RD53B users guide: Introduction to RD53B pixel chip architecture, features and recommendations for use in pixel detector systems.
Version 1.1

RD53 collaboration, jorgen.christiansen@cern.ch as corresponding author

May 10, 2021

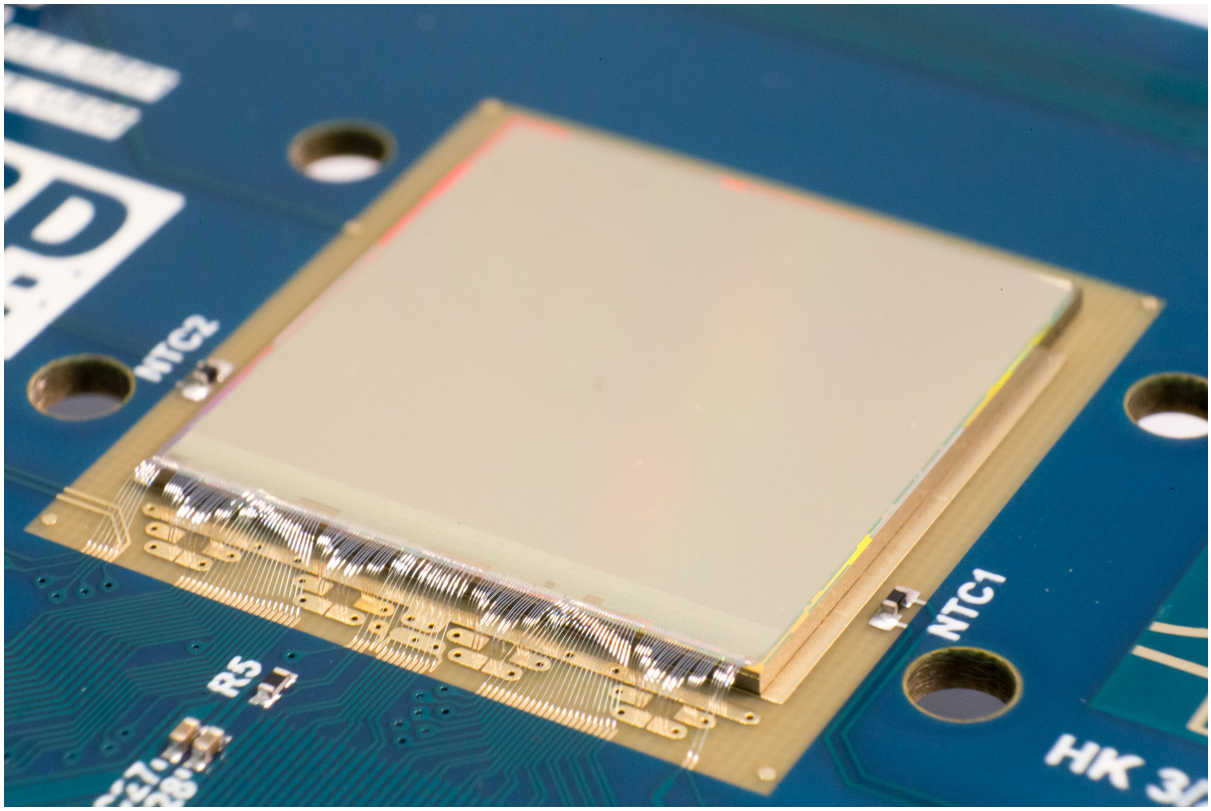


Figure 1: RD53B chip on single chip test card.

This document is a general introduction to the RD53B pixel chips covering its general architecture and features and it includes practical advice and recommendations on how to use the pixel chip on pixel modules as part of a pixel detector system.

It can be useful to read this introduction before:

- Reading detailed chip manuals [5], [7]
- Using the chip at system level,
- Making DAQ software/firmware, etc..

Contents

Chapter 1

Introduction and general requirements

This introduction and overview is not made to replace other sources of more detailed, and continuously updated, reference information about RD53 chips. It is intended as a introduction and tutorial explaining the basic functions and principles of the RD53 pixel chips together with recommendations and practical hints for their use in pixel detector systems. Sources of relevant information are:

- RD53B chip manuals [5], [7]
- RD53 pixel chip requirements [3]
- RD53 web pages [1]:
- RD53 presentations (public access) [2]:
- RD53B testing meetings, Testing TWIKI, test results overview [6].

An extensive list of currently most appropriate papers and presentations with web pointers are given in the reference/bibliography at the end of this document.

Access can be granted to pixel chip RTL source code (system Verilog) and dedicated simulation and verification framework (system Verilog and UVM) on GIT [11].

A Cocotb simulation environment, integrated with BDAQ53 firmware and software, in Python, can also be made available to check and debug BDAQ53 test routines against RD53 pixel chip RTL (Register Transfer Level) code, in a virtual test environment [34].

A normal user of RD53 pixel chips will see functions of the pixel chip, module or system through the specific data acquisition system used. Such a DAQ system will normally deal with all detailed operation issues of the chip (configuration, control and readout links, hit data decoding, monitoring, error detection and reporting, threshold tuning, etc.) via high level software commands to the DAQ system, that then takes care of running the pixel chip via appropriate dedicated FPGA firmware. Chip testing focussed DAQ systems have been developed within RD53, such as BDAQ53 [35] (gigabit ethernet and python based from Bonn) and YARR [36] (PCIe plug in card and C++ based, from LBNL). For experiment specific pixel detector systems the experiments will gradually use their specific DAQ made for large scale multi chip and module systems.

Different chip versions (RD53B-ATLAS and RD53B-CMS) have same basic interfaces and functions and will only have minor differences in specific functions (e.g. related to chip size and analogue front-end). A list of bug fixes and new features in RD53B-CMS relative to RD53B-ATLAS can be found in short form in chapter 17. Changes made to the RD53B-CMS chip will (unfortunately) require a few critical changes to test system firmware and software to work across both chips (e.g. change of begin-End stream bit and changes to configuration registers). The initial RD53A demonstrator chip [8] has the same basic architecture but is missing many of the features available in RD55B generation chips, but had many specific testing features.

1.1 RD53 collaboration and chips

The RD53 collaboration, with 24 institutes, has since 2014 been working on developing pixel chips for the CMS [37] and ATLAS [38] phase II upgrades with a factor ˜4 higher pixel density (˜100x100um^2 ->50umx50um^2 pixels) working with very high hit rates (3-4GHz/cm^2) in an extremely hostile radiation environment (˜1Grad over 10 years). The first years of developments were focussed on radiation tolerance studies of the chosen 65nm technology and implementing and qualifying required radiation hard IP building blocks: DACs, ADC, Analog pixel Front-Ends (AFE), Biasing structures, Bandgap, PLL, IOs, power regulator, temperature and radiation sensors, etc. An appropriate hit processing and readout architecture for the high hit and trigger rates was developed and extensively simulated and verified in a flexible simulation and verification framework [31] based on system Verilog and UVM (Universal Verification Methodology). This simulation and verification framework is used for the critical verification of the pixel chips before their submission for production.

A first 1/2 sized (as sharing reticle with other chips) , but complete, pixel chip called RD53A [8]was submitted in 2017 and has been used extensively over several years for verification of developed IP blocks and general architecture. RD53A has also been instrumental as a test vehicle to test and qualify a large number of different pixel sensors and for system studies, covering serial powering, design and test of pixel modules and test of complete readout systems using LPGBT [39] based optical links to DAQ. A large set of irradiation test campaigns have been made of this chip to get a good understanding of the long term function and reliability of such a complex chip covering: TID effects as function of temperature, Low dose rate effects and initial SEU/SET tests. 3 different analogue fronts-ends were present in this chip together with two different trigger latency buffering schemes to determine the most appropriate implementation for final production chips.

Next generation RD53 chips, named RD53B-XXX, are complete full sized pixel chips made with chosen latency buffer architecture and improved IP building blocks. RD53B generation chips are made specifically for each experiment (RD53B-ATLAS, known as ITK-v1 in ATLAS, and RD53B-CMS, known as CROC-v1 in CMS) as chip size needed to be adapted to the specific needs of each experiment and also uses experiment specific analogue front-ends. These two chips are 99% functionally equivalent with same control and readout interface, with minor specific features related to the analogue front-ends and minor added features and bug fixes. At the point of writing this document, the RD53B-ATLAS chip, with its Differential analogue front-end, has been submitted in is under extensive testing. The RD53B-CMS chip is in the process of being finalized with its linear analogue front-end and a few additional monitoring and debugging extensions.

RD53C-XXX generation chips will be final production version chips, with minor bug fixes and extensions according to exhaustive testing and qualification of the RD53B generation chips and systems.

1.2 Hit and trigger rates

As the LHC luminosity is significantly increased (factor 5 - 10) for the HL-LHC phase 2 upgrades, the required trigger rates have also been increased by a factor ˜10. This , together with using 4 times smaller pixels, results in a required effective readout rate increase of a factor ˜100. Hit rates for the inner layer pixel modules will be as high as 3-4GHz/cm^2 with a gradual hit rate decrease for for outer layers following a general 1/r^2 trend. Required trigger rates are for CMS 750KHz, with pixel modules very close to the interaction point (so very high hit rates) and 1MHz for ATLAS. ATLAS also has an option of a future upgrade with a 2 level tracking trigger where the trigger rate can be as high as 4MHz/1MHz when using the 2 level trigger/readout scheme. Readout link speeds have been increased to 1.28Gbits/s, compatible with max E-link (Electrical

links) rates of the LPGBT and up to 4 such links are available per pixel chip for the highest rate regions. Highly efficient pixel hit data encoding/compression (called Binary tree encoding) is also required to be capable of reading acquired pixel hit data on the smallest number of links possible. For outer pixel layers the required effective readout bandwidth is so low that only a fraction of the bandwidth of a single E-link is needed. To reduce material of readout links in this part the detector (more than 50% of active pixel detector surface) it is possible to merge locally on the pixel module readout data from 2 or 4 chips together into a single readout E-link.

1.3 Radiation requirements

The increase of LHC luminosity also results in significantly increased radiation levels in the pixel detector. For 10 years operation the inner pixel layers should ideally remain functional after ˜1.4Grad (even up to 1.9Grad in the ultimate running scenario of CMS). This has been seen to be a major and difficult challenge for the used 65nm technology and it has been set as a realistic goal to guarantee 500Mrad radiation tolerance when cooled below 0°C, implying exchanging inner pixel layer after 5 years of operation. Analogue IP blocks have been optimized for radiation tolerance and have been shown to have very good radiation tolerance. Compact digital logic, made with small transistors, has been seen to have significant radiation induced delay degradation (100 - 200%). This has been taken into account in the digital design flow of the RD53B chips. Extensive radiation tests of the RD53A chip, with extrapolations of low dose rate effects and cold operation temperature (-10°), have given indications that RD53B chips can potentially remain functional up to the Grad level [**?**].

Radiation induced Single Event Effects (SEE), covering Single Event Upset (SEU), Single Event Transients (SET) and Single Event Latch-up (SEL), are also critical issues to keep a complex pixel chip operating reliably in such an extreme hostile radiation environment. Appropriate design and protection schemes (e.g. TMR: Triple Modular Redundancy) must be applied in critical parts of the design (PLL, configuration, state machine, event data, hit data) to assure sufficiently reliable operation of such a large and complex pixel chip. It is estimated that ˜100 storage elements (flip-flops) in each chip of the inner pixel layers will have their content corrupted every second. More details on radiation tolerance measures in RD53 chips are outlined in chapter 15.

1.4 Powering and cooling

As the pixel detector constitutes the first detection layers in the LHC experiments, it is critical to keep its material budget as low as possible to minimize multiple scattering and particle conversions. A large complex and high rate pixel chip in 65nm CMOS technology will require significant power supply currents, of the order of several amperes per chip, at a low voltage of 1.2V, making it very sensitive to voltage drops and power regulation. Direct powering of more than 10k pixel chips in a pixel detector would lead to an in-acceptable material budget just from basic power cabling. RD53 chips therefore use a novel serial powering scheme, where several (10-20) pixel modules can be powered in series, reusing the injected current between multiple modules and chips in a serial power chain. A specialized serial Shunt Low Drop Output (SLDO) power regulator on the chip itself generates the required stabilized power for the chip and in particular low noise power for the sensitive analogue front-ends in the pixels. A particular novel feature of the used serial powering is that several (2-4) chips on a module can be powered in parallel, as part of a serial power chain, assuring reliable power delivery in case of chip failures. Serial powering with local on-chip regulators can tolerate large voltage drops on power cables (helps to minimize material in power cables at cost of power losses) and also keeps power supply current fluctuations to an absolute minimum, making it optimal for low noise systems.

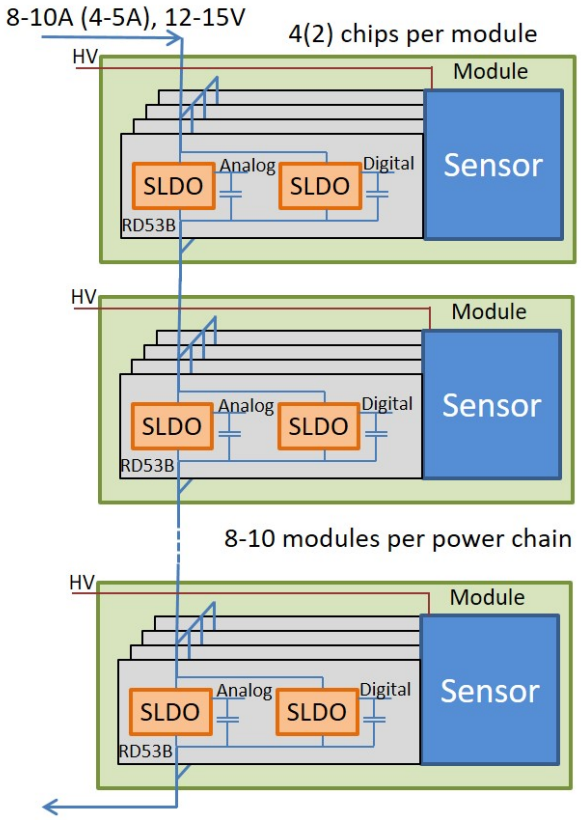


Figure 1.1: Serial powering of pixel modules with multiple chips per module

The advantages of serial powering with on-chip power regulators comes at the cost of power dissipation in the on-chip SLDO regulators (two per chip) that under normal operation can dissipate 20-40% of the power of the active circuitry. Under exceptional conditions (e.g. chip configured to consume little power) the on-chip power regulator may have to dissipate a large fraction (e.g 80%) of the total chip power. It may also have to dissipate significant additional power in case one of the 4 (2) chips in parallel on a module has a power regulator failure. This requires particular attention for the overall powering and cooling systems of a detector using serial powering. Serial powering also requires special precautions for all communication interfaces to the pixel chips (links with AC coupling) and detector grounding (e.g. HV pixel sensor biasing). More details on this is given in chapter 11,

1.5 Control and readout

Control and readout interfaces of RD53 chips are severely constrained from their specific use in an inner layer detector and the required use of LPGBT based optical links to DAQ and control systems of the experiments. Readout and control links must be implemented with the smallest number and lowest mass possible, but must handle significant readout rates (multi gigabit/s) with a high level of reliability in an extremely hostile environment. A highly efficient hit data encoding format (binary tree encoding) has been developed specifically to minimize readout bandwidth. The option of using 1 to 4 readout links per pixel chip, and the option of merging data from 2 or 4 chips into one link, enables the number of readout cables to be optimized for a detector where readout rate requirements decreases with distance to the interaction point ($1/r^2$ dependency). Control and and readout links must use an encoding compatible with AC coupling, required because of serial powering.

A dedicate 160Mbits/s differential control link has been developed to control up to 15 chips

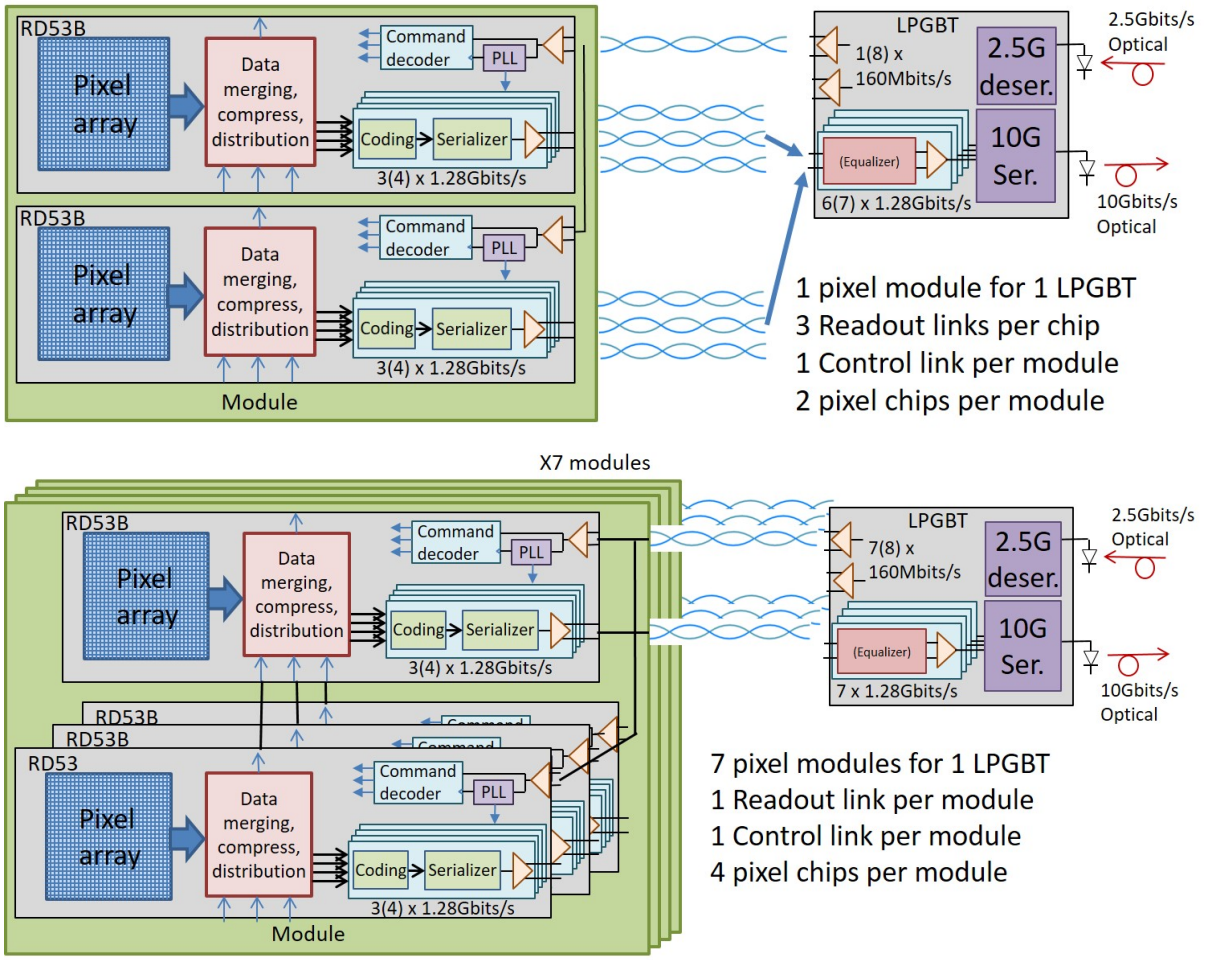


Figure 1.2: Pixel module control and readout. Upper: Inner modules with Multiple (e.g. 3) E-links per pixel chip. Lower: Outer modules with data merging

(1111bin used as broadcast) with embedded reference clock with ns timing control to appropriately align pixel sampling with the bunch collisions. High priority real time commands at 25ns level (e.g. triggers without constraints) together with low priority control and monitoring commands assures that a single control link can fully control multiple pixel chips.

Up to 4 lanes of 1.28Gbits/s are available per chip for reading out acquired and triggered pixel hit data and also reading out necessary monitoring information. Aurora encoding/formatting of readout links has been chosen as it supports all required features (AC coupling compatible encoding, appropriate formatting and framing, data and service type frames, multi lane support, streams with minimized bandwidth overhead, etc.). Aurora is well documented [27] and well supported for FPGAs used in the DAQ systems. It is brought to the attention of the reader that when pixel chips are used together with the LPGBT [39] optical links, two levels of link encoding and formatting (LPGBT encoding on top of Aurora) will be present on the optical links, that must be decoded in the DAQ system FPGAs.

Chapter 2

Architecture and data flow

RD53 pixel chips capture pixel hits across its large (~2x2cm^2) pixel array with appropriate timing related to the bunch crossings to enable a well defined fraction of hits to be accurately triggered for readout over high speed serial links (and LPGBT optical links) to the DAQ. Charge deposited in the pixel sensor, bump bonded to the pixel chip, is amplified and shaped so it can be sampled precisely in the correct bunch crossing, with associated charge information. Sampled and zero-suppressed hit information is stored during the trigger latency (max 12.8us) in latency buffers distributed across the pixel array in small local pixel regions consisting of 4 pixel cells. Transfer of triggered pixel hit data from the pixel regions is organized into 8x8 pixel cores (consisting off 2x8 pixel regions) via columns of pixel core buses to the Digital Chip Bottom (DCB). Triggered hit/event data is in the DCB checked and processed before being queued in derandomizer FIFOs for readout over serial links in their original trigger order.

The RD53 architecture is organized in a hierarchical fashion of which some parts are related to the logical data flow (figure 2.2)while others are related to the physical implementation/floorplan (figure 2.1) of the chip:

- Pixels: Individual pixel with its bump pad, Analog Front-End (AFE) with associated calibration charge injection switches and specific pixel configuration.
- Pixel regions: 4 x 1 pixels with digitization of leading edges and TOT (Time Over Threshold charge information) with shared latency buffering in 8 latency buffer locations.
- Pixel cores: 8x8 pixels, 2 x 8 pixel regions, 64 pixels. Number of pixel cores are diffeent in ATLAS and CMS chips
- Pixel core columns: 50 in ATLAS chip, 54 in CMS chip
- Pixel core rows: 48 in ATLAS chip, 42 in CMS chip
- Pixel array: ATLAS: 8x8x50x48 = 153.600 pixels. CMS: 8x8x54x42 = 145.152 pixels.
- Analogue pixel islands (details in chapter on AFE): 2x2=4 pixel AFEs
- Analogue biasing columns: 2 pixels wide columns for AFE biasing with dedicated bias drivers
- Digital Chip Bottom (DCB): Digital logic outside the pixel array handling all global control/monitoring and readout of event data.
- Analog Chip Bottom (ACB): General Analog circuitry including: Bandgaps, Biasing, ADCs, DACs, PLL, etc.
- Pad frame: IO with wire-bond pads, ESD protection and SLDO power regulator.

