

PCIe400 : Carte de readout générique



J.-P. Cachemiche (CPPM)

Paul Bibron, Julien Languouët, Renaud Le Gac, Frédéric Réthoré

Plan

Besoins

Choix technologiques

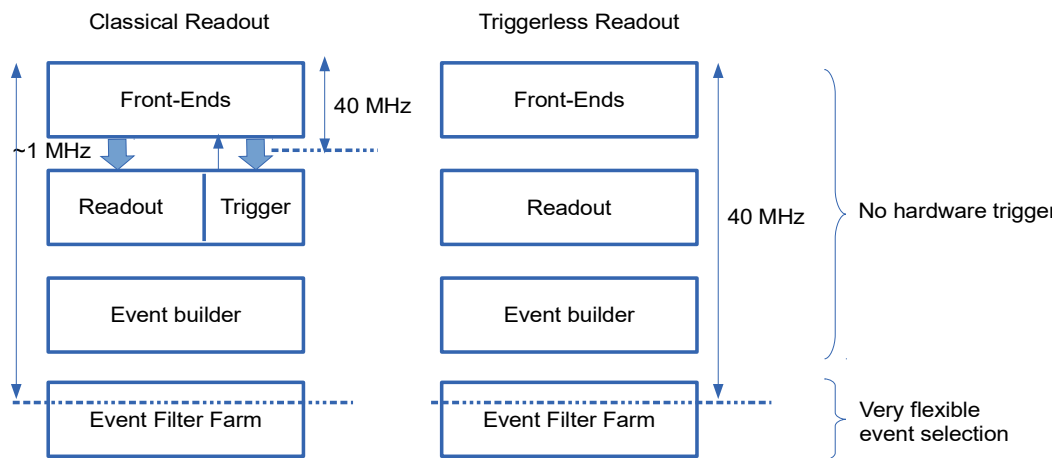
Caractéristiques de la carte

Status

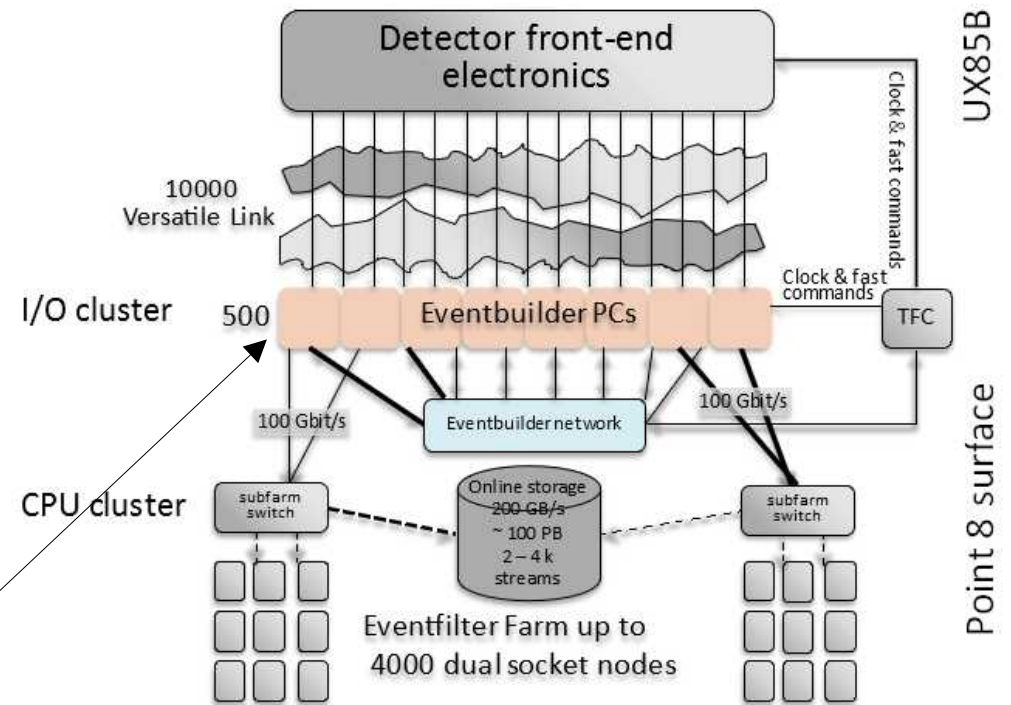
Architecture LHCb

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 algorithmes plus élaborés
- Déjà adopté lors de l'upgrade I



