LHC Run3 に向けた高度化後 ALICE TPC の連続読み出し型データ収集システム

長崎総合科学大[A], 東大CNS [B], 原研[C]

大山 健[A], 荻野 雅紀[A], 田中 義人[A], 浜垣 秀樹[A], 郡司 卓[B], 佐甲 博之[C],

他 ALICE-TPC-CRUグループ

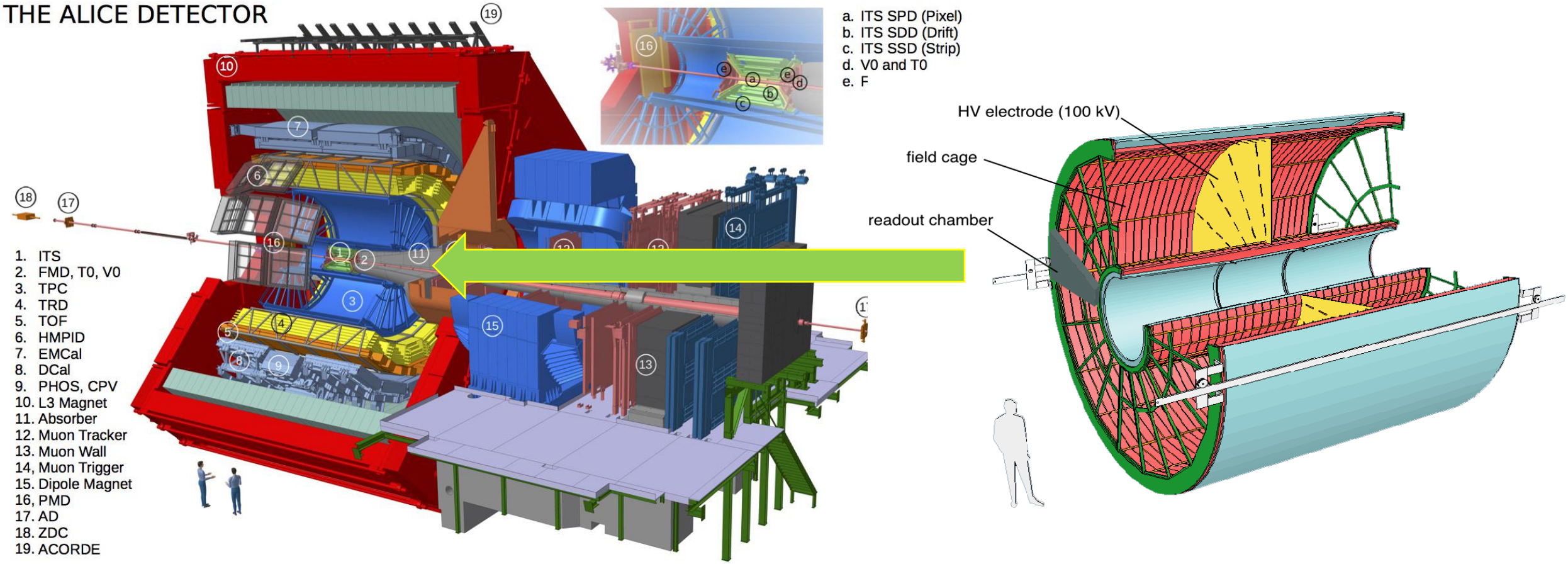
JP17H02903
JP16K13808
JP17K18783

Mar. 16, 2019, JPS Meeting@九州大学

LHC ALICE実験

■ 重イオン衝突実験に特化(高粒子密度: $dN/d\eta$~2000, 低運動量粒子: p_T>150 MeV/c)

■ 中心rapidity領域の主要トラッキングデバイスとしてTPC(世界最大規模・容量 88 m^3)を備える

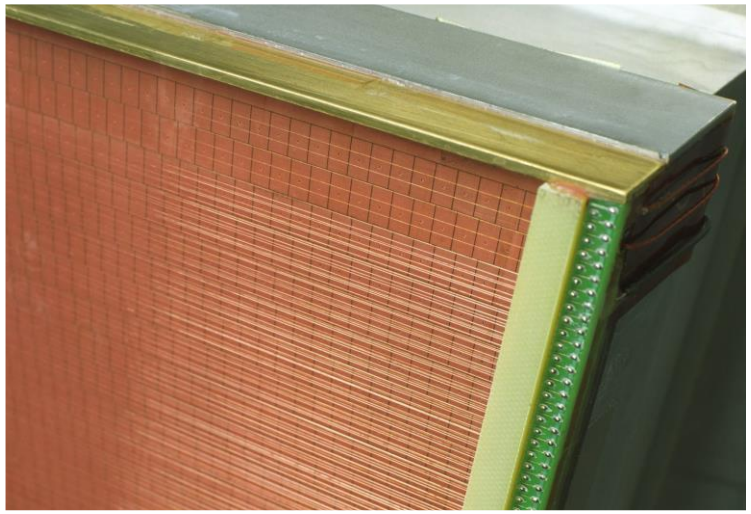


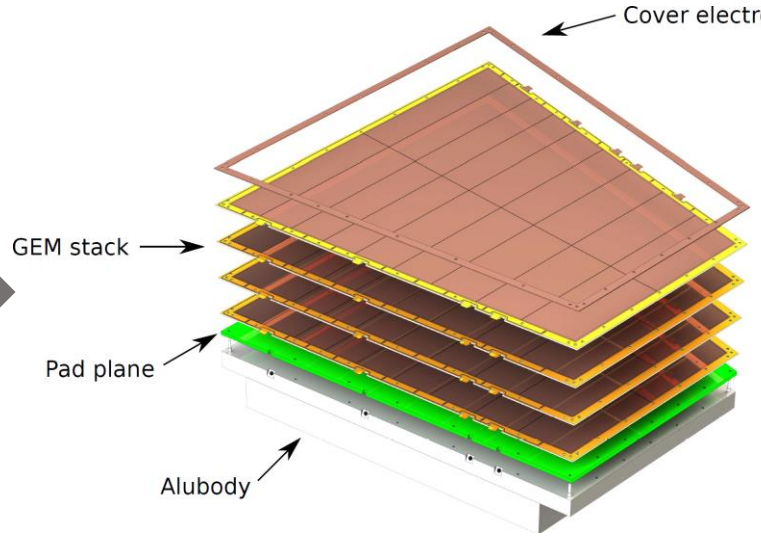
ALICE TPC アップグレード

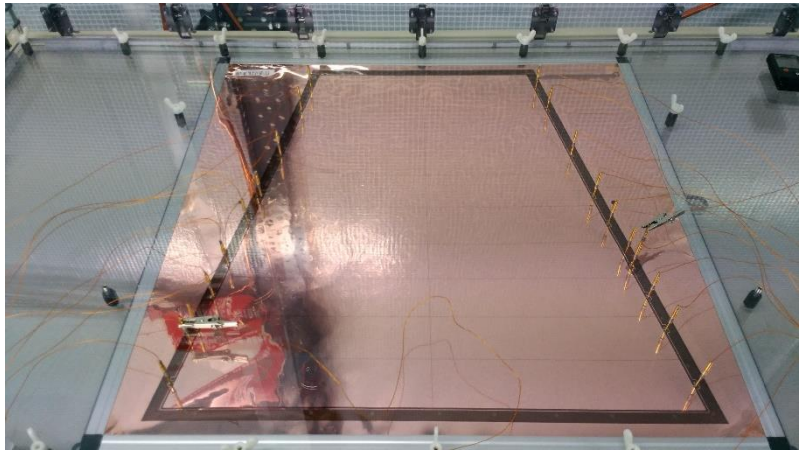
	従来型 (Run1+2; ~2018)	アップグレード後 (Run3+4; 2021~)
増幅・読み出し機構	MWPC・パッド読み出し	GEM (4段)・パッド読み出し
ドリフト時間+イオン除去時間	100 μs + 400 μs	100 μs + **0**
最大トリガレート	1.8 kHz (90% dead-time)	***no limit*（ドリフト終了を待たず連続）**
ADCサンプリング周波数	10 MHz	5 MHz
ADCチャンネル数	56万	56万
データ量(検出器→DAQ@Pb+Pb)	50 GB/s, 50 PB/月	3.5 TB/s, ~3.5 EB/月
データ量(on tape, Pb+Pb)	7~8 PB/月	< 100 PB/月

35:1

注) Pb+Pbでは [/年] = [/月]



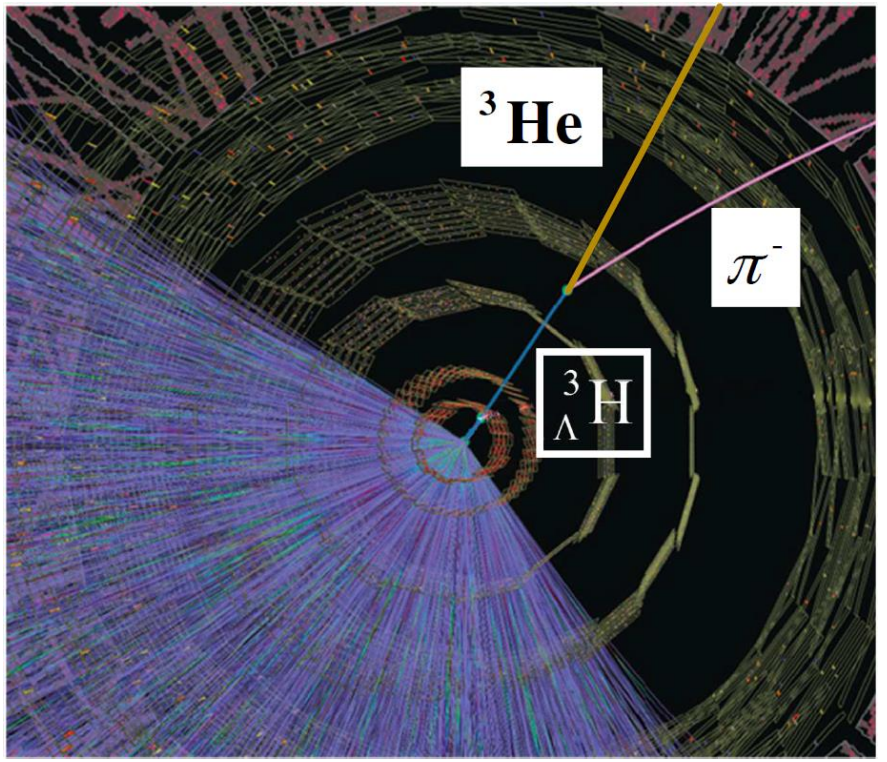




連続読み出し型DAQの必要性

従来

- ハードウェアでトリガ(minimum bias(MB), high-p_T, etc)を構成し面白いイベントのみ読み出す
- 年間データ量は数PB → 後日解析が容易
- pile-up も無く(<10 kHz)、Zero Suppression が容易



