

R&T PCIe400 : Choix Optique



3 mars 2022

Julien Langouët, CPPM



Sommaire

Architecture Transceiver Agilex / F-tile

Comparaison transceivers opto-electroniques

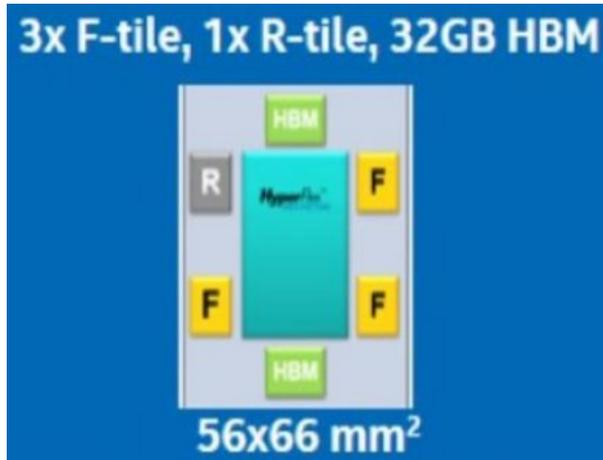
- QSFP-DD
- QSFP112
- SAMTEC FireFly
- Finisar BOA

Propositions de layout sur exemple d'une PCIe AIC full height, double slot, 3/4 length

