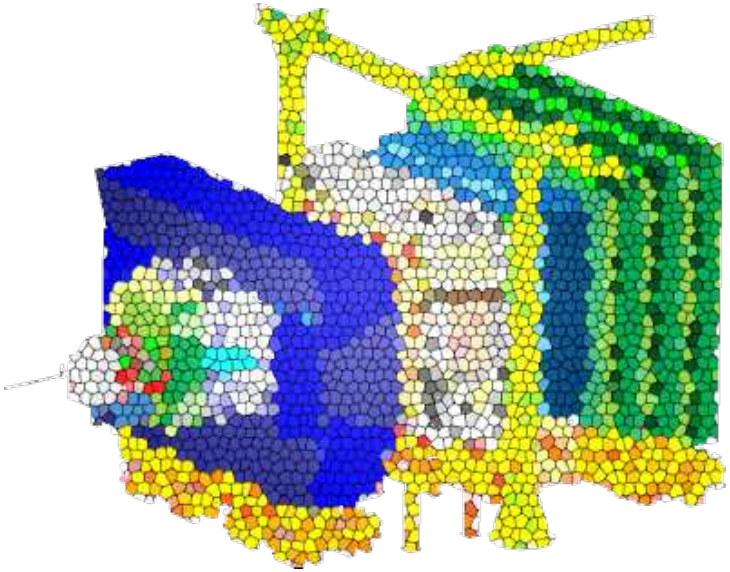
Future DAQ boards : PCIe400



6th Workshop on LHCb Upgrade II

29.03-31.03 2023
Barcelona

IN2P3
Institut national de **physique nucléaire** et de **physique des particules**

Julien Langouët *(CPPM) on behalf of the R&D PCIe400 team*
CPPM, IJClab, IP2I, LAPP, LPCC, LHCb Online
Special thanks to Sophie Baron and Ken Wyllie

Outline

- Baseline and future consideration
- PCIe400 R&D project
- Technical development
- Synthesis

Baseline and future considerations

Readout board goals and rationale

Develop a **common** readout card **aggregating** as many links as possible from font-end with custom protocol (GBT/lpGBT) to output through a **HPC protocol** link (PCIe /Ethernet)

- The adoption of a common front-end serializer (VTRx+) calls for a common readout board across sub-detectors and experiments
- The gateway between custom and HPC protocols allows for flexible and cost effective architecture for event building.
  - Eg : Heterogeneous GPU/CPU event-building on LHCb Upgrade I
- LHCb Upgrade II targets an overall bandwidth increase x5 compared to Run 3 (200Tb/s)

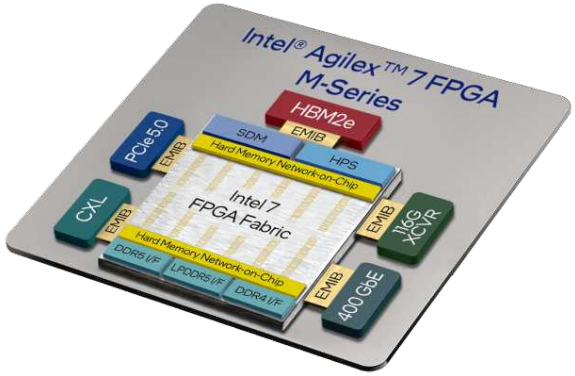
Distribute **high precision clock** to the front-end with **low jitter** and **fine phase control**

- Arise of 4D-techniques for sub-detectors such as the Velo, ECAL, ... push for stringent timing requirements
- Low jitter $\mathcal{O}(10)$ps is required for precision time measurement in sub-detectors
- Phase control resolution $\mathcal{O}(100)$ps is required to synchronize the whole detector

Keeping pace with COTS evolution

FPGA are still reigning for our application

- Growing size FPGA provides flexibility to implement custom protocol and low level data processing (concatenation, primitive)
- New FPGA have real System on Chip with hard processor and high bandwidth memory embedded
- Opens opportunities to process more within the FPGA



HPC driven evolution

- COTS development is driven by big data applications. It is a challenge to shape them for HEP
- New FPGA have higher speed transceivers with reduced number of links
  Eg : Previous generation : 72 XCVR, up to 17Gb/s NRZ
  New generation : 48 XCVR, up to 32Gb/s NRZ, 112Gb/s PAM4
- Ethernet topology enforces symmetric bandwidth while HEP only need high up-link bandwidth for readout and limited down-link bandwidth for fast control $\mathcal{O}$(2Gb/s)

