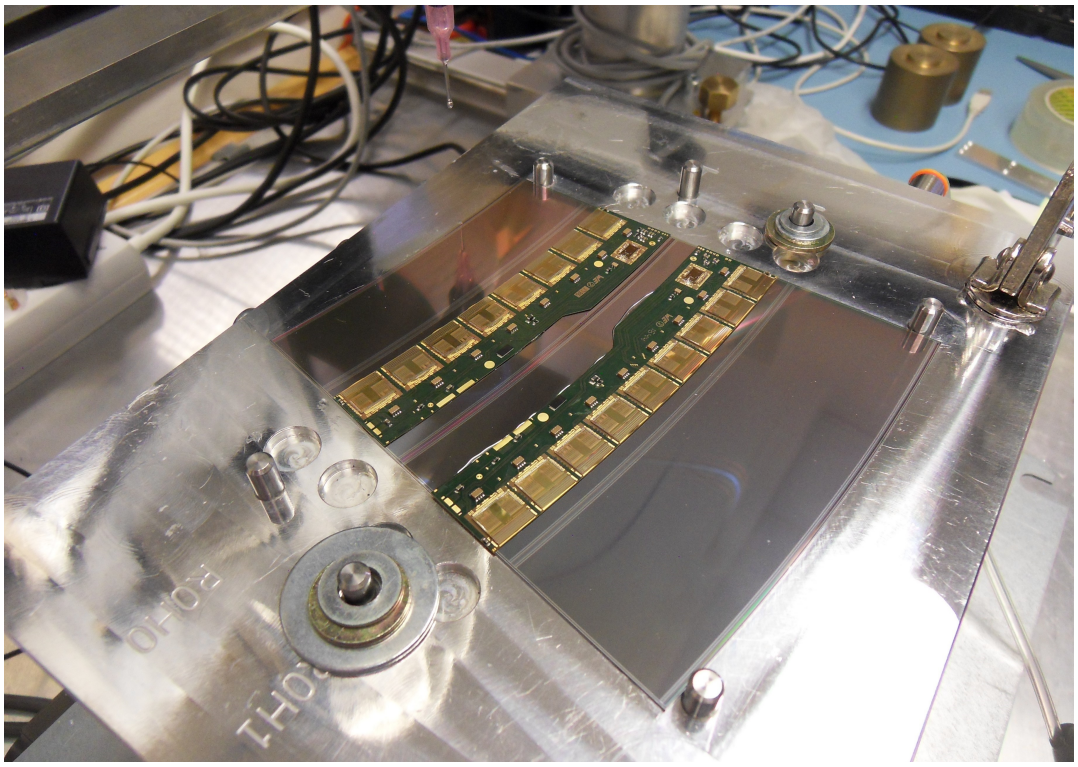
UNIVERSITY OF COPENHAGEN
FACULTY OF SCIENCE



CERN-THESIS-2019-276
22/01/2020

Masters thesis

Jonas Steentoft

Developing Module Assembly and Quality Control Procedures

for the HL-LHC Upgrade of the ATLAS Inner Tracker

Advisor:
Stefania Xella, xella@nbi.dk
Richard A. Brenner, richard.brenner@cern.ch

Handed in: December 31, 2019

Front page image:

The front page image shows the assembly of the first electrical R0 module in the Scandinavian Cluster, with two read-out PCB's (hybrids) attached to a microstrip silicon sensor - following the successful development of a glue robot as described in this thesis.

Contents

Contents

Abstract

To further probe the fundamental structures of matter, the Large Hadron Collider (LHC) at CERN is undergoing a programme of upgrades aimed at increasing the instantaneous luminosity by a factor $5 - 7$. In lieu of this High Luminosity upgrade (HL-LHC), the current ATLAS Inner Detector will be replaced by a new large area all-silicon tracker - the Inner Tracker (ITk). This is a vast and complex undertaking, requiring the involvement of many institutes worldwide. A modular design philosophy is employed, such that the ITk will consist of 19.000 independent sensor modules, facilitating the need for mass-production, with pixel and microstrip n-on-p based sensor technologies being utilised.

The Scandinavian ITk Cluster, consisting of physicists and engineers from Copenhagen, Lund, Oslo and Uppsala University, will be responsible for producing ~ 600 of the microstrip type modules, in a close collaboration between academia and industry. This partnership has required a redesign of many ITk baseline assembly procedures - to better suit the production line-up in industry.

The work carried out during this project touched upon several aspects of the production procedure. The primary product of the research effort described in this thesis, is the development of a robot for high precision high accuracy glue dispensing, both as to regards the placement and amount of glue. The two-component epoxy glue can be delivered with a precision of $2\,\mathrm{mg}$, with clear avenues of improvement, over a time period of $50\,\mathrm{min}$. It will be used in mounting the read-out, powering and control electronics on to the surface of silicon microstrip sensors - a most critical step of the assembly procedure.

In addition to this, a study of early onset micro-discharge, seen in IV curves of the ITk silicon sensors labelled ATLAS12EC, has been carried out. Using NRA analysis techniques, we could show that the top oxide layers of several mini sensors had been contaminated with up to $\sim 7\%$ hydrogen by atomic fraction, most likely due to prolonged humidity exposure - hinting at humidity being the cause for this early onset phenomenon.

Acknowledgments

I would like to extend my sincerest gratitude to the singular entity that helped me solve most of the problems faced throughout this thesis; **tape.** Whether it be regular office tape, scotch tape, gaffa, doubled side, black electrical or the orange stuff also known as kapton. They have all been of paramount importance as regards to the successful execution of many an experimental setup forming the foundation for this body of work.

Besides this most obvious of acknowledgements, I would also like to express a sincere gratitude towards my two supervisors, Associate Professor Stefania Xella, Copenhagen University, and Professor Richard Brenner, Uppsala University. They gave me a magnificent opportunity to contribute in the development of one of the largest and most complex scientific experiments in the world - by providing me with a lab to play in and a guiding hand to follow.

I am furthermore very appreciative of the close day-to-day collaboration with phd. student Eleni Myrto Asimakopoulou during my time at Uppsala University - along with the help provided by postdoc Craig Wigglesworth during my time at Copenhagen University.

Introduction

To further probe the fundamental structures of the universe, the Large Hadron Collider (LHC) at CERN is being upgraded to increase its collision rate significantly, in an attempt to further strip the veil from some of the big open questions in 21th century physics, such as:

- What is the cause behind the observed matter - antimatter asymmetry in the universe?
- What is dark matter?
- What is dark energy?

The High Luminosity upgrade of the LHC (HL-LHC), will require a full replacement of the current ATLAS inner detector (ID), since it was designed with a lifetime of 10 years in the harsh radiation environment of the LHC. The detector's performance also needs to be upgraded, to cope with the increased collision rate of the HL-LHC. The new Inner Tracker (ITk) will be an all silicon detector, consisting of pixel sensors closest to the Interaction Point (IP), and microstrip sensors further away. This will be arranged into two different geometrical structures, the barrel and the end-cap(s), in order to form the hermetic cylindrical shape of the detector.

The ATLAS detector is one of the largest and most complex machines ever built, as such, it is a highly non-trivial task to take an upgrade from the drawing board and into reality - an undertaking which this thesis is a very small part of. This is done by contributing to the development and optimisation of procedures for the production and quality control of the detector sub-modules which the ITk will consist of. It was done as a participant of the Scandinavian Cluster, a gathering of scientists and engineers from the Universities of Copenhagen, Lund, Uppsala and Oslo, along with the Swedish electronics company NOTE [23], collaborating on the production of two types of microstrip modules for the end-cap part of the ITk.

Throughout the time span of my master project, I have been involved in the following aspects of module assembly and quality control:

- Developing a Glue-robot for precise, easy and reliable semi-automated hybrid to sensor assembly, as an alternative method to the ITk collaboration standard method of stencil utilisation - this is the primary product of this body of work.
- Collaborating with industry to optimise automated mounting procedure of ASICs to hybrids.
- investigating abnormal early onset of micro-discharge in IV curves of silicon sensors, causes and solutions.
- Setting up and running electrical tests of fully populated hybrids.
- Developing control software for parts of the box to be used for thermomechanical and electrical quality control (QC) of modules.

The thesis is structured as follows: I will first give a brief introduction on particle physics, CERN, LHC and the HL-LHC upgrade in a Chapter 1. This will be followed by an overview of the ATLAS experiment and the ITk, before going into a brief characterisation of the different components making up the base detector module of the ITk. Chapter 2 also describes the assembly process, as performed in the Scandinavian Cluster, and the different electrical and thermomechanical quality control (QC) tests that individual components and whole modules undergo - during and after assembly. Chapter 3 covers my primary contribution to the Scandinavian ITk effort, the development of a glue robot for

attaching hybrids and power-boards to the sensor surface - with high precision, reliability and ease of operation. The fruits of our labour, the actual assembly and testing of modules will be showcased in Chapter 4, which, due to the level of technical detail has been split from the previous chapter to present a more cohesive narrative.

During the routine quality control of silicon sensors, an interesting malfunction was observed across several different sensors. It was therefore decided to pursue an investigation into the probable causes and solutions to this malfunction. However, this requires that we also cover the relevant semiconductor theory explaining how the ITk sensors function. As the sensors are the key component of the detector, it is vital to have a good understanding of their behaviour, their strengths and weaknesses and the limitations they put on the rest of the process - this will be the focus of the final Chapter 5.

1.1 What is CERN?

The Conseil Européen pour la Recherche Nucléaire (CERN), the European Council for Nuclear Research was founded in 1953 with the intent of investigating the universe at the smallest scales of size, but doing so in an international setting which would promote peace and collaboration across borders. The laboratory was originally founded in the outskirts of Geneva, but quickly grew to expand over the french border, meaning that a typical work day at CERN truly was and is an example of border-less collaboration in the name of science.

The principals of collider based high energy experimental physics can be overarchingly summarised as follow:

1. Build an accelerator to smash small particles together at as high energies as possible.
2. Have a detector recording what happens when particles are smashed together.
3. Develop theory to interpret the results - possibly revealing the substructure of said smashed particles and/or how they combine into/interact with other particles.
4. Use the knowledge gained to reiterate the process with a more powerful accelerator and a better detector, to probe higher energies at higher precision - thereby going to smaller length scales and closer to the fundamental structure of matter.

This scheme, intentionally vague for the sake of brevity, has been incredibly successful w.r.t. uncovering the constituents of matter, from the atomic scale and down - so listing all the related scientific discoveries and technical developments would be an almost insurmountable task.

However, this is of course not a satisfactory answer to the curious of mind, so in the next section I'll give a brief overview of particle physics and some select concepts which are relevant for the operation of high energy physics collider experiments.

1.2 A Brief Introduction to Particle Physics

The modern understanding of nature is based on the Standard Model, which lists four fundamental interactions between a set of matter particles called fermions, defined as having half-integer spin, and with each interaction having one or more mediator particles called bosons, defined as having integer valued spin, which are exchanged between these matter particles during interactions. All the particles of the Standard Model can be seen in Figure 1.1 while the four fundamental interactions are listed below:

- Weak Interaction
  $\rightarrowtail$ Mediated by the electrically charged $W^{+/-}$ bosons and the neutral Z boson - eg. responsible for flavour changing interactions of elementary particles and nuclear processes such as the beta decay.
- Strong Interaction
  $\rightarrowtail$ Mediated by the gluons - responsible for the binding of quarks into hadrons and hadrons into nuclei. The mass of hadronic matter comes primarily from the strong interaction binding energy - not the mass of the constituent particles.
- Electromagnetic Interaction
  $\rightarrowtail$ Mediated by the photon - is responsible for much of the mesoscopic order in the universe by, amongst other things, covalently binding atoms into molecules.

- Gravity
  $\rightarrowtail$ Suggested to be mediated by a graviton - searching for it is ongoing. Gravity is the force behind large scale structure in the universe, eg. in the form of planets, stars and galaxies.

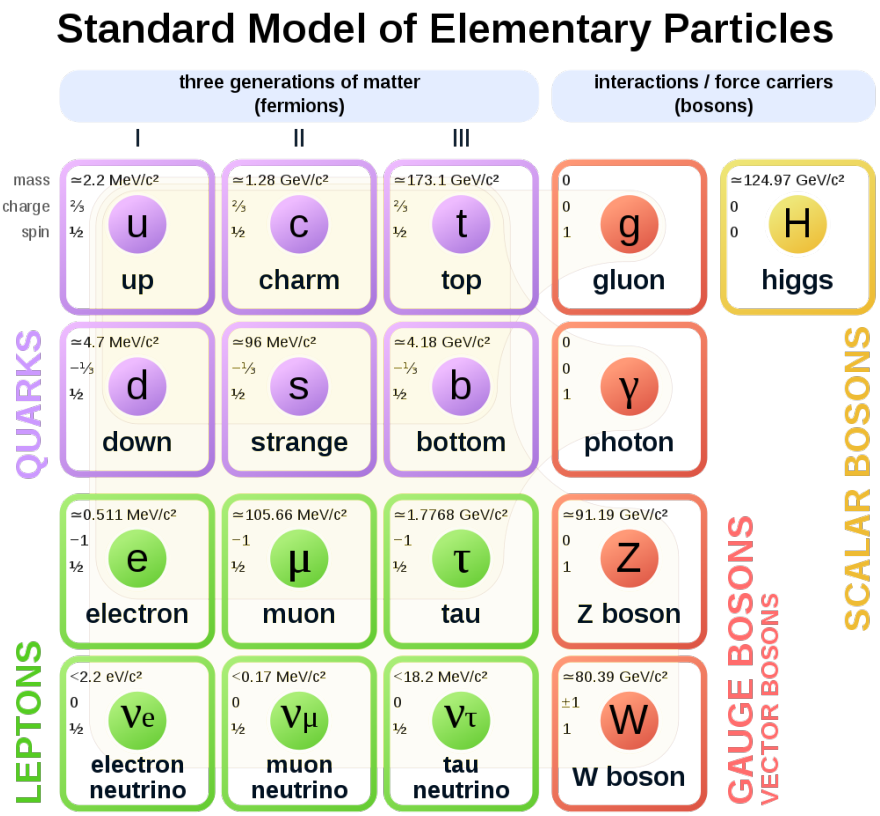


Figure 1.1: The periodic table of the 21th century, showing the different elementary particles of the Standard model and a few select properties like mass, electrical charge and spin. Each fermion also has a corresponding antiparticle with opposite electrical charge - not shown for the sake of clarity.

The fermions of the Standard model differ in mass and which interactions they couple to. Eg. only the quarks have the colour charge needed for strong interactions, and as such, only the quarks form bound structures of 2-5 constituent particles, named, mesons, baryons, tetraquarks and pentaquarks. An overview of the different fermions and their couplings are given in Table 1.1.

Table 1.1: Standard model fermions and their couplings. down/up type quarks refers to weak interaction properties, were down type quarks only decay into up type quarks and vice versa [2].

					strong	electromagnetic	weak
Quarks	down-type	d	s	b	✓	✓	✓
	up-type	u	c	t			
Leptons	charged	e^-	μ^-	τ^-		✓	✓
	neutrinos	ν_e	ν_μ	ν_τ			✓

Also, as a quick side note, in the context of particle physics, gravity is usually ignored for two reasons. The current inability to unite the theoretical framework of the Standard Model and General Relativity in a sensical way, and because of the minuscule effect gravity has on elementary particles, given that the interaction strength scales with the mass of the particle.

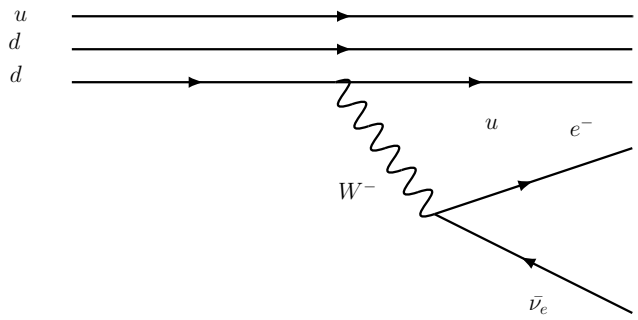


Figure 1.2: Feynman diagram showing the beta decay of a neutron into a proton. The d quarks undergoes a flavour changing decay into an up quark and W^- boson which then further decays into an electron anti-electron neutrino pair.

All macroscopic structures, meaning tortoises, planets etc., are build up from the three lightest types of fermions, excluding neutrinos. The up and down quarks bind together through the strong interaction to create neutrons (ddu) and protons (uud), the constituents of atomic nuclei, which together with electrons, bound electromagnetically to the nuclei, forms the different atoms of the period table of elements. The explanation behind this is the fact that nature is lazy. Put in slightly more scientific way, in a closed physical system with no external forces, the equilibrium state tends to be the minimal energy state of the system. What this means in the context of particles is, that all quarks and leptons will, over time, decay into a lighter version of themselves, relaxing into a lower energy state - and this continues until there are no lighter states to decay into. This is also why unbound neutrons are unstable. With a mass of $939.6\,\mathrm{MeV/c^2}$, it is slightly heavier than the proton, mass $938.3\,\mathrm{MeV/c^2}$, resulting in free neutrons decaying into protons - with a mean lifetime of approximately $15\,\mathrm{mins}$, as seen on Figure 1.2. On the other hand, the mean lifetime of protons only has an experimental lower bound of roughly 10^{29} years, it is the lightest known baryon, meaning a triquark composite particle. It has no allowed decay channels, referring to the conversation laws and underlying symmetries we believe to govern the interactions of elementary particles, and as such, the proton is considered a stable particle - much like the electron [2].

There are many big unanswered questions in particle physics, we'll briefly discuss a few of them here:

Cosmologists have observed that, the overall contents of the universe, divided by energy density, consists of ~ 5 % baryonic matter, ~ 30 % dark matter and ~ 65 % dark energy. This means that, the Standard Model, our microscopic model of fundamental constituents, only includes ~ 5 % of the total amount of "stuff" in the universe - which is not exactly impressive seen from that perspective. Dark matter is thought to be a type matter which only, or primarily, interacts gravitationally, There are currently many different Beyond Standard Model (BSM) theoretical models, eg. heavy right-handed neutrinos, predicting different experimental signatures which are being searched for at ATLAS. Dark energy is the blanket label put onto the unknown driving force behind the universe's accelerating rate of expansion - I'm not aware of any ongoing searches for dark energy at ATLAS.

When looking at the macroscopic world, we see that it is completely matter dominated. Meaning that all baryonic objects, like humans and planets, are made of matter and not antimatter - which is quite the conundrum. In the Standard Model, there's an almost perfect interaction symmetry between matter and antimatter, meaning matter and antimatter being created and annihilated in equal amounts. Asymmetry have been measured in a few cases - through the mechanism of CP-violation, but the measured degree of asymmetry is much smaller than what we see at the macroscopic level. Educated guesses tells us that, during the very early stages of the big bang, something seeded an imbalance in the matter antimatter ratio, and this imbalance grew with the expansion of the universe to the levels we see today. However, no solid evidence have yet been found to properly explain the degree of asymmetry. Personally, I find this to be one of the most intriguing open questions of particle physics,

because it stems on such a spectacular failure of a model that otherwise works so well - in so many aspects of explaining the composition of the physical world.

1.3 The Large Hadron Collider (LHC)

The Large Hadron Collider is the largest particle accelerator in the world, build to further investigate the fundamental structure of matter and its interactions - probing higher energy scales, and thereby smaller length scales, than ever before. The LHC is a circular collider located in a tunnel, $27\,\mathrm{km}$ in circumference, $80 - 130$ m below ground, facilitating collisions of proton-proton, and heavier nuclei, at center-of-mass energies up to $14\,\mathrm{TeV}$ at luminosities around $10^{34}\,\mathrm{cm}^{-2}\mathrm{s}^{-1}$. The machine is a successor to the Large Electron Positron collider (LEP), built in the same tunnel to, amongst other things, find Higgs boson - the only undiscovered particle in the Standard Model at the time of commission. For the proton-proton collisions, it all starts from the same small bottle of hydrogen gas, where the atoms are stripped of their electron prior to injection in the two beam-pipes, and it ends with data generating collisions inside one the many experiments attached to the accelerator. An overview of the entire experimental complex of CERN can be seen on Figure 1.3 - this gives an impression of the size and complexity of the operations carried out.

A simplified view of a particle accelerator like the LHC can be seen on Figure 1.4. Particles are injected into the ultra-high vacuum of the two beam-pipes, where magnets are then used to contain, steer and focus the beam. Acceleration is provided by an electric field cavity. When the two beams have reached maximum energy, and have been focused to have a minimal transverse beam size, to maximise the collision probability, the beams are brought to collision inside the physics experiment were we can record what happened.

In the following sections we'll further explore a few select areas of accelerator physics, to get a better understanding of how this incredible machine functions.

1.3.1 Key Parameters of a Particle Accelerator

There's an unfathomable long list of parameters characterising the operation of a particle accelerator like the LHC, but the most important are probably the center-of-mass (C.O.M.) collision energy, the beam luminosity and the cross sections of the interactions you're searching for.

Collisions energy

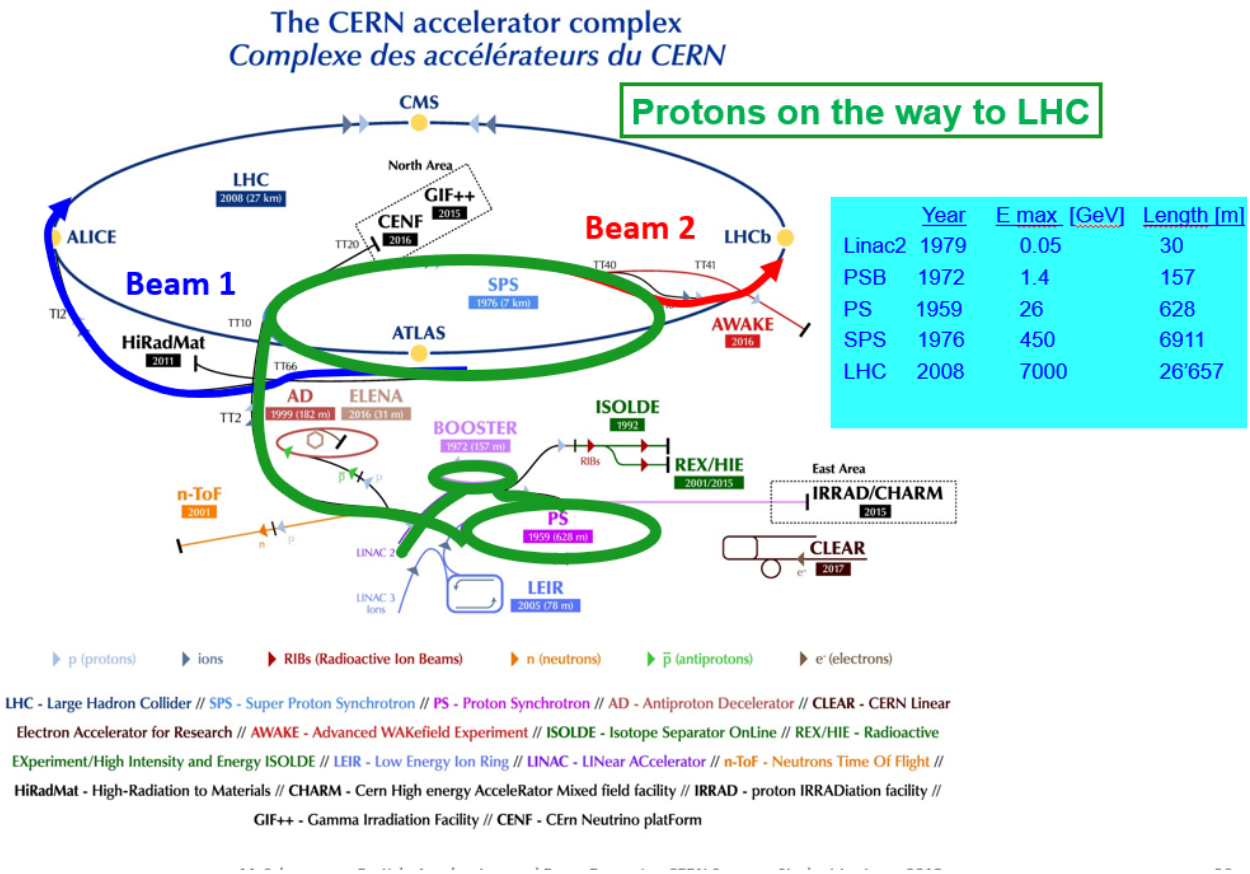
For a pair of interacting particles, each with mass m_i, momentum vector $\mathbf{p_i}$ and total energy E_i we can write up the Einstein energy relation, using natural units of $c = 1$, as

$$\left(\sum_i E_i\right)^2 = \left(\sum_i m_i\right)^2 + \left(\sum_i \mathbf{p_i}\right)^2 . \tag{1.1}$$

This means, that if we want to create a new particle from the collision, either through a scattering or annihilation event, conservation of energy limits us to creating particles of mass m upholding

$$m^2 \leq s = \left(\sum_i E_i\right)^2 - \left(\sum_i \mathbf{p_i}\right)^2 . \tag{1.2}$$

$\sqrt{s}$ is called the center-of-mass energy of the interaction. The magnitude of $\sqrt{s}$ depends on the type of accelerator and the type of particles being accelerated. Using Equation 1.2 one can do a neat little calculation showing the dramatic difference in $\sqrt{s}$ for a fixed target machine and a colliding beam



	Year	E max [GeV]	Length [m]
Linac2	1979	0.05	30
PSB	1972	1.4	157
PS	1959	26	628
SPS	1976	450	6911
LHC	2008	7000	26'657

Figure 1.3: The CERN accelerator complex as of 2017, giving an overview of all the different projects happening at CERN. The paths of protons into the LHC is highlighted, going through the previous genereations of accelerators now acting as part of the LHC injection chain. The maximal energy and size of the difference accelerators are summarised in the table on the right [4].

machine like the LHC - lets say we have a 7 TeV proton colliding with respectively, a proton at rest and another 7 TeV proton.

$$\sqrt{s}_{fixed} = \sqrt{(E+m_p)^2 - (\mathbf{p})^2} = \sqrt{E^2 - p^2 + m_p^2 + 2Em_p} = \sqrt{2m_p^2 + 2Em_p} = 0.115\,\text{TeV}$$
$$\sqrt{s}_{2beams} = \sqrt{(E+E)^2 - (\mathbf{p}-\mathbf{p})^2} = \sqrt{4E^2} = 2E = 14\,\text{TeV}$$

This dramatic difference in $\sqrt{s}$ comes from conservation of momentum requiring a majority of the energy staying in kinetic form during the fixed target collision [2, Chp 1].

One of the real-life limiting factors for the maximally achievable energy in an accelerator is the rate of energy loss. In an circular accelerator like the LHC and LEP (Large Electron Positron collider), the primary mechanism of energy loss for the beam particles is bremsstrahlung, or synchrotron radiation. When charged particles are suspected to external accelerations, they radiate photons in an attempt to shed the energy gained. For a relativistic charged particle kept in uniform circular motion, The total rate of energy loss P can be calculated classically as

$$P = \frac{q^2 E^4 \beta^4}{6\pi\epsilon_0 m^4 c^5 r^2 sin^2(\alpha)}. \tag{1.3}$$

The rate energy loss scales vigorously with the total particle energy to the fourth power but is kept down by the mass of the particle and the magnitude of curvature given by the radius r and the bending

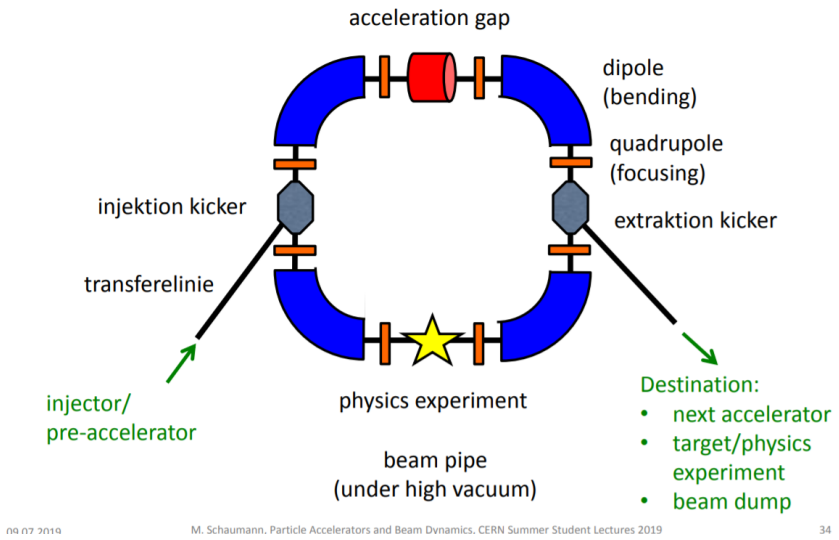


Figure 1.4: Sketch of the basic components of a synchrotron based particle accelerator [4].

angle α [5]. This is what limited LEP, and why it is unlikely that we'll ever see a TeV scale electron positron circular collider. The proton mass is roughly 2000 times greater than the electron mass, meaning that the power loss is suppressed by a factor $(m_e/m_p)^4 \approx 10^{-13}$. For proton-proton circular colliders, it is typically not the energy loss of the beam which sets the limit on achievable top energy, but rather the bending power of the magnets containing the beam. For a magnetic field orthogonal to the plane of motion, as is the case for the dipoles used in the LHC, the trajectory of a charged particle will be bend into a circular motion through the approximative relation

$$p = 0.3BR. \tag{1.4}$$

P being particle momentum in $\mathrm{TeV/c}$, B the magnetic field strength in Tesla and R the bending radius in kilometres. Using this formula, one could produce a rough estimate of the maximum achievable center-of-mass energy of proton-proton collisions at the LHC. However, naively plugging in the typical LHC values will overshoot the actual achievable energies by quite a bit, because only $2/3$ of the ring circumference is actually covered with bending dipoles. The remainder of the space around the ring is used for things such as beam focusing through eg. quadropoles, accelerating RF-cavities, beam diagnostics equipment and the long straight sections, of order several hundred meters, on each side of the actual experiments - like ATLAS.

However, the argument still stands - to reach higher center-of-mass energies than the current $13 - 14$ TeV, better magnets with higher fields strengths or a larger tunnel is required[2, chp. 1].

Luminosity

Luminosity has units of number of particles per area per time, $\mathrm{cm^{-2}s^{-1}}$ and it is the primary machine parameter for describing how many particles are being delivered for collision over time. The event rate for a given particle interaction can be constructed as

$$N = \mathcal{L}\sigma \tag{1.5}$$

where the cross section σ is the quantum mechanical interaction probability, given in units of inverse area, determined by the fundamental physics governing the given interaction. In the LHC, particles don't just collide in one continuous stream, instead they are separated into roughly 2800 bunches of 10^{11} protons, spaced $25\,\mathrm{ns}$ apart at collision. This beam structure is dictated by the fundamental physical constraints for the equipment of the accelerator, like the accelerating RF-cavities. However, one also has to consider the limitations of detectors, eg. the propagation time of read-out signals

being non-instantaneous - meaning that if the collision rate is too high, data will overwritten and lost. Another aspect is the spatial granularity of the detector, if the pile-up, meaning the amount of simultaneous collisions per bunch-crossing, is too high, it will be impossible to properly separate data into the individual events we look for during an analysis - the detector is effectively blinded by too much simultaneous data. The primary determining factors of pile-up are the amount of protons in a bunch, the cross sectional area of the beam and the angle of collision. Typical pile-up conditions in the ATLAS detector so far have been $20 - 50$ collisions per bunch-crossing, increasing over time due to improvements in equipment, operators and analysts performance.

Because the actual amount of collisions depends on the interaction probability of all the different possible interactions, physicist prefer to quantify the LHC "production yield" in terms of time-integrated luminosity, which quantifies the total possible amount of collisions facilitated by the LHC operation. This integrated luminosity is typically given in units of *inverse barns*, $1\text{barn} = 10^{-24}\,\text{cm}^2$, because interaction cross sections are calculated in terms of *barns* - with interesting phenomena like Higgs production having a total cross section on the order of *picobarns* (pb) at the LHC [3].

An estimate of luminosity can be calculated as follows. Assuming a Gaussian beam profile in the horisontal x and vertical y, with transverse beam size σ_x and σ_y, given a collision frequency f and number of particles in the colliding bunches n_1 and n_2, the luminosity is then [2, chp 1]

$$\mathcal{L} = f\frac{n_1 n_2}{4\pi\sigma_x\sigma_y}. \tag{1.6}$$

1.3.2 Beam Control through Magneto-Optics

The storage ring of a particle accelerator has two primary functions, containing the beam, ie. successively bending the beam to follow the machine's radius of curvature, and focusing the beam as much as possible, squeezing the transverse beam size down to maximise luminosity and thereby the data gathering rate. This is typically achieved through the use of the following four generic types of magnets seen on Figure 1.5 - all influencing the beam through the Lorentz Force

$$\vec{F} = m\vec{a} = q\left(\vec{E} + \vec{v}\times\vec{B}\right). \tag{1.7}$$

With the electrical field term being negligible in this regard, since the $8\,\text{T}$ B field of the LHC dipoles generates a force equivalent of a $2400\,\text{MV/m}$ electric field. For reference, the magnitude of such an electric field is roughly a factor hundred bigger than what can be achieved in the state-of-the-art RF-cavities used for acceleration in the LHC.

To achieve a stable beam with a minimal transverse beam size, we need to figure out which types of magnets to use, how many and where to put them around the storage ring. This is done by adopting some clever formalism from the field of geometrical optics. This discipline of physics is concerned with describing the propagation of light through optical elements such as lenses, and it contains a mathematical formalism which facilitates this task in a very elegant way. For every type of optical element, one can calculate a transfer matrix which, when multiplied with the state vector of the light beam, describes the effect of the optical element on the beam. This method generalises to an arbitrary array of optical elements, simply by calculating the total matrix product, in the order in which the beams meets each element, to describe the full beam propagation through the array.

This transfer matrix formalism can also be, and is, widely used to describe the propagation of charged particles through different configurations of B fields - hence the name magneto-optics.

In the following discussion of common magnet types in particle accelerators, **x** will be the horisontal transverse, **y** the vertical transverse and **z** the longitudinal direction of the particle beam -

with the design orbit positioned in the origin of the transverse plane, such that a non-zero transverse coordinate directly gives the amount of deviation from design orbit in that direction[6].

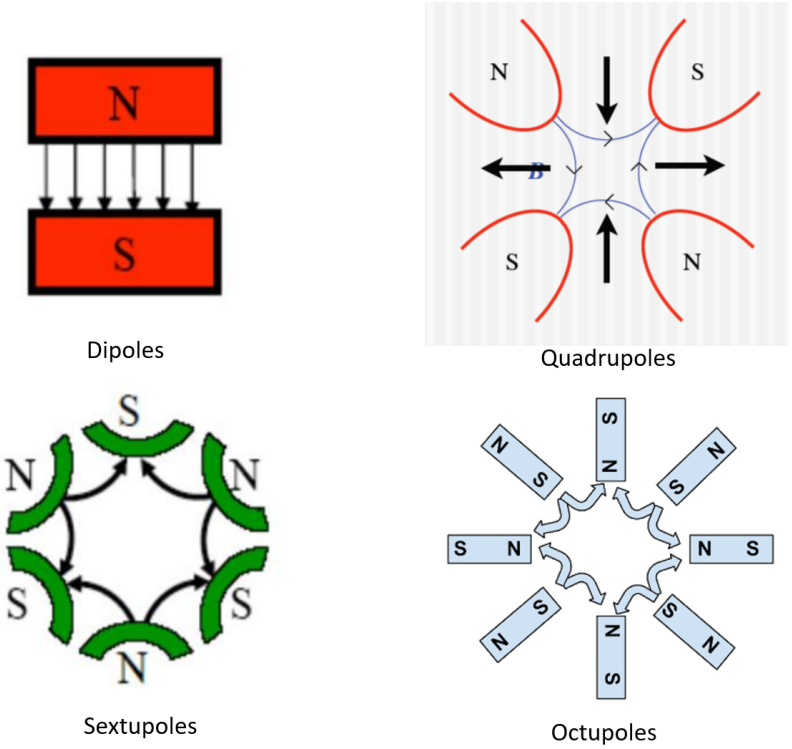


Figure 1.5: Overview of the different types of magnets used in a synchrotron like the LHC. Dipoles for bending, quadropoles for focusing, sextu- and octopoles for higher order corrections such as chromatic dispersion [6].

Dipoles

Dipole magnets are used to bend and steer the beam around the storage ring through a field profile of

$$\vec{B} = (0, B_0, 0) \qquad \rightarrow \qquad \vec{F} = (-qvB, 0, 0) \tag{1.8}$$

resulting in constant bending force in the horisontal direction that leads to circular motion, assuming a velocity purely along **z** - an assumption which is not fully valid, something we'll get back to later.

Quadropoles

Quadropoles are used to focus the beam - with an idealised field of

$$\vec{B} = (Ky, Kx, 0) \qquad \rightarrow \qquad \vec{F} = qvK\,(-x, y, 0) \tag{1.9}$$

and a normalised focusing strength of $g = 0.3K/p$, with the field gradient K in units of $\mathrm{T/m}$ and momentum p in $\mathrm{GeV/c}$. This means that a particle perfectly following the design orbit will be completely unaffected by the quadropole field, while deviating particles will feel a force proportional to the magnitude of deviation. However, as can be seen in Equation 1.9, the Lorentz force will focus in one transverse direction while defocusing in the other transverse. The solution to this problem is using an alternating setup of quadropole rotations, such that a magnet focusing in **x** and defocusing in **y** is followed by a magnet doing the opposite. This alternating setup of focusing and defocusing is commonly referred to as the FODO cell, FOcus-drift-DefOcus. It achieves a net focusing effect of the beam, because once a particle has been properly focused, close to the origin in a given direction, it

will feel a negligibly small defocusing force in the subsequent quadropole, due to the B-field strength scaling with the position deviation in that direction.

The drift space between quadropoles is typically filled with dipole bending magnets, correction magnets like the sextupole described below, beam analysis equipment or collimators. A collimator is a device which pre-emptively cleans the beam of its most deviating parts, simply by absorbing it in some radiation hard material. By doing this before the particles actually diverge, we avoid an unnecessary build-up of radiation damage in all the other sensitive parts of the accelerator.

Sextupoles

In a realistic particle beam, some degree of momentum dispersion will be present, and this in turn leads to a divergence of the transverse beam size due to the quadropole focusing strength dependence on particle momenta. This phenomena is called chromatic dispersion from the corresponding effect in geometrical optics, where the index of refraction has a wavelength dependence, leading to a wavelength based dispersion of the focal length for a lens. However, this effect can be corrected for, and this is the purpose of sextupole magnets,

$$\vec{B} = S\left(2xy, x^2 - y^2, 0\right) \qquad \rightarrow \qquad \vec{F} = qvS\left(y^2 - x^2, xy, 0\right). \tag{1.10}$$

The x^2 (y^2) scaling of the field strength compensates the over(under)shot of the quadropole focusing due to momentum deviations.

Octopoles

Octopole magnets are used as a higher order refinement of the transverse beam profile, having a B-field configuration that results in a combined focusing and chromatic correction of the beam.

1.3.2.1 Beam Stability

From the description of magnet types above, we see that in order to build a storage ring with a stable beam orbiting in it, one needs, as a minimum, a collection of dipole and quadropole magnets spread out across the ring, to bend the beam and keep the transverse beam size from diverging. The design calculations can then be greatly simplified by using a periodic lattice of magnets, such that one only needs to find the transfer matrix of this lattice, to describe the complete orbital motion.

For such a periodic lattice, consisting of dipoles, quadropoles and drift spaces, one can set up a differential equation of motion called Hill's equation, which can be solved to find the conditions for stable orbital motion of a charged particle. For the sake of brevity, this discussion is limited to a conceptual level, and as such we won't dive into the mathematical details, which can be found in eg. [6] or [8], and instead simply focus on the knowledge gained from the solutions to Hills equation.

In order to have a stable orbital motion, accelerator designers need to use a periodic lattice with a transport matrix M upholding $|TrM| < 2$ - the trace of the transport matrix being less than two. This very simple benchmark gives accelerator designers a very straight forward way of evaluating the usability of their design - with the tricky part coming from constructing a transfer matrix accurately describing the physical setup.

Furthermore, the stable trajectories of charged particles around the storage ring, are shown to undergo oscillations in the transverse plane, around a closed orbit through the exact centre of the quadropoles - the so called design orbit. This makes intuitive sense given the focusing-defocusing nature of quadropoles. We call this betatron oscillations, and besides from the momentum dispersion, the oscillation amplitude is primarily determined by two things, the design of the FODO cells and the transverse beam size upon injection into the ring. This is one of the reason that, as a part of the HL-LHC upgrade efforts, LINAC2, the linear accelerator acting as the very first link in the accelerator chain seen on Figure 1.3 will be replaced by the new LINAC4, which will deliver a beam, 3 times

more energetic and 2 times brighter, to the Proton Synchrotron Booster (PSB).

Another very important aspect of betatron oscillations is the frequency, with the number of oscillations per revolution called the machine tune. Under realistic circumstances, you will always have slight flaws present in some of the magnets used throughout the ring, where the B-field at points in space is slightly too large or small, resulting in passing particles getting a wrong kick. Now, if the same particles gets the same wrong kick every time they pass the flawed magnet, the magnitude of the error will accumulate until the particles enter an unstable orbit and diverge from the beam. This would, over time, lead to unnecessary equipment wear through radiation damage, and beam losses limiting the achievable luminosity of the accelerator. This problem can be avoided by using a betatron oscillation frequency which is neither integer nor half-integer valued. Because, this means that a given particle's transverse position, at a given point along the orbit, is different for every full orbit revolution, and with the period of returning to the same transverse position being some very high number - removing the possibility for errors to build-up over time. Typically, the machine tune is controlled through manipulating the quadropole field strength [8].

Transverse Beam Size

Neglecting any dispersion related effects, the transverse beam size in each dimension, at a given point along the ring, can be estimated as

$$\sigma = \sqrt{\beta\epsilon} \tag{1.11}$$

where ϵ is the emittance, explained below, and β being the value of the beta-function at this point along the ring. The beta-function is a part of the solution to Hills equation of motion, and it characterises the relative amplitude of betatron oscillations. You will often hear accelerator physicist talk about the β^*, which is simply the value of the beta-function at the interaction point (IP). At the LHC, this β^* is minimised at the IP through specialised quadropole triplets, however this leads to a large blow-up of the total beam-size before and after the IP - resulting in a balancing act of how small one can get the beam size at the IP without having the consequent blow-up exceeding the beam-pipe diameter [11].

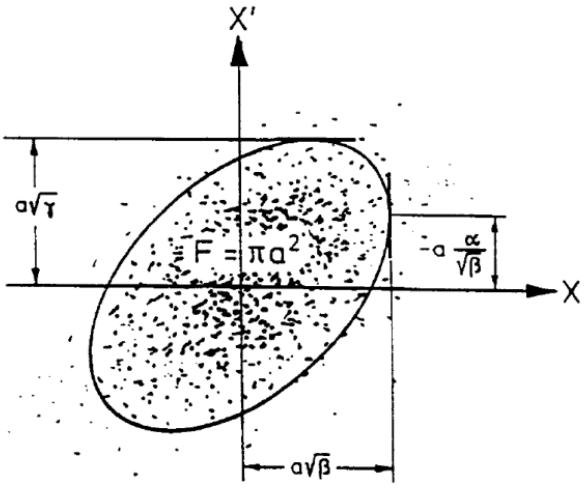


Figure 1.6: Typical distribution of beam emittance in the transverse $(space, velocity)$ phase space. The deviations from the design orbit located in $(0, 0)$ comes from eg. the natural betatron oscillations and real-world imperfections giving rise to the various sources of beam dispersion.

Emittance is a typical way of parametrising the size of the beam profile, and it has several often employed definitions, depending on the specific usage. The one most relevant for our purposes goes as follows. The emittance of a particle beam is, for a given point along the ring and for each dimension of space, defined as the elliptical area of (x, x') phase space containing 95 % of the particles in the beam, x' being the derivate of x. We specifically use an elliptical area because particles in stable betatron orbits map out an ellipse in this phase space.

One should however be careful in distinguishing between transverse and longitudinal emittance, since they're related to different underlying physical mechanisms. Neglecting dispersion effects, The transverse emittance is coupled to the betatron oscillations of the beam, while the longitudinal emittance arises from the use of an AC electrical field for acceleration, leading to so called synchronous oscillations w.r.t. the design orbit - also mapping out a phase space ellipse. A key difference being that the longitudinal phase space ellipse is drawn in the (E_{gain}, ϕ) space, with ϕ being the phase of the accelerating E-field encountered by the particle gaining an amount of energy E_{gain}.

Finally, the concept of a "beam envelope" should be mentioned. It can be understood as the surface of maximal stable transverse deviation from the design orbit - in the (x, x') phase space. As we've seen, deviations can come from many different sources, eg. the natural betatron oscillations, chromatic dispersion or magnetic flaws, but as long as a particle stays within the beam envelope, the particle will stay with the beam. However, if the beam envelope is exceeded, the subsequent bending and/or (de)focusing magnets will send the particle into an unstable orbit, instead of acting as a restoring force.

1.3.2.2 Superconducting Magnets

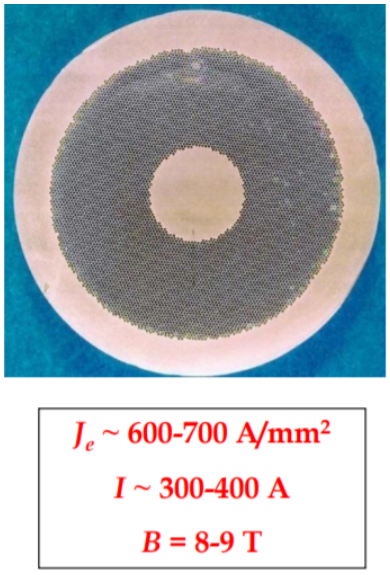


Figure 1.7: Cross section of the superconducting cable that make up the coils of the LHC $8\,\mathrm{T}$ dipole magnets. The cable has a total diameter of $0.85\mathrm{mm}$ and consist of superconducting strands of Nb-Ti with a diameter a $6 - 7\mu\mathrm{m}$, embedded in a matrix of Copper [7].

