PCIe400 : Development status

Julien Langouët *(CPPM) on behalf of the R&T PCIe400 team*
CPPM, IJClab, IP2I, LAPP, LPCC, LHCb Online

Outline

Context and general characteristics

Technical developments

- Cooling solution
- Placement and routing
- Software framework

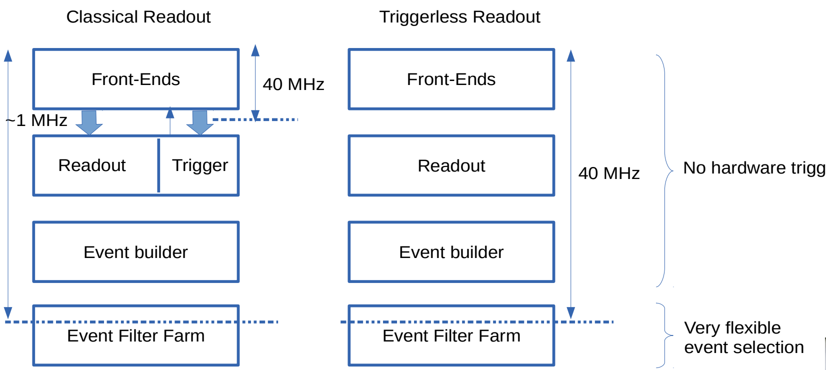
Planning

Synthesis

LHCb acquisition system

“Triggerless” architecture choice

- Resolve trigger saturation issue with increasing luminosity
- Allow for more flexible algorithm
- Adopted during Upgrade I (2019-2022)



Classical Readout

Front-Ends

40 MHz

~1 MHz

Readout

Trigger

Event builder

Event Filter Farm

Triggerless Readout

Front-Ends

Readout

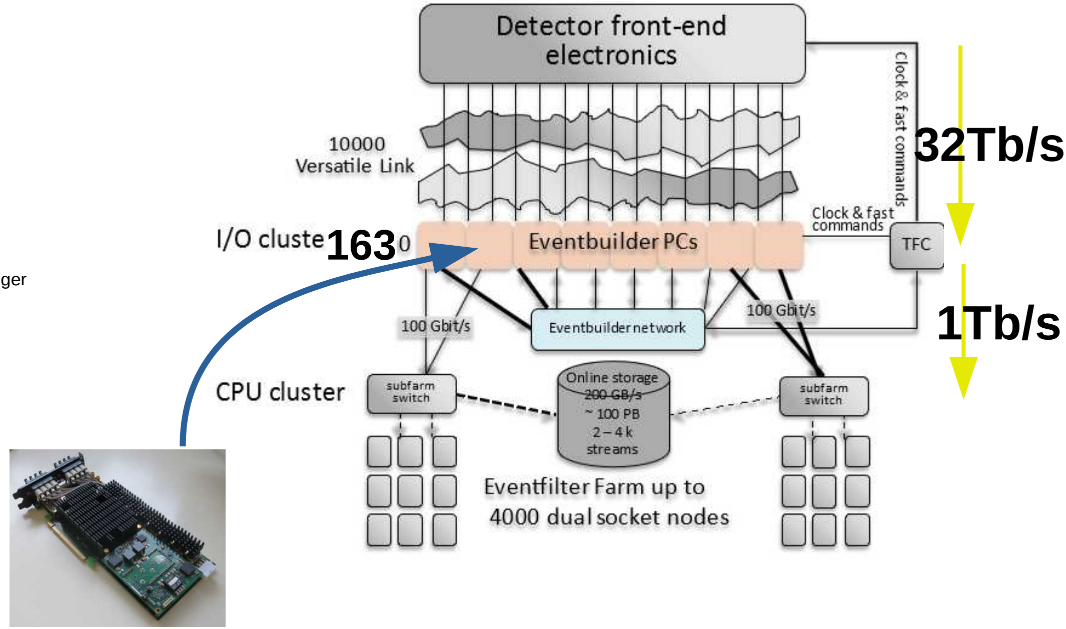
Event builder

Event Filter Farm

40 MHz

No hardware trigger

Very flexible event selection



Context

Goals and rationale

- Similar architecture to Upgrade I : Generic readout DAQ card interfacing 48 custom protocol links (GBT/lpGBT) to 1 commercial protocol link (PCIe Gen5)

- Cope with tighter timing requirement and higher number of channels in LS3 for some subdetectors
  - Bandwidth x4 (400Gb/s)
  - Clock distribution $\mathcal{O}$(10)ps RMS jitter

- Explore experimental path
  - Integrated network interface (400GbE)
  - Integrated complex data processing (neural network)

