Friday 19 May 2022

CMS OT ASICs DESIGN METHODOLOGY

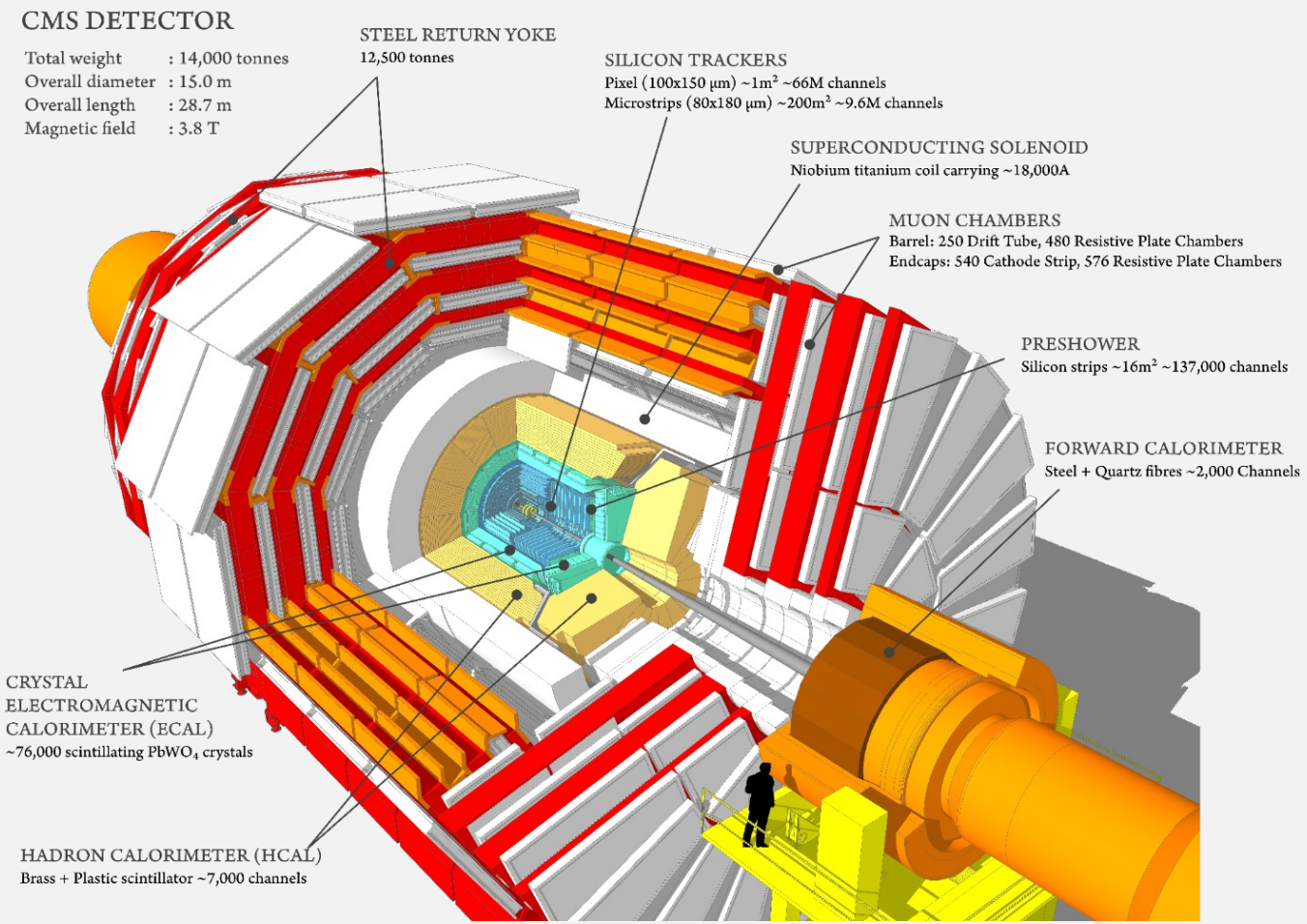
HEP-IC (High Energy Physics Integrated Circuits) workshop

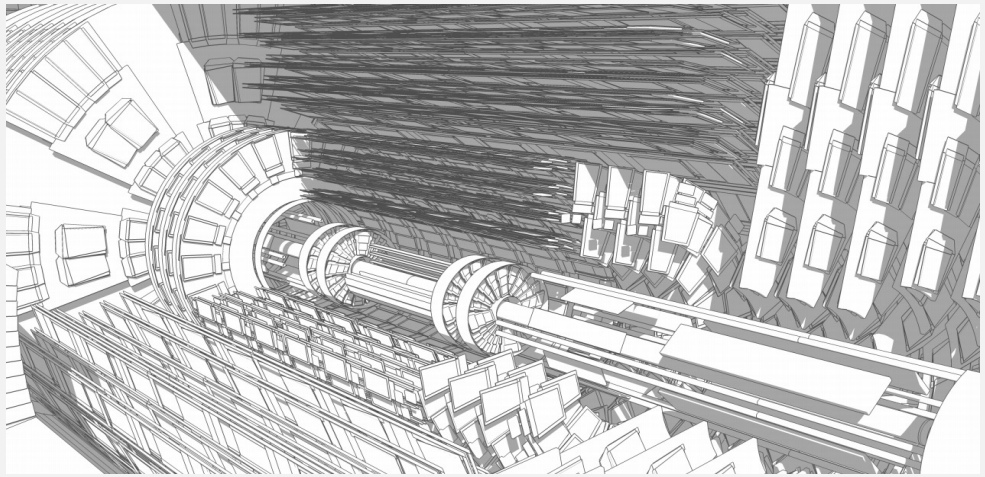
Alessandro Caratelli

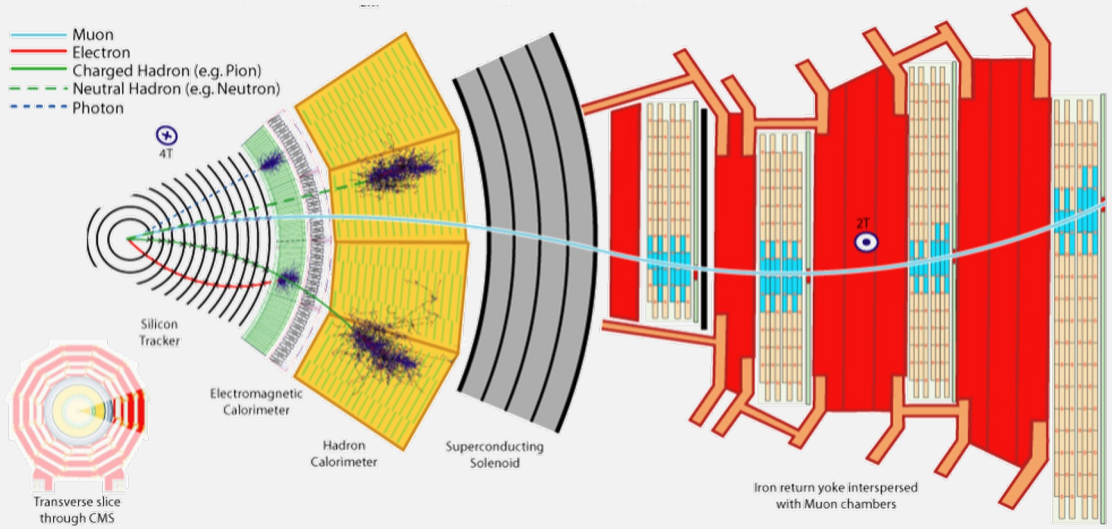
on behalf of the CMS OT ASICs designers team:

CERN: G. Bergamin, A. Caratelli, D. Ceresa, J. Kaplon, K. Kloukinas, A. Nookala, S. Scarfì
IP2I Lyon: L. Caponetto, G. Galbit, B. Nodari, S. Viret

The CMS experiment and the tracker detector







Requirements for the high luminosity tracker upgrade

Phase-II upgrade tracker requirements:

- Higher luminosity
- From 20 to 200 pileup events per BX
- Increase radiation tolerance
- Reduced material budget
- Participate in the L1 trigger
- Improve trigger performance

Tracker module electronics requirements:

- Increase granularity
- Introduction of a pixelated sensor
- Radiation tolerance up to 100 Mrad
- Quick and on-chip particle discrimination
- Higher trigger rate (1MHz) and longer latency (12.5 µs)
- Power density < 100 mW/cm^2
- Add tracking information to the Level-1 trigger decision

A novel particle detector electronic system

The tracker detector can provide for every event additional information for the trigger decision leading to a significant improvement of the particle recognition efficiency

↓

The complete real time tracker readout is not feasible (10 Pb/s)

↓

HOW? The readout electronics in the detector can send pre-selected information for the Level-1 event reconstruction

↙ ↘

Intelligent pixel particle detector capable to locally self-select interesting signatures of particles interesting for the physics, without relying on an external trigger system

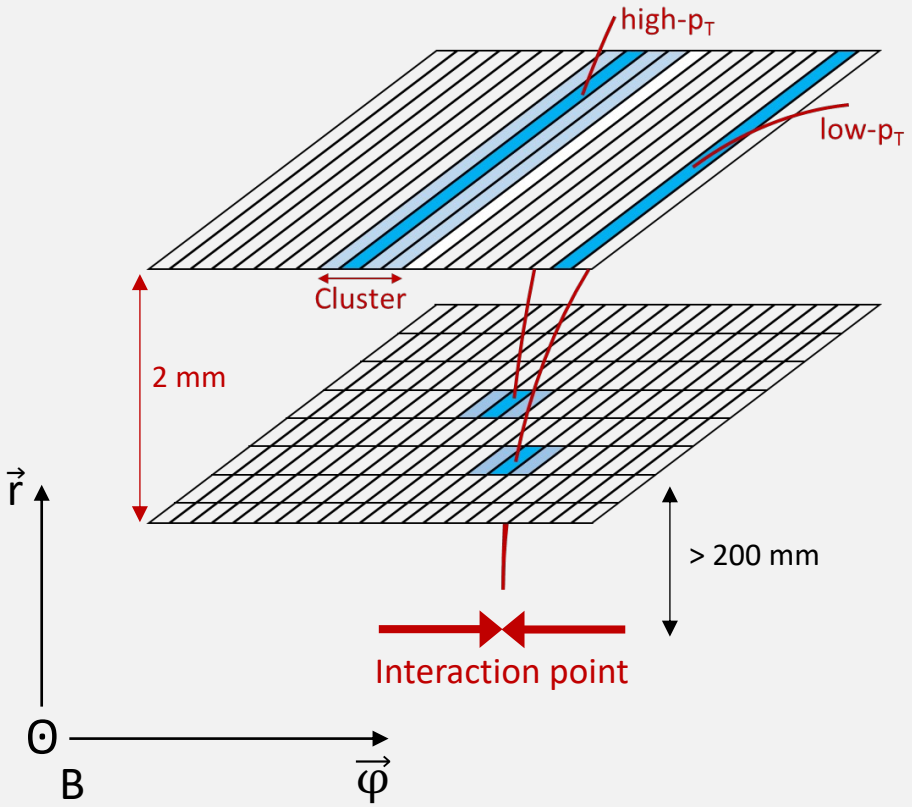
↓

This approach is used for the first time in an high energy physics experiment allows for a **significant data reduction efficiency improvement**

Detector capable of providing particle transverse momentum information in addition to simple geometrical positioning and energy measurements

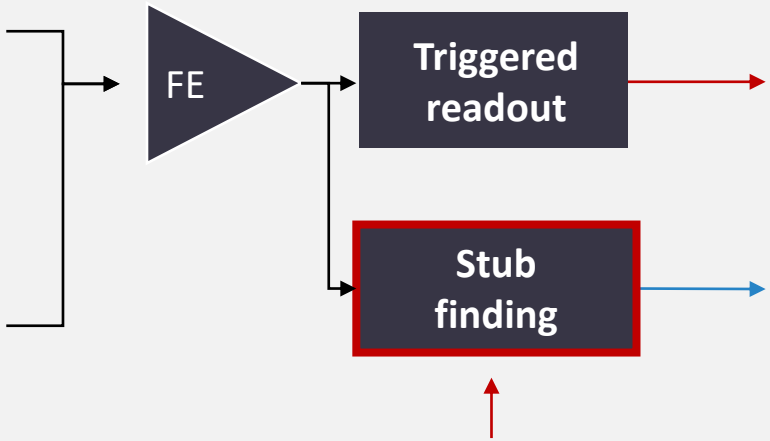
An intelligent particle tracking system based on p_T discrimination

Design of pixel sensors



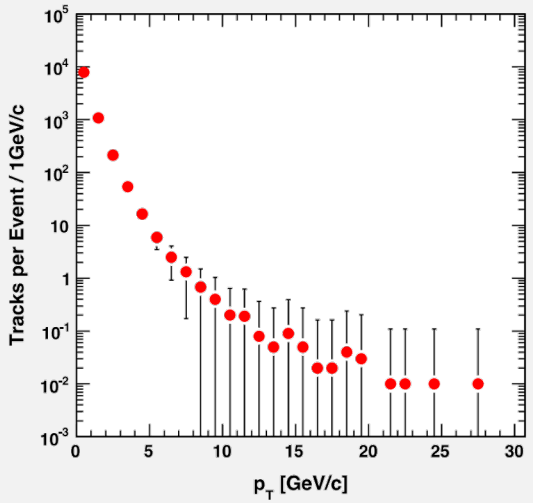
3.8T provided by the superconducting solenoid

Design of readout and data-processing ASICS



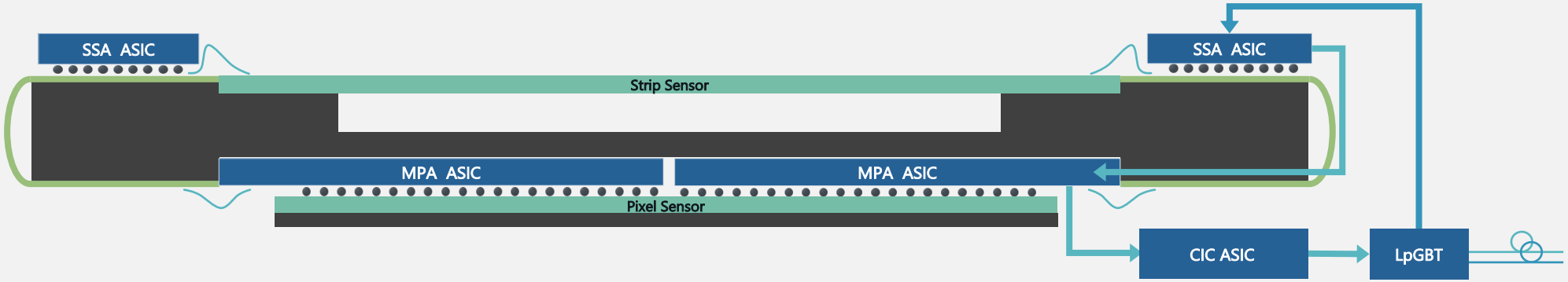
Full raw event at L1 rate (up to 10^6 events/s)

Real-time transmission of encoded information of high transverse momentum particles 40 Mframes per second



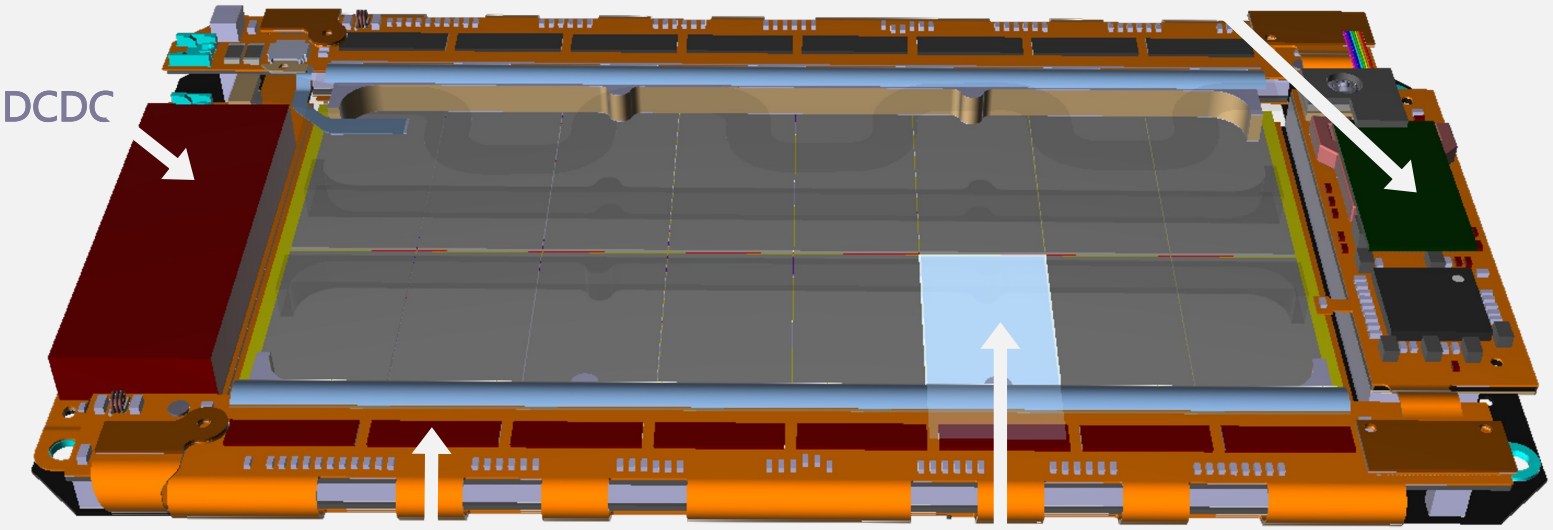
The p_T threshold should be chosen to provide a good rejection of the background and therefore a significant data reducing factor, but not too high to filter the interesting decay products with consequent impact in the analysis performances.