Carte de readout PCIe400



Fonction et besoins futurs

Rôle

- La carte est en fait un concentrateur de données compatible du LpGBT et des futurs sérialiseurs du CERN ou d'autres expériences
- Premier étage de l'évent building : concentration 48 → 1
- Charnière entre le monde "custom" et le monde des réseaux commerciaux standards utilisés dans les data centers

Upgrade II (2030)

- Probablement un facteur 10 requis
- Pas faisable avec la technologie actuelle
- Etape intermédiaire (upgrade 1b : 2026/2027) : facteur 4 sur le débit de sortie et sur la capacité de traitement du FPGA
 - ➔ Bande passante avec le CPU 400 Gbits/s (PCIe Gen5 ou 400 GbE)
 - ➔ FPGA avec environ 4 millions de logic cells
 - ➔ Gestion précise du temps : < 10 ps RMS

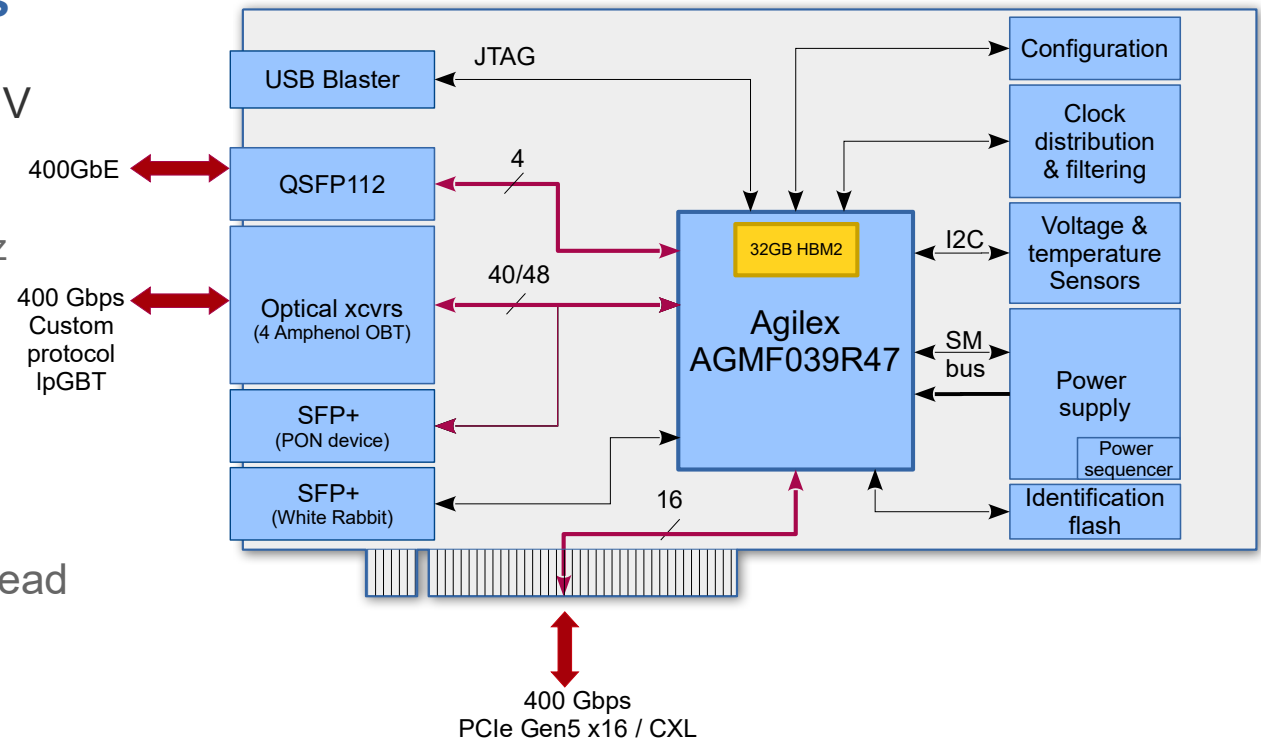
Autres besoins

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

La carte PCIe400

Caractéristiques envisagées

- Agilex AGMF039R47A1E1V
 - 3.9 MLE
 - 32 GB HBM memory
 - Internal speed up to 1 GHz
 - PCIe GEN5 (400 Gbits/s)
 - 32 links at 56G PAM4 or 48 links au 28G NRZ for FE interface
- No DDR memory
 - ➔ Use of PC memory instead
- 28G optics
- White Rabbit clock reception
- PON management
- Gain estimé vs PCIe 40 :
 - Processing : factor 8 to 12
 - Output bandwidth : factor 4



