PCIe400 :
Architecture proposée et challenges technologiques

PCIe400
DAQ Engine

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

J.-P. Cachemiche (CPPM)
Paul Bibron, Julien Languouët, Renaud Le Gac, Frédéric Réthoré

Plan

Introduction

Besoins LHCb

- La carte est en fait un concentrateur de données compatible du LpGBT et des futurs sérialiseurs du CERN avec interfaces réseaux très haut débit utilisés dans les Data Centers
- Premier étage de l'event building : concentration 48 → 1
- Développement stratégique : charnière entre le monde "custom" et le monde des réseaux commerciaux standards utilisés dans les data centers

Autres besoins

- Accent mis sur la généricité pour utilisation dans de multiples contextes

Caractéristiques envisagées (1/2)

Liens d'entrée

- Débit minimal : 400 Gbits/s (PCIe40 x 4)
- Plus si possible
  - Intéressant pour un détecteur de physique dans lequel la bande passante n'est pas répartie uniformément
- Principale différentiation avec cartes du commerce relativement pauvres en nombre de liens

Liens en sortie

- PCIe Gen5 (400 Gbits/s)
  - Dans ce cas aucune mémoire sur la carte : les données sont poussées par DMA dans la mémoire du serveur.
  - Ce dernier envoit les fragments de données dans le réseau
- Optionnellement Ethernet 400G
  - Expédition directe vers le réseau
  - Permet de mettre plus de cartes dans un même serveur
  - Requiert de la mémoire sur la carte

Caractéristiques envisagées (2/2)

Vitesse des liens : jusqu'à 56 Gbits/s PAM4

- Pas vraiment requis pour le LpGBT (10 Gbits/s) mais intéressant pour son éventuel successeur
  - Prépare la génération d'après
- En phase avec les caractéristiques communément trouvées sur les FPGAs de dernière génération

Gestion de la distribution du temps

- Compatibilité avec le prochain standard utilisé par le CERN : White Rabbit

Choix technologiques préliminaires

Spécificités des nouveaux FPGAs

Les liens sériels ne sont plus sur le silicium de la matrice

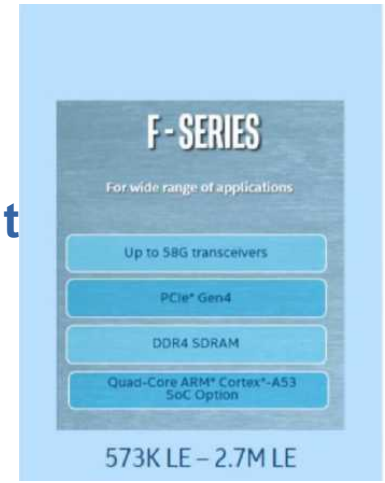
- Notion de Tiles

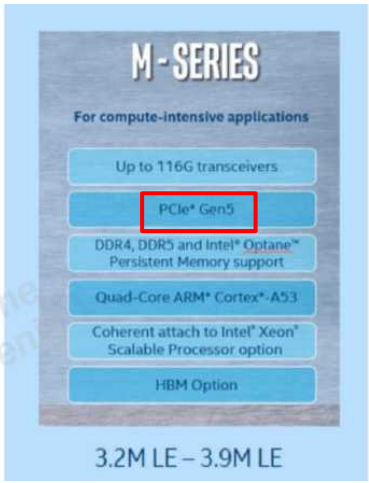
- Flexibilité pour décliner le chip

Choix du FPGA (1/3)

L'utilisation du PCIe Gen5 restreint le choix

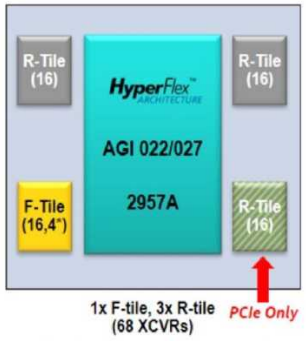
- Pas disponible sur Stratix 10
- Sur l'Agilex, seulement disponible dans les séries I or M

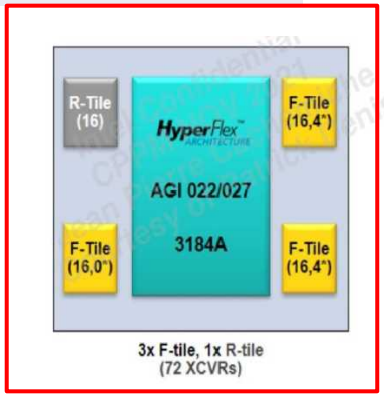


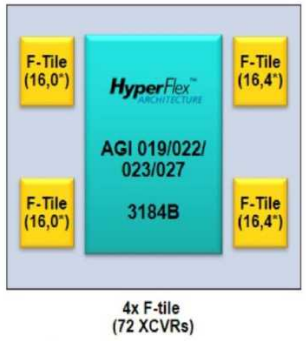


- Dans la série I ou M, seulement disponible dans les R-Tiles

- Dans la série I avec **4 tiles**, la version la plus intéressante est celle-ci :
  1 R-tile, 3 F-tiles







