NUCLÉAIRE & PARTICULES

The DAQ Upgrade II at LHCb

Renaud Le Gac (CPPM IN2P3/CNRS)

Conseil scientifique IN2P3 – 13 février 2025

Plan

- LHCb, LHCb Upgrade I and II
- Genesis of the DAQ project at IN2P3
  - Online system for the LHCb detector
  - Online evolution to LHCb Upgrade I
  - Planned evolution toward LHCb Upgrade II
- The DAQ project for LHCb Upgrade II at IN2P3
- Conclusions

LHCb, LHCb Upgrade I and II

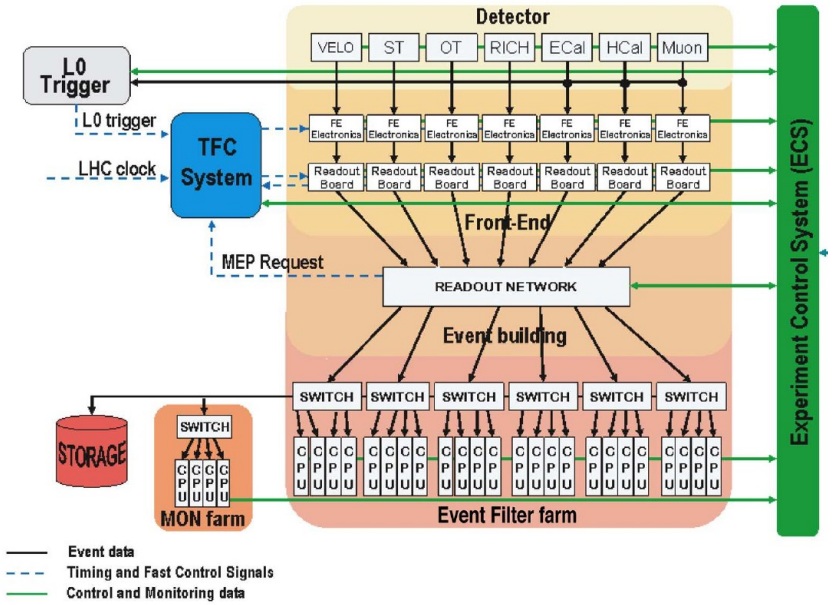
	Date	Runs	Instantané $\mathcal{L}$ [$cm^{-2}\ s^{-1}$]	Intégré $\mathcal{L}$ [fb^{-1}]	*pp* interactions per collision
LHCb	2010 – 2018	Run 1 & 2	4×10^{32}	9	1
LHCb Upgrade I	2022 – 2033	Run 3 & 4	2×10^{33}	50	5
LHCb Upgrade II	2036 – 2041	Run 5	1.5×10^{34}	300	40

Online system for LHCb (2010-2018)

Online overview for the LHCb detector

- The online system consist of three main components:
  - Data Acquisition (DAQ)
  - Timing and Fast Control (TFC)
  - Experiment Control System (ECS)

  Upstream of the DAQ, the Level-0 trigger reduces the collision rate from 40 MHz down to 1 MHz



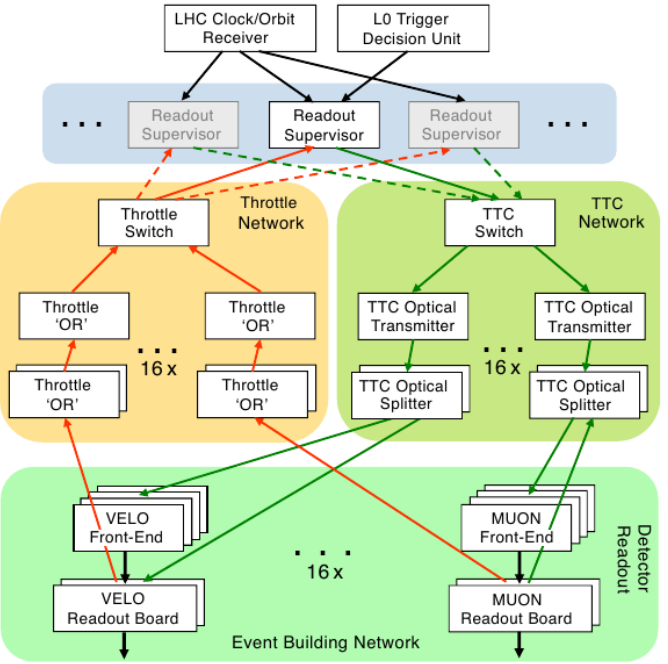
- The DAQ is connected to the front-end electronics via 6000 optical links using the GOL protocol at 1.6 Gbits/s and analogue transmission for the VELO
  - Event building is performed using a common readout card (TELL1) and a Gigabit Ethernet network.
  - For each triggered bunch-crossings, the event builder collects all the fragments, assembles them and distributes the event to the HLT farm.

- The bandwidth of the DAQ is 0.45 Tbits/s

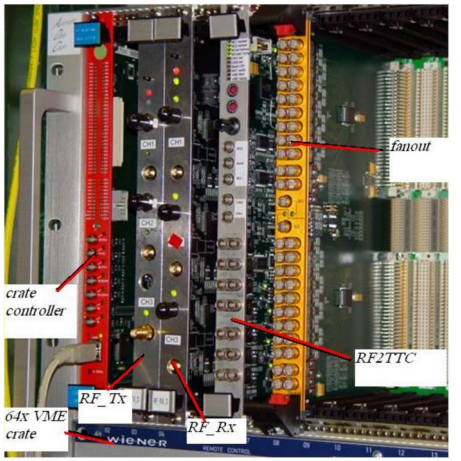
TFC overview for the LHCb detector

- The TFC system drives all stages of the data readout of the LHCb detector between the front-end electronics and online processing.

The TFC distributes the beam-synchronous clock, the L0 trigger, synchronous resets and fast control commands.



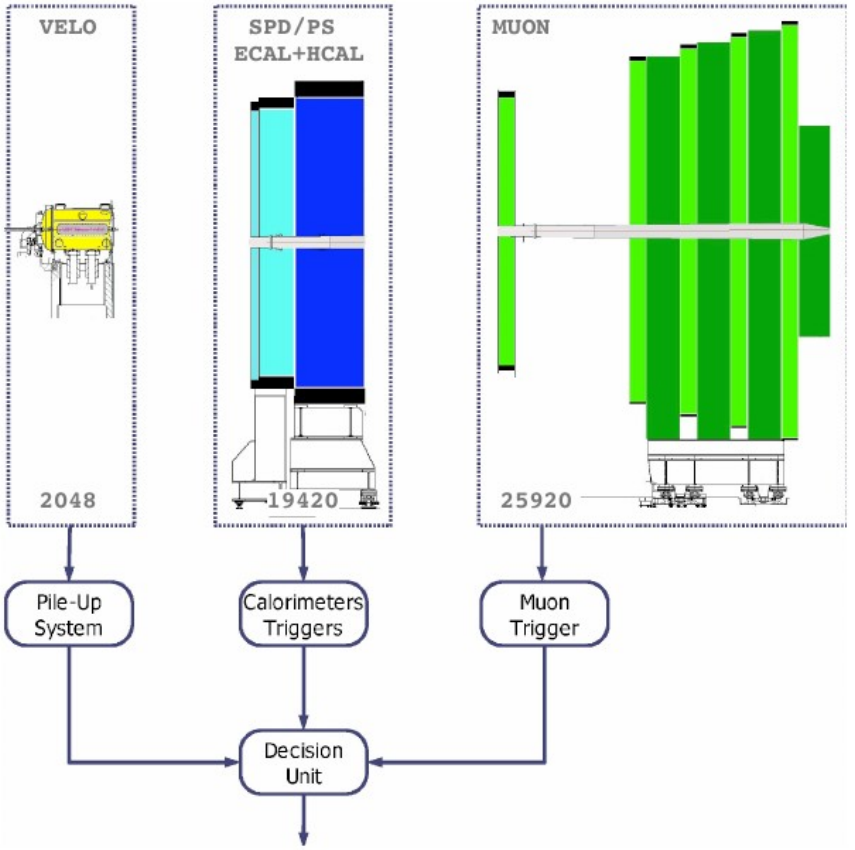
- The TFC relies entirely on CERN custom VME cards (TTCex, TTCtx) and ship sets (TTCrx, QPLL) which disctribute the LHC clock wiht a jitter of 10 ps and a phase determinism of 100 ps



The Level-0 trigger of the LHCb detector

- The Level-0 trigger (L0) reduces the LHC bunch crossing rate of 40 MHz to the rate of 1 MHz with which the entire detector can be read out.

- Due to their large mass, the signature of a *b*-decay is a displaced vertex with particles of large transverse momentum (pT) or energy (ET).
- Threfore, the Level-0 trigger attempts to reconstruct:
  - the highest ET hadron, electron and photon clusters in the calorimeters,
  - the two highest pT muons in the muon chambers.
- A Level-0 Decision Unit collects all the information and derives the final Level-0 trigger decision for each bunch crossing.