Keeping pace with COTS evolution

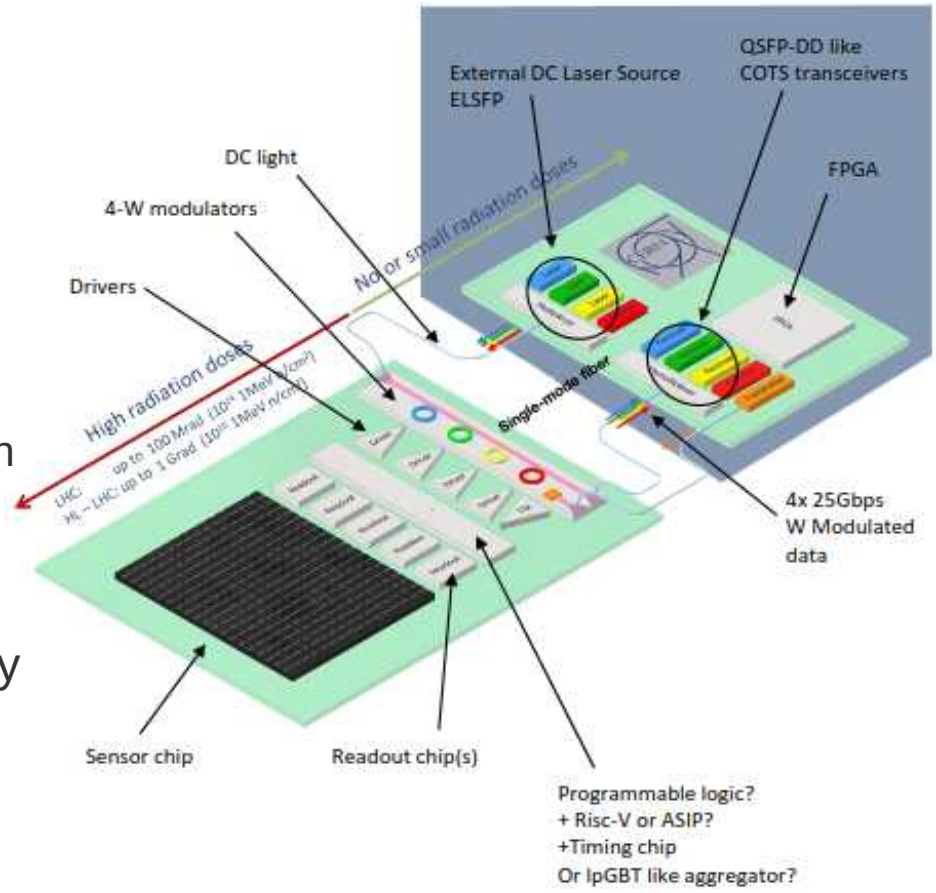
Front-end

- LS3/LS4 sub-detector upgrade targets an increasing number of links with modest bandwidth (10Gb/s) relying on VTRx+ (lpGBT)

- Current R&D program lead by CERN EP-ESE aims at studying emerging technologies :
  - Future of optical links : 100Gb/s aggregated over 4λ (Wavelength Division Multiplexing)
  - Silicon Photonics co-packaging
  - Ethernet-based readout links at front-end -> huge challenges on protocol handling under radiation and LHC-asynchronous links

Choices for future readout boards

- **Aggregator** oriented : acting as protocol gateway
- **Processing** oriented : taking advantage of local resources to process data online



Silicon Photonics readout system example (courtesy of Sophie Baron)

PCIe400 R&D project

R&D PCIe400 project

Goals

- Develop a generic PCIe readout card with up to 48 (GBT/lpGBT) compatible links to PCIe Gen5 or 400GbE
  Output bandwidth x4 compared to previous generation (400Gb/s)
- Explore experimental path to test LS4-oriented features such as
  - Integrated 400GbE network interface,
  - White rabbit node for clock distribution,
  - Cache coherent transaction through PCIe (CXL) for co-processing

Target deployment during LS3 for upgrade sub-detectors

- LHCb (Calo, RICH, Mighty Tracker, Magnet Station)
- Interest from other collaboration Alice, Belle II and CTA

