LHC-ALICE実験でのコンピューティング
(Computing in LHC-ALICE)

75GB/s　生データに挑戦！
次世代 LHC 重イオン衝突実験における
パイプラインデータ処理と世界分散計算機網

中條 達也（筑波大学）

Tatsuya Chujo(for the ALICE collaboration)

March 21, 2015

日本物理学会第７０回年次大会シンポジウム

「実験のための最先端コンピューティング」

早稲田大学

筑波大学
University of Tsukuba

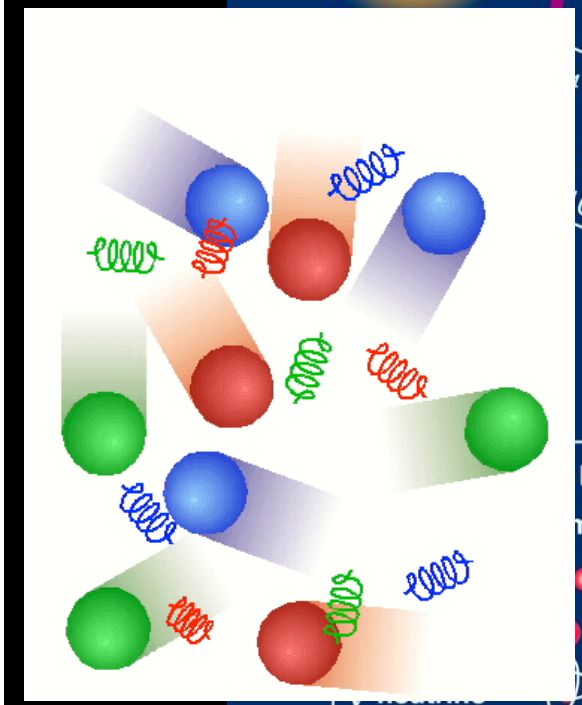
ALICE

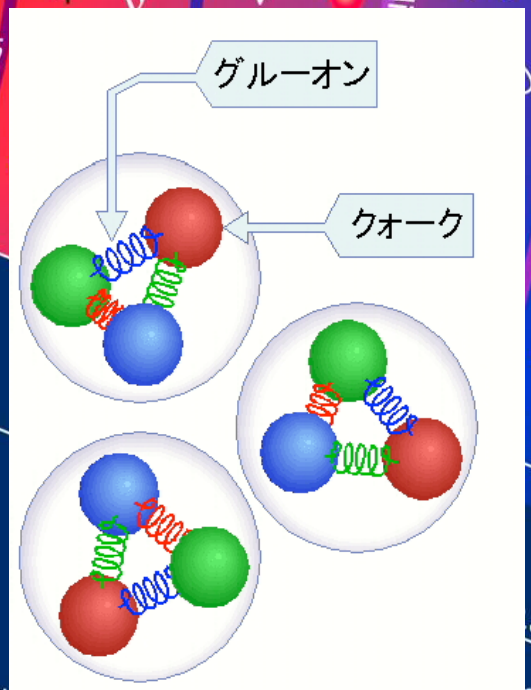
Quark-Gluon Plasma (QGP)

時間:ビックバンから数マイクロ秒後

温度: 2 Trillion K

エネルギー密度

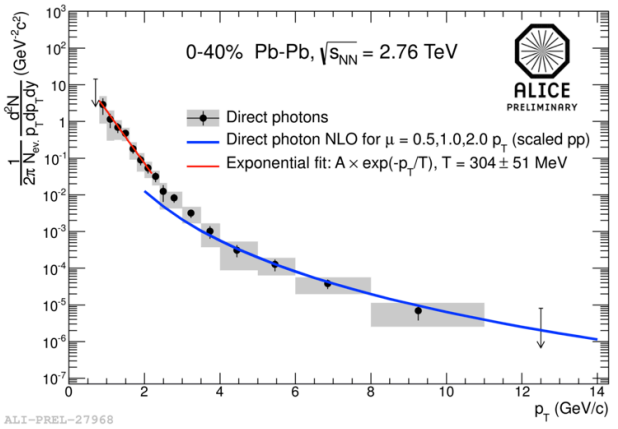


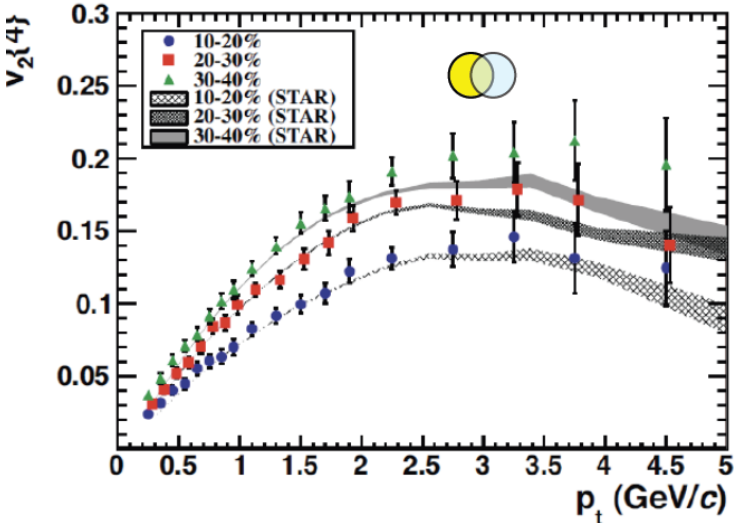


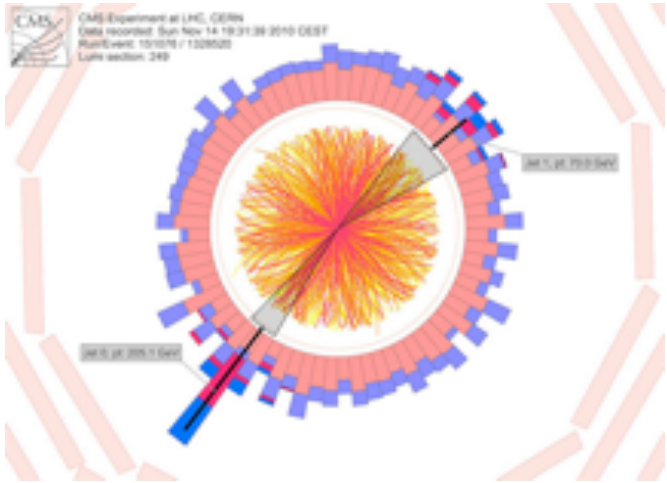
第１期　LHC重イオン衝突実験結果のハイライト

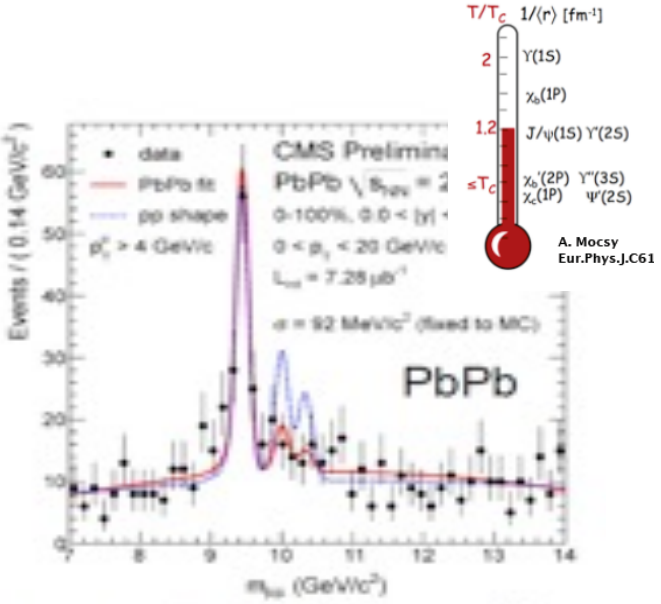
(Run-1 : 2009-2013)

- 初期到達温度: T_{int} ~ 304 ± 51 MeV ~1.4 x T_{int} (RHIC).
- 大きな集団膨張 (等方, 楕円) (ALICE, ATLAS, CMS)
- 大きなジェット抑制効果 (ALICE, ATLAS, CMS)
- ϒ 励起状態の消滅 (高温物質生成の証拠, CMS)

