CERN-RD53-PUB-19-001

Version 3.1, February 22, 2019

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 3.1, February 22, 2019 **2**

2. Introduction **3**

3. Index of changes from RD53A to RD53B **3**

4. Dimensions and wire bond pads **5**

5. Power, environment and operation **6**
- 5.1 Non-uniform heat load 7
- 5.2 Serial chain operation, overcurrent and overvoltage protection 7
- 5.3 Hard and soft failures 8
- 5.4 Radiation 9

6. Trigger and DAQ requirements **9**
- 6.1 Data output 10
- 6.2 Data truncation, filtering, and lossy reduction 11

7. Analog performance requirements **11**
- 7.1 Hit losses 12
- 7.2 Saturation 12
- 7.3 ToT 12

8. Production requirements **12**
- 8.1 Start-up and default configuration 13
- 8.2 Edge pixels 13
- 8.3 Alignment marks 13
- 8.4 Design For Test 13
- 8.5 Serial number 14
- 8.6 Self trigger 14
- 8.7 Cap measure circuit 14
- 8.8 Monitoring 14
- 8.9 Hit-OR outputs 15
- 8.10 Command protocol 15
- 8.11 Sensors 15
- 8.12 Calibration injection 15

9. Additional features for RD53B **15**
9.1 High resolution ToT and ToA 15
9.2 Internal processing 15
9.3 Bump connectivity test 15
9.4 Ability to use special features during operation 15

1. Preface to Version 3.1, February 22, 2019

The initial requirements document release and review by experiments was 6 months before this version. Final requirements could not be defined until now, due to evolution of the system designs. This version is now meant to freeze the requirements ahead of the ATLAS submission of RD53B. The RD53B design has matured in the mean time and many requirements are already built into specific implementations (for example the output data format), making it superfluous to provide many details in this document. The RD53B design manual in preparation will contain the technical details.

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. Taking 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 the RD53B manual and/or behavioral models, to be produced by RD53. Sec. 3 provides a convenient index in the form of modifications and additions to the RD53A design.

The requirements are divided into dimensions (Sec. 4), power, environment and operation (Sec. 5), trigger and DAQ (Sec. 6), performance (Sec. 7), and production (Sec. 8). 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]. 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. Index of changes from RD53A to RD53B

Feature	**Change from RD53A to RD53B**	**Doc/manual section**			**Note**
		here	**RD53A**	**RD53B**	
Size	Increase to full matrix. Reduce top margin.	4		1,2	
Wirebond pads	Remove top pads. Enlarge bottom pads.		4		
Sensor GR bumps	Keep the same		4.1		
Alignment marks	Change (see section)	8.3			
Serial power	Significant evolution of regulator design. Improve startup, add overcurrent, overvoltage protection. Add low power mode (two control options).	5.2	3		
Clock (PLL)	Change design. Now lower jitter and reliable startup to lower voltage		7.1		
Ck phase adjust	Keep the same		7.1		
Output drivers	Keep the same		9.1	10.1	
Input receivers	No change. Already possible to DC or AC couple.				
Current ref	Power from dedicated regulator for safe startup. Increase from 4ua to 20uA for robustness against radiation induced leakages.		8		

Voltage refs	Generate from above common. V-offset common to both regulators.		3, 8		
Front End	covered separately in FE review process				
Pixel core	Keep same size: 8x8		2		
Digital matrix	Functionally the same, but significant code evolution transparent to user. Significantly lower power. Fix anomalies like power increase if comparators set to fired state. Fix injection delay variation across columns (2 ns uniformity).		6		
Pixel config.	Add bit masks or split load line to allow changing only certain bits? Use triple redundancy if area permits		9.3		
Global config.	modify some register assignments.		7.5		
Default config.	keep the same for pixels, change to low power mode for global.	8.1	11.3		
Inj. Fine delay	Increase from 4 bits to 5 bits?		5.4		
ToT	Enhance to 80 MHz count rate with 6 bit to 4 bit compression option	7.3	6		
Hit Or	keep same organization in matrix, but add significant processing in chip bottom. Precision ToT function. Self trigger function. Is an optional ToT cut needed (HitOr = Hit AND ToT_MSB)? Is separate trigger latency register for PTOT needed?	9.1 8.9	6.3	13	
Analog matrix	Add dedicated biases for 2 columns on left, 2 on right, 2 on top and 2 top corners (4 pixels each corner)	8.2	5.5		
Analog inj.	Improve injection voltage distribution to inject into more pixels at a time without systematic offset. Increase granularity (smaller v inj step). Also see Digital Matrix	8.12	5.4, 8.2		
Command input	Keep same protocol, but rearrange some commands. Add trigger tags. Add 2-level trigger. Fix minor bugs in exception handling. Add repeater for command input on general purpose LVDS.	8.10	9.2		
chip ID	Extended to 4 bits. Change broadcast ID from 1xxx to 1111.		9.2		
Trigger proc.	New trigger table and trigger processing state machine. Manages tag book keeping and 2-level. Allows clearing (event truncation). Add self trigger	6 8.6	7.4		
Data flow	Re-coded. Build in new encoding, hit filtering, 4 MHz readout capability, event truncation, etc.				
Data output	New stream encoding with binary tree compression of pixel address. Support for data merging. A separate output mode implementing RD53A encoding is NOT included, but debug mode desirable	6.1	9.4	10	
Data merging	Add new inputs to receive output data from others chips and merge into own output. Operation at 320Mbps (safe) and 640Mbps (goal)	6.1	n/a	11	
Event building	Full chip event building may not be possible for inner layer. Allow split events.				
Output protocol	Keep multilane AURORA. Make sure 3 lane mode works. Keep 4 output lanes.		9.4		
Service records	keep the same		9.4		

