PCIe400 : Development status

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

Outline

Context and organization

PCIe400 : general characteristics

Technical developments

Planning

Synthesis

Context

Goals and rationale

- Gateway between GBT/lpGBT protocol and standard commercial protocols used in data centers: Ensure a cost effective solution
- Multiplex data from 48 → 1 very high bandwidth serial interface
- Event building first stage by taking advantage of power efficient processing resources available
- Generic readout DAQ for multi-context use (Alice, Belle II and CTA)

LHCb Upgrade II (2035)[1]

- New gen DAQ desired on LS3 (2026/2028) and LS4 (2033/2034)
- Data bandwidth requires probably a gain up to x10 compared to Upgrade I
  - Not feasible with current technology
  - Intermediate step : output bandwidth x4, higher processing capacity (~x8)

IN2P3 R&T

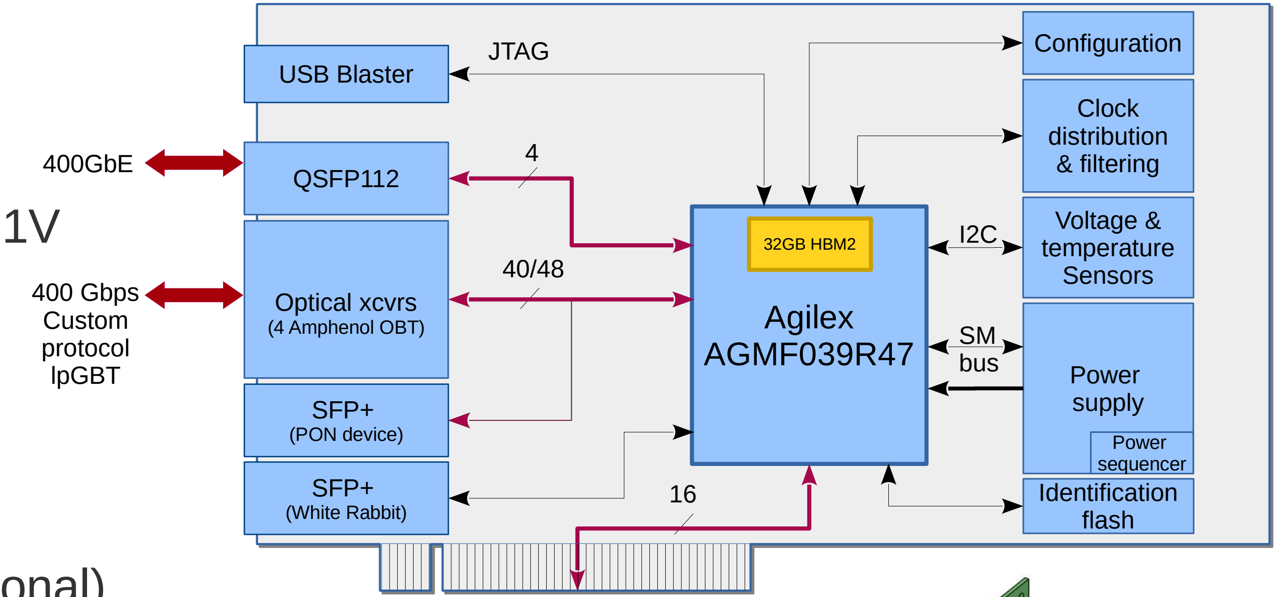
- Project set up to develop the prototype of PCIe40 next generation
- Funded for 3 years from 2022 to end of 2024

[1] LHCb Framework TDR, chapter 5, https://cds.cern.ch/record/2776420/files/LHCB-TDR-023.pdf?version=3

PCIe400

Foreseen characteristics

- Agilex M-series AGMF039R47A1E1V
- No DDR memory
  - Use of PC RAM or HBM2e instead
- Up to 48x26Gbps NRZ for FE
- PCIe Gen 5 / CXL or 400GbE (optional)
- High precision PLL < 10 ps RMS
- White Rabbit clock distribution
- TTC-PON interface for fast control



400GbE

400 Gbps
Custom
protocol
lpGBT

USB Blaster

JTAG

QSFP112

4

Optical xcvrs
(4 Amphenol OBT)

40/48

SFP+
(PON device)

SFP+
(White Rabbit)

16

32GB HBM2

Agilex
AGMF039R47

I2C

SM
bus

Configuration

Clock
distribution
& filtering

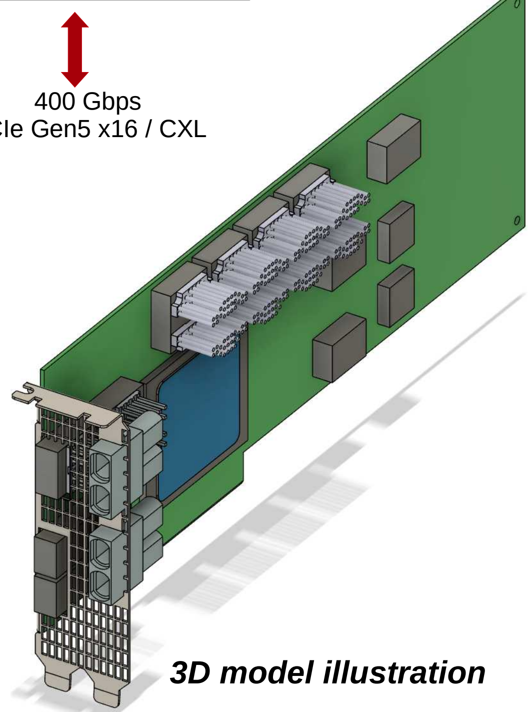
Voltage &
temperature
Sensors

Power
supply

Power
sequencer

Identification
flash

400 Gbps
PCIe Gen5 x16 / CXL



3D model illustration

FPGA resource comparison

Early Access program granted by Intel

- Weekly followup meeting with Intel engineers

New feature

- HBM / NoC (Network on Chip)
  - Facilitate high-bandwidth data movement between core logic and HBM
    - Deep learning acceleration
    - smartNIC to accelerate and offload certain functions from the server

Foreseen gain (compared to PCIe40)

- Processing : factor 8 to 12

	PCIe40	PCIe400
Family	Arria 10	Agilex M-series
Logic elements	1.2 M	3.9 M
DSP	1.5 K	12 K
Frequency (silicon max)	650MHz	1GHz
HBM2e	-	32GB
Hard co-processor	-	Arm Cortex A53 MPCore
Package	2000 pins	4500 pins

