RD53 Pixel chips for ATLAS and CMS phase II upgrades

VERTEX 2022, Tateyama, Japan

Jaya John John (Oxford) on behalf of the RD53 Collaboration

25 October 2022

Outline

- The RD53 collaboration
- Chip design specs and generations
- Chip details
  - Layout
  - Analogue blocks
  - Digital functionality
  - Protection against single event effects
- Test results: irradiation and single event effects
- Summary and outlook

The RD53 collaboration

- Joint effort from ATLAS and CMS institutes to produce Pixel chips for Phase II upgrades, for installation during Long Shutdown 3
- 24 institutes, started in 2013
- In 65nm CMOS, developed a library of radiation-hard analogue IP blocks, through a series of test chips
- Chips developed: see slide 5

Collaboration board chair:
Lino Demaria, Torino

Interface to experiments: Co-spokespersons
Jorgen Christiansen, CERN (CMS) ,
Maurice Garcia-Sciveres, LBNL (ATLAS)

Experiment observers
Duccio Abbaneo, CERN (CMS) ,
Kevin Einsweiler, LBNL (ATLAS)

RD53 design framework: Co-ordination: Bari

Floorplan/integration: Bari

Analog front-ends:
CMS/linear: Bergamo/Pavia:
ATLAS/differential: LBNL

Monitoring: CPPM

IO Pad frame: Bonn

Digital:
RTL, Design flow, P&R, Timing:
Torino, Pisa, CERN, LPNHE (Paris)
Simulation & Verification:
CERN, Bergen, Oxford
Design for testability:
Bari
SEU/SET simulations:
CERN, Seville

Serial Powering:
Dortmund, ITAINNOVA

IPs: Support and possible updates
Current DAC: Bari
Voltage DAC: Prague
Bandgaps: Bergamo
ADC, mux, temp, radiation: CPPM
PLL & serializer: Bonn
Differential IO: Bergamo/Pavia
Power on reset: Seville
Ring oscillator: LAL
Analog buffer: RAL

Testing

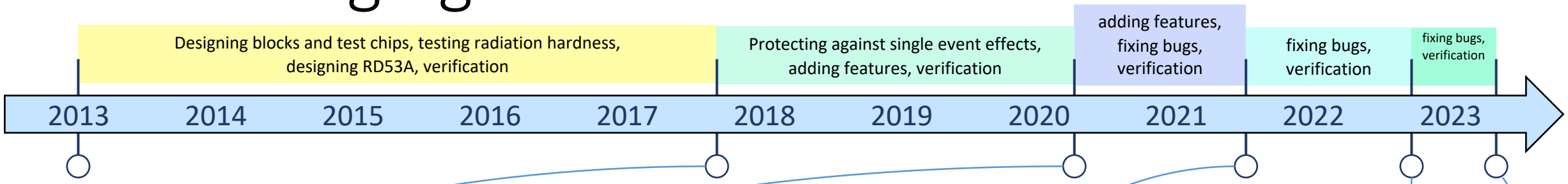
Organization: ATLAS: LBNL, CMS: CERN
Many RD53 and ATLAS/CMS groups: LBNL, Bonn, Oxford, CERN, CPPM, LAL, Torino, Aragon, ETH, Florence, Zurich , , ,
RD53 test systems: YARR (LBNL), BDAQ53 (Bonn)

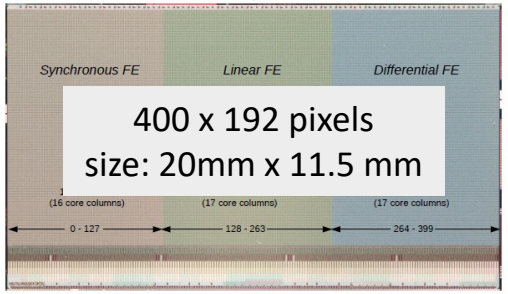
Design requirements and chip specifications

Parameter	Value (ATLAS / CMS)
Max. hit rate	3 GHz/cm^2
Trigger rate	1 MHz / 750 kHz
Trigger latency	12.5 µs
Pixel size (chip)	50 x 50 µm^2
Pixel size (sensor)	50 x 50 µm^2 or 25 x 100 µm^2
Pixel array	400 x 384 pixels / 432 x 336 pixels
Chip dimensions	20 x 21 mm^2 / 21.6 x 18.6 mm^2
Min. threshold	600 e-
Radiation tolerance	1 Grad over 10 years (previously 500 Mrad)
Power delivery	Serial powering
Power	< 1W/cm^2
SEE tolerance	SEU rate, innermost: ~100Hz/chip

SEE = Single Event Effect

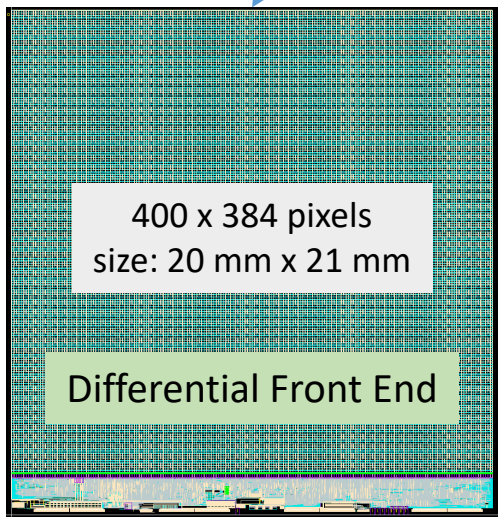
RD53 design generations





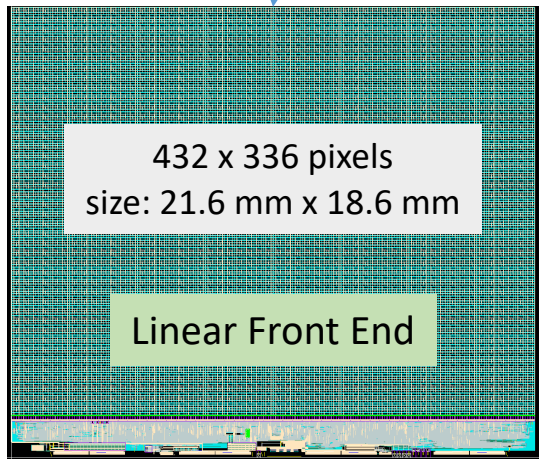
RD53A

- **Half-size demonstrator for:**
  - **evaluating alternatives:**
    - 3 analogue front ends
    - 2 readout architectures
  - **prototyping Pixel modules**
- submitted in August 2017



RD53B-ATLAS (ITkPix-V1)

- **For: ATLAS pre-production**
- Submitted in March 2020
- In October 2020, a mask respin was submitted to patch a serious bug in the pixels' Time over Threshold latches



RD53B-CMS (CROC-v1)

- **For: CMS pre-production**
- Submitted in June 2021

RD53C-ATLAS:
Q4 2022
For: ATLAS production (ITkPix-V2)

RD53C-CMS:
Q1 2023
For: CMS production (CROC-v2)

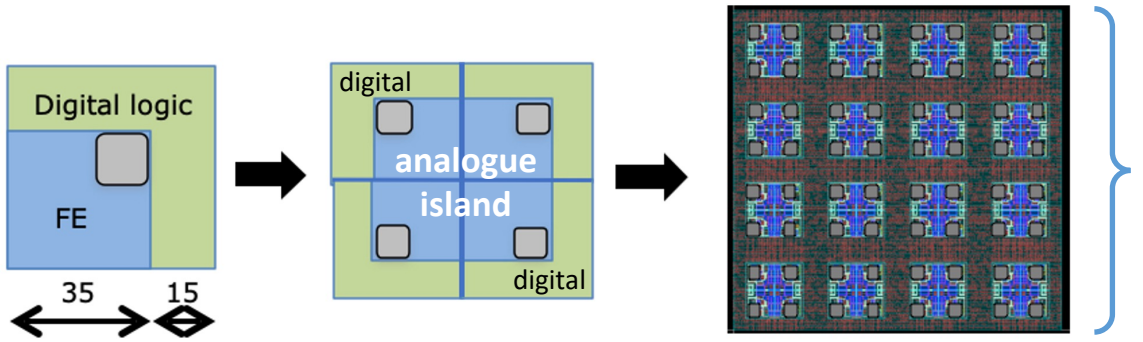
Production chips are the same size as pre-production, but with finalization of features and bug fixes.