To reach the very high B field strengths required at the LHC, superconducting electromagnets are used. The LHC dipoles utilise coils made from cables of small superconducting strands of Nb-Ti, suspended in a very pure Copper matrix. The reason for using many smaller superconducting strands rather than one large strand has to do with system stability. Changes in the B-field will result in temporarily increased local heat loads from the induced Eddy currents, which can be very problematic if this localised heating causes a loss of superconductivity. This can very quickly evolve into a literal

meltdown of the entire magnet, due to the extremely high currents employed and the high normal state resistivity of most superconductors. But, heat transfer scales with the surface area to volume ratio, and by maximising this ratio through using many smaller strands, one minimises the probability of critical hotspots forming, by spreading out the generated heat as quickly as possible.

In case a hotspot does form, and the critical temperature is exceeded, the high thermal and electrical conductivity of the Copper helps spread out the generated heat in the time before Machine Protections Systems kicks in and cuts power to the coil.

1.3.3 Acceleration through RF Cavities

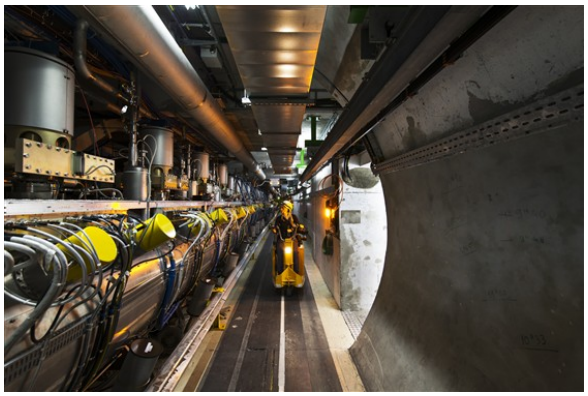


Figure 1.8: Picture of the superconducting RF-cavities used to accelerate particles in the LHC.

To accelerate particles in a synchrotron, one also relies on the Lorentz force, but in this case through electrical fields, because magnetic fields do no work and because they can't provide a force parallel to the direction of motion.

At the LHC, this is accomplished through the utilisation of superconducting Radio Frequency (RF) cavities, seen on Figure 1.8, channelling a $400\,\mathrm{MHz}$ AC current to generate an E-field of $5\,\mathrm{MV/m}$. This allows operators to ramp up the beam energy from its injection value of $450\,\mathrm{GeV}$ to its maximum value of $7\,\mathrm{TeV}$ in roughly $20\,\mathrm{min}$, with an energy gain of $485\,\mathrm{keV/turn}$ and a revolution speed of $11\,245\,\mathrm{turns/s}$ [8].

RF cavities are used over electrostatic acceleration because they can reach higher field values and are much cheaper to operate. But due to the oscillating nature of the field, half its phase space will decelerate rather than accelerate incoming particles. This naturally forces the particles of the beam into longitudinal bunches synchronised with the accelerating phase of the RF-cavities. To optimise the acceleration of the cavity, half the RF period should match the time it takes for a particle to traverse the cavity, such that a particle is affected by the entire accelerating part of the wave and none of the decelerating part.

Protons entering the LHC beam pipe are already highly relativistic, meaning that further increases in momentum have negligible effect on the particle velocity. However it does lead to an increase in the orbital radius, and once momentum dispersion is included in the considerations, one finds certain constraints on the phase usage of the RF-cavity. Looking at Figure 1.9, we see that when designing a phase locked accelerator, ie. particle bunches always entering the cavity at the same phase point of the E-field, that this can either be on the rising or falling side of the AC pulse. Technically you could also aim to hit the peak of the pulse, but it is easier to hit the much wider slopes of the pulse, and it also wouldn't be a stable setup as will explained next. Momentum dispersion means that some particles will arrive respectively earlier or later than intended, and because of the pulsating E-field, they will see a different field strength and thereby achieve different energy gains. In Figure 1.9, we see that if the bunch arrives on the falling slope of the E-field pulse, the blue early particle will receive

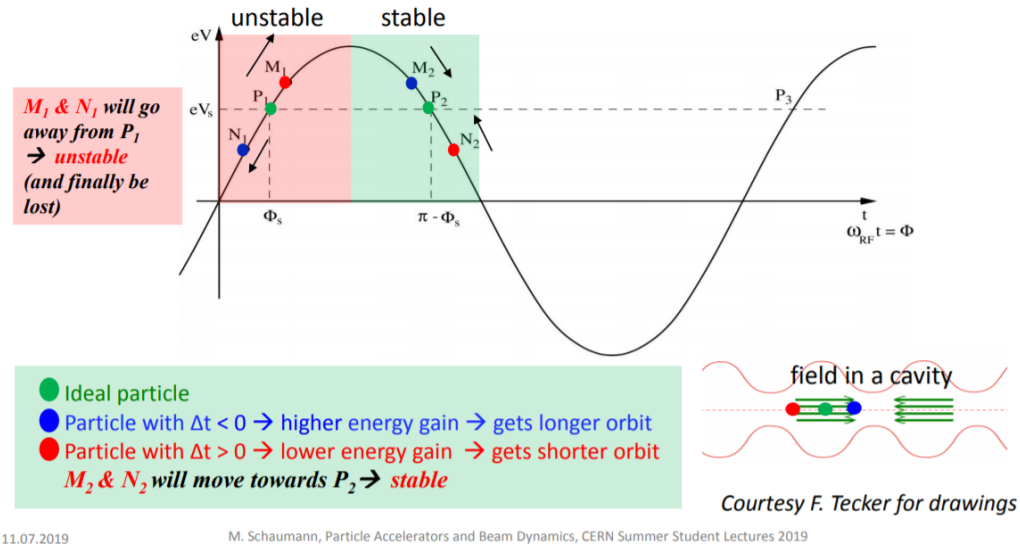


Figure 1.9: Operating principle of RF-cavity acceleration for relativistic particles. The Y-axis signifies the energy gain for a particle while the X-axis gives the time of arrival at a reference point inside the cavity. Keep in mind that both the voltage pulse and the particles traverse the cavity.

a larger energy gain than intended, and, because it is already moving at a highly relativistic speed, the velocity (almost) won't increase, while the orbital length does increase slightly - compared to the design orbit. This leads to a longer orbital period for the blue particle meaning that, when the bunch arrives at the cavity again, the design orbit particles have caught up with the particle previously a head. For particles lacking behind w.r.t. the design orbit, the exact mirrored argument applies - resulting in stragglers catching up to the design orbit - as illustrated by the red particle in Figure 1.9. This leads to oscillations w.r.t. the synchronous particle, always arriving at the design phase point, in the phase space of energy gain and E-field phase. Similarly to the beam envelope of transverse emittance, one can define an area in this phase space, typically called the RF-bucket or the separatrix, which then designates the border between stable and unstable motion.

1.3.4 Lepton vs. Hadron Colliders

A final few notes on the interesting differences between lepton and hadron colliders will follow here. When utilising a lepton collider, the beams consist of fundamental particles, eg. electrons and positrons, so one always know the exact center-of-mass energy and the experimental signatures are very clean, since the collision products always come from either an electron-positron scattering or annihilation process. On the other hand, lepton colliders are limited in energy for the circular setup, due to synchrotron radiation, and limited in statistics in the linear collider setup, since this only allows each particle bunch to be collided once. This is very inefficient since only $\sim \frac{1}{10^9}$ particles in a bunch collide during a bunch crossing. Also, because the collision energy of the two parties is always the same, one has to manually scan the range of energy scales relevant for investigation - further reducing the data gathering rate.

However, to give an example of the extreme precision achievable with a lepton collider, recall the direct scaling between beam energy and radius of curvature given in Equation 1.4. During the operation of LEP, the beam energy measurement could be used to accurately track the lunar phase. This was due to tidal forces contracting and expanding the Geneva underground - causing an ever so slight (de)increase in the circumference of the accelerator [9].

Hadron colliders have polar opposite upsides and downsides. One has much better access to high energies at high luminosities, but the data from collisions are much more complicated to analyse.

Hadrons are composite objects, consisting of a set number of valence quarks along with a Dirac sea of gluons and virtual quarks. This means, that while the hadron as whole has an energy of eg. 7 TeV, the distribution of energy across the fundamental constituents of the hadron is given by a set of probability density functions - with the average $\sqrt{s}$ being a few hundred GeV. It is furthermore impossible to control which constituent particles of the two hadrons interact in a given collisions. So for every single collision there is a huge probability space of gluon-gluon, quark-gluon and quark-quark interactions at different energy scales - making it very difficult to reconstruct what happened at the collision. Also, the cross sections of proton-proton collisions are dominated by uninteresting low energy events, which needs to be sorted out as noise when looking for rare processes. however, this large possibility space of interaction types and center-of-mass energies gives us a helping hand when it comes to looking for the unknown - if one doesn't know where to look, it is best to look in as many places as possible - and this is done automatically in hadron colliders.

So to sum up, hadron colliders, like LHC, are great for discovering new particles, and lepton colliders, like LEP, are great for precision measurements of known particles.

1.3.5 High Luminosity LHC (HL-LHC)

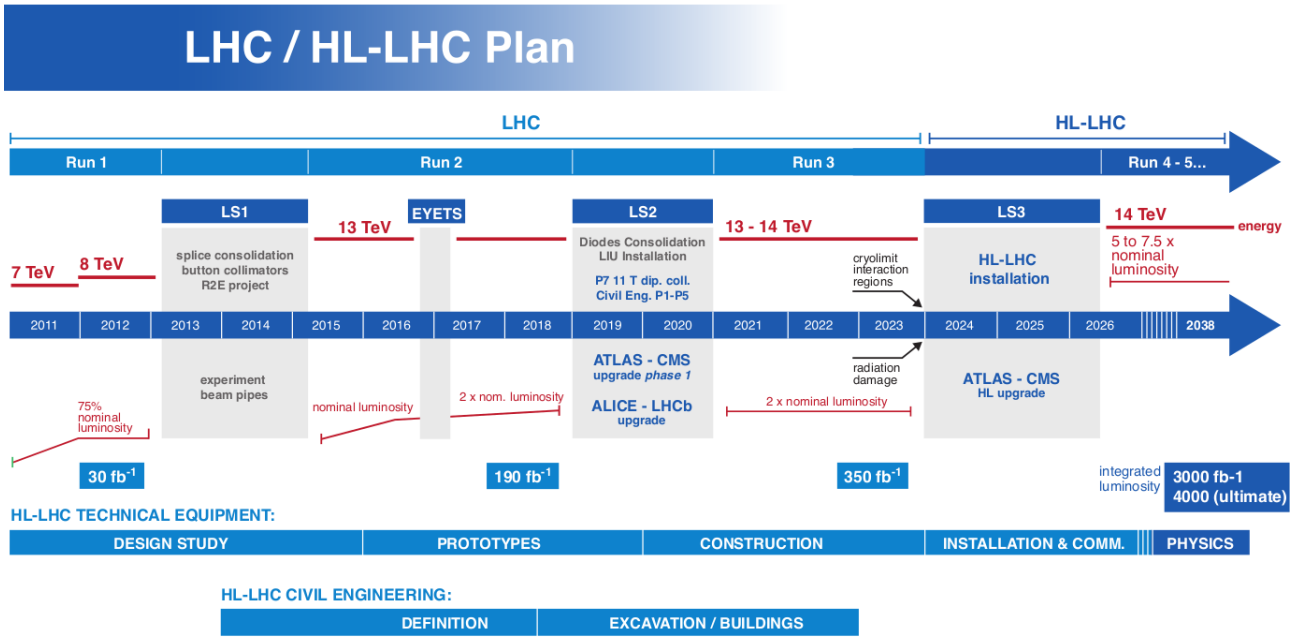


Figure 1.10: General overview of the past and future operational schedule of the LHC

Already back in 2016 operators of the LHC exceeded the nominal design parameters of the machine - delivering an instantaneous luminosity of $1.37 \times 10^{34}\,\mathrm{cm}^{-2}\mathrm{s}{-}1$ [10, chp 2]. However, because we are already operating at the limits of the machines capabilities, the time needed to reduce statistical errors on measurements, through increasing sample sizes, will start growing towards unfeasible timescales. Eg., by the end of Run-3, see Figure 1.10, continuing to run at the same data gathering rates, it would take of order ten years to halve a given statistical error [12]. So, in the spirit of not stalling scientific progress and to "fully exploit the physics potential of LHC" the CERN Council decided to pursue a major luminosity upgrade of the accelerator and detectors, to be implemented during the Long Shutdown 2 (LS2) and LS3. This HL-LHC upgrade will push the instantaneous luminosity up towards $7.5 \times 10^{34}\,\mathrm{cm}^{-2}\mathrm{s}^{-1}$, an increase of a factor $5 - 7$, by, amongst other things, increasing the bunch size and further optimising of the beam profile at the collision point - leading to pile-up

conditions of $140 - 200$ events per collision - compared to the previous level of 30 events per collisions [12]. This increase in pile-up leads to a dramatic rise in complexity of the detector read-out, as seen on Figure 2.3, which, amongst other reasons, necessitates the replacement of the current ATLAS Inner Detector.

Table 1.2 lists some of the key operational parameters for the LHC, how they improved during Long Shutdown 1 and how they will improve once commission of the HL-LHC is finalised.

On Figure 1.12 we get an overview of the major upgrades turning the LHC into the HL-LHC, and we'll briefly expand upon some of these points:

- **Bending magnets** - some of the current Nb-Ti dipoles will be replaced by new Nb-Sn magnets, capable of generating higher B-fields of $13 - 14\ T$. This leads to the same amount of bending power over a smaller distance, freeing up valuable space along the beam-line for the installation of other new equipment.

- **Collimators** - one of the most effective ways of reducing the transverse beam size, both to maximise the luminosity and to keep the radiation damage of magnets etc. to a minimum, is simply to collimate the beam. This is of course a balancing act, if the collimation is too strict the luminosity will go down. However, this is not really a problem, due to the number of particles actually colliding being so much smaller than the beam total - meaning that we can easily allow ourselves to keep only the best parts of the beam.

- **Focusing magnets** - to further the minimisation of the β^*, new quadropoles will be installed both at the ATLAS and CMS experiment. As can be seen in Table 1.2, the β^* is expected to improve with a factor 2.7 when HL-LHC goes into operation.

- **Crab cavities** - inside the experiments of the LHC the two beams don't actually meet head-on, as this would cause downstream disruptions of the beam. However there are, at least two, strong reasons for keeping this crossing angle as small as possible. The beam-beam overlap reduces with the crossing angles, so a small angle maximises luminosity, and from the experimental point of view, it is preferable to analyse collisions which had a cancellation of opposite longitudinal momenta, such that the collision product seen in the detectors can be traced back to a point-of-origin effectively lying still - simplifying the track reconstruction.
  The introduction of Superconducting Crab Cavities is an attempt to somewhat circumvent this balancing act between minimal crossing angle and avoiding downstream beam disturbance, by giving each particle bunch a small transverse momentum kick right before the interaction point - A sketch of what this accomplishes can be seen on Figure 1.11

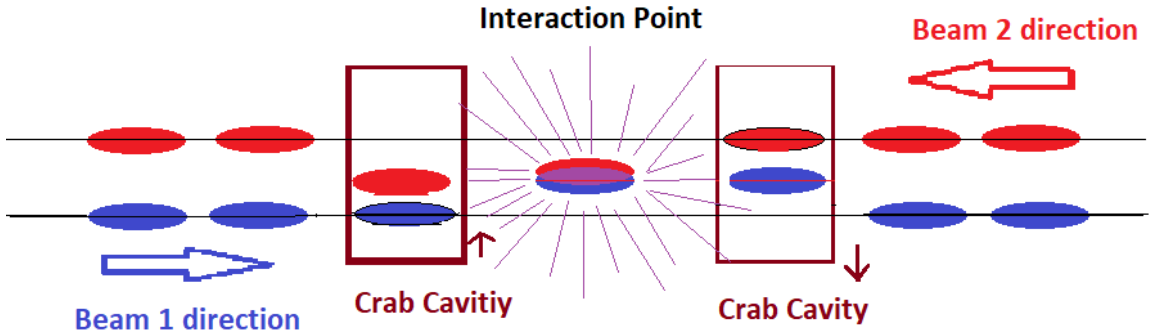


Figure 1.11: Sketch of how the introduction of a transverse momentum component to the beam, just before collision, can help minimise downstream beam disturbance while keeping the beam-beam overlap, and thereby the Luminosity, maximal. The purple lines represent collision products.

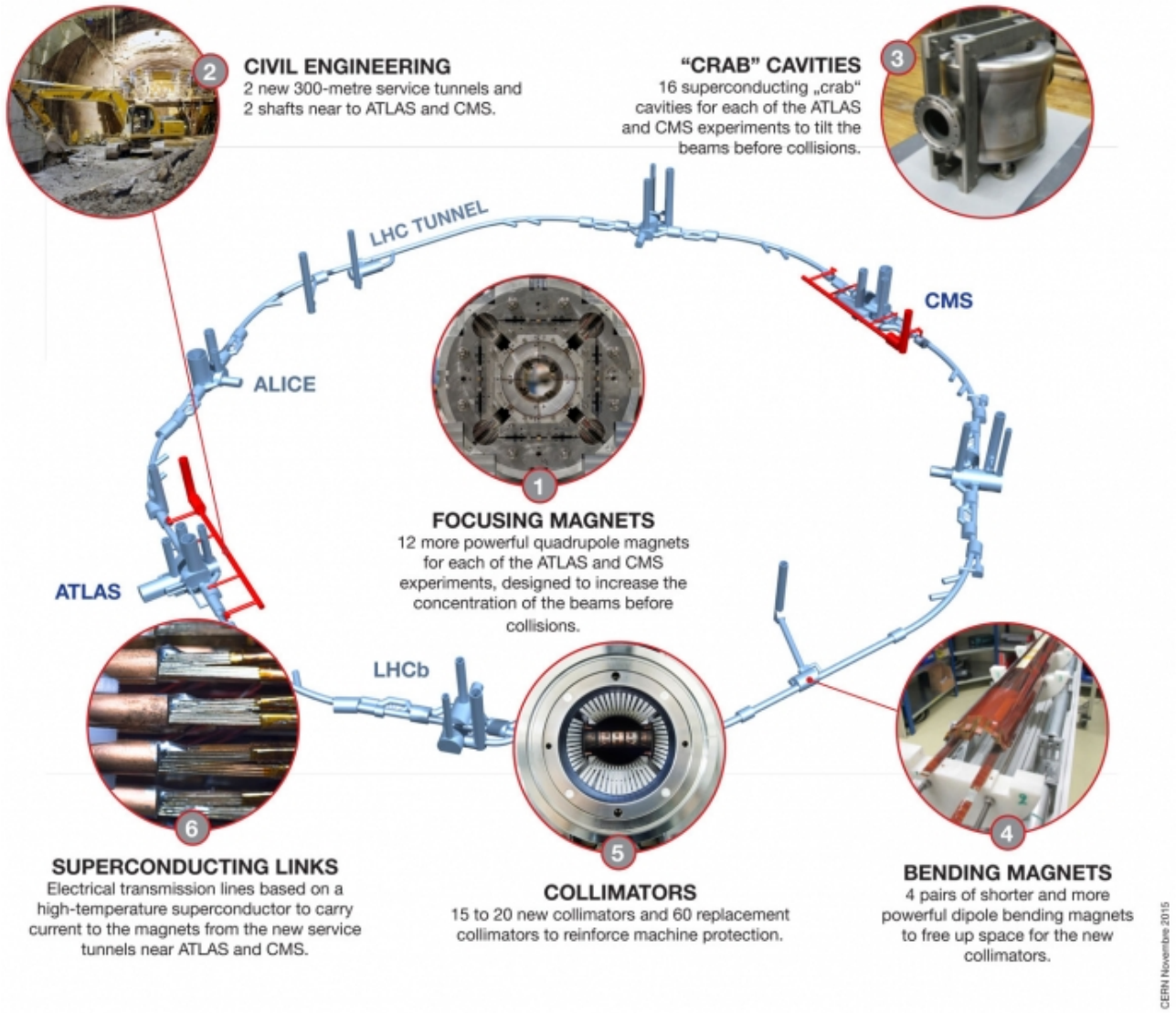


Figure 1.12: Illustrative overview of the different efforts pursued for the HL-LHC upgrade[13].

Parameter	LHC Run-1	LHC Run-2 & 3	HL-LHC
Beam energy [TeV]	0.45–4	6.5–7	7
Peak inst. luminosity [$cm^{-2}s^{-1}$]	$0.8 \cdot 10^{34}$	$(0.7–2) \cdot 10^{34}$	$5 \cdot 10^{34}$ (levelled)
Bunch distance [ns]	50	25	25
Max. number of bunches	1380	2028~2748	2748
β^* [cm]	60	40	15
ε_n [µm]	2.3	2.5–3.5 (2.3 with BCMS)	2.5
Max. num. protons per bunch	$1.7 \cdot 10^{11}$	$1.2 \cdot 10^{11}$	$2.2 \cdot 10^{11}$
Average pileup $\langle\mu\rangle$	21	21~50	140

Table 1.2: The table sums up the improvement in a few of the primary operational parameters going from the LHC to HL-HLC era. The different parameters listed are explained throughout the thesis.

A Toroidal LHC Apparatus (ATLAS)

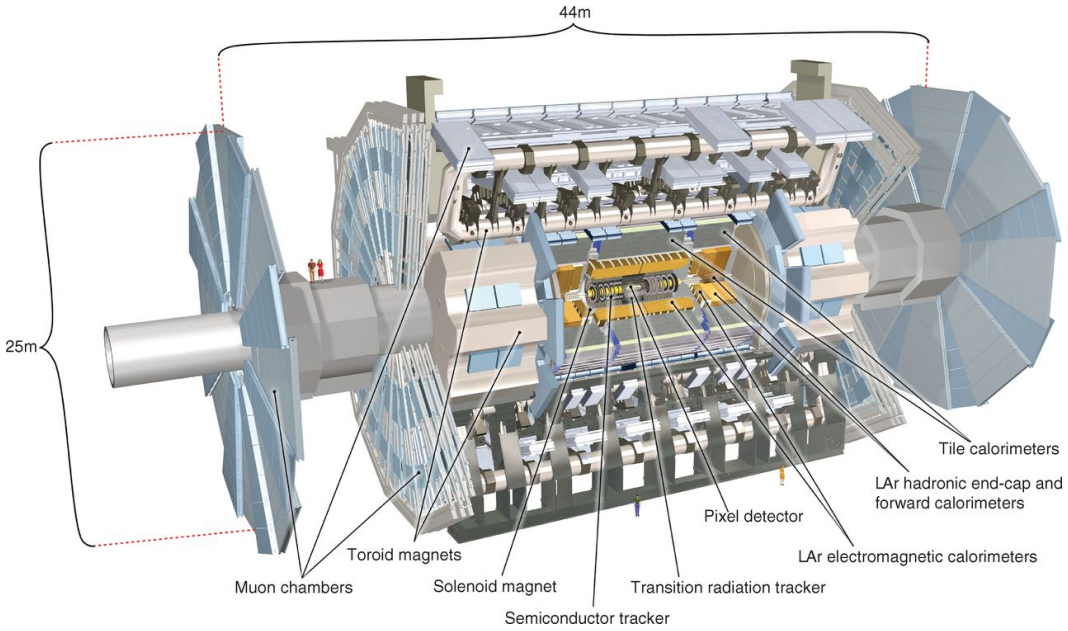


Figure 2.1: 3D animation of the ATLAS experiment showcasing the major components of the detector. Two humans are included for scale between the left-most muon wheel and the muon end-cap of the main body. This thesis is focused on the upgrade and replacement of the entire inner detector, consisting of the Pixel, Semiconductor and Transition Radiation Tracker.

THE ATLAS experiment operates one of two general purpose particle detectors, CMS being the other, attached to the Large Hardron Collider. Standing as a $7000\,\mathrm{t}$ cylinder, $44\,\mathrm{m}$ long and $25\,\mathrm{m}$ in diameter - roughly $100\,\mathrm{m}$ underground - it, along with the rest of the LHC complex, is an awe inspiring marvel of human achievement. The design philosophy was driven by the desire of ensuring either the discovery of the Higgs boson or some other new physics replacing the need of a Higgs mechanism for the Standard Model to work. This first goal was reached in 2012 when, together with CMS, they announced the joint discovery of a new particle consistent with the predicted Standard Model Higgs boson - making the Standard Model a fully self-consistent theory, albeit not capable of explaining external concepts such as dark matter. On Figure 2.1 a 3D model of the detector is sliced open to give an overview of the layered, onion-like, cylindrical structure of the machine and its massive scale.

2.1 Overview

The detector is built as a layered structure, each layer having a unique purpose w.r.t. measuring the products of proton-proton collisions and reconstructing the underlying physical event. The geometry of the detector is such that the layers are concentric and hermetic in the transverse plane w.r.t. the beam-pipe, extending $22\,\mathrm{m}$ along the beam-pipe in each direction from the interaction point. This ensures that only neutrinos and particles flying almost parallel to the beam-pipe escape detection - since you do need a gap in the detector coverage through which the beam-pipe goes in and out.

Referring to Figure 2.2, we'll now go through the different layers of the detector and give a brief overview of their purpose.

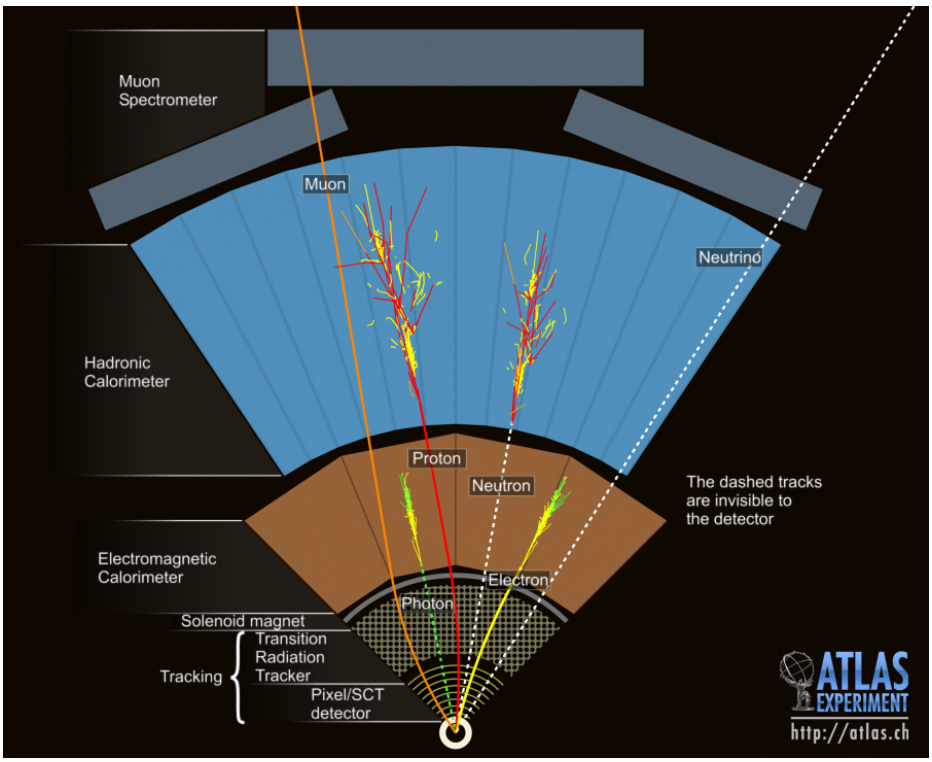


Figure 2.2: Diagram showing the different parts of the ATLAS detector and their functionality w.r.t. finding different types of particles. During a real collision one would see tracks from several hundreds of particles, all mixed together, in a detector slice of this size, due to each of the ~ 30 simultaneous collisions birthing several hundred offsprings during their propagation out through the detector layers

Inner Detector (ID) - The ID consists of three subsystems, the Pixel Tracker, the Semiconductor Tracker (SCT) and the Transition Radiation Tracker (TRT) - along with a solenoid magnet enclosing all of it in a uniform $2\,\mathrm{T}$ field parallel to the beam. The ID records the trajectories of charged particles, which are bend into helical orbits by the Lorentz force proportionally to their momenta. The Pixel Tracker and the SCT are both based on solid-state silicon sensors where planar pieces of silicon, $62.4 \times 22.4\mathrm{mm}^2$ for pixels and $63.6 \times 64.0\mathrm{mm}^2$ for SCT, are divided into electrically isolated sensing elements - either pixels of size $50 \times 300\ \mu\mathrm{m}^2$ or strips with a pitch (inter-strip distance) of $80\,\mu\mathrm{m}$ and a length corresponding to either the half or full sensor length [14].

Particle density decreases with distance from the interaction point, so in order to keep the reconstruction efficiency constant throughout the ID, there are higher requirements on spatial resolution the closer you get to the Interaction Point. However, it was and still would be prohibitively expensive to employ silicon pixel technology for the entire detector, which is why it is only used for the inner most parts of the detector, followed by the silicon microstrip tracker SCT and completed by the gas based Transition Radiation Tracker (TRT). The TRT is a proportional mode drift tube detector, using 300000 thin-walled tubes filled with a mix of Argon and Xenon, with the tubular wall and a central filament serving as the two electrodes of a read-out circuit.

In all three systems, signals are generated by charged particles ionising the medium it transverses - allowing electrodes to gather this ionised charge as an analog electrical signal. On Figure 2.2 we see how the the proton and electron have oppositely curved tracks due to their different electrical charge and that electrically neutral photon and neutron leave no track at all.

In terms of track reconstruction resolution, it is actually possible to achieve values smaller than the spatial granularity of the detector. This is because an event typically leaves a signal in several neighbouring detector cells, allowing for the reconstruction of the most likely event centroid - eg. by assuming an areal Gaussian spread and fitting this against the position of the cells we've associated

Track Reconstruction Resolution µm	ID pixel	SCT	ITk ($p_T = 1\ GeV$)	ITk ($p_T = \infty$)	TRT
Transverse	12	16	< 100	< 8	170
Longitudinal	66(77)	580	< 100	< 50	(3000)

Table 2.1: Track reconstruction resolution for the different parts of the current Inner Detector, compared with the requirements for the ITk, at a pile-up of 200. Disks are the ID equivalent of the ITk Endcap. For the pixel ID, the first longitudinal value refers to the resolution in the barrel and the second in the disks. For the TRT, the resolution is given as one total value per straw - with the parenthesised value being an estimate of the total error based on the minimal amount of straws encountered by a purely transverse event [14] [10].

with a given event. The achievable resolutions depends on many factors, eg. the transverse momentum p_T of the event, because higher p_T events are more distinct from the low energy background of hadron collisions. In Table 2.1, a summary of typical values for the track reconstruction resolution achievable with the current ATLAS ID and the corresponding overall requirements for the ITk to keep the same or better levels of reconstruction efficiency at pile-up levels $7 - 10$ times higher than the limits of the ID.

Calorimeters - Calorimeters function by absorbing the energy of particles and converting it into a signal proportional to the particle energy. This can be done in many different ways but typically a type of scintillator is used. High energy particles lose energy through atomic excitations of the scintillating medium, either via strong or electromagnetic interactions. These excited atoms decays down to their ground state through photon emission, and because the scintillator is transparent to the emitted photon wavelength, these can be collected by photomultiplier tubes (PMTs) and converted into an analog electrical signal, where the amount of induced charge can be directly mapped to the amount of energy deposited in the scintillator by the traversing particle.

When designing a calorimeter, it very important that it is significantly thicker than the mean free path of the particles you hope to collect - to maximise the probability of the particle being fully absorbed in the calorimeter. However, the rate of energy loss and the other underlying processes determining the mean free path differ significantly amongst the particles being created in these collisions. This is why the outer parts of the ATLAS detector is split into three parts, the Electromagnetic Calorimeter (ECAL), which fully absorbs electrons, positron, photons and possibly light low energy mesons like the pion. The Hadronic Calorimeter stops all larger particles except for muons which typically exit ATLAS after interacting with the Muon spectrometer, coincidentally the biggest part of the detector due to the very large mean free paths of muons. It is certainly worth the effort of building the very large Muon spectrometer, because muons provide some of the cleanest experimental signatures to look for, meaning that muon tracks are fully unique and easily identifiable. As seen on Figure 2.2, they leave a tightly contained signal in every part of the detector, due to their low interaction rate with the detector materials, and this makes it much easier to extract accurate information about the event creating the muons.

Besides the solenoid magnet providing the B-field for the inner detector, a set of toroidal magnets are placed along side the muon chambers to soak the rest of the detector volume in a uniform beam parallel $4\,\mathrm{T}$ field. This split magnet concept of enclosing the detector with toroidal magnets also serves as the excuse to make the rather forced acronym of A LHC Toroidal ApparatuS work.

Table 2.2: Listing some of the key operational parameters for the current ATLAS Inner Detector, and the upscaled requirements going into the HL-LHC era of data taking. [16].

Operational Parameter	**ID, LHC, and ATLAS Design Limits**	**HL-LHC Capabilities and Upgraded ATLAS Requirements**
Experiment Lifetime	*10 years*	Current age: *~8 years* Age at upgrade: *14 years* Full duration: *27 years*
Peak Instantaneous Luminosity ($x\ 10^{34}\ cm^{-2}\ s^{-1}$)	*1.0*	Nominal: *5.0* Ultimate: *7.5* **with leveling*
Pileup (*proton collisions per 25 ns bunch crossing*)	Design: *23* Operational Max: *~40 (peak) , ~24 (avg.)*	Nominal: *140* Ultimate: *200* **with leveling*
ATLAS Trigger Rate (*kHz*)	Level-1: *100*	Single Mode; Level-0: *1000* Dual Mode; Level-0: *4000* Level-1 (new): *400-600*
Integrated Luminosity (fb^{-1})	Pixels: *400* SCT: *700* Inserted Beam Layer: *850*	Nominal: *3000* Ultimate: *4000*
Maximum High Energy Particle Fluence ($x\ 10^{15}\ 1\ MeV\ n_{eq}\ cm^{-2}$)	Pixels: *5.0* SCT: *0.2*	ITk Pixel: *18.7* ITk Strip: *1.2*
Maximum Total Ionizing Dose (*MGy*)	Pixels: *3.0*	ITk Pixel: *12.7* ITk Strip: *0.5*

2.2 The ATLAS Inner Tracker (ITk) Upgrade

The ATLAS detector was designed to operate for ten years, being able to properly resolve events with pile-up levels of $20 - 40\ events/collision$. The Inner Detector (ID) was commissioned to cope with the radiation damage induced by collisions products up to integrated luminosities of respectively 400 and $700\ \mathrm{fb}^{-1}$ for the pixel and SCT tracker. To understand the need for an upgrade, we will first need to cover some definitions of nomenclature related to describing high energy particle detectors:

Occupancy ratio - The occupancy ratio refers to the fraction of sensor channels being activated during a bunch crossing. If the occupancy is $100\,\%$, eg. due to inadequate spatial granularity, the entire detector is being lit up simultaneously, and it would be fully impossible to isolate and identify any of the individual collision products. Occupancy is often split up into hit and noise occupancy, and to retain the required hit-finding efficiency of $> 99\ \%$ throughout the lifetime of the experiment, one or two orders of magnitude in difference between the two occupancies is needed. Said in other words, the Signal to Noise Ratio (SNR), which will degrade over time due to radiation damage, should never go below $10 : 1$. Estimates of the ITk end-of-lifetime SNR place it to be roughly twice this minimum value[17].

Reconstruction efficiency - The track reconstruction efficiency is a measure of how often we can accurately reconstruct tracks of a collision event. It is often evaluated through complex Monte Carlo

simulations, where a series of known events are injected into a detailed full detector simulation - followed by a simple count of how often the detector's reconstructions are correct. Many different types of reconstruction efficiencies are used within the detector community, sometimes with overlapping definitions, so it would be a lesson in futility to try and properly define them all.

Radiation damage - The accumulated radiation damage of material is typically evaluated in terms of two different quantities, the Total Ionising Dose (TID) and the Non-Ionising Energy Loss (NIEL):

The TID dose describes the amount of energy deposited in the material due to electromagnetic interactions, and it causes primarily surface damage to the structure. Electronics in radiation heavy environments will experience an initial increase in noise current, the TID bump, followed by a steady state situation, with no significant change in performance as the TID dose accumulates. This TID bump simply needs to be accounted for during the design phase, and is rarely a major limiting factor of operation.

The NIEL can be understand as a ballistic deterioration of the atomic lattice, where collisions with traversing particles displaces atoms from their place in the crystal. The extent of damage caused by NIEL depends both on the total amount of collisions, given by the fluence (time integrated radiation flux), and the momentum of traversing particles - since a more energetic particle can cause more dislocations. This is why NIEL damage is typically quantified in terms of $1\,\mathrm{MeV}$ neutron equivalent fluence - with neutrons chosen as the common reference point, because they are generally the most damaging type of particle encountered. The accumulation of NIEL damage is what limits the lifetime of silicon sensors - something that is elaborated upon in Section 5.2.

As a quick side note, material tolerance of radiation damage, for a given type of particle, is typically listed in terms of an empirically determined constant, *radiation hardness*, with higher values signifying less susceptibility to damage from this type of particle.

The material deterioration of silicon sensors can, to some extent, be counterbalanced by increasing the biasing voltage used to operate the sensor. But in semiconductor devices like these, there's a hard upper limit on the biasing voltage - due to the breakdown mechanisms explained in Section 5.1.3. Also, an additional problem encountered as radiation damage accumulates, is the risk of thermal runaway - where the heat generated by the increasing leakage current exceeds the cooling capacity of the system.

This means that the operational window for biasing the sensors narrows over time, until the sensor becomes useless - the reconstruction efficiency will die out as the Signal-To-Noise ratio becomes too small and the occupancy ratio too high. After the end of Run-3, in 2023 it is expected that the Inner Detector will be too damaged for further operation.

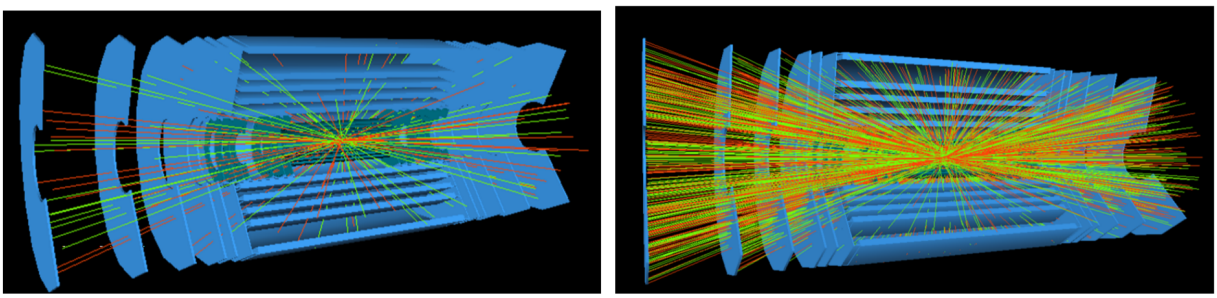


Figure 2.3: Simulation of the data signals generated in the inner tracker during current LHC conditions, with a pile-up of 23 events (left), and the ultimate HL-LHC environment with 230 events (right). This demonstrates intuitively how the difficulty of accurate reconstruction scales with the pile-up of collisions. The detector layout is an old ITk model[15].

Because of this, it is mission critical that new sufficiently radiation hard sensors are developed. They should perform equivalently or better than the current Inner Detector, in terms of track reconstruction efficiency, while operating in the much harsher HL-LHC collision environment, collecting an estimated $3 - 4000\text{fb}^{-}1$ worth of integrated luminosity over the 10 year runtime of the HL-LHC. Due to the increased particle flux, visualised on Figure 2.3, a better spatial resolution and a quicker read-out system, operating at higher bandwidths, is required to keep the occupancy ratio sufficiently low. This is why the entire current Inner Detector will be fully replaced by a new fully silicon based tracker named the Inner Tracker (ITk) - a model of which can be seen on Figure 2.4. The ITk is based on two different types of sensors, micropixel technology being used closest to the interaction point, to maximise the detection resolution and minimise the signal occupancy. Further away, where the collisions products are more spread out in space, microstrip technology is used - a more cost efficient option allowing coverage of a greater active volume while still keeping the occupancy sufficiently low - a requirement for high reconstruction efficiencies.

To summarise just how incapable the current ATLAS ID would be at operating throughout the HL-LHC campaign, we refer to Table 2.2, comparing the design limits of the ID with the requirements for HL-LHC operation.

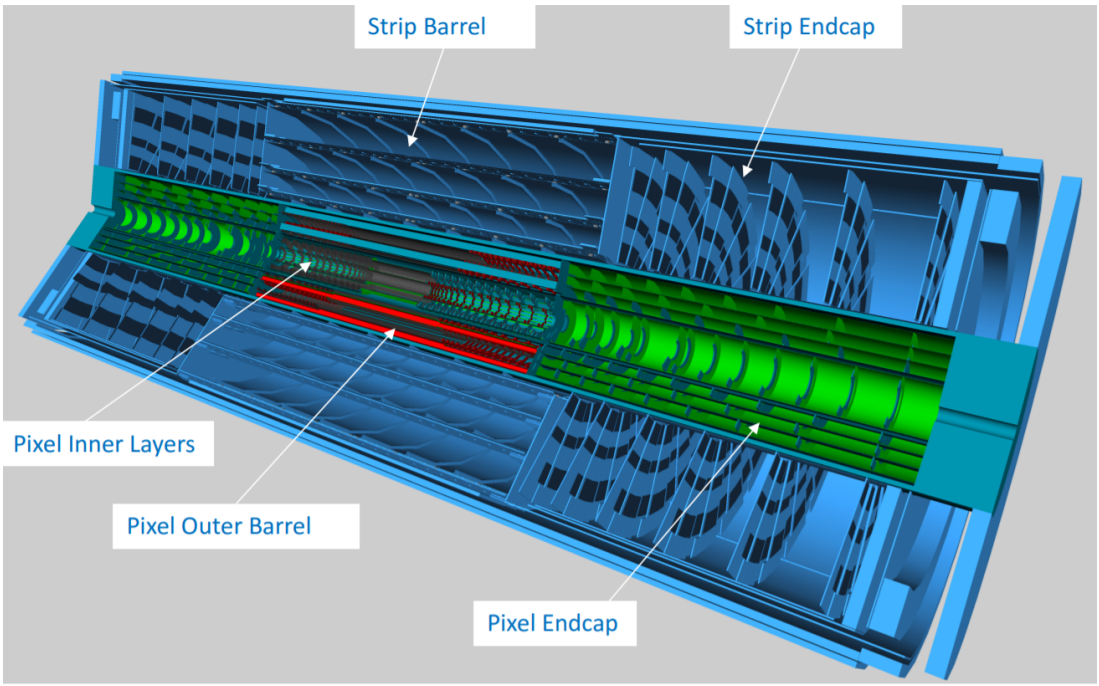


Figure 2.4: 3D model of the entire ITk detector. The detector is geometrically split up into a barrel section and two end-caps, and it will employ two different types of silicon sensors. Pixel based technology, offering the best spatial resolution, is used closest to the interaction point, while microstrip technology is sufficient further away [21].

Figure 2.5 shows a diagram of the division into micropixel and strip sensors, shown for one quadrant of the detector, with the rest of the detector being mirror images of this layout - to ensure the rotational symmetry of the detector. The micropixel and strip tracker have quite different technical requirements, so for the sake of efficiency, it was decided to completely split the development and production of the two trackers into separate task forces - overlapping only when strictly necessary. Furthermore, within each of the two task-forces, there is a division based on the geometrical structure of the ITk into an end-cap and a barrel community. However, there is a much greater level of cooperation between eg. the barrel and end-cap strip communities, since they share a greater load of technical challenges compared to the pixel vs strip division. I have been solely involved with the

development of the end-cap microstrip tracker throughout this thesis, and as such, there will be very sparse mention of pixel sensors and the barrel going forward.

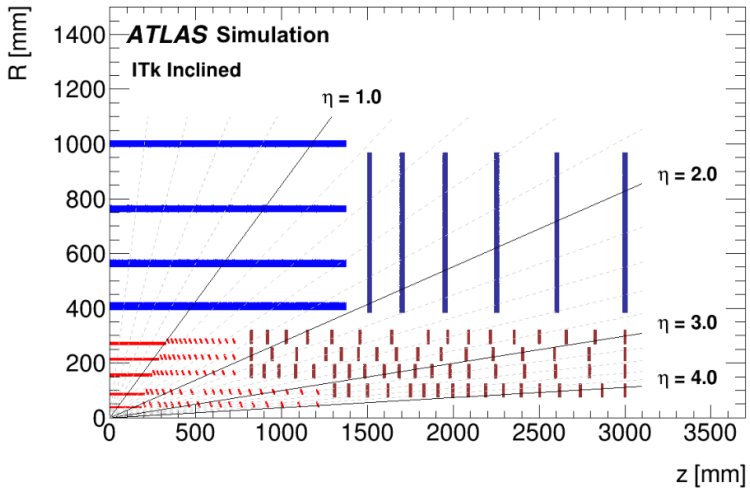


Figure 2.5: Showing the structure of the upper right quadrant of ITk. Origin is placed at the interaction point (IP), the Z axis runs parallel with the beam-pipe, R is radial distance from the IP. Reds are pixel detectors and blues are strip detectors. η is the pseudo-rapidity, a convenient coordinate transformation of the polar angle which is infinite when parallel to the beam-pipe and zero when orthogonal to the beam-pipe. [10, chp 3].