IN2P3 R&D

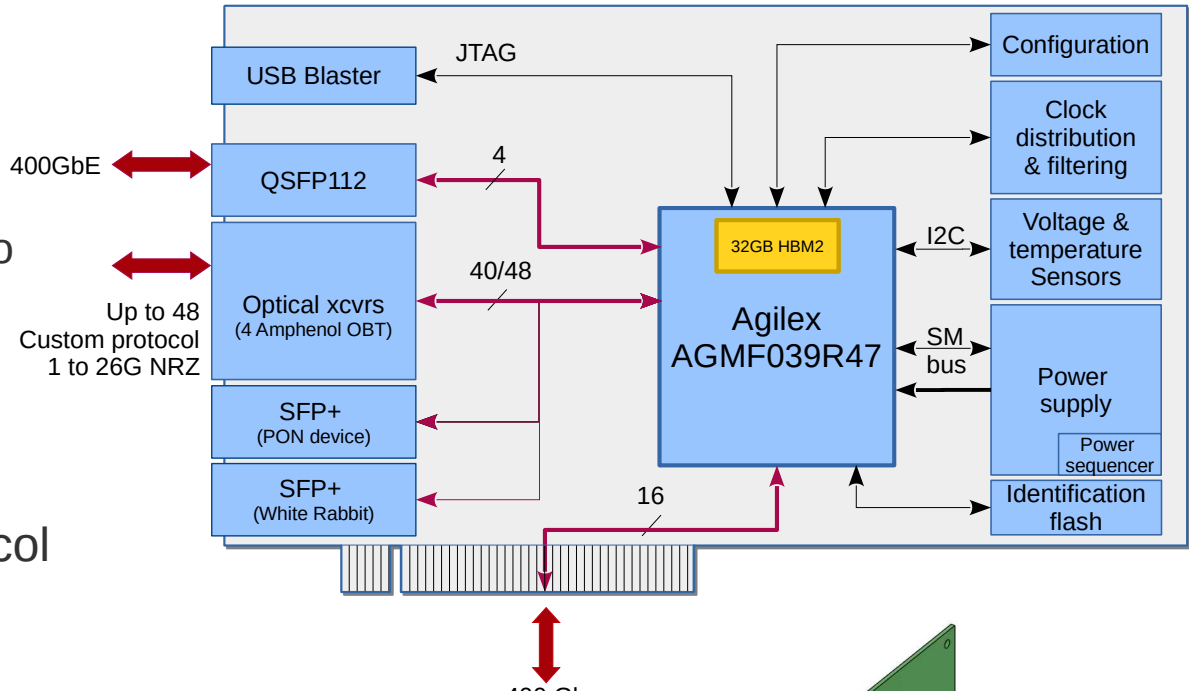
- Project funded for 3 years from 2022 to end of 2024 covering prototyping phase
- Potential production is anticipated taking benefits from PCIe40 production and support experience involving several labs in IN2P3 and LHCb online team

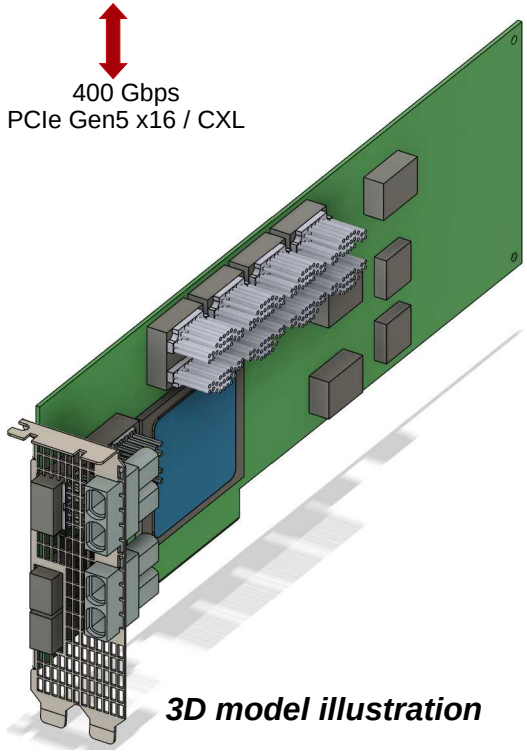
PCIe400

Characteristics

- Agilex 7 M-series AGMF039R47A1E2V
  - Processing capabilities x8 - 12 compared to previous generation FPGA (Arria 10)
- No DDR memory
  - Use of server memory or HBM2e instead
- Up to 48x26Gbps NRZ for custom protocol
- PCIe Gen 5 / CXL
- QSFP112 for 400GbE (experimental)
- 2 SFP+ for White Rabbit clock distribution or PON fast control