Target deployment of PCIe400 is during LS3 for upgraded detectors

- Possible interest from Belle II, CTA and Alice collaborations

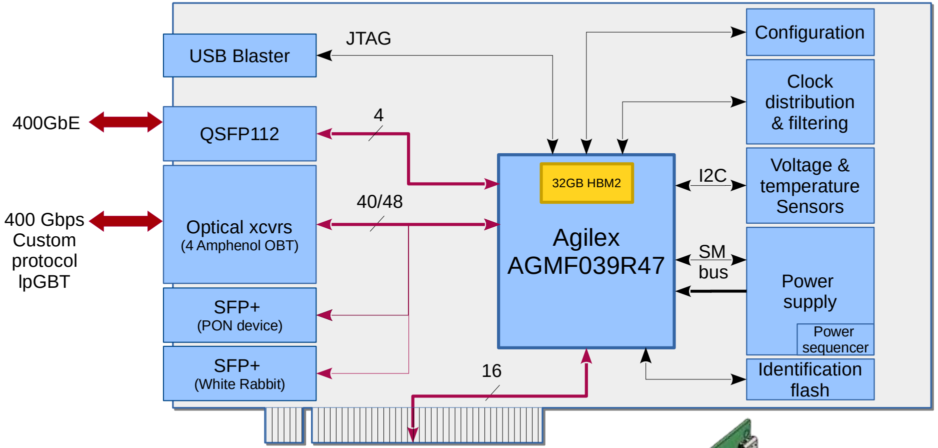
IN2P3 R&D

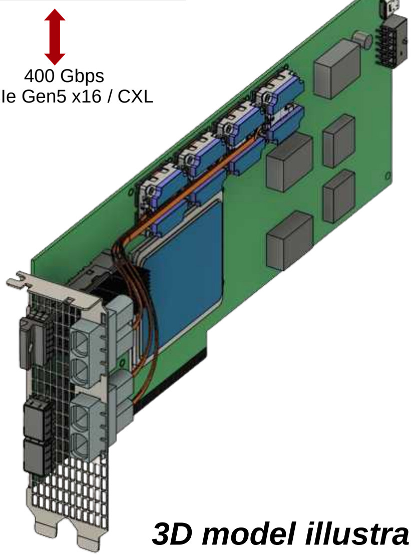
- Project set up to develop the prototype of PCIe40 next generation
- Funded for 3 years from 2022 to end of 2024

PCIe400

Foreseen characteristics

- Agilex M-series AGMF039R47A1E2V
- 32GB HBM2e memory on board
  - No need for DIMM DDR memory
- Up to 48x26Gbps NRZ for FE
- PCIe Gen 5 / CXL
- QSFP112 for 400GbE (experimental)
- 2 SFP+ for White Rabbit clock distribution or PON fast control
- High precision PLLs <100fs RMS





3D model illustration

Optical interface

4x Amphenol OBT

- MPO x24 12 duplex channels
- 1.25G to 26.3G NRZ (match lpGBT requirements)

2x SFP+ (10Gb/s)

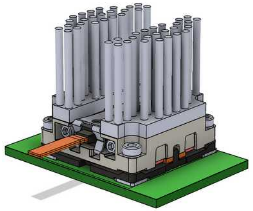
- TTC-PON OLT/ONU for fast control
- White rabbit for clock distribution

QSFP112

- 4x112G PAM4
- Direct Attach Cable <3m or opto <100m

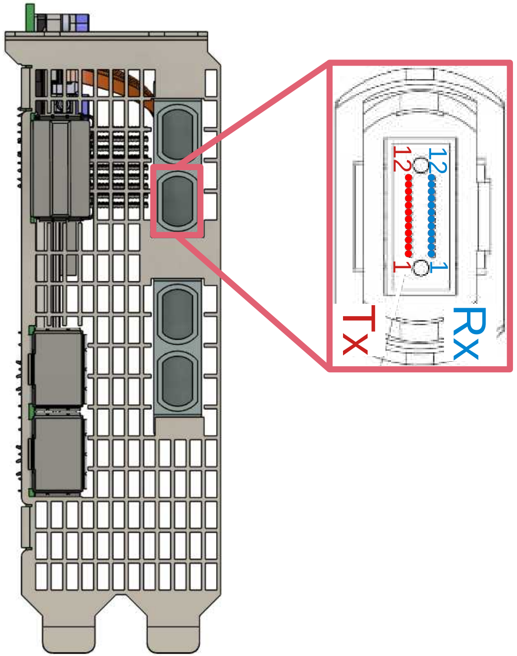
Some links are multiplexed

Usage	# FE links
No TFC/WR/400GbE	48
WR	47
2 TFC (OLT + ONU)	46
2TFC + 400GbE	38



Amphenol OBT
1.25G à 26.3G NRZ

QSFP112
106.25Gb/s PAM4



PCIe400 front-view

Cooling the board

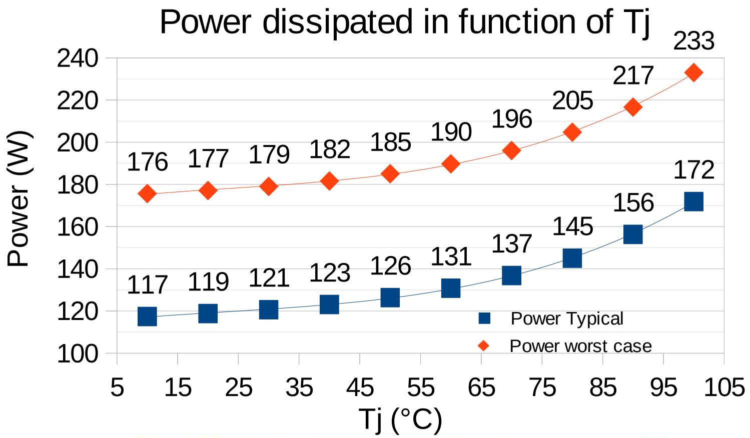
Power dissipation

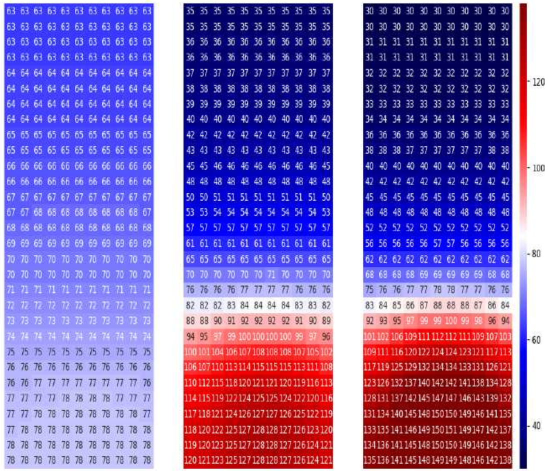
FPGA total power dissipated (TDP)

