Generic readout board : PCIe400

European edition of the International Workshop on CEPC
9th April 2024 – Marseille

Julien Langouët *(CPPM) on behalf of the R&T PCIe400 team*
CPPM, IJClab, LP2IB, LAPP, LPCC, LHCb Online, Subatech

Outline

Context

PCIe400 overview

Hardware technical challenges

- Cooling the board
- Power distribution
- Signal integrity
- Routing

Synthesis

LHCb data processing system

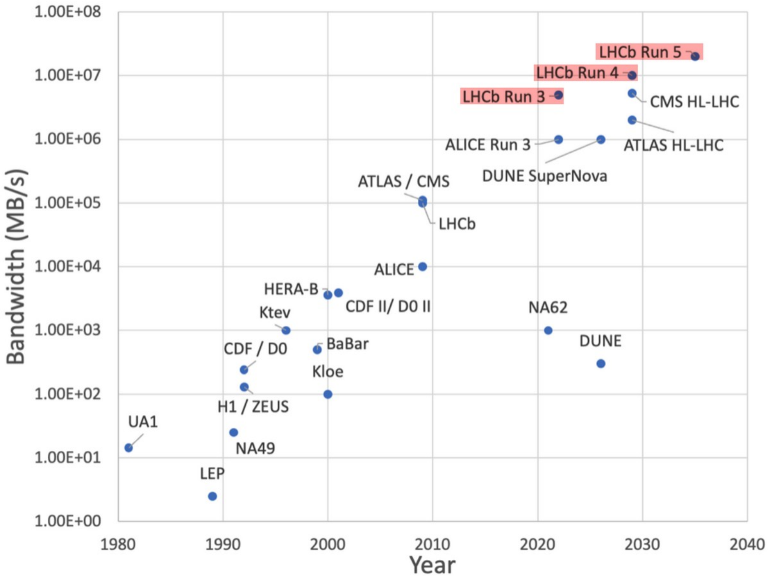
With increasing luminosity classical trigger strategies show low efficiencies especially for LHCb physics studies

- Impossible to use a subset of detector information to select interesting event for offline analysis
- All data must be readout, reconstructed and selected at LHC full collision rate

LHCb data processing is one of the biggest challenge in HEP

Run5 / Upgrade II requirements vs Run 3 / Upgrade I

- **4x number of front-end links**
  11k → 40k
- **2x bandwidth** of front-end serial links
  5 Gbps → 10 Gbps
- **5x data throughput**
  40 Tbps → 200 Tbps



Data processing rates history in HEP experiments A. Cerri

Requirement of data acquisition system

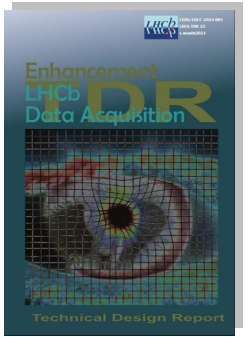
Common generic readout DAQ card interfacing custom protocol from front-end to commercial protocol back-end system

Intermediate enhancement in LS3 [2026 – 2028] of some sub-detectors

Opportunity to develop a new generation of board

- Generic readout DAQ card interfacing up to 48 front-end links to 1 commercial protocol link PCIe Gen5 with a **bandwidth x4**
- **Explore** experimental path to prepare Upgrade II
  - Integrate a **network interface** such as 400Gbps (RoCE) RDMA over Converged Ethernet in the FPGA
  - Integrate **complex data processing** such as tracks primitive reconstruction
- **Distribute LHC master clock** with tighter timing requirement : reproductive phase determinism **$\mathcal{O}$(10)ps RMS**

LHCb TDR 23



LHCb TDR 25

PCIe400 overview

Designed around latest Intel/altera FPGA Agilex 7 M-series

- 4 Million of logic elements and <1 GHz internal frequency

32GB integrated High Bandwidth Memory (HBM)

