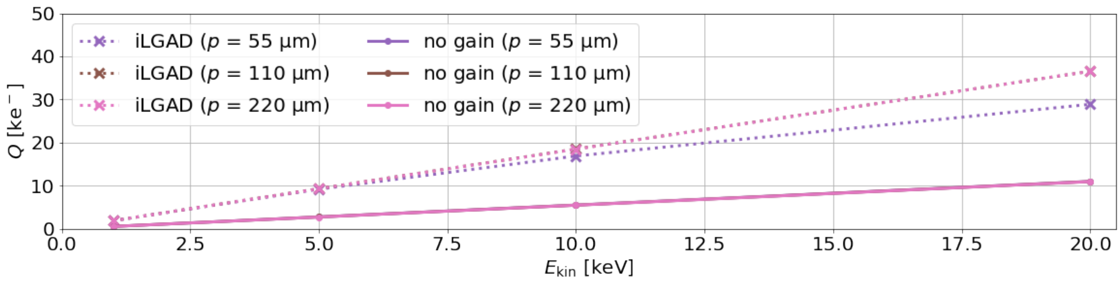


(c) : $E_{\mathrm{kin}} = 10\,\mathrm{keV}$

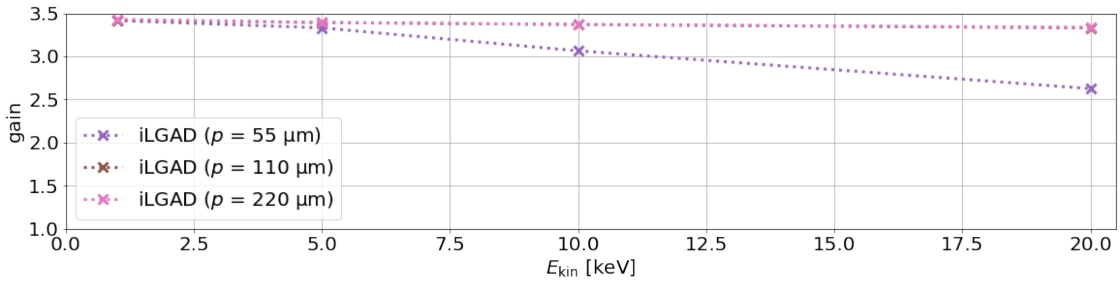


(d) : $E_{\mathrm{kin}} = 20\,\mathrm{keV}$

Figure 6.15: Current profiles registered by the central electrode of 250 µm thick sensors. The vertical lines correspond to the 2D profiles shown in the Figs. 6.13 and 6.14.

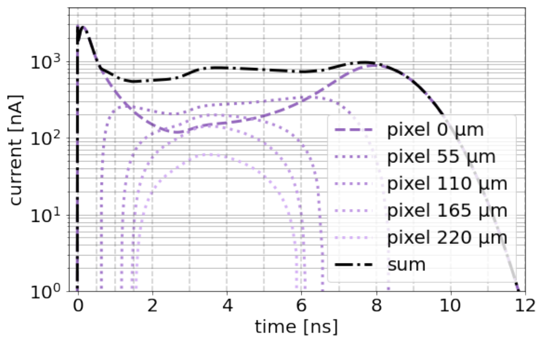


(a) : Collected charge.

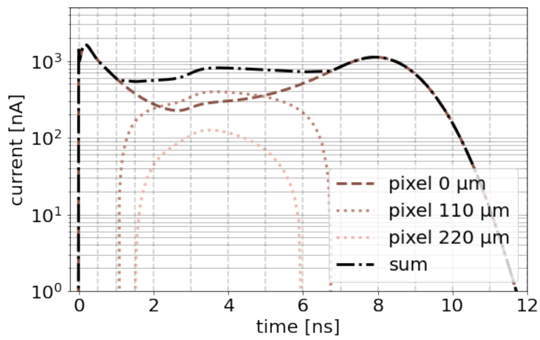


(b) : Gain calculated as a ratio of the integrated collected charge in iLGAD and no gain sensors.

Figure 6.16: Collected charge and gain as a function of the injected charge for the 250 µm thick sensors. The no gain data overlap in all cases while for the iLGAD only the $p = 110\,\mu\mathrm{m}$ and $p = 220\,\mu\mathrm{m}$ do.



(a) : Pitch $p = 55\,\mu\text{m}$.



(b) : Pitch $p = 110\,\mu\text{m}$.

Figure 6.17: Induced current of the neighbouring channels for a different distance from the central point. Because of the symmetry around zero, the channel response is summed.

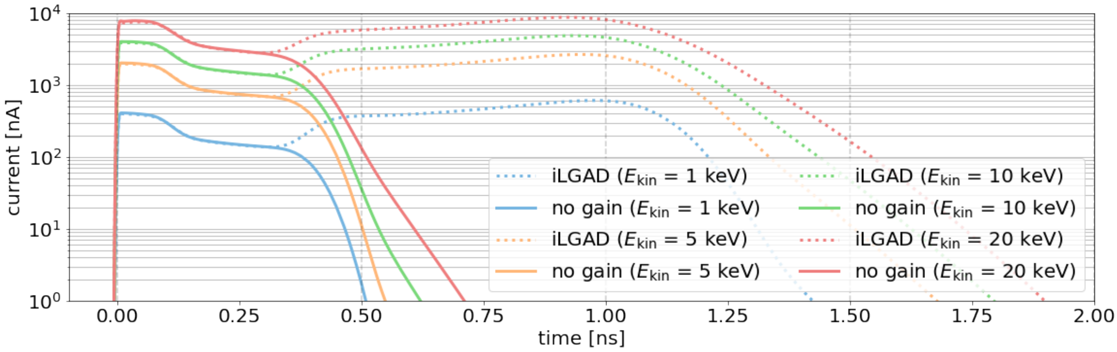
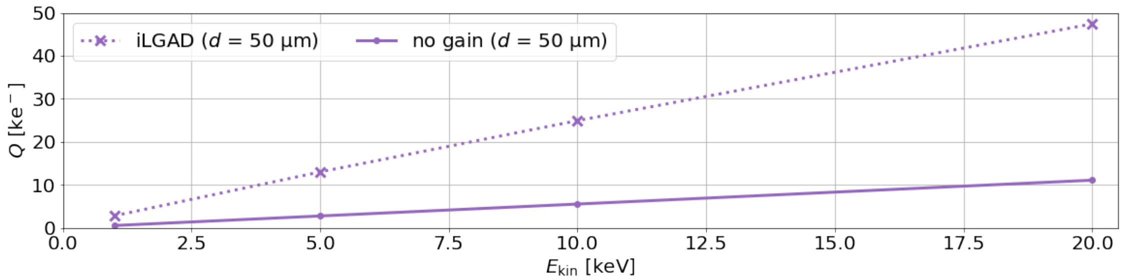
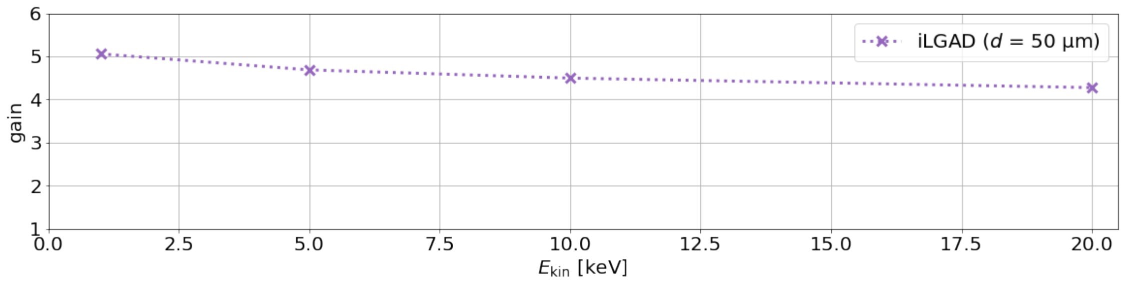


Figure 6.18: Current profiles registered by the central electrode of $50\,\mu\text{m}$ thick sensors.



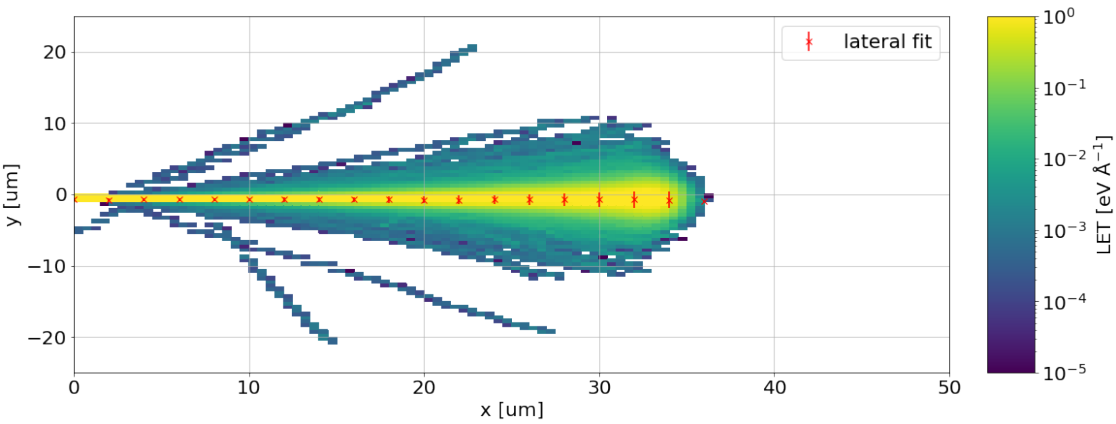
(a) : Collected charge.



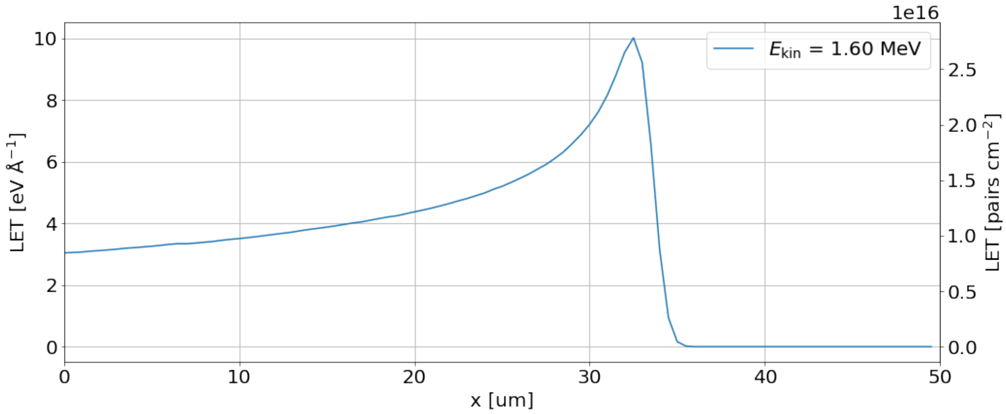
(b) : Gain calculated as a ratio of the integrated collected charge in iLGAD and no gain sensors.

Figure 6.19: Collected charge and gain as a function of the injected charge for the $50\,\mu\text{m}$ thick sensors.

Similarly to the previous simulations, the same charge injection has been simulated, the results for a pitch $55\,\mu\text{m}$ are shown in Fig. 6.18.



(a) : Heatmap projection of the LET into the x-y plane. The red lines represent the $(\mu \pm \sigma)$ of the Gaussian fits in the lateral direction.



(b) : Dependency of the LET on the distance travelled. Right axis is converted to the format imported into the TCAD.

Figure 6.20: LET of 1.6 MeV in silicon, simulated in SRIM.

The results show a much higher current values achieved over a shorter time duration. While the 250 µm thick sensor had the signal from the multiplied charge occur at 3 ns, for the thin devices this current already starts to dominate at around 0.75 ns. However, as was mentioned in the Sect. 6.2.2, the actual precision of the measurement depends on the readout circuitry.

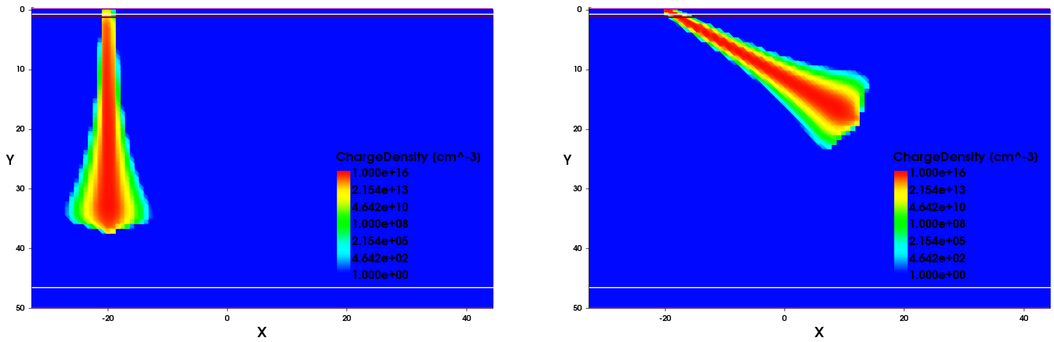
Looking in more detail at the collected charge and gain factors (see Fig. 6.19), the gain of the iLGADs is again decreased based on the injected charge, however, the effect is much smaller compared to the thicker sensors. The results of the thin sensors could be affected by the proportionally larger reverse bias voltage compared to the thicker sensors, however, the performance of the thinner devices is still expected to be better due to a shorter charge travel distance and an improved weighting field potential.

6.4.4 Test beam simulation

In order to verify the performance of the samples produced, a test beam measurement has been planned and performed in RBI Zagreb, using proton micro beam. While the actual data analysis is not part of this work, the executed scans were simulated for a

Table 6.1: Parameters defining pair creation from 1.6 MeV proton in silicon.

x [cm]	LET [pairs/cm^2]	σ [cm]	x [cm]	LET [pairs/cm^2]	σ [cm]
1.00e-04	2.95e+16	1.14e-05	2.10e-03	1.00e+16	4.84e-05
3.00e-04	1.99e+16	1.73e-05	2.30e-03	9.15e+15	5.63e-05
5.00e-04	1.94e+16	1.83e-05	2.50e-03	8.58e+15	6.43e-05
7.00e-04	1.74e+16	2.12e-05	2.70e-03	8.00e+15	7.55e-05
9.00e-04	1.62e+16	2.32e-05	2.90e-03	7.77e+15	8.70e-05
1.10e-03	1.54e+16	2.53e-05	3.10e-03	7.95e+15	1.01e-04
1.30e-03	1.43e+16	2.82e-05	3.30e-03	9.20e+15	1.15e-04
1.50e-03	1.33e+16	3.16e-05	3.50e-03	2.93e+15	1.18e-04
1.70e-03	1.21e+16	3.63e-05	3.70e-03	4.40e+12	5.68e-06
1.90e-03	1.11e+16	4.17e-05			



(a) : 0°.

(b) : 60°.

Figure 6.21: Example of the TCAD implemented single proton energy deposition. The kinetic energy of the proton beam is 1.6 MeV.

better understanding of the obtained results. The devices in this case were simple diodes without any segmentation.

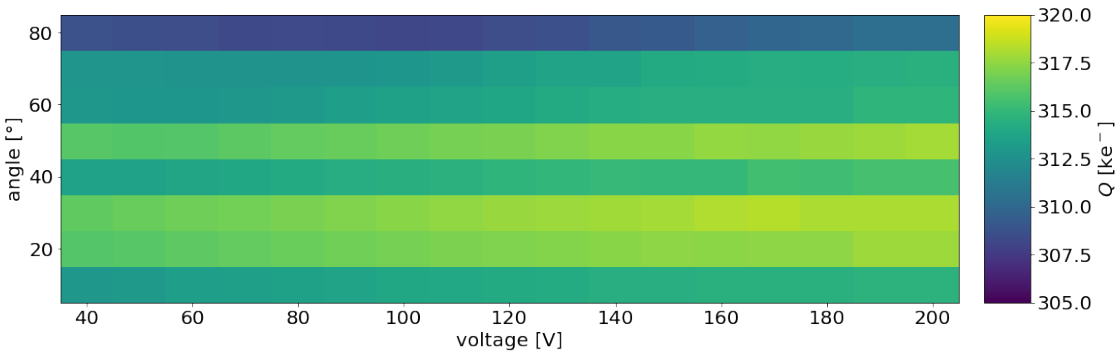
6.4.4.1 Energy deposition

Even though the Sentaurus TCAD suite provides many options to simulate the energy deposition in the device, the tools cannot create precise profiles of the energy loss. For this reason, the SRIM tool has been used to provide an estimate of the spatial distribution of the e-h pairs generated by the proton beam.

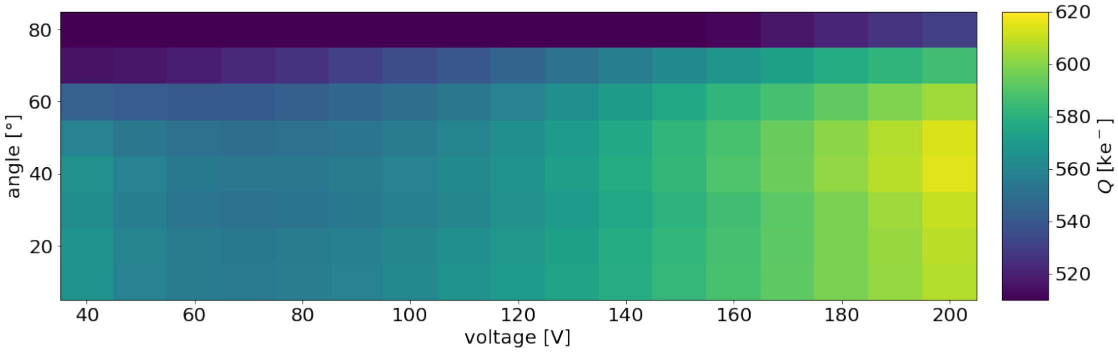
The simulation was done using a beam of 1.6 MeV protons, matching the test beam measurement setup. It was performed using 10000 particles, the resulting LET representing the energy loss in the material is plotted in the Fig. 6.20.

In order to import the results into the TCAD, the 2D spatial profile was fitted in the lateral direction using a Gaussian function every 2 µm. The peak LET, at every point was calculated from the total integral, the final values used are summarised in Table 6.1.

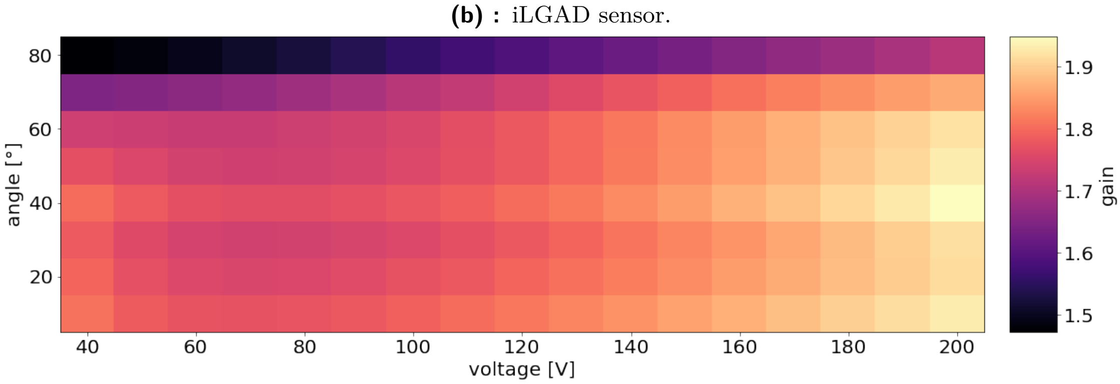
When importing custom energy deposition profile, the sdevice tool performs a linear interpolation between the provided points, matching the mesh of the device. It is possible



(a) : Sensor with no gain.