- Estimation at early stage with limited gateware inputs from developers
  -> risk of over-designing cooling solution
- Estimated between 120W to 230W
- Need for high performance cooling solution

Cooling solution study (with LAPP)

- Specification of the airflow and ambient temperature in PCIe server
- Model of a vapor chamber because of low spread resistance and high surface available using COMSOL
- Comparison of an active and passive heat-sink
- Thermal mock-up developed to verify simulation model





1. Vapor chamber + copper fin

2. Copper base + copper fin

3. Aluminum base + aluminum fin

Heat spread on heatsink base (Heatsink)

Cooling solution design

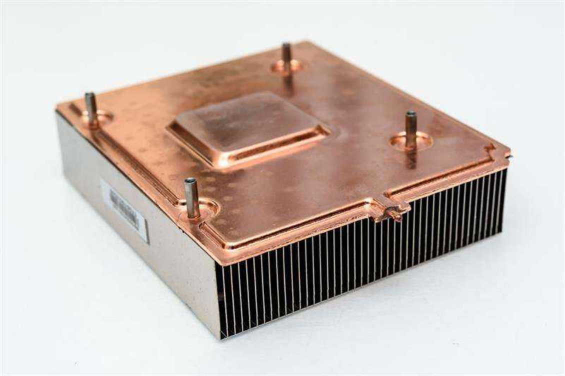
Outsourcing design prototype

- 3 companies responded with interest on our project out of 11 companies contacted

Cecla Metal Process

- Direct access to thermal engineers
- Low cost of heat-pipe technology compared to vapor chamber with high NRE
- Preliminary CFD study undergoing to find maximum power dissipable with such technology
- Mechanical study for fixation points of heat-sinks

Heatpipe heatsink illustration



Vapor Chamber heatsink illustration

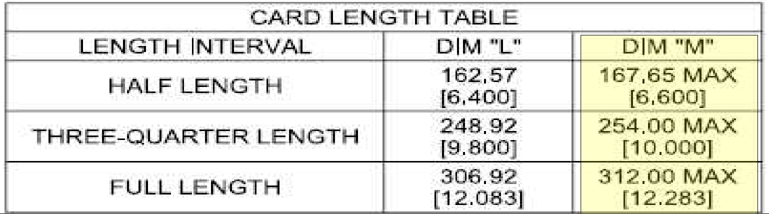
Routing the board

Placement overview

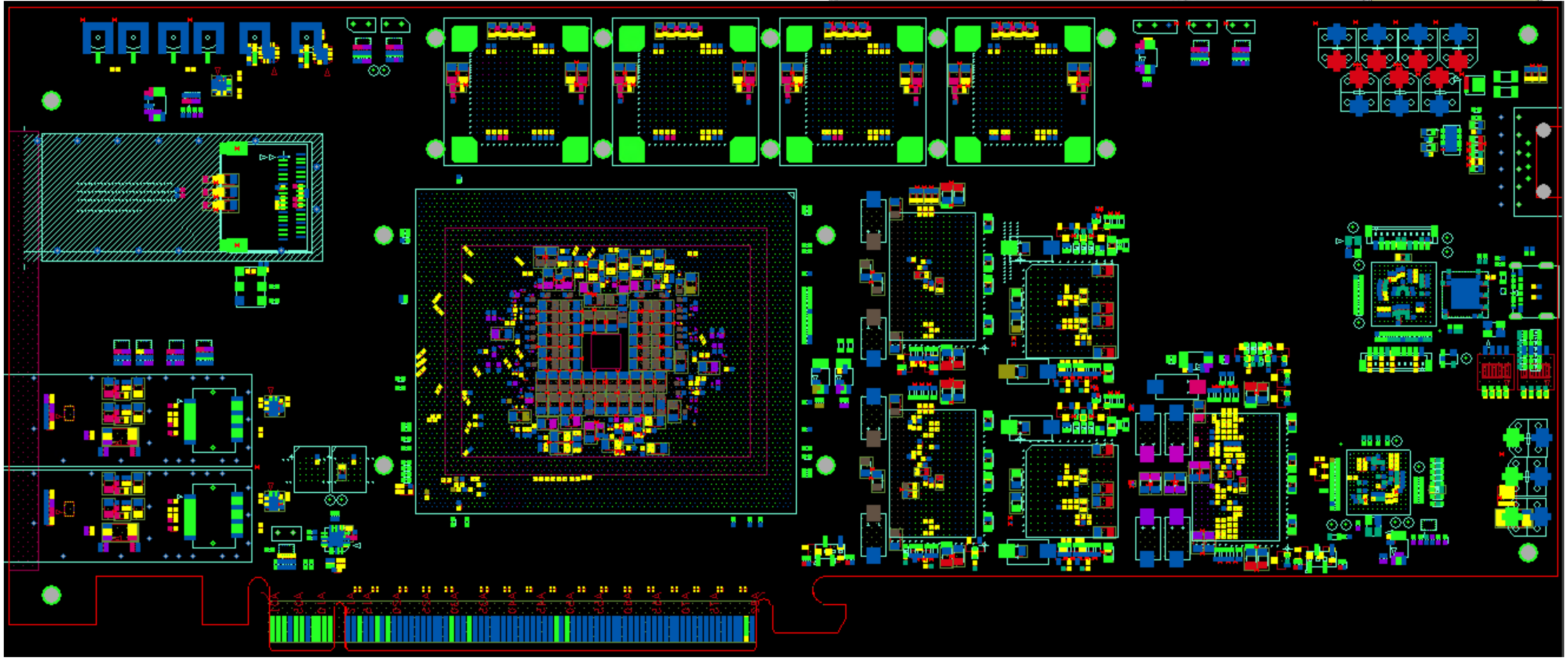
PCB dimension : 270 x 100mm

2300 components on board

PCIe specification



CARD LENGTH TABLE		
LENGTH INTERVAL	DIM "L"	DIM "M"
HALF LENGTH	162,57 [6,400]	167,65 MAX [6,600]
THREE-QUARTER LENGTH	248,92 [9,800]	254,00 MAX [10,000]
FULL LENGTH	306,92 [12,083]	312,00 MAX [12,283]