The Pixel-Strip module



LpGBT & VTRX+

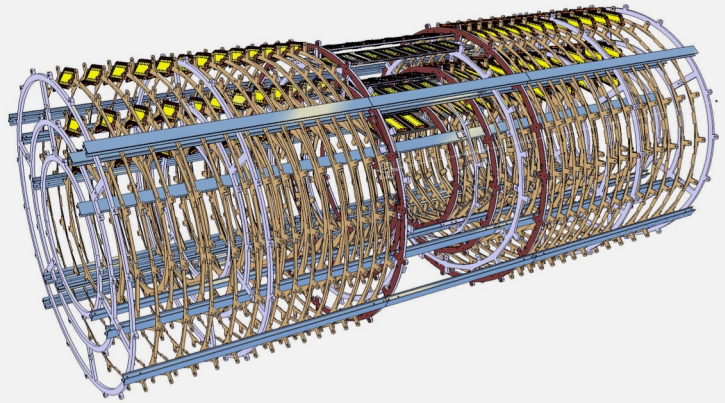
DCDC



16 x SSA ASICs (Strip ROC)

16xMPA (Pixel ROC + stub finding)

13296 Modules
44 M strip + **174 M pixels**
200 m^2 of Silicon Area



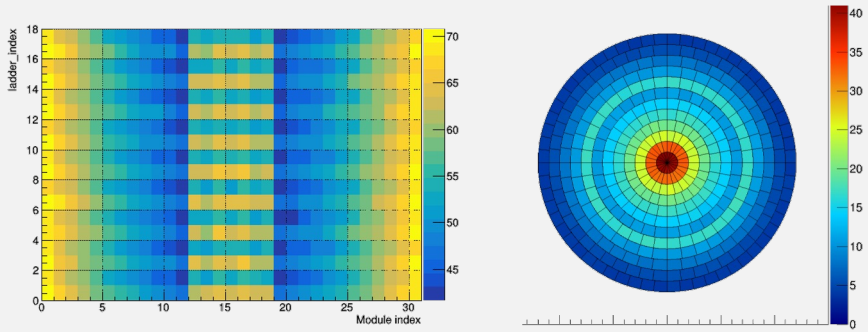
MPA and SSA ASICs: system level architecture choices

Open design choices:

- Define which functionality should be implemented in the SSA and which in the MPA
- Minimise system power requirements
- Minimize bandwidth requirements
- Maximize the particle recognition efficiency
- Bandwidth among ASICs
- Data encoding or unsparsified data transmission
- Transmission FIFOs depth
- Data compression
- Particle hit clustering at SSA level
- Several others

The optimal solution needs to take into consideration:

- 3 different ASICs to be designed
- 18 chips communicating to implement the algorithm
- 2 readout data paths
- Functionality and the efficiency depends on physics statistics, particle rates and hit occupancy (no simple test vectors)
- Minimize power consumption



MPA and SSA: system level architecture choices

Becomes necessary a Simulation framework capable of:

System Studies and performances evaluation

Design Verification

- **Study and compare** different system implementation
- Evaluate **tradeoff between performances and power optimization**
- Extract and report **efficiency parameters** by comparison with an **ideal system reference** model
- Evaluate the **efficiency of the particle recognition** algorithm and of the data readout
- **Realistic stimuli generation** from Monte-Carlo simulations of complex interactions in high-energy particle collisions

MPA and SSA: system level architecture choices

Becomes necessary a Simulation framework capable of:

System Studies and performances evaluation

- **Verify** the RTL implementation and the **chip-set functionalities**
  - To address the simulation on specific subsystem functionalities
  - To verify their effects at module level
- Generation of **realistic activity information** for power analysis
  - In the final ASICs, the power consumption is strongly dependent from the sensor-input signals and the architecture is optimized accordingly

Design Verification

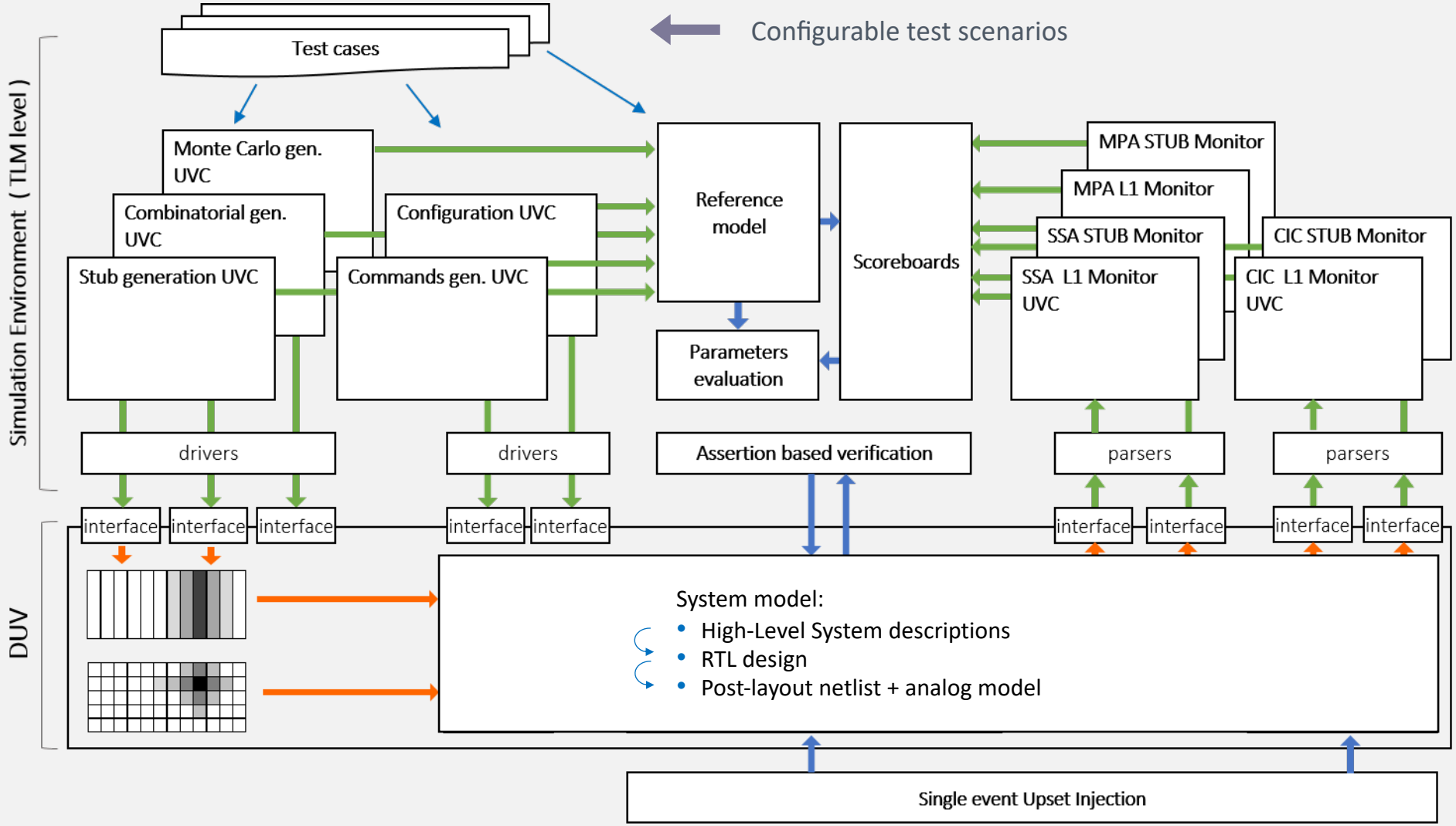
- Verify at clock-cycle level precision:
  - The **ASICs functionalities**
  - The **subsystems integration**
  - The communication between modules and the **communication protocols** between ASICs
- After physical implementation, **functional verification** of the netlist with back annotated delays

System level simulation framework implementation

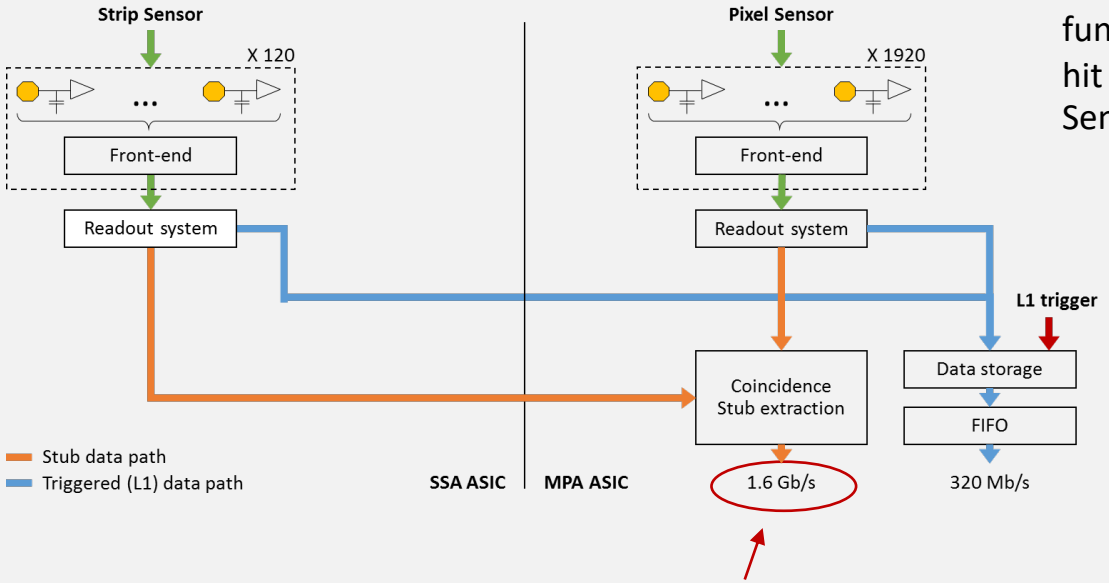
Implemented in:

SystemVerilog/UVM + Python

Randomized stimuli emulating particle detector hits, noise, loppers and jet emulation according the CMS tracker geometry

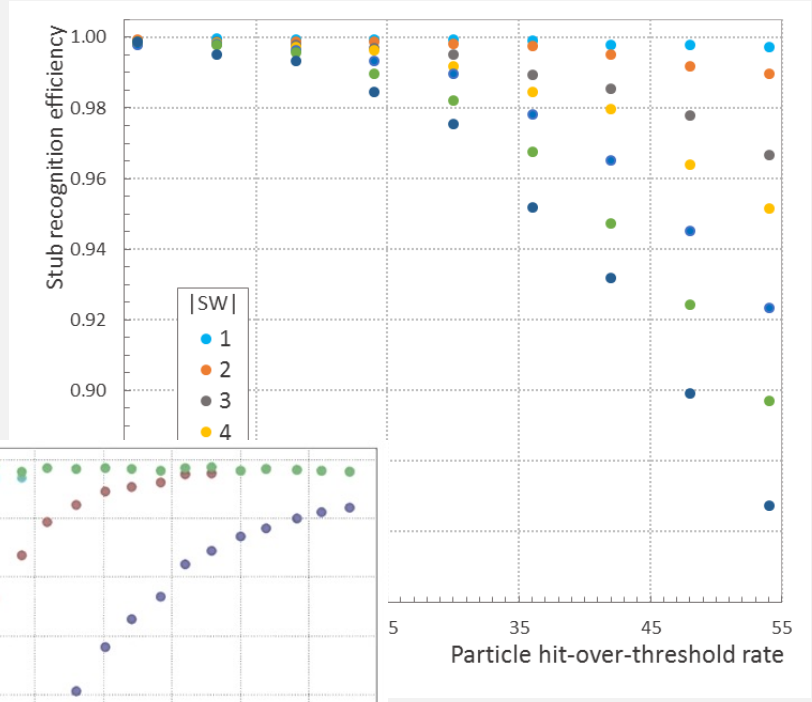


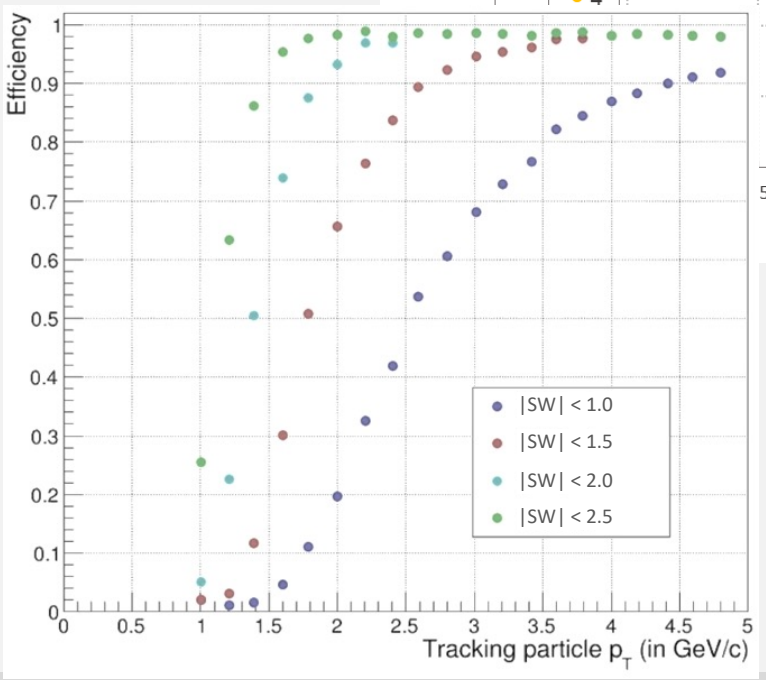
PS module – Stub efficiency and p_T window cut



