PCIe400 :
Carte de readout générique

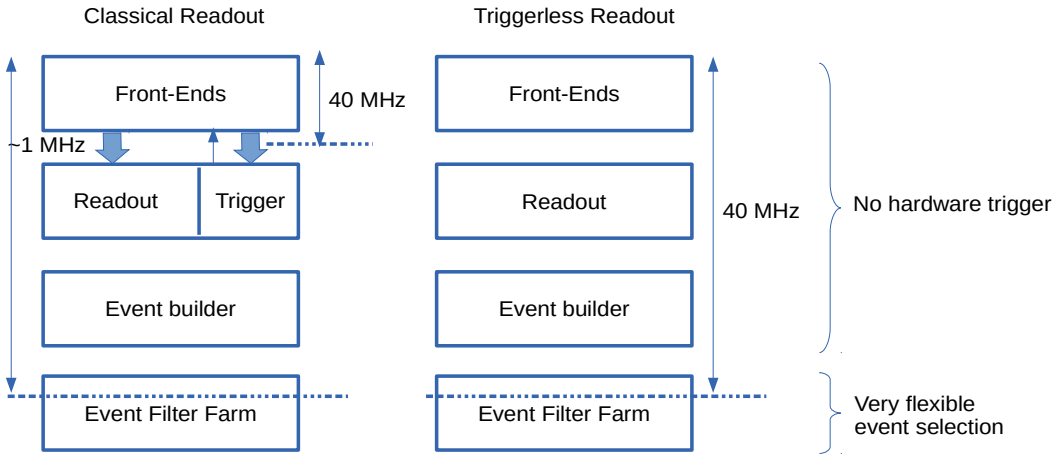
Julien Langouët CPPM
Paul Bibron, Jean-Pierre Cachemiche, Renaud Le Gac, Frédéric Réthoré **CPPM**
Paolo Durante, Antoine Junique **CERN**
Antoine Bach, Christophe Beigbeder, Daniel Charlet, Chafik Cehikali,
Christelle Soulet, Monique Taurigna, Souhir Elloumi, Eric Plaige, Xavier Lafay **IJCLab**
Pierre Delebecque, Jean-Marc Nappa, Sebastien Vilalte, Guillaume Vouters **LAPP**
Thomas Chabaud, Frederic Druriolle, Patrick Hellmuth, Abdel Rebii **LP2I**
David Etasse **LPC Caen**

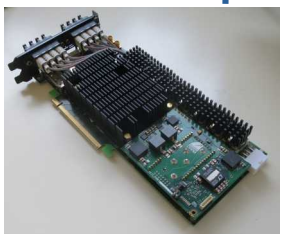
Plan

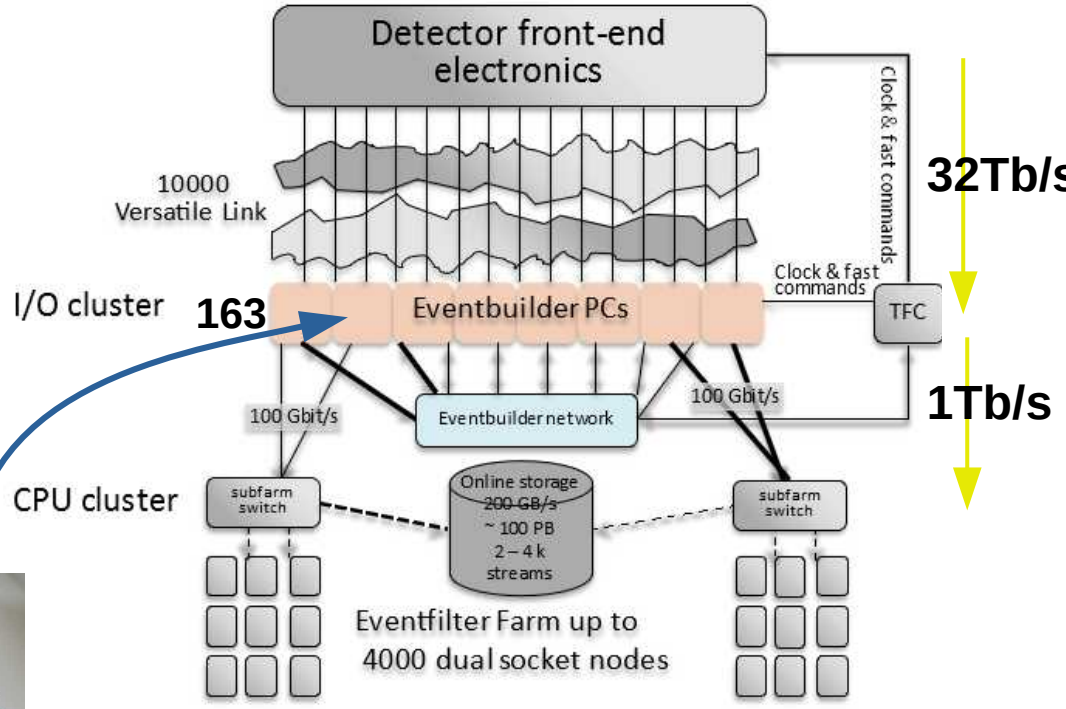
Architecture LHCb

Choix d'une architecture « triggerless »

- Résoud le problème de la saturation des triggers sous l'effet de la forte luminosité.
- Permet de mettre en place des algitithmes plus élaborés
- Déjà adopté lors de l'upgrade I







