CERN-RD53-NOTE-18-001

Version 2.1, Aug 1, 2018

RD53B Design Requirements

ABSTRACT: Common requirements for the ATLAS and CMS HL-LHC pixel readout chips. A common design framework called RD53B will be used to generate the production chip layouts for both experiments, with minor customization affecting primarily the chip size, which will be different for ATLAS and CMS, but is simply a parameter in the RD53B framework. The RD53B framework will be heavily based on the RD53A readout chip design, as many requirements are already met by RD53A.

Contents

---

1. Preface to Version 2.1, Aug 1, 2018

RD53A revised version to combine ATLAS and CMS requirements. Includes three rounds of comments from RD53 members (28 CDS exchanges available upon request). This version is now a starting point for the experiments to check that nothing is missing or wrong for their particular detector design. A final document will be produced after collecting this feedback. Both experiments must approve the final document.

2. Introduction

The RD53 collaboration was established in 2013 to design a hybrid pixel readout chip for the high rate and radiation expected in the ATLAS and CMS phase 2 upgrades [1]. The goal was to deliver in a 3-year time frame the elements required for ATLAS and CMS to produce readout chips. The RD53A integrated circuit [2] embodied these deliverables. RD53A demonstrated in a large format IC the suitability of the chosen 65nm CMOS technology (including radiation tolerance), the stable low threshold operation, and the high hit and trigger rate capabilities, required for HL-LHC upgrades of ATLAS and CMS. RD53A was not intended to be a final production IC for use by the experiments, and contains design variations for testing purposes, making the pixel matrix non-uniform. The RD53 collaboration mandate has now been extended to design and deliver the production chips for both ATLAS and CMS, which will be based on RD53A, but must now meet all production needs of both experiments.

This document collects the requirements that ATLAS or CMS production chips must respect. The intent is to take the most stringent needs from each experiment, such that the resulting requirements always meet the needs of both experiments. This document is meant to be brief, leaving detailed technical specification of chip functions to dedicated specification documents and/or behavioral models, to be produced by RD53 or the experiments as needed.

The requirements are divided into dimensions (Sec. 3), power, environment and operation (Sec. 4), trigger and DAQ (Sec. 5), performance (Sec. 6), and production (Sec. 7). Many features have already been implemented in RD53A. Where those features meet requirements or have been accepted by the experiments, they are expected to carry over to RD53B and are documented in detail in the RD53A manual [2]. Sec. 8 collects RD53A features and expected updates or changes. Additional features not implemented in RD53A or covered in the previous sections are given in Sec. 9. Chip design constraints are imposed on the experiments based on the RD53A experience and are given in subsections where appropriate.

3. Dimensions

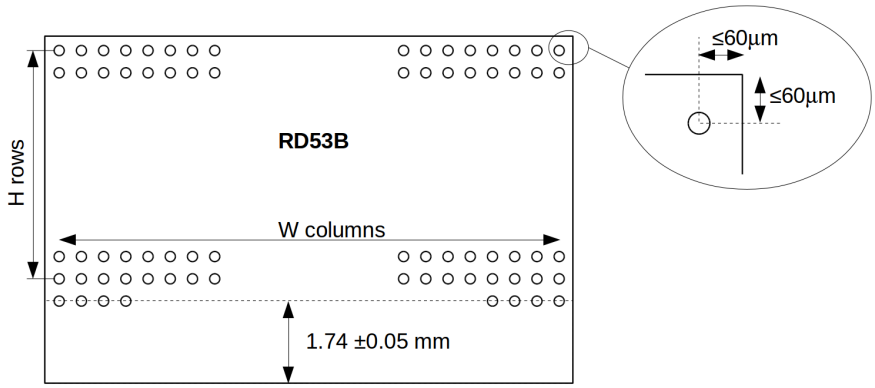

Figure 1: Diagram showing pixel matrix size and margin requirements.

Parameter	ATLAS	CMS
Pixel bump pitch	50 μm $\times$ 50 μm	
pixel rows (H)	384	328
pixel columns (W)	400	440
Last pixel to chip edge	60 μm	

Table 1: Chip size values.

Two different pixel matrix sizes must be fabricated using the same design. The I/O, power circuits, hit processing and readout, and performance are expected to be identical for the two sizes, with only the number of pixels changing. The chip size is shown schematically in Fig. 1. The number of pixels is W×H as shown in Table 1. In addition to the W×H pixel inputs, there are 8 bump pads at the bottom of the array to route sensor bias and guard ring contacts to wire bond pads, exactly as in RD53A. The bottom of chip height should remain within 100 μm of the RD53A value as shown in the figure. Also shown is the margin from last bump to chip seal ring on the three abutable sides. The air-gap between chips in a multi-chip module depends not only on this margin, but also on the dicing of the physical chips, which leaves some additional silicon beyond the seal ring. This margin requirement assumes the dicing streets are 80 μm wide, as was the case for RD53A. (The dicing street width sets an upper bound on the additional silicon beyond the seal rings).