PCIe400 synoptic





3D model illustration

Optical interface

Limited number of FPGA transceiver

- Number of serials links for front-end depend on configuration

4x Amphenol OBT for custom protocol

- 12 duplex channels (MPO-24)
- Compatible with VTRx+

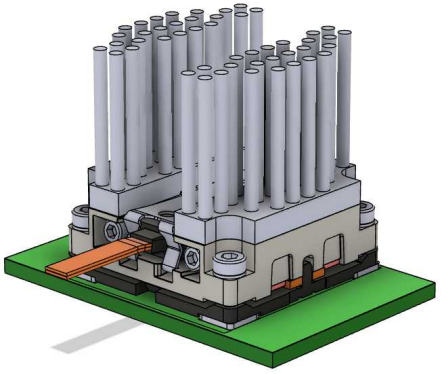
2x SFP+ 10G for TFC / White Rabbit

- Anticipated evolution of White Rabbit over 10GbE

QSFP112 for 400GbE (4x112G PAM4)

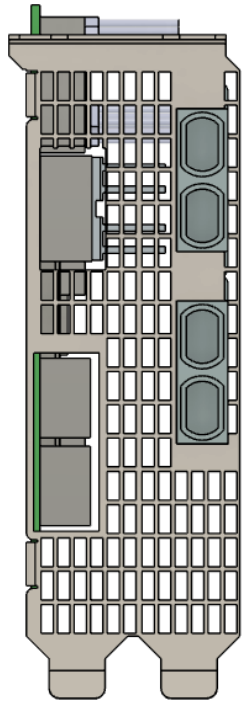
- Backward compatible with QSFP modules
- Natural evolution from QSFP56 (200G)
- Optical modules slowly become available

	# FE links
No TFC/WR/400GbE	48
WR	47
TFC (TTC-PON)	46
TFC (TTC-PON) + 400GbE	38



Amphenol OBT
1.25G à 26.3G NRZ

QSFP112
106.25Gb/s PAM4



PCIe400 front-view

Technical development

Clock tree

Clock tree based on 2 jitter cleaner external PLL

- Clock scheme simplified and adjusted from PCIe40 design
- Component selection with help from CERN EP-ESE