RD53 chip layout

1 pixel,
50 x 50µm

4 pixels
= Pixel Quad

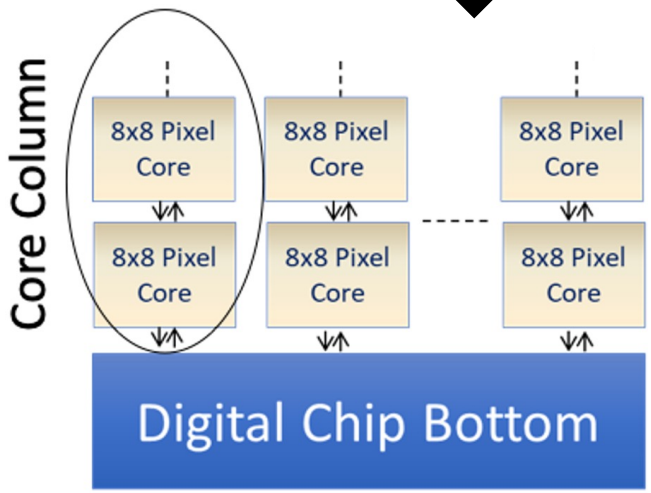
8 x 8 pixels
= 1 Pixel Core



fundamental tiling unit

Pixel Cores are grouped into Core Columns

Each Pixel Core is identical

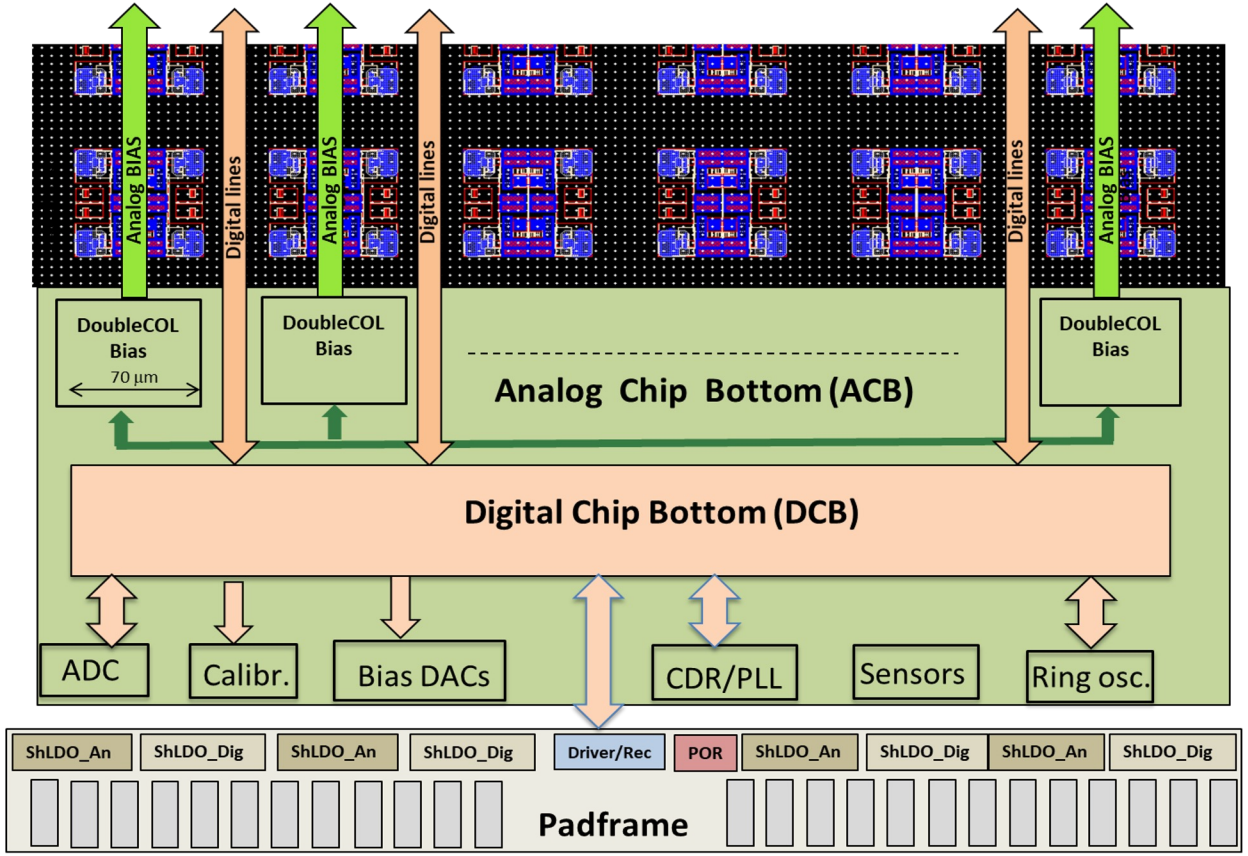


Below the pixel array is the periphery and pads:

- Analogue IP blocks — see slides 7 to 11
- Digital circuitry — see slides 12 to 18

Analogue bias and digital signals flow vertically along Core Columns

- Pixel Cores are abutted to minimise routing



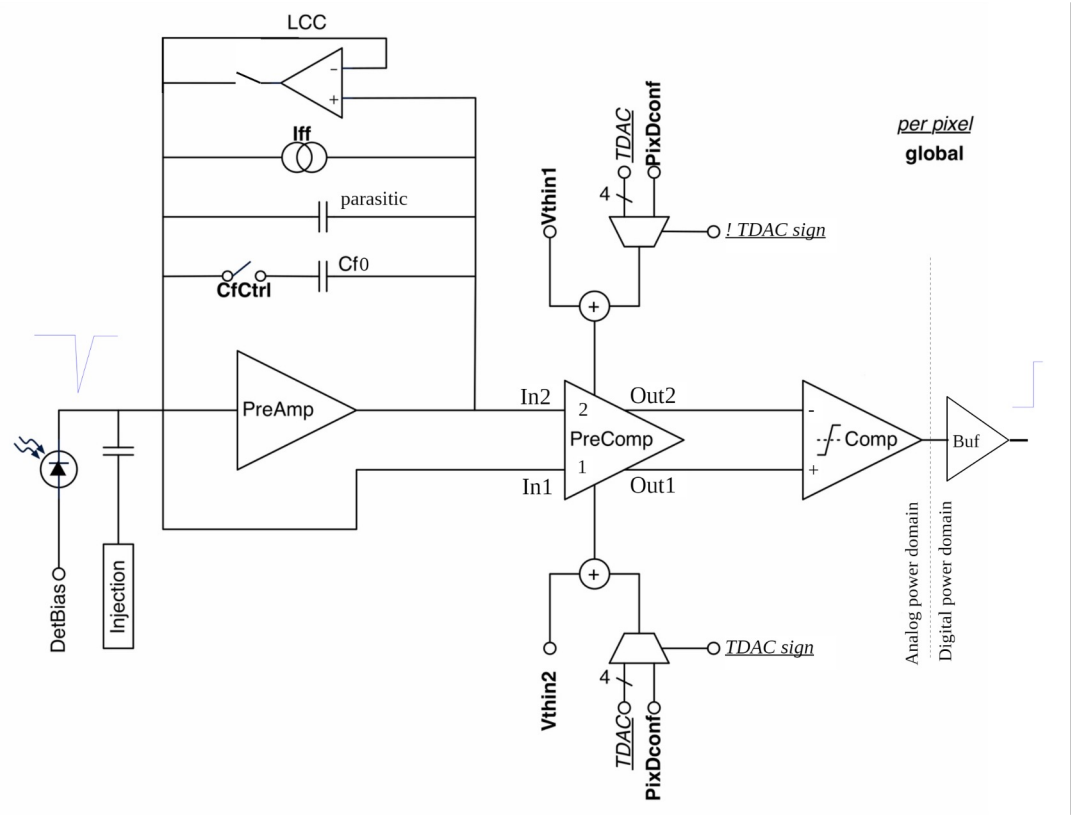
Analogue blocks: overview

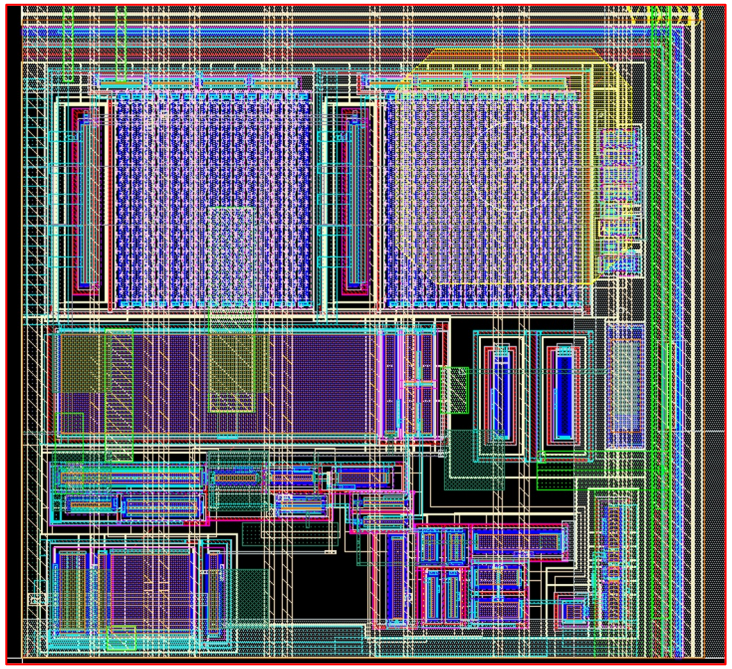
Block	Description	Slide
Analogue front end	The ATLAS chip versions use a differential front end. The CMS chip versions use a linear front end.	8 9
Shunt LDO	Enables start-up and serial powering. Constant input current shared between chips, modules on serial chains. 1 LDO for digital power, 1 LDO for analogue power.	10
Clock & Data Recovery (CDR)/PLL	Recovers a 160MHz clock and command/trigger stream. The PLL generates internal clocks: 160 MHz, 64 MHz, 640 MHz and 1.28 GHz.	11
Bias circuit	Provides biases to the pixel array. Based on bandgap references.	
Calibration circuit	Injects hits into the pixel array, to calibrate its response.	
Monitoring block	Digitises analogue quantities using a voltage mux, current mux and 12-bit ADC	
Temperature and Radiation sensors	Temperature sensors: polysilicon resistors. Radiation sensors: based on PMOS devices with a linear variation in voltage in the dose range 10 - 1000 Mrad.	
LVDS pads/drivers	Pads and drivers for differential inputs/outputs	

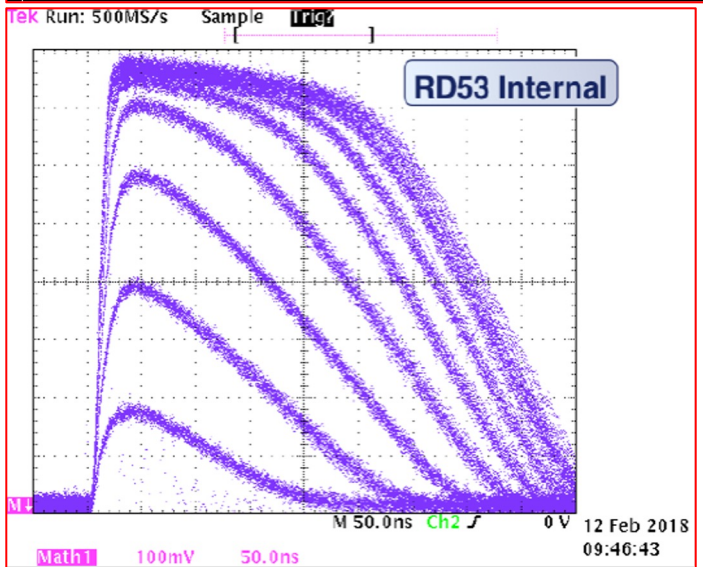
PLL = Phase Locked Loop

Analogue front end: differential (ATLAS/ITkPix)

- Charge sensitive amplifier
- Leakage current compensation circuit
- Continuous reset integrator, with tuneable feedback current (global setting)
- DC-coupled pre-comparator stage:
  - 10-bit DAC for global threshold.
  - 4+1 bit local trimming DAC for threshold tuning.
- Fully differential input comparator