(b) : iLGAD sensor.



(c) : Calculated gain.

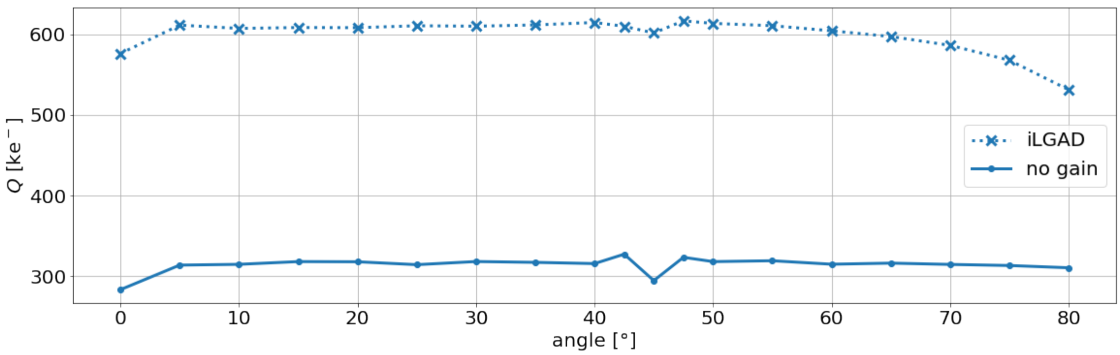
Figure 6.22: Comparison of the collected charge Q for sensors without gain and iLGADs, based on the bias voltage and incidence angle of the proton beam of energy 1.6 MeV.

to calculate the direction of the particle using the externally provided angle, an example of 0° and 60° is shown in Fig. 6.21. It is important to note that a finer mesh had to be created in order to match the small variations of the energy deposition.

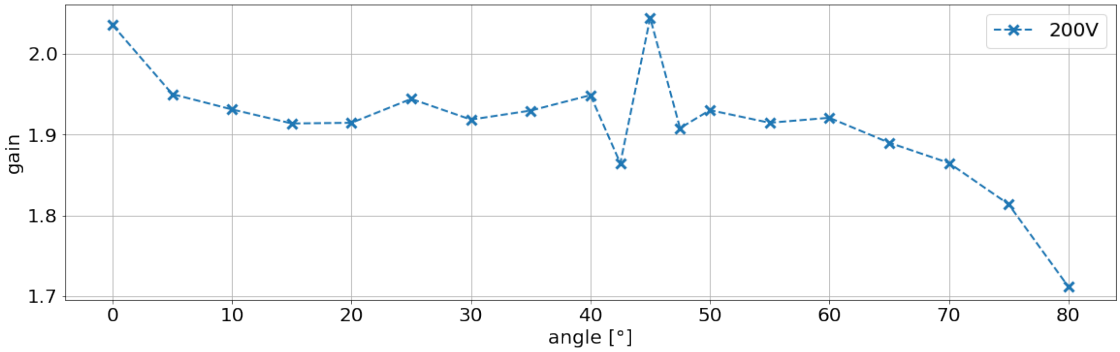
6.4.4.2 Angle-voltage scan

Using the defined simulation parameters, the scan has been performed for the range of voltage values from 40 V to 200 V. For each of the voltage steps, the incidence angle has been varied from 10° to 80°, interacting at the multiplication side. The resulting dependency as a 2D heatmap of the total collected charge and gain is in Fig. 6.22.

The charge collected by the electrode in the sensor without the internal gain provides



(a) : Collected charge.



(b) : Gain calculated as a ratio of the integrated collected charge in iLGAD and no gain sensors.

Figure 6.23: Collected charge and gain as a function of incidence angle for protons with $E_{\text{kin}} = 1.6\,\text{MeV}$ and bias voltage 200 V.

a verification baseline for the scan. As expected, the diode collected the same amount of charge regardless of the angle or the voltage applied, the small inconsistencies in the angular scans are probably caused by the meshing of the device. The only exception is the scan at 80°, where the collected charge is slightly smaller due to the fact that the interaction occurs in a dead layer of the device. This is even more prominently visible in the iLGAD sensors.

The gain is obtained using the ratio of the two results, a higher gain being reached when biasing with a higher voltage. The results indicate a stable gain up to a large incidence angle (around 70°) before it decreases. The decrease could be explained by the charge screening in the multiplication layer as well as the decreased travel distance along which the charges can multiply. This is more visible in a dedicated scan for 200 V in Fig. 6.23, the variation in the results (namely for the 0° and 45°) caused by the meshing or edge effects of the simulated region.

6.5 Summary

Increase of the average proton-proton collisions per single bunch crossing provided by the HL-LHC requires novel approaches for the detectors to separate the observed tracks. A preferred solution is an addition of a timing information, either on a per track or a per hit basis. This can be achieved by the UFSD used in the tracking stations, mainly built as silicon devices with low gain of $\mathcal{O}(10)$. Such detectors with internal multiplication region are currently undergoing an extensive R&D programme, focused on the timing performance and fill factor studies. Simulation of different aspects (IV, dependency of gain on voltage and impact angle) of recently produced iLGAD sensors is presented in the text.

Chapter 7

Summary and Conclusions

7.1 LHCb Upgrade I

The upgrades of the LHCb experiment present a complex challenge from both technological and engineering points of view. For the VELO Upgrade, the UoM heavily contributes through assembly and QA of the modules. As the precision requirements are challenging, different techniques were developed in order achieve the best possible results.

The module assembly steps may be split into three groups – starting with the construction of a bare module used as a mechanical support; followed by the alignment and gluing of tiles, hybrids and their connection using wire-bonds; and finally the rest of the data and power delivery cabling. A number of metrology procedures have been deployed in order to verify the outcome of each of the assembly steps, providing a quick feedback in the case where issues arise. The modules are then electrically tested, mimicking the operating conditions of the VELO while recording their electrical and thermal performance. In an ideal case, a single module is made and qualified in about 9 days (six days for the building with metrology tests and three days for the electrical testing), however, it is possible to parallelize the production and finish on average a module every three days. The duration from start to finish of the UoM produced modules is visualised in Fig. 7.1.

The information about the state of each module is frequently uploaded to the database, automatically analysed and visualised with numerous scripts for further inspection. Such an approach simplifies the data processing by the module builders, minimises the chance of an error and provides a safe backup of the data. The data is safely stored to be cross-referenced with the final device at any time during its expected decade of operations. As every individual step contains different information and data formats, the analysis of each of them required a dedicated development of processing algorithms. Cross-relating the values and tracking the evolution over time has been useful in understanding and correcting any issues present. This information has also been used to assign grades and quality to the tests as well as finalised modules.

By the 31 August 2021, the UoM produced: 20 grade A, 17 grade B, two grade C and D and a single grade F module. Modules graded A, B, C are suitable for installation in the experiment with full operational performance. The whole VELO sub-system requires

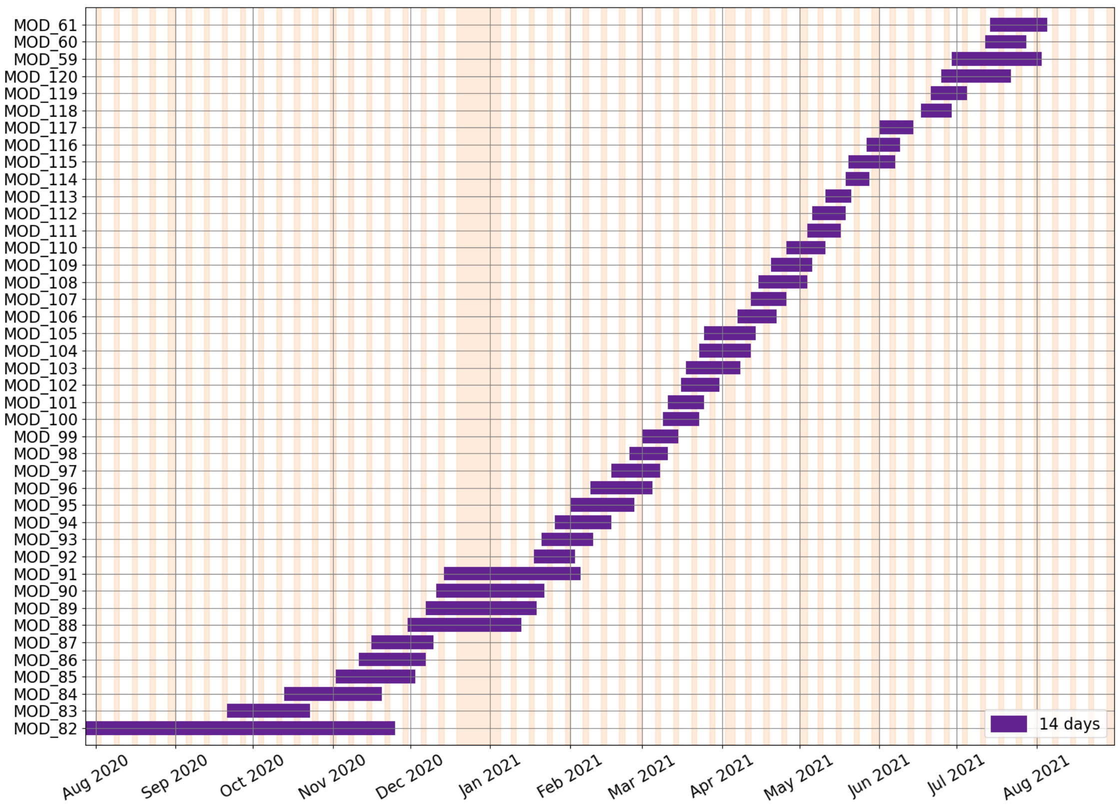


Figure 7.1: Duration of the module assembly and testing done at the UoM, plotted as of 31 August 2021.

52 modules in total, the production at both UoM and Nikhef is still ongoing at the time of writing. Afterwards, the integration of the modules into the VELO halves takes place at the UoL and a subsequent installation and commissioning of the whole detector system is planned at CERN.

Since the silicon sensors will operate in harsh radiation areas, their performance has to be monitored over time and the reverse bias voltage increased accordingly. This is normally achieved by performing a CCE scan where the depletion depth is estimated from the collected signal. In order to speed up the process in the new VeloPix ASIC, a novel 2D threshold-voltage scan method has been proposed and verified using simulated data. The results indicate that the algorithm should provide all of the information required, given a few degree cut on the angle of the track is applied. These conclusions are to be verified using the full LHCb tracking software, where one in five modules is scanned and the rest is operating at nominal conditions and used for the tracking and extrapolation of the interactions.

The duration of the new CCE scan has been estimated based on the number of scan points and voltage range, with a single subset (a fifth) of the modules taking ≈8 min. Due to the increased luminosity, the scans would have to be performed on a bi-weekly to monthly basis in order to match the same step of about $0.5\,\mathrm{fb}^{-1}$ as was done with the former VELO detector. Since the data come from the nominal LHCb physics conditions, it has been suggested to scan a different subset of modules in each session. This approach would provide the required scan frequency while keeping the impact on the data taking minimal.

7.2 LHCb Upgrade II

The LHCb Upgrade II is proposed for the LS 4 of the LHC. The framework TDR, giving potential options, is currently under review and over the next few years the R&D must be performed to select the preferred options for the construction. As it will be more difficult to distinguish the tracks due to the increase of the number of proton-proton collisions per single bunch crossing, it is envisioned that a precise time-stamping information be added.

One type of devices considered for the trackers is based on the charge multiplication in the silicon sensor (such as UFSD). However, these devices have two main issues that need to be mitigated – fill factor and radiation tolerance. While there are methods which increase the fill factor for small pitch devices (trench isolation, AC coupling or use of iLGADs), the radiation hardness has not yet been improved.

The TCAD simulations performed in this work provide an insight into the charge collection and multiplication of the recent small-pitch iLGAD sensor production. The results from simulating of the interaction of $E_{\mathrm{kin}} = 1.6\,\mathrm{MeV}$ protons for different impact angles indicate an angular dependency of the gain. The decrease of the gain for larger angles (above 60°) could be explained by the charge screening and by the lowered travel distance of the charge carriers in the multiplication layer. This is yet to be compared with the data taken during a test beam performed at RBI, Zagreb.

These devices could eventually be used to provide the timing information $\mathcal{O}(50\,\mathrm{ps})$ in the second VELO upgrade, constructed in a form of timing layers or, in the case where the radiation hardness is improved, a full 4D tracking system. The addition of the precise time-stamps could also be utilised by other LHCb tracking stations (the UT or the inner part of the Mighty Tracker).

Blank page

Bibliography

[1] P. Svihra, *The design and construction of lhcb velo upgrade modules*, , 2019. proceedings of 2019 IEEE Nuclear Science Symposium and Medical Imaging Conference (NSS/MIC), doi: 10.1109/NSS/MIC42101.2019.9059693.

[2] C. Ianzano *et al.*, *Fast camera spatial characterization of photonic polarization entanglement*, Scientific Reports **10** (2020) p. 6181, doi: 10.1038/s41598-020-62020-z.

[3] R. Sen *et al.*, *Mapping O2 concentration in ex-vivo tissue samples on a fast PLIM macro-imager*, Scientific Reports **10** (2020) p. 19006, doi: 10.1038/s41598-020-75928-3.

[4] P. Svihra *et al.*, *Multivariate discrimination in quantum target detection*, Applied Physics Letters **117** (2020) p. 044001, doi: 10.1063/5.0012429.

[5] L. A. Zhukas, P. Svihra, A. Nomerotski, and B. B. Blinov, *High-fidelity simultaneous detection of a trapped-ion qubit register*, Phys. Rev. A **103** (2021) p. 062614, doi: 10.1103/PhysRevA.103.062614.

[6] A. Roberts *et al.*, *First demonstration of 3D optical readout of a TPC using a single photon sensitive Timepix3 based camera*, JINST **14** (2019) pp. P06001–P06001, doi: 10.1088/1748-0221/14/06/p06001.

[7] A. Zhao *et al.*, *Coincidence velocity map imaging using Tpx3Cam, a time stamping optical camera with 1.5 ns timing resolution*, Review of Scientific Instruments **88** (2017) p. 113104, doi: 10.1063/1.4996888.

[8] F. Marcastel, *CERN's Accelerator Complex. La chaîne des accélérateurs du CERN*, CDS (2013), General Photo.

[9] L. Evans and P. Bryant, *LHC machine*, JINST **3** (2008) pp. S08001–S08001, doi: 10.1088/1748-0221/3/08/s08001.

[10] O. Aberle *et al.*, *High-Luminosity Large Hadron Collider (HL-LHC)*, Tech. Rep. CERN-2020-010, CERN, Geneva, 12, 2020. doi: 10.23731/CYRM-2020-0010.

[11] ALICE Collaboration, K. Aamodt *et al.*, *The ALICE experiment at the CERN LHC*, JINST **3** (2008) pp. S08002–S08002, doi: 10.1088/1748-0221/3/08/s08002.

[12] ATLAS Collaboration, G. Aad *et al.*, *The ATLAS experiment at the CERN large hadron collider*, JINST **3** (2008) pp. S08003–S08003, doi: 10.1088/1748-0221/3/08/s08003.

[13] ATLAS Collaboration and CMS Collaboration, G. Aad *et al.*, *Combined Measurement of the Higgs Boson Mass in pp Collisions at $\sqrt{s} = 7$ and 8 TeV with the ATLAS and CMS Experiments*, Phys. Rev. Lett. **114** (2015) p. 191803, doi: 10.1103/PhysRevLett.114.191803.

[14] CMS Collaboration, S. Chatrchyan *et al.*, *The CMS experiment at the CERN LHC*, JINST **3** (2008) pp. S08004–S08004, doi: 10.1088/1748-0221/3/08/s08004.

[15] LHCb Collaboration, A. A. Alves *et al.*, *The LHCb detector at the LHC*, JINST **3** (2008) pp. S08005–S08005, doi: 10.1088/1748-0221/3/08/s08005.

[16] LHCb Collaboration, R. Aaij *et al.*, *Measurement of CP Violation in the Phase Space of $B^{\pm} \to K^{+}K^{-}\pi^{\pm}$ and $B^{\pm} \to \pi^{+}\pi^{-}\pi^{\pm}$ Decays*, Phys. Rev. Lett. **112** (2014) p. 011801, doi: 10.1103/PhysRevLett.112.011801.

[17] LHCb Collaboration, R. Aaij *et al.*, *Measurement of CP violation in $B_s^0 \to \phi\phi$ decays*, Phys. Rev. D **90** (2014) p. 052011, doi: 10.1103/PhysRevD.90.052011.

[18] LHCb Collaboration, C. Elsässer, *$\bar{b}b$ production angle plots*, `https://lhcb.web.cern.ch/lhcb/speakersbureau/html/bb_ProductionAngles.html`, 07, 2021.

[19] LHCb Collaboration, I. Bediaga *et al.*, *Framework TDR for the LHCb Upgrade: Technical Design Report*, Tech. Rep. CERN-LHCC-2012-007, LHCb-TDR-12, CERN, Geneva, 04, 2012.

[20] LHCb Collaboration, P. R. Barbosa-Marinho *et al.*, *LHCb VELO (VErtex LOcator): Technical Design Report*, Tech. Rep. CERN-LHCC-2001-011, LHCb-TDR-5, CERN, Geneva, 2001.

[21] M. Agari *et al.*, *Beetle—a radiation hard readout chip for the lhcb experiment*, Nucl. Instrum. Methods. Phys. Res. A **518** (2004) pp. 468–469, Frontier Detectors for Frontier Physics: Proceeding, doi: 10.1016/j.nima.2003.11.058.

[22] R. Aaij *et al.*, *Performance of the LHCb vertex locator*, JINST **9** (2014) pp. P09007–P09007, doi: 10.1088/1748-0221/9/09/p09007.

[23] LHCb Collaboration, R. Aaij *et al.*, *LHCb detector performance*, IJMPA **30** (2015) p. 1530022, doi: 10.1142/S0217751X15300227.

[24] LHCb Collaboration, R. Aaij *et al.*, *LHCb VELO Upgrade Technical Design Report*, Tech. Rep. LHCb-TDR-013, CERN, Geneva, 2013.

[25] LHCb Collaboration, R. Antunes-Nobrega *et al.*, *LHCb reoptimized detector design and performance: Technical design report*, Tech. Rep. CERN-LHCC-2003-030, LHCb-TDR-9., CERN, Geneva, 2003.

[26] LHCb Collaboration, S. Amato *et al.*, *LHCb magnet: Technical Design Report*, Tech. Rep. CERN-LHCC-2000-007, LHCb-TDR-1, CERN, Geneva, 2000.