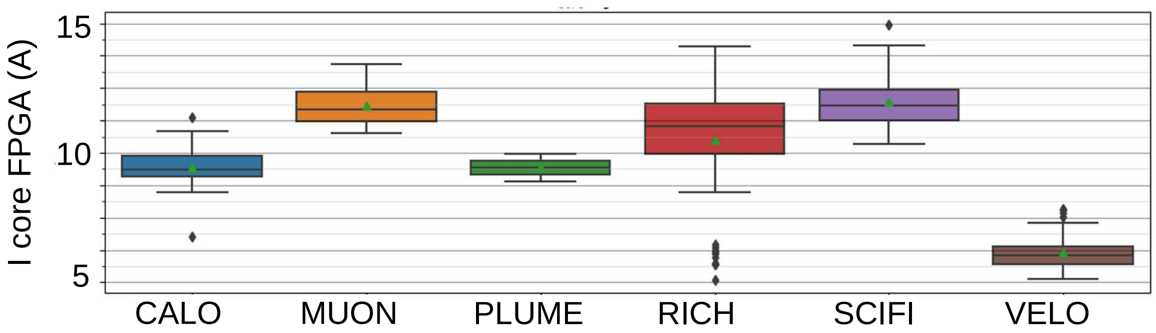
Technical developments

Toggle rate estimation

Estimation method

- Power model for Agilex M-series not available before Q4 2022
  - Post fit simulation of firmware with dummy logic and adjustable toggle rate
  - Extrapolation from Agilex I-series on Quartus Power and Thermal Calculator

- Definitive firmware not available for hardware design
  - Use of deployed PCIe40 for LHCb to measure core FPGA current and die temperature
  - Use of Quartus Power analyzer flow to retrofit a mean toggle rate estimation



FPGA Core current (A) PCIe40 for different LHCb subdetector flavors

Toggle rate is typically <12.5%

- 12.5% corresponds to a typical toggle rate (max 15%) considered by Intel
- Simplify the FPGA decoupling scheme as Intel recommendation can be applied with security margin

Power dissipation

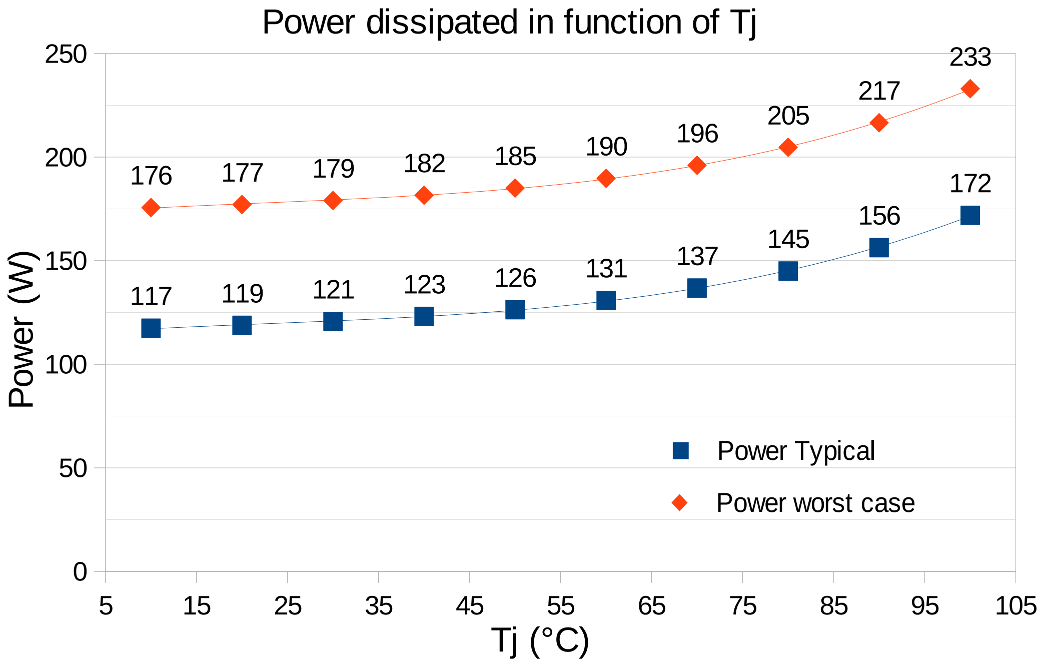
Estimation accross whole FPGA

- Preliminary Agilex M-series power model available from Q4 2022
  - 12.5% toggle rate (from previous study)
  - 640MHz (x2 to previous generation)
  - Use of Quartus Power and Thermal Calculator
  - Use of PCIe40 firmware deployed for LHCb to evaluate resource occupation
- Definition of a typical and worst case
  - Typical case:
    - 60 % logic
    - 80% RAM
    - 48 links at lpGBT speed
  - Worst case:
    - 80% logic
    - 100% RAM
    - 40 links 25G + 400GbE
  - Missing HBM use evaluation (WIP)

First estimation give 120 W to > 200 W

- Preliminary models are pessimistic
- Static power increase with T° junction (Tj)
- Dynamic power increase with frequency and toggle rate