Fonction et besoins futurs

Rôle

- Charnière entre protocoles « *custom* » et protocoles commerciaux standards utilisés dans les data centers
- Concentrateur de données 48 → 1 interfaces sériels ~10G compatibles avec les futurs sérialiseurs du CERN vers interfaces très haut débit ~400G
- Premier étage de l'event building

LHCb Upgrade II (2035)[1]

- Déploiement LS3 (2026/2028) et LS4 (2033/2034), prise de données 2035
- La bande passante requiert probablement un facteur 10
  - Pas faisable avec la technologie actuelle
  - Etape intermédiaire (LS3 2026/2028) : facteur 4 sur le débit de sortie et sur la capacité de traitement du FPGA
    - Bande passante avec l'event builder 400 Gbps (PCIe Gen5 ou 400 GbE)
    - FPGA avec environ 4 millions de locic cells
    - Gestion précise du temps : < 10 ps RMS

Autres besoins

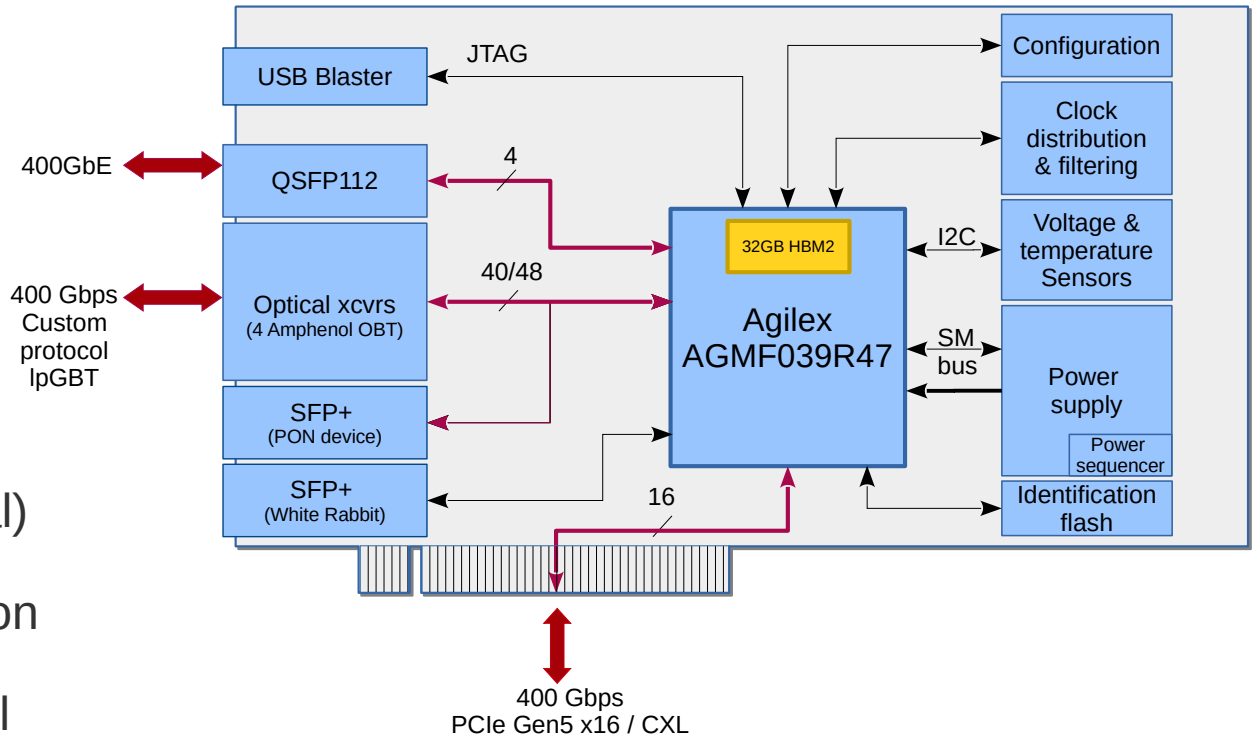
- Généricité pour utilisation dans de multiples contextes
  - Déjà des intérêts par ALICE, BelleII et CTA

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

La carte PCIe400

Caractéristiques envisagées

- Agilex AGMF039R47A1E1V
  - 3.9 MLE
  - 32 GB HBM2 memory
  - Internal speed up to 1 GHz
  - PCIe Gen5 / 400GbE hard IP
  - 32 links at 56G PAM4
    or 48 links au 32G NRZ
    for custom protocols
- No DDR memory
  - Use of PC memory instead
- 26G NRZ optics for FE
- 112G PAM4 for 400GbE (optional)
- White Rabbit clock synchronisation
- PON management for fast control
  - TTC-Pon



Gain estimé vs PCIe40 :

- Processing : factor 8 to 12
- Output bandwidth : factor 4

Organisation

Projet de R&T PCIe400

Motivations

- Equipe réduite en fin de projet LHCb upgrade I (PCIe40)
- Départ à la retraite de 2 ingénieurs du CPPM ayant conçu le hardware la carte PCIe40
- Départ à la retraite d'un informaticien
  - Besoin de ressources supplémentaires

- Intérêt d'autres groupes dans d'autres laboratoires : ALICE, BelleII, CTA

Mise en place du projet de R&T PCIe400

- Projet démarré en 2022
- Durée 3 ans
- Objectif : développer le hardware, les firmwares et logiciels de base génériques
- Industrialisation et production non incluses