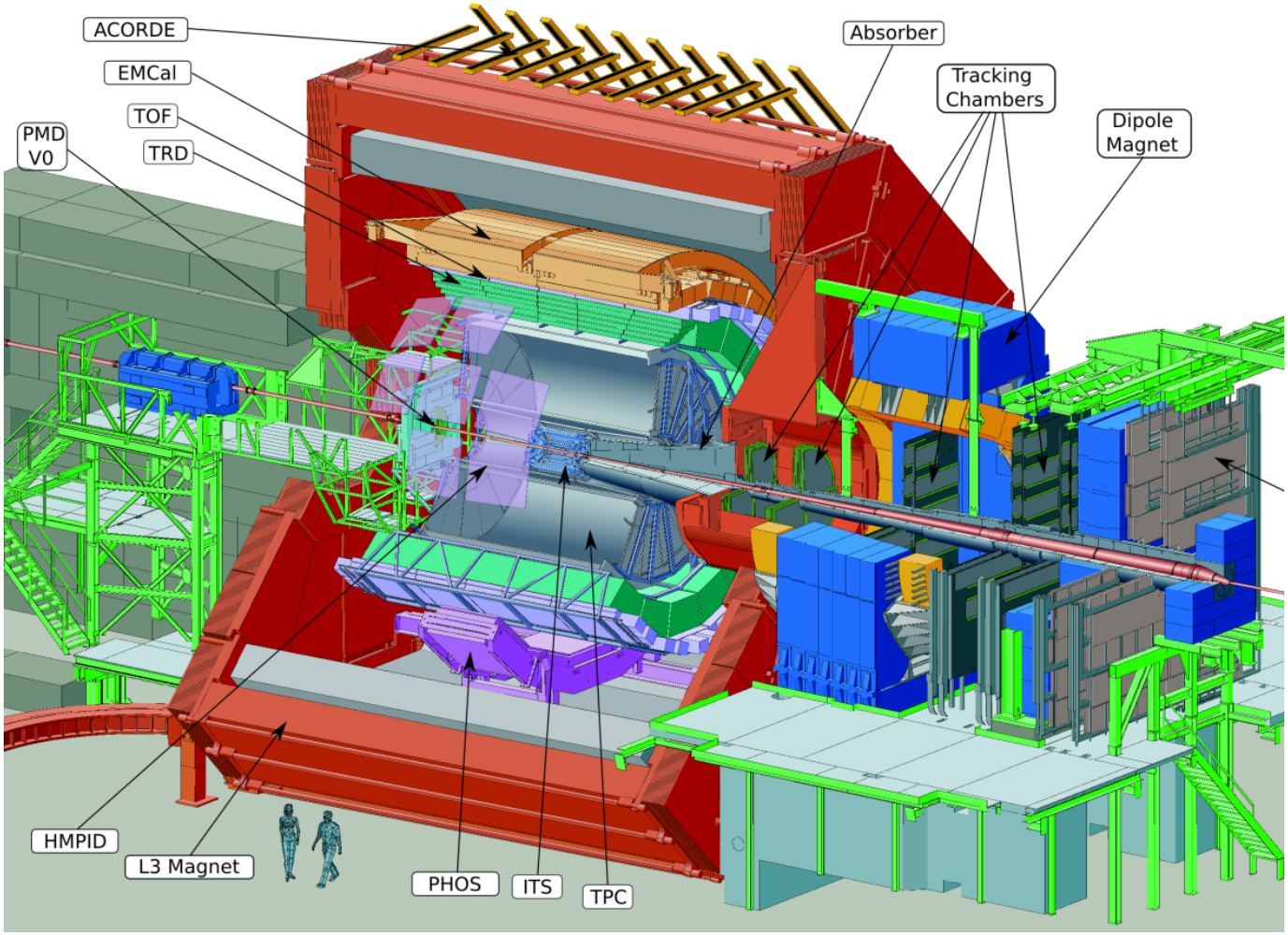




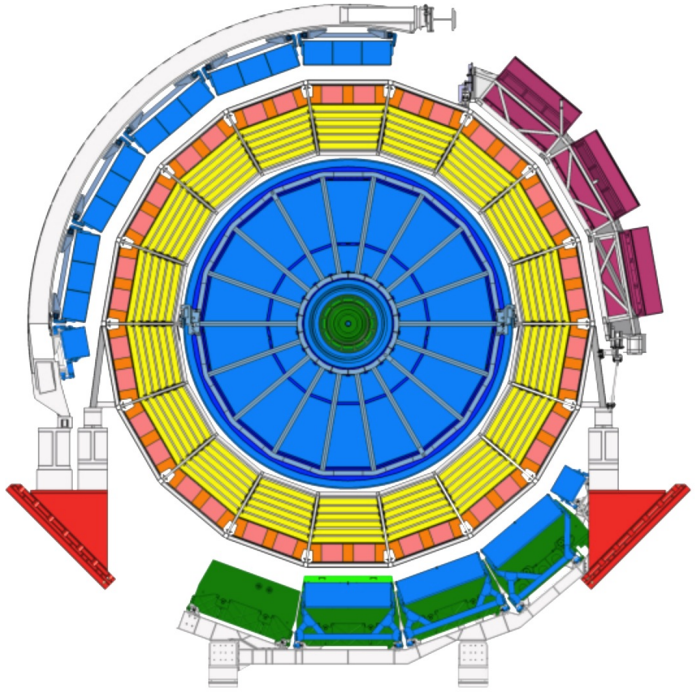




LHC-ALICE実験



16m x 16m x 26m, 10,000 tons



18 の異なるサブ検出器

飛跡再構成、粒子識別、
エネルギー測定、
イベントトリガー

1400 名以上の研究者、149 研究機関、40カ国

LHC で唯一、重イオン実験とQGP研究に特化した実験

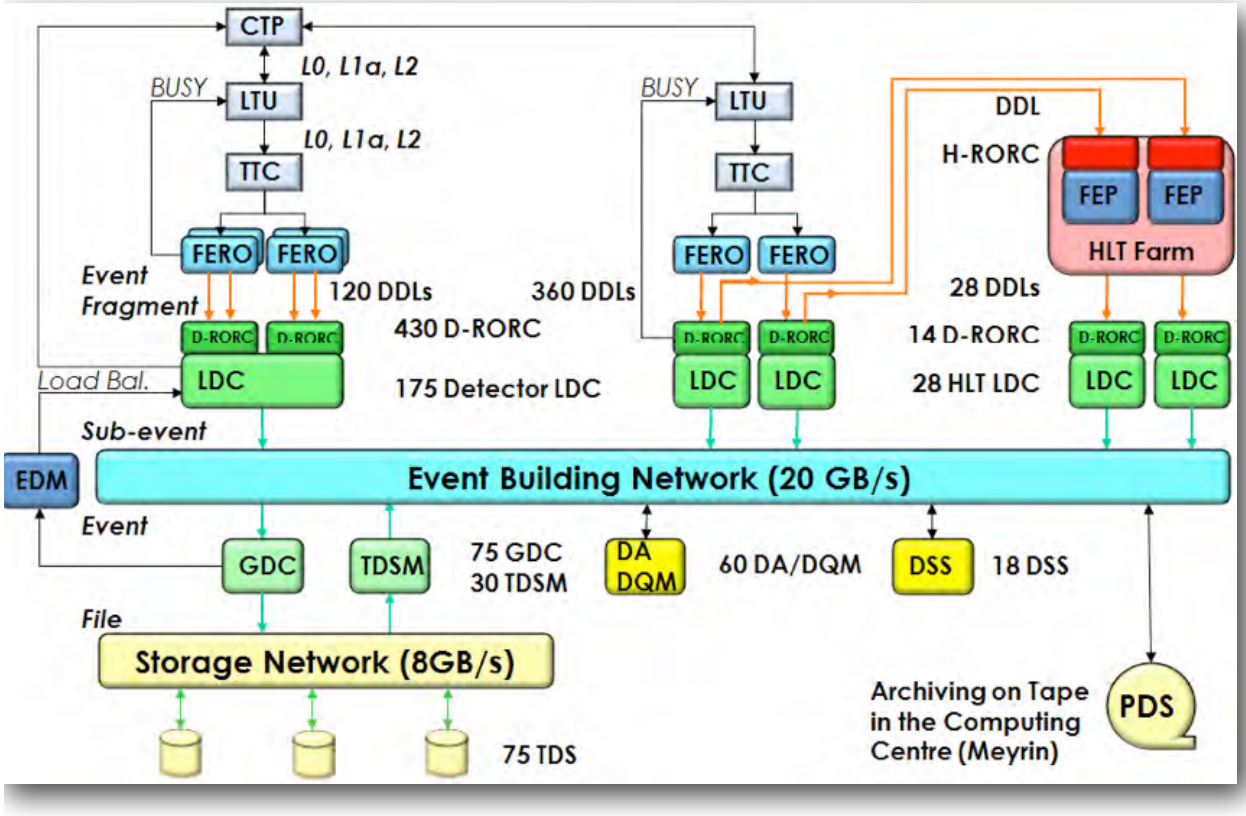
日本の参加機関：広島大、長崎総科大、東大CNS、筑波大

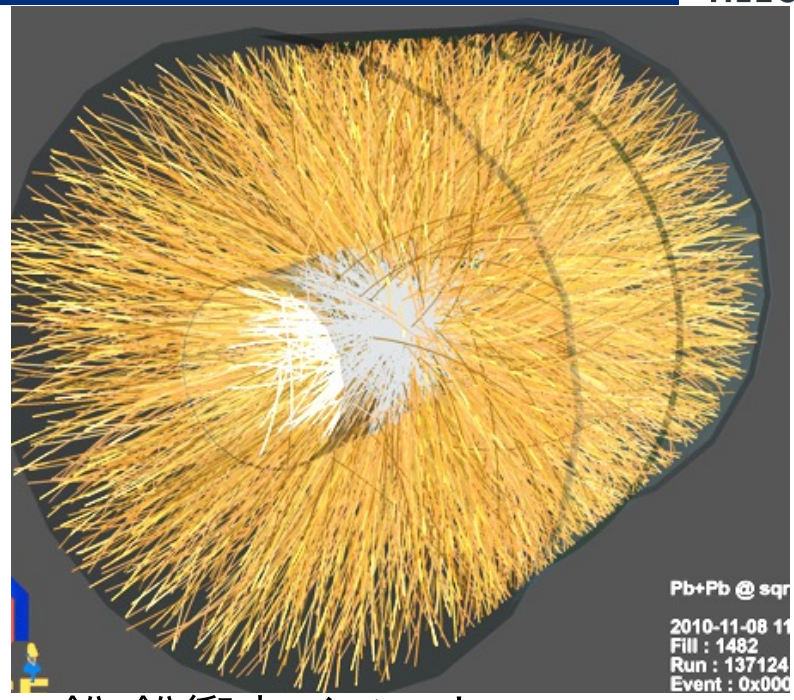
ALICE 実験のデータ収集 (現状)

通常のLHC ビーム crossing: 40 MHz

ALICE:
マルチレベルトリガー (L0, L1, L2)

40 MHz → ~1 kHz





鉛-鉛衝突イベント

($\sqrt{s_{NN}}$ = 2.76 TeV, Pb-Pb)

<u>オンライン</u>

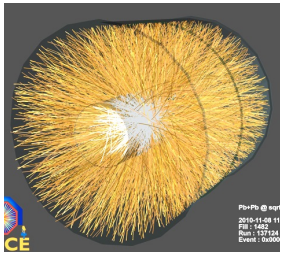
1) ビームバックグラウンド除去

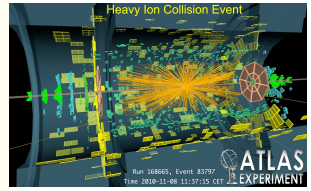
2) 興味のあるイベント選択

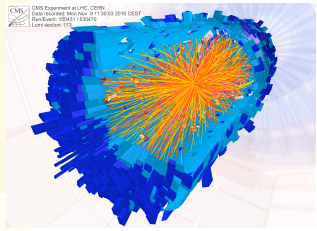