Développement

Projet de R&T PCIe400

Motivations

- Equipe réduite en fin de projet
- 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

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 IN2P3

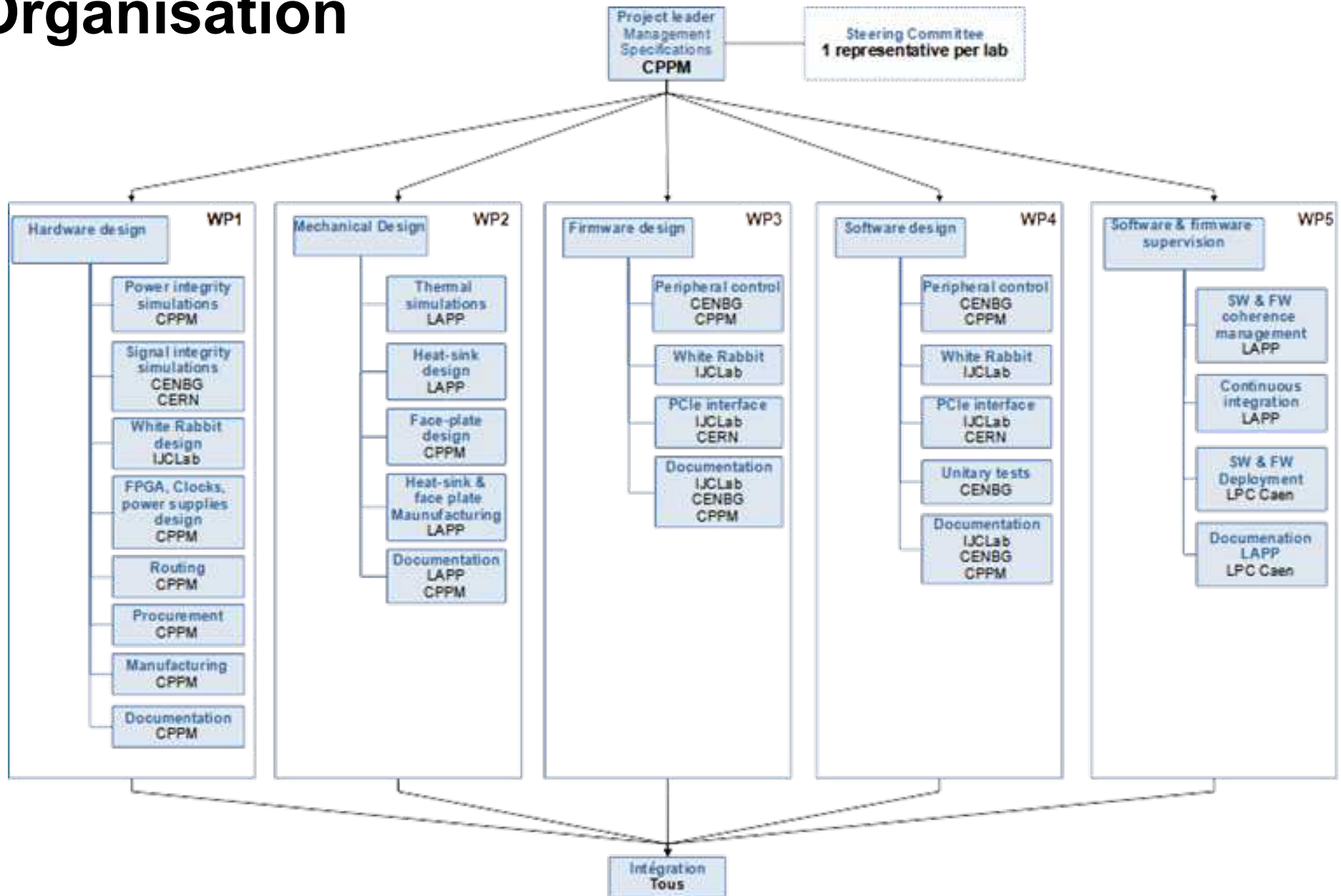
Nom des personnes	Statut	2022	2023	2024	Total (FTE)
CPPM		185%	100%	100%	3.85
Jean-Pierre CACHEMICHE	IRHC	60%	0%	0%	
Frédéric RETHORE	IR	25%	0%	0%	
Paul BIBRON	CDD IR	100%	100%	100%	
Kevin ARNAUD	IE	50%	30%	0%	
LAPP		30%	30%	15%	0.75
Guillaume VOUTERS	IR	15%	15%	15%	
Sebastien VILALTE	IR	5%	5%	0%	
Jean 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