Stack-up choice

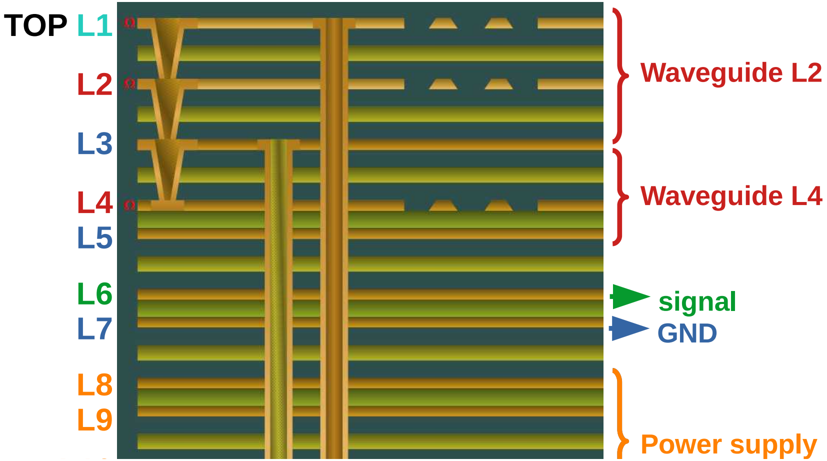
Highlight constraints

- >100A on FPGA Core and ~20 power rails from 0.8V to 5V
- 108 differential pairs at 10, 28 and 112Gb/s (PAM4)
- 0.9mm pitch BGA on FPGA
- 1.57mm thickness imposed by PCIe
- Controlled impedance 7 % for 400GbE diff. pairs

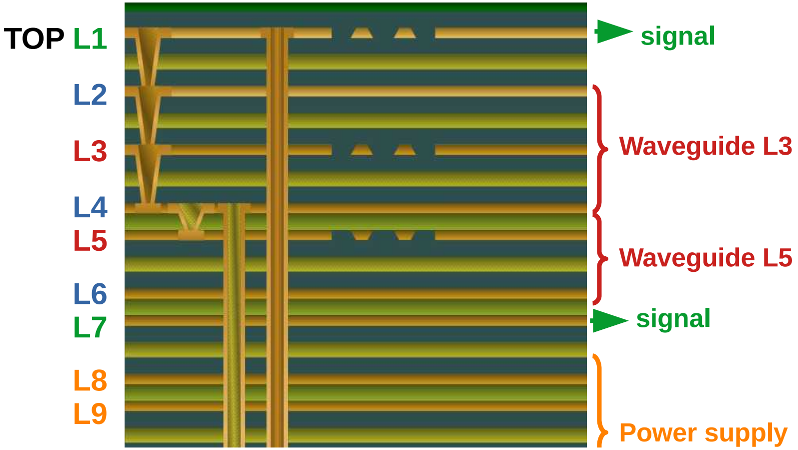
2 solutions envisaged : 18 layers

Criteria	3 level µvias	4 level µvias
Lower cost	☺	
Reduced number of µvia in path	☺	
Lower constraints on placement		☺

3 level µvias



4 level µvias



Power integrity simulation

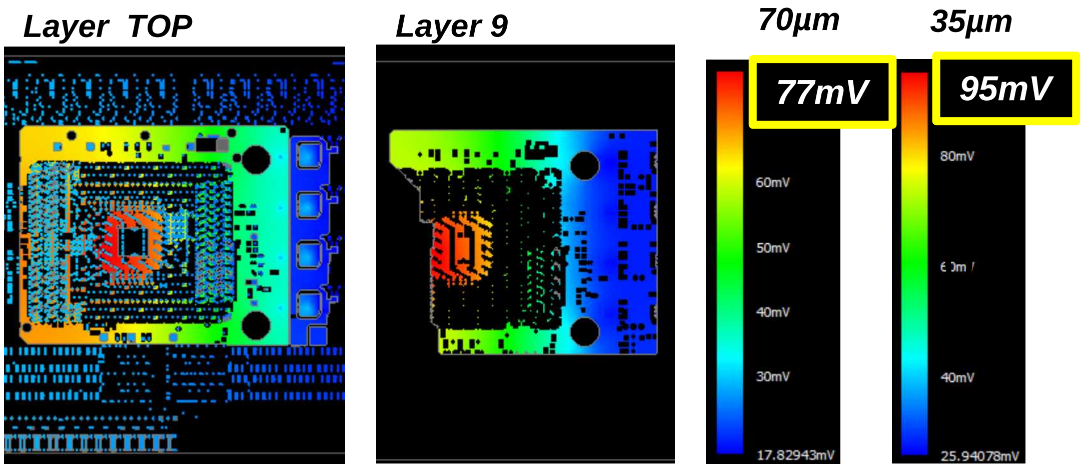
PCB thickness should not exceed 1.70mm (1.57 ±0.13mm) PCIe spec

- 4 layers of 70µm copper foil for power planes results in **1.78mm PCB finished**

Solution is to reduce power planes thickness from 70µm to 35µm

- Simulation of FPGA core voltage rail (0.8V @200A) using Intel FPGA Agilex I-series layout (4 power stage + controller)

Voltage drop in power planes



Power dissipation within power plane

Power dissipated	70µm	35µm	Δ
Layer TOP	9.6W	11.0W	+14 %
Layer 9	4.9W	6.8W	+38 %
TOTAL	14.5W	17.8W	+23 %