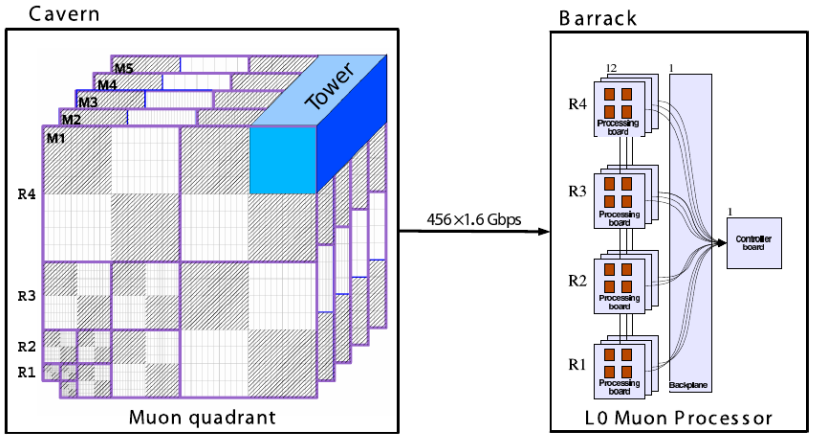


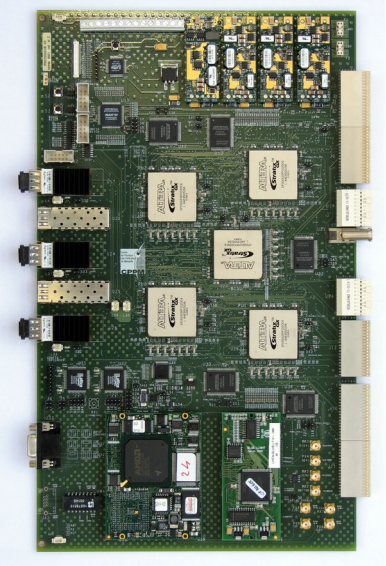
Focus on the Level-0 Muon custom processor

CPPM

- The L0 muon processor reconstructs track segments in the muon detector and measures their transverse momentum

Level-0 Muon	
FPGA:	Stratix GX
Optical I/O:	1864 @ 1.6 Gbit/s
Bandwidth:	2.7 Tbits/s
Processing time:	1.2 µs
Form Factor:	9U: 400x220 mm





48 Processing boards

in the cavern

IN2P3 contributions to the Level-0

- The IN2P3 Institutes were the main contributors to the Level-0 trigger

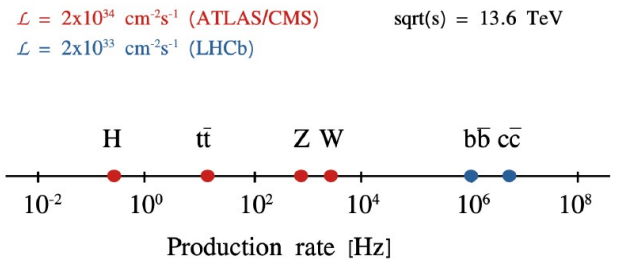
Component	Institute
L0 Pile up	NIKEF
L0 Calorimeter	IN2P3-LAL
L0 Muon	IN2P3-CPPM
L0 Decision Unit	IN2P3-LPC
Project Leader	*IN2P3-CPPM*

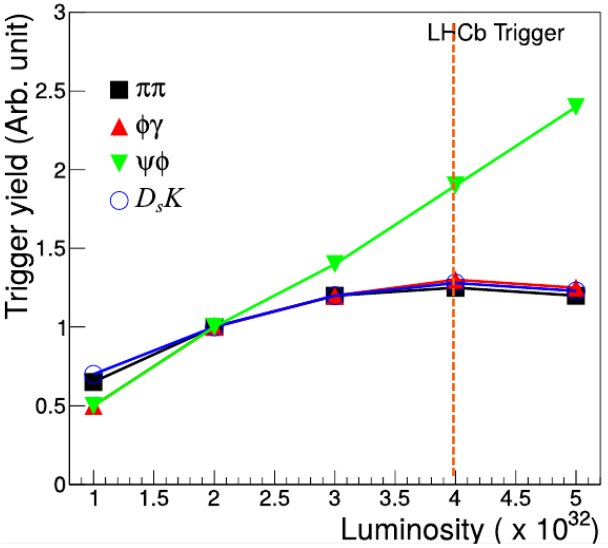
- With the removal of the Level-0 trigger in the *LHCb Upgrade I*, the CPPM migrated to the design, production and maintenance of the generic and multi-purpose readcout card called PCIe40

Online evolution to LHCb Upgrade I (2022 – 2033)

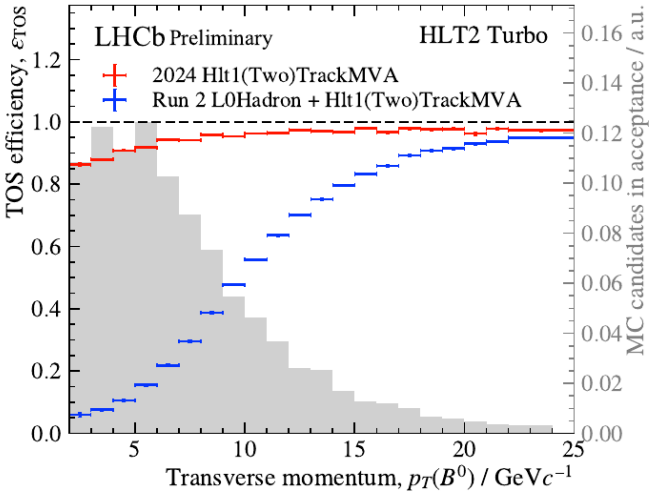
Increasing the instantaneous luminosity

- The Level-0 trigger rate staturates for hadronic final states at the luminosity of Run 1 and 2 ($4 \times 10^{32}\,cm^{-2}s^{-1}$)
- High rate of beauty and charm decay at $2 \times 10^{33}\,cm^{-2}s^{-1}$





- In order to efficiently select a large number of interesting *b* and *c*-decays, the optimal approach is to:
  - remove the level-0 trigger
  - move to a new paradigm where all bunch-crossings are assembled, reconstructed with ultimate precision and then event selected in near real time.
- Selection efficiency increases between 2 and 3



(b) TOS efficiencies in $B^0 \to D^- (K^+\pi^-\pi^-)\pi^+$.

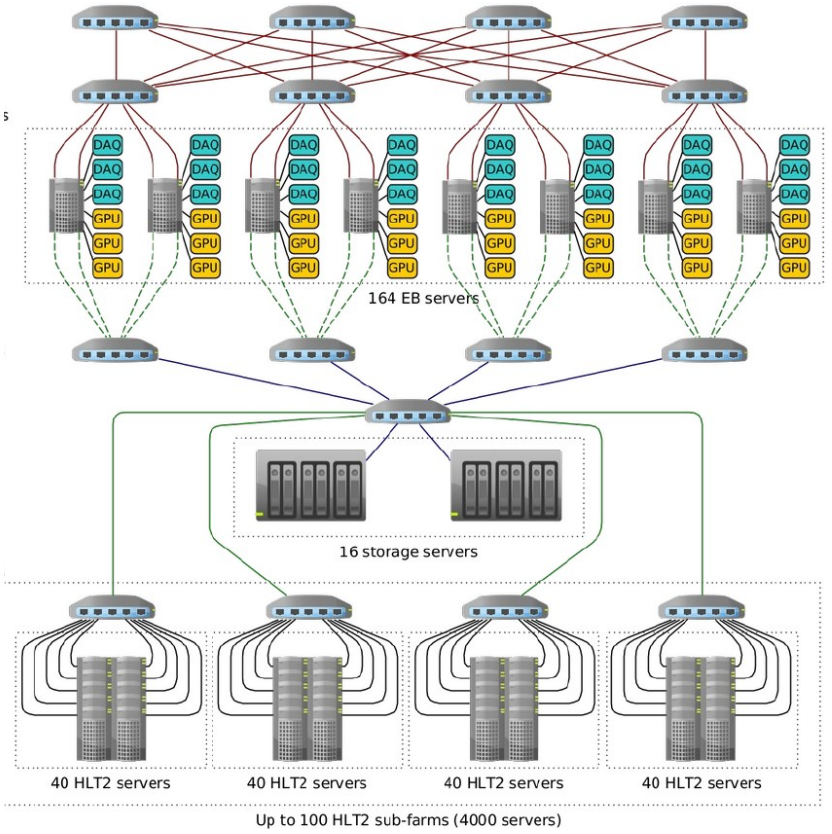
DAQ architecture for the LHCb Upgrade I – Run 3

- The DAQ uses standard protocol and COTS components as far as possilble :
  - PCi Express, Infiniband network
  - PC-server designed for up to 8 GPUs
  - DMA transfert

- The DAQ is connected to the front-end electronics via 11 000 optical links using the GBT protocol at 4.8 Gbits/s

- The first stage is a PC-server housing 3 readout cards, PCIe40 and 3 GPUs for HLT1

- Event building is performed for each bunch-crossing using a 200 Gigabit Infiniband network.