Organisation



Avancement

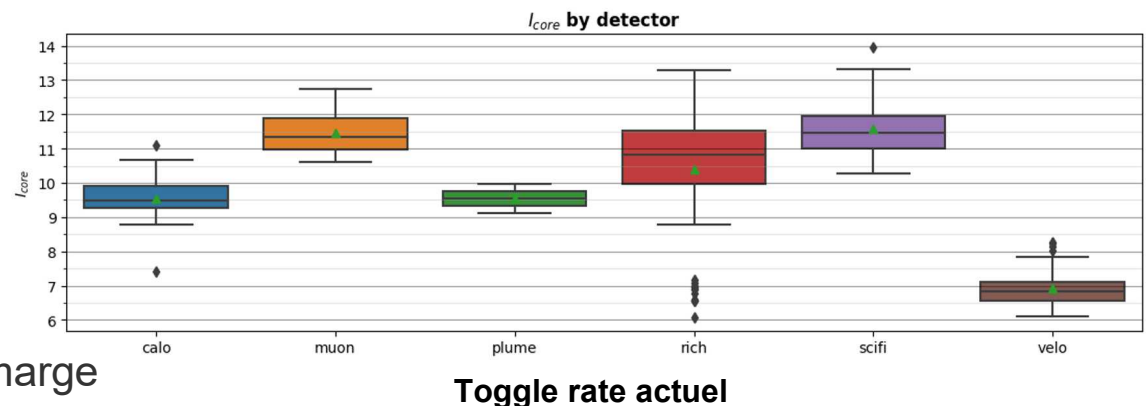
Dimensionnement alimentations

Méthode de simulation

- 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
 - Difficulté : nouveau FPGA non encore disponible
 - Outils non disponibles
 - ▷ Approximations avec un FPGA similaire

Consommation estimée pour le coeur du FPGA

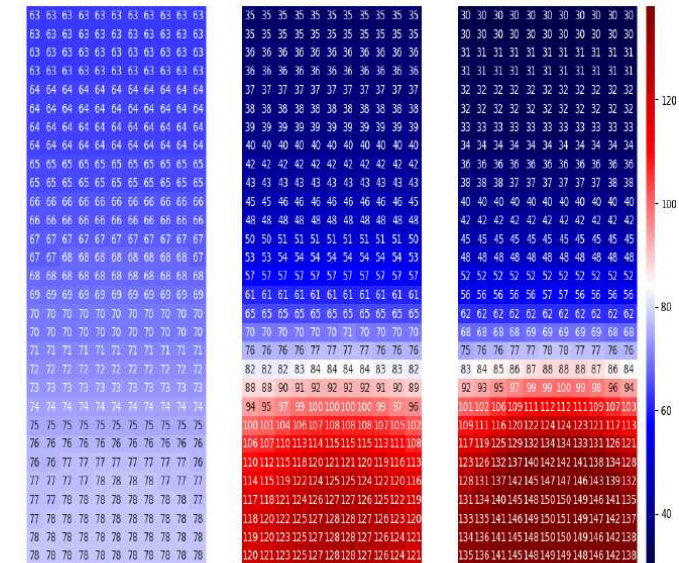
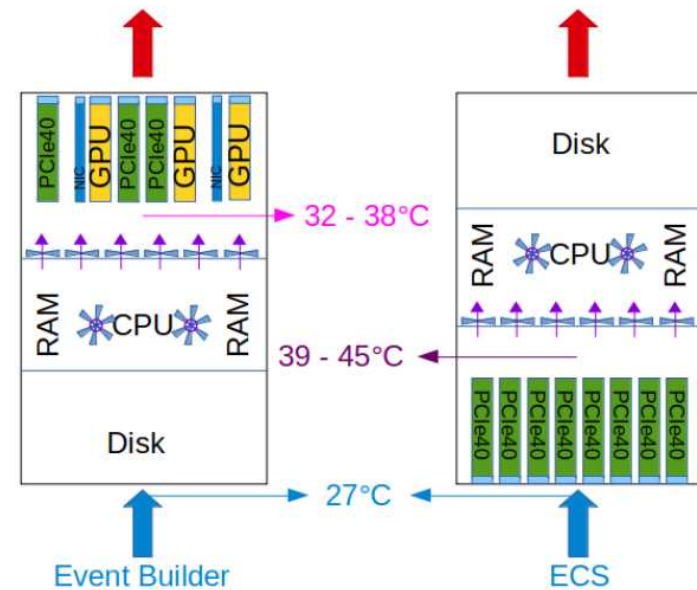
- Cas typique :
 - 85 W avec toggle rate de 12.5 %,
 - 70 % de taux d'occupation,
 - 45°C température de jonction
- Pas encore toutes les informations des détecteurs actuels
- Réflexion en cours pour établir la marge