Choix du FPGA (2/3)

Sortie récente de nouvelles versions avec Plus de liens série et plus de cellules logiques

- Jusqu'à 120 XCVR et jusqu'à 3.9 MLE
- Cependant pas utilisables car seulement équipé de F-Tiles
  - Les F-Tiles ne gèrent pas le PCIe Gen5

4x F-Tile
(120 XCVRs)

Meilleur choix jusqu'à présent

- PCIe Gen5 x 16
- 48 liens jusqu'à 32G NRZ
- or 32 liens jusqu'à 56G PAM4

3x F-tile, 1x R-tile
(72 XCVRs)

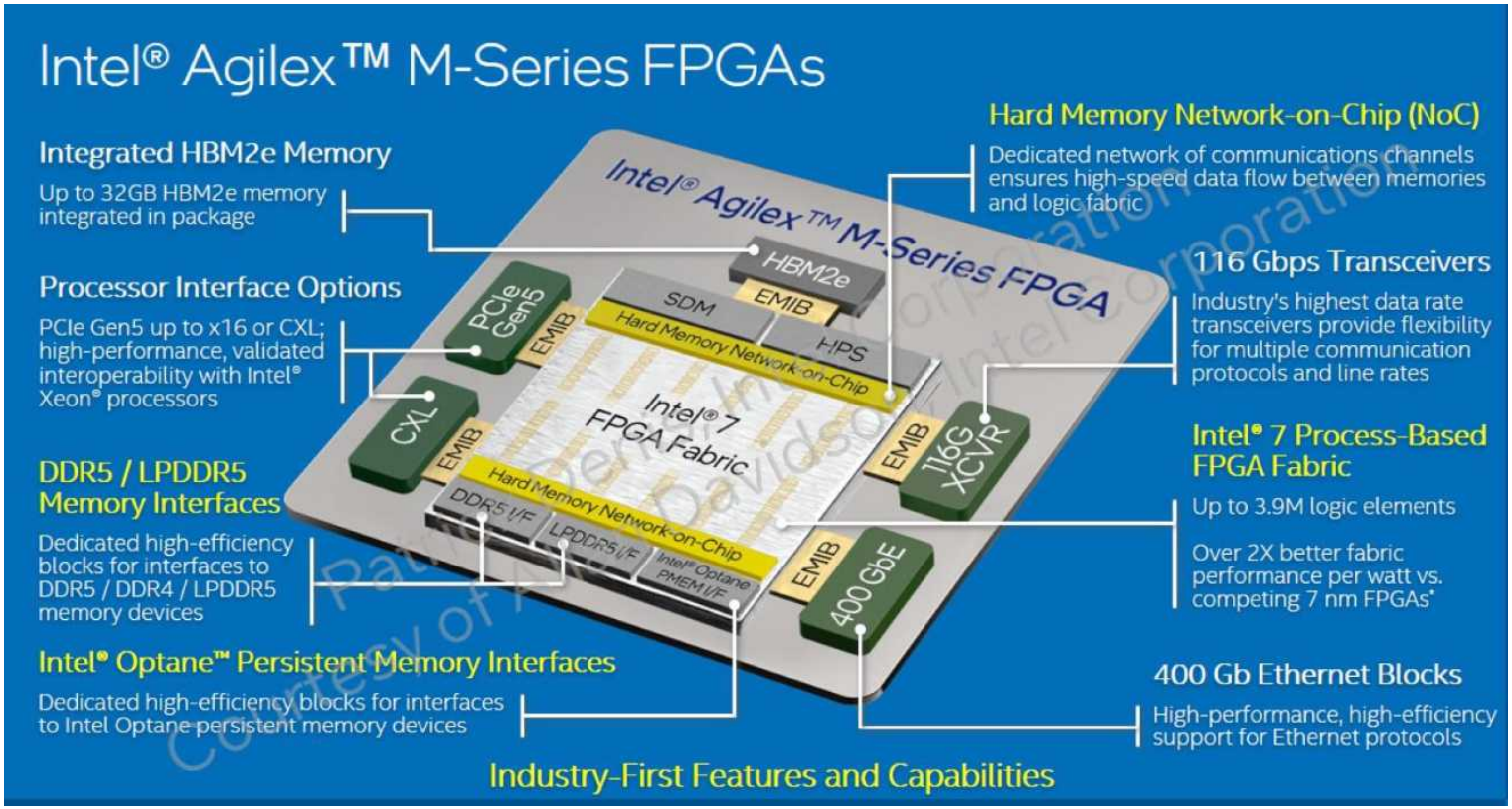
- Cependant, modification des dates de sortie chez Intel : **retard de 6 mois** (Q2'23)
  - Pas un problème à l'échelle du projet
  - Mais compliqué par rapport aux dates de retraite de 2 ingénieurs hardware (Avril and Sept 2022)
- Pinout pas disponible avant **Q4'22**