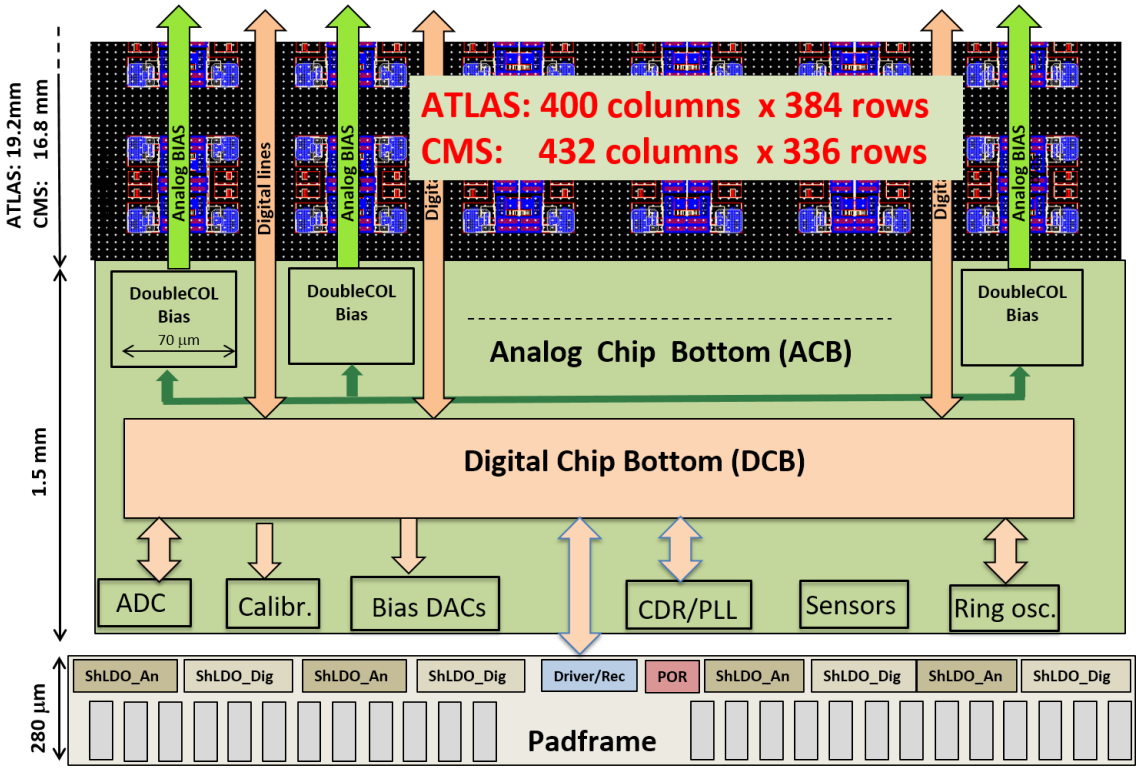


Figure 2.1: Physical floorplan of RD53B chip with DCB functional view showing analog pixel islands, with their biasing, surrounded by digital pixel region logic

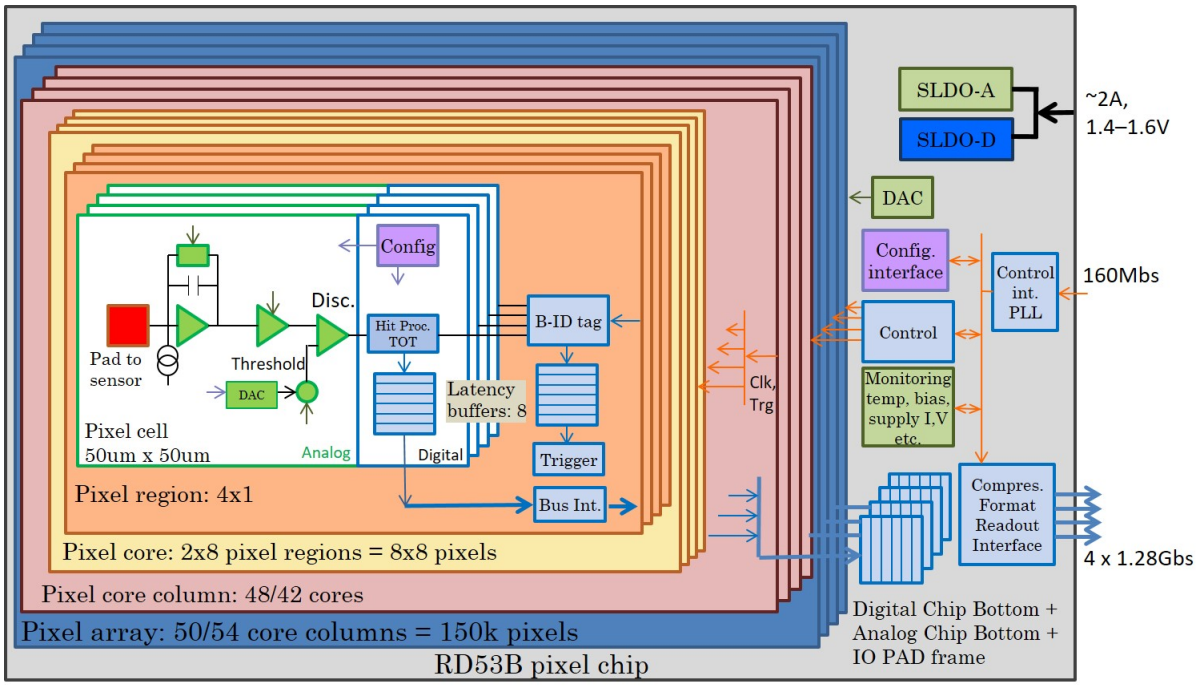


Figure 2.2: Data flow architecture of RD53 chip

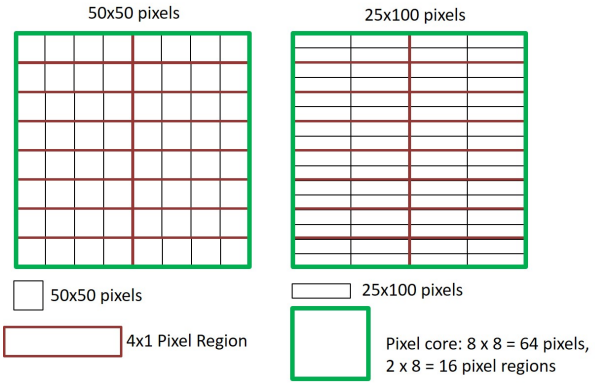


Figure 2.3: Pixel region configuration with 50x50um^2 and 25x100um^2 pixels

It is important to notice that the effective organization of pixels into pixel regions, with shared latency buffer, depends on the organization and routing on the pixel sensor. For square 50x50um^2 pixels the pixel sensor array and the pixel chip have the same pixel array structure and the pixel regions are seen as 4x1 pixels covering 50x200um^2. Such an elongated pixel region shape is advantageous for latency buffer sharing at the end of the pixel barrel, where pixel clusters are elongated because of particle angle and active pixel sensor thickness. When having elongated pixels of 25x100um ^{2} on the pixel sensor, the pixel region is effectively seen as a 2x2 pixel region covering the same 50x200um^2 pixel region area.

Chapter 3

Command, control and timing interface

All command, control and timing of the pixel chip is performed via the single 160Mbits/s differential control link that can drive/address up to 15 chips (4 bit addressing in slow control commands + broadcast). The function of the pixel chip itself is fully bound to the 40MHz bunch crossings of the LHC and the 160Mbits/s control link speed is defined from having to transfer concurrently the 40MHz reference clock to the chip (40MHz clock encoded in link) and being capable of transferring 40MHz clock synchronous control signals/commands (trigger, resets, etc.) and in addition loading required configuration and request readout of specific monitoring information.

The custom command link encoding is optimized for its specific purpose. It assures a max run length of 2 with full DC balance over 8 bits (sync command exception with DC balance over 16bits) for efficient and low jitter AC coupling and PLL clock recovery. The phase alignment to the 40MHz system reference clock is performed with dedicated sync frames that must be sent at regular intervals to assure correct alignment to the 40MHz system clock. All clocks in the pixel chip are derived from a central 1.28GHz PLL using transitions in the control link as synchronization references. The PLL uses both rising and falling edges during lock acquisition, and can then be configured to use either both edges or only rising edges, being less sensitive to duty cycle distortion on the control link (RD53B-CMS uses by default only rising edges). All clocks in the chip are derived from this 1.28GHz reference and appropriately phase aligned with the 40MHz system clock, signalled by dedicated sync frames.

To accurately phase align the pixel chip to the actual bunch collisions (e.g. sampling of the hit signals in the pixel array) the absolute phase of the 40MHz clock has a configurable phase with 0.78ns resolution assuring appropriate phase alignment to the LHC collisions. Alignment at the level of clock cycles to the hits (e.g. triggers) must be done in the control link transmitter (DAQ module) together with the configurable trigger latency of the chip. The pixel chip assures constant latency of all real time commands (also when doing power cycling and and link re-sync). The phase alignment and stability of the sampling clock in relation to the received hits requires significant attention at the system level. Longer term phase stability will depend fully on phase drifts in the timing distribution system of the experiment. Phase drifts in the pixel chip itself (clock delays but also delay in analogue front-ends) will depend on temperature stability (and also power supply) and is estimated to be better than 100ps per degree change (To be verified). Phase drifts in the pixel chip will also depend on induced radiation effects, but should remain stable (at the level of 100ps) during a single run where up to 1Mrad TID can be accumulated for inner layer detectors (To be verified).

At startup the PLL will go through a locking phase to correctly generate its 1.28GHz reference clock (requires valid control link sync/idle frames) and will then go through a 40MHz clock alignment sequence requiring appropriate sync frames. During normal operation received frames are continuously checked and after a given number of illegal frames will try to re-synchronize (but not re-initialize PLL as this may de-sync the whole readout chain). A complete control link re-

sync and chip reset can at anytime be enforced by a **special link reset** command/sequence, by applying for a given time (10us) a dedicated bit sequence with more than 2us between transitions (so not complying with normal control link encoding) Correct AC coupling of this also needs to be assured, but jitter is in this case not critical. Single bit command errors can be detected (but not corrected) and if a corrupted command frame is detected, the pixel chip will ignore such a command. Number of detected corrupted commands can be monitored. After a configurable number of detected corrupted commands/frames the control link is assumed to have gotten out of sync and the chip will try to re-sync to the command frames.

The RD53A demonstrator chip was sometimes seen to have issues to obtain and keep appropriate PLL, control link and clock synchronization, so this has been significantly improved in RD53B generation chips.

Appropriate transmission and termination of the control link is critical to obtain a stable working pixel chip and pixel detector. Jitter must be minimized as this will propagate to the high speed serial readout links. The on-chip PLL has jitter filtering capability for high frequency jitter above the effective PLL control loop bandwidth (˜1MHz). Low frequency jitter/ phase variations can not be filtered by the PLL. The PLL has the option of using both transitions on the control link for frequency locking (more transitions) or only use the leading edge transitions during operation (less transitions but then not sensitive to possible duty cycle distortion on the differential control link transmission path).

Control links must be appropriately AC coupled to the pixel chip (normally on the pixel module end). When having multiple chips on the pixel module connected to same control link the signal distribution on the HDI (High Density Interconnect) must be designed with great care to minimize reflections. A parallel fan-out is not recommend as it will give significant transmission discontinuities at the fan-out point and is impossible to terminate correctly (depends on stub lengths but even stubs of 1-2cm have been seen to cause reflection issues that can be very hard to measure and diagnose). It is recommended to have a serial chain distribution from chip to chip on the module with proper termination at the end of the differential distribution chain. The effective time skew between chips is small (100ps) and not critical for the correct capture of pixel hits in the 25ns bunch crossing period. The differential control link receiver in the pixel chip does not have on-chip termination resistors, so this can be optimized at the module level to the transmission cables used and the distribution network on the pixel module. The receivers have an active DC biasing circuit to assure appropriate DC biasing after AC coupling consisting of an impedance of ˜4Kohm to the appropriate common mode level (Vdd/2). (The RD53B-ATLAS has DC biasing set to 0.26V). Multiple receivers with DC-biasing can be present on a control link. Effective receiver input capacitance (800fF to chip gnd per pin) and wire-bonding inductance (1nH per 1mm wire-bond) must be taken into account when optimizing control link distribution on the pixel module. AC coupling of the control link on the pixel module can be done with a single set of AC coupling caps common for all chips on the module or with individual AC coupling per chip.

The PLL is known to be sensitive to power supply noise so is powered via dedicate power pads. It is recommended to connect the PLL powering to the analogue power domain from the analogue SLDO regulator. An off-chip passive power filter can if needed be implemented on the pixel module.

The pixel chip has the option of doing control link forwarding to other pixel chip(s) on the module or neighbouring module(s). The control link data from the differential receiver is simply driving differential driver(s), without any active jitter filtering, as on-chip PLL is not part of this command forwarding signal path (PLL will obviously be present on pixel chip getting control link from the control link forwarding pixel chip).

Clock synchronous real time commands, consisting of 16bit frames (covering four 25ns clock cycles) have absolute priority and can interrupt low priority slow control commands (loading config, request for monitoring info from ADC, etc), that simply resumes their previous

command sequence after high priority real time command frames. As real time commands covers 4 LHC clock cycles, the **trigger command** has a 4 bit field indicating which of the 4 bunch crossings have active triggers (no constraints on consecutive triggers). Each trigger command also has a user defined 6bit trigger tag (can only use 54 of 64 values because of encoding constraints) that will be assigned to the event when being read out (different than traditionally use of BX-ID and L1-ID used in current generation front-end chips). This **trigger tag** will be appended a **2bit sub-tag** defining in which of the four bunch crossings it belongs to, adding up to a trigger tag of 8 bits in the event readout. It is important that the pixel chips are not given overlapping trigger tags as they are the basis for event handling in the chip. A trigger tag can only be re-used when a previous event with the same tag has been read out. Trigger tags will normally be issued in straight sequential order making it equivalent to a limited range event ID (e.g. 6LSB of a 16bit event ID). The main difference to the classical combination of event-ID and BX-ID is that the trigger tags are not based on local counters in the pixel chip, that can get out of sync. Event-ID and BX-ID can optionally be added to event data read out (not available in RD53B-ATLAS).

Other real time commands (e.g fast clear, resets, cal pulse injection) can not be sent at the same time as trigger commands and the overall system will have to take this into account. It can be mentioned that the pixel chip does not require regular bunch count resets, per machine cycle, as the latency buffer does not rely on this (would actually prevent appropriate function of latency buffers across machine cycles). The **clear command** is in particular implemented to be capable of doing a fast clear of all data buffers and state-machines in case normal event processing has been seriously compromised by a SEU/SET (See chapter 15).

As real time commands (e.g. triggers) have absolute priority, and can temporarily interrupt any **slow control command** (handled in real time by DAQ module FPGAs), it is possible to continuously send slow control commands, for continuously re-writing configuration registers (called continuous/trickle re-configuration) and is highly recommended for the pixel configuration when being used in a hostile radiation environment. (See chapter 15).

Chapter 4

Analogue front-end and hit capture

The Analogue Front-Ends (AFEs) are grouped in 4 pixels and implemented as small analogue islands in a sea of pixel array logic, as illustrated in figure. 2.1. The grouping of 4 AFEs into analogue islands, where the AFE layout has been mirrored/flipped to fit together, can cause minor systematic mis-match differences because of the different orientation of analogue transistors. After appropriate threshold adjust, such systematic AFE differences will be very small. The analogue islands are isolated from the surrounding noisy pixel logic using the deep Nwell (triple well) available in the 65nm technology. AFEs also use a separate analogue power supply (together with Analogue Chip Bottom: ACB) coming from a dedicated analogue SLDO regulator.

Induced charge from traversing particles in the pixels of the pixel sensor are transferred to the pixel chip via fine pitch bumps. The analog front-end in each pixel integrates deposited charge by a pre-amplifier stage, with charge integrating feedback, followed by appropriate signal buffering/shaping. The discharge of collected integrated charge is done with a programmable discharge current (sometimes referred to as Krummernacher current), resulting in an analog pulse width proportional to charge (linearity of this is front-end specific). The linear AFE in the CMS chip is linear within few % where as the differential AFE for the ATLAS chip is not made to be linear. Amplified/shaped signal, fast rising edge and slowly decreasing falling edge, as illustrated in figure 4.2, is transformed into a 1 bit digital hit signal by a discriminator with programmable threshold. A simplified block diagram of a generic AFE is indicated in figure 4.1. The AFEs work independent of the 40 MHz bunch crossing clock and hits are only synchronized to the 40MHz clock at the entry to the digital pixel hit processing in the pixel regions. Effective threshold per pixel is determined by a global threshold bias together with a 5bit threshold adjust per pixel, to compensate for threshold dispersion among pixels. The dynamic range of the threshold adjust is defined by a global programmable bias to allow dynamic range and resolution to be optimized to the actual threshold dispersion in the pixel array.

Several biasing levels for the AFEs are set via global configuration registers connected to biasing DACs driving the analog pixel array:

- Pre-amplifier biasing: Determines effective speed of AFE charge integration (but also affects effective gain, noise and dispersion). Major contributor to AFE power consumption.
- TOT discharge current: Defines discharge rate of integrated charge thereby defining TOT resolution and effective dynamic range.
- Discriminator biasing: Determines effective speed (and time-walk) of discriminator.
- Global threshold: Global chip threshold on-to which threshold adjusts will be applied in each pixel.
- Threshold adjust range: Defines range of local threshold adjust DAC so this can be made to cover observed threshold dispersion

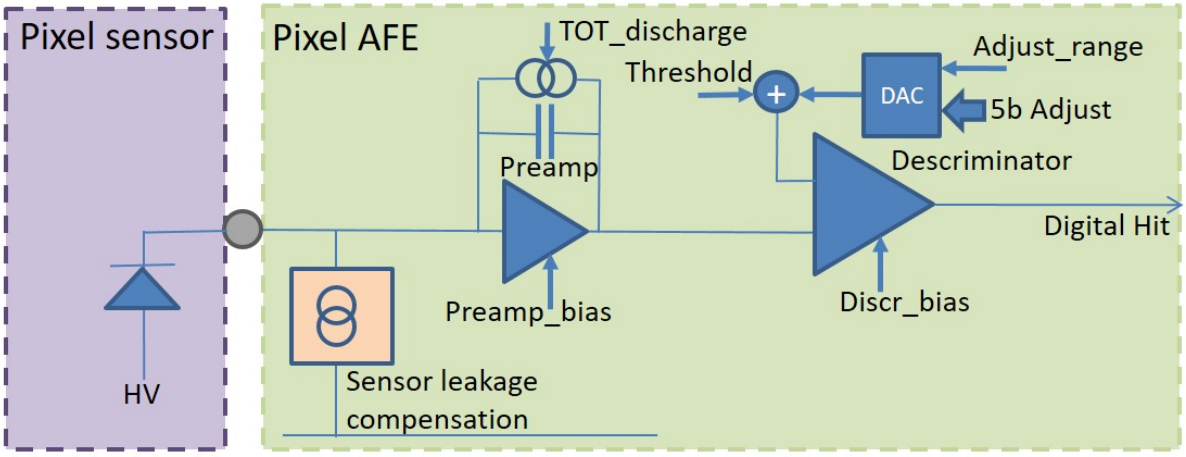


Figure 4.1: Block diagram of generic AFE

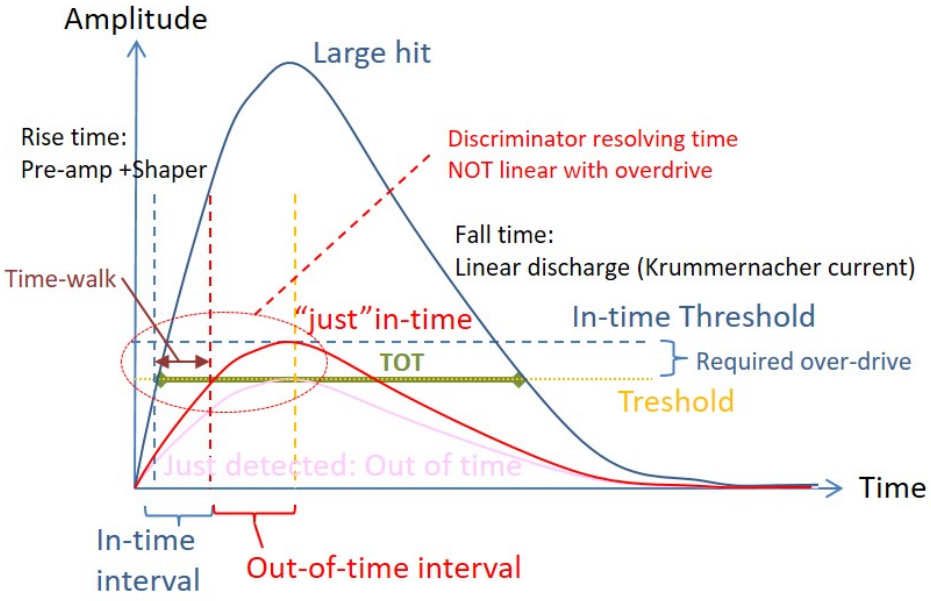


Figure 4.2: Charge measurement with TOT and indication of time walk for just above threshold signals

Biasing of the AFEs have significant effects on their behaviour and power consumption and the reader is referred to dedicated information on this for the specific AFE (Linear AFE [18], Differential AFE [17]) . Optimization of best possible AFE performance, at a given allowed power budget, is a delicate optimization that must be done according to pixel sensor characteristics, hit rates and accumulated radiation effects in both pixel sensor and the pixel chip itself. Biasing to the AFEs are distributed to the pixels according to the physical layout floorplan of the AFEs in analogue islands consisting of 2x2 pixels (not same structure as pixel regions that are 4x1 pixels) with biasing lines organized in columns with width of 2 pixels, having dedicated biasing drivers per biasing column (can eventually also give some minor mismatch behaviour/structure). All biasing is driven by global biasing DACs.

Edge pixels have separate pre-amplifier biasing as edge pixels on the sensor typically are 2-4 times bigger and therefore also bigger capacitance. This specific biasing is organized in 6 groups having specific pre-amp biasing, so can be adapted to different pixel module configurations (single, dual, quad chip modules):

- Central/main pixels
- 2 left side pixels
- 2 right side pixels
- 2 top pixels
- 2x2 top left corner pixels
- 2x2 top right corner pixels

For the specific biasing of the analog front-end and its effective behaviour, the reader is referred to detailed documentation and characterization of each AFE.

Leading edge of discriminator hit signal is a measure of arrival time of particle (with associate time walk) and pulse width is proportional to deposited charge. Leading edge is synchronized to the 40MHz sampling clock and pulse width is measured with single or both edges of the 25ns sampling clock (40 and 80MHz effective pulse width sampling) for a Time Over Threshold (TOT) measurement of 6 bits. The 6 bits TOT counting can be mapped directly into 4 bit TOT (ignoring 2 MSB bits) or the 6 bits can be mapped into 4 bits, with a dual slope mapping. With dual slope mapping the full TOT resolution is maintained in the first half of the 4 bit dynamic range whereas high dynamic range is obtained in the second half of the 4 bit range (with 1/4 resolution) having an effective conversion gain of 1/4 in this part of the range, as illustrated in figure 4.3. Dual slope assures good position interpolation at the edge of pixel clusters, where collected charge is normally small, and high dynamic range for dE/dx measurements of pixel clusters that can be used to identify highly ionizing particles and contribute to general particle identification. The TOT counting in the pixel will incur an effective pixel TOT dead-time given by the average charge deposit and used TOT resolution (Typically 4 - 8 25ns clock cycles = 100 - 200ns). More information on TOT dead time is given in chapter 16 .

