



Acquisition de données

Exemple de l'expérience LHCb et prospective

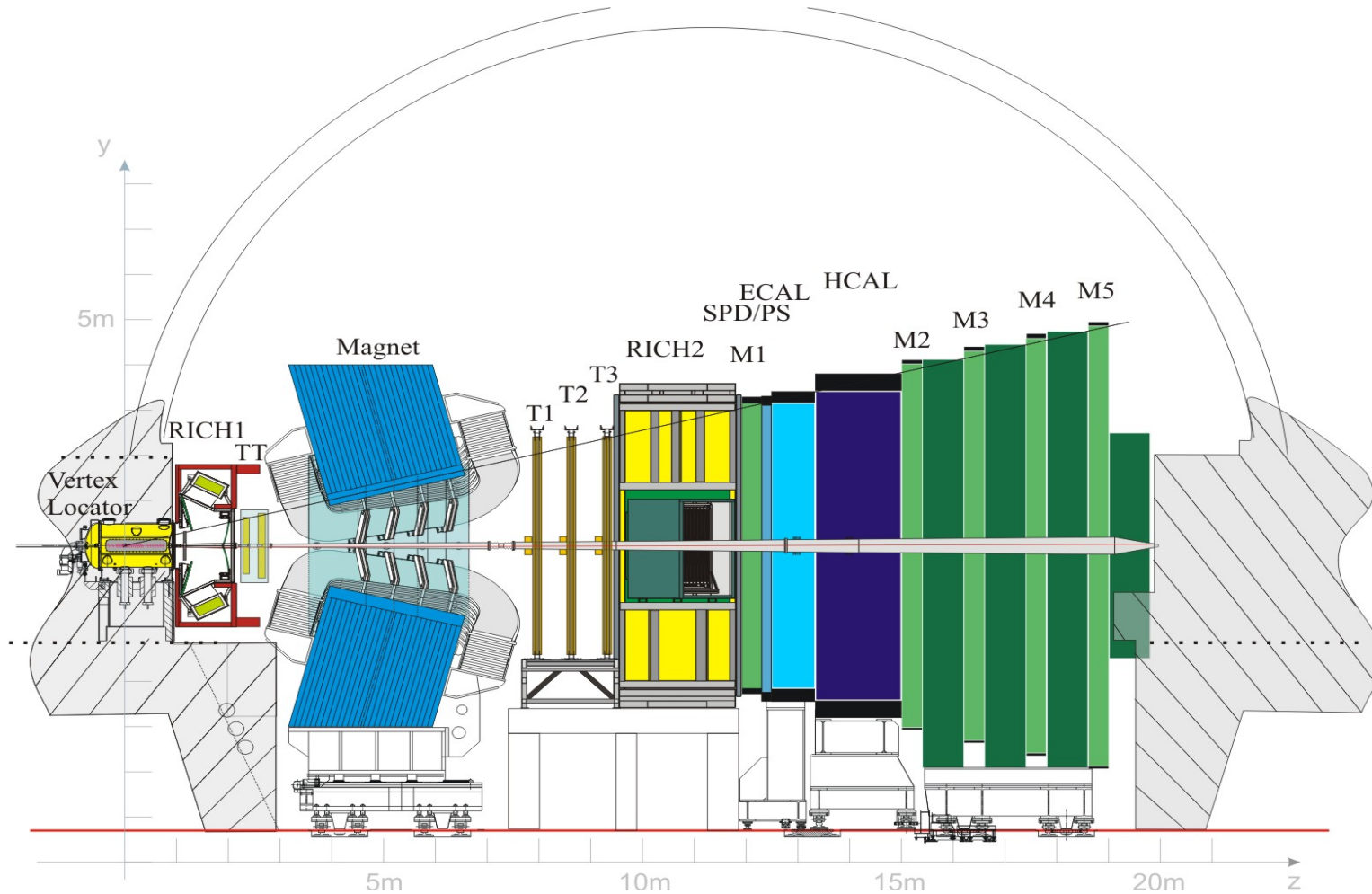


J.P. Cachemiche
Centre de Physique des Particules de Marseille

Plan

- **Présentation du trigger LHCb**
- **Démarche de réalisation du trigger à muon**
- **Techniques de base pour le parallélisme**
- **Evolution de l'architecture d'acquisition LHCb pour 2019**

Le détecteur LHCb



Etude des asymétries matière/anti-matière dans la physique du méson B

Fonctions du système

- Lecture d'environ 1 million de canaux
- Production de 100 000 paires $b\bar{b}$ par secondes
- Recherche des événements impliquant un méson B
- Acquisition
- Identification
- Stockage (quelques kHz)

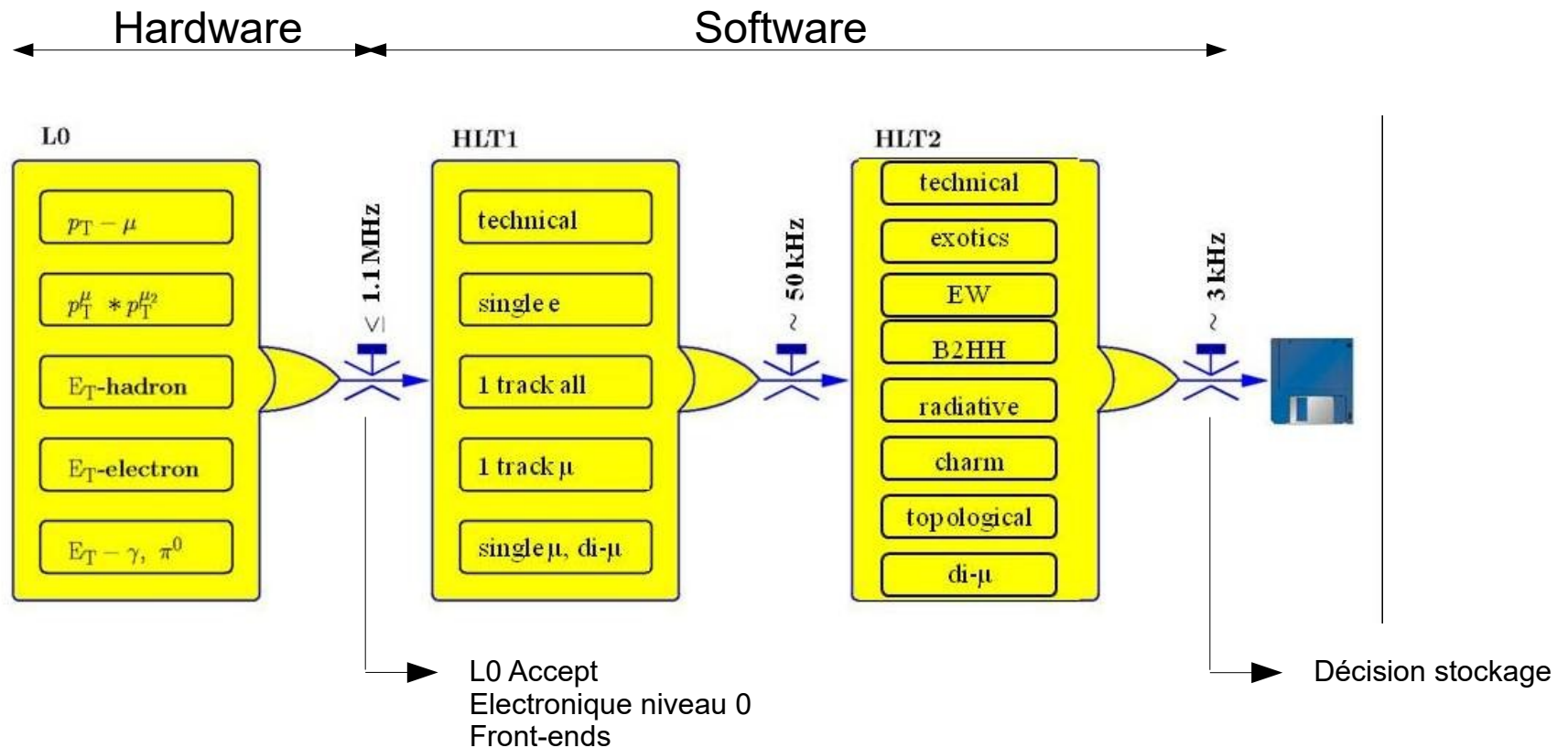
Decay Modes	Visible Br. fraction	Offline Reconstr.
$B_d^0 \rightarrow \pi^+ \pi^- + \text{tag}$	0.7×10^{-5}	6.9 k
$B_d^0 \rightarrow K^+ \pi^-$	1.5×10^{-5}	33 k
$B_d^0 \rightarrow \rho^+ \pi^- + \text{tag}$	1.8×10^{-5}	551
$B_d^0 \rightarrow J/\psi K_S + \text{tag}$	3.6×10^{-5}	56 k
$B_d^0 \rightarrow \bar{D}^0 K^{*0}$	3.3×10^{-7}	337
$B_d^0 \rightarrow K^{*0} \gamma$	3.2×10^{-5}	26 k
$B_s^0 \rightarrow D_s^- \pi^+ + \text{tag}$	1.2×10^{-4}	35 k
$B_s^0 \rightarrow D_s^- K^+ + \text{tag}$	8.1×10^{-6}	2.1 k
$B_s^0 \rightarrow J/\psi \phi + \text{tag}$	5.4×10^{-5}	44 k

Expected numbers of events reconstructed offline in one year (10^7 s of data taking) with an average luminosity of $2 \times 10^{32} \text{ cm}^{-2} \text{ s}^{-1}$, for some channels.

Trigger

Filtrage des événements

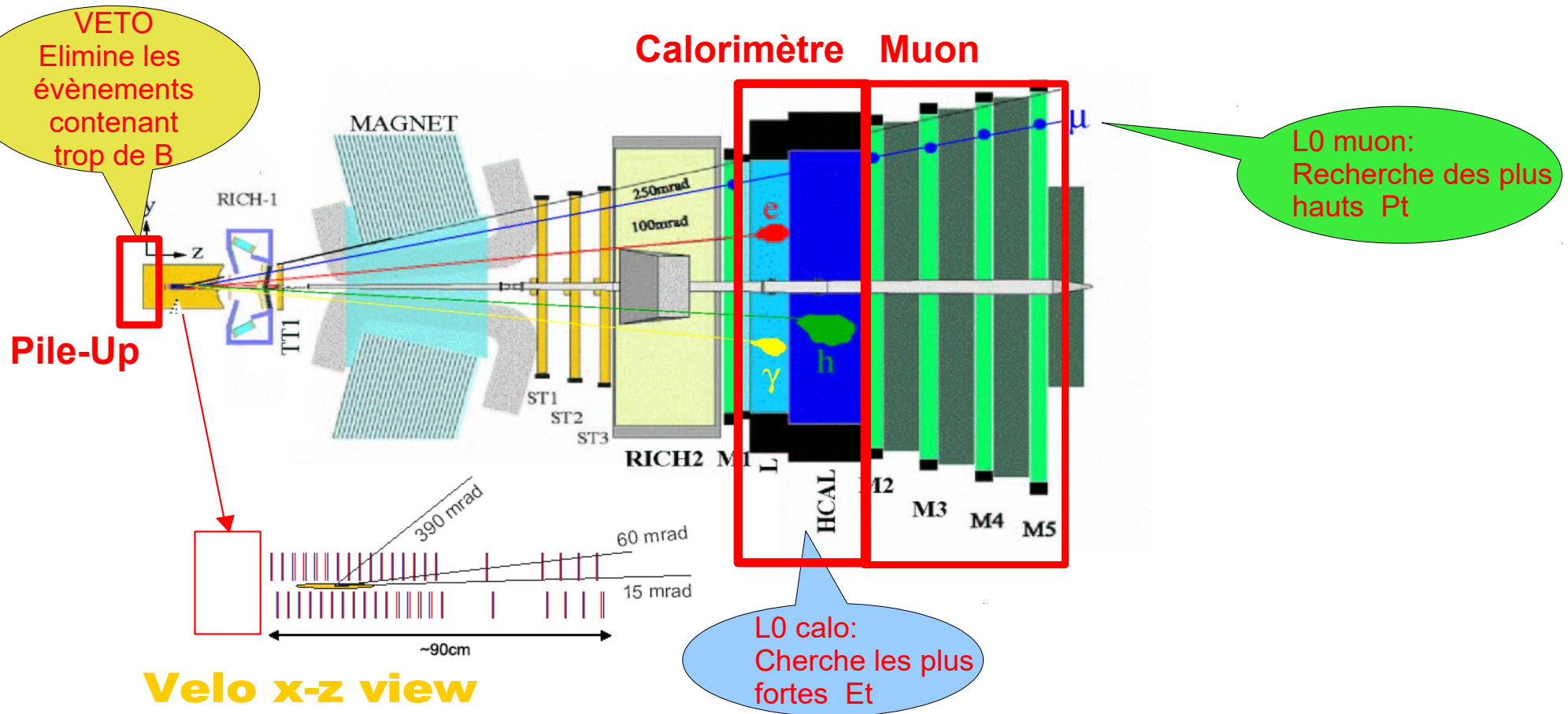
3 étapes de réduction :



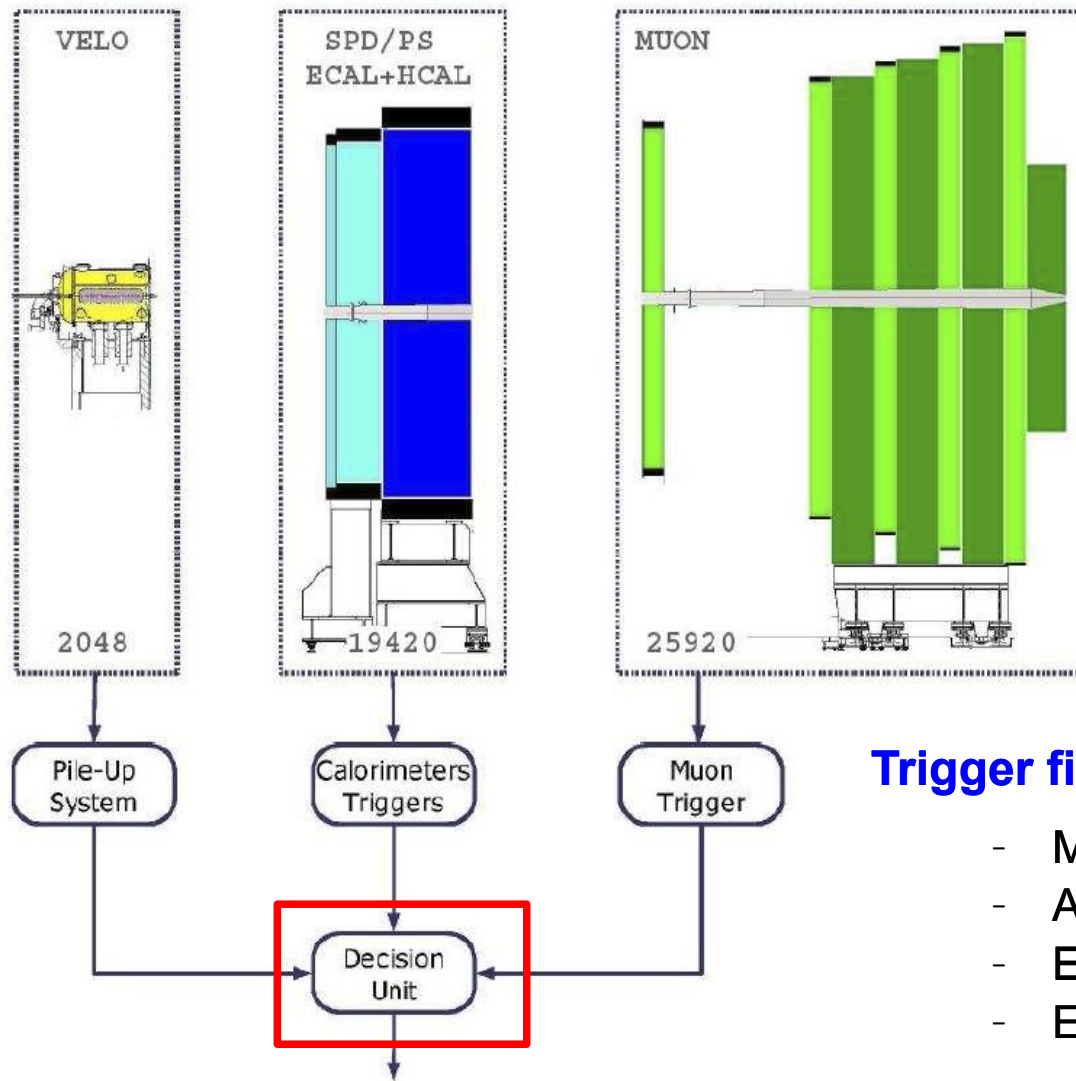
Sous-détecteurs participant au trigger de niveau 0

3 détecteurs

- Calorimètres, muons et pile-up veto

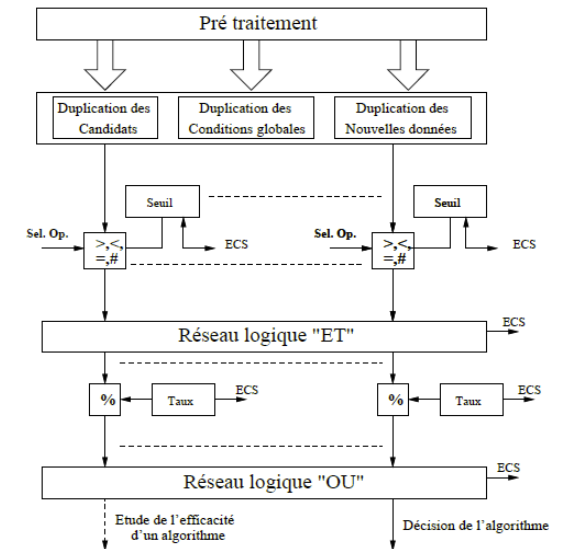


Unité de décision



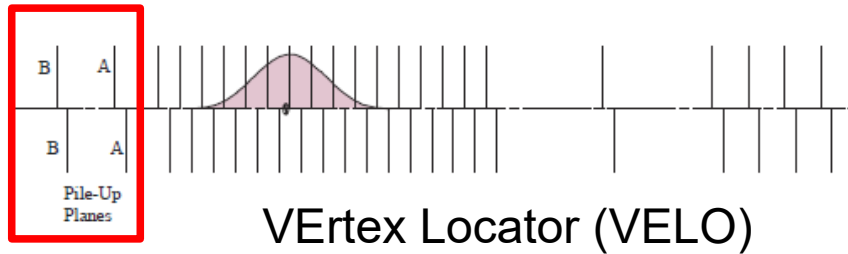
Trigger final réalisé par l'unité de décision

- Mise en temps
- Algorithme trigger global
- Envoi décisions au TFC supervisor
- Envoi décision au readout ainsi qu'au HLT



Algorithme unité de décision

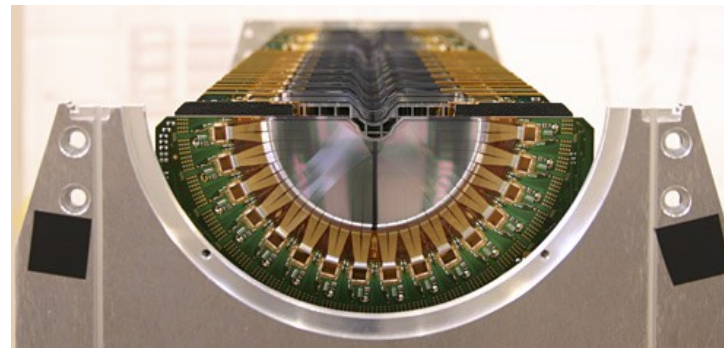
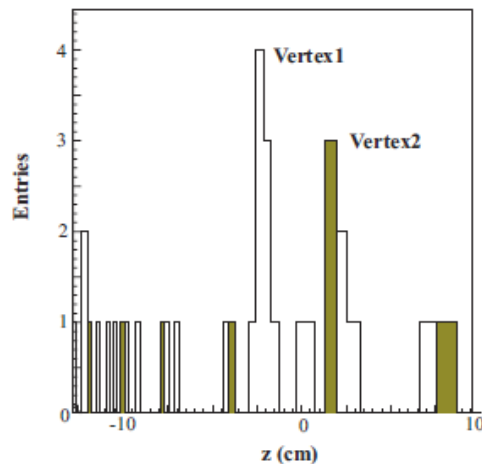
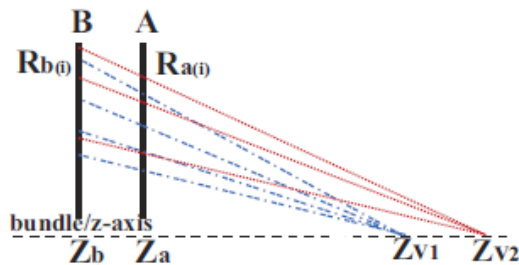
Pile-up veto



