Bundesministerium
für Bildung
und Forschung

ALICE

BMBF Forschungsschwerpunkt
ALICE Experiment
202
ALICE

Data Distribution and Load Balancing for the ALICE Online-Offline System

G. Nešković for the ALICE Collaboration

DRAFT: ALICE CHEP18 Rehearsals
25.06.2018

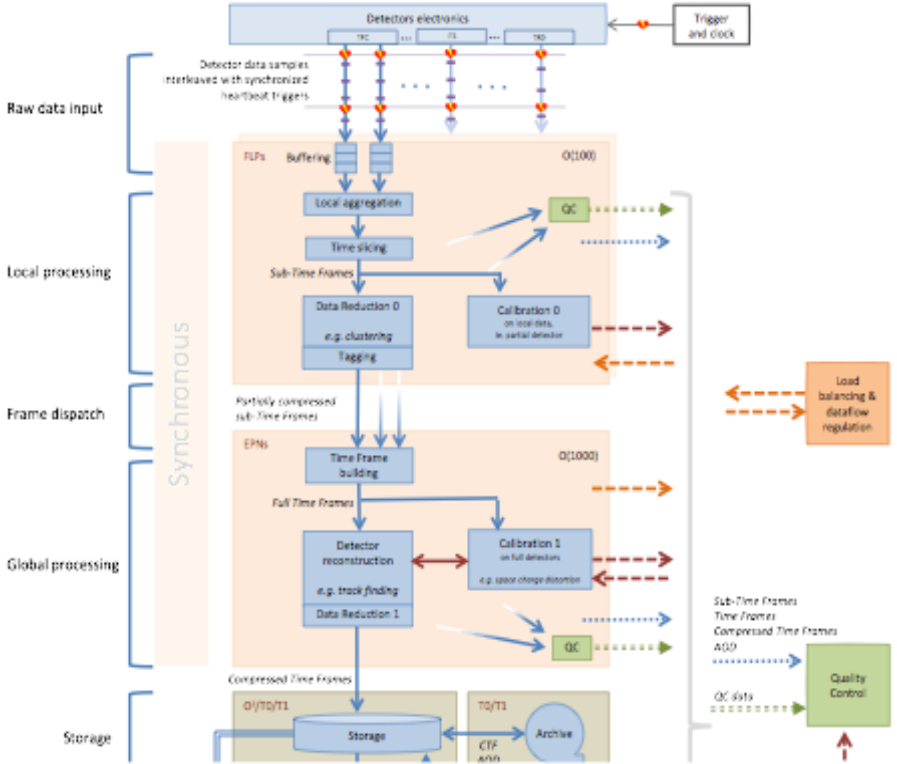
FIAS Frankfurt Institute
for Advanced Studies

GOETHE
UNIVERSITÄT
FRANKFURT AM MAIN

ALICE O²
Data Flow in Synchronous Processing

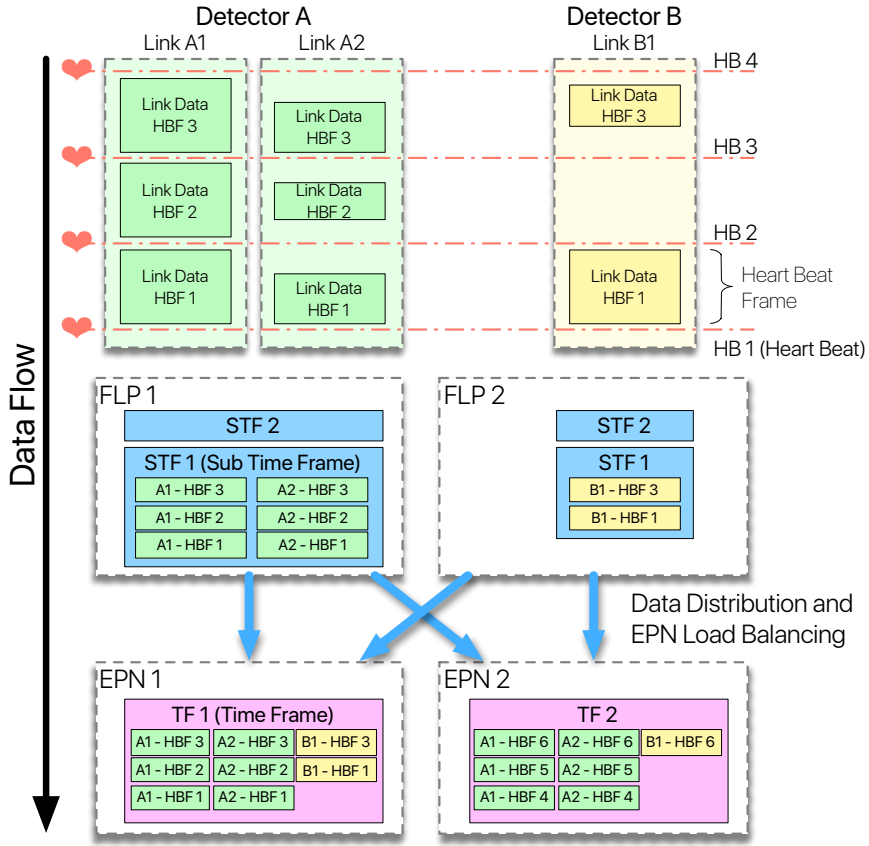
- Scope of the talk:
  - ALICE O² Data Flow during the Synchronous Processing
- Stages of the Synchronous Processing :
  - Raw Detector Data Recording
  - Local Processing
  - Global Data Aggregation and Load Balancing
  - Global Processing