Simple PCB thermal model gives a ~30°C temperature rise in PCB with 17.8W over 33cm²

- 4 layers of 35µm copper foil for power planes results in **1.57mm PCB finished**

Signal integrity simulation

Goal

- Check impact of topologies (stacked/staggered µvias, stitching vias, ...)
- Ensure functionality without overdesign

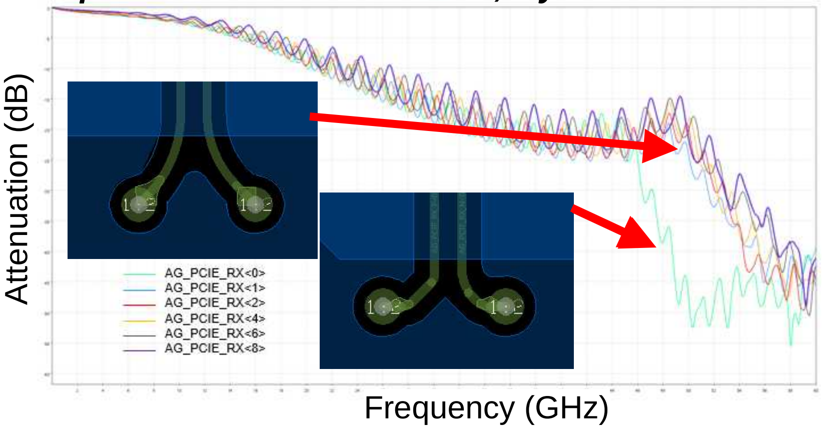
S-parameter extraction of striplines

- Rapid process to observe reflection due to bad geometry in layout

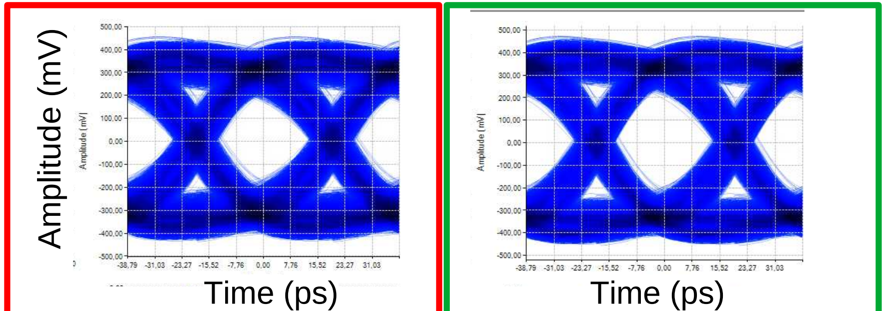
Temporal simulation

- Verify in depth transmission lines functionality

S-parameter PCIe Rx ch0 to 8, layout defect observed



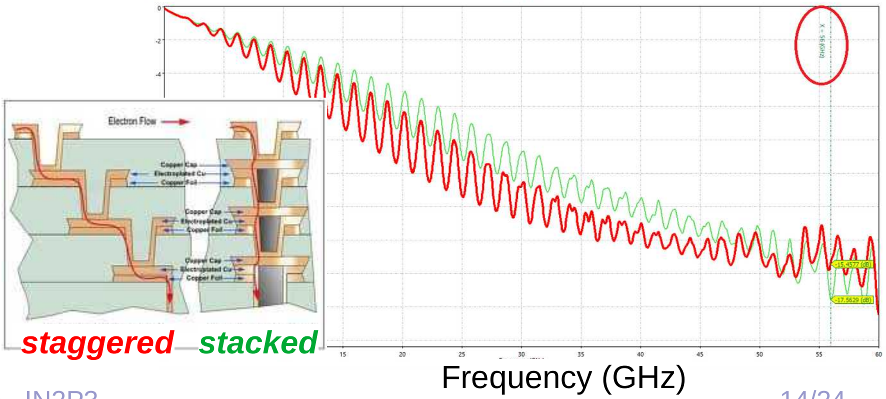
Comparison 2 layers µvias stacked or staggered



H = 321 mV L = 0.59 UI

H = 382 mV L = 0.62 UI

Comparison 2 layers µvias stacked or staggered



Monitoring the board

Several access path

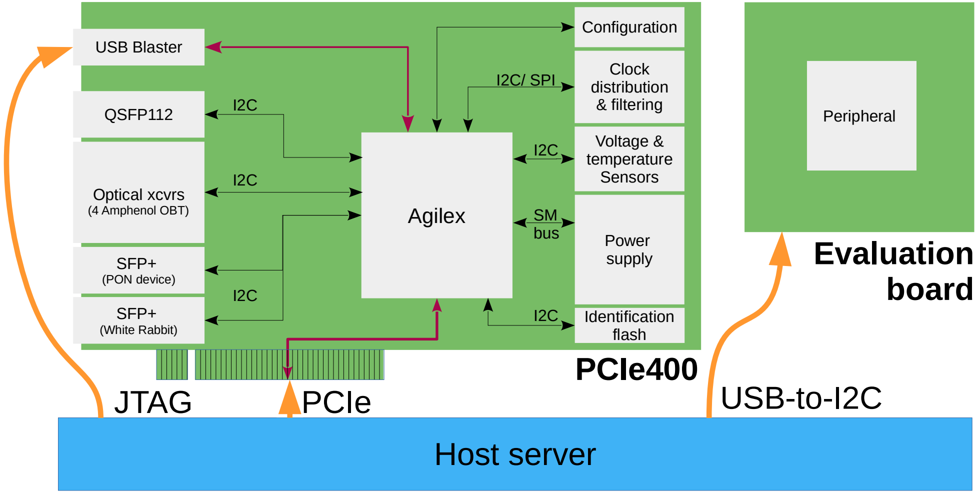
Goal

- Implement abstraction layers : ‘Low Level Interface’ to help developers focus only on their core skills
- Ensure testability for reliable production and long term support