4. Power, environment and operation

The readout chip must work in the ATLAS and CMS detectors, with the constraints imposed by integration into the full systems, under their data taking conditions, for the lifetime of the experiments. Table 2 quantifies what that means. Since the same chip will be used everywhere in the detector, the design must respect the most severe requirements. For example, the inner layer conditions dictate radiation tolerance and SEU requirements. (Also for hit rate, but that is covered in Sec. 5 and 6). There are differences in the choice of parameters for power and operation relative to the RD53A specifications [3], which reflect a better understanding of how the detectors will function.

The power will be supplied to chains of multiple stages in series, where each stage has 4 chips in parallel except in the CMS inner layer where each stage has 2 chips in parallel. The detectors must design services, cooling, and system for each serial chain, leading to a serial chain current requirement. Specification of individual chip current and subcircuit consumption, all the way down to individual pixel current, flows down from these requirements taking into account functional and circuit design details. This specification will be done by RD53. It will take into account balancing of current among the 4(2) chips and among analog and digital regulators within each chip, overhead needed for serial power regulation, startup, possible fault conditions, etc. Specific values for each of these details are not required by the experiments. For operating the detector, what matters is the current supplied to the full serial chain. Prescribing details of how this current distributes itself to eventually end up powering individual pixels would unnecessarily constrain design options without any benefit to the experiments. Therefore, Table 2 gives only the serial chain current requirements

Parameter	Value	=RD53A	Comments
Operating Temperature	-40°C to +40°C	y	performs in this range
Survival Temperature	-55°C to +250°C	N	not degraded by exposure
Serial chain nominal I, inner	6(3) A	N	for 4(2) chips in parallel
Serial chain nominal I, rest	5 A	N	always 4 chips in parallel
Serial chain max. I, inner	7(3.5) A	N	for 4(2) chips in parallel
Serial chain max. I, rest	6 A	N	always 4 chips in parallel
Single chip absolute max. I	4 A	y	see 4.3
Current transients	<5%	y	of chain I. See 4.1, 4.2
Startup serial I	1.5 A (placeholder)	N	4 chips in parallel. Chip responsive at this I. To be checked and updated.
Radiation TID	500 Mrad	y	delivered at -15°C. 25°C anneal only
Ionizing dose rate	<15 rad/s	N	delivered at -15°C.
TID gradient	$\leq$(3:1)	N	TID ratio across chip width or height
Annealing	<1 month per 100 Mrad	N	at 25°C.
NIEL fluence	10^{16} 1 MeV n.eq./cm^2		
Flux of >1 GeV particles	$<2\times10^8$/cm^2	N	
Flux of HEH particles	TBD	N	

Table 2: Power and environmental requirements. The flux of high energy particles (>1 GeV) will cause single event upsets in logic and memory and performance requirements must be met in this flux. The column "=RD53A" indicates if the requirement is unchanged relative to the RD53A specifications (y), or is different or new (N).

as opposed to individual pixel current or power as was done in the RD53A specifications. Note that higher current is allowed for the inner layer, which represents a small fraction of the total system power, and so can afford to "burn" some extra power to meet the most demanding performance goals. The rest of the detector, on the other hand, dominates the system power, but has lower hit rate and so can more easily meet performance with lower current.

4.1 Serial chain operation

A serially powered detector has not been operated so far. Experience with serial chains comes only from prototyping activities. Therefore, cautionary requirements are placed on serial chain functionality. Perhaps some of these will turn out to be too conservative, but reliable operation is too critical to use looser requirements based on untested expectations about how a large system will function. The current setting and voltage offset for serial operation should be controlled by external resistors allowing the user to precisely fix them.

A new requirement that was not specified for RD53A is the suppression of current transients, given in Table 2 as a fraction of the serial chain current. This ensures their effective load represented by one or more serial stages does not fluctuate during chain operation, as this would lead to potentially dangerous voltage transients in the the other stages. Such fluctuations are observed in serial power system tests with FE-I4 chips and have varied causes, not all well understood yet. For visualization purposes, the RD53A ShuLDO operation can be pictured as steering current between the chip core and a dummy load (the shunt element) such that the total current remains constant.

The user fixes the value of the dummy load resistance with an external setting (a ratio resistor), and the circuit acts to steer current away from this load, towards the chip core, as needed. If the chip core is completely off then all the current flows in the dummy load, given by the voltage drop over dummy load resistance. This is the total current that remains constant as current is steered to power the core. The "no transients" requirement means that the dummy load must always have nonzero current passing through it. If the chip core ever demands more current than is available from reducing the dummy load current, RD53B is required to "refuse" that demand, limiting the core current draw so that transients do not propagate to the serial power chain. This limitation should not disable the chip or lose configuration for small transients (microsecond scale duration and few <20% of normal current in amplitude), while extreme overcurrent events (milliseconds to permanent and large amplitude) may, if deemed appropriate, trigger reset/disable actions. Reset/disable actions should in any case be internal to the affected chip and not propagate to the serial chain. A voltage difference between the analog and digital domains should not cause a high current state.