- QSFP-DD
- QSFP112 + FireFly
- QSFP112 + BOA

Conclusion et perspectives

Architecture XCVR Agilex / F-tile

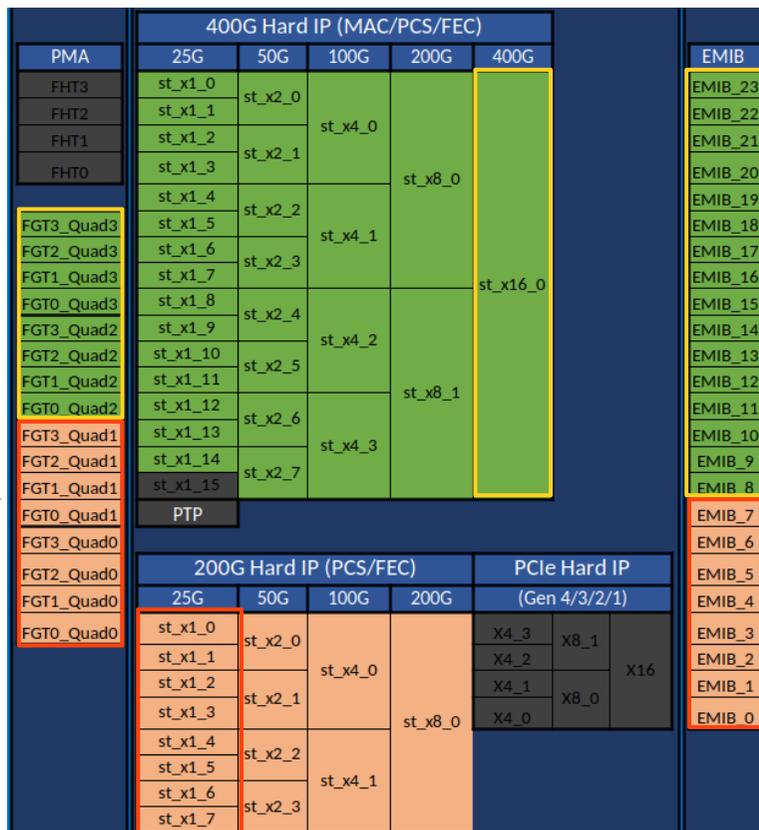


AGM039

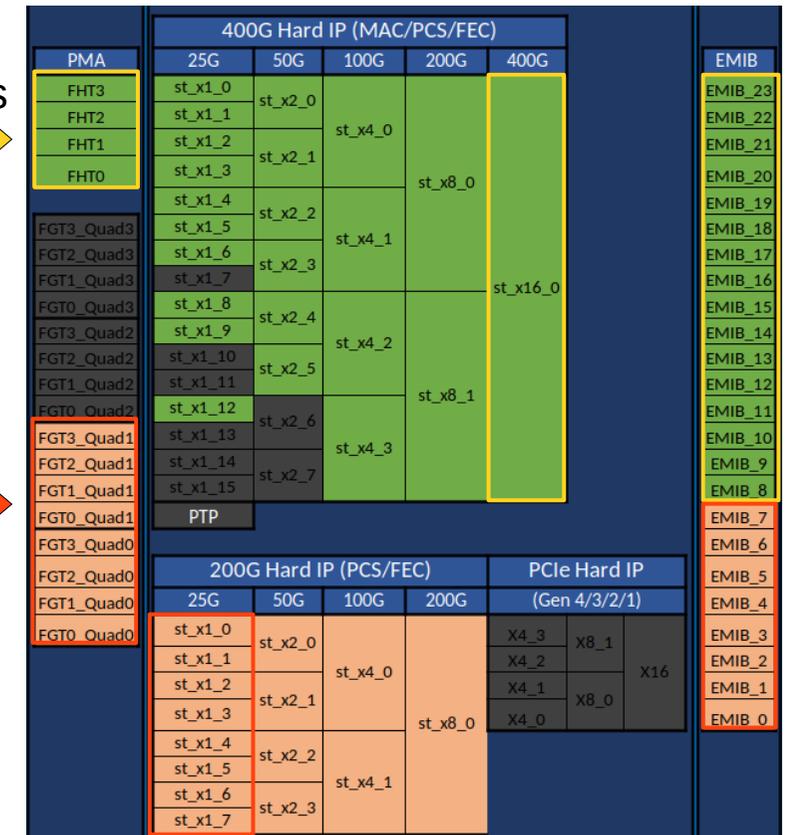
- R-tile -> dédiée au PCIe Gen 5 x16
- F-tile -> 400GbE / general purpose
 - 4x4 FGT PMA duplex : 1 à 32 Gb/s NRZ ou 20 à 58 Gb/s PAM4
 - 4 FHT PMA duplex : 24 à 58 Gb/s NRZ ou 48 à 116Gb/s PAM4
 - 24 EMIB (Embedded Multi-Die Interconnect Bridge) XCVR/Core bridge

Objectif : Avoir le maximum de PMA pour FE et intégrer optionnellement le 400GbE pour préparer le futur

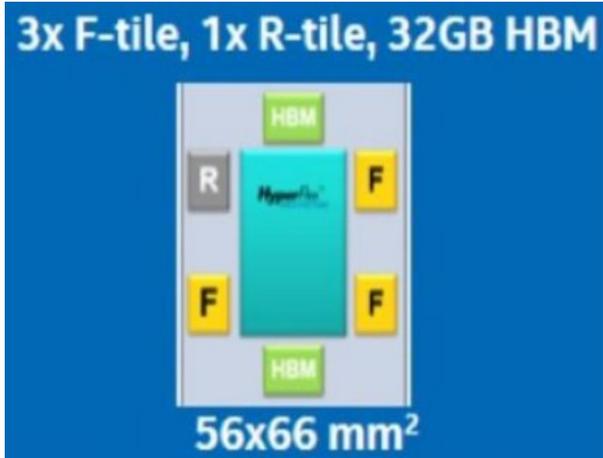
400GbE-8 + 8x 1 à 32Gb/s



400GbE-4 + 8x 1 à 32Gb/s



Architecture XCVR Agilex / F-tile



- R-tile -> dédié au PCIe Gen 5
- **F-tile** -> 400GbE / general purpose
 - **4x4 FGT PMA** : 1 à 32 Gb/s NRZ ou 20 à 58 Gb/s PAM4
 - **4 FHT PMA** : 24 à 58 Gb/s NRZ ou **48 à 116Gb/s PAM4**
 - 24 EMIB (Embedded Multi-Die Interconnect Bridge)
XCVR/Core bridge