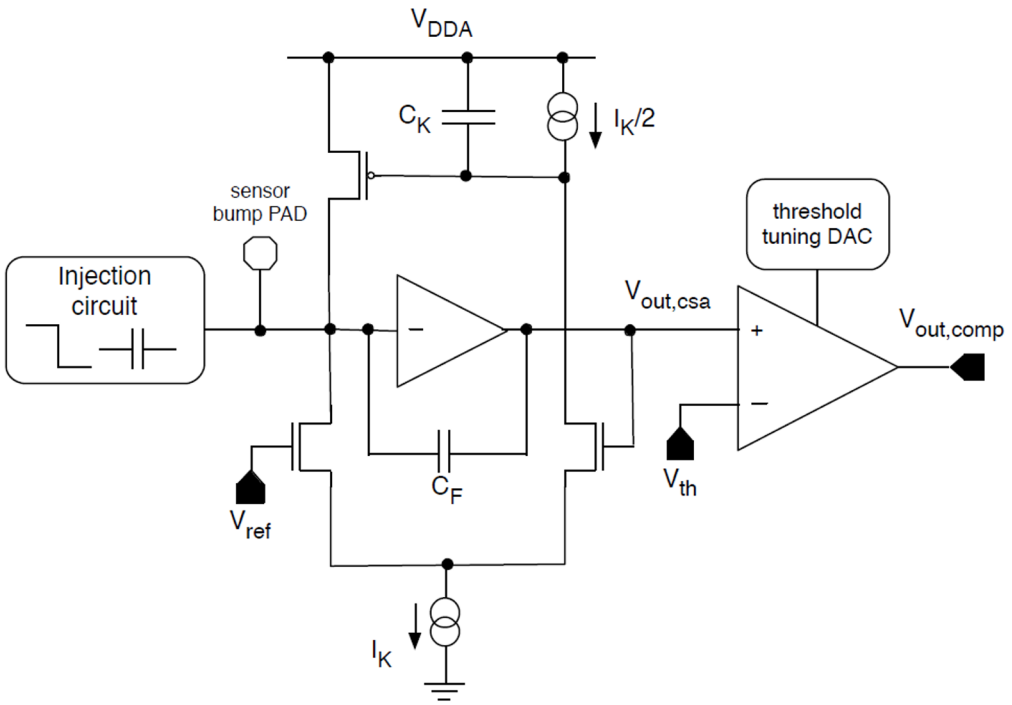


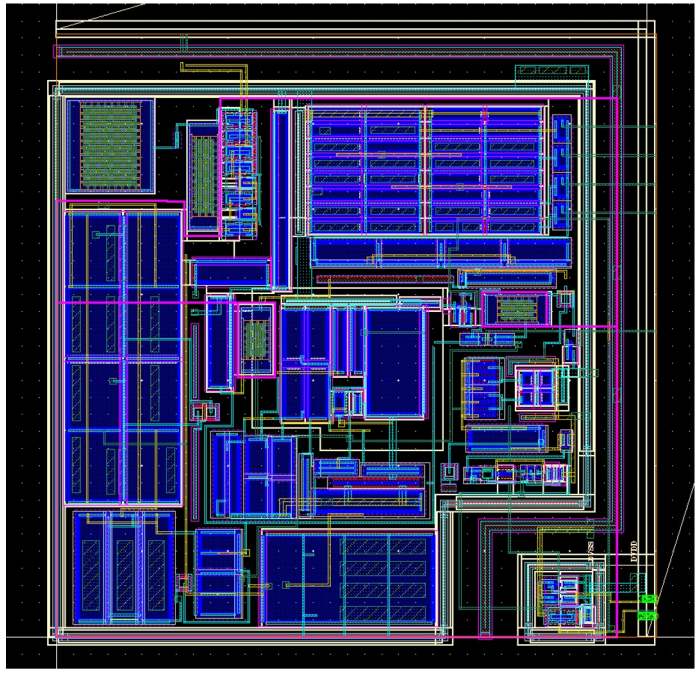


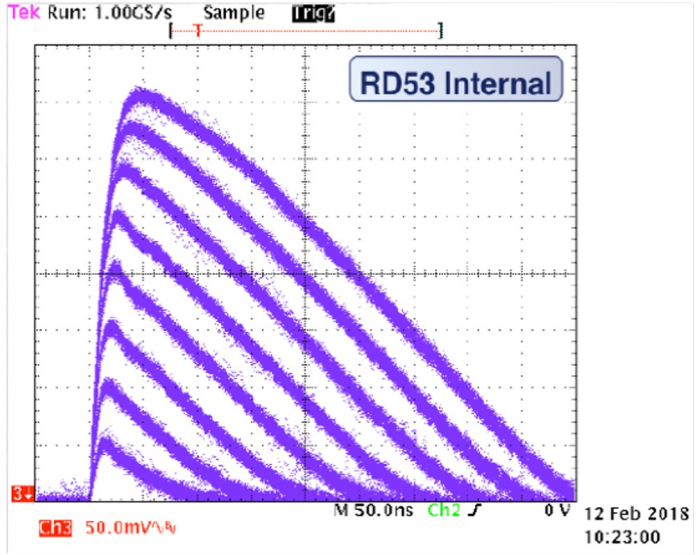


Analogue front end: linear (CMS/CROC)

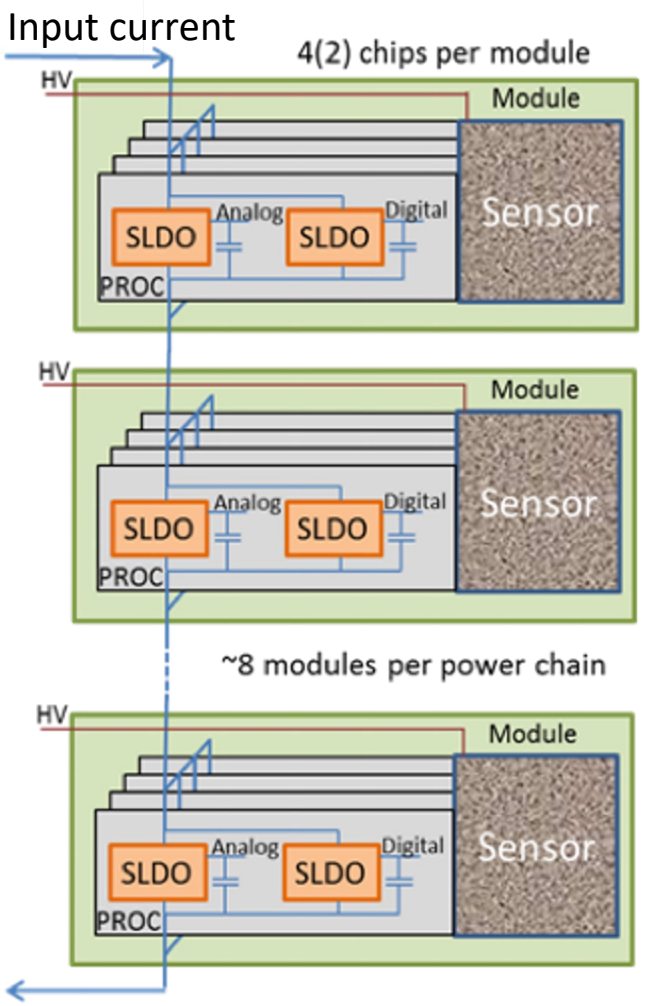
- Charge sensitive amplifier
- Krummenacher feedback for return to baseline and leakage current compensation
- Comparator:
  - 10-bit DAC for global threshold
  - 5-bit local trimming DAC for threshold tuning







Shunt LDO – for serial powering



- Serial powering: a constant input current passes through all modules in a chain (typically 8 modules)
- On the module, the current is shared between 2 to 4 chips
- Each chip has 2 ShuntLDOs: 1 for analogue, 1 for digital
- In case of chip failure, the other chips of the module need to absorb its current in their ShuntLDOs
- Features:
  - On-chip regulated supply voltages, low noise
  - Protection against overvoltage and under-shunt conditions
- Design challenges solved through simulations and iterations (partial list):
  - Radiation hardness
  - Start-up when cold
  - Accuracy of current measurement

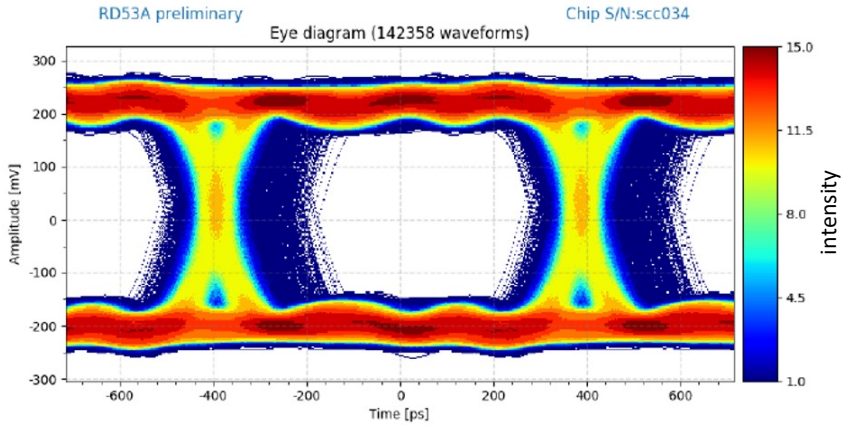
LDO = Low DropOut regulator

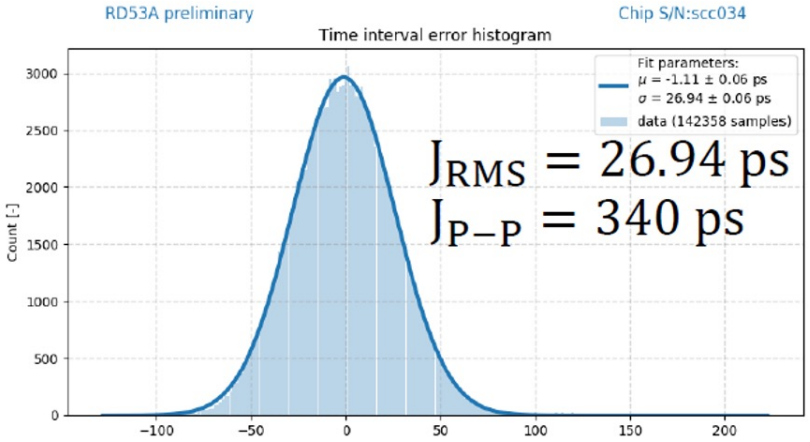
Clock and Data Recovery (CDR) / PLL

- Role: recover the 160MHz clock and command stream from the transitions on the Control link input (see slide 13)
- The PLL then generates internal clocks: 160, 64, 640, 1280 MHz

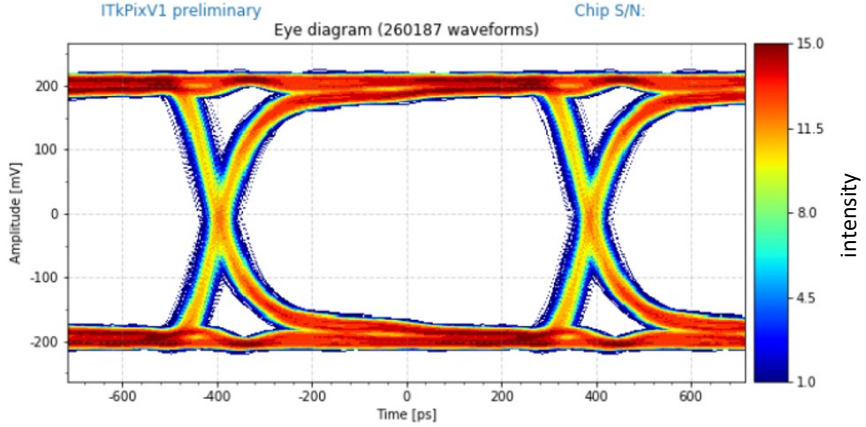
CDR/PLL greatly improved compared to RD53A:
- Reduced jitter
- Better start-up reliability