Run3以降の高輝度LHC重イオン衝突実験

- dN/dη~2000 → 低運動量レア-イベントは最早トリガできない
- 高度なイベント選定にはtracking+再構築が必須
- pile-up問題 (50 kHz → 平均 5 event が多重発生)
  - 平均 occupancy 30%
- Zero-Suppressionが出来ない(GEMによる, 後述)

結論　ADCデータを全て連続読み出しし、オンラインでデータリダクションを行うしかない

ALICE TPCのデータ処理システム

検出器
GEM+Pad

アナログ
(560k ch.)
数cmのワイヤ

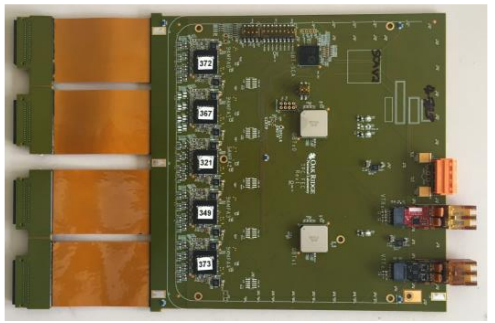
Front-End Electronics

amp

A/D
5 MHz
10 bit

DSP + Zero Suppression

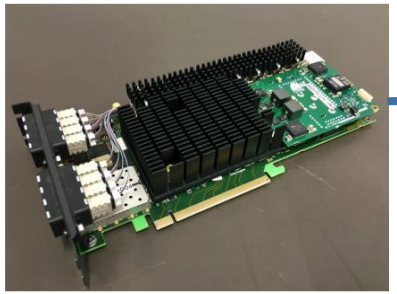
GBT



光ファイバ 4 Gbps x ~7000本 → 28 Tbps (3.5 TB/s), 距離~100 m

地下実験施設 ↑

地上 ↓



FLP(First Level Processor)ノード

CRU(Common Readout Unit)
FPGA Arria 10 (GTX1150, 10AX115S3F45E2SG)

x20
80 Gbps
(10 GB/s)
1600 ADCs

GBT
x20

ソート | DSP処理 | クラスタリング前処理 | フォーマット

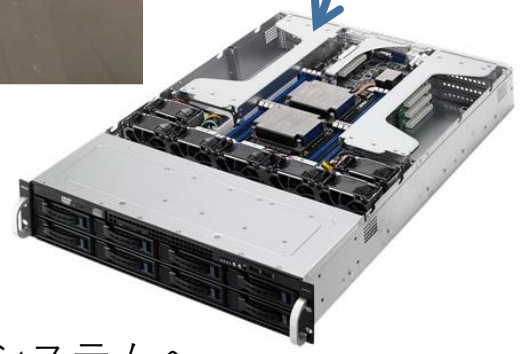
16 Gbps
(occupancy ~30%を仮定)

クラスタリング
後処理, (圧縮,)
フォーマッティング

x86 CPU

PCI Express Gen3

ネットワークIF

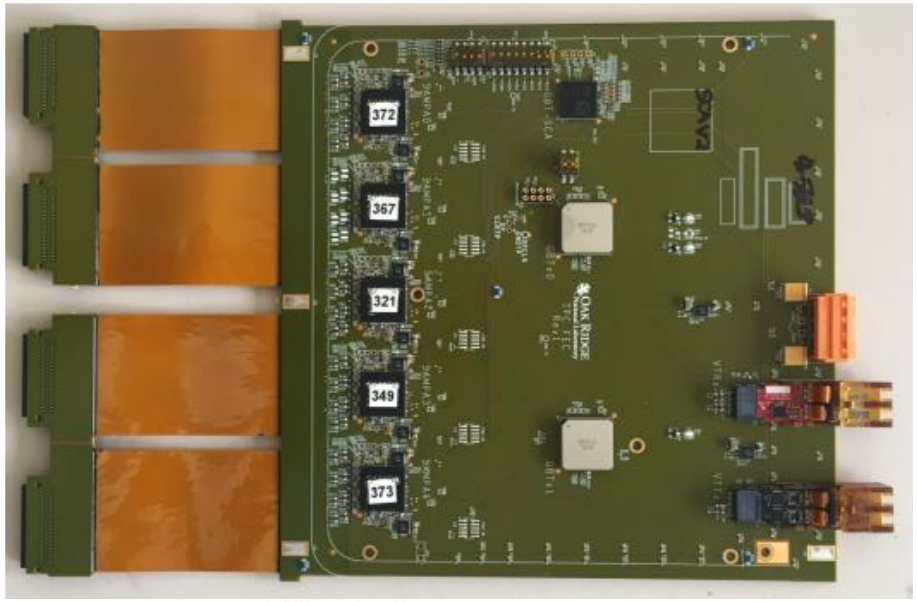


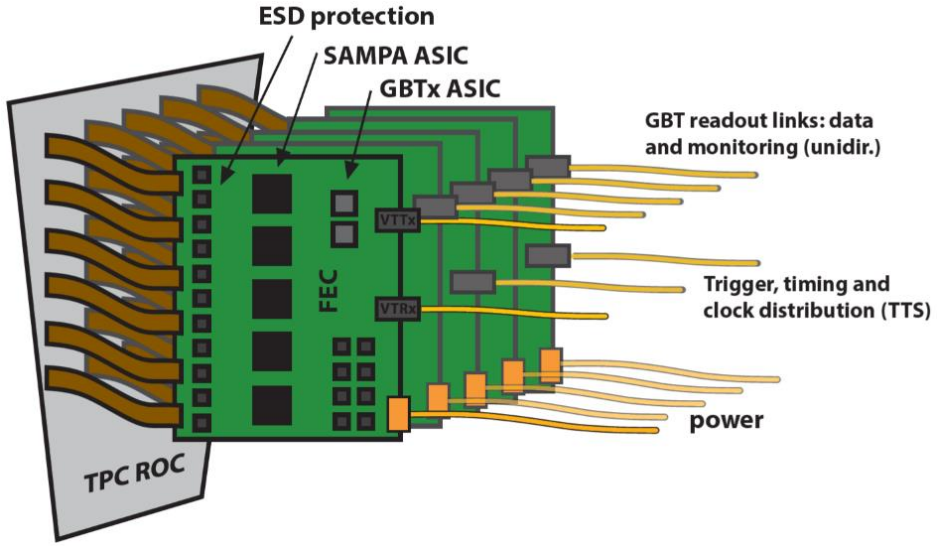
他のノードへ
全ノード数~360

後段システムへ
(calibration, tracking)

16 Gbps/ノード, 5.6 Tbpsトータル

TPC front-end readout





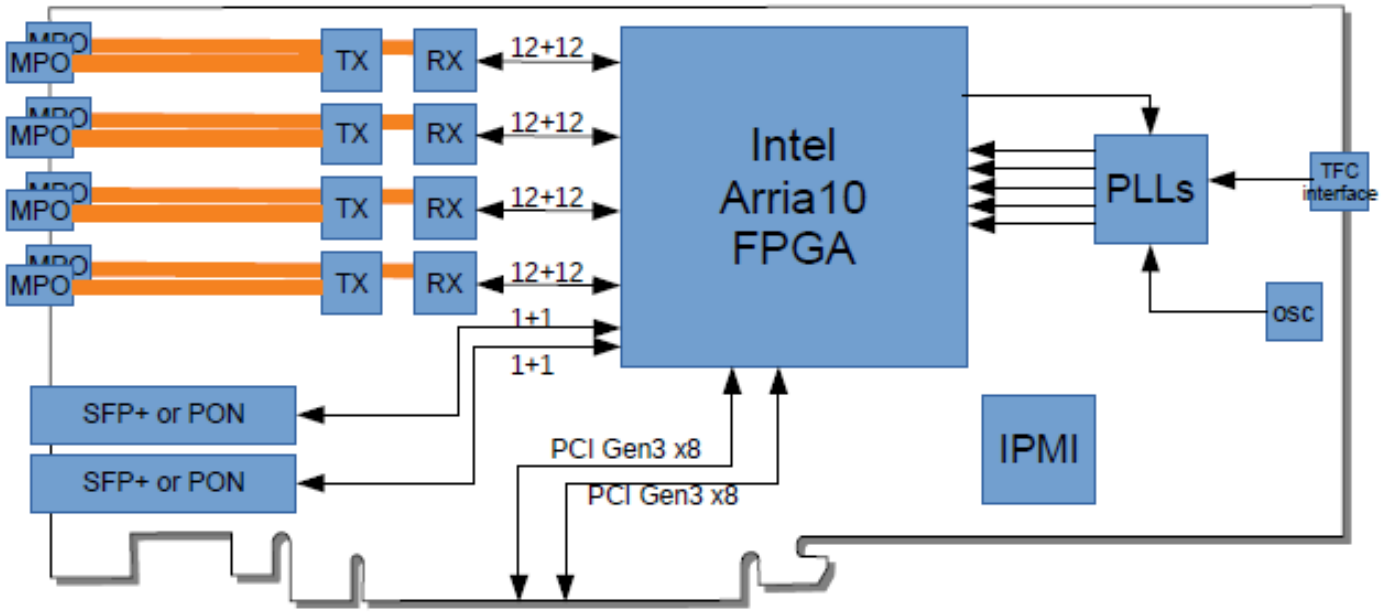
FEC (Front-end Card)

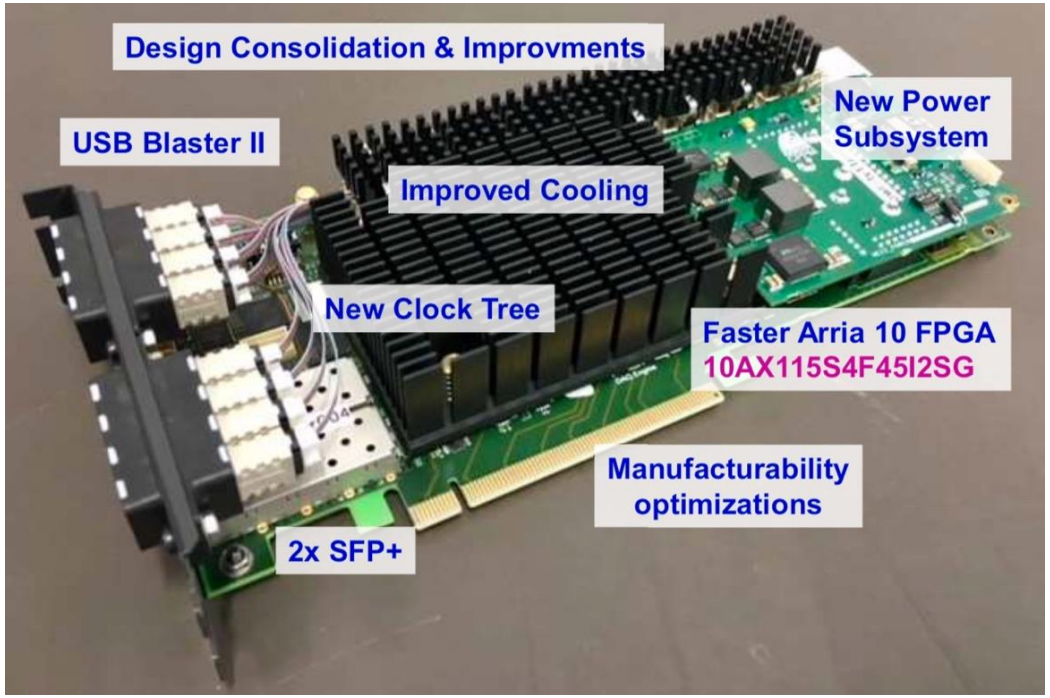
- 5 SAMPA ASIC (32 ADC, 連続読出) = 160 ADCチャンネル/カード
- 10 bit ADC, 5 MHz operation
- 2 GBTリンクでデータ出力: 4 + 4 Gbps
- 設計: オークジッリ国立研究所
- 3276枚のカードを生産・インストール中