[27] LHCb Collaboration, P. R. Barbosa-Marinho *et al.*, *LHCb inner tracker: Technical Design Report*, Tech. Rep. CERN-LHCC-2002-029, LHCb-TDR-8, CERN, Geneva, 2002. revised version number 1 submitted on 2002-11-13.

[28] LHCb Collaboration, P. R. Barbosa-Marinho *et al.*, *LHCb outer tracker: Technical Design Report*, Tech. Rep. CERN-LHCC-2001-024, LHCb-TDR-6, CERN, Geneva, 2001.

[29] LHCb Collaboration, R. Aaij *et al.*, *LHCb Tracker Upgrade Technical Design Report*, Tech. Rep. CERN-LHCC-2014-001, LHCB-TDR-015, CERN, Geneva, 02, 2014.

[30] LHCb Collaboration, S. Amato *et al.*, *LHCb RICH: Technical Design Report*, Tech. Rep. CERN-LHCC-2000-037, LHCb-TDR-3, CERN, Geneva, 2000.

[31] LHCb Collaboration, R. Aaij *et al.*, *LHCb PID upgrade technical design report*, Tech. Rep. CERN-LHCC-2013-022, LHCB-TDR-014, CERN, Geneva, 11, 2013.

[32] LHCb Collaboration, S. Amato *et al.*, *LHCb calorimeters: Technical Design Report*, Tech. Rep. CERN-LHCC-2000-036, LHCb-TDR-2, CERN, Geneva, 2000.

[33] LHCb Collaboration, P. R. Barbosa-Marinho *et al.*, *LHCb muon system: Technical Design Report*, Tech. Rep. CERN-LHCC-2005-012, LHCb-TDR-4-add-2, CERN, Geneva, 2001.

[34] Siemens, *SIMATIC WinCC/OA*, webpage, 2021. `https://www.winccoa.com/`.

[35] LHCb VELO, group. Internal communication, 2018-2021.

[36] O. A. D. A. Francisco, *Microchannel Cooling techniques at LHCb*, POS **Vertex 2017** (2018) p. 024, doi: 10.22323/1.309.0024.

[37] A. Mapelli *et al.*, *Microfluidic cooling for detectors and electronics*, JINST **7** (2012) pp. C01111–C01111, doi: 10.1088/1748-0221/7/01/c01111.

[38] A. Bitadze and K. Akiba, *A Silicon Microchannel bi-phase CO2 Cooling for LHCb*, 2019 Nuclear Science Symposium and Medical Imaging Conference (NSS/MIC) (2019) pp. 1–5, doi: 10.1109/NSS/MIC42101.2019.9060021.

[39] O. A. D. A. Francisco *et al.*, *Evaporative CO2 microchannel cooling for the LHCb VELO pixel upgrade*, JINST **10** (2015) pp. C05014–C05014, doi: 10.1088/1748-0221/10/05/c05014.

[40] T. Poikela *et al.*, *VeloPix: the pixel ASIC for the LHCb upgrade*, JINST **10** (2015) pp. C01057–C01057, doi: 10.1088/1748-0221/10/01/c01057.

[41] E. Buchanan, *The LHCb vertex locator (VELO) pixel detector upgrade*, JINST **12** (2017) pp. C01013–C01013, doi: 10.1088/1748-0221/12/01/c01013.

[42] T. Poikela *et al.*, *The VeloPix ASIC*, Journal of Instrumentation **12** (2017) pp. C01070–C01070, doi: 10.1088/1748-0221/12/01/c01070.

[43] A. F. Prieto, *Development of Electronics for the VELO Upgrade Detector*, PhD thesis, Universidade de Santiago de Compostela, 2020, CERN-THESIS-2020-216.

[44] V. Gromov *et al.*, *Development of a low power 5.12 Gbps data serializer and wireline transmitter circuit for the VeloPix chip*, JINST **10** (2015) pp. C01054–C01054, doi: 10.1088/1748-0221/10/01/c01054.

[45] P. Moreira *et al.*, *The GBT Project*, CDS (2009) p. 5, doi: 10.5170/CERN-2009-006.342.

[46] A. Fernández Prieto *et al.*, *Phase I Upgrade of the Readout System of the Vertex Detector at the LHCb Experiment*, IEEE Trans. Nucl. Sci. **67** (2020) pp. 732–739, doi: 10.1109/TNS.2020.2970534.

[47] A. Caratelli *et al.*, *The GBT-SCA, a radiation tolerant ASIC for detector control and monitoring applications in HEP experiments*, JINST **10** (2015) pp. C03034–C03034, doi: 10.1088/1748-0221/10/03/c03034.

[48] E. L. Cid *et al.*, *Production, measurement and simulation of a low mass flex cable for multi gigabit/s readout for the LHCb VELO upgrade*, JINST **8** (2013) pp. C01018–C01018, doi: 10.1088/1748-0221/8/01/c01018.

[49] B. Allongue *et al.*, *Custom DC-DC converters for distributing power in SLHC trackers*, CDS (2008) p. 5, doi: 10.5170/CERN-2008-008.289.

[50] K. Wyllie *et al.*, *Electronics Architecture of the LHCb Upgrade*, Tech. Rep. LHCb-PUB-2011-011, CERN-LHCb-PUB-2011-011, CERN, Geneva, 06, 2013. Previous version: 01/03/2013 Current version: 18/06/2013 version 2.6) based on LHCb-INT-2011-006 (removed from CDS).

[51] P. Durante *et al.*, *100 Gbps PCI-express readout for the LHCb upgrade*, JINST **10** (2015) pp. C04018–C04018, doi: 10.1088/1748-0221/10/04/c04018.

[52] Particle Data Group, P. A. Zyla *et al.*, *Review of Particle Physics*, PTEP **2020** (2020) p. 083C01, doi: 10.1093/ptep/ptaa104.

[53] H. A. Bethe and J. Ashkin, *Experimental nuclear physics*, Wiley, New York, 1953.

[54] B. B. Rossi, *High Energy Particles*, Prentice-Hall, Inc., Englewood Cliffs, NJ, 1952.

[55] H. Bichsel, *Straggling in Thin Silicon Detectors*, Rev. Mod. Phys. **60** (1988) pp. 663–699, doi: 10.1103/RevModPhys.60.663.

[56] L. D. Landau, *On the energy loss of fast particles by ionization*, J. Phys. **8** (1944) pp. 201–205, doi: 10.1016/B978-0-08-010586-4.50061-4.

[57] P. V. Vavilov, *Ionization losses of high-energy heavy particles*, Sov. Phys. JETP **5** (1957) 749.

[58] G. R. Lynch and O. I. Dahl, *Approximations to multiple coulomb scattering*, Nucl. Instrum. Methods. Phys. Res. B **58** (1991) pp. 6–10, doi: 10.1016/0168-583X(91)95671-Y.

[59] S. M. Sze and K. K. Ng, *Physics of Semiconductor Devices*, John Wiley & Sons, Ltd, 2006.

[60] M. L. Cohen and J. R. Chelikowsky, *Electronic Structure and Optical Properties of Semiconductors*, Springer, Berlin, 1 ed., 1988.

[61] G. Lutz, *Semiconductor Radiation Detectors*, Springer-Verlag, Berlin and Heidelberg, 01, 2007.

[62] L. Rossi, P. Fischer, T. Rohe, and N. Wermes, *Pixel Detectors: From Fundamentals to Applications*, Springer-Verlag, Berlin Heidelberg, 2006.

[63] G. F. Knoll, *Radiation Detection and Measurement*, John Wiley & Sons, New York, 4 ed., 2010.

[64] C. Jacoboni, C. Canali, G. Ottaviani, and A. Alberigi Quaranta, *A review of some charge transport properties of silicon*, Solid-State Electronics **20** (1977) pp. 77–89, doi: 10.1016/0038-1101(77)90054-5.

[65] M. A. Green, *Intrinsic concentration, effective densities of states, and effective mass in silicon*, JAPIAU **67** (1990) pp. 2944–2954, doi: 10.1063/1.345414.

[66] W. Shockley, *Currents to conductors induced by a moving point charge*, JAPIAU **9** (1938) pp. 635 – 636, doi: 10.1063/1.1710367.

[67] S. Ramo, *Currents induced by electron motion*, , vol. 27, IEEE, 09, 1939. doi: 10.1109/JRPROC.1939.228757.

[68] G. Kramberger *et al.*, *Timing performance of small cell 3D silicon detectors*, Nucl. Instrum. Methods. Phys. Res. A **934** (2019) pp. 26 – 32, doi: 10.1016/j.nima.2019.04.088.

[69] C. Kenney *et al.*, *A prototype monolithic pixel detector*, Nucl. Instrum. Methods. Phys. Res. A **342** (1994) 59.

[70] H. J. Barnaby, *Total-Ionizing-Dose Effects in Modern CMOS Technologies*, IEEE Trans. Nucl. Sci. **53** (2006) pp. 3103–3121, doi: 10.1109/TNS.2006.885952.

[71] C. T. Sah, *Origin of interface states and oxide charges generated by ionizing radiation*, IEEE Trans. Nucl. Sci. **23** (1976) pp. 1563–1568, doi: 10.1109/TNS.1976.4328540.

[72] H. E. Boesch and T. L. Taylor, *Charge and interface state generation in field oxides*, IEEE Trans. Nucl. Sci. **31** (1984) pp. 1273–1279, doi: 10.1109/TNS.1984.4333495.

[73] R. Wunstorf, *A Systematic investigation of the radiation hardness of silicon detectors for high-energy physics experiments*, PhD thesis, University of Hamburg, 10, 1992, DESY-FH1K-92-01.

[74] V. A. J. Van Lint, T. M. Flanagan, and R. E. Leadon, *Mechanisms of radiation effects in electronic materials v.1*, Wiley, Hoboken, 1980.

[75] M. Moll, *Radiation damage in silicon particle detectors. Microscopic defects and macroscopic properties*, PhD thesis, University of Hamburg, 12, 1999, DESY-THESIS-1999-040.

[76] J. Lindhard and V. Nielsen, *Nuclear collisions and ionisation fluctuations in charged particle detectors*, Physics Letters **2** (1962) pp. 209–211, doi: 10.1016/0031-9163(62)90229-9.

[77] RD50 Collaboration, A. Vasilescu and G. Lindstroem, *Displacement damage in silicon, on-line compilation*, `https://rd50.web.cern.ch/niel/`, 07, 2021.

[78] G. Kramberger *et al.*, *Modeling of electric field in silicon micro-strip detectors irradiated with neutrons and pions*, JINST **9** (2014) pp. P10016–P10016, doi: 10.1088/1748-0221/9/10/p10016.

[79] V. Eremin, E. Verbitskaya, and Z. Li, *The origin of double peak electric field distribution in heavily irradiated silicon detectors*, Nucl. Instrum. Methods. Phys. Res. A **476** (2002) pp. 556–564, Proc. of the 3rd Int. Conf. on Radiation Effects on Semiconductor Materials, Detectors and Devices, doi: 10.1016/S0168-9002(01)01642-4.

[80] G. Kramberger *et al.*, *Effective trapping time of electrons and holes in different silicon materials irradiated with neutrons, protons and pions*, Nucl. Instrum. Methods. Phys. Res. A **481** (2002) pp. 297–305, doi: 10.1016/S0168-9002(01)01263-3.

[81] Fisnar, *Fisnar F5200N*, webpage, 2021. `https://www.fisnar.com/products/robotics/benchtop-robots/f5200n-2-compact-gantry-benchtop-robot/`.

[82] S. De Capua *et al.*, *Shear and Thermal Testing of Adhesives for VELO Upgrade*, Tech. Rep. LHCb-PUB-2016-026, CERN-LHCb-PUB-2016-026, LHCb-INT-2015-001, CERN, Geneva, 11, 2016.

[83] Henkel, *Stycast 2850FT*, webpage, 2021. `https://www.henkel-adhesives.com/ch/en/product/potting-compounds/loctite_stycast_2850ft.html`.

[84] OGP, *SmartScope*, webpage, 2021. `https://www.ogpnet.com/products/metrology-systems/multisensor-metrology-systems/smartscope-video-multisensor-systems/smartscope-zip-multisensor-systems/smartscope-zip-300/`.

[85] OGP, *Measure-X software*, webpage, 2021. `https://www.ogpnet.com/products/software/metrology-software/measure-x/`.

[86] Edwards-vacuum, *GasCheck G3 leak detector*, webpage, 2021. `https://www.edwardsvacuum.com/en/our-products/leak-detection/gascheck`.

[87] B. Hommels and S. Sigurdsson, *NTC interlock box*, webpage, 2021. `https://www.hep.phy.cam.ac.uk/~hommels/interlock/TemperatureInterlock.htm`.

[88] Tektronix, *Keithley 2400 series*, webpage, 2021. `https://www.tek.com/keithley-source-measure-units/keithley-smu-2400-standard-series-sourcemeter`.

[89] Qt, *Qt software*, webpage, 2021. `https://doc.qt.io/`.

[90] A. Affolder *et al.*, *Radiation damage in the LHCb vertex locator*, JINST **8** (2013) pp. P08002–P08002, doi: 10.1088/1748-0221/8/08/p08002.

[91] LHCb Collaboration, *The Boole Project*, webpage, 2021. `https://lhcbdoc.web.cern.ch/lhcbdoc/boole/`.

[92] GEANT4, S. Agostinelli *et al.*, *GEANT4: A Simulation toolkit*, Nucl. Instrum. Meth. Phys. Res. A **506** (2003) pp. 250–303, doi: 10.1016/S0168-9002(03)01368-8.

[93] D. Hynds, *Resolution Studies and Performance Evaluation of the LHCb VELO Upgrade*, PhD thesis, University of Glasgow, 2014, CERN-THESIS-2014-292.

[94] G. Sarpis, *Search for CP Violation in Baryon Decays and Radiation Damage of the LHCb VELO*, PhD thesis, University of Manchester, 2020, CERN-THESIS-2020-051.

[95] LHCb Collaboration, R. Aaij *et al.*, *Framework TDR for the LHCb upgrade II opportunities in flavour physics, and beyond, in the HL-LHC era*, Tech. Rep. LHCB-TDR-023-001, CERN, Geneva, 07, 2021.

[96] LHCb Collaboration, R. Aaij, B. Adeva, M. Adinolfi *et al.*, *Physics case for an LHCb Upgrade II*, Tech. Rep. LHCb-PUB-2018-009, CERN-LHCb-PUB-2018-009, CERN, Geneva, 05, 2018.

[97] H. F.-W. Sadrozinski, A. Seiden, and N. Cartiglia, *4D tracking with ultra-fast silicon detectors*, Reports on Progress in Physics **81** (2017) p. 026101, doi: 10.1088/1361-6633/aa94d3.

[98] M. Ferrero *et al.*, *An Introduction to Ultra-Fast Silicon Detectors*, CRC Press, Boca Raton, 07, 2021.

[99] G. Pellegrini *et al.*, *Technology developments and first measurements of low gain avalanche detectors (LGAD) for high energy physics applications*, , vol. 765, 2014. HSTD-9 2013 - Proceedings of the 9th International Hiroshima Symposium on Development and Application of Semiconductor Tracking Detectors, doi: 10.1016/j.nima.2014.06.008.

[100] P. Webb and A. R. Jones, *Large area reach-through avalanche diodes for radiation monitoring*, IEEE Trans. Nucl. Sci. **21** (1974) pp. 151–158, doi: 10.1109/TNS.1974.4327455.

[101] S. Cova *et al.*, *20-ps timing resolution with single-photon avalanche diodes*, Review of Scientific Instruments **60** (1989) pp. 1104–1110, doi: 10.1063/1.1140324.

[102] C. Allaire *et al.*, *Beam test measurements of Low Gain Avalanche Detector single pads and arrays for the ATLAS High Granularity Timing Detector*, JINST **13** (2018) pp. P06017–P06017, doi: 10.1088/1748-0221/13/06/p06017.

[103] W. Maes, K. De Meyer, and R. Van Overstraeten, *Impact ionization in silicon: A review and update*, Solid-State Electronics **33** (1990) pp. 705–718, doi: 10.1016/0038-1101(90)90183-F.

[104] R. Van Overstraeten and H. De Man, *Measurement of the ionization rates in diffused silicon p-n junctions*, Solid-State Electronics **13** (1970) pp. 583–608, doi: 10.1016/0038-1101(70)90139-5.

[105] Y. Okuto and C. R. Crowell, *Threshold energy effect on avalanche breakdown voltage in semiconductor junctions*, Solid-State Electronics **18** (1975) pp. 161–168, doi: 10.1016/0038-1101(75)90099-4.

[106] D. J. Massey, J. P. R. David, and G. J. Rees, *Temperature dependence of impact ionization in submicrometer silicon devices*, IEEE Transactions on Electron Devices **53** (2006) pp. 2328–2334, doi: 10.1109/TED.2006.881010.

[107] A. G. Chynoweth, *Ionization rates for electrons and holes in silicon*, Phys. Rev. **109** (1958) pp. 1537–1540, doi: 10.1103/PhysRev.109.1537.

[108] W. Shockley and W. T. Read, *Statistics of the recombinations of holes and electrons*, Phys. Rev. **87** (1952) pp. 835–842, doi: 10.1103/PhysRev.87.835.

[109] G. Paternoster *et al.*, *Novel strategies for fine-segmented low gain avalanche diodes*, Nucl. Instrum. Methods. Phys. Res. A **987** (2021) p. 164840, doi: 10.1016/j.nima.2020.164840.

[110] N. Moffat *et al.*, *Low gain avalanche detectors (LGAD) for particle physics and synchrotron applications*, JINST **13** (2018) pp. C03014–C03014, doi: 10.1088/1748-0221/13/03/c03014.

[111] I. Cortes *et al.*, *Gain estimation of RT-APD devices by means of TCAD numerical simulations*, , 03, 2011. Proceedings of the 8th Spanish Conference on Electron Devices, CDE'2011, doi: 10.1109/SCED.2011.5744152.

[112] J. F. Ziegler, *SRIM-2003*, , vol. 219-220, 2004. Proceedings of the Sixteenth International Conference on Ion Beam Analysis, doi: 10.1016/j.nimb.2004.01.208.

[113] Synopsys, *Synopsys TCAD*, webpage, 2021. `https://www.synopsys.com/silicon/tcad.html`.

[114] E. Currás *et al.*, *Inverse Low Gain Avalanche Detectors (iLGADs) for precise tracking and timing applications*, , vol. 958, 2020. Proceedings of the Vienna Conference on Instrumentation 2019, doi: 10.1016/j.nima.2019.162545.

Blank page

Appendices

Blank page

High Energy particle physics Ventilator

The onset of the global COVID-19 pandemic resulted in the disruption of the VELO Upgrade module production. Since most of the work presented is based on the analysis and testing of the new modules, the author of this thesis decided to contribute to a CERN based initiative, centred around members of the VELO group, of developing an open-source reliable high-end ventilator that could be deployed for the patients suffering with pulmonary diseases. Most of the project development took about half a year during which a number of fully functioning prototypes have been developed and tested, more details are provided in the attached pre-print.

The main contribution of the author of this thesis was dedicated towards the development and testing of the core software. The software chain can be split based on the processing unit – an ESP32 micro controller used to control the hardware and a Raspberry Pi used for user input/output and data storage; the two linked via USB.