Fixed by the LpGBT tranceiver

Stub finding efficiency at the MPA output as function of the particle hit occupancy and the Sensitive Window cut



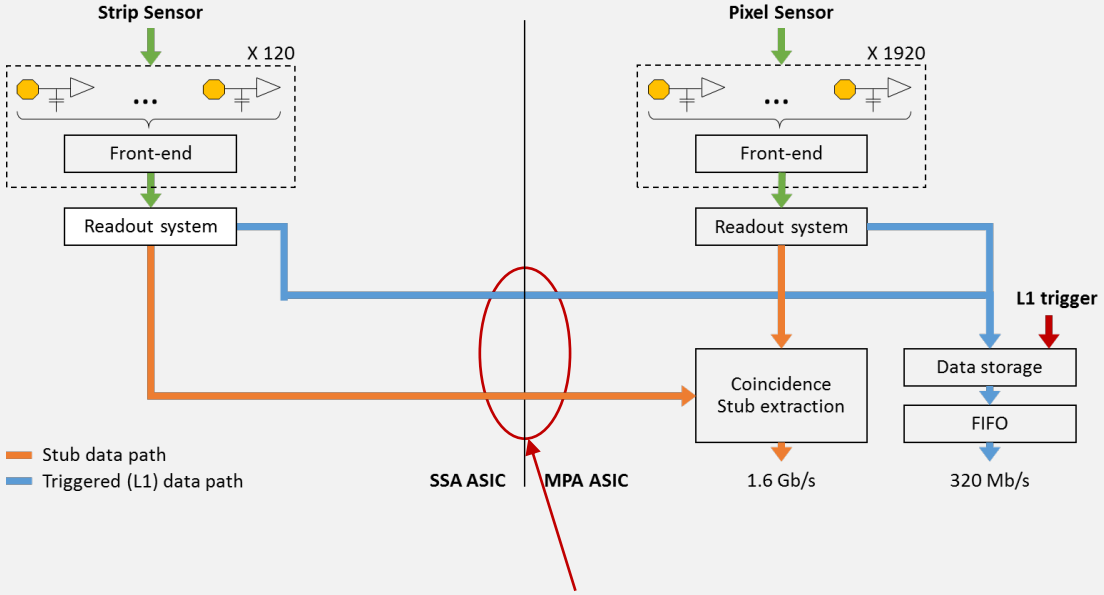


Tracker reconstruction efficiency for different window cuts, assuming no losses at module level.

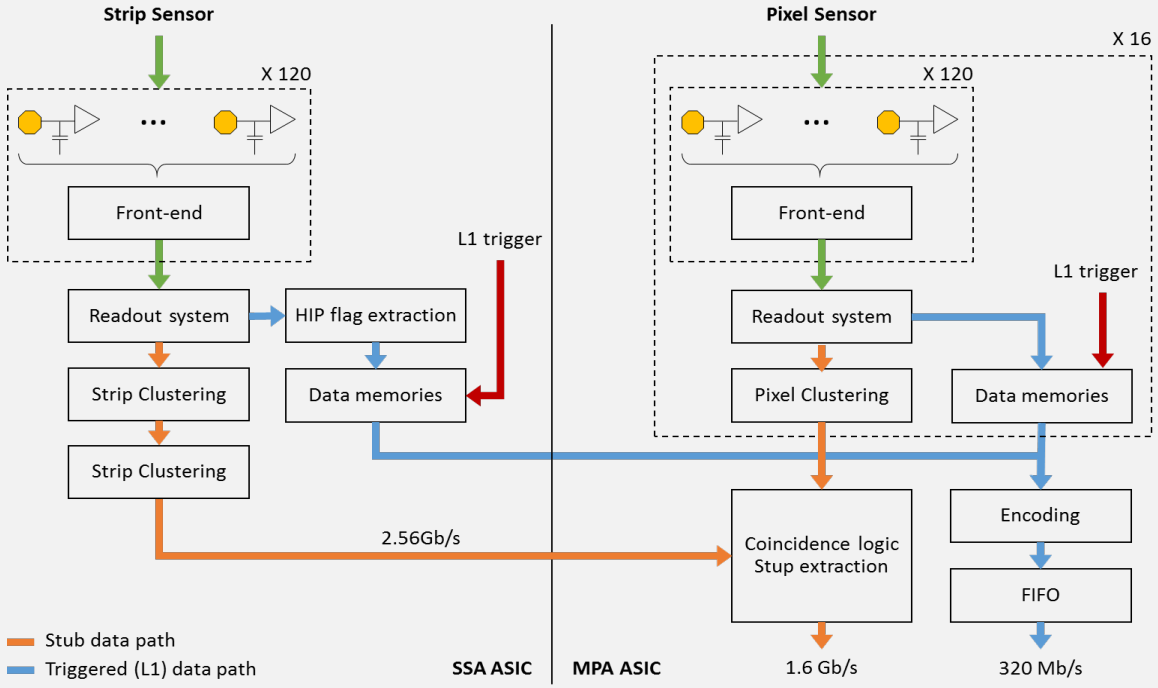
$$p_T = \frac{\sqrt{3}}{3} \cdot q \cdot r \sqrt{\left(\frac{\sin\theta_0}{\cos(\theta_0 - \alpha)}\right)^2 \cdot \frac{d^2}{(S_W \cdot l)^2}} = \frac{B \cdot r_C \cdot c \cdot d}{2\delta_\theta} \sqrt{1 + \frac{\delta_\theta^2}{d^2}}$$

Defines the window cut to be applied in the chip

PS module – Architecture optimization

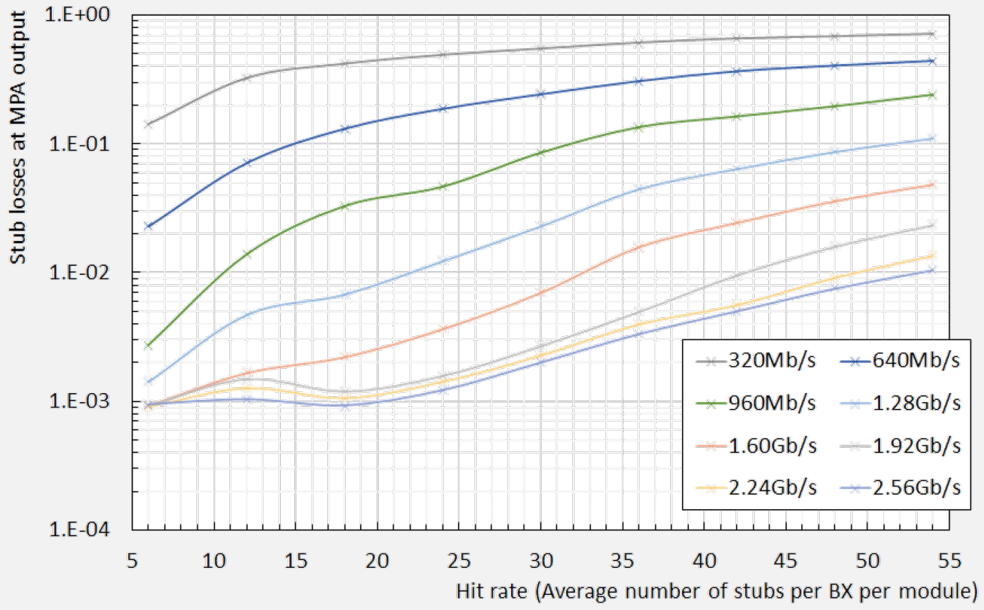


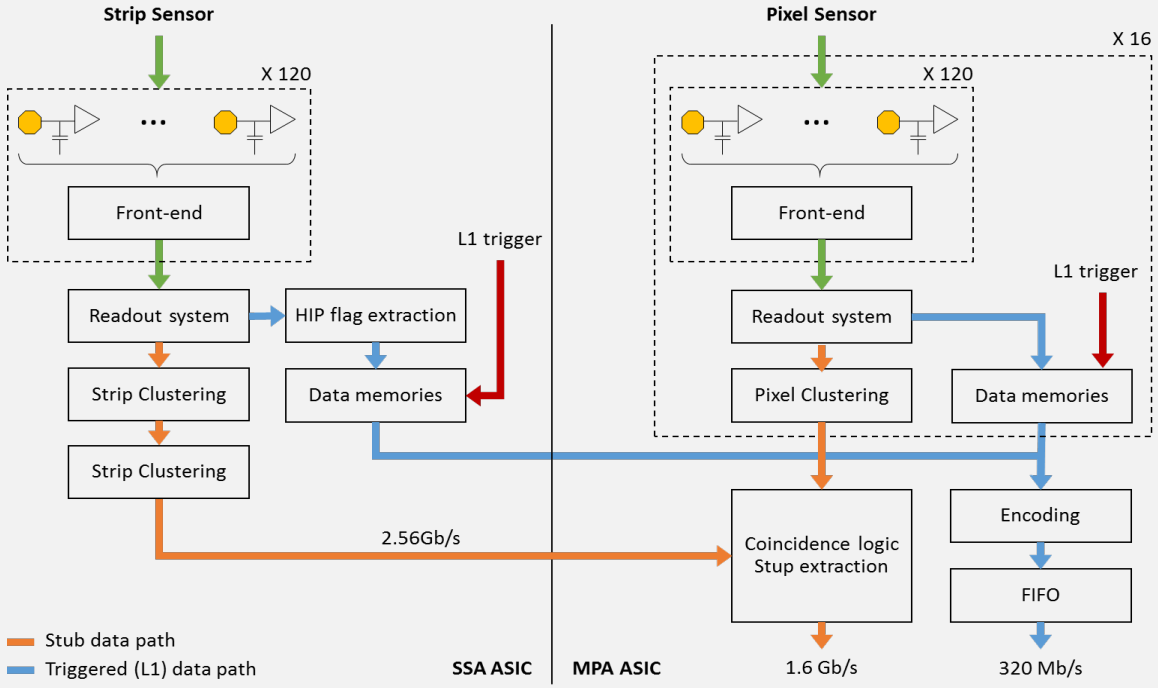
Significant contribution to the system power consumption.
Unsparsified transmission requires 16 x 7.2 Gb/s



- More complex SSA design requiring
  - Additional embedded memories in the SSA
  - Clustering logic and filters to further reduce bandwidth.
  - Encoding logic
- Save 71.4% of IO transmitters power consumption (267 mA per module) compared to unsparsified solution

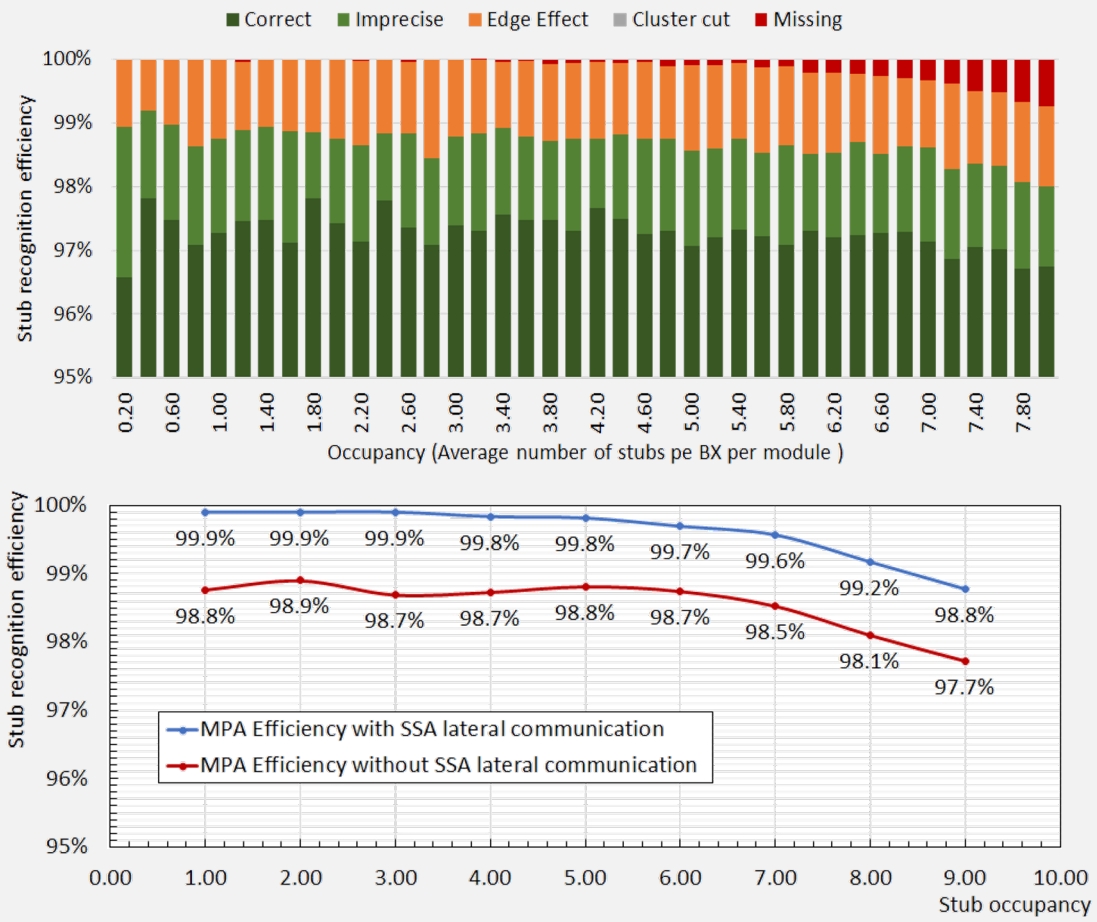
PS module – Architecture optimization

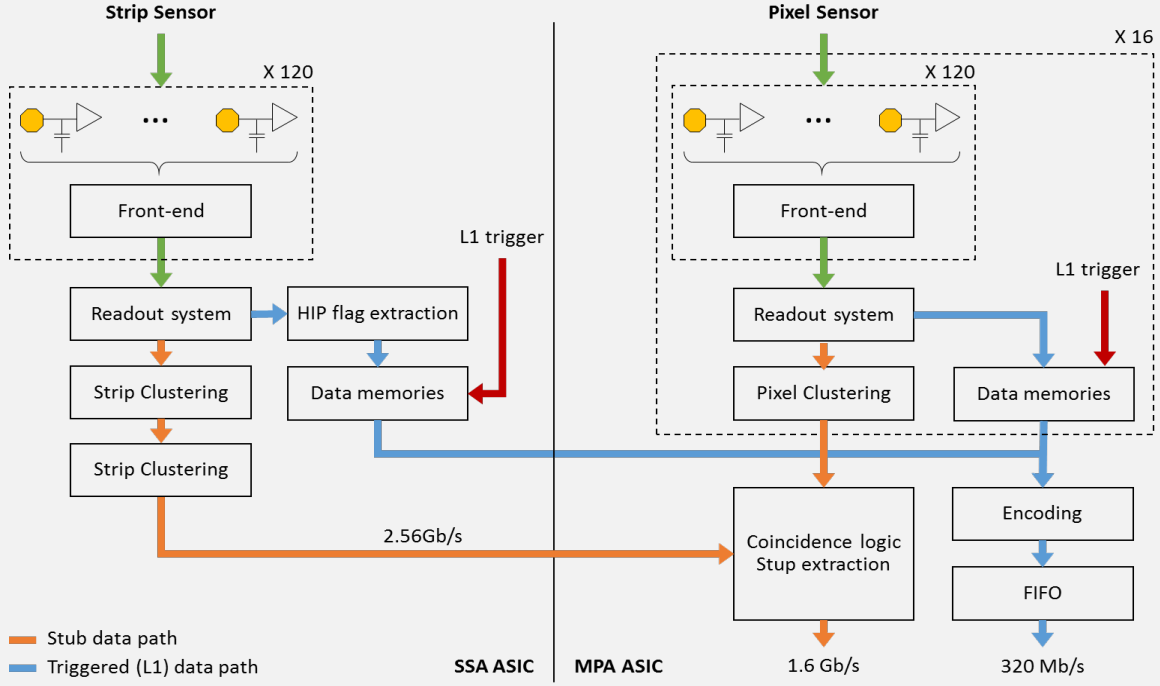




- One **SSA** should provide at least **6 particle clusters** per BX event
- One **MPA** should provide **5 stubs every 2 events** (time averaging)