3) 専用計算機クラスターによるデータボリュームの圧縮

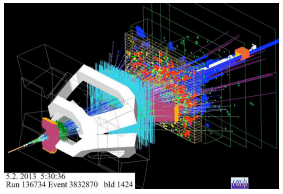
データレート・データボリューム (現状)

	衝突ビーム	Recording (Events/s)	Recording (MB/s)	データアーカイブ (PB per year)
ALICE	**Pb-Pb (1ヶ月)**	**200**	**1250**	**2.3**
ATLAS	**p-p (~8ヶ月)**	**100**	**100**	**6**
CMS	**p-p (~8ヶ月)**	**100**	**100**	**3**
LHCb	**p-p (~8ヶ月)**	**200**	**40**	**1**



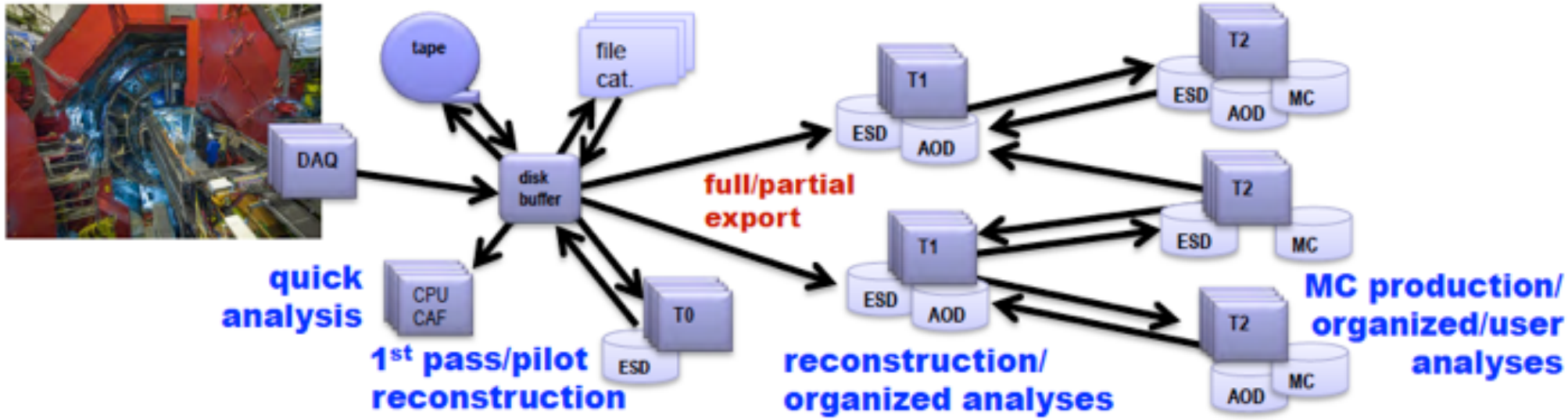






- Pb-Pb イベントサイズ > p-p イベントサイズ
- オンライン事象選択 + オフラインプロセシング →最終データをアーカイブ
- 各実験、それぞれのオンライン・オフライン処理方法

コンピューティングモデル　(Run1 and Run2, -2018)

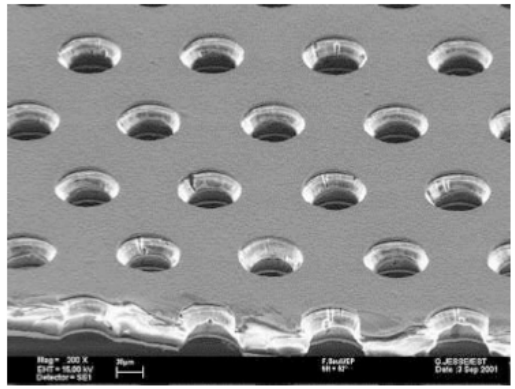


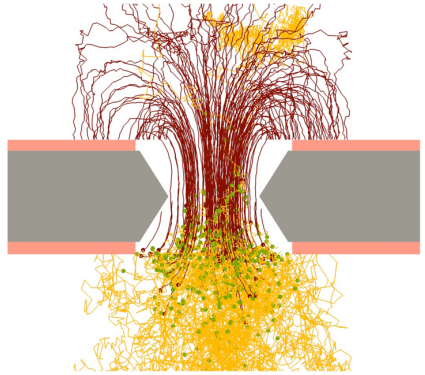
slide by T. Sugitate

ALICE 高度化計画 (Run-3, 2018-)