Power cycling of individual chips or modules in a serial chain is not possible. Power cycling the entire chain is assumed to be a major disruption of operation. Therefore, a reset mechanism is required that allows returning a chip to its default configuration and communication state without cycling the power and without affecting the serial chain current.

4.2 Design constraints

The absolute maximum regulator input voltage is 2.0 V. For cooling system design this gives an absolute maximum power per single chip and per module using the required currents listed in Table 2. Note that the maximum module power is given by the serial chain current, and not by the maximum single chip current. The latter can only occur when other chips in the same module have failed open as indicated in Sec. 4.3. See [2] for more details.

The suppression of current transients is not achieved only within the chip, but also depends on off-chip capacitance at both the input and output of each of the power regulators. Currently these capacitors are assumed to be 2 μF each. Stored charge in these capacitors and time constants for active regulators may prevent a perfect control of serial chain current transients. The Table 2 requirement on transients is, therefore, somewhat flexible.

It should be noted that power dissipated in the ShuLDO regulators will be concentrated in the chip periphery. This is an issue that should be considered for the thermal design of modules and supports. This power will be large at startup or after a reset, when the chips have their default configuration, and will decrease after the chips are configured. In of a chip failure, the remaining good chips in the module will take extra current leading to a constant high heat load in their periphery.

4.3 Hard and soft failures

A single chip hard failure may occasionally occur, for example an internal short or disconnection of wire bonds. These will be rare events and therefore are allowed to disrupt an entire serial chain, including full shutdown. However, they must not trigger a chain reaction causing other chips to also have hard failures (for example due to overvoltage). The requirement is that a sudden hard failure (open or short) of a chip within a serial chain should not lead to a dangerous transient for any other chip in the chain, either in the same module or other modules. The possibility of open circuit hard

failures within a module leads the maximum single chip current requirement of Table 2, as the remaining chip(s) after a failure must pass the full serial chain current. See [2] for more details.

A Soft failure means that a chip temporarily behaves outside limits, but can be recovered through reset or other means. Soft failures can be much more frequent than hard failures and can take place in multiple chips at the same time, depending on the causes. Clearly, a soft power failure in a single chip would not cause damage to other chips if the above hard failure requirement has been met. But the same cannot be said for a soft failure in N chips at once. For larger and larger N within the same serial chain, eventually it will not be possible to avoid dangerous transients on the "bystander" chips. Therefore, the requirement on soft failures is the "no transients" requirement of Table 2. This essentially means that RD53B is required to never behave like a short or an open except in the case of a hard failure. No possible command sequence, reset, hit occupancy pattern, triggers, etc. should cause the chip to draw more or less current in serial power configuration. In system tests using FE-I4B chips, soft failures can be caused by invalid command sequences that arise on e-links during GBT reset. For RD53B it must not be possible to cause a soft power failure regardless of what arbitrary waveform, bit sequence, or lack of defined voltage levels is presented to serial command input.

4.4 Radiation

The radiation requirements include total ionizing dose (TID) and instantaneous particle flux. The latter causes single event upsets (SEU). The chip must operate reliably, meeting efficiency requirements in the presence of this flux. SEU can be viewed as reducing efficiency, by disabling individual pixels, groups of pixels, and corrupting triggers and data. It is left up to the designers to make specific choices for how to maintain operation and efficiency within this flux. The tools that can be used are well known, including SEU hardening of circuits, use of self-healing protocols such that errors (for example in state machines) do not persist, gradual reconfiguration during data taking (trickle configuration), etc. Note that both the flux of non-relativistic nuclei and of high energy particles are important. Both cause localized upsets of single gates, but only the latter cause extended charge deposits that can upset spatially separated redundant elements.

TID damage is better understood now than when the RD53A specifications were written. Assuming low temperature operation and annealing at room temperature only with power off, it seems possible to increase the 500 Mrad tolerance specified for RD53A, with the caveat that enhancement of damage due to low dose rate (as will be experienced in the experiment) is still being investigated. As both experiments have designed their inner 2 layers or more to be serviceable, the requirement on TID tolerance for RD53B remains 500 Mrad. Nevertheless, it is understood that the designers will strive to improve TID tolerance relative to RD53A, even if RD53A already meets the 500 Mrad requirement.

A Non-Ionizing Energy Loss (NIEL) value is given in addition to ionizing dose because some special devices, such as parasitic bipolar transistors, are affected by NIEL.