Gen. ADC	Keep the same. Add external input with reference current supply (can read NTC)		8.1		
T-Rad sensors	Keep the same IP, 3 in total: 1 close to analog ShLDO, 1 close to digital ShLDO, 1 in central location of chip bottom, near Ring oscillators	8.11	10.6		
T. sensors	Add "thin" resistive sensors for top/bottom differential measurement	8.11	10.6		
Ring osc.	Update design, tune to 1 GHz nominal, move to central location	8.7	10.7		
Cap measure	Move to chip bottom. Add internal clock generation and measurement		10.9		
E-fuses	Add e-fuses for serial number programming	8.5	n/a		
Gen. LVDS out	Make tristable. Add CMD repeat function	8.10	10.4		

Table 1: Index of modifications and additions to the RD53A design needed for production.

4. Dimensions and wire bond pads

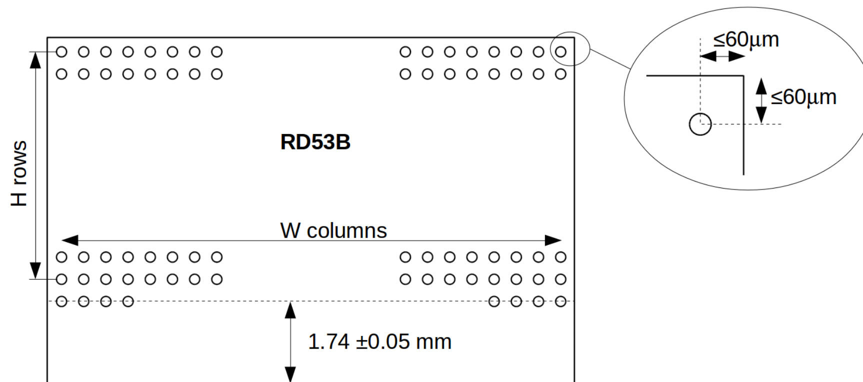

Figure 1: Diagram showing pixel matrix size and margin requirements. The 8 bump pads for sensor guard ring connection can be seen at the bottom of the bump matrix, four on the left and four on the right. They are on the same 50x50 grid as all other bumps.

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

Parameter	**ATLAS**	**CMS**
Pixel bump pitch	$50\,\mu$m $\times$ $50\,\mu$m	
pixel rows (H)	384	336
pixel columns (W)	400	432
Last pixel to chip edge	$60\,\mu$m	

Table 2: Chip size values.

80 number of pixels is W×H as shown in Table 2. 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 four on the left (right) should be connected to a wire bond pad on the left (right) that has no ESD protection or other contact with active circuitry. The bottom of chip height should remain within 100 μm of the RD53A value as shown in the figure. Also shown is the
85 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).

90 The wire bond pads should be increased from their RD53A size to allow wire bond rework. As this can affect performance due to the capacitance of I/O pads, the exact increase is to be determined by RD53 studies. A target size of 120 μm tall by 70 μm wide is suggested. The center to center pitch shall remain 100 μm.

5. Power, environment and operation

95 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 3 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
100 in Sec. 6 and 7). 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. It is also better understood that the detector design must pay attention to heat load non-uniformity (Sec. 5.1). Robustness against Single Event Latch-up (SEL) should be evaluated and built into the design. SEL issues have been seen in other chips in various technology nodes,
105 and the 65 nm bulk process used by RD53 is not inherently more or less SEL safe.