ALICE 高度化; 高レートへの対応

1. GEM-TPC による高レート連続読出し
2. ITS シリコン検出器 高レート読出し
3. DAQ 高度化 (CRU 等)



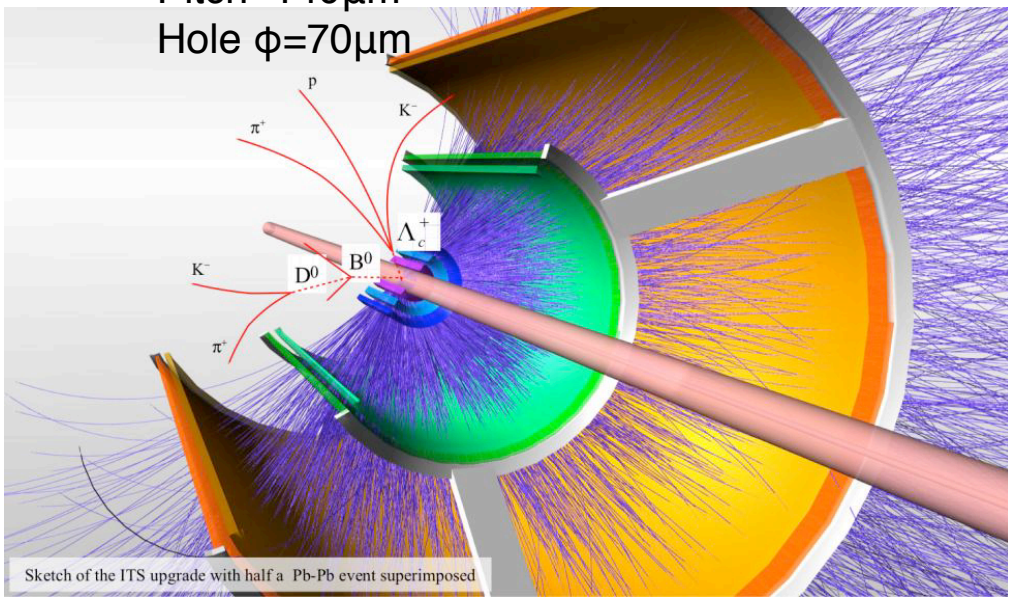


Standard GEM
Pitch=140μm
Hole ϕ=70μm

LHC 高輝度化に対応した Pb-Pb @50kHz 連続読み出しを実現する

最小バイアスイベント (MB) を全て記録、Run-1 の統計量の１００倍（ALICE実験でのみ可能）

→ **高精度、稀事象プローブへアクセス可能**



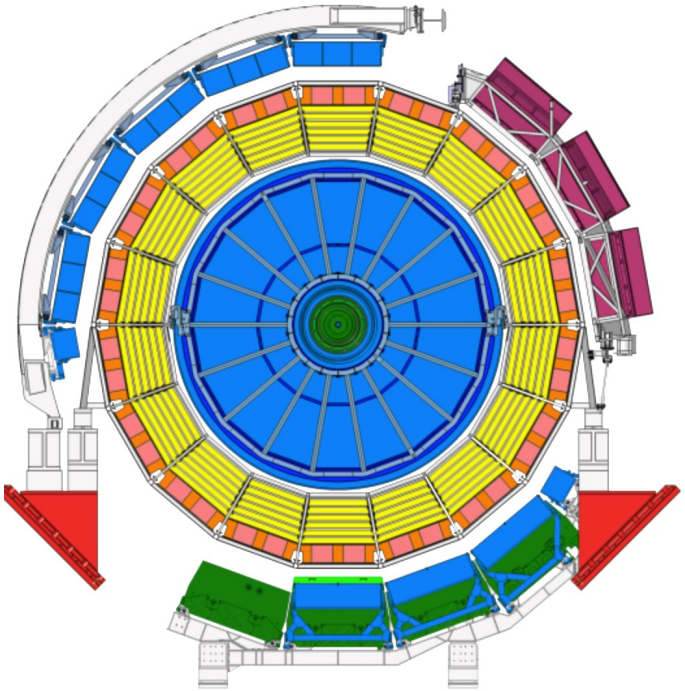
Sketch of the ITS upgrade with half a Pb-Pb event superimposed

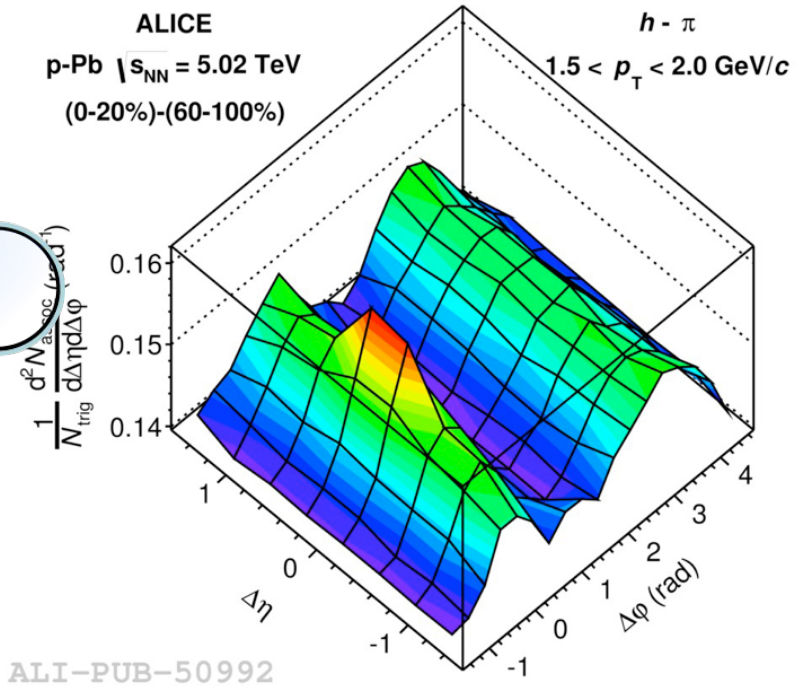
物理ゴール：QGP物性の精密測定

例）

- 重クォーク、光子、希少プローブの方位角異方性
- レプトン対（低 S/N比）→トリガーレス、高統計データ取得

The ALICE Online-Offline (O2) プロジェクト





- 検出器 Raw データから物理解析へ
- 最適なコンピューティング設計が重要

- 予想されるデータレート (Raw data)： **>1 T Byte /s**
- 検出器の連続読み出しをサポート
- オンライン・データ再構成によりデータ量を圧縮
- オンライン、オフライン, DAQ, HLT (High-Level Trigger)チームによる共通ハードウェア・ソフトウェア開発