TODO: Make a more suitable Figure to illustrate the data distribution process more clearly



ALICE O² Synchronous processing
Data Flow

- Heart-Beat Frame (HBF):
  - Detector data recorded in-between two HBs
  - For both contiguous and triggered readout detectors
- Sub Time Frame (STF):
  - Subset of Detector Data recorded on a single First Level Processor (FLP) node
  - Accumulated during a time period (~20ms)
  - Joined with any results of Local Processing on the FLP
  - The size highly depends on the Detector, and geographical region of the links
- Time Frame (TF):
  - Complete Set of all Detectors Data, recorded for the same TF interval
  - Size ~12 GB/TF
  - Input for the Global Synchronous Reconstruction performed for additional data reduction



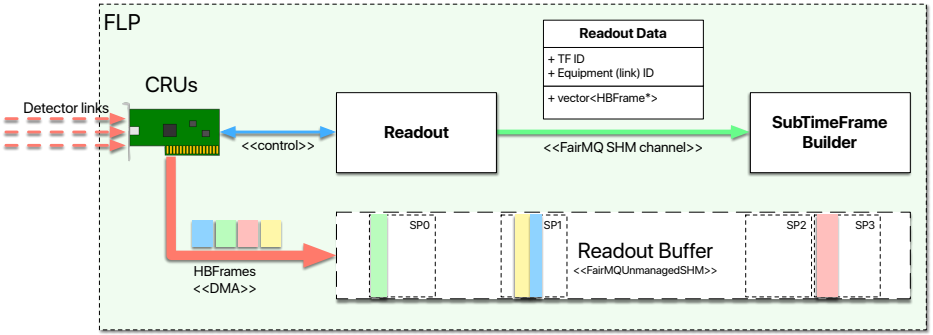
ALICE O^2 Data Distribution
Load Balancing Requirements

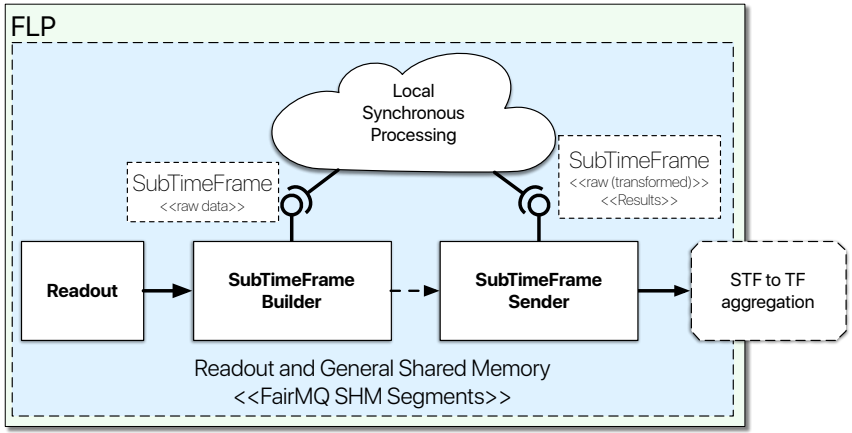
- Global Detector Readout Load Balancing (FLP domain):
  - Detector Tests and Commissioning
  - Maximize likelihood of simultaneous data recording for as many Data Links as possible
  - Coordinated by the Central Trigger Processor (CTP) :
    - Fixed: Deterministically reject HBFrames (data rate throttling)
    - Automatic: Globally evaluate all recorded HBFrames and discard STFs on negative decision
- Network Traffic Shaping and Congestion Avoidance (Network domain):
  - Maintain the model of the Network Topology and FLP->EPN Links
  - Use the link utilization to schedule the TF aggregation minimizing the Network congestion
- EPN Load Balancing (EPN domain):
  - Collect available processing resources from EPNs participating in the run
  - Evenly distribute TFs to available EPNs