Require particular attention on : power integrity and cooling



Power integrity

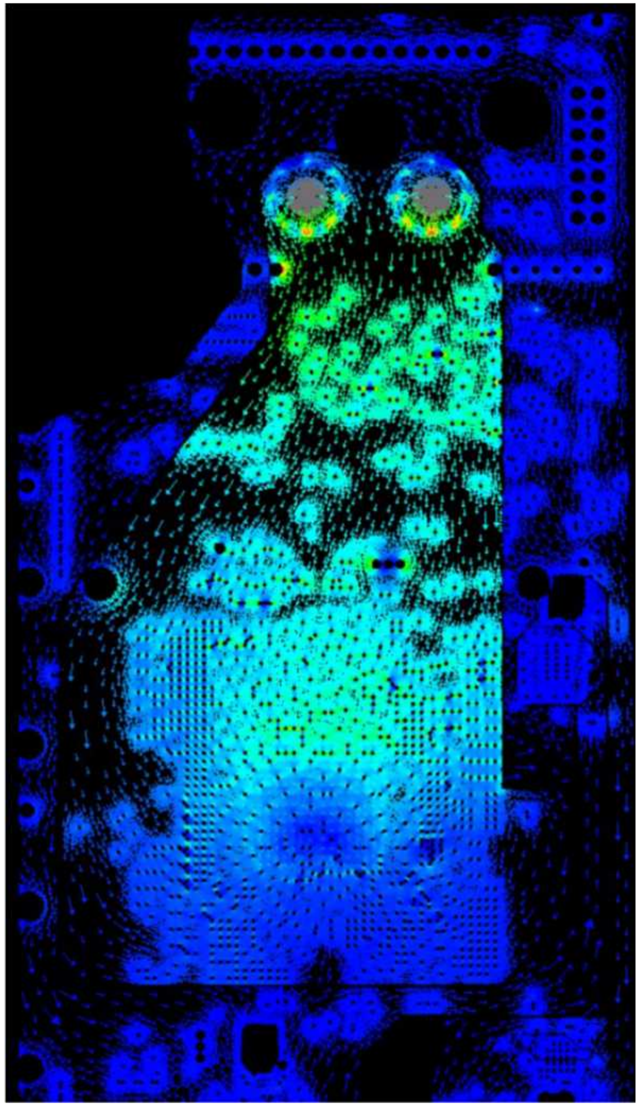
Power rails

- Between 55 and 100 A I_{core} (worst case)
- > 20 power rails from 0.8 V to 1.8 V
  - Careful power plane design
  - Caution on vias current

Cadence PowerDC / OptimizePI simulations

- Static current analysis
  - Voltage drop and current bottleneck check
- Dynamic current analysis
  - Decoupling capacitor placement check

Planned during routing phase



Current flow simulation illustration

Cooling consideration

Technological study

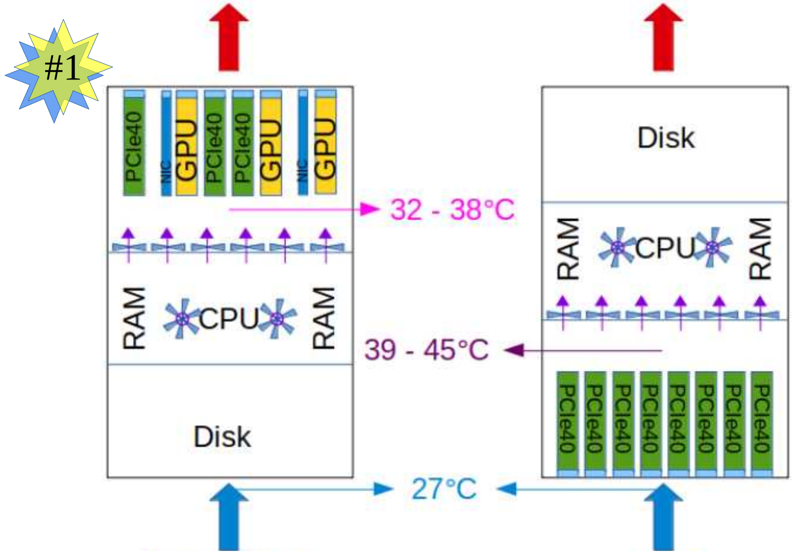
- 2 Airflow architecture identified
- Few specification on PCIe SIG
  - ambient T° between 20°C and 60°C

Heatsink types

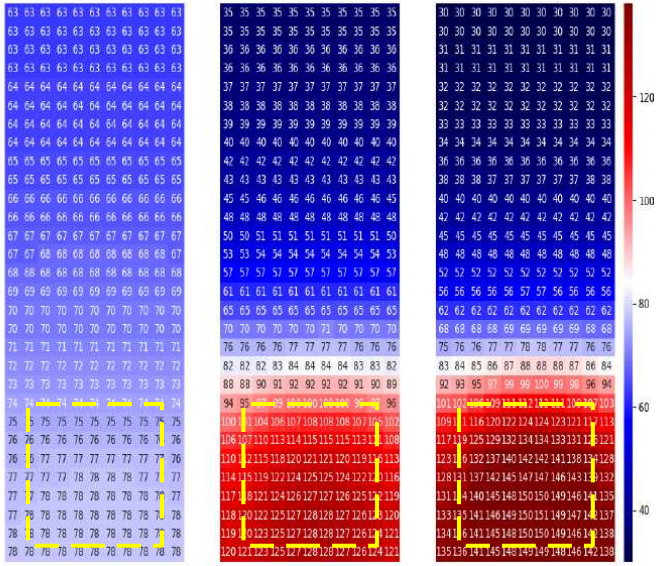
- Comparison with Heatscape models
  - Aluminum base + zipper fins
  - Copper base + *zipper fins*
  - Vapor chamber + *zipper fins*
- Vapor chambers has lower heat spread resistance making fins more efficient

Approximate thumb rule

- < 200 W: vapor chamber
- > 200 W: liquid cooling (infrastructure challenge)



Server architecture (top view)