ALICE O2 システムの要求される機能

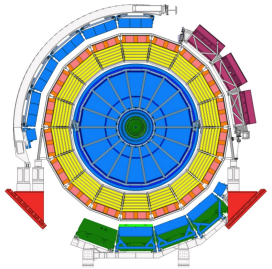
✓ **1 TB/s のデータ量をハンドル**

- データストレージ前にデータをフルに圧縮
- 品質の良いデータ較正によるデータ再構成
- ローカルストレージのフル活用

✓ **Grid 計算機網の活用**

- 解析に使う再構成データをストレージに入れると同時に、Grid 上にキープ

O2 システムの基本思想



検出器読み出し装置

RAW データのトリガーレス連続読み出し

1.2 TB/s

読み出し, 分割, 結合
パターン認識・データ較正
ローカルデータ圧縮

圧縮 *Sub-Timeframes*

データ結合 (aggregation)
同期グローバル再構成
データ較正、データ圧縮

圧縮 *Timeframes*

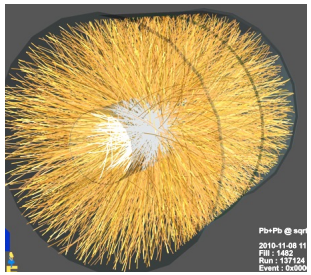
60-75 GB/s

データストレージ・アーカイブ

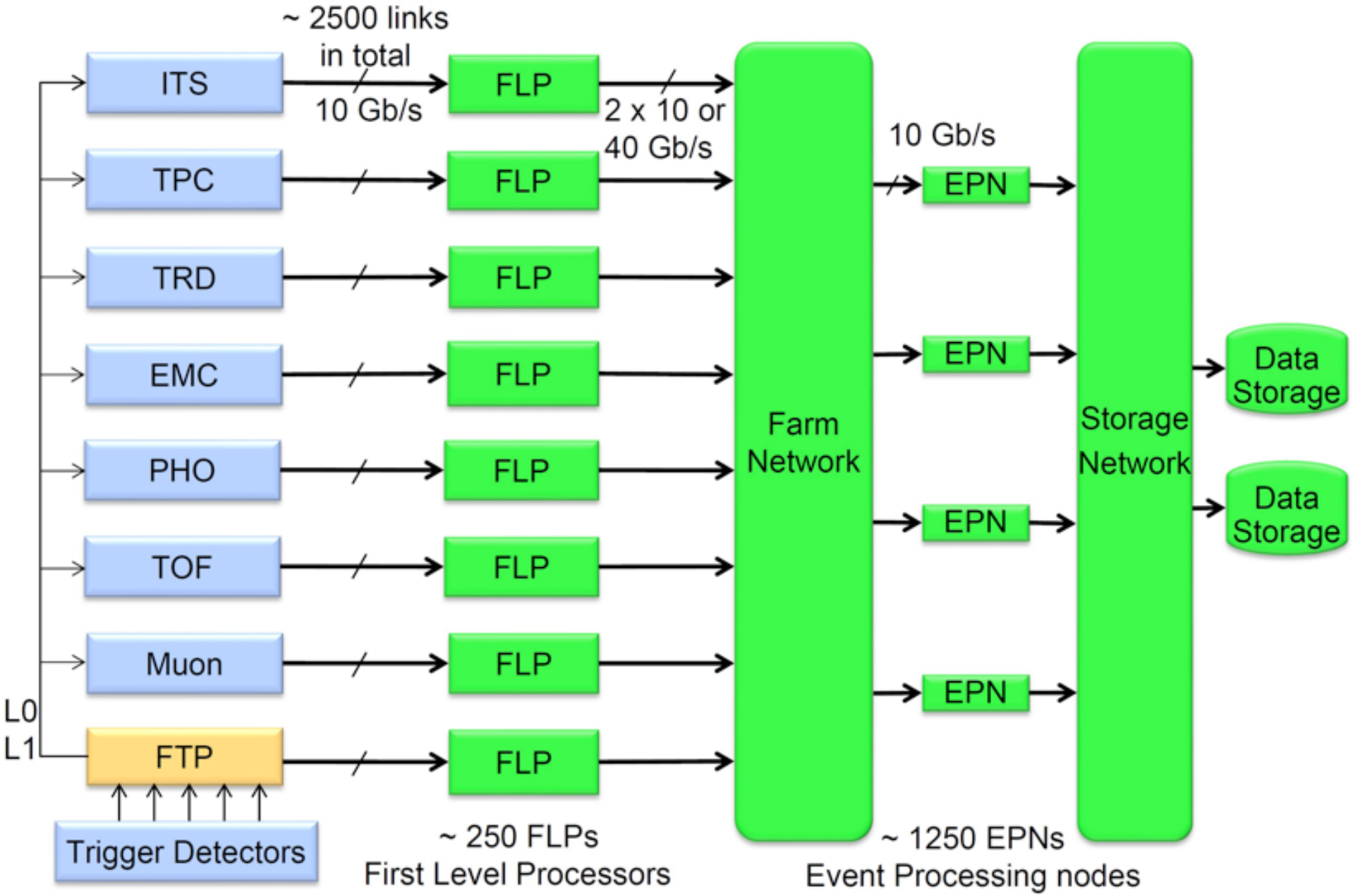
再構成イベント

圧縮 *Timeframes*

非同期再構成、精密データ較正
データ品質管理、イベント抽出



ALICE O2 ハードウェア設計

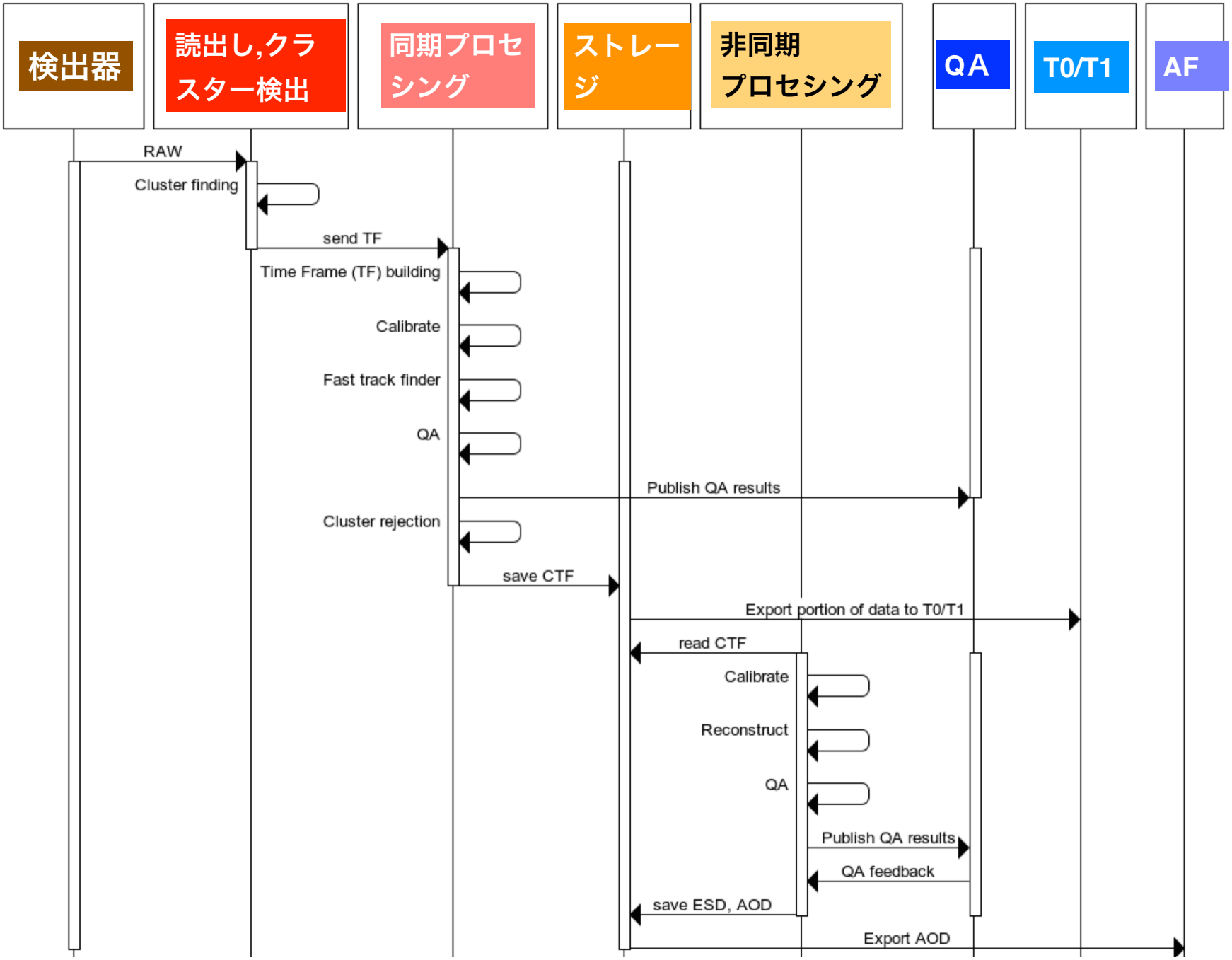


コンピューティングモデル (データフロー)

略称	説明	
RAW	検出器から来るRaw データ	
CTF	**C**ompressed **T**ime **F**rame クラスター情報と飛跡情報を持った検出器読み出しデータ 「中間ファイル」 （O(100 ms) の検出器情報を保持）	
ESD	**E**vent **S**ummary **D**ata	
AOD	**A**nalysis **O**bject **D**ata （物理解析用）	
HISTO	ヒストグラム情報 (ある特定解析用 AODサブセット)	
MC	モンテカルロ・シミュレーション	