CRU FPGA

ALICE共通データ前段処理FPGAボード

- LHCb+ALICEによる共同開発(HW製造: マルセイユ, TPC FW開発: 長崎総合科学大, ハイデルベルグ)
- 48 GBT duplexリンク → 最大4.48 Gbps x 48 = 215 Gbps
- Intel/Altera Arria 10 FPGA → Xeon CPU core 比 **240倍のアクセラレーション(TPC CF) (tbc)**
- 標準 PCI Express 3 x 16Lane バス(128 Gbps) → 実用最大~90 Gbps



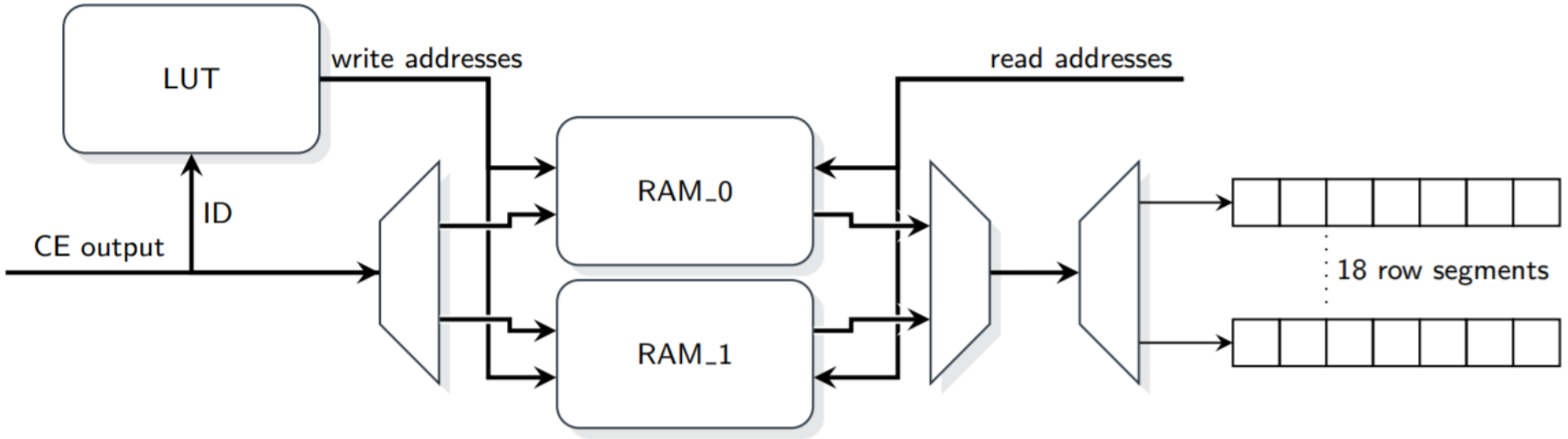


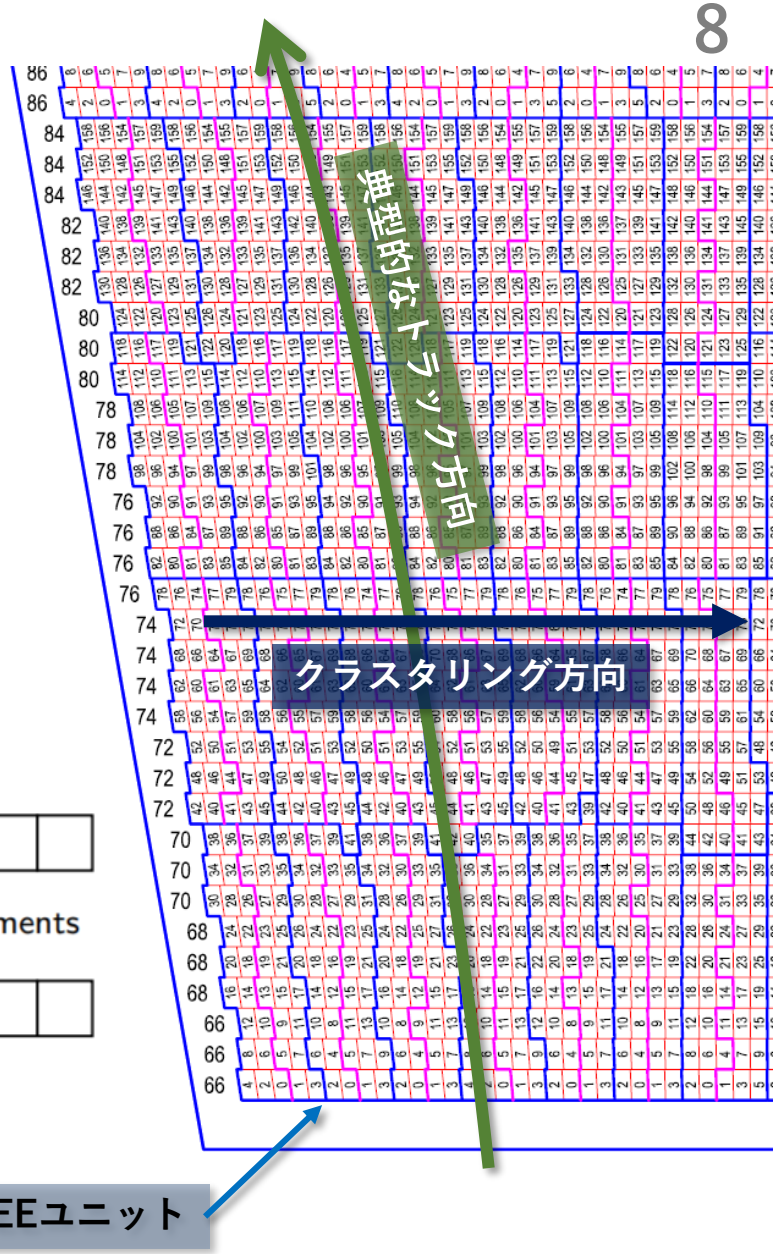
Drawing and photo by Kiss Tivadar

ソーティング

■検出器の形状とFEEの物理的制約から、SAMPA ASICのチャンネルの並びと、クラスタリングを行いたいパッドの方向が一致しない

- 巨大ルーティングマトリクス(1600-to-1600)をFPGA内に実装
- ノード毎に異なるconfigurableマトリクス
- FPGA内メモリを活用





Sebastian Klewin, Dec. 2017

コモンモード除去フィルタ

■ GEMとパッドプレーンは、平行平板コンデンサ
- 容量カップリングでコモンモードノイズ（クロストーク）が拭えない

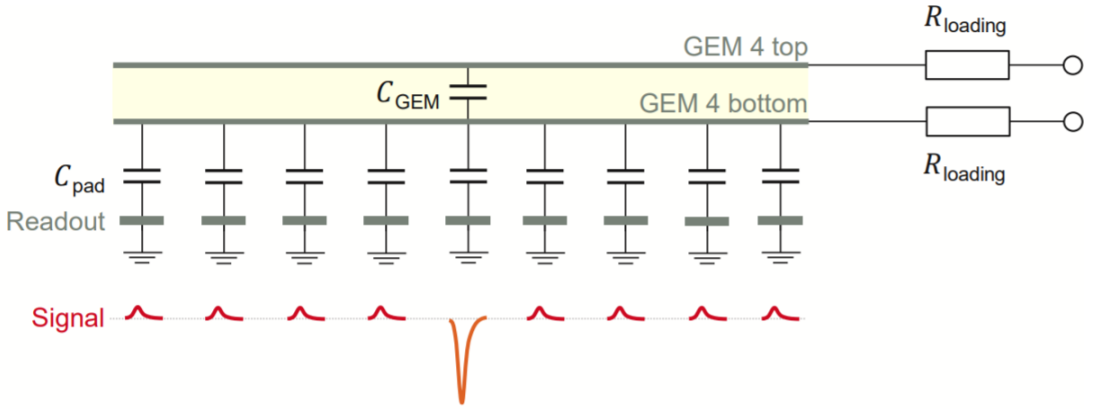
■ FPGA内に多チャンネル再帰型Adaptive Filterを構築
- 200 ns 毎に1600個のADC値の平均値を計算
(SAMPA内の32チャンネルだけでは不正確)

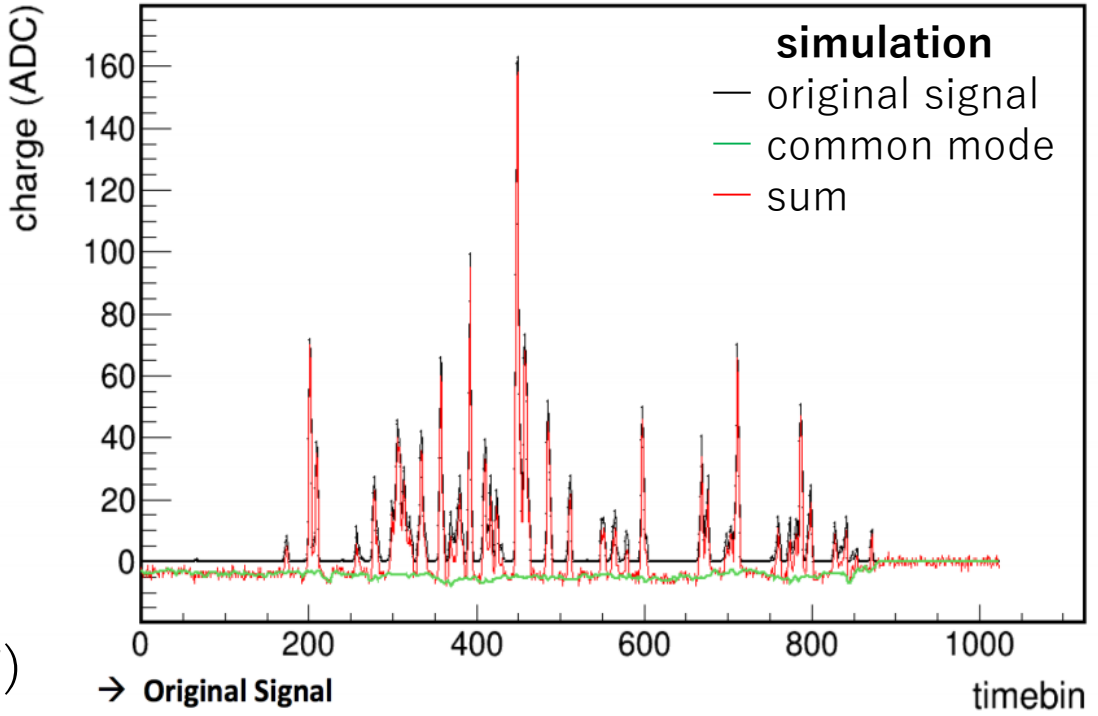
$$O_j = I_j - I_{CM} \quad , \qquad I_{CM} = \frac{\sum I_i}{N_{cont}}$$