16 x 1 à 32Gb/s

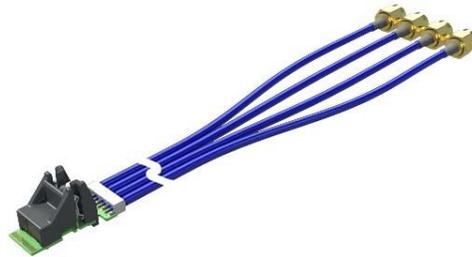
16 x up to 32Gb/s NRZ

PMA	400G Hard IP (MAC/PCS/FEC)					EMIB	
	25G	50G	100G	200G	400G		
FHT3	st_x1_0	st_x2_0	st_x4_0	st_x8_0	st_x16_0	EMIB_23	
FHT2	st_x1_1	st_x2_1				st_x4_1	EMIB_22
FHT1	st_x1_2		st_x2_2				st_x4_2
FHT0	st_x1_3	st_x2_3				st_x4_3	
FGT3_Quad3	st_x1_4		st_x2_4	st_x4_2			EMIB_19
FGT2_Quad3	st_x1_5	st_x2_5				st_x4_1	EMIB_18
FGT1_Quad3	st_x1_6		st_x2_6	st_x4_3			EMIB_17
FGT0_Quad3	st_x1_7	st_x2_7				st_x4_0	EMIB_16
FGT3_Quad2	st_x1_8		st_x2_4	st_x4_2			EMIB_15
FGT2_Quad2	st_x1_9	st_x2_5				st_x4_1	EMIB_14
FGT1_Quad2	st_x1_10		st_x2_6	st_x4_3			EMIB_13
FGT0_Quad2	st_x1_11	st_x2_7				st_x4_0	EMIB_12
FGT3_Quad1	st_x1_12		st_x2_0	st_x4_1			EMIB_11
FGT2_Quad1	st_x1_13	st_x2_1				st_x4_2	EMIB_10
FGT1_Quad1	st_x1_14		st_x2_2	st_x4_3			EMIB_9
FGT0_Quad1	st_x1_15	st_x2_3				st_x4_0	EMIB_8
FGT3_Quad0	PTP		st_x2_0	st_x4_1	EMIB_7		
FGT2_Quad0	200G Hard IP (PCS/FEC)	st_x2_1			st_x4_2	EMIB_6	
FGT1_Quad0			200G Hard IP (PCS/FEC)	st_x2_2		st_x4_3	EMIB_5
FGT0_Quad0	200G Hard IP (PCS/FEC)	st_x2_3			st_x4_0		EMIB_4
PCIe Hard IP (Gen 4/3/2/1)			200G Hard IP (PCS/FEC)	st_x2_0		st_x4_1	X4_3
	X4_2	X8_0			EMIB_2		
	X4_1	X8_0			EMIB_1		
	X4_0				EMIB_0		

Comparaison XCVR opto-elec

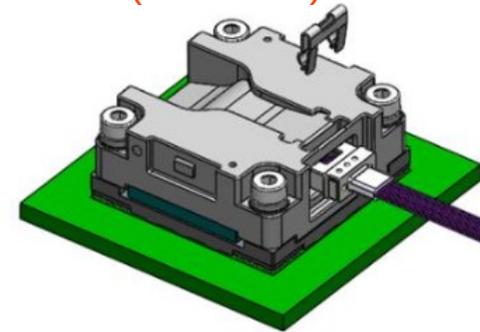
Samtec FireFly ECUO

- 70m OM3 850nm
- 12 simplex ou 4 duplex
- *Ribbon* MPO-12 ou MPO-24
- 3.3V (+1.8V si duplex 28G)
- 14 / 25 / 28 Gb/s
- 240mm² footprint
- **300€ T12 | 200€ R12 | 300€ B04 (14G)**
- **600€ B04 (28G)**