- Up to 5.2 Tbps exchange between Fabric and HBM

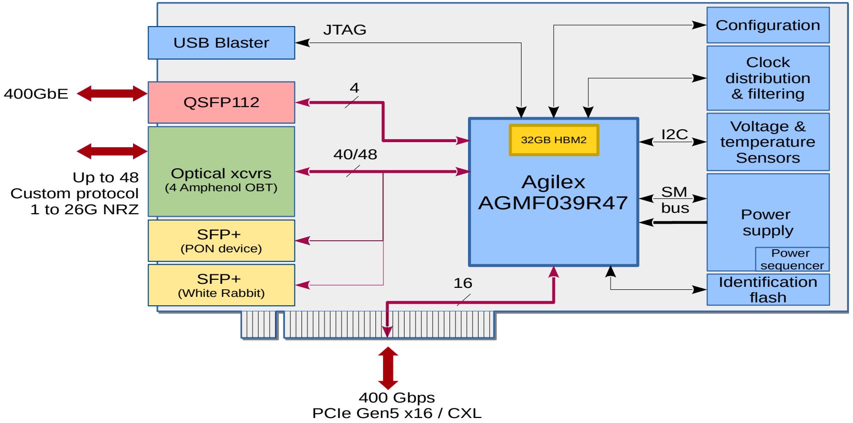
Hard processor inside Arm Cortex 4 cores @ 1.2Ghz

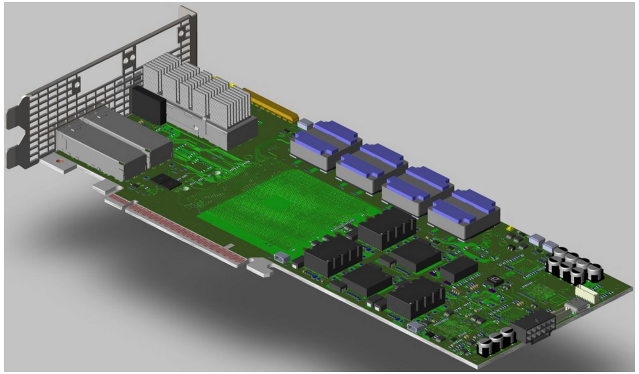
High bandwidth I/O

- 48 bidirectional links with front-end at up to 25Gbps
- 2 SFP+ for Time Fast Control system
- PCIe Gen 5 x16 with 400 Gbps output bandwidth - Compatible with Compute Express Link for cache coherent transactions
- 4x bidirectional 112 Gbps for network interface

Time distribution

- High precision PLL with <100 fs jitter intrinsic





3D rendered view of PCIe400

Cooling the board

Power dissipation

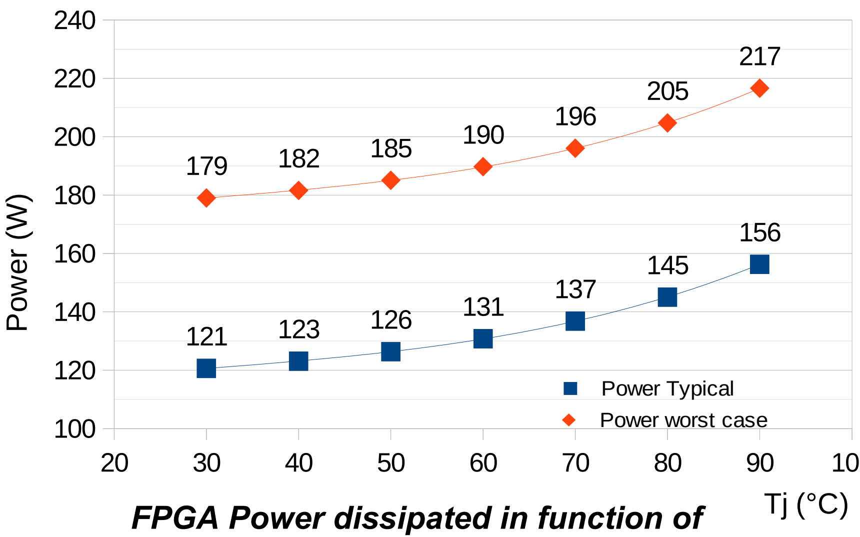
FPGA total dissipated power (TDP)