2.2.1 Overview of the End-Cap Strip Tracker

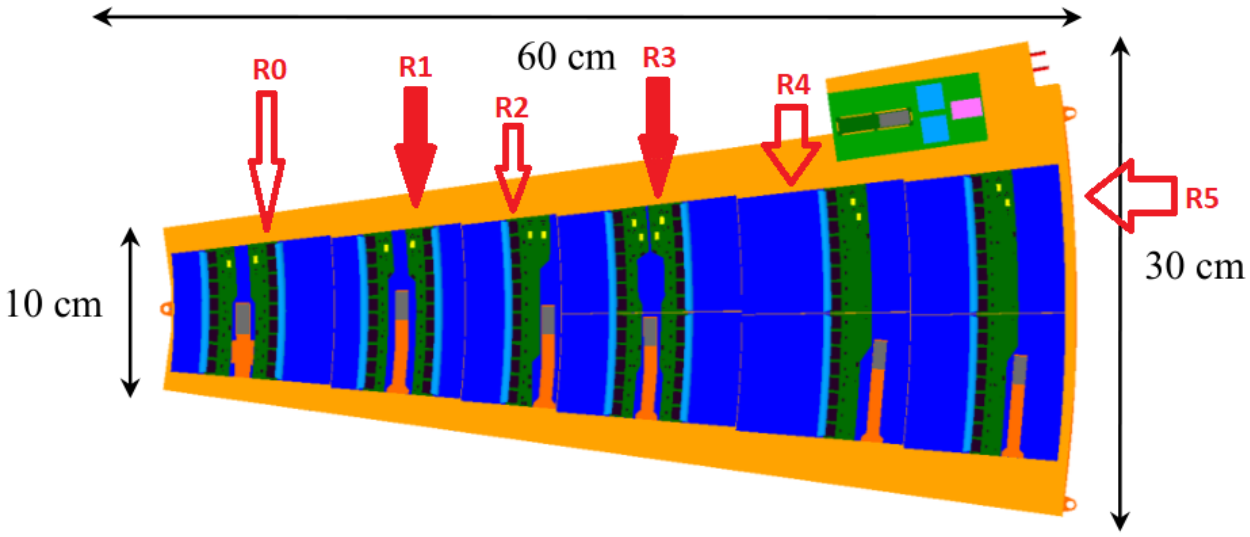


Figure 2.6: Showcasing the concept of a petal. The six rings of the end-cap are split up into petals, with each petal consisting of back to back single-sided modules named R#, differing in their proximity to the beam-line. The device sticking out on the top side is the End Of Structure (EOS) module, which serve as a communication link between the sensors of the petal and the detector as a whole. The Scandinavian Cluster will produce R1 and R3 modules.

As can be seen on Figure 2.4, the two parts of the end-cap consist of six wheels with sensors mounted on both sides of the wheel support structure. The sensors are mounted such that the strips

points towards the radial center of the ring, except for a $40\,\mathrm{mrad}$ offset in the stereo angle between the back-to-back sensors, allowing for a precise measurement of the second transverse coordinate - along the elongated direction of the strips. Intuitively one might think that a stereo angle of $90\,\mathrm{deg}$ would be ideal, but in the case of simultaneous traversing particles, this would lead to ambiguities in particle localisation, due to the much higher number of intersections between strips at this angle - Illustrated on Figure 2.7.

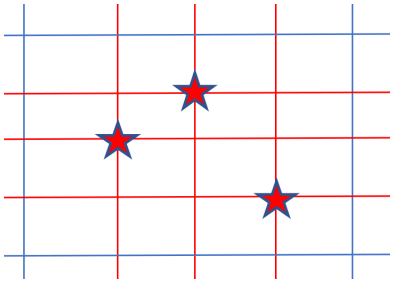


Figure 2.7: Illustrating the concept of ghost hits. The stars represents traversing particles and the red lines are the strips activated by these three particles. Any intersection between two red strips could potentially be a real measurement, there's no way for us to distinguish between the six "ghost hits" and the three real ones. The amount of intersections, and thereby the probability of seeing ghost hits reduces with the stereo angle.

To keep the occupancy below ~ 1 % throughout the detector volume, the strip length decreases with proximity to the IP, being $19.2\,\mathrm{mm}$ closest to the IP and growing to $60.2\,\mathrm{mm}$ furthest away. This, combined with the circular nature of the end-cap, resulted in the development of six different module geometries, the seemingly best solution for handling the increase in circumference with radial distance - while keeping the coverage hermetic.

Each wheel of the end-cap is split up the in azimuthal direction into smaller blocks named petals - see Figure 2.6. These petal have 6 modules on each side, using a total of 9 sensors per side, since some of the modules contain two sensors. The modules are the basic building blocks of the ITk, containing a sensor and the required electronics to operate it, with the electronics mounted directly onto the surface of the sensor. It might seem counter-intuitive to place the electronics on top of the sensor, but it is actually a clever way of optimising the material budget, by utilising the mechanical rigidity of silicon. The onboard electronics needs to be mounted somewhere, so by using the sensor as the mounting structure, we avoid having to implement additional "dead" material used to support the electronics - thereby helping minimise the material budget of the detector. The material budget of a detector is always optimised to be as low as possible, since any interactions with collision products outside of the active sensing volume, be it tracker or calorimeters, results in lost or misrepresentative information [10, chp 3&5].

However, no matter where the on-board electronics are located, one has to consider the probability of collisions products interacting with the registers and memories of the Application Specific Integrated Circuits (ASIC)s used in the read-out electronics. This can lead to digital bit flips in the memory bank or data flow of the ASIC, a type of stochastic radiation damage called Single Event Upsets (SEU), which, over time, can completely corrupt the programming of the device - rendering it non-functional. There are two primary ways of combatting this phenomenon in the ITk: developing sufficiently radiation hard ASIC's, such that the rate of bit flips is kept at a manageable level - while at the same time fully reconfiguring the ASICs every so often from an off-detector repository. For the current Inner Detector, this reconfiguration is done roughly once a second during operation.

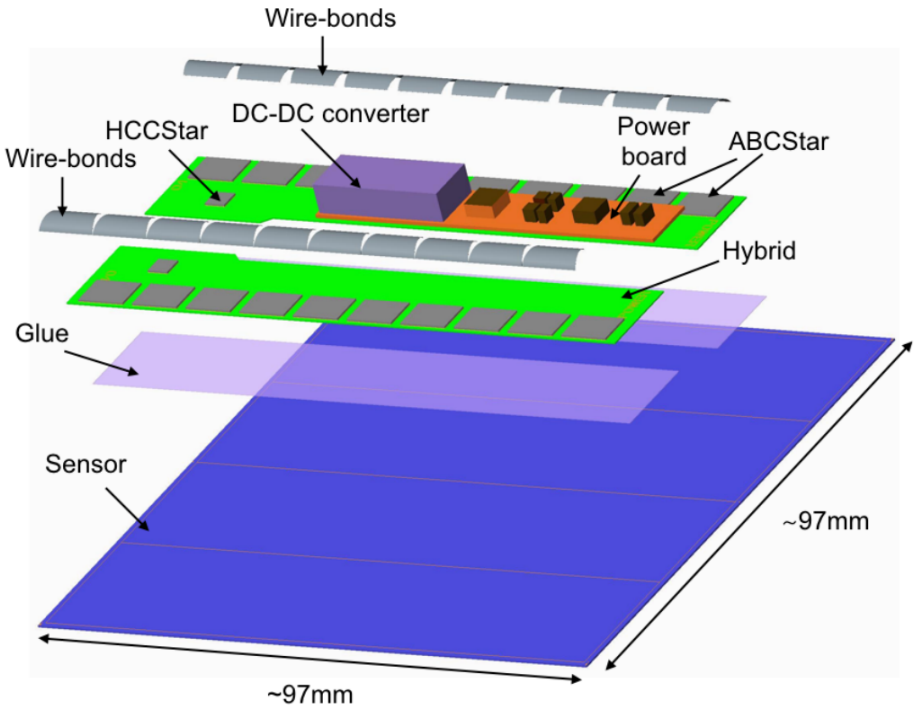


Figure 2.8: Schematic representation of the components of a barrel type strip module. ASIC's are attached to the hybrid, and the hybrid is glued to the sensor surface - along with the powerboard. The end-cap modules are designed in a slightly different manner, with the powerboard being on a separate PCB, independent from the hybrid(s) [10].

2.3 The ITk Microstrip Module Effort

The fundamental unit of the ITk detector will be the module, see Figure 2.8, which is a structure consisting of:

- At least one silicon sensor, using a microstrip geometry, based on n-on-p lithographic technology.
- An on-board electronic readout system, called the hybrid, consisting of a PCB with two types of ASICS attached, the ATLAS Binary chip (ABC,) interfacing directly with the sensor, and the hybrid Control Chip (HCC), which processes the data signals from the ABCs and handles external communication with the EOS board. One or two hybrids are used per sensor, with a naming scheme of "R#H##" being used when labelling hybrid types. R# refers to the sensor type, and H## is added in cases of multiple hybrids per sensor - eg. the R0 module has two hybrids, the R0H0 and the R0H1.
- A dual function powering and monitoring board, the powerboard, it supplies the powering for all module components and monitor their vitals.
- a High voltage line (HV tab) for the reverse biasing of the silicon sensor. It is not present on Figure 2.8 but would be attached to the aluminium backplane of the sensor.

2.3.1 The Silicon Sensor

To avoid unnecessary repetition only a brief feature overview of the silicon microstrip sensor used for the ITk are given here, with a more in-depth characterisation given in Chapter 5. The sensor uses very pure high-resistivity silicon as a base material for creating an array of, sufficiently, electrically isolated

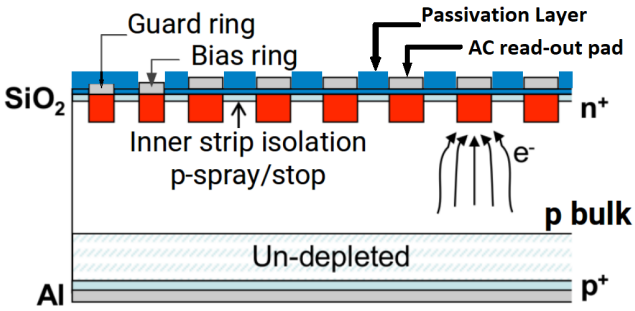


Figure 2.9: Cross section of the primary features of the n-on-p sensors used for the ITk. The passivation layer is a protective oxide layer covering the entire surface except for slots at the strip probe & bonding pads as well as the biasing pads of the bias ring - to allow for proper electrical contact. This 3-D feature is difficult to convey through a 2-D drawing and is therefore mentioned explicitly.

pn-junctions. This is done through a lithographic doping scheme performed at single micron precision in placement. The pn-junction array is then operated in a collective reverse biasing mode, to achieve a full depletion of the structure, thereby minimising the free charge density of the semiconductor in equilibrium. When an energetic charged particle traverses the reverse biased sensor, it will ionise electrons from the valence band into the conduction band, typically around $\sim 10^4$ electrons for a Minimum Ionising Particle (MIP). This localised free electron cloud then drifts to the nearest positive contact terminal, see Figure 2.9, were it will induce a corresponding amount of charge across the capacitor like structure of the AC read-out pad. This charge is then funnelled through the attached electrodes into the read-out circuit, see Figure 2.11 - starting with the generation of a voltage pulse in the amplifier - proportional to the amount of charge ionised in the first place.

For the ATLAS12EC R0 sensor, a prototype design of the ITk microstrip sensor, the pn-junction array is in the form of four segments of strips, $\sim 3\,\mathrm{cm}$ long and $18\,\mu\mathrm{m}$ wide, with an inter-strip distance, or pitch, of 73.5 to $84\ \mu\mathrm{m}$ increasing radially - resulting in 4360 unique channels across the $\sim 90\,\mathrm{cm}^2$ active area of the sensor.

The silicon sensors of the current ATLAS ID are based on p-on-n technology, but the entire ITk, both pixels and strips, will instead be based on n-on-p technology, meaning that the silicon bulk is p doped while the strips(pixels) are formed using localised n+ doping. n-on-p is more radiation hard than the p-on-n doping scheme, but twenty years ago, during the design and production of the ATLAS ID, it was not feasible to utilise n-on-p technology for such a herculean project.

In Chapter 5, we will go more into detail on why n-on-p is more radiation hard than p-on-n.

2.3.2 The Onboard Electronics

ABC ASIC - The ATLAS Binary Chip (ABC) is the ASIC directly connected to the sensor - a picture of the ABC130 version of the chip can be seen of Figure 2.10. Each ABC has 256 input pads which are wirebonded to individual microstrip channels on the sensor. Due to the SiO_2 layer between the silicon and the aluminium strip bond pad, the sensor is capacitively coupled to the read-out ASIC, suppressing the steady-state leakage current from being read-out, by acting similar to a high-pass filter. The steady state leakage current is drained via the biasing resistor, out through the bias ring, while the transient charge pulses pass across the Metal-Oxide-Semiconductor (MOS) structure into the amplifier of the ABC. The voltage pulse generated in the amplifier is then shaped and discriminated into a binary output, meaning that we only end up with information on whether, at a given time, there was a voltage pulse higher than a certain threshold value in a given strip channel. The final version of the chip, the ABC*, will use discriminator values of 0.5-$1\mathrm{fC}$ in operation. If the full signal pulse was read out, instead of just a binary value, one could also gain information on the energy of the particle. The

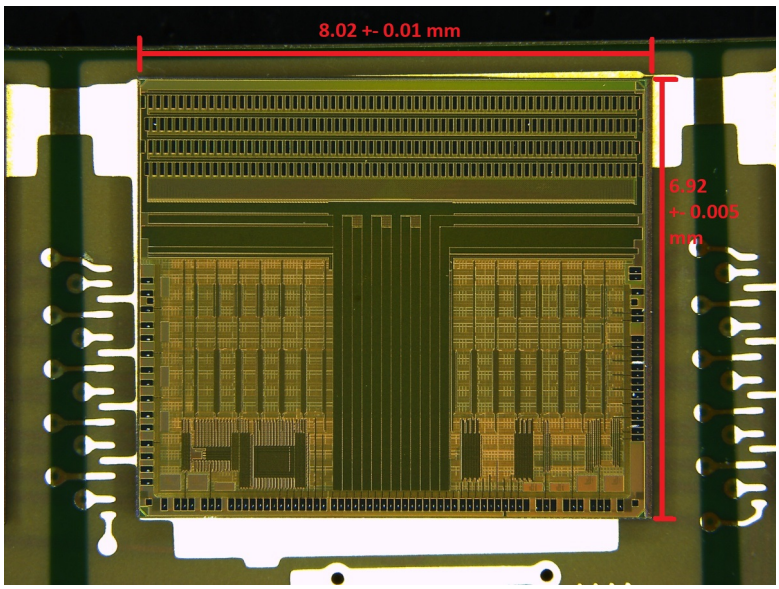


Figure 2.10: Picture of the ABC130 mounted on a R0H0 hybrid, taken under a microscope. The top four rows of pads are where the 256 sensor strip channels will be connected using aluminium wire-bonds $25\,\mu\mathrm{m}$ in diameter. The dimensions are $8.02 \times 6.92\ \mathrm{mm}^2$.

primary reason for using a binary read-out is to minimise the bandwidth requirements for the read-out electronics. One could consider simply increasing the amount of onboard read-out electronics to gain the needed bandwidth for a non-binary read-out system. However, then we would run into problems w.r.t. increased multiple coulomb scattering, obscuring the tracks of collision products, making them harder to reconstruct. Also, secondary particle creation, from interactions between collision products and the detector material, would effectively add additional noise when looking for high energy collisions. This is obviously quite undesirable, which is why so much effort is spend on minimising the non-active material budget of the detector - eg. through the use of a binary read-out system for the tracker.

A simplified diagram showing the analog part of the ABC circuitry can be seen on Figure 2.11, with an example of what is not shown being the complex array of buffering and trigger logic also implemented. The ABCStar will be capable of holding up to 86 events for $128\,\mu\mathrm{s}$ in memory, while the triggers evaluates if the event information is worth passing on or if it should be dumped. Each ABCStar will be capable of transmitting data to its communications relay at rates up to $160\,\mathrm{Mb/s}$ [10, Chp. 6.2].

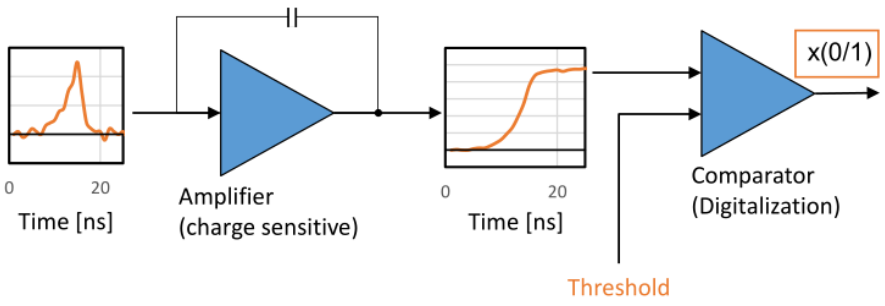


Figure 2.11: Simplified circuit diagram of the analog functionality of the ABC chip. It takes in the raw electric charge generated in the sensor, amplifies and shapes the input to a square pulse based on its peak height and then compares it to a threshold value to produce a binary output signal.

HCC - The hybrid Control Chip, See Figure 2.12, is the onboard communication relay between the individual module and the EOS, which then connects the petal to the detector as a whole. ATLAS does not want to rely on partial read-out, so, every time the hardware triggers are activated somewhere in the detector, a read-out command is propagated out to perform a read-out of the entire detector. This is an example of why buffers are needed to temporarily store data, namely to allow for the propagation time of potential read-out signals to reach any individual module from anywhere else in the ATLAS detector. HCC*, the production version of the chip, has, amongst other things, an onboard $40\,\mathrm{MHz}$ clock, matching the collision frequency of the LHC [10, Chp. 6.2].

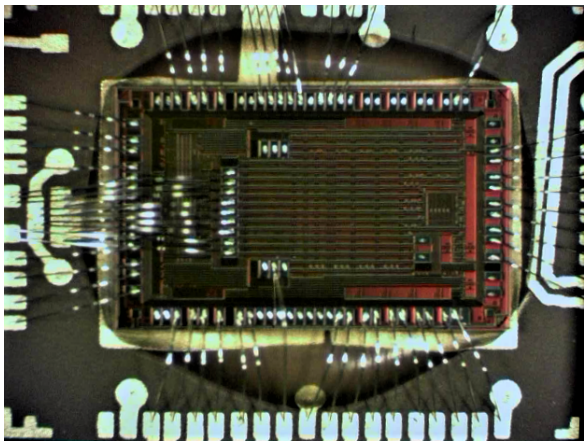


Figure 2.12: Picture of a wirebonded HCC130 chip mounted on a R0H1 hybrid - the dimensions are $4.75 \times 3.05\ \mathrm{mm}^2$. The wirebonds are made of aluminium and are $25\,\mu\mathrm{m}$ in diameter.

Powerboard - By reusing the present low voltage supply lines in the current Inner Detector, significant savings in terms of cost, complexity and time can be achieved. This is why the powerboard has an onboard DC-DC converter, down-stepping the incoming $12\,\mathrm{V}$ supply to eg. the $1.6\,\mathrm{V}$ each HCC and ABC need to operate. The Autonomous Monitor And Control chip (AMAC) also monitors the leakage current in the range $10\,\mathrm{nA}$ to $5\,\mathrm{mA}$, and module temperature through a NTC thermistor with a $0.5\,^{\circ}\mathrm{C}$ resolution.

As a preventive measure against SEU's, all essential programmatic logic are designed in a triplicated manner. Whenever a function is called, all three versions can be loaded and compared, if two versions are identical but not the third, it is probably due to a SEU, and the erroneous logic will be overwritten with the correct version. It is very unlikely that different SEUs affect multiple versions of a logic function with identical errors - and it has been proved in test-beam campaigns to be a very effective counter measure to the SEU problem. This scheme of triplicated essential logic is employed in all of the onboard ASIC's, not just the AMAC.[10].

HV-tab The high voltage (HV) circuit of the module consists of the following: A kapton enclosed aluminium flat-wire $2 - 3\ \mathrm{mm}$ wide called the HV-tab, which will be ultrasonically welded to the aluminium backplane of the sensor, and with the circuit leaving the module through the HV switch located on the powerboard. Due to its delicate nature, the presence of the HV-tab complicates handling during module assembly quite a bit. In the current ATLAS SemiConductor Detector (SCT), the high voltage is supplied to the sensor backplane through an electrically conducting silver epoxy, but this is no longer a viable solution, because of too high contact resistance across the backplane-epoxy interface.

2.4 Quality Control

Ongoing quality control (QC) is an essential part of any large scale production. The Scandinavian Cluster performs as both a hybrid and module assembly site, and as such will have to implement the required QC procedures related to both of these production efforts. The cluster will receive silicon sensors, fully functional Powerboards, hybrids with SMD components attached and the two types of ASICs to be mounted on the hybrids. An overview of the production flow from individual components into fully qualified modules can be seen on Figure 2.18 - prefaced by a text explaining the different QC procedures employed during production.

Visual Inspection - At every step of production the first QC evaluation is a thorough visual inspection by eye or microscope, looking for obvious faults such as

- Missing or damaged Surface Mounted Devices (SMD) on the hybrid - eg. resistors and capacitors.
- Glue leakage covering the guard ring of the sensor or any of bond-pads on the hybrid or sensor.
- Physical contaminants, eg dust, on components.
- Missing or loose wirebonds.

Metrology - Metrology refers in this context to measurements performed using a Smartscope - a non-contact optically based machine capable of measuring position in all three axes with single micron precision. It is used to evaluate eg. the accuracy and precision with which components are placed during the assembly processes.

Sensor Exclusive QC Upon reception, the sensor is visually inspected for any obvious signs of mechanical damage. Then an IV curve is performed, measuring the leakage current as a function of the reverse biasing voltage, typically up to the maximum operational voltage of $600\mathrm{V}$, with a successful IV curve showing a leakage current density $I \leq 10^{-7}\frac{A}{cm^2}$ [10]. An IV curve is a very simple and efficient diagnostic tool, establishing if a sensor is within specifications, with the typical causes of failure including but not limited to:

- improper edge isolation, allowing current to bypass the pn junction and run from the backplane up through the edge and out through the surface readout pads.
- insufficient inter-strip isolation, short circuiting neighbouring strips.
- surface contamination, either in the form of loose particulars, eg. dust, or uncontrolled absorption of substances leading to altercations of the electrical properties of the silicon. It is theorised that this is how high humidity increases the leakage current.
- mechanical damage in the form of eg. scratches or errors in the dopant profiling during sensor production.

Since we know very well, from a theoretical point of view, how the IV curve of a perfect reverse biased pn-junction should look, too large deviations from this expected behaviour tells us that the structure is no longer operating as a pn-junction, and as such can't be used as charged particle tracker.

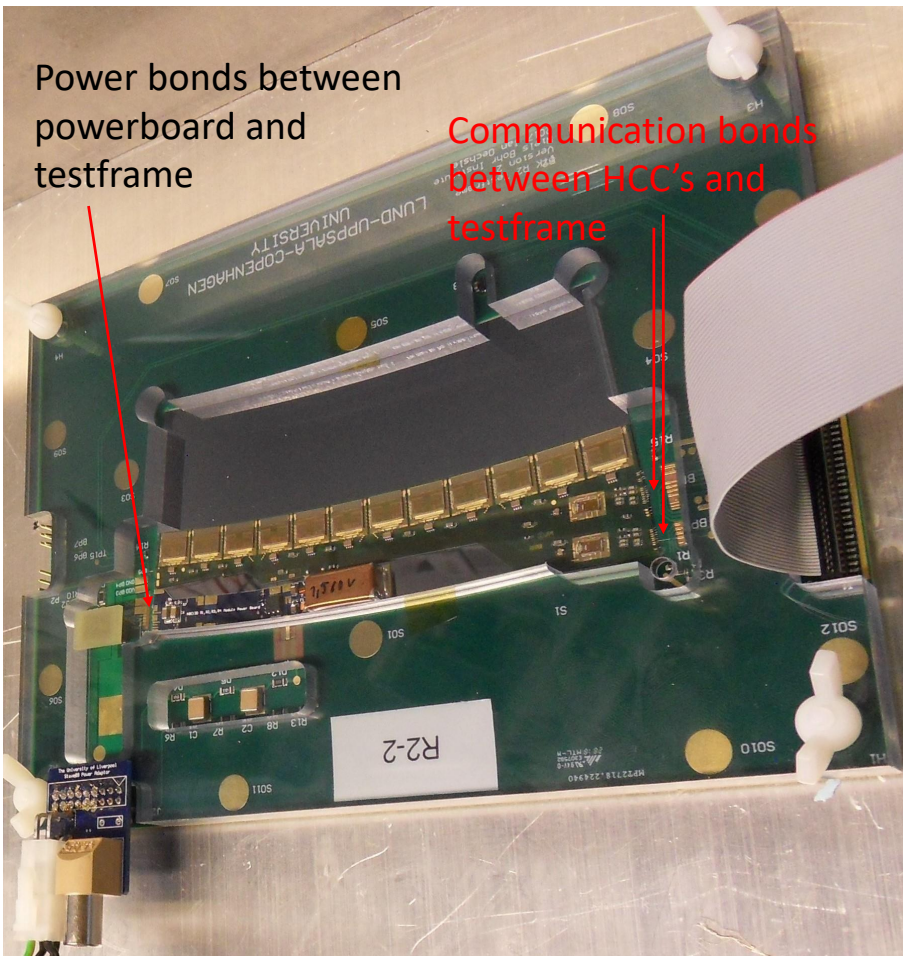


Figure 2.13: One of two R2 semi-electrical modules produced in the Scandinavian Cluster in early 2019, undergoing electrical testing. The cable on the left supplies the power while the other is a communications cable connecting to a FPGA board - which then interfaces through an Ethernet cable to a computer.

2.4.1 Electrical Tests

After ASICs have been bonded to the hybrid, and after the hybrid has been bonded to the sensor, a sequence of electrical calibrations and tests are performed. This is done to optimise and equalise the behaviour of the module across all of its channels, and to evaluate important operational parameters, like input(output) noise levels and signal gain - meaning the amplification factor used to boost an input signal before it is fed into the discriminator - see Figure 2.11.

These tests are performed by attaching communication wirebonds between the hybrid or module and a test-frame, See Figure 2.13, to enable communication between the device and a computer running the custom made testing software ITSDAQ. To perform these tests, an individualised configuration file is used to instruct ITSDAQ in how it should attempt to communicate with the device. However, this is production testing, meaning that things are bound to not always work, and in those cases it can be very difficult to properly troubleshoot the system. A communication failure can be due to a hardware error, eg. under/over powering of the device, missing, misplaced, broken or touching(short-circuiting) wirebonds or simply a loosely connected communication cable. It could also be due to a software error like, a wrongly tuned configuration file, which has a very long list of ways to mess up, bugs in the setup of the High Speed Input/Output (HISO) com's port, connecting the PC to the Field Programmable Gate Array (FPGA) board through a $100\,\mathrm{Gbit/s}$ ethernet cable - or some convolution of said hardware and software errors.

As such, a lot of time and effort has been spent on careful and meticulous troubleshooting of

the ITSDAQ setup in Uppsala, to enable us to successfully test the hybrids we produce at NOTE and in-house during the prototyping of our production procedures. A brief summary of the different electrical tests, how they function and what their purpose is follows.

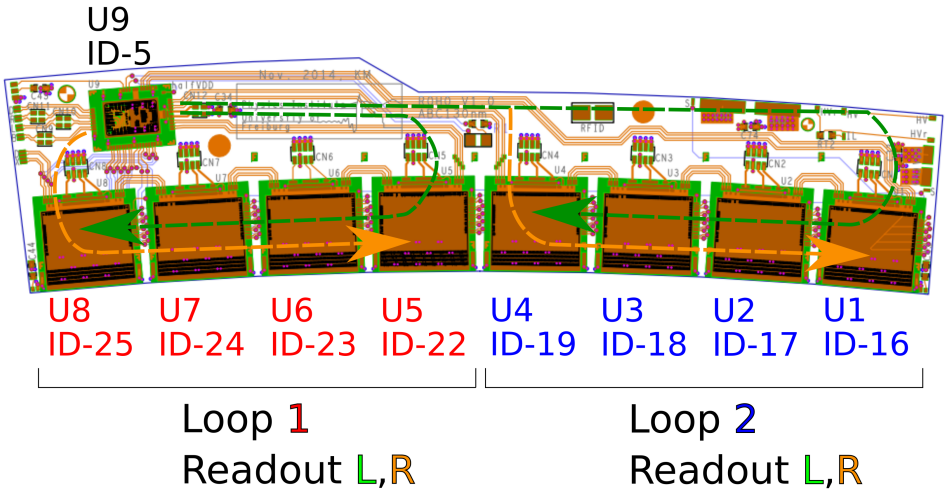


Figure 2.14: Schematic of how the HCC130 communicates with the ABC130 chips on the R0H0 - one of two hybrids used on the R0 module. The ABC130 chips are grouped into two bidirectional loops of four chips each.

ASIC communication - The first two tests are always:

- Request and obtain a successful ID return from the HCC('s) on the hybrid.
- Request and obtain a successful ID return from the ABC's on the hybrid.

This serves as a basic test to evaluate if the communication to and from the hybrid is working as intended, with the ID's determined by a pattern of wirebonds acting as a kind of physical bit register, creating a binary number based on which ID related bond-pads have and haven't been bonded for each ASIC.

A frequently encountered problem was successfully retrieving the HCC ID's, but only getting nonsensical signals from the ABC's - This could eg. be due to errors in the ABC ID wirebond configurations, or the read-out chain wirebonds - eg only allowing communication to flow in one direction, instead of the intended bidirectional flow as seen on Figure 2.14. On the software side of things, the 130 generation of hybrid designs combine ABC's together in loops of four chips, and in the configuration file for a hybrid, $3 + 2N$ settings (N being the number of ABC's) needed to be in exact agreement, both w.r.t. internal consistency and w.r.t. reality, before the ABC ID request would return meaningful and consistent information. This clumping together of ABC's increases the difficulty of troubleshooting, because a single error can mess up large parts of the hybrid, instead of just the single ABC where the error occurs. In the production chip set using the Star architecture, this read-out loop has been removed, with each ABC* directly connected to the HCC* instead - something that should simplify testing procedures like these significantly.

Typically, once the the HCC and ABC ID's are consequently returned correctly, all other subsequent tests will also execute properly - but whether the hybrid(module) fails or passes these tests is different story.

Threshhold scan - Due to the binary output of the ABC's, it somewhat tricky to evaluate eg. the gain of the chip, since we can't directly compare the input and output signal. The way to circumvent this problem is through threshold scans. This is a recurring part of the different tests performed, so to avoid repetition we'll define it here - before describing the actual tests.

A threshold scan characterises the response of a binary system, by scanning through a range of threshold values, and recording the binary response to an identical input. To cope with the noise of the system, many measurements, typically a few hundred, are performed at each threshold value. The V_{t50} is a common metric for a threshold scan, which gives the voltage of the input signal resulting in an acceptance rate - or efficiency of 50 % in the discriminator. Also, the plot of efficiency vs threshold value is typically called the response curve of the system.

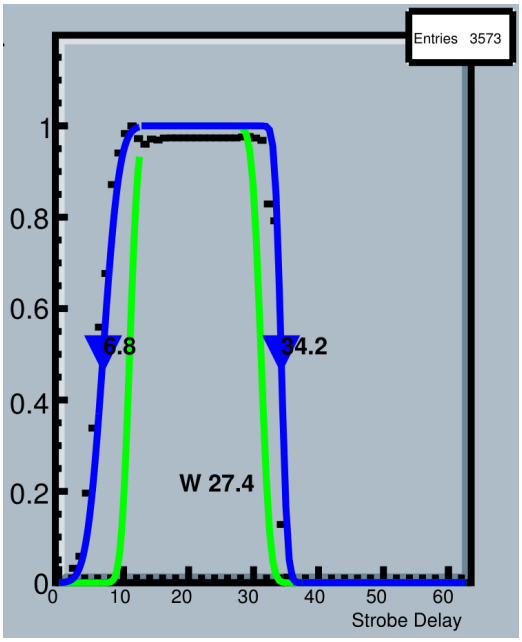


Figure 2.15: Strobe delay plot for one of the ABC's on the R0H0 hybrid used in the first electrical module built in Scandinavia. Efficiency on the Y-axis and strobe delay setting on the X-axis. The W value is the peak width and the other two numbers are the 50% efficiency values on each side of the peak. It is not clear from the internal documentation what the difference between the blue and green curve is.

Strobe delay - The discriminator activates based on the internal clock cycle of the ABC, meaning that, eg. every $50\,\mathrm{ns}$ the discriminator will compare, whatever integrated input is available, with the threshold, and generate a 1 or 0 output value - depending on the amplitude of the input. However if this clock is not properly synchronised with the incoming signal, it can lead to false negatives due to comparing the tails of the signal to the threshold rather than the peak value. To minimise this offset, a strobe delay is introduced, being the delay between sending the command to inject a calibration pulse, and the actual arrival time of the injected pulse. First a threshold scan is performed, so a threshold value can be correctly set to eg. $2\,\mathrm{fC}$, such that a calibration pulse of $4\,\mathrm{fC}$ always results in a binary output of 1 - if it's read correctly by the discriminator. Then we scan through the possible values of the strobe delay, inject the calibration pulse 200 times at each delay setting and record the efficiency of the discriminator at this delay setting. We except the efficiency to be 0 when the discriminator is asynchronous and 1 when it is in sync - with an example of the expected step function achieved by this shown on Figure 2.15. The strobe delay is then set to a value well inside the plateau peak of the step function, typically values being 25 % or 40 % inside the plateau.

Three point gain - The 3pt Gain test is performed by doing a threshold scan at three different values of injected charge, eg. 0.5, 1.0 and $1.5\,\mathrm{fC}$, and then plotting the threshold voltage vs injected charge, for some fixed value of the efficiency common across the different input charges - eg. the V_{t50} value. Using a linear approximation, the gain of the read-out chain, see Figure 2.11, can then be found as the slope of this linear fit. The output noise level can be estimated from the spread in the discriminator output, using the V_{t50} threshold but without injecting any signal charge - while the input noise level can be calculated as the output noise divided by the gain - with examples of this shown on Figure 2.16.

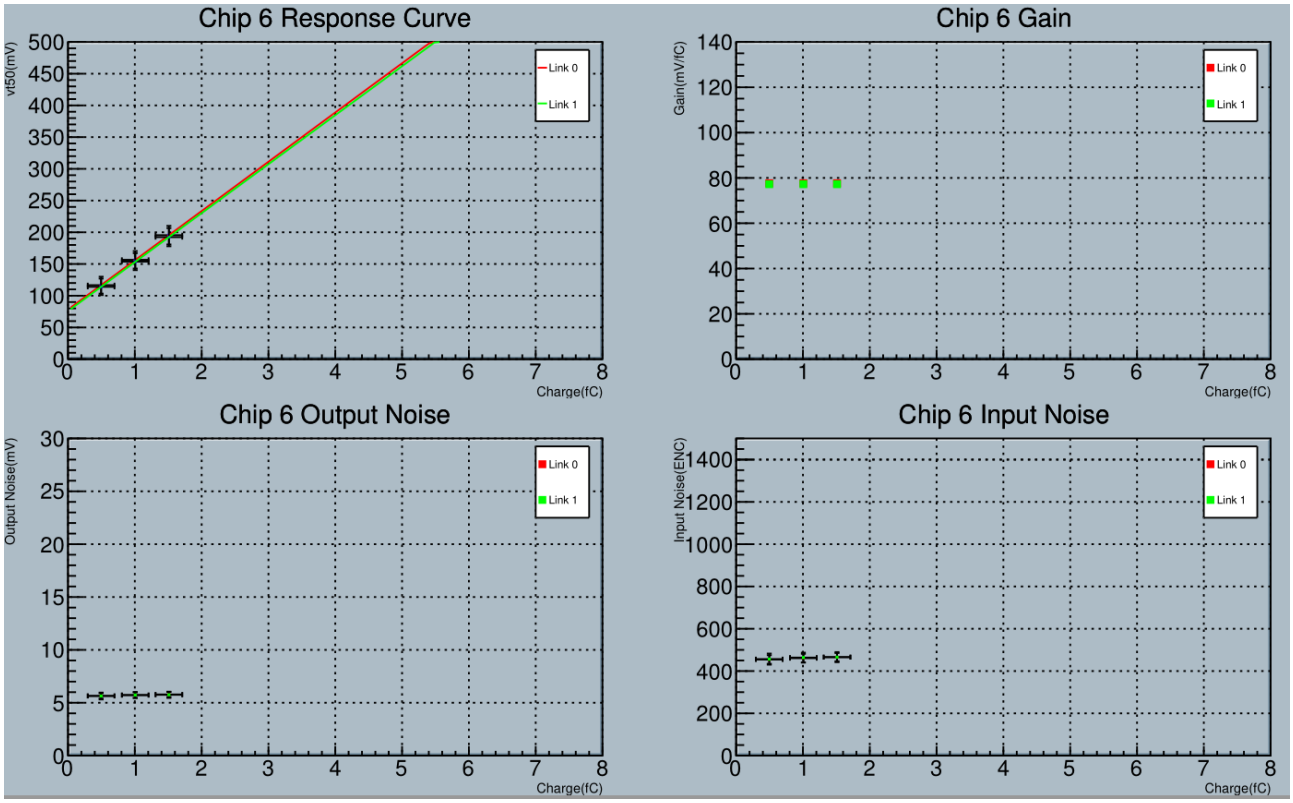


Figure 2.16: Test results from the 3pt Gain test, showing the gain of the ABC chip along with input and output noise. The X axis is always injected charge in fC. From the top left to bottom right the Y-axis is; V_{T50} in mV, Gain in $\mathrm{mV/fC}$, noise in mV and noise in ENC - Equivalent Noise Charge, meaning number of electrons needed to generate a voltage pulse of a given size.

Trim range - Each ABC has 256 individual channels, and as such there are bound to be some amount of variations in system response across different channels for identical inputs. The Trim Range test performs a series of threshold scans to evaluate which settings result in a minimal variance across the channels of each chip - and then apply these settings for future use. This test is only performed on full modules - it is pointless to optimise the individual channel response before the front-end wirebonds attach the ABC's to the sensor.

Response curve - The purpose of this test is identical to the 3pt Gain test, except that 10 different values of input charge are used, in order to perform a more accurate second degree polynomial fit. The 3pt gain test is much quicker than the Response Curve, and is used to test lone hybrids. The Response Curve is only used for full modules, because the higher fidelity characterisation is unnecessary for sole hybrids.

Noise occupancy - After the hybrid(module) has been fully calibrated by the previous tests - a final noise occupancy evaluation is performed. This is essentially a threshold scan without any input signal - to quantify what threshold value one should use to maintain a given Signal-to-Noise ratio.

2.4.2 Thermal Cycling

After a module has successfully passed the electrical tests, it undergoes thermo-mechanical stress testing, through repeated cycling of module temperature, to provoke thermal contraction and expansion. This is done to eg. identify poorly attached wirebonds, which then fully disengage during thermal cycling, such that the module can be send back for rework - instead of simply malfunctioning inside of ATLAS during operation. The thermal cycling procedure is visualised in Figure 2.17 and can be summarised as follows:

- At least 10 cycles over 12 hours, with the module being powered the entire time. A cycle is defined as going from $-35\,°\mathrm{C}$ to $40\,°\mathrm{C}$ and back again to $-35\,°\mathrm{C}$ - with the powerboard NTC defining the module temperature.
- After the initial cool down and the final cycle, a full module characterisation, as described in the previous section, should be performed.
- After each cycle, a simple set of confirmation tests are run to evaluate if any damage has occurred or if the module is still fully functional:
  - ID return for HCC's, ABC's and AMAC, turn Powerboard and sensor biasing voltage off/on.
  - Strobe delay - the strobe delay optimisation is very temperature sensitive, so it is recalibrated every time before the 3 point gain test.
  - 3 point gain - Checking if the noise and/or gain of the module has changed.
  - AMAC read-out of sensor temperature and leakage current.
- After the cycling has finished, a final two hour test at $20\,°\mathrm{C}$ is started - measuring the high voltage stability of the module.

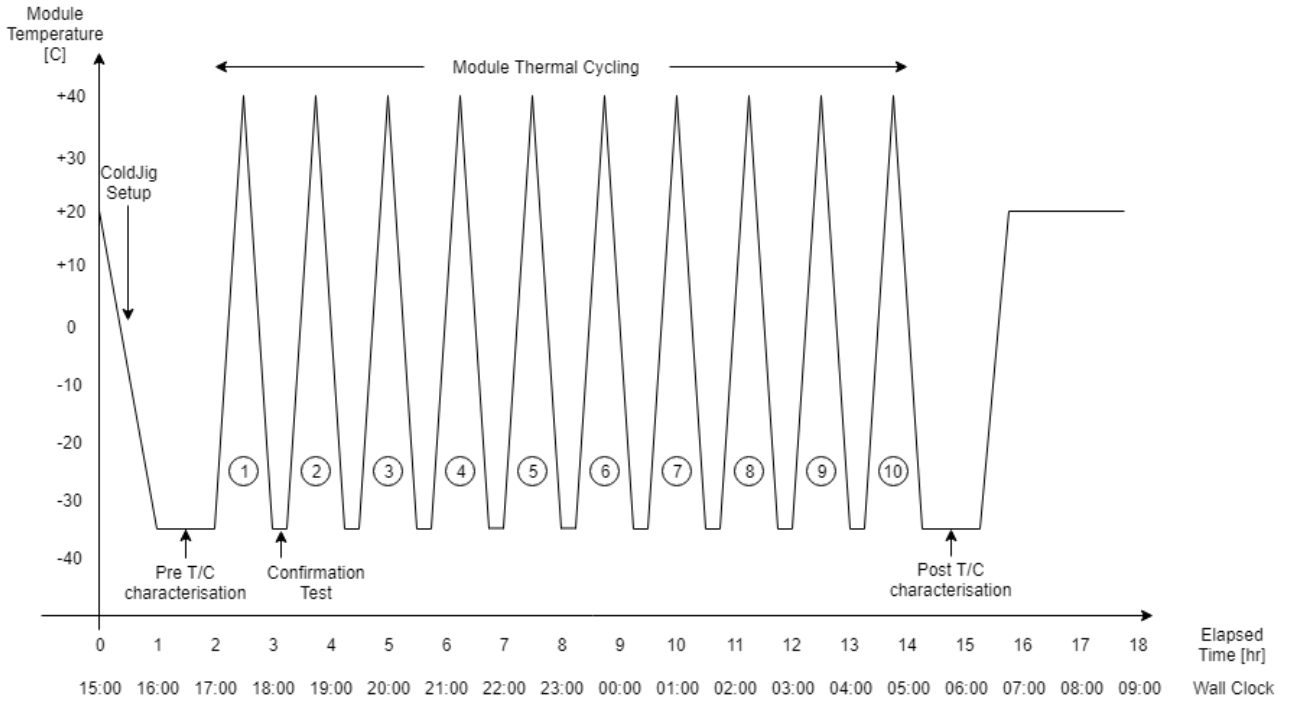
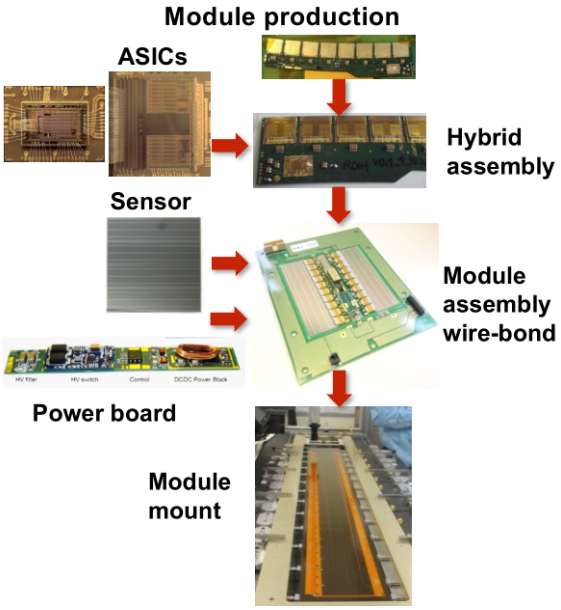


Figure 2.17: Showcasing the steps of the thermal cycling procedure. Module characterisation refers to an IV curve and a full suite of ITSDAQ tests. During the cycling, a less time consuming set of electrical tests are used.

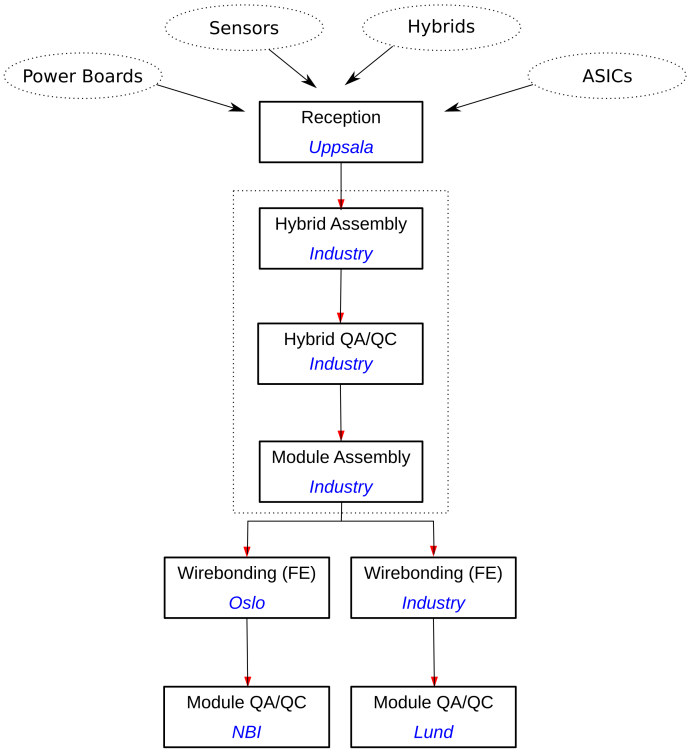
2.5 Module Assembly in the Scandinavian Cluster

The Scandinavian Clusters consists of scientist and engineers from four different universities, namely Copenhagen (NBI), Lund, Oslo and Uppsala university - which have banded together in order to contribute to the production effort of the ITk project. We are one of several end-cap module assembly sites, receiving batches of individual components, developed elsewhere, and combining them into fully functional detector modules - an overview of how and where this is done can be seen on Figure 2.18. The name ”Industry” seen on the figure refers to the fact that Uppsala intends to hire an electronics company, NOTE [23], to carry out the majority part of their assembly efforts. This does not mean that the Uppsala team, which I was part of for eight months, simply have handed the company a list of requirements and have had them figure out how to do things on their own. Instead we’ve spend our time and effort developing and testing the production procedures, both working in-house and at NOTE, combining our skills with the industry technicians to optimise the development cycle. This also means, that for the production, some of the needed machines, which NOTE does not posses, will be loaned to them by Uppsala University.