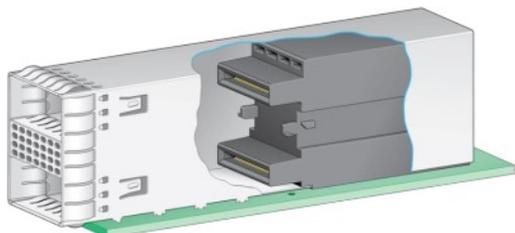
II-VI/Finisar BOA

- 70m OM4 850nm (possible 70m OM3)
- 12 duplex
- MT-24 connector (no ribbon)
- 6W (2.5V + 3.3V rails)
- 25Gb/s
- 625mm²
- CFP MSA slow control interface
- **600€ (no ribbon)**



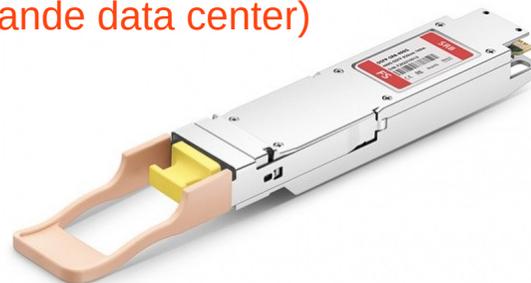
QSFP-DD (Amphenol 400GBASE-SR8)

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



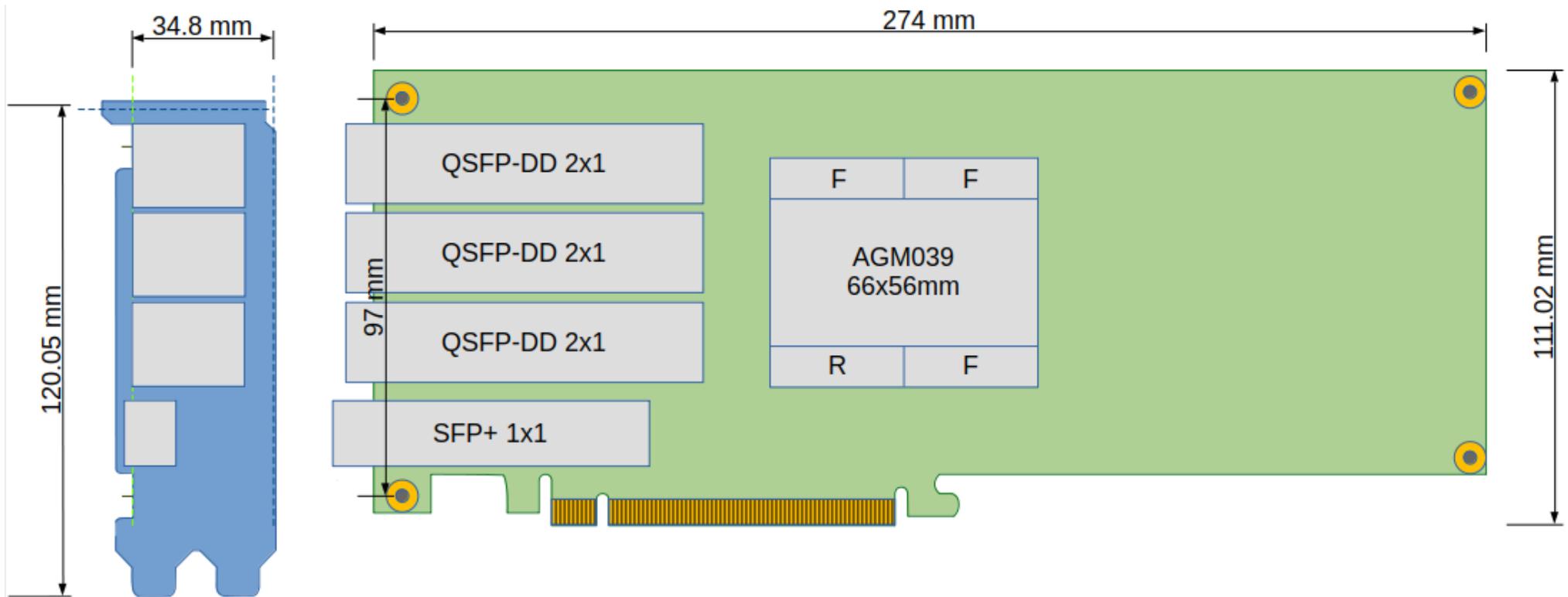
QSFP112 (II-VI/Finisar 400GBASE-SR4)