Access to board peripherals are centralized on the FPGA

- Several bus to access FPGA : JTAG, PCIe
- Use of USB-to-I2C cable for early development phase on peripheral evaluation board

PyPi package to manage different bus drivers with same code



Software framework

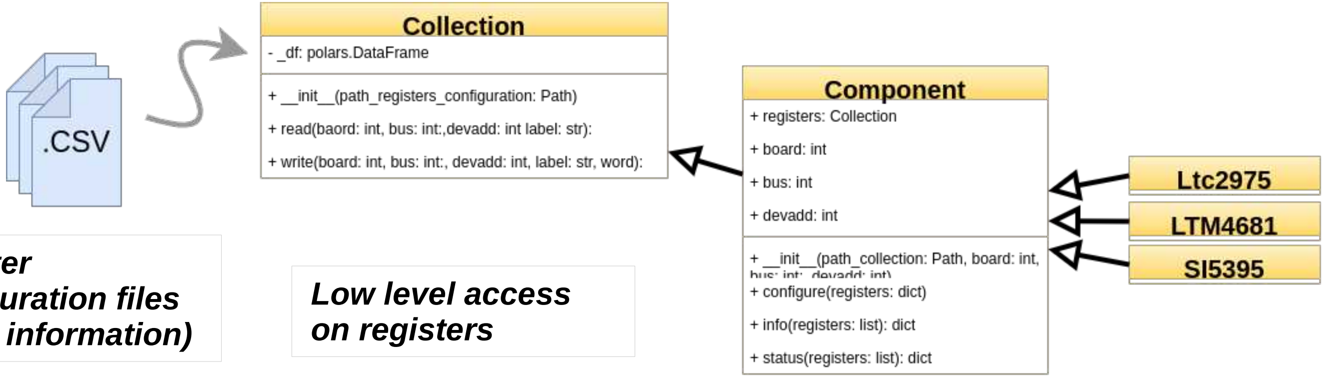
Development with help from LP2i Bordeaux

Peripheral components

- Each component can be described by a list of registers with limited number of fields

Implementation

- Take advantage of Dataframes and its manipulation methods in order to efficiently access any register fields.



Collection

- _df: polars.DataFrame

+ __init__(path_registers_configuration: Path)

+ read(baord: int, bus: int:,devadd: int label: str):

+ write(board: int, bus: int:, devadd: int, label: str, word):

Component

+ registers: Collection

+ board: int

+ bus: int

+ devadd: int

+ __init__(path_collection: Path, board: int, bus: int, devadd: int)

+ configure(registers: dict)

+ info(registers: list): dict

+ status(registers: list): dict

Ltc2975

LTM4681

SI5395

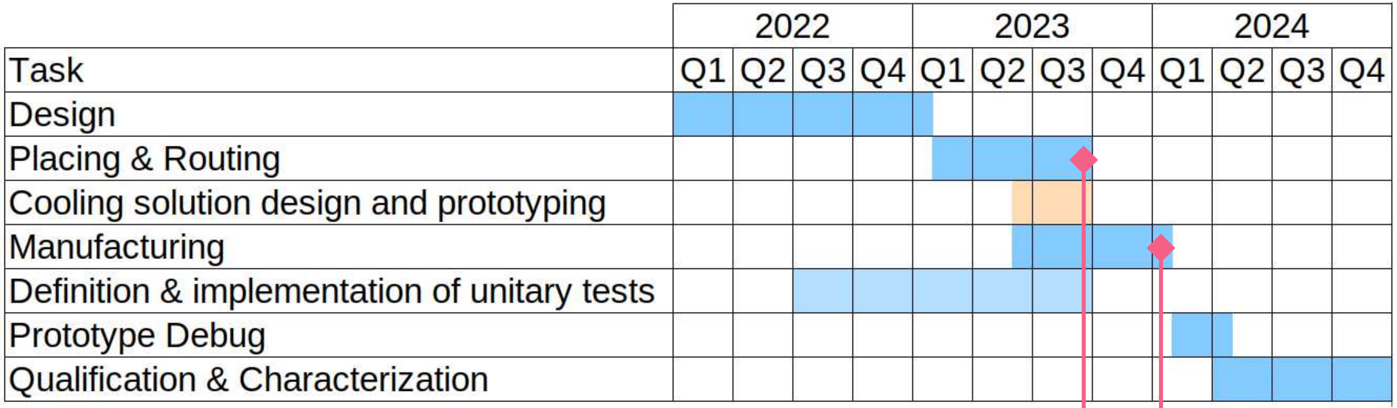
Register configuration files (static information)

Low level access on registers

Provides common methods to access to batches of registers

Component model : Provide high level specific methods for usefull actions

Planning



	2022				2023				2024			
Task	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4
Design	🟦	🟦	🟦	🟦	🟦							
Placing & Routing					🟦	🟦	🟦 ◆					
Cooling solution design and prototyping						🟧	🟧					
Manufacturing						🟦	🟦	🟦	🟦 ◆			
Definition & implementation of unitary tests			🟦	🟦	🟦	🟦	🟦					
Prototype Debug									🟦	🟦		
Qualification & Characterization										🟦	🟦	🟦

Prototype available Jan. **2024**

Routing review September **2023**

Placing is planned to be finished in June 2023

- Delay due to reassignment of pins on FPGA to maximize HBM internal bandwidth

Routing is scheduled to be finished in September 2023

- Delay according to initial planning due to difficult stack-up choice

PCB manufacturing is longer than expected with european manufacturer

Synthesis

Hardware

- A passive cooling solution is under discussion with outsource specialists for future design and prototyping
- Stack up is under review with several PCB manufacturer to ensure manufacturing feasibility
- Routing can not start as long as topology is not validated
- Power and Signal integrity simulations are vital to validate layout