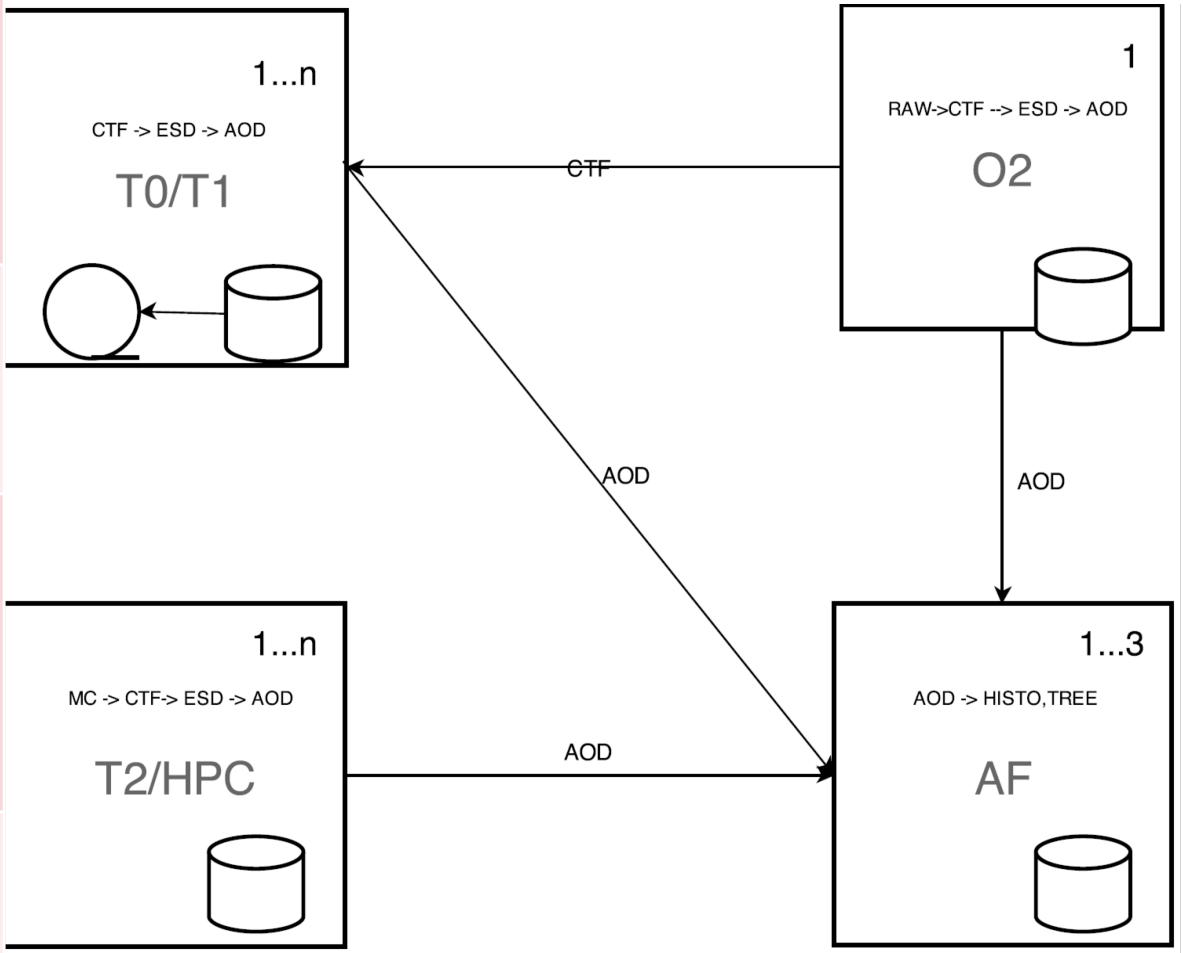
by Pierre Vande Vyvre

データフロー


検出器
読出し,クラスター検出
同期プロセシング
ストレージ
非同期プロセシング
QA
T0/T1
AF
RAW
Cluster finding
send TF
Time Frame (TF) building
Calibrate
Fast track finder
QA
Publish QA results
Cluster rejection
save CTF
Export portion of data to T0/T1
read CTF
Calibrate
Reconstruct
QA
Publish QA results
QA feedback
save ESD, AOD
Export AOD

コンピューティングモデル (O2 プロセシングフロー)

施設	機能
O2	**ALICE Online-Offline Facility (P2、新規建設)** オンライン再構成 (Run 期間中) データストレージ供給 Run終了後:データ較正・再構成ジョブ
T0	**CERN Computer Center** CPU, ストレージ、アーカイブを供給
T1	T0 に高バンド幅 **(100+ Gb)で繋がる Grid サイト** CPU, ストレージ、アーカイブを供給 再構成・データ較正ジョブ
T2	(10+ Gb)で繋がるGrid サイト; **シミュレーションジョブ**
AF	専用解析サイト (e.g. HPC) **Analysis Facility** AOD を収集、アーカイブ、解析ジョブ （物理ワーキンググループで集約）



- 現有のGrid の機能、集約的解析ジョブスキームを維持
- ユーザーに対しオープン、かつ高効率化

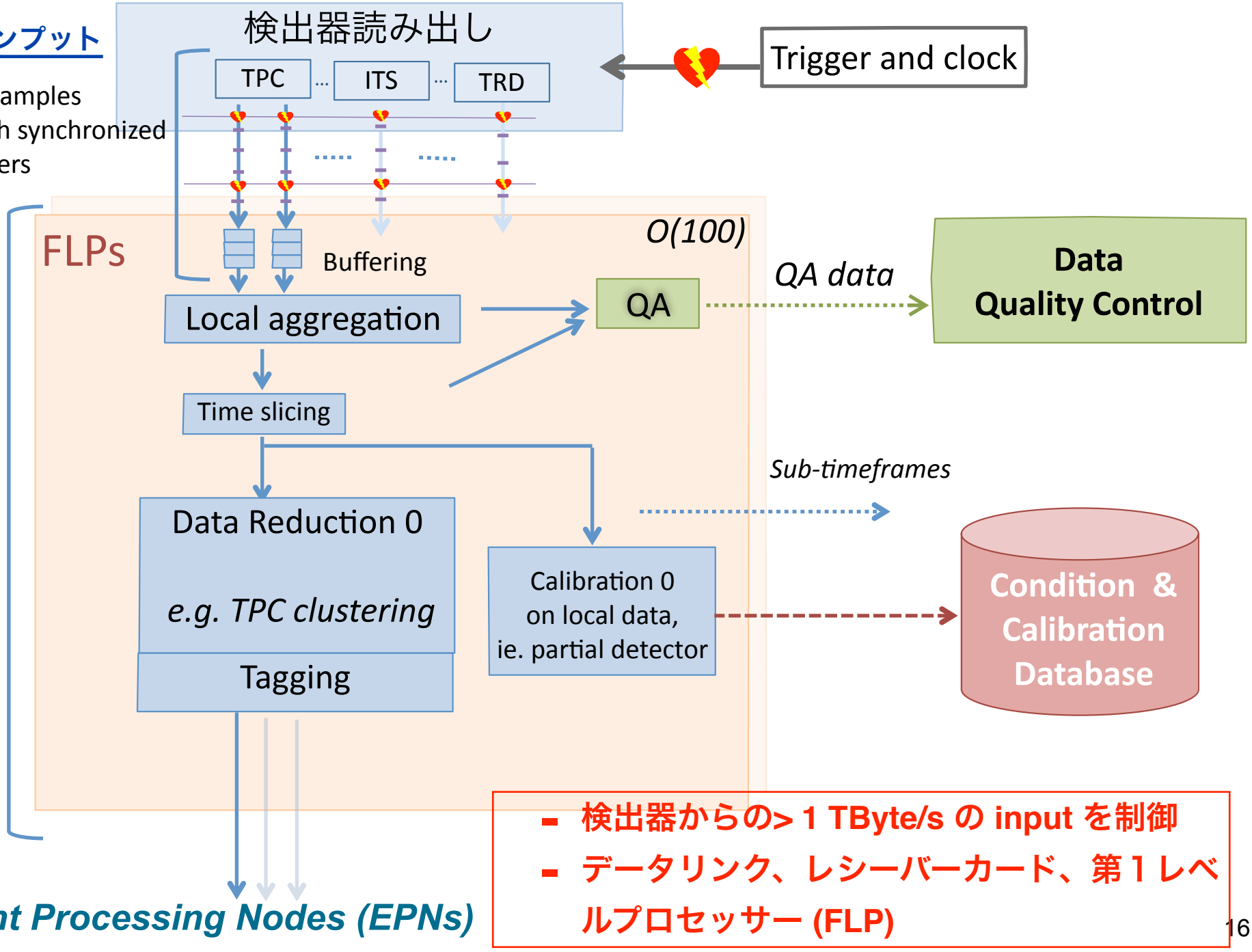
by Pierre Vande Vyvre

ALICE O2: データインプット

by Pierre Vande Vyvre (modified)