- 100m OM4
- 4 duplex
- MPO-12
- 106.5Gb/s
- CMIS slow control interface
- **?? € (demande data center ?)**



Proposition 1 : QSFP-DD

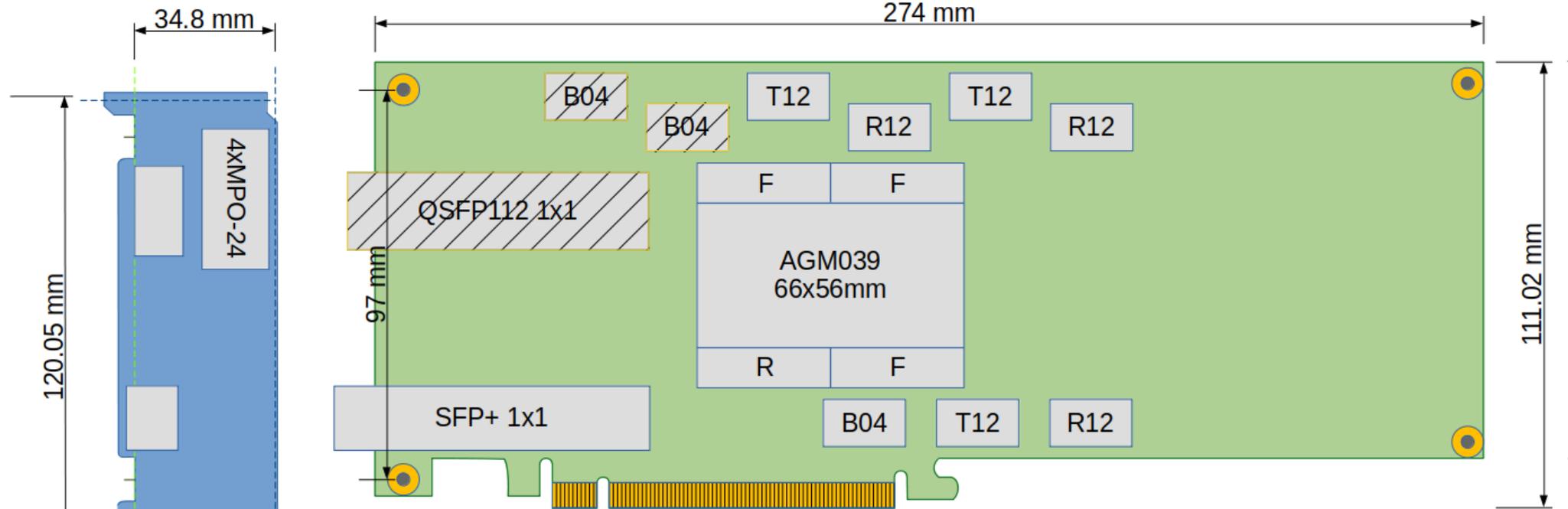
[400GbE + 40@53Gb/s PAM4 duplex] ou [48@53Gb/s PAM duplex FE]