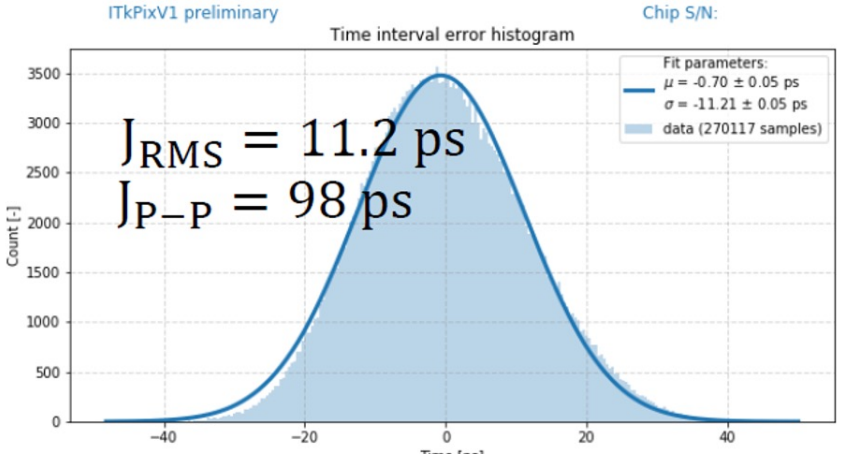
RD53A





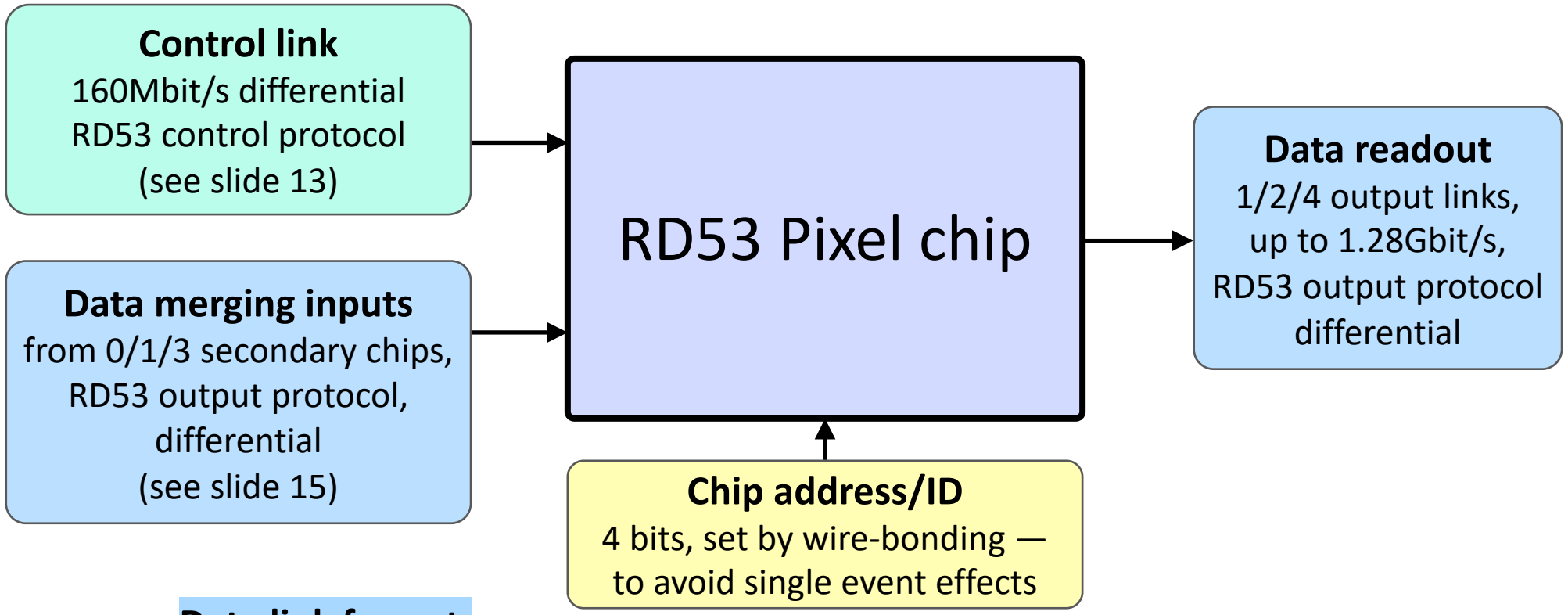
RD53B-ATLAS / ITkPix-V1





PLL = Phase Locked Loop

Digital chip interfaces: overview



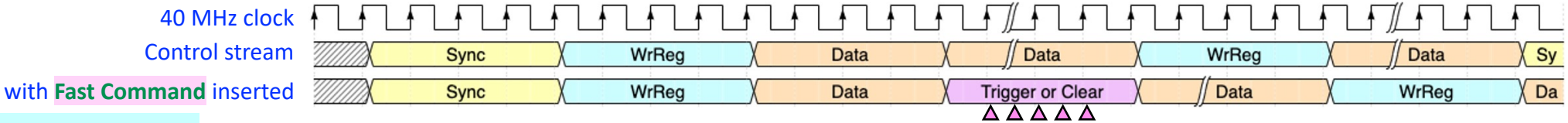
Data link format:

- Aurora 64b/66b encoding
- Inside the Aurora frames: custom RD53 output protocol (see slide 14)
  - This represents both Physics events (hits) and register readout data
- Used for both the chip readout and the data merging inputs

Chip interfaces: control link

Fast Commands = 16 bits, 100 ns to send
Slow Commands = variable length, up to some ms to send

- The RD53 control protocol was designed to meet these needs:
  - **High radiation environment:**
    - The protocol symbols are a Hamming distance of 2 bit flips apart, allowing error detection
    - To speed up recovery from errors, the **Clear** command, which resets the datapath, is a minimum-sized **Fast Command**
    - Fast Commands can be interleaved by DAQ during **Slow Commands** such as Write Register



    - Write Register commands have features to auto-increment register addresses and to broadcast to up to 15 chips on a shared bus, to speed up chip configuration
    - These together mean that the chip can be continuously reconfigured in ~100ms, call it 5 times/s
  - **High trigger rate:**
    - Trigger commands are minimum-sized **Fast Commands**
  - **Minimise cabling and material in the detector:**
    - Both commands and the main 160MHz clock are recovered from this 1 differential pair
    - The periodic Sync command guarantees sufficient transitions to recover the clock

Chip interfaces: data output and formats

Hit data readout and compression

- Triggers cause hit data to be read out from the Pixel Array using tokens.
- Hit data from columns is compressed using Binary Tree Encoded (up to factor of 2)

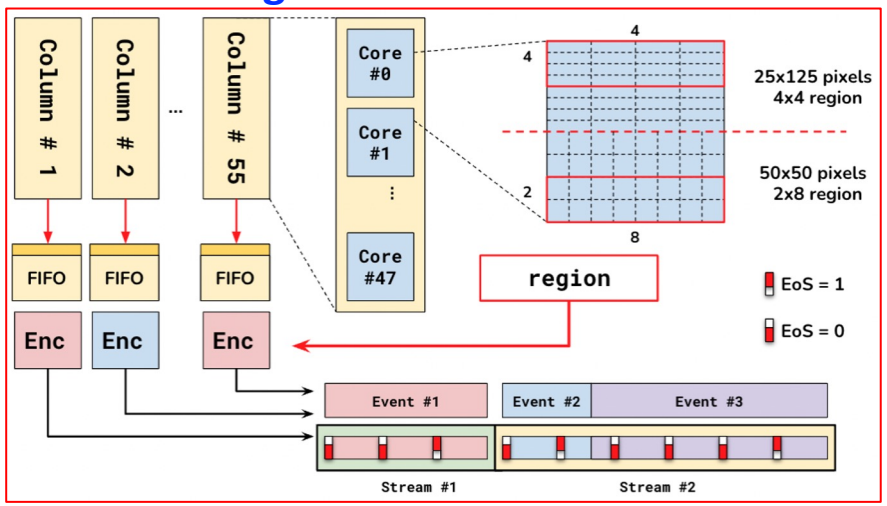
Event building and Streams

- All hits belonging to 1 trigger are formed into 1 event, by Trigger Tag
- Events are encapsulated into a variable length Stream
- Streams allow splitting of data over multiple Aurora frames of 64 bits

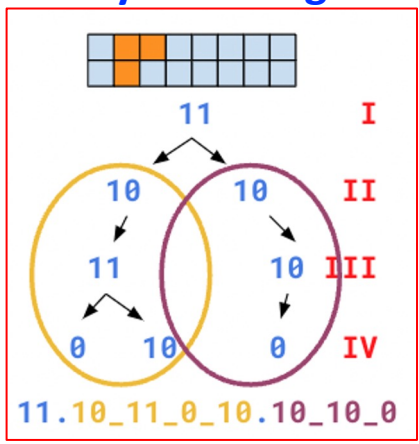
Formatting within Streams

- Streams are split into 64-bit words (Aurora pixel data frames)
- The first bit shows if the current frame is the last of the steam (End of Stream/EoS bit)
- The next 8 bits are the Event Tag
- Next: 6-bit address of the core column (“ccol” in the diagram)
- Next: 6-bit address of the core row (“crow”)
- The core is divided into regions of 2x8 pixels.
- Each region’s hit data is Binary Tree Encoded (“BT” in the diagram)
- Next: the Time Over Threshold data (4 bits) for each hit pixel (“ToT”)
- Adjacent events are separated with a 3-bit separator (111)
- After the last event, any unused space in the final Aurora frame is filled with zeroes.

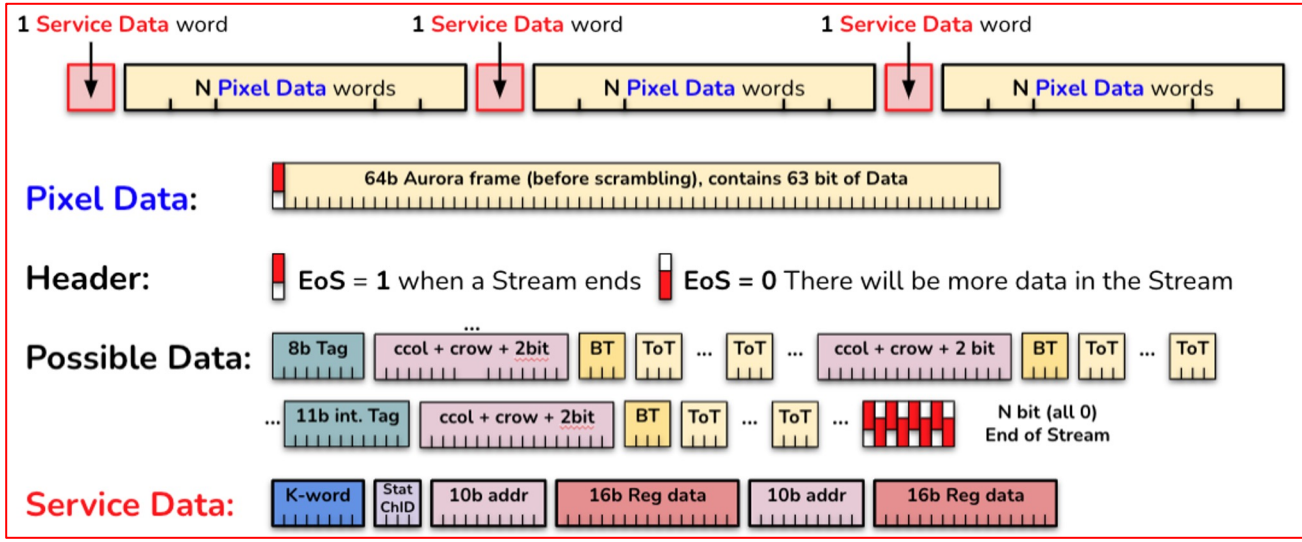
Event building



Binary encoding



Encoding of Physics and Register (“Service”) data



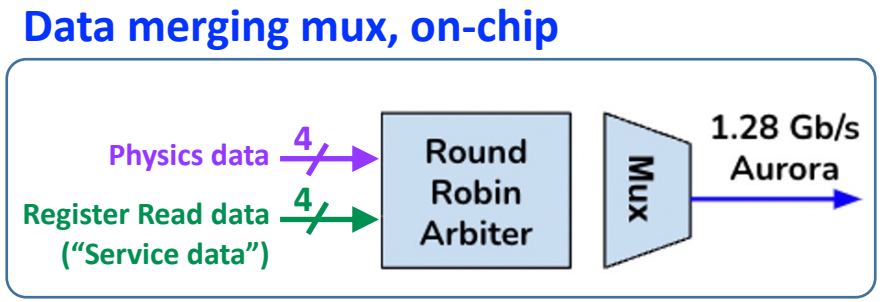
Chip interfaces: data merging

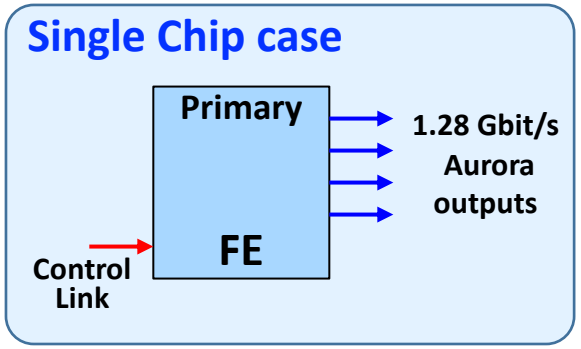
Data merging is used to optimise the number of links sent off a module.