■ シグナルにより、この値は強くバイアスを持つ

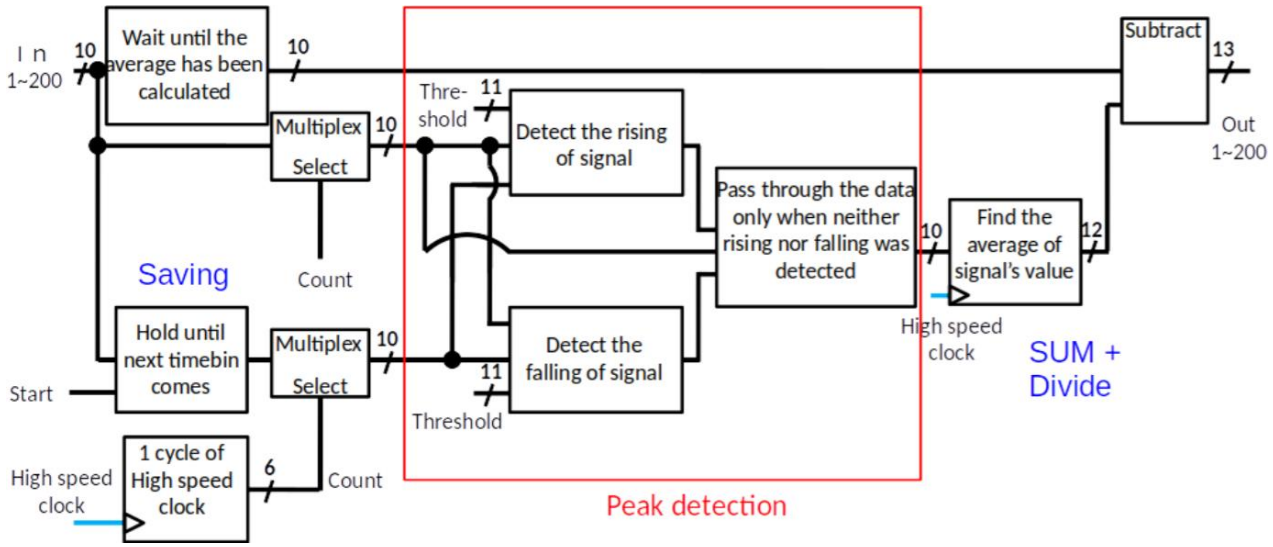
プランA: Rising edge, Falling edge でピーク検出・除外

プランB: 中央値計算(5 MHzでヒストグラムを作成・評価)

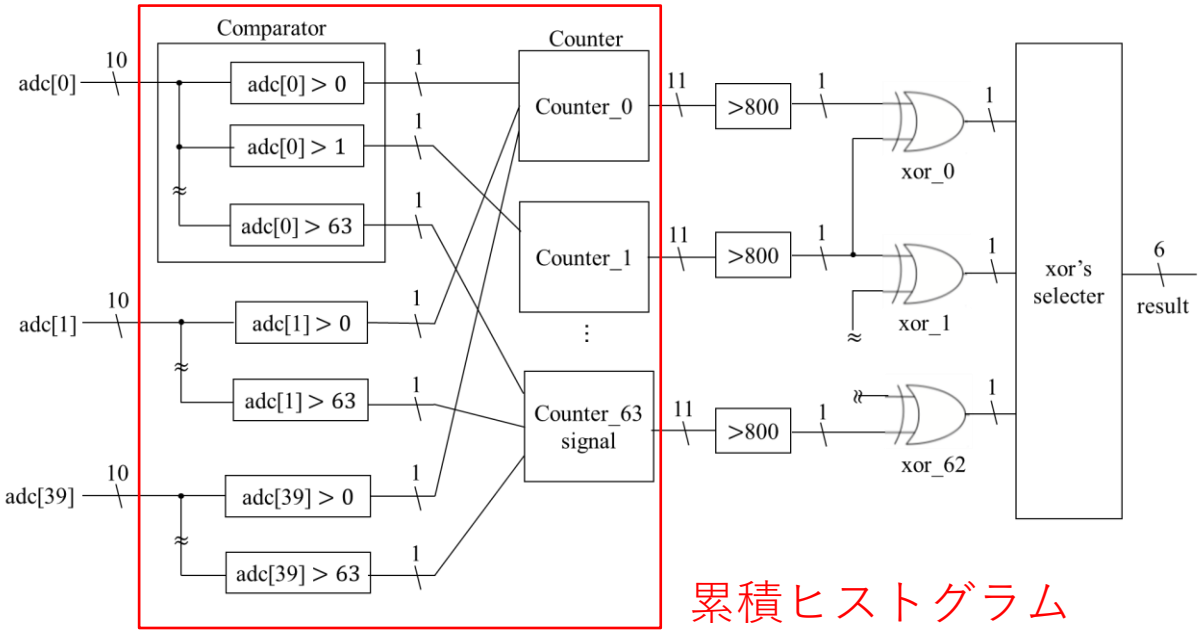




コモンモード除去フィルタ(続)



[peak rejection by Y. Takeuchi]



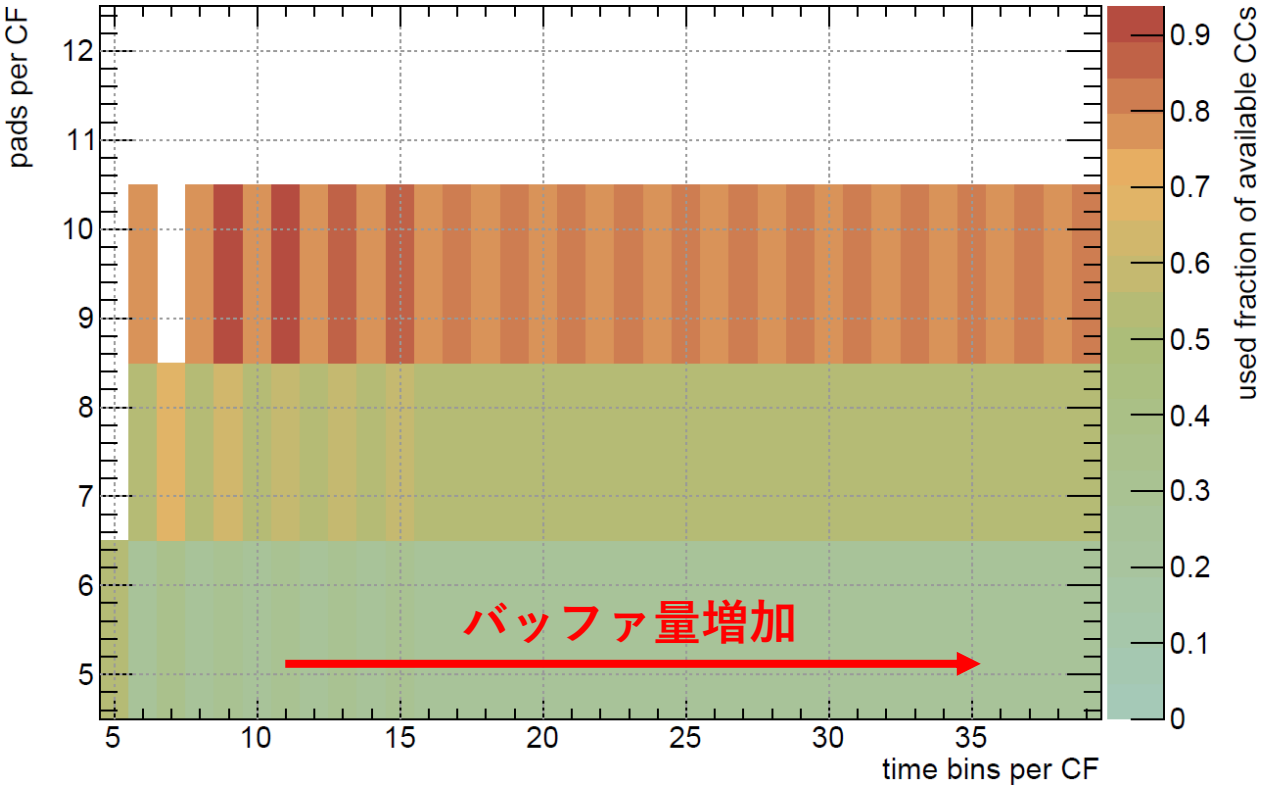
[median by Y. Matsuyama]