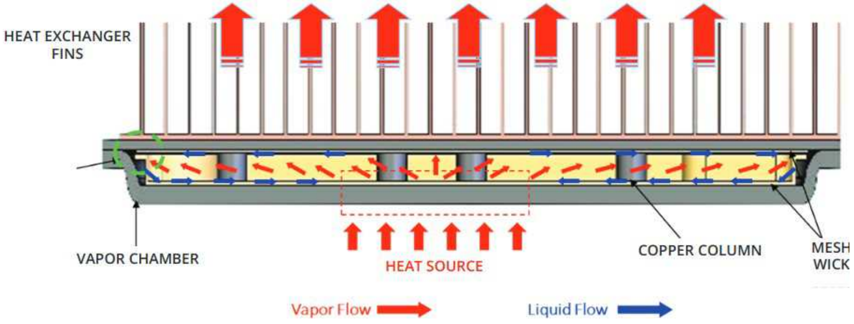
1. Vapor chamber + copper fin

2. Copper base + copper fin

3. Aluminum base + aluminum fin

Heat spread on heatsink base comparsion

Zipper fin



vapor chamber principle

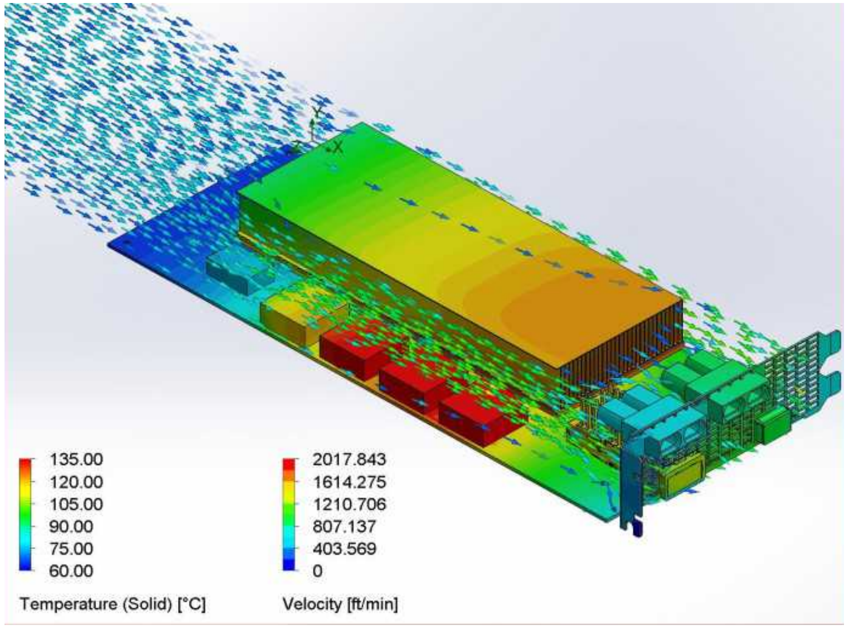
Cooling design

Computational Fluid Dynamics (CFD) simulations

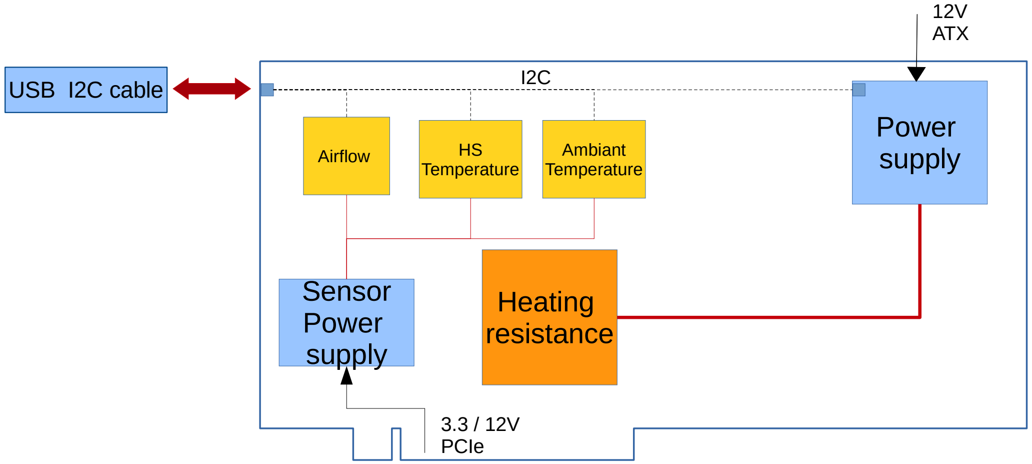
- Vapor chamber model construction under COMSOL
  - Rapid geometry variation
  - Optimization of fins height and width
  - Rapid airflow variation

Therma400

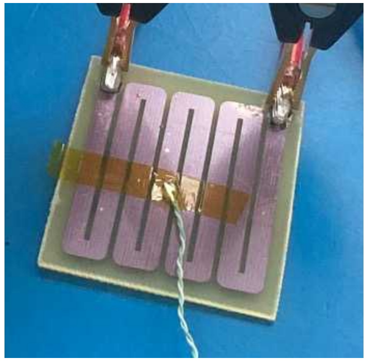
- Instrumented mockup board with temperature and anemometers
- Designed for CFD model verification and Prototype heatsink tryout
- FPGA emulation with a heating resistor
- Cabling is undergoing



CFD example illustration



Therma400 synoptic



Prototype heating resistor

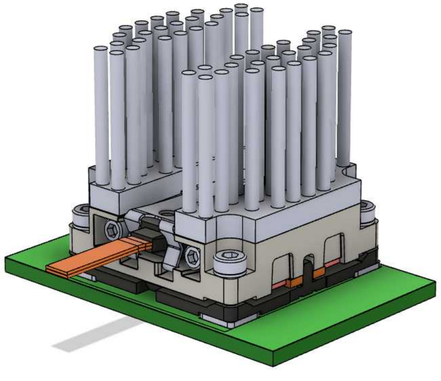
Optical interface

Retained solution with CERN consultation

- 4x Amphenol OBT
  - 12 duplex channels
  - 1.25G to 26.3G
  - 100m OM4, 850nm
  - MPO x24

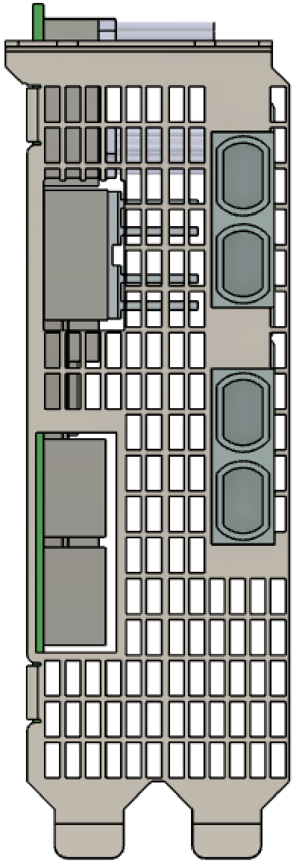
- 2x SFP+ for 10GPON / White Rabbit
  - TTC-PON OLT/ONU for fast control
  - White rabbit for clock distribution