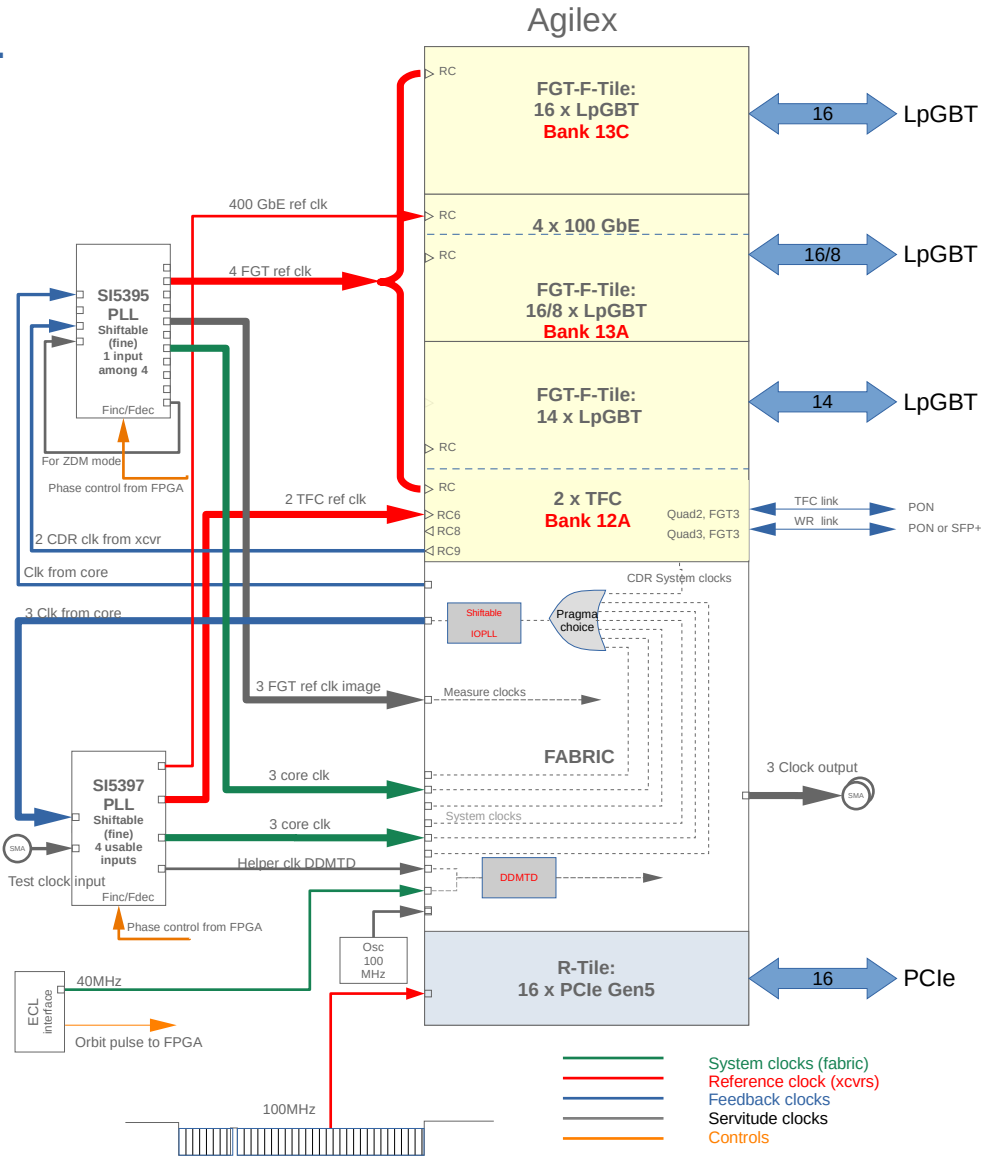
Several clock source scheme

- External PLL as clock generator
- LHC clock from connector (ECL)
- Recovered clock from SFP+

Several phase control schemes

- External PLL registers
- Internal FPGA PLL registers
- Controlled ppm frequency shift with external PLL (experimental)

Effort is required during board characterization to ensure timing requirements on a realistic test bench