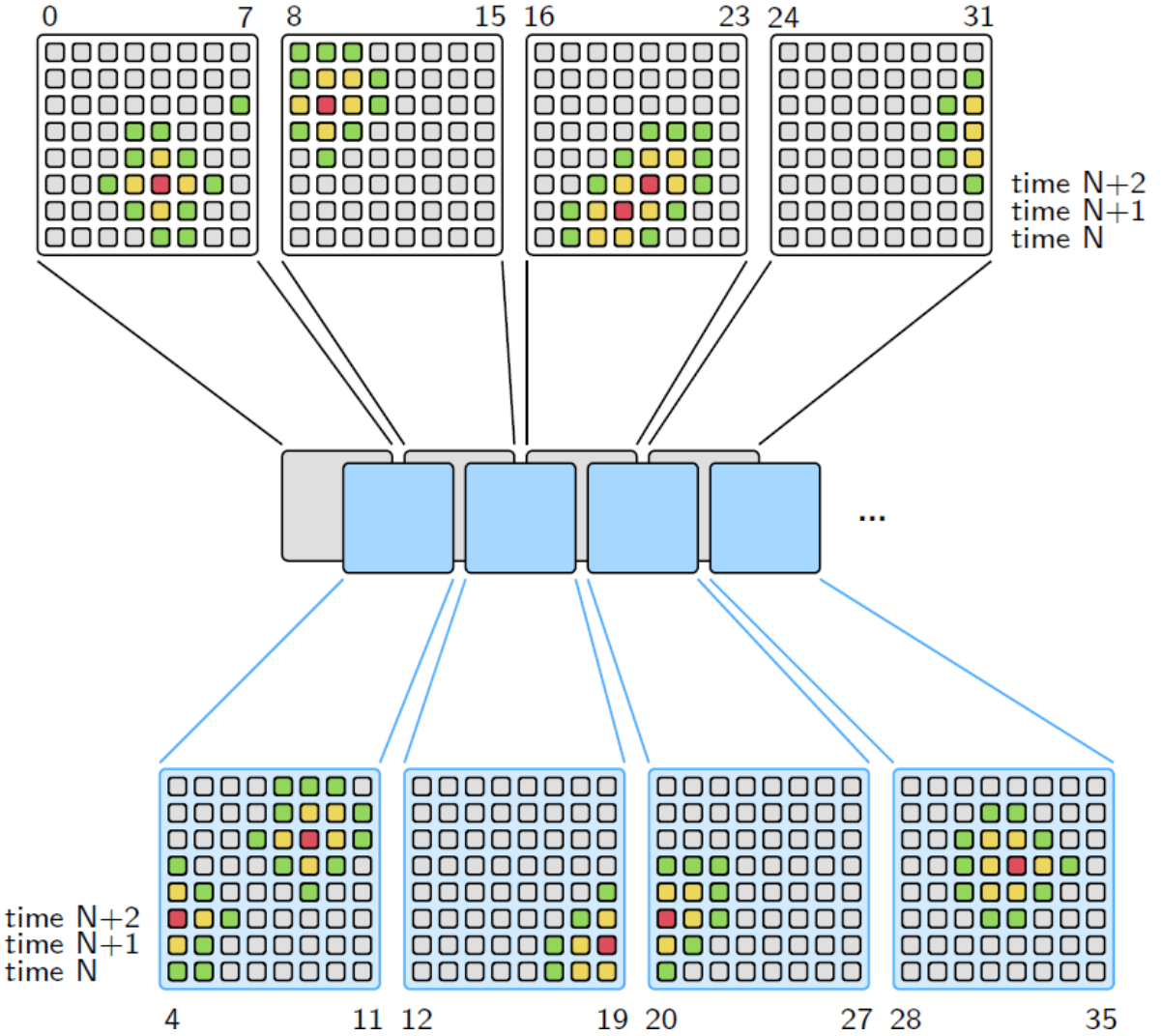
■現在二プランを比較検討中

Type	プランA: ピーク除去式	プランB: 中央値式
ALM使用率	2%	**15%**
バイアス	劣る(1-2 ADC値バイアス)	優れる(~ゼロバイアス)
occupancy耐性	50%以上でも動作継続	50%で破綻

クラスタ発見アルゴリズム

- pad-timebin二次元空間内でlocal maximaを探す
- 小さい領域をスキャンするモジュールを多数 *or* 大きい領域をスキャンするモジュールを少数





- 見つけたピークを含む 5x5領域を後段に出力

drawings by Sebastian Klewin

クラスタリング前処理

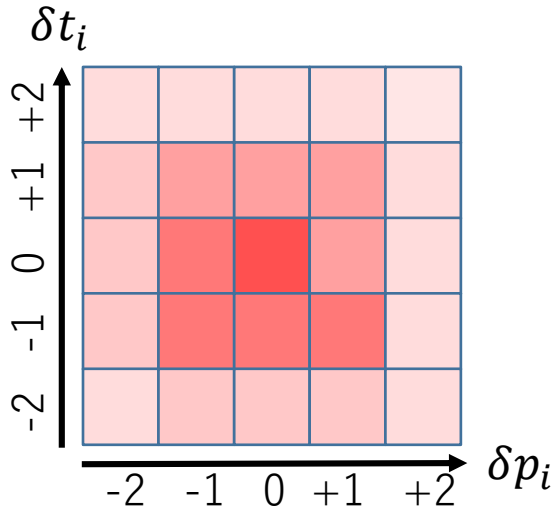
■ 5x5 pad-timebinデータに対し、クラスタ・パラメタを計算

ローカル座標系でのクラスタ・パラメタ計算

- $q_{tot} = \sum q_i$
- $\mu_x = x + \frac{\sum q_i \, \delta x_i}{q_{tot}}$
- $\sigma_x^2 = \frac{\sum q_i \, \delta x_i^2}{q_{tot}} - (x - \mu_x)^2 = \frac{\sum q_i \, \delta x_i^2}{q_{tot}} - \left(\frac{\sum q_i \, \delta x_i}{q_{tot}}\right)^2$

$i = 1 \ldots 25$, x: pad, timebinインデックス

$\delta x_i = \{-2, -1, 0, +1, +2\}$ との積=ビットシフト



■ FPGA実装: 割算を避け、前処理のみを行い、最終的な割算はCPUに任せる

FPGAフレンドリな計算
(和とビットシフトのみで前処理)

- $q_{tot} = \sum q_i$
- $\hat{\mu}_p = \sum q_i \delta p_i$
- $\hat{\mu}_t = \sum q_i \delta t_i$
- $\hat{\sigma}_p = \sum q_i \delta p_i^2$
- $\hat{\sigma}_t = \sum q_i \delta t_i^2$

データ転送, PCI Express

250 bit → 160 bit にパッキング

CPUフレンドリな計算(FLPで後処理)

- $\mu_p = p + \hat{\mu}_p / q_{tot}$
- $\mu_t = t + \hat{\mu}_t / q_{tot}$
- $\sigma_p^2 = \hat{\sigma}_p / q_{tot} - \left(\hat{\mu}_p / q_{tot}\right)^2$
- $\sigma_t^2 = \hat{\sigma}_t / q_{tot} - (\hat{\mu}_t / q_{tot})^2$

S. Klewin's PhD thesis coming soon

FPGAリソース使用率

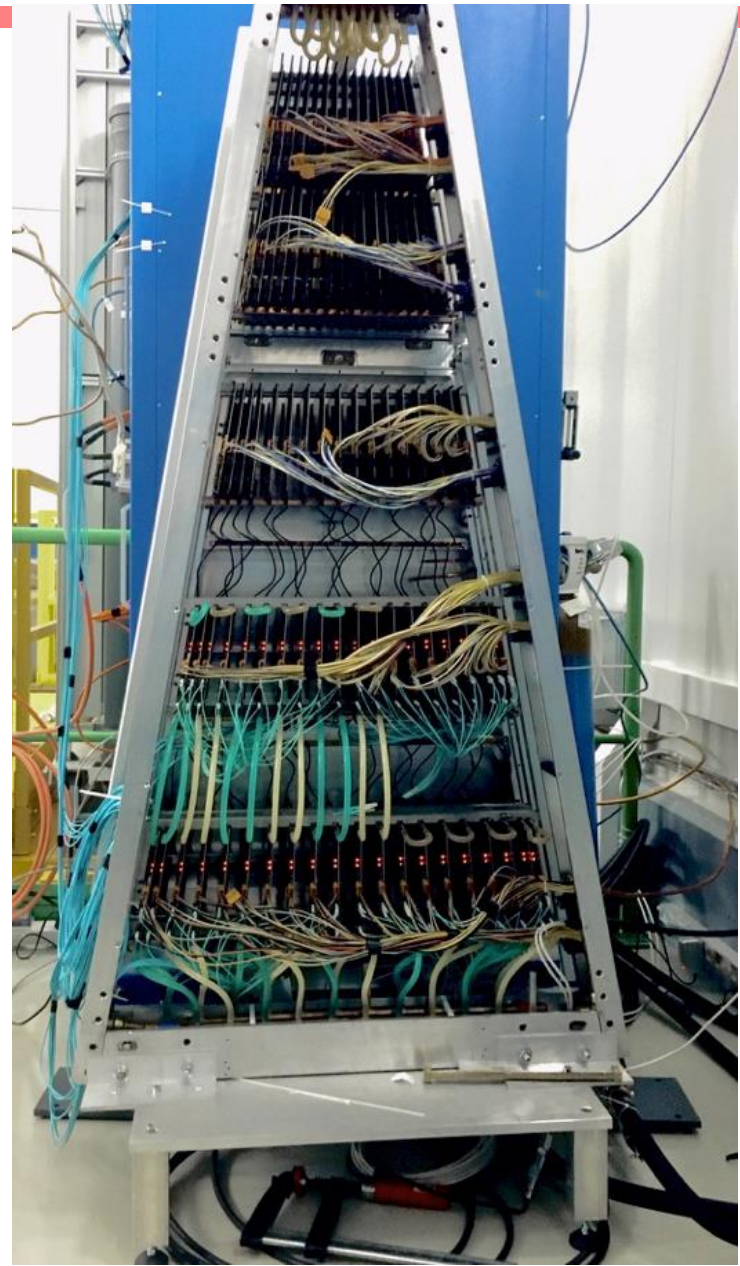
■ Arria10 10AX115S3F45E2SG

Module	ALMs	%	M20Ks	%	DSPs	%
周辺ロジック	7144	1.7	116	4.3	54	3.6
GBTデコード	10264	2.4	0	0	0	0
ソーティング	43956	10.0	40	1.5	0	0
クラスタリング	167905	39.3	906	33.4	362	23.8
読み出し部分	3890	0.9	0	0	0	0
コンフィギュレーション	10024	2.3	0	0	0	0
ユーザーロジック合計	**243184**	**57**	**1062**	**39**	**416**	**27**
共通ロジック	119762	28	1252	46	0	0
全ファームウェア合計	**362946**	**85**	**2314**	**85**	**416**	**27**

コモンモードフィルタは現在組み込み作業の途中(ALM 15% からダウンサイズの努力中)