The Scandinavian Cluster are involved solely in the end-cap part of the ITk project, and will be producing roughly 10% of the total number of End-cap modules, $\sim$ 600, split across the two types R1 and R3[10, pg. 95]. However, up to now, tooling and components have, for the entire ITk collaboration, been limited to only the R0 architecture - meaning that prototyping efforts could only be done using R0 components. This explains why the thesis is focused on R0 development instead of R1 and R3. It also means, that considerations had to be made, to ensure that any procedure being developed, was sufficiently general to allow for a relatively seamless transition into the R1 and R3 scheme. Otherwise an unnecessary extra work burden would be created - limiting the quality of my contributions to the production efforts of the Scandinavian Cluster.



(a) Imaging the different steps of stave assembly



(b) Division of labour wihtin the cluster.

Figure 2.18: Showcasing the different steps of module assembly and where it is done within the Scandinavian Cluster. The left-side image compilation is for a barrel stave assembly, but the principals are the same for the end-cap petals.

2.5.1 Summary of Contributions

Now that scope of the ITk project and the production effort undertaken by the Scandinavian Cluster has been clearly laid out, it is time for brief overview of my contributions to the cluster, before going into details during the following chapters:

- The main contribution is the development of a glue robot and corresponding procedures, for the mounting of hybrids and power-boards onto the sensor. The robot can deliver a fixed amount of glue within a 50min time window after mixing the two-component epoxy, with sufficient precision to fulfil the technical requirements set forth by the ITk collaboration. This body of work is described in Chapters 3 and 4.
- silicon sensor studies, evaluating causes of abnormal IV curves and testing solutions to make sensors usable for production again. These investigations are detailed in Chapter 5.
- Performing electrical testing of hybrids, evaluating eg. gain, timing, noise and occupancy of readout electronics. This has already been described in the previous Section 2.4.
- Working at industry, to assist in optimising the ASIC-to-hybrid assembly, using their DATACON 2200 evo pick&place machine [24] - described in Chapter 4.

Development of a Glue Robot

3.1 Introduction

In order to have a functional detector, the silicon sensor needs to be joined to its readout electronics and powering distributor, the hybrid and the powerboard. Within the wider ITk collaboration, the baseline plan is, at the time of writing, to use a vacuum pick-up tool which places the hybrid into a holding jig with the hybrid bottom side facing up, then applying a two-component epoxy using a stencil. After this, the glued hybrid is transferred to the sensor mounting jig, where the sensor and hybrid are placed accurately w.r.t. each other through the usage of precision pins and slots in respectively the jig and the pick-up tools.

The mounting procedure can be seen on Figure 3.1a and 3.1b, while the stencil used for gluing a R0H0 hybrid is shown on Figure 3.3. The metal structure seen in the background of Figure 3.1 is the sensor mounting jig - an aluminium plate machined to a flatness of order $\sim 10\,\mu\mathrm{m}$, with in-build vacuum channels to keep the sensor in place during assembly - along with precision pins for alignment. The sensor is aligned by pushing it up against the three corner pins and then activating the vacuum stabilisation, while the pins along the sides of the sensor are used for alignment of the vacuum pick-up tools used to place hybrids onto the sensor.

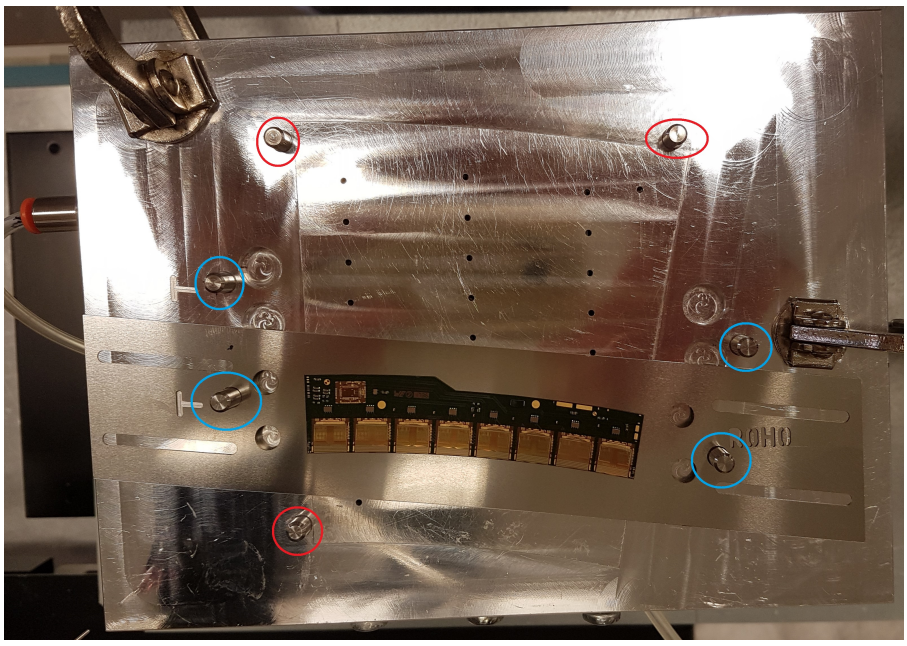
In the Scandinavian Cluster, the module assembly will take place at NOTE [23], our industry partner, and we foresee some unique challenges in this regard - motivating us to redesign the baseline assembly procedures to better suit our workflow. This development of a new assembly procedure, replacing the hybrid-to-sensor stencil approach with a high precision glue robot, has been the main project of this thesis. In the remainder of this section, we will go through the following items to properly set the scene for the development and calibration of our glue robot:

- Cover the technical requirements for the glue used in hybrid-to-sensor mounting.
- List the technical specifications defining a successful hybrid-to-sensor gluing operation.
- Describe the stencil based approach and its short comings w.r.t. our situation.
- Describe the problems to be solved, in order to develop a functional glue robot for our purposes.

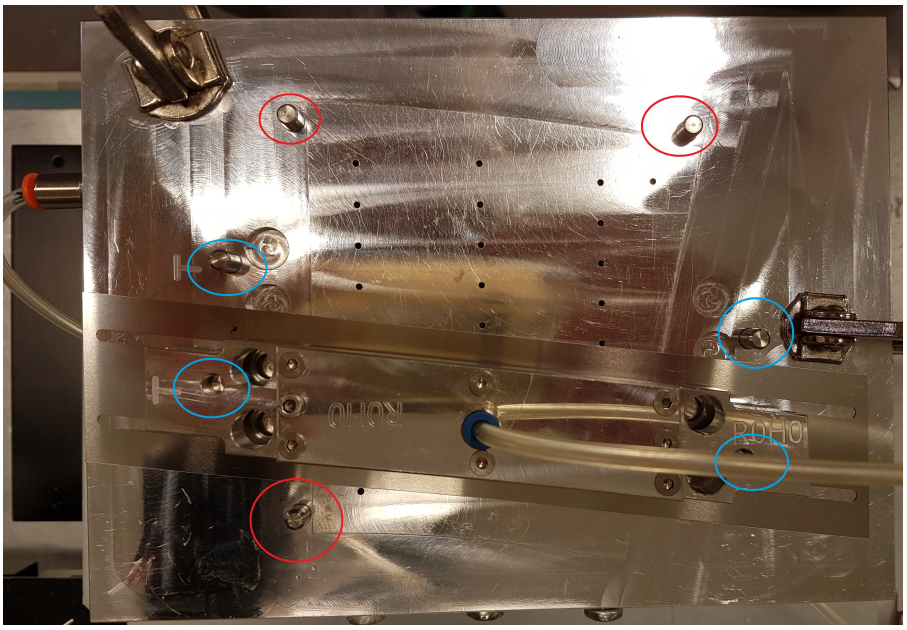
3.1.1 Technical Requirements of the Glue

There is a long list of technical requirements that the glue used in this project has to fulfil, the primary being:

- Low shrinkage during curing, this could ruin the very precise alignment of components and(or) induce mechanical stress on the sensor surface - ultimately degrading its electrical performance.
- High thermal conductivity, to prevent the build-up of hot spots leading to thermal damage of module parts.
- Sufficient bonding strength, both at low ($-35\,^{\circ}\mathrm{C}$) and high ($40\,^{\circ}\mathrm{C}$) temperatures.
- Sufficient radiation hardness, so the glue sticks properly throughout the planned 10 year lifetime of the detector - or longer.
- Sufficient Mechanical stability for wirebonding, one should be able to press down on a bonding surface, without it giving way.



(a) Stencil to correctly position the R0H0 hybrid for pick-up by the vacuum pick-up tool.



(b) The vacuum pick-up tool for the R0H0 hybrid - picking up the R0H0 hybrid placed in the corresponding alignment stencil as seen to the left.

Figure 3.1: Overview of sensor mounting jig and the hybrid vacuum pick-up tools. Below the hybrid and stencil, one can see the vacuum channels meant to keep the sensor in place during assembly. The orange outlet for the sensor vacuum can be seen on the upper left side of aluminium plate. Highlighted in red are the corner alignment pins for the sensor and in blue the pins which the vacuum pick-up tools for the R0H0 and the R0H1 slot into.

The first glue that was fully qualified for hybrid-to-sensor assembly was a two component epoxy, EPOLITE FH-5313 [19], and this was the only glue available throughout the bulk of the work done in this project. However, this glue has gone out of production, so Polaris PF 7006A will be used instead for the actual module production. This glue became available in the Scandinavian Cluster in August 2019, at which point all usage of Epolite was discontinued.

As a side-note, we cannot use the UV-curing glue utilised in ASIC-to-hybrid assembly. This is

because of an infeasibility in getting UV light to properly penetrate through the entire glue layer between the hybrid and sensor - leading to risks of the glue centre not being properly cured.

Towards the end of the project, the calibration procedure of the robot was redone for the Polaris glue. This turned out to be a golden opportunity to fix fundamental errors discovered during development of the gluing procedure.

3.1.2 Specifications for Successful Hybrid-Sensor Assembly

The specifications for a successful gluing operation of the hybrid to sensor attachment are:

- Height of the glue layer between hybrid and sensor $z = 120 \pm 40 \mu m$.
- No glue seepage onto the sensor front-end bond pads or the guard ring.
- Sufficient glue coverage below the areas were wirebonds are placed on the hybrid - to provide mechanical stability. This is quantified by requiring a filling factor of ~ 60 %. [1]

These requirements are motivated below, while a visual reference for the glue placement is provide in Figure 3.2:

The thickness of the glue layer is a balancing act between conflicting interest. If the layer is too thick, it would act as an unnecessary thermal barrier, between the heat generated in the ASIC's, and the cooling system situated in the mechanical support structure beneath each sensor. It would also complicate the front-end ASIC-to-sensor wirebonding. If the layer is too thin, the mechanical strength of the bond would be insufficient. Furthermore, additional noise would be generated in the read-out circuit - due to a capacitative coupling between the sensor strips and the grounded backplane of the hybrid.

The sensor edge is electrically connected to the high voltage bias applied at the backplane, meaning that any glue seeping out over the guard ring, see Section 5.2.2 for a definition of this, could possibly short circuit the entire system. Also, any glue coverage of the sensor bond pads would make them unbondable, resulting in a loss of resolution for the sensor.

Wirebonding is basically a form of microscopic welding, involving some amount of downward force being applied to the bonding surface. This means, that if the bonding surface is not sufficiently supported from below, it will flex downwards when force is applied - making wirebonding almost impossible to accomplish. This is why we need the glue to spread out sufficiently - to ensure no bondable parts of the hybrid are hanging free in mid-air [10].

All of the these requirements are accomplished for the R0H0 and R0H1, by using a total of $\sim 135(10)\,\mathrm{mg}$ of glue - spread out in the form of two four-line patterns. Any competitor to the baseline stencil approach needs to be capable of dispensing with the same or higher degree of precisions - both w.r.t. the total glue mass and the placement of it.

This is the challenge we face in developing a glue robot. It needs to have a precision in mass of $\sim 2.5\,\mathrm{mg}$, for each of the four lines in the R0 patterns, and a precision in glue placement $\lessapprox 1\,\mathrm{mm}$ - in units of absolute deviation.

[1] A filling factor is defined as the fraction of volume between the hybrid and sensor filled by glue.

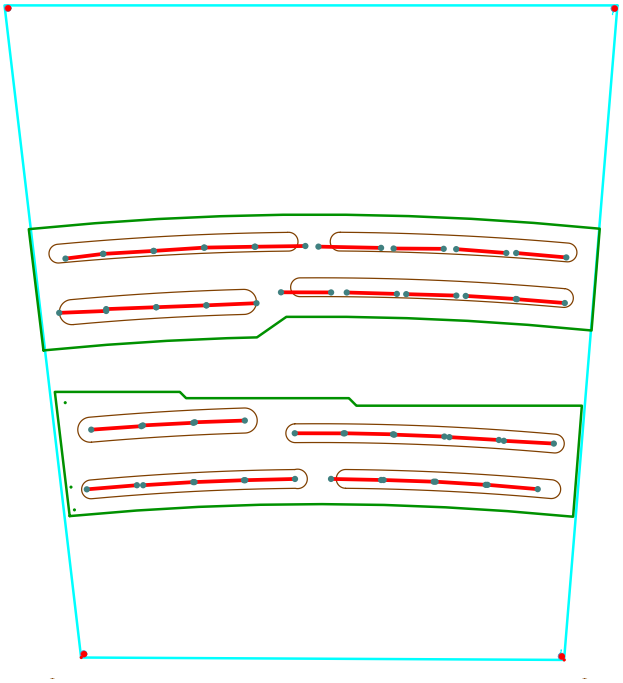


Figure 3.2: Drawing showing the outlines of the R0 sensor, the R0H0 (lower) and the R0H1(upper) hybrid along with the corresponding glue stencil outline. The dimensions of the sensor are $\sim 105 \times 86$ mm^2 - with $\sim 10\,\mathrm{mm}$ difference in width from top to bottom. The red lines make up the final pattern used by the glue robot during assembly of the first electrical R0 module in the cluster. The R0H1 lines were manually downshifted slightly, due to observations made during the prototyping studies. Given that the R0H1 lower edge was the only place we observed glue seeping out during assembly, this manual change should probably be rolled back in future iterations of the pattern.

3.1.3 Why Replace the Baseline Approach with a Glue Robot?

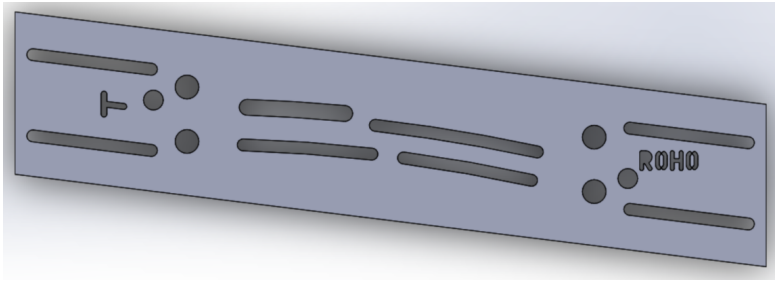


Figure 3.3: Drawing of the R0H0 hybrid-to-sensor gluing stencil. The stencil is placed on top of the hybrid backside, glue is spread out over the stencil and the excess is swiped away - leaving behind the pattern of glue defined by the cut-outs in the stencil. The two biggest circular holes, seen at each end, are used to slot the stencil into the precision pins of the hybrid holding jig, guaranteeing proper alignment of were the glue is applied.

Figure 3.3 shows the R0H0 glue stencil and outlines how the gluing procedures is carried out. There are many relevant factors determining the accuracy and precision with which one can operate the stencil, including but not limited to:

- The type of swiper used to wipe of excess glue - eg. stiff plastic or soft rubber.
- The angle one holds the the swiper at.

- How one holds and applies pressure on the swiper.
- How much, and how consistently one applies pressure throughout the swipe.
- How close the stencil sits to the hybrid surface - microscopic gaps, eg. due to cured glue residues from previous usage, will allow for glue seepage.

All of the above issues can be solved, to some extent, through sheer repetitive training. Private correspondence with other module assembly sites, using the baseline stencil approach, report an achievable precision in the hybrid-to-sensor glue amount of ~ 10 %. However, in our cluster, module assembly will take place in industry, and as such, we would have to pay for the many hours industry technicians would have to spend simply training to build up their precision. This would also cause a severe dependence on the industry technician(s) trained to do this, risking high delays in production if that key employee becomes unavailable. By developing a robot, we hope to achieve similar or better precision than a well trained stencil-user, but in a manner that is as operator independent as possible.

Another aspect to consider, is the amount of hybrid handling required in the baseline procedure. We are especially concerned for the steps of flipping the hybrid upside-down and vice versa. We believe that this creates unnecessary risks of damaging functional wirebonds, due to accidental mishandling. As such, we would prefer a method were glue is deposited directly onto the sensor surface, thereby minimising the overall number of handling steps for the sensitive components.

So - by spending our time and effort on developing a flexible semi-automated procedure, we're attempting to floor the learning curve - such that any operator can successfully assemble different types of modules (R1 and R3) - with minimal time spend training.

3.1.4 Feature Requirements for the Glue Robot

The initial idea for the glue robot was to write a Python script utilising the RS-232 protocol for serial communication between a pc and the two actuators of the setup, the XY-table and the high pressure air valve attached to the glue syringe, by giving the program sets of (x,y) coordinates to follow, while dispensing the glue along these predetermined paths. The motivation for using Python was simply that the language had all the needed functionality and that the writer was already familiar with working in the Python environment.

This home-made setup was originally intended to serve as an initial training platform - while the purchase and delivery of a commercial glue robot were being carried out. However, after receiving our commercial machine, we realised that the functionality of the provided control software was severely lacking - the machine was simply not usable for our purposes. As such, the home-made setup that I had developed, was upgraded from training platform to intended production platform.

Throughout the development process more features were incrementally added as the need for them became apparent during testing and troubleshooting of the setup. Listed below is a final list of the features implemented, to ensure a successful mounting operation using the glue robot - an image of which can be seen on Figure 3.4.

- A terminal-operated python program sufficiently user friendly for out-of-the-box operation by technicians in industry.
- RS-232 serial communication used to interact with the actuators of the setup.

- Input is given through .txt files with sets of (x,y) coordinates, creating a glue pattern to be layed down, and a desired amount of mass for each glue line. Corrections of translation and rotation have to be performed, to properly overlay the sensor coordinate system with the XY-table coordinate system.
- Dynamic calibration correcting the dispensing parameters w.r.t. the change in glue viscosity over time, due to curing, - to allow for flexibility w.r.t. delays during operation and the ability to use the same glue pack for several mountings.
- Overall calibration, such that the robot can deliver x mg of glue over a distance of xx mm within the required precision - and with the speed of the XY-table being the only feasible way to directly manipulate the amount of glue being dispensed.
- Automated logfile generation of the dispensings, with overwrite protection implemented so previous data isn't lost - used for calibration and quality control of the robot performance.
- Safety features such as emergency termination of glue flow, and ensuring the program doesn't crash in cases of trivial errors eg. due to bad user input.
- Calibration procedure developed in a sufficiently general way, to allow for a, hopefully, low effort transition between the R0 pattern using during development and the R1 and R3 patterns to be used in production.

The final setup for our semi-automatic glue robot consists of a Märzhäuser-Wetzlar XY-table [20], along with high precision single use syringes, connected to an electrically controlled pneumatic valve and mounted in a manual Z-adjustable vice, controlled by the mentioned custom made python framework - with the robot speed being the controller of how much glue is being dispensed.

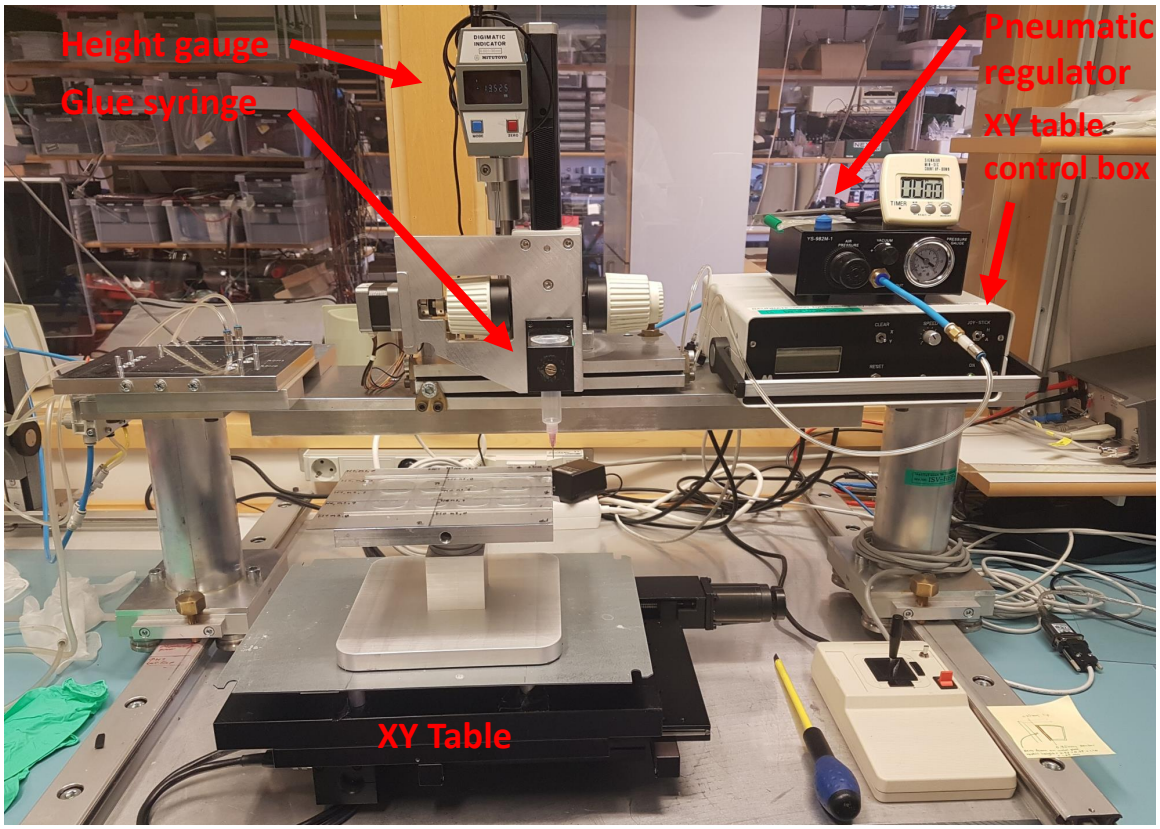


Figure 3.4: Image of the glue robot prior to a test run - explaining why the blue/white high pressure air hose hasn't been connected to the glue syringe yet.

3.2 Complications of the Robot Gluing Procedure

In this section we'll present a brief summary of the complications encountered in calibrating the glue robot, to provide the reader with a minimal frame-of-reference while progressing through the technical details of the robot development in the remainder of the chapter.

We are dispensing glue by using pressurised air to squeeze it out of a stationary syringe - while the dispensing surface is being moved around by the XY-table. So, for a fixed dispensing pressure and dispensing length, there will be an inverse relationship between the robot speed and the amount of glue being dispensed. This is because, as the velocity grows, the total dispensing time decreases - resulting in less glue. However, there are many different factors in play, determining how the dispensed mass scales with the robot speed.

Time Calibration:
The glue, both the Epolite FH-5313A-A-PAK and the Polaris PF 7006A, are two component epoxies delivered to production sites in pre-proportioned bags - see Figure 4.3a. They both have a very skewed mixing ratio, eg. Epolite having roughly 13% hardener and 87% resin. This makes it very difficult to do a live mixing of the glue during dispensing, the small amount of hardener would very easily get lost in the tubing of the dispenser. Taking a large quantity of these bags and pouring the components into two different large containers also seemed problematic, it would inevitable lead to some spillage of resin or hardener, which would then throw off the precisely tuned mixing ratio. This will never occur with the A-PAK's because they're intended to be mixed in the the bag, so any eventual spillage will be of the mixed epoxy [19].

The solution for the glue robot was to mix the glue in the bag, as intended, and then pouring it into a syringe for automated dispensing. However, this creates a different problem in that, the glue starts curing as soon as mixing commences, meaning that our setup, using pressurised air to squeeze glue out of the syringe, will deliver less glue over time, when using identical dispensing settings, due to an increase in the viscosity of the glue. This is obviously not acceptable, delays due to sporadic machine or operator failure can and will occur, so our setup needs to be capable of delivering a fixed amount of glue, independent of the time of dispensing after mixing - to some reasonable extent of course.

Target Mass Calibration:
We need to figure out how to translate between a desired target mass, in mg, and the robot speed setting which will deliver this amount of glue. This mapping between target mass and robot speed is correlated with the time dependency of the glue describe above - eg. speed setting 50 will give significantly different amounts of glue if used $30\,\mathrm{min}$ apart in time.

Length Calibration:
There's a significant difference in the intended pattern shape, when asking for eg. $20\,\mathrm{mg}$ of glue, over respectively $15\,\mathrm{mm}$ and $30\,\mathrm{mm}$ of length. This creates a correlation between the length setting and the target mass calibration - because the speed setting correctly giving $20\,\mathrm{mg}$ over $15\,\mathrm{mm}$ will give $>20\,\mathrm{mg}$ over $30\,\mathrm{mm}$ due to the longer dispensing time. The glue per length ratio differs, and this needs to be considered in order to chose the proper speed value - allowing the robot to complete both tasks successfully.

We're intending to follow the official glue patterns of the ITk collaboration, which, for the R0 pattern, contain ~ 6 different line lengths - and we expect the R1 and R3 patterns to have a similar structure of multiple line lengths.

As such, what we end up with is a four dimensional calibration problem for our setup. The delivered glue mass depends on: the time since mixing, the length of the glue line, the table speed and

the pressure of the air flow. The variables which we can control in a satisfactory manner are the length of glue lines and the table speed - the air pressure is kept constant because it is manually regulated.

That there are so many correlated variables to consider, should motivate to the reader why this is a rather non-trivial problem to solve. However, we believe that, the ease with which any operator can use a, properly calibrated, glue robot to successfully assemble modules, is worth the time and effort spend solving these problems.

3.3 Time Dependency of Glue Application

The first aspect of the robot calibration we'll investigate, is how we can time-stabilise the dispensing - by compensating for the change in glue viscosity over time, due to its curing. To do this, we measured $m(t)$, the change in mass over time, using identical settings for the air pressure, length of the glue line, speed of the xy table, needle diameter and height over the dispensing surface.

It turned out to be quite a challenge, producing multiple data series of mass vs time, that could reasonably be argued to all come from the same underlying distribution. The causes of inconsistent behaviour in the measurements have not been precisely identified, but educated guesses, and discussing with collaborators from other institutes, tells us that variations in glue temperature, quality and time of mixing eg. through varying concentrations of air bubbles in the glue, and, perhaps most importantly, variations in procedure due to human error being responsible. To solve these initial troubles, we invented a proper check list, to be followed when preparing for a gluing run, which helped the operator minimise procedural variations between measurement runs. The list varied somewhat depending on the circumstances of the specific run, but would, in general, contain the following items:

- Remove the glue pack from cold storage ($-9\,^{\circ}\mathrm{C}$), warm it by hand massage and let it sit at room temperature for at least 1 hour before gluing, to ensure thermal equilibrium of the glue with the very stable room temperature of the laboratory in which the dispensing is carried out.
- Go through a (sub)check list to ensure all machinery and components are ready, functional, and initialised with the relevant parameters of operation, before starting the glue mixing.
- Do a dry run of the desired operation. (This might be skipped later on in production, once we've gained sufficient confidence with the setup, but should always be done during development.)
- Start the robot program and an external stop watch, when the green plastic spacer is removed and mixing is commenced - to ensure accurate timing data.
- Use a fixed mixing time.

Evaluating the Robot's Performance

Perhaps the most important improvement after the initial troubles, comes not from something directly stated in the list above - but rather the implementation of a semi-automated log-filing system into the robot framework. During the execution of a gluing operation, the robot will record relevant informations for every line in the pattern and save it as a txt file. The filename is auto-generated using a tag related to the type of gluing operation and a timestamp, with a overwrite protection feature checking if the filename is already taken within the working directory, and if so, generates an available filename.

Before dispensing, the robot prompts the user for input on the current dispensing pressure and needle height over dispensing surface. This is done manually to force the user to verify these settings - thereby reducing the chance of using wrong and potentially dangerous settings. Initially, the needle gauge was also prompted for, but this is now hardcoded into the log-file, given that we stuck to using

a gauge 20 pink needle tip since the early prototyping days of the project. This information is written in the first line of the log-file, along with the name and units of the data.

The following data is recorded:

- $t_{initial}$ and t_{final} in $\mathrm{minutes}$, for the dispensing.
- length of the line in mm.
- target mass in mg.
- robot speed setting.

After dispensing had finished, the array of glue lines, see Figure 3.5, is cured, typically in an accelerated manner by baking it at $55\,^{\circ}\mathrm{C}$ for $\sim 45\ \mathrm{min}$. Then each glue line is weighed independently, using a $0.1\ \mathrm{mg}$ precision scale - in chronological order to ensure that the n'th entry in the logfile corresponds to the n'th entry in the record of masses. To ensure that the glue didn't stick too well to the plastic dispensing surface, we typically coated the dispensing surface in a thin layer of dishwasher soap - making the glue bond primarily to the thin film of soap. However, it still happened occasionally, that excessive force was used when prying a glue line from the dispensing surface, making it jump up vigorously and escape into the abyss of the lab floor. In these cases we simply noted down a "0" in the mass record, such that the corresponding entry in the logfile could be correctly disregarded.

This semi-automated log-filing and weighing procedure forms the basis of the datasets used for all the analyses carried out in this chapter.

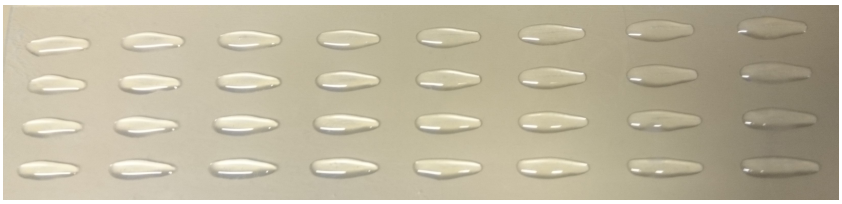


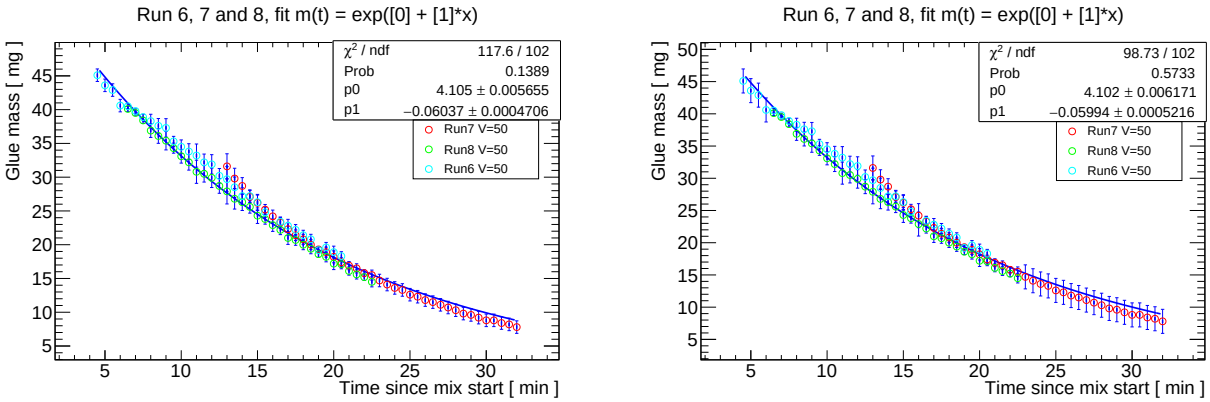
Figure 3.5: Typical picture of an array of glue lines, from a measurement run investigating the dependency between the dispensed mass and some relevant variable.

Using the check list and proper logfiling procedure, we managed to produce three measurement series which could reasonably form the basis of a fit to parametrise the mass vs time dependency - see Figure 3.6a and 3.6b.

The displayed fit is a simple exponential function of the type

$$m(t) = e^{p_0 + p_1 \cdot t} \tag{3.1}$$

Going forward, this p_0 and p_1 will be referred to as respectively the constant $A = e^{p_0}$ in units of mg and the slope b in units of min^{-1}. The minor difference in the slope of Run 7 compared to Run 6 and 8 might be explained by a slight difference in the pressure. The regulator is a manual dial, and these can deviate from a set value over time, due to leakage in the piping. Otherwise it might be due to a work hardening of the Run 6 and 8 glue compared to the Run 7. Other ITk module assembly groups have done similar measurements where they saw a difference in glue behaviour at the same points in time, depending on if you had been dispensing before this point or if it was the first dispensing done with that batch of glue. Their interpretation was that applying pressure to the glue would compress and, for a lack of a better term, work harden it, leading to a slower flow compared to fresh glue at a given point in time. However while this explanation also fits the measurements presented here, the effect hasn't been thoroughly investigated, and as such is closer to speculation than fact.



(a) Average standard deviation is used here, see text below for explanation.

(b) Maximum standard deviation is used here, see text below for explanation.

Figure 3.6: The variation in amount of glue being dispensed over time, due to increase in glue viscosity from curing. The robot settings were, speed $V = 50$, pressure $P = 3\,\mathrm{bar}$, path length $L = 20\,\mathrm{mm}$, needle gauge 20 and time between dispensings $t = 30\,\mathrm{s}$.

The only difference between the two plots are the assigned error bars, which comes from a best effort approach w.r.t. to the limited dataset we're working with. For any given point in time on the graph(s), there are between 1 and 3 data points. When $N_{points} > 1$ we calculate the standard deviation of this set, and combine it in quadrature with the precision of the scale used to weigh the glue lines as shown in Equations 3.2 and 3.3 [25].

$$\sigma_{robot} = \sqrt{\frac{\sum_{i=0}^{N} (x_i - \mu)^2}{N - 1}} \tag{3.2}$$

$$\sigma_{mass} = \sqrt{\sigma_{robot}^2 + \sigma_{scale}^2} \tag{3.3}$$

Where $\sigma_{scale} = 0.2\,\mathrm{mg}$, estimated by weighing several individual glue lines, in the mass range $[4; 60]\mathrm{mg}$, ten times in a row and then taking the mean value of the different standard deviations found. As a side note, one can consider σ_{robot} to be the systematic error of the robot dispensing.

Now, in the cases where $N = 1$, we can't calculate a standard deviation and we also can't naively set $\sigma_{mass} = \sigma_{scale}$. This would imply that we have better information (reflected in a smaller error) about the underlying mass distribution at a given time, in the cases were we have the least amount of data available - which is false. The solution we chose for this conundrum was to take the average (maximal) σ_{mass} calculated among the $N > 1$ cases and using this as the error for the $N = 1$ cases. This might not be the optimal solution, but it is simple and works reasonably. Due to limitations on our stock of glue, and very limited possibilities for obtaining more, we had to be conservative with our glue usage and not repeat the same measurements more than strictly needed. We ended up using the average error, $\sigma_{robot}^{avg} = 0.9\,\mathrm{mg}$, instead of the maximal, $\sigma_{robot}^{max} = 1.8\,\mathrm{mg}$, it seemed like the more reasonable choice, though in this specific case the choice has rather low impact - given that the values of the fitted parameters are the same within 1σ.

The error bar chosen for the singular points on Figure 3.6a, as explained above, will be used as an estimate of the error due intrinsic variations, on all measurements going forward in the Epolite studies. The total error will in these cases be

$$\sigma_{intrin} = \sqrt{{\sigma_{robot}^{avg}}^2 + \sigma_{scale}^2} = 0.92\,\mathrm{mg} \tag{3.4}$$

We also tried including an additional constant to the fitting function

$$m(t) = Ae^{b \cdot t} + C \tag{3.5}$$

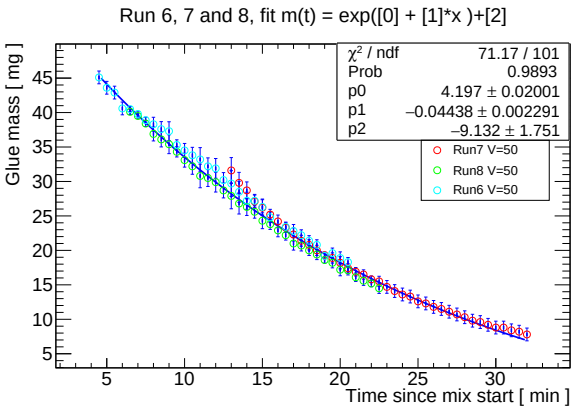


Figure 3.7: Glue mass vs time, due to curing effects - with additional fitting parameter.

which can be seen on Figure 3.7. While this leads to better GoF values, the fact that $p2 < 0$ means that the fit predicts a negative mass of glue being dispensed for large times t, when $e^{p_0+p_1*t} < p_2$. This is of course non-physical, and in practice it severely limited the window of opportunity for dispensing the glue. The fit could only be used for calibration in the time period where it predicted $m(t) > 0$. This is very unpractical, and led to discarding the fit in favour of the one presented in Figure 3.6a.

3.3.1 Developing a Viscosity Correction

In the previous section, we established how the mass changes with time, now we wish to develop a feature for the robot to cancel out this behaviour. As a reminder, the only way our setup can automatically change the amount of glue being dispensed, is through changing the speed of the XY-table.

For a fixed path length and dispensing pressure, a lower velocity will result in longer dispensing time, thereby more dispensed mass, a type of inverse scaling. We see on Figure 3.6a that, As time progresses, the dispensed mass per time decreases, meaning that the robot speed should decrease to counteract the effect and achieve a constant amount of glue over time. To begin with, we can match the decay rate of the table speed with the fitted decay rate from Figure 3.6a - this must be the rate at which the speed needs to change.

$$V(t, M_{tar}) = \frac{V_0}{M_{tar}} \cdot Ae^{bt}. \tag{3.6}$$

However, since Equation 3.1 has units of mg, a scaling factor of V_0/M_{tar} is also added. V_0 is the robot speed setting used during the $m(t)$ measurement runs - the robot speed register accepts integer values in the range $[1; 100]$, with the number being proportional to the number of revolutions per second of the two spindles moving the XY-table. The intent was for M_{tar} to have units of mass and be our input handle to the robot - meaning that instructions for a gluing operation should include a (x, y) coordinate set to determine the pattern, and values of M_{tar} to give the amount of mass desired in each line of the pattern. This fraction is chosen simply because it's the simplest mathematical construct that matches units, a first order approximation of sorts, which will evolve throughout the thesis.

As a quick side note; while implementing this speed calculation in the robot software, two additional considerations had to be made:

The Y-axis motor of the XY-table had a tendency to get stuck at high speeds, so a speed limit $V < 70$ was implemented. This restrains the time period of operation, since the minimum achievable, correctly dispensed, mass will come from dispensing at $V = V_{max}$. So, depending on the desired amount of glue, one might have to wait until the glue is sufficiently thick for the maximal speed to result in a dispensing of this targeted amount.

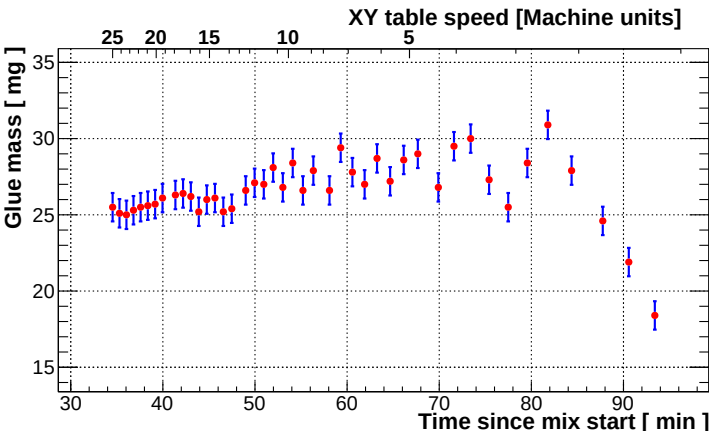
Also, because the XY-table only accepts integer values of speed setting, this sets an inherent limit on the achievable accuracy on the dispensed mass.

3.3.2 Evaluating Viscosity Correction

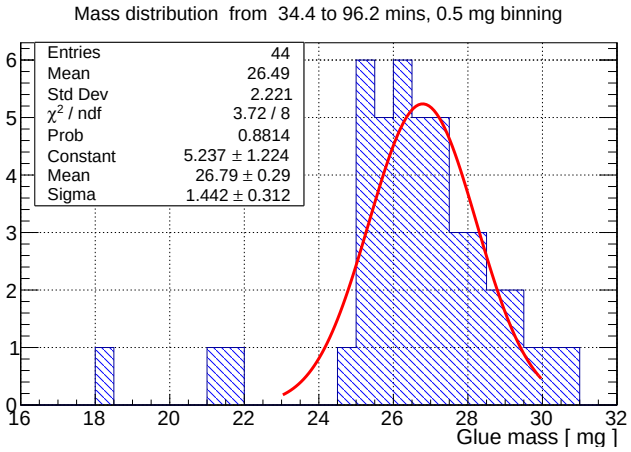
To test the quality of the $V(t, M_{tar})$ viscosity correction given by Equation 3.6, we chose some value of M_{tar} and measured how well the robot could deliver this mass as time progressed. From Figures 3.8a and 3.8b, we see acceptable control of the dispensed mass until 50-60 min after mixing, depending how low a tolerance for deviations we desire. If we limit the time to $t \leq 50\,\mathrm{min}$, see Figures 3.8c and 3.8d, and do a linear fit to the mass vs time data, we find the slope to be within $1\ \sigma$ of zero, and a standard deviation on the mass distribution of $0.57\,\mathrm{mg}$.

This shows That we can achieve a very nice control of the delivered mass up to $50\,\mathrm{min}$ after mixing. Also, accumulated experience tells us, that a time window of $50\,\mathrm{min}$ would allow us to glue $3-4$ modules using the same bag of glue - once the production flow is properly streamlined.

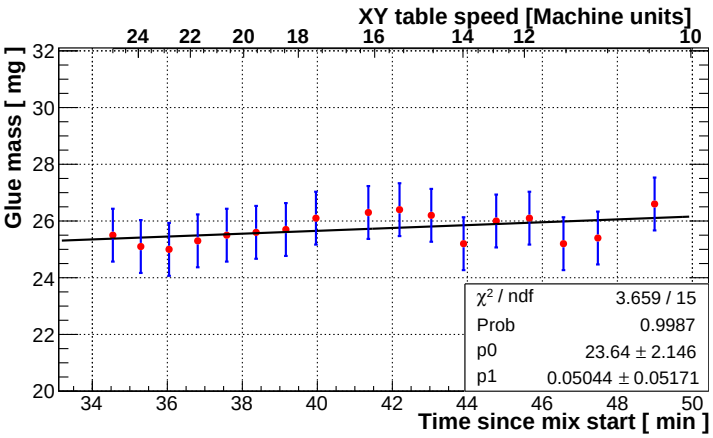
Unless explicitly stated otherwise, the viscosity correction provided by Equation 3.6 has been utilised in all data taking going forward.



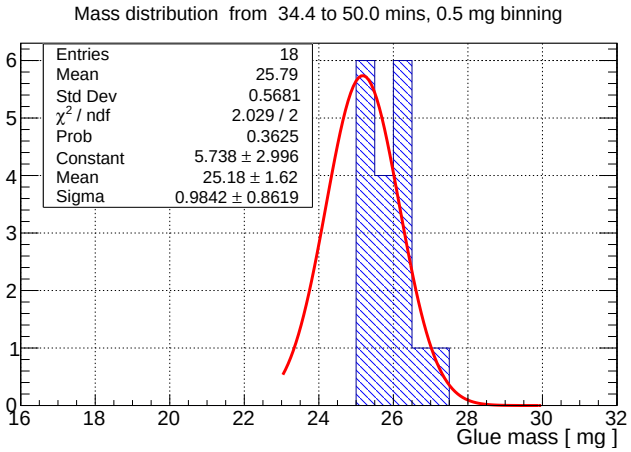
(a) Stabilised time evolution of dispensed mass.



(b) Histogram showing the spread of the dispensed mass, $0.5\,\mathrm{mg}$ binning, fitted with a Gaussian distribution.



(c) Zoomed version of the above, $t \leq 50\ \mathrm{min}$, fitted with a straight line.



(d) Distribution of dispensed mass, $0.5\,\mathrm{mg}$ binning, using the cut in time seen on to the left. Fitted with a Gaussian.

Figure 3.8: Testing the time stability of the dispensing with $M_{tar} = 15$. The time axis starts around 33 mins because other test were done beforehand, using the same batch of glue. The upper X-axis is a logarithmic scale, displaying the robot speed settings utilised, nicely showcasing how implementing Equation 3.6 enforces an exponential decay of the speed as time progresses. The error bars are $0.92\,\mathrm{mg}$ and is the one assigned to singular points in Figures 3.6a and 3.7.

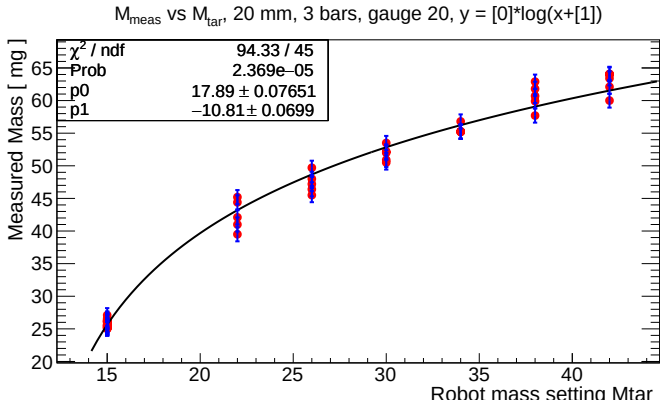
3.4 Robot Mass and Length Calibration

In the previous section, we demonstrated sufficient time control over the glue amount being dispensed, but poor control of the actual amount, given that $M_{tar} = 15$ resulted in roughly $26\,\mathrm{mg}$ being dispensed. Furthermore, all the tests shown so far were done using a path length of $20\,\mathrm{mm}$, whereas the path lengths for the R0 glue pattern are, respectively, 33, 41.4 and $24.6\,\mathrm{mm}$ long. The reason for working with $20\,\mathrm{mm}$ lines is that, before the R0 campaign, the Scandinavian Cluster were involved in producing two R2 semi-eletrical modules, where we created a glue pattern consisting solely of $20\,\mathrm{mm}$ lines.