A communication protocol based on the High-level Data Link Control (HDLC) data format has been proposed, providing modularity and supporting different types of data payloads. This way the priority of data transfers could be defined, handling alarms or a simple setting read-back in a different manner. Dedicated communication libraries have been developed both in C (ESP32 micro controller) as well as Python (Raspberry Pi), all implemented in a non-blocking and thread-safe way. For the safety of the operation, the data format contains counters and checksums which ensure the integrity and correct alignment of the communication.

Additional work on the ESP32 firmware has been done, focused on the command processing and alarm propagation within the event loop. This has been frequently tested with the real prototypes, rapid development often required as this linked the operation between the user interface and hardware. The author also contributed towards the development of the touchscreen based GUI system, controlled by the Raspberry Pi.

At the time of writing, the HEV has been adapted and expanded upon by an UK collaboration HPLV, aiming to provide a better baseline needed for the medical certification. It has also been licensed by Swiss (Jean Gallay SA) and Indian (AXIS enterprises PVT. LTD.) companies.

EUROPEAN ORGANIZATION FOR NUCLEAR RESEARCH (CERN)

CERN-EP-TECH-NOTE-2020-002
July 24, 2020

The HEV Ventilator

HEV Collaboration†

Abstract

HEV is a low-cost, versatile, high-quality ventilator, which has been designed in response to the COVID-19 pandemic. The ventilator is intended to be used both in and out of hospital intensive care units, and for both invasive and non-invasive ventilation. The hardware can be complemented with an external turbine for use in regions where compressed air supplies are not reliably available. The standard modes provided include PC–A/C (Pressure Assist Control), PC–A/C–PRVC (Pressure Regulated Volume Control), PC-PSV (Pressure Support Ventilation) and CPAP (Continuous Positive Airway Pressure). HEV is designed to support remote training and post market surveillance via a web interface and data logging to complement the standard touch screen operation, making it suitable for a wide range of geographical deployment. The HEV design places emphasis on the quality of the pressure curves and the reactivity of the trigger, delivering a global performance which will be applicable to ventilator needs beyond the COVID-19 pandemic. This article describes the conceptual design and presents the prototype units together with their performance evaluation.

†Authors are listed at the end of this paper.

Contents

1 Motivation **1**

2 Overview of HEV functionality **1**
2.1 HEV operation modes . . . 1
2.2 Specifications . . . 3

3 HEV Conceptual Design **4**
3.1 Central Pneumatic Unit . . . 4
3.2 Alternative air supply . . . 8
3.3 Accessories and disinfection protocols . . . 8
3.4 Electrical design . . . 10
3.5 Control and User Interface . . . 10
3.5.1 Control . . . 10
3.5.2 Microcontroller functionality . . . 11
3.5.3 Raspberry Pi functionality . . . 12
3.5.4 Communication . . . 14
3.5.5 Software Practice . . . 14
3.5.6 User Interface . . . 14

4 HEV Prototypes **16**
4.1 Prototype mechanical design . . . 17
4.2 Alternative mechanical design . . . 19

5 Prototype test results **20**
5.1 Setup . . . 20
5.2 Target Pressure Performance . . . 21
5.3 Inhale Trigger Peformance . . . 25
5.4 Oxygen mixing test . . . 27

6 Conclusion **29**

References **30**

1 Motivation

The worldwide medical community currently faces a shortage, especially in low and middle income settings, of medical equipment to address the COVID-19 pandemic [1]–[5]. In particular this is the case for ventilators, which are needed during COVID-19 related treatment both in the acute phase, when invasive fully controlled ventilation is needed, and also in the subacute phase during the weaning from mechanical ventilation, which can last for an extended time period. Companies are scaling up production [6], but this will not be sufficient to meet the demand according to the current forecasts. There currently exist a wide spectrum of ventilatory support devices, ranging from highly sophisticated through to simpler units [7], [8]. In the context of the COVID-19 pandemic a large number of proposals are already circulating for devices which can be quickly manufactured cheaply and on large scale [9]–[15].

The pandemic has also drawn attention to the lack of ventilation equipment in low and middle income countries. Globally, pneumonia is the most common infectious cause of death [16]–[18], and the need for adequate respiratory equipment for treatment and management of pneumonia patients will persist even as the COVID-19 pandemic wanes.

In this paper we describe the design of a high-quality ventilator, named HEV (High Energy particle physics Ventilator), which is intended to provide full functionality while being capable of manufacture at relatively low cost, and show the results obtained from the prototype testing. This is an update from the original proposal [19]. The HEV design [20] is based on readily available and inexpensive components, and the ventilator is intended to be used both in and out of the intensive care unit environment, and for either intubated or non-intubated patients. The design began in March 2020, using as a starting point the set of MHRA (Medicines and Healthcare products Regulatory Agency) guidelines provided by the UK government regarding Rapidly Manufactured Ventilator Systems [21]. As the project evolved, regulations and recommendations from other bodies, including the EU, AAMI and WHO [22]–[24] were taken into consideration, the team was reinforced by an international advisory body of clinicians, and advice, collaboration and equipment was provided from local hospitals. The design and prototypes described here are not, at this moment in time, a medically approved system and will need a process of verification according to medical certification. However, the design is sufficiently advanced that the functional results can already be shared, as the HEV collaboration moves forward towards certification and manufacture.

2 Overview of HEV functionality

2.1 HEV operation modes

Patients affected by COVID-19 face serious issues of lung damage, and the ventilatory equipment must be able to handle situations of rapidly changing lung compliance as well as potential collapse and consolidation. It is critical that the ventilator is able to deliver protective ventilation to patients with nearly normal as well as patients with low compliance. The driving pressure of the ventilator is a crucial factor for patient outcomes [25]. In particular, when a low tidal volume is used, the driving pressure is an important variable to monitor and assess the risk of hospital mortality [26]. In light of the prolonged recovery/weaning phases involved in COVID-19 critical care cases, there is a need for ventilators which are able to deliver protective controlled ventilation but also are able to delivery assisted ventilation to be efficient for the ventilator weaning process.

Following the COVID-19 guidelines, the HEV development has prioritised pressure modes, and aims to offer these modes in the simplest possible format. The HEV ventilation modes are divided into two groups: pressure-assist-controlled modes, where the patient effort plays no role or a partial role in the ventilation, and pressure-supported modes, where the patient breathing is spontaneous but supported by the ventilator[1]. The former group includes PC–A/C (Pressure Control – Assist Control) and PC–A/C–PRVC (Pressure Control – Assist Control – Pressure Regulated Volume Controlled), while the latter includes PC-PSV (Pressure Control – Pressure Supported Ventilation), and in addition provides CPAP (Continuous Positive Airway Pressure). The CPAP mode is also available, as for lower-cost devices which offer less versatility than the HEV ventilator. This mode is included with the HEV modes in order to provide the widest range of support throughout COVID-19 treatment, and may be a crucial option for selection in low resource settings [28].

The modes supported by HEV are summarised in table 1. Note that in the interests of simplicity and ease of operation the patient inhalation trigger is to be set for all modes, although the exact trigger levels can be adjusted by the clinician. This is important to avoid isometric contractions of the diaphragm against a closed valve that could promote diaphragm injuries and prolonged duration of ventilation. The HEV design also allows PEEP (Positive End-Expiratory Pressure), which is not a ventilation mode in itself but is designed to support steady low positive pressure to the lungs to avoid alveolar collapse.

HEV nomenclature	**Patient trigger**	**Inhalation start**	**Exhalation start**	**Respiratory rate (RR)**	**Comment**
Pressure Control modes					
PC-A/C	on	Machine/ patient triggered	Machine cycled	Minimum RR programmed, patient effort can increase RR	p_{inh} constant
PC-A/C-PRVC	on	Machine/ patient triggered	Machine cycled	Minimum RR programmed, patient effort can increase RR	Volume guarantee via p_{inh} variation
Pressure Support modes					
PC-PSV	on	Patient triggered	Patient cycled	Spontaneous In case of apnea fail-safe to PC–A/C	p_{inh} constant
CPAP	on	Constant Positive Airway Pressure Spontaneous breaths In case of apnea fail-safe to PC–A/C			

Table 1: Summary of HEV basic ventilation modes.

[1]For a pedagogical discussion of ventilator modes, see [27].

The PC–A/C mode supplies a defined target pressure to the patient, with a PEEP defined and set by the clinician. In the case where there is no patient effort, the breathing rate and the inhalation time are fully defined by the parameters entered by the clinician. The mode also allows the start of inhalation to be triggered by patient effort, detected from the air flow measurement. The PC–A/C–PRVC mode is an extended mode of PC–A/C ventilation, where the ventilator aims to provide a set tidal volume at the lowest possible airway pressure. This works by ventilating at an initial set pressure, and if the tidal volume is not achieved, (due to, for example, changes in the patient airway resistance or lung compliance), the ventilation can be gradually adjusted. In both of these modes the exhalation part of the cycle is triggered by timing, defined by the parameters set by the clinician.

In PC-PSV mode, the inhalation and exhalation parts of the breathing cycle are defined by patient effort and lung mechanics. The inhalation is triggered as for the PC–A/C modes, and the exhalation is triggered when the flow drops to a pre-defined fraction of the peak value – typically 25%, but can be set by the clinician.

In all modes it is possible to measure the plateau pressure and intrinsic PEEP in order to provide clinical diagnoses and estimation of the patient static lung compliance or to detect AutoPEEP. The ventilator operation sequence includes a pause time at the end of the inhalation phase during which the valves are closed, normally set to an imperceptibly short time of 5 ms. The pause time can be increased, for a few breaths, to a few hundred milliseconds, in order to accurately measure to plateau alveolar pressure at zero flow. The intrinsic PEEP at the end of the exhale phase can be measured in the same way during the pre-inhale state. These measurements are manual operations performed typically over the course of three or four breaths, and can be used to track clinical parameters.

Particular attention has been paid during the design of HEV to the ability of the machine to synchronise to the patient and to offer suitable pressure profiles, issues which directly influence the patient comfort [29]. This is discussed more completely in section 5 where the HEV prototype measurements are presented.

2.2 Specifications

The HEV ventilator is designed to meet the main specifications defined in table 2.

Specification	**Characteristics**
Operation modes	PC–A/C, PC–A/C–PRVC, PC-PSV, CPAP, as defined in the text
PEEP installed	Adjustable in the range $5 - 20\,\mathrm{cm\,H_2O}$, in increments of $5\,\mathrm{cm\,H_2O}$
Inhalation airway plateau pressure limit	Set to $35\,\mathrm{cm\,H_2O}$ by default, with an option to increase to up to $70\,\mathrm{cm\,H_2O}$ in exceptional circumstances and by positive decision and action by the user
Mechanical fail-safe valve	At $80\,\mathrm{cm\,H_2O}$
Mechanical fail-safe valve to connect patient to atmosphere in case of failure	Present
Minute volume flow capability	Up to 20 L/min
Inspiratory flow capability	Up to 120 L/min
Respiratory rate	$10 - 30$ breaths/min, adjustable in increments of 2
Inhalation:Exhalation time ratio	1:2 provided as standard, adjustable in the range 1:1–1:3
Inhalation time	Optional setting
Tidal volume setting	Provided in the range $250 - 2000$ mL in steps of 50 mL
Inhaled oxygen proportion FIO_2	Adjustable between 21% and 100% in 10% steps
Gas and power supply inlets	Set according to the MHRA standards [21]
Mandatory alarms	Gas or electricity supply failure Machine switched off while in mandatory ventilation mode Inhalation airway pressure exceeded Inhalation and PEEP pressure not achieved Tidal volume not achieved or exceeded Hypoventilation and high leakage
Monitoring of set parameters	Ventilation mode, Respiratory rate, Tidal volume, PEEP, and FIO_2
Monitoring of actual measured parameters	Airway pressure, Respiratory rate, Achieved tidal volume, PEEP, FIO_2, and real time confirmation of each patient breath in pressure support mode

Table 2: Summary of the HEV main specifications.

3 HEV Conceptual Design

3.1 Central Pneumatic Unit

The HEV conceptual schematic is shown in figure 1. The design is based around a central buffer which pneumatically decouples the ventilator circuit into two, almost independently

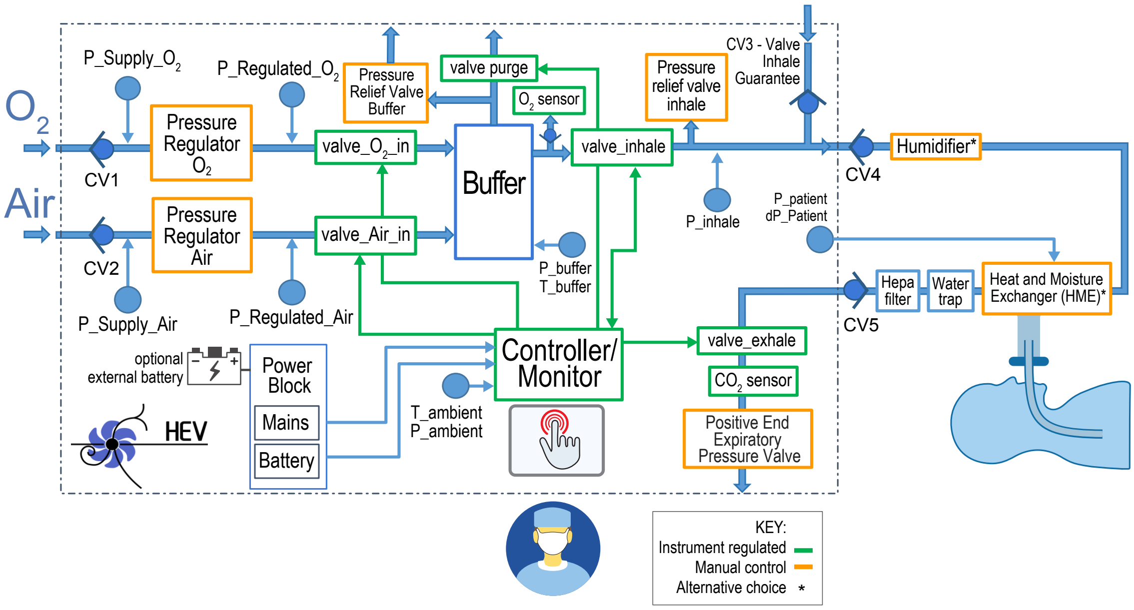


Figure 1: Conceptual design of the HEV ventilator.

functioning parts, relating to the filling of the buffer and the gas supply to the patient, respectively.

On the filling side of the circuit, two inputs are provided for air and oxygen. These may be supplied via the standard compressed hospital air and oxygen supplies, in which case typical input pressures will be between 2 and 5 bar. Alternatively, the input may be supplied via a compressor, oxygen concentrator, or turbine, in which case lower pressures can be expected. In either case, regulators ensure a step down of pressure to the buffer, and the input valves are used to control the gas flow into the buffer. The individual flows are passively mixed inside the buffer, which is filled to a target pressure. Once this has been achieved, the buffer output valve is opened, initiating the respiratory cycle. This valve is controlled by a PID (ProportionalIntegralDerivative) controller, which allows a stable delivery of pressure and a fine tuning of the pressure rise time. The controller takes as input the inhale pressure measurement and regulates the valve opening to maintain the inhale pressure at the target value. This has the advantage that the pressure delivery is independent of flow and buffer pressure. At the end of the inhalation phase there is a possibility to set a wait time, during which both the inhale and exhale valves are closed. After the inhalation phase finishes, the exhale valve is opened for exhalation and the buffer is re-filled for the next breath cycle. The standard state diagram illustrating these actions is shown in figure 2. The state changes are controlled via the microcontroller and may be time or condition driven. For CPAP operation the input valves to the buffer are kept open and the PID valve is regulated to supply a constant level of pressure. The PID algorithm ensures that the system is robust against fluctuations in flow or gas supply pressure.

The buffer concept presents many operational advantages. In general, the separation of

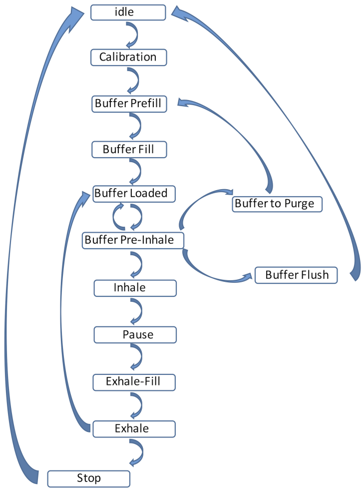


Figure 2: HEV Ventilator State Chart.

the fill and exhale cycle into two separate circuits makes the design, control and component selection more straightforward, and allows less expensive components to be selected. The initial step-down of the pressure between the supply and the patient introduces safety and robustness against variations in the gas supply. It also makes the fine tuning of the precise pressure control which is required for the patient side of the circuit more readily accessible. The buffer volume also avoids that the O_2 and air delivery systems need to be able to handle the peak flow rates of up to 120 L/min needed in the inhale phase, and the patient air supply is protected. The mixing of the gases, which is provided in a natural way inside the buffer, avoids the need to purchase an external gas blender. In addition the measurement of the O_2 concentration, which can be done by spying on the static gas volume, is an inherently more precise measurement than measuring on a gas stream, and does not require a fast reaction time of the meter, nor a medically compatible meter. Should the design need to be adapted to a more extreme (very hot, or very cold) environment, thermal control of the gas in the buffer is straightforward. In addition there is a monitoring advantage: the delivered tidal volume can be calculated from the pressure drops in the buffer. This provides a precious monitoring cross-check in addition to the standard tidal volume measurement, reinforcing the safety of the design.

To ensure better oxygen intake and to prevent lung collapse, the exhalation branch is fitted with a PEEP valve. The respiratory rate, inhalation time and pause time are all controllable. If a PEEP pressure is set, then the pressure in the lungs will not fall below the minimum of the PEEP pressure. For normal vent mode operation the pause time is set to be imperceptibly short (~ 5 ms), however longer values can be set, for instance when the care giver wishes to measure the lung static compliance. The design concept provides inherent safety in that there is a step down of pressure from the regulator before the air reaches the buffer, and in addition the inhale valve is only opened once the pressure is stabilised. Via the buffer step the system is also inherently protected from variations in the inlet pressures.