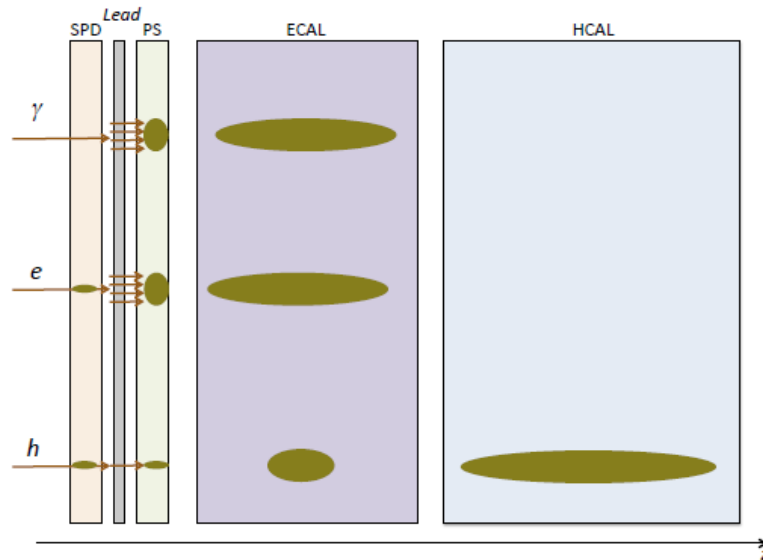
Détection des croisements contenant trop d'interactions

→ Événements trop difficiles à analyser

- Détection de tous les vertex déterminés par les hits des plans A et B
- Elimination des hits correspondant aux 2 vertex de plus grande énergie
- S'il reste un ou plusieurs vertex, élimination de l'événement (VETO)



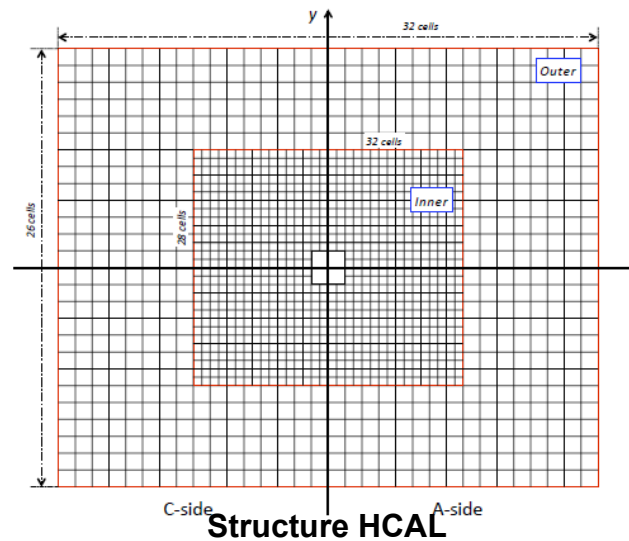
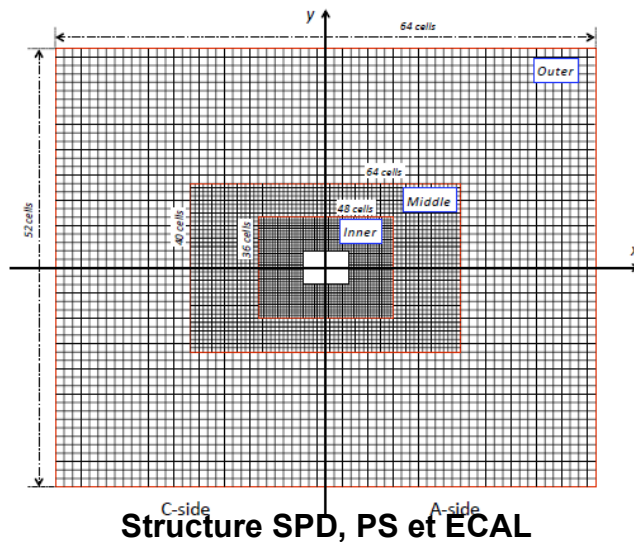
Calorimètres



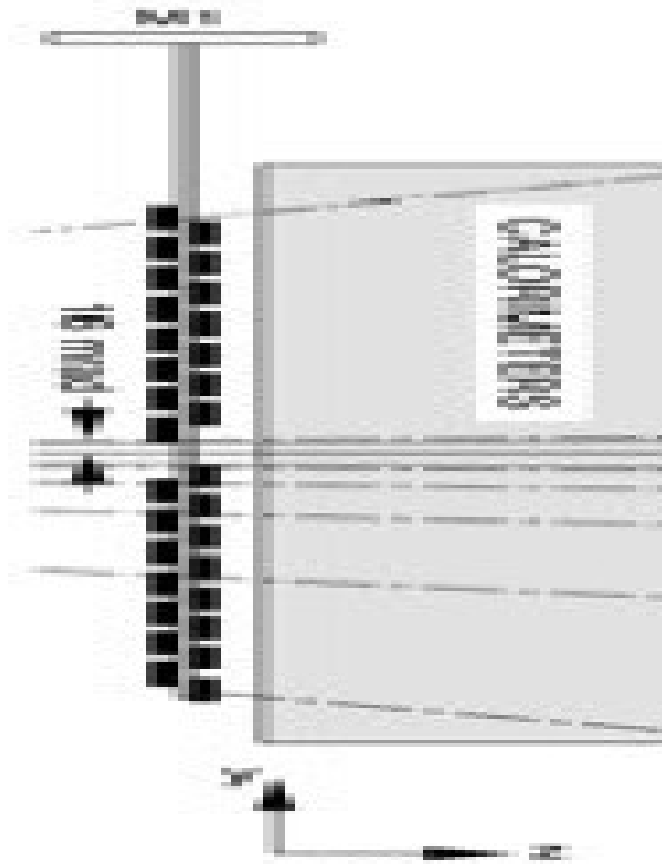
Détection des particules avec une E_T élevée

Plusieurs sous-systèmes :

- **SPD** (Scintillator Pad detector)
 - ➔ Identifie les particules chargées et sépare les électrons des photons
- **PreShower (détecteur de pied de gerbes)**
 - ➔ Identifie les électrons et photons
- **Calorimètre Electro-magnétique**
 - ➔ Mesure l'énergie des électrons et photons
- **Calorimètre Hadronique**
 - ➔ Mesure l'énergie des hadrons



Trigger à muon



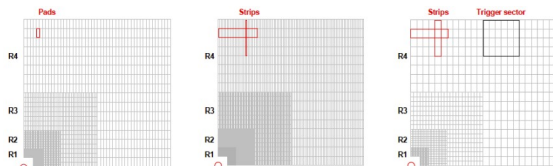
Détection des muons avec une impulsion transverse (P_t) élevée

- 1400 GEM et MWPC répartis sur 5 plans
- 120000 canaux
- 435 m²
- 2.5 millions de cables

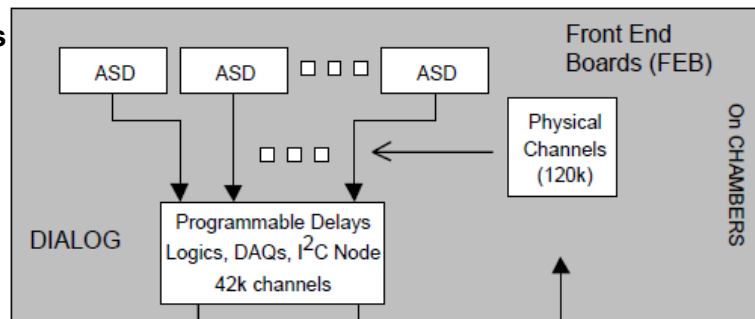


Slice muon chambers

Données muons

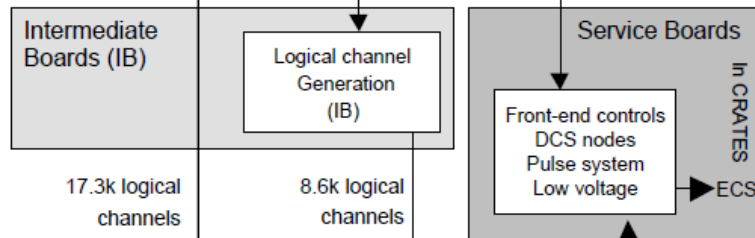


~7600 boards



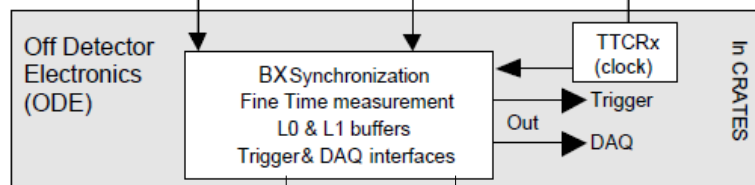
~125 000 Physical channels

~ 150 boards



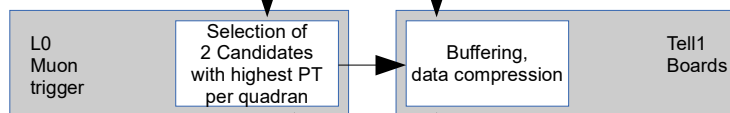
~42 000 LVDS channels

~ 150 boards



~26 000 Logical channels

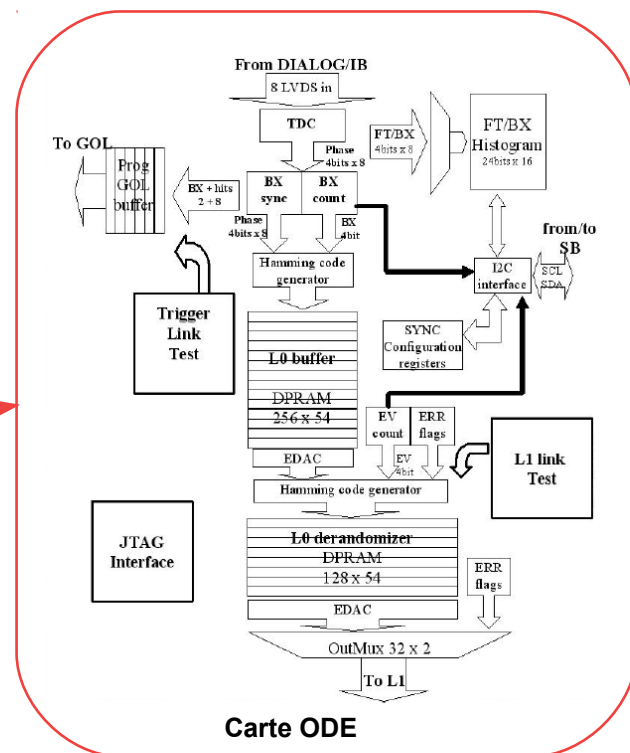
~ 52 boards



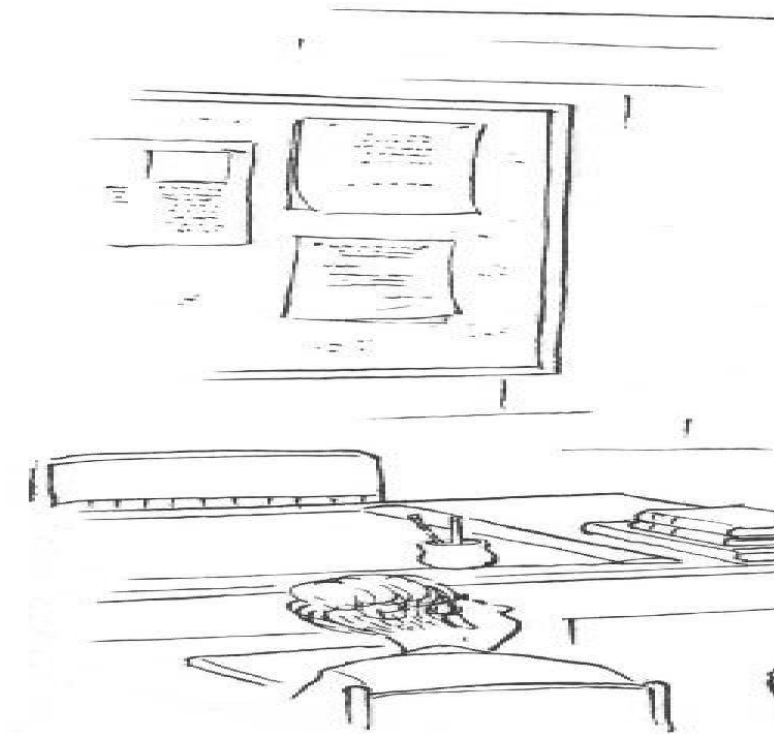
~1400 optical channels

~ 10 boards

L0 Decision Unit
Computer farm

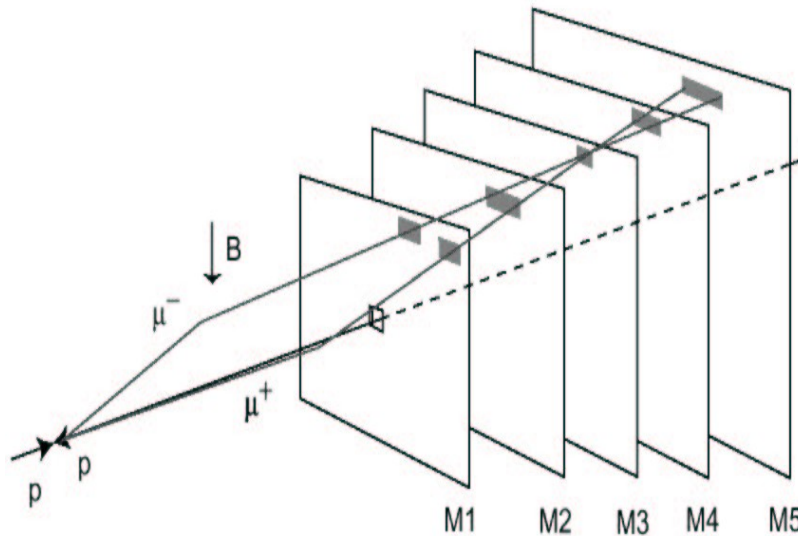


Carte ODE



Démarche de réalisation du trigger à muon

Recherche des candidats



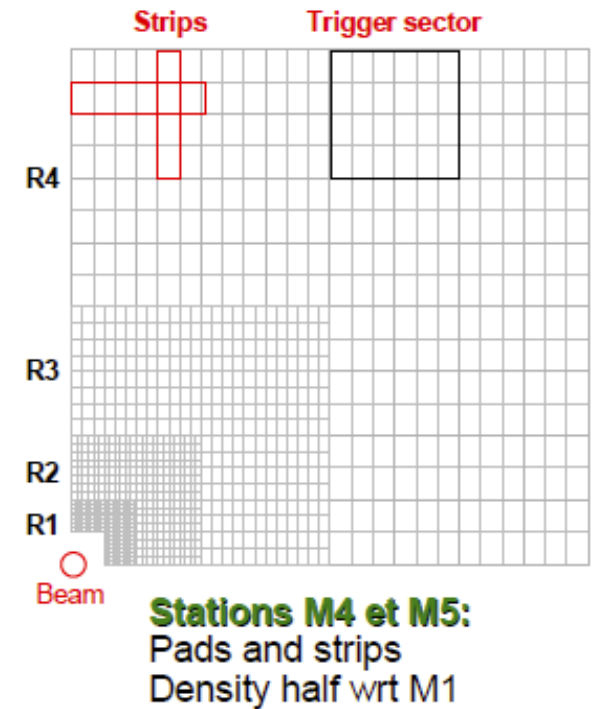
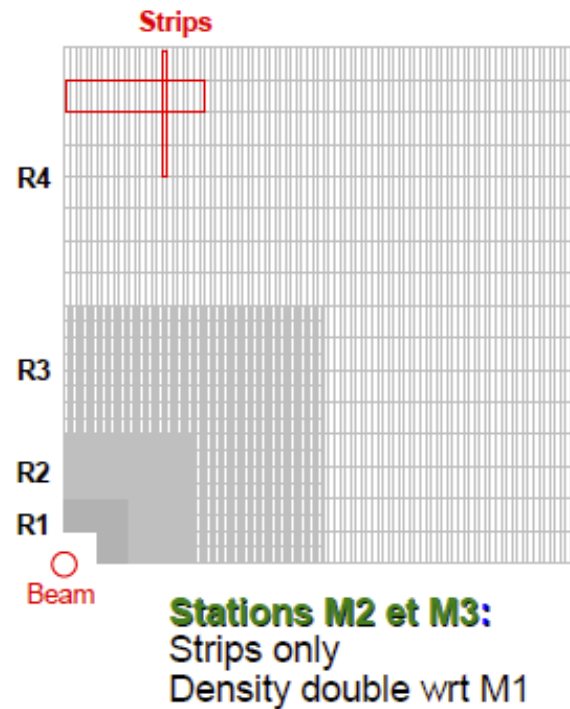
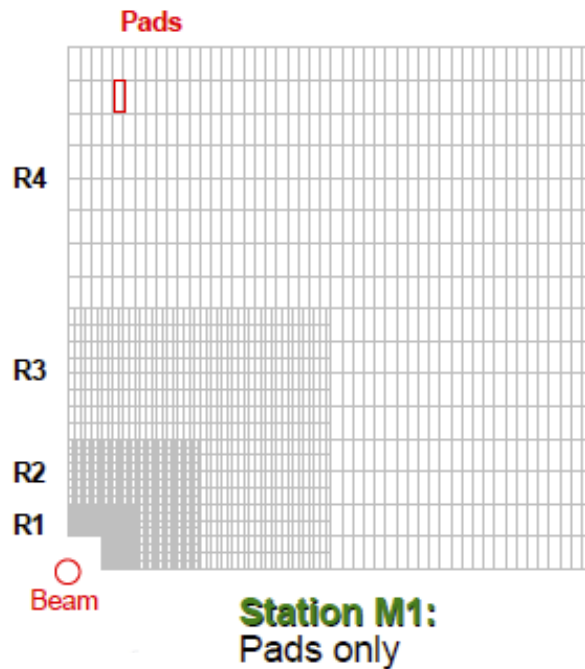
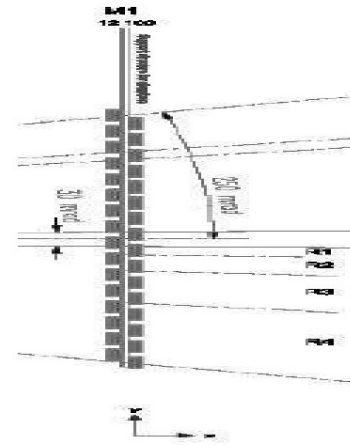
Principe de l'algorithme:

- 1- Trouver un pad touché en M3
- 2- Définir un axe de recherche centré sur le PAD
- 3- Ouvrir 2 cones le long de cet axe
- 4- Sélectionner une trace si un pad est touché dans le couloir dans les plans M5 et M4 et M2
- 5- Le point de passage en M1 est extrapolé en suivant la droite partant de M3 et passant par le pad touché de M2
- 6- Recherche d'un hit dans la zone extrapolée
- 7- Ce point dans M1 donne l'angle de la trace par rapport au faisceau donc P_T (impulsion transverse)

Chambres à muons

5 stations, 4 régions par station :