- QSFP112
  - 4x112G PAM4
  - Direct Attach Cable <3m or opto <100m



Amphenol OBT
1.25G à 26.3G NRZ

QSFP112
106.25Gb/s PAM4



PCIe400 front-view

	# FE links
No TFC/WR/400GbE	48
WR	47
TTC OLT + ONU	46
TTC OLT + ONU + 400GbE	38

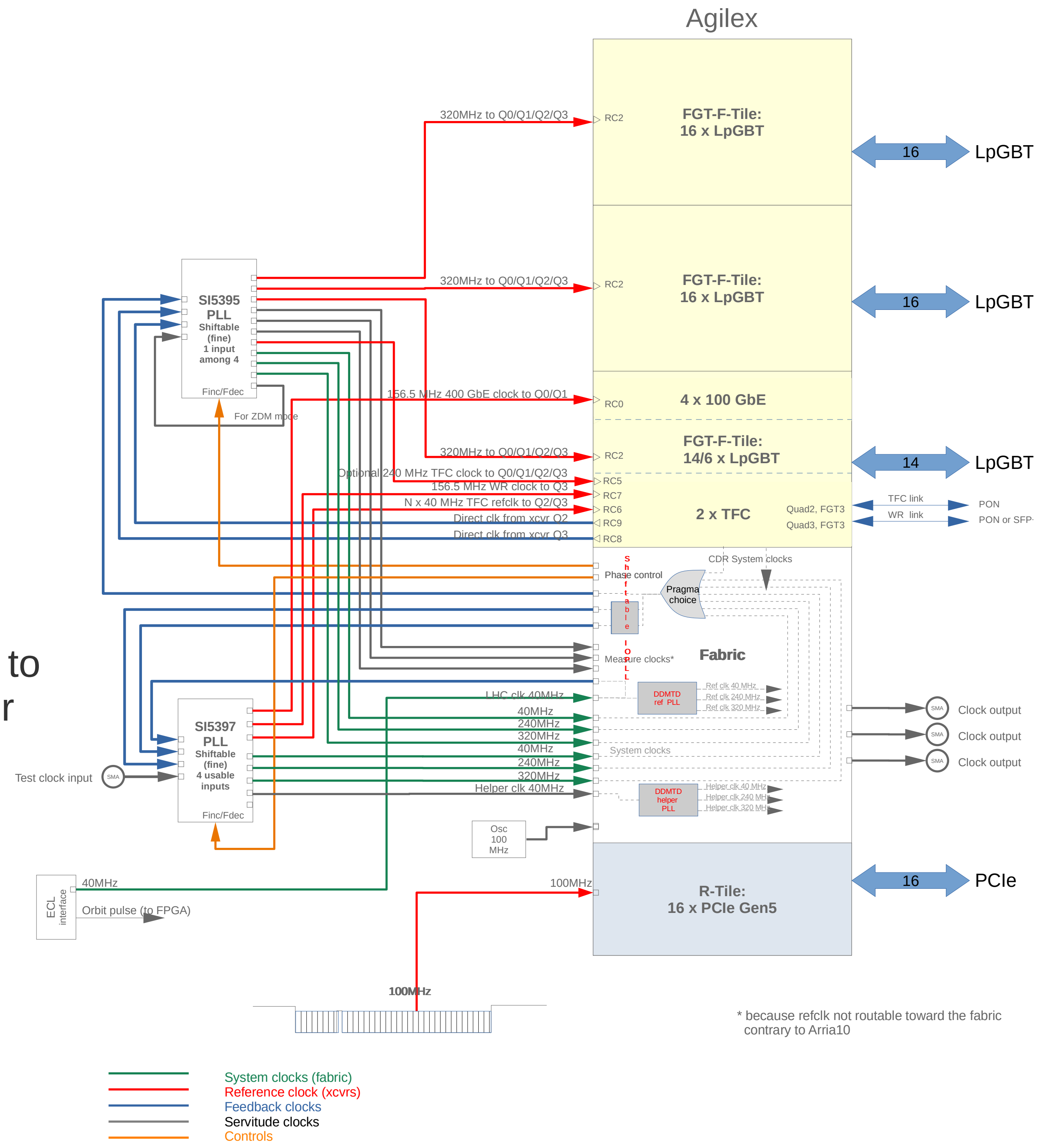
Clock tree

Simplification of PCIe40 clock tree

- 2 external PLL
- No clock buffer or crystal required
  - LHC frequency generated by PLL

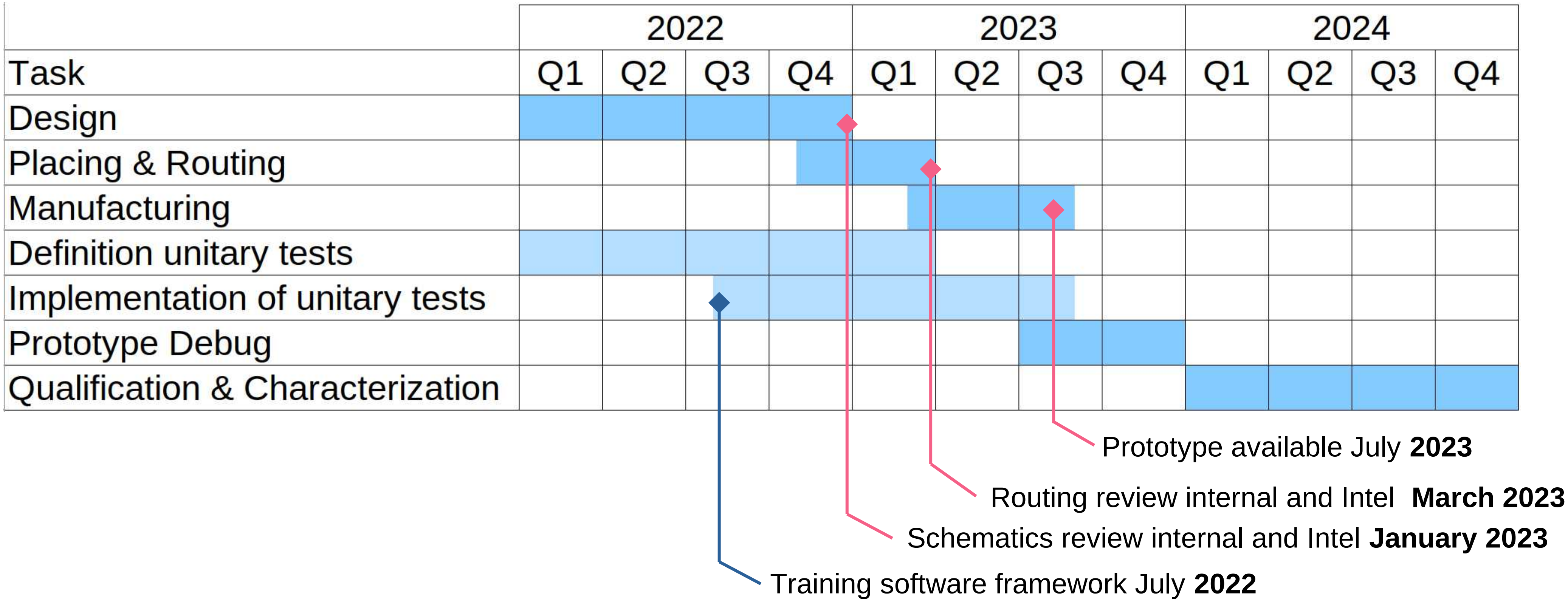
New features

- < 100 fs jitter RMS
- Recovered clock feed external PLL to Clean reference clock to transceiver
- Several schemes for phase control
  - Internal to PLL
  - Through DDMTD + internal PLL
  - Through DDMTD + external PLL