1st Device	Pkg	ES	PRQ
AGI 027	3184A	n/a	Q2'23

Choix du FPGA (3/3)

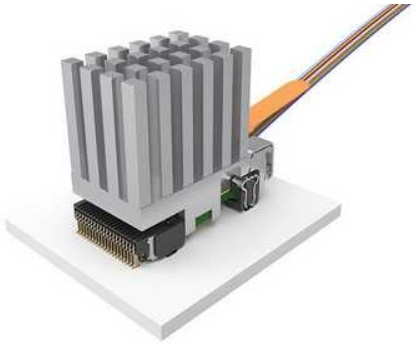
Choix final

- Intel propose en avance de phase un **circuit en série M** avec 4 MLE + 32G HBM Au même prix !
- Disponible en **Q3'22** (9 mois plus tôt que le précédent)
- Mais seulement supporté par Quartus en **Q2'22**



Liens optiques (1/2)

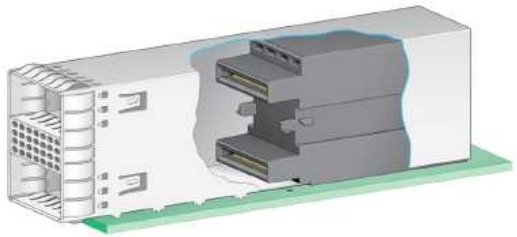
Principalement 3 solutions



- Samtec firefly
  - Performance jusqu'à **28 Gbps**
  - Modules x4 duplex and x12 sir
  - Coûts ~600 € x4 duplex 25G
    ~500 € x12 simplex 25G
    ~200 € x12 simplex 14G

- Finisar BOA
  - Disponibles en modules 12-voies **full-duplex**
  - Portée maximum 70m à **25 Gb/s** sur OM4 MMF
  - Débit adaptable de 1 Gb/s up à 28.1 Gb/s par voie
  - Coût ~700 €

- Finisar 400G SR8 QSFP-DD (de nombreuses autres sources)
  - Disponible en modules 8-voies **full-duplex**
  - Portée 70m sur OM3 MMF ou 100m sur OM4 MMF
  - **8x50G** PAM4 retimed 400GAUI-8 electrical interface
  - Coût ~ 600 $
    - Prix qui va probablement chuter car très utilisé dans les data centers
- La plupart des QSFP-DD ne peuvent pas fonctionner en 10G NRZ sauf le chip Amphenol
  - à vérifier



Liens optiques (2/2)