- Chambres à fils
 - Détection de muons à P_T élevé

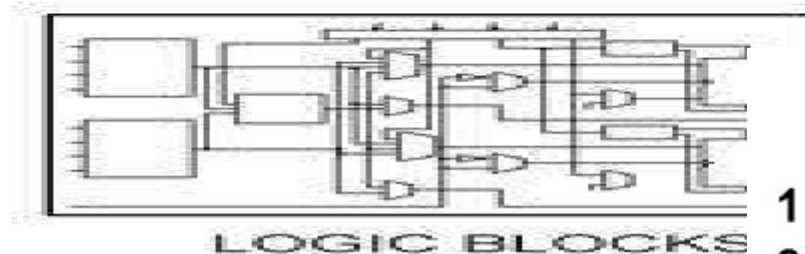
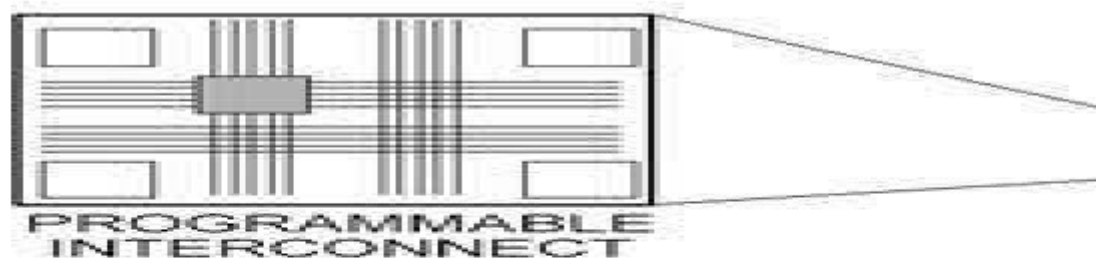


- Granularité proportionnelle à la densité de particules
- Vue simplifiée : nombreuses exceptions topologiques

Unité de traitement : un FPGA

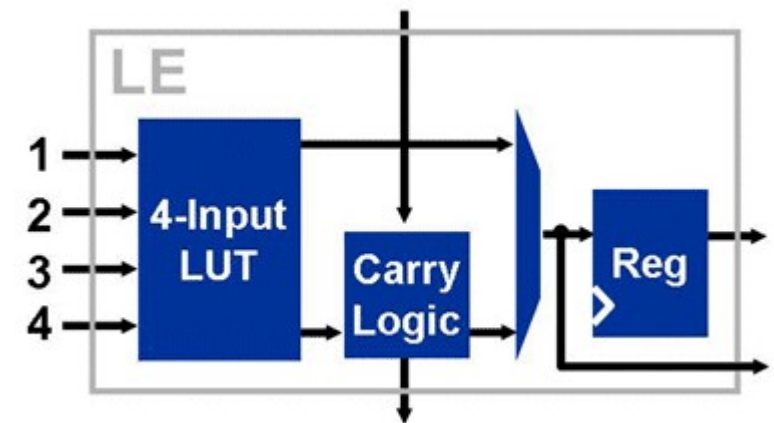
Field Programmable Gate Array

- **Matrice de cellules logiques**
interconnectables de façon programmable



Source NI

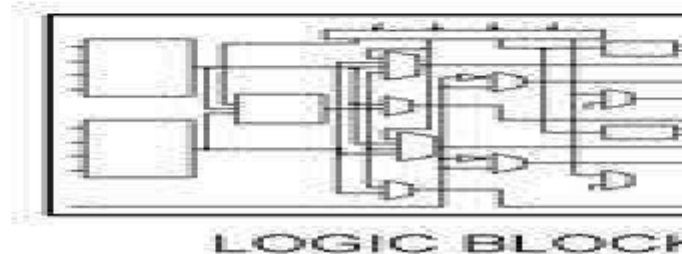
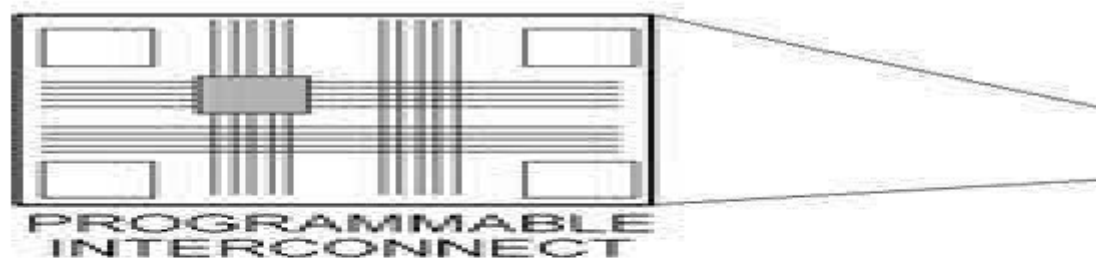
- 2006 :
→ Plusieurs dizaines de milliers de cellules logiques



Unité de traitement : un FPGA

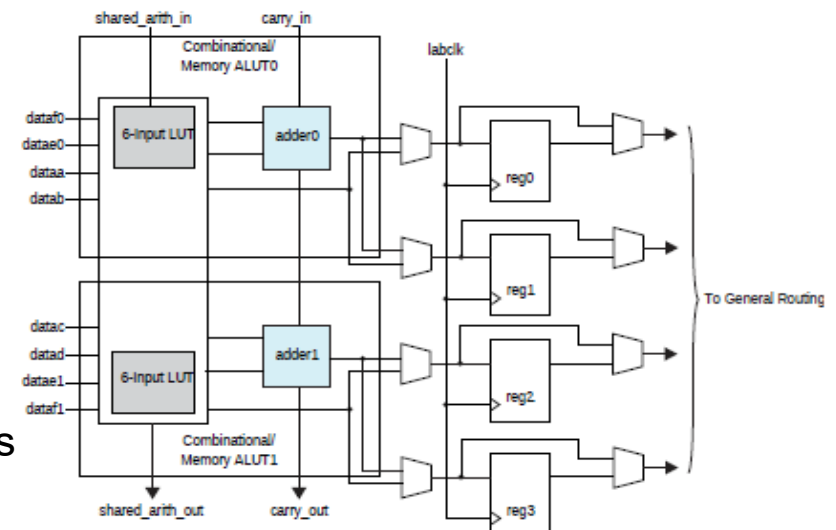
Field Programmable Gate Array

- **Matrice de cellules logiques**
interconnectables de façon programmable



Source NI

- 2018 :
→ Plusieurs **millions** de cellules logiques



Caractéristiques des FPGA

Entrées sorties programmables

- LVDS, CML, HSCL, CMOS, SSTL, LVPECL
- Avec des fonctions de filtrage : préaccentuation, égalisation

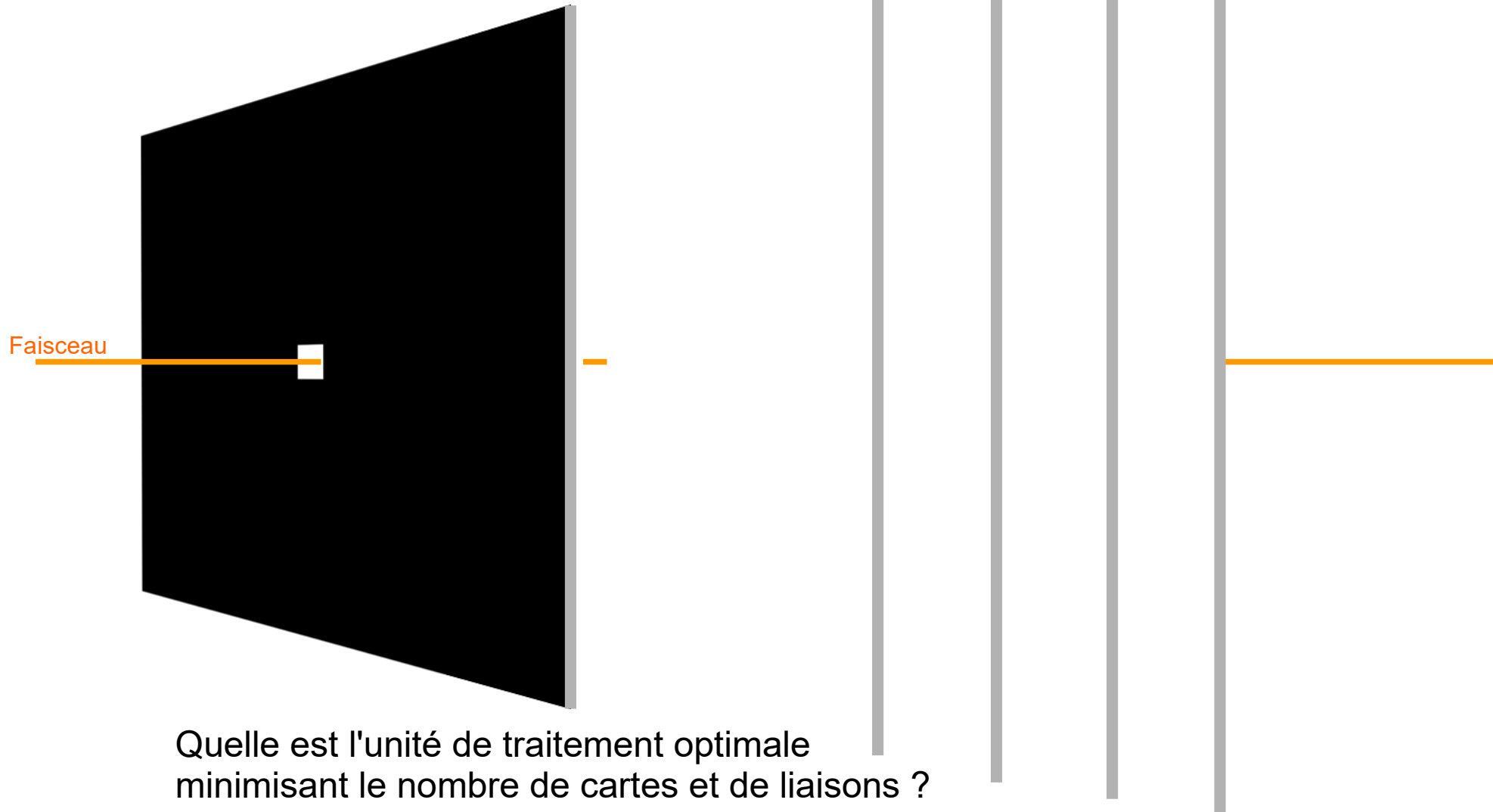
Contient également des structures câblées

- Mémoires
- PLLs
- Cellules DSP
- Sérialiseurs/désérialiseurs multigigabits
- Hardware IP blocks
 - interfaces mémoires : DDR3, DDR, QDR, ...
 - Interfaces protocoles de communication : PCIe, GbE, Interlaken, ...
- Hardware CPU : ARM
- Convertisseurs Analogiques Digitaux

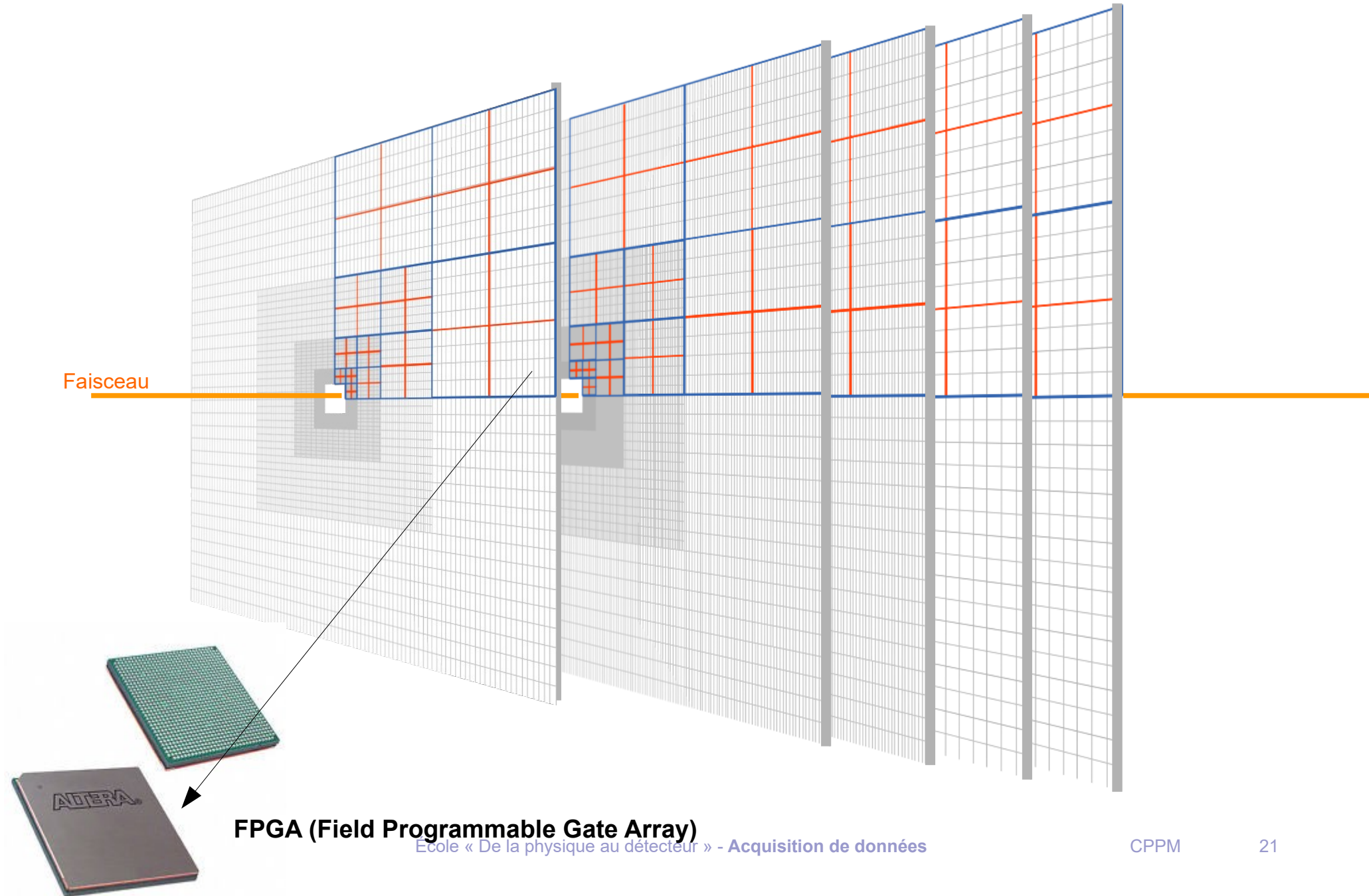
Peut contenir des fonctions autrefois dédiées aux instruments de mesure