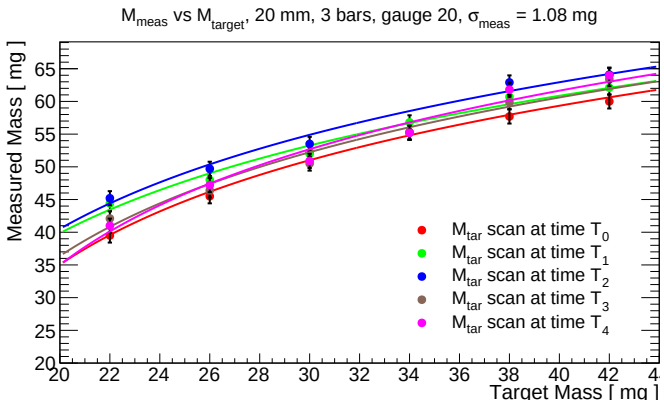
This section is therefore spend on investigating how to calibrate the mass and length scaling of the robot, such that we can ask for X mg of glue over xx mm - and actually dispense this with reasonable precision.

To measure how the dispensed mass scaled with M_{tar} and path length we devised the following tests:

We scanned through a range of M_{tar} values $[42, 38, 34, 30, 26, 22]$, while keeping $L = 20$ mm constant, and a range of lengths, close to the actual R0 pattern, $[22, 28, 34, 40, 46]$ mm, while keeping $M_{tar} = 31$ constant. Each scan was repeated five times, alternating between M_{tar} scan and L_{scan} - to also use this as an opportunity for evaluating the quality of our time correction. The $M_{tar} = 31$ was chosen for the length scan because it was the lowest value of M_{tar} that didn't require long waiting times due to the table speed limit of $V \leq 70$. In hindsight we should have instead chosen a lower value, giving a dispensed mass closer to the $20 - 30\,\mathrm{mg}$ we need for the R0 pattern.



(a) Relationship between the measured mass of glue and the M_{tar} parameter



(b) Left hand plot divided into five different fits, base on the five periods in time in which they were dispensed.

Figure 3.9: To properly use the $V(t, M_{tar})$ equation, we need to determine how the mass input parameter M_{tar} scales with the dispensed mass. The data was generated in the time range of $7 - 29$ minutes after mixing.

Fit #	p0	p1	χ^2/ndf	P value
T_0	$18.00 + -0.3$	$-12.99 + -0.6$	$3.04/4$	0.55
T_1	$18.06 + -0.3$	$-10.91 + -0.8$	$5.65/4$	0.23
T_2	$18.75 + -0.3$	$-11.32 + -0.7$	$5.65/3$	0.13
T_3	$18.33 + -0.3$	$-12.73 + -0.6$	$10.71/4$	0.03
T_4	$18.83 + -0.2$	$-13.57 + -0.5$	$12.08/4$	0.02

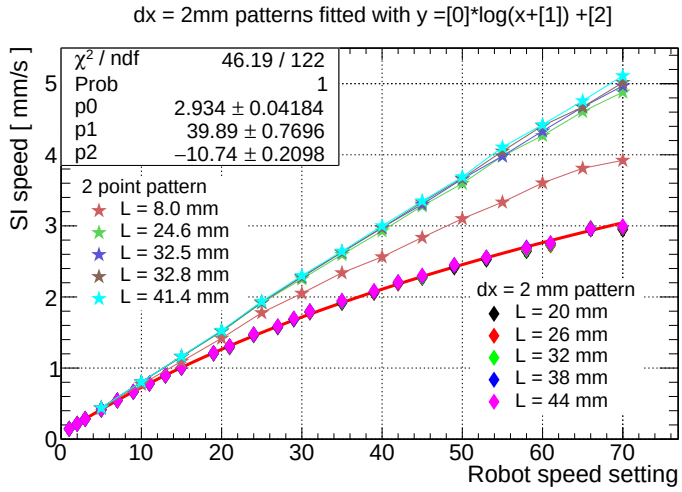
Table 3.1: Fit results when dividing the M_{tar} data series by the 5 different sections in time where they were taken. The data was generated in the time range of $7 - 29$ minutes after mixing.

Figure 3.9a contains data both from this M_{tar} scan, and from time stability test seen on Figure 3.8c, dispensed at $M_{tar} = 15$, and we see a clear logarithmic dependency. On Figure 3.9b, the M_{tar} scan data has been split up into the five individual measurements series, done at different points in time, but with the viscosity correction implemented. Table 3.1 shows the results of the five fits, and we see that they all agree within $1 - 2\,\sigma$ - another proof that our viscosity correction is working quite nicely. The error bars in the Y axis are calculated as

$$\sigma = \sqrt{\sigma_{intrin}^2 + \sigma_{t50}^2}. \tag{3.7}$$

Where σ_{intrin} comes from Equation 3.4, and σ_{t50} refers to the error related to the time-stabilisation - given by the standard deviation seen on Figure 3.8d.

It was quite unexpected that the scaling between dispensed mass and M_{tar} was logarithmic in nature, and a lot of effort was spend trying to understand the physical mechanism behind this. The solution came, when we saw, on Figure 3.10a, that there's also a logarithmic scaling between the robot speed setting and the actual speed in SI units. This scaling can be explained as an artefact of how the glue pattern is created. Each line is created as a set of dots with a spacing of $2\,\mathrm{mm}$, meaning that, for every $2\,\mathrm{mm}$ the robot travels, it accelerates to the given speed setting, stays at that speed for a short while, then decelerates to a full stop before proceeding to the next point. This means that the total dispensing time is not simply set by the linear fraction of $length/speed$, because some amount of time is spend on the (de)acceleration cycles, with this effect increasing in significance as the speed setting, or the peak velocity, increases - because then less time is spend moving at the peak velocity compared to the time spend on (de)accelerating.



(a) robot speed in SI as a function of speed setting. Timing information was used to match a given speed setting to its corresponding value in $\mathrm{mm/s}$. The red line is a fit to all of the $dx = 2\,\mathrm{mm}$ data - demonstrating the logarithmicity of the speed scaling for this type of line construction. For the 2-point patterns of corresponding length, we can see, by eye, the expected linear scaling. Linear fits are not shown for the sake of brevity.

(b) Illustration of line constructions using differing amount of points. The robot moves from point to point, with a full stop after reaching any one point.

Figure 3.10: Showcasing how the average robot speed scales with the robot speed setting, for different line lengths and types of lines.

Figure 3.10a came from investigating the relationship between the actual robot speed, in SI units, and the robot speed setting - for different line lengths and constructions. We investigated the relevant length scales for the R0 pattern $24 - 44\,\mathrm{mm}$, along with $L = 8\,\mathrm{mm}$, something that will be expanded upon in Section 3.5. W.r.t. line constructions, we used two different architectures, The standard for the Epolite studies, being the set of dots spaced $2\,\mathrm{mm}$ apart, and the other being simply two dots spanning

the entire length of the given line. The SI speed was calculated by using the timing and length data recorded in the logfile for a given line - as such, it is the average speed of the robot from the start to finish of a given line. The lengths were double checked by also having the robot draw the lines, with a pen attached instead of a syringe, and measuring the drawn length with a $0.5\,\mathrm{mm}$ precision ruler - giving us a standard deviation of $\sigma_L = 0.5\,\mathrm{mm}$, an error which is most likely dominated by precision of the measurement equipment.

The fit displayed on Figure 3.10a comes from combining all the $dx = 2\,\mathrm{mm}$ data and confirms the highly logarithmic nature of the speed scaling. Standing opposite to this is the clear linearity of the data when only using a single set of coordinates, defining the start and ending point of the line. Also, the $8\,\mathrm{mm}$ line, constructed using only two points, sits nicely in between the two extremes, consistent with our understanding of shorter path-length leading to a more non-linear dispensing time.

These observations all seem to confirm, that the total dispensing time, parametrised in terms of average SI speed, scales logarithmically with the robot speed setting. This behaviour then gives us a physical mechanism that can, at least partially, explain why we see a logarithmic scaling between dispensed mass and M_{tar}.

As a side-note, the R0 glue pattern was initially created by tracing it off from official technical drawings of the stencil, without much thought given to the process other than a consistent $2\,\mathrm{mm}$ spacing. We simply didn't realise the importance of how the pattern is constructed, before being directly confronted by the corresponding consequences. This logarithmic scaling is an additional complication of the robot calibration. So when this work is redone for the Polaris glue, it would be beneficial to reconstruct the glue patterns using only 2 or 3 points per line in the pattern - so as to minimise this undesirable effect.

W.r.t. the mass-length scaling, we expect a linear scaling based on $m(x+dx) = z \cdot y \cdot (x+dx) \cdot \rho$, with z, y, x, describing the dimensions of the glue line and ρ the density of the glue. On Figure 3.11 a decent linear fit to the data is performed, though the p-value is on the low side.

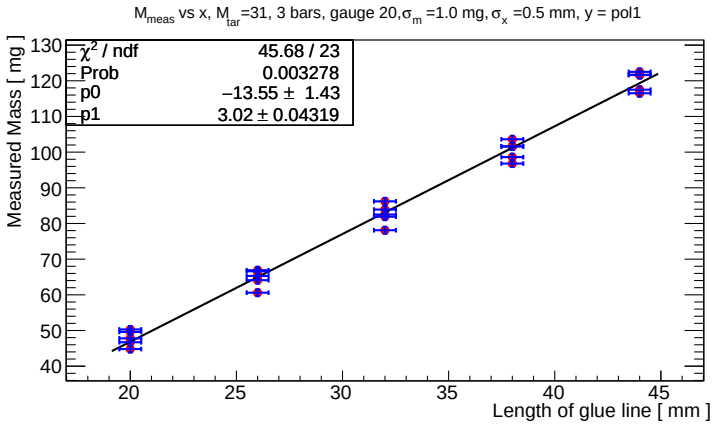


Figure 3.11: Dispensed mass scaling with length, for $M_{tar} = 31$, performed in the time range $5 - 26$ min after mixing the glue.

Initially the idea was, base on Figures 3.9a and 3.11, to create a new speed calculation of the type

$$V(t, M_{tar}(m), L) = V1(t, M_{tar}(m)) \cdot V2(L) \tag{3.8}$$

However, this would require the M_{tar} and length calibration to be uncorrelated, which they are not - a change in M_{tar} alters the glue per length ratio thereby creating a new length scaling between mass and length. So, ignoring the length correction for now, we only focus on the M_{tar} calibration. From

Figure 3.9a we can derive the following expression for $M_{tar}(m)$ with m being mass in mg.

$$m_{meas} = k_0 ln\left(M_{tar} + k_1\right) \tag{3.9}$$

$$\updownarrow$$

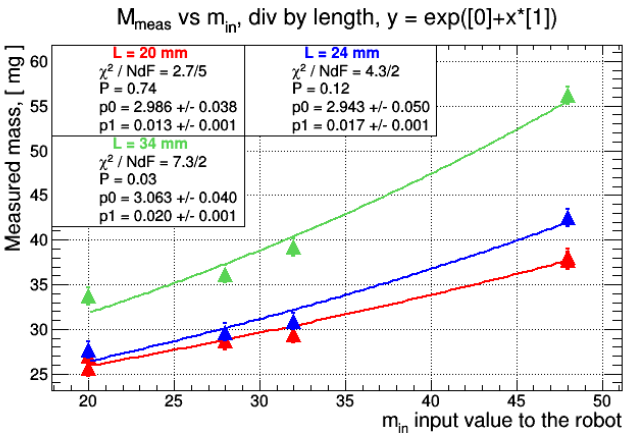
$$M_{tar}(m_{meas}) = e^{m_{meas}/k_0} - k_1 \tag{3.10}$$

$k_0 = 17.89$ and $k_1 = -10.81$ being the fit parameters from Figure 3.9a. Implementing this in the robot speed calculation we get

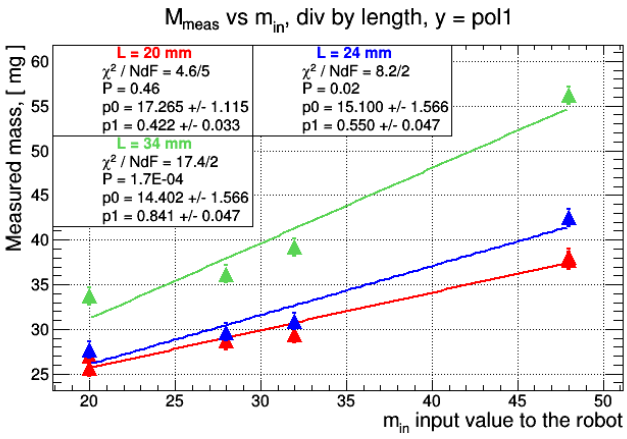
$$V(t, m_{in}) = \frac{V_0}{M_{tar}(m_{in})} e^{A+bt} = \frac{V_0}{e^{m_{in}/p_0} - p_1} e^{A+bt}. \tag{3.11}$$

where m_{in} is introduced to represent the the mass input command in units of milligrams.

On Figure 3.12, we test the performance of Equation 3.11. The first thing to be noted is, that Equation 3.10 was derived solely from data at $l = 20\,\mathrm{mm}$, which means it is quite unreasonable to expect Equation 3.11 to work as intended at other length scales, due to the correlation of the variables M_{tar} and $length$. While we don't see the desired $y = x$ line, it is possible to do a nice linear fit on the $20\,\mathrm{mm}$ data, but less so for the other length scales - as seen on Figure 3.12b. This is a problem since we need to develop an approach that flexible both in length and mass range. Since the viscosity correction is also based on $L = 20\,\mathrm{mm}$, We decided to backtrack our analyses and reinvestigate the mass vs time scaling for different sets of initial conditions - to get a better understanding of the systems underlying behaviour.



(a) Fitting and exponential to the data.



(b) Linear fit performed.

Figure 3.12: Testing the mass calibration of the glue robot based on Equation 3.11. A successful test would have been a $y = x$ diagonal line, such that the input mass command matches the produced output. The data on the X-axis is the input values to Equation 3.11, m_{in}, rather than values of M_{tar} as was used in Figure 3.9a.

3.4.1 Revisiting the Viscosity Correction

To get a better idea of how one might go about implementing a velocity calibration that properly takes time, target mass and target length into account - we decided to go back to basics and perform more raw measurements of mass vs time. This time however, it was done at three different fixed speeds, $[10, 50, 70]$ and at four different path lengths, $[20, 24, 32, 42]\ \mathrm{mm}$, the R0 pattern and $20\,\mathrm{mm}$ for comparability with previous data. This gave 12 different datasets, all of which were fitted with an exponential decay, see Figure 3.13.

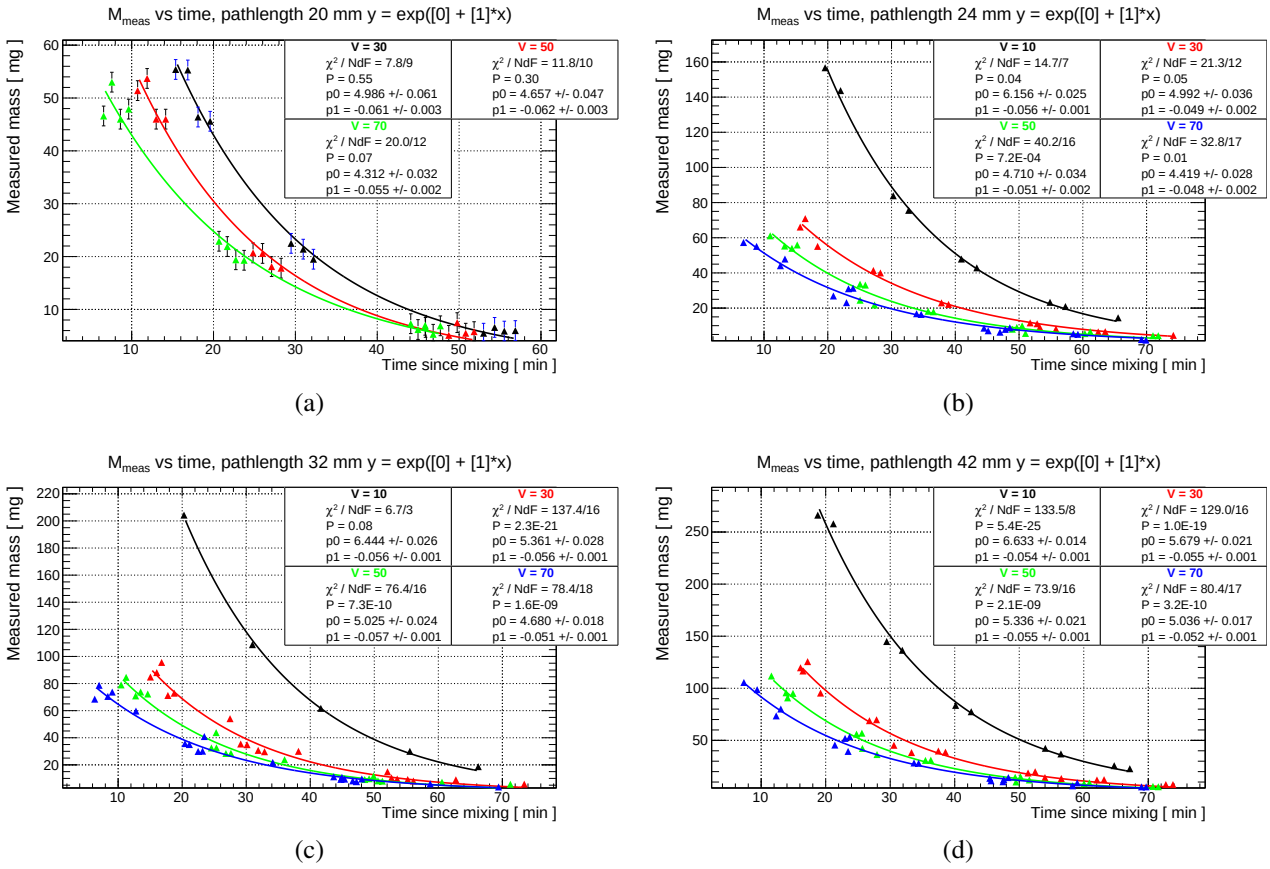


Figure 3.13: Glue mass vs time, divided by speed setting within the plot and by path length amongst the four plots. There are errors on all data points.

The purpose of these tests were to evaluate how correlated the parameters of the fit were on the speed and length settings used to generate the data. A-priori we expect the following connections:

- A positive linear dependence between the constant scaling factor, p_0 and path length.
- A type of inverse scaling between p_0 and the speed setting.
- No scaling between the slope, p_1 neither with path length or speed setting.

It is especially important that the slope is independent from the speed settings used, since we want to alter the speed, based on the slope, to correct for the curing effect. W.r.t. the path length, the no-correlation assumption is simply based on a lack of physical motivation as to why there should be any connection between path length and slope.

To evaluate how the dependencies of the decay parameters are, we create a new set of four plots, showing the values of the fitted parameters as a function of path length and speed setting, this can be seen on Figure 3.14. For the constant, p_0, the results are more or less as expected, but for the slope, p_1, the plots are quite hard to interpret, but it does seem that slope is not nearly as independent of length and speed setting as we would like, and this is probably part of the explanation of the limited success achieved on Figure 3.12.

In an attempt to quantify whether or not the slope is independent of length and speed setting, We took the full dataset of slopes and performed a Kolmogorov Smirnoff (KS) test on it. Assuming no correlation, we expect the fluctuations of slope values to be of Gaussian nature - therefore a Gaussian distribution is used as the reference sample for the KS test. To improve the accuracy of this reference sample, we use a weighted mean μ_w, and the error on this weighted mean σ_w, as the input to the reference Gaussian. The weights w_i used are the inverse variances obtained from the error on fitted

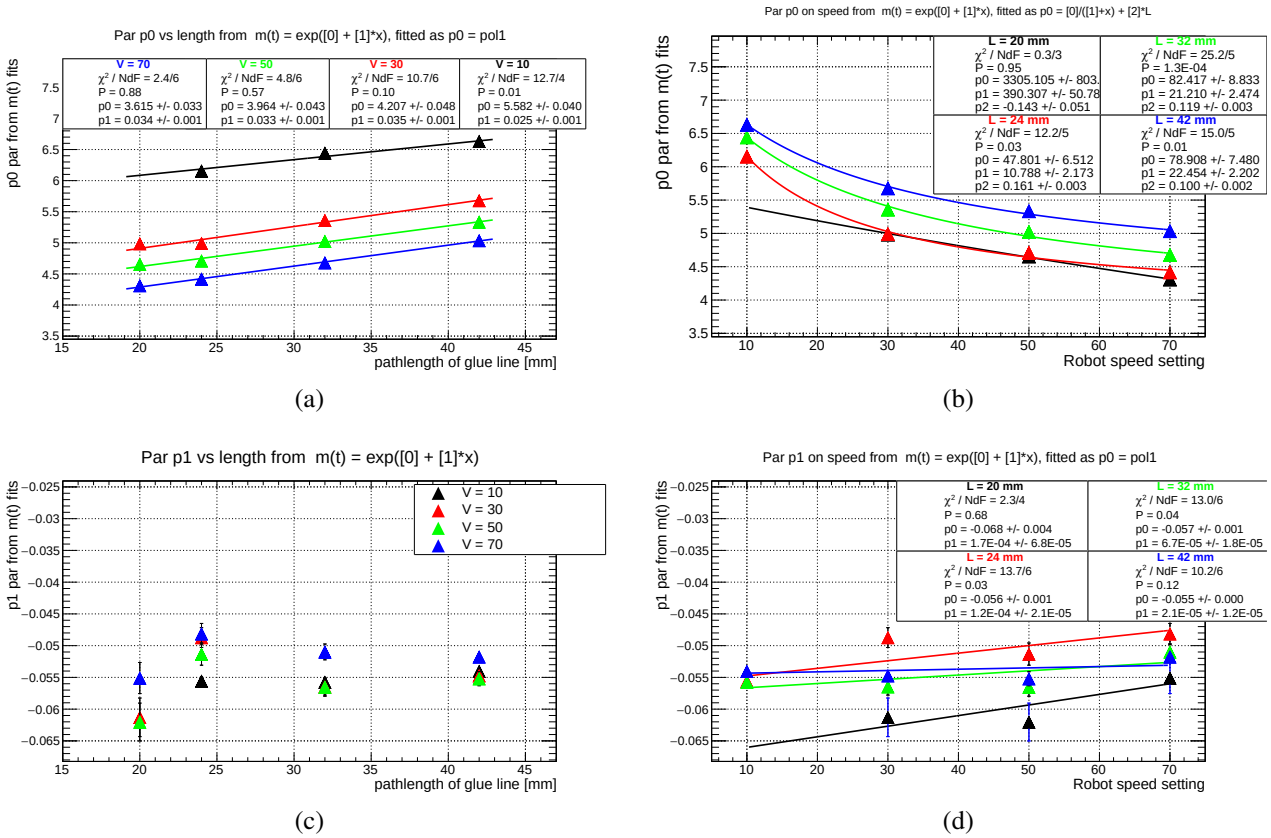


Figure 3.14: Fitted parameters of glue mass vs time, $m(t) = e^{p_0+p_1 t}$. The top row is values of p_0 and the bottom row values of p_1. The reason for no fit on Figure (c) is that I don't know what to fit it with.

values of the slopes, this is an elegant and simple way of boosting the importance of data points with the lowest associated errors [26].

$$w_i = \frac{1}{\sigma^2_{p_i^1}} \qquad \mu_w = \frac{\sum_{i=0}^{n}(p_1^i w_i)}{\sum_{i=0}^{n} w_i} \qquad \sigma_w = \sqrt{\frac{1}{\sum_{i=0}^{n} w_i}} \tag{3.12}$$

Doing this we found a p value of $p = 1.67E-09$ and KS test value of $D = 0.5996$, showing that the fluctuations in p_1 cannot be attributed to Normal statistical variance.

Due to several reasons, both logistical and scientific, it was decided to not pursue the calibration of the Epolite dispensing any further. Instead starting over with glue Polaris, implementing a few crucial differences from the beginning - hoping to solve the problems encountered in the Epolite studies.

3.5 Polaris Studies

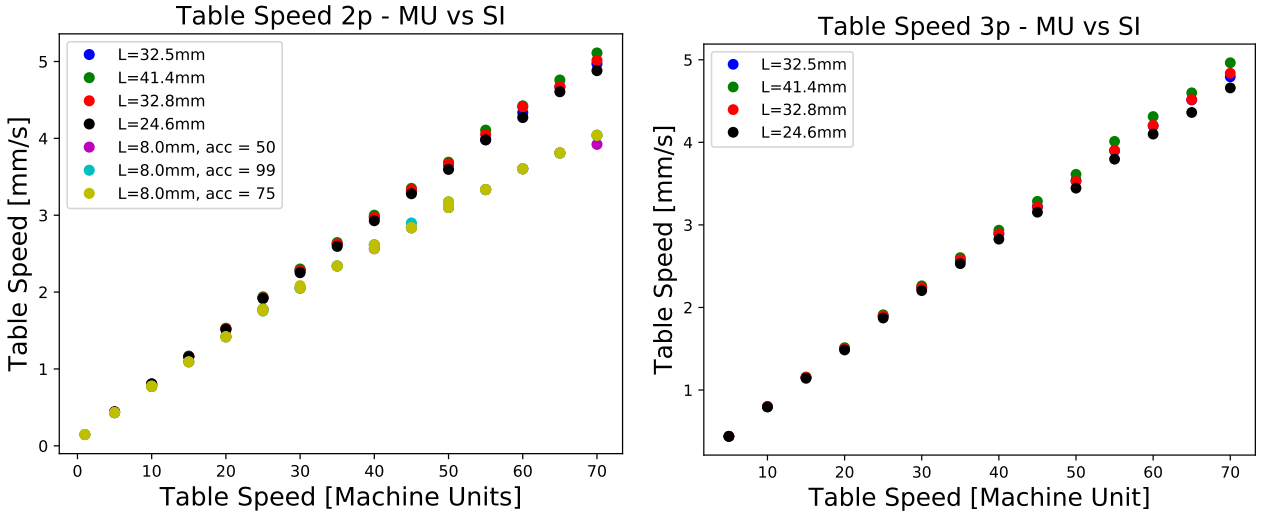
As mentioned in the beginning of this chapter, Epolite was the first glue qualified for module production, but it has gone out of production, and as such, a replacement glue was needed. This replacement is Polaris PF 7006A - it fulfils all the same technical requirements as the Epolite, but none of the data gathered in the Epolite studies can be reused since the two glues are fundamentally different substances.

This section will cover the chronological development of the glue robot into the final version used to successfully produce the first fully functional electrical module in the Scandinavian Cluster.

3.5.1 Investigation of Speed Scaling

During the final phases of the Epolite studies, we suspected the logarithmic scaling between robot speed setting and actual speed, Figure 3.15, of contributing to the troubles faced in calibrating the mass scaling of the robot. Therefore, the first thing to investigate this time around was how this speed scaling looked when the glue pattern was changed to only use two or three dots to make a line - compared to the prior method of having a dot per two mm of line length. This was done in Figure 3.20 for the four lengths in the R0H0 pattern, where we see that the degree of linearity in the scaling is coupled to the length of the line. This is expected, the robot speed setting is proportional to the number of revolutions per second performed by the two spindles which the table moves along. So, only when the time spent travelling at constant speed dwarfs the time spend (de)accelerating, will the speed scaling be properly linear. We also see a slightly more curved scaling for the three point pattern compared to the two point pattern. The official R0 patterns has some small amount of curvature in the lines, and this would require a minimum of three points to implement, since you can't define a curved line using only two points. However, given that the glue spreads out to $\sim 3\,\mathrm{mm}$ in width directly after deposition, we decided it was impotent to implement a curvature of $\sim 1\mathrm{mm}$. Therefore the two point pattern was selected - to maximise linearity in the speed scaling moving forward.

We also investigated changing the acceleration setting of the XY-table, as seen on Figure 3.15a, however changing this value to its maximum seemed to have negligible effect on the speed scaling - so for the sake of simplicity we decide to leave the acceleration setting at its default value of 50.



(a) speed scaling for two point pattern.

(b) speed scaling for three point pattern.

Figure 3.15: Testing the scaling between Robot speed setting and measured speed in SI units. "acc" refers to the acceleration setting of the XY-table, $acc = [1; 99]$, with 50 being the default value. "MU" stands for Machine Units, commonly referred to as the robot speed setting in this text.

The $8\,\mathrm{mm}$ lines also present in Figure 3.15, were added after a discussion on the transferability of the prototyping efforts, using the R0 architecture, to the production of R1 and R3 modules. We had already decided to not attempt a dynamic length calibration for the glue robot, in order to reduce the workload and optimise our chances of success. This means, that the targeted result for this calibration effort is the ability, for several different fixed lengths, to dispense any given mass, within a relevant range, at single milligram precision, for a time range allowing us to produce more than one module per bag of glue or to allow for flexibility in case of practical problems during mounting. If we did this simply using the lengths of the R0 pattern, the work done here would have to be repeated for each different pattern to be dispensed in Uppsala. This would be a very unpractical, and frankly quite unintelligent, approach.

Instead, we opted to use a single smaller path length, $8\,\mathrm{mm}$, a sub-line of sorts, which could then be used to construct the different lines of the real pattern, by adding $8\,\mathrm{mm}$ lines together, with customised small spacings between each $8\,\mathrm{mm}$ line to reach the targeted length of eg. the R0H0 and R0H1 pattern. The idea is, that in using this approach we have a much more flexible solution, which can rather easily be adapted to different glue patterns. Furthermore, it becomes much more efficient to gather data for the mass calibration when there is only one length scale to deal with - both w.r.t. time spend and amount of glue needed to gain sufficient statistics. However, one concern with this approach is, the non linear speed scaling at the $8\,\mathrm{mm}$ length scale. On the other hand, the length of the sub-line needs to be sufficiently small to allow for flexibility in building up the larger glue patterns - so settling on a length for the sub-line is a balancing act between two important opposing features of the system. In the end, the $8\,\mathrm{mm}$ length was chosen primarily for its near integer division with the path lengths of the R0H0 pattern. The final pattern lengths used in the first actual mounting procedure can be seen in Table 4.1.

3.5.2 Estimating Glue Amount for R0 Assembly

Besides the actual calibration of the glue robot, enabling it to deliver a given mass at a given time, we also need to figure out how much glue we actually need, for a successful hybrid-to-sensor assembly. An estimate of this was made by calculating the volume of the different stencil slots, and it could alternatively have been made from the requirements on glue height and filling factor. However this can only be treated as a rough estimate, so while doing the mass calibration studies, we also made trial runs of the mounting procedure, gluing transparent plastic hybrids and dummy hybrids to dummy sensors, to evaluate the performance of a given amount of glue. The amount of glue under the hybrids were estimated by dispensing the same pattern just before and after the actual mounting, but keeping the glue lines to weigh after curing, instead of using it to assembly dummies. For these trial runs, the $120\,\mu\mathrm{m}$ height requirement was enforced by placing two strands of fishing line, with exactly this thickness, in between the ”hybrid” and ”sensor”. This fishing thread method is not perfect, as it somewhat hinders the spread of the glue, but it is a decent approximation of how the glue flows out when the gap is at the correct height - which is very valuable information for the calibration effort. This approach can be seen on Figure 3.16.

During curing, a big flat metal plate was placed on top of the modules - to ensure the hybrids were pressed flat against the fishing thread, to get the height correct. This type of study were performed several times, making adjustments of length, position and mass in-between - until we believed we had a set of settings with a reasonable change of success. These settings can be seen in Table 4.1 and the pattern itself in Figure 3.2.

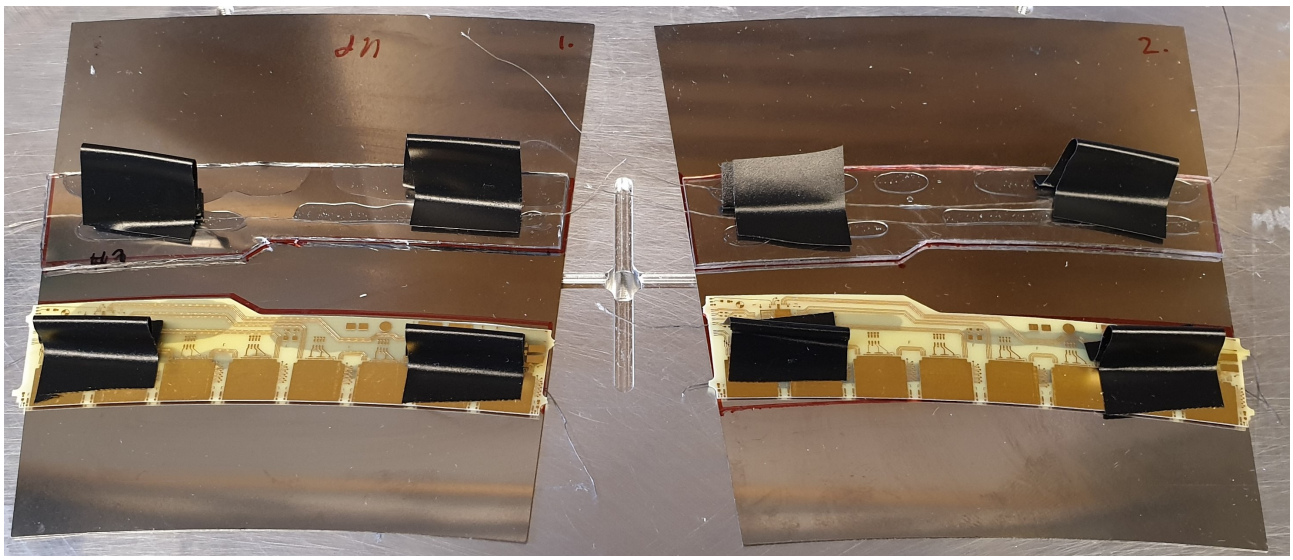


Figure 3.16: Investigating how a given amount of glue spread out between the hybrid and sensor dummies. The black tape was used as handles to place down the hybrids instead of the vacuum pick-up tools used for the real hybrids. The hybrids were placed by freehand using the red marker outline as a placement guide.

3.5.3 Mass vs Time and Speed

After settling on an approach for handling the length calibration, we started doing mass vs time measurement series again, a selection of five different speed settings were chosen to get a representative set of measurements over the span of speed settings available to us. Throughout three weeks a total of six measurement runs were done, to gather statistics for the calibration effort, though they were scattered across the time axis due to the need for building dummy modules, in order to establish proper target values of the glue amount. Meaning that in between gathering $8\,\mathrm{mm}$ mass vs time data, we also had to estimate how much glue, in each line of the pattern, was actually needed to fulfil the technical specifications set forth by the collaboration. The final result of this can be seen on Figure 3.17. Slight

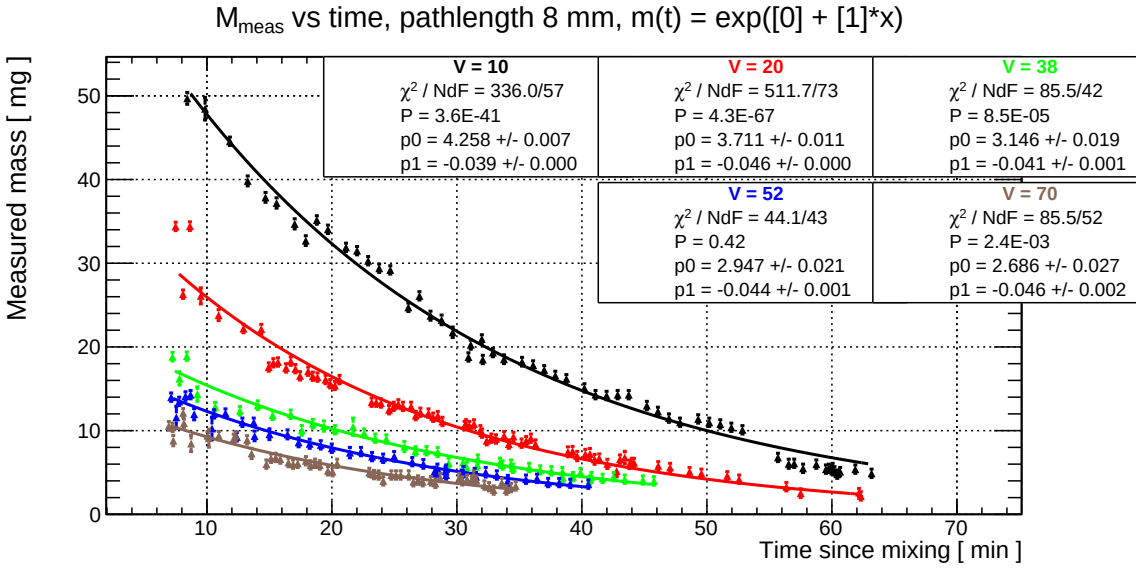


Figure 3.17: Studies of the time dependency of dispensed glue mass, performed at five different fixed speeds, to determine eventual correlations between the slope of the fit and the speed setting used. $L = 8\,\mathrm{mm}$ is the only length used in the Polaris studies.

variations between runs, coming eg. from differing concentrations of air bubbles in the glue, can lead to the discrepancies seen between some of data points close in time but far off in mass. The magnitude of the air bubble issue turned out to be significantly larger than previously assumed, but this wasn't discovered until late in the studies due to the use of black syringes instead of transparent ones.

The error bars seen on Figure 3.17 and 3.18, are quadratic sums of the standard deviation, in case of multiple measurements of the same glue line, and a flat systematic error of $0.5\,\mathrm{mg}$. The size of this systematic error was based on two things, the standard deviation seen on Figure 3.8d showcasing the quality of viscosity correction achieved with the Epolite, and iterating how low we could set this value during fitting, without completely flooring the GoF values. Ideally, we should have done multiple sets of measurement spanning the entire time interval at all five speed settings, but when there are strict limits on time and glue supplies, a certain degree of compromise as well as some improvisation is needed to succeed. We should probably have used the error estimate of $\sigma_{intrin} = 0.92\,\mathrm{mg}$ from Equation 3.4, instead of the more conservative estimate of $0.5\,\mathrm{mg}$. But as we shall see later, this would not actually affect the in-use mass calibration of the glue robot and as such, the only impact of this would be better GoF values on the plots - coincidentally making them misleading w.r.t. the information we had available to base decisions on during the development phase.

Another noteworthy thing we investigated, was the parametrisation of the raw, uncorrected scaling of mass as a function of speed setting - as seen on Figure 3.18. In order to do this, we needed to minimise the time dependency from the viscosity effect, and this was done by dispensing a single glue line at each of the five speed settings in direct succession, with no enforced pausing in between. This was then repeated five times, to see if we would get any consistency between the results. The function which best fitted the data was determined through iterative informed guessing. It is interesting to see that, to the extent of our knowledge, there is indeed a direct inverse scaling between mass and speed - but the system is too complex to be described by an integer power.

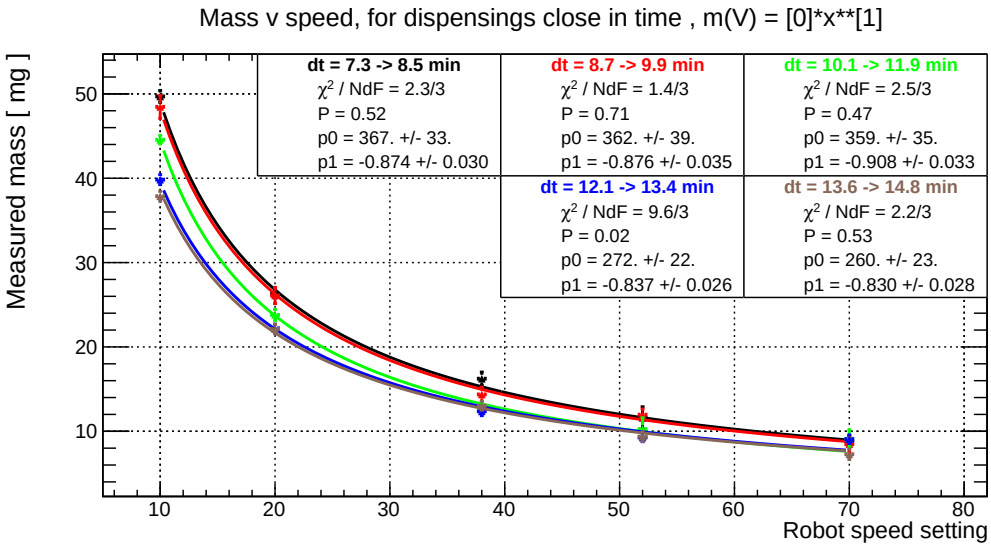


Figure 3.18: To investigate the scaling of mass vs speed, free from the time dependency, sets of data at different speeds was generated closely together in time, dt giving the start and finish time of each data set. Note how the p_0 decreases with time after mixing, which makes perfect sense - the absolute amount is expected to go down over time due to the curing effect.

After doing these initial measurements, we now have two different parametrisations of how dis-

pensed mass scales with respectively time and speed.

$$m(t, V_0) = e^{A(V_0)+b\cdot t} \tag{3.13}$$

$$m(V, t') = \frac{p_0(t')}{V^{p_1}} \tag{3.14}$$

where the mass vs time fit $m(t, V_0)$ depends on the fixed speed V_0 at which the data was gathered, and the mass vs speed fit $m(V, t')$ depends on the time after mixing t' at which the data was generated. Due to these cross dependencies, also seen in Section 3.4.1, we cannot use separation of variables to build the final function of $V(t, m)$ for finding the speed which results in a given mass at a given time. One option was to attempt a multi-dimensional fit, trying to find a unified parametrisation of $V(t, m)$ based on the above equations - but getting a 2D fit to converge nicely is notoriously difficult. Based on experience gathered during the Epolite studies, it would most likely be very time consuming fiddling with the parametrisation and starting values, and with a high probability of never really converging to something useful. So instead, we tried to come up with a solution which was simpler, more robust and less time demanding in implementation - leading to the construction of a look-up table.

3.5.4 The $V(m, t)$ Look-Up Table

Instead of building an analytical fit, mapping any value of mass and time into a value of speed, we decided to go for a more direct and data driven approach, collecting all the data gathered in the previous section into a look-up table with a set granularity. This means that we created a three dimensional table, the first index being a mass range and the second index being a time range, with the value(s) inside this combination of indices being the speed setting(s) resulting in an amount of glue being delivered within that time and mass range. It is basically an approximation of an analytical fit - directly using the collected data instead of a parametrisation of it, to map mass and time into speed. From the studies of mass estimation for the R0 pattern, see Figure 3.16, we concluded that the relevant range, plus safety margin, for the mass of an $8\,\mathrm{mm}$ line was $\Delta m \in [4; 11]\mathrm{mg}$. The time range was similarly determined as $\Delta T \in [7; 67]\ \mathrm{min}$. Based on our experience working with both the Polaris and Epolite glue, we decided that a granularity, or bin width, of $1\,\mathrm{mg}$ and $2\,\mathrm{min}$ was a decent compromise between precision and filling out the table in a reasonable manner - given the scope of our dataset.

The idea for implementing this was to write a function which took a desired mass, a bin width and a range - and then converted these informations into a corresponding integer index. Using the numbers listed above, each entry was supposed to match a mass range of $m \pm 0.5\mathrm{mg}$ with $m \in [4.5; 1; 10.5]$.

The second index of the table, took the time after mixing, in minutes, as an input, and converted it to a list index as follows:

$$ind_t = int\left(Round\left(\frac{t_{input} - t_{min}}{\text{bin width}}\right)\right). \tag{3.15}$$

We shift the zero-point to the minimal value of the range and then divide by the bin width. For a bin width going towards infinity, we expect all inputs to be put in the same index, zero, and for a bin width going towards zero we expect the opposite, that any difference, no matter how minute, will lead to a different index being assigned. This is exactly accomplished by dividing with the bin width. A similar implementation was used for the mass index. A visualisation of the $V(m, t)$ table can be seen in Figure 3.19.

In case the user ask for a mass or time value outside the accepted ranges, the ”value getter” function will print an alert and then resort to using the nearest extremum value of the range - as an approximation of what the user asked for. If multiple speed values were sorted into the same table entry, the average of these will be returned.

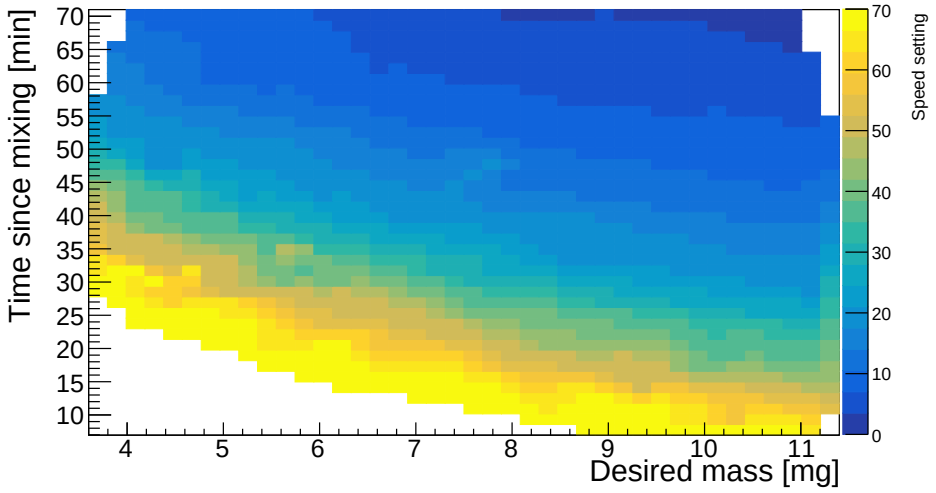


Figure 3.19: $V(m,t)$ look-up table, $L = 8\,\mathrm{mm}$. The colour axis represents the speed setting of the robot. To avoid operational errors related to empty entries, the table contains unusable speed values, $V > 70$ and $V \approxeq 0$, but these are not shown in the plot for the sake of clarity.