Circuits 56G

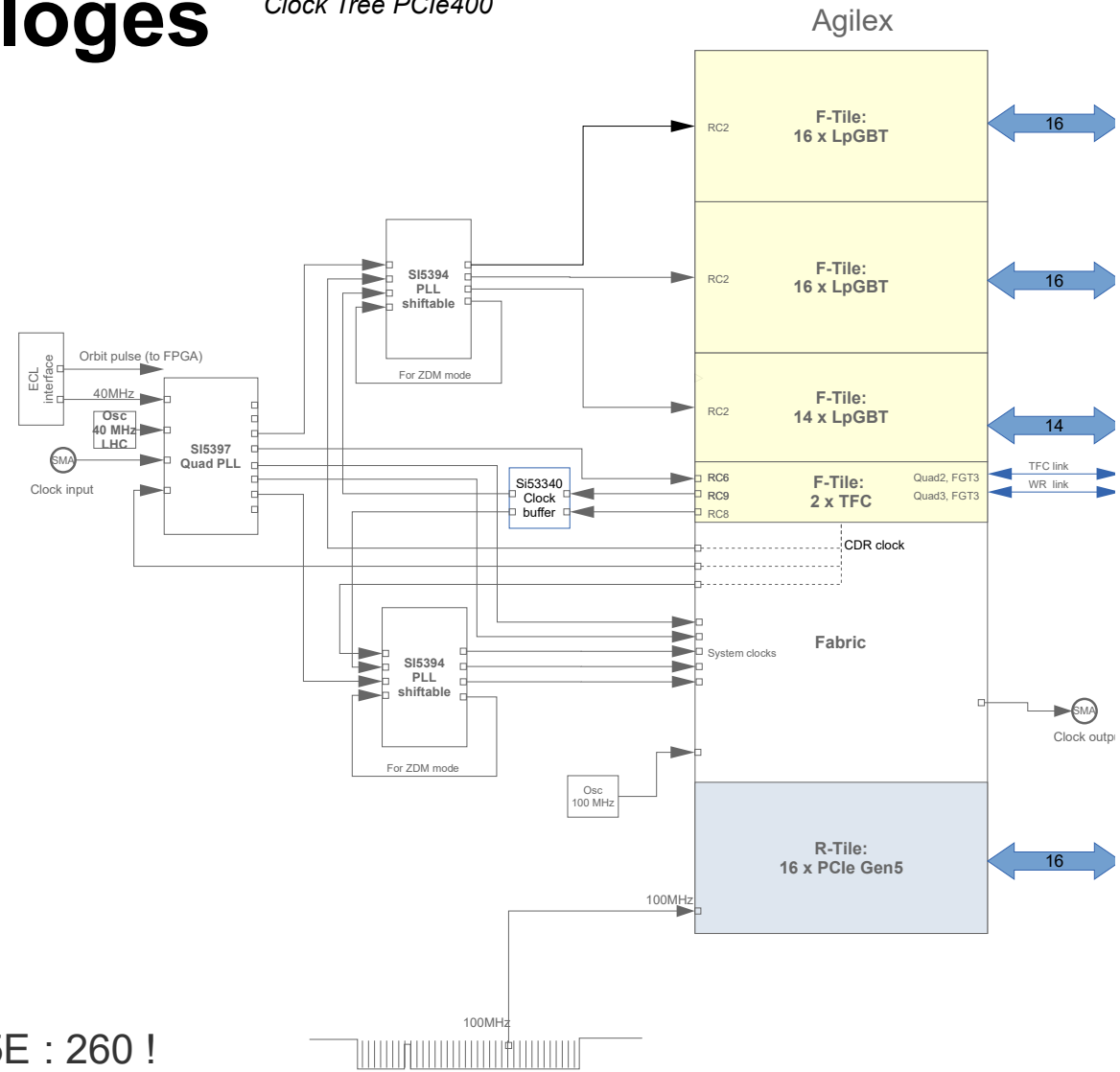
- Souhaité pour tester complètement le FPGA et être (peut-être) compatible de la prochaine version de concentrateur sériel développé par le CERN CERN concentrator
- But not required in the medium term

Choix pas complètement finalisé

- Circuits bidirectionnels Samtec
  - Assez onéreux en $ par lien
- Circuits unidirectionnels Samtec
  - Solution similaire avec ce que nous aviopns avec les minipods
  - Attente d'information sur les circuits futurs en version 56G
- BOAs
  - Impossibilité de sous-équiper la carte avec par exemple des Rx only, mais surcoût relativement faible
  - Couplage RX/TX sur MPO 24 → Splitters entre la carte et le patch pannel
  - Mais configuration uniforme
- QSFP-DD
  - Encombrement, impact sur refroidissement
  - Faisabilité à démontrer

PLLs et arbre d'horloges

Clock Tree PCIe400



Défini avec le CERN

- Autorise un maximum de flexibilité

Difficulté d'approvisionnement des composants

- Minimum quantities for Si5395E : 260 !
  - Cost for a prototype : ~7000 €

- Finalement trouvé une version avec un peu plus de jitter chez Mouser, mais 52 semaines de délai

Quelques règles générales

Restriction du nombre de composants différents

- Ex : Fonctions annexes
  - Power sequencer
  - USB Blaster
  - Monitoring température et sécurité
  - Contrôle de la flash de programmation
  - White Rabbit
    - Un seul composant pour chaque instance (Max10)
  - → Evite de multiplier les outils de développement

- Une seule marque de composants pour les DC/DC ou bien les PLLs
  - → Facilite la programmation