- Analyse logique,
- Serial Data Analyser

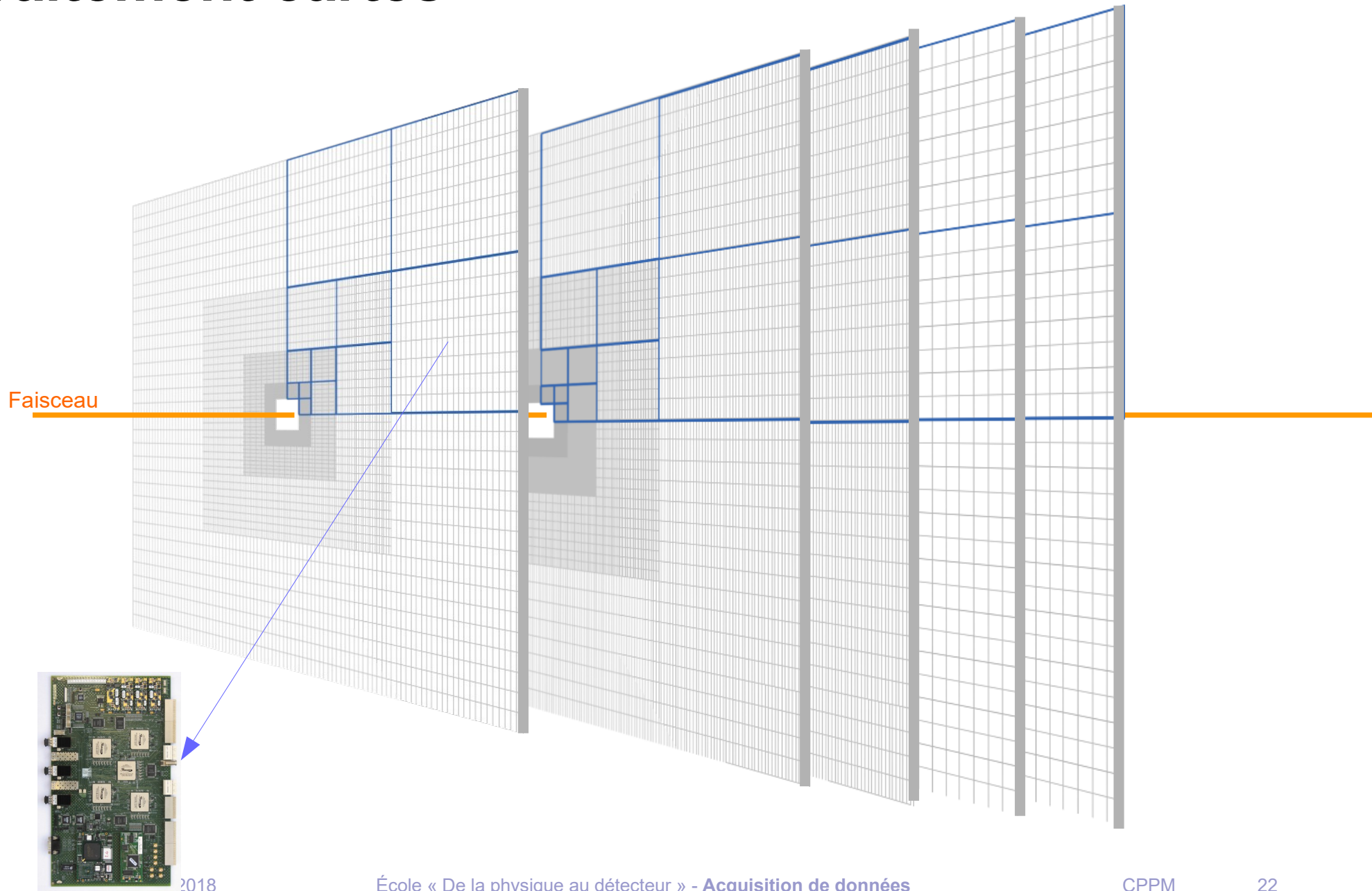
Segmentation du traitement



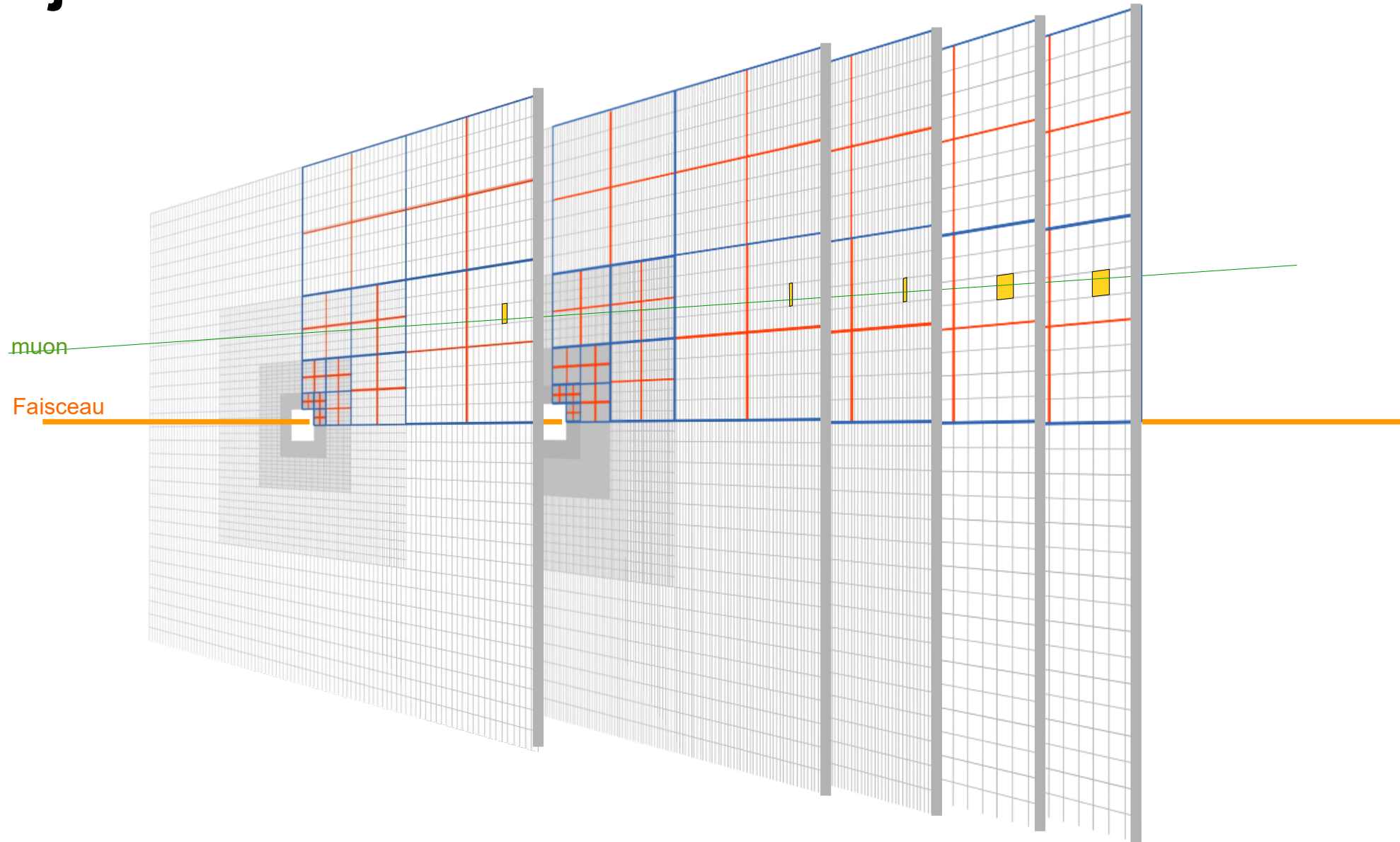
Traitement FPGAs



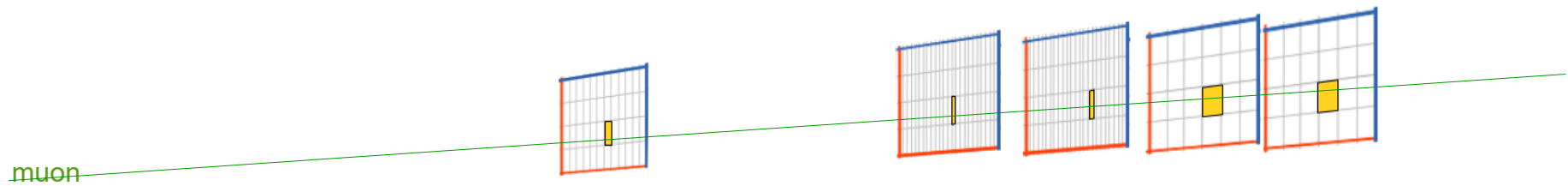
Traitement cartes



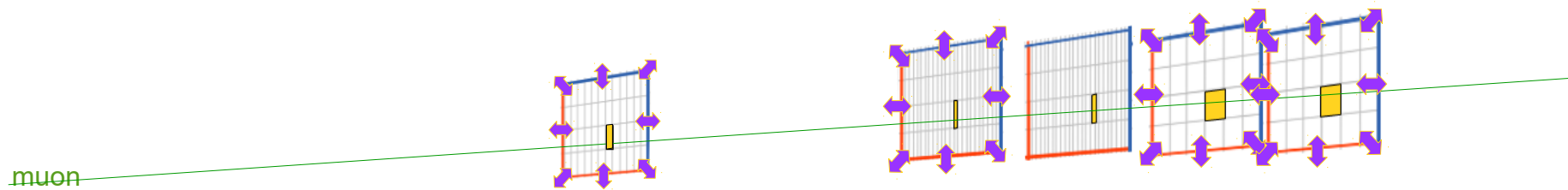
Trajectoire muon



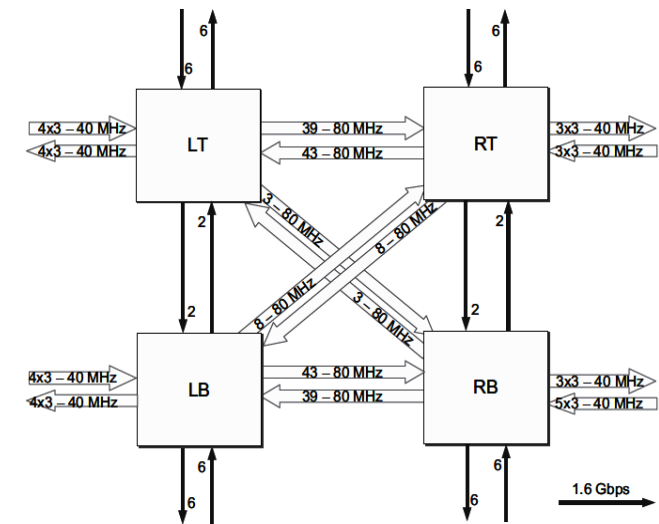
Ce que le FPGA doit voir



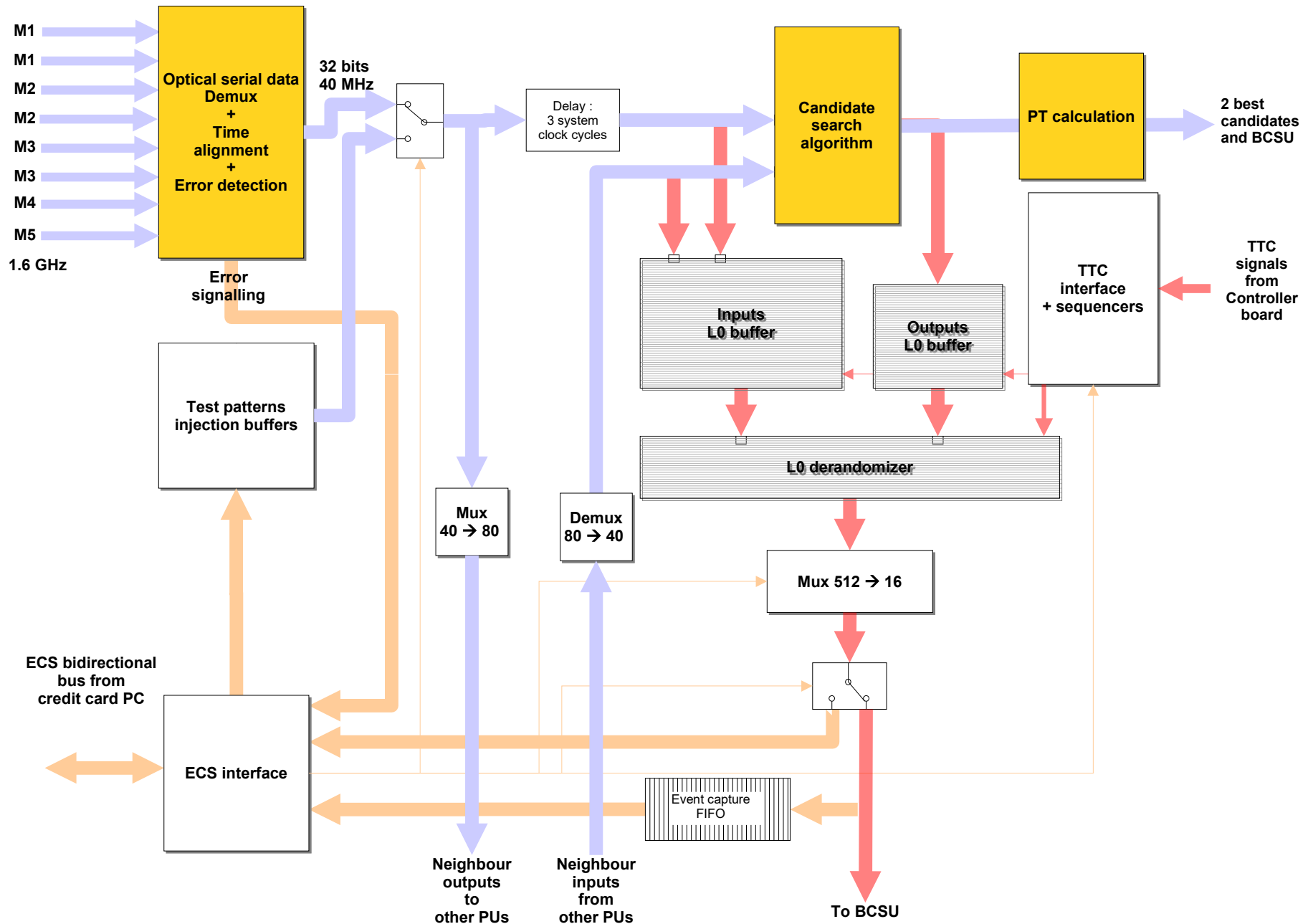
Echanges de données



**Nombreuses communications
pour traiter les détections aux limites**

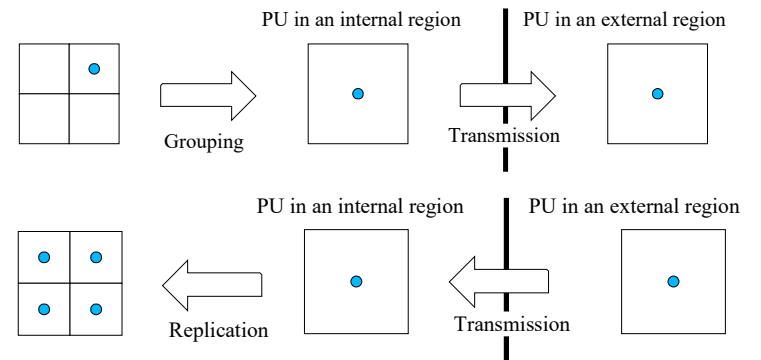
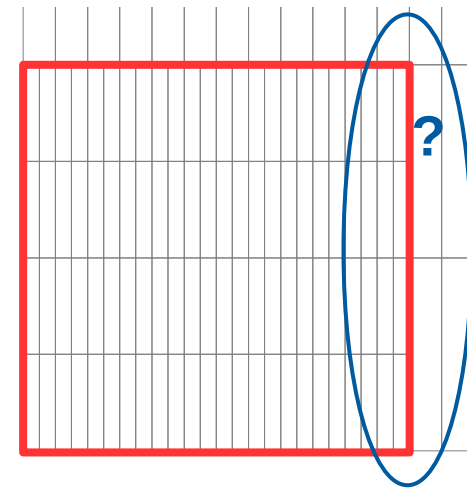
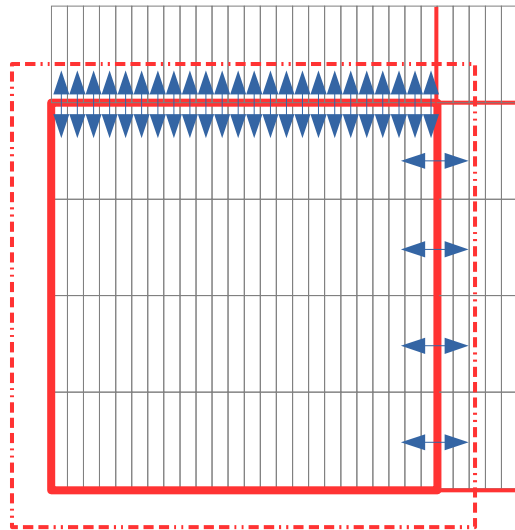


Traitement de données



Homogénéisation de l'espace de travail

Échanges de données

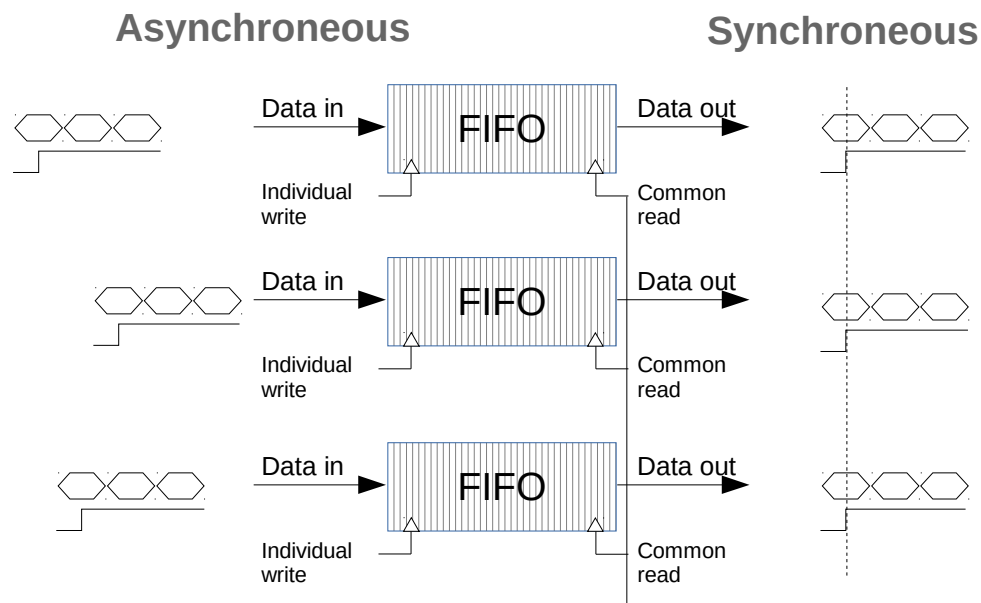


Mise en temps

Toutes les données arrivent décalées

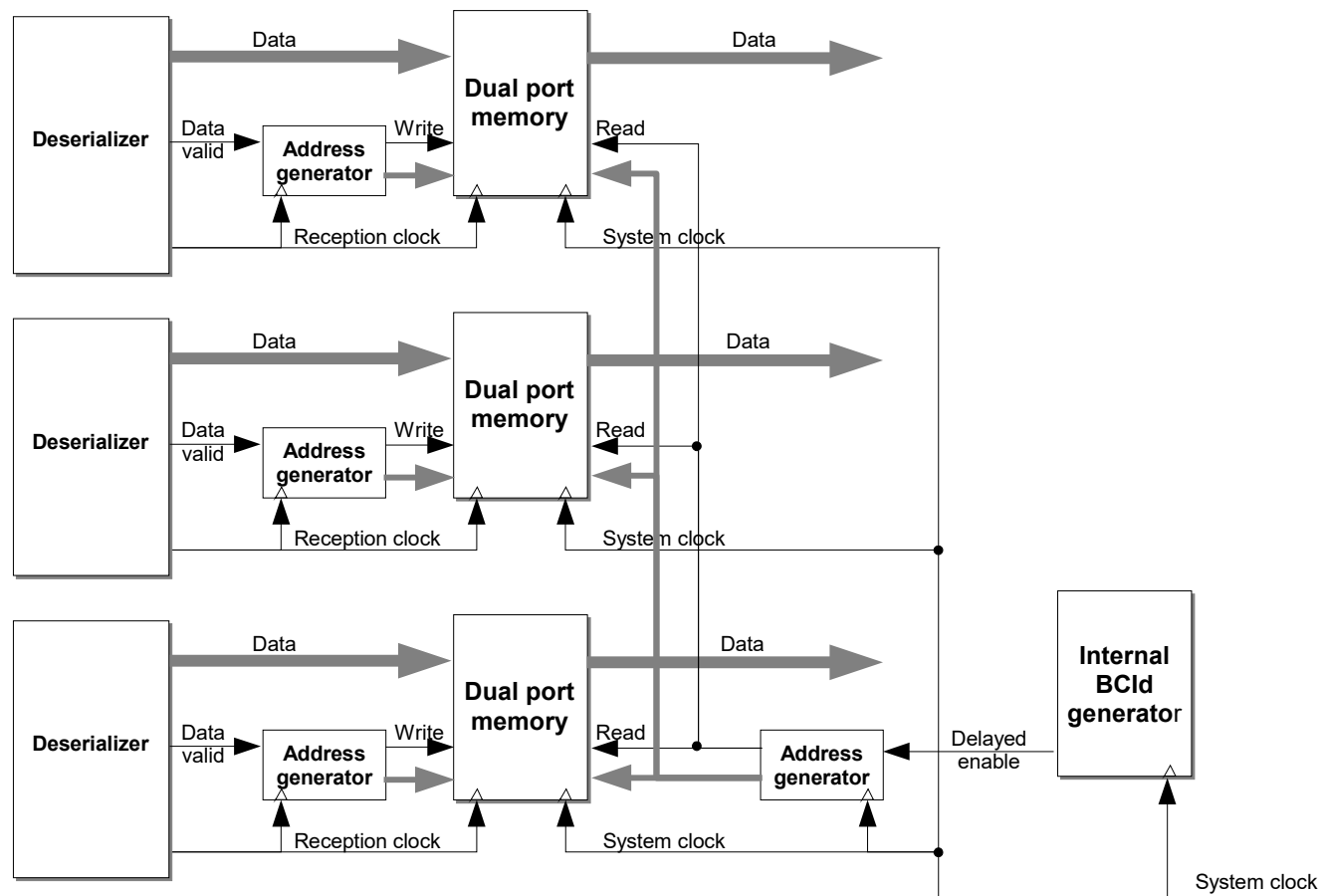
- Chemins de longueur différente
- Dérives thermiques
- Données de voisinage

Synchronisation



Mise en temps

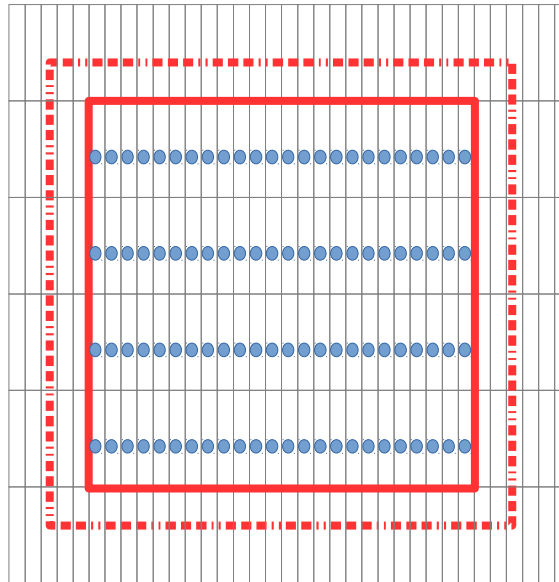
Implémentation effective



Traitement

Parallelisme massif

- Un résultat à donner toutes les 25 ns
- Pas le temps de chercher séquentiellement
 - ➔ L'algorithme de recherche est effectué sur l'ensemble des cellules simultanément



48 algorithmes

M3 seed

Traitement

Structure pipeline

- Le temps de recherche est supérieur à l'écart entre deux collisions (25 ns)
- Multiplication des unités de traitement trop coûteuse
 - On adopte une structure pipeline



	F1	F2	F3	F4	F5	F6	F7	F8
T0	Dn	Dn-1	Dn-2	Dn-3	Dn-4	Dn-5	Dn-6	Dn-7
T0 + 25	Dn+1	Dn	Dn-1	Dn-2	Dn-3	Dn-4	Dn-4	Dn-6
T0 + 50	Dn+2	Dn+1	Dn	Dn-1	Dn-2	Dn-3	Dn-4	Dn-5
T0 + 75	Dn+3	Dn+2	Dn+1	Dn	Dn-1	Dn-2	Dn-3	Dn-4
T0 + 100	Dn+4	Dn+3	Dn+2	Dn+1	Dn	Dn-1	Dn-2	Dn-3
T0 + 125	Dn+5	Dn+4	Dn+3	Dn+2	Dn+1	Dn	Dn-1	Dn-2
T0 + 150	Dn+6	Dn+5	Dn+4	Dn+3	Dn+2	Dn+1	Dn	Dn-1
T0 + 175	Dn+7	Dn+6	Dn+5	Dn+4	Dn+3	Dn+2	Dn+1	Dn

Operation	Estimated time [ns]	Estimated number of clock periods
Time of flight to M5	63	13
FE board processing	70	
Transmission to IB and ODE (15 m)	105	
IB processing	40	
ODE processing	30	
Transmission to processing (100 m)	600	24
Muon processing	1200	48
Transmission to L0 decision Unit	50	2
L0 Decision Unit processing	525	21
L0 Decision Unit distribution	800	32
Contingency	500	20
Total	3983	160