- A trigger-free architecture with 40 Tbits/s of total bandwidth

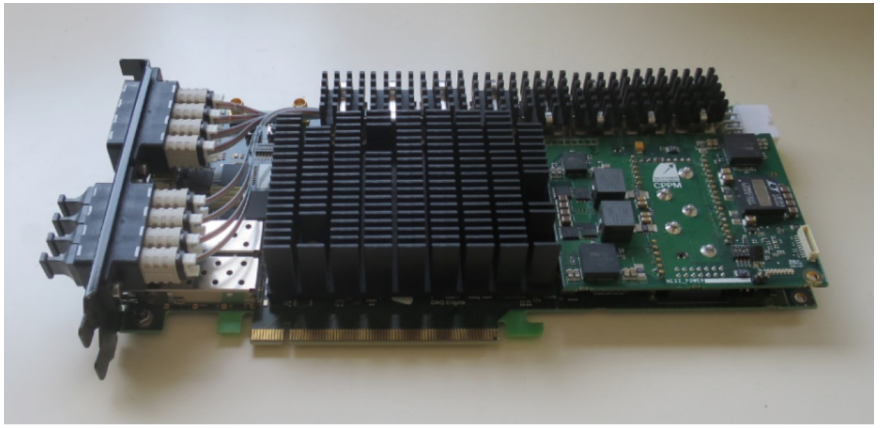
Generic and multi-purpose PCIe40 card

CPPM

- LHCb has chosen to use a single, custom-built card for data acquisition, TFC and slow control: the PCIe40 card

- 1250 cards were produced between June 2019 and May 2020 for the LHCb, ALICE, Mu3e and Belle II collaborations

- Built around the Arria10 FPGA, the different functionalities are achieved by loading a variety of gatewares onto the FPGA



PCIe40	
FPGA	
Family:	Arria 10
Logic Elements:	1 150 kLE
Max frequency:	644 MHz
RAM:	53 Mb
Optical I/O	
bidirectional links 10 Gbps:	up to 48 Gbps
TFC PON:	2
CPU I/O	
PCIe GEN3:	16
Output bandwidth:	100 Gbps
Form Factor	
PCIe:	241 x 111 mm

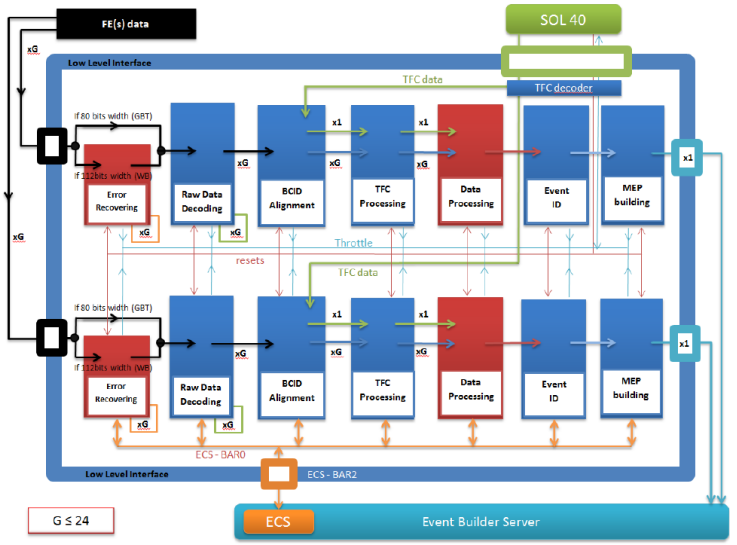
User gateware and continous integration

LAPP

- User gateware dedicated to triggerless DAQ faces significant challenges:
  - data frames arrive asynchronously over all optical links
  - data frames need to be decoded, realigned according to their BXID, assembled into an event packet. Event packets will then be sent to the DAQ network.
  - Aggregated data can also be manipulated into primitive such as local clusters.

- A framework has been introduced in which blocks common to all subdetectors (blue) are developed centrally. This allows subdetector developers to focus only on specific blocks (red) for their system.

- The online system required 21 gatewares. These are built, verified and versioned using the GitLab's Continuous integration tools

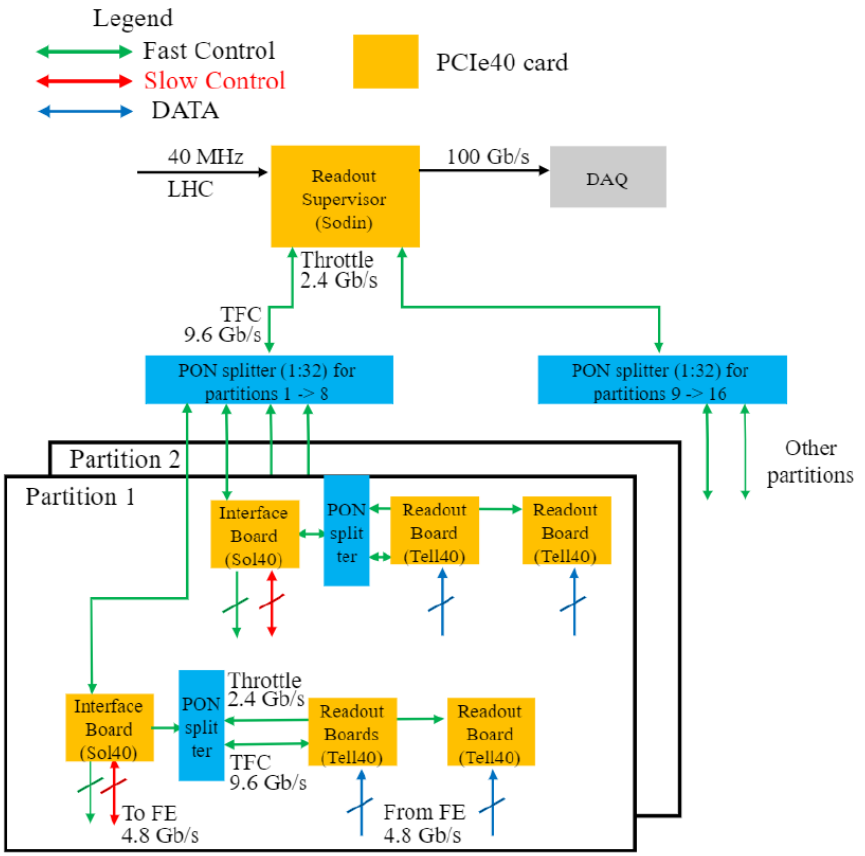


TFC architecture for the LHCb Upgrade I – Run 3

- The TFC relies entirely on the PCIe40 card. It is a complex architecture with multiple hubs

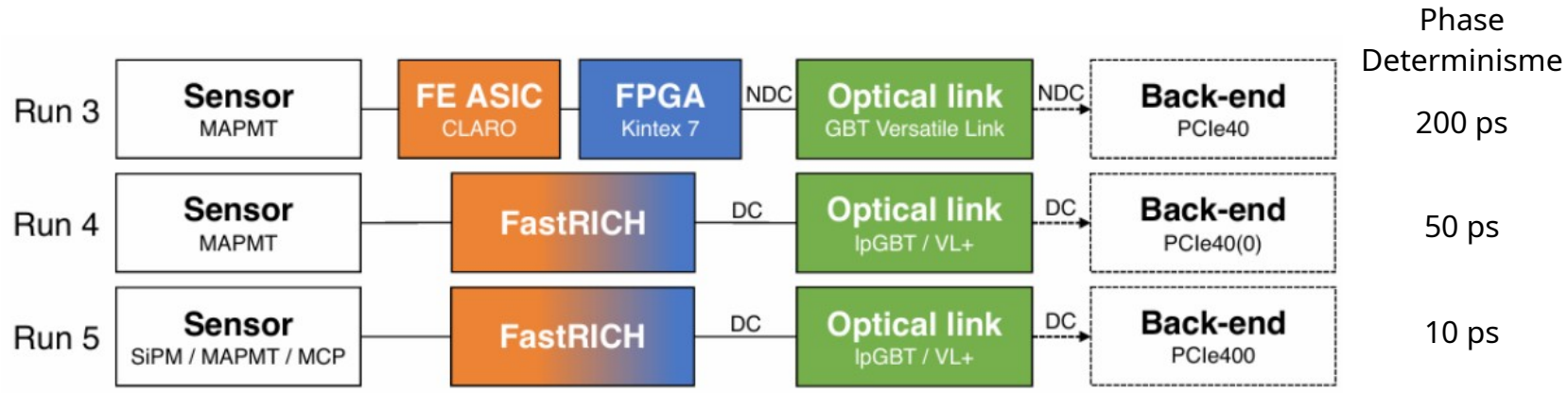
- The PCIe40's clock tree uses high-precision PLLs to ensure good jitter of the bunch clock recovered at the front-end electronics.

- The phase between the bunch clock recovered at the front electronics and the bunch crossing must remain the same after a reset or a power cycle.
  - This is known as *phase determinism*.
  - It is lost in FPGA transceivers.
  - A complex mechanism has been implemented in the FPGA Arria 10 to achieve an accuracy of 200 ps.
  - Mechanism depends on Vendor and FPGA familly



Planned evolution towards LHCb Upgrade II

DAQ relevant enhancements to LS3



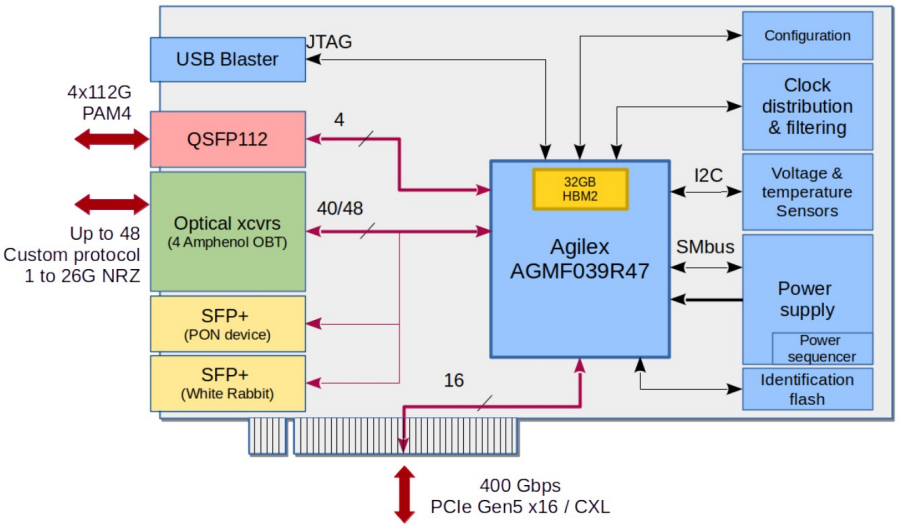
- For Run 4, the RICH detector will deploy a new front-end electronics during LS3 :
  - New FE ASIC with timing measurement capabilities $\mathcal{O}$(25 ps)
  - Optical link protocol lpGBT instead of GBT
  - Increase the output bandwidth by ~30 %
  - Phase determinism at the level of $\mathcal{O}$(50 ps) matching the MAPMT jitter of 150 ps

- The RTA will experiment with a custom DWT processor. It reconstructs trace segments in the SciFi using retina architecture that requires a network of powerful FPGAs.