Ressources

5 laboratoires + CERN

- Groupe important mais ne représente qu'environ 4.5 FTE
  - Demande une organisation rigoureuse
- Fragilités :
  - Hardware
    - 2 départs à la retraite imminents pour le CPPM
    - 1 CDD à prolonger ou stabiliser
  - Passage de relais au niveau de l'IJCLab

Nom des personnes	Statut	2022	2023	2024	Total (FTE)
CPPM		10%	10%	10%	0.30
Renaud LE GAC	DR1	10%	10%	10%	
TOTAL (FTE)		10%	10%	10%	30%

Chercheurs

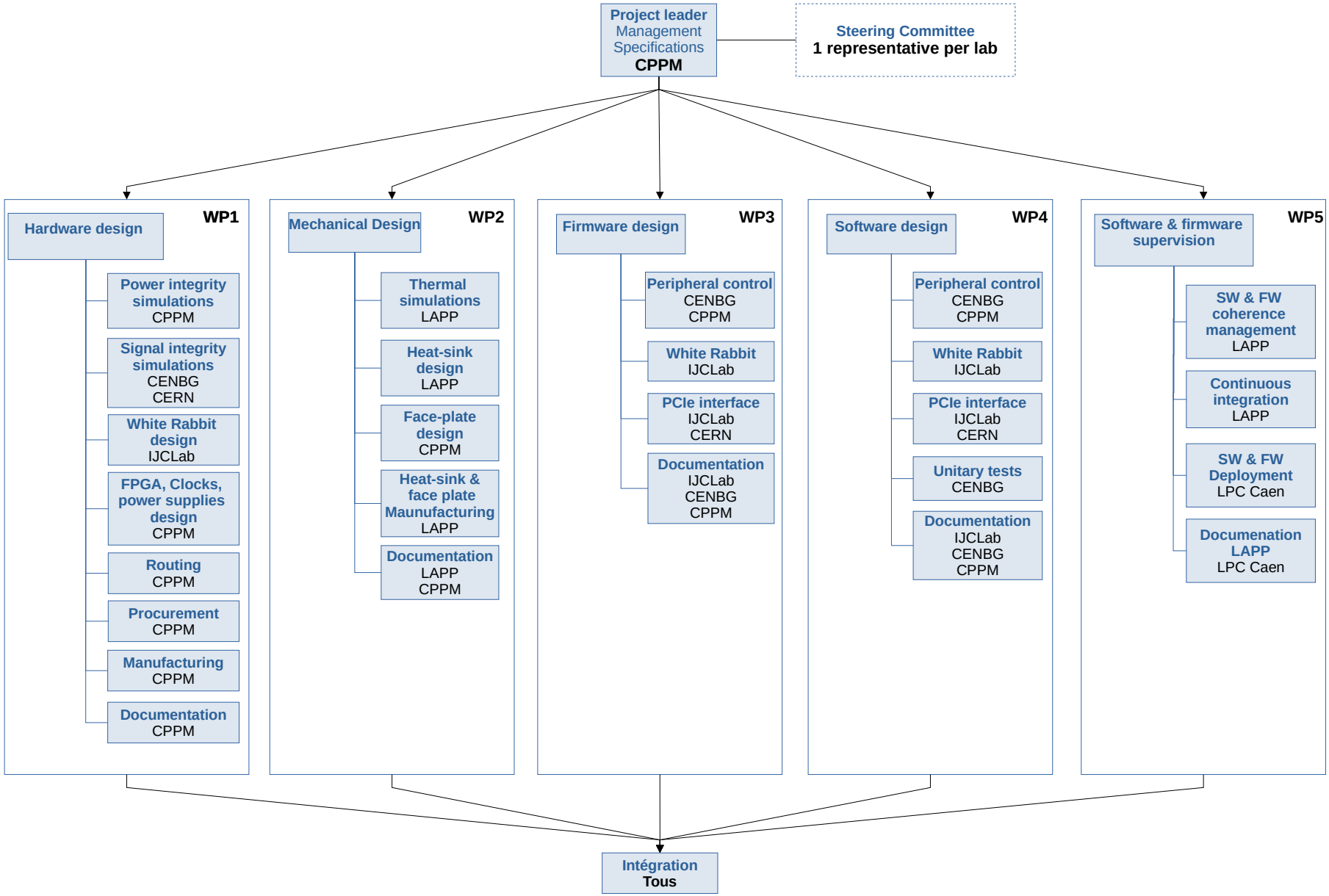
Nom des personnes	Statut	2022	2023	2024	Total (FTE)
CPPM		185%	100%	100%	3.85
Jean-Pierre CACHEMICHE	IRHC	60%	0%	0%	
Frédéric RETHORE	IR	25%	0%	0%	
Paul BIBRON	CDD IR	100%	100%	100%	
Kevin ARNAUD	IE	50%	30%	0%	
LAPP		30%	30%	15%	0.75
Guillaume VOUTERS	IR	15%	15%	15%	
Sebastien VILALTE	IR	5%	5%	0%	
Jean Marc NAPPA	IE	5%	5%	0%	
Pierre DELBECQUE	IR	5%	5%	0%	
CENBG		60%	95%	95%	2.50
Frédéric DRUILLOLE	IRHC	10%	10%	10%	
Patrick HELLMUTH	IR	10%	15%	15%	
Abdel REBII	IR	30%	50%	50%	
Thomas CHABAUD	AI	10%	20%	20%	
IJC lab		160%	270%	220%	6.50
Christophe BEIGBEDER	IRHC	10%	10%	10%	
Daniel CHARLET	IR	30%	30%	10%	
Chafik CHEIKALI	IE	10%	10%	10%	
Christelle SOULET	IR	10%	10%	10%	
Monique TAURIGNA	IE	50%	50%	10%	
Souhir ELLOUMI	IE	10%	50%	50%	
Eric PLAIGE	IE	10%	10%	10%	
Xavier LAFAY	IE	10%	20%	30%	
CDD	IE	20%	80%	80%	
LPC Caen		15%	15%	15%	0.45
David Etasse	IR	15%	15%	15%	
TOTAL (FTE)		4.35	4.95	4.30	13.60