- Profondeur du pipeline muon trigger LHCb : 48 coups d'horloge

Calcul du P_T

Formule simple

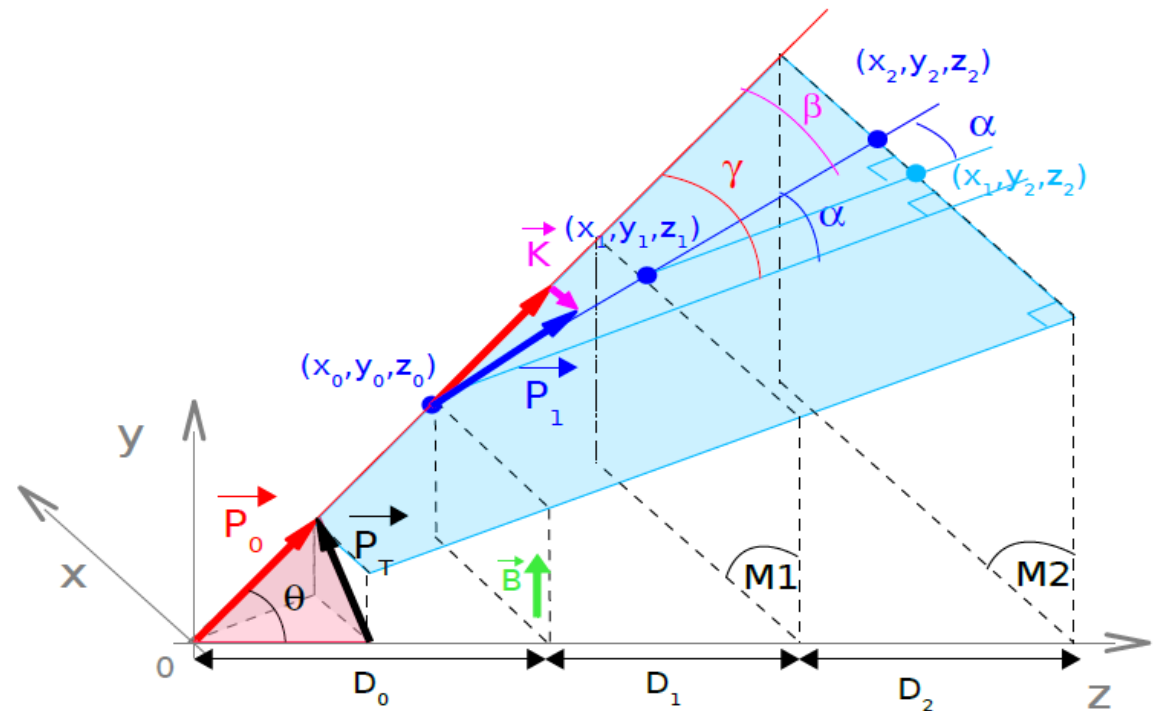
$$P_T = P_0 \sin(\theta)$$

Calcul du P_T

Formule simple

$$P_T = P_0 \sin(\theta)$$

sauf que



- P_0 is obtained from β , the deflection angle and \vec{K} :

$$P_0 = \frac{K}{2\sin(\frac{\beta}{2})} = \frac{K}{2\sin(\frac{\gamma-\alpha}{2})}$$

where $\tan(\gamma) = \frac{x_0}{\sqrt{D_0^2 + y_0^2}}$ and $\sin(\alpha) = \frac{x_2 - x_1}{\sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2 + D_2^2}}$

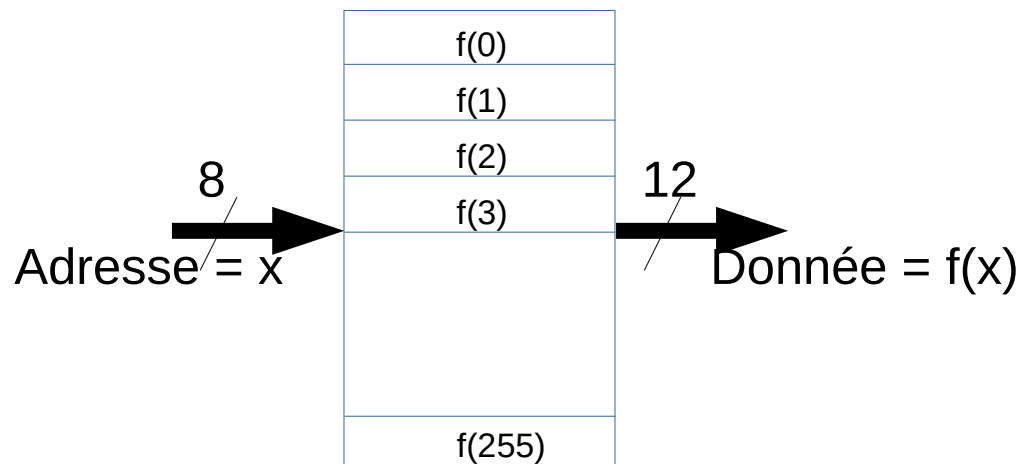
- $\sin(\theta)$ is given by:

$$\sin(\theta) = \frac{R_0}{\sqrt{(R_0^2 + D_0^2)}} = \frac{\sqrt{x_0^2 + y_0^2}}{\sqrt{x_0^2 + y_0^2 + D_0^2}}$$

Calcul du P_T

Utilisation de LUT

- Permet de calculer n'importe quelle fonction de type $f(x)$ même complexe en un coup d'horloge



- Faisable tant que range (x) reste faible

Quelques chiffres

Trigger muon LHCb

- Avec 240 FPGAs interconnectés, ceci permet de réaliser **740 milliards** d'algorithmes de recherche par seconde.

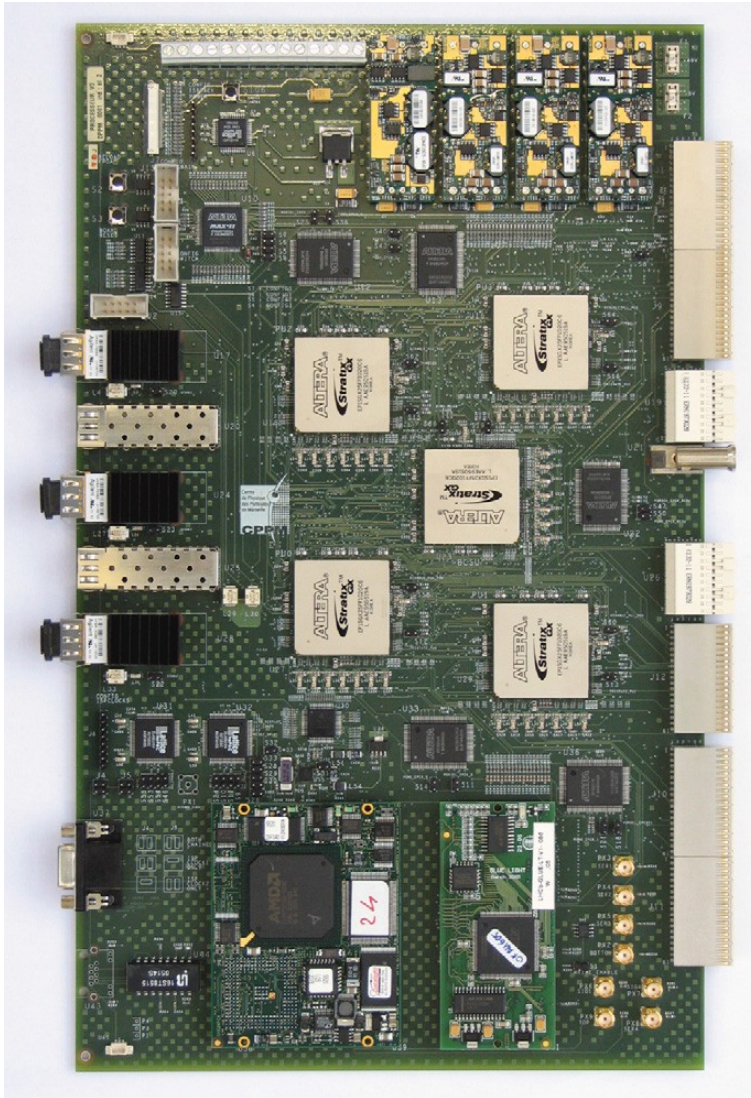


Testabilité

Primordial de comprendre les anomalies quand elles surviennent

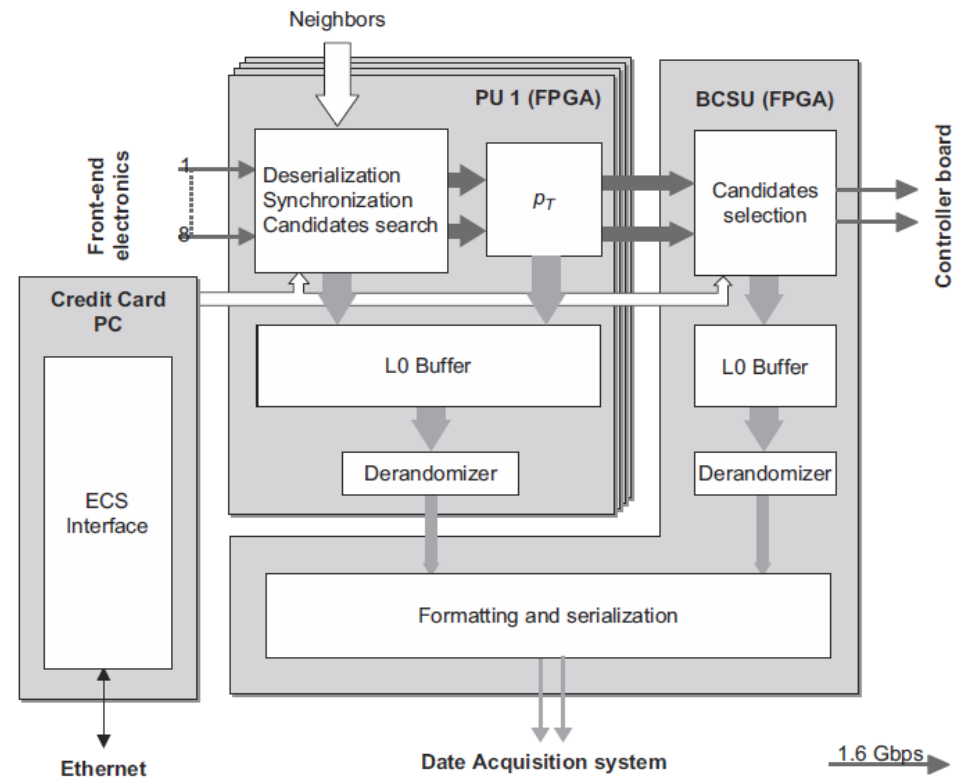
- L'algorithme ne représente pas plus de 50 % de l'occupation du FPGA
- Le reste est occupé par des fonctions de test et de monitoring
 - Injection de données simulées
 - Capture d'événements au vol
 - Relecture à différents endroits de la chaîne de traitement.

Carte trigger

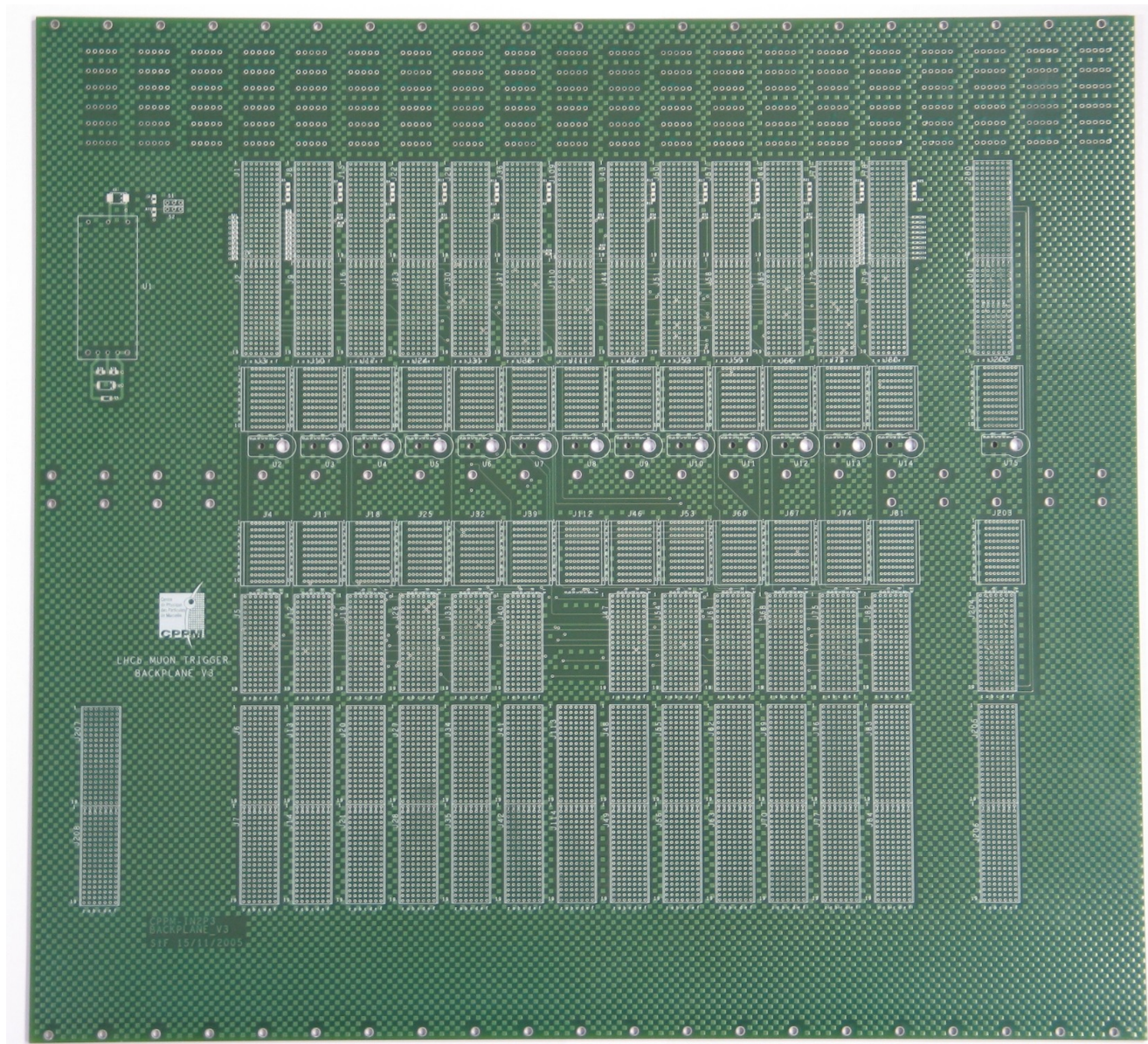


Carte générique

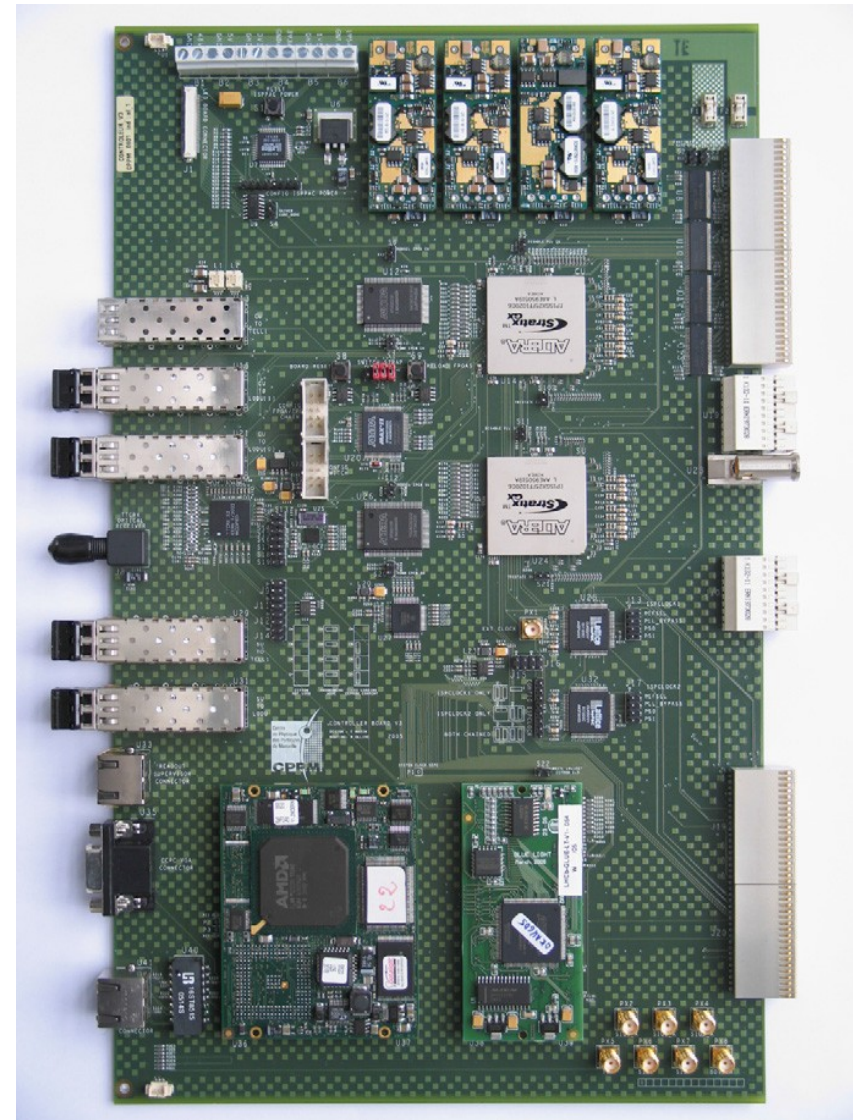
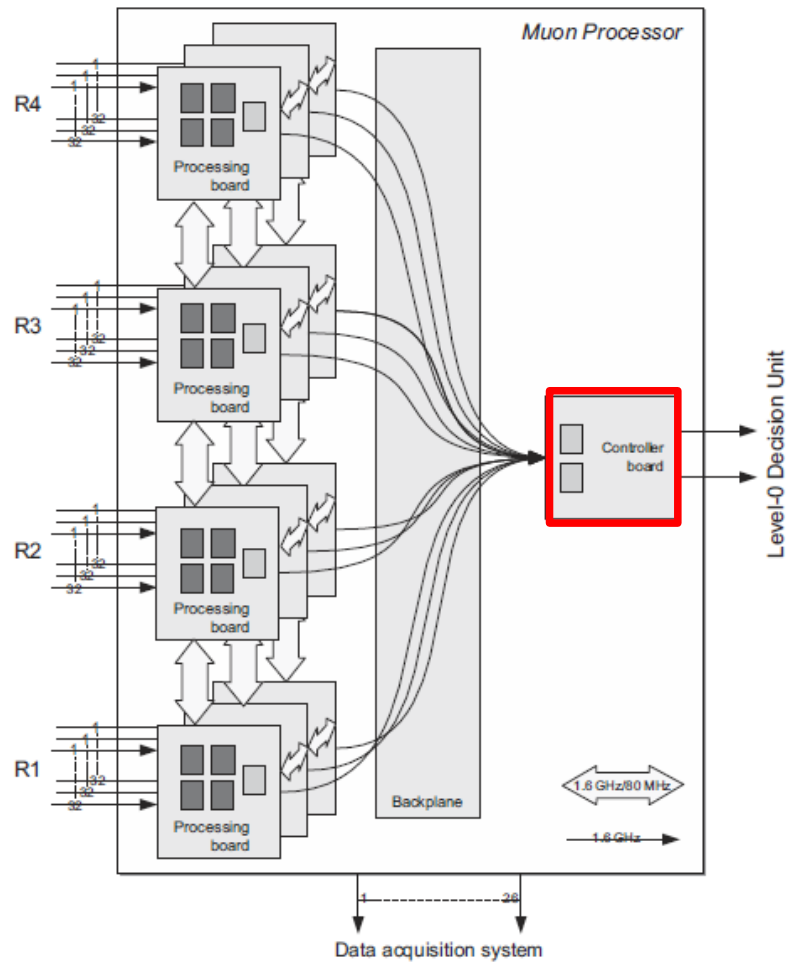
- Mais 48 configurations de FPGA différentes