Generic and multi-purpose card PCIe400

PCIe400	
FPGA	
Family:	Agilex 7M
Logic Elements:	3 851 kLE
Max frequency:	1 GHz
RAM:	311 Mb
HBM:	32 GB
ARM Cortex processor:	4 core
Optical I/O	
bidirectional links 25 Gbps:	up to 48
bidirectional links 112 Gbps:	up to 4
TFC PON or White Rabbit:	2
CPU I/O	
PCIe GEN5:	16
Output bandwidth:	400 Gbps
Timing	
high precision PLL with jitter < 100 fs	
White Rabbit node	
Form Factor	
PCIe:	268 x 111 mm

exploratory features

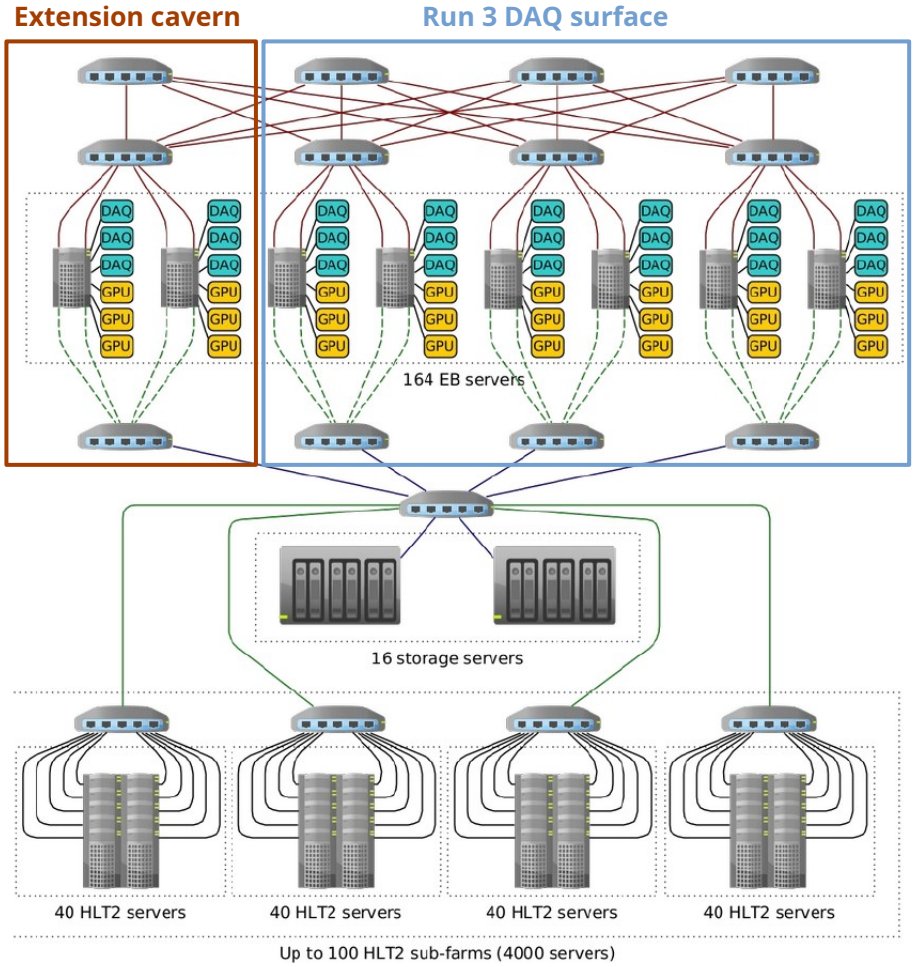


First prototype – January 2025

Intersested collaborations: Belle II, ...

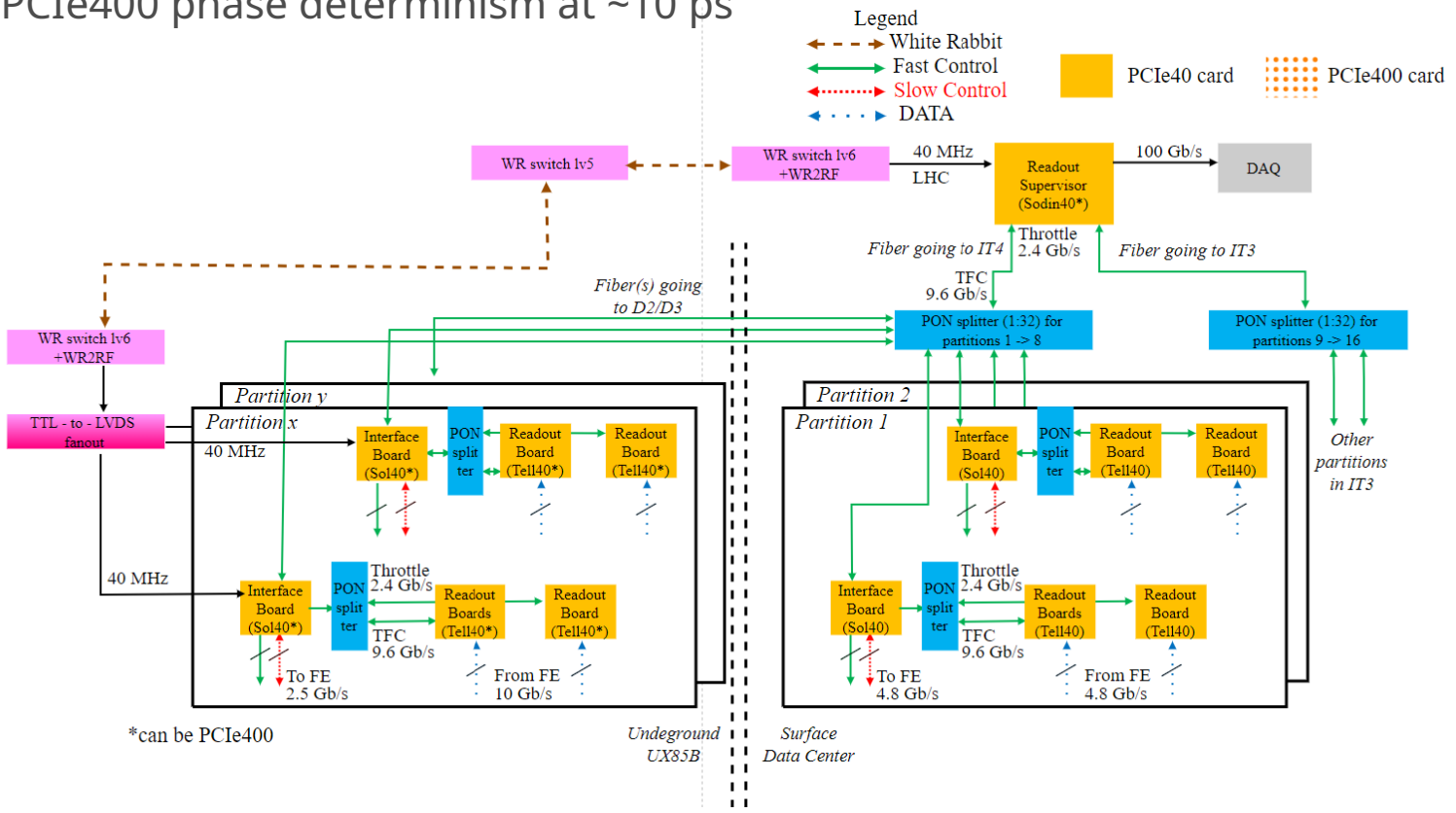
DAQ architecture for the LHCb Upgrade I – Run 4

- The architecture of the DAQ system remains unchanged with respect to the Run 3.
- However, due to the limited length of the lpGBT connections (≤ 150 m), the DAQ must be extended, in the cavern not at the surface.



TFC architecture for the LHCb Upgrade I – Run 4

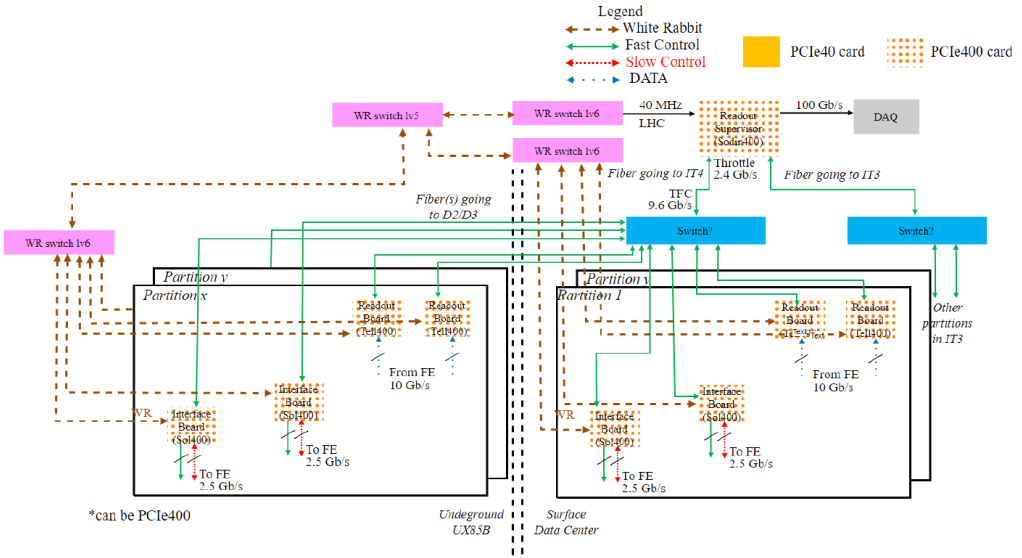
- The TFC architecture has to evolve:
  - New protocol to receive the bunch clock, White Rabbit (**)
  - Improved phase determinism from ~200 ps down to ~50 ps
  - Experience PCIe400 phase determinism at ~10 ps



(**) Expertise at IJCLab

Toward the Online for Upgrade II

- A trigger-free architecture with a total bandwidth of 200 Tbits/s is under discussion for Upgrade II



- Many unkowns:
  - evolution of the computing market, form factor of GPU, PC-server, etc
  - roadmap for the future generation of FPGA,
  - processing power required to prepare the sub-detector's raw data for HLT.
- The collaboration setup a small group of experts who are estimating the processing power required in the back-end and studying the optimum data acquisition system from the front-end electronics up to the HLT1 decoding.
  - It will have to make recommendations in 2026 in order to outline the DAQ architecture in the *TDR Data Processing Framework* by the end of 2026
- Specifications of the multi-purpose card for the Upgrade II in 2027

The DAQ project for LHCb Upgrade II at IN2P3...