5. Trigger and DAQ requirements

The trigger and DAQ must be compatible with two different operating modes, which can be understood as a separation of the trigger and readout functions. While for RD53A "trigger" means both

Parameter	Value	=RD53A	Comments
Trigger rate	$\leq$4 MHz	N	trigger only without readout
Trigger latency (L0)	$\leq$12.5 μs	y	programmable
Auto read rate	<10 MHz	N	reads all triggers. output bandwidth permitting
Manual read rate	$\leq$1 MHz	N	read cmd. L1 latency after trigger
Readout latency (L1)	$\leq$25 μs	N	latency for above command
Trigger command	tagged	N	each trig. command contains an identifier (tag)
Number of unique tags	>50	N	
Trigger buffer depth	>100	N	note each trig. command can specify 1-4 triggers
Read command	auto or tagged	N	see text

Table 3: Trigger and DAQ requirements. The column "=RD53A" indicates if the requirement is unchanged relative to the RD53A specifications (y), or is different or new (N).

select the data from a specific bunch crossing and read it out, we now make a distinction between selecting the data (the trigger proper) and the readout action, which may or may not happen for every single trigger. The RD53A operating mode is the default for both experiments, and we refer to it as auto-read. Additionally, ATLAS requires a 2-level trigger mode, in which triggers are first sent and only some of them are later read out. This requires separate trigger and read commands. In both cases all hits older than the programmed latency that are not triggered should be automatically erased from the pixel buffers. Table 3 lists the required trigger parameter values.

The trigger commands will be synchronous with a pre-programmed latency (look-back time). Each trigger will have an identifier called a tag associated with it. The trigger tag is provided by the DAQ to the chip and has no meaning within the chip- it is an arbitrary code that the chip must return with the data fragments corresponding to that trigger. In case of 2-level trigger mode (manual read as opposed to auto read), there must be read command that identifies a prior trigger by its tag. In auto read mode the trigger and read rates are obviously the same and the sustainable rate can be as high as the installed output bandwidth allows (this is how RD53A operates). The highest rates correspond to ATLAS planned installation high bandwidth readout in the outer layers (4 MHz) and CMS plans to use certain low occupancy modules for luminosity monitoring (as high as possible up to 10 MHz). In manual read mode the trigger and read are decoupled. Only ATLAS plans to use this mode with the highest read rate as given (which is lower than that the trigger rate- otherwise one would just run in auto read mode). The reason is that in the inner layers not enough output bandwidth is possible for auto read at the highest trigger rate.

5.1 Data output

The AURORA protocol used by RD53A should be kept. The number of active, bonded lanes should be selectable between 1, 2, 3, or 4. The RD53A lane rate of 1.28 Gbps should be kept, programmable down to 160 Mbps. A required new feature is the addition of 3 input lanes, nominally at 320 Mbps so that one chip can be configured as master and either 1 or 3 chips as slaves. In the

Requirement	Value	Comment
Bits/hit barrel inner layer	$\leq$13 bits	excl. overheads. To be confirmed
Bits/hit other	$\leq$15 bits	excl. overheads. To be confirmed
Per trigger overhead	$\leq$25 bits	
T99 for 1 MHz	2.0 μs	(ATLAS only) to be confirmed
T99 for ATLAS 4 MHz	1.0 μs	(ATLAS only) to be confirmed

Table 4: Readout encoding requirements as discussed in text. This applies to lossless encoding including ToT information. Numbers to updated with final values from ATLAS and CMS simulations.

case of 1 slave chip the a transfer bandwidth of 640 Mbps is required, which can be implemented as one lane or two lanes at 320 Mbps each. The data from each slave chip is to be combined by the master, along with its own, into a single 1.28 Gbps output.

The RD53A chip output encodes data using 33 bits per hit region (after 64b/66b) plus 50 bits per trigger on average. There is no data compression or suppression of zero ToT values within hit regions. This was simple, robust encoding suitable for RD53A, but the production chips must increase encoding efficiency in order to fit the number of data links allocated in the experiment designs. The required encoding efficiency is shown it Table 4. In high hit rate regions the bits/hit dominate the output bandwidth, but in low hit rate regions and high trigger rates (ATLAS two level trigger or CMS luminosity monitoring) the event/trigger overhead can be large. Both terms are important and the chip encoding must be efficient in both cases. Data encoding must be efficient for large clusters (e.g. end of barrel) and small clusters (e.g. end caps or inclined modules). Final verification of readout formatting efficiency and data rates will have to be based on MC studies/simulations (both at software and chip implementation level) covering different locations in the pixel detectors of the two experiments. Such simulations will also determine what effective link occupancy factor can be achieved (we expect this will be no higher 80%). This is related to readout latency and event truncation. ATLAS has specific constrains on the readout latency (T99 as shown in table 3), as data read out is used in a second level trigger system with a constrained maximum decision latency. CMS does not have such a specific constraint and readout latency is simply assumed constrained by the available output buffering (assumed programmable with a given maximum) in the pixel chip.