Firmware and software

- A software framework has been implemented for efficient board monitoring based on PCIe40 experience
- Test firmware should be available as soon as prototypes are available

Project development

- 3 month delay due to stack-up choice and FPGA pinout optimization after Intel schematics review
- Technical design review with CERN planned by end Summer 2023
- LHCb-internal review planned for Q1 2024

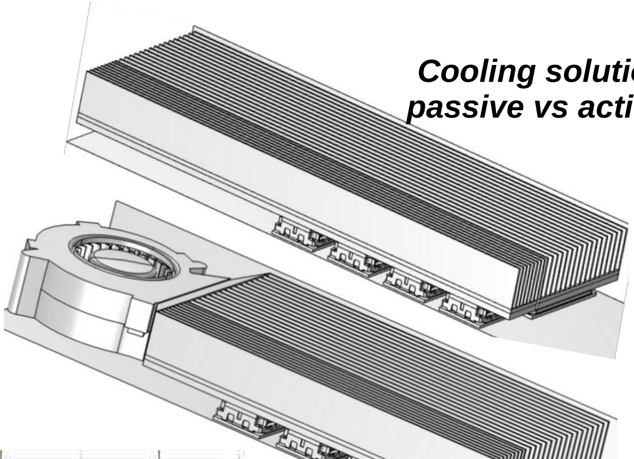
Backup

Cooling solution study

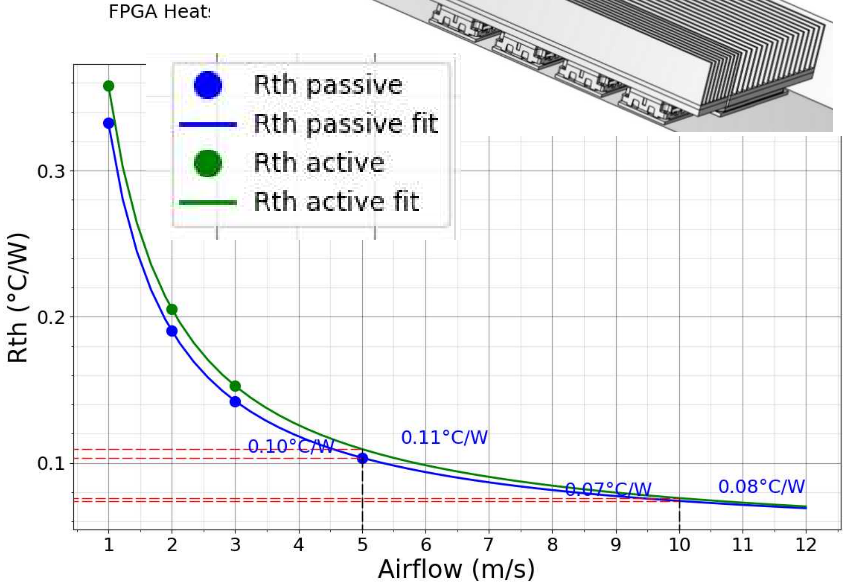
Air cooling study

- FPGA 160W nominal power dissipation
- OBT Optical transceivers 30W must be studied : high power concentration, placement constraints
- Passive and active (with largest fan that fits >6m/s) comparison
- Nominal ambient temperature 38°C
- Up to 5m/s (980LFM) simulated servers may reach higher flow ?

Passive cooling solution show better results and is less constraining mechanically

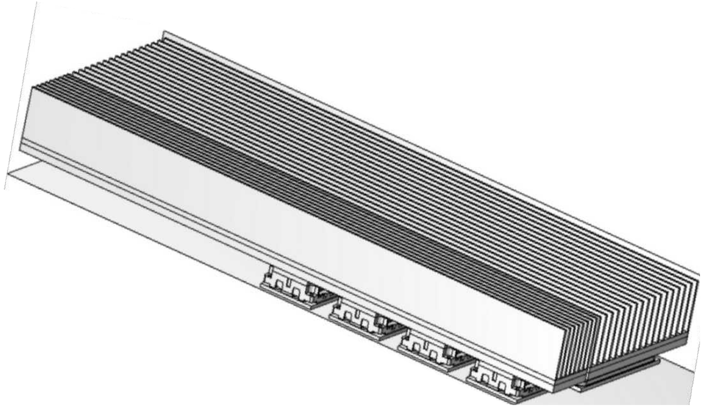


Cooling solution passive vs active



Heatsink Thermal resistance

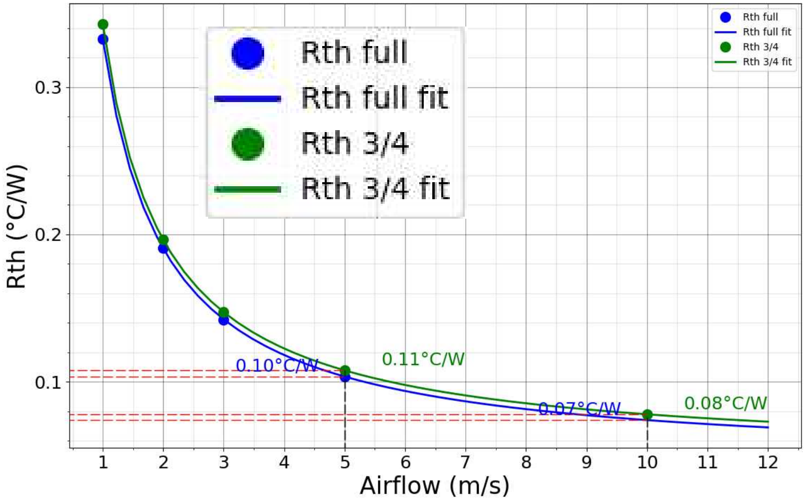
Cooling solution : Length impact



Low impact of heat-sink length

- Full : PCB 312mm / Heat-sink 237x70x23mm
  3/4 : PCB 254mm / Heat-sink 179x70x23mm
- ‘GPU’ : PCB 268mm / Heat-sink 193x70x23mm