Ingénieurs IN2P3

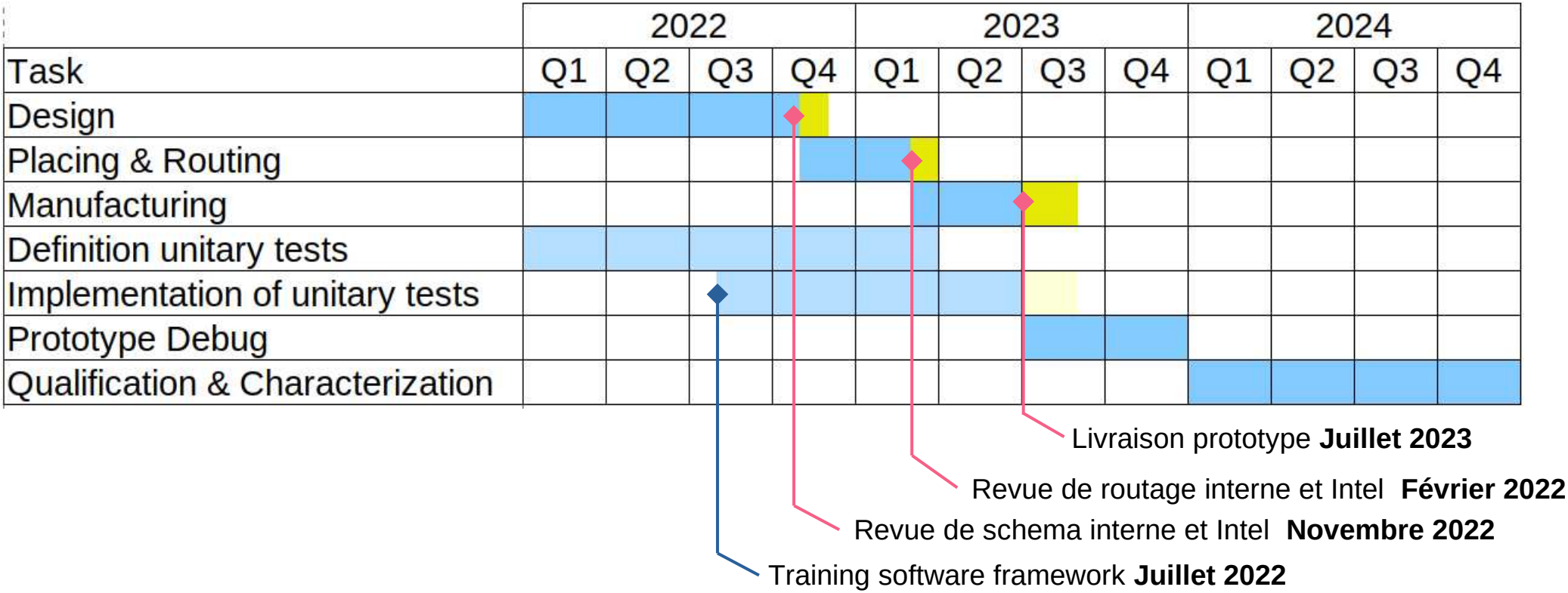
Nom des personnes	Statut	2020	2021	2022	Total (FTE)
CERN		45%	45%	30%	1.20
Antoine JUNIQUE	IR	15%	15%	0%	
Paolo DURANTE	IR	30%	30%	30%	
TOTAL (FTE)		0.45	0.45	0.30	1.20

Ingénieurs CERN

Tâches


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

Planning



Contraintes

- Livraison tardive de la documentation Intel pour le FPGA choisi
  - Collaboration étroite avec spécialistes Intel (Programme Design Engagement Flow)
- Ressources de routage limitées en interne au CPPM
  - Formalisation des points sensibles au routage
  - Travail préparatoire sur le placement déjà engagé