まとめ、現状と今後

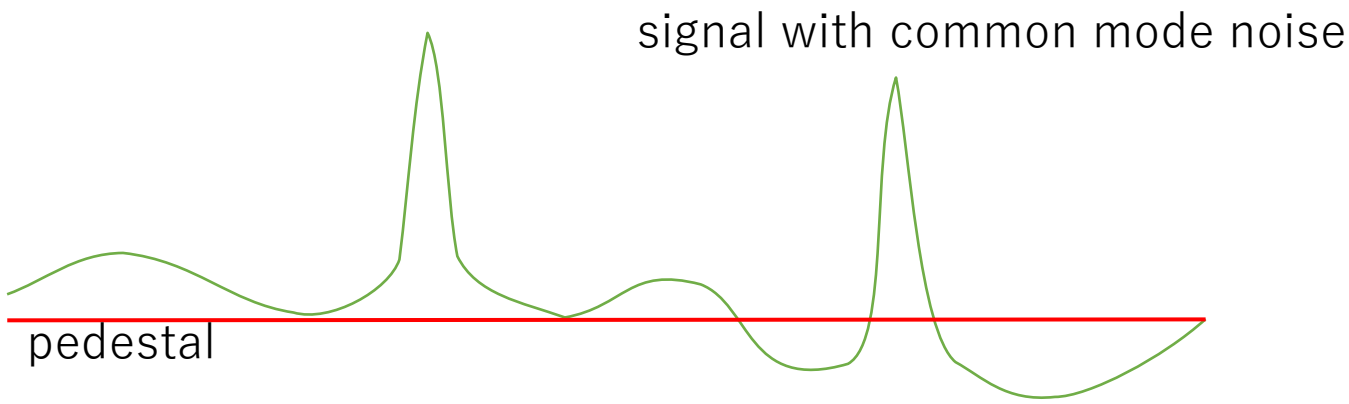
- ALICEアップグレードでは、トリガ無し連続読み出し型GEM-TPCの実現を目指す
- 要求: データスループットを1/35に圧縮 (3.5 TB/s → 100 GB/s)
  - 且つCPUコア数を抑える
  - 前段データ処理にArria 10 FPGAを採用 (ここで1/5に圧縮)
- 現状
  - 主要なモジュールの実装完了
  - ロジックシミュレーション(RTL)は良好
  - フィッティングは上限に達しつつあるので、スリム化を実施中
  - 特にコモンモードフィルタのロジック使用量のスリム化が課題
- 今後
  - 実機（CERN）におけるテスト
  - さらなる1/7の圧縮 … トラッキングにGPUを導入予定



Other filters

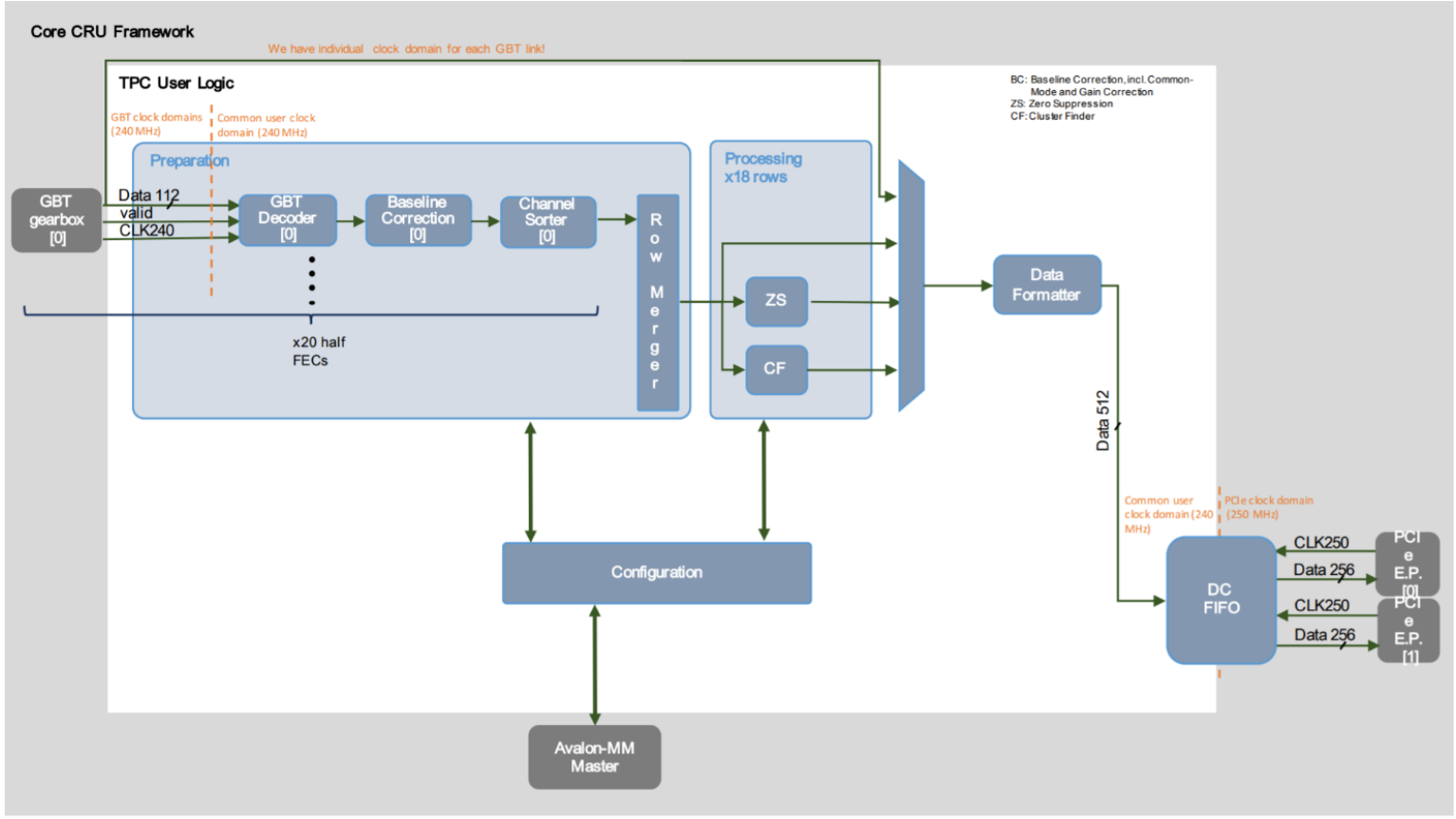
■ Pedestal subtraction filter
- TPC decided to do NOT subtract pedestal on SAMPA but do that on CRU
- subtracting pedestal → chop negative values (unless we introduce sign flag → data increase)
- with common mode, this problem will be significant
- pedestal value can be represented finer (fixed point number with half and quad LSB bits)

■ Gain correction
- not done, optimistically



TPC User Logic

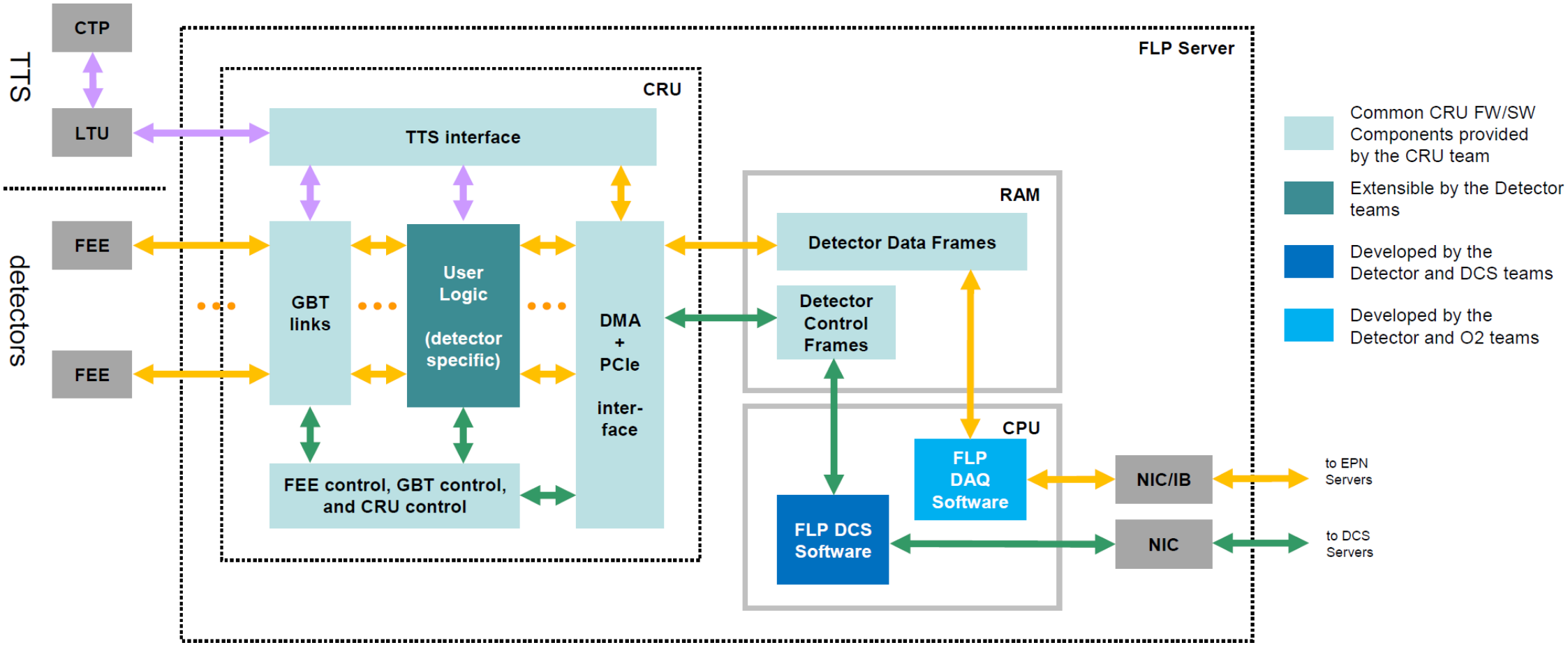
- raw data processing
  - channel sorting / pedestal subtraction / common mode rejection / clustering /data formatting
- DCS: forwarding DCS control command & data
  - Power / SAMPA & GBT configurations / CRU FPGA setup parameters



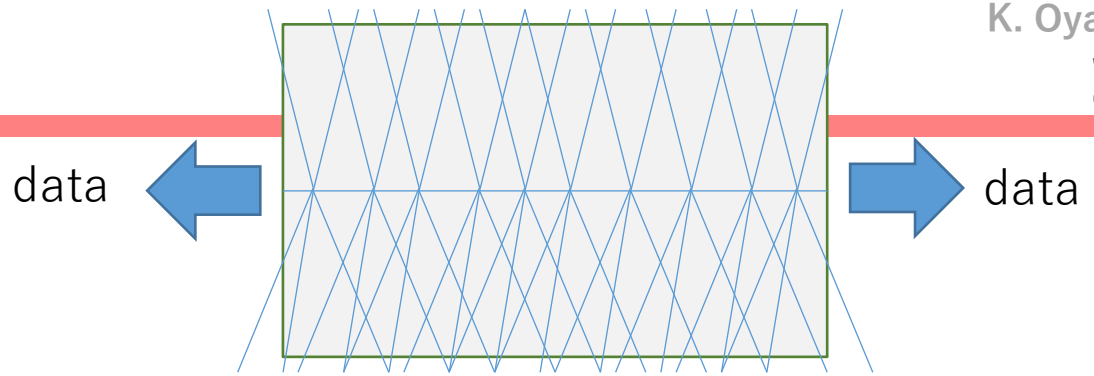
CRU FPGA 内部ロジック開発体制

■ Central CRU チーム(Grenoble)がペリフェラルロジック(ALICE共通)を開発

■ 検出器CRUチームはUSER-Logicを開発(TPCの場合 Frankfurt, Heidelberg, Nagasaki-IAS)



TPC Upgrade (cont.)