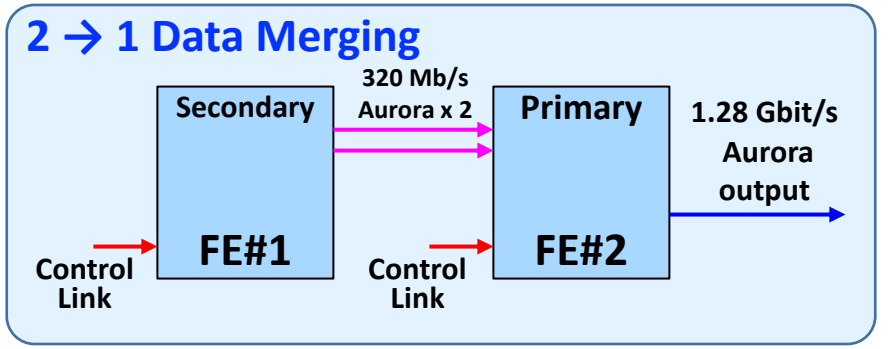
In the outer layers, if there were one 1.28 Gbit/s per chip, it would be under-utilized.

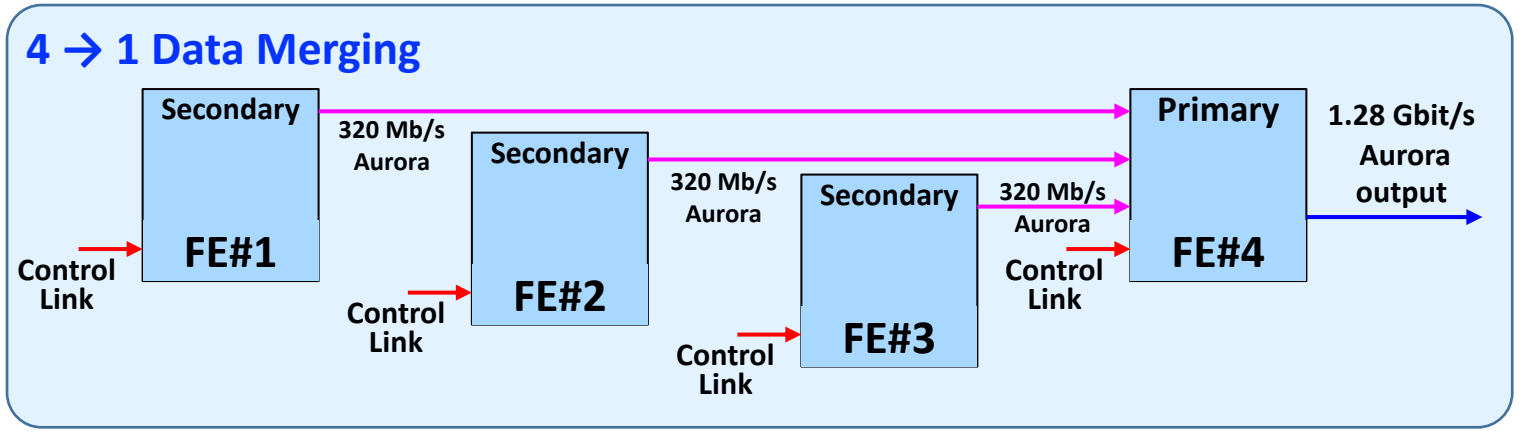
There are 2 defined modes of data merging, per the figures.

Data from the secondary chip(s) is merged with data from the primary chip in a simple **round robin** (top right)









Digital: hit processing and data reduction

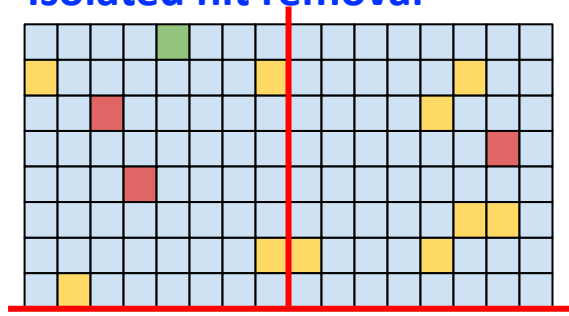
There are several ways to **reduce** the amount of Physics data:

- **"Binary readout":** leave out the Time Over Threshold bits, ~30% bandwidth saved
- **Event truncation:** for each Core Column, if the size of the event is over a threshold, trim it
- **Timeout truncation:** for each event, if it takes too long to read out, trim it using a timeout
- **Isolated hit removal:** remove single pixels from events, favouring clusters

There are **debug options** to add information:

- **Raw output:** output a simple hit map (do not use the Binary Tree compression)
- **CRC:** append a 32-bit Cyclic Redundancy Check to the end of a Stream
- **BCId and Level1 ID:** add 16 bits containing the Bunch Crossing ID, or Level 1 ID, or both.

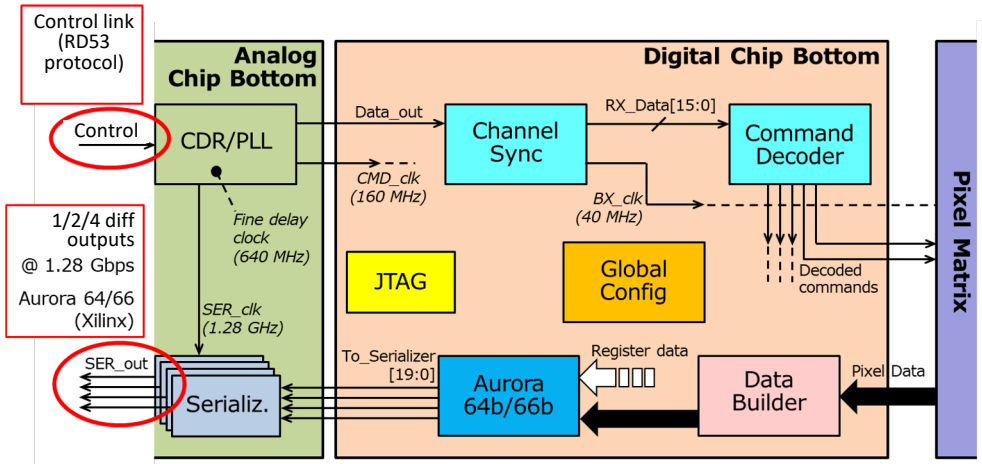
Isolated hit removal

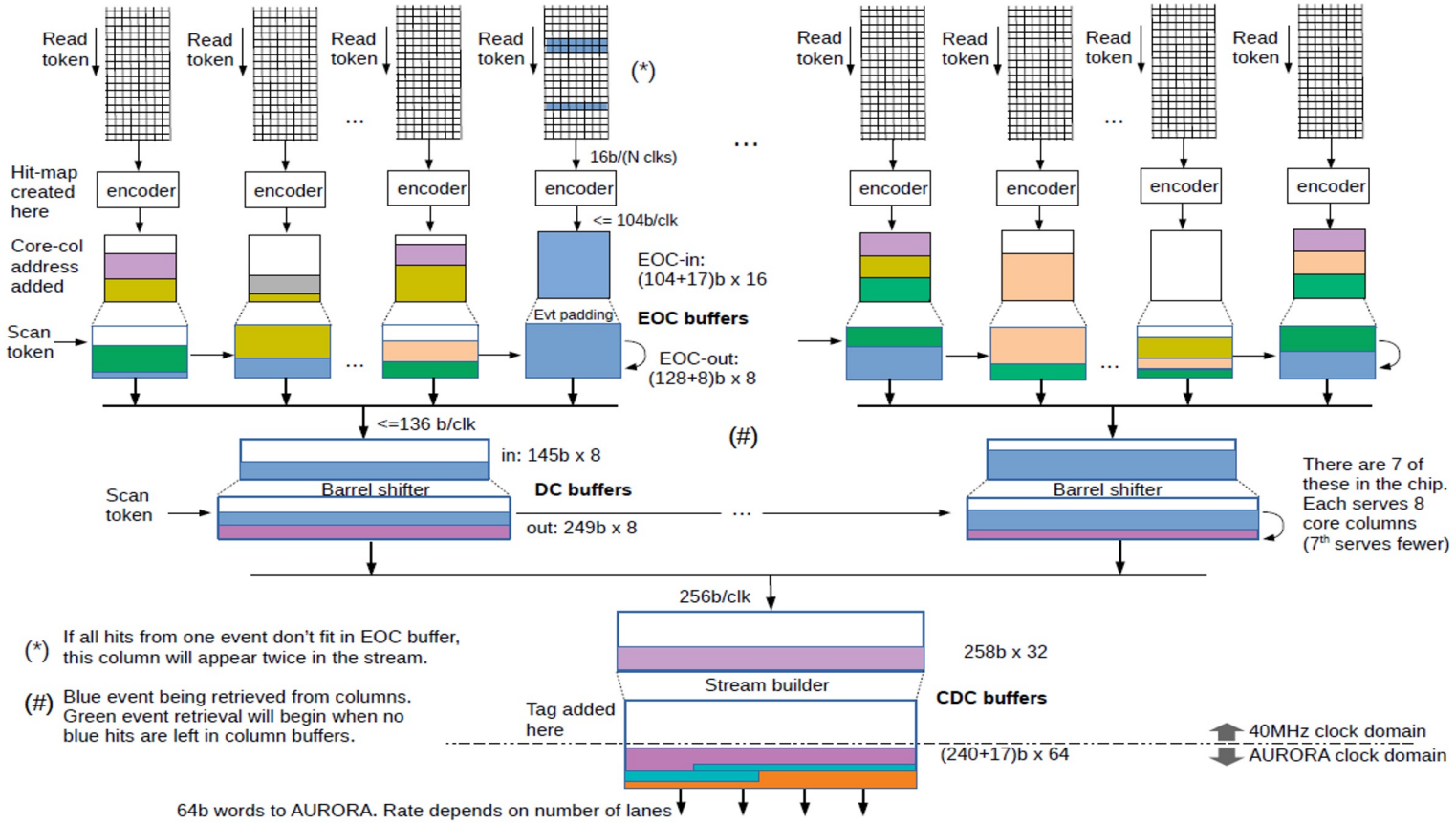


Pixels in red are considered isolated.