Choix technologiques

Dimensionnement alimentations

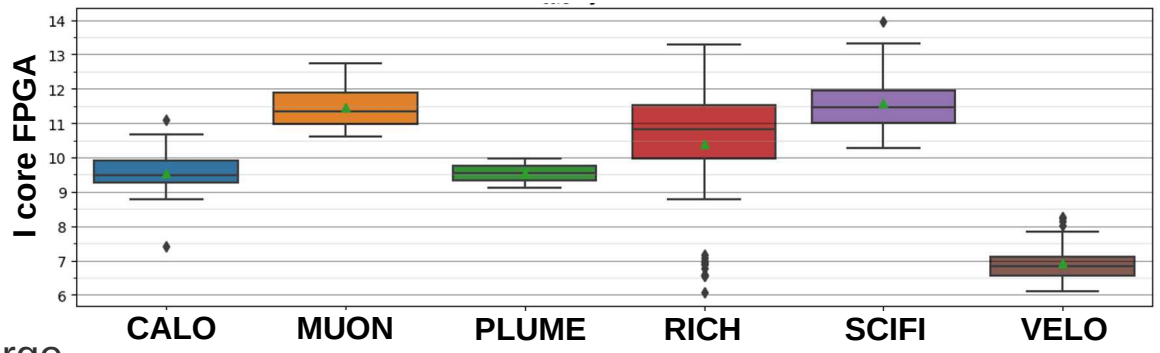
Méthode de simulation

- Modèle de simulation disponible à partir de Q4 2022
  - Extrapolation depuis un Agilex I-series (dernière génération disponible)

- Firmware définitif non disponible
  - Utilisation de random pattern generators pour émuler la charge
    - Ajustement d'un toggle rate virtuel par comparaison avec les consommations enregistrées sur des firmwares réels
    - Simulations post-fit (peut prendre jusqu'à 30 heures)
  - Exportations de fichiers VCD vers Quartus Power Analyzer

Consommation estimée pour le coeur du FPGA

- Cas typique 85W :
  - 12.5 % toggle rate,
  - 70 % de taux d'occupation,
  - 45°C température de jonction
- Pas encore toutes les informations des détecteurs actuels
- Réflexion en cours pour établir la marge



Courant consommée (A) dans coeur FPGA PCIe40 classé par Firmware de sous-détecteur LHCb

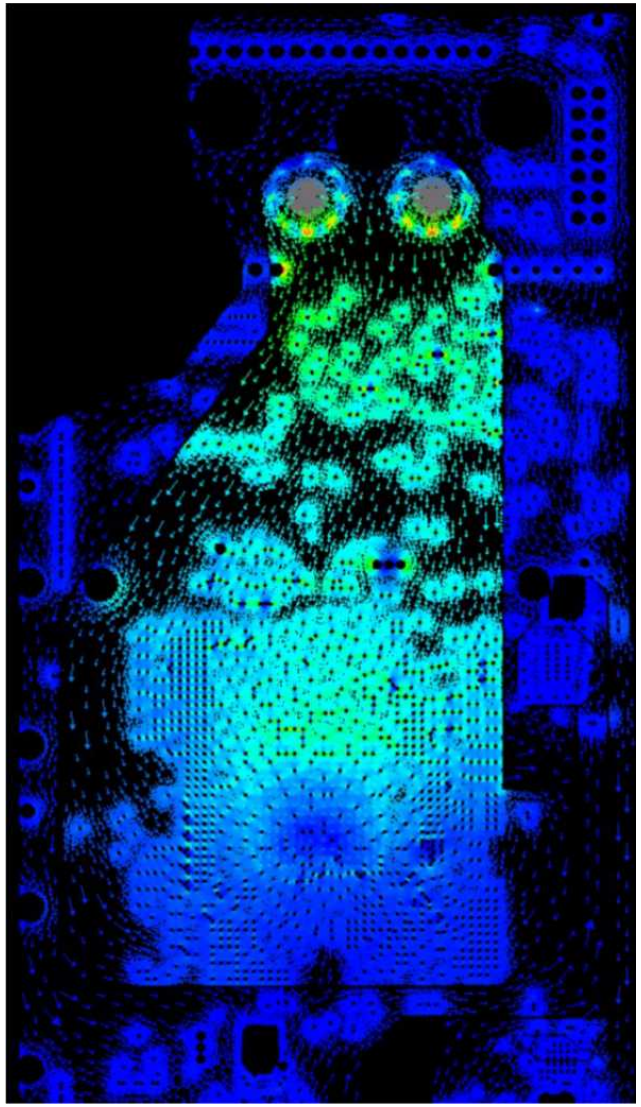
Power integrity

Alimentation du FPGA

- Entre 55 et 100A dans le coeur FPGA (pire cas)
  - >20 rails d'alimentation de 0.8V à 1.8V
  - Attention particulière aux géométries des plans d'alimentations
  - Attention aux densités de courant dans les vias

Simulations Cadence PowerDC / OptimizePI

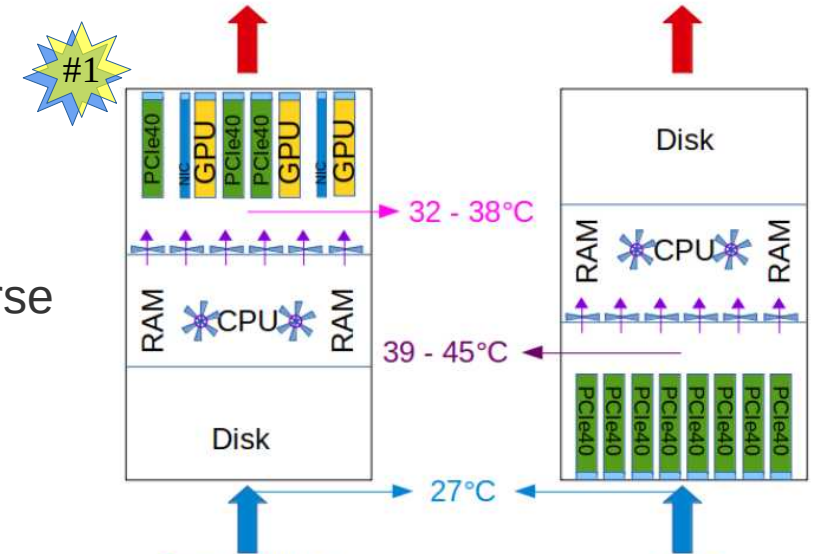
- Analyse en courant statique
  - Vérification du drop et étranglements
- Analyse en courant dynamique
  - Vérification du découplage jusqu'a F_{target}