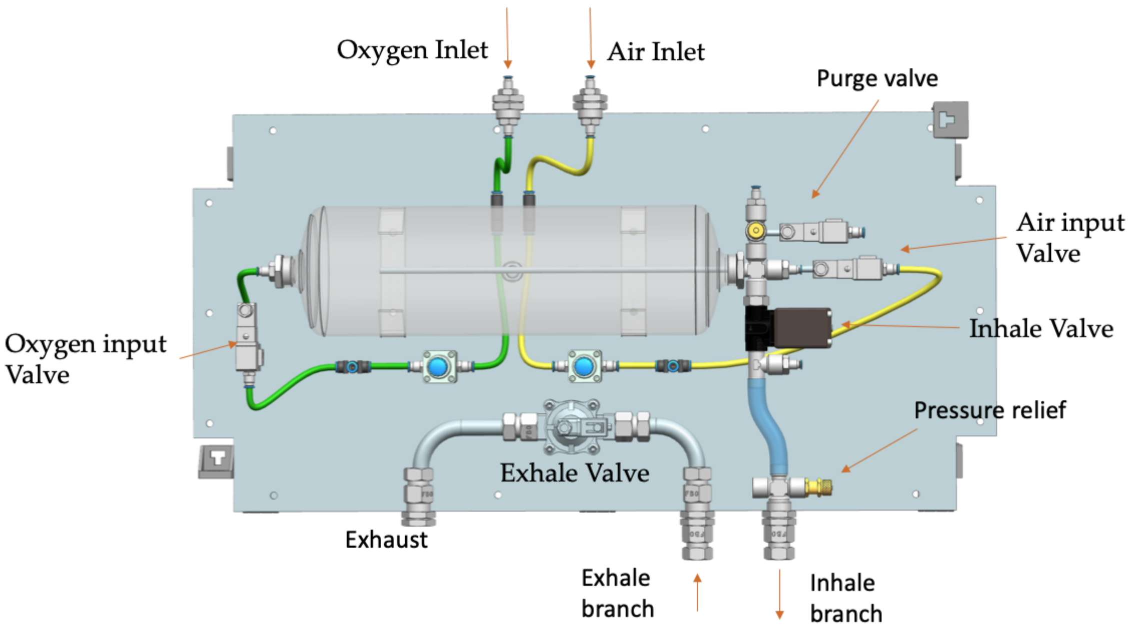


Figure 3: Schematic of the pneumatic part of the HEV system.

The volume of air taken by the patient (tidal volume) can be calculated knowing the buffer pressure measurements, the PEEP value, and the fixed volumes of the buffer and tubes in the system. The volumes of the tubes, typically between 1 and 2 L, including any equipment such as humidifying devices, may be routinely confirmed, along with checking for leaks and measuring tube compliance, before the ventilator is connected to the patient.

The patient is directly protected from over-pressure via the pressure relief valve, which opens at $50\,\mathrm{cm}\,H_2O$. In addition, the pressure sensor in the buffer continuously monitors for over-pressure. In case of over-pressure in the buffer, the electro-valve is opened to purge its contents and refill it to the correct level. In case of failure of the output valve of the buffer or a power cut, the system switches to exhalation mode, and the patient is connected to atmosphere via the inhale guarantee valve, which opens mechanically.

A schematic of the central pneumatic part of the system is shown in figure 3. This diagram highlights how pure oxygen and air are injected separately into the buffer by using two different valves. To achieve a good mix, a tube is inserted into the tank from the air side, taking air to be injected close to the oxygen injection point. Due to the high flow and the transition of the tube into the larger buffer a turbulent flow is generated, which mixes the oxygen and air. The mixture then travels to the other side of the buffer where it is removed via the proportional valve. The concentration of the oxygen is measured and in case it is not correct it is possible to vent the buffer and the cycle can start over. The oxygen levels can be mixed from 21 to 100 percent. The oxygen fraction is measured with zirconia sensors inserted after the oxygen tank which are reliable, inexpensive and readily available. The sensors are installed after a bypass, which enables them to be independent of the patient supply and to use a low flow of gas. As an additional option, a CO_2 sensor can be installed at the exit of the of the respirator and read out from the controller. CO_2 sensors based on non-dispersive (NDIR) technology are inexpensive, have good selectivity, and are no affected by the oxygen level.

3.2 Alternative air supply

The HEV collaboration is actively investigating alternatives to the compressed air supply available in hospitals. The major requirement is that the system should be able to fill the HEV buffer in less than one second.

The first option considered is a relatively small and transportable system, based on turbine blowers. A prototype has been built with 3 small turbines in series, powered by 24 V. The speed is controlled by a 0 to 5 voltage level. The prototype is illustrated in figure 4. The air is filtered by a HEPA (high-efficiency particulate air) filter. The different parts are interconnected by pieces of 1 inch plastic pipe. Before reaching the outlet, the air is transported by a corrugated steel pipe, about 1 meter long, acting as an intermediate reservoir. In addition, the corrugated steel pipe allows thermal exchange with the environment, reducing the temperature of the air which is heated in the turbines. In the system, the temperatures of the turbines and of the compressed air are continuously monitored with thermal sensors. Pressure sensors are used to maintain the required air flow by acting on the turbine speed. The prototype has been tested by connecting to a $\sim$10 L buffer acting as a HEV simulator, with one input valve and one output valve alternatively working on a 3 to 4 seconds cycle. The test shows that the turbines are able to fill the buffer in the required time and are thermally stable. The change of the temperature in the reservoir was less than 0.5 °C, after one hour of continuous work.

For increased independence from the hospital setting, it would be ideal to include an external battery which could power the turbine and the HEV. Suitable candidates, of the order of 80 Ah, 24 V can be found on the market, which also include the option to be recharged by solar panel. A relatively low-cost option which will be tested in the next prototype is based on a battery from the e-bike market. Figure 4, right, shows the concept for a complete turbine-based system. The system is divided in two parts separated by the support for the turbines. The corrugated steel pipe, temperature and pressure sensors, and the outlet connector are at placed behind the support. The air filter, power supply, and a box with a controller are on one side of the structure. An optional battery system can be installed at the other side. A system of low-power fans is used to cool the turbines.

An alternative to a custom designed turbine unit would be a commercial oil-free compressor. Suitable candidates have been identified with high availability and low cost and are being tested in parallel with the turbine system.

3.3 Accessories and disinfection protocols

The choice of breathing circuit is an element of the system which is considered as an accessory which can be finalised in the manufacturing stage of the ventilator. HEV has been tested

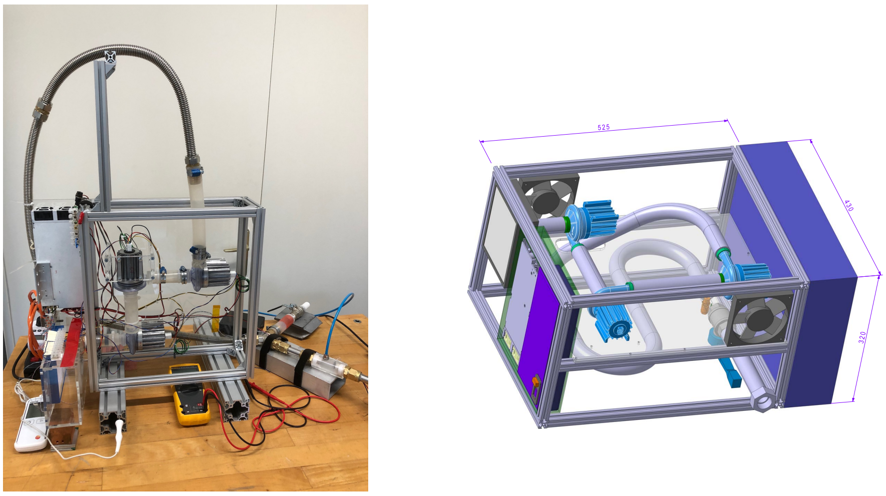


Figure 4: Left: prototype of the turbine system proposed as an alternative to the compressed air supply. Right: drawing of the concept for a complete system which is a box divided in two parts: the top contains the turbines, the bottom contains the corrugated steel pipe, the thermal and pressure sensors and the outlet connector. Mounted on the left side of the box are the air filter, the power supply and the box containing the controller. The blue box on the right provides an enclosure for the optional battery system.

with concentric tube geometry and a double limb circuit, and could be supplied with an adapter to use either method.

The breathing circuit can also be equipped with humidifier or a heat and moisture exchanger (HME) filter by choice, which is necessary to protect the patient, who may be dehydrated, from the dry medical or ambient air. The additional volume and the small pressure drop introduced by the humidifier device will be taken into account during the calibration setup phase of HEV and do not affect the operational parameters in any way. Alternatively, an HME filter might be used to maintain humidity levels and provide particle filtering, particularly if there is a worry that condensation in the flow pressure sample lines can affect the monitoring. The HEV will likely feature a HEPA filter to be used before the exhale valve to filter 99.9% of bacteria and virus, allowing the exhale to be released to atmosphere. In this case there would not be a need for a scavenger circuit.

All equipment that comes in contact with the patient needs to be either changed or disinfected and sterilised after every patient. There are multiple options to do this and currently autoclave cleaning is supported, i.e all the materials which will be reused need to withstand a temperature of 132°C for up to 4 minutes. The entire exhaust block may be easily dismounted and swapped with a spare block, so that the ventilator can continue to be used for the next patient while the block undergoes steam or autoclave sterilisation.

The PEEP valve is a purchased accessory. Care must be taken to select a valve with sufficient range and precision and priority will be given to reusable components which can be cleaned by chemical disinfection or autoclave procedures.

3.4 Electrical design

The electrical design has focused on rapid production for the HEV prototype. It has been implemented as two parts to allow flexibility in design choices and to simplify future modifications: a motherboard for hosting a microcontroller, interfacing to valves, sensors, and connecting to the user interface provided by a Raspberry Pi via a touchscreen; and a power system, including an uninterruptible power supply (UPS), to provide the power to the motherboard and components connected to it.

The motherboard is the core of the system with the following features. A connector serves to mount and power the Raspberry Pi and access the signals on its general-purpose-input-output (GPIO) bus. The board also has connectors to mount and power the ESP32 microcontroller, and access the input/output signals for valve control, sensing and other ancillary functions. There are integrated circuits to drive control signals to valves and the corresponding connectors for cabling to the valves. These circuits are driven by the ESP32 using pulse-width-modulation appropriate to the choice of valve. Further connectors serve for the cabling carrying signals from sensors. The signals are attenuated on the motherboard before connection to the ESP32 for digitisation. The motherboard also provides a number of embedded sensors for monitoring and debugging, light-emitting-diodes (LEDs) and a buzzer for user information and alarms, spare channels for additional valve and sensor connections, and connectors to allow the powering of fans and the touchscreen.

The motherboard (figure 5) has dimensions $322\,\mathrm{mm} \times 160\,\mathrm{mm} \times 60\,\mathrm{mm}$, and is compatible with the mechanical design of the prototype HEV system. It runs on a nominal 24 V DC $(18 - 36\,\mathrm{V}$ DC) power supply, and contains local power regulation to generate the lower voltages required for the hosted components.

The power system is modular and uses commercially available components mountable on standard DIN-rails. The main stage is a stabilised power supply unit with battery management. This generates the 24 V DC for the motherboard and is compatible with $110 - 230\,\mathrm{VAC}$. It is connected with a UPS which will take over if mains power is lost and allows autonomous use of the system for more than 20 minutes. Configuration of the battery management is done by the Raspberry Pi through the motherboard via a specific connector and cable to the UPS. Connection to the mains power is with an IEC (International Electrotechnical Commission) inlet filter and power is turned on and off with a rotary switch. The DIN-rail mounting is compatible with the mechanical design of the prototype HEV system.

3.5 Control and User Interface

3.5.1 Control

The control concept, based around the embedded controller receiving the signals from the sensors and valves, and the touchscreen interface to the clinician, is illustrated in figure 6. The control software is implemented directly on the embedded controller, which fully controls the ventilator operation. The controller uses feedback from sensors to control valves in order to achieve the desired pressures.

The general overview of the control system and the user interface is shown in figure 7.

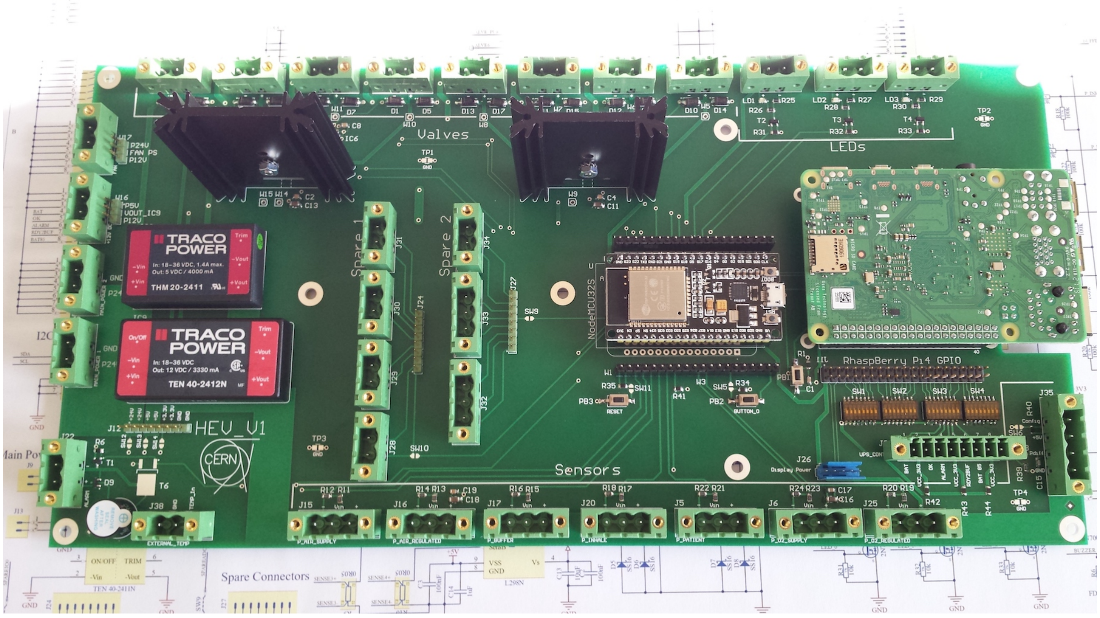


Figure 5: Photograph of the PCB motherboard as implemented for the HEV prototype.

Central to the operation and monitoring is the Microcontroller. An ESP32 Microcontroller chip has been chosen at the time of prototyping, due to its high availability and low cost. This is a dual core, single process Arduino compatible controller with additional CPU power and memory. It runs with no operating system which can be an advantage from the point of view of reducing complexity and promoting stability. Several alternatives exist and could be used depending on local availability in different geographical locations. All of the primary functions corresponding to the breathing function of the patient are controlled by the Microcontroller. The Microcontroller connects to the electronics to read the pressures, flows and temperatures. It controls the opening and closing of the valves and provides a simple status/warning via a speaker and a set of LEDs. Interconnect electronics is provided via the intermediate PCB (not shown) which steps up and down voltages where necessary.

The Microcontroller is connected to a Raspberry Pi which handles communications to the touch screen, WiFi and Ethernet, and controls the displays. If the communications are interrupted, the ventilator continues to run normally. A web server is also provided such that display information can be seen remotely (although this is not considered an essential function for operation).

3.5.2 Microcontroller functionality

The Microcontroller software design consists of three distinct threads of processing:

- A breathing loop responsible for operating the valves in a manner corresponding to the ventilator modes, such as PC–A/C or PC-PSV, as described in section 2.1.

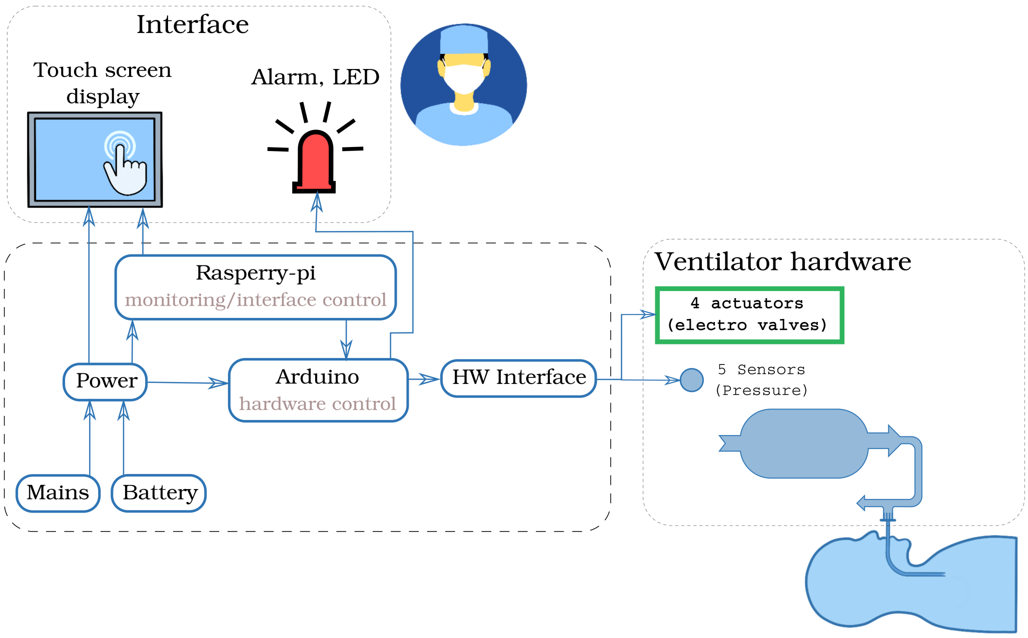


Figure 6: Conceptual layout of the controls and user interface.

- A User Interface (UI) loop for relaying the current status and readings to the Raspberry Pi, and for accepting commands for setting modes and parameters.
- A Safety Loop which is responsible for raising alarms when patient or system readings deviate from acceptable limits.

For all loops a finite state machine is employed to ensure operation is achieved in a repeatable and deterministic manner. The state machines operate semi-independently, with defined interfaces for information to pass between them. For example, this allows the Safety Loop to know both the current parameters set by the UI loop, and the current breathing loop ventilation mode. This allows each state machine to be tested individually by expressly checking the state response to each input at the interface.

The Microcontroller design allows the breathing and safety loops to continue to operate in the event of loss of connection to the Raspberry Pi via the UI loop. Visual and audible alarms signify a loss of connection but breathing and safety functions continue whilst the connection is being re-established. The lack of an underlying operating system in the Microcontroller reduces the probability of system lock-up/failure caused by other processes running in the background.

3.5.3 Raspberry Pi functionality

The Raspberry Pi software is designed to display the readings from the Microcontroller in a UI on the touchscreen. Both the webserver and the native displays receive the same information.