It is only easy to look within the data for one Core Column at a time, so pixels touching vertical boundaries are not removed.

Digital: data flow





- **Hits in the Pixel matrix are stored locally** upon hit arrival and **associated to a timestamp**.
- 6-bit counter with only 4-bits sent out from the matrix.
- **Each pixel has 8 Time Over Threshold (ToT) memories, each 4 bits**.
  - Option for 6 to 4 bit mapping (dual slope)
  - Counting with 40 MHz or 80 MHz clock
- Time stamp **memory shared between 4 pixels**
- **Token based readout** of hits in parallel for each CoreColumn as soon as a valid Trigger is received
- **Multiple levels of data processing**, event building, data buffering and formatting
- Registers are read on command and periodically **("Service data"),** interleaved with Physics data in a set ratio (default: 1 in 50)
- **Aurora frames are built** and sent to the high speed serializers

Digital: clocks and resets

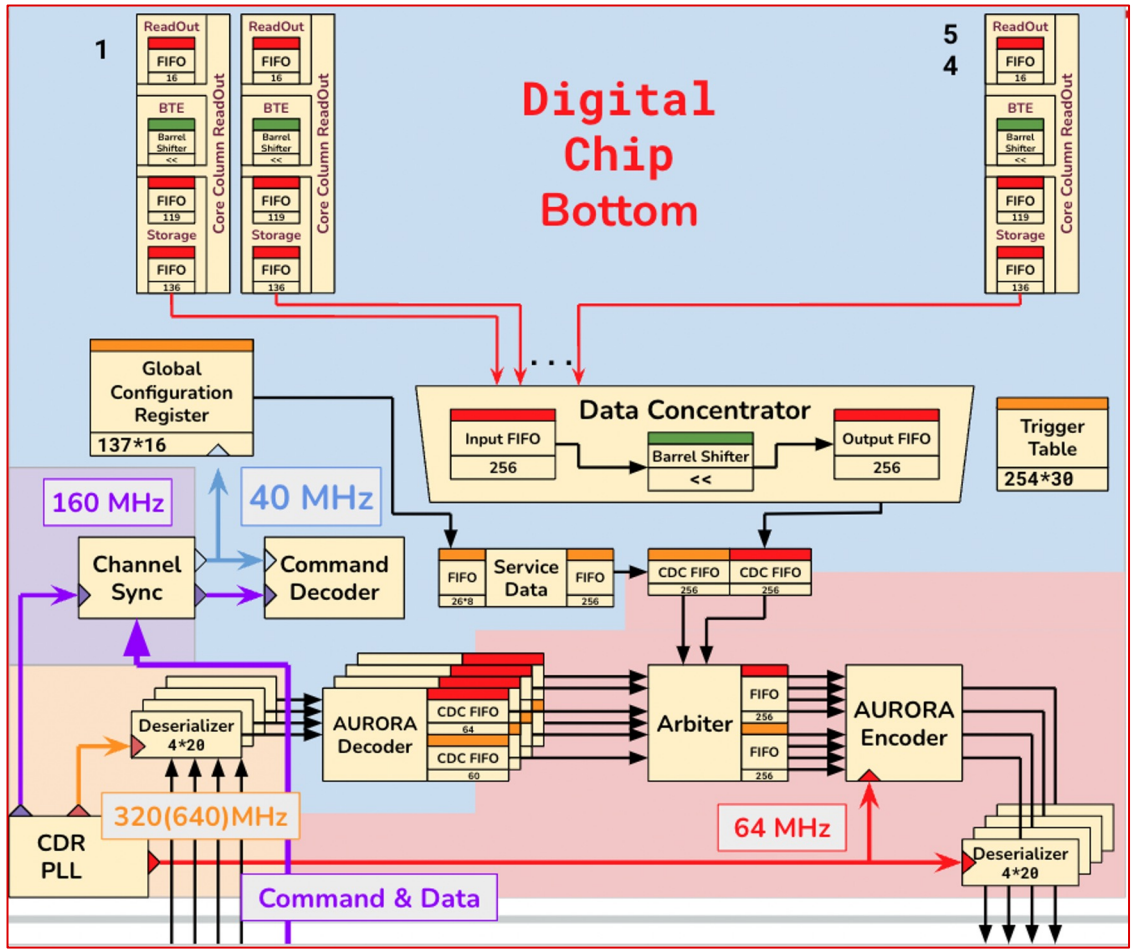
Clock generation and distribution:

- **CDR/PLL** recovers the 160 MHz main clock from the control link and generates internal clocks (160, 64, 640, 1280 MHz).
- The Channel Synchronizer generates the 40 MHz master clock.
- 4 different clock domains are dealt with using **Clock Domain Crossing** design techniques – checked with formal tools.
- **Clock gating** is used to reduce power consumption, for blocks not in use and in the Pixel Matrix.

Reset strategy:

- To minimise risk of false resets due to SEEs:
  - There is no reset pin, nor Power-Up Reset
- 3 kinds of reset are possible, however:
  1. Individual blocks can be reset using the **Global Pulse** command
  2. The **Clear** command resets the datapath
  3. On the control link, if **Sync** commands are not sent for ~1µs, the CDR/PLL is reset, allowing resynchronisation.
- All pixel configuration and global registers are supplied from muxes. At power-up, the muxes supply hard-coded defaults. Then DAQ may programme all configuration to the intended values. To switch the muxes to use the new values, magic numbers need to be written to two key registers.

Data flow and clock domains



Single event effect (SEE) protection

Single Event Upsets (SEUs) = permanent bit flips in memories

Single Event Transients (SETs) = temporary inversion of a signal, returning to its original state

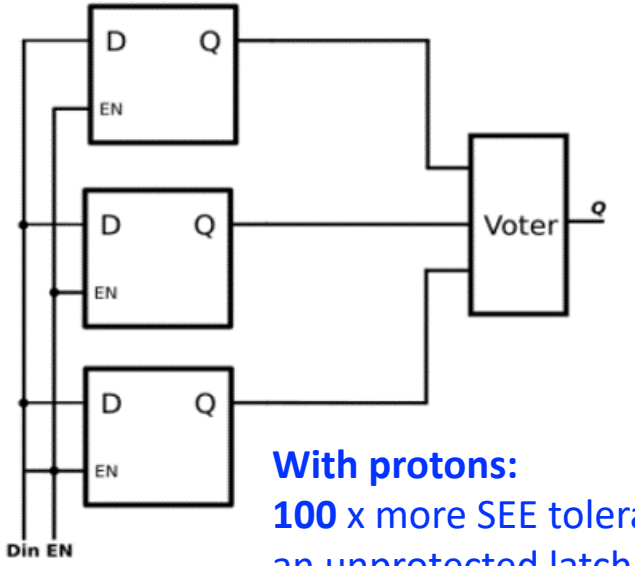
The issue: for the innermost layer, the expected rate per chip of Single Event Upsets (SEUs) is 100Hz – must protect!

Principle: protect vital circuitry and data as far as feasible within space limits

Priorities: protect configuration data, state machines, memory pointers

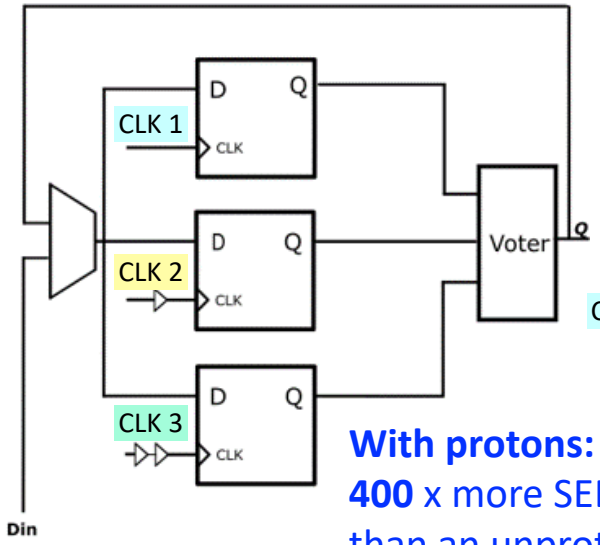
At these densities of single event effects, no protection is perfect so the command protocol includes a fast **Clear** of the datapath and provision for **continual reprogramming** of global registers and pixel configuration (slide 13)

Pixel Configuration

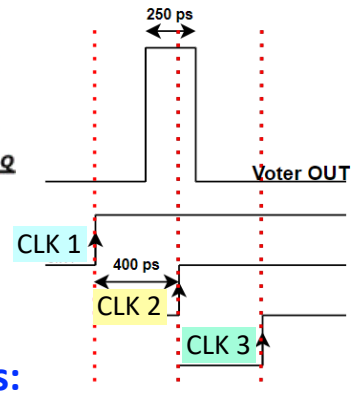


With protons:
100 x more SEE tolerant than an unprotected latch
No feedback – lack of space

Global Configuration (registers)

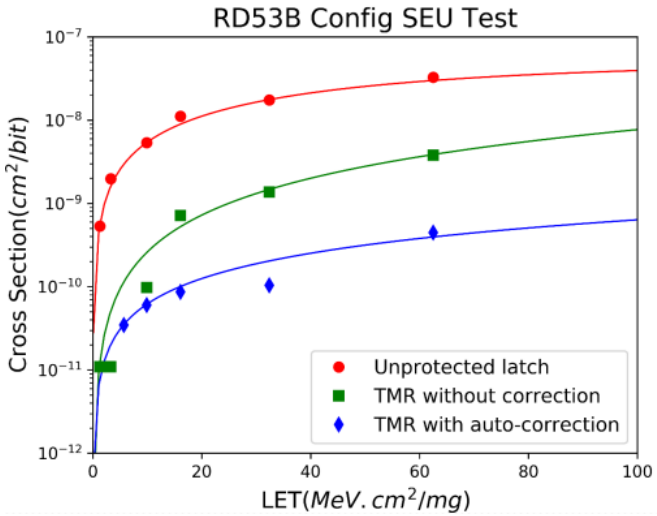


SET filtering



With protons:
400 x more SEE tolerant than an unprotected latch
Has **feedback** for correction

SEU cross-section comparison from heavy-ion testing:



Digital verification: approach

- RD53 verification framework replaced since June ’21 to reflect state of the art:
  - Metric-driven verification: simulate until the desired functional coverage is reached
  - Functional coverage means: the features of the chip
    - Use this in addition to the usual metrics such as: lines of code covered and logical expressions covered
  - Use constrained randomization to test all feasible combinations of configuration and inputs
    - Randomize sets of tests and learn which set is optimal (fewest runs) to reach coverage goals
  - Previously we had good randomization of hits and triggers (to evaluate architectures) but we needed to also randomise configurations — among many improvements