Refroidissement

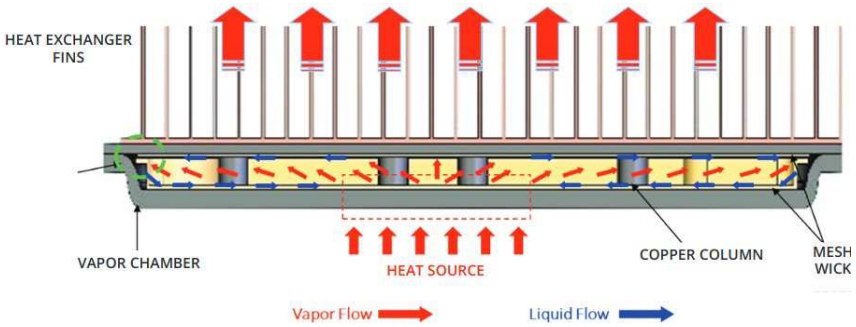
Étude technologique

- 2 architectures de serveurs avec flux d'air inverse
  - Peu de spécifications sur la norme PCIe
    - Température entre 20 et 60°C
      - Mesure au CERN : entre 32 et 38°C
    - Flux d'air supérieur à 1m/s (200LFM)
- Radiateurs
  - Simulations réalisées avec Heatscape
  - Plusieurs types de radiateurs évalués
    - Base métalique + *zipper fins*
    - Base « *vapor chamber* » + *zipper fins*
  - Montrent qu'une technologie avec vapor chamber plus efficace

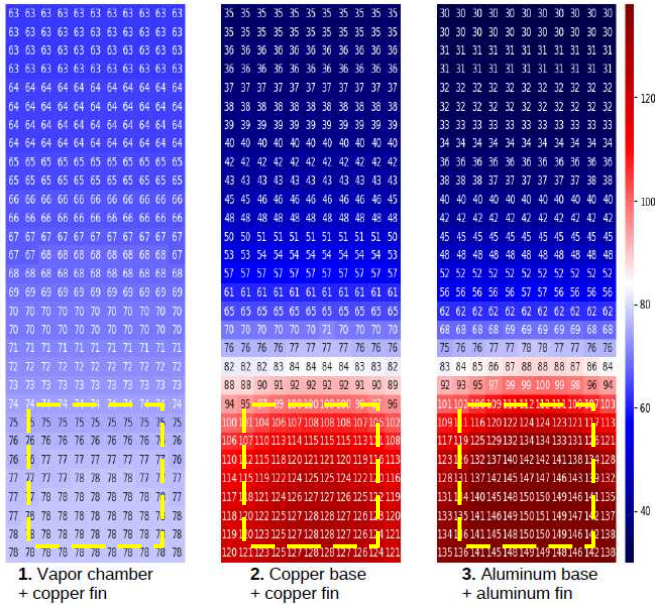


Architectures serveurs

Zipper fin



Principe du « *vapor chamber* »