To fill out the look-up table we did an additional measurements series, gradually scanning through the speed settings $[5; 70]$, in steps of two speed units, while slowly decreasing the speed as time progressed. Still, due to the limited size of our dataset, and the impracticality of trying to obtain data at every single entry, we had a $V(m,t)$ look-up table with roughly half of the entries being empty. Half again of these empty entries are irrelevant combinations of mass and time - eg. it is not possible to actually get $4\,\mathrm{mg}$ at $7\,\mathrm{min}$ - the glue is too runny and the robot can't move fast enough. However, it would be quite bad if an empty value was returned by the "value getter" - the robot control software would most likely crash at an arbitrary place in the pattern, something that would require trying to clean the partially dispensed glue off of the sensor. Even if it didn't, either the speed wouldn't update or it would revert to its default setting, when an empty set-speed command was sent to the XY-table. All of these events would most likely result in a failed production.

To avoid this, we used 1D fits to interpolate speed values for all the empty entries in the look-up table. A series of 2D histograms were created, speed on the Y-axis with a binning of 3 units of speed and time on the x axis with a binning of $1\,\mathrm{min}$ - chosen such to optimise the GoF values. Each histogram represents a given mass range corresponding to the granularity of the look-up table. An exponential decay was fitted to each histogram, it was the parametrisation closest to matching the data across the series of histograms - see Figures 3.20. This parametrisation was then implemented in the table generator, such that for each empty entry, it would use the appropriate fit to interpolate a speed value to be used. In cases where the interpolated value are larger than the XY-table's speed limit, the robot goes into a waiting loop of pausing for $2\,\mathrm{s}$ then recalculating the speed value, until $V \leq 70$. Due to working with low statistics, the fits are quite rough, leading to discrepancies in the $V(m,t)$ table - where neighbouring entries jump up and down in speed value, because they were sourced respectively from the data and the fits. This leads to a balancing act between setting small bin widths in the $V(m,t)$ to improve its resolution, but on the other hand limiting the number of entries filled by the less trustworthy interpolation.

An advantage of the $V(m,t)$ approach is that it improves with usage. Meaning that, the data generated to eg. evaluate the performance of the method can be fed back into the table and interpolation fits - improving the fidelity of the mass delivery in futures dispensings. This was not possible with the viscosity correction methodology used during the previous Epolite studies - so another positive feature of the $V(m,t)$ approach.

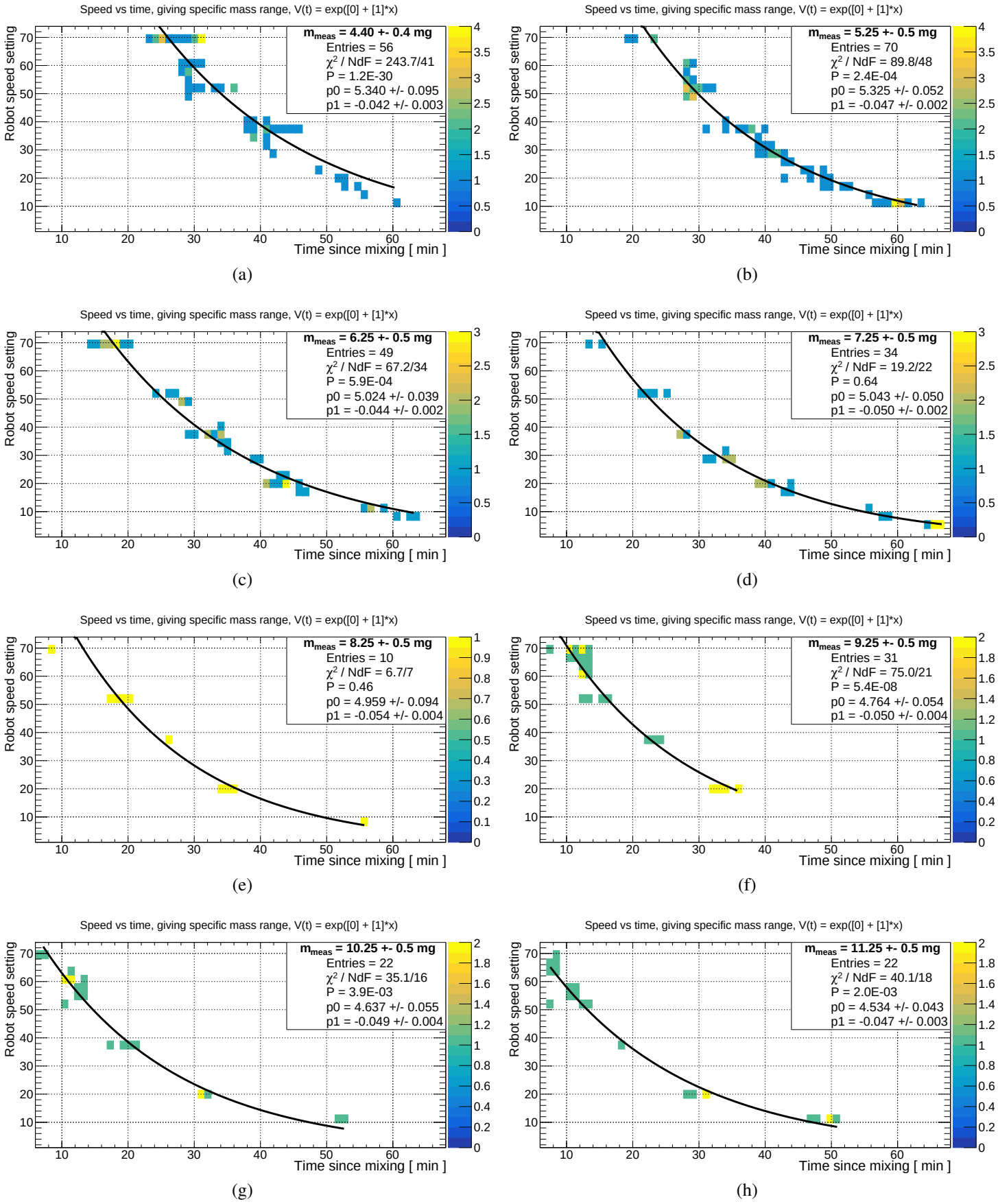


Figure 3.20: Sorting of data based on a mass range, such that one can do a direct speed vs time fit, to find the speed that, at a given time, should result in a dispensed mass within the specified range. A simple exponential decay function is used for all the fits.

3.5.4.1 Evaluating the $V(m,t)$ Performance

A measurement series was designed to evaluate the quality of the $V(m,t)$ table concept. The test consisted of repeatedly dispensing the full R0 pattern, at three different sets of mass settings, for as long as possible. This would allow us the find the time range over which the viscosity effect is properly stabilised, and what the mass accuracy of dispensing is.

The mass settings used were m_{nom}, the nominal R0 values, along with $m_{nom} \pm 2\text{mg}$. W.r.t. the total mass of the paths in the R0 patterns, This corresponds to scanning a mass range of $[15; 42]$ mg, since each path consists of $3-5$ of the $8\,\text{mm}$ lines.

During the initial phase of the Polaris studies, there was a concern that the mass of a glue path consisting of N $8\,\text{mm}$ lines would not simply be $m_{tot} = N \cdot m8_{tar}$ - with $m8_{tar}$ being the target mass of the single $8\,\text{mm}$ line. This would be due to dynamic effects like the glue string between surface and needle not breaking in between lines, leading to an excess of glue because of high surface tension pulling more liquid out of the syringe. However, preliminary measurements showed negligible difference between m_{tot} and $N \cdot m8_{tar}$ and this is also backed up by the data presented here. To analyse the data generated using the $V(m,t)$ table, we want to investigate the deviation in mass from the attempted target - Δm. Calculating this Δm can be done in two different ways, either expressing the difference in terms of the total path mass or the target line mass

$$\Delta_{tot} = M_{meas} - N \cdot m8_{tar} \tag{3.16}$$

$$\Delta_8 = \frac{M_{meas}}{N} - m8_{tar}. \tag{3.17}$$

Observe that $\Delta_8 = \Delta_{tot}/N$, meaning that if the path total mass scales dominantly with the number of lines, we expect to see the same distribution shapes, irrespective of plotting Δ_8 or Δ_{tot}, but shifted along the Y-axis by a constant factor. The paths in the R0 pattern used for these measurements have $3-5$ lines in them, with the average number of lines being $\mu_N = 4.2$. Looking at the graphs produced from this evaluation run, we see behaviour matching these expectation very nicely, further supporting the claim of $m_{tot} = N \cdot m8_{tar}$. In the plots below we only show Δ_{tot}, for the sake of brevity.

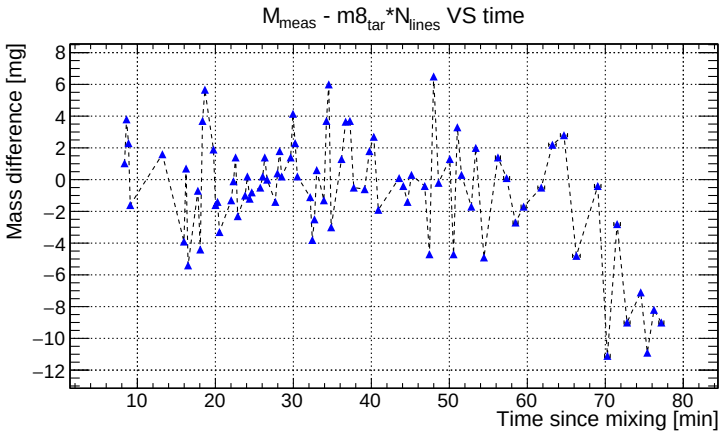


Figure 3.21: Showcasing the time stability of the glue delivery. Random fluctuations seems to be the dominating feature up until the $70\,\text{min}$ mark. The sudden decrease in mass after this point can be attributed to an insufficient viscosity correction.

Looking now at Figure 3.21, we see that the time dependence of the target mass deviations are dominated by what appears to be random fluctuations up to $70\,\text{min}$ after glue mixing. Given the lack of data at these very large times, Figure 3.17 data stops around $\sim 63\,\text{min}$, it is actually rather impressive that the viscosity correction is successful up until the $70\,\text{min}$ mark. For the following plots we cut away the data with dispensing time $t < 70\,\text{min}$.

Now, cross referring Figure 3.19 and Table 4.1, we see that, for the current R0 pattern, a glue pack is workable for a period of $40 - 50$ min, starting shortly after the 20 min mark and ending at the 70 min upper limit.

On Figure 3.22 we investigate correlations between the mass deviation and the target mass - with results being more or less as expected. The central range, $6 - 9$ mg, where the table has the most data, also has the smallest fluctuations in the Δ_{tot}. The $m8_{tar} = 3.1$ mg is below the minimum entry in the table, 4 mg, which is used as the target instead - resulting in a consistently positive Δ_{tot} for this $m8_{tar}$ - as expected. The same in reverse holds for the other extrema $m8_{tar} = 11.4$ mg. Besides these observations, only random fluctuations seems present in the data, with the amplitude of these most likely being reducible by the inclusion of more data to build the $V(m,t)$ table from - or by increasing the granularity of the table. For the following plots, we cut away the data with $m8_{tar}$ values out-of-bounds.

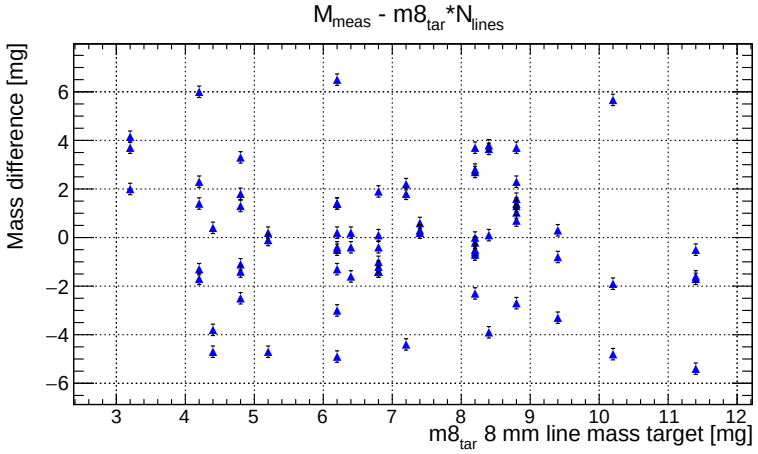
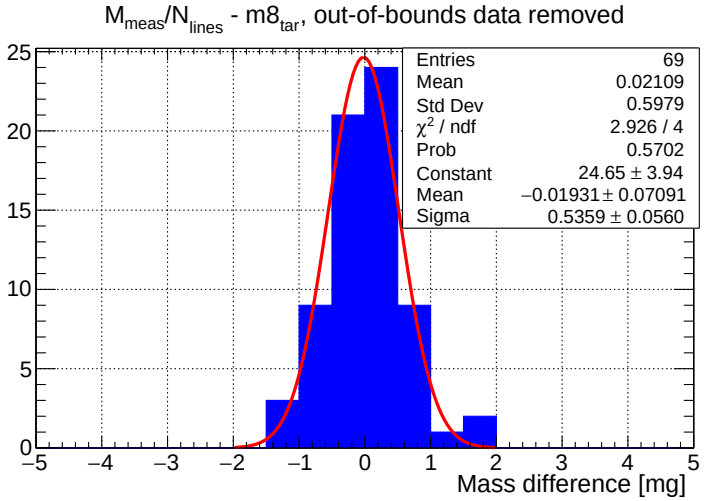
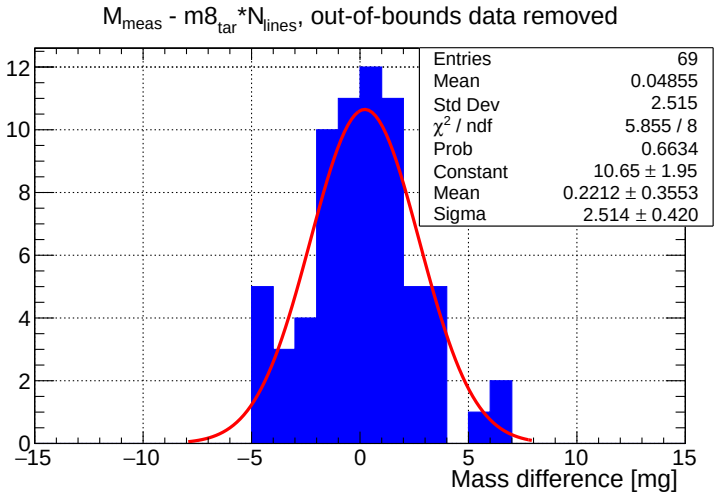


Figure 3.22: After cutting way the the $T > 70$min data, the data was evaluated for any correlation between target mass, $m8_{tar}$ and deviations from this target.



(a) Mass difference expressed in terms of the 8 mm line mass. 0.5 mg bin width.



(b) Mass difference expressed in terms of the total path mass. 1 mg bin width.

Figure 3.23: Showcasing the distribution of the deviation from the target mass for the Glue-robot using the $V(m,t)$ calibration approach. the closer this distribution is to a dirac delta centered at zero, the better performing our robot is. Both histograms have been fitted with a Gaussian distribution.

After seeing that there are no, at least obvious, unexpected tendencies in the data, we can move ahead with evaluating the overall accuracy and precision of dispensing using the $V(m,t)$ approach

to calibrate the Glue-robot. This is done by filling two histograms with respectively Δ_{tot} and Δ_{m_8}, See Figure 3.23, the accuracy can be evaluated based on how close to zero the distribution mean is and the precision based on how small the standard deviation is. It would be nice to have a smaller amplitude on the random fluctuations, but having achieved proper viscosity stabilisation for such a long period of time while also being able to dispense masses in the range $[15; 42]\text{mg}$ with very nice accuracy and reasonable precision - is a quite impressive feat. Remember also that all the data shown here, generated to evaluate the Glue-robots performance, is being fed back into the the $V(m, t)$ table to improve it.

The ITk specifications for hybrid-sensor do not directly focus on the glue amount, but rather on the glue thickness being within $120 \pm 40\mu\text{m}$, using an absolute uncertainty, and having no glue problematic glue leakage. This corresponds to a height tolerance of 30 %, and if you fit a Gaussian to the distribution of normalised mass deviation, shown on Figure 3.24, we see that a 30 % deviation is a 3.7σ event. However, it is not correct to assume a direct correspondence between glue variation and height variation, due to the usage of spacers and weights under curing. The spacers set the height while the robot determines the quality of the glue spread. Keeping that in mind, this is still a very promising result, with room for improvement - as will be discussed in the next section.

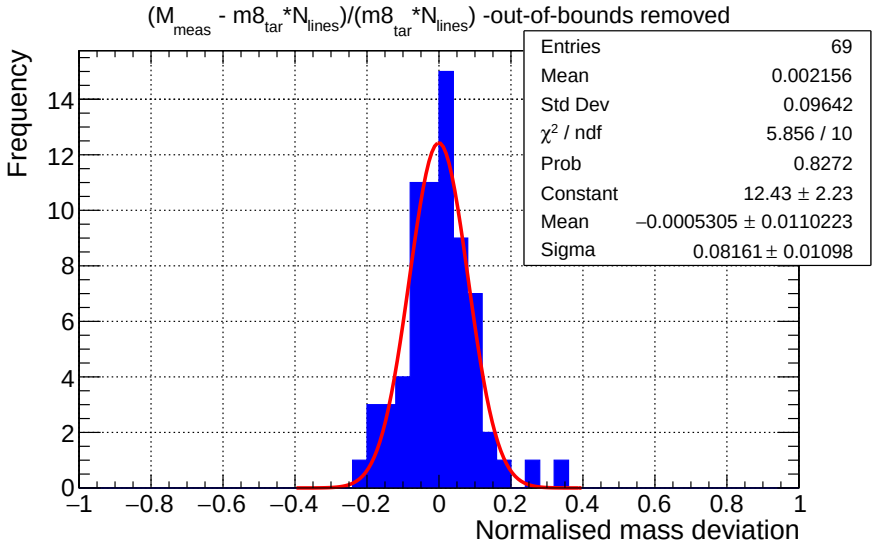


Figure 3.24: Normalised mass deviation with a bin width of 0.04 - fitted with a Gaussian distribution.

As a final evaluation of the robot performance, we look at the distribution of measured mass vs total targeted mass seen on Figure 3.25. The total target mass consists of N $8\,\text{mm}$ glue lines, each with an uncertainty corresponding to the mass granularity of the look-up table, $m_{granularity}$. This means that the error in the Y-axis is simply the scale precisions σ_{scale} while the error in the X-axis is calculated as

$$\sigma_{m_{tar}} = m_{granularity} \cdot \sqrt{N} = \frac{\sqrt{N_{lines}}}{2} \; mg. \tag{3.18}$$

While the precision showcased on Figure 3.25 is not as impressive as what we achieved using the analytical approach of the Epolite studies, see eg. Figure 3.8a, one must remember that the $8\,\text{mm}$ line look-up table approach is a fully finished method, encompassing a viscosity correction, a mass calibration and a length scaling - something we never managed to do with the analytical approach.

Furthermore, due to a very rushed development of the initial $V(m, t)$ code, the mass index ind_m was erroneously calculated as

$$ind_m = int\left(m_{input}\right) - m_{min} \tag{3.19}$$

effectively flooring the input value and shifting the zero point to the m_{min} value. This was not discovered until after producing the first electrical module and gathering the validation data presented in this section, which makes the quality of the performance demonstrated so far even more impressive.

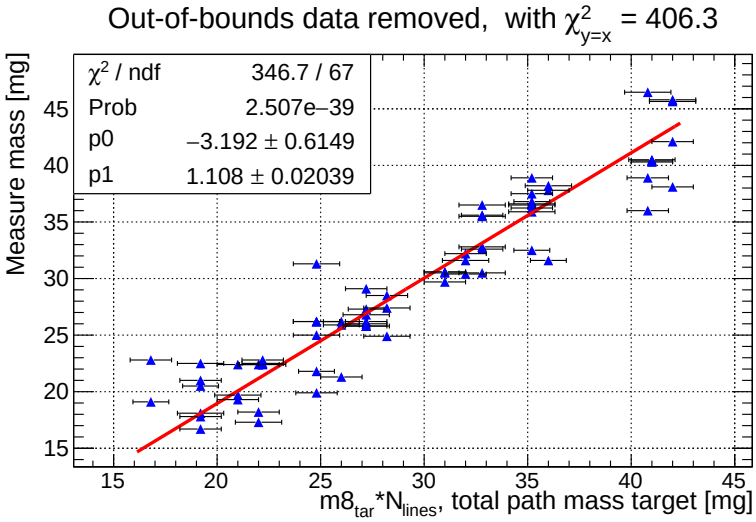


Figure 3.25: Measured mass vs targeted mass. To quantify how far away the data is from the ideal $y = x$ scaling, we compare the χ^2 of the best fitting straight line to $\chi^2_{y=x}$, the chi-square calculated for the $y = x$ case.

3.6 Summary of Development

The development of a Glue-robot for hybrid to sensor mounting, reached the first fully functioning version with the production of the first electrical module in the Scandinavian Cluster. The robot is, at the time of writing, capable of dispensing within a stabilised time window of $\sim 50\,\mathrm{min}$, with a precision in mass of 8% and for any path length within a few mm of an integer multiple of $8\,\mathrm{mm}$. This combination of flexibility and precision, to be improved as laid out below, should guarantee that the glue robot can be used in future assembly of R1 and R3 modules, only requiring that the official glue patterns be transformed into $8\,\mathrm{mm}$ equivalent lines, and a minor study to determining the accompanying $m8_{tar}$ values needed to fulfil the technical requirements, as listed in Section 3.1.2.

The performance of the robot is already quite good, and the following clear avenues of further improvement have been identified:

- Fixing the bug in the mass index generator.
- Include the data generated in the evaluation run, both directly into the look-up table and also in the interpolation fits used to fill out empty entries.
- Optimise the interpolation fits - eg. by writing a code that could find the combination of bin-widths for the three variables optimising the GoF values, or by investigating if plotting speed vs mass, for small time ranges, is better than the current implementation of speed vs time, for small ranges in mass.

Given the successful execution of this project, it is, at the time of writing, expected that the glue robot will used by our partners in industry during production - unless Uppsala acquires a better performing machine in the future.

Tales of Production

To finish off the parts of the thesis related to the assembly of detector modules, we will in this chapter give a practical overview of, how module production is actually carried out in the Scandinavian Cluster - excluding only the procedures that have yet to be fully developed at the time of writing. The intent is to give a walk through of the process and as such won't contain much data analysis - in contrast to the other chapters of the thesis.

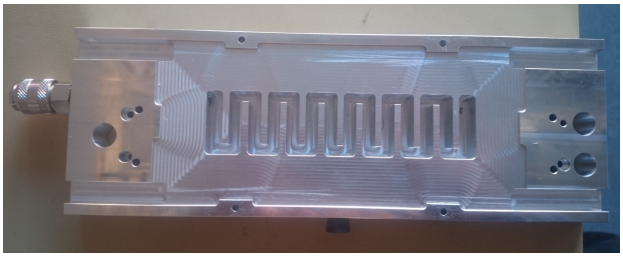
Throughout the duration of this project, I've been directly involved in producing the following:

- Two semi-electrical R2 modules - meaning functional populated hybrids mounted onto a dummy sensor, a piece of silicon cut to size and with fiducial markers for petal assembly. The purpose of this was to provide the petal assembly sites with modules to test their procedures on - Figure 2.6 explain what a petal is.
- The first electrical R0 module, which if successfully passing all quality control criteria, will be used as a part of demonstrating to the wider ITk collaboration, that the Scandinavian Cluster is capable of producing modules within specifications.

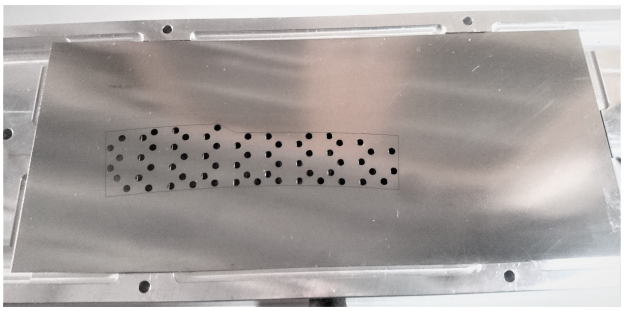
For the ASIC-to-hybrid assembly, I helped with the practicalities of production and with data evaluation of the prototyping efforts, while phd. student Eleni Myrto Asimakopoulou served as the principal developer along with the industry representative. For the hybrid-to-sensor assembly, I served as the principal developer, both w.r.t. the glue robot machinery and of the overall assembly - with generous assistance provided by phd. student Eleni Myrto Asimakopoulou.

4.1 ASIC to Hybrid Assembly

Uppsala will collaborate with a company, NOTE[23], to produce modules and as such, we needed to develop tooling compatible with using their DATACON 2200 evo pick&place machine[24]. Since we'll be producing both R1 and R3 modules, it would furthermore be nice if the tooling was developed in a rather general way, allowing for easy switching between the different types of hybrids. With this in mind, the vacuum chuck and corresponding vacuum stencils were designed - See Figures 4.1a and 4.1b. This stencil also allows for safe and easy handling of the hybrid: Simply align the hybrid on the vacuum stencil and secure it with a piece of kapton tape - then the hybrid can be transported in and out of the DATACON without risk of dropping it, bending it or touching the ASIC's after mounting. During operation, the vacuum serves a dual purpose of keeping the hybrid in place and sucking it flat - the latter being rather important when aiming for a glue layer of thickness $120 \pm 40\mu m$. While the target glue thickness is the same for ASIC-to-hybrid and hybrid-to-sensor mounting the glue used is not. For gluing ASICs, the UV activated glue Loctite AA 3525 is used.



(a) Multi purpose vacuum chuck.



(b) Vacuum stencil for the R0H0 hybrid, to be placed in the chuck.

(c) LED bar, which attaches to the vacuum chuck and is used to cure the ASIC-to-hybrid glue. Metal blinders are mounted to contain the UV light - protecting the glue in the nearby semi-transparent syringe tip from prematurely curing.

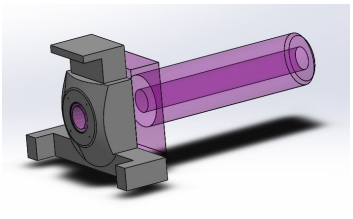
(d) The entire ASIC-to-hybrid assembly jig during final curing, after being removed from the DATACON.

Figure 4.1: The vacuum jig developed for mounting ASICs to hybrids using the DATACON 2200 evo pick&place machine at NOTE[24].

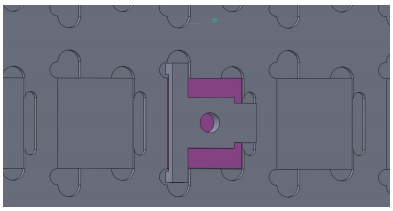
The procedure for ASIC to hybrid mounting is shown on Figure 4.2 and can be summed up as follows:

- A hybrid is attached to the matching vacuum stencil and ASIC's are placed in a chip tray. Both are loaded into the DATACON apparatus and the vacuum is activated.
- The DATACON apparatus utilises image recognition to find the hybrid fiducial marks and thereby localise the ASIC mounting pads. An ASIC is similarly picked up by locating the corners of the chip - using the custom made vacuum pick-up tool.
- The glue pattern is dispensed on the proper ASIC pad of the hybrid, the ASIC is lowered onto the glue at the proper height, and the glue is stabilised by lightly curing it for $5\,\mathrm{s}$, using the UV LED bars seen of Figure 4.1.
- After all the ASICs have been placed, the glued is cured further with the UV LED bars for $10\,\mathrm{min}$.
- If the gluing was successful, the hybrid continues on to be wirebonded.

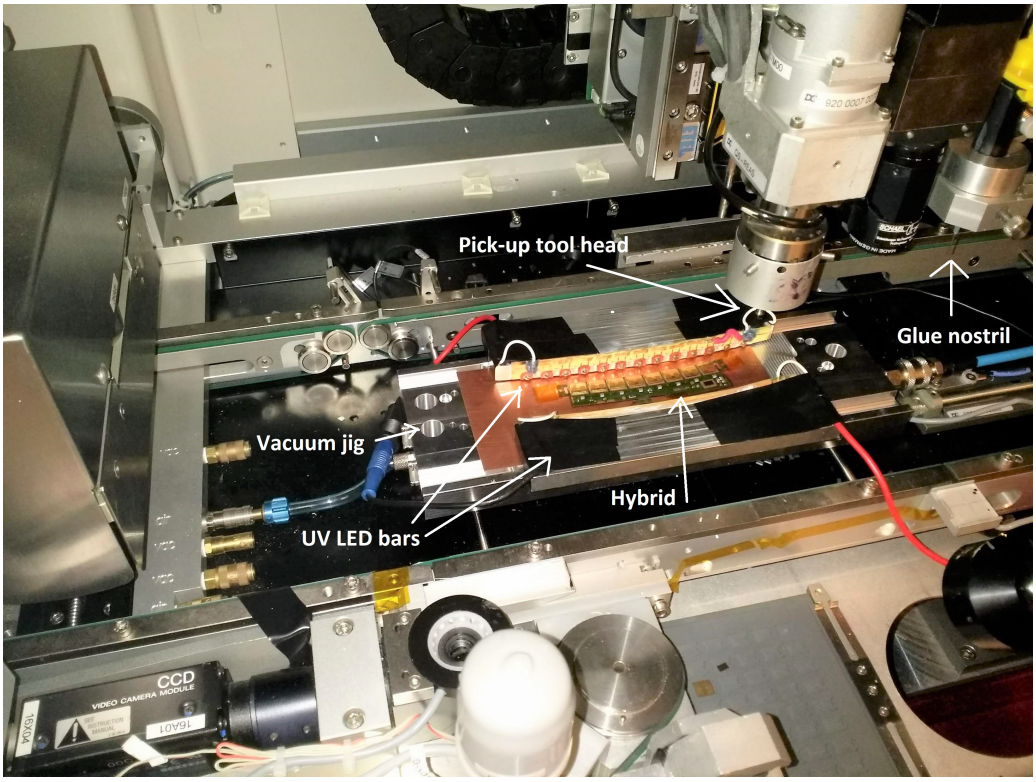
Due to machine limitations, it is not possible for NOTE's DATACON to do a touchdown measurement of the local hybrid height before placing each individual ASIC. This led to quite a bit of difficulties in getting a consistent glue height within the specifications. The glue has so low viscosity, that the machine has to hold the chip at the proper height above the hybrid surface, while doing a $5\,\mathrm{s}$ pre-cure of the glue underneath, to harden it. Also, even when the hybrid was sucked flat, we measured the ASIC pads to vary in height up to $150\,\mu\mathrm{m}$ in the worst case. The intended solution of



(a) Custom made ASIC pick-up tool, to ensure proper thickness of glue layer.



(b) Matching ASIC tray, to hold the ASICs during mounting.



(c) Picture of the DATACON apparatus in action, using image recognition and fiducial markers on the hybrid and ASIC to correctly glue and place the chip.

Figure 4.2: Visualisation of how automated ASIC placement is done at NOTE, using the DATACON 2200 evo pick&place machine[24]. The custom vacuum pick-up tool and matching tray was designed by phd student Eleni Myrto Asimakopoulou.

this problem is to use a custom made pick-up tool, see Figure 4.2a. The three legs of the tool have a precisely tuned height such that the chip hangs the intended $120\,\mu\mathrm{m}$ above the surface when the legs are touching down. Furthermore, the legs are touching down upon the hybrid bond pads - thereby limiting the possibility for glue to leak out and contaminate the bonding area for the ASIC-to-hybrid wirebonds.

Besides the height requirement, it is very important that the rotational deviation of the ASIC corners are kept within $100\,\mu\mathrm{m}$ from nominal placement [10, pg. 157]. If the ABC [1] is rotated w.r.t. to the hybrid it will complicate the front-end wirebonding between the ABC and the sensor, and in the worst cases there won't be space to bond all the 256 channels per ABC, resulting in added dead material in the detector, simply because the sensor channels aren't being read out.

[1] The ABC is the binary read-out chip directly connected to the sensor - described in Section 2.3.2.

After production is done, the hybrid undergoes quality control in the form of metrology and electrical testing. The electrical testing ensures that proper communication is established for all the HCC's and ABC's and evaluates if they perform within specifications - see Section 2.4.

4.2 Module Assembly

Before being used for module assembly, a sensor undergoes visual inspection to locate any obvious damage occurred during transport. An IV curve is also done to ensure the leakage current being within the acceptable limit - if it fails the sensor undergoes further investigation in an effort to bring the leakage current down to usable levels. After a successful IV curve has been measured, the hybrid(s) and powerboard are mounted onto it.

4.2.1 Complications of Working with high-Viscosity Liquids

When gluing the hybrid to the sensor using the two-component epoxy Polaris PF 7006A, we encountered yet another problem in preparing the glue for dispensing. As can be seen on Figure 4.3, the glue comes in pre-proportioned bags of $\sim 25\,\mathrm{g}$, with the idea being that you remove the green plastic divider and mix it in the bag before using it. This mixing step inevitably leads to the formation of some amount of air bubbles in the glue. (These problems were also relevant when using the Epolite glue, as it was packaged in a similar manner)

A problem with filling the syringe by pouring the glue in from the top, is the presence of these air bubbles suspended in the glue. This leads to a systematic variance in the amount of glue being dispensed under otherwise identical circumstances, making it an effect we obviously want to minimise. We tried vacuuming of the air bubbles, but the presence of a plunger and the high viscosity of the glue made this a problematic operation. We ended up moving away from the vacuuming idea, in favour of simply dispensing onto a sheet of clean room paper for $10 - 15\ \mathrm{s}$ while holding the syringe upside down and tapping it - until the first major spurt of air bubbles occurred. This method is not perfect, with significant amounts of bubbles still present in the glue afterwards. However, due to using black plastic syringes instead of transparent ones, as is needed for the ASIC-to-hybrid UV sensitive glue, we didn't realise the extent of this problem until time constraints hindered us from developing a better solution. As such this method might be improved by others at a later point in time.

Another issue is the dispensing height over the surface of the sensor. Due to the glue quickly becoming very viscous, it tends to ball up on the needle point of the syringe, instead of falling straight down - the "ball-up" continues until the amassed glue is heavy enough for gravity to win. This is a problem when one cares about the precise placement of the glue. If one brings the needle point closer to the dispensing surface, less "ball-up" will occur; the glue is more quickly contacting the surface - after which surface tension smoothly draws down subsequent glue coming out from the needle. However, this increases the risk of the metal needle tip directly contacting and damaging the highly sensitive sensor surface - so once again a balancing of opposing concerns had to be carried out.

(a) Example of the pre-proportioned bags of Epolite FH-5313. The user removes the green plastic spacer and mixes the two component epoxy inside the bag.

(b) Glue is poured in from the top of the syringe, then a plunger is added. Without the plunger the glue would just be pushed to the walls of the syringe instead of out through the needle.

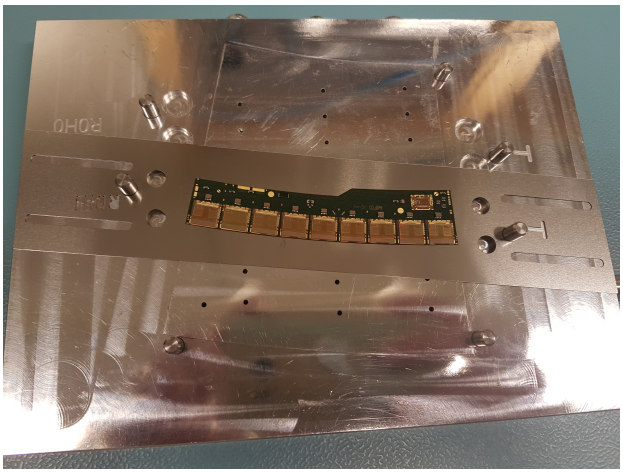
Figure 4.3: Details of how the two-component glue is packaged in a pre-calibrated manner (a), and how the glue is prepared for dispensing (b).

4.2.2 Hybrid-to-Sensor gluing Procedure

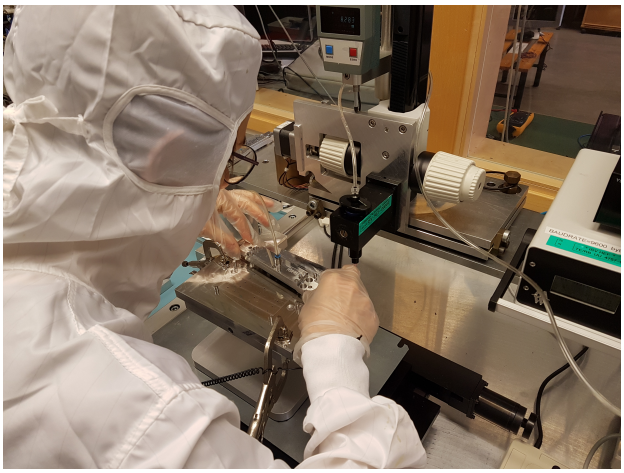
The following is a chronological list of the procedure for hybrid-to-sensor attachment, established by me and colleagues in Uppsala during the prototyping efforts - with the method being principally the same for mounting hybrids and powerboards: [2]

- Glue is taken from long term storage, a freezer kept at $-8.8\,^{\circ}\mathrm{C}$ and $29\,\%$ relative humidity, to climate controlled room temperature of $22\,^{\circ}\mathrm{C}$ - roughly $60\mathrm{min}$ before dispensing.
- The hybrids are placed into the corresponding alignment stencil and picked up using the matching vacuum pick-up tool - See Figures 4.4a and 4.4b. (The powerboard is placed by hand because the tooling wasn't yet developed.)
- The sensor is placed into the assembly jig, held in place with vacuum and cleaned using compressed air - being careful to never have the air gun closer than $\sim 10\,\mathrm{cm}$ from the sensor, to avoid random hand movement making the metal air gun scratch the sensor surface.
- An external stop watch and the robot control software is started simultaneously, the glue is mixed by hand for $3\,\mathrm{min}$, squeezing it back and forth inside the bi-pack using the plastic divider pin and a table top surface. One should minimise direct fondling of the glue, in-so-far as possible, to avoid transferring erratic quantities of body heat into the glue - thereby changing its viscosity.
- The glue is poured into a $10\,\mathrm{cc}$ syringe with a $20\,\mathrm{GA}$ metal precision tip already attached. A plastic plunger is put into the syringe from the back, serving as a solid contact surface for the high pressure air to press down on.

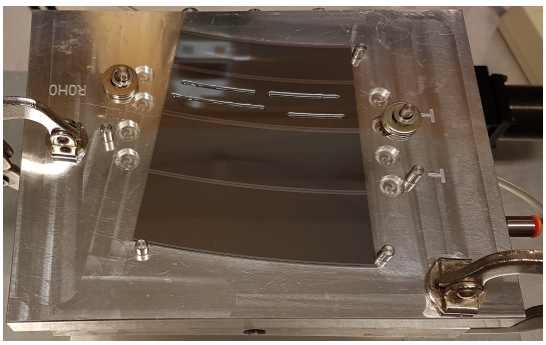
[2] The HV-tab is supposed to be attached to the sensor prior to hybrid mounting, but at the time of this assembly, the HV-tab attachment procedure was not yet implemented in Uppsala.



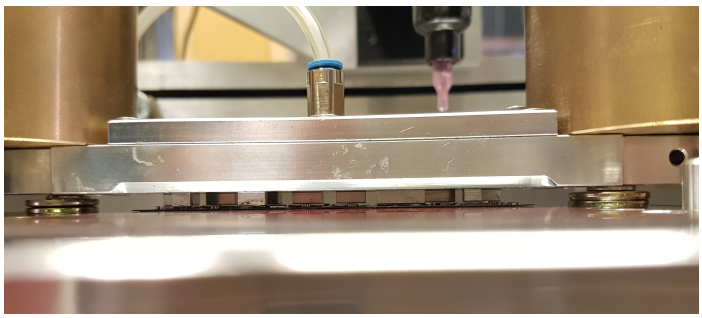
(a) Aligning the hybrid for pick-up using a stencil slotted into the precision pins of the sensor mounting jig.



(b) Using the vacuum pick-up tool to handle the hybrid.



(c) The R0H0 glue pattern dispensed by the robot.



(d) The pick-up tool is weighed down onto spacers calibrated to give a glue thickness of 120 microns after curing for 12 hours. The gap between hybrid and sensor is present but very hard to see.

Figure 4.4: The four primary steps of attaching a hybrid to a sensor

- With the syringe held upside down, glue is dispensed onto a sheet of clean room paper, while tapping the sides of the syringe, until at least one major burst of air bubbles is spurted out.

- The syringe is mounted in the Z adjustable vice, The dispensing height is calibrated by touching down on the jig surface and then setting the height which corresponds to the needle tip being $1.1\,\mathrm{mm}$ above the sensor surface.

- The XY-table and sensor coordinate system are aligned by setting the origin to the lower left corner of the sensor and using the lower right corner for rotational correction. Both of these placements are done by eye, test shows that the eye-alignment is consistent within an average of $0.5\,\mathrm{mm}$ in each direction.

- The operator commands the robot, developed by me, to dispense the relevant glue pattern, see Figure 4.4c.

- keeping the vacuum on, the mounting jig should be removed from the robot before placing down the hybrid(s) and powerboard - to free up the robot for gluing the next module. [3]

[3] This was not done for the first electrical module, seeing as we were only gluing that single module.

- The hybrid(s) are placed down with the pick-up tool(s) resting on spacers to ensure the targeted glue thickness - as seen on Figure 4.4d. weights are placed on top of the pick-up tools, to press them down onto the spacers and the module is left to cure for 12 hours.
- The powerboard is attached in a subsequent but similar procedure, due to the physical size of the pick-up tools being incompatible with mounting both of the hybrids and the powerboard at the same time. However, for the first electrical R0 module, the powerboard was attached in an ad-hoc manner - because the powerboard tooling was still under development at the time of assembly.

After module assembly, visual inspection is performed to check for obvious glue leakage in problematic places. The powerboard and hybrid are then wirebonded to a test frame, such that another round of electrical testing can be done, to detect any possible damage done to the hybrids during assembly. IV curves are also performed, to see if the assembly has had adverse affects on the leakage current, both directly after gluing and after front-end bonding has been completed. If the initial post-assembly test are successful, the module is passed on to front-end wirebonding, were the 4360 channels of the R0 sensor are connected to the ABC ASICs, with each chip handling 256 of these channels [10, pg.104]. After front-end bonding has been completed the module undergoes a final set of quality control tests in the form of the thermal cycling procedure.

Examples of finished modules can be seen on Figures 4.6, 4.8 and 4.7. For the first electrical module glued in Uppsala, the final values for pattern length and glue mass can be seen in Table 4.1. In Figure 4.5 we see the results of module metrology, evaluating the thickness of the glue layer between sensor and hybrids. As a reminder, the thickness of the glue layer is exclusively defined by the height of the spacers which the hybrid pick-up tools rest on during curing. The total amount of glue dispensed is determined by the glue robot, and whether the glue spreads out in an undesirable manner, making the module unusable, is determined by the total glue amount and to some extent the height of the spacers.

The spacers used in this build were sets of washers - as can be seen on Figure 4.4. Originally we planned to use 3-D printed plastic spacers, but we couldn't achieve the required precision with this method, whereas washers are machined to a very high degree of flatness, making them a usable last-minute ad-hoc solution. In the future this will be improved upon - eg. by machining custom spacers in metal instead of 3-D printed plastic.

The measurements of Figure 4.5 show the difference in height between the top side of the ASIC's and down to the surface of the sensor. To extract the glue height one calculates

$$z_{glue} = z_{meas} - (\text{PCB height } 0.25 \text{ mm}) - (\text{ASIC to hybrid glue } 0.12 \text{ mm}) - (\text{ASIC height } 0.32 \text{ mm})$$
$$z_{glue} = z_{meas} - 0.690\,\text{mm}$$

which gives us a hybrid-to-sensor glue height of respectively $40\,\mu\text{m}$ and $50\,\mu\text{m}$ for the R0H0 and R0H1. This is quite a bit below the requirement of $120 \pm 40\ \mu\text{m}$, a miscalibration easily fixed in future builds, simply by increasing the spacer height. Given the ad-hoc nature of the spacer implementation for this build, it is quite impressive that the R0H0 and R0H1 glue height was consistent within $10\,\mu\text{m}$.