- Estimation at early stage with limited gateware inputs from developers
  → risk of over-designing cooling solution
- Estimated between 120W to 230W
- Need for high performance cooling solution

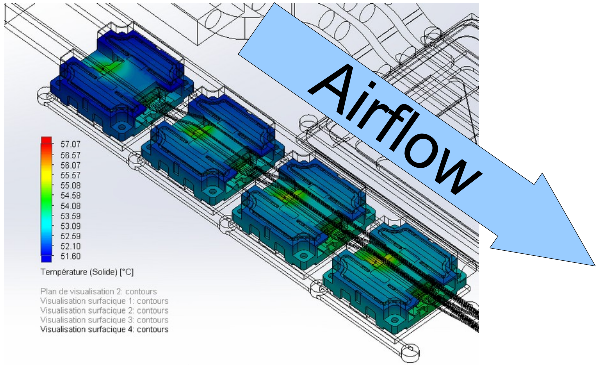
Opto-electronic transceiver

- Estimated as constant 30W

High constraints on placement due to form factor



FPGA Power dissipated in function of junction temperature


Airflow
57.07
56.57
56.07
55.57
55.08
54.58
54.08
53.59
53.09
52.59
52.10
51.60
Température (Solide) [°C]
Plan de visualisation 2: contours
Visualisation surfacique 1: contours
Visualisation surfacique 2: contours
Visualisation surfacique 3: contours
Visualisation surfacique 4: contours

Opto-electronic transceiver CFD simulation heatmap

Cooling solution

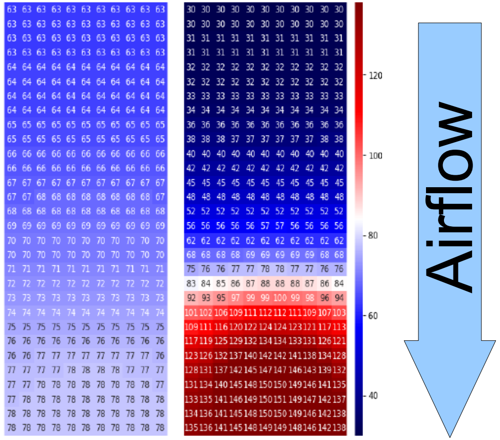
Air cooling solution privileged for more flexibility

- Vapor chamber show high performance for our application but high NRE cost
- Instead heat-pipe heat-sink with skived fins outsourced design

Nominal performance validated in simulation @ 38°C ambient and 5m/s airflow

- FPGA is maintained at 85°C with 160W
- Opto-electronic transceivers are maintained <60°C at 30W
- QSFP112 is maintained at 75°C with 12W dissipation
- SFP+ are maintained <60°C