LHCb Data Acquisition Enhancement TDR

- The LHCb collaboration submit to the LHCC, the *LHCb Data Acquisition Enhancement Technical Design Report* describing the evolution of the DAQ for Run 4 :
  - New PCIe400 readout card for RICH channel expansion
  - Implement an hardware processor based on the RETINA architecture and PCIe400 boards to perform real-time track reconstruction in SciFi
  - Learn how to deploy, step by step, a clock targeting an ultimate jitter and phase determinism of $\mathcal{O}$(10 ps)

- The LHCC notes that these development represents an important and effective step towards the future Upgrade II and recommends approval to allow resources to become available and MoUs to be signed.

- Approved by the CERN Research Board in December 2024

Proto-project PCIe400 at IN2P3

- We propose to IN2P3 a *proto-project* in the continuity of the PCIe400 R&D

- Mid-September 2024, the IN2P3 reviewed the proto-project in order to assess its maturity and whether IN2P3 has the resources to commit with confidence

Institut national de physique nucléaire & de physique des particules

Fiche projet de développement

PCIe400 DAQ Engine

PCIe400/Renaud Le Gac

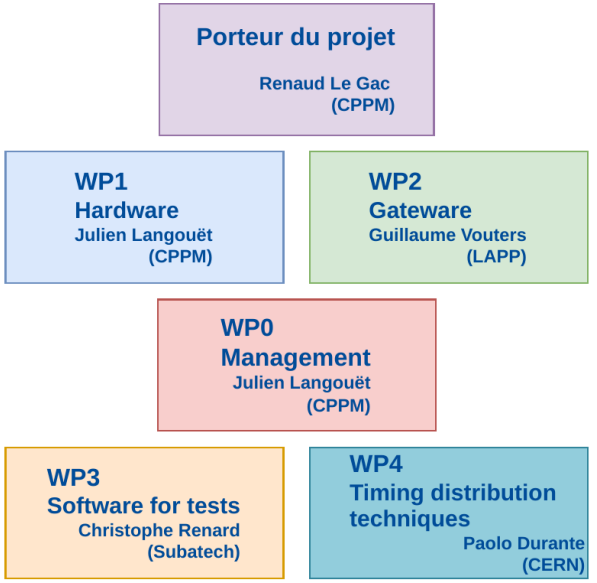
Réf Interne :	Référence interne laboratoire
Réf ATRIUM :	ATRIUM-XXXX
Titre / Title :	Fiche Projet « PCIe400 »
Résumé / *Summary:*	Le document décrit le projet de développement et la production d'une nouvelle carte de lecture générique : PCIe400 dans le cadre de l'Upgrade II de LHCb. *Elle sera installée durant l'arrêt LS3 du LHC (2026-2028).* Ce projet s'étend de 2025 à 2028 et fait suite au projet de R&T.

Nom du porteur du projet : Renaud Le Gac Email : legac@cppm.in2p3.fr	Centre de Physique des Particules de Marseille CPPM Site Web Labo : http://www.cppm.in2p3.fr
Signature numérique apposée sur version finale en PDF	Signature numérique apposée sur version finale en PDF

- Preminary recommendations:
  - Identified as IN2P3 flagship project with high stakes and commitment
  - Extend the R&D phase until the end of 2025 to allow time for the project to mature and overcome key technological barriers

Organisation of the proto-project PCIe400 – LS3

- Supported by 5 IN2P3 institutes: CPPM, IJCLab, LAPP, LPNHE, SUBATECH
  as well as by the LHCb / Online team
  - WP1 Hardware:
    the design and production of the boards as well as preparatory work for the LS4 version
  - WP2 Gateware:
    implementation of the framework and the continuous integration of the gatewares
    + specific developments
  - WP3 Software:
    software development and a series of tests for the qualification and production of the PCIe400 cards
  - WP4 Time distribution / system integration:
    LHCb testing and qualification



- *Coordinators* for each WP are idendified as well as the *Technical and Physics Coordinators*

Contributors LS3 – from KDP2 review Sept 2024

- Physicists and engineers from IN2P3

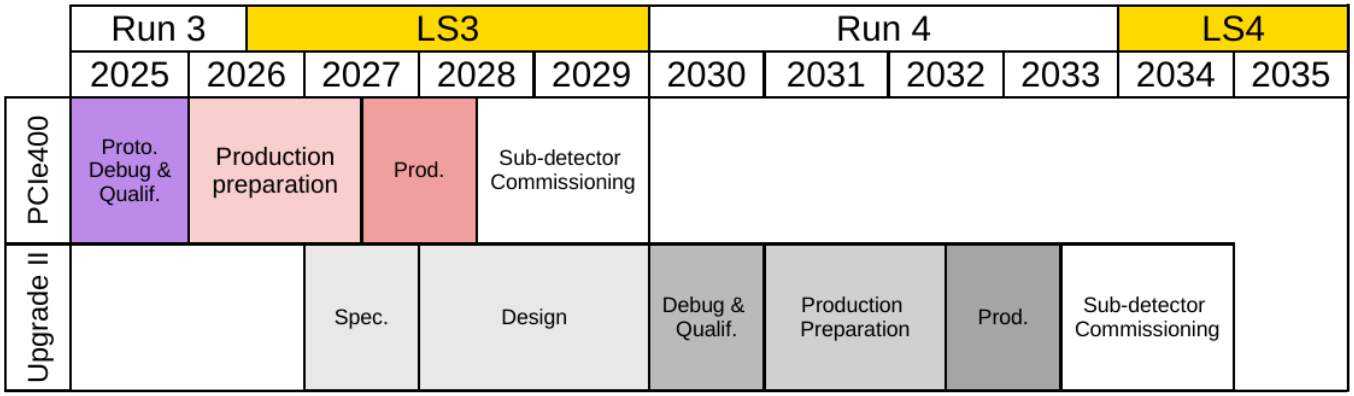
Nom des personnes	Statut
CPPM	
Renaud LE GAC	DRCE

Nom des personnes	Statut
CPPM	
Julien LANGOUËT	IR
Frédéric HACHON	IR
Kévin ARNAUD	IR
Paul BIBRON	IR
Costy NASSIF-MATTAR	APPRENTI
IJClab	
Daniel CHARLET	IR
Antoine BACK	IR
LAPP	
Guillaume Vouters	IR
LPNHE	
Jean-Luc MEUNIER	IR
Stefano RUSSO	IR
Gabriel Degret	IR
SUBATECH	
Christophe RENARD	IR
Amaury HERVO	IR

- From others institutes

Nom des personnes	Statut
CERN	
Paolo DURANTE	IR
TBD	Etudiant
TBD	IR
TOTAL (FTE)	

Master plan



- A card of similar complexity to the PCIe400 was assumed for the Upgrade II timeline.

- Decision on the Upgrade II version should be taken by end 2026.
  They are several possibilities:
  - Only PCIe400
  - PCIe400 + Concentrator on top of it
  - Mixture of PCIe400 and Upgrade II version
  - Only Upgrade II version
  - New idea which can appear in the future

Deliverables – LS3

New date taking into account the update of LHC schedule

Date	Deliverable
December 2026	Performance reports on LpGBT and phase determinism
July 2028	Delivery of a hundred PCIe400 cards to the LHCb Online team with Low Level gateware and software

	PCIe400
RICH	55
DWT axial	64
DWT stereo	32
Total	151 + Spares

CERN-LHCC-2024-001

Deliverables – LS4

Table 5: Number of LpGBT links and their occupancies for LHCb Upgrade II [3]

Sub-detector	DAQ links	Total bandwidth in Tbit/s	Average bandwidth per lpGBT link in Gbit/s
VELO	3 400	34	10.0
UP	1 888	7	3.7
Magnet Station	1 400	5	3.6
Mighty Tracker	9 500	30	3.2
RICH	5 700	30	5.3
TORCH	4 312	27	6.3
PicoCal	2 360	21	8.9
MUON	1 576	16	10.2
Total	30 136	170	

- The number of boards to be produced for the Upgrade II system can be estimated from the number of lpGBT links, as they all need to be connected to a board.
  - The total is $\mathcal{O}(750)$ modules without spares
  - This is comparable to the number of PCIe40 cards produced for LHCb Upgrade I

- The delivery date is under discussion. It will depend on the needs of the sub-detectors to build and commission their system, most likely in the '30s.