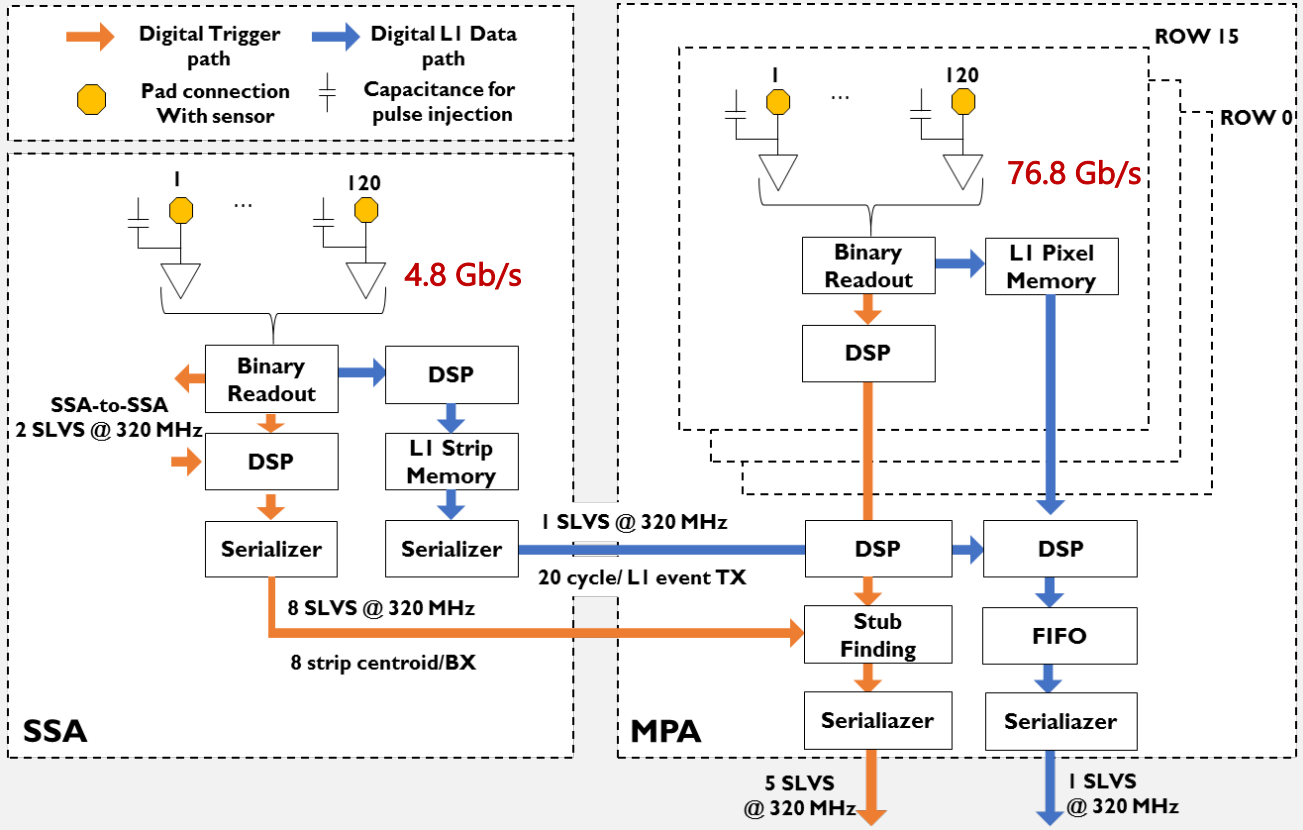
PS module – Clustering among neighbor ASICs

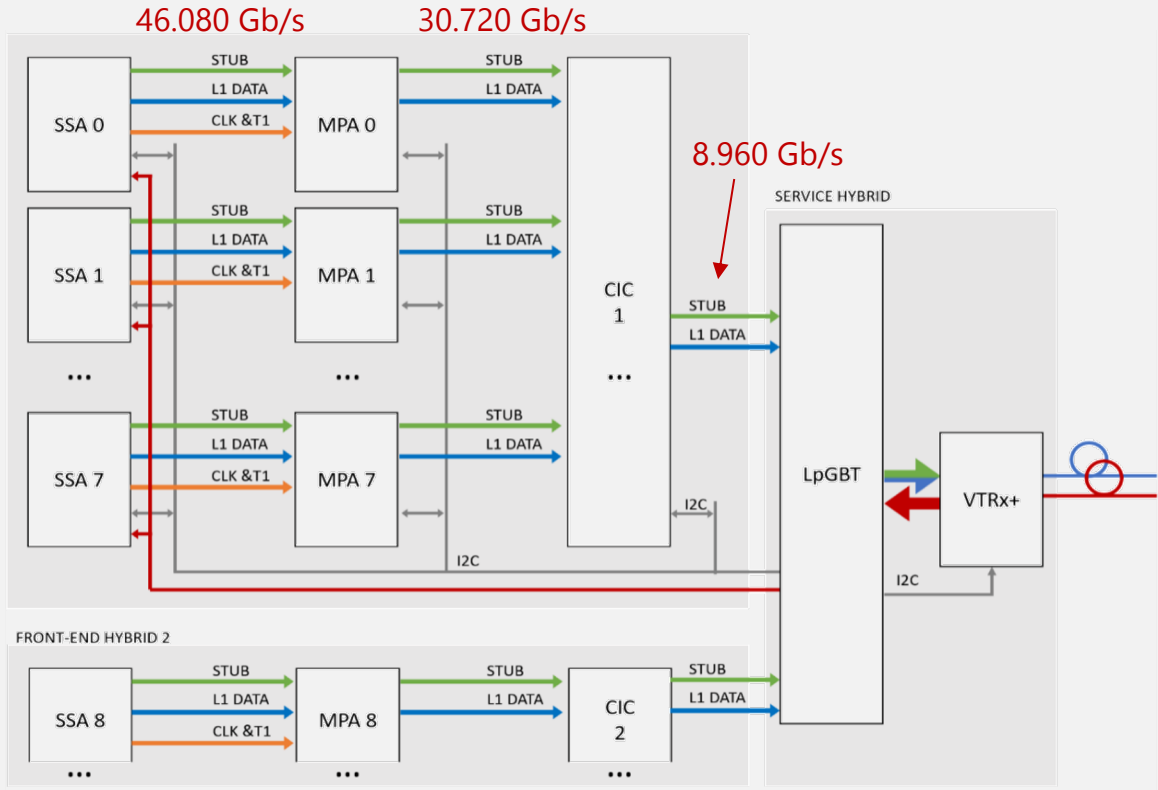




- Lateral communication in SSA → +5 mW/chip +1% efficiency
- Lateral communication in MPA → +20 mW/chip -0.15% efficiency
- Module-to-module communication would give a simulated benefit below 0.5% while drastically increase assembly complexity

System architecture definition

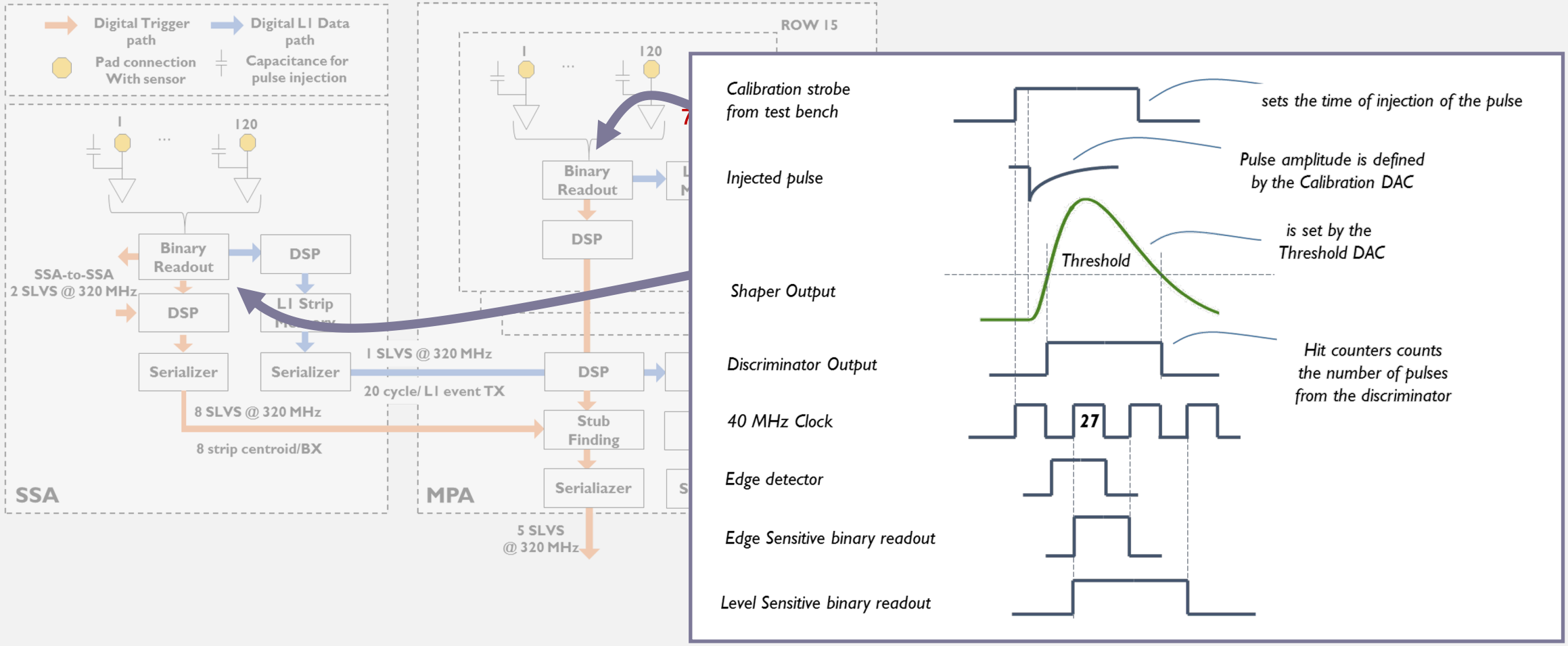




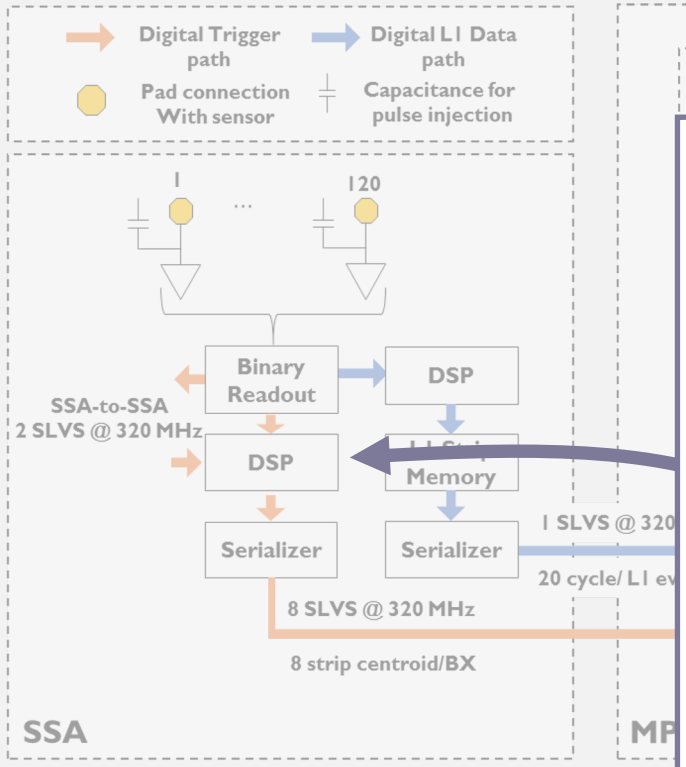
40 MHz rate for high-pT particles
1.6 Gb/s
very low latency: < 500 ns

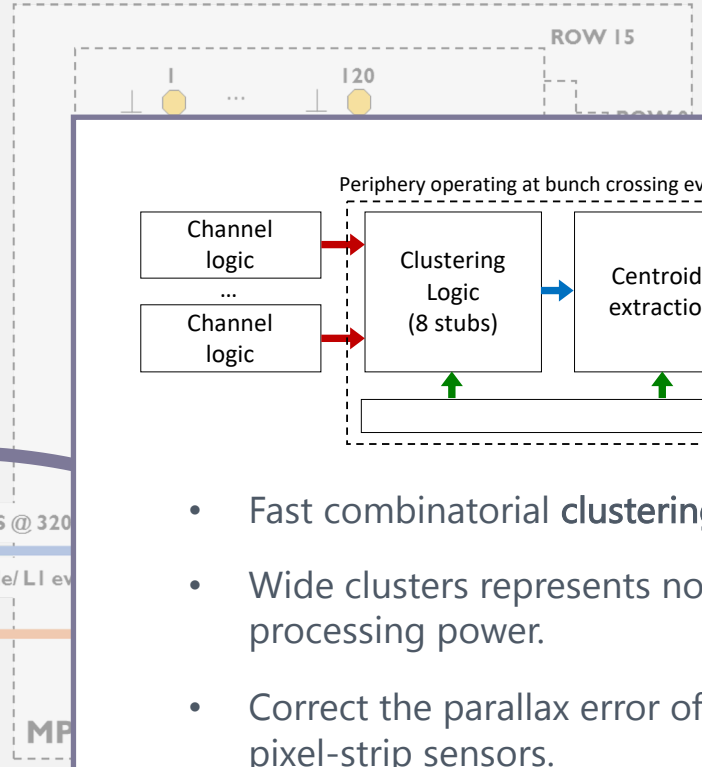
Full frame at L1 trigger rate (1MHz)
320 Mb/s