Final cooling solution will be decided after tests on prototype



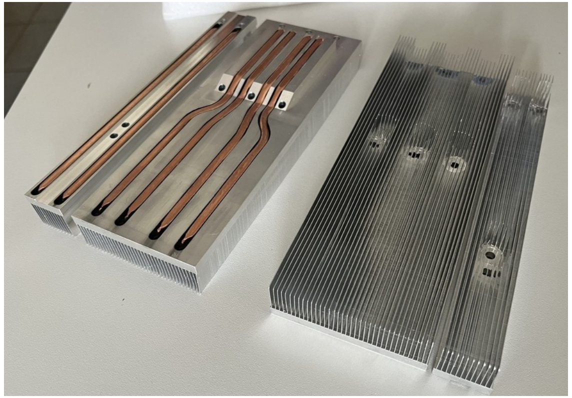
Vapor chamber

63	63	63	63	63	63	63	63	63	63	63
63	63	63	63	63	63	63	63	63	63	63
63	63	63	63	63	63	63	63	63	63	63
63	63	63	63	63	63	63	63	63	63	63
64	64	64	64	64	64	64	64	64	64	64
64	64	64	64	64	64	64	64	64	64	64
64	64	64	64	64	64	64	64	64	64	64
64	64	64	64	64	64	64	64	64	64	64
65	65	65	65	65	65	65	65	65	65	65
65	65	65	65	65	65	65	65	65	65	65
66	66	66	66	66	66	66	66	66	66	66
66	66	66	66	66	66	66	66	66	66	66
67	67	67	67	67	67	67	67	67	67	67
67	67	68	68	68	68	68	68	68	68	67
68	68	68	68	68	68	68	68	68	68	68
69	69	69	69	69	69	69	69	69	69	69
70	70	70	70	70	70	70	70	70	70	70
70	70	70	70	70	70	70	70	70	70	70
71	71	71	71	71	71	71	71	71	71	71
72	72	72	72	72	72	72	72	72	72	72
73	73	73	73	73	73	73	73	73	73	73
74	74	74	74	74	74	74	74	74	74	74
75	75	75	75	75	75	75	75	75	75	75
76	76	76	76	76	76	76	76	76	76	76
76	76	77	77	77	77	77	77	77	77	76
77	77	77	77	78	78	78	78	77	77	77
77	77	78	78	78	78	78	78	78	78	77
77	78	78	78	78	78	78	78	78	78	78
78	78	78	78	78	78	78	78	78	78	78
78	78	78	78	78	78	78	78	78	78	78

Solid metal

30	30	30	30	30	30	30	30	30	30	30
30	30	30	30	30	30	30	30	30	30	30
31	31	31	31	31	31	31	31	31	31	31
31	31	31	31	31	31	31	31	31	31	31
32	32	32	32	32	32	32	32	32	32	32
32	32	32	32	32	32	32	32	32	32	32
33	33	33	33	33	33	33	33	33	33	33
34	34	34	34	34	34	34	34	34	34	34
36	36	36	36	36	36	36	36	36	36	36
38	38	38	37	37	37	37	37	37	38	38
40	40	40	40	40	40	40	40	40	40	40
42	42	42	42	42	42	42	42	42	42	42
45	45	45	45	45	45	45	45	45	45	45
48	48	48	48	48	48	48	48	48	48	48
52	52	52	52	52	52	52	52	52	52	52
56	56	56	56	56	57	57	56	56	56	56
62	62	62	62	62	62	62	62	62	62	62
68	68	68	68	69	69	69	69	69	68	68
75	76	76	77	77	78	78	77	77	76	76
83	84	85	86	87	88	88	88	87	86	84
92	93	95	97	99	99	100	99	98	96	94
101	102	106	109	111	112	112	111	109	107	103
109	111	116	120	122	124	124	123	121	117	113
117	119	125	129	132	134	134	133	131	126	121
123	126	132	137	140	142	142	141	138	134	128
128	131	137	142	145	147	147	146	143	139	132
131	134	140	145	148	150	150	149	146	141	135
133	135	141	146	149	150	151	149	147	142	137
134	136	141	145	148	150	150	149	146	142	138
135	136	141	145	148	149	149	148	146	142	138

Heat spread on heatsink base comparsion



Prototype heat-sinks

Power distribution

Power integrity

22 power rails with high accuracy in voltage and high current up to 100A

- Use 70µm thick copper planes to reduce voltage drop and power dissipation in planes

18 layers PCB restricted to 1.57mm in thickness due to PCIe edge connector specification

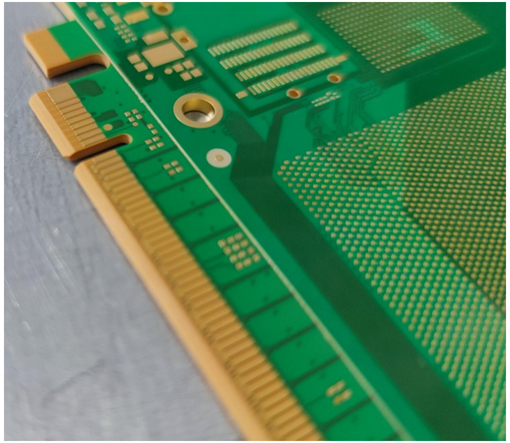
- Design of a PCB thinner on connector zone