Raw データインプット

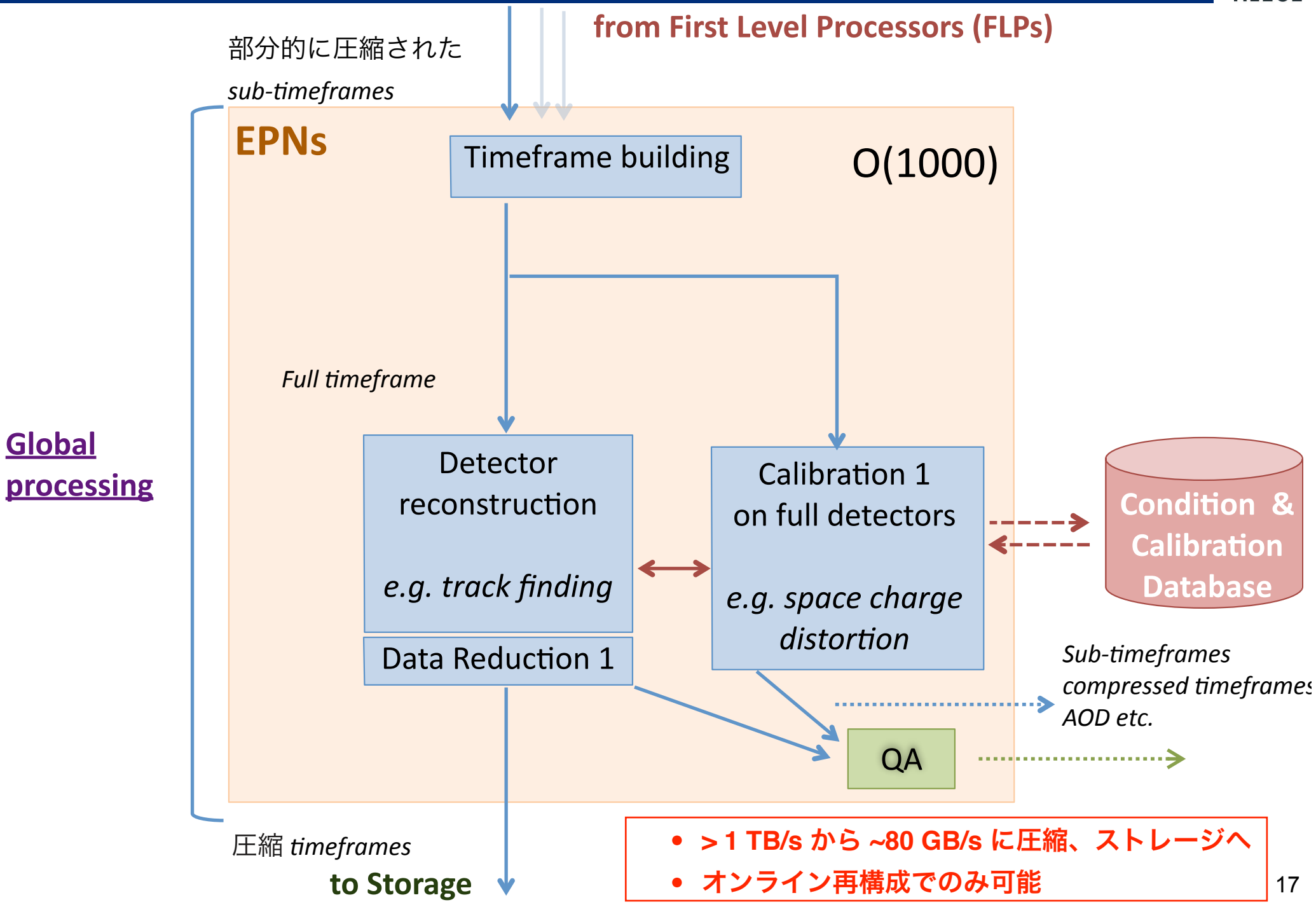
Detector data samples interleaved with synchronized heartbeat triggers



- 検出器からの> 1 TByte/s の input を制御
- データリンク、レシーバーカード、第１レベルプロセッサー (FLP)

ALICE O2: データ圧縮

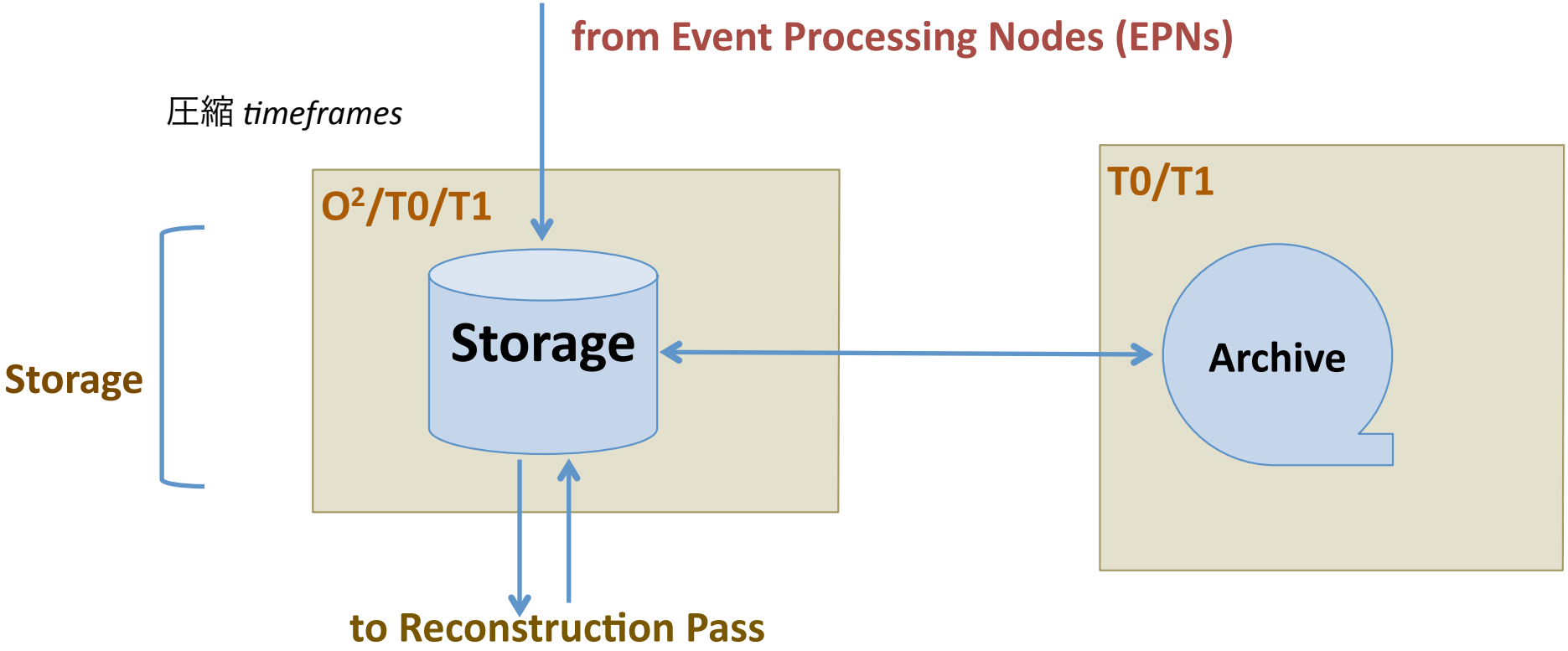
by Pierre Vande Vyvre (modified)



- > 1 TB/s から ~80 GB/s に圧縮、ストレージへ
- オンライン再構成でのみ可能

ALICE O2: データストレージ

by Pierre Vande Vyvre (modified)



•中間データフォーマット（物理解析では直接使わない）：

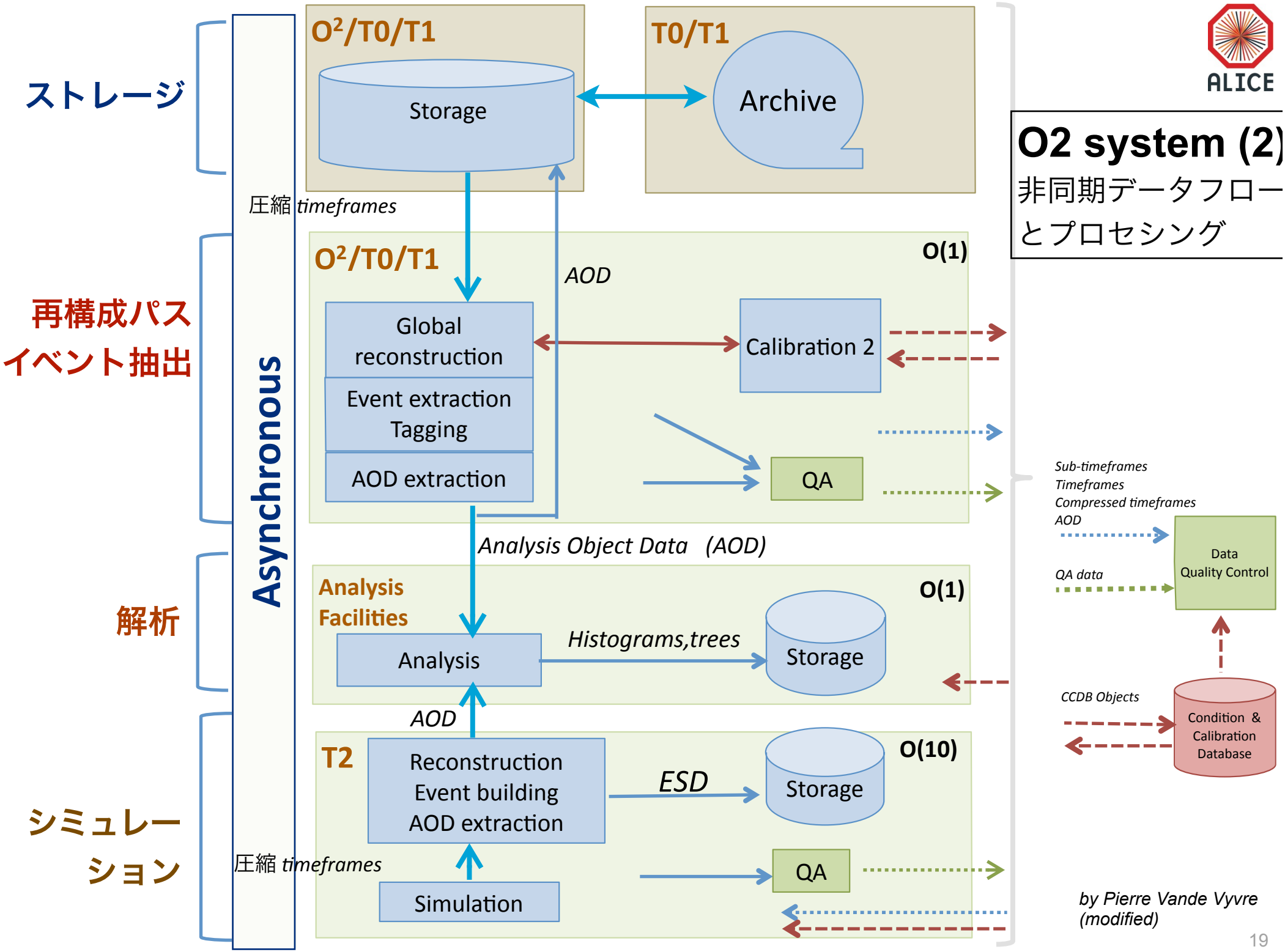
- 80 GByte/s (max.) を制御, ~1,250 超のノードに分散
- 平均負荷：15 GByte/s
- O2 のローカルストレージを利用
- 恒久ストレージはT0 (T1)へ