- Technologie** → Incertitude QSFP-DD compatibles NRZ avec débit custom ? 10.24Gb/s ou 2.5Gb/s pour VL+
- En accord avec le marché des data center, éventuellement multitude de fournisseurs : Économie d'échelle sur les QSFP-DD
- Mécanique** → Forte contrainte de refroidissement. Solution : cheminée pour diriger le flux d'air ? Refroidissement liquide ?
- Face avant non conforme avec PCIe SIG
- Routage** → Complexité de routage : 16 paires diff. À 25Gbd PAM4 sur ~35mm + 80 paires diff. À 10Gbd NRZ sur ~75mm
- Test des paires diff. cuivrées : QSFP-DD -> SMA breakout board ?
- Optique** → Requier patch panel relativement complexe (séparation des Tx et Rx sur chaque MPO-16)
- Homogénéité de l'interface optique 6x MPO-16

Proposition 2 : QSFP112 + FireFly

[400GbE + 40@28Gb/s NRZ duplex] ou [48@28Gb/s NRZ duplex]



Technologie

- Incertitude de l'arrivée sur le marché QSFP112, coût ?
- Dépendance avec SAMTEC, couvert par le contrat du CERN pour l'exploitation des VL+ ?

Mécanique

- Refroidissement à air faisable a priori
- Possible de sous-équiper la carte. Par ex : pour des configuration 48Rx/0Tx

Routage

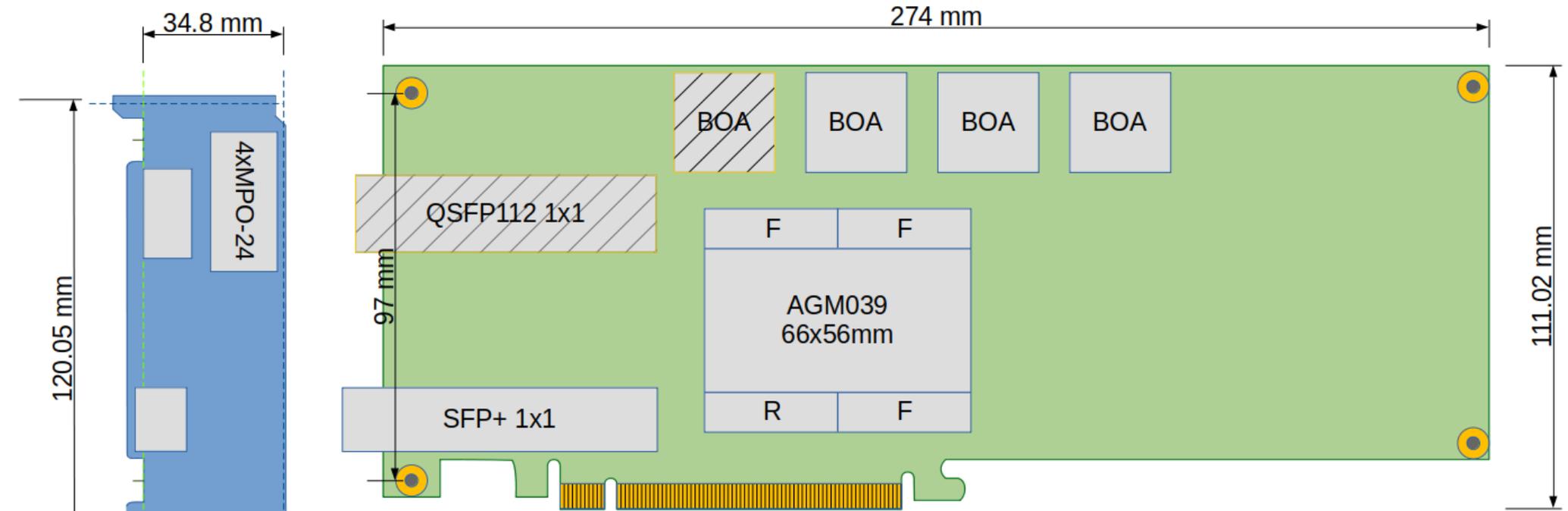
- Simplicité de routage : modules on-board au plus près des XCVR FPGA
- Test des paires diff. Cuivrées : Utilisation de FireFly ECUE vers SMA ?