Simulation heatscape

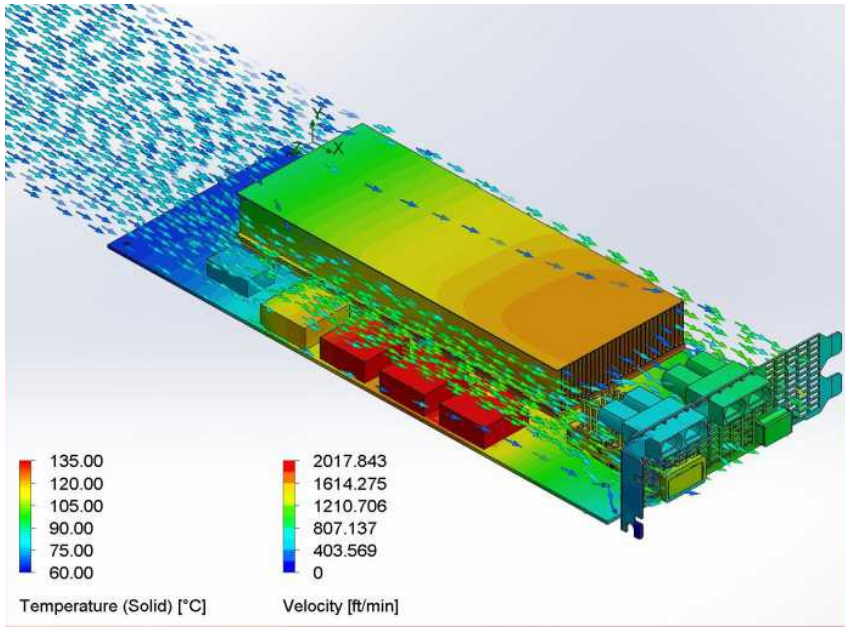
Refroidissement

Réalisation d'un modèle de « vapor chamber » avec COMSOL

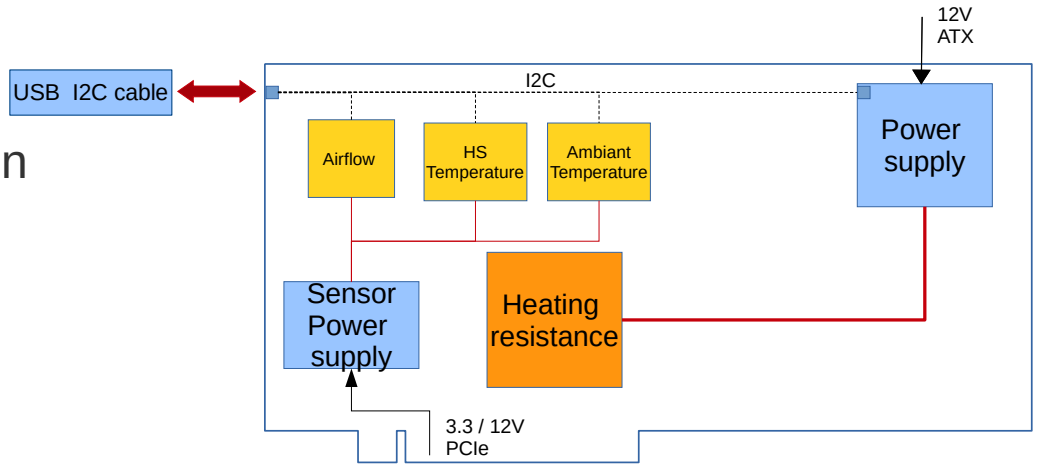
- Variation géométrie
- Hauteur et largeur des ailettes
- Sens et vitesse de l'air

Therma400

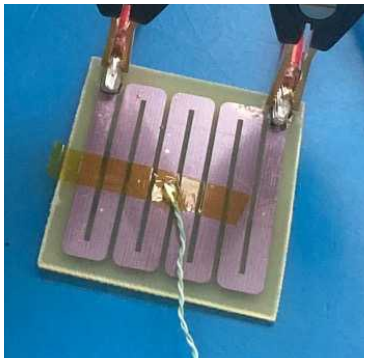
- Design d'une carte instrumentée en température et mesure de vitesse de l'air
  - Permet :
    - Vérification de simulation CFD
    - Essayer un radiateur
- Emulation du FPGA par une résistance en serpentin
- En cours de fabrication



Simulation CFD



Synoptique Therma400

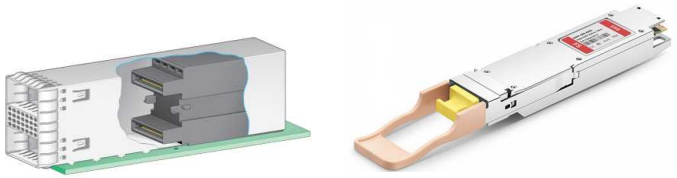


Prototype de résistance en serpentin

Transceiver opto-électroniques

Plusieurs solutions envisagées pour FE

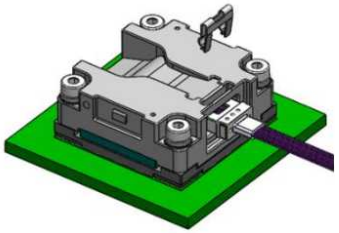
- Anticipation des développements de sérialiseurs du CERN
  - Transceiver « On-Edge » (QSFP-DD)
    - Obturation du flux d'air en face avant
    - Compatibilité custom protocol CERN
  - Firefly Samtec
    - Faible densité en nombre de canaux
  - En concertation avec le CERN, choix porté sur BOAs de Finisar
  - Finalement remplacés par OBT d'Amphenol
    - Compatible avec Finisar
    - Meilleur support
    - Solution de refroidissement fournie
    - Socket fournis



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

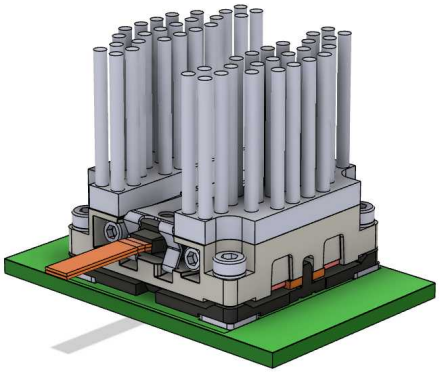
QSFP112
106.25Gb/s PAM4

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



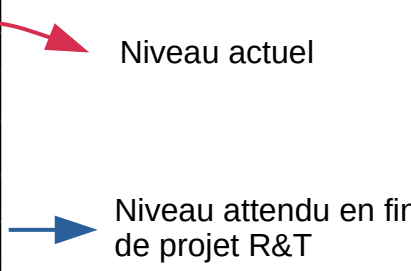
Standard QSFP112 pour 400GbE

- MSA task force regroupe les FF
  - QSFP-DD/QSFP-DD800/QSFP112
- QSFP112 introduit en 2021
  - Preuve de concept
  - Cages et Direct Attach Cable disponibles

Statut

Tous les principaux choix technologiques finalisés

Niveau	Nom synthétique	Interface optique FE (~10G)	Interface PCIe Gen5	Interface réseau 400GbE	Cooling	Gestion précise du temps ~10 ps
TRL1	Principe de base					
TRL2	Application formulé					
TRL3	Preuve du concept					
TRL4	Validation fonctionnelle					
TRL5	Modèles à échelle réduite					
TRL6	Validation de la conception					
TRL7	Qualification d'un modèle					
TRL8	Qualification du système réel					
TRL9	Opération du système réel					