Capture and synchronization of the discriminated hit signal can be done in two alternative ways: Simple sampling where short hits not present at the rising edge of the sampling clock will be missed. Latched sampling where a short discriminator pulse is kept high until it has been captured by first coming 40MHz sampling clock edge. Latched sampling is guaranteed to capture short hit pulses, but will have slightly higher sensitivity to noise hits (by definition short) as indicated in figure 4.4 . The RD53A chip was implemented with simple sampling and the RD53B-ATLAS chip has been made with the latched sampling scheme. For the RD53B-CMS chip (and later versions to come) it has become possible to configure the sampling mode used.

Sampled hit signals are processed in small local pixel regions consisting of 4 pixels. When one (or multiple) pixels in a pixel region has a hit, binary hit information (hit or no hit) plus 4

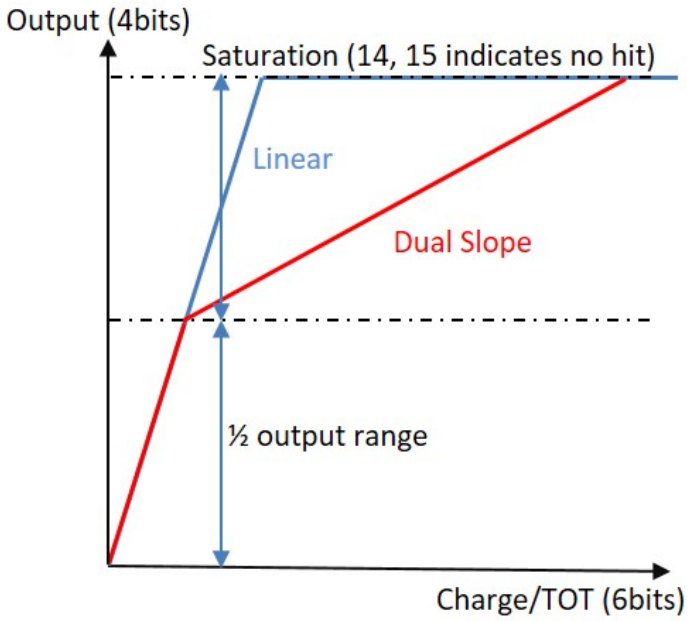


Figure 4.3: Linear and Dual slope mapping TOT

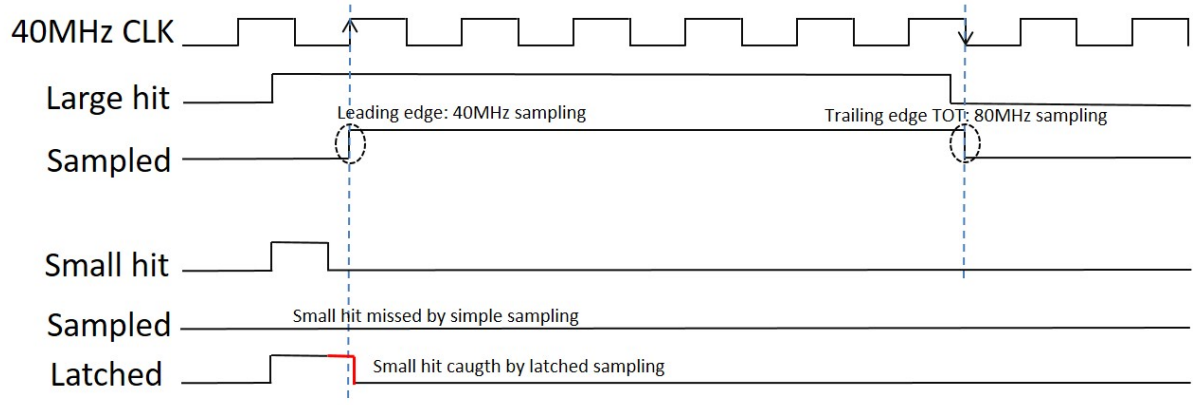


Figure 4.4: Hit sampling mode with basic sampling and latched sampling where small (out of time) hit is extended until captured by 40MHz sampling clock. Indication of 80MHz TOT sampling

bit TOT (TOT=1111bin used to indicate no hit) of the four pixels are stored in a local latency buffer location where it awaits a trigger coming with a configurable latency. This is implemented by storing together with the hit/TOT data a 9 bit Bunch ID, from a central 40MHz Bunch-ID counter and distributed to the whole pixel array. The bunch ID of pixel hits in the pixel region is determined by the first leading edge. Writing to a buffer location is not finalized before the TOT measurement(s) are finalized (takes minimum 2 clock periods and up to a maximum of 64 clock cycles determined by the largest charge/TOT among the 4 pixels). The pixel region hit capture and buffering is made in a non-blocking fashion where a new hit arriving in following clock cycles on a pixel, not part of first cluster, will be captured in the next free buffer location. Each pixel region has local latency buffers with storage for up to 8 pixel region hits (that can each have from 1 -4 hits with TOT). The grouping of 4 pixels into local pixel regions enables a significant reduction in required storage for the latency buffer as hits are normally clustered from traversing particles typically making multiple hits in neighbour pixels. The cluster size and shape depends on location of traversing particle (e.g. middle of pixel or between two pixels), angle of particle combined with pixel size and pixel sensor thickness, magnetic field and its direction, and finally radiation effects in the pixel sensor that tends to spread/diffuse charge over larger area when having significant radiation damage. If all latency buffer locations in a pixel region are occupied, additional incoming hits are simply ignored. Hit losses from this at the absolute highest hit rates (3.5GHz/cm^2) have been modelled and simulated with Monte Carlo hit data and shown to be well below 1% [32]. More information on this can be found in chapter 16 .

The digital logic in the pixel array, capturing and storing hit information during the trigger latency, use clock gating to make significant power savings in the large amount of pixel array logic. Pixel region hit capture logic only has active clocking during the capture window of a hit (window depends on length of TOT) . This gives significant power savings in the pixel logic but making power consumption dependent on hit rates.

Chapter 5

Latency buffering

Traditional storage of detector hits during the trigger latency is made as a simple clocked pipeline (power hungry) or as a circular buffer (much less power and using high density dual port SRAM). This is simply excluded for use in a pixel detector chip with extremely small area available per pixel and a relatively long trigger latency. For a detector with low occupancy per channel (<1 %) it is much more efficient to store zero-suppressed hit information with an associated time tag (bunch ID). As indicated in figure 5.1 for a single channel with 4 bits of charge information only 104/2k = 5% storage is required for a 12.8us latency. When having a latency buffer organized in 4 channel regions this is getting as low as 8*(9+16)/4*2k= 2.5%, as bunch ID time tag shared among 4 pixels in a pixel region. To this has to be added the overhead of handing the 8 buffer locations and the bunch ID comparison logic.

A latency buffer with 8 buffer locations per pixel region has been verified to be sufficient to guarantee a low risk (<1 %) of the buffer running full at the absolute highest hit rates (3.5GHz/cm^2). This has been verified with statistical modelling and with simulations with realistic Monte Carlo hit data [32] and results from this can also be seen in chapter 16 .

When a latency buffer location is in active use, the stored Bunch ID is continuously compared to a global latency counter, also distributed across the pixel array. This trigger latency counter runs with a relative offset to the hit bunch ID counter, defining the effective trigger latency. When the stored hit bunch ID matches the trigger latency counter the hit information have been waiting for a time equal to the trigger latency. If an active trigger is generated at this moment, the hit information is flagged as triggered and have to be read out. Otherwise hit information is discarded and buffer location is released. When pixel region hit data is tagged as triggered, the Bunch ID information is exchanged with a trigger event ID. Multiple pending triggered hits will then await its corresponding readout from the pixel array.

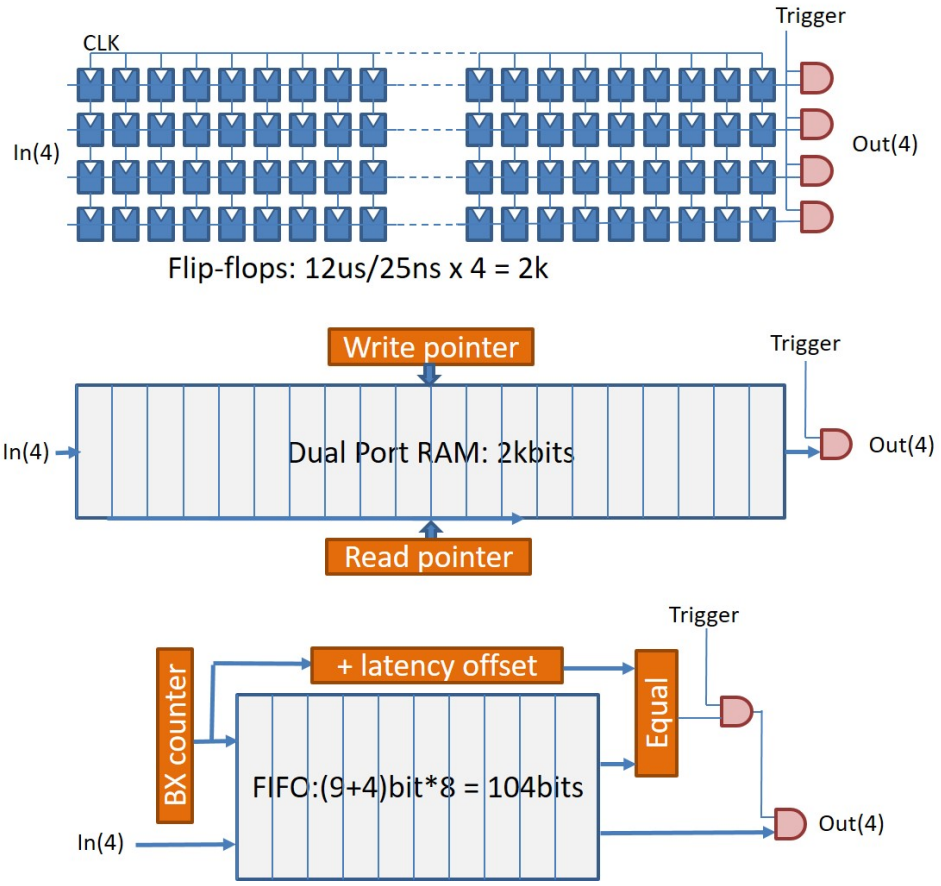


Figure 5.1: Latency buffering as pipeline, circular buffer or zero-suppressed with a time/bunch tag. Required storage indicated for 12.8us latency

Chapter 6

Readout from pixel array

Readout of triggered hit data from the local pixel regions (awaiting in the pixel region latency buffers and having been flagged for readout) is controlled from a core column controller at the end of each core column bus. Pixel cores, consisting of 2x8 pixel regions (8x8 pixels), shares a core column readout bus, with its associated controller in the DCB. The readout from the pixel array of triggered pixel region hit data is initiated by the core column controller signalling to all its pixel cores (and its associated pixel regions) the event ID of the event to read out and sends a readout token to the top of the column. Pixel regions having readout flagged hit data with an event ID matching the event currently under readout, will await the arrival of the readout token. When the token arrives to a pixel region with awaiting hit data, it will assert its data on the shared readout bus together with its pixel region address and indicate the presence of active readout data. It then passes the token to the next pixel region with hit data for the same event. The sequence of pixel cores/regions being read from the pixel array is therefore given by their physical location in the core column. When the token finally returns to the pixel core column controller, all event data in the core column for this event have been collected.

Each pixel core column has its independent readout controller, so different pixel core columns can be in the process of reading out different events at any point in time. This improves the effective readout rate from the array when having multiple pending triggers/events, at the cost of needing local lists of triggers awaiting readout. As a pixel core column bus is covering a very large number of pixel regions (48x16 = 768), the effective readout speed on this long bus is strongly constrained by the speed of the bus (affected by radiation) and the associated token passing mechanism. The effective readout speed on the core column bus is therefore two clock cycles per pixel region data-set, that with radiation degradation can get as long as 3(4) clock cycles (configurable).

It is possible by configuration to constrain the maximum number of pixel regions to read out from each core column per event, to prevent possible readout congestion from events with excessive number of hits. If the max hit region count is reached in a column, remaining pending hits in the pixel regions are simply cleared and the event is flagged as having reached the max hit count for this core column. It is also possible to constrain the maximum time available to readout all pixel core columns (thereby effectively constraining the number of hits per event).

A central trigger table keeps track of events awaiting readout from the pixel array, and small distributed copies of this information is used in core column readout controllers to handle the local readout in each column (as not event synchronized across all core columns).

To support a future ATLAS upgrade option with a two level trigger scheme, the generic single trigger scheme has been extended. In the normal one level trigger scheme, hit data flagged as triggered by the L0 in the latency buffers will initiate immediately their readout from the pixel array (with some possible waiting time when pixel core bus busy reading out previous events). For the 2 level trigger extended scheme, L0 triggered hits remain in the pixel region latency buffer for a configurable time-out period (max 25.6us). During this time-out period (L1 trigger

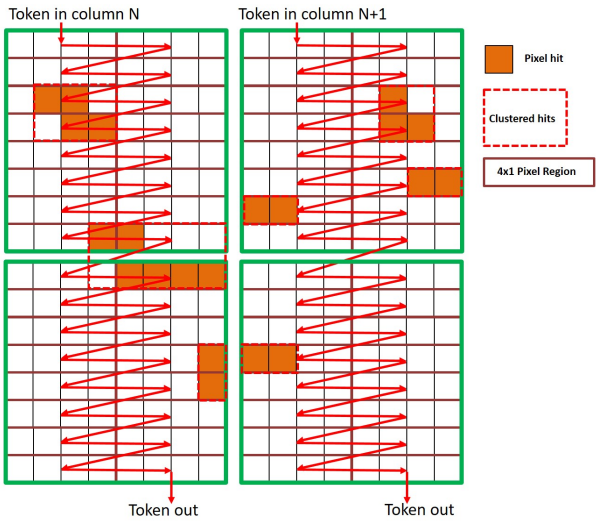


Figure 6.1: Pixel array readout organized in pixel core column buses with token based readout and individual controllers in the DCB

latency), they can via a dedicated command, with associated trigger tag (6 bit + 2 bit extension), be flagged to be read out (accepted by L1), or otherwise by default being rejected (L1 rejected), at the end of the time-out period . The L1 trigger accept/read command can be a broadcast to all chips on a control link or can be addressed individually to each chip on the control link. A fraction of hit data (given by L0 trigger rate over bunch crossing rate. e.g. 10%) may therefore occupy a few latency buffer locations in the pixel regions during the additional time-out period, that will reduce the effective number of buffer locations available to store hit data during the L0 latency (10 - 20% effect depending on L0 accept rate and time-out period) .

Chapter 7

Data processing and buffering

Event data accepted for readout will go through multiple levels of processing, event building, buffering and formatting before being ready for final readout via the serial readout links. Multiple levels of data buffering is used in the overall processing to serve multiple purposes: Align events from independent core columns, event building, handling different processing times, crossing clock domains and finally as derandomizers before the readout bottleneck of the readout links. Effective de-randomization of event sizes and instant trigger rates are required to be capable of effectively using up to an average of 70 - 90% of the available maximum readout bandwidth (more info in chapter 16).

To prevent excessively large events to clog up the whole readout, it is possible to constrain the number of pixel hits (regions) per core column and the maximum time to readout all hits from the pixel arrays (see previous chapter). If hits have been discarded, it is appropriately flagged in the event.

Excessive amounts of noise hits (or other sources like photon induced hits) can optionally be removed by a single isolated hit removal function (real particle hits are for a large majority making multiple hits in clusters).

Finally hit data can be formatted with a very efficient binary tree encoding scheme, particularly suited for clustered hit data, reducing the required readout bandwidth to ˜1/2 compared to simple zero-suppressed single hit encoding. It is also possible to suppress the TOT charge information, having only binary hit information, giving a data rate reduction of ˜30%. Readout encoding with raw pixel region hits with TOT is also available for debugging purposes (4x4bits TOT + 12bits pixel region address).

The total amount of data buffering before the final readout is of the order of 25kBytes in the DCB plus event data buffering that effectively takes place in the pixel array from a trigger is given until hit data has been read out of the pixel array.

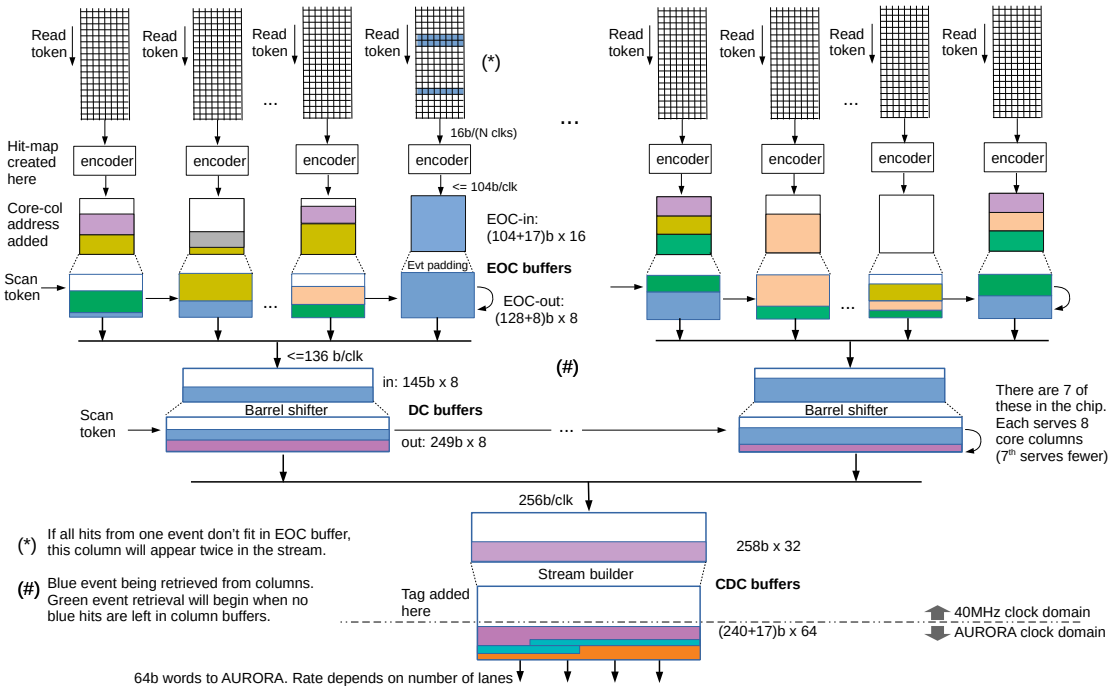


Figure 7.1: Processing and buffering of event data before readout. From [5].

Chapter 8

Readout formatting and serialization

Variable length event data (as zero-suppressed hits with binary tree encoding) are finally formatted, framed and encoded for appropriate transmission on 1 - 4 serial readout links with configurable serialization speed. Readout framing and formatting is done in accordance with the Aurora links definition from Xilinx [27], and therefore readily available as interface IPs for Xilinx FPGAs for use in RD53 test and DAQ systems (not appropriate for final systems as LPGBT coding on top). Aurora line encoding is done with standardized 64B/66B encoding, based on scrambling, compatible with AC coupling required when using serial powering. It has a built in distinction between normal data (event data) and service data (monitoring and config. read-back) defined by a 2bit header in the encoded 66bit frames. The bandwidth fraction allocated to service data is configurable and will typically be set to 1-2% of the available bandwidth.

64/66 bit data words/frames can be organized into event streams to maximise data bandwidth utilization. The number of events to put in a single event stream is configurable, so can be optimized for the specific application. When having multiple events in a stream the bandwidth utilization is the best possible, as frame padding is done at the end of the stream, not for each event. When having single individual events per stream, event padding is needed for each event (Average padding = half a frame = 33bits). At a 1MHz trigger rate a bandwidth of 33Mbits/s per lane is lost to padding: 33Mbits/s / 1.28Gbits/s = 2.5%, when having single events per stream. If having data merging from 4 chips on one readout link the padding overhead becomes 4x 2.5%=10% or bigger if a majority of events have no hits (so many empty events with lots of padding). This can be reduced proportionally when having multiple events per stream. When programmed for a given number of events per stream, a stream may still have less events than this in case no more events awaits readout.

If a bit transmission error occurs in a stream, then multiple events can become corrupted (because of link encoding scrambling and binary tree encoding). Using single events per stream, only one event will be affected by a single bit transmission error. Please notice that Aurora does not have Forward Errror Correction (FEC), as is the case for the LPGBT links. Effective event data corruption is estimated to be low enough that this is considered acceptable for a pixel detector where effective readout bandwidth is critical.

Any number of readout links can be used (1,2,3,4), all having same speed. Multiple lanes carrying a single data stream uses strict lane alignment (all having either data or service frames). It is advisable (but not required), to use the lowest numbered links (e.g. link 1 and 2 and not link 3,4 when using 2 links). From the RD53B-CMS an option of having a CRC (Cyclic Redundancy Check), for transmission error detection, at the end of a stream is available.

Serialization of readout data is done with the internal PLL clock, locked to the control link. Readout link jitter will therefore depend on effective jitter on the control link. High frequency jitter (>1MHz) will to some extent be filtered by the PLL whereas the PLL for low frequency phase drifts will simply follow. This is of critical importance for the E-link interface to the

LPGBT, as it does not have phase following PLLs per E-link input, but relies on phase stability relative to its own internal PLL, also used for the control links to the Pixel chips (assuming using same LPGBT to drive control link as used to receive readout E-links). If this is not the case, then relative phase drifts between the control link (e.g. one LPGBT) and the readout link (e.g. another LPGBT) must be extremely well controlled and stable.

Serialized data are driven by dedicated differential CML (Current Mode Logic) drivers with configurable drive currents (max 14mA) and configurable pre-emphasis. The driver has an effective output differential impedance of 100ohm (50 ohm single ended) assuring best possible matching to 100ohm differential electrical cables (twisted pair, flex micro strip-lines, twinax). The 100ohm driver impedance assures that possible transmission cable reflections are terminated when coming back to the driver, and therefore prevents multiple reflections on the cable. It should be noticed that the driver current have to drive both the driver self-termination and the receiving end termination (effective resistance per line of 25ohm).

Data transmission must be appropriately AC coupled when used in a system with serial powering. This AC coupling must be done with an appropriate time constant for the used serialization speed and the DC balance given by the 64B/66B encoding with scrambling (based on statistical DC balancing and a max run length that can be as long as 64bit). High speed ceramic 100nF capacitors have in general been seen to be appropriate for this. These AC coupling capacitors must have sufficient voltage rating to deal with voltage levels encountered in a serial powering system (20V) and it can be recommended to have a 2x safety factor on this as isolation failures of these capacitors can have severe consequences. AC coupling also requires an appropriate DC biasing network for the receiver to maintain absolute common mode voltage levels compatible with the effective common mode range of the differential receiver. This will typically consist of a 5 - 100kohm resistive impedance to the the mid point of the differential receiver (e.g. Vdd/2). AC coupling is normally done on the receiving end of the link in combination with the termination, but can in principle also be done on the transmitting end (on pixel module) if dictated by particular grounding issues. Optimization of readout link transmission, attenuation, reflections, AC coupling and appropriate reception (e.g. jitter and phase tracing) requires careful attention, simulations and systems tests to be assured to be reliable. The reader is referred to consult specialized literature on this delicate and critical issue. In the ATLAS detector with long link readout cables (6m) a dedicated cable equalizer ASIC will be needed before the LPGBT to assure appropriate data transmission at 1.28Gbits/s.