Person-power needed

- Fields of interest per institute:

Work Packages	Institutes
WP1 Hardware	CPPM, LHCb/Online
WP2 Gateware	CPPM, LAPP
WP3 Software	CPPM, SUBATECH, LHCb/Online
WP4 Timing distribution techniques	CPPM, IJCLab, LPNHE, LHCb/Online

- Required person-power for the PCIe400 version:

		Long shutdown 3			
Work Package	**2025**	**2026**	**2027**	**2028**	**Total (FTE)**
Management	0.7	0.4	0.3	0.3	1.7
Hardware	1.0	1.2	2.1	2.0	6.3
Gateware	2.4	1.7	1.9	0.6	6.6
Software for tests	1.2	1.1	0.4	0.2	2.9
Timing distribution techniques	2.5	2.5	1.0	1.0	7.0
Total (FTE)	**7.9**	**6.8**	**5.6**	**4.1**	**24.5**

The comparison between required and available FTE shows that the project in unserstaffed by 20 %. The shortfall is mainly in the *time distribution techniques* and *exploratory features*

- Required person-power for the Upgrade II version:

The person-power will depend on the option chosen. It will require the same commitment, $\mathcal{O}(6)$ FTE per year, if the Upgrade II version is similar in complexity to the PCIe400.

Costs of PCIe400 and Upgrade II developements

- The production of the boards is funded by the subdetector projects.
- IN2P3 will fund the pre-production boards and the tooling required for the qualification and burning tests of the boards at the manufacturer.

Item	Run 3		LS3			Run 4				LS4		Total
	2025	2026	2027	2028	2029	2030	2031	2032	2033	2034	2035	
Test bench	10											10
Study liquid cooling		16										16
Production tools at EMS	20	35										55
PCIe400 production pre-series		150										150
Development kits	16	8	2									26
Prototypes Upgrade II version				10	300							310
Production tools at EMS						60						60
Upgrade II production pre-series							240					240
Total per year	46	209	2	10	300	60	240					867

- Projection per version:

LS3	231 k€
LS4	636 k€
Total	867 k€

Risks

- Retirement of the Scientific Coordinator

  Additional resources made available by CPPM or project taken on by someone from another participating institutes.

- Departure of a project member with a key skill.

  Follow up on the recruitment strategies in the light of the identified competences with the participating institutes and IN2P3 management

- Reduction in workforce due to non-extension of fixed-term contracts or non-appointment of permanent staff

  Reallocation of tasks to other members of the project team or identification of a new member of staff through networks.

- No phase determinism for Agilex transceivers

  The PCIe40 might be a backup solution for Run 4. Another approach is to use competitor board for the time distribution

- Non-functional PCIe400 prototype

  Reorganizing the schedule to bring forward the revision phase

- Delays in development and qualification due to technical difficulties

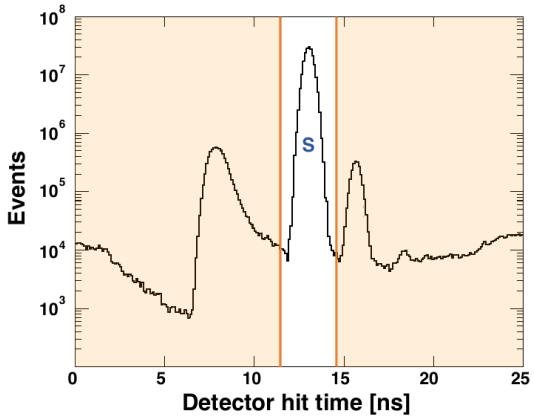
  No more duplication of prototype boards. PCIe Gen5 servers will be qualified using the four prototype boards, while RICH sub-detector developers will use the PCIe40 board

Conclusions

- The proposed project is a continuation of our contributions over more than 25 years, in particular the level-0 muon processor and the multipurpose card, pcie40.
- The IN2P3 is supporting a flagship development in electronics through the R&D of the PCIe400 card (2022 - 2025). The LHCC recognises its importance by approving the LHCb Data Acquisition Enhancement TDR.
- The PCIe400 card is a key component of the LHCb experiment. The collaboration has recognised its importance by creating a sub-project in the Online and by appointing a deputy in charge of the backend for the Upgrade II.
- The project is supported by 5 institutes of IN2P3 and the LHCb Online group. The participating institutes are encouraged to increase their contribution to this large project.
- The PCIe400 is the cornerstone of keeping pace with technology and preparing for the final version of the multipurpose card, which is expected to be specified in 2027.
- Finally, we believe that our major contributions will improve the physics output in the flavour and heavy ion sectors and open the doors to new ideas that can be exploited by future experiments.

Backup...

The RICH time gate...



[LHCb-TDR-024]

	Sensor [σ]	ASIC time walk	FE time gate	TDC time bin
LHC Run 3	150 ps	< 4 ns	6.25 ns	None
LHC Run 4	150 ps	CFD correction	2 ns	25 ps
HL-LHC Run 5	~ 50 ps	CFD correction	150 – 300 ps	25 ps

[LHCb-TDR-026]

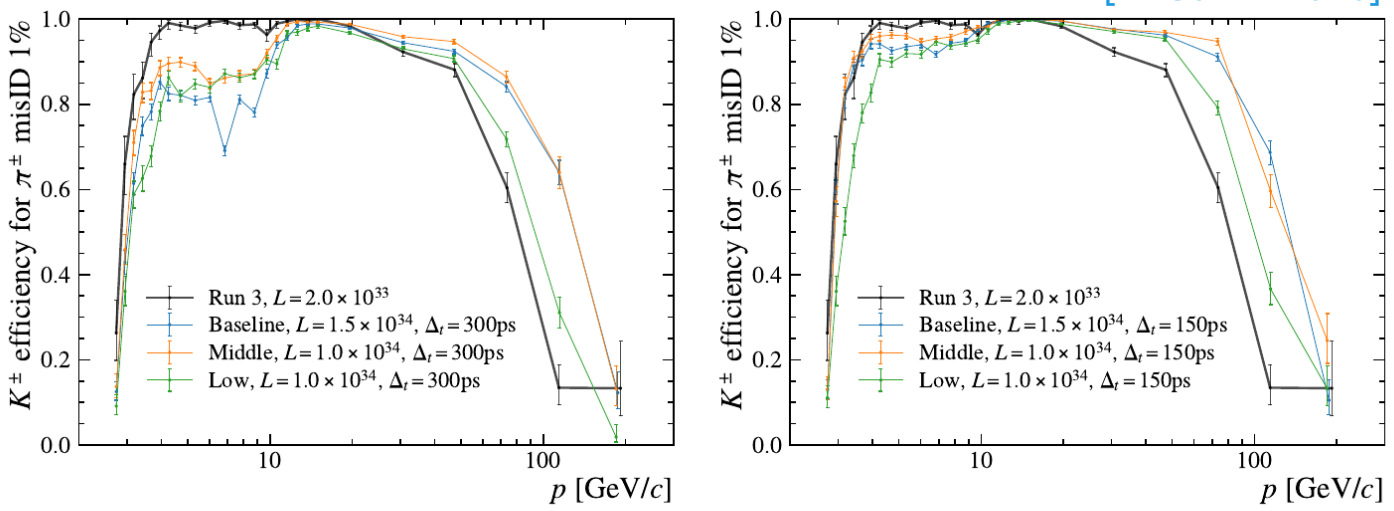
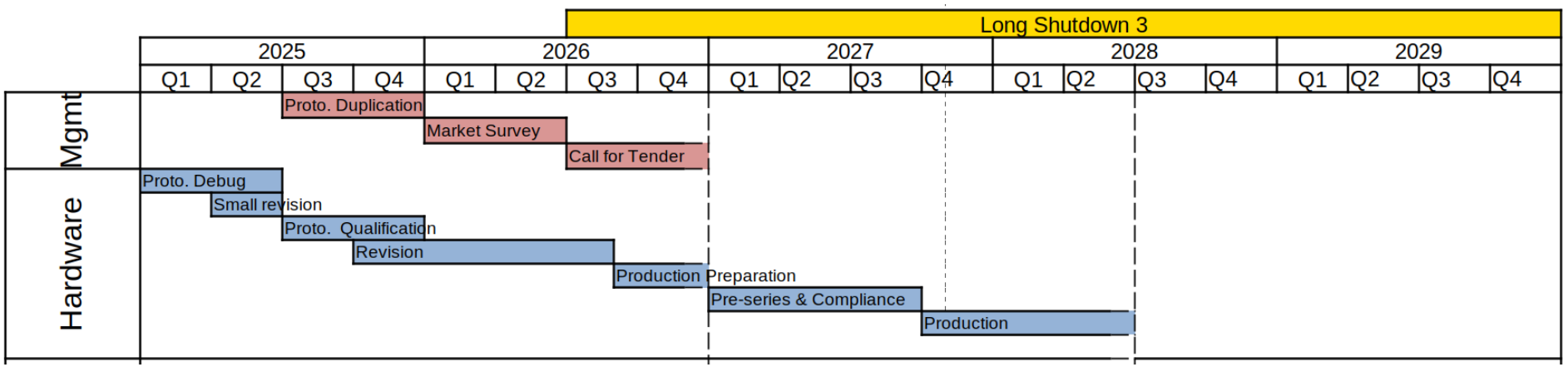


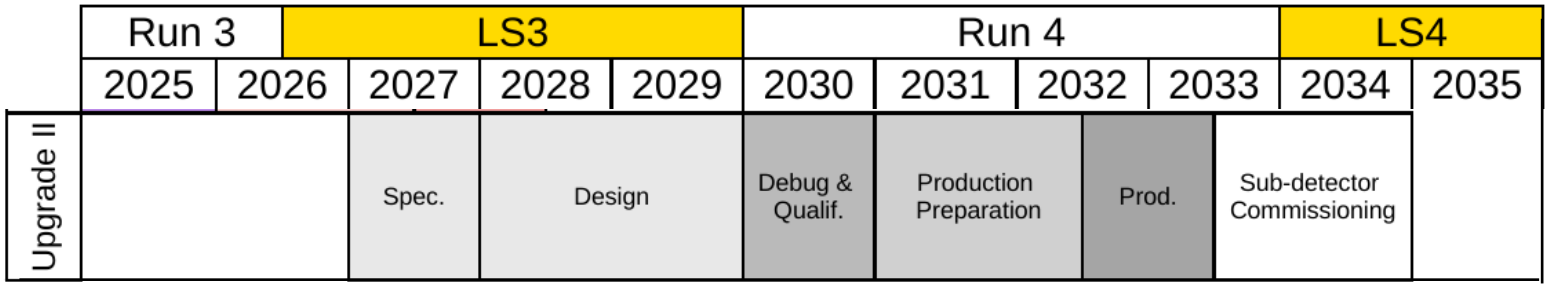
Figure 28: Kaon efficiency for 1% pion misidentification versus momentum, for the Run 3 and the three RICH Upgrade II options at different luminosities, for (left) 300 ps and (right) 150 ps time windows, assuming the same average photoelectron yields.