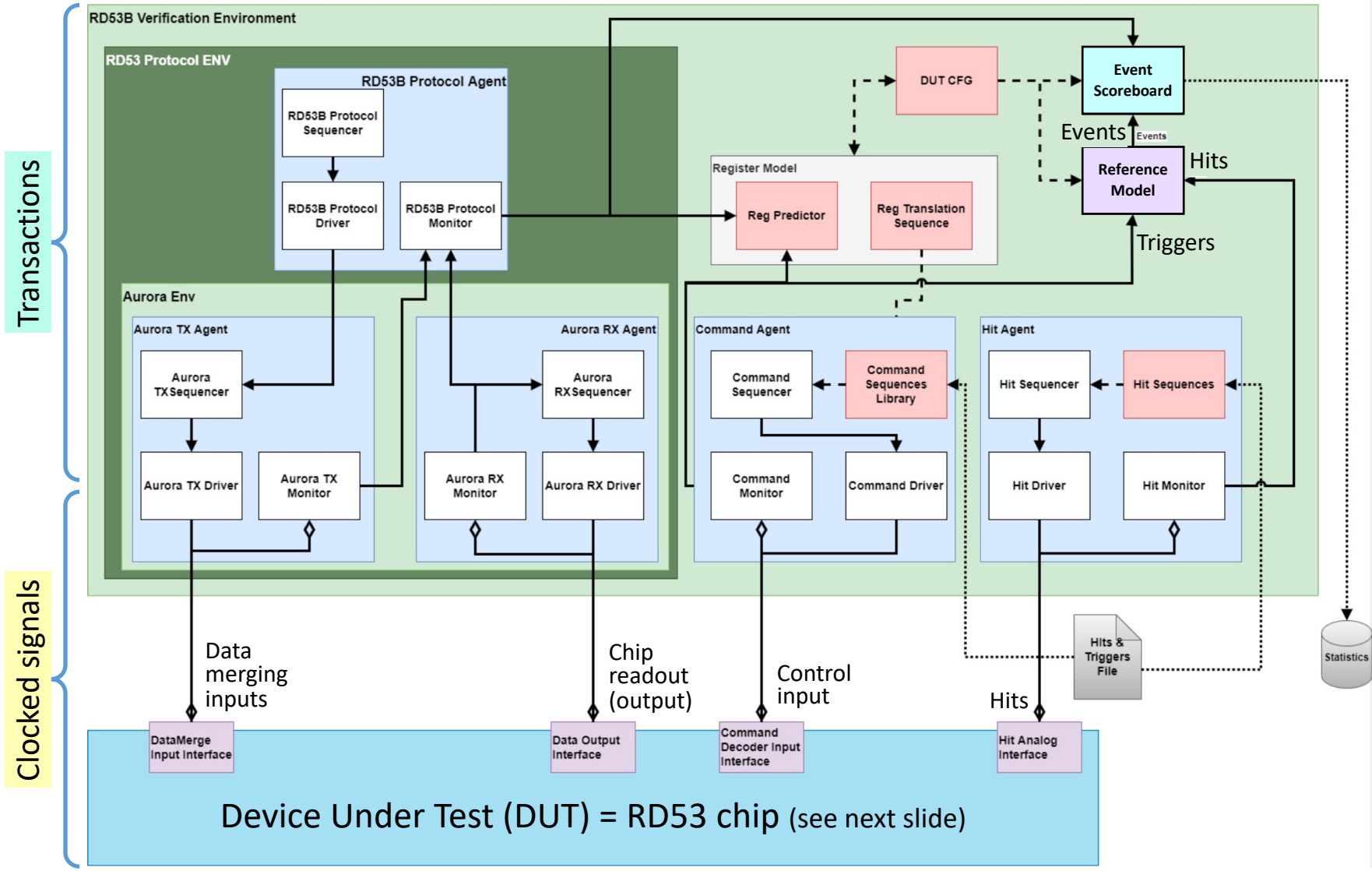
- Verification standard used:
  - Universal Verification Methodology (UVM) [IEEE 1800.2-2020]
  - Industry standard, well-supported by tool vendors

Digital verification: framework

How it works:

- Overall goal: verify that chip outputs match outputs predicted by the verification framework, for randomised configurations and inputs
- For each chip input, have a **driver** which converts from transactions* to clocked signals (*to reduce simulation time)
- For each chip output, have a **monitor** which converts from clocked signals to transactions
- Present the same inputs – register writes, triggers, hits, data from other chips – to both the simulated chip (Device Under Test/DUT) and the **Reference Model**
- Strings of chip inputs are called **sequences** and are "played out" by **sequencers.**
- Compare the events in the chip output to those predicted by the **Reference Model** in the **Event Scoreboard**

RD53 verification framework (UVM)



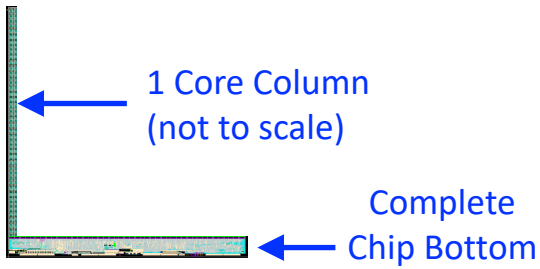
Digital verification: devices under test

- To manage simulation time, want to simulate using the smallest possible “chip” while meeting functional coverage goals
- Do this by simulating a range of Devices Under Test:

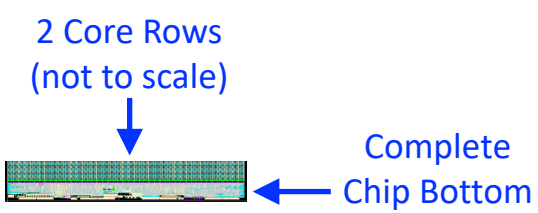
Slower to simulate →

Slower to simulate ↓

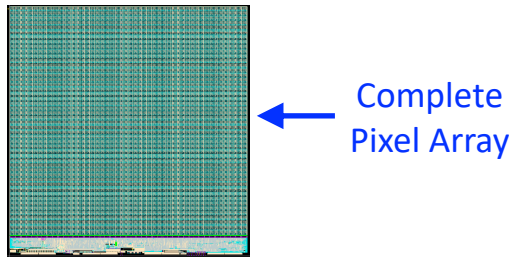
RTL Single Core Column
8 x 384 pixels
Obtain: majority of coverage



RTL “Baby Chip”
400 x 16 pixels
Exercise all columns



RTL Full Chip
400 x 384 pixels
Exercise filling of FIFOs



Gate Level Single Core Column
8 x 384 pixels
Obtain: additional timing checks

Gate Level “Baby Chip”
400 x 16 pixels
Exercise all columns

Gate Level Full Chip
400 x 384 pixels
Exercise filling of FIFOs

RTL = Register Transfer Level, before any placement or routing takes place, faster to simulate
Gate Level = after placement and routing, includes parasitics, slower to simulate

Test results: irradiation of AFE

- Performed threshold scans after irradiation to 1 Grad for a large range of parameters, to compare with pre-irradiation measurements
- Number of failed pixels < 5% for most parameter settings

→ **Chip operational at most chip settings up to 1 Grad**

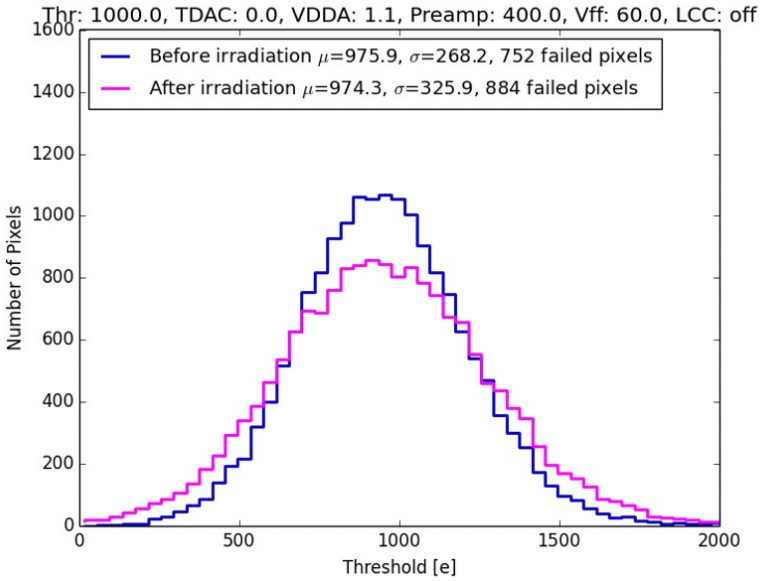
- At chip settings given by simulations, decrease analog voltage (VDDA) on irradiated chip

→ **Even after irradiation to 1 Grad, ITkPixV1 is operational down to 1.0 V**

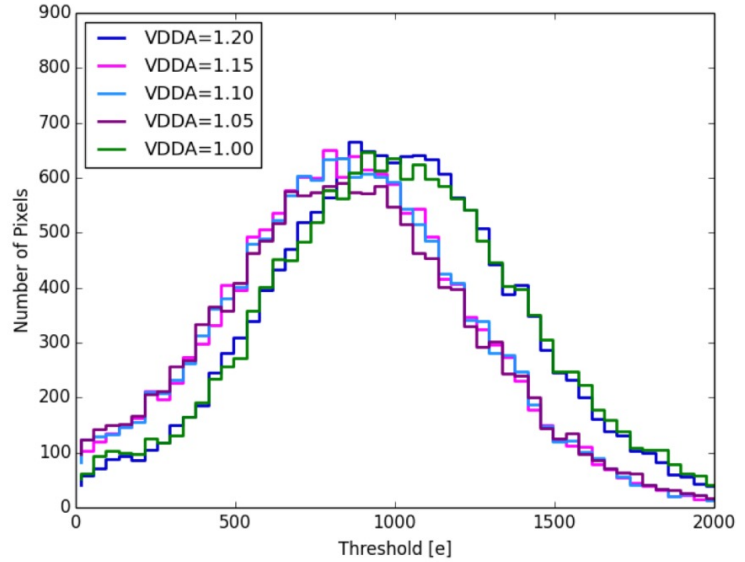
→ Significant improvement compared to RD53A