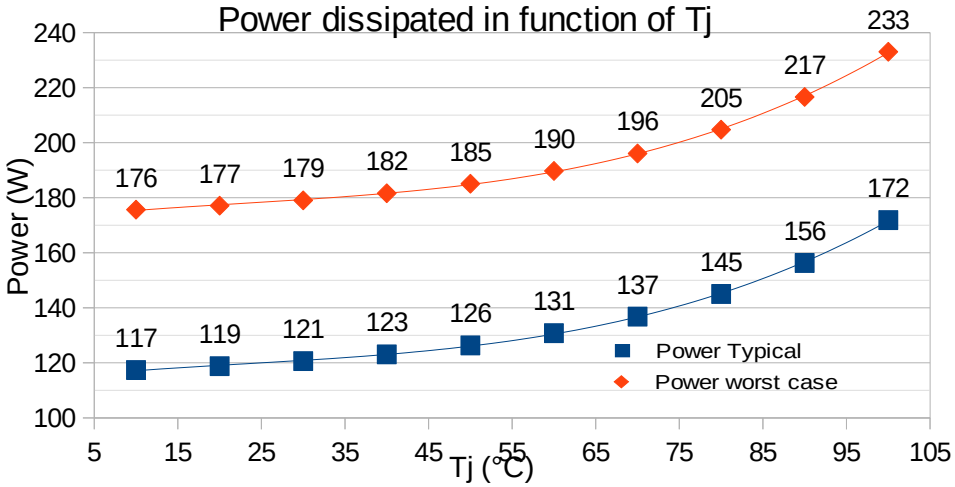


Simplified Clock tree synoptic

Power dissipation

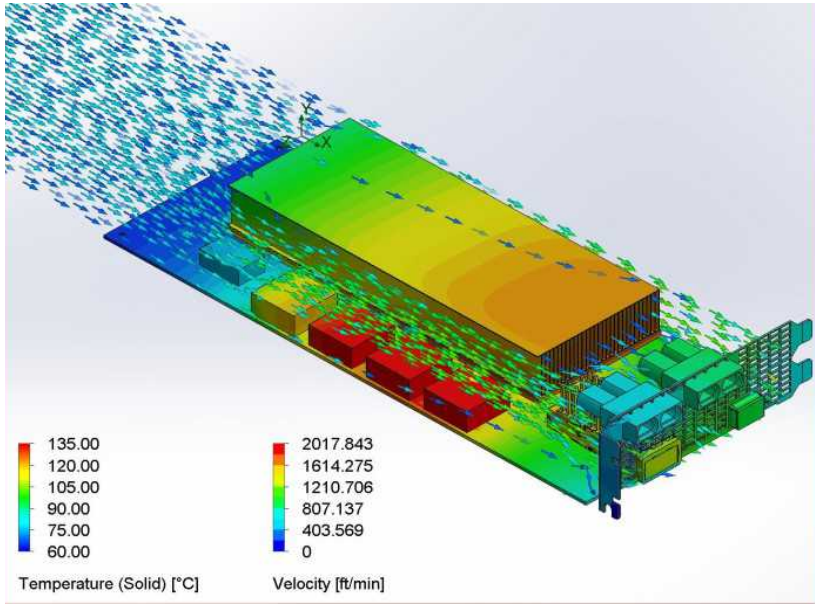
FPGA total power dissipated (TDP)

- Estimation at early stage with limited gateware inputs from developers
  -> risk of over-designing cooling solution
- Estimated between 120W to 230W
- Up to 100A current for FPGA core
- Need for high performance cooling solution



Cooling solution

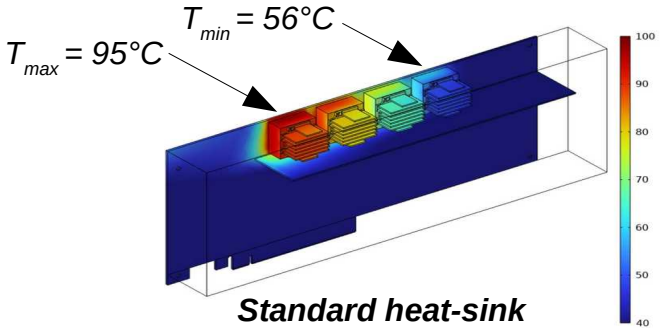
- PCIe form factor requires compact cooling solution
- PCIe specification are blurry regarding airflow in chassis compared to ATCA
- CFD simulations to study air cooling feasibility with vapor chamber heat-sink
- Thermal mock-up developed to measure airflow in situ



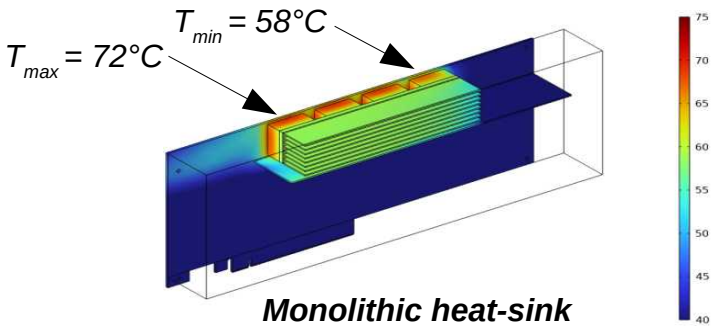
Cooling solution architecture choice

Standard heat-sink for optical transceivers is not appropriate

- Monolithic vapor chamber heat-sink to optimize airflow over optical transceiver