* because refclk not routable toward the fabric contrary to Arria10

Planning



	2022				2023				2024			
Task	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4
Design	🟦	🟦	🟦	🟦								
Placing & Routing				🟦	🟦							
Manufacturing					🟦	🟦	🟦					
Definition unitary tests	🟦	🟦	🟦	🟦	🟦							
Implementation of unitary tests			🟦	🟦	🟦	🟦	🟦					
Prototype Debug							🟦	🟦				
Qualification & Characterization									🟦	🟦	🟦	🟦

Prototype available July **2023**

Routing review internal and Intel **March 2023**

Schematics review internal and Intel **January 2023**

Training software framework July **2022**

Schematics almost done

2 prototypes expected for hardware debug

Synthesis

Hardware

- Schematics should be finalized by the end of the year
- Component procurement anticipated
- Several studies undergoing for cooling solution
- 2 prototypes are expected by summer 2023

Firmware and software

- Development starting now for Low Level Interface (LLI)
- Test firmware should be available as soon as prototypes are available
- Virtual mockups are used for development (Eval. cards + USB/I2C bridge)

Generic board with several application possible

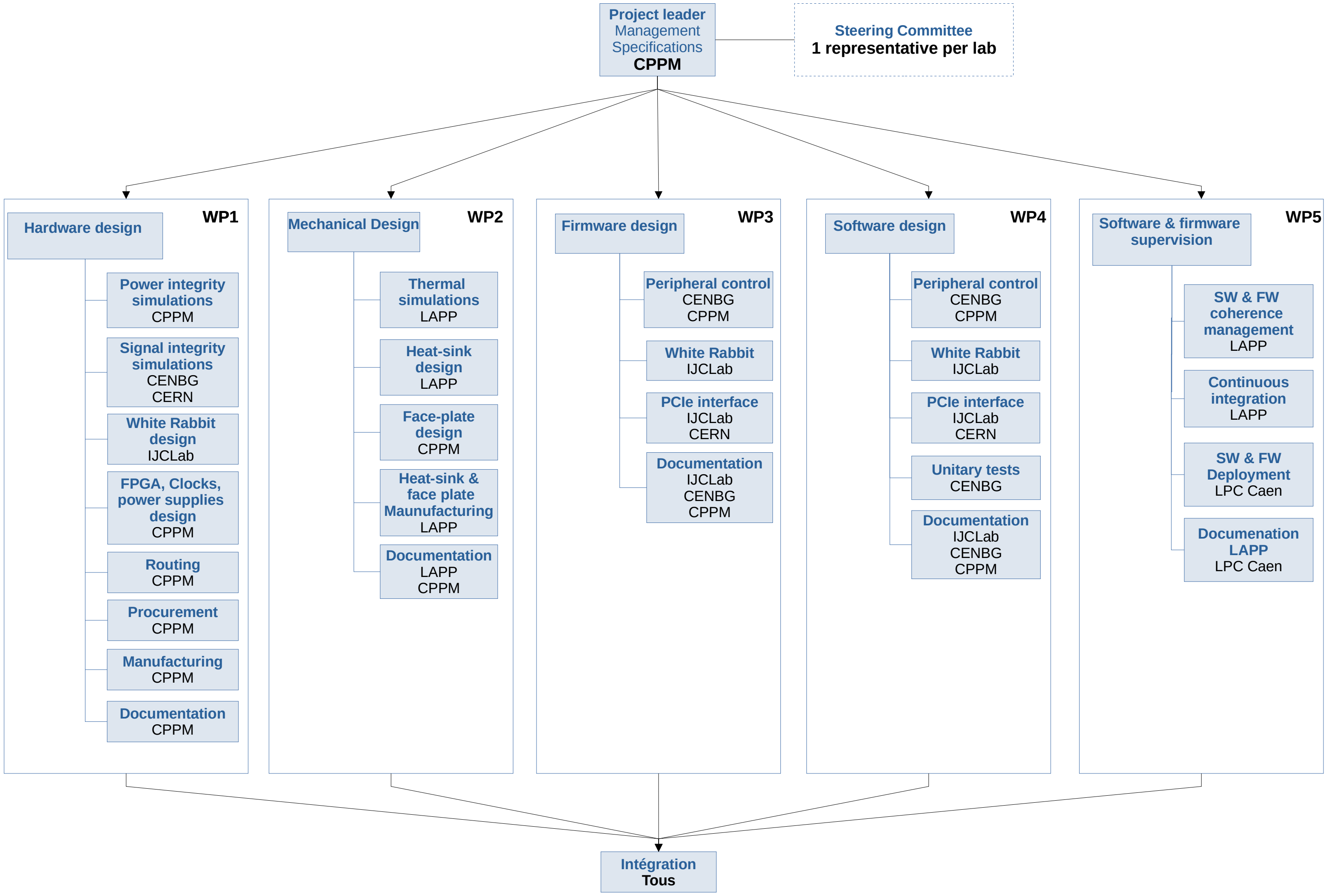
- New FPGA features allows to explore new functionalities
  - Integrated network interface (smartNIC)
  - Interconnect network with 400GbE
  - Serial links at 26Gbps and beyond
  - Hard co-processor for fast monitoring
  - ...