Safety margin planning LS3



- Safety margin ~12 months :
  - Revision of 12 months is for of full redesign of the card
    Minor revision can be done in about 3 to 6 months
  - Production of 150 unit should not exceed 3 months.
    It mainly depends on the contract signed with the EMS
- In case of disaster the PCIe40 is the backup

Safety margin planning LS4



- Safety margin ?
  - Depend on the complexity of the design and on the delivery date
  - The PCIe400 took a long time to specify and design because both the FPGA and the card were in development at the same time. For the Upgrade II version, we hope to use a well-known FPGA that is already in production. This will reduce the specification and design steps by a year
  - Adapting the project to ensure that it is delivered safely and on time
- In case of disaster, the PCIe400 could be the backup.

Possible concentrator for Run 5...

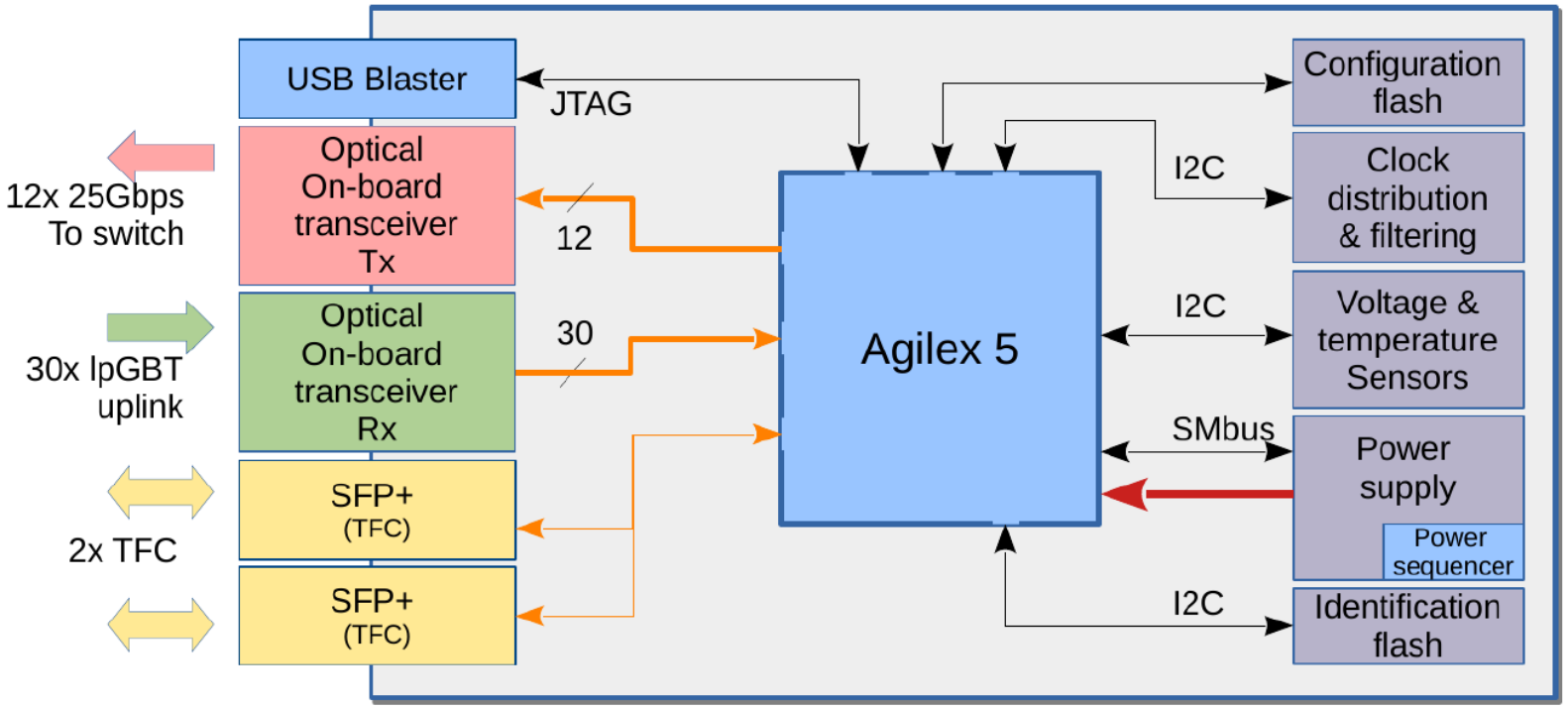
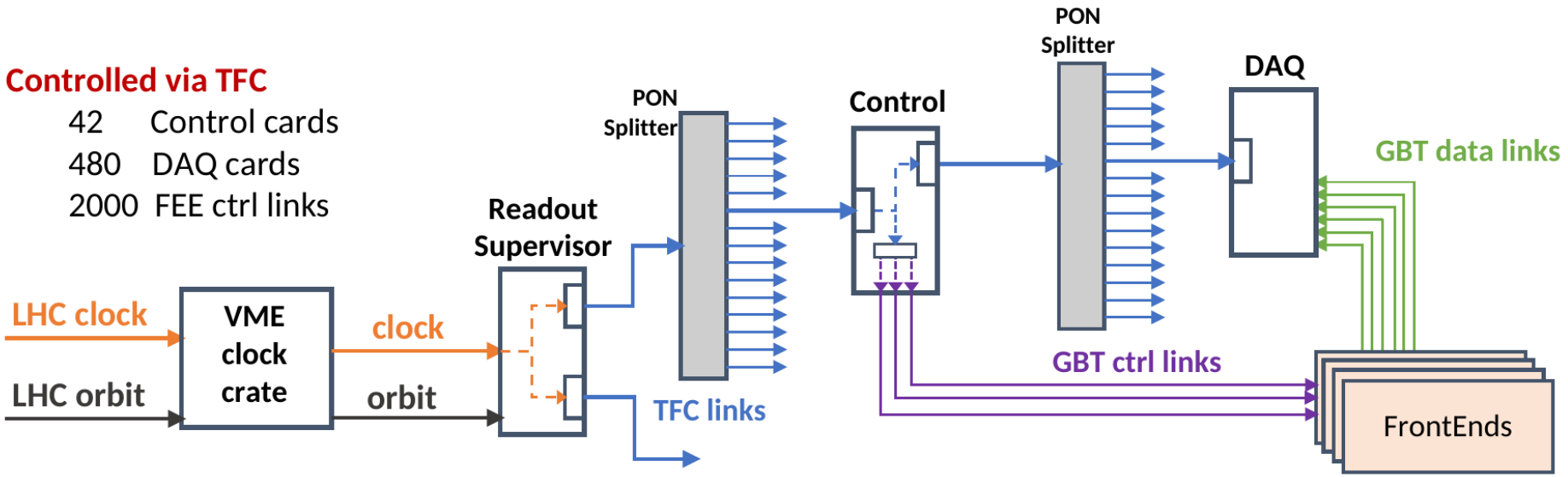
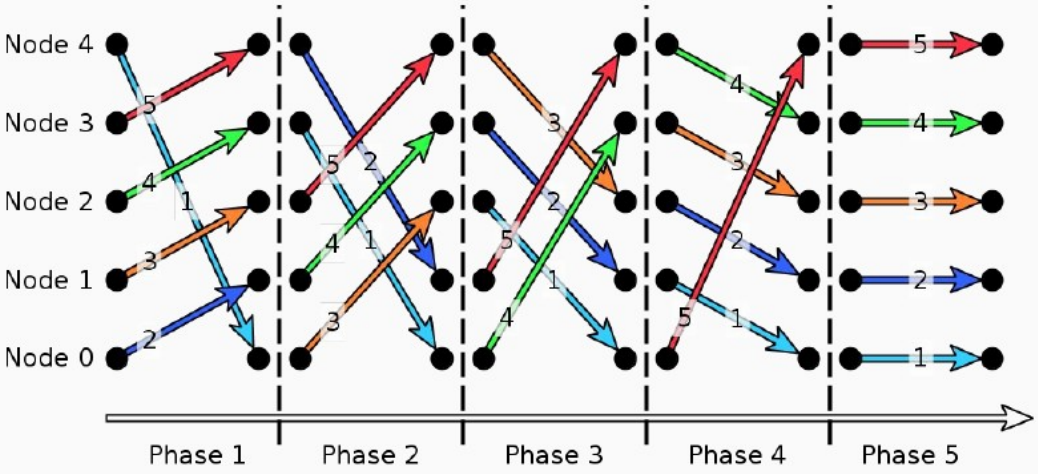


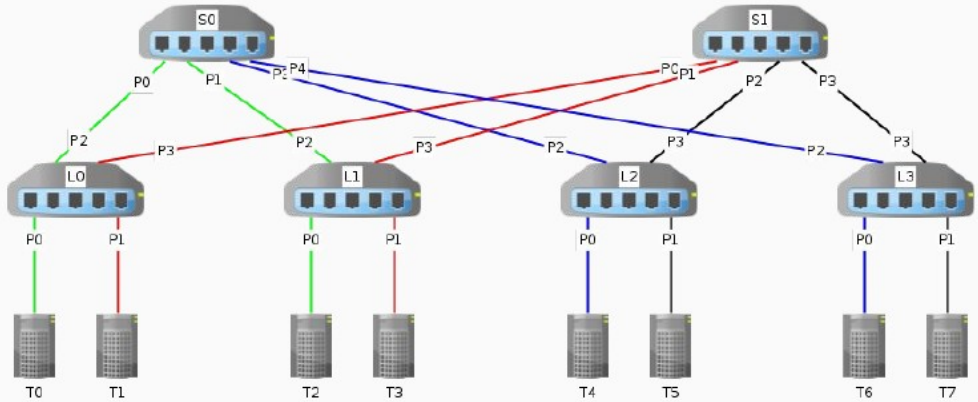
Figure 19: First sketch of a concentrator board upstream to the PCIe400 board.