- No issues observed in tuning thresholds after irradiation without sensor

Threshold distributions **before** and **after** irradiation to 1 Grad



Threshold distributions with reduced VDDA @ 1 Grad



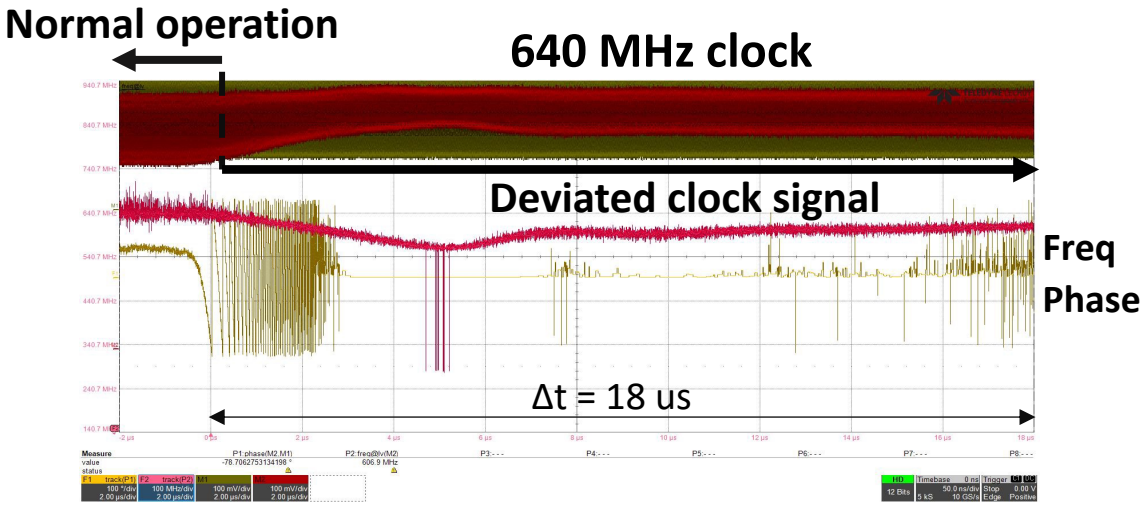
Test results: Single Event Effects (1/3)

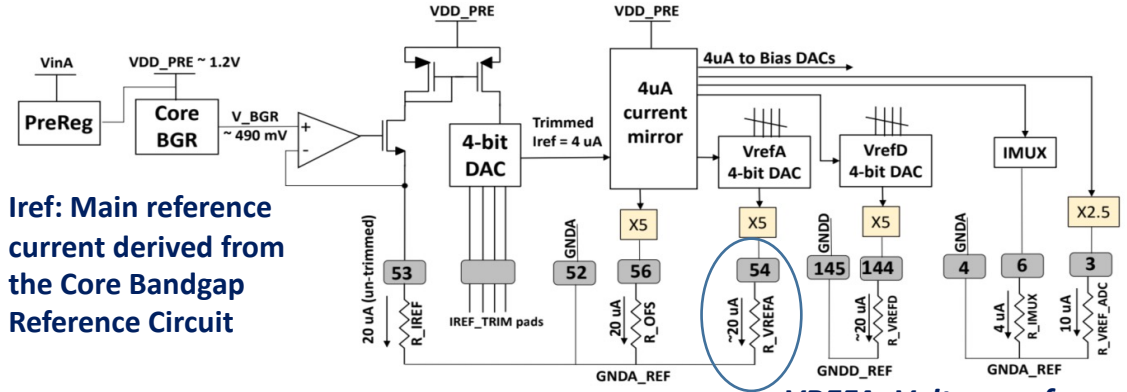
- ItkPix-V1 was tested with 480 MeV protons (TRIUMF, Canada), heavy ions (CRC Louvain) and 24 GeV protons (PS, CERN)
- Campaigns largely successful but observed repeated readout dropouts, causing DAQ resync
  - Projected to occur 5 times/s, innermost layer

Suspicion: CDR/PLL bias looks to be varying.

The CDR is powered from the analogue regulator

- whose voltage derives from the analogue reference voltage
- which in turn derives from the main current reference and the Core Bandgap





Iref: Main reference current derived from the Core Bandgap Reference Circuit

VREFA: Voltage reference for analog bias (regulator input)

Readout issue investigations with laser injection

How to localise the issue?

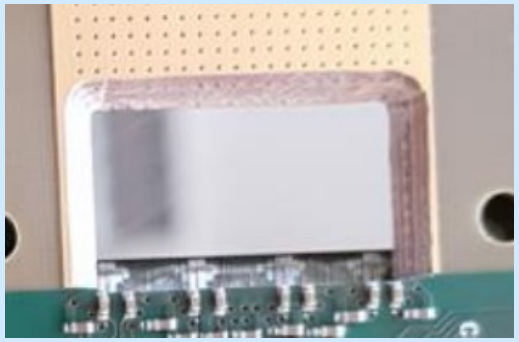
- Two Photon Absorption allows a geographic scan across the chip to identify sensitive nodes
- Uses near-infrared, focussed through substrate
- Charge collection occurs only at beam focus
- Requires optical access to the back of the chip

Chip Preparation

Cutting the copper from the testing card to expose the chip bottom (all ROIs in the chip bottom)

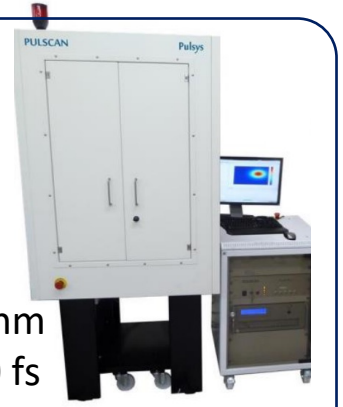
Shifting the chip up by extending the wire bonds

Ready for laser injection



Two-Photon Absorption (TPA) laser setups

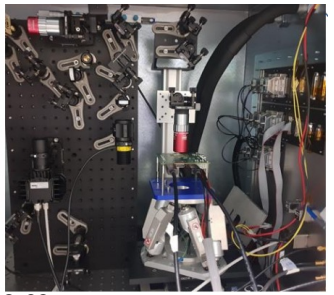
@ KU Leuven, Campus Geel



PULSCAN
- Wavelength 1550 nm
- Pulse duration 450 fs
- Pulse energy up to 50 nJ

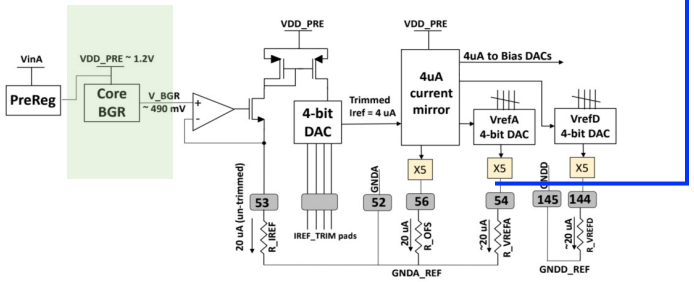
@ CERN

Many thanks to the SSD lab of the EP-DT group (CERN)

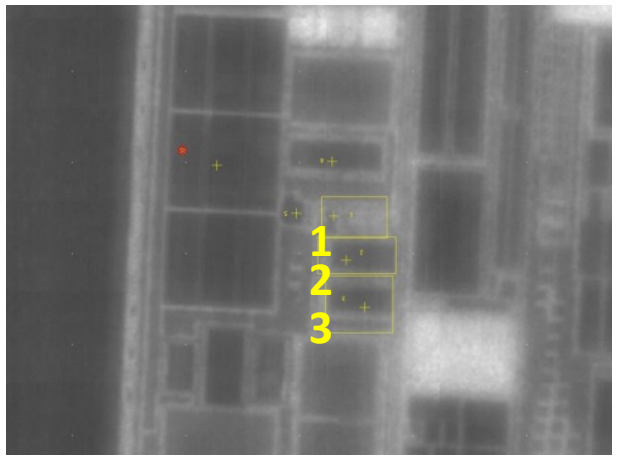


FYLA
- Wavelength 1550 nm
- Pulse duration 430 fs
- Pulse energy up to 2.2 nJ

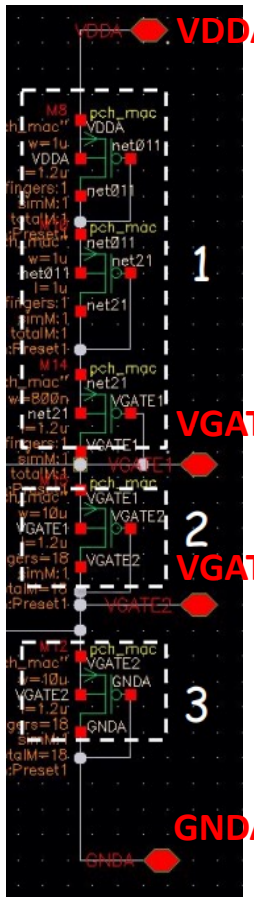
Issue identification and resolution

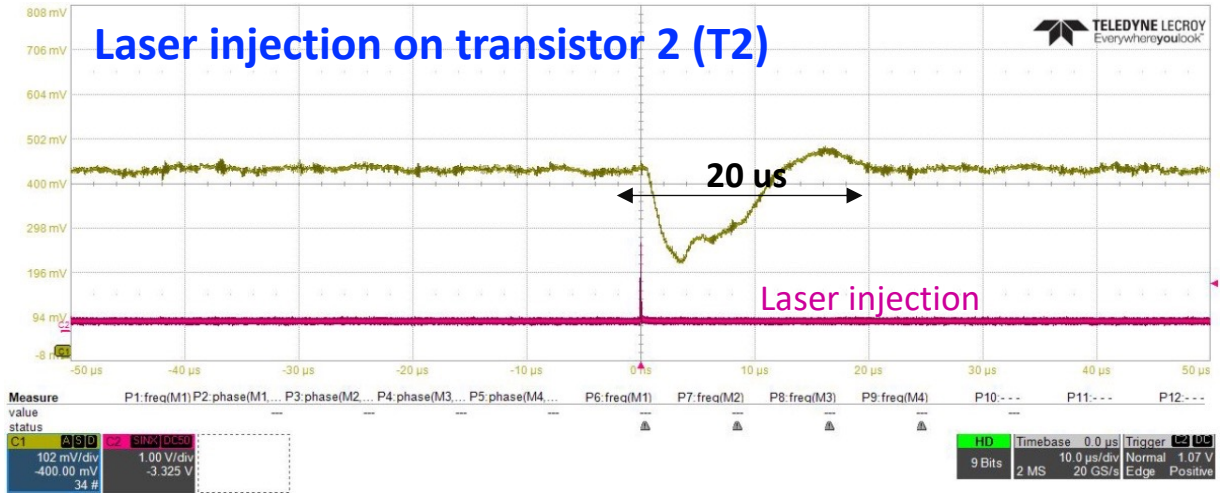


IR image of the Bandgap Reference circuit. Sensitive areas (transistors) are marked.



VREFA





- Laser injection shows the Bandgap Reference is the origin.
- Problem is reproduced in the Single Event Transient SPICE simulations
- SEE Compensating capacitors are added (VGATE1 to GDNA, VGATE2 to GNDA)
- SET simulations show that the revised design mitigates previous SET sensitivity
- RD53C is now expected to have a much lower cross-section of dropout of the readout

The critical IPs in the analog chip bottom are all characterized against SETs by laser injection and the expected SET robustness is confirmed.

Summary and outlook

- RD53 is a collaboration of ATLAS and CMS institutes to develop Pixel chips for the High-Luminosity upgrade of LHC, working together for 9 years
- RD53 has developed, in 65nm, radiation hard:
  - building blocks
  - a demonstrator chip (RD53A)
  - and pre-production chips (RD53B) for ATLAS and CMS
- We plan to submit the final production chips (RD53C) for:
  - ATLAS in Q4 this year
  - CMS in Q1 2023

Thank you for listening. Questions?