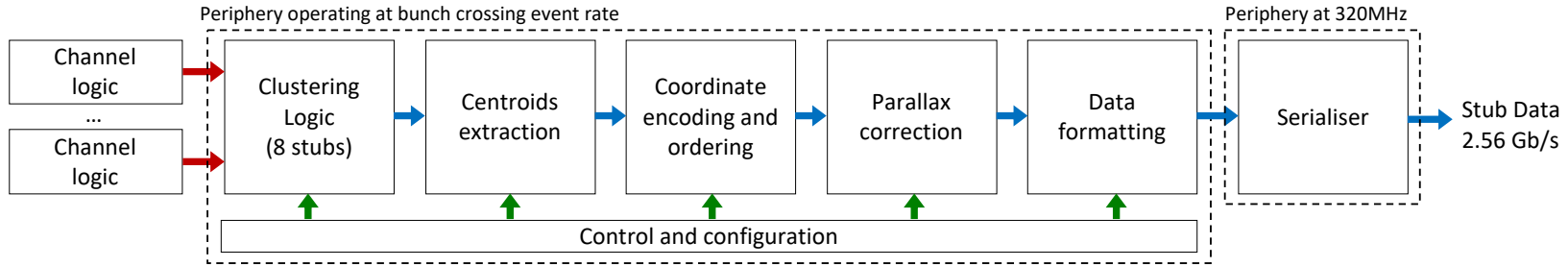
System architecture definition



System architecture definition

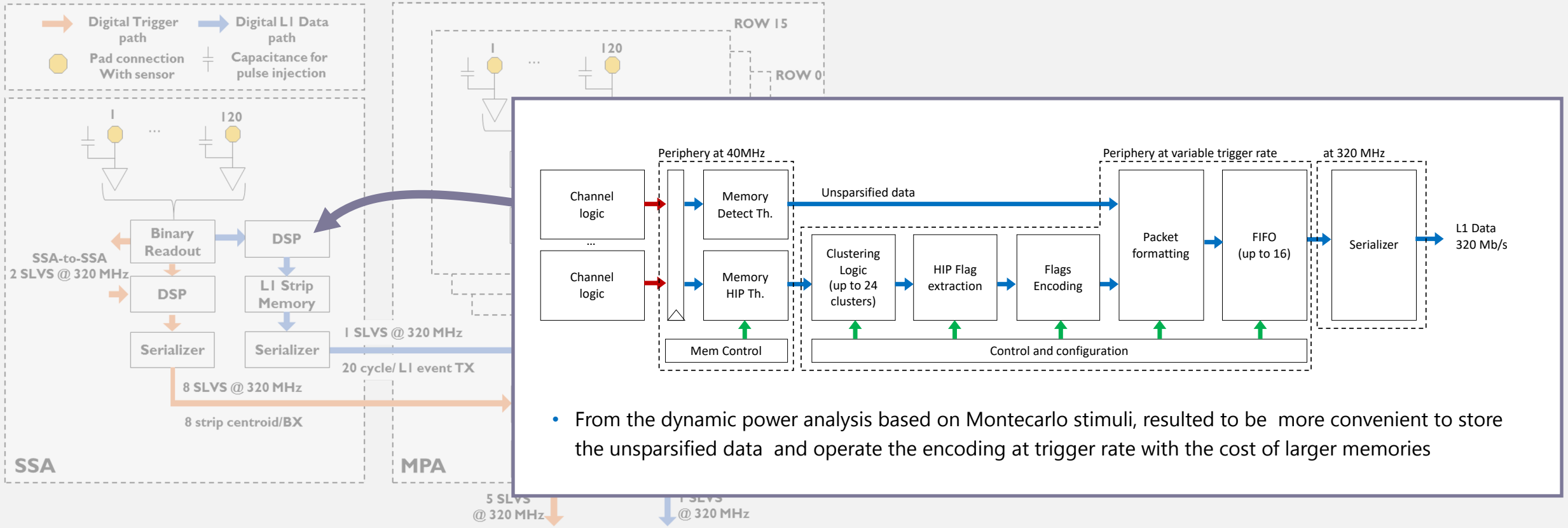






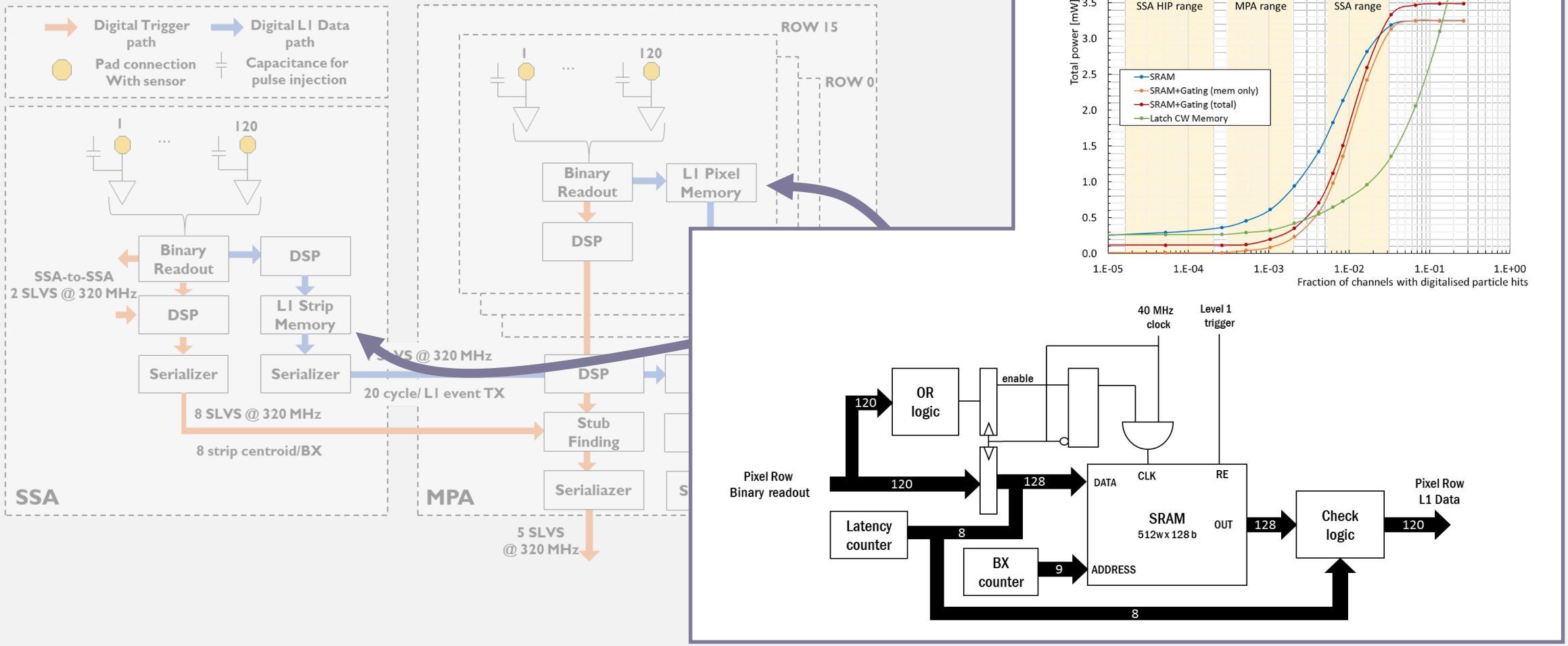
- Fast combinatorial **clustering** at event rate to limit cross-talk effect
- Wide clusters represents not interesting events: are **filtered** to optimize bandwidth and processing power.
- Correct the parallax error of approximating the cylindrical geometry with planar pixel-strip sensors.
- Up to 8 Cluster Centroids coordinates are transmitted per every event to the MPA coincidence logic for the **Stub generation** and the **transverse momentum discrimination**

System architecture definition



- From the dynamic power analysis based on Montecarlo stimuli, resulted to be more convenient to store the unsparsified data and operate the encoding at trigger rate with the cost of larger memories

System architecture definition

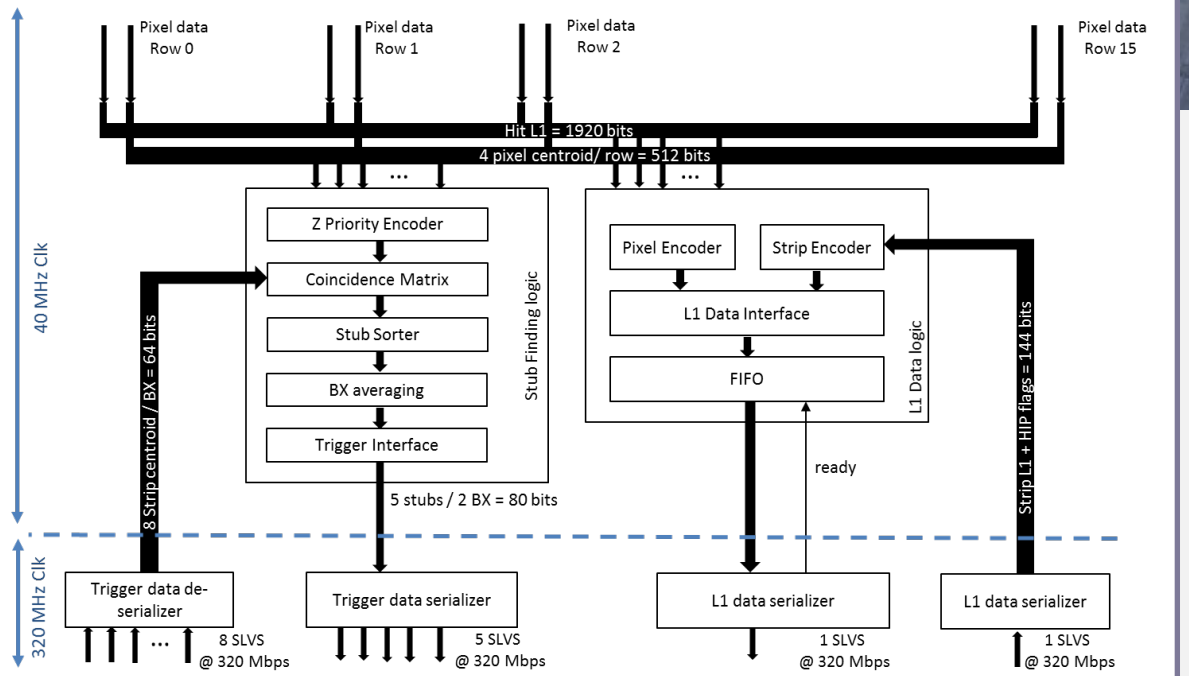


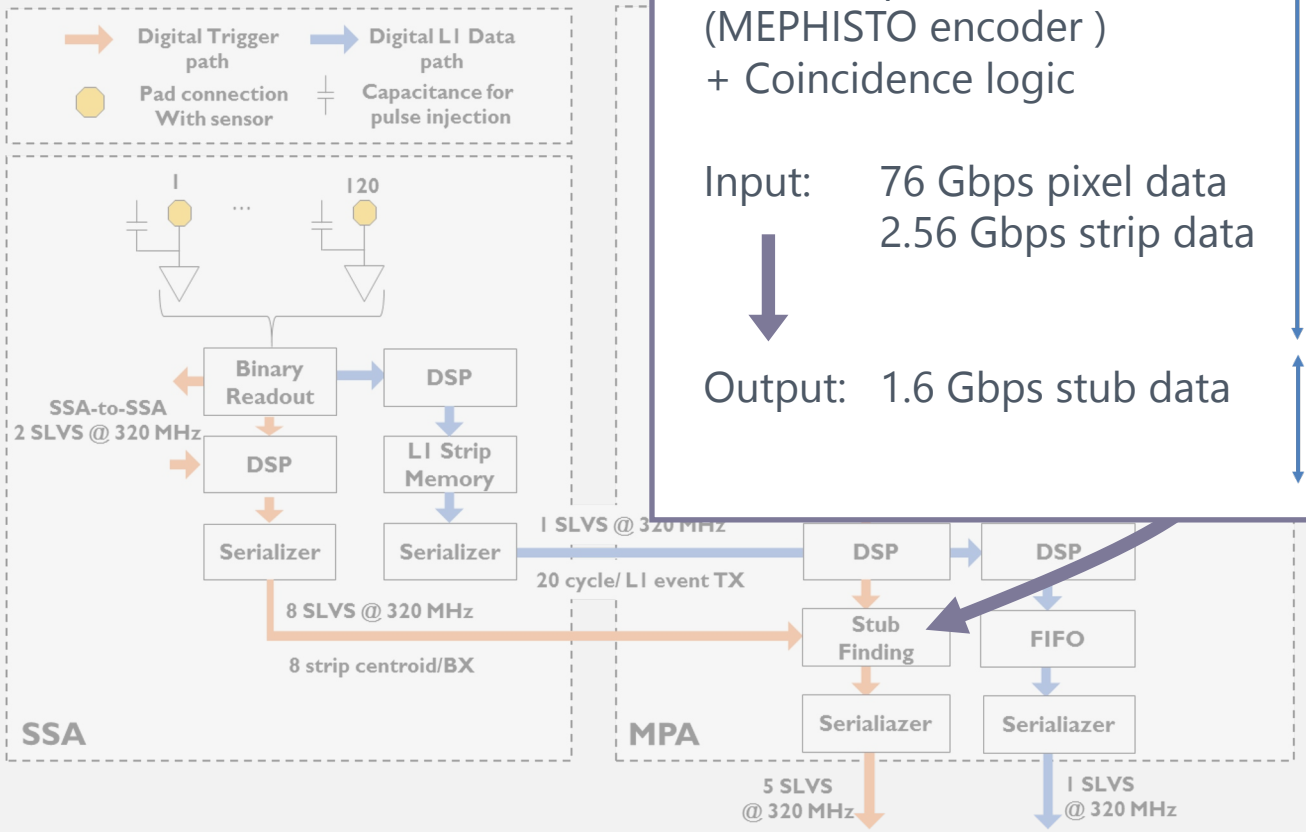
Stub Finding Logic

Zero compression
(MEPHISTO encoder)
+ Coincidence logic

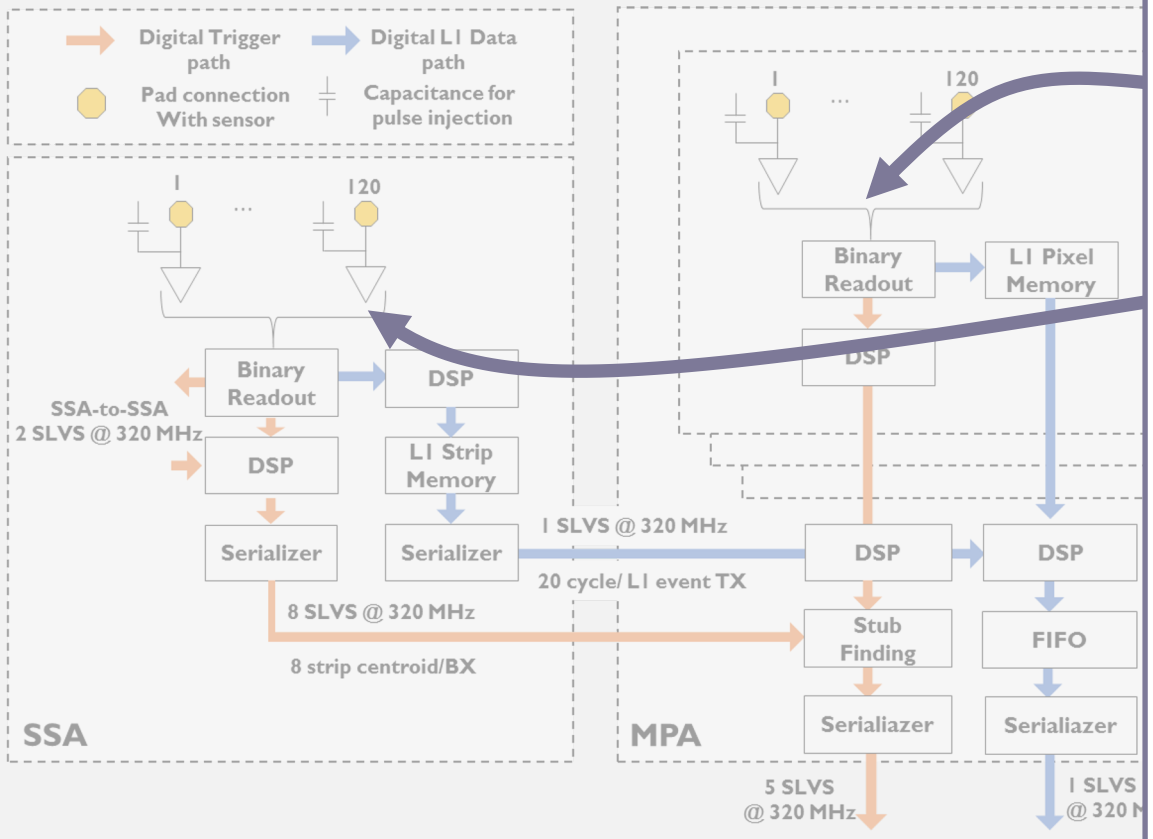
Input: 76 Gbps pixel data
2.56 Gbps strip data