Finally, data integrity requirements must be respected by the implemented encoding. The use of tags discussed earlier must be supported within the above bounds. Corruption of a fragment of the output data of any size must not lead to loss of more than 100 hits beyond the corrupted fragment. Typically the loss should be significantly less than a small number of hits ($<<100$) following a corrupted fragment. Furthermore, the decoding of the output data stream should automatically recover without any explicit DAQ action in response to detected corruption.

5.2 Data truncation, filtering, and lossy reduction

Anomalous events with a large number of hits can occasionally happen. A programmable truncation capability is required so that readout of large events can be terminated and their data flushed.

Pixel type	Capacitance		Leakage I		Hit rate	
	min.	max.	min.	max.	min.	max.
Inner normal	20 fF	75 fF	0	10 nA	0	75 kHz
Inner edge	30 fF	100 fF	0	15 nA	0	113 kHz
Outer normal	20 fF	50 fF	0	10 nA	0	15 kHz
Outer edges	40 fF	100 fF	0	20 nA	0	30 kHz

Table 5: Characteristics of different pixel sensor types. "Inner" pixels are those in the inner layer, which are expected to be 3D sensors with single chip modules for ATLAS and possibly 2-chip modules for CMS. "Outer normal" are the pixels away from the chip edges in the multi-chip modules for the various barrel layers and endcaps other than the inner. All are expected to be planar. "Outer edges" are ttab:perfhe larger pixels needed to span gaps between chips in multi-chip modules.

Bits per hit requirements are based on simulation. While the simulation is considered reliable for collision particles and secondaries produced in detector material, it does not include all possible backgrounds including unexpected transient noisy pixels, afterglow, x-rays, out of time hits, etc. Validation with present detector data cannot probe the lower thresholds that will be used in ITk. Therefore, an on-chip, programmable hit filtering capability should be implemented. For example discard isolated hits (1-hit clusters) below a given ToT, which would target noise and x-ray/afterglow backgrounds. This requirement is to be refined after further studies.

It should be possible to suppress ToT information and read out only binary data. This is both a last resort safety margin and useful for applications such as luminosity monitoring where charge information is not needed and the highest possible trigger rate is desirable. Binary readout should reduce the data volume by 4 bits per hit in the case of uncompressed ToT and somewhat less if ToT has been compressed to achieve the targets of Table 4.

6. Performance requirements

In simplest terms, the hybrid modules (readout chip plus sensor) must record the hit pixels and crossing times of 99% of incident charged particles, must hold this information until a trigger decision is received (according to Sec. 5), and must read out the triggered information without loss, negligible fakes, and in a short enough time as needed for higher level triggering. This performance must be delivered within the system and environmental constraints of Sec. 4.

While the experiments plan to use one readout chip throughout the entire pixel detector, the critical performance requirements are defined by the innermost layer, which will be at a radius of 3-4 cm from the HL-LHC interactions. Use in outer layers has generally lower performance requirements, and the operating settings must allow varying the performance to cover the range.

The chip has to work with the signal provided by the sensor and pull this signal away from the sensor capacitance. Thus, the chip design must assume something about the sensor, and at the same time chip design considerations place constraints on sensor features. The sensor properties assumed here must still be verified on RD53A module prototypes. The single pixel sensor properties are given in Table 5.

Perf. metric	Value	=RD53A	Comment or Test Conditions
Input polarity	Negative	y	
Min. in-time threshold	900 e^-	N	50% of 900 e^- hits are in time
Charge above threshold resulting in <25 ns time walk	300 e^-	N	see Sec. 6.1
Min. stable threshold	600 e^-	y	see Sec. 6.1
Hit loss from in-pixel pileup	$\leq$1%	y	See Sec. 6.2
Hit loss from all other sources	$\leq$1%	N	See Sec. 6.2
Recovery from saturation	<1 μs	y	See Sec. 6.3
Noise occupancy	$< 10^{-6}$	y	Fraction of hits pixels w/o charge injection in any 25 ns interval
Single pixel noise (ENC)	design-dependent	y	See Sec. 6.1
Pixel-pixel BX clock skew	$\leq$4 ns	y	diff. between slowest and fastest
Hit delay dispersion	$\leq$4 ns	y	diff. between slowest and fastest
Charge ADC modes (selectable)	ToT4, ToT6/4	N	4-bit counter or 6-bit ctr. compressed to 4 bits (6.4)
ToT speed	40 MHz or $\geq$80,MHz	N	selectable
ToT readout modes	4b or 1b	N	before compression, 1b=binary
ToT scale uniformity	<15% RMS	N	scale dispersion among pixels
Charge scale shift	<2% / Mrad	y	change in mean with radiation
Threshold scale shift	<15 e^- / Mrad	y	change in mean with radiation
Threshold dispersion shift	<60 e^- / Mrad	y	added in quadrature
Charge meas. dispersion shift	<0.1 MIP/Mrad	y	added in quadrature

Table 6: Performance metrics and required values. These requirements apply universally to all pixel types (there are no bare chip performance requirements as the detectors only use nodules, not bare chips). The column "=RD53A" indicates if the requirement is unchanged relative to the RD53A specifications (y), or is different or new (N).