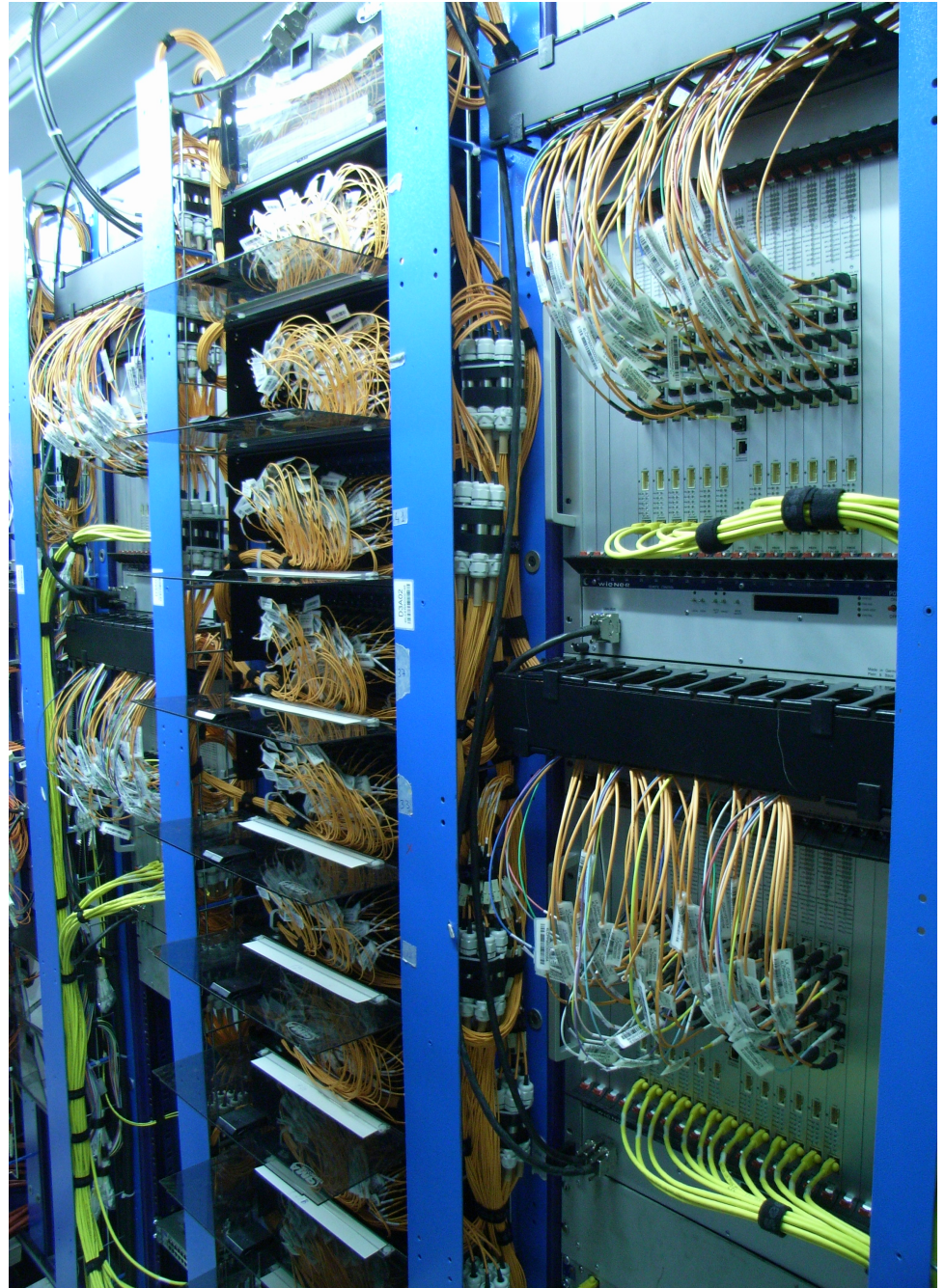
Custom backplane



Carte de contrôle



Le trigger à muons

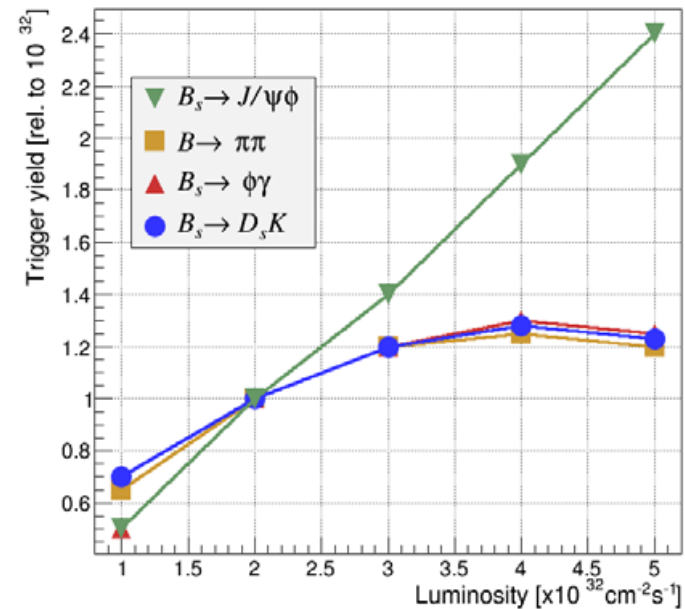


Evolution du détecteur : l'upgrade

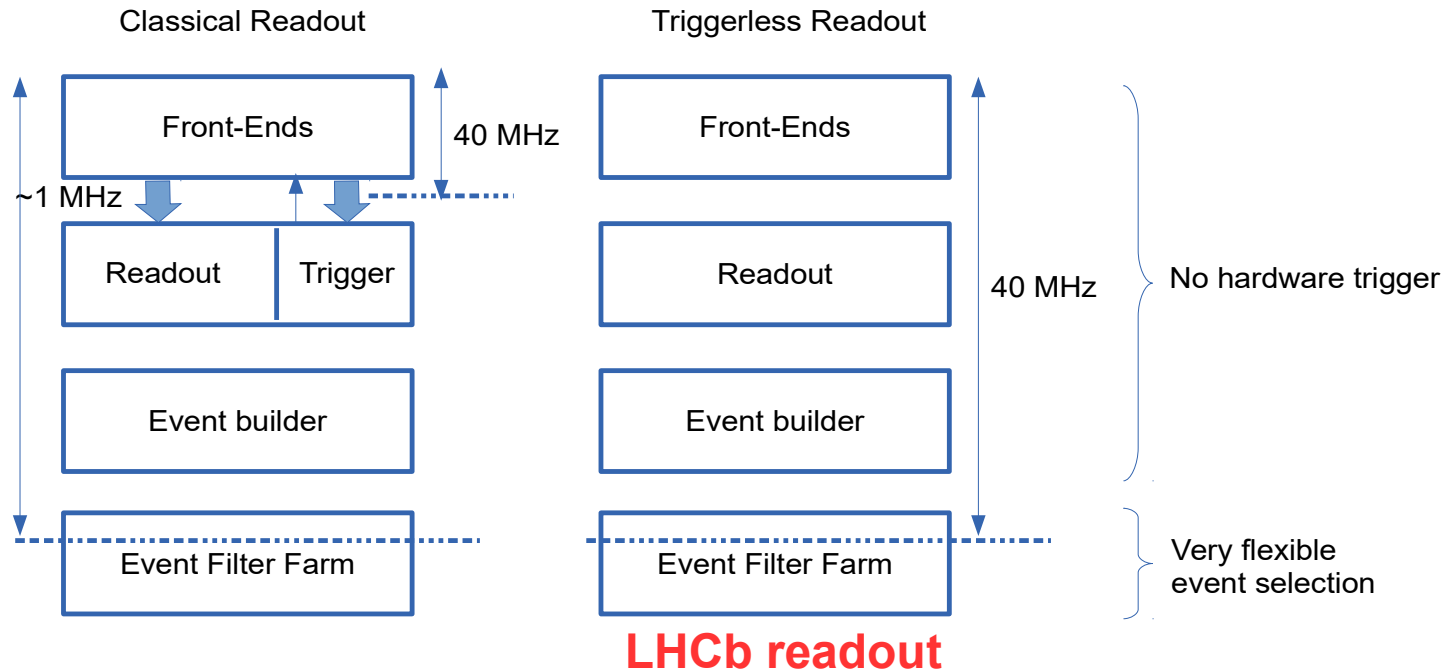
LHCb Upgrade

Motivation

- Luminosité maximale sous 5 ans : 5 fb^{-1}
- Au rythme actuel la précision statistiques des mesures varie très lentement
- En augmentant la luminosité de 2×10^{32} à $10^{33} \text{ cm}^{-2}\text{s}^{-1}$
 - Parvenir à une luminosité cumulée supérieure à 50 fb^{-1}
- Saturation du trigger sur les canaux hadroniques



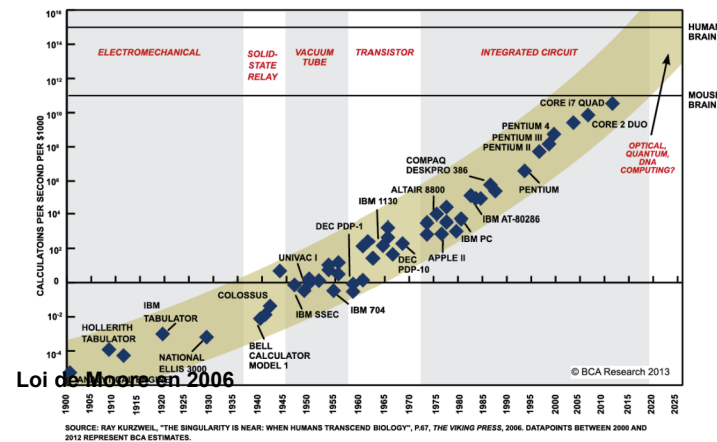
Principe de l'upgrade LHCb



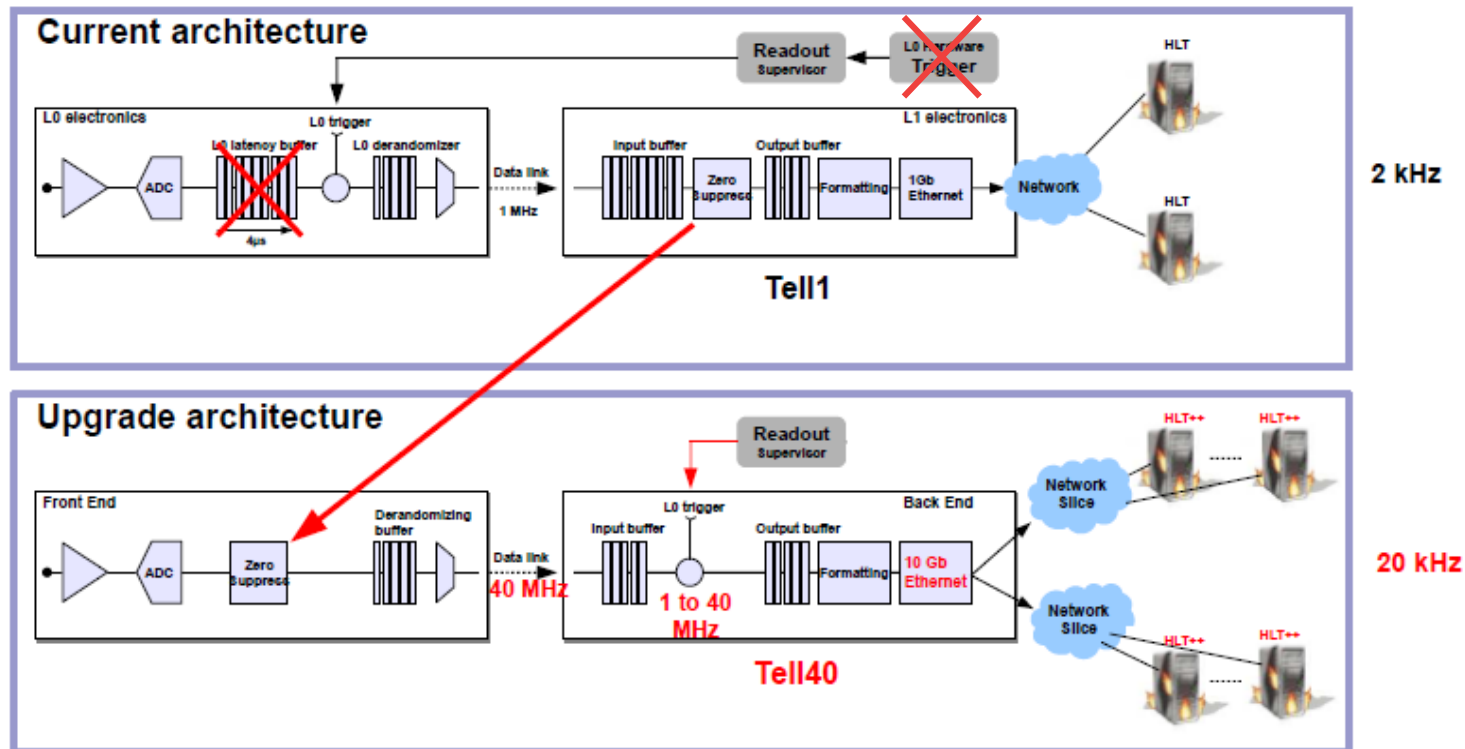
- Triggerless readout
- Tous les fragments d'événements sont routés à 40 MHz vers la ferme

Loi de Moore

- Si la ferme peut traiter les événements à 1 MHz en 2008
Elle doit pouvoir traiter à 40 Mhz entre 2018 et 2020



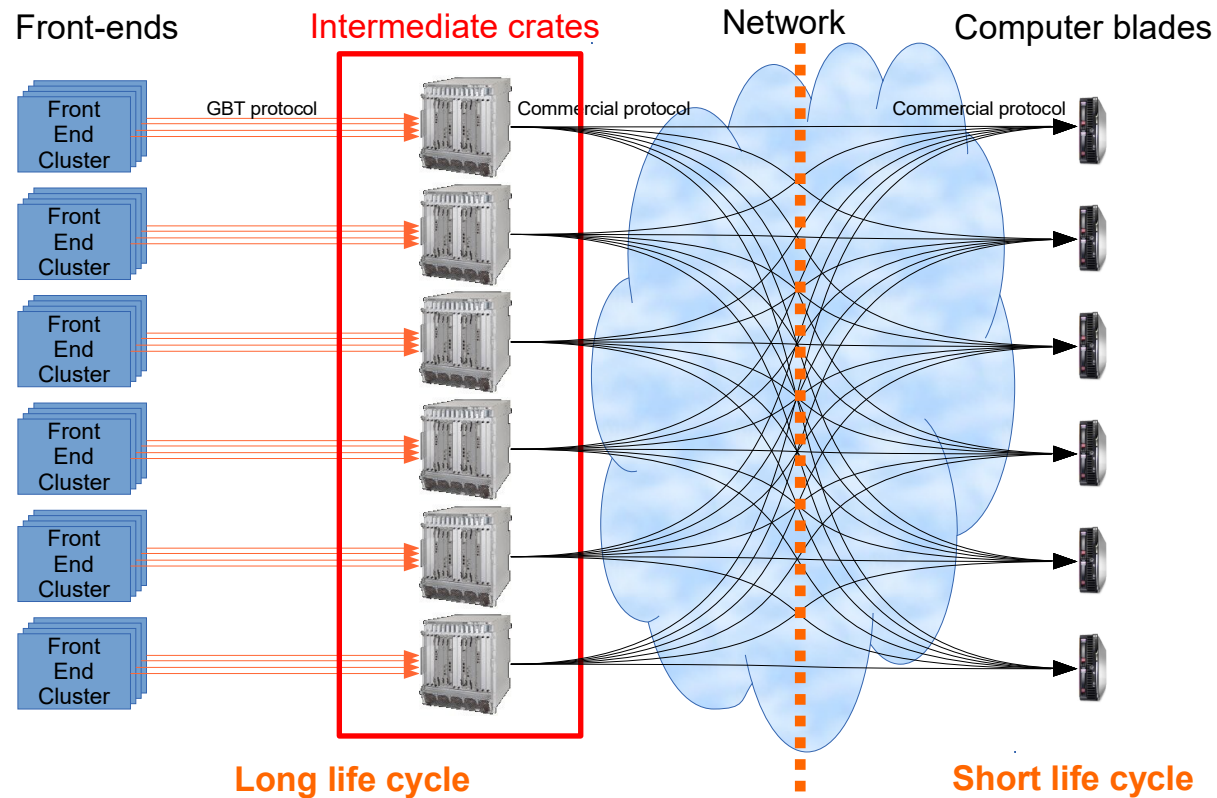
Upgrade : triggerless readout



- Compression dans les front-ends pour diminuer le nombre de liens optiques
- Débit de readout multiplié par 40

Choix d'architecture initiale du readout LHCb

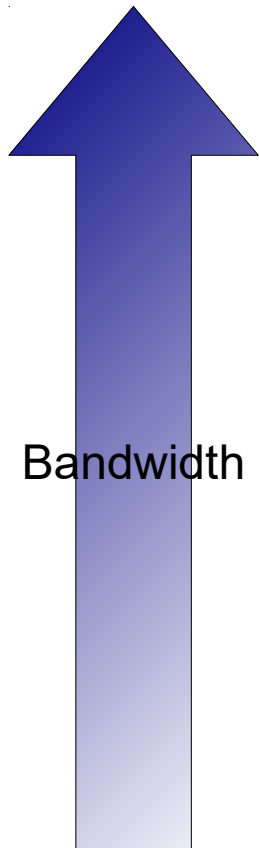
Systeme éprouvé distinguant le back end des fermes à courte durée de vie



Standard mécano-électrique

VME vieillissant

→ Besoin d'un nouveau standard

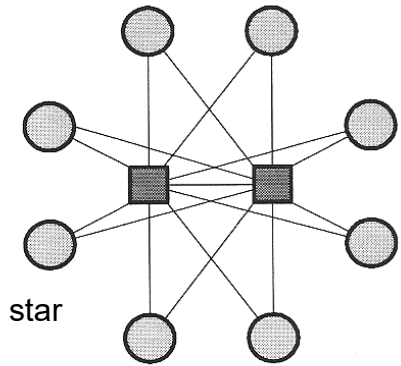


Standard	Bandwidth in Mbytes/s
ATCA 40Gb	1 820 000
ATCA 10Gb	455 000
VPX (VITA46)	112 500
VXS (VITA 41)	20 000
SHB Express	17 500
Compact PCIe/PSB	5 000
PCI 64 x 33 Mbits/s	533
VME 320	320
VME64x	160
PCI 32 x 32 Mbits/s	133
VME64	80
VME32	40
VME16	20

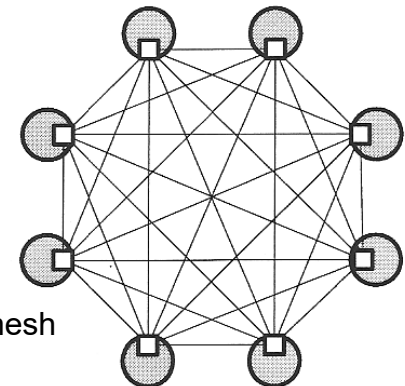
Plus de backplane custom

Utilisation du standard ATCA

- Nombreux avantages :
 - Bien adapté aux composants récents
 - Plus de place pour les radiateurs
 - Alimentation jusqu'à 3kW/crate
 - Refroidissement adapté
 - Backplane standard
 - Topologie basée sur des liens sériels
 - Mezzanines normalisées
 - Coûts similaires au VME
 - Redondance
 - Système normalisé de surveillance de l'état du système (IPMI)