The readout latency (time from trigger to whole event has been read out) will have a complex dependency on statistical fluctuations in hits, triggers and the available readout bandwidth. The readout latency will become particular critical if the available readout bandwidth is close to the average required bandwidth (e.g. 80 - 90%). Readout latency studies have been made with Monte Carlo hit data and the reader is referred to such specific studies (Experiment specific and strongly dependent on pixel layer and number of readout links used). Examples of typical readout latency distributions are given in chapter 16.

A highly optimized binary tree hit encoding scheme is used to format readout data to minimize required readout bandwidth. This format is particularly efficient for clustered hit data. TOT information is conveniently put after the binary tree encoding of hits to enable easy and efficient removal of TOT information for binary pixel readout. With the binary tree encoding there is not a fixed number of bits per hit as multiple hits in a local cluster is encoded to together for simple and optimal data size reduction. In practice the number of bits needed per hit has in simulations with realistic Monte Carlo hit data been seen to be in the range of 10 - 15bits per hit (for more info look in chapter 16). If using the raw readout format (available as debugging option), as coming from pixel array, the number of bits per hit is in the range from a absolute maximum of 28bits/hit (single hits in pixel region) to ˜14bits/hit (if having ˜2 hits per pixel region).

From the RD53B-CMS chip an optional CRC checksum, at the end of each event stream, is

available according to the Aurora standard definition of this.

Stability of the Aurora readout lanes depends strongly on signal transmission and signal quality on the electrical differential cables used for this (and jitter on control link). More information and measurements of this can be found in [28].

Chapter 9

Data merging

Data merging is made available for low rate outer pixel layers, to merge readout data from 2 or 4 chips on a pixel module into a single readout lane, thereby making significant reductions to the number of required readout links and related material. A master chip is driving a single readout link and 1 or 3 slave chips, on the same physical control link (to have same clock time reference), drives their low rate readout data on a single or dual lane local link to the master chip. As the data bandwidth from the slave chips in this configuration is by definition low, the local slave to master links only needs to run at relatively modest speed. The local slave to master links are constrained to work at 1.28Gbits/s /4 = 320Mbits/s thereby limiting speed issues (an option of 640Mbits/s local links is available but is not recommended to be used before careful test and qualification of this has been made). In case of 2 chip merging, the slave chip drives two 320Mbits/s lanes to the master chip. Master and slave chips are assumed to run fully synchronous with very similar and stable clocks phases as driven from same control link (with embedded clock) and is assumed to have similar temperature and radiation damage (affects timing significantly).

It was decided to keep Aurora readout encoding on these local links (could have been simplified as AC coupling not needed and all chips in tight synchronisation). The master chip has 3 low speed Aurora receivers, that can be connected to any of four local link inputs that allows to make pixel modules that can be configured to use different data merging configurations. The slave to master Aurora formatted data is descrambled and 66bit data frames are stored in small FIFOs in the master chip, until it can be transmitted on the master readout link. Merging from the four (2) chips (3/1 slaves plus data from master itself) is done in a simple round robin fashion among the four at the frame level. To identify the chip source of each frame, the 64 bit data words includes a 2 bit chip identifier followed by 62 data bits (only needed when doing data

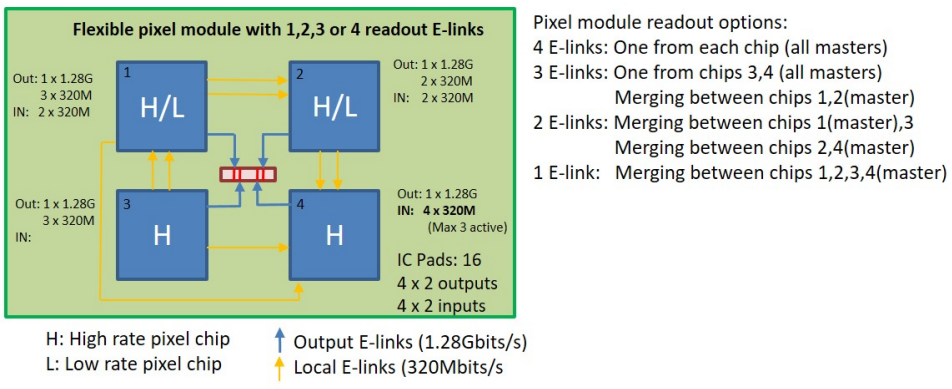


Figure 9.1: Data merging between master and slave chips on flexible quad module that can have 4, 3, 2 or 1 readout links

merging). The insertion of the 2 bit chip ID in each data word is done by each chip itself, based on its chip ID (from wire bonding). As the effective master link bandwidth is exactly 4 (2) times the bandwidth from its slaves, a simple round robin frame merging can keep up, without need of supplementary data buffering in the master chip. The input of the data merging in the master chip is over-sampled with a 640MHz clock, using both clock edge (1.28GHz sampling), from the PLL and the most appropriate sampling phase is determined by a signal transition detection circuit on each local link. This can handle +/- 1ns phase variation on the data merging inputs. If a larger phase change has built-up over time (e.g. non uniform radiation) the data merging between chips may need to be resynchronized.

It is important to notice that the simple basic frame based merging with chip IDs does not assure that the chips sharing a link are actually in the process of reading out the same event !. The DAQ system receiving event data on a readout link with merged data must therefore consider this as being 4 (2) separate event flows that must be processed as coming from four separate chips.

For the local data merging lines it is not recommended to use AC coupling (as not needed) and for the short chip to chip connections the driver can be configured to run with relatively low currents (no pre-emphasis needed). From the RD53B-CMS chip it has been decided not to have on-chip differential termination on the data merging inputs (so RD53B-ATLAS chip has on-chip differential termination). As the serial data output drivers are self-terminated, it is normally not needed to also terminate the data merging receivers at this modest speed and short distance. This implies that the effective signal amplitude is doubled on these local links for a given driver current. If differential receiver termination is needed, it must be added on the pixel module.

Chapter 10

Monitoring

The pixel chip has extensive monitoring capabilities of its environment and its own internal functions. A 12 bit ADC is used for analog monitoring. This ADC is a relatively slow successive approximation ADC that requires to go through a conversion cycle of several clock cycles. This conversion sequence must be initialized by a slow control command, using the generic pulse generator, and then the conversion result can be read out. Monitoring information is on the readout link put in special service frames (marked in 2bit Aurora frame header) that are available at regular intervals with a programmed fraction of the link bandwidth.

Multiple temperature sensors are located within the chip to be capable of monitoring the temperature and gradient over the pixel array and monitor the temperature of the chip bottom close to the SLDOs, that are known to be particular hot spots (in particular in certain failure modes). An external NTC (Negative Temperature Coefficient) or PT1000 temperature sensor can also be biased with a programmable current and be read via the ADC (or used to monitor an external voltage when source current set to zero). An issue with unstable readings from the two temperature sensors at top and bottom of pixel array have been addressed in the RD53B-CMS.

SLDO power supply currents at the input and in the shunt can be monitored. The effective load current must be be calculated as difference between input current and shunt current. Supply voltages can be monitored together with internal references and biasing voltages/currents. An issue with non-linear shunt current measurements in RD53B-ATLAS has ben resolved in RD53B-CMS.

Radiation effects can be monitored via a set of digital ring oscillators based on different gate types and transistor types (Normal Vt and Low Vt). These ring oscillators have been seen to be particular useful in radiation testing as gives a direct measure of radiation induced delays in digital circuits (which is known to be the main radiation effect in the chip). It should be mentioned that the ring oscillators are quite sensitive to supply voltage so this much be taken into account when making radiation testing with these. Basic radiation effects at transistor level can also be monitored (but in a rather complicated fashion) via two different temperature sensor structures, that have different radiation sensitivity.

Counters are available to collect statistics on detected error conditions (e.g. corrupted commands).

An auto read feature is available to readout the content of selected status and configuration registers at regular intervals, without making specific read request via the control link. In the current version of chips this can not be used to get regular analogue monitoring information from the monitoring ADC as an ADC conversion needs to be triggered by specific control commands.

From the RD53B-CMS chip version (so not in RD53B-ATLAS version), basic counting of SEUs in the triplicated global configuration and in triplicated pixel configuration register bits will be available (see more under radiation effects).

Chapter 11

Powering and cooling issues

RD53 pixel chips have been made specifically to be powered with serial powering, but also supports direct powering and powering with on-chip LDO (Low Drop Output) regulator. Used with direct powering the power regulator is completely bypassed by wire-bonding. In LDO mode the shunt regulator is disabled and the SLDO (Shunt LDO) works as a classical LDO power regulator. Direct powering is particular useful for specific characterization tests, but is not recommended to be used in a pixel detector system. The SLDO (or LDO) is critical to assure stable and low noise power supplies within the chip. The chip has separate power regulators for analogue and digital functions to prevent noisy digital logic to be capable of disturbing the sensitive analogue parts (in particular AFEs and PLL). In particular the PLL has its own power pads, so an external power filter can be made for this if required (PLL should be powered from low noise analog power domain). It can also be mentioned that both power domains are isolated from the chip substrate, that are then connected together to the local pixel module ground and the grounds of analogue and digital power (no absolute ground available when using serial powering). The chip substrate is available as a separate pad that must also be connected to the local module ground.

Effective decoupling of power supply variations is of particular importance for serial powering (but also for LDO and direct powering) as the chip will need, for short periods, larger currents than what is actually supplied to the pixel module. This refers to short power peeks within the 25ns clock cycle but also longer power variations across multiple clock cycles, as the pixel chip uses extensive clock gating to obtain vital power savings in its digital logic. All digital pixel, pixel region and pixel core logic uses clock gating. The hit sampling and TOT counting in a pixel region is only activated when a hit is present. Latency buffer locations and their readout from the pixel array are only active when active hits are accessed. Digital processing, buffering and formatting in the chip bottom also use clock gating. It is therefore critical to have sufficient local decoupling capacitances to serve as small local energy reservoirs. This is needed at both the input to the SLDOs and at the outputs of the SLDOs. Decoupling at the SLDO output is in particular needed to handle high frequency variations (within 25ns clock cycle) and the decoupling at the SLDO input is mainly to handle lower frequency variations, but also to reduce noise coupling in the serial power chain between analogue/digital, chips on same module and between modules. The overall time constant of this two level decoupling must be of the order of microseconds.

A large number of wire-bond pads are used for the SLDO input current and the connection to external decoupling of the SLDO output, to be compliant with the required current levels and, even more critical, to have very low inductance to the external decoupling caps. High speed (Low ESR) capacitors must be used for decoupling (e.g. ceramic capacitors) and the routing between these and the chip must have minimal inductance (so short) and when possible have power and ground planes (small but very fast capacitance for high frequency components). It is recommend to have 6uF decoupling for serial power input and have 2uF for each power domain

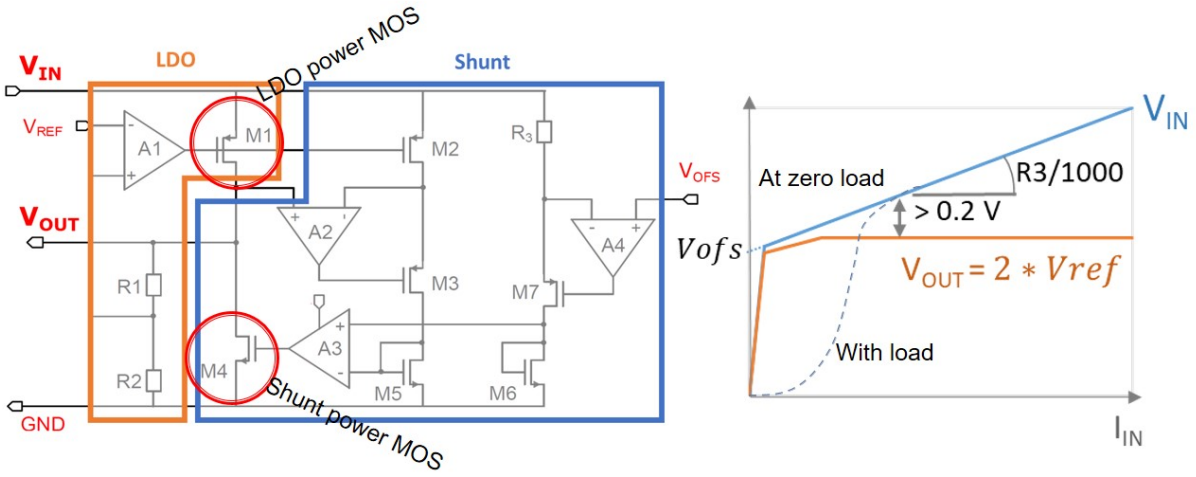


Figure 11.1: SLDO power regulator functional diagram with critical LDO and SHUNT power transistors

output (digital, analogue). It is advantageous to have multiple smaller distributed capacitors (e.g. 4 x 2.2uF for power in), distributed along the power bus along the chip. The material budget of these decoupling caps can be significant so optimization of this will be required based on realistic system tests with representative hit and trigger rates (and their fluctuations). It can be mentioned that the effective regulation bandwidth of the SLDO and decoupling will be in the same range as the effective trigger rates (MHz) so this must be carefully verified in system tests.

The output voltage of the SLDO/LDO regulators can be tuned by configuration within a limited/safe voltage range (1.1 - 1.3V) to allow to compensate for possible chip to chip differences and to compensate for possible (unexpected) long term radiation and ageing effects. At startup the middle of this range is used (nominal 1.2V) which is in general the optimal supply voltage of the chip. The output voltage is defined by an internal bandgap reference that has shown excellent immunity to temperature variations, TID and variations of input voltage. The digital supply voltage can eventually be programmed to be higher than its nominal value to compensate for radiation effects, slowing the digital circuitry, at the cost of increased power dissipation (potentially same can be applied for the analogue).

Precise references and biases are critical for a large number of analogue functions in the pixel chip. In the RD53B generation chips (was different in RD53A and had some issues), all references and biases are derived from a single high precision bandgap driving a reference current through an external high precision resistor. This enables all references to be independent of radiation (bandgap obviously made to the extent possible to be independent of radiation, temperature, voltage, etc.) and allows possible adjustments/corrections to be made with the external resistor. In addition the reference current in the reference resistor can be adjusted via 4 wire-bonds. The use of these wire bonds or/and adjusting external resistor allows specific chip and module tuning to be done based on chip wafer probing characterization (but can not be adjusted during the lifetime of the module).

11.1 Serial powering

RD53 chips have very specific serial powering features that makes it significantly different than traditional serial powering (e.g. simple shunt regulators used on transatlantic communication cables). The RD53 SLDO regulator allows analogue and digital power domains to have independent SLDO's with very good noise isolation. It is also possible to have multiple chips powered in

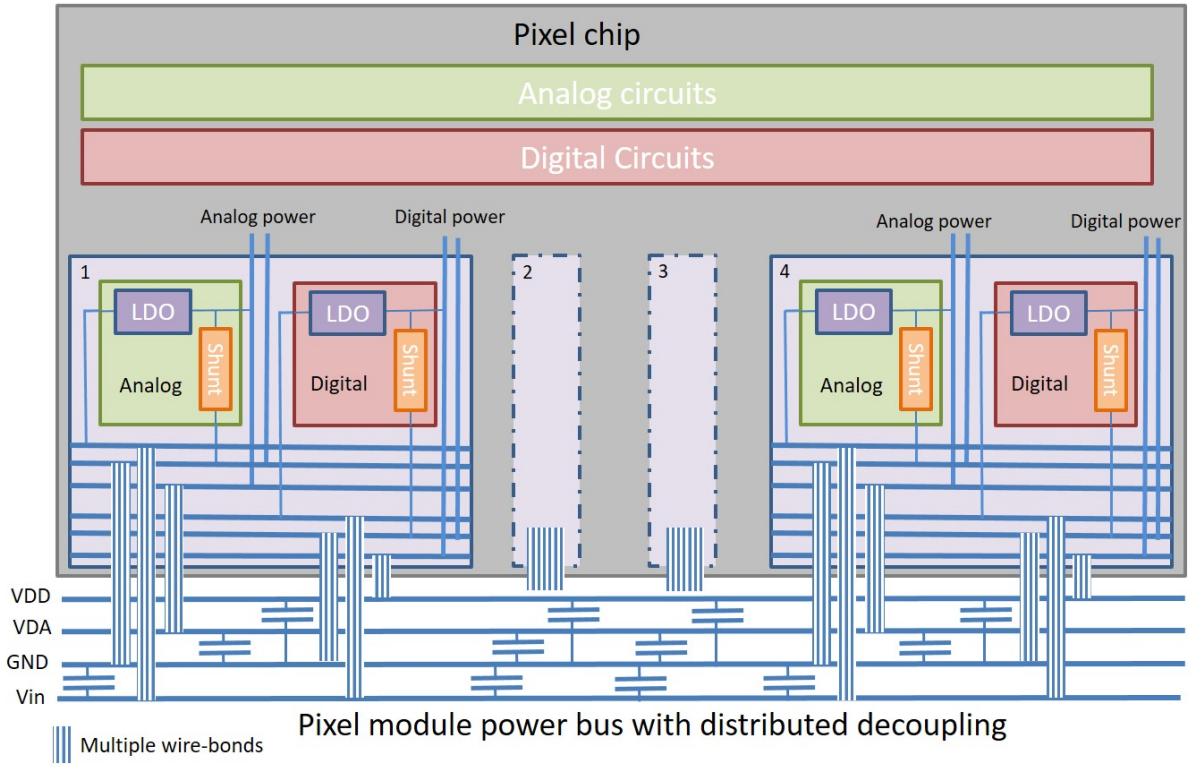


Figure 11.2: Distributed decoupling at both SLDO inputs and outputs

parallel (e.g. 4 chips per module) ending up with 8 SLDOs in parallel as outlined in figure 1.1. Current injected in a serial power chain is passing multiple (e.g. 10) pixel modules and within pixel modules the current sharing between chips/SLDOs in parallel requires special attention (classical shunt regulators can not do proper current sharing).

For a serial powering system it is critical to inject more than the average current needed by the loads. A current head-room of 10 - 30% is required to assure that there is sufficient available current for the loads (the active part of the chip) at all times taking into account a set of critical factors. Chips will have chip to chip differences in required current and injected current must cover the most demanding in the chain (e.g. from radiation effects). Current fluctuations over time must be covered by local decoupling and margins are required for this to be efficient as indicated on figure 11.3. Current sharing between chips in parallel will not be perfect and margins for this mis-match will also have to be covered.

Current sharing is obtained by having each SLDO appearing (from the power input side) as a constant voltage drop (source) plus a well defined resistive impedance making it appear as a voltage clamp/zener (offset voltage) with a defined resistive impedance as indicated on figure 11.5. The output of the SLDO is a voltage source stabilized by a LDO (independent of current). The RD53 SLDO regulator is implemented with two active control loops where these three key parameters can be defined separately (Output voltage, Input resistance and offset voltage). The voltage to which one sees the input resistance is termed Voff (offset). The input characteristics is (in principle) defined by a shunt regulator and the output by a LDO regulator. The two regulators though work closely together via two coupled control loops. The LDO output voltage is defined by an on-chip bandgap and can be adjusted from a configuration register. The input impedance and related Offset voltage are defined by external resistors. The input characteristics are not adjustable via configuration registers in the chip as changes to these will affect significantly the full serial power chain and could potentially become dangerous for the system (SEU/SET and mis-configuration). Optimization of the input characteristics must be done as function of required operation current by the chips and the required current sharing

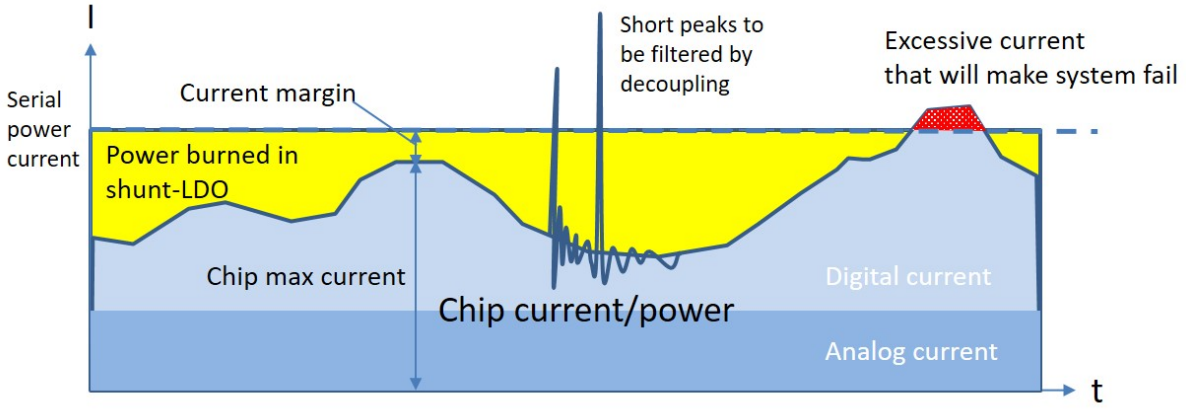


Figure 11.3: Serial powering constant current supply and issue with varying load current

between the analog and digital part of the chip. The offset voltage is the same for the analog and digital regulators (and normally also between chips). Current sharing between analog and digital is then defined by the difference in effective input resistance. For the RD53 chips the required current for analog and digital domains are similar and can in first approximation be assumed to be the same (in the 40-60% to 60-40% range), depending on chosen AFE biasing, effective hit and trigger rates and number of used readout links.

To operate a serial powering scheme with acceptable power losses in the SLDO (see below) and being capable of operating over a given current range, it is quickly realized that it is best to have an effective Voffset in the range of 1 - 1.2V. At lower offset voltage, and higher resistance, modest changes in injected current will imply the input voltage to increase significantly when increasing current (giving significant power losses). Use of higher Voffset, and lower resistance, gives significant issues assuring appropriate current sharing when having minor differences between chips. The SLDO regulator requires a minimum voltage difference between input and output of 0.2V at full load. To have some margins on this for a nominal output voltage of 1.2V, it can be recommended to have a input voltage of 1.5 - 1.6V at nominal working conditions (minimum is 1.4V). Default values of Voff=1.0V and effective input resistance of 600mOhm (possibly 500mohm) gives a solid working point of 1.6V input voltage at a current of 1A per SLDO (total of 2A). This can/should be optimized for specific systems and specific working conditions.

When analysing current sharing mis-match between chips in parallel is it quickly realized that it is particularly critical to have good matching of Voff among SLDOs/chips in parallel. Voff mismatch between chips is estimated to be of the order of 2-5% (depending if they are from same wafer or from different wafers and lots). Good Voff matching can be obtained by matching chips on same module based on wafer characterization (chips can be divided into batches depending on Voff). Alternatively one can adapt the Voffset external resistor to each chip, based on wafer probing characterization. It is also an option to adjust Voff indirectly by modifying the common current reference of the whole chip, but this obviously affects all references on the chip and have to be done with great care and a good understanding of side effects. Finally the RD53B generation chips also have a new option of having a common Voff among chips on same pixel module, via a simple external resistor network between separate Voff output and input terminals. For this one though have to consider possible system effects of having chip failures (see below).