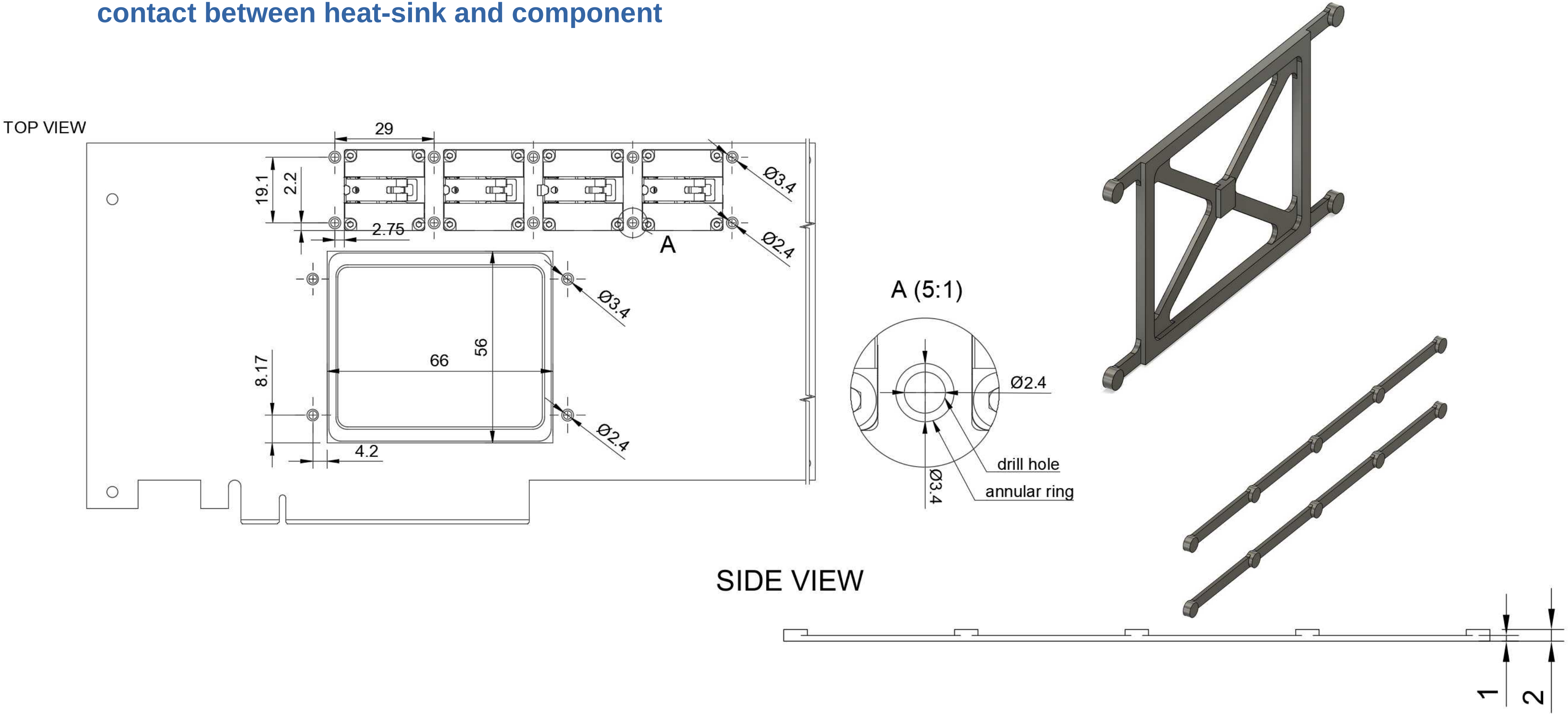
FPGA RthJA comparing 'full' and '3/4' length



Heatsink Thermal resistance

Heatsink mouting points

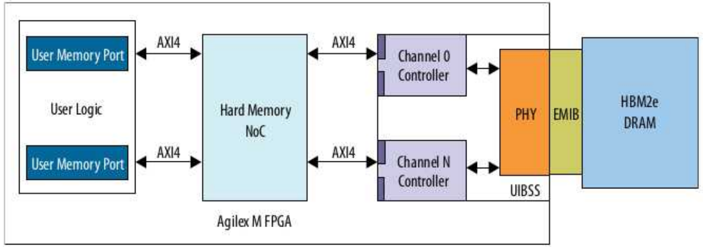
A backplate can be used on PCB bottom side to mitigate PCB warpage and ensure uniform thermal contact between heat-sink and component



Maximize HBM bandwidth

HBM is connected to the FPGA fabric thanks to AXI4 bus

- Each HBM pseudo-channel is connected to the NoC through ‘targets’
- Each User memory port are called ‘initiators’
- Targets and initiators can all be connected thanks to the NoC



HBM maximum bandwidth (on 1 HBM die) is 358GB/s

- 16 initiators are required to saturate HBM bandwidth (256b @700MHz)
- 1 supplementary initiator might be required for HBM monitoring

Initiator hardware resource are multiplexed with some GPIO buffers

- Limits the number of GPIO usable with related initiator

Illustration of Assigned GPIO pins preventing use of multiplexed GPIO with Noc initiators using Interface Planner