Simplified view of the TFC – Run 3



Event Builder: Traffic scheduling





- The processing of **N** events is divided into **N** phases (**N** is the number of EB nodes
- In every phase one RU sends data to one BU, and every BU receives data from one RU
- During phase **n** RU **x** sends data to BU (**x** + **n**)%**N**
- All the units switch synchronously from phase **n** to phase **n** + 1

Congestion-free traffic on "selected networks" (e.g. fat-tree networks)

Retina architecture

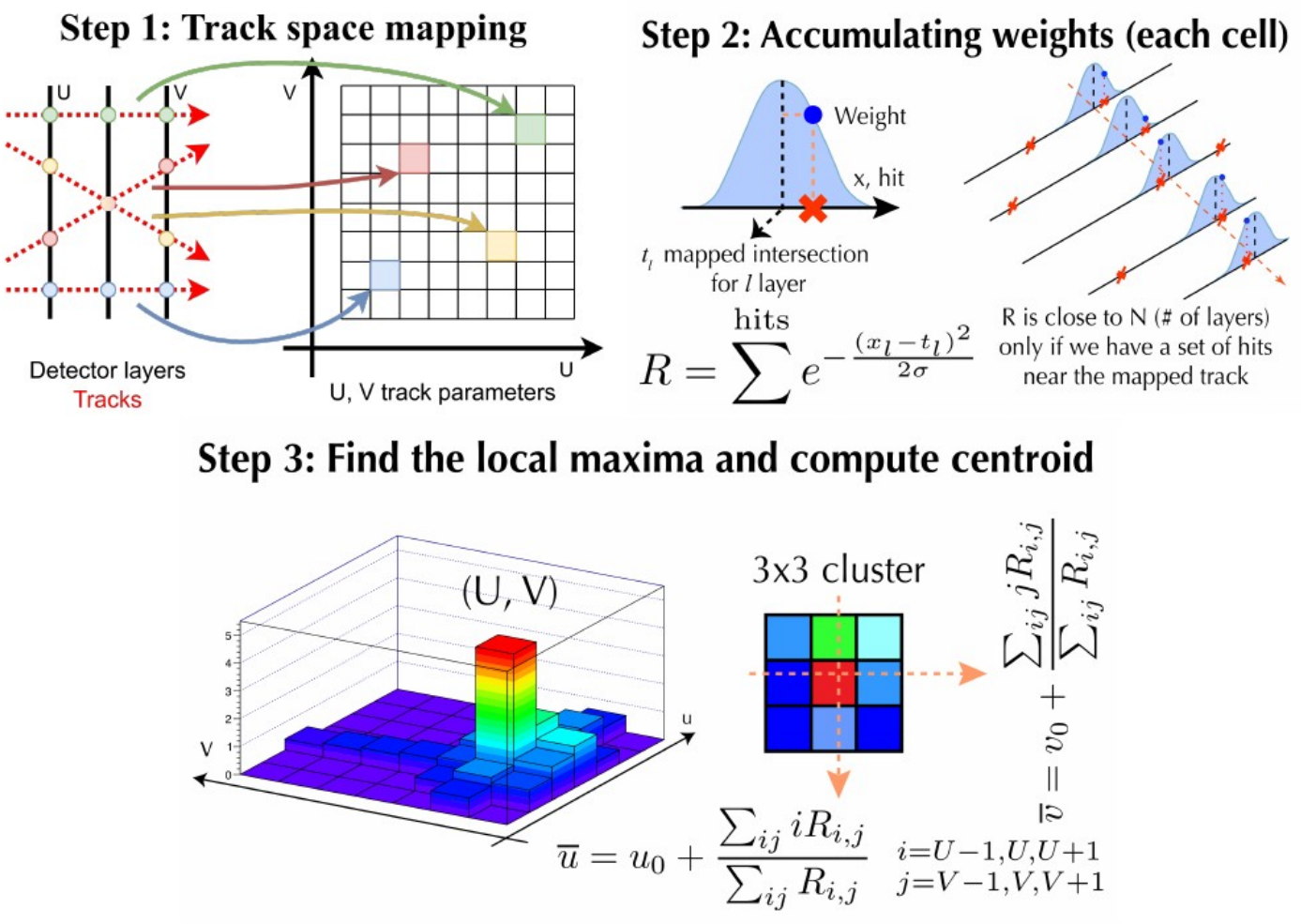


Figure 18: Track reconstruction steps with the Retina architecture.

DWT processor – Run 4

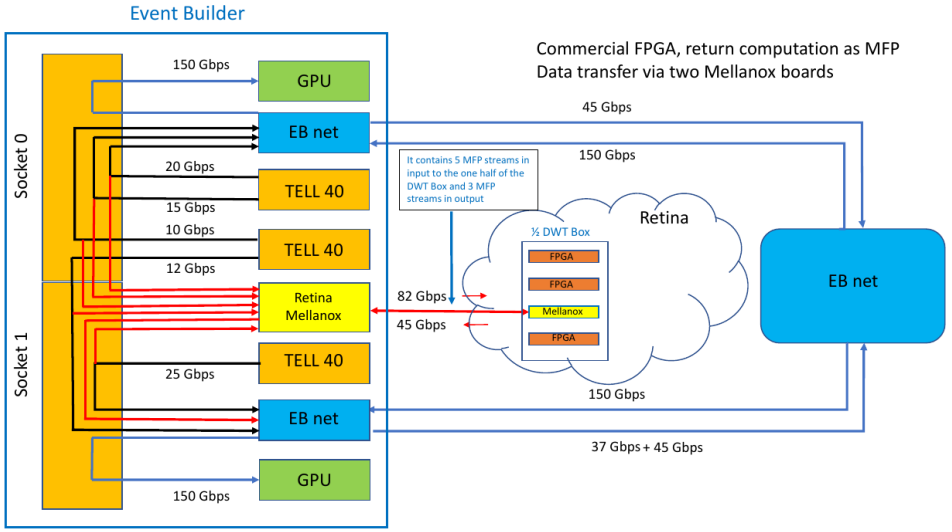


Figure 28: Connection scheme of the DWT processor.

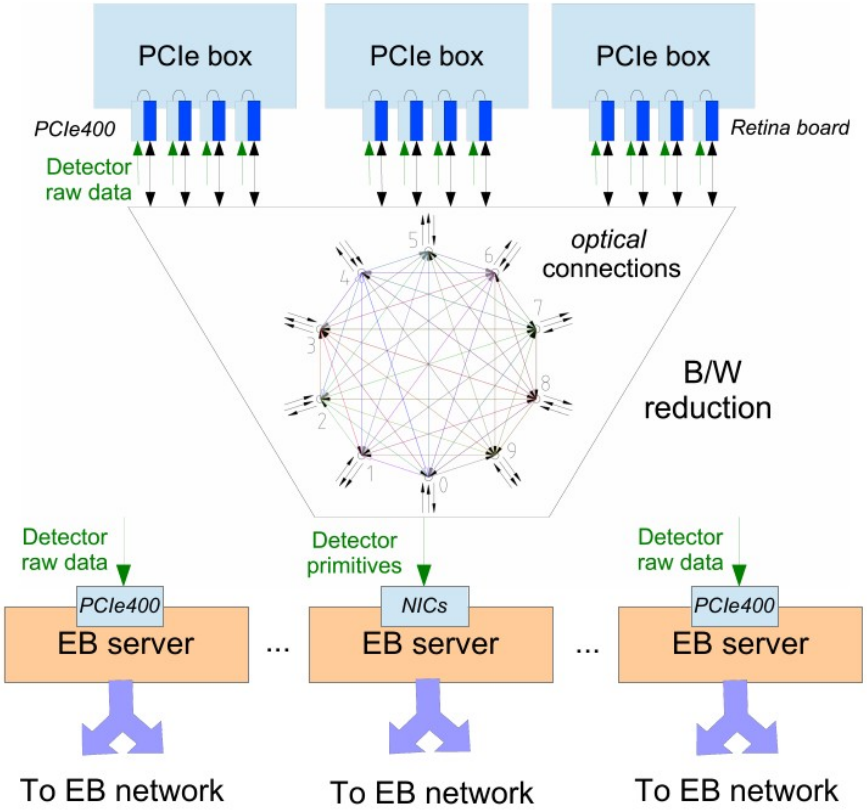


Figure 24: Possible data flow in Upgrade II.

Evolution of switch bandwidth