Matching of the input resistance is in practice less critical but can obviously also be obtained adapting the external resistor defining this to each chip (or have chips sorted according to this). It can be mentioned that a big advantage of the RD53B generation chips is that SDLO input and shunt currents can be monitored with the on-chip ADC (this was not the case for the RD53A chip and made it very difficult to determine input current sharing between multiple chips on a

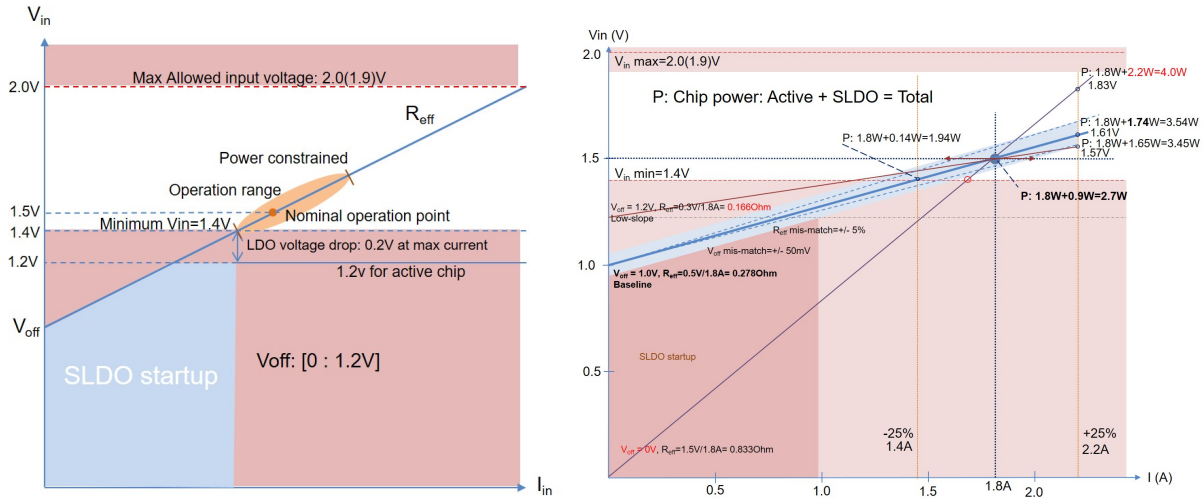


Figure 11.4: Serial power working point optimization. From [21]

pixel module).

In the RD53B generation chips a high current voltage clamp protection is available to limit the input voltage to below 2V, which is the absolute maximum voltage that the SLDO can safely handle (limited by 65nm CMOS technology). This can be for short voltage spikes that can occur during dynamic events (power up, power down, system failures, etc.) and for static failure conditions. It must be mentioned that if the input voltage gets close to 2 volt during extended periods it will normally imply excessively large power dissipation in the SLDOs, as SLDO operation is normally configured for a nominal input operation voltage of 1.4V - 1.6V.

In case a chip consumes more current than what is available in the serial power chain (failure, mis-configured, insufficient current in serial power chain, etc.), the SDLO output and input voltages will gradually collapse according to how the load current depends on supplied voltage. When having multiple chips with appropriate current sharing and current headroom, a single chip consuming more current than nominal, will first eat into its own current head room, then start to consume current headroom from the other chips in parallel and will eventually cause its own output voltage and the shared input voltage to collapse. The other chips in parallel will also become affected by this input voltage collapse. When such a local voltage collapse occurs, it will in principle not have direct effects on the pixel chips on other modules in the serial power chain, as the serial chain current is kept constant by the external constant current type power supply. The local voltage collapse on a module will though result in the serial power chain total voltage to decrease and can therefore be observed by voltage monitoring of the serial chain power supply. The locally seen SLDO output voltages on the other modules will basically not be affected by this and can continue to work as before. It must though be kept in mind that the voltage collapse of a single module will affect the absolute voltage level of modules, after the affected module in the chain. This will affect HV biasing of the sensor (referring to global ground and not local grounds) and also absolute voltage levels across AC coupling capacitors for the control and readout links, and the system will likely have abnormal function (events with many noise hits, bit corruption on readout links, etc.) during a short time window where the local voltage collapse occurred (system tests required to get a better understanding of such system effects).

RD53B chips have an optional (not enabled by default) under-shunt protection feature. This is basically like an indirect output current limitation, implemented by checking that there is a shunt current in the local SLDO. If the local SLDO shunt current gets very small it implies that either the load is consuming too much current or that there is a major issue with the current sharing between chips. If the under shunt current protection sets in, the output voltage is reduced

(assuming that output current will then decrease) until a shunt current of more than 10mA is present. Because of technology constraints (max source drain voltage) the output voltage can not be lowered to less than 0.7V. This in practice implies that the under-shunt protection option can not be used to protect against hard shorts on the chips. The under-shunt protection scheme has been seen in some system simulations to cause current oscillations between parallel chips, that needs to be checked with realistic system tests.

Detector grounding is a delicate issue that must be dealt with with great care when making a serial powering system. Local module grounds MUST be isolated from global system ground. Serial power chains must be connected to global system ground at one single well chosen ground reference point (e.g. entry to pixel detector Faraday cage). All communication links must have AC coupling, with appropriate DC biasing after AC coupling capacitors. Location (pixel module, DAQ module, LPBT interface module) of AC coupling is critical for potential ground loops and uncontrolled current paths. AC coupling will required some time to stabilize if DC voltage levels are changing (power up, local pixel module power failures, etc.). EMC coupled noise to common levels will also have to considered. Special attention has to be taken for any cable shield that in general can only be connected at one end DC wise and the other end may have to be connected capacitively to local grounds (of left floating). Power(current) supplies must have floating outputs but still be appropriately connected to system/safety ground. HV biasing for the pixel sensors is a particular issue, especially when biasing all pixel modules in a serial power chain with a common HV supply. Each pixel module will have its specific effective HV biasing level, affected by the absolute voltage level on each module (Can be up to ˜20V difference). Particular issues have also been seen to arise if the HV is off and the LV is on (and the opposite) where detector leakage currents can lead to unfortunate forward biasing of some detector modules. Other, possible overlooked, pixel module connections (temperature sensors for detector safety system, etc.) can not be connected to the local pixel module ground planes. Any overlooked (mechanics, cooling, etc.) or accidental (when probing, failing AC coupling capacitor, etc.) connection from local pixel module ground to global system ground can have severe consequences.

11.2 Power dissipation and cooling

The vital advantages of serial powering (significantly lower material in powering cables, low noise, etc.), comes at the cost of higher power dissipation in the on-chip power regulators (and grounding issues). In nominal operation the SLDO regulator will have to dissipate power related to the required voltage head room (e.g. 1.5V - 1.2V = 0.3V) and the required current head room (10 - 30%). In nominal operation this results in a power dissipation in the SLDOs of the order of 20 - 40% of the power required by active chip circuitry. The SLDOs in the RD53 chips are located between the chip bottom and the wire-bonding pad frame, distributed along the full width of the chip to spread this power dissipation to the extent possible and keep some distance to the AFEs in the pixel array. It is critical that the SLDOs are appropriately cooled on low mass pixel modules and that this does not cause a major temperature gradient across the pixel array (e.g. $<10°C$).

When a chip is not in nominal operation mode, serial powering has the (unfortunate) characteristics that the chip must/will consume the same current/power being injected in the serial power chain. If digital logic is shut down (by disabling clocking) and/or the analog front-ends are configured for very low power, then the power that are normally consumed/dissipated here will have to be consumed/dissipated in the SLDOs. In absolute worst case all chip power may have to be consumed in the SLDOs and it is critical that this can not lead to destructive local overheating of the chip/module. Such a case can occur if the chip by mistake has been configured to have such a low power configuration, while injecting the full serial chain current. Measures have been implemented to prevent such mis configurations by having a constrained configuration

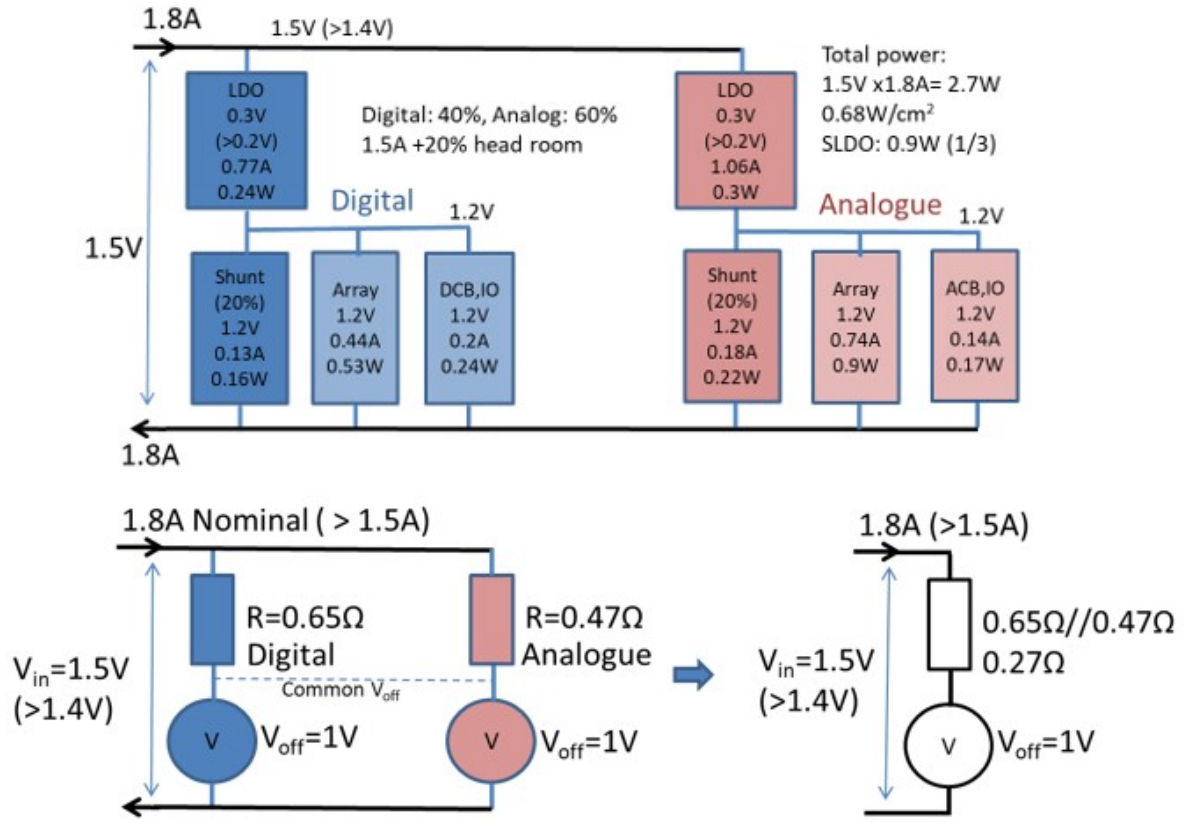
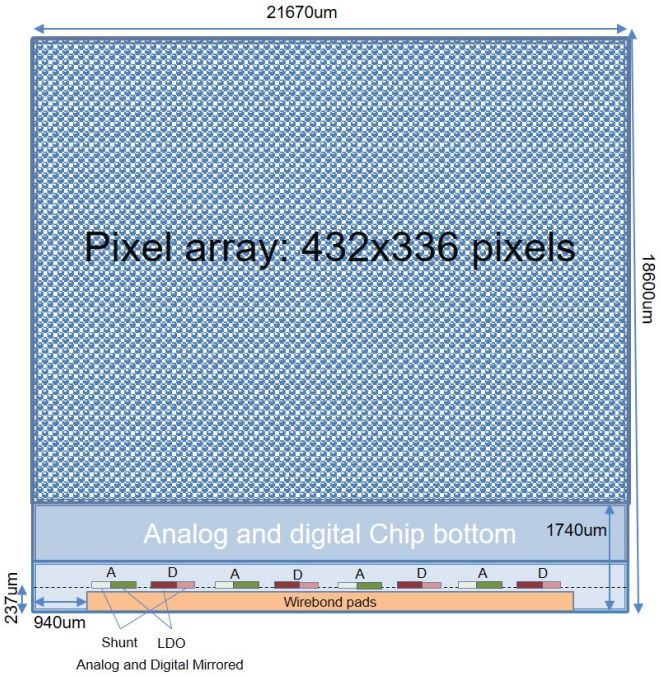


Figure 11.5: Simplified power model of pixel chip in normal operation mode with assumed 60% analog power and 40% digital power and a current headroom of 20%. From [21]

range for certain critical configuration registers. Such constrains can though be removed by a special configuration bit, in case it is really needed to go outside this pre-defined safe range.

A particular issue is the default chip configuration at start up, when applying power. If the chip supports a low power mode (ATLAS chip, see below) the chip will startup with low active power consumed (e.g. 10 - 20%), and most injected power will be consumed/dissipated in the SLDOs, until chip appropriately configured. Depending on how the SLDO is hardware configured (with external resistors) it can be possible to startup with reduced injected current, and then gradually ramp this up as chip gradually gets configured for normal operation. If low power mode is not supported (not enabled in CMS chip) then default configuration is such that the active part of the chip will consume a bit less than nominal power at startup (e.g. 80% of nominal).

If a single chip has a major SLDO failure, and consumes reduced/no power, then the injected serial current will have to be shared among remaining chips in parallel (e.g. 3 chips of 4 or 1 chip of 2). The SLDOs have been designed to have significant extra shunt current capability (2x current). The critical issue is the localised power dissipation in the SLDOs that can eventually overheat. It is also known (chapter 15) that radiation effects will be significantly increased if the SLDO hot spot gets hotter than 0°C for extended time periods (detrimental annealing). If the local cooling can not sustain such additional localized power dissipation in the SLDOs, the whole serial power chain will have to be powered down before remaining chips on the module risks to get damaged. Failure scenarios therefore have to be an integral part of the thermal design of the pixel detector and pixel modules. In case a chip failure makes a power short, then the other chips in parallel on the same module can not be correctly powered (as impedance to low to develop required input voltage). The other pixel modules can in this case still get their required current and the whole serial power chain will just have to run with slightly reduced



Power MOS	Size (um) X,Y	Location (um) X,Y, lower left
Ana_SHUNT_1	87+257; 27	1346; 237
Ana_LDO_1	401; 27	1690; 237
Digi_SHUNT_1	87+257; 27	3113; 237
Digi_LDO_1	401; 27	2709; 237
Ana_SHUNT_2	87+257; 27	6545; 237
Ana_LDO_2	401; 27	6890; 237
Digi_SHUNT_2	87+257; 27	8313; 237
Digi_LDO_2	401; 27	7909; 237
Ana_SHUNT_3	87+257; 27	11745; 237
Ana_LDO_3	401; 27	12089; 237
Digi_SHUNT_3	87+257; 27	13513; 237
Digi_LDO_3	401; 27	13109; 237
Ana_SHUNT_3	87+257; 27	17045; 237
Ana_LDO_3	401; 27	17389; 237
Digi_SHUNT_3	87+257; 27	18813; 237
Digi_LDO_3	401; 27	18409; 237

Figure 11.6: SLDO power transistors locations and sizes for the RD53B-CMS chip (RD53B-ATLAS very similar with slightly narrower chip). From [21]

total voltage. Power estimations of the SLDO in different conditions are given in [21] and also gives relative position of the power dissipating devices in the SLDOs (also shown in figure 11.6).

11.3 Low power mode

In case the pixel chip supports low power mode (ATLAS chip), the default power-on configuration of the chip is such that the chip consumes minimal power with operational control, monitoring and readout interfaces (pixel array powered down). To be capable of operating a chip in low power mode and not being constrained by the hardwired SLDO configuration (by external resistors). It is possible to force the SLDO into a specific low power mode, before applying serial power, by applying an external AC signal (to allow AC coupling). In this dedicated low power mode, the SLDO Voffset is raised to 1.4V, to enable a reduce serial chain current (10-20% of nominal) to generate sufficient input voltage to power the control, monitoring and readout interface (single lane). It is obviously not possible to have the chip fully working in this low power mode and one should not inject the normal operation current as the SLDO will in this case dissipate excessive power (high current combined with increased input voltage). Such a low power mode is available to be capable of testing interface connections (control and readout links) with reduced power dissipation (e.g. with simple forced air cooling), but requires to route the additional low power enable AC signal in the pixel detector.

An alternative low power mode, where the SLDO it-self goes into low power mode when reduced current is injected, was prototyped but not implemented in final chips because of system and radiation reliability worries and insufficient time for thorough testing and qualification.

The CMS chip also has the low power mode available in the SLDO, but must be disabled by wire-bonding and will at power up not start up with low power but at a power slightly lower than nominal (80%).

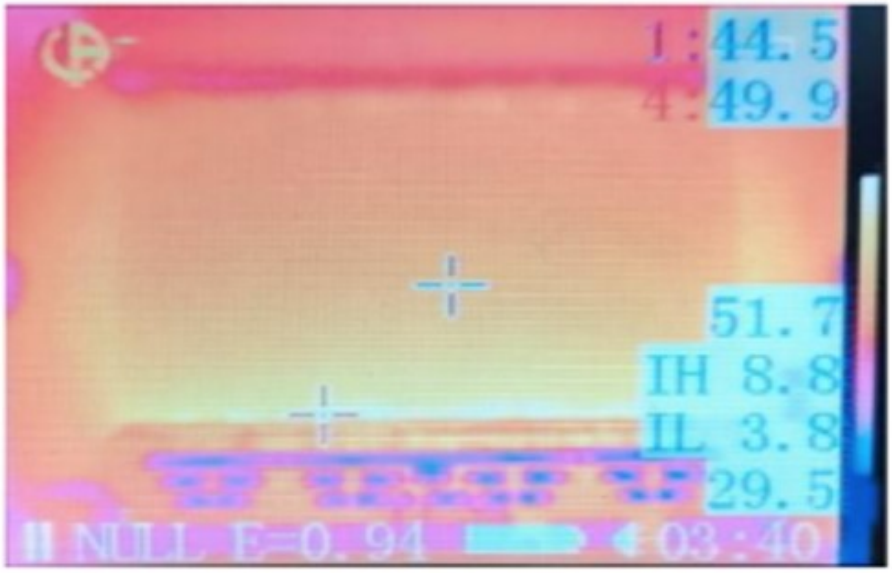


Figure 11.7: Thermal image of RD53A with SLDO hot spot

Chapter 12

Timing issues

The correct function of the pixel chip in a high rate experiment relies on well aligned and stable timing. All timing in the chip is determined by the encoded clock in the 160Mbits/s control link. Transitions on the control link are used to lock the 1.28GHz PLL to the encoded clock in the control stream. A 160MHz clock is derived from this for the control decoder logic that based on specific sync frames generates the 40MHz hit sampling and processing clock. The 1.28GHz clock of the PLL is used for the serialization (1.28Gbits/s and below) on the serial readout links and also to do oversampling on serial links between slave and master chips when using multi-chip data merging.

It is critical for the optimal/reliable operation of the chip that the control link has high quality differential signal transmission with low jitter to assure a stable time reference for the different parts of the chip. The most critical timing will be when using 1.28Gbits/s readout links where jitter and phase stability will be critical at the sub 100ps level. The PLL has jitter filtering capability for high frequency jitter (>˜1MHZ) but will follow low frequency phase drifts/jitter.

Sampling of the discriminated hits in the pixel array is performed with the 40MHz master clock that is distributed to the pixel array with maximum skew of +/-1ns (can with radiation above 100Mrad become as big as +/- 2ns). The master clock can be skewed in time steps of 0.78ns over a full 40MHz clock cycle. At the system level the clocks and real time commands must be distributed and aligned to the LHC collisions by an appropriate timing and control distribution system. When changing the relative phase of the master clock of the pixel chip, the phase of readout links will not change and LPGBT E-links will not need to be phase re-aligned (TO BE VERIFIED).

Absolute phase and stability of sampling clocks and serialization clocks will depend on the timing and control system of the experiment and phase shifts (e.g temperature dependent) in the control path (global timing distribution system, control and DAQ module, optical links, LPGBT, electrical links, etc.). The pixel chip itself also has delay paths dependent on supply voltage, temperature and accumulated radiation effects. The phase drifts of the pixel chip itself is not yet determined but is expected to stay stable at the sub-ns level when temperature is stabilized within a few °C. Supply voltage to the active circuits are stabilized by the on-chip power regulator. Significant phase shifts (tens of ns) must be expected from radiation effects over the life time of the experiment.

Chapter 13

Calibration and threshold adjust

Calibration charge injection into the input of the AFEs is used for threshold tuning across the pixel array and to determine absolute threshold and effective charge conversion gain via the TOT. Charge injection into the AFE is accomplished by applying well defined voltage steps via an injection capacitor directly to the input of the AFE. Tree different voltages: Vcal-high, Vcal-med and analog gnd are routed to the charge injection capacitor in each pixel via digitally controlled analog switches. Vcal-high and Vcal-med are voltages defined by two 12bit voltage DACs and can be checked with the monitoring ADC. The injection switches are controlled by a configurable pulse generator that is triggered by the calibration injection command. The injection of charge can be enabled/disabled per pixel. Injected charge is given by the voltage step when switching between the three voltage levels and the value of the injection capacitance. When doing charge injection by switching between Vcal-med and Vcal-high the effective charge injection is independent of voltage drops in the analogue ground distribution to the pixel array. Switching between the three voltage levels can be done with a single injection sequence enabling two consecutive charge injections to be made. It is also possible to inject different charges in neighbouring pixels for detailed cross-talk studies. The timing relation between the sampling clock and the calibration pulse injection switches is critical and is made to have a max time skew of +/- 1(2)ns (RD53A chip had timing skew issues for the calibration pulse injection). The phase of this can be phase shifted relative to the sampling clock with 0.78ns time steps.

RD53B chips have a dedicated circuit to measure the absolute value of the injection capacitance (as can have chip to chip, wafer to wafer and lot to lot variations) by measuring a current when charging and dis-charging a small array of such capacitors at 10MHz. The effective charge/discharge current can be measured with the on-chip ADC or via an external pin (to get better absolute precision during wafer probing).

Calibration pulse injection can be done on multiple pixels concurrently to speed up threshold tuning, but needs to be done with care because of limited drive capability of the two calibration injection voltages from the DACs and the resistive - capacitive nature of their distribution across the array (bottom to top together with AFE biasing). Such constrains are normally taken care of by appropriate threshold tuning routines in the RD53 test and DAQ systems (must be checked if problematic results are obtained using threshold tuning and calibration). When changing voltage levels via the DACs it is also required to wait some time (order of 1us) for the highly capacitive distribution network to settle with required precision.

It can be noticed that it is critical to have a good and stable threshold tuning to get stable and reliable charge/TOT measurements. When appropriately tuned the charge/TOT dispersion can still be 5-10% across the array, as TOT discharge current is not tuned per pixel. This can if needed be compensated for in off-line tracking/reconstruction.

The calibration injection command can also be used for digital pixel injection which is very useful for functional and system testing and debugging.

Chapter 14

Special test, characterization and debugging options

The RD53B pixel chips contain a set of test and debugging features that can be highly valuable during test beams, chip tests, system tests and debugging. Such debug and test features can eventually also be used during normal system running. Finally there are dedicated production test features (DFT = Design For Test) that allows exhaustive production tests to be made of digital logic during wafer probing.

Direct powering is available for specialized powering tests (see chapter 11),

A digital injection function is available with configurable enable/disable per pixel. This is a subset of the analogue calibration injection scheme described in chapter 13. This is extremely useful for quick and easy functional testing at both pixel chip and system level.