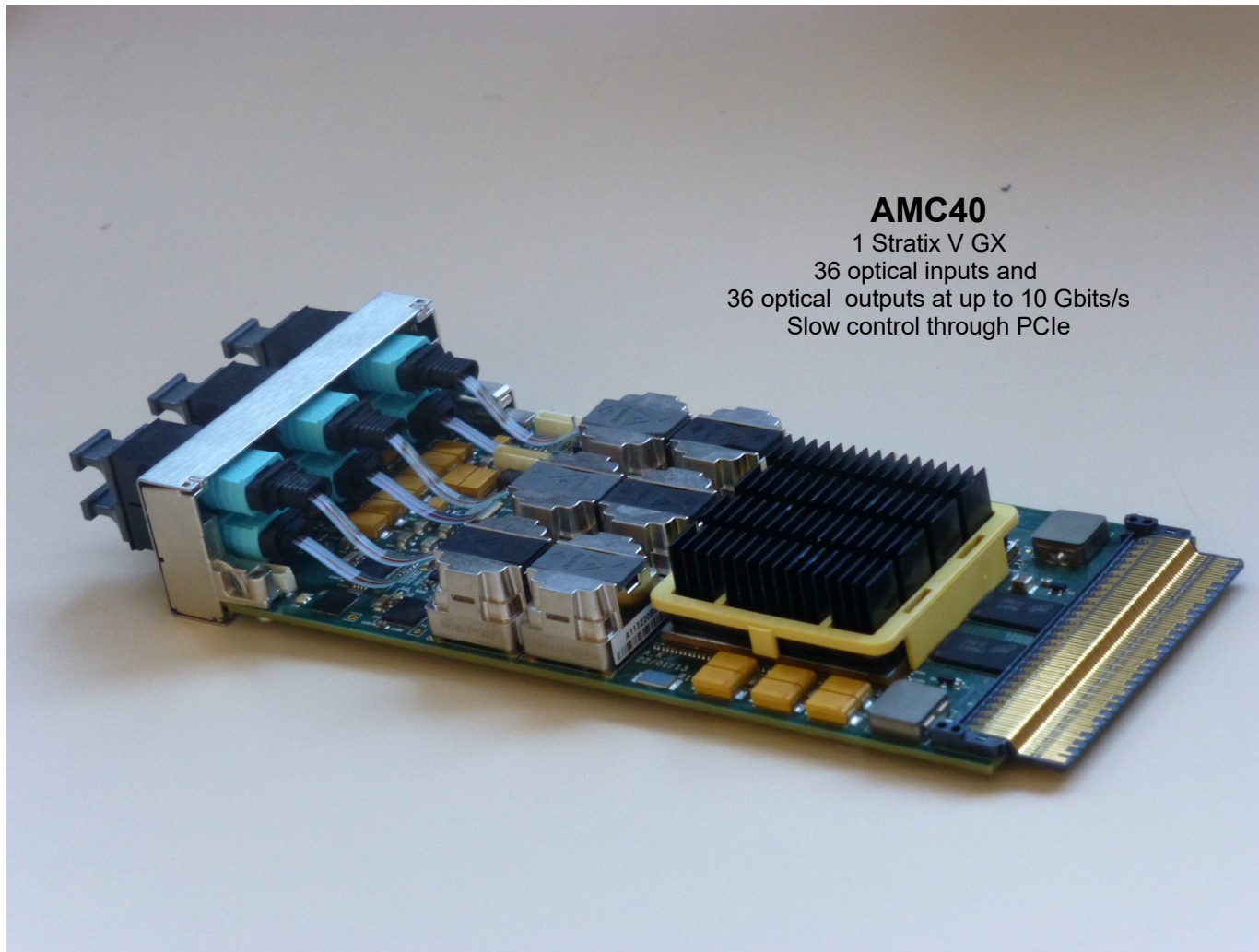
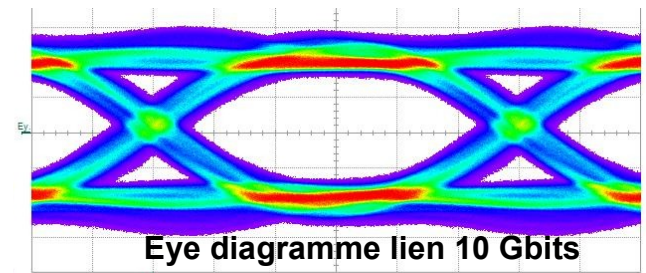


Topologie dual star

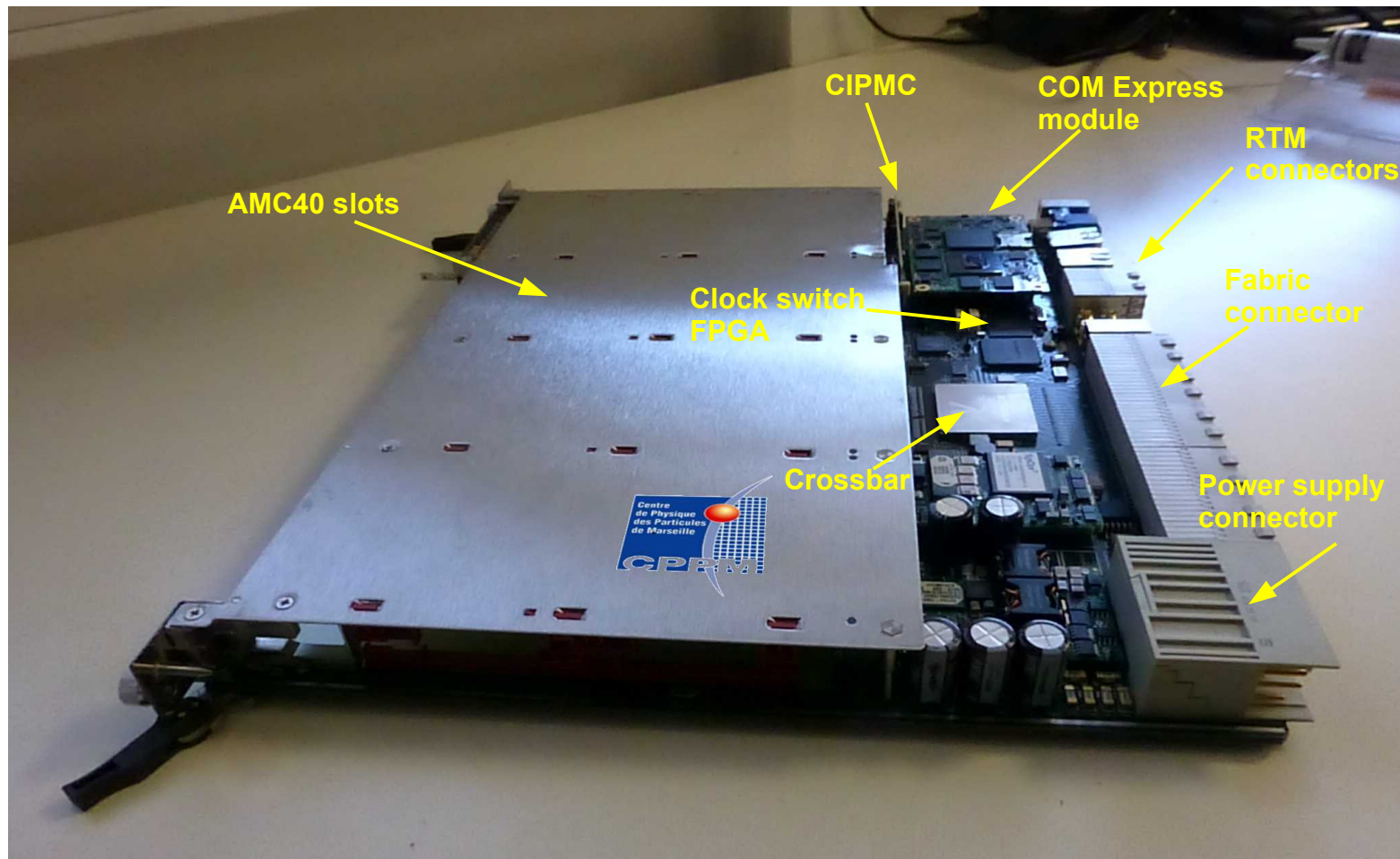


Topologie full mesh

Carte AMC40



Carte ATCA40



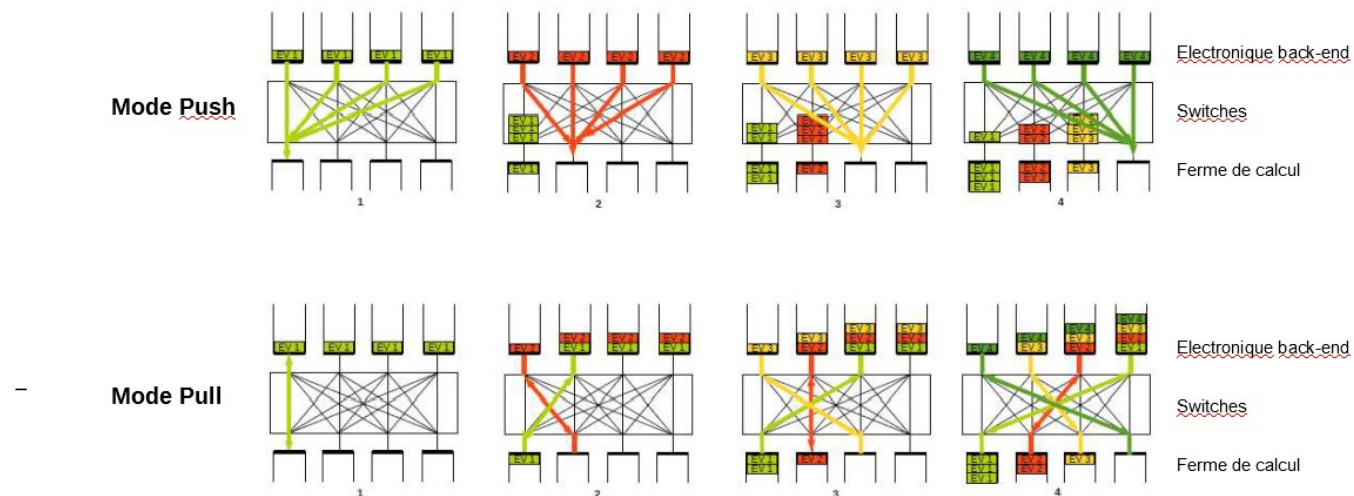
Problème mémoire

Architecture Push

- Requiert mémoire dans les switches
- Coût des switches très élevés (facteur 3)

Architecture Pull

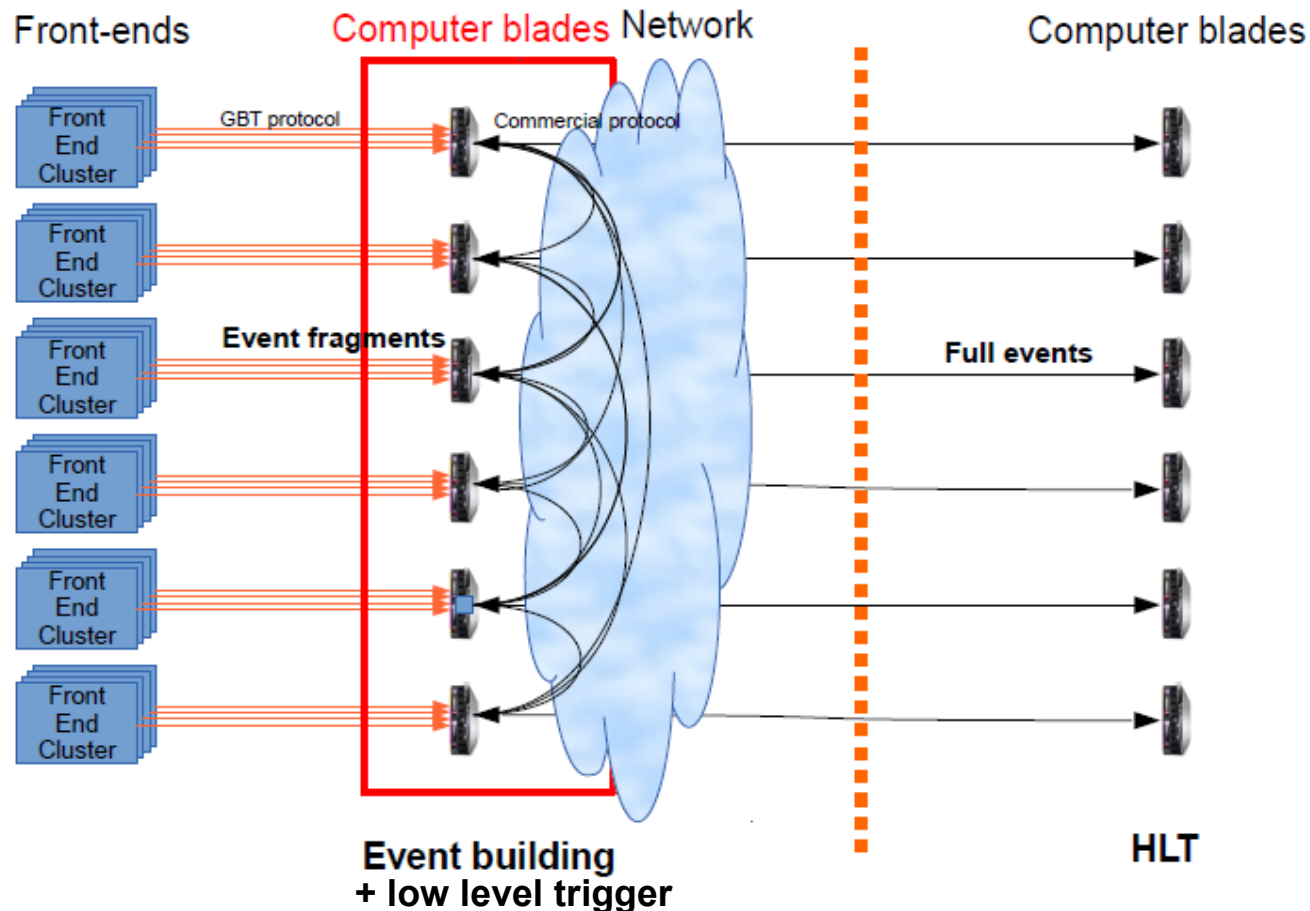
- Impossibilité de rajouter de la mémoire sur les cartes



Nouveau schéma de readout

Déplacement des FPGAs back-ends dans les fermes de calcul

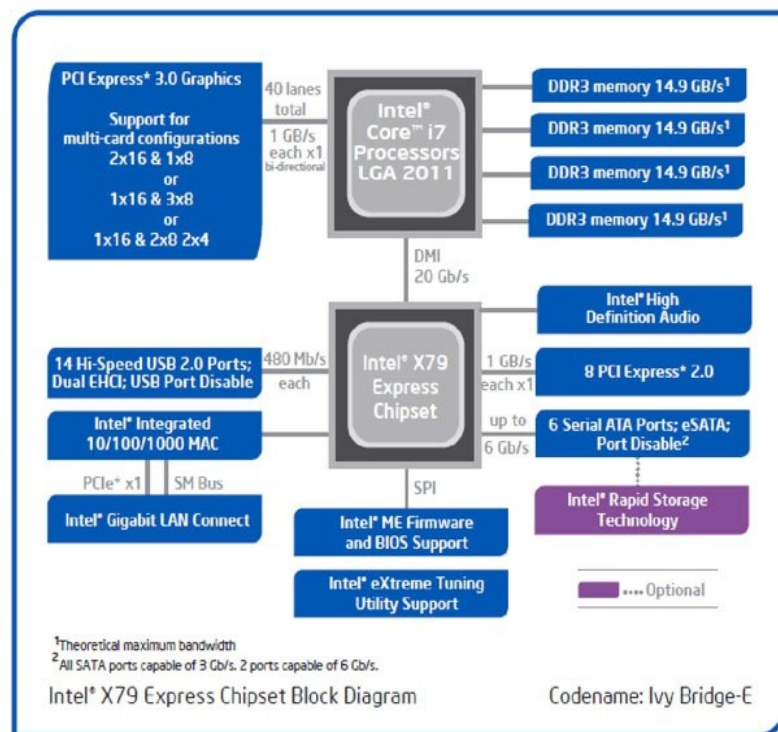
- Utilisation de la mémoire des CPUs pour bufferiser les données
- Implique également de réaliser l'event building dans les serveurs



Amélioration architecture interne des CPUs

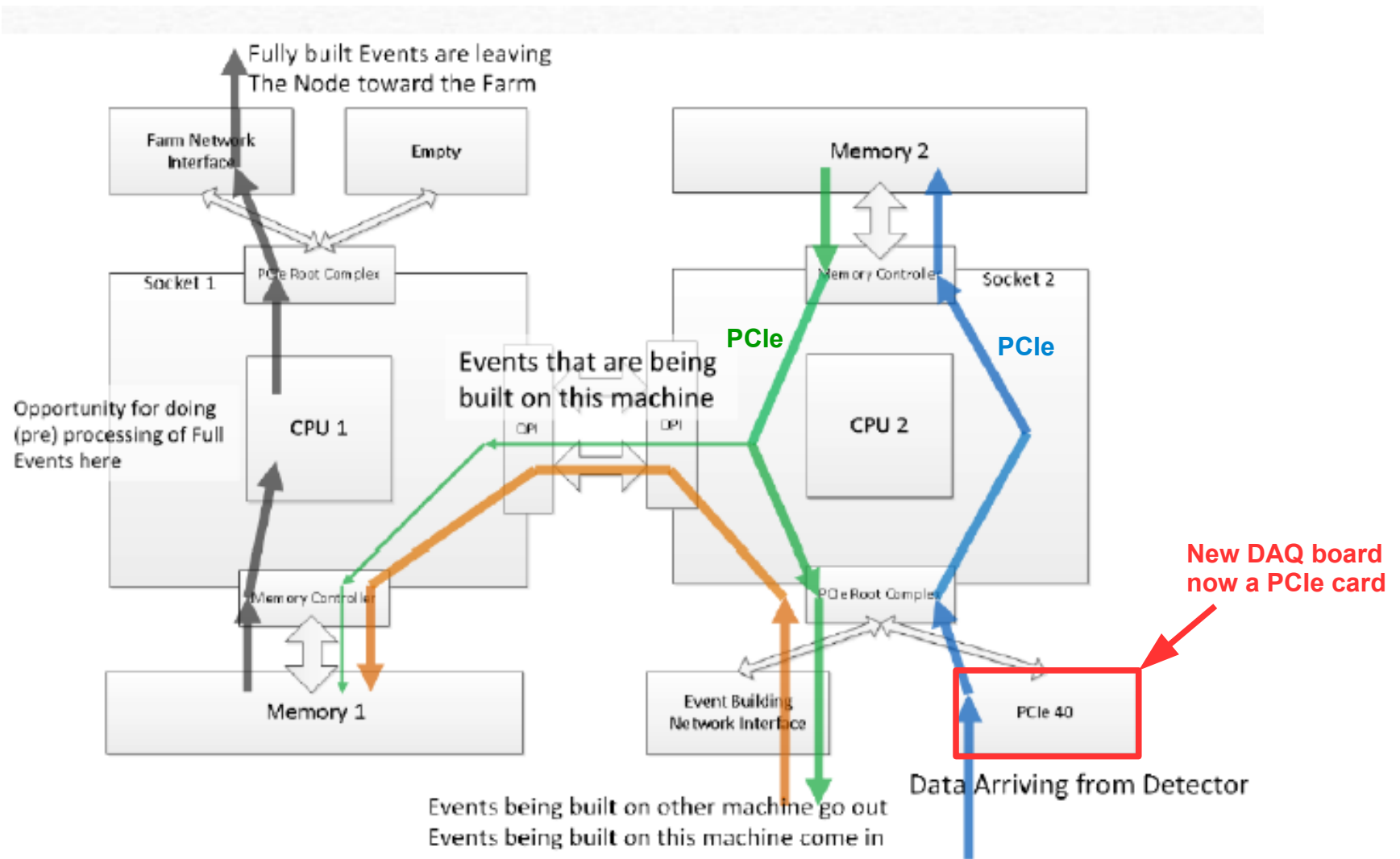
Libération de la bande passante des CPUs à partir de la génération Ivy Bridge d'Intel

- 40 canaux PCIe GEN3 à 8 Gbits/s
- Accès à la mémoire ne passe plus par le processeur