- LHC will provides above 50 kHz Pb+Pb event rate after upgrade (20 μm average event interval)

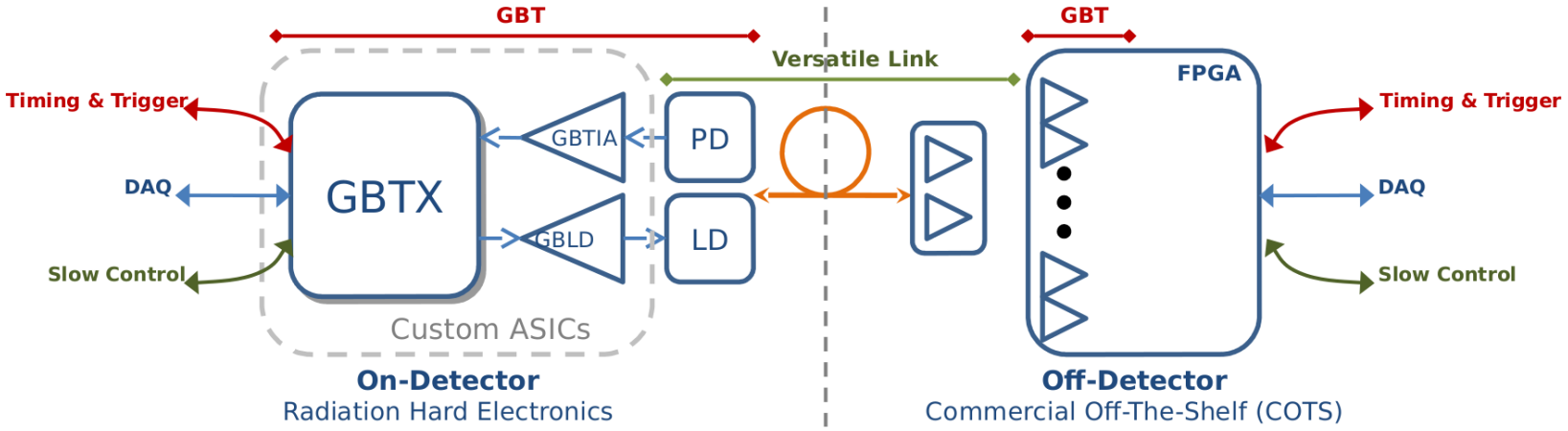
- TPC drift time (100 μs)
  - large pile-up
  - average 5

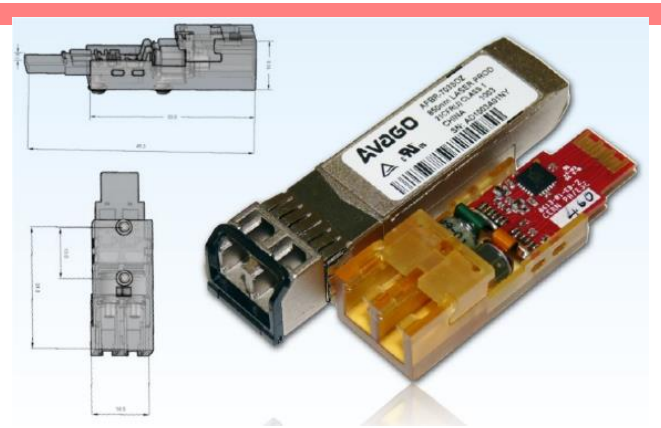
- Continuous (triggerless) data taking

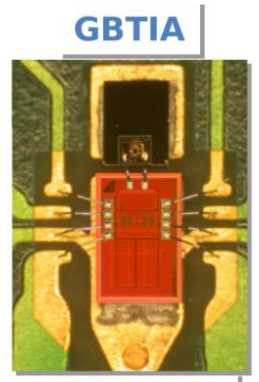
- 3.5 TB/s data rate
  - large data reduction is required

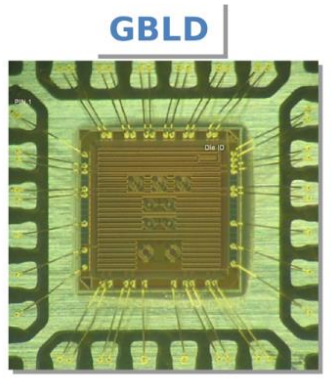
Clock trigger 分配とデータ読み出し(GBT)

■ GBT: CERN and LHC 共通開発プロジェクト(SERDES + optical link)

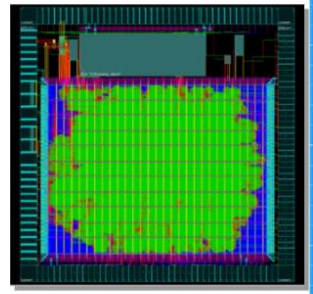




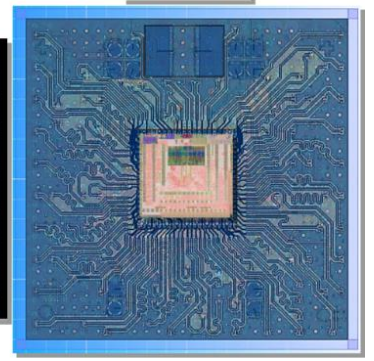




GBT - SCA



GBTX



■ 最大データ転送速度: 4.8 Gbps data + リアルタイムトリガ・クロック転送

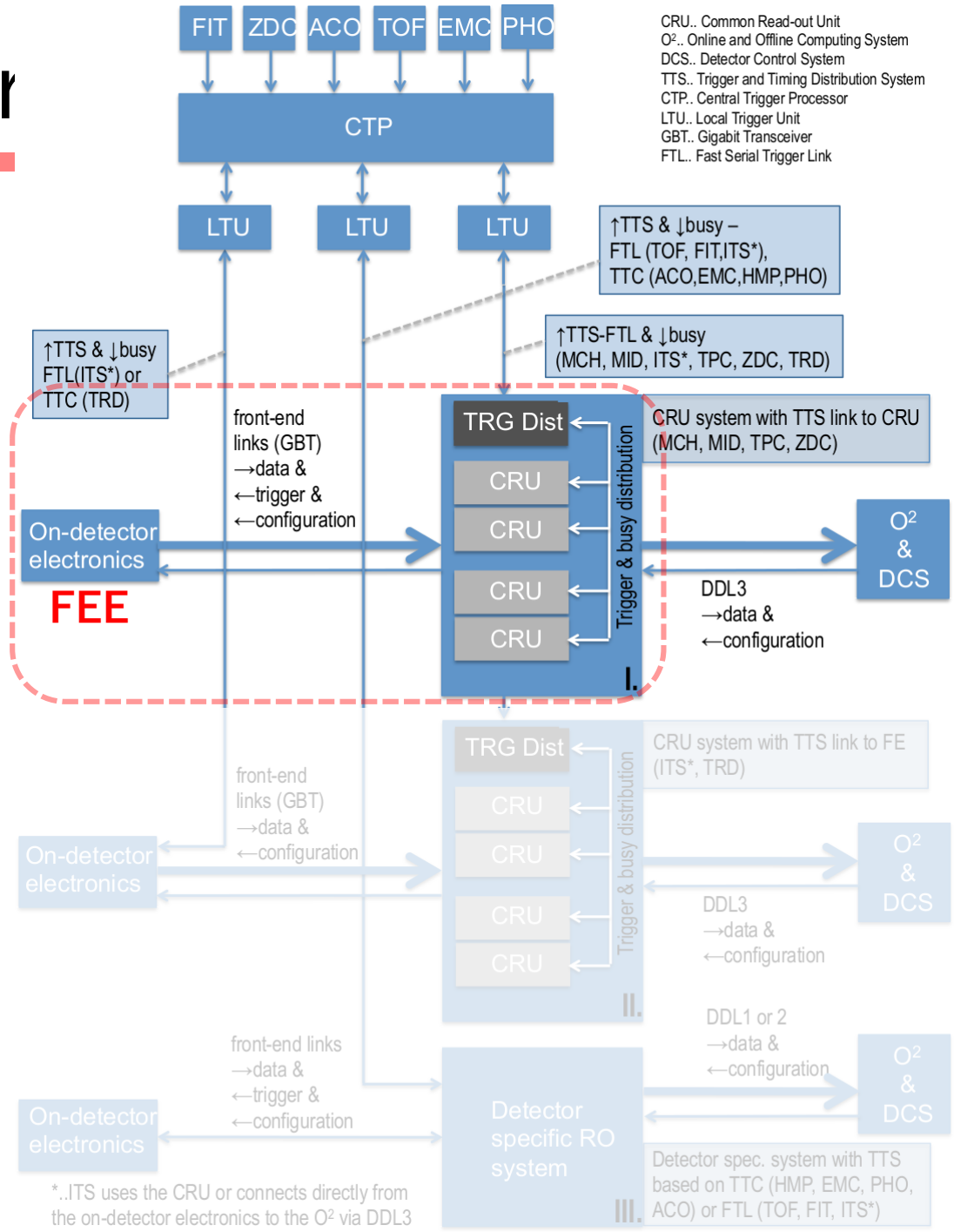
■ 耐放射線
- GBT*: ~ 100 Mrad
- SFP transceiver: 50 Mrad

■ FPGA IP ライブラリの整備

■ 検出器コントロール・プロトコル(遠隔コンフィギュレーション・モニタリング)