A flexible auto trigger function is available where the chip can self trigger on its own hits with appropriate/configurable trigger latency and a defined window of multiple triggers. This is based on a hit OR network from the pixel array. The hit OR network consists of four hit OR lanes per core column with a fixed mapping such that neighbour pixels are mapped on separate hit OR lanes with an enable per pixel. At the end of the core columns it can be configured which of the four lanes contribute to a global hit OR signal.

The hit ORs are also routed to a high time resolution leading edge and TOT measurement unit in the chip bottom (PTOT=Precision TOT). Here leading edge and TOT are digitized with 1.56ns time resolution and can be triggered for readout via the normal readout path (with specific data formatting). It must be taken into account that the large hit OR networks in the pixel array have significant delays (˜10ns) depending on the location of the actual pixel being hit, so use of this will require timing calibration and appropriate off-line corrections.

During normal event readout, events are identified by trigger/event tags that have been sent to the pixel chip together with the trigger. It will from the RD53B-CMS chip (so not in RD53-ATLAS chip) also be possible to include either a 16bit bunch crossing ID or 8bits BX-ID plus 8bit event count ID.

SEU counting features will be added to the RD53B-CMS chip (so not in RD53B-ATLAS chip) covering SEUs in the global configuration and TMRed pixel configuration bits. The SEU counting of TMR protected pixel configuration bits are done via the hit OR network from the pixels and is only active when appropriately configured and has some intrinsic limitations in counting of multiple bit upsets within the pixel cores. It will also for special SEU immunity verifications be possible to disable any on the triplicated clocks in RD53B-CMS.

A set of dummy global configuration registers will from the RD53B-CMS be made without TMR protection so they can be used as a reference to measure cross-section of unprotected registers (these registers were by mistake also made with TMR protection in the RD53B-ATLAS version).

For basic/easy readout link verification it will in the RD53B-CMS chip be possible to send

repeatedly a constant user defined serial word (without scrambling and Aurora formatting) or a standardized CRC pattern.

The general purpose differential output can be programmed to show status of specific internal signals. The RD53B-CMS will have an extended set of signals related to data merging as it has been seen to be particularly delicate to debug in test systems.

Finally for specialized tests (e.g. SEU tests) it is possible to bypass the PLL and supply directly required clocks to the chip.

Chapter 15

Radiation effects

The required radiation tolerance to extremely high radiation dose and immunity to Single Event Effects (SEE) is one of the major challenges to make the RD53 pixel chips. It requires all IP blocks and all digital logic to be designed and tested specifically for this. This increases required development and testing time and resources by a factor 2-4 compared to a similar chip where radiation issues are not of concern.

15.1 TID

In the inner pixel layers the accumulated Total Ionising Dose (TID) over 10 years is estimated to be of the order of 1.2Grad. This is a factor 10 higher than any previous ASIC used in HEP (or any other community covering space, military, nuclear power, etc.). Deep submicron CMOS technologies can in general be made to stand 10 - 100Mrad when choosing appropriate technology and using certain design tricks (enclosed layout for analogue transistors, not using minimum size transistors for digital, account for delay degradation, etc.). Above the 100MRad level, severe radiation degradation is observed with complex dependability on technology, dose rates, biasing, temperature and annealing. An extensive radiation evaluation program has been made in the first years of RD53 to characterize this for the chosen 65nm technology and determine appropriate design precautions. Severe radiation damage has been seen above 100Mrad when chips are under bias (powered) and at room or higher temperatures, incompatible to make complex mixed signal pixel chips for the radiation levels required for the phase 2 pixel upgrades. In particular small/digital PMOS transistors have severe degradation of current capability. It has been determined (and confirmed) that an acceptable (but still big) radiation degradation can be constrained when chips are cooled while being irradiated (pixel detectors will be cooled to -10 - -30 °C) and not biased/powered when not cooled. If highly irradiated chips are powered at higher temperatures than 0°C then a large detrimental annealing will gradually occur and chips may become non functional. The time constant of this detrimental annealing at room temperature is of the order of days/weeks with a strong temperature dependency. This detrimental annealing only occurs when chips are powered as the electrical fields within the transistors will then make radiation induced trapped charges move into critical active parts of the MOS transistors (spacers and gate oxide) [12]. Approximative transistor radiation models have been developed in RD53 that have been extensively used for the design and optimization of all building blocks (Extended radiation models now available from CERN ASIC services as part of their ASIC design kits). All analogue IP blocks have been designed specifically for radiation tolerance, using relatively big transistors, and have all been extensively radiation tested/qualified.

Digital logic, made with small and compact transistors, are particularly sensitive to radiation effects above the 100Mrad level. A pixel chip for required rates and complexity requires a very large number of high density logic and data buffers (total chip is ~1/2 billion transistors, equivalent to ~100million logic gates) so it is not possible to use logic gates with large transistors as it

will simply not fit. An appropriate digital gate library has been chosen as a compromise between fitting required logic and acceptable radiation damage. Certain gate types are excluded from use as having problematic radiation behaviour. Even maintaining chips cold during irradiation, logic gate delays can increase by as much as 200%. This has been taken into account in the digital design flow using appropriate gate level radiation timing models.

Low dose rate radiation effects are a particular delicate point as the 65nm technology have been seen to have significantly increased radiation damage at low dose rates. Extensive low dose rate tests have been performed on dedicated radiation test chips and on the RD53A prototype. These have confirmed that low dose rate irradiation give significantly increased (factor 2-4) degradation of digital gates, compared to high dose rate tests. Projections from low dose rate tests indicate that RD53 pixel chips can be designed to handle 500Mrad and can most likely also remain functional up to 1Grad. Fortunately digital transistors during normal operation are under biasing conditions where induced radiation damage is significantly less than under worst case biasing conditions.

The 65nm technology has (fortunately, as other technologies have major issues with this) been seen to only have very small radiation induced increase of leakage currents, so the pixel chip is not expected to have a significant power consumption increase. Margins for some possible radiation induced power consumption increase (10 - 20%) should though be taken into account for powering and cooling systems.

Short term radiation effects during a typical physics run at LHC (24hours) can also be expected to affect the behaviour (timing alignment, threshold adjust, etc.) of a well calibrated pixel detector system. At high luminosity running an incremental irradiation of up to 1Mrad can be expected for the inner-most layers during a run (much less for outer layers). It has been seen that the effect of incremental irradiation increase is the largest at low irradiation levels, so it must be anticipated that regular calibration and threshold adjusts can be needed during initial runs (if at full luminosity).

It must also be mentioned that only limited experience is currently available on the issue of possible differences of radiation tolerance between different wafer lots and from different TSMC fabrication facilities. To the extent possible it must be assured that all prototype and production wafers are coming from the same fab. It is known that each fab have small differences that can have a major impact on radiation tolerance (that is not guaranteed by the manufacturer).

Extensive radiation testing of the RD53A chip and test chips can be found in [12], [14], [13] and further test results will become available as RD53B generation chips have been radiation tested.

15.2 Single event effects: SEL, SEU, SET

Single Event Effects (SEE) are the other major challenge making a complex chip work reliably in an extremely hostile radiation environment.

Single Event Latchup (SEL) is prevented by having a high density of substrate contacts in all IP blocks and digital gates. A gate library has been chosen where each gate has substrate and well contacts to assure localized low resistance contact to the local substrate and wells (in particular as extensively using triple well isolation for minimal noise coupling from digital logic to analogue front-ends). No SELs have been seen in heavy ion test beams of test chips, the RD53A demonstrator chip and the RD53B-ATLAS chip.

Particle induced Single Event Upsets (SEU) in digital storage elements (flip-flops, memories) are for the inner pixel layer estimated to happen at a rate of 100 upsets per chip per second. Single Event Transients (SET) in logic signals and distribution networks (e.g. clocks, reset, etc.) will also generate upsets/glitches in the digital logic. It requires a well defined strategy together with appropriate design techniques to assure sufficiently reliable working of 10K pixel chips in such a hostile environment for many hours. Critical information and signals in the chip are

protected with Triple Modular Redundancy (TMR) to make it immune to upsets in critical parts of the logic. Using TMR comes with major area and power consumption overheads (200%) that are not compatible with the required small pixel size, data buffering requirements and power dissipation constraints. Only critical information is protected with TMR (configuration, state machines, buffer pointers, trigger tables and critical event information). Oher information will be allowed to have SEU bit flips as long as it does not seriously affect the function of the chip. Occasionally corruption of hits and event data will be allowed to occur as long as it only rarely affects a single (or a few) hits/event(s) and the pixel chip self-recovers its correct function.

Critical chip configuration registers have been made with TMR protection. All global configuration registers have full TMR protection with auto correction. (a few dummy registers on purpose without TMR in the RD53B-CMS chip). For the large number of individual pixel configuration registers it has not been possible to fit full TMR protection. Pixel configuration bits that affects the operation of the pixel (threshold adjust and enable) are protected with TMR, but without auto correction. This implies that SEU bit flips in the 3 TMR bits are allowed to accumulate over time and if two bit flips occur among the three, the configuration bit will become corrupted. It must be mentioned that a corruption of a pixel configuration bit will only have limited effects. A pixel can get disabled/enabled or a threshold adjust bit can get changed, resulting in a possibly noisy or inefficient pixel. The statistical risk of having multiple bit flips in the same three TMR bits is extremely low (even in our hostile radiation environment). Pixel configuration registers can/should be continuously re-configured via the pixel chip control link. Continuous reconfiguration of all pixel configuration registers can be done at a rate up to ~10 times per second, as the control link has a significant bandwidth (but DAQ/control module must be capable of doing this). Precautions have also been made to prevent configuration registers to become accidentally reset to default configuration values by SET glitches. Configuration registers uses synchronous reset that have a much smaller risk of being upset by a SET glitch than asynchronous reset. It is advised to finish a complete pixel configuration reload with a write to a non existing pixel such that an SET glitch on the critical load signal can not affect any pixel configuration.

As mentioned, hit data storage in pixel regions, for the latency buffers, and further data buffers in the DCB do not have triplicated data storage. Pixel region logic (incl local clock gating) and latency buffering have been verified with extensive SEU injection simulations that SEUs here only provokes low rate of possible accidental hits or loss of hits. Exceptionally the readout of an event from the pixel array to the DCB can become corrupted, but processing of following events will be processed normally. The same is the case for event processing and buffering in the DCB. All buffer pointers and state-machines are triplicated together with critical event information. SEU injection simulations have been used to verified that SEUs in non protected storage elements will only cause partial corruption of an event and will not prevent following events to be processed correctly.

It is known from practical experience that TMR will in practice not give 100% protection but will reduce the effective sensitivity by 2-3 orders of magnitude (depending on delicate implementation issues). A dedicated fast clear command is therefore available that will initialize all state machines, data buffers, etc. such that the pixel chip can immediately recover normal functionality in case it has exceptionally gotten stuck. DAQ systems in the final experiments must be capable of using this when needed and be capable of quickly re-integrating such a pixel chip in the normal data flow.

If a fast clear command does not recover normal operation of a pixel chip, the control link can be used to force pixel chips on a given control link (typically 2-4 chips on a pixel module sharing a control link) into a special reset state, from where the chips will have to be re-configured and re-initialized (e.g. PLL) to become operational again. This is equivalent to a soft power-on reset, without having to go through a power cycle of a serial power chain.

TMR protection of critical registers and logic can be made in different fashions. In the RD53B

chips TMR protection is implemented by triplicated storage elements, but no triplication of related logic. TMR registers are driven by triplicated clocks with time skews of ~200ps, thereby being capable of filtering SET logic glitches shorter than this (only one TMR bit will be affected and this will be corrected for by majority voting among the three). Extensive SEU/SET tests will be made of the RD53B chips to confirm appropriate SEU/SET protection of critical functions in the chip. In the RD53B-ATLAS chip the triplicated skewed clocks can not be enabled/disabled which makes triplication verification difficult. Enabling/disabling of individual triplicated clocks will become available in the RD53B-CMS version to allow verification of triplication circuitry without making dedicated tests for this in ion and proton beams.

The RD53B-CMS chip will have SEU counting capability for TMR protected global configuration registers and a limited number of pixel config bits, using the hit OR network.

Estimates of different classes of SEU/SET failures have been made based on measured SEU cross-sections [16]. These estimates will be updated when more extensive SEU tests have been made under realistic operation conditions (running with high hit and trigger rates). It is critical that DAQ systems, receiving data from pixel chips located in a radiation environment with an estimated flux of high Energy Hadrons (HEH) of up to 1GHz/cm^2 in a inner layer pixel detector, can efficiently handle this locally. It must be well prepared to receive SEU induced spurious hits, corrupted hits and events. Relative spurious and corrupted hits can be estimated to be at the very low level of 10^{-9}. Occasional corrupted event data will occur and this can be estimated to be at the level of 10^{-8}, where event data may not be appropriately decoded because of the highly compressed binary tree hit encoding. In this absolute worst case radiation scenario it has been seen in dedicated SEU tests that 1.28Gbits/s links can loose link synchronization every minute, caused by short phase jumps induced in the critical on-chip PLL driving the high speed link serializers. The DAQ system must be capable of re-synchronizing to the link from each chip when needed. Depending on the time (1 - 100 events) needed to obtain fast link re-synchronization, it can be estimated to give a worst case data loss at the level of 10^{-5} - 10^{-7} for inner layer chips, and a factor 100 less for outer layers. Final ATLAS and CMS pixel systems, with LPGBT optical links from pixel chips to the DAQ system, may have additional link synchronization issues. The trigger tag based event identification should ease such pixel chip and link resynchronization in a reliable fashion, as event identification is not based on local chip counters affected by SEUs. If a pixel chip has been seen to not respond to triggers in a reliable fashion the fast clear command must be issued ASAP (should be FPGA firmware driven to be fast) to recover correct participation in data taking. Continuous re-configuration is also required to assure that pixel configuration and global chip configuration are not corrupted. It is highly recommended to do continuous reconfiguration at a rate of 0.1 - 10Hz for the extreme radiation environment in a HL-LHC pixel detector.

Chapter 16

Typical behaviour and performance in HL-LHC environment

The behaviour and performance of the RD53 pixel chip in a given experiment depends on the environment (hit rates), system parameters (trigger rate, trigger latency, number and speed of readout links, data merging between chips, etc.) and pixel chip settings (threshold, TOT pulse width, binary or TOT readout mode, etc.). It is a relatively complex task to predict well the performance and effective limitations of using a RD53 pixel chip in a given system. This is particularly the case for the high hit and trigger rates required in the HL-LHC upgrades combined with minimizing the number of readout links, to have an acceptable material budget. Constraints and limitations can in general be grouped into four major aspects:

- 1: Analog front-end charge capture and discrimination. Depends on silicon sensor and AFE settings.
- 2: TOT dead-time losses. Depends on hit rate and AFE TOT settings.
- 3: Trigger latency buffer losses. Depends on hit rate and trigger latency.

  (3B: In case using two level trigger: Effective L0 latency hit buffer size will effectively be slightly smaller as a small part of pixel region latency buffer will also be used for hit storage during L1 latency).
- 4: Readout delay/latency and possible event truncations. Depends on hit rate, trigger rate and available readout bandwidth.

These 4 aspects are to a large extent independent of each other, but can under certain conditions have interdependencies. If the analog charge capture and discrimination (threshold and threshold tuning per pixel) is not set appropriately, hits will be missed or there may be an excessive amount of noise hits, which can seriously affect the following steps. If a large fraction of hits are lost because of TOT dead-time then there will obviously be less hits to store during the trigger latency so the losses from latency hit buffering will decrease slightly when TOT dead-time losses increase, at given hit and trigger rates. If the readout is saturated, hits/events will pileup in the readout FIFOs. For extreme cases this will eventually back-propagate to the process of extracting hits from the buffers in the pixel regions. The RD53B chip have programmable options controlling when hit information will be discarded, to prevent the chip to end up in an oversaturated state (hits are discarded but events are to the extent possible maintained to keep event synchronization at the system level).

The effective losses and typical performance characteristics are illustrated with two examples that have been simulated with Monte Carlo hit data for 150um thick, 25x100um^2 pixel sensor covering 3 cases for typical use in HL-LHC with a baseline trigger rate of 750kHz and 12.8us latency (CMS case but results can be scaled to ATLAS and other conditions).

- A: Inner layer module with 3.4GHz hit rate per cm^2 (85kHz per pixel) using 3 readout links at 1.28Gbits/s per chip. For both a centre module (small hit cluster sizes) and a module at the end of the inner barrel (slightly larger hit cluster sizes)
- B: Outer layer module with 270MHz hit rate per cm^2 (7kHz per pixel) merging readout data from 4 pixel chips into a single 1.28Gbits/s readout link.

16.1 TOT Dead time losses

In the RD53 pixel chips, the collected charge of individual hits are digitized with a TOT conversion scheme. Collected charge is converted into a pulse length by the AFE that is then measured with a digital counter (40MHz or 80MHz). The extension in time of the analogue hit signal implies that a new hit arriving on the same pixel during this period, is lost (in practice charge from this is added to first hit and affects TOT charge measurement). Synchronization and sampling of the TOT pulse has a small additional digital dead time of 1 -2 clock cycles depending on when the hit signal occurs relative to the 40MHz reference sampling clock. The charge distribution/spectrum on each pixel typically has a modified Landau shape (as charge normally shared among 1-4 pixels for each traversing particle) as shown in figure 16.2.

At the relatively modest hit rates per pixel (lower than 100kHz), the effective dead-time losses can be estimated by the following simple linear equation: Dead-time-loss = ˜Hit-rate x Average-TOT-time. This is under the assumption that hits arrive randomly (exponential distribution of time between hits) and does not include effects from the LHC beam structure with regular injection and abort gaps (in general only ˜80% of 25ns periods have collisions in the LHC experiments). It is also under the condition that the TOT spectrum has a simple flat charge distribution, as illustrated in figure 16.1, which in general is not the case for deposited charge per pixel. In case the analog front-end has a nonlinear charge to TOT mapping (e.g. differential AFE in the ATLAS chip) the effective dead-time will also be affected. A realistic TOT pulse width (and TOT dead-time) is illustrated in figure 16.2 for typical charge deposition in pixel sensors (Monte Carlo simulation data).

Inner pixel layers of a HL-LHC detector can have hit rates as high as 3(4)GHz/cm^2 (notice that each particle typically generates 1-4 pixel hits) giving an equivalent hit rate per pixel of 75(100)kHz for 50x50um^2 (or 25x100um^2) pixels. Having an average effective TOT pulse width of 8x25ns = 200ns results in a dead-time loss estimation of 75KHz * 200ns = ˜1.5%. For the specific case of a flat TOT spectrum (as shown in 16.1) with an average TOT of 200ns and hit rate of 3.5GHz/cm^2 the estimated hit loss of 87.5KHz * 200ns = 1.75% fits well simulated losses of 1.83% as shown in figure **??**.

This is for most applications too high hit loss, so for inner high rate layers the AFE will typically be configured to have a shorter average TOT pulse width of 4x25ns=100ns (or 50ns) to limit dead-time losses to be below the acceptable ˜1% level. This reduces effective charge resolution which can be compensated for by using the 80MHz TOT counting mode. It has been seen that limiting the charge resolution, for the inner layer only, normally does not result in a significant track resolution degradation. For outer pixel layers, with much lower hit rates, the dead time losses are normally negligible. It should be noted that the pixel region concept, used for hit buffering, has been implemented such that effective dead-time is per individual pixel (and not per pixel region that would have allowed simplifications in the pixel logic and TOT charge digitization).

The result of simulating analog TOT dead-time losses with Monte Carlo hit data for the high and low rate cases are shown in table 16.1 for different charge to TOT conversion scales of 12ke (short) and 5.2ke (long) charge for 8 counts at 80MHz (100ns). As expected, analogue TOT dead time losses increase linearly with average TOT pulse width (and hit frequency), so for high rate applications (e.g. inner layer) it is highly advantageous to configure the AFE to

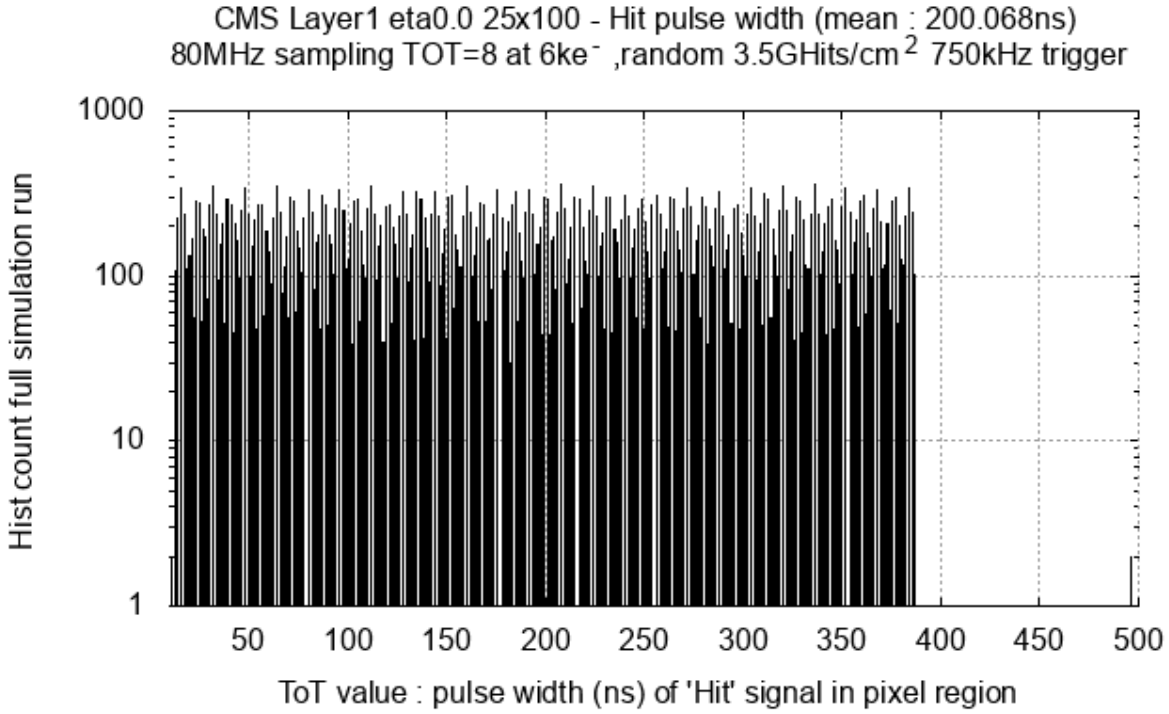


Figure 16.1: Generated flat TOT spectrum at 3.5GHz/cm^2. Large average TOT generating excessive TOT dead time losses of 1.8% as shown in figure **??**.