The chips in all layers must have equally good performance, that will be easier or harder to achieve because of the characteristics in Table 5. For example, the hit loss from in-pixel pileup is more difficult to achieve for Inner pixels, where it will demand a faster return to baseline, than for outer pixels, which can therefore operate with a relaxed return to baseline. An exception is made for Inner edge pixel performance, which may not meet all requirements, for example in pixel pileup inefficiency. The required performance is given in Table 6. Some of the required values will differ from the RD53A specifications. This follows evolution of sensor development as well as influence of the achieved performance of the RD53A design. The following should be particularly noted.

6.1 Threshold, noise, and noise occupancy

The 600 e^- stable threshold specification of RD53A is preserved as a requirement. However, it is a secondary requirement depending on achieved time-walk. The primary requirement for data taking is the in-time threshold, which is required to be 900 e^- in order to be efficient for signals from 75 μm MIP path length, after 50% charge loss due to sensor radiation damage (this implies a Landau peak of approximately 3000 e^-). The amount of timewalk will depend on the front

end type, and for a faster front end a 600 e^- absolute threshold may not be needed to achieve the 900 e^- in-time threshold primary requirement. Based on simulations and measurements all candidate front ends are expected to meet a 300 e^- time-walk metric, which is better than was specified for RD53A. This is small enough that a late hit recovery method may not be required in the readout. It is critical for the experiments to make an assessment about this using the timewalk figures measured in RD53A prototype modules, as implementation of late hit recovery functionality requires compromises elsewhere. No late hit recovery was implemented in RD53A.

As was the case for RD53A, there is no explicit requirement on input referred pixel noise. The important parameters for operation are minimum threshold and noise occupancy at that threshold. Different trade-offs between single pixel noise and other parameters can be used to achieve the same result as explained below. The noise occupancy is the dark count probability in an arbitrary 25 ns interval. A 10^{-6} noise occupancy means 0.1 noise hits per bunch crossing in a 10^5 pixel chip, while the number of real hits from tracks in such a chip will be of order 100 (10) in the inner (outer) layer(s).

It is reasonable to ask what equivalent input noise charge (ENC) is needed to achieve 10^{-6} noise occupancy at 600 e^- threshold. Clearly one must have ENC«126 e^-, since a 10^{-6} probability corresponds to a Gaussian tail beyond 4.75σ, but this must be corrected to take into account the in-time noise sigma, not the total noise sigma, hence the use of "«" to indicate an upper bound rather than an acceptable level. But even using the in-time noise sigma is not sufficient: it is the quadrature sum of the ENC, the static threshold dispersion, and the RMS threshold fluctuations vs. time that must fall below the 126 e^- upper bound on equivalent input charge. Simply assuming that these three contributions are equal results in a 73 e^- ENC upper bound, which is a reasonable target, but it is not a strict requirement, as the importance of the other two contributions can be traded off. Measurements of static and dynamic threshold dispersion in RD53A modules will further inform this optimization.

The static threshold dispersion can be small right after tuning, but there may be a significant time and radiation dose in between calibrations. A single run may deliver a 1 Mrad dose in the inner layer. Therefore, the bottom 3 rows of Table 6 place requirements on how much thresholds can shift and dispersion can grow with radiation dose. These requirements may be avoided in case of a continuous tuning capability, rather that tuning being limited only to calibration intervals between runs.

6.2 Hit losses

A main physics performance requirement is 99% cluster efficiency for minimum ionizing particles. It is not trivial to translate this into a single pixel hit loss requirement. Random dead or temporarily disabled pixels can bias multi-pixel clusters, but not completely erase them. Therefore, we expect that higher than 1% single pixel loss can be tolerated while still meeting the physics requirement. At the same time, lowering hit loss has a high cost, for example in terms of power. We therefore want to be careful to not place a higher than needed hit loss threshold. We also note that loss mechanisms are most significant for the inner layer, so all other layers will automatically exceed this requirement.

Hit losses are divided into in-pixel pile-up and other sources, with a lower than 1% loss requirement for each. The latter category includes dead pixels, disabled or inefficient pixels due to

SEU, digital buffer overflow, and data transmission losses (due to SEU or otherwise).