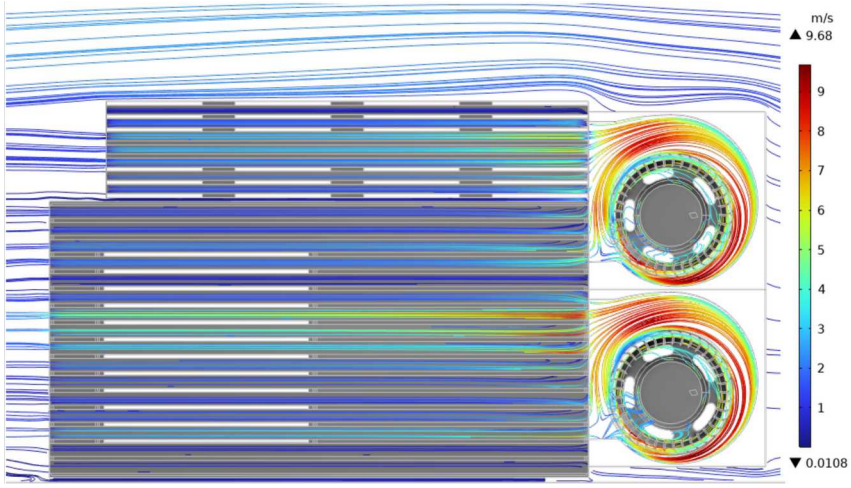
Standard heat-sink



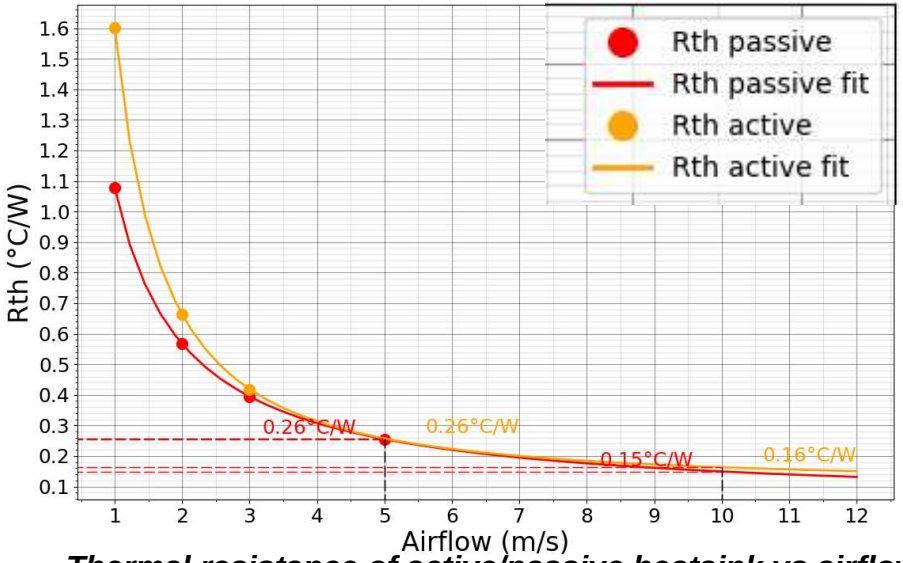
Monolithic heat-sink

Active solution to mitigate dependence to chassis ?

- Blower fans have non uniform airflow over venting width and induce high placement constraints
- No large improvement of thermal resistance compared to passive heat-sink