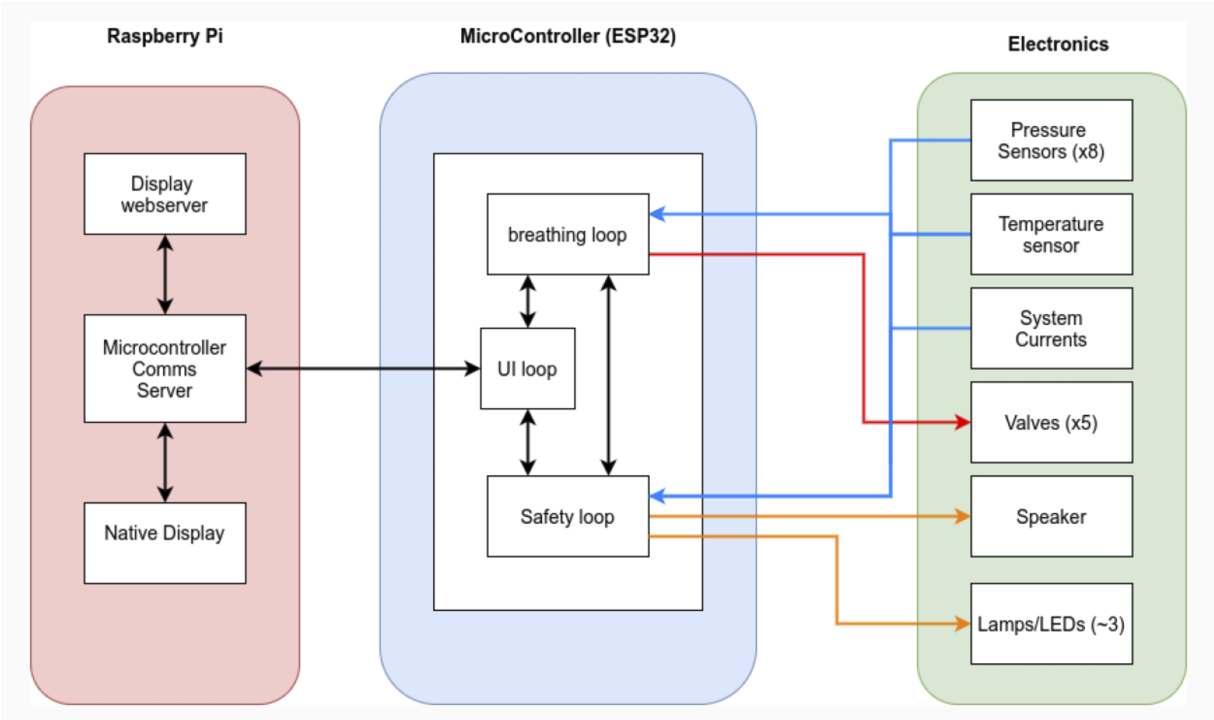


Figure 7: Diagram of the location of the different elements of hardware and software in the two computers (Microcontroller and Raspberry Pi) and the circuit board attached to the Microcontroller.

Both can be used to send control messages to the Microcontroller. However, in the case of the webserver, the controls may be disabled for reasons of security and safety. Both interfaces employ the same layout, such that they appear as similar as possible. Following the advice of clinicians, the user interface has been designed based on the following concepts:

- Clear text, symbols and graphs which can be seen from the end of a hospital bed through PPE.
- Neutral colours when all is OK, and by contrast, flashing indicators and messages when there are alarms. In general, safe and unsafe status should be easily distinguished.
- Screen locking/unlocking feature (with a timeout) to prevent accidental touchscreen presses.
- Confirmation of all parameter changes (two parameter changes would require two separate confirmations).
- Setting and reading back information (so that they can be compared).
- Simple navigation: no setting/parameter should be more than two clicks away. Normal operation should be separated from calibration/expert testing, to maintain an uncluttered interface.

- Interface should be touchscreen friendly: items should be placed far enough apart to minimise accidental mis-clicks.
- Familiarity: designed to look familiar to clinicians, similar to already existing ventilator interfaces.

3.5.4 Communication

Communication between the Microcontroller and the Raspberry Pi is achieved using the High-level Data Link Control Protocol (HDLC) [ISO 13239:2002]. A subset of the full protocol functionality was implemented, including supervisory (ACK/NACK frames) and information (INFO) frames. Information frames are subdivided into Data, Command and Alarm subtypes. Error detection is provided by a checksum (CRC16-CITT). More details are available in [30]. The protocol implementation is provided as a set of libraries on both the Microcontroller and the Raspberry Pi.

On the Raspberry Pi the protocol library is employed by a dataserver which broadcasts the data and alarm messages from the Microcontroller to the user interface processes. Similarly, Command messages are relayed from the user interfaces in the reverse direction. This is the expected method of operation, however, any message type can be sent in either direction. JSON is used as the communication format for messages from the dataserver to the user interfaces. These messages are built from the data, command, and alarms formats defined in the HDLC protocol library such that duplicate definitions are not required in JSON. More detail on the dataserver is given in [31].

3.5.5 Software Practice

The software development has been done in a robust and flexible manner, prioritising considerations of component failure from the start. The development team has been organised into pairs of developers working in the same domains, with strong familiarity with the code such that no individual is irreplaceable. Readability and simplicity of code has been favoured over complexity, for ease of bug tracing and testing. For the prototype, the guidelines of IEC 62304 have been used, and internal checking/assertions are included in the software. The communication protocols follow the High Level Data Link Control recommendations from ISO/IEC 13239:2002. In particular, data transmission is done with acknowledgement, and checksums are performed to confirm the data integrity. While the software system in place for the prototype is not yet fully qualified to ISO standards, the structure has been set up in such a way as to show that this is possible.

3.5.6 User Interface

The User Interface (UI) is designed to be familiar and readily usable in a medical environment, follows industry standards and conventions, and respects the regulatory guidelines. Two interfaces are be provided: a Native User interface and a Web User Interface. The Native User Interface runs on the touchscreen integrated into the ventilator unit. It continuously displays settings, measurements, patient waveforms and accepts instructions from the clinician. The displays are implemented in large clear fonts that can be seen easily from a distance of about 2 m away. Furthermore, the buttons or control elements are spaced well enough apart to accommodate thumb sized control. Figure 8 contains examples of the Native UI display. Remote access is provided via the Web UI, accessible via WiFi or Ethernet connection to the ventilator on computer screens or mobile devices. In this way the data of one or more patients can be collected and displayed at the nurses' station. This also opens up the possibility of remote consulting, which can be very useful for training or patient management in remote settings. The web interface can be configured so that full control, partial control, or no control is possible remotely. Access rights will be configurable as required. A Web UI example can be found in figure 9) The Native UI is automatically displayed on power up of the device. At start-up, a mode selection screen allows for selection of one of the modes described in section 2.1. Associated with each mode are the set parameters to be chosen for that mode. The parameter settings can be revisited at any time during operation. Each setting change requires a confirmation from the user. The UIs are designed with safety in mind. A locking feature of the Native UI prevents accidental touches. Inactivity will activate automatic locking. Locking can be enabled/disabled by holding the lock button for 3 seconds. Visual feedback is given during the locking/unlocking procedure. Indicators for the power source (mains or battery) are given on the UI, including residual charge indicators (e.g. charge $< 85\%$, 30 mins remaining).

Figure 8: Example displays of the Native User Interface. Numbers and graphs are purely indicative.

A "homepage" is shown for the current ventilation mode with the most important settings, parameters, waveforms and control buttons. Similar to the locking feature, the Native UI defaults to presenting the homepage after a period of inactivity by the user. A selection is provided for time ranges of the waveforms (i.e., showing the last 5, 15, 30 or 60 seconds). Historical data are recorded for up to 10 days. Encryption will be an option for stored patient data. Settings (or target values) and measured values are clearly distinguishable in the UI. The HEV Graphical User Interface is organised to respect all of the requirements listed in [21], [23], [24] on both implementations, and includes the following items on the homepage:

- Ventilation mode.

- Working inhalation pressure setting.
- Respiratory rate setting.
- Inhalation time setting.
- FIO_2 setting.
- Monitored plateau pressure.
- Monitored respiratory rate.
- Monitored PEEP.
- Monitored FIO_2.
- Monitored exhaled tidal volume.
- Monitored exhaled minute volume.

The following items are plotted as a continuous graphical display, with an option given to pause the display should the clinician wish to study the waveforms more closely.

- Volume vs time.
- Pressure vs time.
- Flow vs time.

Alarms are displayed in an intuitive way at the top of the screen at all times. A dedicated alarm page gives more detail on the alarms. An ordered list of the last ten alarms is shown, ordered by alarm priority, and current and historical alarms are easily distinguishable. It is possible to reset or silence alarms (for a period of time). The user interfaces match the "traffic light" lamps in terms of on-screen visualisation.

Finally, more technical details of the internal operation and calibration of the ventilator are provided on a separate "expert" page. A visual warning indicates that this is not for normal usage. The aim of the prototyping was to put in place all underlying software flexibility to be able to freely implement the desired UI. Before final manufacture a full useability study will be performed in order to optimise the UI.

4 HEV Prototypes

Three prototypes were constructed with an identical mechanical design in order to allow work in parallel on different aspects of the design implementation. The prototypes (figure 10) are designed as standalone units, each with two separate compartments, pneumatic and electrical, separated by a support plate. The mechanical description below as well as the functional tests described were all performed on these prototypes.

Figure 9: Example displays of the Web User Interface. Numbers and graphs are purely indicative.

In parallel, a fourth prototype, illustrated in figure 11, was developed at the Galician Institute of High Energy Physics at the University of Santiago de Compostela (IGFAE/USC) with a two-fold goal. Firstly, we wished to demonstrate that such a respirator could be developed easily in other places than CERN. Secondly, having successfully reproduced the design within a short timescale, we were able to launch additional development and contributions to the control software. In addition, this triggered discussions with local hospitals and physicians, further improving the design of the device.

4.1 Prototype mechanical design

The prototype mechanical design was created with the aim of enabling the production of a fully specified and functional ventilator. Additional space has been deliberately added, so that the working prototypes could be built as quickly as possible, and to ensure that there will be no problem to interchange components for others with similar properties if this is necessary for full medical compatibility. The resulting design iis a unit with approximate dimensions $500 \times 500 \times 1350$ mm. Note that due to this choice, the machine is larger than necessary and not in its optimal packaging; this can be easily optimised at the manufacturing stage.

The simple mechanical design was chosen to feature standard components, not commonly used in respiratory equipment. This would enable it to be produced locally by adapting the parts to the availability of local suppliers. The ventilator is simple to construct and requires little or no machining. A list of the typical parts required includes:

- A proportional solenoid inhale valve with 8 mm internal diameter.
- Two fast acting solenoid valves for inlet of air and oxygen into the buffer volume.
- A solenoid valve for purging. It could be low volume, that in case of over-pressure only some air is released.
- An exhale ON/OFF valve with large diameter internal orifice.
- Five check valves.
- Two pressure relief valves, one for the patient circuit and one for the buffer.

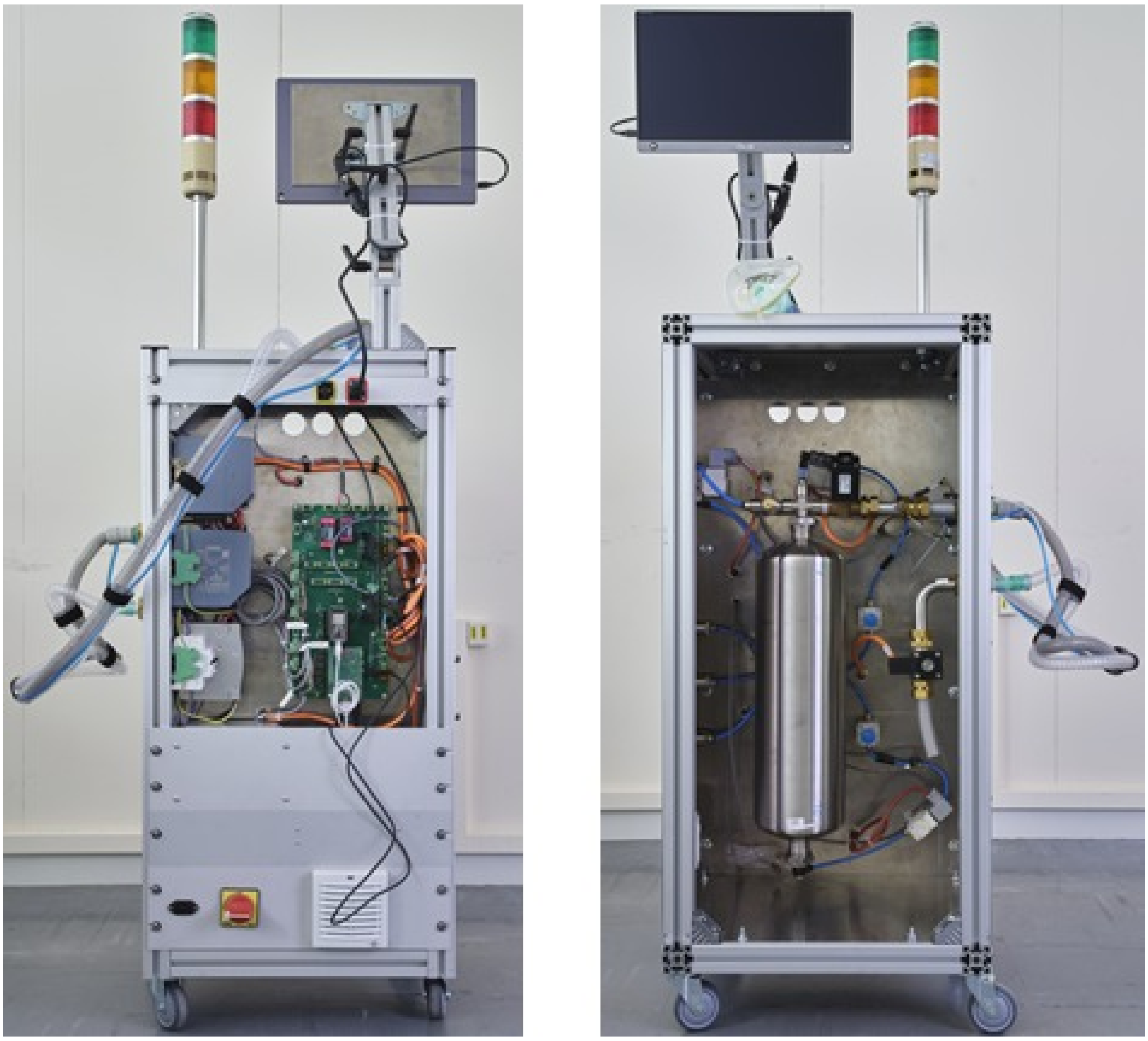


Figure 10: HEV prototypes as tested for full functionality.

- A 10 L buffer container.
- Two pressure regulators, able to operate up to a maximum pressure of 10 bar, one for oxygen and one for air.
- Eight pressure sensors.
- A bidirectional differential pressure sensor for the flow measurement.
- Two temperature sensors.

In addition to these parts, there is a list of required and optional accessories, described in section 3.3, such as the PEEP valve, breathing circuit, filters and humidifiers.

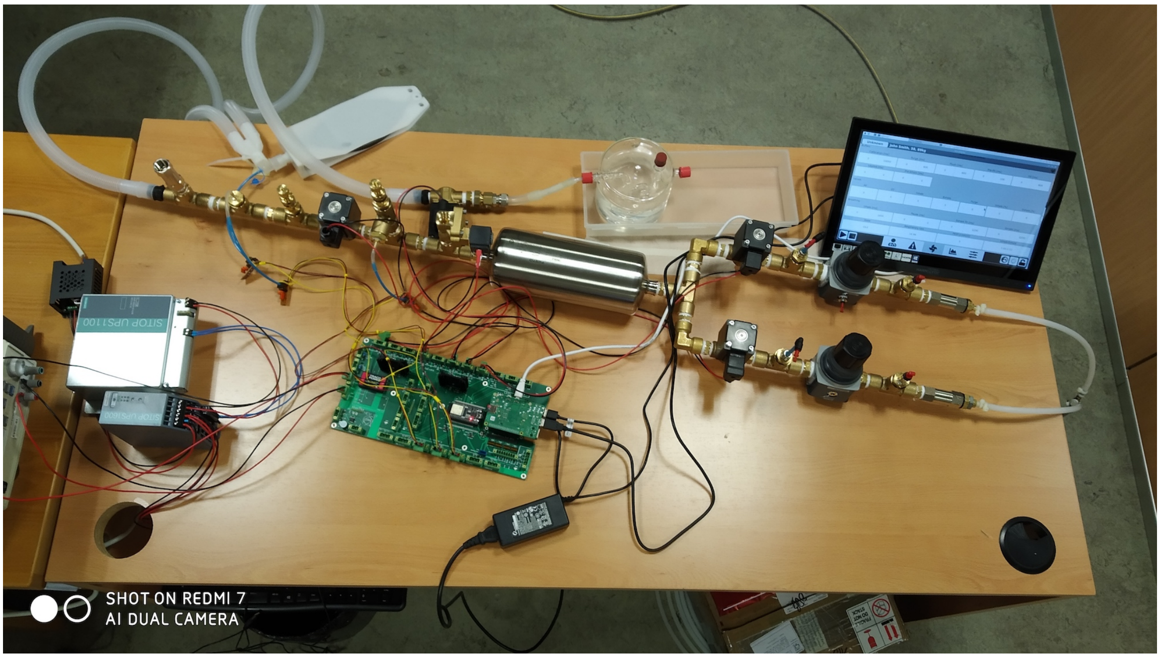


Figure 11: Functional table top prototype produced for collaboration and development purposes.

Functionally, the prototype designs follow the concept described above. The resulting cabinet is mounted on wheels, can easily be moved by one person, is very stable, and provides a convenient surface to mount the display at head height. The cabinet is closed with doors, so that easy access for cleaning is possible. The cabinet is subdivided internally into two separate compartments, front and back, housing the pneumatic and electronics components separately, which provides protection against explosion risk from potential oxygen leaks. The air tubes connect through a standard bulkhead thread connector on the outside. In this way it is easily replaceable to match hospital connection standards around the world.

4.2 Alternative mechanical design

The prototypes have been built with a deliberately large amount of space to allow rapid development and exchange of parts. The final ergonomics of the HEV may look quite different depending on the requirements in the region of deployment and the accessories included. The HEV collaboration has provided two different mechanical designs to which the HEV design could be adapted to fulfill different needs, which are illustrated in figure 12. Option A is more compact and can be mounted on wheels or a trolley. Space is provided to support oxygen and compressed air bottles, as well as the turbine system, such that the entire system and accessories can be provided as one integrated unit, which can be desirable for certain geographical locations. Option B is a still more compact and light version, for which the total dimensions are comparable to existing commercial ventilators and the weight is targetted to be around 25 kg. The touch screen can be folded away for transport and the ventilator easily

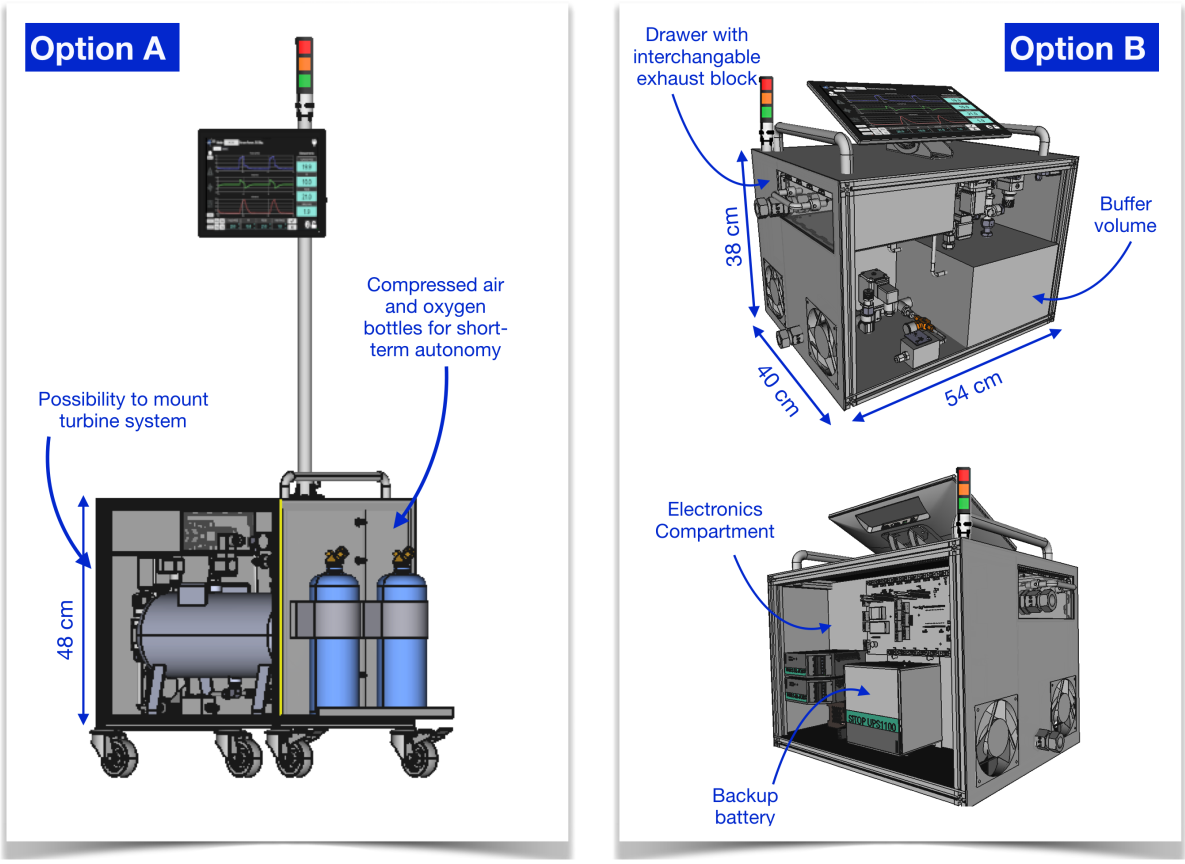


Figure 12: Potential alternative mechanical designs with an identical functionality to the HEV prototypes.

mounted on a trolley. Both options are identical in functionality to the HEV prototypes which have been built and tested.