Target deployment of PCIe400 is during LS3 for new detectors

Backup

Task organization

- Unite the workforce of 5 labs from IN2P3 as well as LHCb online team @CERN
- Overall resource are ~4.5 FTE per year
  - 2 full time engineers + Jean-Pierre Cachemiche (retirement) at CPPM
  - ~2.5FTE distributed over 10 people



Project leader
Management
Specifications
CPPM

Steering Committee
1 representative per lab

WP1 – Hardware design
- Power integrity simulations – CPPM
- Signal integrity simulations – CENBG, CERN
- White Rabbit design – IJCLab
- FPGA, Clocks, power supplies design – CPPM
- Routing – CPPM
- Procurement – CPPM
- Manufacturing – CPPM
- Documentation – CPPM

WP2 – Mechanical Design
- Thermal simulations – LAPP
- Heat-sink design – LAPP
- Face-plate design – CPPM
- Heat-sink & face plate Maunufacturing – LAPP
- Documentation – LAPP, CPPM

WP3 – Firmware design
- Peripheral control – CENBG, CPPM
- White Rabbit – IJCLab
- PCIe interface – IJCLab, CERN
- Documentation – IJCLab, CENBG, CPPM

WP4 – Software design
- Peripheral control – CENBG, CPPM
- White Rabbit – IJCLab
- PCIe interface – IJCLab, CERN
- Unitary tests – CENBG
- Documentation – IJCLab, CENBG, CPPM

WP5 – Software & firmware supervision
- SW & FW coherence management – LAPP
- Continuous integration – LAPP
- SW & FW Deployment – LPC Caen
- Documenation LAPP – LPC Caen

Intégration
Tous

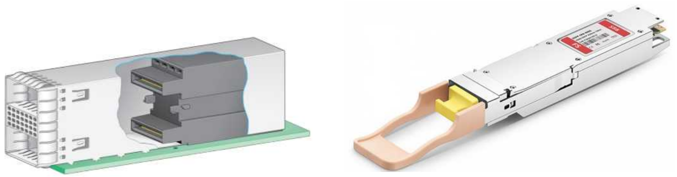
Optical interface

Several solution compared

- Anticipated CERN serializer development
  - Transceiver « On-Edge » (QSFP-DD)
    - Clogged I/O braket for airflow
    - Compatibility for custom CERN protocols
  - Firefly Samtec
    - Low channel density (simplex modules)
  - After consulting CERN, retained choice Finisar/Coherent BOA
  - Finally replaced by OBT Amphenol
    - Compatible with Finisar
    - Better technical support
    - Cooling solution
    - Socket reference available

Standard QSFP112 for 400GbE

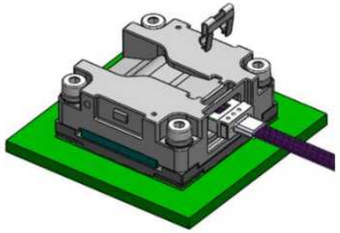
- MSA task force groups several FF
  - QSFP-DD/QSFP-DD800/QSFP112
- QSFP112 introduced in 2021
  - Proof of concept
  - Cages and Direct Attach Cable available



QSFP-DD
53.125Gb/s PAM4 (lower rate NRZ possible?)

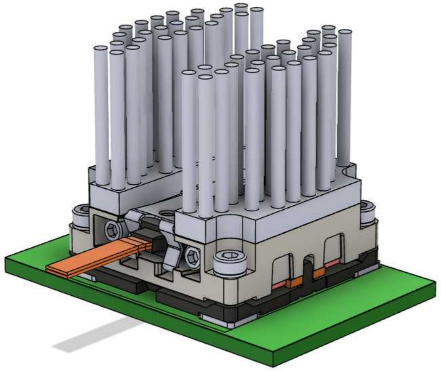
QSFP112
106.25Gb/s PAM4

Samtec FireFly ECUO
14 / 25 / 28 Gb/s NRZ



Coherent / Finisar BOA
1 à 28.1Gb/s NRZ

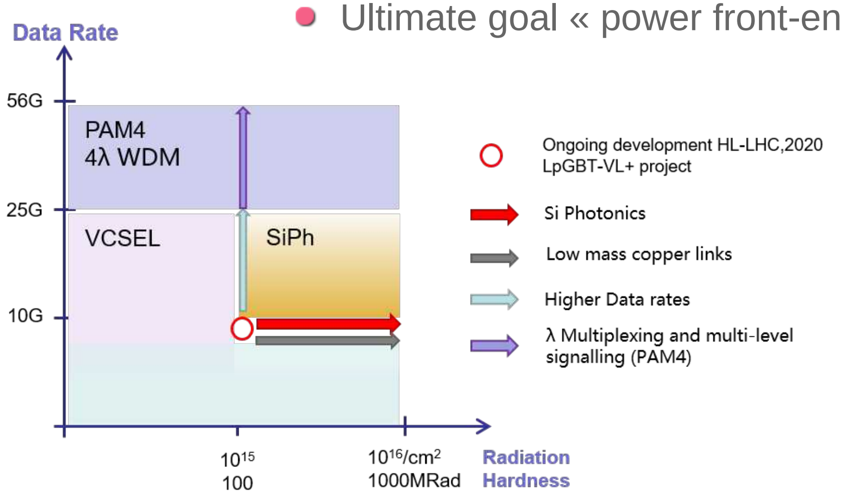
Amphenol OBT
12 duplex
1.25G à 26.3G NRZ
~6W 3.3V



R&D CERN serializer

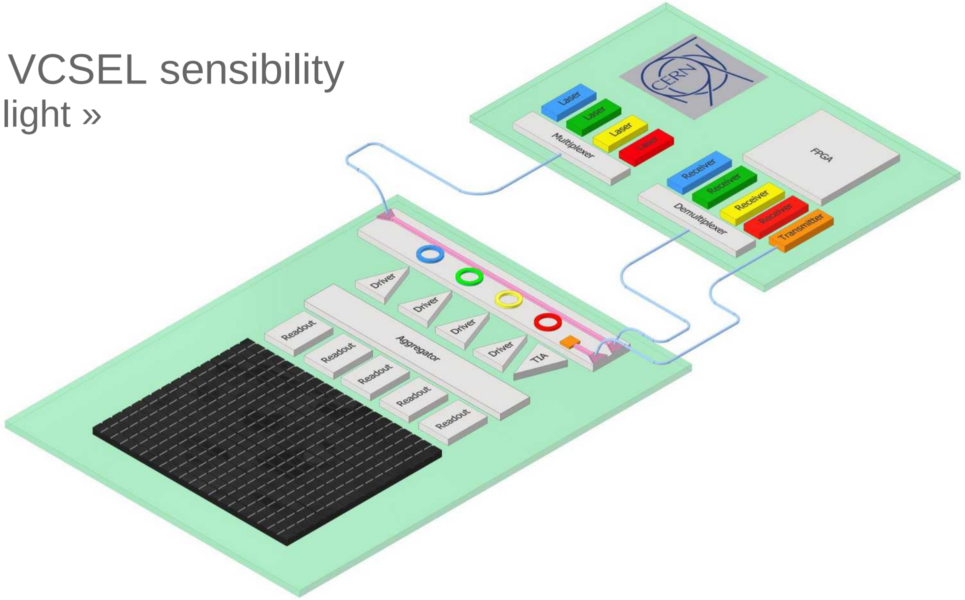
WP6 of EP-R&D CERN

- Goals and rationale
  - Higher data transfer rate of serial links
    - Considered solution : Amplitude modulation (PAM4), Wavelength multiplexing (WDM)
  - Higher radiation tolerance for front-end
    - Silicon photonics + WDM because of VCSEL sensibility
    - Ultimate goal « power front-end with laser light »