Top view airflow speed through heatsink with active blower fans



Thermal resistance of active/passive heatsink vs airflow

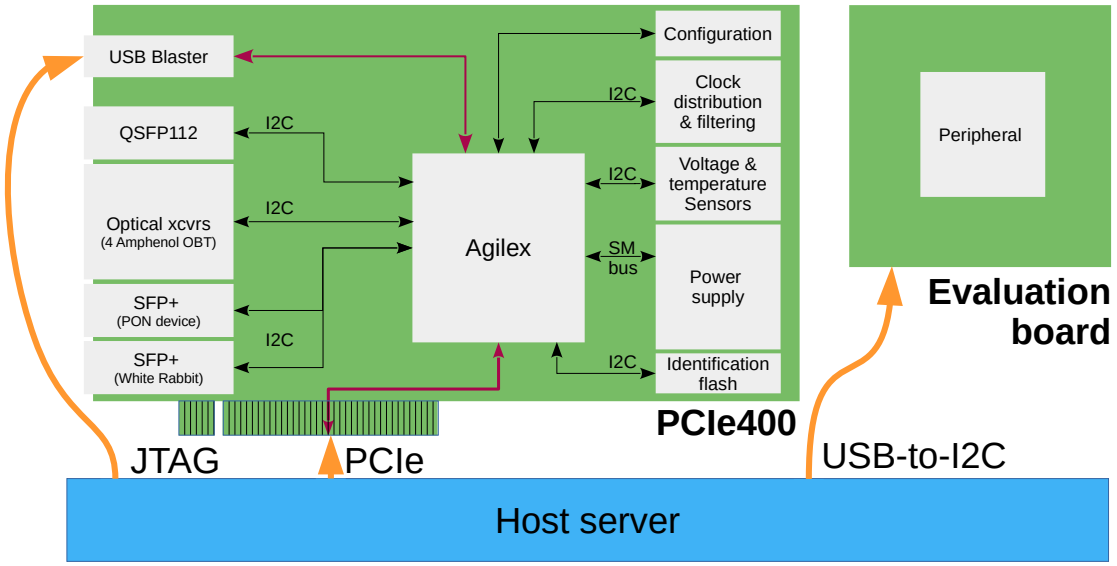
Software and Gateware development

Goal

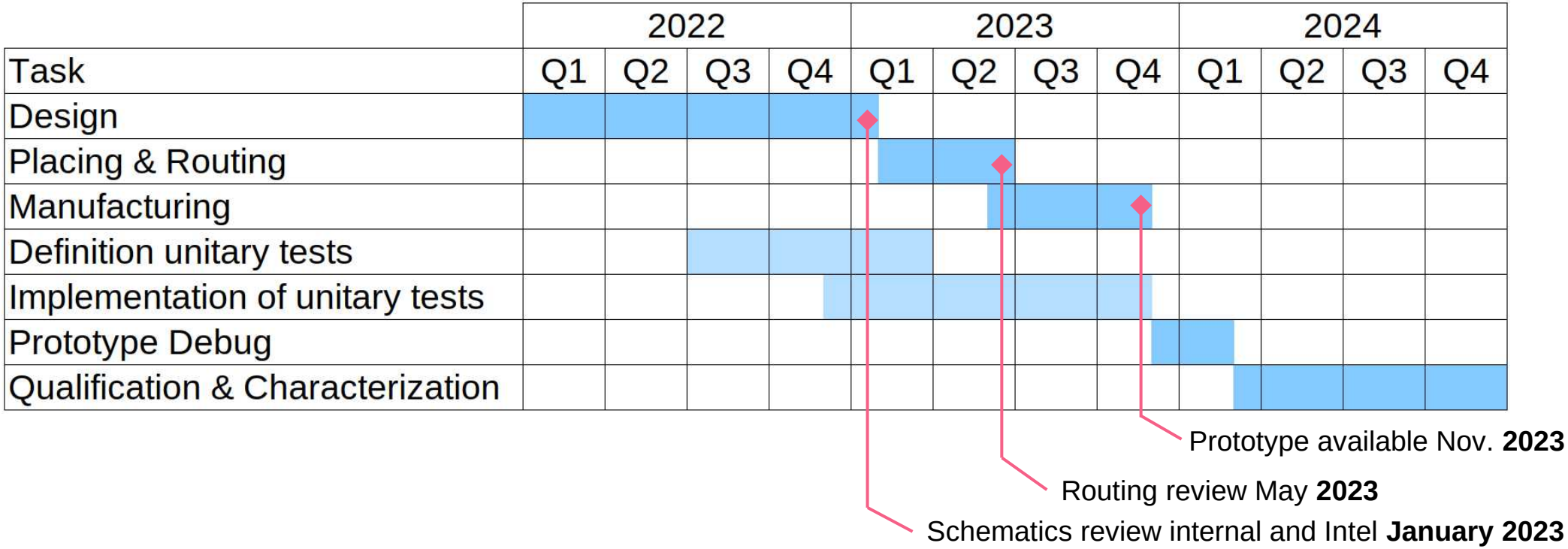
- Implement abstraction layers : ‘Low Level Interface’ to help developers focus only on their core skills
- Ensure test-ability for reliable production and long term support

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



Planning



Current phase : Placement and routing

- Draft PCB stack-up to refine with manufacturer
- Early simulation for high speed serial link signal integrity to define routing topology
- Mechanical constraints specified
- Power integrity simulations planned to ensure limited voltage drop by design and distributed via current

Synthesis

PCIe400 : On-going development

- Evolution of PCIe40 to accommodate with higher bandwidth and tighter timing requirement
- Target integration on few sub-detectors for LS3 on LHCb and maybe others
- Good inputs from sub-detectors teams is essential to review the specifications, understand the quantities and prepare the infrastructure
- Technical design review with CERN planned for Summer 2023
- LHCb-internal review planned for Q1 2024

PCIe400 is a stepping stone to prepare for future generic readout board

- PCIe form factor fits well with foreseen back-end architecture based on heterogeneous COTS
- Ethernet is another promising topology for interconnecting with COTS on a scalable system
- Experimental features to mature choice on future DAQ system architecture

Beyond PCIe400

- New design required targeting LS4 with adequate technology available at the time
- Aim at doubling at least the output bandwidth to pursue connectivity with data center standards
- Recent kick-off meeting chaired by CERN to discuss a future generic readout board.
  LHCb was represented by Online, Electronics Coordinator and CPPM

Thank you for your attention !