Refroidissement

Premières simulations

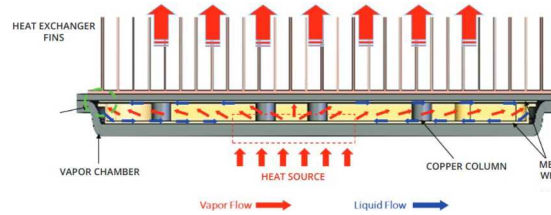
- Réalisées avec Heatscape
- Grande variabilité de conditions
 - Norme PCIe spécifique jusqu'à 60°C
 - Mesures faites au CERN plutôt de l'ordre de 45°C
 - Sens du flow
- Plusieurs types de radiateurs évalués
- Montrent qu'une technologie avec vapor chamber plus efficace



Cross cut



Zipper fin



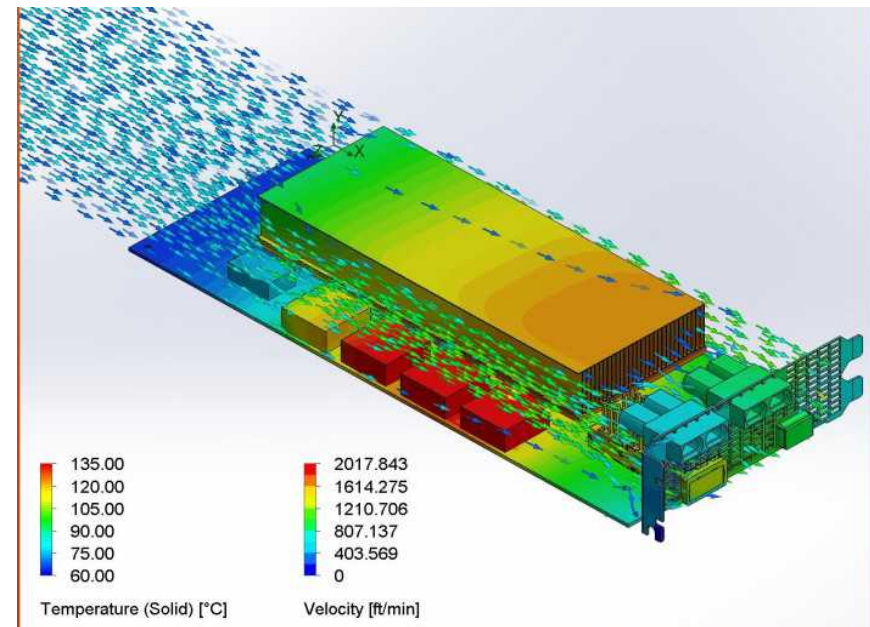
Vapor chamber

Simulation heatscape

Refroidissement

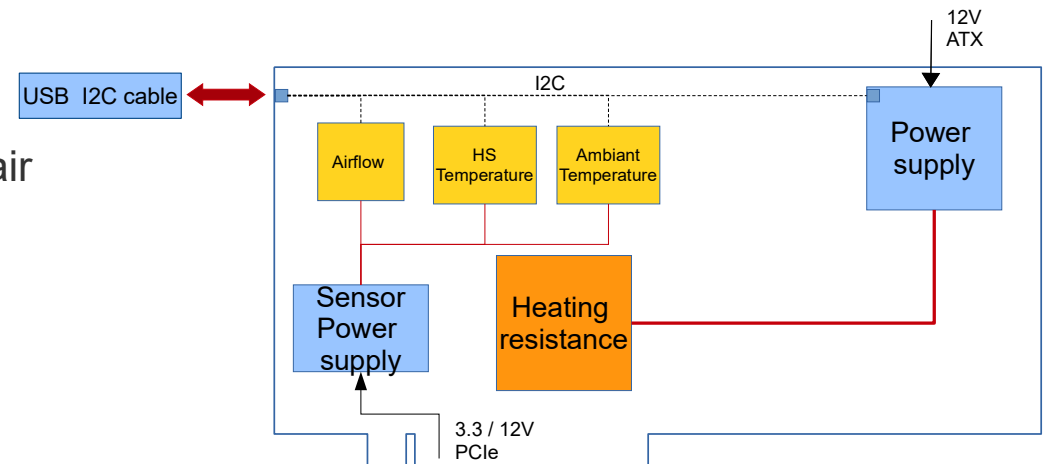
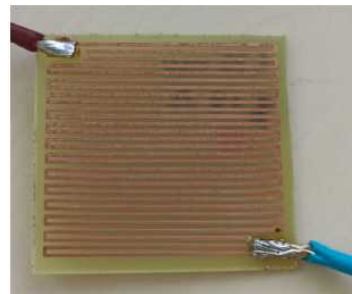
Afinage simulations avec COMSOL

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



Vérification des simulations avec maquette thermique

- Design d'une carte instrumentée en température et mesure de vitesse de l'air
- Permet de
 - vérifier la simulation
 - d'essayer un radiateur
- Emulation du FPGA par un serpentin



Optique

Plusieurs solutions envisagées

- Cages en face avant éliminées pour ne pas obturer la ventilation
- 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
 - Supports fournis

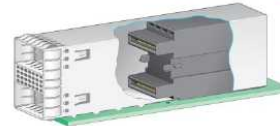
Samtec FireFly ECUO/ECUE

- 70m OM3 850nm
- 12 simplex ou 4 duplex
- Ribbon MPO-12 or MPO-24
- 3.3V (+1.8V if duplex 28G)
- 14 / 25 / 28 Gb/s NRZ
- 240mm² footprint
- 950€ T12 | 700€ R12 (25G)