5 Prototype test results

5.1 Setup

The HEV prototype is tested by connecting it to a lung simulator through a coaxial breathing circuit set, as shown in figure 13. The coaxial breathing circuit set[2] has a differential pressure based flow sensor[3] whose readout is embedded in HEV (`dP_patient`). Alternatives to this sensor are available with other manufacturers [4], and a specific breathing circut for HEV in order to decrease reliance on the existing supply chain can also be an option. The lung

[2]Hamilton PN 260128.
[3]Hamilton PN 281637.
[4]Intersurgical P/N 2072000

simulator is a TestChest light [5][32], which allows for the change the mechanical parameters of the lung (resistance of the airways, compliance of the lung) as well as generates adjustable spontaneous breathing. The setup is equivalent to the MHRA [21] and ISO 80601-2-12 figure 201.102 [33].

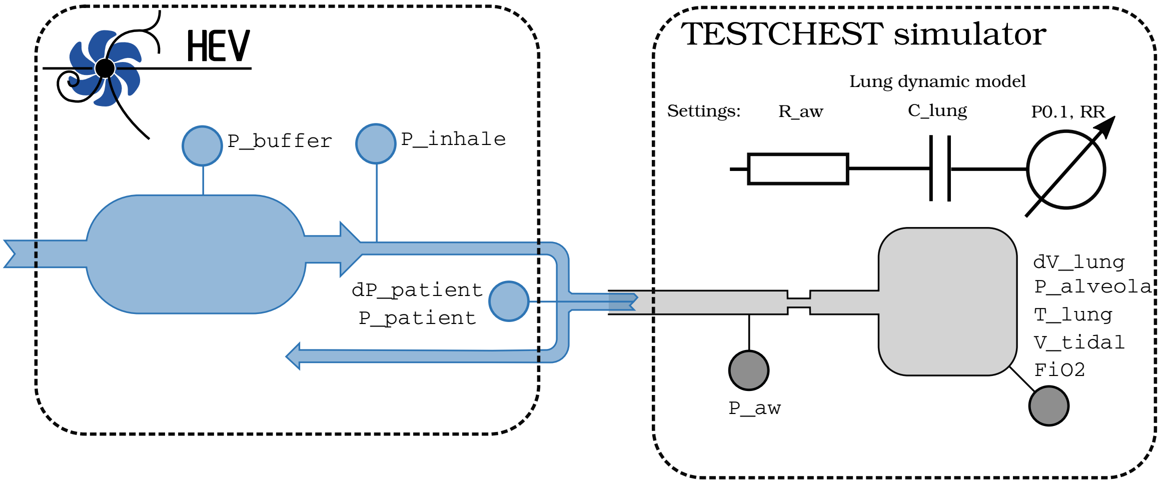


Figure 13: Test setup schematics.

5.2 Target Pressure Performance

As discussed in section 3.1, when the inhale starts, the proportional valve `valve_inhale` opens in a controlled way in order to reach the target pressure in a given time. This is done through a PID controller that monitors the `P_inhale` pressure measurement as input for the inhale valve opening control. As illustrated in figure 14, any pressure level within the setting range can be reached and the system stays stably locked to this value with an uncertainty below 3% in most of the cases.

Further tests will be performed in presence of leaks, but even with the highest leak setting of the TestChest, no difference in the pressure profile is visible.

The time needed to reach the maximum of pressurisation can be tuned between the fastest setting of about 50 ms and about 300 ms, as illustrated in figure 15. This gives the clinician the flexibility to use the fastest response of the ventilator, typically for intubated, sedated patients, through to slower rise times which may be useful at other periods of recovery.

In figures 16, the response of the HEV ventilator to patients with compliance varying from 10 to 100 ml/mbar and with resistance from 5 to 50 mbar/l/s is shown. The PEEP values range from 5 to 15 cm H_2O and target pressure up to 45 cm H_2O are tested. The behaviour of the ventilator is illustrated in figure 16.

When the patients shows no airway resistance or when their lung compliance is low enough, the flow can fully develop during the allocated inhalation time such that the tidal

[5]Organis GMBH, Landquart, www.organis-gmbh.com

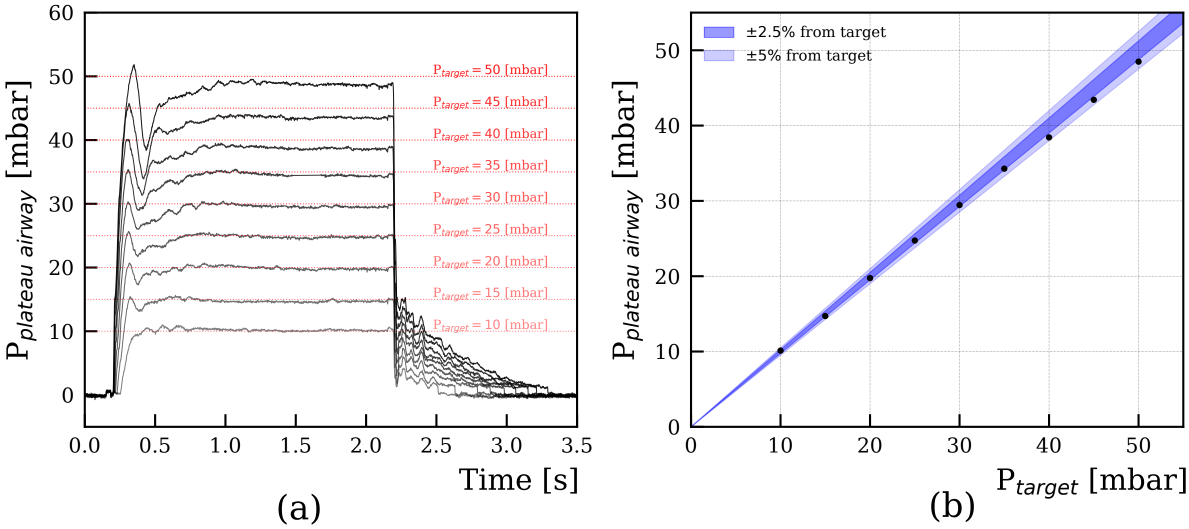


Figure 14: Left: pressurisation of a 50 ml/mbar compliance, 5 mbar/l/sec resistance patient with various target pressures. The inhalation time is set to 1.5 seconds, and is followed by a pause of 0.5 seconds. Right: the deviation of the inhale pressure from the target is computed during the pause.

volume can be computed from the product of the compliance and the differential pressure. The slowing down of the flow with the increased resistance and increased compliance is also well visible. Lower compliance patients could not be tested with this lung simulator but will be the subject of a dedicated test later; the performance is expected to be good.

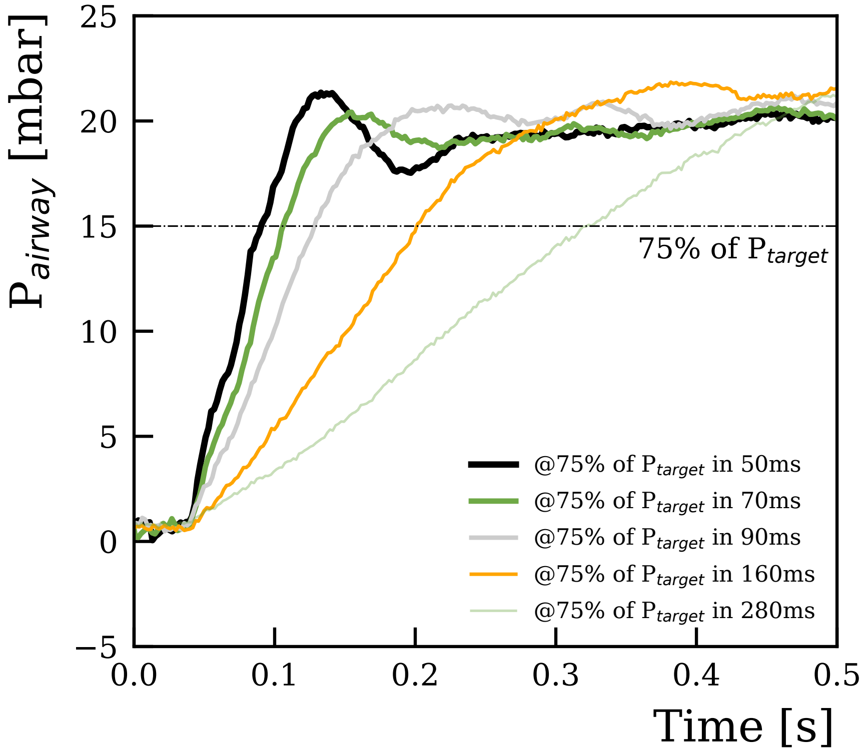


Figure 15: Rising edge of the inhalation for various rise time setting. The lung compliance here is 50 ml/mbar for a resistance of 5 mbar/l/s

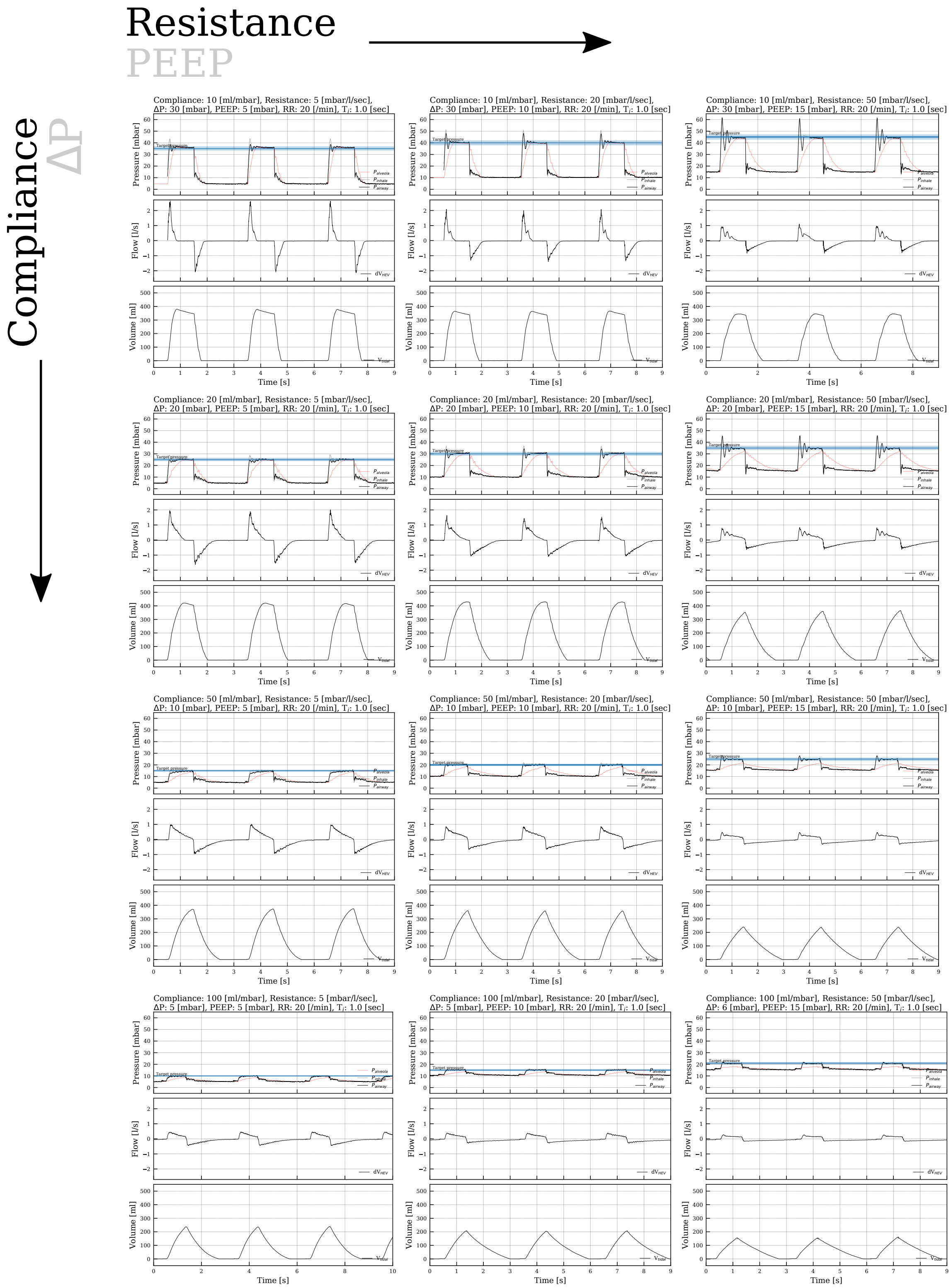


Figure 16: Pressure, flow and volume registered for patient configurations with increasing lung compliance and airway resistance. As expected, the flow is more peaked with reduced resistance and compliance.

5.3 Inhale Trigger Peformance

The inhale and exhale trigger are essential to guarantee the comfort of the patients and to ensure fast recovery time. The trigger functionalities were developed with this aspect in mind and particular effort was made to qualify them.

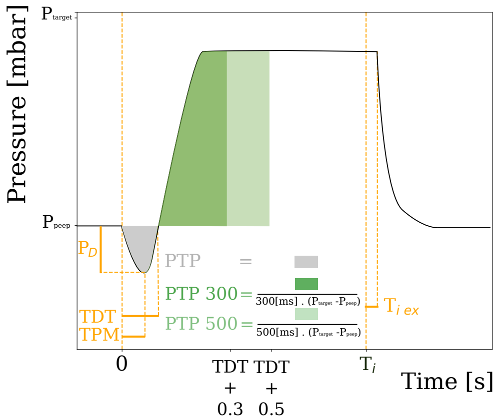


Figure 17: Measured parameters to qualify the inhalation trigger. The variables used in this work are based on the work presented in [34] and [35]

Inhale trigger algorithm description Whenever the patient initiates a breath, the proximal flow sensor sees an increased flow and an under-pressure. The increase in flow is used as a trigger for the inhale sequence. The inhale trigger algorithm works in the following way. Whenever the flow reaches 10% of the maximal exhale flow, the window allowing for a inhale trigger is opened. This condition is by definition always met when the patient initiates an inhalation. The expected flow at a given instant is computed by a linear regression from a given time window before that point. It provides the baseline, to which the measured flow is compared. If the measured flow, corrected for the baseline, is above a threshold that can range from 0.2 l/min to 20 l/min, then the inhalation starts. Lower thresholds are sensitive to noise, in particular that induced by the heart-lung interactions. This threshold is set by the clinician.

Inhale trigger qualification To qualify the performance of the inhale trigger, the variables defined in [34] are used. Figure 17 illustrates those variables: the Time to Minimum Pressure, TPM, is defined as the time between the beginning of the inhalation effort and the minimum value measured reached by `P_patient`, and the Trigger Delay Time TDT is defined as the time between the beginning of the inhalation effort and the moment the pressure returns to zero. Values of TPM in commercial ventilators typically vary between 50 and 150 ms depending on the inhalation effort, see for example [34], while TDT, which should ideally be below 150 ms so as not to be felt by the patient, varies for commercial ventilators in practice between 90 and 250 ms [34]. The pressure-time product during trigger, PTP, is represented in figure 17 by the grey area and represents the effort until the pressure is effective. It ranges

$P_{0.1}$ [mbar]	2						4					
PEEP [mbar]	0			5			0			5		
ΔP [mbar]	10	15	20	10	15	20	10	15	20	10	15	20
TPM [ms]	100	99	79	121	104	74	118	69	74	95	95	72
TDT [ms]	130	105	101	139	112	98	150	105	98	135	107	96
PD [mbar]	2	1.9	1.6	1.9	2	1.7	4.7	3.7	4	4.1	3.9	2.6
PTP [mbar]	0.09	0.03	0.02	0.12	0.08	0.07	0.24	0.15	0.13	0.28	0.2	0.18
PTP300 [%]	28.8	41.4	43.8	29.1	38.6	46.4	24.8	36.6	41.0	29.8	36.1	43.5
PTP500 [%]	39	53.8	55.6	43.7	51.2	58.8	34.6	44.6	49.1	38.2	44.	52.2

Table 3: Results of the inhale trigger characterisation.

	No leaks	Small	Medium	Large
TPM [ms]	74	72	70	70
TDT [ms]	98	94	94	94
PD [mbar]	4	4	3.7	3.4
PTP [mbar.s]	0.13	0.13	0.11	0.1
PTP300 [%]	41	40.5	41.4	41.1
PTP500 [%]	49.1	48.5	49.3	49.1

Table 4: Inhale trigger sensitivity to different leakage level for lungs with compliance of 50 mbar/ml and resistance of 5 mbar/l/s

from 0.02 to 0.3 mbar.s in commercial ventilator [34]. The ideal PTP300 percentage and ideal PTP500 percentage,referred to as PTP300 and PTP500 in the following, are respectively the ratio of the pressure integral over the 300 ms (500 ms) following the trigger delay (the area of the green regions in figure 17) and the ideal PTP at 300 and 500 ms. It should be as large as possible, with typical commercial ventilators exhibiting values between 10 and 50% (20 and 75%) for PTP300 (PTP500) depending on the inhalation effort [34].

To test the inhalation trigger, the same lung parameters, pressurisation parameters and inhalatory effort than in [34] were used. TestChest was set with a compliance of 50 mbar/ml and a resistance of 5 mbar/l/s, the breath cycles at 12 respiration per minutes consisted in a 1 s inspiration with a constant inspiratory flow giving a occlusion pressure at 100 ms ($P_{0.1}$) of $2\,\mathrm{cm\,H_2O}$ (low effort) and $4\,\mathrm{cm\,H_2O}$ (high effort) and several pressurisation parameter: ΔP of 10,15 and 20 mbar with a PEEP of 0 and 5 mbar. The inhale trigger threshold is set to 0.5 L/min.