Power dissipation within power plane

Power dissipated	70µm	35µm	Δ
Layer TOP	9.6W	11.0W	+14 %
Layer 9	4.9W	6.8W	+38 %
TOTAL	14.5W	17.8W	+23 %

PCB T° rise	17	30

Illustration of a PCIe PCB thinner on connector

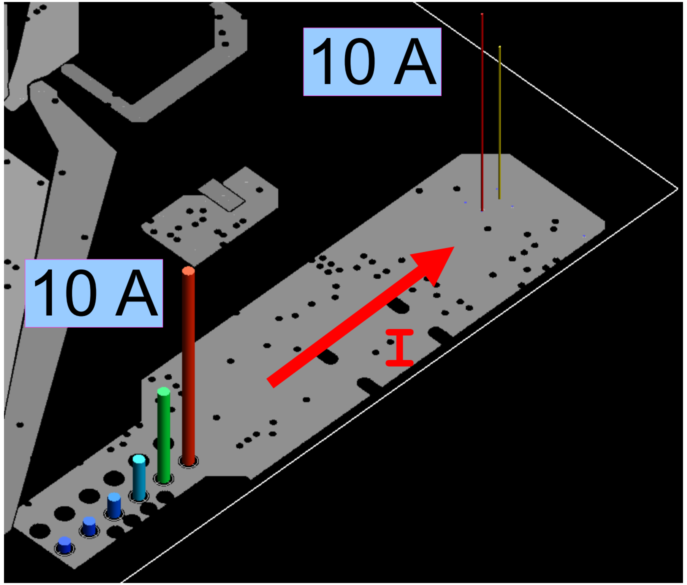


Power integrity simulations

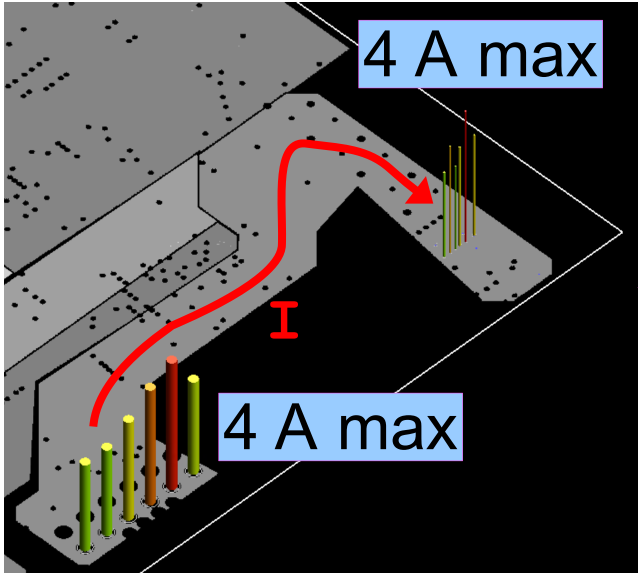
Optimize power plane geometry

- Reduce high current in vias
- Uniformize current in power planes to reduce heat dissipation in the PCB