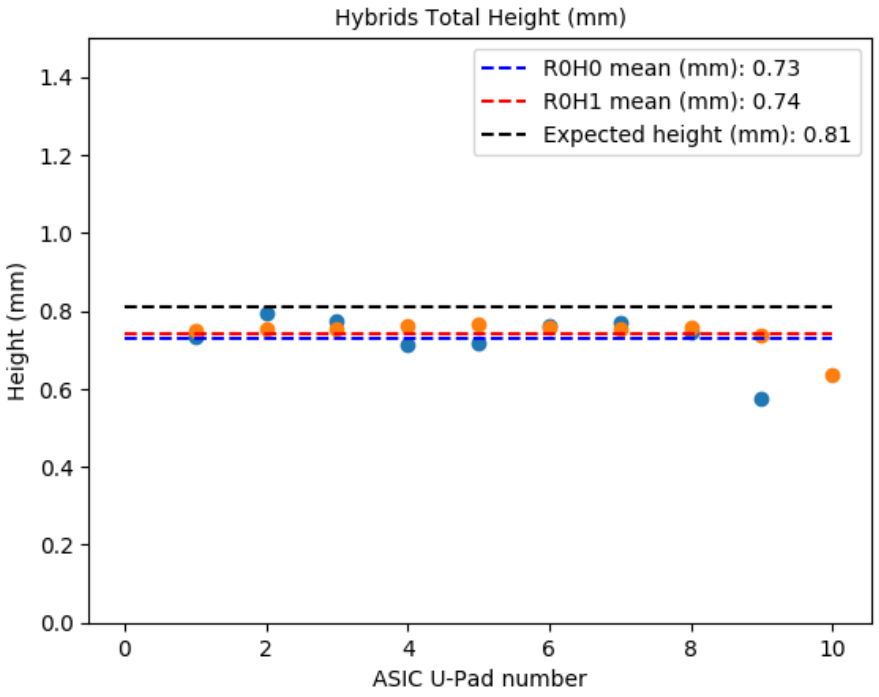


Figure 4.5: Module metrology showing the thickness of the glue layer under the two hybrids of the first electrical R0 module. The Y axis is the surface height of the ASIC's attached to the hybrid - w.r.t. the sensor surface. The X-axis enumerates the ASIC's of a R0 hybrid w.r.t. the naming scheme seen on Figure 2.14. The glue height is respectively 40 and 50 µm for the H0 and the H1 hybrid. The two outlying points are the two HCC's, having a smaller total height compared to the ABC's. They are shifted in the X-axis because the H1 have an additional ABC compared to the H0. (Figure courtesy of collaborator phd. student Eleni Myrto Asimakopoulou - Uppsala University).

	dx - mm	dx / # 8	mass -mg	mass / dx_8
			R0H0	
p0	33.3	4.2	28.0	6.7
p1	41.4	5.2	33.0	6.4
p2	33.0	4.1	28.0	6.8
p3	24.6	3.1	29.0	9.4
			R0H1	
p0	38.4	4.8	30.4	6.3
p1	45.3	5.7	29.0	5.1
p2	39.6	5.0	30.4	6.1
p3	31.5	3.9	32.2	8.2

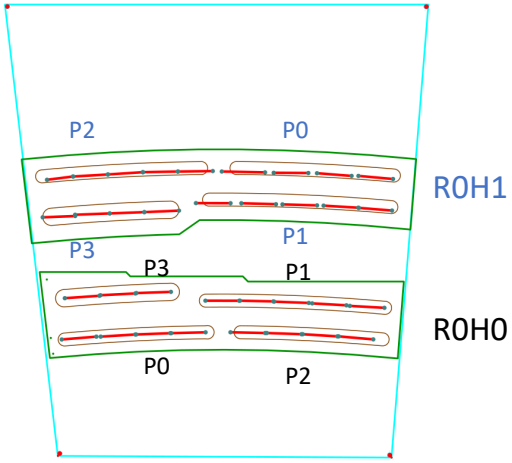


Table 4.1: Final values used for the first gluing of a electrical hybrid to sensor in Uppsala, with a visual reference of the two patterns provided to the right. $p\#$ is the line number in the glue pattern, dx the length of the line, $dx/\#8$ the number of $8\,\mathrm{mm}$ lines needed to fill out the line, followed by the total mass of each line and the mass required per $8\,\mathrm{mm}$ line. To cope with the non-integer amount of $8\,\mathrm{mm}$ lines required to match the pattern lengths, either small gaps are added in between the $8\,\mathrm{mm}$ lines or an extra $8\,\mathrm{mm}$ line is added.

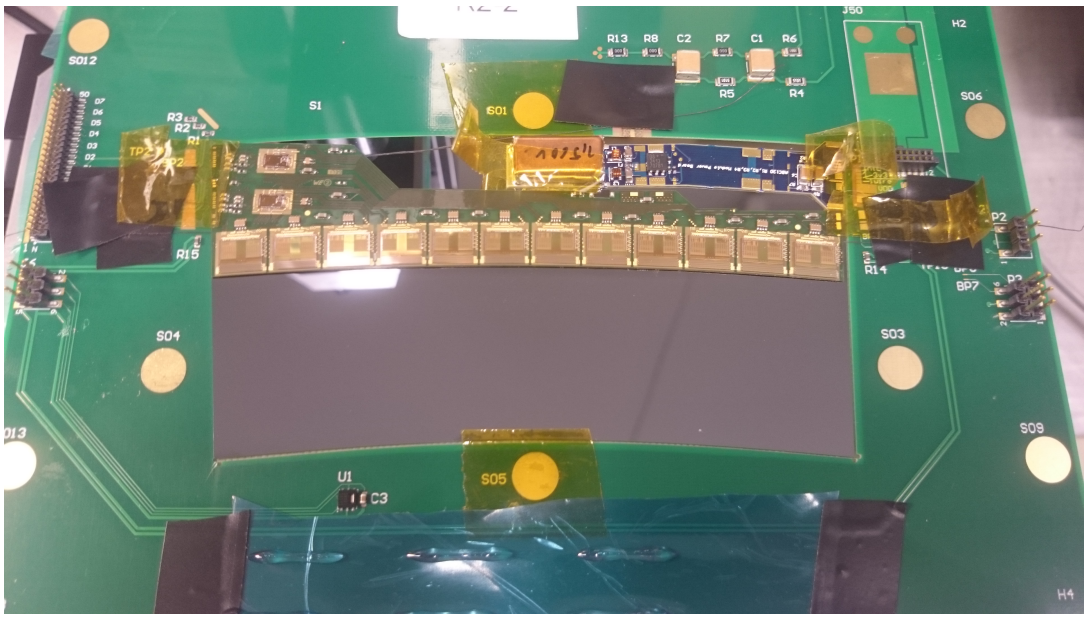


Figure 4.6: One of two semi-electrical R2's made by the Scandinavian Cluster. Dummy sensors were used, so a minimal effort approach of using fishing thread with diameter $120\,\mu\mathrm{m}$ was used to set the glue height. The blue PCB is the power-board. The glue dispensed on the blue plastic were quality control lines dispensed before and after the deal dispensing, so we had a way to verify the actual amount of glue put onto the silicon.

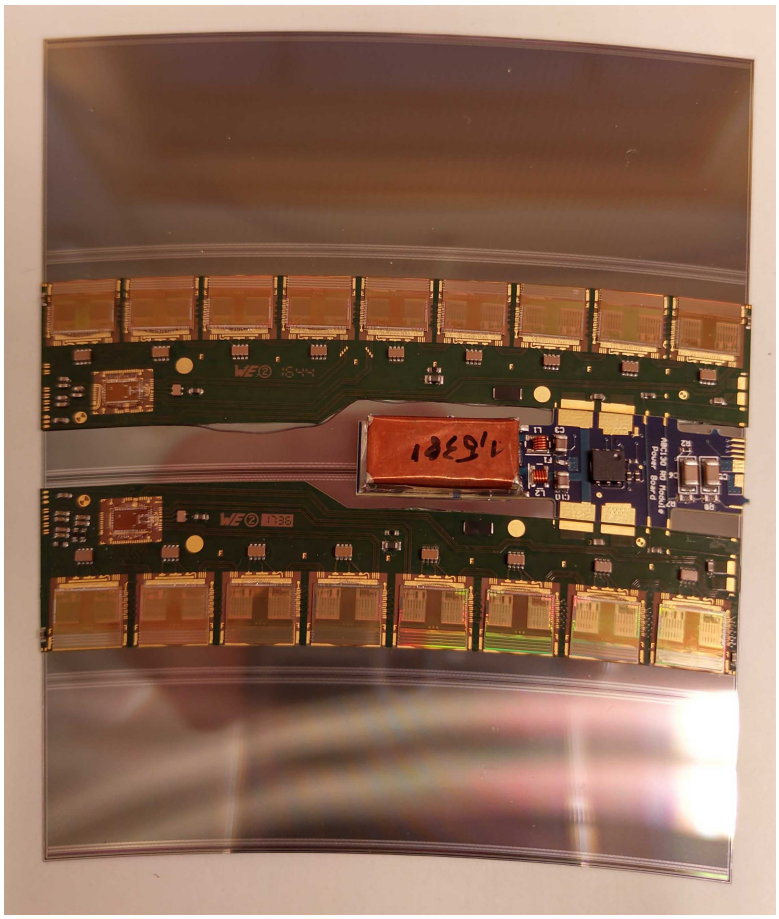
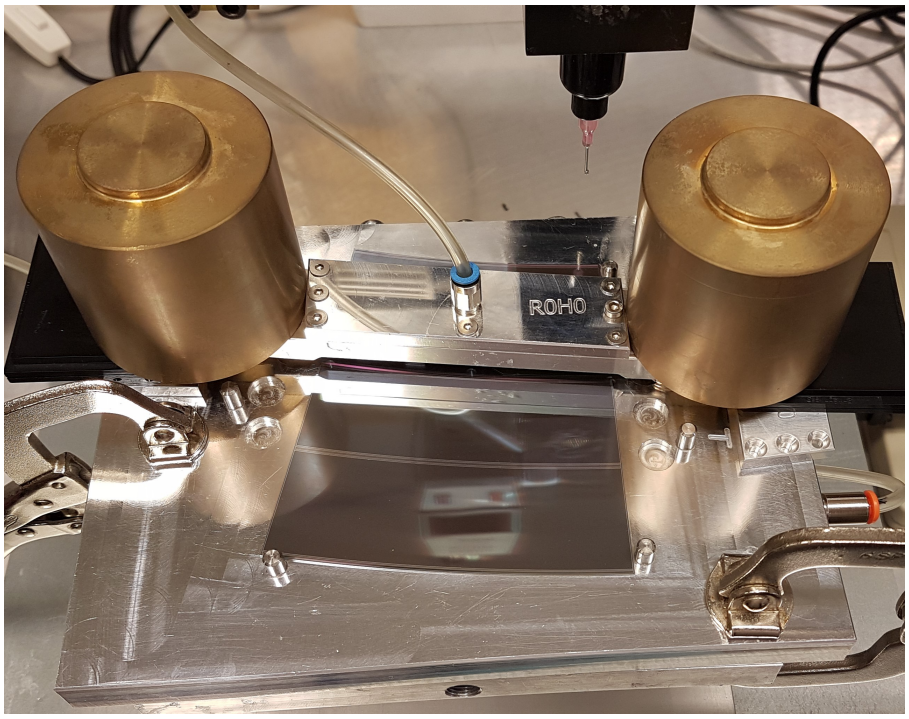
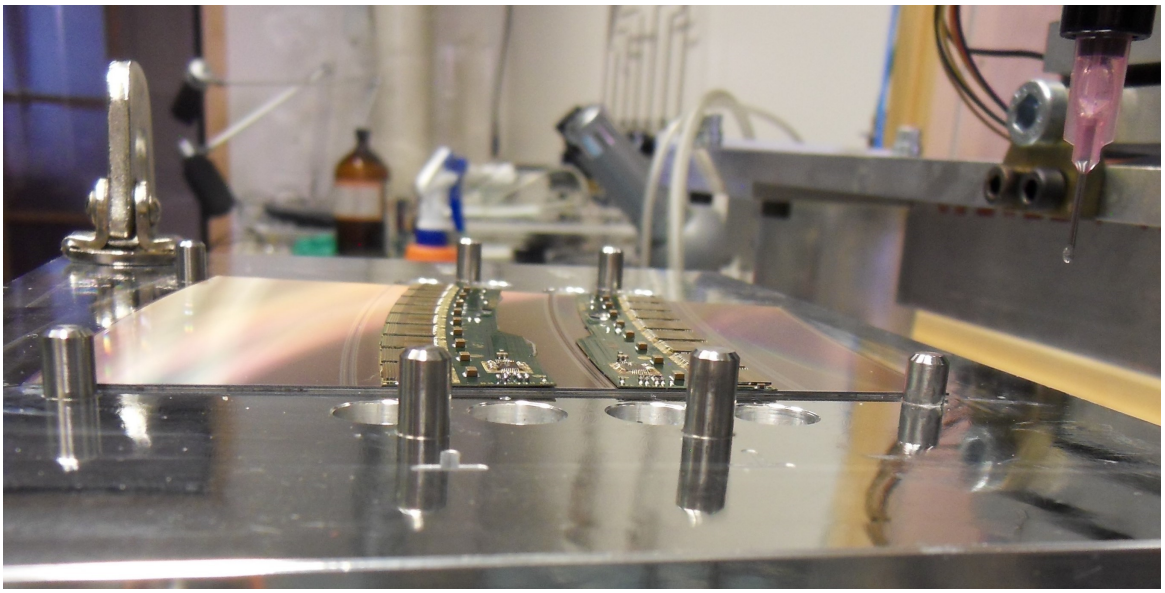


Figure 4.7: Top side view of the eletrical R0, including the powerboard, prior to being shipped to Oslo for front-end wirebonding. The strips are going in the vertical direction. The segmentation of the four sets of strips can be seen by the lighter grey bands running horisontally across the sensor - with each hybrid being wirebonded to two of the segments.



(a) Curing stage of R0H0 mounting. The two brass weights press the vacuum pick-up tool, holding the R0H0 hybrid, down onto two spacers, such that the hybrid is held the intended height above the surface while the glue cures.



(b) Side view of the electrical R0.

Figure 4.8: First electrical module assembled in the Scandinavian cluster. an ATLAS12EC type sensor[29] was used along with the 130 generation of ASIC's [10, Chap 6.2]. There's a slight glue leakage under the H1 hybrid, to the left on (b), but it is not problematic since the front-end ASIC-sensor bonds are on the other side of the hybrid. The sensor placement is aligned by pushing it against the three corner pins seen, while the hybrid placement is handled by the pick-up tools slotting into the pins on each side of the sensor

Sensor Studies

The silicon sensor is the primary component of the modules which make up the ITK, and in this thesis they have been studied in the context of quality control during module production. In this chapter we'll first review relevant semiconductor theory, before proceeding on to the studies of early-onset behaviour, referring to low voltage breakdowns in sensor current, unexpectedly seen in full size R0 sensors and the investigation of probable causes and solutions of this problem.

5.1 Semiconductor Theory of Silicon Particle Detectors

When it comes to electrical properties of materials, we generally divide solids into one of three categories as listed on Figure 5.1 - based on their degree of conductivity. The microscopic basis for this sorting comes from the energy states of respectively bound and free electrons inside solids. The electrons of an atom occupies sets of discrete energy states, also known as orbitals, but when $N \propto 10^{25}$ atoms are combined into a macroscopically sized material, these discrete energy states undergo a N-fold degeneracy, due to the Pauli exclusion principle. These minutely different discrete states effectively create continuous bands of electron energy states, centred around the energy levels of the single atom orbitals.

A bound state refers to an electron constrained in space by the electromagnetic potential of its parent nucleus, or possibly the potential of a few nearby nuclei - if the electron is part of the covalent bonding between these neighbouring nuclei. A free state refers to an electron sufficiently energetic to escape the localised potential well of its parent nucleus, but not energetic enough to escape the total superimposed potential of the solid - meaning that the free state describes electrons effectively propagating freely inside the material. The energy required for an electron to jump from the valence to the conduction band typically comes from thermal excitations or external sources such as ionising particles traversing the material.

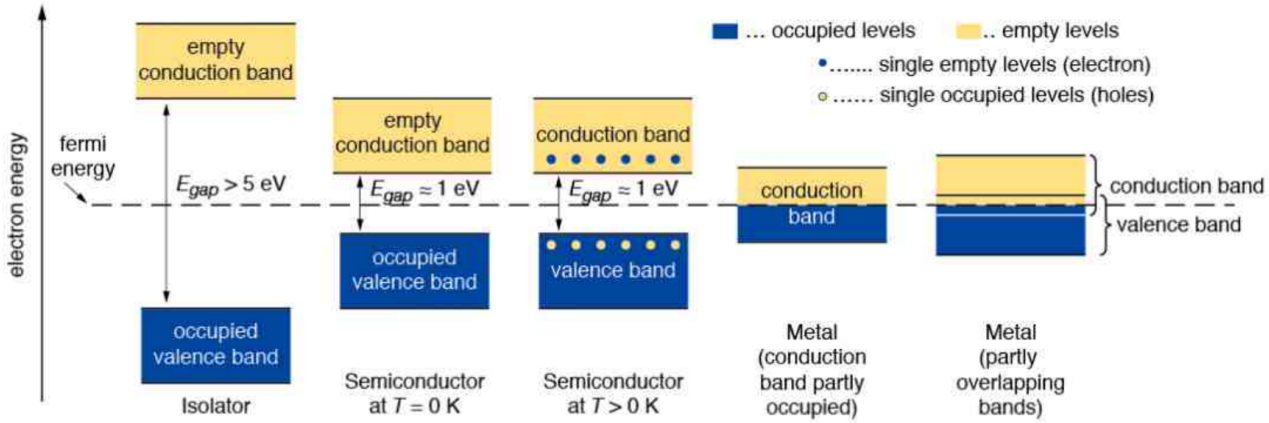


Figure 5.1: Band gap model for insulators, conductors and semiconductors. The Valence band is the most energetic filled band while the conduction band is the least energetic free state band [27].

The division of solids into insulators, semiconductors and conductors is quantified based on the band-gap of the material - the energy required to excite an electron from a bound state at the top of the valence band into a free state in the bottom of the conduction band - the minimal difference in energy between bound and free electron states. Said in other words, the energy required to mobilise a charge carrier inside the material, as the band-gap increases the electrical conductivity decreases. Semiconductors are a class of materials with band-gaps in between that of proper insulators(conductors) and

this allows us to very handily manipulate semiconductor systems to either allow or block the flow of currents - depending on external inputs to this system.

Silicon is a tetravalent indirect crystalline semiconductor and the primary focus of this theory review. Crystalline refers to the macroscopic structure being composed of a repeating lattice of silicon atoms, while tetravalency refers to the four electrons in the outermost orbital - utilised in covalent bonding with four neighbouring atoms as seen on Figure 5.2. In an energy band diagram, see eg. Figure 5.1, the horisontal axis oftens details the crystal momenta of particles, a handy parametrisation of particle motion in the periodic potential of the solid. Crystal momentum is conversed in the same manner is proper momentum, and this leads to the distinction between direct and indirect semiconductors. In a direct semiconductor the minima of the conduction band and the maxima of the valence band is located at the same value of crystal momenta, allowing an electron to transfer directly from band to band - given an excitation energy of E_{gap}. In an indirect semiconductor, these energy extrema are located at different values of crystal momenta leading to required excitation energies larger than simply the band-gap value E_{gap}. To conserve crystal momentum, electrons either use intermediate states, coming from crystal defects, to eg. shed excess momentum, or a direct transfer occurs, but at the non-minimal energy band-gap. This is one of the primary reasons as to why the band-gap in silicon is $E_{gap} = 1.15eV$, but the mean ionisation, or rather the electron-hole pair creation, energy is $3.6eV$ [18, chp 4].

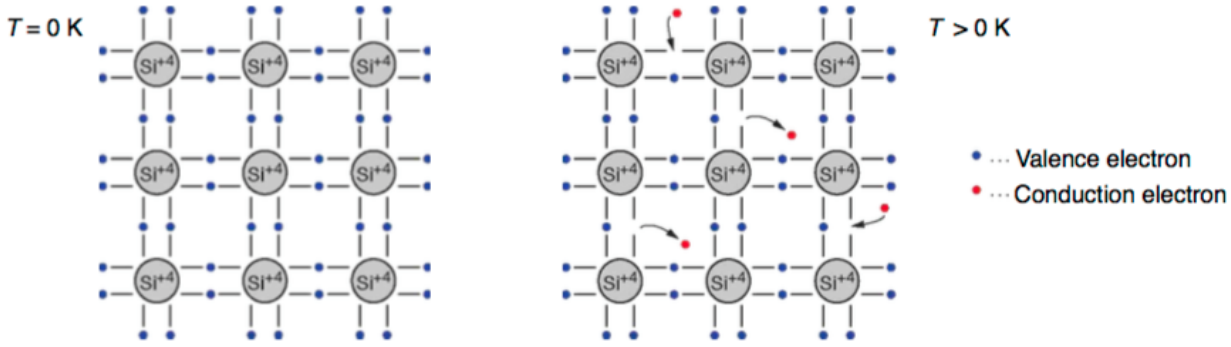


Figure 5.2: 2D projection of the tetravalent structure of a silicon crystal, also showing how thermal agitation causes valence electrons to be ionised[27].

When a valence band electron is ionised, the local charge neutrally balanced by electrons and protons is disrupted - there's now a lack of negative charge, something that can also be perceived as a localised positive charge. This localised positive charge, also called a hole, is mobile in the way that it can attract a nearby valence electron, effectively switching places. Hole's are a type of pseudo-particle, they cannot exist in a vacuum, but it can be shown that for many intents and purposes, electrons in the conduction band and holes in the valence band, can be treated as actual free electrons - except for a different effective mass, and the charge sign of holes, with this effective mass representing the affects of the background periodic potential from the atomic lattice. Also, both particles can propagate freely through the lattice, though at different speeds due to their varying effective mass, meaning that both parties can conduct a current. Whenever a valence electron is ionised into the conduction band, a hole is also created in the valence band - meaning that electrons and holes are always created in pairs. It is of course also possible for a conduction electron and a valence hole to recombine into a valence electron, leading to an equilibrium value being established

for the free charge densities of the lattice

$$n = p = n_i = \sqrt{N_c N_V} e^{-\frac{E_{gap}}{2k_b T}} \tag{5.1}$$

$$N_{C/V} = 2\left(\frac{2\pi m_{n/p} k_b T}{h^2}\right). \tag{5.2}$$

The free electron(hole) density is labelled by $n(p)$, $N_C(N_V)$ are the density of states in respectively the conduction and valence band, T the temperature in Kelvin, $E_{gap} = E_C - E_V$ the band-gap at $T = 0\,\mathrm{K}$, given by the maximum of the valence band E_V and the minimum of the conduction band E_C, with Boltzmann's constant k_b and Planck's constant h. These equations are derived based on Fermi-Dirac statistics and an assumption of a lattice consisting solely of silicon atom - at $T = 300\,\mathrm{K}$ one can estimate the intrinsic charge carrier concentration to be $n_i = 1.45 \cdot 10^{10}\ cm^{-3}$.

5.1.1 Doping

By intentionally introducing specific impurities into the silicon lattice, a process known as doping, we can manipulate the free charge carrier densities of the material, and thereby its electrical properties. There are two types of doping, n and p doping - based on the two types of charges carriers in semiconductors. In n-doping, the dopant element is selected based on having a surplus of electrons in the valence orbital compared to the bulk material, and a valence band just below the conduction band of the bulk material. This surplus of electrons is not part of any covalent bonding between atoms, and they are very easy to excite into the conduction band, meaning that you effectively increase the free electron density proportionally to the employed doping concentration. p-doping is the reverse, the dopant element has fewer valence electrons than the bulk material, and the dopant conduction band sits just above the valence band of the bulk, attracting bulk valence electrons into its vacancies and creating an increase in the free hole density of the material.

Typical elements used to dope tetravalent silicon is, for n-doping the pentavalent Phosphorous with a with valence band $45.3\,\mathrm{meV}$ below the silicon conduction band, and for p-doping the trivalent Boron with a conduction band $45\,\mathrm{meV}$ above the silicon valence band. Typically we call the n-doping elements donors and the p-dopants acceptor, because they either donate or take an electron from the bulk lattice. The scale of doping is typically in the range $10^{12} - 10^{18}\mathrm{atoms/cm^3}$, and if it is above $10^{16}\mathrm{atoms/cm^3}$ we label it as n+ or p+ doped, to signify the high dopant concentration.

For a homogeneously doped semiconductor, we can estimate the free charger carrier densities very well by equating the dopant concentration with the corresponding free carrier density. This relation fails at very high dopant concentration and for mixed doping scenarios, eg because at sufficiently high doping, each dopant atom won't necessarily replace a bulk atom in the lattice. The total free charge densities, n and p, can then be evaluated as

$$n = N_D = n_i e^{\frac{E_F - E_i}{k_b T}} \tag{5.3}$$

$$p = N_A = n_i e^{\frac{E_i - E_F}{k_b T}} \tag{5.4}$$

$$E_i = \frac{E_C + E_V}{2} + \frac{3k_b T}{4} ln\left(\frac{m_p}{m_n}\right). \tag{5.5}$$

$N_D(N_A)$ is the donor(acceptor) dopant concentration, E_F is the Fermi energy, the energy at which the occupation probability of a possible electron state is $\frac{1}{2}$, a parameter modified by the introduction of dopants, while E_i is the energy level in the middle of the band-gap scaled by the difference in effective mass if free electrons and holes.

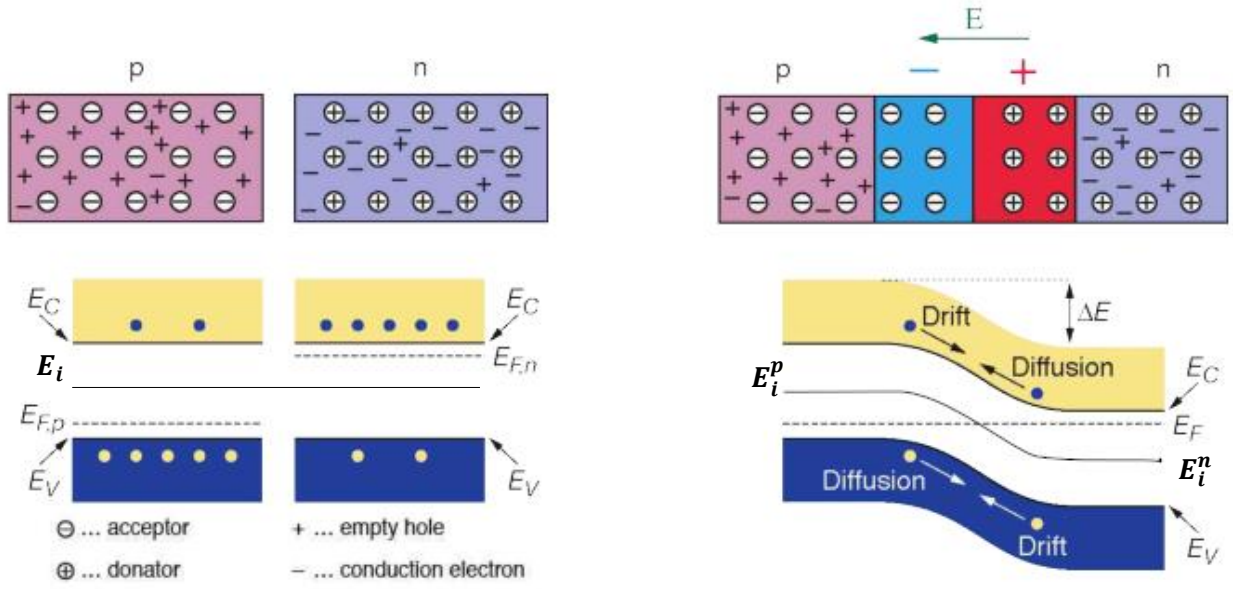


Figure 5.3: pn-junction visualised through the Band gap model, the balancing act of the majority carrier diffusion current and the minority drift current is highlighted. [27]

5.1.2 The pn-Junction

The pn-junction is the theoretical basis for using semiconductors as radiation detectors, it describes a semiconductor where one part of the solid has been n-doped, while the other part is p-doped, leading to a junction between the two forming. In the n-doped part there will be an excess of free electrons and in the p-doped part an excess of free holes. Thermal agitation will instigate a diffusion of the free charges, causing a recombination of electrons and holes. But since the dopant elements were originally electrically neutral, this recombination results in a net polarisation of the lattice near the junction interface, the acceptor atoms becoming negatively charged due to the absorption of an electron - and vice verse for the donors. This build-up of charge difference across the junction generates an electric field with a direction such that opposes the diffusion of majority charge carriers - eg. electrons flowing from the n-doped side towards the p-doped side. The E-field also separates and ejects any electron-hole pairs eg. thermally generated inside the field region - meaning the field is self-sustaining since the excess nuclear charge won't be neutralised over time by mobile charge carriers. The magnitude of the electric field grows with the degree of lattice polarisation until an equilibrium is reached where the majority carrier diffusion current is balanced by the minority carrier drift current - with the minority charge carrier drift current being eg free electrons, accelerated by the E-field, to drift from the p-side to the n-side of the material. As a quick side-note, the concentration of minority charge carriers can be estimated from considering the mass-action law stating that, in thermal equilibrium, the product of two interacting concentrations should equal a constant - or, an increase in the majority carrier is accompanied by a decrease of the minority carrier - since some of the added majority charges will recombine with the minority free charges. Algebraically, this can be stated as

$$n_n p_n = n_i^2 \tag{5.6}$$

with n_n being the dominating free electron density in the n-doped semiconductor and p_n being the minority free hole density.

The concept of a pn-junction is visualised on Figure 5.3, where two, initially separate, n and p doped pieces of semiconductor are brought into contact. When adding n(p) dopants we except the Fermi level to go up(down) because more electron states become available at higher(lower) energies. In thermal equilibrium we except the Fermi level to be identical across the material, eg. for reasons of continuity, which in the case of a pn-junction leads to a gradient in the conduction and valence band energy across the junction, with the change in band energy, ΔE on Figure 5.3, proportional to the electric potential across the junction,

$$V_0 = \frac{\Delta E}{q} = \frac{E_i^p - E_i^n}{q}. \tag{5.7}$$

where q is the elementary charge, E_i^n and E_i^p follows the notation of Figure 5.3 and is the, almost, halfway point of the band-gap in respectively the n and p doped parts of the junction - see Equation 5.5. Using Equations 5.3 and 5.4, we can estimate the internal potential across the junction V_0.

$$N_A N_D = n_i^2 e^{\frac{E_F - E_F + E_i^p - E_i^n}{k_b T}} \leftrightarrow \tag{5.8}$$

$$E_i^n - E_i^p = k_b T ln\left(\frac{N_A N_D}{n_i^2}\right) \leftrightarrow \tag{5.9}$$

$$V_0 = \frac{k_b T}{q} ln\left(\frac{N_A N_D}{n_i^2}\right) \tag{5.10}$$

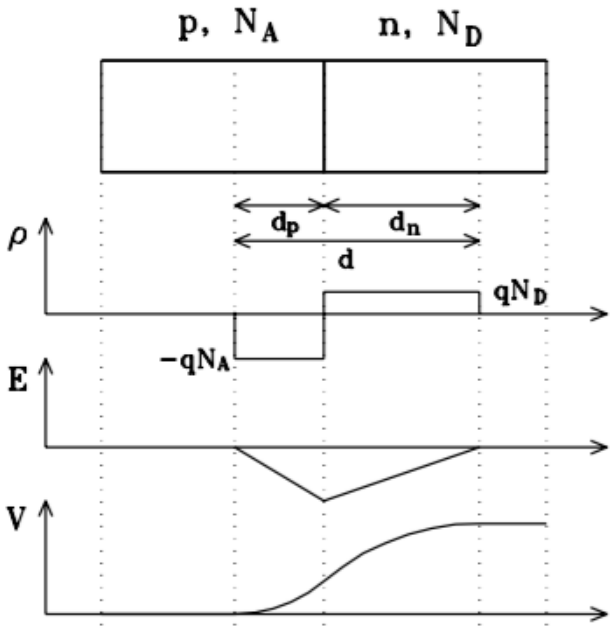


Figure 5.4: Visualisation of the assumption of abrupt charge distribution change, and the resulting electrical field, potential and charge distribution across the pn-junction[18]

Now, using the abrupt change assumption, meaning that the charge distribution across the junction is modelled as a step function - See Figure 5.4, and expecting the electric field to die out at the boundaries of the space charge region, since it is generated by these immobile charges, we can calculate an estimate of the maximal E-field across the junction, and, perhaps even more important, the extension of the space charge region into the n and p doped parts of the semiconductor, which combined is typically referred to as the depletion width of the pn-junction.

The expectation for zero E-field at the space charge boundaries means the net charge of the region should be zero, implying that $N_D d_n = N_A d_p$, where d_n (d_p) is the extent of the depletion zone into

the n(p) doped region of the semiconductor. Given these conditions, the maximum of the electric field is simply

$$\mathcal{E}_{max} = \frac{qN_D d_n}{\epsilon} = \frac{qN_A d_a}{\epsilon} \tag{5.11}$$

with ϵ being the permittivity of the semiconductor. Rewriting this into equations for the electrostatic potential in respectively the n and p doped part of the solid, recalling Equation 5.10 and doing a bit of algebraic gymnastics we arrive neatly at an expression for the total depletion width of the pn-junction

$$d = d_n + d_p = \sqrt{\frac{2\epsilon\,(N_A + N_D)}{qN_A N_D} \cdot (V_0 + V_{ext})} \tag{5.12}$$

An important feature of this equation for depletion width is that in the limit eg. $N_D >> N_A$, you'll find that $d_p >> d_n$, meaning that the depletion zone prefers to extent into the lighter doped part of the semiconductor. This feature is exploited when using planar fabrication technologies to produce semiconductor radiation detectors, eg. the case of microstrip trackers, where very thin sections of highly doped silicon are used to deplete the entire bulk of the sensor.

It can be shown that if you apply an external voltage across the junction, this will either amplify or diminish the depletion width, and while this external contribution to the depletion width is obviously not part of the equilibrium derivation of Equation 5.12, the V_{ext} term was added for the sake of brevity - so as to not list essentially the same equation twice in a row. If one applies a positive terminal to the n-side and a negative terminal to the p-side, the majority charge carriers will be attracted away from the junction, thereby increasing the width of the depletion zone - this is called reverse biasing and we always aim to operate a semiconductor radiation detector at sufficiently high reverse bias as to deplete the entire volume of the sensor.

5.1.3 Leakage Current

A reversely biased pn-junction should ideally block any and all current from passing through it, but in real devices like the ITK strip sensors you still observe a small current, on the order of $\mathrm{nA/cm^2}$, a phenomenon named leakage current, which is caused by a combination of bulk and surface related effects. The bulk leakage current comes from the thermal generation of electron hole pairs, either in the neutral n or p regions or inside the depletion zone. For pair creation in one of the two neutral regions, the generated majority charge will be blocked by junction E-field, but the minority charge, once it has diffused from its point of creation over to the space-charge boundary, will be accelerated across the junction and contribute to the minority drift current of the device. when the pair creation happens inside the depletion zone, the electron and hole are automatically separated and ejected from the depletion zone by the E-field - we call this the volume current of the device. While the electron and hole currents are opposite in direction, due to their opposite charge sign, they add up positively when calculating the total current. The volume current is by far the dominating source of bulk leakage current because the minority drift current is suppressed by the diffusion time required to charges to reach the space-charge boundary. Reminding ourselves that, for an unbiased pn-junction we except zero total current, a simple an elegant estimate of the volume current density is

$$J_v = qG_{th}\,(d - d_{V_0}) = q\frac{n_i}{\tau_g}\,(d - d_{V_0})\,. \tag{5.13}$$

The rate of thermal pair generation G_{th} times the increase in the depletion width due to reverse biasing, with n_i the intrinsic free charge density and τ_g the average lifetime of thermally generated free charges.

While the bulk leakage currents are due to the inherent physics of the system, the surface leakage currents are more often related to the imperfections of the system. This could be due to contamination of the surface, eg. dust or the absorption of condensable vapours like water, altering the electrical characteristics of the surface. Surface effects like these are generally very tricky to model theoretically, but since surface defects can easily be the dominating feature of real life devices, as we'll see later, it is very important to gain a understanding of how to avoid or mitigate these problems when constructing a semiconductor based detector.

In general, we expect a relation between the biasing voltage and the leakage current to follow a trend like what is seen on Figure 5.5. In forward biasing, shown as a positive voltage on the graph, the external voltage opposes the intrinsic one, turning the junction into a regular conductor, and with reverse biasing the depletion region growths and only the small leakage current passes through the junction. However if one increases the reverse biasing voltage too much, a sudden and large increase in the current, called a breakdown, occurs, because the device once again acts more like a normal conductor. There are two common types of breakdown mechanism, in a Zener breakdown, the magnitude of the electric field is large enough to ionise significant quantities of the valence electrons, and possibly even the electrons of the inner shells, which are normally constrained in the potential wells of their parent nuclei. In an avalanche breakdown, the free electrons are so energetic that cascades of secondary ionisation starts taking place.

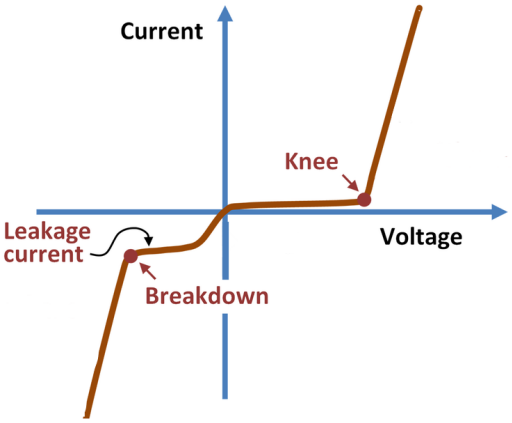


Figure 5.5: Sketch of idealised IV curve for a pn-junction device[28].

5.2 Semiconductors as Charged Particle Tracking Detectors

For high energy particle tracking, one typically uses semiconductor or gas based systems, the former providing the highest spatial resolution and fastest response time but also a higher material budget and a much higher price tag than the gas based solution.

For the ITk, we'll solely use semiconductor tracking sensors - the expected particle flux will simply be too high for the slower gas based systems. As regards to particle-detector interactions, we want traversing particles to consistently leave a minimal amount of energy in the radially subsequent layers of sensors such that we can reconstruct the trajectories accurately without interfering too much with the following measurement of particle energy - any energy deposited outside the calorimeter increases the systematic error of the energy measurement.

When using a pn junction as a detector, high energy particle traverse the fully depleted sensor, ionising a bunch of electrons and holes, which the internal E-field accelerate to different ends fo the crystal where a read-out circuit is attached to pick up the generated charge. We can try to estimate the current pulse generated by a traversing Minimum Ionising Particle (MIP) and compare it to an

estimate of the steady state leakage current, given by Equation 5.13. For a silicon pn-diode $1\,\mathrm{cm}^2$ in area and $300\,\mu\mathrm{m}$ thick at room temperature $300\,\mathrm{K}$, with a mean energy loss per length of $dE/dx = 3.87\mathrm{MeV/cm}$ and the mean pair creation energy $E_{pair} = 3.6\mathrm{eV}$, we find the number of generated electron hole pairs to be [27]

$$N_{pairs} = \frac{dE/dx \cdot thickness}{E_{pair}} \approx 3.2 \cdot 10^4 \tag{5.14}$$

To estimate the timescale of this charge generation we make the following considerations:

- The MIP is basically travelling at the speed of light, so it traverses the diode in $1\,\mathrm{ps}$
- Any electron, generated by the MIP, with energy significantly higher than E_{par} will ionise the crystal further, releasing more electron hole pairs. This kind of secondary ionising particles are typically called δ electrons. The electron maximal lattice velocity in silicon is $1 \times 10^7\,\mathrm{cm/s}$ - leading to a total traversal time of $3\,\mathrm{ns}$.

Assuming that the δ electrons provide a significant amount of the total charge created, a minimal estimate of the generated current pulse is

$$I_{mip} = \frac{2qN_{pairs}}{t} = \frac{2 \cdot 3.2 \times 10^4 \cdot 1.62 \times 10^{-19}\,\mathrm{C}}{3 \times 10^{-9}\,\mathrm{s}} = 3.2\mu\mathrm{A} \tag{5.15}$$

In detector grade silicon, meaning high-resistivity and ultra pure, the generation lifetime is typically in the $5 - 30\,\mathrm{ms}$ range, using this we can estimate the steady state bulk leakage current as

$$I_{leak} = q\frac{n_i}{\tau_g}(d - d_{V_0}) \cdot A \simeq \frac{1.62 \times 10^{-19}\,\mathrm{C} \cdot 1.45 \times 10^{10}\,\mathrm{cm}^{-3} \cdot 3 \times 10^{-2}\,\mathrm{cm} \cdot 1\,\mathrm{cm}^2}{20 \times 10^{-3}\,\mathrm{s}} = 3.5\mathrm{nA} \tag{5.16}$$

This corresponds to a very impressive Signal-to-Noise-Ratio (SNR) of 1000. However, in a realistic silicon sensor read-out system, the leakage current is not actually the dominating source of noise, and as such a SNR of $20 - 100$ is a more accurate expectation of a typical system.

It is not good practice to directly compare the transient signal pulse with the constant leakage current, since they behave inherently different. The purpose of this calculation was simply to show that, under normal circumstances, it would be very unlikely to mistake a random fluctuation in the leakage current as a signal pulse.

5.2.1 Radiation Tolerance

One of primary differences between the current silicon sensors used in the ATLAS ID and the new sensors to be used in the ITk, is that the ID sensors are based on p-on-n technology, meaning that the bulk is n-doped while the thin strips(pixels) are $p+$-doped - while the ITk will deploy n-on-p sensors. One of the primary types of radiation damage, called NIEL, is when traversing particles collide with lattice atoms and dislocate these from their place in the crystal allowing them to diffuse away. The voids left behind in the lattice will be perceived by its surroundings as immobile negative space charges, due to the lack of the positively charged nuclei, effectively increasing the acceptor dopant concentration of the material. This means that if your sensor is based on p+ strips and a n-doped bulk, this pn junction will disappear with the accumulation of radiation damage, simply because the n-doped bulk is slowly being turned into a p-doped bulk, and with the pn-junction deteriorating, the free charge carrier concentrations will increase - degrading the signal-to-noise-ratio. However, because the backplane is n+ doped, to ensure ohmic contact with the Al coating, a pn-junction is formed between the n+ backplane and the p-bulk - after the onset of this so called type inversion. This means

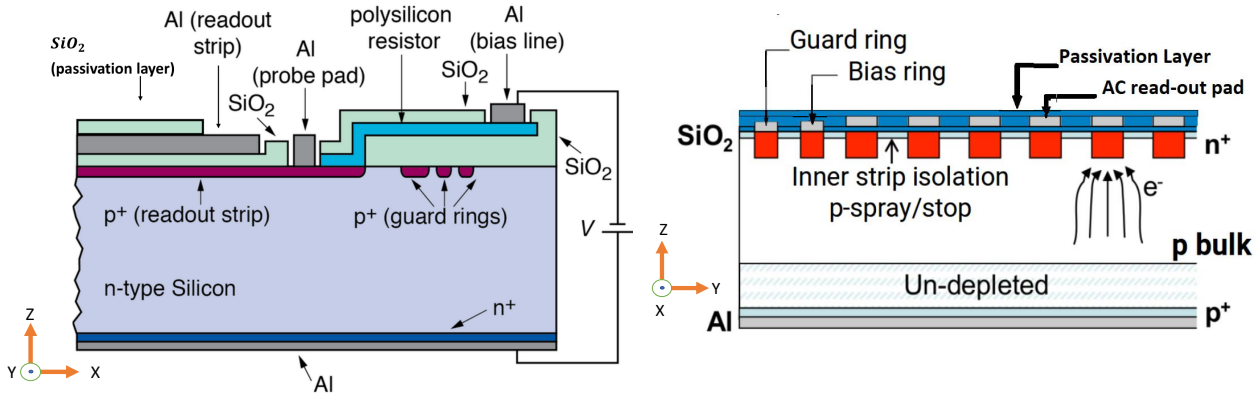
that our sensor goes from having N independent read-out channels, N being the number of strips, to having one independent channel - effectively reducing the spatial resolution by a factor $\sim 10^6$ [1].

This type inversion effect can be avoided by utilising n-on-p base sensors instead, the $n+$ doped strips(pixels) have a very small volume compared to the bulk, along with a much higher doping concentration, increasing the time scale of type inversion to such a degree that the n-on-p sensors are practically immune to the problems of type inversion.

The ATLAS ID didn't use n-on-p sensors because the technology wasn't economically and practically feasible at the mass production scale required, back twenty years ago when the original detector was being designed. As a quick side note, one of the many benefits of working with silicon sensors is that we can take advantage of the developments in industrial fabrication techniques designed for producing the Integrated Circuit Chips found en-masse in all computers.

5.2.2 ATLAS12EC case Study

Now that a overview of the theory behind semiconductor radiation detectors have been established, we will look at a specific example of how this is carried out in real life, namely the ATLAS12EC sensor, a prototype of the microstrip sensors to be used in the end-cap parts of ITk. As seen on Figure 5.6a, the microstrip design entails splitting the surface area of the sensor up into many individual pn-junctions, electrically isolated from each other, with the strip dimensions being of order $\mu\mathrm{m} \times \mathrm{cm}$, a compromise between spatial resolution, occupancy ratio and cost. When an ionising particle traverses then sensor, the generated electrons will drift to the nearest n+ doped strips, where the MOS, (Metal-Oxide-Semiconductor) structure acts similarly to a parallel plate capacitor, and the sudden build-up of charge on one plate leads to a voltage pulse being induced across the capacitor. This voltage pulse is what the pick-up electrodes then transfer into the read-out circuit of the ABC's - the front-end ASIC's.



(a) The figure shows a p-on-n sensor, but the structural principles are the same for the n-on-p devices used in the Itk. The structure continues outside the left edge of the drawing. [27].

(b) The left corner of the sensor is shown, many more strips and a symmetric edge-termination structure follow outside the right edge of the drawing. This is a cross section in the middle of the sensor, where the top passivation layer covers everything.

Figure 5.6: Two different cross sections of microstrip sensor, rotated $90\,\mathrm{deg}$ w.r.t. to each other and showing the common features of such a device - the dimensions of the features w.r.t. each other are not to scale. The defined coordinate system shows how the two drawings are rotated w.r.t. each other.

[1] This number comes from a ratio of the typical spatial resolution for a functional microstrip sensor vs. the total active area of the sensor

Following below is a short summary of the different primary components of a microstrip sensor, like the ATLAS12EC, and their functionality - for the visual counterparts, the reader is referred to Figure 5.6.

Base structure - The sensor has a p doped bulk and $n+$ doped strips, with a implant width of $16\,\mu\mathrm{m}$ and aluminium read-out strips of width $22\,\mu\mathrm{m}$. The metal layer is slightly wider than the implants in an attempt to minimise high E-field values in the transition region - which could otherwise be the seed for a breakdown of the leakage current. The distance between neighbouring strips, the pitch, is $75\,\mu\mathrm{m}$, and is maybe the most important parameter in determining the spatial resolution of the sensor. As a quick side note, There's no direct benefit, w.r.t. to spatial resolution, in reducing the pitch below $50\,\mu\mathrm{m}$, due to the spatial extent of the charge cloud generated by the traversing ionising particle [29] [9].

p stops - Positive charges in the SiO_2 atop the silicon will attract electrons to the SiO_2-Si interface in a homogenous manner. This interface layer of electrons would end up connecting the $n+$ strips to each other, essentially forming one big $n+$ layer across the device, counteracting the intent of having individual strips. To avoid this, a small p implant, $6-8\mu\mathrm{m}$ wide, called the p-stop, is deposited between each strip implant, such that free electrons, attracted to the surface interface in the regions in-between strips, will recombine with holes from the p-stop, rather than short circuiting the strips. The inter-strip resistance is required to uphold $R_{interstrip} > 10 \cdot R_{bias}$ at $300\,\mathrm{V}$ - R_{bias} being the resistance the polysilicon biasing resistors.