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
110 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
115 the current supplied to the full serial chain. Prescribing details of how this current distributes it to eventually end up powering individual pixels would unnecessarily constrain design options without any benefit to the experiments. Therefore, Table 3 gives only the serial chain current requirements 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
120 power, and so can afford to “burn” some extra power to meet the most demanding performance

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 5.3
Current transients	<5%	y	of chain I. See 5.2, **??**
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
Annealing	<1 month per 100 Mrad	N	only at 25°C with power off
Ionizing dose rate	<15 rad/s	N	delivered at -15°C.
TID gradient	$\leq$(3:1)	N	TID ratio across chip width or height
NIEL fluence	10^{16} 1 MeV n.eq./cm^2		
Flux of >1 GeV particles	<2×10^8/cm^2	N	
Flux of HEH particles	TBD	N	

Table 3: 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).. (*) The +250°C temperature will be experienced during solder bump reflow.

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.

5.1 Non-uniform heat load

The power regulators will be located at the chip bottom, distributed as evenly as possible along
125 the width. Therefore, the power density in the chip bottom will be significantly higher than in the pixel matrix, and this should be considered for cooling and resulting sensor temperature. While the regulators can shunt a significant current during start-up (before the chip configuration has been loaded) or in exception conditions, the main concern for operation is the steady state heat load in normal data taking. In this case the shunt current must be by design small compared to the
130 total chip current, but still the chip bottom power will be high due to the linear regulator drop out voltage. The extra power is due to voltage, not current. For a 300 mV drop-out, 1.2 V regulated internal voltage, and 300 mA periphery plus shunt standing current, the chip bottom will dissipate approximately 40% of the total chip power, yet it only comprises 10% of the chip area.

5.2 Serial chain operation, overcurrent and overvoltage protection

135 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 active suppression of current transients, given in Table 3 as a fraction of the serial chain current. (Fast transients (μs scale) must be suppressed with off-chip decoupling capacitors, as usual.) Transient suppression ensures the 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. The suppression should act on all transients, regardless of cause (including SEU).

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 (<ms 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 within the chip should not cause a high current state.

Since the maximum safe input voltage will be 2 V, and the voltage is dynamically generated in a serial power system depending on the supplied current, overvoltage protection should also be a feature of the production design. If too much current is supplied the voltage drop across the chip should be clamped at 2 V.

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.

5.3 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 3, as the
180 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
185 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 3. 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,
190 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.

195 **5.4 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
200 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
205 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.
210 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
215 special devices, such as parasitic bipolar transistors, are affected by NIEL.

6. 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	≤4 MHz	N	trigger only without readout
Trigger latency (L0)	≤12.5 μs	y	programmable
Auto read rate	<10 MHz	N	reads all triggers. output bandwidth permitting
Manual read rate	≤1 MHz	N	read . L1 latency after trigger
Readout latency (L1)	≤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 4: 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
220 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
225 erased from the pixel buffers. Table 4 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
230 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
235 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.

6.1 Data output

240 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 or 640 Mbps, so that one chip can be configured as master and either 1 or 3 chips as

slaves. 320 Mbps Must be guaranteed to work, while 640 Mbps is a best effort requirement. The data from each slave chip is to be combined by the master, along with its own, into the master's normal AURORA output.

The RD53A chip hit data encoding had 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 by at least 20% in order to fit the number of data links allocated in the experiment designs. A new encoding has been developed for RD53B and simulated but both experiments and exceeds this 20% improvement. The details of this encoding are documented in the RD53B manual and considered accepted. The RD53A output encoding will not be maintained as an optional output mode, as this would require substantial duplication of logic and buffering in the chip bottom. However, a debug mode is desirable, in which each tag would be followed by internal counter or register values before the start of hit data. At a minimum these should be the bunch crossing counter value and the trigger counter value. The bandwidth utilization of this extra information is not a concern since this mode would only be enabled as a debugging feature.

6.2 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.

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 some large fraction of 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. Since the purpose of the filter is to attenuate backgrounds so that they do not significantly increase the data rate, it need not reject 100% of the background but just a large fraction.

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 **??**.

7. Analog performance requirements

These requirements have been covered by the front end selection review process, which the experiments have been involved in. The selection of front end an direct interaction to define changes needed relative to RD53A take the place of a detailed requirements document. The selected front end will be documented in the RD53B manual and the measured performance in RD53A and test chips is documented in the review report and review materials.

7.1 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 1-2% 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 of these two categories. 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.

7.2 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 $100\,\mathrm{ke}^-$ or even $1\,\mathrm{Me}^-$. 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. The goal is $<1\,\mu$s recovery time following a charge up to $1\,\mathrm{Me}^-$.

7.3 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 $3\,\mathrm{ke}^-$ minimum and $12\,\mathrm{ke}^-$ 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 (pixel dead-time).

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.

8. 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 3.

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.

8.1 Start-up and default configuration

The startup serial current of Table 3 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).