Example of current in vias optimization on 12V Auxiliary power plane at 18 A



Current in vias
Original power plane design



Current in vias
updated power plane design

Signal integrity

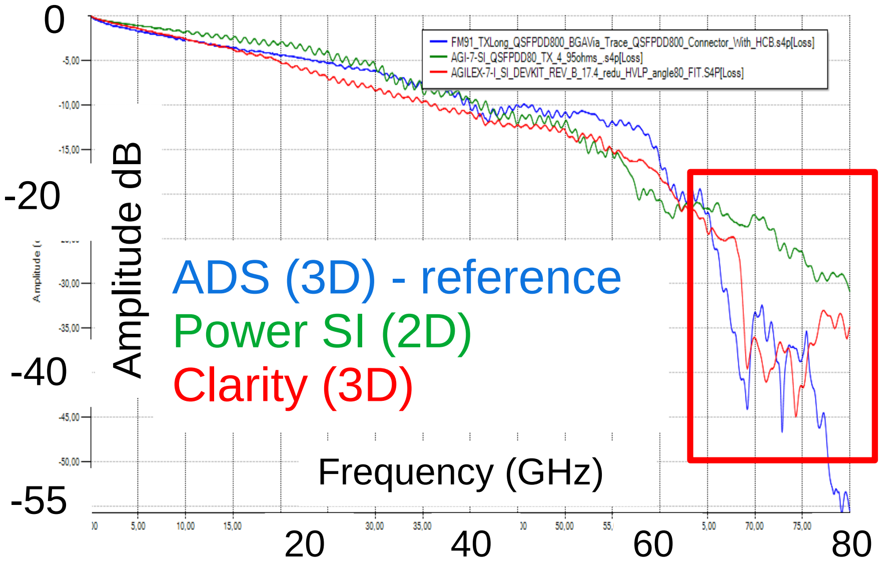
Simulation tools

108 differential pairs at up to 112Gbps PAM4

- 84GHz bandwidth
- Need to take into consideration vias 3D geometry

Simulation tools take a lot of computational resources

- S-parameter extraction of a single differential pair takes ~8h on a 48 cores @3.2GHz machine



SDD21 depending on simulation tool

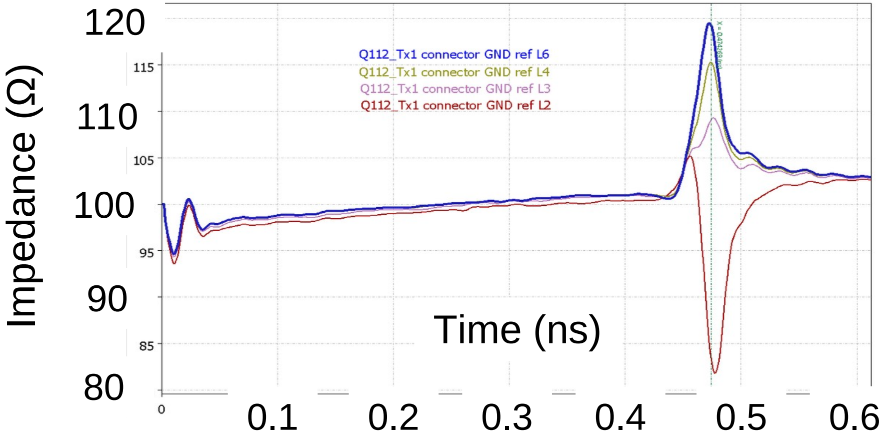
Simulation performed

Several iterations are required to reach specification

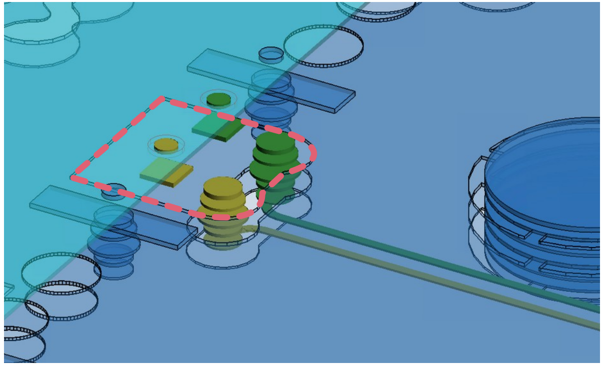
- 112Gbps traces require controlled impedance with **7% tolerance** while standard is 10%

With 84GHz bandwidth, small detail can lead to large impedance mismatch

- Fanout
- Vias structures for current return path
- Openings on adjacent planes
- Openings of planes under connector pads
- Trace length matching



112Gbps simulated TDR showing impedance mismatch depending on distance to GND reference plane