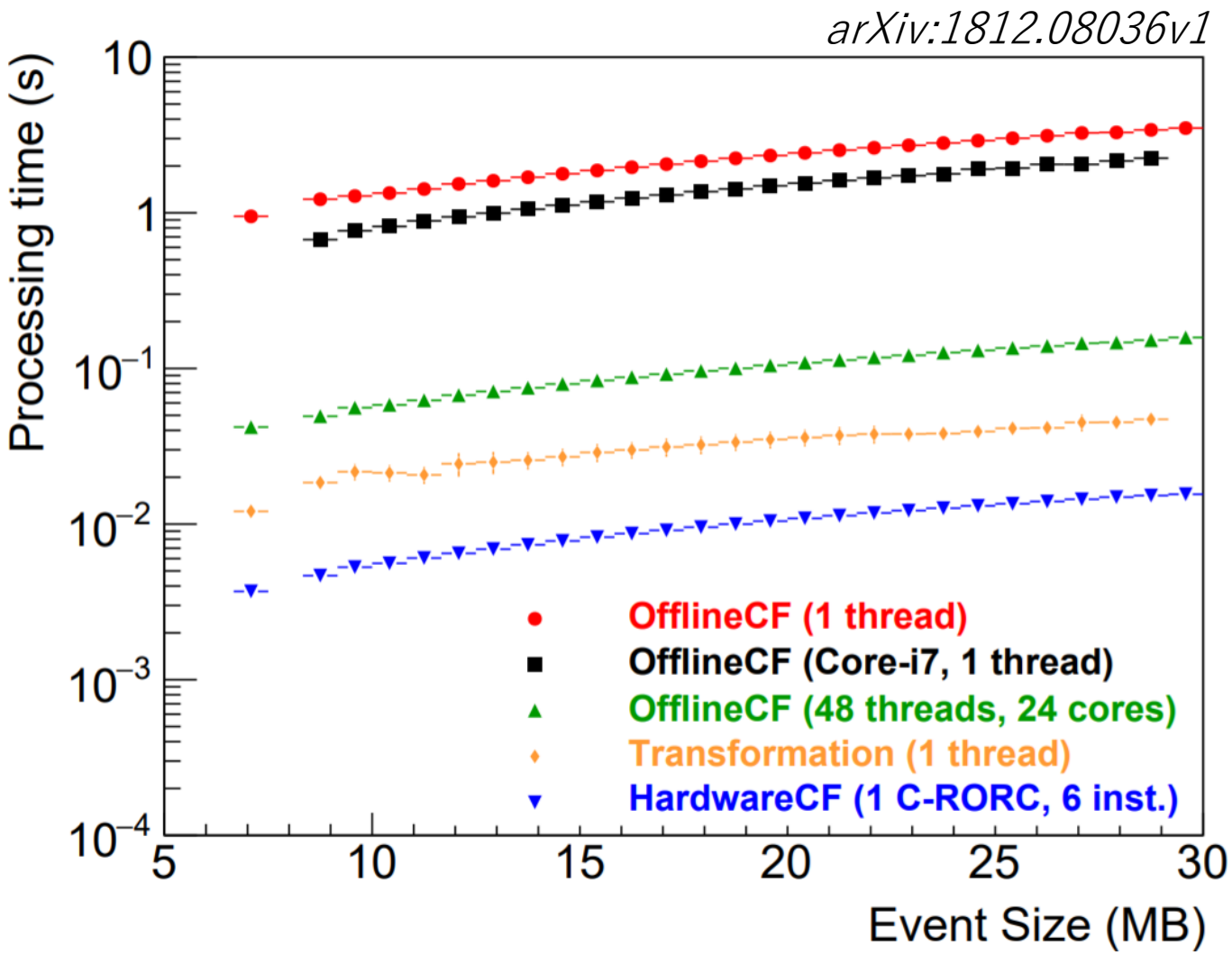
ALICE readout system after

- On-detector electronics
  - controlled via GBT, sends data via GBT
  - front-end electronics needs only GBT duplex fiber interface and power & cooling services
- Common Readout Unit (CRU)
  - common design for all new detectors incl. FoCal
  - max. 48 duplex GBT connections
  - placed in a PC server (FLP), communicate with CPUs via PCI express bus
- trigger and machine clock distribution is also via GBT
  - CTP sends trigger and fast control to CRU
  - then CRU forwards it to front-end
- detector control is also via GBT
  - DCS system will configure & acquire status from front-end via CRU and GBT



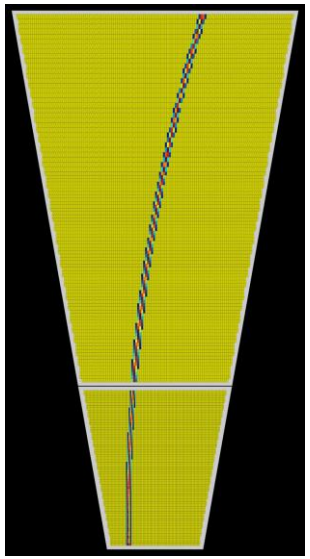
FPGA acceleration factor for TPC clustering

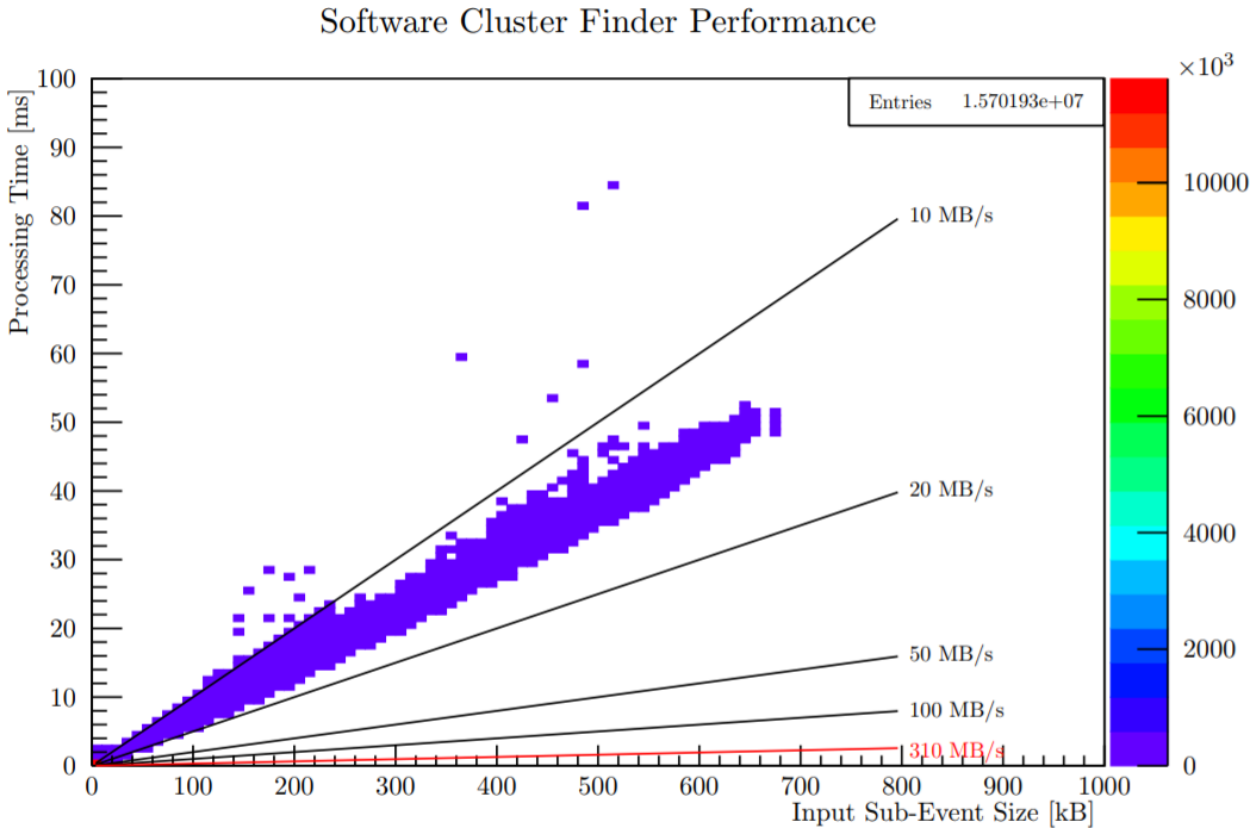
- HLT Xeon E5-2697 nodes and newer CPU (Core-i7, black recutangle) compared to hardware on C-RORC
- above factor 10 improvement compared to 48 thread (full power of dual Xeon) software processing
  - means FPGA corresponds to 240 cores, 480 threads
- transformation: only coordinate transformation



Cluster Finder Acceleration

- Software processing time v.s. event size shows about 15 MB/s processing data rate
  - XEON E5-2697
  - 32 core
- 205 Pb-Pb events
- HW CF (red) is 20 times faster than software processing
  - 320 MB/s RCU2
  - XILINX Virtex 6

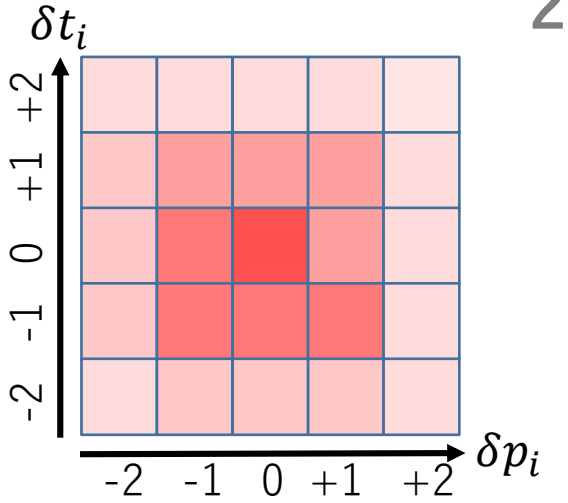




クラスタリング前処理

■ 5x5 pad-time データに対し、クラスタ・パラメタを計算

・ 割算を避け、前処理のみを行い、最終的な割り算はCPUで行う



グローバル座標系

$i = 1 \ldots 25$, x: pad, time-binインデックス

- $q_{tot} = \sum q_i$
- $\mu_x = \frac{\sum q_i x_i}{q_{tot}}$
- $\sigma_x^2 = \frac{q_i(x_i-\mu_x)^2}{q_{tot}} = \frac{\sum q_i x_i^2}{q_{tot}} - \mu_x^2$

ローカル座標系

- $q_{tot} = \sum q_i$
- $\mu_x = x + \frac{\sum q_i \, \delta x_i}{q_{tot}}$
- $\sigma_x^2 = \frac{\sum q_i \, \delta x_i^2}{q_{tot}} - (x - \mu_x)^2 = \frac{\sum q_i \, \delta x_i^2}{q_{tot}} - \left(\frac{\sum q_i \, \delta x_i}{q_{tot}}\right)^2$

$\delta x_i = -2, -1, 0, +1, +2$ … ビットシフトで済む

FPGAフレンドリな計算(和とビットシフトのみで前処理)

- $q_{tot} = \sum q_i$
- $\hat{\mu}_p = \sum q_i \delta p_i$
- $\hat{\mu}_t = \sum q_i \delta t_i$
- $\hat{\sigma}_p = \sum q_i \delta p_i^2$
- $\hat{\sigma}_t = \sum q_i \delta t_i^2$

データ転送, PCI Express

250 bit → 160 bit にパッキング

CPUフレンドリな計算(FLPで後処理)

- $\mu_p = p + \hat{\mu}_p / q_{tot}$
- $\mu_t = t + \hat{\mu}_t / q_{tot}$
- $\sigma_p^2 = \hat{\sigma}_p / q_{tot} - \left(\hat{\mu}_p / q_{tot}\right)^2$
- $\sigma_t^2 = \hat{\sigma}_t / q_{tot} - (\hat{\mu}_t / q_{tot})^2$

S. Klewin's PhD thesis coming soon