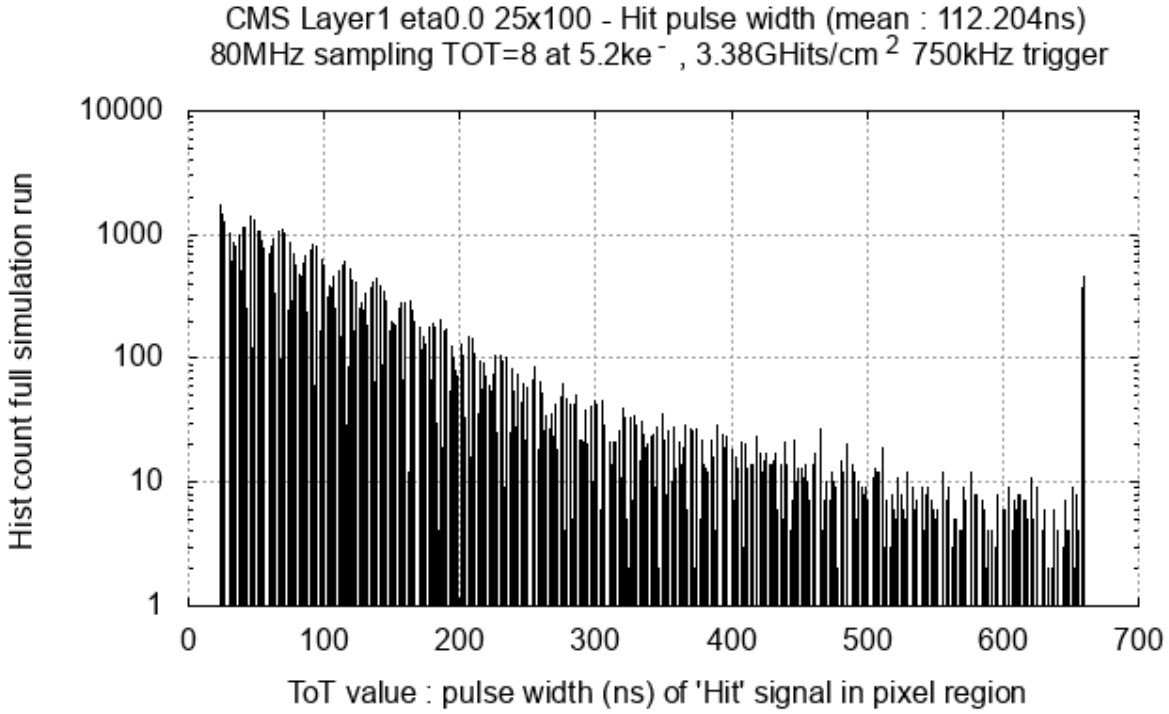


Figure 16.2: Typical Monte Carlo TOT spectrum from charge deposition in pixel silicon sensor at 3.4GHz/cm^2. Long/overflow accumulated in last bin.

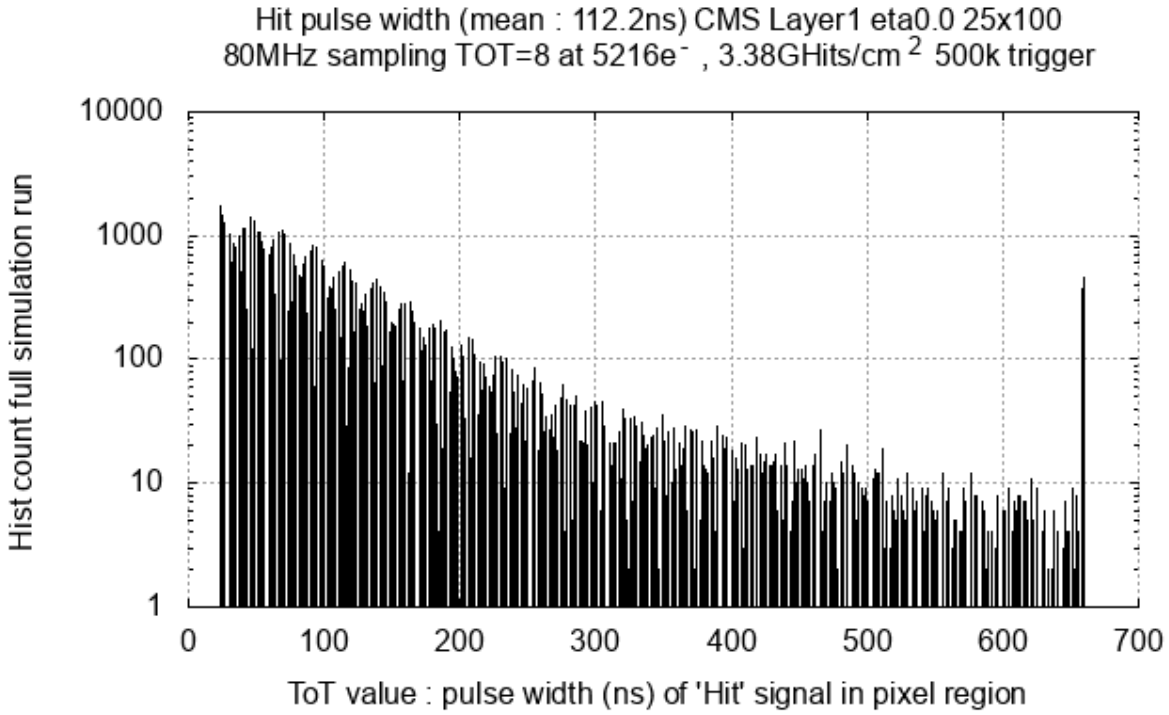


Figure 16.3: TOT spectrum for long TOT (5.2ke = 8x12.5ns) at high hit rate of 3.4GHz/cm^2.

generate short TOT pulses (at cost of reduced charge resolution).

16.2 Latency buffer losses

Storing hits during the L0 trigger latency is a significant challenge for the large number of small pixels (160k pixels per chip), the high hit rates, and the extended trigger latency required for the HL-LHC experiment upgrades. Only a few hit buffer locations can fit in the area of each pixel. Grouping 4 pixels into pixel regions with shared hit buffers, containing hit time tag and TOT values, has allowed to fit just enough buffering (8) for HL-LHC pixel detectors. It is relatively straight forward to calculate the average number of hits that needs to be stored in the small pixel region buffers (Hit-rate x Trigger-latency). It is though non-trivial to estimate the effective hit losses for a given hit buffer size. The strong time correlation between hits in neighbour pixels, from pixel clusters from same particle, is critical for the good efficiency of the shared buffers, but difficult to include in estimates for effective hit buffer losses.

A basic cross check estimate can be made to determine if potential losses from this will be acceptable. For the simple case of generated single isolated hits (TOT shown in figure 16.1) the average buffer occupancy for the 4 pixel region can be estimated to be: 4 pixels/region x 87.5KHz/pixel x 12.8us = 4.5. A bit less than half of the buffer locations (3.5) are therefore available for statistical variations. In this case the effective hit loss from the pixel region buffer of 8 is of the order of 0.9%, as shown in the simulation summary table 16.1. As can be seen from the same summary table, the latency buffer losses for same hit rate is reduced to 0.2%, when having clustered Monte Carlo hit data. From this it can be summarized that for typical pixel sensor hit data the latency buffer loss is of the order of 0.1 - 0.2% when the average buffer filling is half, from simple estimate based on raw hit rate. It should here be mentioned that this is without taking into account that only 80% of LHC machine bunches are filled, so latency buffer losses in HL-LHC will in practice be below 0.1% for the inner most layers.

Simulated hit buffer losses for the different cases are shown in figure **??** together with the

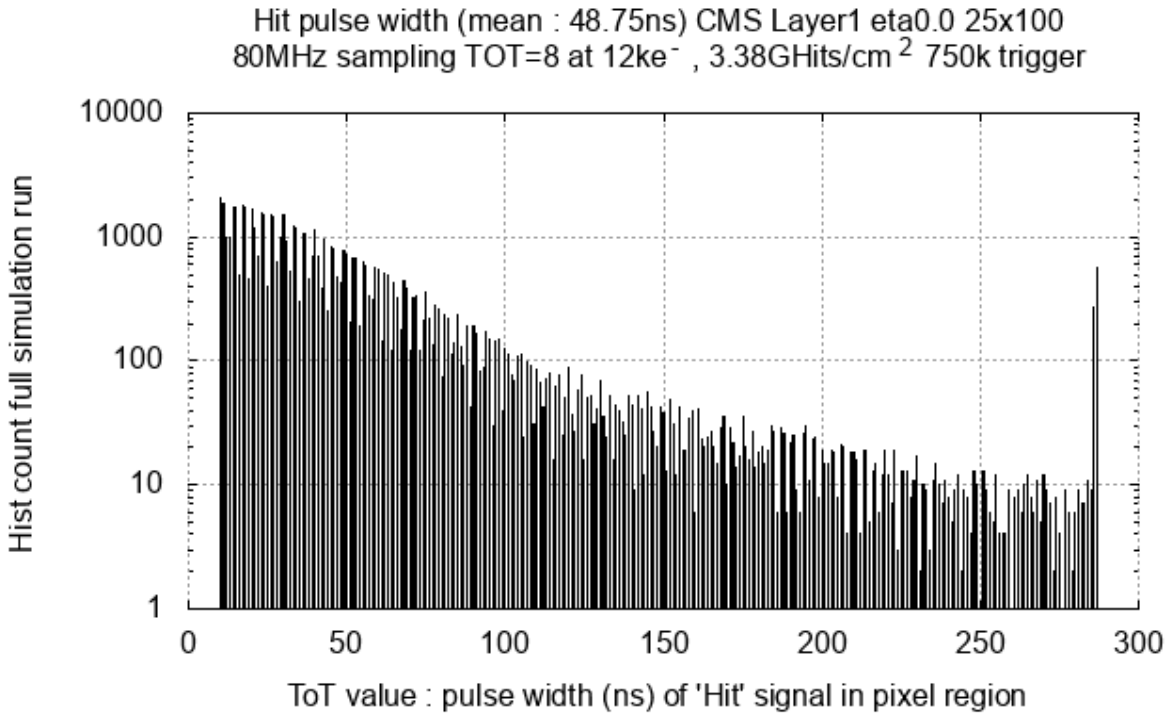


Figure 16.4: TOT spectrum for short TOT (12ke = 8x12.5ns) at high hit rate of 3.4GHz/cm^2.

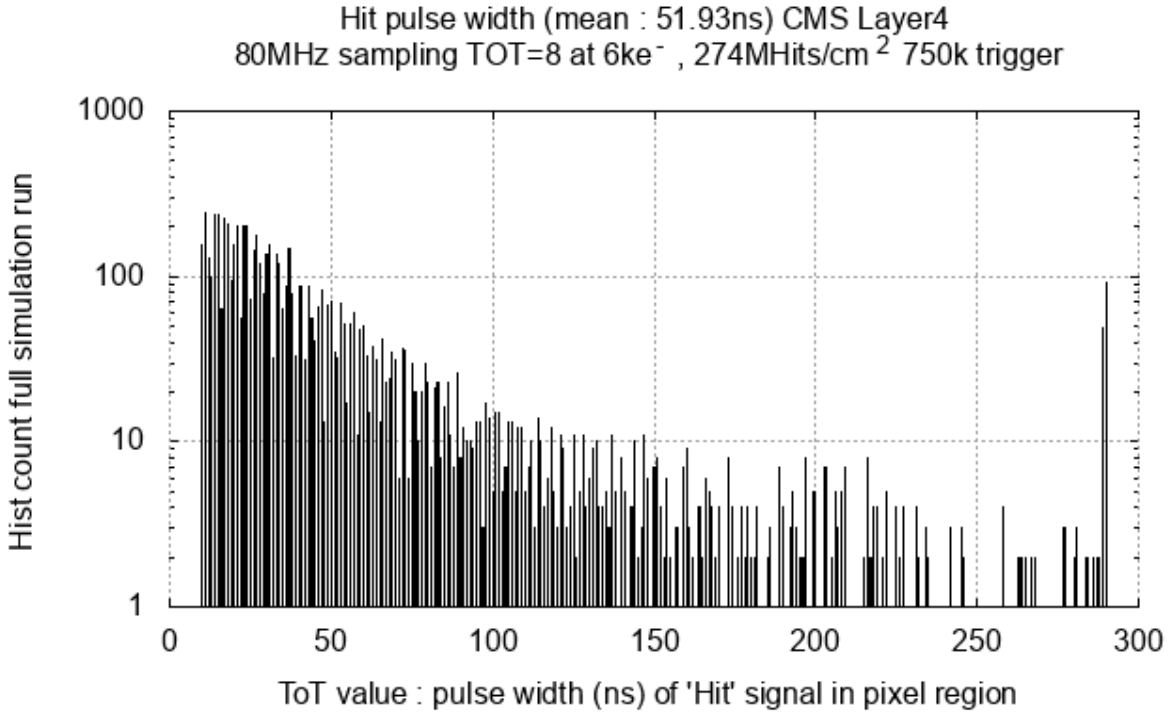


Figure 16.5: TOTspectrum for long TOT (5.2ke = 8x12.5ns) at low hit rate of 270MHz/cm^2.

Position Trigger rate Readout events	Readout latency		Rate per lane (Gbps)	Latency buffer Occupancy ToT (Fig ref)	Hit Statistics		
	US	% Events	%age of 1.28 Gbps	% of occupied location	% Hit loss ratio	Simulated Hits / event	Simulated hits / cm² Total hits (sub Matrix)
BARLYR1_CENTER(eta 0.0) TRG: 750 KHz MC Events: 935 Monte Carlo TOT, ?ns Avg	11.2 12.8 16.3	98.2 99.0 100	0: 0.953168 (74.46%) 1: 0.937941 (73.27%) 2: 0.932935 (72.88%)	**8**:1.65%, **7**:3.39%, **6**: 7.69%, **5**:14.48%, **4**: %, **3**: % **2**: %, **1**: %	Total classified hit : 0.8261 % Dead-time : 0.6314 % PR buffer overflow : 0.1947 %	307.01	3384203483.245150 ~ **3.38 G/cm2**
BARLYR1_CENTER(eta 0.0) TRG: 750 KHz Events: 905 Flat TOT, 200ns Avg.	12.4 13.0 15.1	98.2 99.0 100	0: 1.005304 (78.53%) 1: 0.992354 (77.52%) 2: 0.988932 (77.26%)	**Not recorded** **(Fig ref: 16.1)**	Total classified hit : 2.7140 % Dead-time : 1.8340 % PR buffer overflow : 0.880 %	319.35	3520189594.356261 ~ **3.52 G/cm2**
BARLYR1_CENTER(eta 0.0) TRG: 750 KHz MC Events: 935 Monte Carlo TOT, 112ns Avg	11.2 12.8 16.3	98.1 99.0 100	0: 0.954433 (74.56%) 1: 0.939364 (73.38%) 2: 0.934094 (72.97%)	**Not recorded** **(Fig ref: 16.2)** **TOT=8 at 5216e⁻**	Total classified hit : 0.7347 % Dead-time : 0.5349 % PR buffer overflow : 0.1997 %	307.01	3384203483.24 ~ **3.38 G/cm2**
BARLYR1_CENTER(eta 0.0) TRG: 750 KHz MC Events: 935 Monte Carlo TOT, 48ns Avg	11.3 12.9 16.3	98.0 99.0 100	0: 0.957345 (74.79%) 1: 0.942540 (73.63%) 2: 0.937482 (73.24%)	**Not recorded** **(Fig ref: 16.4)** **TOT=8 at 12ke⁻**	Total classified hit : 0.4770 % Dead-time : 0.2625 % PR buffer overflow : 0.2145 %	307.01	3384203483.24 ~ **3.38 G/cm2**
BARLYR1_EDGES (eta 2.5) TRG: 750 KHz MC Events: 935 Monte Carlo TOT, 112ns Avg	9.3 10.2 12.2	98.0 99.1 100	0: 0.858281 (73.27%) 1: 0.834578 (73.27%) 2: 0.823359 (73.27%)	**8**: (0.70%) **7**: (1.79%) **6**: (4.94%) **5**: (11.17%)	Total classified hit : 0.7658 % Dead-time : 0.6927 % PR buffer overflow : 0.0731 %	296.72	3270737213.403880 ~ **3.27 G/cm2**
BARLYR4_CENTER(eta 0.0) TRG: 750 KHz MC Events: 935 Monte Carlo TOT, 52ns Avg	5.1 5.8 7.0	98.1 99.0 100	**0**: 0.257292 (20.10%)	**4**: (0.17%) **2**: (16.46%) **1**: (75.65%) **(Fig ref: 16.5)** **TOT=8 at 6ke⁻**	Total classified hit : 0. 0474 % Dead-time : 0. 0474 % PR buffer overflow : 0.0 %	24.82	273590608: ~ **273.59 M/cm2**

Table 16.1: Simulation summary table with used readout bandwidth, readout latency, Latency buffer occupancy, TOT deadtime losses and Pixel Region (PR) buffer losses. (Latency buffer occupancy numbers to be added when available)

TOT dead time losses. It can be seen that in general the hit buffer losses are smaller than the TOT dead-time losses and to a large extent independent of the used TOT pulse width. At increasing dead time losses there is a small, but visible, reduction in buffer losses as dead-time losses helps to remove close coming hits in bursts, that poses the biggest challenge to the small pixel region hit buffers.

For the special case of using two level triggering (L0 and L1), the pixel region hit buffer will have to buffer hits during both the L0 trigger and the L1 trigger latencies. In a well designed trigger architecture only a small fraction should be buffered during the L1 trigger. For the case where 10% of the bunch collisions will have to be stored during the L1 trigger (implies L0 trigger rate of 40MHz/10 = 4MHz) and a L1 latency twice as long as L0 latency, it can be estimated that the following fraction of the buffer will be needed for L1 hit storage: L0-rate/BX-rate * (L1-latency/L0-latency). For the given example, this becomes 4MHz/40MHz * 2 = 20%. For a pixel region buffer size of 8, 1.6 buffer locations will be used for L1 storage and 6.4 buffer locations for L0 storage. In reality the sharing of the Latency buffer between L0 and L1 hit storing is done dynamically according to instant needs. Differently phrased, the effective L0 buffer size is reduced by ~20%, to also accommodate L1 hit storage in the given example. Detailed pixel chip simulations, with appropriate hits and triggers, is required to determine effective hit losses dynamically.

16.3 Readout latency and buffering

It is obviously vital to have sufficient readout bandwidth to transfer triggered hit information (event data) to the DAQ system of the experiment. If only half of the available readout bandwidth is used in average, one can be confident that the readout will work efficiently, and maximum readout latency of events will be determined by just a few (1-2) events queuing up in the output buffers of the RD53 pixel chip. This has some dependencies on the statistical properties of both hits and triggers (e.g. appearance of particular bursts of hits and triggers).

For applications where the number of readout links (and/or readout speed) must be min-

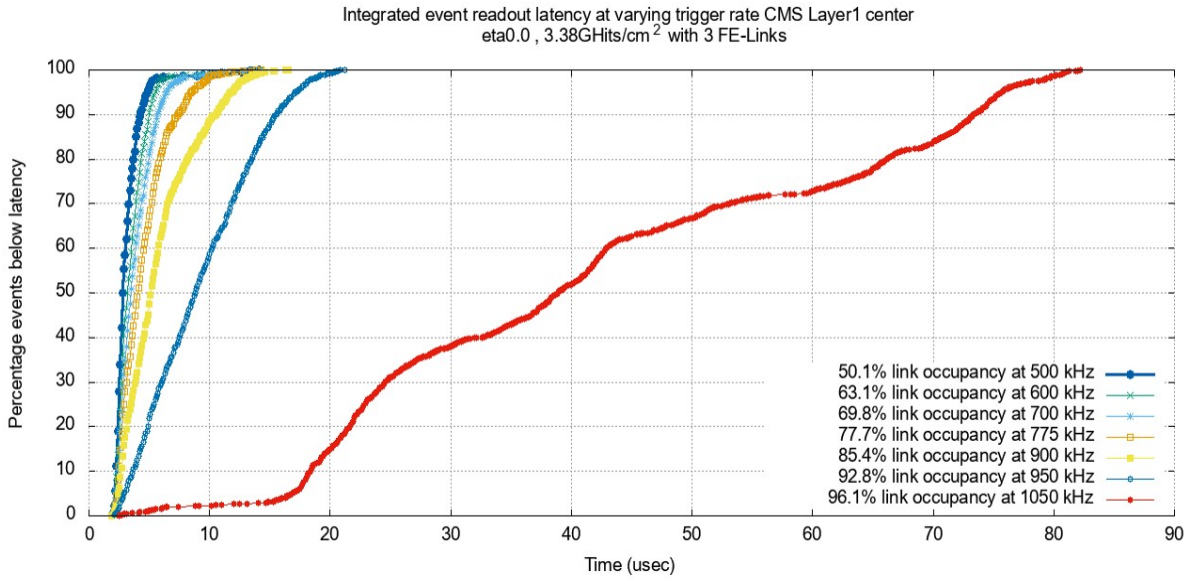


Figure 16.6: Readout latency at high hit rate: 3.4GHz/cm^2 for single chip with 3 readout links at 500 - 1050kHz trigger rate

imized, and/or readout latency is a critical system parameter, it is required to simulate this with appropriate Monte Carlo hit data and realistic trigger sequences (random triggers, bursts of triggers, implementation of central trigger burst filters, trigger bias to select large or small events, etc.). In many pixel detectors the number and material budget of readout links must be minimized , even for outer layers where hit rates are normally 10-100 times lower than the inner-most layer. For an inner layer pixel detector (Case A) 3 out of 4 links is used for readout. For an outer pixel layer module, event data from 4 chips share a single readout link (Case B). Approaching the available readout bandwidth is done by running at increasing trigger rate until reaching excessive data pileup in the readout buffers of the pixel chip.

Simulation results are shown in figures 16.6, 16.7, 16.8, 16.9, showing readout latency and average readout link utilization as function of trigger rate. In both cases the readout works smooth and efficiently until the average readout bandwidth reaches 85-90% of available bandwidth. Above this level, it becomes clearly visible in the readout latency that event data piles up in the readout buffers in the pixel chip and it can not recover from this. Increasing trigger rates further, the readout buffers will get full and the pixel chip will be forced to truncate event data according to its configured options for this.

To make basic estimates of required readout bandwidth, it can be mentioned that for typical Monte Carlo hit data used in these simulations (25x100um^2 pixels, 150um thickness) the implemented binary tree pixel address encoding scheme uses of the order of 12.3bits per hit when having 4 bit TOT per pixel (including Aurora formatting overhead). For cases where hits are part of large clusters (e.g. end of barrel) the average number of bits per hit can decrease to 11-12bits/hit. In case the chip is configured to only read out binary hit data, with binary tree pixel address encoding, the required number of bits per hit is decreased by 4 bits (30% reduction). This is for the case where each event contains hundreds of hits and the event framing overhead (and end of event padding) is therefore insignificant. For cases where only small number of hits are contained in each event (or event stream), event framing overhead, and end of event/stream padding, must be taken into account. For the outer layer case with smaller number of hits per event (˜25 per chip) and overhead related to data merging between 4 chips (2 bit per 66bit Aurora frame to indicate pixel chip ID) the average number of bits per hit increases to ˜14.5, including 4bit TOT.