•最終データフォーマット（物理解析で使用, AOD, ESD）：

- GRID ストレージへ、ユーザーがアクセス可能な状態

O2 system (2)

非同期データフローとプロセシング

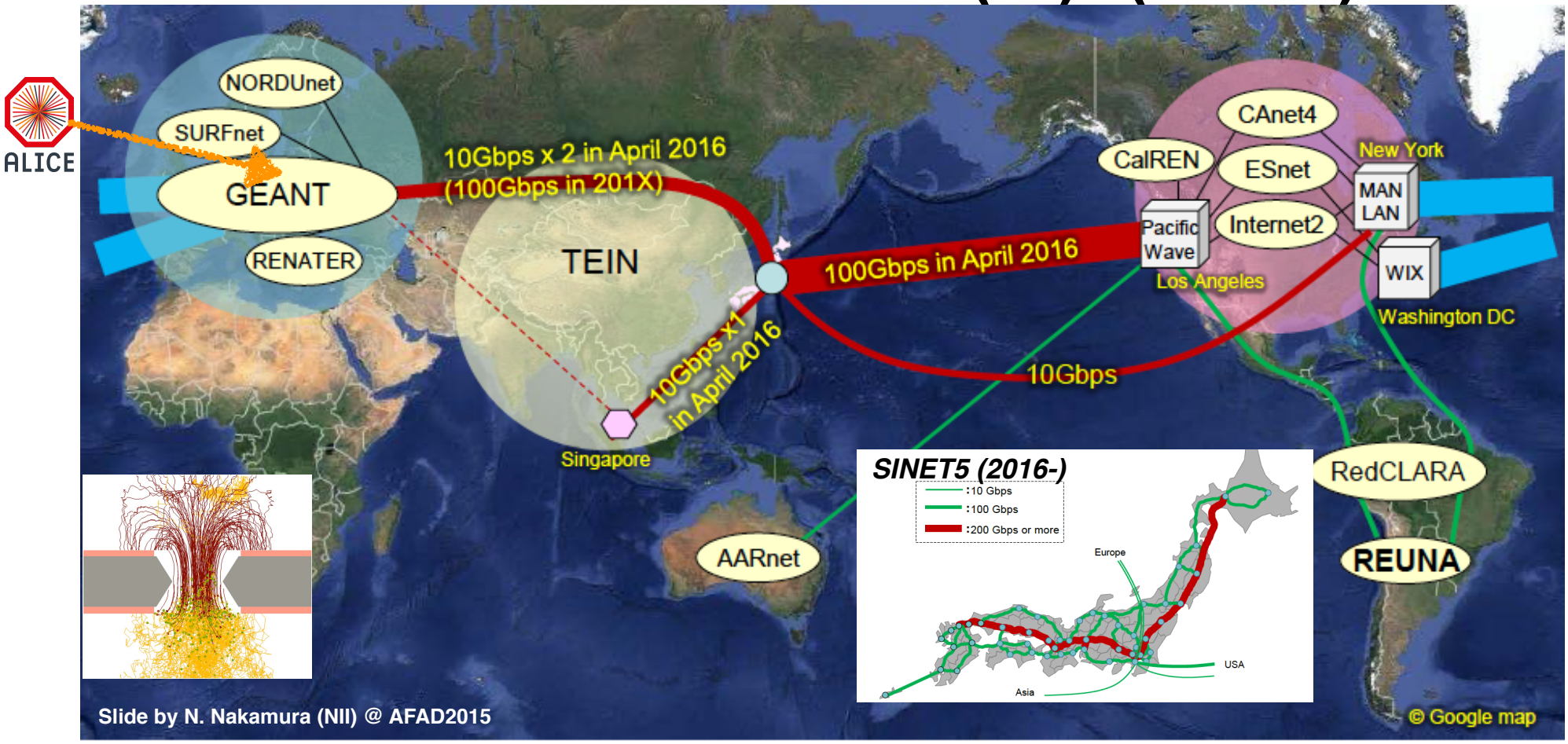


by Pierre Vande Vyvre (modified)

ALICE O2；まとめ

- **2015 - 2017:**
  - LHC Run2 データ収集, Run-1 の２倍以上の重イオンデータ
  - データ収集と同時にRun-2 (+Run-1) の物理解析を同時進行
  - コンピューティングモデル：Run-1 のスキームからほとんど変更なし
  - データ量増大に伴い、アジアでのデータ量増大 (x2?)
- **2018 - :**
  - LHC Run3 データ収集, Run-1 の１００倍の重イオンデータ
  - ALICE O2 の設計を適用
    - O2, T0/T1/T2, AF のスキーム
  - 大幅なデータ圧縮、主に O2 でデータ再構成、アーカイブは T0/T1、
  - 解析は AF、シミュレーションは T2
  - アジアにおける AF?

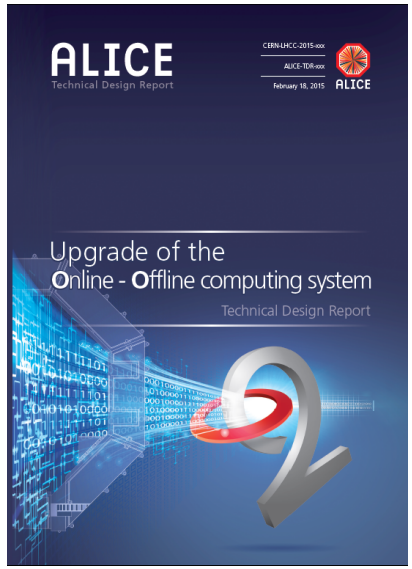
ネットワークと ALICE(-J)（2016-）



- 世界規模 Grid 計算網モデル再構築：新 ALICEデータセンター O2 (@CERN)と解析センター (AF)の建設,運用
- ALICE 日本グループ：Run-3 に向けたコンピューティングへの貢献が期待されている
- **ALICE-J Tier-2**
  - 広島大（現 ALICE T2 拠点）に加え、筑波大に Tire 2 の構築を検討
  - HPC, SINET5 (2016-, 国内200Gbps のバックボーン, 100 Gbps日 ⇄米国、欧州) の利用検討

まとめ

- **ALICE におけるオンライン・オフライン・コンピューティング高度化が進行中 (2018 年稼働予定)**
  - **50 kHz で鉛-鉛衝突（最小バイアス）事象を連続読み出し**
  - **1 TB/s raw データ** が検出器から生成、大幅なデータ圧縮が不可欠
  - 80 GB/s に圧縮してストレージへ送り、物理アウトプットを効率よく導出
  - O2 スキーム: O2 データセンター（ALICE実験施設の近辺に設立）(T0/T1)におけるオンラインデータ再構成とデータ較正、解析ファーム (AF)、シミュレーション (T2)
- O2: テクニカル・デザイン・レポート(TDR)の ALICE内でレビュー中、LHCC レビュー (2015年4月)
- ALICE 日本グループ web ページ (日本語)
  - http://alice-j.org




Thank you for your attentions.