In-pixel pile-up occurs when a new hit arrives while the discriminator is still high from a prior hit. The probability for this scales with the hit rate and the return to baseline of the front end. A fast return to baseline reduces the effective gain of the front end and constrains ToT performance. Detailed front end specifications will flow from this requirement and sensor signal models.

6.3 Saturation

Very large signals can occur not only due to hits in operation, for example nuclear fragments, but also for other reasons in testing. It is hard to define a very large signal size- could be 100 ke^- or even 1 Me^-. Such events will be very rare and therefore it is not necessary to consider them in estimating efficiency. The important thing is that a pixel should eventually recover from such an event. We have specified <1 μs as a recovery time, but this is not a strict limit and even longer times would be acceptable. The important thing is to check that the design is compatible with such rare events.

6.4 ToT

The ToT gain is determined by the number of electrons input charge corresponding to 8 counts (half the range of 4-bits). The desired gain range is 3 ke^- minimum and 12 ke^- maximum. Note that this is determined by the front end return to baseline and the counter rate, both of which are adjustable. However, the return to baseline for the inner layer is constrained by the single pixel pileup hit loss requirement.

The case of a 6-bit counter compressed to 4-bit is to be implemented as dual slope, 1:1 between 0 and 7 and 1:4 above 7. Thus the maximum number of LSB's counted would be 39, reaching 5 times the ToT range of x electrons for 8 counts. This is a dynamic range of 5.3 bits.

7. Production requirements

Module construction and test, and detector integration and test, give rise to several requirements. In particular, it will be necessary to test and operate individual components and track them. The ability to test during integration steps is needed to verify integrity of assemblies before they are blocked from access by other items integrated after them. During this time there will be no evaporative cooling present, requiring low power operation as indicated by the startup serial current of Table 2.

Note that there is no use case for reduced operation in low power mode followed by a transition to normal power (full) operation. Low power mode is needed because during integration there is no proper cooling in place yet, and thus normal power mode is forbidden (it would lead to high temperature and trigger safety interlocks). Once propose cooling is in place, there is no longer any need for low power mode. The only two use cases are, therefore, turn on in low power mode, or turn on in normal power mode. There is no requirement for a direct transition between these modes.

7.1 Start-up and default configuration

The startup serial current of Table 2 is meant to allow minimal electrical tests to verify integrity. The serial chain must be able to operate at this low current. As the current is set from an external

power source, it is always possible to supply a low current to a serial chain. What is required is that the chip I/O functions at this low current and the chip responds to commands. The ShuLDO regulators that define the impedance model of the chip as seen from outside must generate enough voltage at this low current. Clearly, the chip internal current consumption at startup must be below this low current value, so the default global and pixel configurations must be chosen accordingly. This is different from RD53A, where the default configuration was chosen to allow as much testing as possible in case of problems writing configuration to the chip (the purpose of RD53A was to demonstrate performance, not to build a detector).

7.2 Edge pixels

The outer edge type pixels in Table 5 are needed to span the gap between adjacent chips in quad or dual chip modules. To be able to span a distance greater than 100μm without having very large individual pixels, it is proposed to stretch the pixels in the last two or four rows or columns, as shown in Fig. 2. The chip design must provide dedicated bias control to the two or four columns (selectable at layout synthesis time) on either side of the chip, the two or four rows at the top, and the two top corners. ATLAS and CMS must each specify whether they want two or four columns/rows in the edge bias zones. The different zones are needed so that those pixels may be biased differently in order to cope with greater capacitance and leakage current than the normal pixels. To cover all the possible module use cases, size independent bias controls are needed, as shown in Fig. 3.

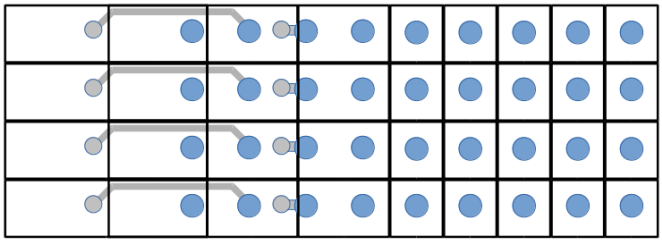

Figure 2: Diagram showing four columns of larger pixels to the left of five columns of normal pixels. The larger blue circles are on a $50 \times 50 \mu m^2$ grid matching the RD53B bump pad locations. The light gray shade represents sensor metal connecting each pixel to a chip input.

7.2.1 Design constraints

While top edge and side edge pixels will have separately controlled bias, corner pixels cannot have their own dedicated bias. They must be treated as top or edge pixels (to be decided during implementation), so a non-optimal bias must be accepted for them.

7.3 Alignment marks

Alignment marks are needed for flip chip bump bonding. Two marks should be included at the extremes of bottom of chip periphery. Marks at the top of the chip are not be possible given the bump pad density. The exact layout of the marks should be coordinated with bump vendors.