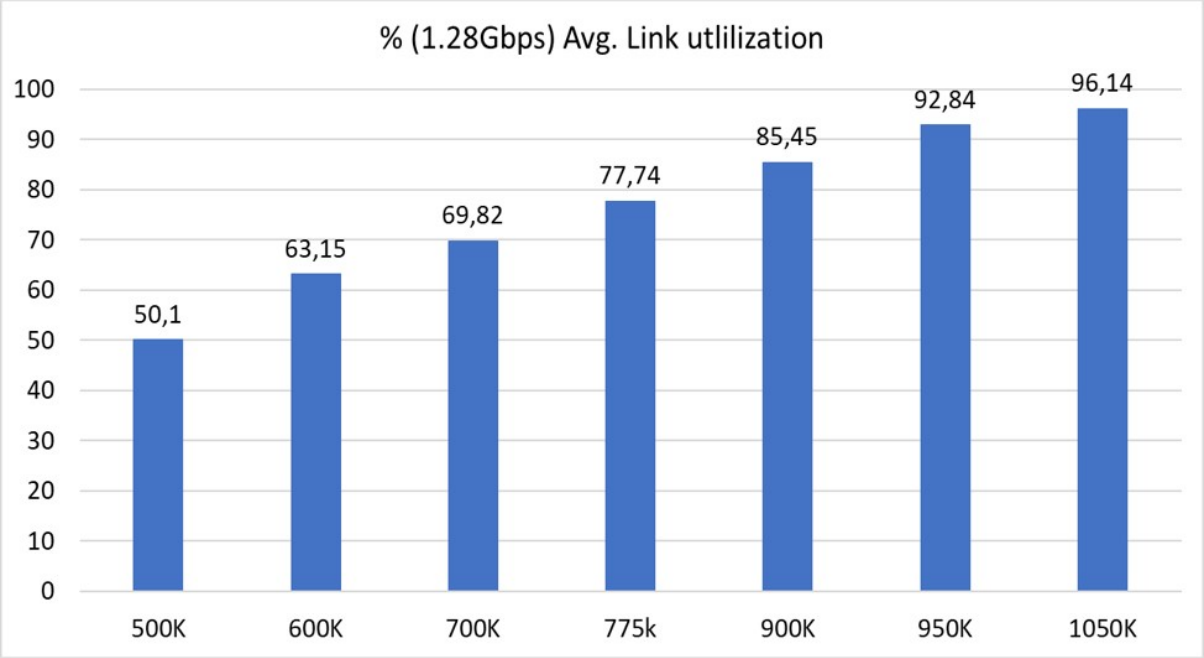


Figure 16.7: Readout link utilization at high hit rate: 3.4GHz/cm^2 for single chip with 3 readout links at 500 - 1050kHz trigger rate

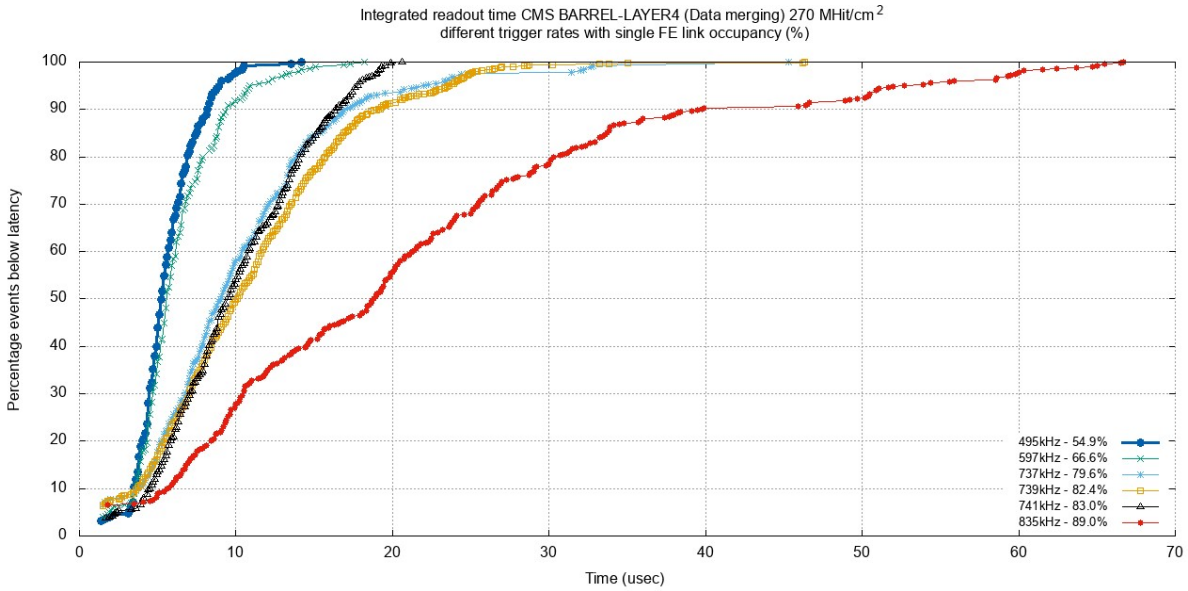


Figure 16.8: Readout latency at low hit rate: 270MHz/cm^2 for 4 chips with data merging into 1 readout link at 495 - 835kHz trigger rate

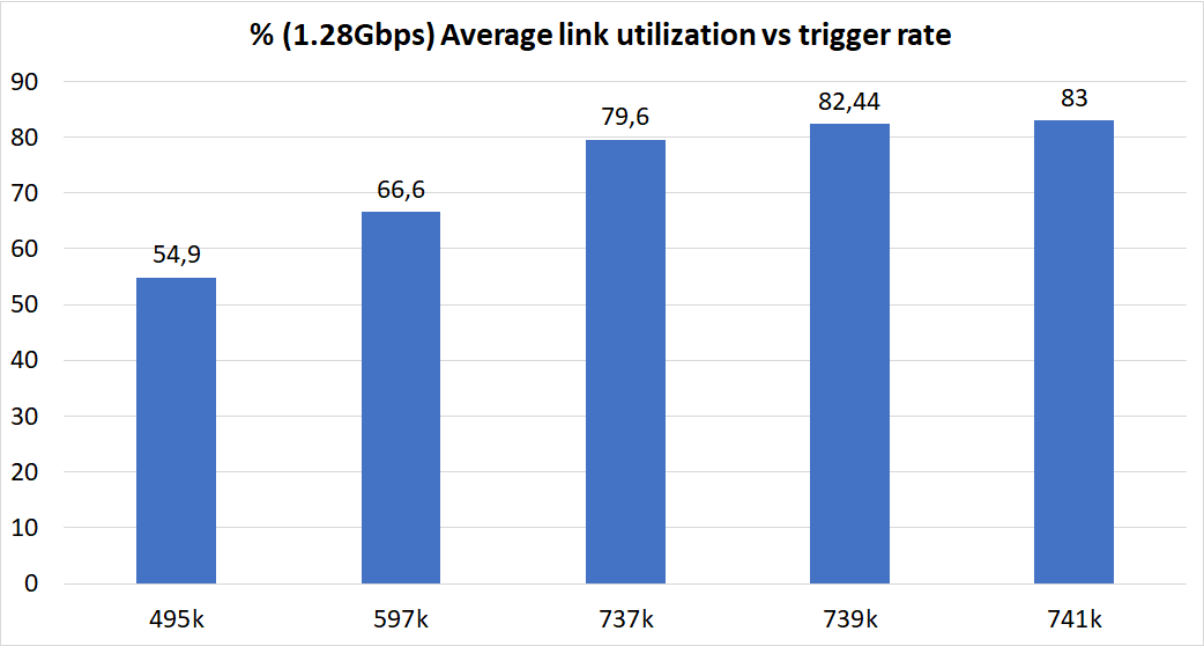


Figure 16.9: Readout link utilization at low hit rate: 270MHz/cm^2 for 4 chips with data merging into 1 readout link at 495 - 835kHz trigger rate.

Chapter 17

Bug fixes and new features in RD53B-CMS

From the initial RD53B-ATLAS chip to the RD53B-CMS, the following set of bug fixes and new features have been implemented. This is in addition to specific changes made to adapt the common architecture to the Linear analogue front-end used in the CMS chips and the slightly different pixel array size. For details refer to RD53B-CMS specific manual [7]. **It should be noted that configuration register addresses, and bit mapping within configuration registers, may change between chip versions. This may also become the case for final production versions so test and DAQ systems and test routines should to the extent possible be made with this in mind (e.g. using configuration register names and bit names, not absolute addresses and bit positions). A required change of begin/end stream bit implies that test and DAQ system firmware and software must be made according to the RD53B-ATLAS or RD53B-CMS chip being tested.**

17.1 Bug fixes in RD53B-CMS

The following major bug fixes have been implemented in the RD53B-CMS:

- Non functional 4 bit TOT latch corrected to have functional TOT charge mesurement.
- Dummy global configuration registers without TMR (for SEU cross-section measurements).
- Shunt current measurement corrected for scaling factor and observed non-linearities
- Data merging input polarity selection

17.2 Changes and new features RD53B-CMS

The following changes and new features have been implemented in the RD53B-CMS:

- Pixel array size changed from ATLAS: 8x8x50x48 = 153.600 pixels to CMS: 8x8x54x42 = 145.152 pixels
- Pixel data stream format has been modified to have an end of stream/event bit, instead of a begin of stream/event bit, to clearly identify when all data from an event/stream has been read out.
- Global configuration registers have different address mapping (and possibly also bit mapping).

- Differential analogue front-end replaced with Linear analogue front-end with different configuration parameters/registers.
- Pixel hits from analogue front-end can be detected with edge or level sensitive circuit.
- PTOT data encoding has been changed to be more efficient.
- Time resolution of general pulse generator changed (50ns in ATLAS chip, 25ns in CMS chip). Also affects internal resets as based on pulse generator.
- PLL uses by default only rising edges for locking as will have less jitter (not sensitive to pulse width distortion).
- Control link AC coupling bias voltage changed to VDD/2 = 0.6V
- Data merging inputs without termination.
- SEU counting from 4 configuration bits per Pixel. HitOr buses in each column are used to signal SEU occurances to a central 16bits counter.
- SEU counting in SEU protected global configuration registers
- Optional BCID and L0ID in event readout data.
- Optional 32bit CRC at end of event stream.
- Optional constant serial test pattern (without scrambling).
- Optional disable of triplicated clocks (for SEU TMR verification and testing).
- More debug signals on general purpose output related to data merging.
- Possibility to disable triplicated clocks
- Dummy global registers without TMR
- Dummy global registers with TMR, without clock skew on triplicated clocks
- Improved circuit to measure injection caps
- Removed option to overwrite range limitation of CDR bias

Bibliography

[1] RD53 collaboration web pages: http://cern.ch/RD53

[2] RD53 conference presentations: https://indico.cern.ch/category/5598/

[3] RD53 pixel chip requirements, CERN-RD53-PUB-19-001: https://cds.cern.ch/record/2663161

[4] RD53B users guide (this document): https://cds.cern.ch/record/2754251

[5] RD53B-ATLAS manual, CERN-RD53-PUB-19-002: https://cds.cern.ch/record/2665301

[6] RD53B test results:

RD53B weekly testing meetings: https://indico.cern.ch/category/9316/

RD53B testing TWIKI: https://twiki.cern.ch/twiki/bin/viewauth/RD53/RD53BTesting

RD53B test results overview: https://docs.google.com/spreadsheets/d/1yvjTcFx8udD4Cv6ow3FE0gAU5CmF79ldRCxXjDHUFJg/edit?usp=sharing

Test results of RD53B chips for CMS and ATLAS phase-2 pixel upgrades, Domink Koukola et al., TWEPP2021. To come

[7] RD53B-CMS manual, To come

[8] The RD53A Integrated Circuit, CERN-RD53-PUB-17-001:https://cds.cern.ch/record/2287593

[9] RD53 overview presentations:

RD53 pixel chips for ATLAS and CMS pixel upgrades, Flavio Loddo et al. TIPP 2021, To come

RD53 pixel chip developments for the ATLAS and CMS High Luminosity LHC, FLavio Loddo et al, 16th Trento workshop : https://indico.cern.ch/event/1007887/

[10] RD53 LHCC status reports:

2020: https://indico.cern.ch/event/939299/contributions/3946845/

2019: https://indico.cern.ch/event/835603/contributions/3502836/

2018: https://indico.cern.ch/event/726320/contributions/3005309/

[11] RD53 RTL repository on GIT (restricted access): https://gitlab.cern.ch/rd53.

[12] 65nm radiation tolerance papers and presentations:

Radiation-Induced Short Channel (RISCE) and Narrow Channel (RINCE) Effects in 65 and 130 nm MOSFETs: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7348757&tag=1

Dose-Rate Sensitivity of 65-nm MOSFETs Exposed to Ultrahigh Doses: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8340851

65nm TID and SEU experience, ACES 2020:https://indico.cern.ch/event/863071/contributions/3738770

Total ionising dose radiation damage studies of the RD53A chip for the ATLAS and CMS upgrades, IWORD 2019: https://indico.cern.ch/event/848367

DRAD results obtained during irradiation campaings, CERN-RD53-PUB-20-001: https://cds.cern.ch/record/2725573?

Characterization of radiation effects in 65 nm digital circuits with the DRAD digital radiation test chip, TWEPP2016: https://iopscience.iop.org/article/10.1088/1748-0221/12/02/C02039

[13] CERN Radiation tests and models (restricted access):https://espace.cern.ch/asics-support/tsmc65/_layouts/15/start.aspx#/SitePages/TSMC%2065nm%20-%20TID%20models.aspx

[14] RD53 TID and Low dose rate test results:

How low is low dose rate: To come ((https://indico.cern.ch/event/1023841)

1Grad TID: https://indico.cern.ch/event/1011941/contributions/4272881

X-ray dose correction: https://indico.cern.ch/event/993420/contributions/4230896

X-ray irradiation Bonn: https://indico.cern.ch/event/993419/contributions/4227891

X-ray irradiation: AFE and low dose rate: https://indico.cern.ch/event/993417/contributions/4201214

Low dose xray irradiations: https://indico.cern.ch/event/907954

Update from Kr85 irradiation: https://indico.cern.ch/event/907954

Low dose rate irradiations in Bonn: https://indico.cern.ch/event/907954

[15] RD53 SEU tests:

Single Event Effects on the RD53B Pixel Chip Digital Logic and On-chip PLL, TWEPP2021: To come

RD53B SEU test results, TIPP2021: To come

RD53B SEU test results, internal presentations:

Heavy ion test GANIL: https://indico.cern.ch/event/1023838/contributions/4330987

Heavy ion tests Leuven: https://indico.cern.ch/event/974714/contributions/4126245/attachments/2155756/3636160/03122020ITkPixV1_SecondSEUCampaing_PreliminaryResults.pdf

Proton test TRIUMF: https://indico.cern.ch/event/988704/contributions/4169315/attachments/2167429/3658550/210106_ITkPixV1_TRIUMF_SEU.pdf

Proton test TRIUMF: https://indico.cern.ch/event/1011938/contributions/4246422

RD53A SEU measurements: https://indico.cern.ch/event/907940/

SEU test chip: Characterization of Soft Error Rate Against Memory Elements Spacing and Clock Skew in a Logic with Triple Modular Redundancy in a 65nm Process https://pos.sissa.it/343/029/pdf

[16] SEU/SET estimates for RD53 pixel chip: https://indico.cern.ch/event/1004379/contributions/4225595/

[17] Differential AFE:

RD53A test TWIKI (restricted access): https://twiki.cern.ch/twiki/bin/view/RD53/RD53ATesting

DIFF AFE usersguide for RD53A: https://twiki.cern.ch/twiki/pub/RD53/RD53ATesting/Diff_userguide.pdf

For RD53B-ATLAS: To come

[18] Linear AFE:

CMS analog front-end: simulations and measurements, Luigi Gaioni, CERN-RD53-PUB-20-002: https://cds.cern.ch/record/2746420

From RD53A test TWIKI (restricted access): https://twiki.cern.ch/twiki/bin/view/RD53/RD53ATesting

LIN AFE guidelines for RD53A: https://twiki.cern.ch/twiki/pub/RD53/RD53ATesting/LIN_AFE_guidelines.pdf

LIN AFE tuning example for RD53A: https://twiki.cern.ch/twiki/pub/RD53/RD53ATesting/LIN_AFE_tuning_example.pdf

For RD53B, Update on improved version for RD53B-CMS: https://indico.cern.ch/event/932714/contributions/3919467/attachments/2066404/3467920/LIN_AFE_summary_June_2020.pdf

[19] Tests of AFEs and RD53A:

IV of pixels inputs and breaking limits: To come (https://indico.cern.ch/event/1023841

Comparative evaluation of analogue front-end designs for the phase-2 upgrade of the CMS inner trcaker, Natalia Emriskova: Submitted to JINST

Test results from the RD53A pixel readout chip and design status of its successor, Hiroshima 2019: https://indico.cern.ch/event/803258/contributions/3582903

RD53 analog front-end processors for the ATLAS and CMS experiments at the High-Luminosity LHC, Vertex 2019: https://indico.cern.ch/event/806731/contributions/3503810

Mark Standke master thesis, Characterization of the Joined ATLAS and CMS RD53A Pixel Chip: http://cds.cern.ch/record/2717863

Analog front-end characterization of the RD53A chip, TWEPP 2019: https://pos.sissa.it/370/021/pdf

RD53 AFE review (Access limited to RD53 members): 1st Meeting Dec. 2018: https://indico.cern.ch/event/769894/, 2nd Meeting Jan. 2019: https://indico.cern.ch/event/790036/

[20] SLDO:

TIPP2021: TO COME

An Integrated Shunt-LDO Regulator for Serial Powered Systems, M. Karagouins: Proc. of IEEE ESSCIRC '09, (2009).

[21] Serial powering presentation on issues and optimization: https://indico.cern.ch/event/928451

[22] Serial powering in CMS pixel detector:

Presentation at Hirosima 2019: https://indico.cern.ch/event/803258/contributions/3582853/attachments/1962394/3262057/267-Orfanelli-CMSInnerTrackerv1.pdf

Serial Powering for the Tracker Phase-2 Upgrade (vertex 2019): Presentation: https://indico.cern.ch/event/806731/contributions/3507505/attachments/1927839/3192084/Vertex2019_SerialPowering_Koukola.pdf Proceedings paper: http://cds.cern.ch/record/2712278/files/CR2019_300.pdf

CMS serial powering overview: https://indico.cern.ch/event/755140/contributions/3130561/attachments/1733961/2803724/SerialPowerOct2018v1.pdf

[23] Serial powering in ATLAS pixels:

ATLAS serial powering overview: https://indico.cern.ch/event/755140/contributions/3130562/attachments/1733857/2803593/Serial_Powering_Meeting_ITk_Overview.pdf

[24] RD53B Iref and serial power measurements:

RD53B serial power testing meeting: https://indico.cern.ch/event/1023842

Quand module serial powering: https://indico.cern.ch/event/1023838/contributions/4330802

Corrected wafer probing SLDO measuremnts: https://indico.cern.ch/event/1023837/contributions/4312809

RD53A Quad serial power chain: https://indico.cern.ch/event/1023835/contributions/4298473

Low power mode test: https://indico.cern.ch/event/1023835/contributions/4301728

SLDO IV curve measurements: https://indico.cern.ch/event/1011942/contributions/4291435

Iref trimming on quad: https://indico.cern.ch/event/993419/contributions/4227892

[25] Monitoring:

Calibration of temperature sensors: https://indico.cern.ch/event/1011941/contributions/4278988

[26] PLL and CDR papers and presentations:

A Clock and Data Recovery Circuit for the ATLAS/CMS HL-LHC Pixel Front End Chip in 65 nm CMOS Technology, TWEPP 2019 paper: https://pos.sissa.it/370/046/pdf

TWEPP 2019 presentation: https://indico.cern.ch/event/799025/contributions/3486150

ATLAS ITK week upgrade talk (access limited to ATLAS members): https://indico.cern.ch/event/728934/#56-rd53b-cdr

Internal RD53 CDR review (access limited to RD53 members having signed NDA): https://indico.cern.ch/event/811387

Internal RD53 SEE measurements (access limited to RD53 members having signed NDA): https://indico.cern.ch/event/832515/#10-io-padframe-pll-cml

PHD thesis of Piotr Rymaszewski: To come

[27] Xilinx, Aurora 64B/66B Protocol Specification, SP011 (v1.3) October 1, 2014.

[28] Measurements on CMS short (max 2m) E-link cables : https://indico.cern.ch/event/1002882/contributions/4220779/attachments/2184538/3691067/210204_elinks_evaluation.pdf Measurements on ATLAS long (max 6m) E-link cables: TO FIND reference

[29] Test and characterization of CMS E-link cables with RD53A: https://indico.cern.ch/event/1002882/contributions/4220779/attachments/2184538/3691067/210204_elinks_evaluation.pdf

[30] Wafer probing and production testing

Wafer probing Bonn: https://indico.cern.ch/event/993417/contributions/4201241

[31] Sara Marconis PHD thesis, Design and optimisation of low power hybrid pixel array logic for the extreme hit and trigger rates of the Large Hadron Collider upgrade: http://cds.cern.ch/record/2630094?ln=en

A UVM simulation environment for the study, optimization and verification of HL-LHC digital pixel readout chips: https://iopscience.iop.org/article/10.1088/1748-0221/13/05/P05018

Advanced power analysis methodology targeted to the optimization of a digital pixel readout chip design and its critical serial powering system: https://iopscience.iop.org/article/10.1088/1748-0221/12/02/C02017

The RD53 Collaboration's SystemVerilog-UVM Simulation Framework and its General Applicability to Design of Advanced Pixel Readout Chips: http://cds.cern.ch/record/1750098

[32] Simulation of hit losses from pixel region size and latency buffer depth: https://indico.cern.ch/event/785024/contributions/3265573/attachments/1780753/2896895/190117_BufferingEfficiencyUpdate.pdf

[33] Readout rate studies and FIFO-sizing from chip simulation: https://indico.cern.ch/event/831161/contributions/3481457/attachments/1874881/3086835/190704_CMS_IT_ASIC_DataRates_FIFOs.pdf

[34] Cocotb simulation environment: https://gitlab.cern.ch/silab/bdaq53/-/wikis/Simulator-Setup

[35] BDAQ53, a versatile Readout and Test System for Pixel Detector Systems for the ATLAS and CMS HL-LHC Upgrades: https://arxiv.org/abs/2005.11225

The BDAQ53 test and DAQ system WIKI: https://gitlab.cern.ch/silab/bdaq53/-/wikis/home

[36] The YARR test and DAQ system: https://yarr.readthedocs.io/en/latest/

[37] CMS pixel detector and electronics:

CMS pixel TDR: https://cds.cern.ch/record/1481838

CMS pixel overview presentation (ACES 2020): https://indico.cern.ch/event/863071/contributions/3738839/attachments/2045658/3427296/IT_ACES2020.pdf

Presentation at Hirosima 2019: https://indico.cern.ch/event/803258/contributions/3582853/attachments/1962394/3262057/267-Orfanelli-CMSInnerTrackerv1.pdf. Proceedings paper to come.

[38] ATLAS pixel detector and electronics:

ATLAS pixel TDR: https://cds.cern.ch/record/2285585

ATLAS pixel overview presentation (ACES 2020): https://indico.cern.ch/event/863071/contributions/3738836/attachments/2045603/3427086/ATLAS-ITk-Pixel_Phase2-ACES-2020_FH.pdf

[39] LPGBT, Low Power GigaBit Transfer link chip: https://lpgbt.web.cern.ch/lpgbt/

LPGBT presentation at ACES 2020: https://indico.cern.ch/event/863071/contributions/3738814/attachments/2044931/3425692/lpGBTstatusAndPlansACES20200526.pdf