Optique

- Bretelle optique non remplaçable (FireFly + bretelle optique solidaires)
- Inhomogénéité de l'interface optique (modules duplex et simplex)

Proposition 3 : QSFP112 + BOA

[400GbE + 40@25Gb/s NRZ duplex] ou [48@25Gb/s NRZ duplex FE]



Technologie

- Incertitude de l'arrivée sur le marché QSFP112, coût ?
- Dépendance avec II-VI/Finisar BOA

Mécanique

- Refroidissement à air faisable a priori
- Impossible de sous-équiper la carte. Par ex : pour des configuration 48Rx/0Tx

Routage

- Complexité de routage : modules on-board proches XCVR FPGA
- Test des paires diff. Cuivrées ?

Optique

- Homogénéité de l'interface optique : bretelle custom MT-24 -> MPO-12 ou MPO-24

Conclusion et perspectives

Pas de solution ultime. Les critères mécaniques (thermique) et de routage sont cruciaux

Quels sont nos moyens pour la simulation ?

- Thermique
- Intégrité de signal : établir un plan de travail pour éclairer le choix de l'optique ?

Discussions avec fournisseurs

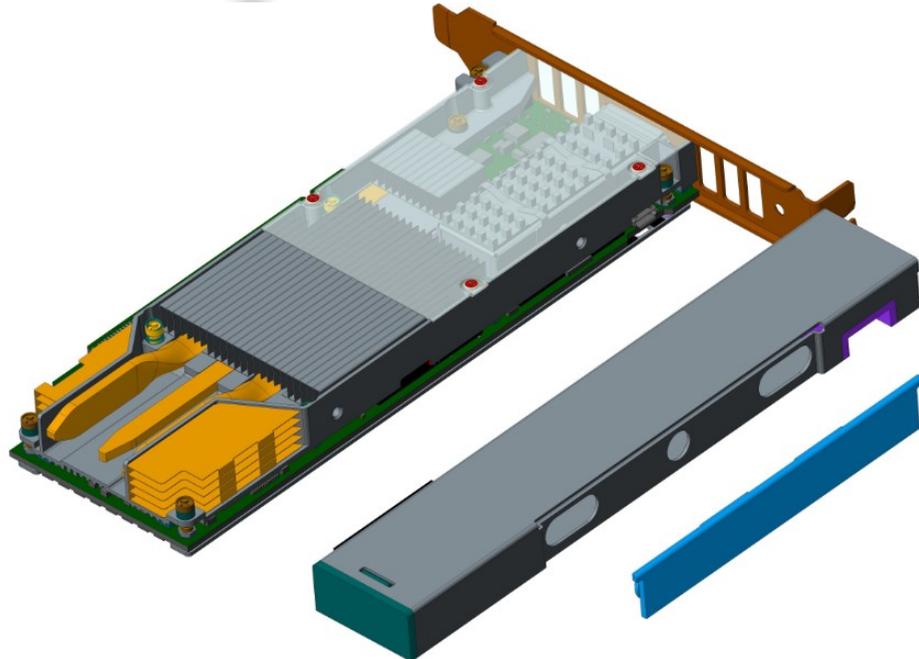
- Discussion prévue avec **SAMTEC** (demain)
 - Roadmap
 - Coût de la solution envisagée
 - Testabilité des paires diff. Cuivrées
- Discussion à prévoir avec **Sylex** pour solution bretelles optiques avec BOA
 - Coût des bretelles
- Discussion à prévoir avec **Amphenol/Finisar**
 - Roadmap
 - Relancer la discussion sur la compatibilité des QSFP-DD à débit réduit (bypass du CDR)

Back -up

Refroidissement



Exemple de cheminée, « air duct »
Intel acceleration card Intel Arria 10



Exemple bretelle optique pour solution BOA