Serial link roadmap

P. Moreira & al. https://cds.cern.ch/record/2649646/files/CERN-OPEN-2018-006.pdf



System view

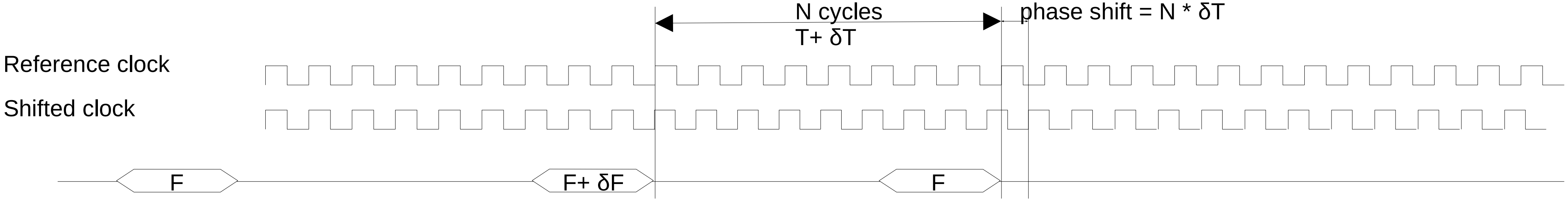
Current status

- Technological choices for ASIC DART28 (transceiver/driver SiPh)
  - 25.65Gbps NRZ (multiple LHC bunch Clock)
  - Equivalent FEC to lpGBT
  - 28nm CMOS
    - Prototyping on FPGA
- Probable compatibility with PCIe400 ± WDM demux

Phase control

Precise phase adjustment

- Frequency increment or decrement during a controlled amount of clock cycles

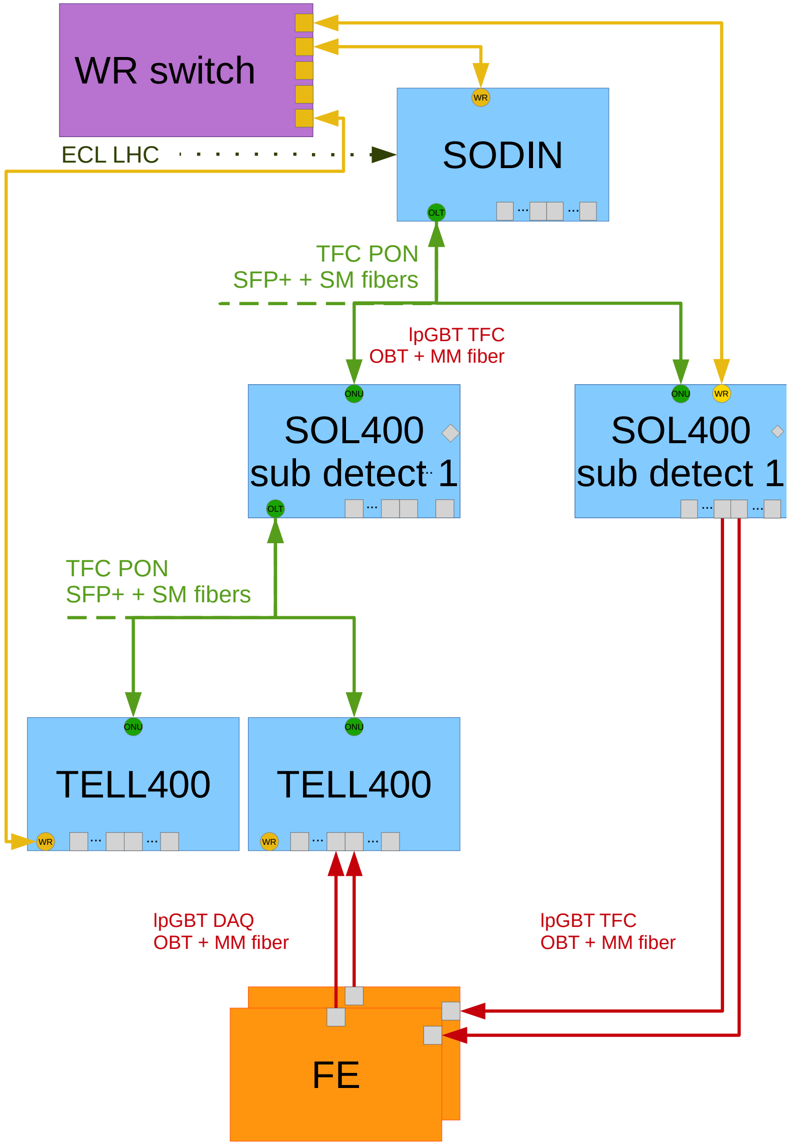


Exemple :
F = 100 MHz
δF = 0.010001MHz
δT = 1/100000000 – 1/100010001 = 1ps
Programming of 100.010001 MHz during 500 cycles → phase shift = 500 ps

PCIe400 : DAQ architecture example

PCIe400 can be used as

- Clock distribution SODIN
- Fast control SOL400
  - To mitigate the fact that only 2SFP+ links, more SOLL400 are required to benefit from WR clock distribution to FE boards
- Readout TELL400


WR switch
ECL LHC
SODIN
WR
OLT
TFC PON
SFP+ + SM fibers
lpGBT TFC
OBT + MM fiber
SOL400
sub detect 1
ONU
OLT
SOL400
sub detect 1
ONU
WR
TFC PON
SFP+ + SM fibers
TELL400
ONU
WR
TELL400
ONU
WR
lpGBT DAQ
OBT + MM fiber
lpGBT TFC
OBT + MM fiber
FE