Bias ring - All the strips are connected in parallel, through a polysilicon resistor, to the bias ring, which provides a common ground for each channel of the sensor and the read-out electronics. There are several slots in the passivation layer along the bias ring, to establish electrical connection for eg. a probing needle or the read-out electronics.

Guard ring - The guard ring serves to electrically isolate the active area of the sensor from the edges of the devices - since the crystal defects at the cutting edges lead to very high leakage currents. This is handled by having a DC coupled electrode connected to a n+ implant encircling the area, with the electrode running straight into the common ground. This way, any current generated at the edges will find the path of least resistance being to flow through the guard ring and into the ground, instead of going through one of the actual strips.

Polysilicon resistor - To avoid having the generated signal charge flow into the biasing ring, and to provide a drain for the steady state leakage current, each strip is connected to the bias ring through a $1.5\,\mathrm{M}\Omega$ resistor, realised by the deposition of a silicon channel through the oxide with a scrambled crystal structure [29]. This randomised crystal structure of the polysilicon results in an increased electrical resistance, compared to a single crystal structure.

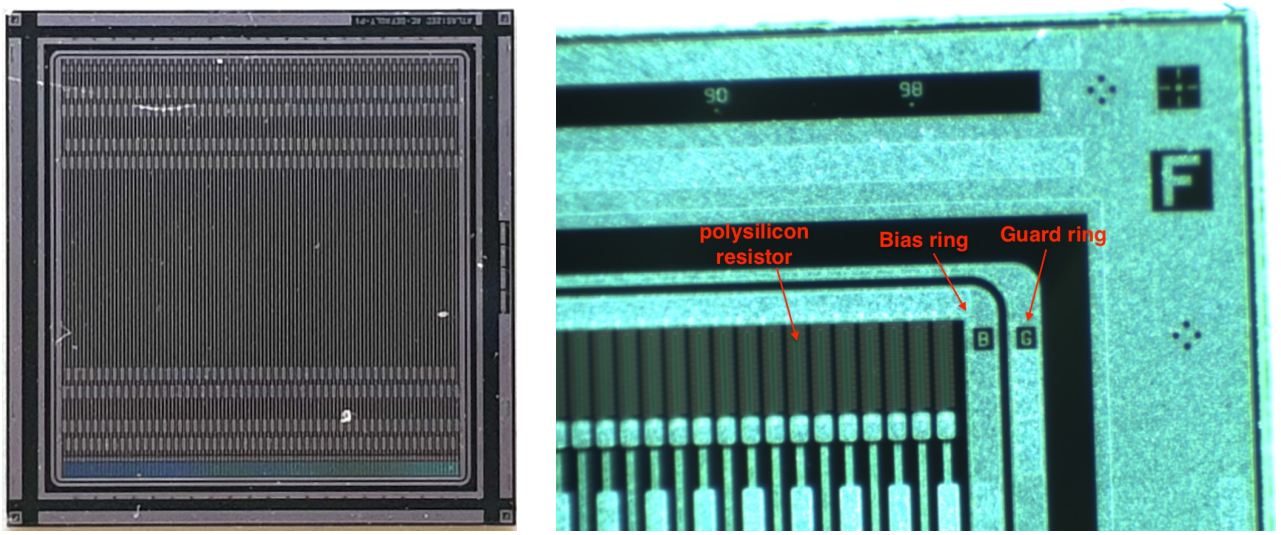
AC coupled bonding pads - By using a MOS structure for the read-out of strips with an oxide layer between the strip implant the read-out metal, we get a degree of high-pass filtering, minimising the noise comming from the constant leakage current, while still being maximally sensitive to the transient signals generated by passing charges.

PTP structure - We employ a binary read-out chain, and as such, don't really care about the amplitude of signals, as long as they're above the threshold value. Furthermore, if the generated current pulse is too big, eg. in the case of pinhole damage to the coupling oxide layer, it could damage the

read-out electronics. To avoid this, we include a Punch-Through-Protection structure in the strip read-out circuitry. The PTP structure is a type of MOSFET, and it acts as an overflow protection valve. If the voltage difference across the strip exceeds some upper threshold, the PTP opens up a direct channel to the biasing ring, circumventing the polysilicon resistor, and allowing charge to be drained away into the common ground through the biasing ring.

Backplane - The two terminals of the sensor biasing circuit are the bias ring and the aluminised backplane. To ensure a good ohmic connection across the silicon aluminium interface, the end of silicon bulk is highly p doped - typically above $1 \times 10^{19}\,\mathrm{cm}^{-3}$. At this very high concentration of free charges, the semiconductor locally approaches the band structure of a metallic conductor - and the characteristic resistance across the Si-Al junction goes towards zero.

To conclude this brief overview of the functionality found in the ITk strip sensors, we refer to Figure 5.7 - showing how these devices look in reality. In the next section, we will dive into a discussion of the misbehaving sensors encountered during the prototyping efforts of module assembly in the Scandinavian ITk cluster.



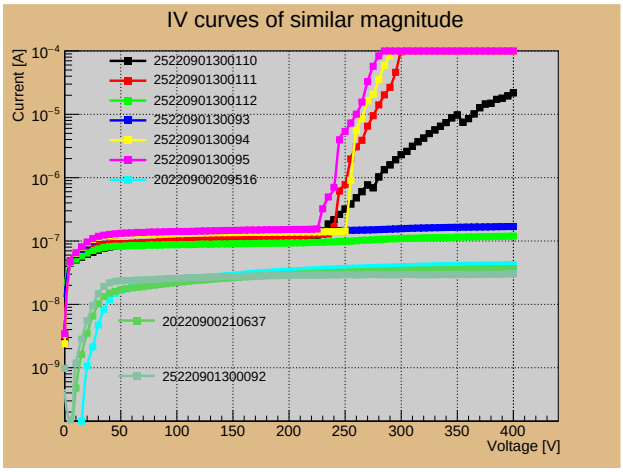
(a) Picture of an ATLAS12EC type mini sensor of size $1 \times 1\,\mathrm{cm}^2$ - an ITk prototype. The "dirt" in the photo is on the lens of the microscope, not the sensor.

(b) Zoomed picture of a ATLAS07 mini sensor, the type currently used in the SCT.

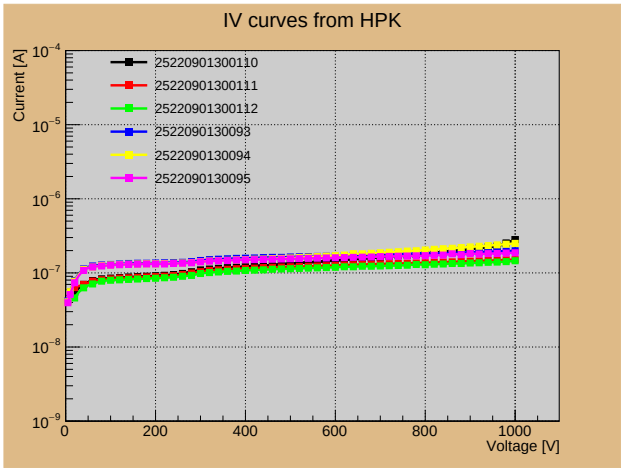
Figure 5.7: Pictures showing how the features of a strip sensor look in reality. The strip width is $\sim$20 µm and the distance between adjacent strip centers, the pitch, is respectively 75 and 80 µm in the left and right hand side picture.

5.3 Early-Onset Behaviour

Upon reception of new sensors, module assembly sites are required to perform a reception test, meaning a visual inspection and a IV curve, to evaluate if the sensor was damaged during shipping. The sensors are intended to start operation at a bias voltage of 500 V, going up 600 V after 10 years of usage due to radiation damage - while the supplier, Hamamatsu Photonics (HPK) tests them up to 1000 V. The operational requirements are a leakage current density below $100\,\mathrm{nA/cm}^2$ and a micro-discharge onset voltage $V_{MD} > 600\mathrm{V}$. A micro-discharge is when the leakage current diverges, below



(a) IV curves for some of the R0 sensors present in Uppsala. Measurements performed at $22-23°C$ with a current limit set at $100\,\mu A$, the humidity was not recorded.



(b) IV curves done by HPK, measurement done at $25\,°C$.

Figure 5.8: Showing the difference in reception IV curves in Uppsala, and a corresponding measurement performed by the suppler right before shipping them out.

the voltages required for an actual breakdown and they are caused by physical imperfections of the sensor. While real breakdowns are consistent in their onset voltage, and signify the hard physical limits of the system, micro-discharges are erratic phenomena which can be cured, partially or fully, if one can identify and alleviate the cause(s).

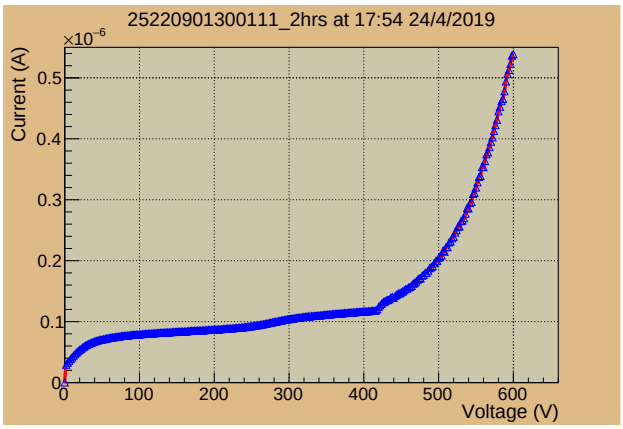
In Figure 5.8, a set of reception IV curves, done in Uppsala, are shown and compared to the corresponding measurement done by the supplier right before the sensors were shipped out. The measured leakage currents varies with temperature according the charge carrier densities, and often when comparing IV curves done under different environmental circumstances one normalises the leakage current to a common reference temperature. The Uppsala lab temperature is stabilised at $22-23°C$ through central climate control and the HPK curves were done at $25°C$ - so it seemed unnecessary to perform a temperature normalisation. We see that four of the sensors which, in the HPK test had a decent leakage current up to $1kV$, have since then developed an early-onset feature. This can be explained as being due to a change, eg chemical or mechanical, of the sensor, differences in the measurement setup impacting the results or a combination thereof. It is also known that sensor exposure to high levels of humidity, both during storage and measurement, can lead to early micro-discharging onset of the leakage current, something we'll return to later. As such, for the sake of comparability, an IV curve should always contain information about the environmental conditions during measurement. However, humidity was not monitored at the time of measurement for the IV curves seen in Figure 5.8a. However, these IV curves were measured during winter time in Uppsala, were the ambient humidity is consistently low, due to the cold weather.

As a side note, it is quite peculiar how the batch of sensors with names ending in 92-94 have two well behaved sensors, the "92" and "93", but with a factor five difference in their plateau leakage current.

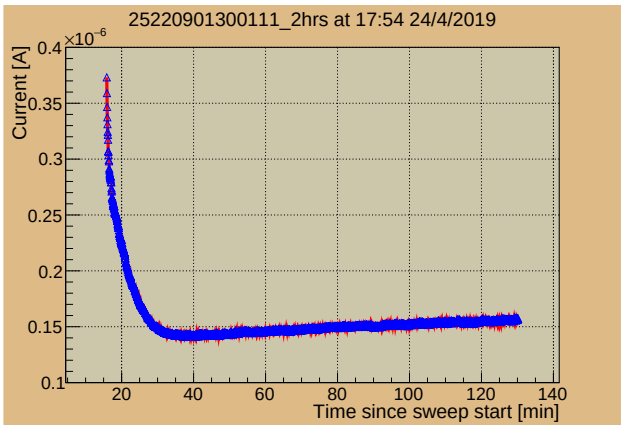
We decided to initially limit our investigation of this early-onset behaviour to one sensor, to limit the risks of mechanical damage due to excessive handling. We chose, for no particular reason, the sensor named 25220901300111, from now on referred to as "111".

The first thing was performing some 15 IV curves while varying the voltage step size and the equalising time between ramp-ups in voltage, to see if we were simply using unfortunate settings in our IV procedure setup. While we didn't observe any dramatic improvements of the leakage current

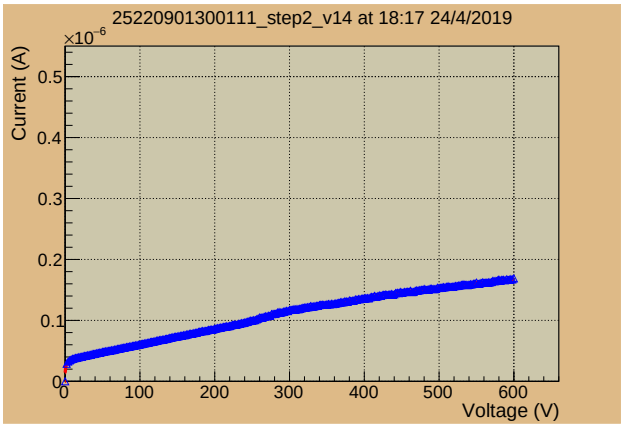
with the different sets of settings attempted, we did see a slow but consistent increase in the onset voltage of the micro-discharge, seemingly only dependent on the amount of IV curves performed. Based on this observation we hypothesised that training the sensor might alleviate the early-onset behaviour. Training means ramping up the voltage as normal, and then leaving the sensor at the highest biasing voltage for an extended period of time, of order several hours, while continuously monitoring the leakage current of the device. A seemingly positive test of this hypothesis can be seen on Figure



(a) IV curve at the beginning of training



(b) current vs time training curve



(c) IV curve directly after training finished.

Figure 5.9: The evolution of the 111 sensor's leakage current, before, during and after training.

5.9. It is not very well understood, from a theoretical point of view, how this training works - due to the complexity of the system, but one can make educated guesses as to how this annealing like effect works. One example could be the case of crystal defects acting a trapping centres, causing a local build-up of charge which leads to a region of very high electrical field, capable of seeding a micro-discharge. In this case, the extended high voltage biasing could potentially act as to slowly drain away the build-up of free charges, or "refill" these trapping centres such that they revert back to a charge neutral state.

Three weeks later, a new set IV curve was done, see Figure 5.10, to investigate if the early-onset had returned, and the results were a bit unexpected. The onset voltage was down to $V_{MD} \approxeq 150\text{V}$, lower than anything seen before, but it improved to $V_{MD} \approxeq 280\text{V}$ after a single subsequent measurement, where previously we saw V_{MD} typically increase by a few volts between consecutive measurements. At this point it would have been very interesting to repeat the training, both on the 111 sensor and the sensors suffering from early-onsets in their IV curves - trying to establish a decay constant for the positive benefits of the training phenomena. This did not happen due two reasons, an equipment mal-functioning and/or bug in the control software leading to the current vs time measurement being

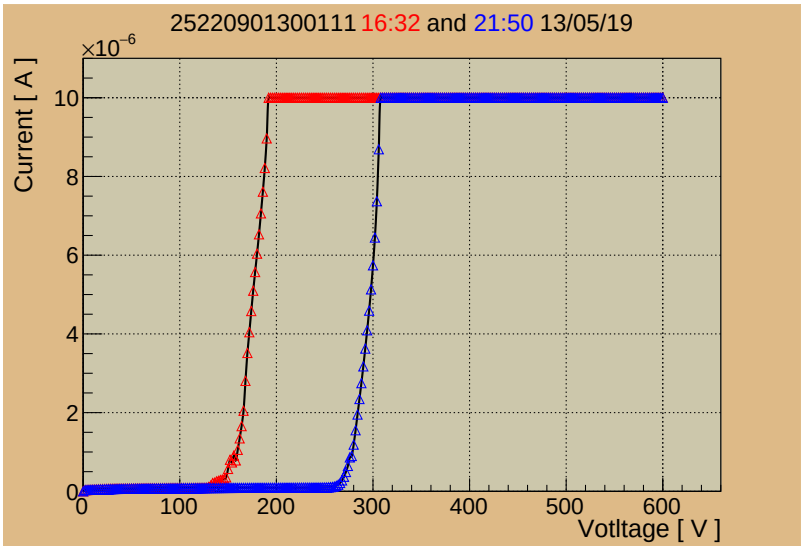


Figure 5.10: IV curves 3 weeks after initial training, the early-onset is back, at lower voltages than ever before, but also seems to improve much more rapidly than previously seen.

unavailable, combined with insufficient time to fix this due to a need for me to shift my focus from sensor studies and back to glue robot development. As such I didn't have time to properly continue these studies before leaving Uppsala.

However, $4.5\,\mathrm{months}$ after these training studies were conducted, the 111 sensor was measured again, before being used in the assembly of the first electrical R0 module produced in the Scandinavian Cluster. During this time period, the sensor was stored in $\lesssim 5\,\%\mathrm{RH}$ dry storage

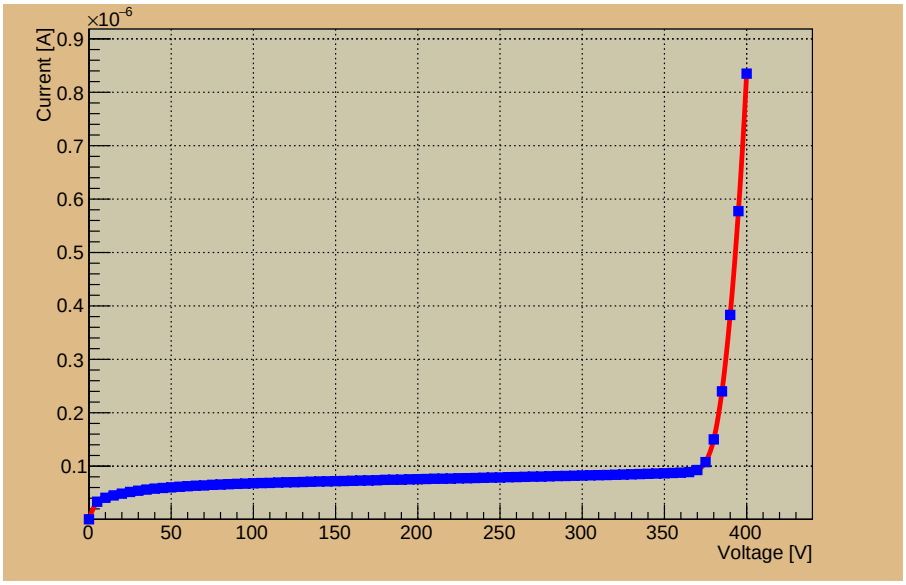


Figure 5.11: IV curve of the 111 sensor prior to module assembly. $4.5\,\mathrm{months}$ of dry storage at $\lesssim 5\,\%\mathrm{RH}$, has passed between this IV curve and the prior one seen on Figure 5.10. The onset voltage has improved significantly, and was even $\sim 150\,\mathrm{V}$ better than at the reception test shown on Figure 5.8a.

5.3.1 NRA Studies of Humidity Sensitivity

Going into the fall of 2019, many of the ITk institutes working with sensors had reported observations on the seemingly adverse affects of humidity on the sensor leakage current - primarily manifesting itself in the form of early-onset behaviour similar to what we see on Figure 5.8a. Studies have shown that both long term exposure during storage, as well as too high levels of humidity during the IV measurement can lead to early onsets of micro-discharge. For the long term sensitivity, this was investigated in the typical manner of storing sensors in respectively wet and dry conditions while performing IV curves in between - with the environmental conditions tightly controlled during the IV measurements. As a side note, when doing many repeated IV curves at different levels of humidity, one has to be careful with not confounding the humidity dependence and the training effects seen on Figure 5.9.

However, while the adverse affects of high humidity are well known, the physical mechanisms responsible for this sensitivity are not very well understood within the ITk collaboration. In the case of high humidity during measurement, we believe that micro-condensation might be the sinner, creating a conductive thin-film across the surface. However, the cause(s) of the long term humidity sensitivity remain an open issue within the collaboration.

In an attempt to shed some light on this, we utilised the Nuclear Reaction Analysis (NRA) facility available at the Uppsala University Tandem Accelerator complex[30]. NRA can be used to measure the depth profile of hydrogen concentration in solid samples, eg a mini silicon strip sensor, and we assume that a presence of hydrogen in our sensors would be primarily due to the absorption of water into the structure.

When performing NRA measurements, we exploit the nuclear interaction between ^{15}N ions and H atoms, having a large and narrow resonance at $E_{res} = 6.385\,\mathrm{MeV}$ of width $\sigma_{res} = 1.8\,\mathrm{keV}$, resulting in the emission of photons at $E_\gamma = 4.43\,\mathrm{MeV}$. By subjecting the sample to a ^{15}N ion-beam with a beam energy gradually increasing from a starting point at the resonance, $E_{beam} \geq E_{res}$, the energy loss per penetration depth will ensure that resonance occurs gradually deeper inside the sample, thereby allowing us to probe the hydrogen concentration as a function of depth into the sample. The hydrogen concentration in terms of atomic fraction, $\rho(E_{beam})$, is evaluated from the ratio between the integrated beam current N and the generated photon current Y- seeing as the number of interactions increases with the density of hydrogen in the sample. Furthermore, a hydrogen-implanted Si reference sample, having a concentration of $\rho_{ref} = 18.5\ \%$, is used to calibrate the depth profile of hydrogen in an arbitrary sample as follows

$$\rho(E_{beam}) = \frac{Y(E_{beam})}{N} \cdot C_{ref} \tag{5.17}$$

Here C_{ref} is a material constant transforming the normalised signal into atomic fraction of hydrogen. The value of C_{ref} was provided by the material scientists operating the NRA facility, and is calculated based on the material properties of respectively the H-implanted Si reference sample and the SiO_2 layer we're probing in this measurement. This is of course assuming that we are not probing deeper into the sensor structure than the top passivation layer - see Figure 5.6. If we do probe deeper into the sensor, one would expect to see a mixed volume of aluminium along the strips and SiO_2 in between - something we would need to take into account when calculating the hydrogen profile. The aluminium parts of the sensor are assumed to not contain any hydrogen, such that a higher volume of aluminium means a direct decrease in the measured signal. This means, that when the beam is penetrating a mixed material volume of SiO_2 and Al, one needs to correct for the artificial decrease in signal, due Aluminium coverage, to extract the true hydrogen concentration in the oxide parts of the sample. This

correction is done as

$$\rho_{cor} = \frac{\rho}{\frac{A_{ox}}{A_{tot}}} = \frac{\rho}{1 - \frac{A_{Al}}{A_{tot}}} \approxeq \frac{\rho}{0.7} \tag{5.18}$$

where A_{ox} is the area covered by oxide, A_{al} the area covered by aluminium and $A_{tot} = A_{ox} + A_{al}$ the total area of the sample. This estimate is only valid where the beam hit, in the middle of the mini sensors, where there are only strips and no other structural features.

To convert the employed beam energy into a measurement of depth into the sample, we perform the following calculation

$$z(E_{beam}) = \frac{E_{beam} - E_{res}}{S}. \tag{5.19}$$

Where S denotes the energy loss per length, or the stopping power of the material. It should be mentioned that S actually depends on the energy scale of the beam, but the change across the working range of our experiment is so small that we can reasonably approximate S as a constant.

5.3.1.1 Results

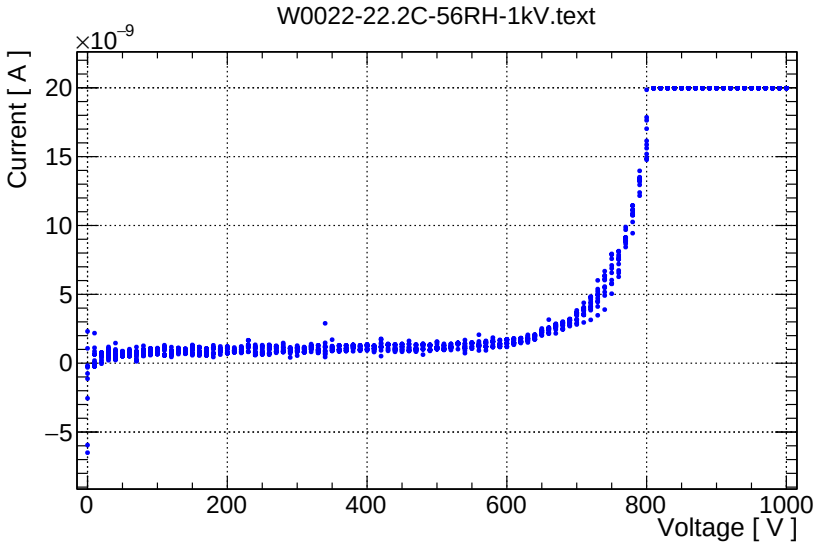


Figure 5.12: IV curve of the ATLAS12EC mini sensor named VPX22728-W022 - before the NRA study. That it is similar to the full sized sensors, can be seen from its serial code ”2522 0901 300 022”, where the first 11 digits are identical to that of the 111 full sized sensor we’ve studied.

Of four available mini sensors, from the same batch type as the 111 full size sensor (ATLAS12EC), only one of them, named ”VPX22728-W022” shortened to ”W022”, had a current breakdown, shown on Figure 5.12. This breakdown actually occurs above the sensor specification limit of $V_{MD} > 600$V, but it still remains the best test candidate for this study. It is believed that mini’s are much less prone to surface related breakdowns than the full sized sensors, simply because of their much smaller surface area reducing the probability for bad things to happen. It would of course have been nice to use all of the available mini’s for the NRA study, but this was not possible due to time constraints. The operators of the NRA facility did measure one other mini sensor, a DC-coupled ATLAS12A mini named ”VPX12318-W621”, as an initial test of the technique’s feasibility. This allowed us to study the effect of month long storage in a dry cabinet, $RH \sim 5\%$, and to have a control sample for the annealing procedure.

The results from the NRA study can be seen on Figures 5.13a and 5.13b, with the annealing procedure described in Table 5.1. We clearly see a consistent presence of hydrogen basically throughout the entire oxide layer. While we see the hydrogen content in W022 decrease to half its original value after annealing, we are sadly not able to measure if this has a beneficial effect on the leakage current. The NRA measurement turned out to be destructive, in that the beam implanted the sensor with nitrogen ions, ruining the microstrip doping profile and leading to six orders of magnitude increase in the current passing through the sensor - due to the nullification of the depletion mechanism. We feared that this might be the case, prior to performing the experiment, but did not know for certain until after performing the post-NRA IV curves.

An unintended but very interesting feature of the data seen on Figures 5.13a and 5.13b, is the very sudden and sharp drop-off in hydrogen concentration at 2100 $arb\ units$ into the sensor. After cross referencing with microscopy observations, we believe this drop-off happens at the passivation to metal strip interface, where the material volume changes from pure SiO_2 to a mix of Al strips and SiO_2 in-between the strips. This means that our humidity study has incidentally revealed the thickness of the passivation layer, something that is a trade-secret for the manufacturer of the sensors, which ATLAS is contractually bound to not reveal. This is why all the measured hydrogen depth profiles show here are given in terms of arbitrary units of depths, instead of eg. $nanometers$ - such that we can show the results without revealing any confidential information. Also the found thickness of the passivation layer is consistent with what other institutes have found, in dedicated measurements of this structure.

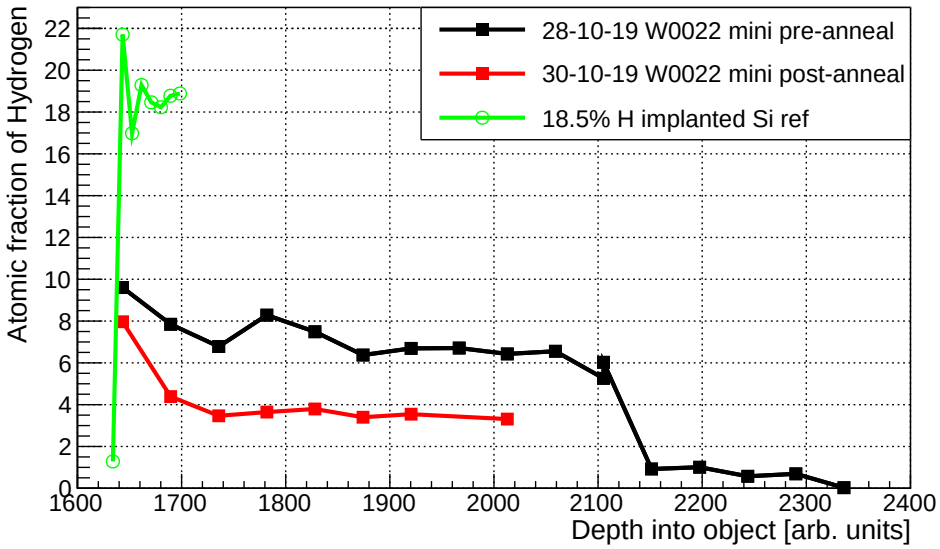
For the first sensor studied using NRA, VPX12318-W621, we don't see the same sudden drop off in hydrogen concentration, see Figure 5.14a. It is a DC coupled mini, meaning that there is a direct aluminium to silicon interface, with no oxide layer in between - unlike the AC coupled W022 mini. As a side note, the coupling oxide layer is typically significantly thinner than the passivation oxide layer. This lack of a sudden drop off seems to indicate that W621 has no passivation layer, however we still expect it to have oxide in between the aluminium strips. This means that we are potentially probing a mixed metal/oxide volume from the very surface and down - in this case the hydrogen concentration correction from Equation 5.18 should be applied across the entire depth profile, which is done in Figure 5.14b.

It would have been very interesting to try and take these measured hydrogen concentration profiles and implement them in TCAD simulations of a sensor-like structure - to investigate how one would expect this contamination to affect the electrical performance of the device. While time didn't permit us to do these kind of studies, we can still present our working theory on the subject. We believe that the hydrogen acts as an excess of positive charge in the oxide layers, causing an accumulation layer of electrons at the SiO_2-Si interface. This accumulation layer degrades the p-type edge isolation, allowing the large edge currents, caused by crystal defects, to more easily flow into the, otherwise isolated, pn-junction circuit.

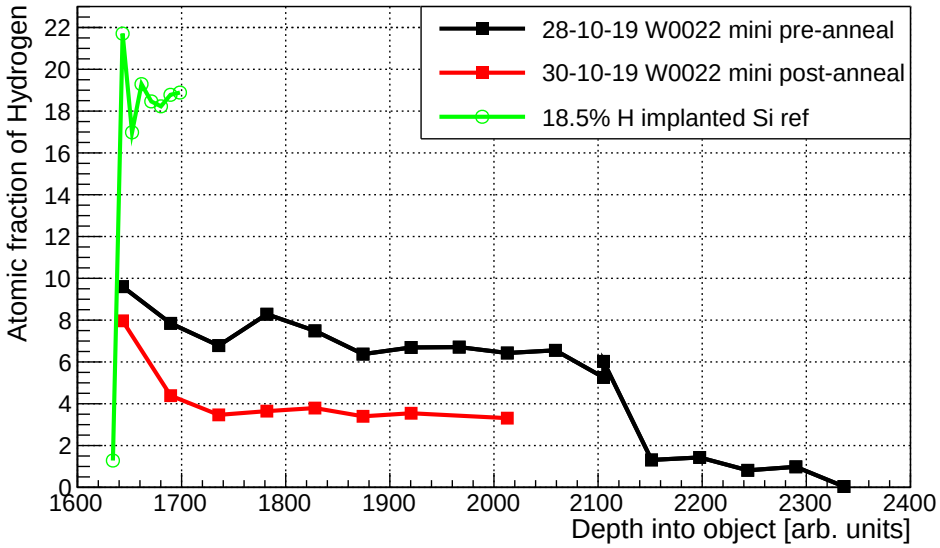
While we weren't able to test this hypothesis, it is reported in literature that the accumulation of hydrogen at the SiO_2-Si interface of MOS devices can cause electrical degradation[32, Sec. 3.3].

time period	storage conditions
Time immemorial to 28-10-19	Ambient conditions
28-10-19	Baked for 3 hours in vacuum chamber, temperature slowly ramped to 200°C across the entire time period.
28-10-19 to 30-10-19	Stored in vacuum chamber

Table 5.1: Storage record of VPX22728-W022 an ATLAS12EC AC-coupled mini sensor. "Time immemorial" refers to the storage conditions prior to the start of this experiment.



(a) No aluminium correction used in this version of the plot.

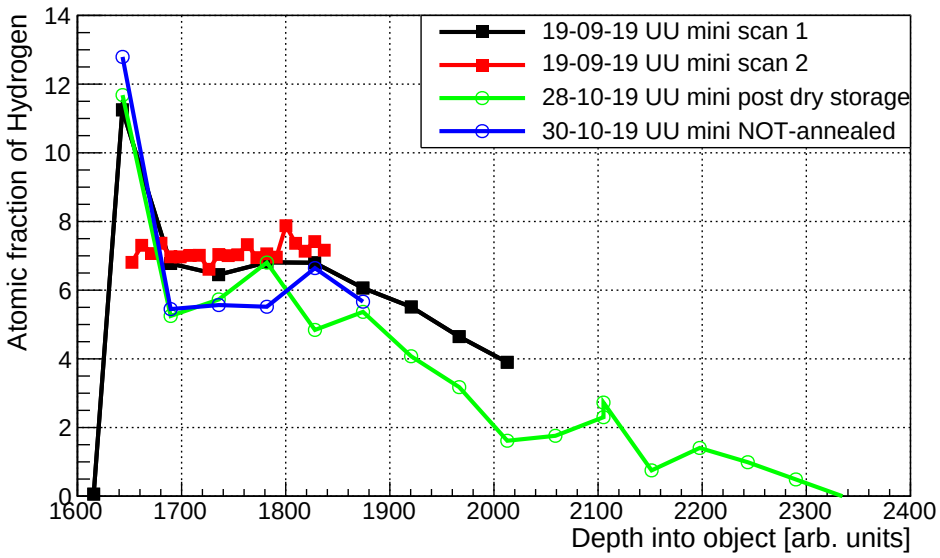


(b) Aluminium correction applied - scaling the Y data by a factor of $1/0.7$ for $x > 2130$, with 0.7 being the fractional surface coverage of the oxide.

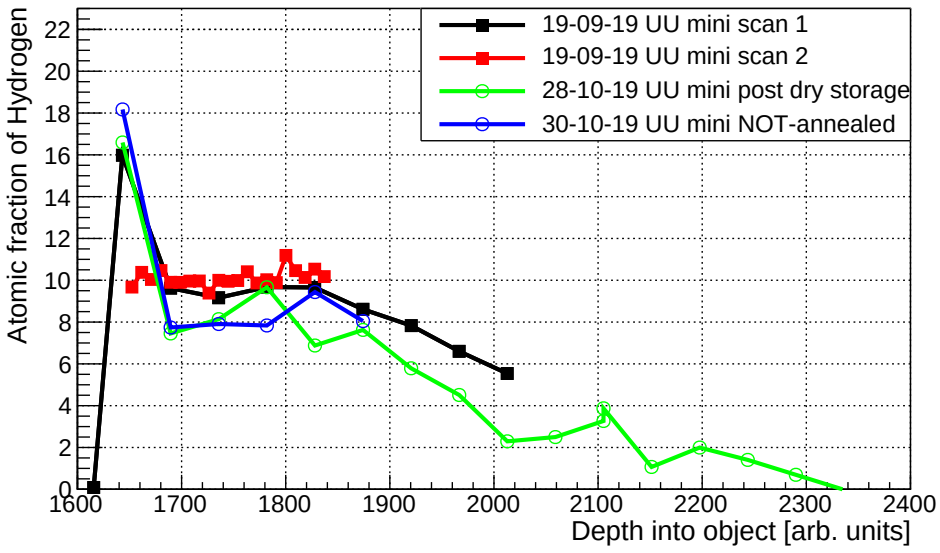
Figure 5.13: Results from NRA investigation showing the depth profile of hydrogen concentration in VPX22728-W022, an ATLAS12EC AC-coupled mini sensor stored in ambient conditions prior to the study. The depth profile is shown with and without the aluminium coverage correction applied, in case the assumption of "no hydrogen in the aluminium volume" is later proven wrong.

Time period	Storage conditions
Time immemorial to ∼9-9-19	Dry cabinet storage $RH < 5\%$
∼9-9-19 to 19-9-19	Ambient conditions
19-9-19 to 28-10-19	Dry cabinet storage $RH < 5\%$
28-10-19 to 30-10-19	Stored in vacuum chamber but not annealed

Table 5.2: Storage record for the VPX12318-W621, a DC coupled ATAS12A mini, used for the initial test-of-feasibility and as a control sample w.r.t. the effects of annealing. ”Time immemorial” refers to the storage conditions prior to the start of this experiment.



(a) No aluminium correction used in this version of the plot.



(b) Aluminium correction applied - scaling all of the Y data by a factor of $1/0.7$, with 0.7 being the fractional surface coverage of the oxide.

Figure 5.14: Results from NRA investigation of VPX12318-W621, a DC coupled ATAS12A mini. The two initial scans give a sense of the precision achievable. We see a small but definite decrease in hydrogen concentration after drying out the sensor for a month - showing just how slowly the device dries out. This sensor has a different design than the ATLAS12EC type full sized sensors otherwise investigated. It was used for the very first feasibility study, due to being the only mini sensor available in Uppsala at the time. The depth profile is shown with and without the aluminium coverage correction applied, in case the assumption of ”no hydrogen in the aluminium volume” is later proven wrong.

Summary and Conclusion

The LHC will be upgraded to the High Luminosity LHC, increasing the data gathering rate by a factor $5 - 7$, allowing us to further probe the mysteries of the universe. The current ATLAS Inner Detector needs to be replaced, due to accumulated radiation damage of the current detector, and its inability to cope with the much higher pile-up conditions of HL-LHC operation. The new Inner Tracker (ITk) will be based on a modular design, splitting the tracker up into ~ 19.000 individual silicon based detector modules. Due to the extreme performance requirements, needed to properly operate a detector in the HL-LHC environment, along with the overall size of the detector, the successful (mass)production of detector modules becomes a highly non-trivial effort.

The Scandinavian Cluster, consisting of Copenhagen, Lund, Oslo and Uppsala University, is one of many assembly sites, and will both be performing hybrid and module assembly - responsible for producing ~ 10 % of the total module amount needed for the microstrip end-cap part of the ITk.

This thesis has been devoted to preparing the Scandinavian Cluster for production, working on the implementation of existing procedures for assembly and quality control - along with the development of alternative approaches, better suited to our unique situation of collaborating with an industry partner, carrying out significant parts of our production flow. I have been involved in almost every step of the production flow, from bare components to finished detector module, but my main contributions lie in the hybrid-to-sensor assembly program. Due to our collaboration with industry, we wanted to replace the manual baseline procedure for the hybrid-to-sensor glue attachment, with a more automated approach - better suited for the conditions of production flow in an industrial setting. As such, we set out to develop a glue robot, which could dispense the two-component epoxy glue with sufficient precision in the dispensed mass and placement of glue, such as to comply with the technical requirements of hybrid-to-sensor assembly. While the placement of glue was easily controlled using an XY-table with high precision stepper motors, the calibration of dispensed mass turned out to be a highly non-trivial problem. The speed of the XY-table was our only way of controlling the amount of glue being dispensed, and this had to calibrated w.r.t. three different codependent input variables. The time elapsed since mixing the two-component epoxy, the target mass and the length over which to deposit this mass. It was not possible to do an independent analytical parametrisation of each variable. Instead we built a data-driven look-up table of speed as a function of time and mass, for fixed path length of $8\,\mathrm{mm}$, allowing us to build larger glue patterns by stacking $8\,\mathrm{mm}$ lines together.

The successful development of this home-made glue robot, allowed us, in the fall of 2019, to produce the first electrical R0 module in the Scandinavian Cluster - and it is currently foreseen that this robot will be used by our industry partner during production.

Aside from working on module assembly, I've also been engaged in the quality control of components, both w.r.t. to electrical testing of the read-out electronics situated on the hybrids, and of the silicon sensors - being the primary component of the tracker. This led to the discovery of abnormal behaviour in the IV curves of ATLAS12EC sensors, suddenly suffering from much earlier onsets of micro-discharging, than what was found during quality control performed by the supplier. This is a very hot topic within the ITK sensor community, with the working hypothesis being that the early onset is related to humidity exposure - both during long term storage and during IV measurement. We've contributed to the knowledge base within the community by performing Nuclear Reaction Analysis (NRA) on several mini sensors. This allowed us to map the depth profile of hydrogen concentration down through the surface of the sensor, and we found an almost homogenous concentration of ~ 7 % hydrogen, by atomic fraction, throughout the entire top oxide passivation layer. If time had allowed, it would have been very interesting to further investigate, using TCAD simulation tools, how this hydrogen presence might affect the electrical performance of the sensor.

Bibliography

[1] CERN - Visited on 07-08-2019.
https://home.cern/about

[2] Mark Thomson, *Modern Particle Physics*, (Cambridge University Press 2016)

[3] Higgs production cross section - Visited on 21-11-2019.
https://twiki.cern.ch/twiki/bin/view/LHCPhysics/CERNYellowReportPageAt1314TeV2014

[4] Michaela Schaumann, CERN Summer Student Lecture 1 - 2019.
https://indico.cern.ch/event/817563/attachments/1872833/3090826/SummerStudentLectures_2019_IntroductionToAccelerators_Schaumann_L1.pdf

[5] Synchrotron radiation - Visited on 24-10-2019.
https://en.wikipedia.org/wiki/Synchrotron_radiation

[6] Introduction to Accelerator Optics.
https://uspas.fnal.gov/materials/11ODU/Into.pdf

[7] Daniel Schoerling, CERN Summer Student Lecture 1 2019.
https://indico.cern.ch/event/817556/attachments/1874695/3086450/Schoerling_SummerStudents_PARTI.pdf

[8] Andrzej Wolski, *Beam Dynamics in High Energy Particle Accelerators*, (Imperial College Press 2014)

[9] Peter Hansen, *Particle detectors and accelerators Lecture notes - second edition*, (Polyteknisk Boghandel & Forlag 2015)

[10] ATLAS ITk Collaboration, *Techical Design Report for the ATLAS Inner Tracker Strip Detecto* - CERN-LHCC-2017-005 ATLAS-TDR-025, CERN, April 1, 2017.

[11] Michaela Schaumann, CERN Summer Student Lecture 3 2019.
https://indico.cern.ch/event/817568/attachments/1865537/3094692/SummerStudentLectures_2019_IntroductionToAccelerators_Schaumann_L3.pdf

[12] HL-LHC schedule - Visited on 25-10-2019.
https://project-hl-lhc-industry.web.cern.ch/content/project-schedule

[13] HL-LHC list of upgrades - Visited on 03/11/2019.
https://home.cern/resources/faqs/high-luminosity-lhc

[14] ATLAS ID Community, *Inner Detector Technical Design Report Volume I* - CERN-LHCC-97-16 ATLAS-TDR-4. CERN, April 30, 1997.

[15] Peter Vankov, ATLAS Upgrade for the HL-LHC: meeting the challenges of a five-fold increase in collision rate - *arXiv:1201.5469*

[16] Steven Juhyung Lee, *Development and Evaluation of ATLAS Inner Tracker Strip Sensors and Strip Detector Modules*, (msc. thesis 2018)

[17] https://cds.cern.ch/record/2658150/files/ATL-ITK-PROC-2019-002.pdfMartin Sykora, *ITk Strip Module Design and Performance* - ATL-ITK-PROC-2019-002.

[18] Gerhard Lutz, *Semiconductor Radidation Detectors* , (Springer 1999)

[19] Epolite glue.
https://twiki.cern.ch/twiki/pub/Main/AtlasEdinburghGroupMaterialStudies/Fuller_Epolite_FH-5313_epoxy.pdf

[20] Märzhäuser-Wetzlar MCL 2 axis table documentation - Visited on 16-12-2019.
https://www.marzhauser.com/nc/en/service/downloads.html?tx_abdownloads_pi1%5Baction%5D=getviewcatalog&tx_abdownloads_pi1%5Bcategory_uid%5D=27&tx_abdownloads_pi1%5Bcid%5D=365&cHash=8ef1218985bb62762f2d8ce73fa0581a

[21] Susanne Kuehn, CERN, *The Upgrade of the Inner Tracker of the ATLAS experiment for the High-Luminosity LHC* - Visited on 04-12-2019.
https://indico.cern.ch/event/865308/attachments/1947167/3244336/DetSeminar-ATLASITk-SK-vF.pdf

[22] https://cds.cern.ch/record/2304805/files/ATL-ITK-PROC-2018-009.pdfFirst bulk and surface results for the ATLAS ITk Strip stereo annulus sensors - ATL-ITK-PROC-2018-009.

[23] NOTE, electronics company - Visited on 10-10-2019.
https://www.note.eu/en/

[24] DATACON 2200 evo machine - Visited on 20-12-2019.
https://www.besi.com/products-technology/product-details/product/datacon-2200-evo/#tabs-57

[25] R.J. Barlow, *Statistics - A Guide to the Use of Statistical Methods in the Physical Sciences*, (John Wiley & Sons Ltd. 1999)

[26] Weighted mean - Visited on 12-09-2019.
https://en.wikipedia.org/wiki/Weighted_arithmetic_mean

[27] Manfred Krammer, *Presentation on Silicon Detectors*.
https://indico.cern.ch/event/124392/contributions/1339904/attachments/74582/106976/IntroSilicon.pdf

[28] IV curve illustration - Visited 19-12-2019.
https://en.wikipedia.org/wiki/P%E2%80%93n_diode#/media/File:Nonideal_diode_current-voltage_behavior.png

[29] ATLAS Upgrade Strip Sensor Collaboration, *Technical Specifications for Supply of ATLAS12EC Silicon Microstrip Sensors. – Version 2.1*.
https://indico.cern.ch/event/448475/contributions/1113817/attachments/1174155/1696556/ATLAS12ECTechnicalSpecs..v2.1.pdf

[30] Uppsala University Tandem Laboratory - Visited on 07-12-2019.
https://www.tandemlab.uu.se/infrastructure/Accelerators/pelletron/t1/

[31] K. Komander, M.V. Moro, S.A. Droulias, J. Müggenburg, G.K. Pálsson, T. Nyberg, D. Primetzhofer, M. Wolff, *Hydrogen site location in ultrathin vanadium layers by N-15 nuclear reaction analysis* - Nuclear Inst. and Methods in Physics Research B 455 (2019) 57–60. https://doi.org/10.1016/j.nimb.2019.05.033

[32] Markus Wilde, Katsuyuki Fukutani, *Hydrogen detection near surfaces and shallow interfaces with resonant nuclear reaction analysis* http://dx.doi.org/10.1016/j.surfrep.2014.08.002