ALICE O² Data Distribution
FLP and Intra-Node Data Transport

- Efficient Data Transport on a single node:
  - Solution: Make the CRU-DMA engines stream Data to the *Shared Memory Segment*
  - Data Block are *never* copied by the CPU:
    - No memcpy(), memmove(),...
    - Leaves CPU cycles and memory bandwidth available for the Local Processing tasks

- Multi-Process approach:
  - Provided by the ALFA Framework
  - New Shared Memory transport in FairMQ for Intra-Node communication*
  - Data flows from process to process via exchange of SHM Messages (handles)
  - Objects automatically reclaimed when not needed (e.g. when STF is transmitted)

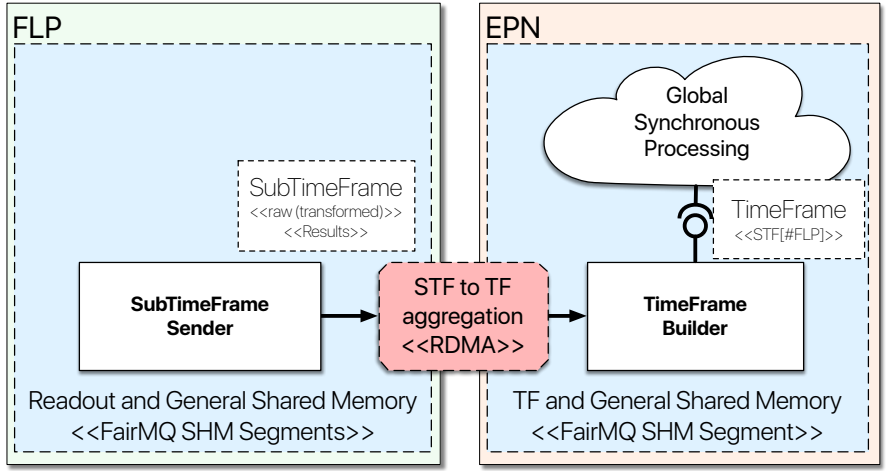




ALICE O² Data Distribution
Inter-Node Data Transport: Remote Direct Memory Access

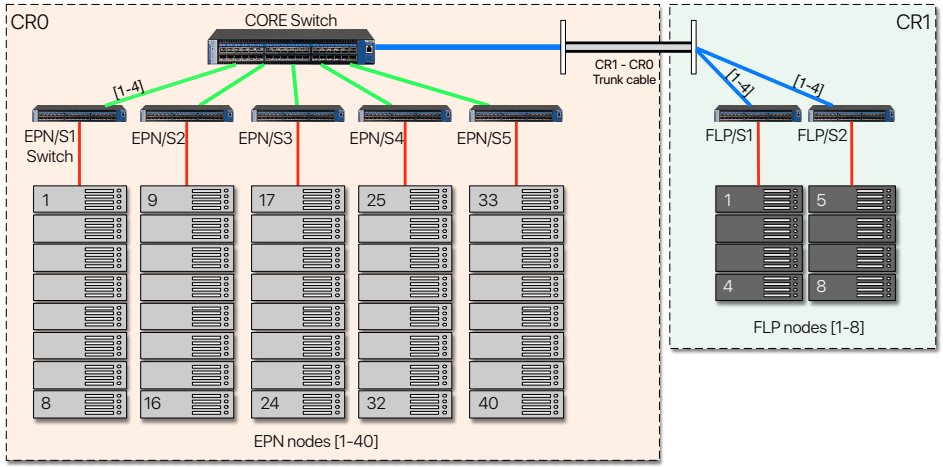
- Extend efficient inter-node data transport onto the Network:
  - Supported by modern HPC interconnects:
    - InfiniBand
    - RoCE (requires *link-level flow* and *congestion control* for reliable operation)
  - Use network hardware to move data out of the node (RDMA)
    - Higher bandwidth and lower latencies with minimal CPU overhead
  - Avoid expensive TCP/IP stack overhead