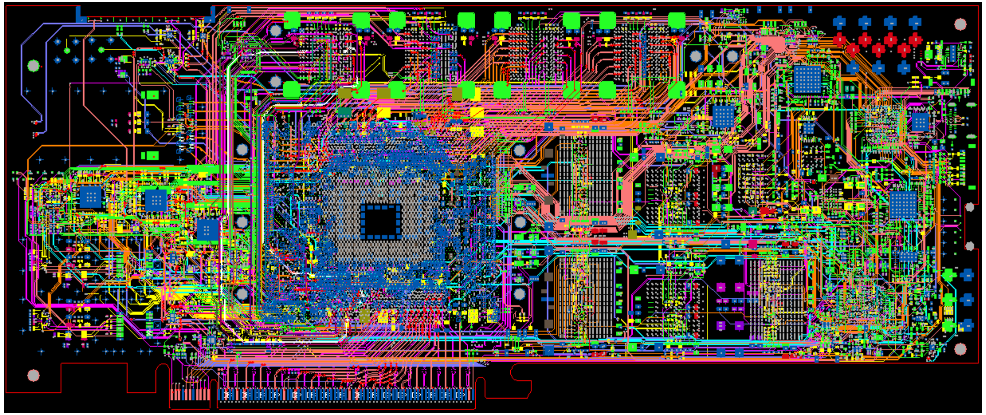
Example of connector opening on L2

Routing

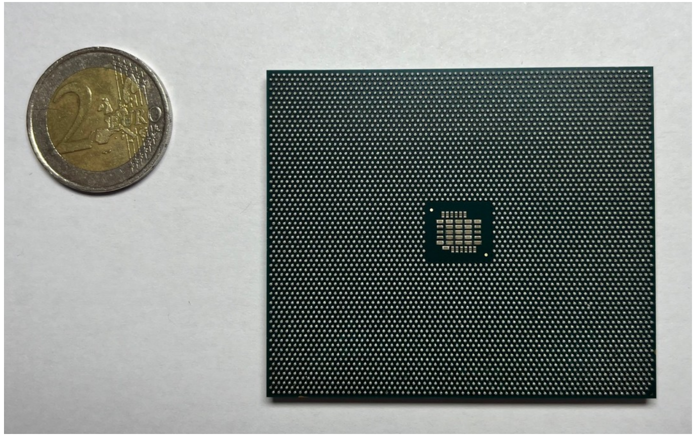
Very dense routing

- 270 x 110 mm PCB (GPGPU form factor)
- 2500 components on-board

4500 pins FPGA with 0.9mm pitch with hexadecimal structure

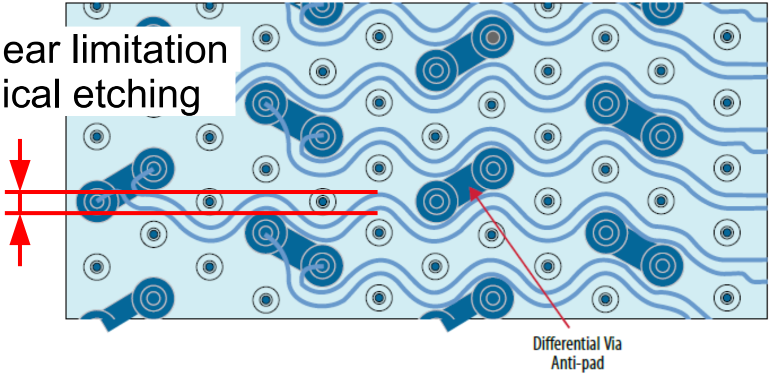


PCIe400 Layout illustration



FPGA pin photo

80µm : near limitation for chemical etching



Hexadecimal pin breakout

Synthesis

PCIe400 is a R&D development pursued by IN2P3

- 400Gbit/s output bandwidth per board with up to 48 bidirectional interfaces for front-end
- Baseline solution for LHCb future upgrade
- Generic design that can suit several application (Belle II, Alice, CTA)

It also paves ways to explore future DAQ topologies

- 400Gbit/s network interface allowing switch based interconnections or process pipelining between boards
- Integration of a white rabbit node for future generation of precise clock distribution

Hardware design show many technical challenges in cooling, power distribution and signal integrity

Several technical challenges yet to overcome with modern SoC FPGA

- Phase deterministic clock distribution
  - ECFA DRD7 7.3c Timing distribution techniques
- Optimizing input data bandwidth to efficiently use FPGA ressources
  - 7.5b 100GbE from front-end to back-end

First prototype boards to be tested in June 2024