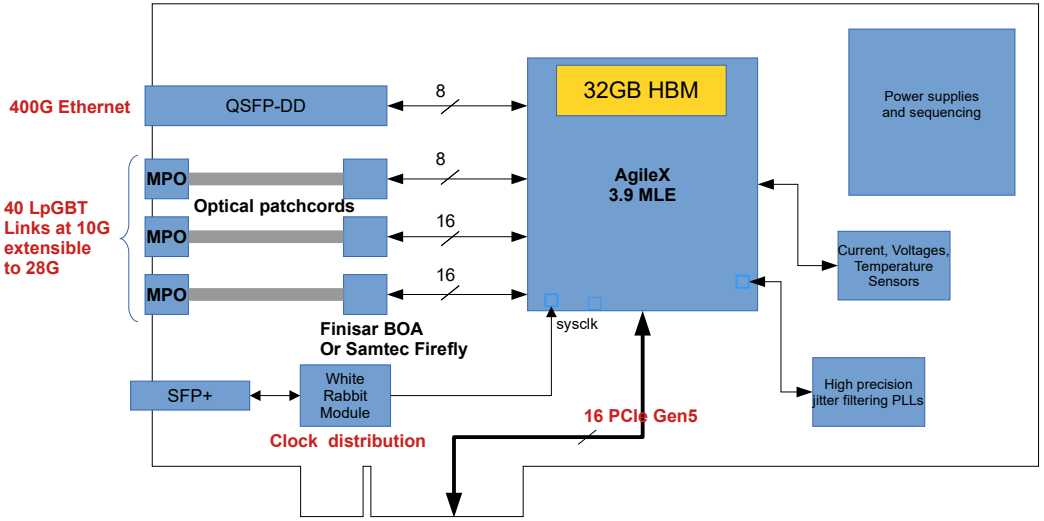
Au final : gain de temps pour le développement

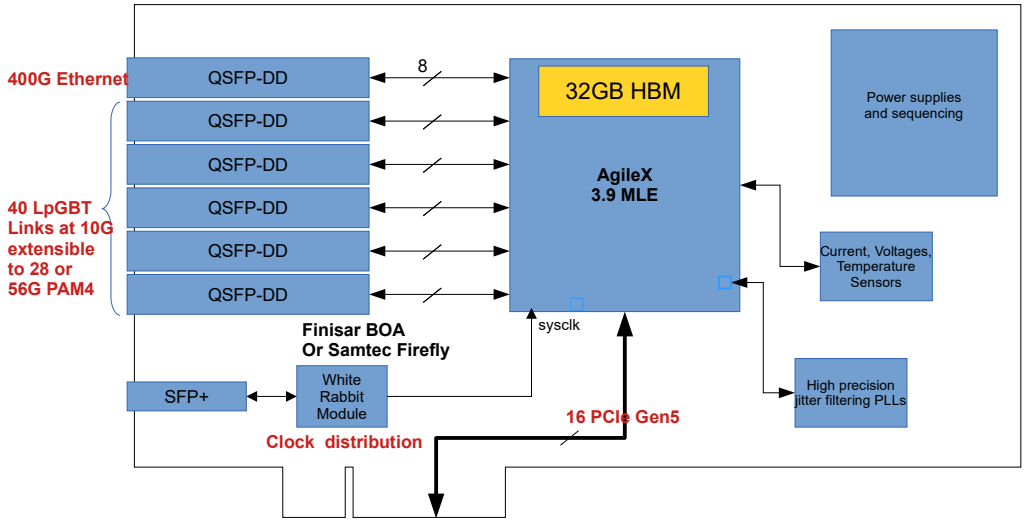
- Moins d'outils de programmation hétérogènes
- Moins d'outils à s'approprier

Architectures possibles

Caractéristiques envisagées

- Agilex AGMF039R47A1E1V
  - 3.9 MLE
  - Mémoire 32 GB HBM
  - Fréquence interne : jusqu'à 1 GHz
  - PCIe GEN5 (400 Gbits/s)
  - 32 liens 56G PAM4
    ou 48 links en 28G NRZ
    pour interfaces FE
- Pas de mémoire DDR
  - Utilisation de la mémoire du PC
- 28G or 56G optics
- White Rabbit clock reception
- Gain par rapport à la carte PCIe 40 :
  - Processing : facteur 8 to 12
  - Bande passante en entrée : facteur 2.8 to
  - Bande passante en sortie : facteur 4
  - 32 Gb de mémoire intégrée sur le chip





Conclusion

Conception en cours depuis Avril 2021

Encore quelques choix technologiques à fixer

- Optique, technologie, impact sur le refroidissement
- Nombre de lien rellement utilisables
- Utilisabilité de l'Ethernet 400G avec une mémoire de 32 GB

Difficulté d'approvisionnement anticipés par commandes des principaux composants critiques