Schématique en cours

- Basée uniquement sur bibliothèque IN2P3 pour faciliter les échanges

Difficultés actuelles

- Documentations du FPGA disponibles Q4 2022
  - Schématique des périphériques saisie en amont
  - Réunions de travail en étroite collaboration avec Intel
- Approvisionnement composants
  - Large anticipation des commandes, mais risque subsiste sur petits composants

Conclusion

Conception en cours depuis Avril 2021

Hardware

- Principaux choix finalisés
- Le routage devrait démarrer courant novembre
- Prototype attendu en mi-2023

Firmwares et softwares

- Démarent actuellement
- L'objectif est qu'ils soient en grande partie disponibles pour le debug
- Seront testés par une maquette virtuelle de la carte
  - Cartes d'évaluations connectées par des interfaces USB/I2C ou USB/SPI

Carte relativement générique pouvant être utilisée dans de nombreux domaines

Difficultée : ressources humaines

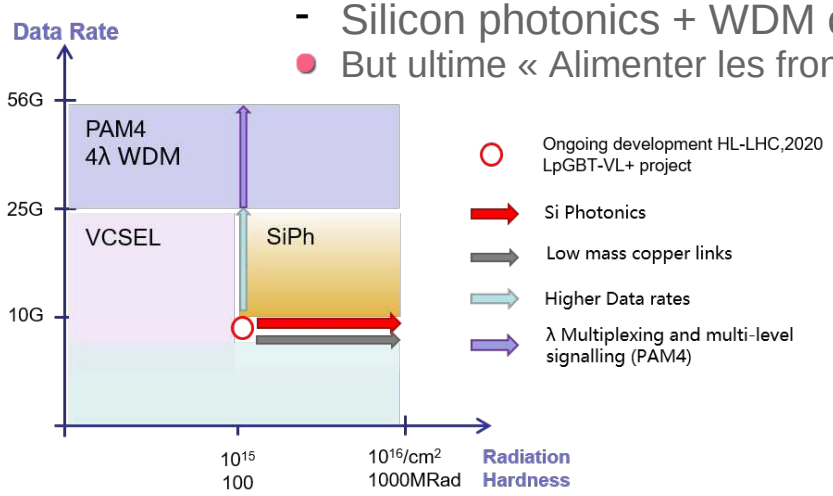
- Départs à la retraite
- CDD à prolonger ou stabiliser en conception hardware
- Ressources de routage limitée

Backup

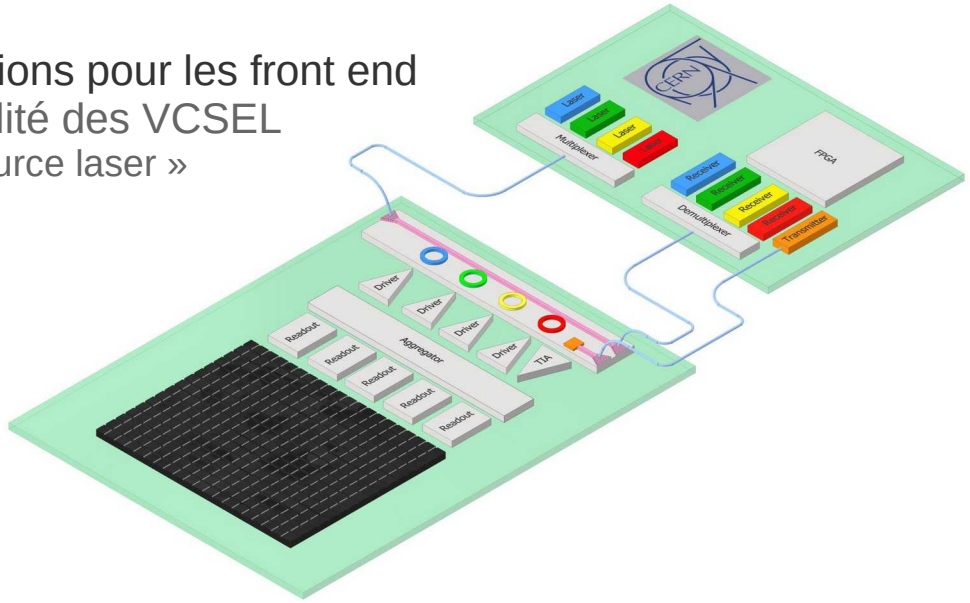
R&D sérialiseurs CERN

Programme inscrit dans WP6 du EP-R&D CERN

- Objectif et motivations
  - Augmentation du débit de données des liens sériels
    - Solution envisagée : Modulation en amplitude (PAM4), Multiplexage en longueur d'onde (WDM)
  - Augmentation de la tolérance aux radiations pour les front end
    - Silicon photonics + WDM car sensibilité des VCSEL
    - But ultime « Alimenter les front-end en source laser »



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



Vue d'ensemble système

Statut actuel

- Choix technologiques ASIC DART28 (transmetteur/driver SiPh)
  - 25.65Gbps NRZ (multiple LHC bunch Clock)
  - FEC équivalent au lpGBT
  - 28nm CMOS
    - Prototypage sur FPGA
- Probable compatibilité avec PCIe400 ± demultiplexeur WDM