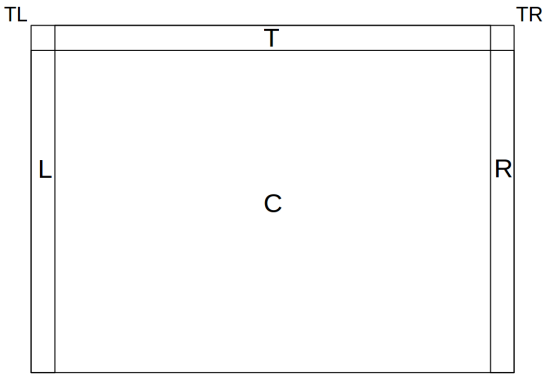

Figure 3: Diagram showing the six zones for independent pixel bias adjustment to handle edge and corner pixels. They are: C=center, L=left, R=right, T=top, TL=top left, and TR=top right. The width of the edge zones can be two or four pixels, to be selected at layout time.

7.4 Design For Test

Effective structural testing of th periphery with high fault coverage must be possible in a short time (order 1 minute or less). Testing must cover triple redundant logic (that is intrinsically immune to single bit faults). In the pixel matrix structural tests are not required. Testability of the pixel matrix logic is left for the designers to define.

7.5 Serial number

The chip should have a PROM readable as a global configuration register that can be used to program a serial number. Radiation tolerance is desirable but not mandatory.

7.6 Self trigger

An internal self trigger function is needed to facilitate source scans for every module during production. Modules will not have HitOR outputs that can be used on test boards to externally implement a self-trigger.

7.7 Cap measure circuit

This circuit is present (multiple copies) at the top of RD53A and requires separate non-overlapping clocks and dedicated current inputs. Only one circuit is needed in the production chip, but it must be in the bottom pad frame. Furthermore, internal generation of the control signals is desirable to avoid the need for special test pins.

8. RD53A features and expected updates/changes

8.0.1 Monitoring

RD53B should maintain the extensive monitoring features from the RD53A (temperature, radiation monitoring, biasing, calibration voltages, etc.). This should be extended with SLDO current monitoring (that was not included in RD53A due to lack of time).

RD53A Feature	action	Comments
Default configuration	change to low power	see Sec. 7.1
Wire bond pad design	keep	to be confirmed
Sensor bias ring bumps and wire bond pads	keep	
Alignment marks	change	see Sec. 7.3
Top wire bond pads	remove	
Hit OR outputs	change	See 8.1
Multilane AURORA output	keep	verify 3 lane mode
CML output drivers	keep	
Command protocol	update	See 8.2
Differential receivers	keep	
Cap measure circuit	move to bottom	
Monitoring MUX, ADC	keep	
Radiation & Temp. sensors	relocate	See 8.3
Ring oscillators	relocate	See 8.3
Configuration write/read	keep	
Pixel masking	keep	
Calibration injection	update	See 8.4
Clock phase adjustment	keep	
Service/status data frames	keep	

Table 7: RD53A features and translation to RD53B.

8.1 Hit-OR outputs

RD53A has four differential outputs that allow separate HitOr maps in parallel. The outputs should be reduced to one in RD53B, with internal logic to manage the combination if the four maps for output and for internal self triggering.

8.2 Command protocol

A new command must be added to implement the ATLAS 2-level trigger. An command symbol with maximum number of transitions that can be used as idle must be defined.

8.3 Sensors

8.4 Calibration injection

Reduce LSB size. The minimum injection step size in RD53A is slightly too large for accurate S-curve sampling.

9. Additional features for RD53B (desirable but not mandatory)

9.1 High resolution ToT and ToA

Based on HitOr. Ideally for each matrix Hit Or output line there should be a precision leading edge time stamp and a precision pulse width measurement. This information could be saved in registers

and/or sent out in the regular data stream, if selected, for example using a virtual row number N+1, where N is the number of physical pixel rows.

9.2 Internal processing and histogramming

Programmable pattern search to count occurrence frequency in un-triggered data. For example histogramming of cluster length distribution. Histograms could be read out at will the same as configuration registers. Internal processing can also be used to feed internal self-trigger. For example allowing logic combinations of the various Hit-ORs. This could allow for example triggerless luminosity measurement.

9.3 Bump connectivity test

Appropriate features to allow efficient testing of bump-bonding (shorts and opens) before being assembled into modules (without sensor bias) and without radioactive source testing.

References

[1] RD53 Collaboration, "RD Collaboration Proposal: Development of pixel readout integrated circuits for extreme rate and radiation," CERN-LHCC-2013-008 ; LHCC-P-006 (2013).

[2] RD53 Collaboration, "The RD53A Integrated Circuit," CERN-RD53-PUB-17-001 (2017).

[3] RD53 Collaboration, "RD53A Integrated Circuit Specifications," CERN-RD53-PUB-15-001 (2015).