8.2 Edge pixels

Edge pixels 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 rows or columns.. The chip design must provide dedicated bias control to the two columns on either side of the chip, the two rows at the top, and the two top corners (4 pixels in each corner). 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, six independent bias controls are needed, as shown in Fig. 2.

In addition to independent biasing, it is desirable to have separate threshold DAC's for the left and right columns (not possible for top), so that the threshold can be set differently for edge pixels if needed.

8.3 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.

8.4 Design For Test

Effective structural testing of the 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

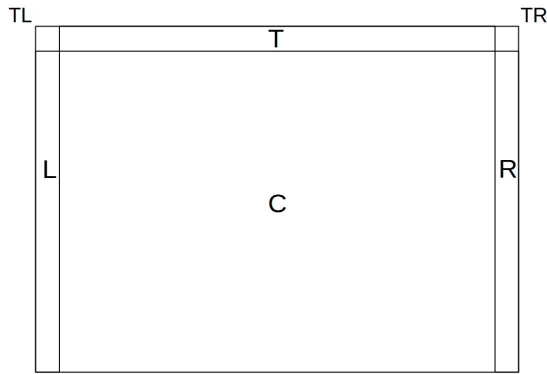

Figure 2: Diagram showing the six zones for independent pixel bias adjustment to handle edge and corner pixels. They are: C=center, L=left, R=right, T=top, TL=top left, and TR=top right. The width of the edge zones can be two or four pixels, to be selected at layout time.

logic is left for the designers to define, with a strong requirement of total chip test time under 10 min. and a goal of under 5 min.

8.5 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.

8.6 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. Self trigger should have significant configuration options, like which core columns to include, which lines to combine, consecutive bunch crossings to trigger, programmable delay and wait time following a trigger, and ideally ToT and ToA programmable thresholds.

8.7 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. Readout via built in ADC is desirable.

8.8 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). The location of temperature, radiation and ring oscillators should be optimized to measure absolute values as wells as gradients across the chip (top/bottom and side/side).

8.9 Hit-OR outputs

RD53A has four differential outputs that allow separate HitOr maps in parallel. The outputs can be
380 reduced to one in RD53 if needed, with internal logic to manage the combination if the four maps
for output and for internal self triggering.

8.10 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. Ability to use a
385 general purpose LVDS output as a command repeater should be added.

8.11 Sensors

8.12 Calibration injection

Reduce LSB size. The minimum injection step size in RD53A is slightly too large for accurate
S-curve sampling. The RD53A column dependence of the injection pulse delay should be fixed.
390 The goal is 2 ns dispersion of the injection timing across the full matrix.

9. Additional features for RD53B

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
395 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

Some internal processing is needed to feed the internal self-trigger. For example allowing logic
combinations of the various Hit-ORs. This could allow for example triggerless luminosity mea-
400 surement.

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.

9.4 Ability to use special features during operation

405 Some of the test and extra features of the previous section could have interest physics benefits
if they can be used during operation. Special consideration should be given to this, and operation
compatibility should be implemented to the extent possible. New capabilities can lead to significant
physics gains and often all it takes is a little extra thought and care in the implementation of a new
feature. The example of precision ToT (PTOT) is given here.

410 The PTOT uses the HitOr outputs from each core column. During data taking, if all pixels are
included in the HitOr, those outputs will be always high even at the outer layer hit rate. On the

other hand, certain physics analyses search for highly ionizing particles, and depend on accurate measurement of high charges. In particular, charges that saturate the normal pixel ToT can be of great interest. By adding a new mode on the HitOr signal, where instead of the raw pixel hit output being ORed, only those outputs exceeding ToT cut are included, can enable measurement of high charges during operation. Because the high ToT hit rate is very small compared to the total hit rate, these high charge HitOr outputs will not longer be saturated and the PTOT circuit will function well. At the same time, since multiple high charge hits in coincidence are very unlikely, it will be easy to associate the PTOT hits with the correct pixel hits.

However, because applying a ToT cut to the hit outputs effectively introduces a delay equal to the number of clocks needed to reach the ToT cut value, the PTOT hits will have a shorter trigger latency than regular hits. Therefore, to make this mode usable (which would be very interesting for physics), the PTOT modules should use a different (separately programmable) latency value than the pixel matrix. These features are not a complex to implement, but without them the PTOT could not be used during data taking and would not open up a new physics capability. This illustrates how a small amount of special consideration can turn test features into new capabilities at almost no cost.

Other features to consider using during operation include the TOA capability and the self-trigger. If The hit signal is ANDed the negated pixel ToT counter enable signal, only a very short leading edge pulse will be produced, allowing to measure time of arrival at fairly high data rate. Thus a self-trigger on significantly delayed hits could be implemented. Small hits delayed by time walk will generally not spoil this function, because they will be neighboring a large hit and the TOA of the large hit will be seen too.

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).