Output: 1.6 Gbps stub data

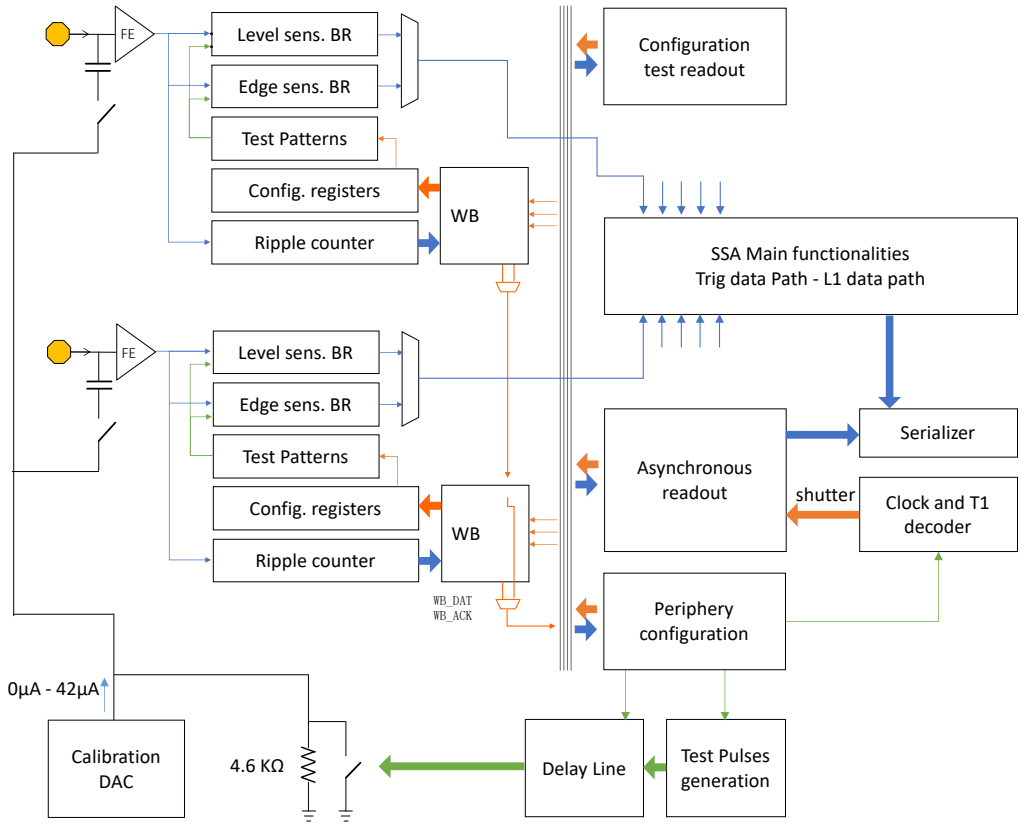




System architecture definition



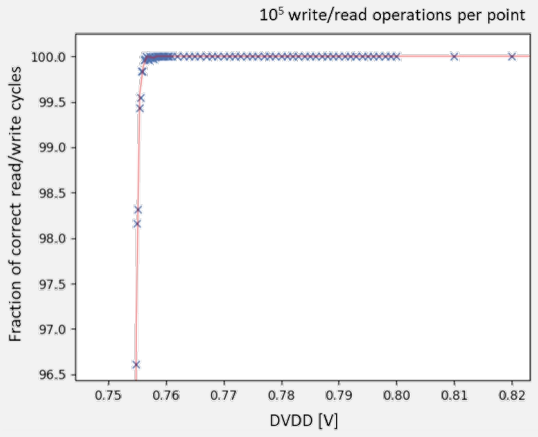
Calibration circuit

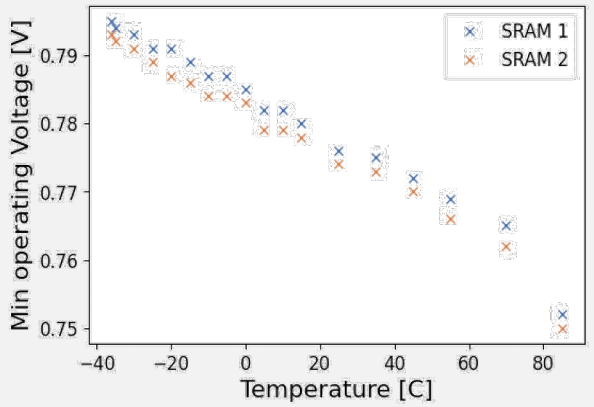


Design for Testability

Memory Built-In-Self-Test

- Test full memory functionality in <1 ms
- Results saved in internal registers accessible via slow-control
- Few additional hardware self contained hierarchical block
- Clock gating during normal operation (only leakage power)





Periphery Scan Chain

- FSM Easy to access – standardized approach with TRL
- 92% of fault coverage in SSA
- ~95% of fault coverage and 25k ff in MPA
- Test all periphery in less than a second

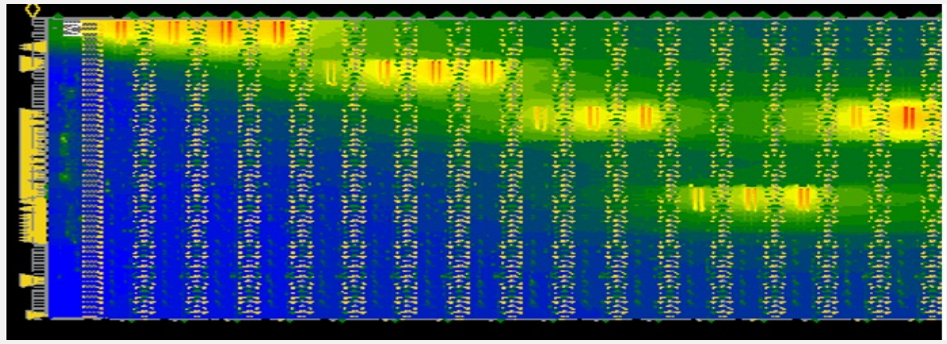
Logic Built-In-Self-Test for pixel array

- FSM embedded in Pixel Array logic and vectors from configuration
- Requires dedicated development
- Requires compression logic
- ~90% coverage

Power Reduction Methodology

Power optimization

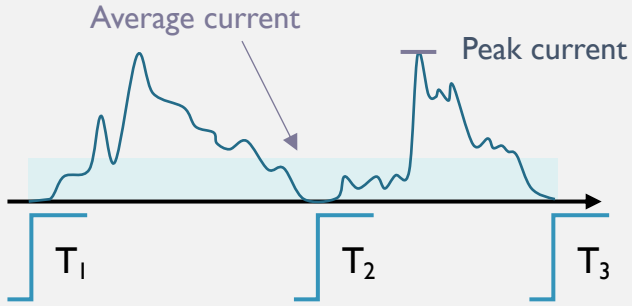
- Clock gating in all configuration registers and logic
- Architecture studies to minimize power consumption
- Use Multi-VT standard cells
- Use latches for FIFO
- Use gated SRAM blocks
- Multi-supply voltage (1.0V – 1.2V)
- Find and optimize power hungry and low activity blocks



Dynamic IR-drop

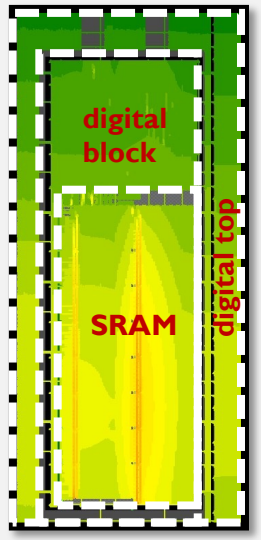
Power study:

- Static and **Dynamic power** analysis
- Find power hungry and low activity blocks
- Voltage drop analysis on different scenario
- Power-Grid-View (PGV) for Macro



Dynamic power analysis

IR-drop study w detailed PGV for SRAM macro



Cumulative radiation effects

Digital domain:

- **9-tracks library** selected as compromise between power consumption and radiation tolerance considering the operating range $-40°C$ / $0°C$
- **Characterization of the digital cells** parameters (prop. delay, transition time, setup/hold, etc.) for radiation corner
- **Increased margins for TID degradation (setup uncertainty jitter + additional 8% of clock period – reduced max transition – derate factors)**
- Due to **narrow channel** effects → Avoided usage of small width buffers (D0, D1) and delay elements
- **Thin oxide IO pads** → 1.2V (CMOS IO and SLVS)
- Custom **ESD structures** latch-up resistant

Memory tolerance:

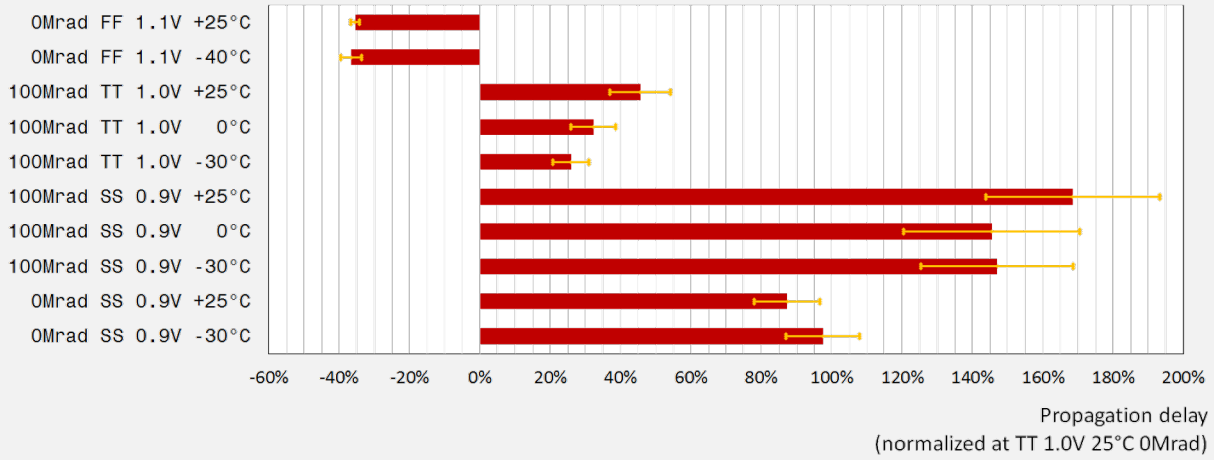
- A custom memory compiler allowed to generate a **SRAM** with cell transistor featuring nMOS W > 200 nm pMOS W > 500 nm
- Protection against latch-up is reached by placing p^+ guard bands between n^- regions.

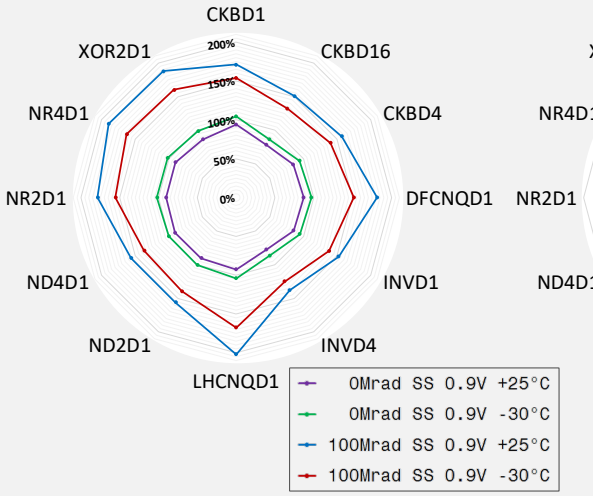
Usage of ELT devices in input stage:

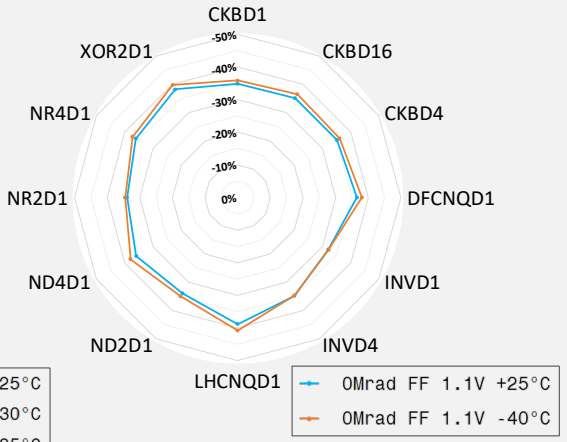
- To prevent the **radiation induced** drain-to-source **leakage current increase** due to the charge trapped in the shallow trench isolations (STI).
- To **mitigate the 1/f noise increase** on irradiated devices due to side-effects of the STI region in nMOS operated at low drain current.

Digital library choice and delay corner comparison

- Supply voltage scaling
- 9 tracks library chosen as compromise between power consumption and radiation tolerance
- Temperature inversion effect prevent the SSA from using a high-Vt library cells at 0.9V.
- Mix of standard-Vt and low-Vt digital cells at 1.0V+10%
  as compromise of power consumption, memory operation and propagation delay at $-40°C$







SEU tolerance

State machines

- Triple module redundancy (FULL)
- Triplicated Clock-trees
- Triplicated Reset distribution
- FF minimum distance 15um

Latch FIFOs

- Control and header fields triplicated
- Data latches not protected

Data pipeline

- No SEU protection applied due to limited power budget

Clock tree

- Clock tree triplicated
- The non-triplicated logic uses the voted clock in critical areas
- The non-triplicated logic uses one of the branches in non-critical areas:
  - Simplify scan-chain insertion
  - Helps in reducing buffering for hold fix (power)
  - Allow for CPPR on the 3 branches

Triplicated pads for

- Clock
- Reset
- Control
- Scan-Chain IOs

Configuration registers

- Triple module redundancy with error detection and self-correction
- Clock enabled only during
  - asynchronous readout operation,
  - configuration operations
  - self-correction

Glitch filters

- Reset inputs
- TEST-MODE signal
- Scan-chain TEST POINTS control (on the control of the system clock / test clock selection multiplexers)

SEU tolerance

Physical implementation

- Use of instance space groups among triplicated registers
- Avoid logic simplification by synthesis and P&R flow
- Spacing for clock and reset:
  - After CTS locate the critical cells and impose a minimum distance
  - procede in successive ECO placements and ECO routing steps