Table 3 summarises the results of the inhalation trigger qualification. Comparing the results to the ventilator studied in [34], HEV inhale trigger appears to perform very well.

The trigger response is studied again in presence of leaks. Only one set of pressurisation is used, with PEEP at $0\,\mathrm{cm\,H_2O}$ and ΔP=$20\,\mathrm{cm\,H_2O}$. Four settings are tried: no leaks, weak, medium and strong leaks as set by TestChest, and results are reported in table 4. No significant difference is observed.

Exhale trigger algorithm description The implementation of the exhale trigger is more straightforward. When the inhale flow decreases down to a fraction of the maximum inhale flow, the exhale phase is triggered.

Exhale trigger qualification In order to qualify the performance of the exhale trigger, $T_{I_{ex}}$ defined as the duration of pressurisation by the ventilator in excess with respect to T_I, which represents the true duration of inhalation by the patient, is measured as in [35]. This is illustrated in figure 17. It is not a property of the ventilator per se, but by appropriate tuning of the exhalation trigger it should be possible to bring it to below 100 ms. Values of $T_{I_{ex}}$ below 10 ms are achieved.

5.4 Oxygen mixing test

Because the mixing is performed in the buffer in a phase which is physically uncorrelated to the patient the breath cycle (i.e. during patient exhalation when the buffer is disconnected from the patient), we perform the mixing test independently from the test of the ventilator modes.

The O_2 concentration in the buffer is controlled by changing the relative opening time of `valve_O2_in` and `valve_Air_in`. For a given O_2 concentration setting, the opening times can be computed. After stabilisation of the measured O_2 concentration (FIO_2) in the lung simulator, the FIO_2 is compared to the set value. Further control will be introduced in the future by regulating the opening time from a feedback of the measured O_2 concentration in the buffer.

The measured FIO_2 as function of the expected O_2 percentage as calculated from the relative time opening of `valve_O2_in` and `valve_Air_in` is shown in figure 18. The measured FIO_2 in the lung is within 5% of the set value, which is an acceptable performance. Further tuning to the valve opening time can be done in order to correct for the small non-linearity in the response.

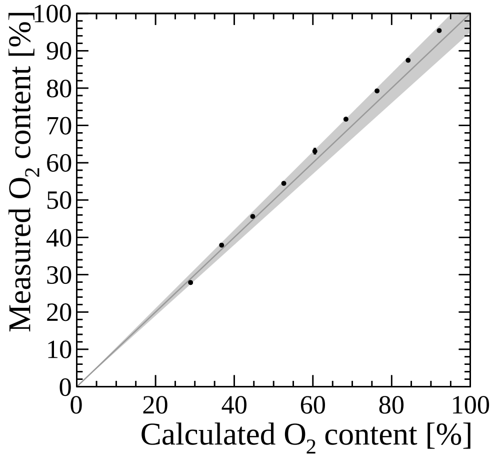


Figure 18: Measured FIO_2 as function of the expected O_2 percentage as calculated from the relative time opening of `valve_O2_in` and `valve_Air_in`. The grey band represents the region within $\pm 5\%$ of the set value.

6 Conclusion

HEV has been developed to be a high-quality, low-cost ventilator, suitable for use in a hospital setting. The design is intended for easy and fast manufacturing that can be performed in a decentralised way with affordable and readily available parts. The central concept of the design with a gas accumulator gives many advantages in terms of robustness, safety, affordability and precise ventilation behaviour. The electrical design is conceived in a modular way for quick prototyping and deployment, which facilitates mass production. The design is intended to be robust and adaptable for a wide range of geographical deployment, including in regions where compressed air may not be readily available and a turbine alternative can be used. Three prototypes have been manufactured and have been tested in situ under clinical supervision with the full range of simulated patients defined in the MHRA specifications and the results are presented in this paper. HEV has also been tested at the ETH Zurich Chair of Product Development and Engineering Design Ventilator test rig. In pressure control mode HEV accurately achieves the target pressures, with fast rise time which is tuneable to slower times on clinician request. Special attention has been paid to the inhale and exhale triggers to optimise patient comfort. The inhale trigger, based on the flow measurement, accurately reacts to the patient effort, with short rise times and excellent PTP values. The system displays and monitoring use concepts familiar to particle physics such as the possibility for remote monitoring from screens or mobile devices, data logging for quality control and performance monitoring, and remote training.

As far as production is concerned, it is foreseen, on the one hand, to enable this through providing partner academic institutions with the detailed design for these institutions to follow up in accordance with local possibilities and standards; on the other hand, directly through industry, non-governmental, governmental and international organizations, such as the World Health Organisation (WHO), for which purpose discussions are ongoing and contacts have been established with potential partners. Every effort is being made to finalise the design of the HEV in accordance with the state-of-the-art best practices and standards, but the formal certification process should of course be initiated by the parties that decide to place this device on the market. The hardware and software design has been done in a flexible way which allows the development of different modes of operation, for instance volume control modes which in principle can be developed and applied as a firmware update. In addition, the HEV prototypes can be used as a testbench to quickly implement and test novel algorithms or hardware updates, and in this way could provide a fresh avenue for medical research.

Acknowledgements

We express our gratitude to Giovanni Anelli, Amy Bilton, Paola Catapano, Manuela Cirilli, Paolo Chiggiato, Beniamino Di Girolamo, Doris Forkel-Wirth, Benjamin Frisch, Samuel Herzog, Lucie Pocha, Javier Serrano, Erik Van Der Bij, Jens Vigen and Maarten Wilbers from CERN, Geneva, Switzerland, the CERN against COVID-19 task force, and to Simon Cohen from Monash Children's Hospital, Melbourne, Australia, for many illuminating discussions and much practical support.

References

[1] L. Rosenbaum, "Facing Covid-19 in Italy – Ethics, Logistics, and Therapeutics on the Epidemic's Front Line," *New England Journal of Medicine*, vol. 382, no. 20, pp. 1873–1875, 2020, PMID: 32187459. DOI: `10.1056/NEJMp2005492`. eprint: `https://doi.org/10.1056/NEJMp2005492`.

[2] S. Kliff, A. Satariano, J. Silver-Greenberg and N. Kulish, "There aren't enough ventilators to cope with the coronavirus," [Online]. Available: `https://www.nytimes.com/2020/03/18/business/coronavirus-ventilator-shortage.html`.

[3] M. L. Ranney, V. Griffeth, and A. K. Jha, "Critical Supply Shortages – The Need for Ventilators and Personal Protective Equipment during the Covid-19 Pandemic," *New England Journal of Medicine*, vol. 382, no. 18, e41, 2020. DOI: `10.1056/NEJMp2006141`. eprint: `https://doi.org/10.1056/NEJMp2006141`.

[4] M. Malta, A. W. Rimoin, N. A. Hoff, and S. A. Strathdee, "The 2019-ncov pandemic in the global south: A tsunami ahead," *EClinicalMedicine*, p. 100 384, 2020, ISSN: 2589-5370. DOI: `https://doi.org/10.1016/j.eclinm.2020.100384`. [Online]. Available: `http://www.sciencedirect.com/science/article/pii/S2589537020301280`.

[5] E. J. Emanuel, G. Persad, R. Upshur, B. Thome, M. Parker, A. Glickman, C. Zhang, C. Boyle, M. Smith, and J. P. Phillips, "Fair allocation of scarce medical resources in the time of covid-19," *New England Journal of Medicine*, vol. 382, no. 21, pp. 2049–2055, 2020. DOI: `10.1056/NEJMsb2005114`. eprint: `https://doi.org/10.1056/NEJMsb2005114`.

[6] T. Netland, "A better answer to the ventilator shortage as the pandemic rages on," [Online]. Available: `https://www.weforum.org/agenda/2020/04/covid-19-ventilator-shortage-manufacturing-solution/`.

[7] Centre de formation Urgences-santé, "Oxylator EMX^{tm}," [Online]. Available: `https://www.urgences-sante.qc.ca/wp-content/uploads/2014/03/Oxylator-EMX.pdf`.

[8] MIT E-Vent, "MIT Emergency Ventilator (E-Vent) Project," [Online]. Available: `https://e-vent.mit.edu`.

[9] C. Galbiati *et al.*, “Mechanical Ventilator Milano (MVM): A Novel Mechanical Ventilator Designed for Mass Scale Production in Response to the COVID-19 Pandemics,” 2020. arXiv: 2003.10405 [physics.med-ph]. [Online]. Available: https://arxiv.org/abs/2003.10405.

[10] Mechanical Ventilator Milano, “Mechanical Ventilator Milano,” [Online]. Available: www.mvm.care.

[11] Take a Breather Team, “Attachment to Automate the Bag Valve Mask in Low Resource Settings,” [Online]. Available: http://oedk.rice.edu/Sys/PublicProfile/47585242/1063096.

[12] VENTILAID Team, “Open-Source solutions to fight the Coronavirus pandemic,” [Online]. Available: https://www.ventilaid.org.

[13] OxVent Team, “OxVent,” [Online]. Available: https://oxvent.org.

[14] RTBF avec Belga, “Coronavirus: des ingénieurs belges développent un prototype de respirateur,” [Online]. Available: https://www.rtbf.be/info/belgique/detail_coronavirus-des-ingenieurs-belges-developpent-un-prototype-de-respirateur?id=10463991.

[15] R. e. a. Read, “COVID-19 Ventilator Projects and Resources with FAQs,” [Online]. Available: https://github.com/PubInv/covid19-vent-list.

[16] JustActions, “ The Missing Piece. Why Continued Neglect of Pneumonia Threatens the Achievement of Health Goals.,” [Online]. Available: https://stoppneumonia.org/wp-content/uploads/2018/11/The-Missing-Piece_-0611_Spread.pdf.

[17] H. J. Zar, S. A. Madhi, S. J. Aston, and S. B. Gordon, “Pneumonia in low and middle income countries: Progress and challenges,” *Thorax*, vol. 68, no. 11, pp. 1052–1056, 2013, ISSN: 0040-6376. DOI: 10.1136/thoraxjnl-2013-204247. eprint: https://thorax.bmj.com/content/68/11/1052.full.pdf.

[18] S. J. Aston, “Pneumonia in the developing world: Characteristic features and approach to management,” *Respirology*, vol. 22, no. 7, pp. 1276–1287, 2017. DOI: 10.1111/resp.13112. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/resp.13112.

[19] J. Buytaert *et al.*, “The HEV Ventilator Proposal,” 2020. arXiv: 2004.00534 [physics.med-ph]. [Online]. Available: https://arxiv.org/abs/2004.00534.

[20] HEV project, “High Energy physics community proposes HEV ventilator to help combat COVID-19,” [Online]. Available: https://hev.web.cern.ch.

[21] UK Medicines and Healthcare products Regulatory Agency, *MHRA Specification for Rapidly Manufactured Ventilator System (RMVS), version 4.0*, 2020. [Online]. Available: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/879382/RMVS001_v4.pdf.

[22] EU, *Guidance on medical devices, active implantable medical devices and in vitro diagnostic medical devices in the COVID-19 context*, 2020. [Online]. Available: https://ec.europa.eu/docsroom/documents/40607?locale=en.

[23] Association for the Advancement of Medical Instrumentation, *Emergency use ventilator (EUV) design guidance, revision 1.2*, 2020. [Online]. Available: https://www.aami.org/docs/default-source/standardslibrary/200410_cr501-2020_rev1-2.pdf?sfvrsn=699e62b7_2.

[24] World Health Organization, *Technical specifications for invasive and non-invasive ventilators for COVID-19: Interim guidance, 15 April 2020*, Technical documents, 2020. [Online]. Available: https://apps.who.int/iris/handle/10665/331792.

[25] Clinical Experts Group, Hamilton Medical, "Driving pressure in ARDS patients," [Online]. Available: https://www.hamilton-medical.com/de/E-Learning-and-Education/Knowledge-Base/Knowledge-Base-Detail~2017-05-08~Driving-pressure-in-ARDS-patients~09161a01-6223-4668-8fa8-d906319d64f9~.html.

[26] M. B. Amato, M. O. Meade, A. S. Slutsky, L. Brochard, E. L. Costa, D. A. Schoenfeld, T. E. Stewart, M. Briel, D. Talmor, A. Mercat, J.-C. M. Richard, C. R. Carvalho, and R. G. Brower, "Driving pressure and survival in the acute respiratory distress syndrome," *New England Journal of Medicine*, vol. 372, no. 8, pp. 747–755, 2015, PMID: 25693014. DOI: 10.1056/NEJMsa1410639.

[27] Karin Deden, *Ventilation modes in intensive care*. [Online]. Available: https://www.draeger.com/Library/Content/rsp-new-nomenclature-ventilation-modes-icu-booklet-9066477-en.pdf.

[28] R. Inglis, E. Ayebale, and M. J. Schultz, *Optimizing respiratory management in resource-limited settings*, 10.1097/MCC.0000000000000568, 2019. [Online]. Available: https://journals.lww.com/co-criticalcare/Fulltext/2019/02000/Optimizing_respiratory_management_in.8.aspx.

[29] L. Brouwer, A. Hoedemaekers, and J. van der Hoeven, "Effect of inspiration rise time on work of breathing and patient comfort during pressure support ventilation," *Critical Care*, vol. 10, no. 1, P38, 2006. DOI: 10.1186/cc4385. [Online]. Available: https://doi.org/10.1186/cc4385.

[30] HEV project, *Communication protocol*, Mar. 2020. [Online]. Available: https://ohwr.org/project/hev/wikis/Software-communication-protocols-for-HEV/.

[31] ——, *Data server specification for the HEV*, Mar. 2020. [Online]. Available: https://ohwr.org/project/hev/wikis/Software-communication-protocols-for-HEV/.

[32] J. Brunner, *Testchest physiological model*. Dec. 2017, ISBN: 978-3-9524884-0-9.

[33] International Organization for Standardization, *Medical electrical equipment - Part 2-12: Particular requirements for basic safety and essential performance of critical care ventilators*, ISO 80601-2-12:2020, Feb. 2020. [Online]. Available: https://www.iso.org/obp/ui#iso:std:iso:80601:-2-12:ed-2:v1:en.

[34] C. Delgado, J. E. Romero, J. Puig, A. Izquierdo, C. Ferrando, F. J. Belda, and M. Soro, "Performance of the new turbine mid-level critical care ventilators," *Respiratory Care*, vol. 62, no. 1, pp. 34–41, 2017, ISSN: 0020-1324. DOI: `10.4187/respcare.04938`. eprint: `http://rc.rcjournal.com/content/62/1/34.full.pdf`. [Online]. Available: `http://rc.rcjournal.com/content/62/1/34`.

[35] L. Vignaux, D. Tassaux, and P. Jolliet, "Performance of noninvasive ventilation modes on icu ventilators during pressure support: A bench model study," *Intensive care medicine*, vol. 33, pp. 1444–51, Sep. 2007. DOI: `10.1007/s00134-007-0713-0`.

HEV collaboration

J. Buytaert[1,*], A. Abed Abud[1,2], P. Allport[26], A. Pazos Álvarez[11], K. Akiba[3], O. Augusto de Aguiar Francisco[1,4], A. Bay[10], F. Bernard[10], S. Baron[1], C. Bertella[1], J. Brunner[21], T. Bowcock[2], M. Buytaert-De Jode, W. Byczynski[1,8], R. De Carvalho[23], V. Coco[1], P. Collins[1,*], R. Collins[18], N. Dikic[1], N. Dousse[23], B. Dowd[16], R. Dumps[1], P. Durante[1], W. Fadel[1], S. Farry[2], A. Fernández Prieto[11], G. Flynn[16,25], V. Franco Lima[2], R. Frei[10], A. Gallas Torreira[11], R. Guida[1], K. Hennessy[2], A. Henriques[1], D. Hutchcroft[2], S. Ilic[7], A. Jevtic[7], C. Joram[1], K. Kapusniak[1], E. Lemos Cid[11], J. Lindner[9], R. Lindner[1], M. Milovanovic[1,6], S. Mico[1], J. Morant[1], M. Morel[1], G. Männel[20], D. Murray[4], I. Nasteva[5], N. Neufeld[1], I. Neuhold[1], F. Pardo-Sobrino López[19], E. Pérez Trigo[11], G. Pichel Jallas, E. Pilorz[1], L. Piquilloud[17], X. Pons[1], D. Reiner[13], C. Roosens[24], P. Rostalski[20], B. Schmidt[1], E. Saucet[23], F. Sanders[1], C. Sigaud[1], B. Schmidt[1], P. Schoettker[17] R. Schwemmer[1], H. Schindler[1], A. Sharma[1], P. Svihra[4], J. van Leemput[24], L. Vignaux[22], F. Vasey[1], H. Woonton[14,15], K. Wyllie[1].

[1]*European Organization for Nuclear Research (CERN), Geneva, Switzerland*
[2]*Oliver Lodge Laboratory, University of Liverpool, Liverpool, United Kingdom*
[3]*Nikhef National Institute for Subatomic Physics, Amsterdam, Netherlands*
[4]*Department of Physics and Astronomy, University of Manchester, Manchester, United Kingdom*
[5]*Universidade Federal do Rio de Janeiro (UFRJ), Rio de Janeiro, Brazil*
[6]*Deutsches Elektronen-Synchrotron (DESY), Platanenallee 6, 15738 Zeuthen, Germany*
[7]*University of Niš, Niš, Serbia*
[8]*Tadeusz Kosciuszko Cracow University of Technology, Cracow, Poland*
[9]*University of Applied Sciences Offenburg, Offenburg, Baden-Wuerttemberg, Germany*
[10]*Institute of Physics, Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland*
[11]*Instituto Galego de Física de Altas Enerxías (IGFAE), Universidade de Santiago de Compostela, Santiago de Compostela, Spain*
[13]*John Curtin School of Medical Research, Canberra, Australia*
[14]*Monash Health, Melbourne, Australia*
[15]*Dandenong Hospital, Melbourne, Australia*
[16]*Prince of Wales Hospital, New South Wales, Australia*
[17]*Centre Hospitalier Universitaire Vaudois, Lausanne, Switzerland*
[18]*College of Veterinary Medicine, Cornell University, Ithaca, NY, USA*
[19]*Anesthesiology-Reanimation and Pain Therapuetics Service, Lucus Augusti University Hospital, Lugo, Spain*
[20]*Institute for Electrical Engineering in Medicine, University of Lübeck, Lübeck, Germany*
[21]*Neosim AG, CH-7000 Chur, Switzerland*
[22]*Cardio-Respiratory Units, Hôpital de La Tour, Meyrin, Switzerland*
[23]*Hôpitaux Universitaires de Genève, Genève, Switzerland*
[24]*GZA hospitals, Antwerp, Belgium*
[25]*University of New South Wales, Sydney, Australia*
[26]*Particle Physics Group, School of Physics and Astronomy, University of Birmingham, United Kingdom*
[*]*Corresponding authors: Jan.Buytaert@cern.ch and Paula.Collins@cern.ch*