- 300€ T12 | 200€ R12 | B04 300€ (14G) | B04 620€ (28G)



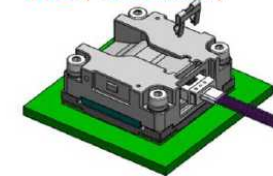
QSFP-DD (Amphenol 400GBASE-SR8)

- 70m OM3 850nm
- 8 duplex
- MPO-16
- 10W (single 3.3V)
- 53.125Gb/s PAM4 (lower rate NRZ possible?)
- CMIS slow control interface
- 600€ (↘↘ strong demand from data center)



II-VI/Finisar BOA

- 70m OM4 850nm (70m OM3 on demand)
- 12 duplex
- MT-24 F connector (no ribbon)
- 6W (2.5V + 3.3V rails)
- 28.1Gb/s NRZ
- 625mm²
- CFP MSA slow control interface
- 600€ (+150€ ribbon)



QSFP12 (II-VI/Finisar 400GBASE-SR4)

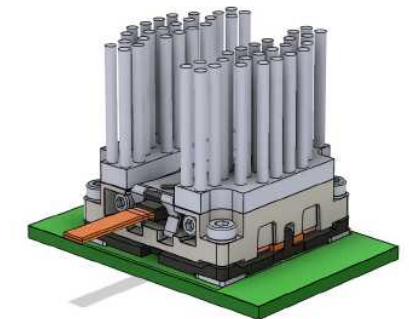
- 100m OM4
- 4 duplex
- MPO-12
- 106.25Gb/s PAM4
- CMIS slow control interface
- ?? € (demand from data center ?)



Optics choice

Amphenol OBT (10124588-314)

- 12 duplex 1.25G to 26.3G NRZ (CDR bypassable)
- MT ferrule x24
- 100m OM4 850nm
- Inbuilt PRBS-9 generator and checker
- ~6W power consumption
- Double 3.3V power supply (Tx and Rx separated)
- Use of socket, no through hole required
- Heat sink provided



Status

Tous les principaux choix technologiques finalisés

Schématique en cours

- Basée uniquement sur bibliothèque IN2P3 pour faciliter les échanges

Difficultés actuelles

- Attente de la version Quartus permettant de générer le FPGA
 - ➔ Devrait arriver dans quelques jours
- 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