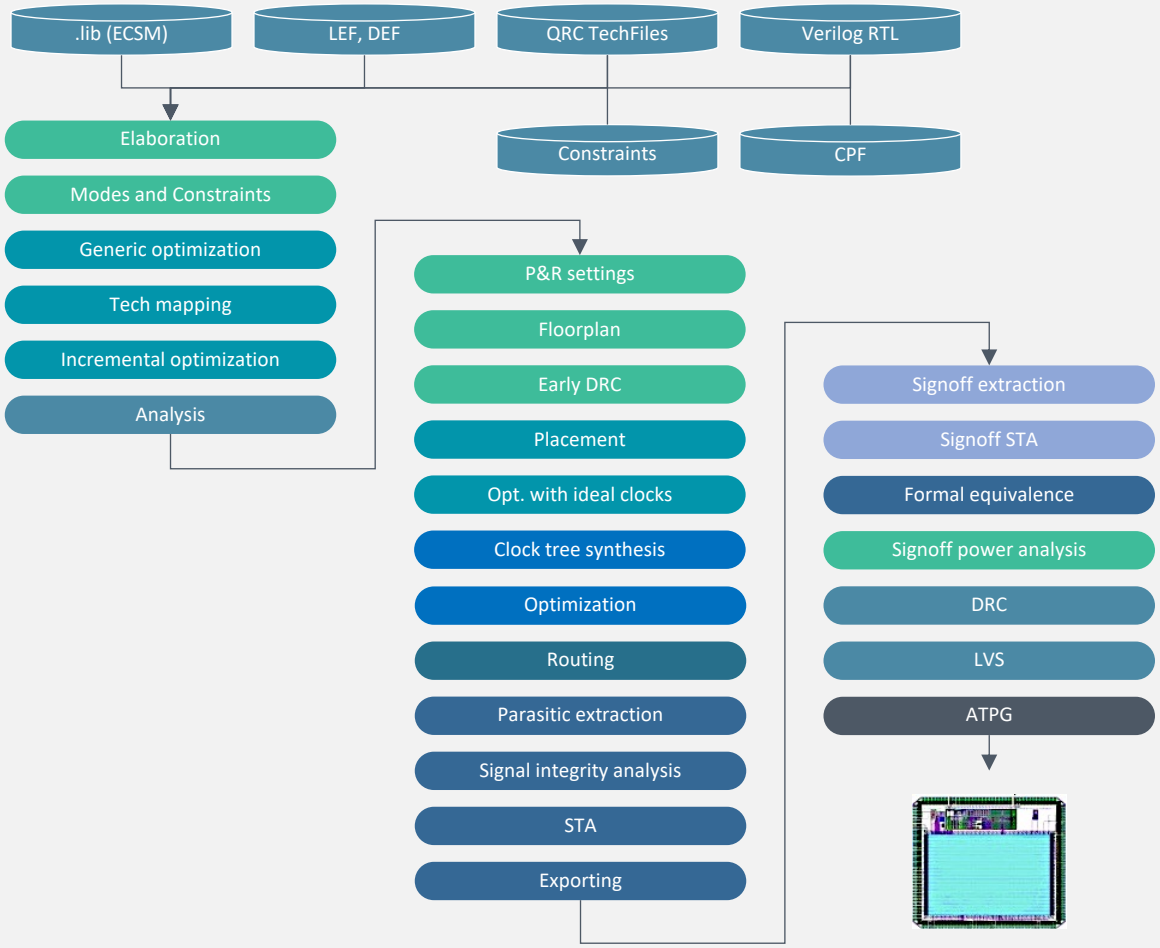
Functional simulation

- System Verilog UVC for randomize the injection (constrained from the specific test case)
- The randomization is constrained accordingly: Error probability, average SEE rate, minimum time split, etc..
- Injection of single event effects in multiple ASICs at the same time to evaluate the consequences that SEE in an ASIC have on the other ASICs part of the chipset
- Possibility to focus the SEU injection on particular module or subsystem and evaluate the effect at system level
- Possibility to inject SEU in hundred of cells per clock cycle (register grouped in non-interacting categories)

Additional checks

- Script to verify that no triplicated instance is optimized out
- Script to verify placement constraints after chip assembly

Physical Implementation flow



- Digital-on-top design flow
- Hierarchical implementation
- Multi supply voltage 1.0 V ± 10% – 1.2 V ± 10%
- 3 independent power and ground domains to reduce noise coupling with guard-ring isolation
- Multi-Vt design (Low-Vt used only in critical timing arcs)
- C4 bump floorplan + wirebond for wafer probing
- Complex CTS and timing closure due to triple clock tree balancing and SEU hardening
- Complex constraints for TMR and digiatal cells placement
- Skew balancing among triplicated and voted clock trees
- Strip cell sampling clock guarantees <200ps skew in all corners
- Non-default CTS rules to mitigate cross-coupling
- QRC extracted information already at the optimization stage due to design size

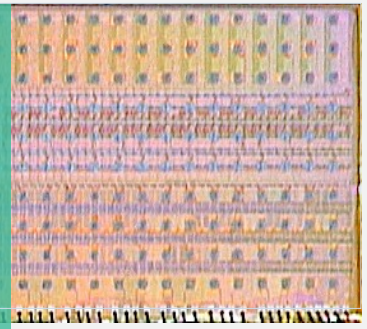
The ASICs prototypes

SSA2 ASIC

Short-Strip readout ASIC
for the CMS OT

DESIGN OR TESTING TEAM:

G. Bergamin, A. Caratelli, D. Ceresa,
J. Kaplon, K. Kloukinas, S. Scarfì



MPA2 ASIC

Pixel readout ASIC for the CMS OT

DESIGN OR TESTING TEAM:

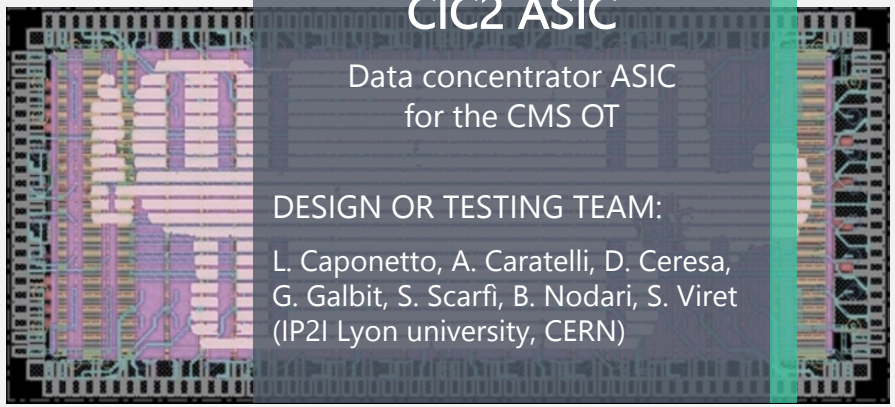
G. Bergamin, A. Caratelli, D. Ceresa, J. Kaplon, K. Kloukinas, S. Scarfì

CIC2 ASIC

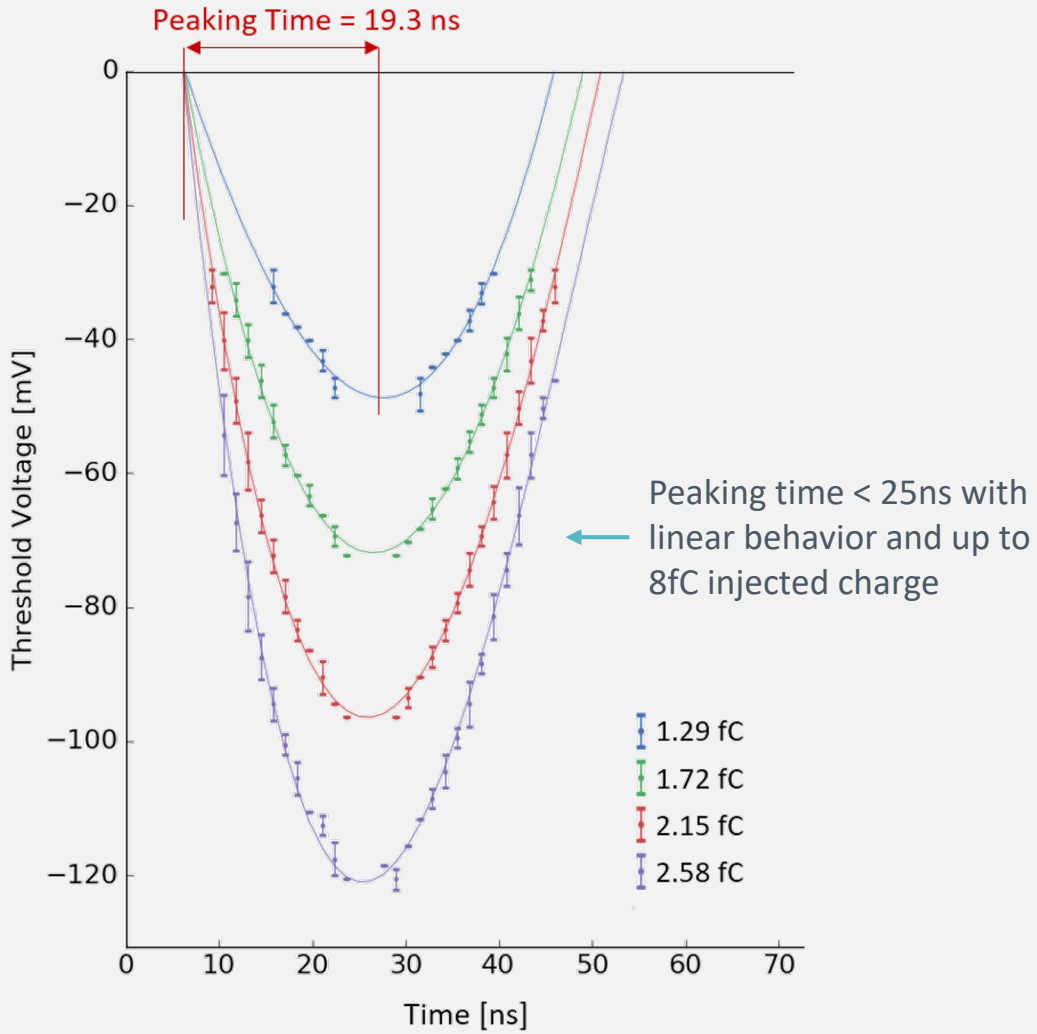
Data concentrator ASIC
for the CMS OT

DESIGN OR TESTING TEAM:

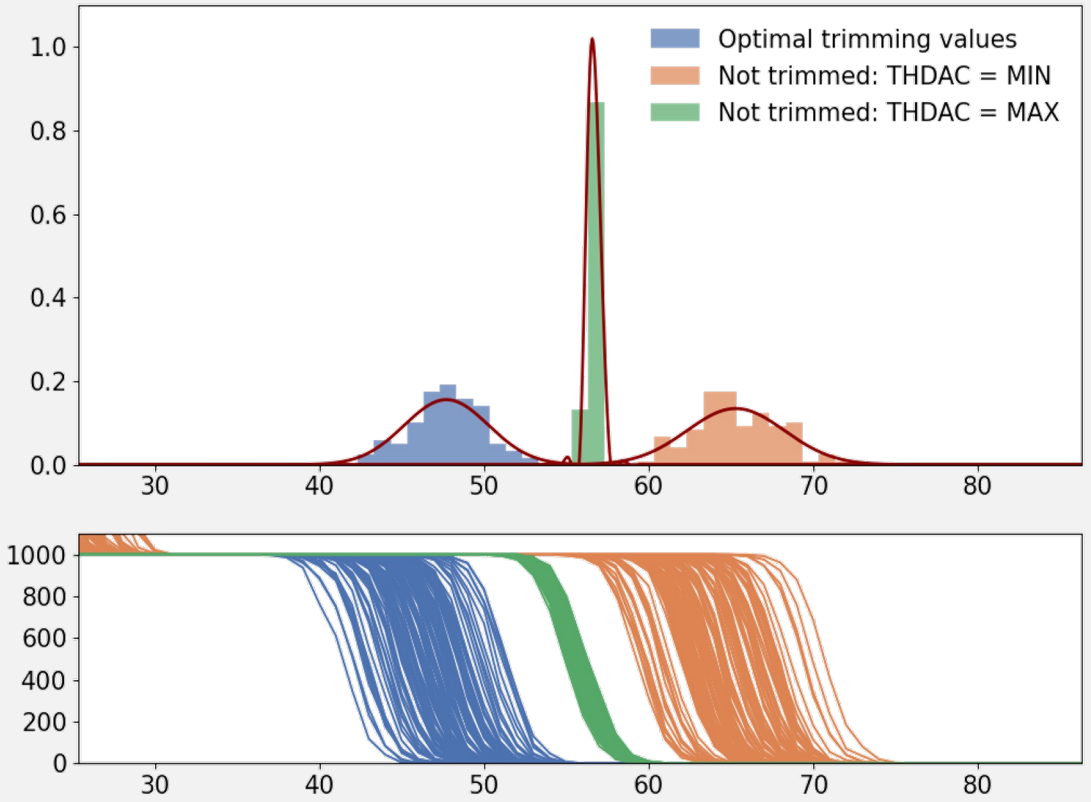
L. Caponetto, A. Caratelli, D. Ceresa,
G. Galbit, S. Scarfì, B. Nodari, S. Viret
(IP2I Lyon university, CERN)



Front-end electronics characterization



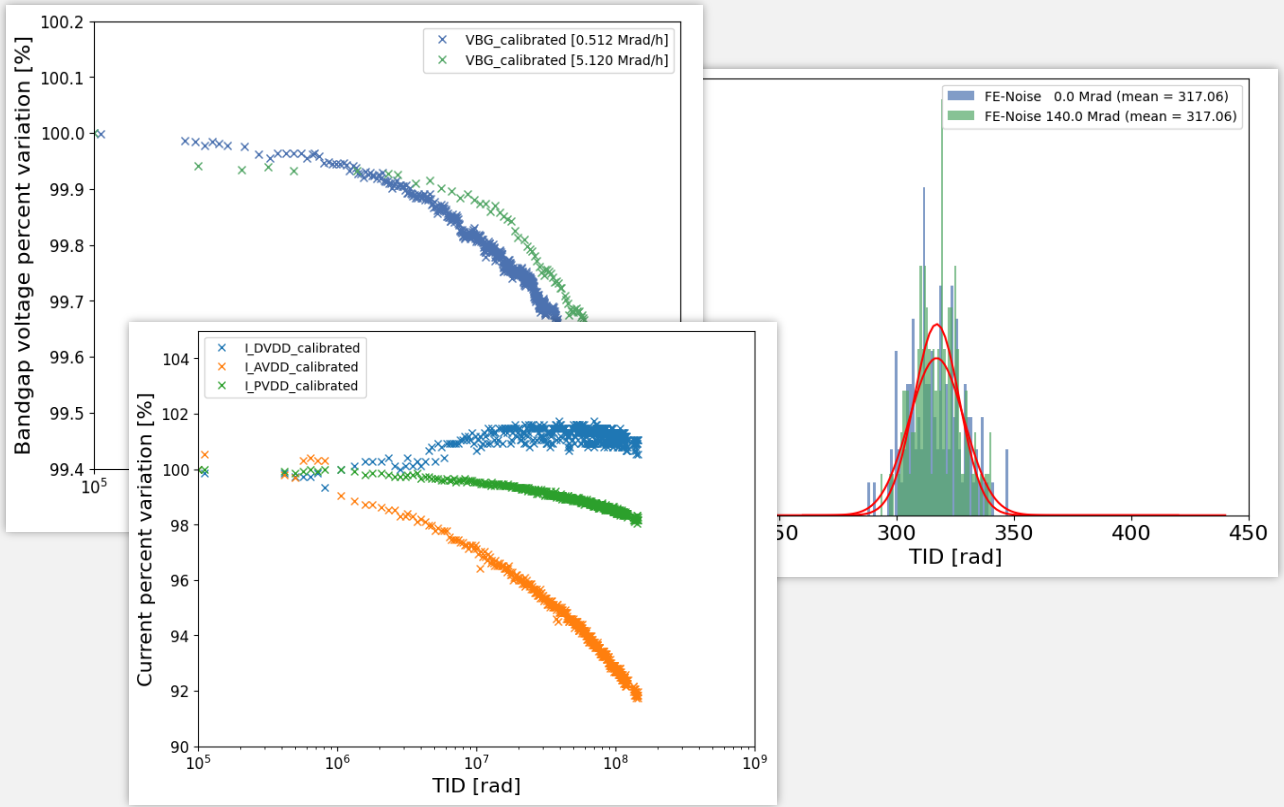
Peaking time < 25ns with linear behavior and up to 8fC injected charge