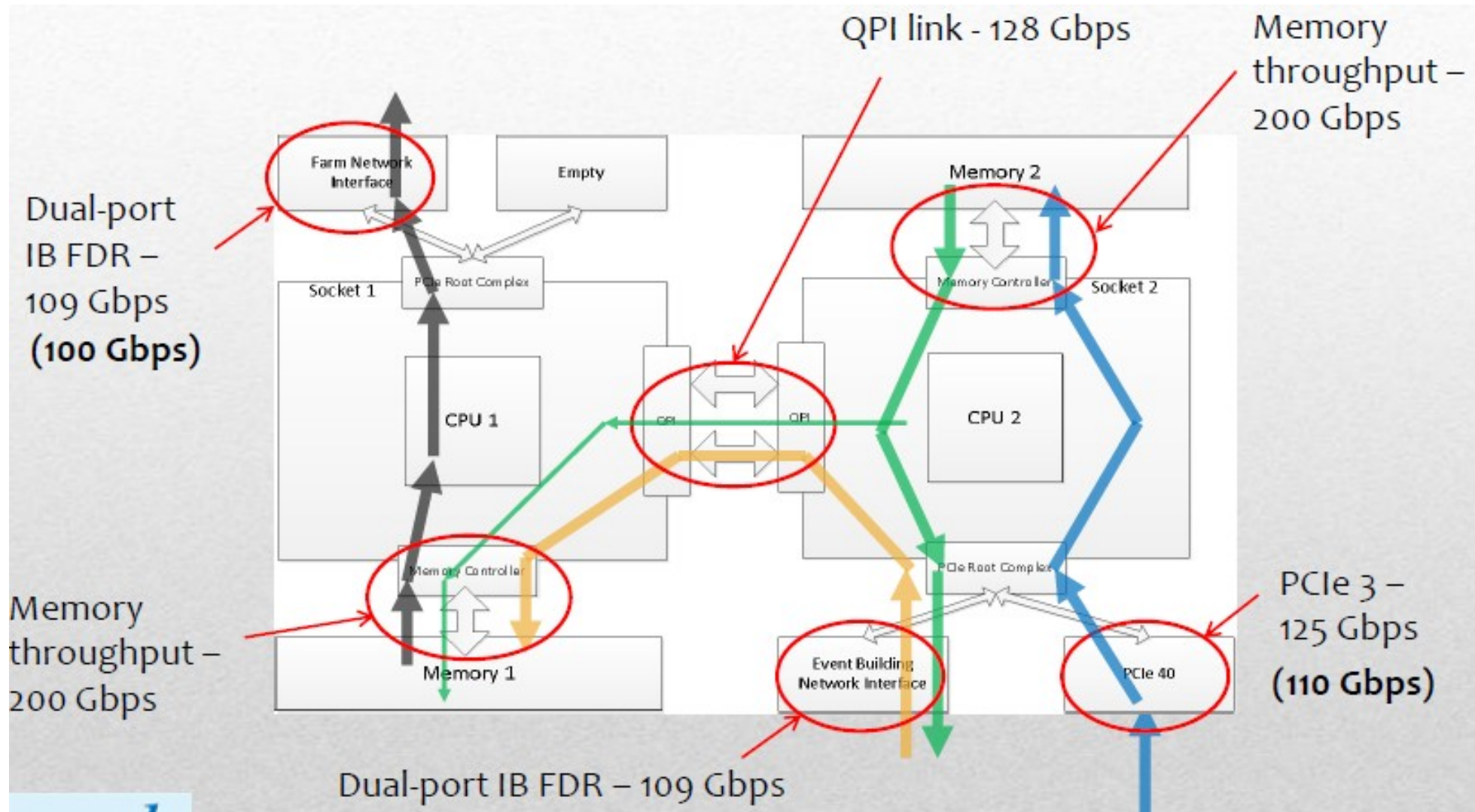


→ Donne la capacité au CPU de prendre en charge l'Event Building complet

Data path



Bandwidth



Avantages et inconvénients

Résout le problème de la bufferisation

Coûts

- Beaucoup de mémoire dans le CPU → carte d'acquisition plus simple et switches moins chers
- Plus de crates intermédiaires
- Moins de liens optiques
- Possibilité de faire tourner partiellement le HLT dans les CPUs d'event building
→ Plus de 80 % de la puissance inoccupée

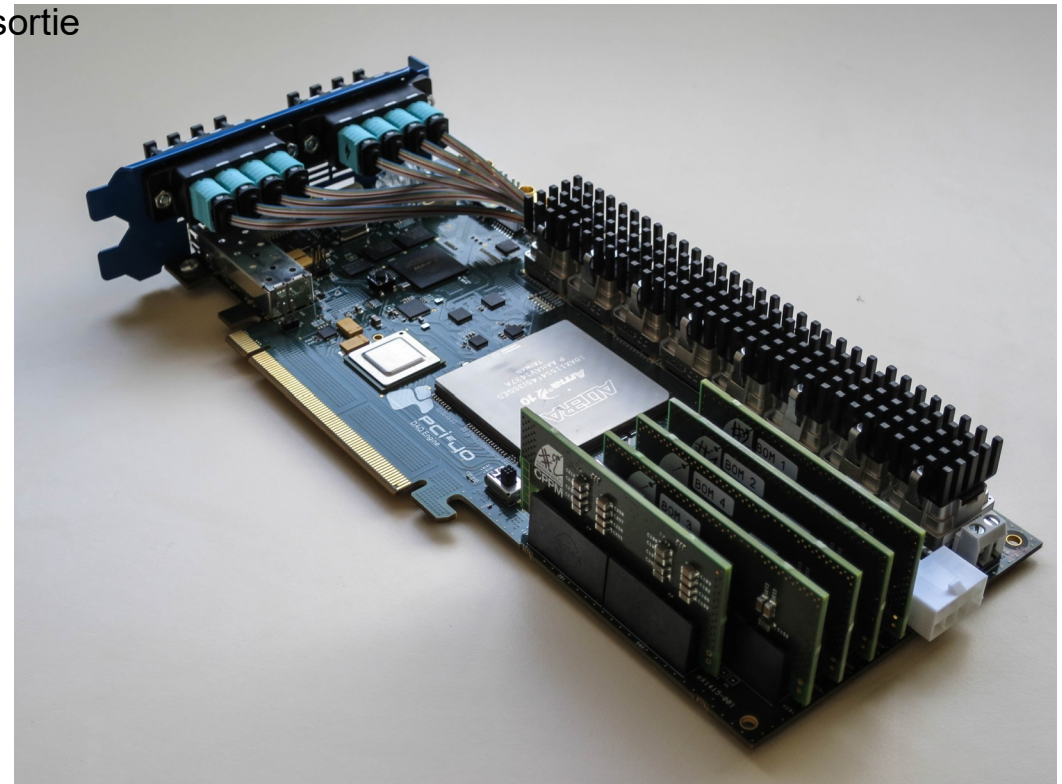
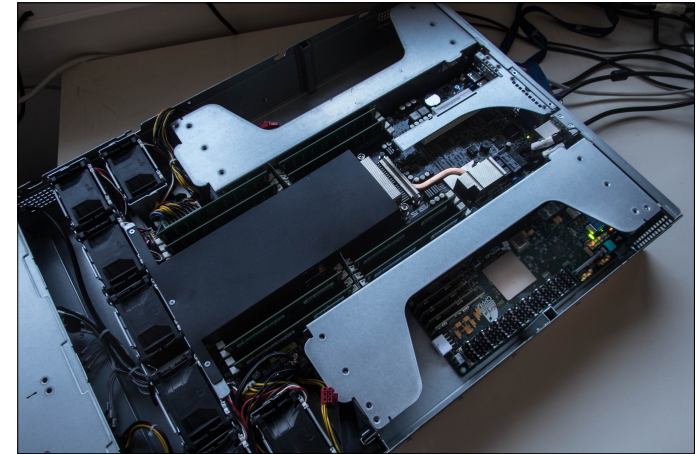
Durée de vie du système

- Vie moyenne d'un PC = ~4 ans (jusqu'à 8 selon statistiques du CERN)
Maintenance plus compliquée

Nouvelle carte de readout

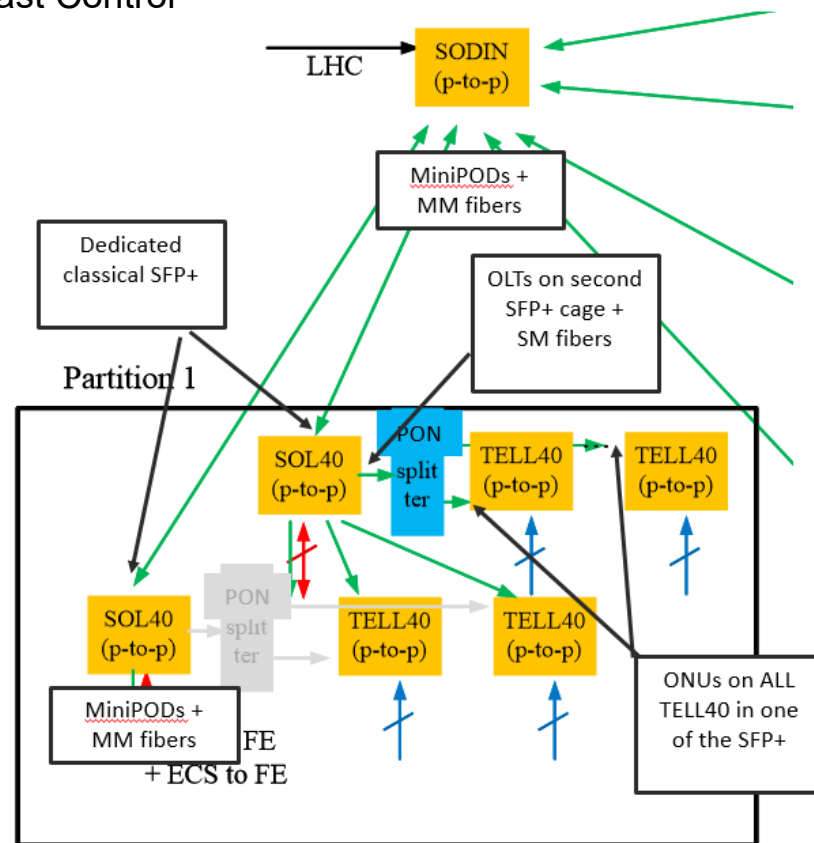
PCIe40

- Arria10 - Technologie 20nM - 1980 pins
- 1.15 millions de logic cells
- 72 liens 10 Gbits/s
- Bande passante :
 - Optique 500 Gbits en entrée, 500 bits en sortie
 - PCIe 100 Gbits en entrée et en sortie
- 50 fois plus de logic cells que FPGAs du trigger à muons
- 60 A sur le core sous 0.9V



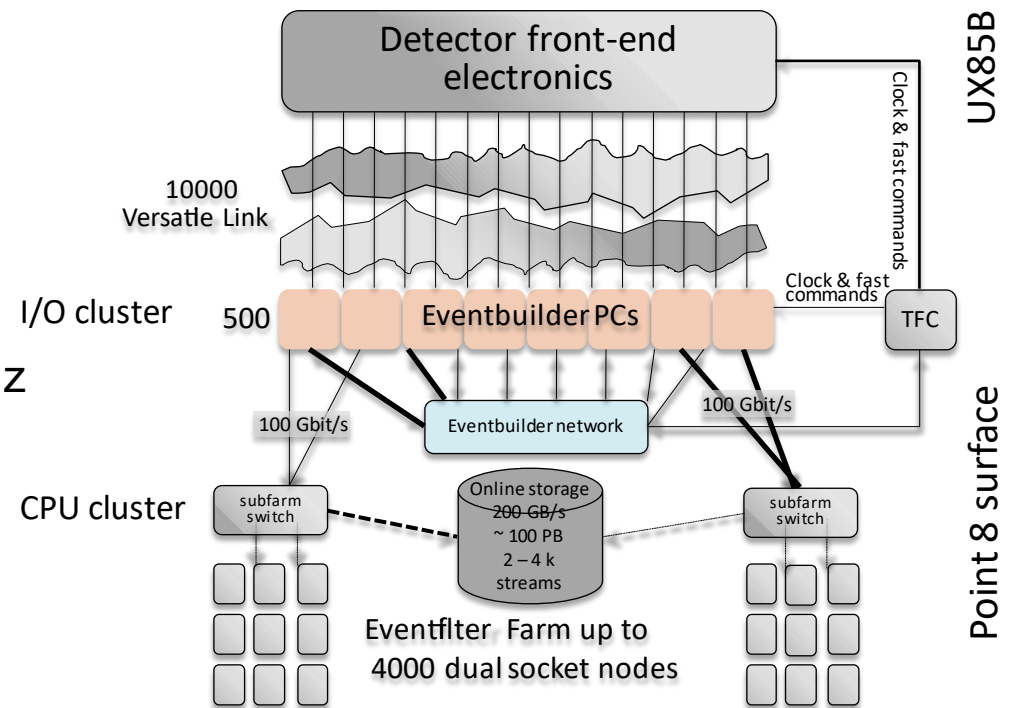
Architecture globale

- Peut assumer plusieurs fonctions dans le système par reprogrammation du FPGA
- Plusieurs noms selon sa fonction :
 - SODIN : Timing distribution and Fast Control
 - SOL40 : Slow control
 - TELL40 : Acquisition



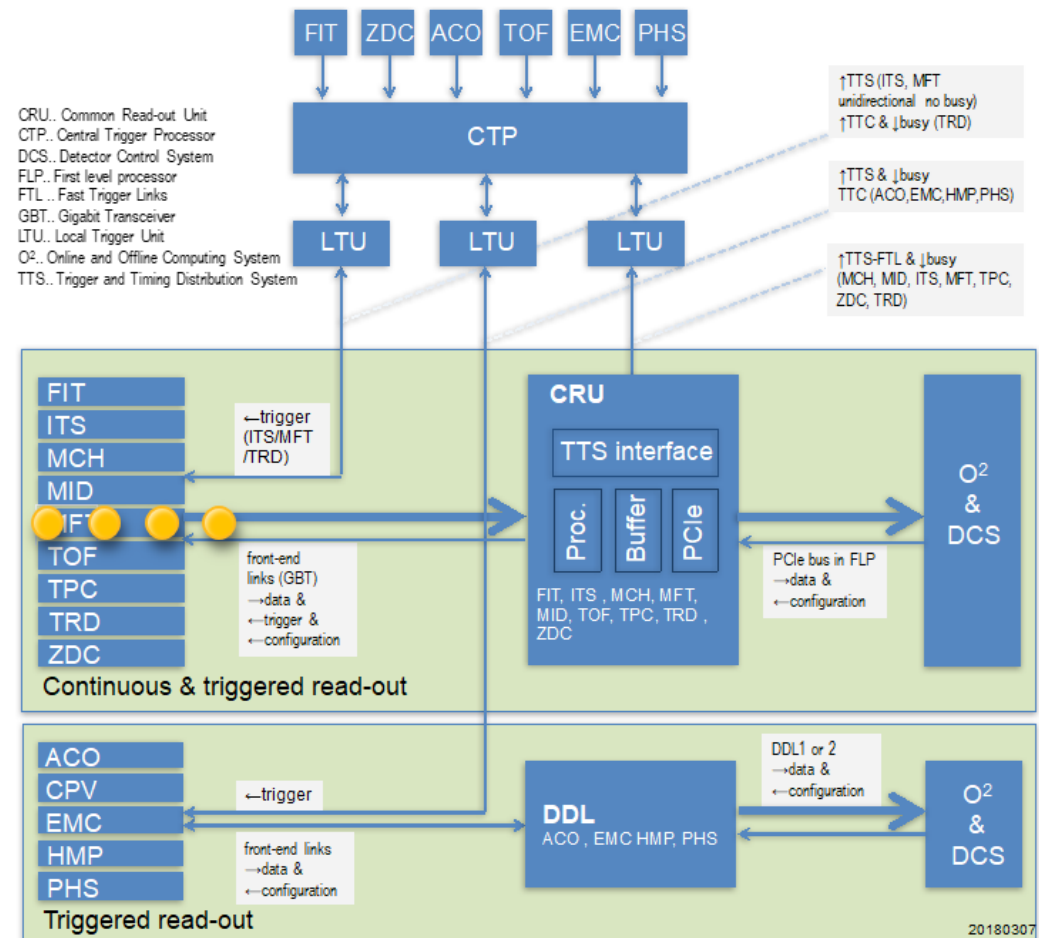
LHCb architecture

- Readout located on surface
 - o Distance between FE and RO : ~350m
- ~ 10000 optical links
- ~ 500 readout boards
- ~ 100 TFC/ECS cards
- ~ 100 kBytes per event at 40 MHz
- ~ 32 Tb/s aggregate bandwidth
- ~ 4000 dual CPU nodes



ALICE architecture

- Readout located on surface
 - o Distance between FE and RO : ~120m
- ~ 9000 optical links
- ~ 540 readout boards
- ~ 68 MBytes per event at 50 KHz
- ~ 27 Tb/s aggregate bandwidth
- ~ 1500 GPU based event processing nodes



Courtesy Alex Kluge

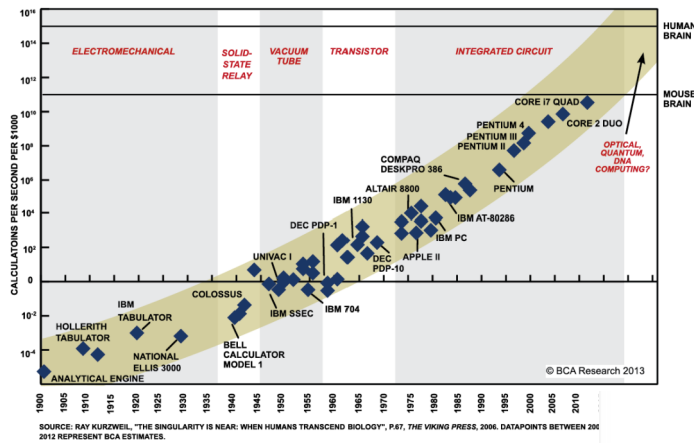
Trigger ou triggerless ?

Choix effectués pour les systèmes de readout du CERN

	ALICE	LHCb	CMS	ATLAS
Hardware trigger	No	No	Yes	Yes
Software trigger input rate	50 kHz Pb-Pb 200 kHz p-Pb	30 MHz	500/750 kHz for PU 140/200	0.4 MHz
Baseline processing architecture	CPU/GPU/FPGA/ Cloud&Grid	CPU farm (+coprocessors)	CPU farm (+coprocessors)	CPU farm (+coprocessors)
Software trigger output rate	50 kHz Pb-Pb 200 kHz p-Pb	20-100 kHz	5-7.5 kHz	5-10 kHz

Tenue du temps réel

- Loi de Moore

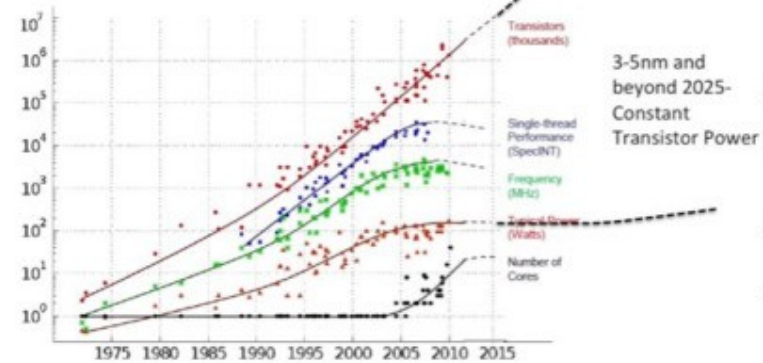


Loi de Moore en 2006

sauf que :

The trends will change past End of Moore's Law

35 YEARS OF MICROPROCESSOR TREND DATA



- ~2004 Dennard scaling, perf+ = single thread+ = transistor & freq+ = power+
- 2004~2015 feature scaling, perf+ = transistor+ = core# +, constant power
- 2015~2025 all above gets harder
- 2025~ post-Moore, **constant feature&power = flat performance**

End of massive parallelism and many-core speedup eras

Temps réel

- Le calcul pur ne suffit plus
 - Utilisation d'un ensemble de technologies
 - Calcul GPU sur carte graphique
 - Calcul GPU sur FPGA
 - Coprocesseurs
 - Neuronal
 - Deep learning
 - Vectorisation (SIMD)
 - Parallélisation des algorithmes
 - Many-cores

Conclusion

Tendances

- Architectures :
 - apparition des concepts triggerless
 - requiert un travail considérable d'optimisation pour maintenir les coûts
 - Données à 40 MHz : très challenging pour les switches
 - Pas forcément généralisable à tout type de machine
- Standards :
 - Quelle que soit l'architecture, adoption progressive du standard xTCA par les expériences
 - ATCA pour ATLAS, μ TCA pour CMS
 - mais aussi du PCIe
 - LHCb, Alice
 - Coexistence probable des deux types de solutions

	Event-size [kB]	Rate [kHz]	Bandwidth [Gb/s]	Year [CE]
ALICE	20000	50	8000	2019
ATLAS	4000	200	6400	2022
CMS	2000	200	3200	2022
LHCb	100	40000	32000	2019

Future DAQ in the LHC

Niko Neufeld, CERN

Pub !



Ecole « Technologies émergentes pour les Systèmes DAQ »

- Du 11 au 15 Novembre 2018
- Lieu Fréjus
- Programme :
 - Transmissions radios dans les détecteurs
 - Technologies optoélectroniques
 - Couplage efficace FPGA/CPU
 - Langages de haut niveau pour FPGAs
 - Technologies many-cores
 - Calcul GPU sur carte graphique
 - Calcul GPU sur FPGA
 - Réseaux neuronaux et deep learning
 - Réalisations effectives à l'aide de ces technologie
 - Comparaisons
- Présence d'Intel et Kalray