SSA-MPA STATUS AND TEST RESULTS

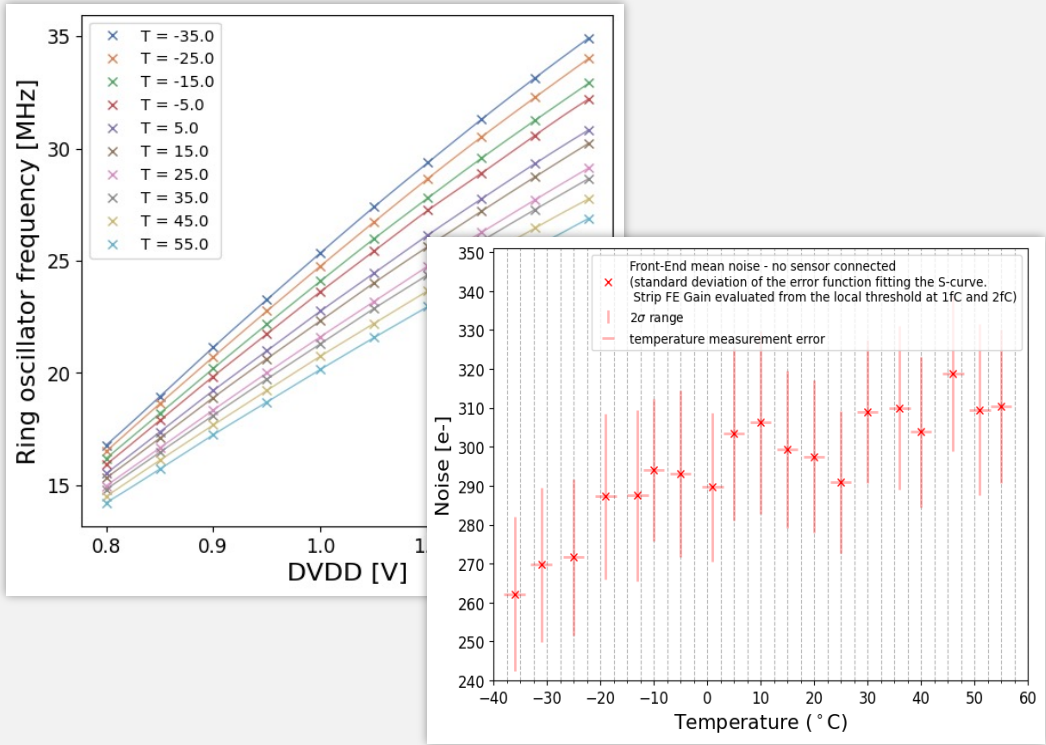
TID Characterization

Tested up to 200 Mrad at different dose-rates.



Temperature measurements

Tested from -40C to 60C



SEE test with heavy ions (MPA-SSA)

No hard errors observed:

- No loss of control observer
- No loss of synchronization observed

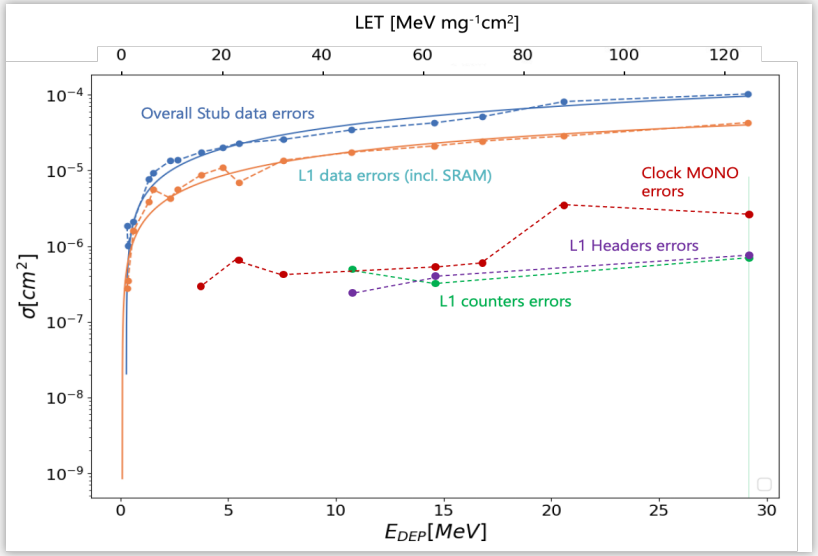
Control/Config path (SEE protected) is error-free:

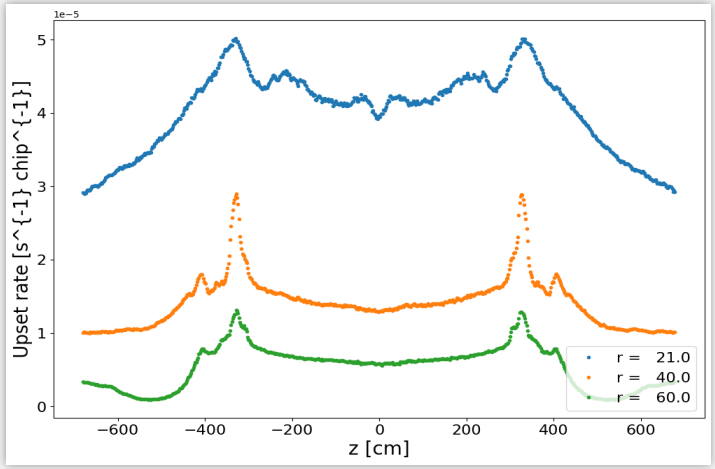
- Verified by readout and comparison of full chip configuration at each test iteration
- No I2C errors for total fluence of 170E6 particles → Limit cross section 5E-9 cm^2

Data path (not protected) Error Rate estimation for HL Tracker Environment:

- Upset Rates at Hadron flux of $4 \cdot 10^7$ cm^{-2} s^{-1}

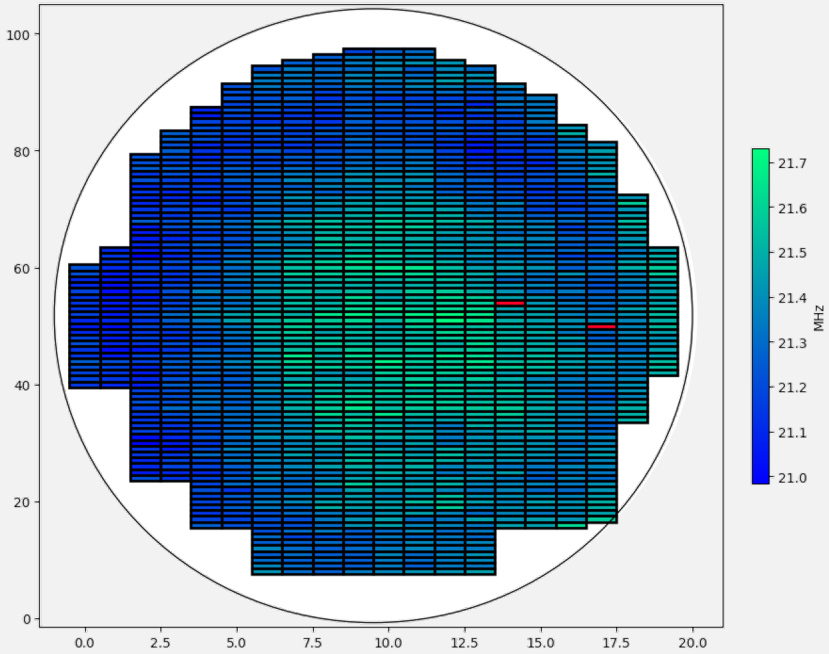
SEU Counter	Stubs	L1
3.34e-3 s^{-1} $chip^{-1}$	5.5e-3 s^{-1} $chip^{-1}$	2.87e-3 s^{-1} $chip^{-1}$





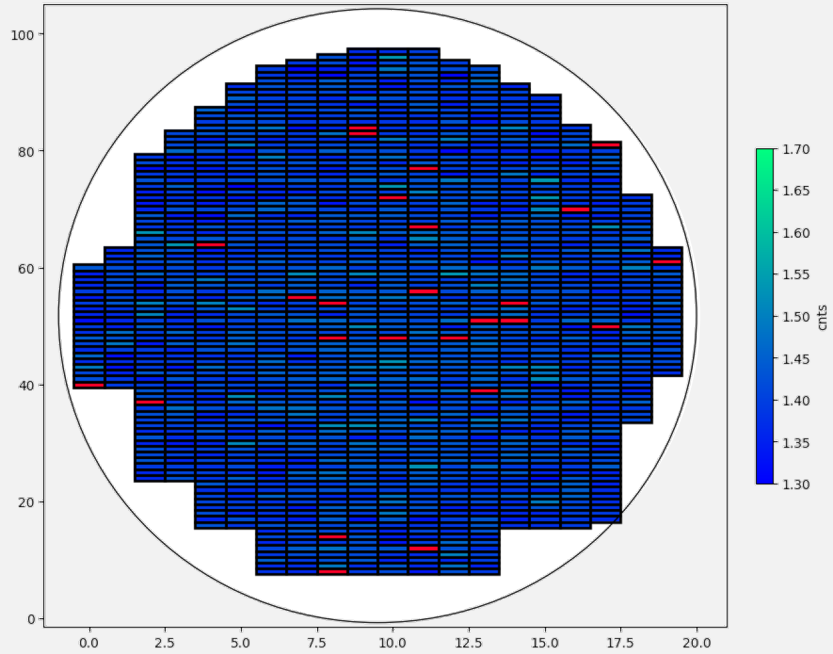
Wafer probing (SSA)

Ring oscillators Frequency



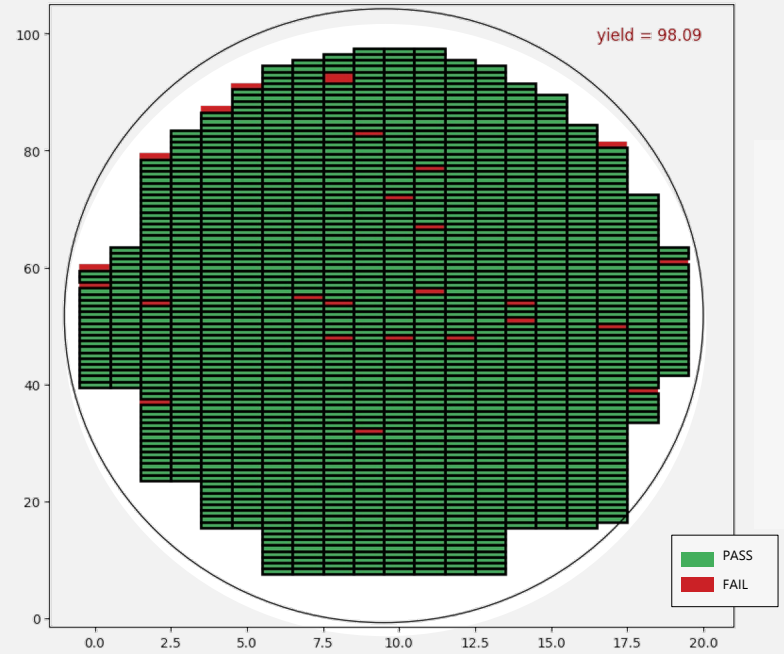
- The SSA includes different types of ring oscillator to monitor variations in:
  Process – Temperature – Total Ionizing Dose

FE Noise Performance Tests



- Map of the average FE noise
- Cut criteria noise < 1.7 LSB

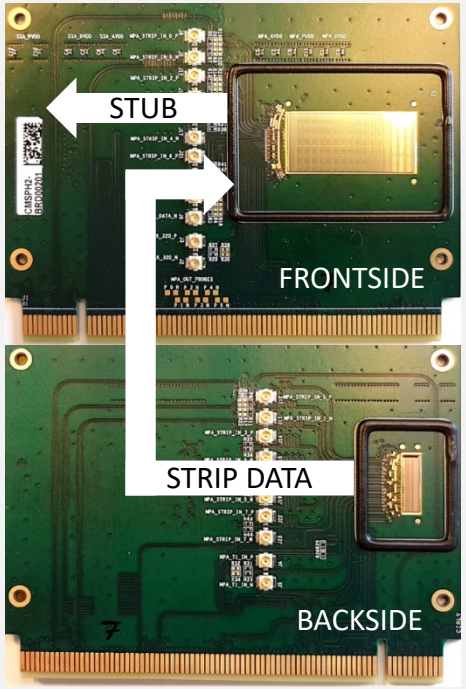
Total yield map



- Stub data and L1 data
- Configuration and memories
- Scan chain and BIST
- Analog bias calibration
- FE functionality, threshold trimming and noise

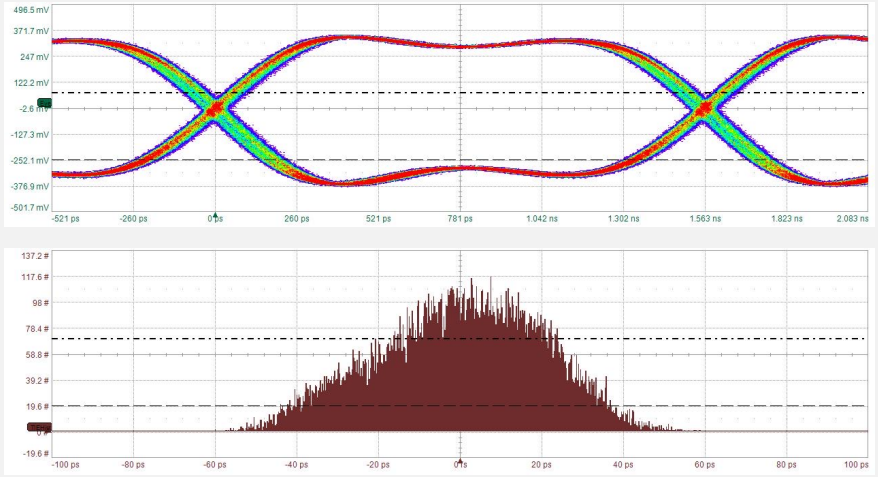
Chip-to-chip communication and **power consumption**

- Real-Time 2.56 Gbps parallel bus provides Strip data from SSA to the MPA.
- MPA 1.6 Gbps output parallel bus provides only selected particle information (Stub).
- High Momentum particle information latency = 500 ns



Eye Diagram – Clock 320 MHz at SSA output – MIN Current setting

Average Eye amplitude: **583.6 mV .** Average Eye Width: **1.442ns** .
Average cross: **49.60 %**



Jitter histogram – Clock 320 MHz at SSA output – MIN Current setting

Power Consumption:

- Configured w/o activity:
  - MPA: 184 mW
  - SSA: 67 mW
- Increase activity:
  - MPA: < +10 mW
  - SSA: < +2 mW
- MAX power for MPA + SSA
  - < 275 mW
  - < 100 mW/cm^2

Summary

- System-level studies allowed to define the architecture of the PS-Module ASICs
- The First full-size and full-functionality MPA, SSA and CIC prototype chips demonstrated the feasibility of on-chip particle discrimination with a power density < 100 mW/cm^2
- The final version of the ASICs (MPA2, SSA2 and CIC2) have been tapeout for a full mask-set engineering run
- The tests on the final version of the chips show results in agreement with the expectation
  - Front-end performances fulfil specifications
  - X-Ray TID test confirms radiation harness up to 200 Mrad
  - Heavy Ion test confirms the functionality of the chosen hardening strategy
  - Climatic chamber tests shows a parameter variation within the calibration range
- The Additional tests are currently ongoing.
- Wafer-level testing show a high yield, which allow us to move to the next step of ordering the production wafer and define the automated production test procedure.

Conclusions