- New FairMQ transport for RDMA*:
  - STF aggregated inside a SHM segment of EPNs
  - Suitable for Data Flow in Synchronous Processing using FairMQ SHM channels
  - No explicit CPU data copies, *end-to-end!*



ALICE O² Data Distribution
Network Load-Balancing

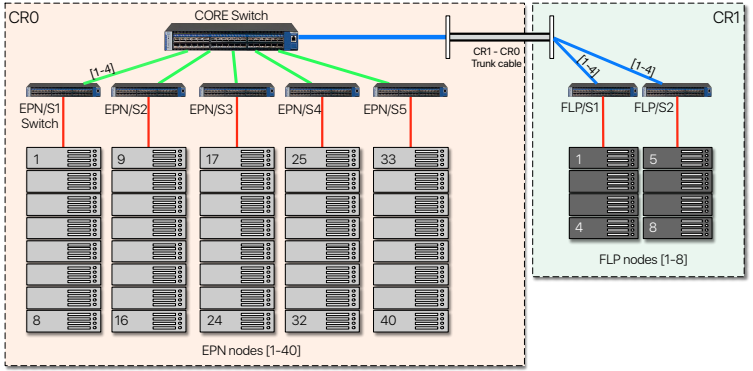
- Network Requirements for Data Distribution:
  - DR_{FLP} = 4Tb/s FLP->EPN
  - Uneven FLP data rates
  - Fat-tree like network with different blocking ratios for EPN and FLP nodes
  - Support for staged deployment
- Objectives for Load-Balancing:
  - Steer data to available EPNs while avoiding network congestion
  - Simple Round-Robin EPN selection is not suitable:
    - Variable processing times
    - Unpredictable data paths through the core links
  - *Solution:* Maintain the Model of Network Link utilization



ALICE O² Data Distribution
TF Scheduling

- TF Schedule preparation:
  - Create the Connection List for all FLP and EPN pairs:
  - Possible with source based routing of InfiniBand
  - Assign initial scores for for each segment (link bandwidth)

- TF Distribution Schedule properties:
  - Contains more TFs than **DataRate$_{FLPs}$ / Bandwidth$_{EPN}$**
  - EPNs do not repeat
  - Keep the remaining score of core links as equal as possible (congestion avoidance)
  - Distribute TFs with *the most* processing resources (EPN load balancing)



FLP 1 - FLP/S1	FLP/S1 - CORE [1]	CORE-EPN/S1 [1]	EPN/S1 - **EPN 1**
FLP 1 - FLP/S1	FLP/S1 - CORE [1]	CORE-EPN/S1 [1]	EPN/S1 - EPN 2
FLP 1 - FLP/S1	FLP/S1 - CORE [2]	CORE-EPN/S1 [2]	EPN/S1 - EPN 3
FLP 2- FLP/S1	FLP/S1 - CORE [1]	CORE-EPN/S2 [1]	EPN/S2 - EPN9
FLP 2 - FLP/S1	FLP/S1 - CORE [1]	CORE-EPN/S2 [1]	EPN/S2 - EPN10
FLP 2 - FLP/S1	FLP/S1 - CORE [2]	CORE-EPN/S2 [2]	EPN/S2 - **EPN11**
FLP 3- FLP/S1	FLP/S1 - CORE [1]	CORE-EPN/S3 [1]	EPN/S3 - EPN17
FLP 3 - FLP/S1	FLP/S1 - CORE [3]	CORE-EPN/S3 [1]	EPN/S3 - **EPN18**
FLP 3 - FLP/S1	FLP/S1 - CORE [4]	CORE-EPN/S3 [1]	EPN/S3 - **EPN19**
FLP 4- FLP/S1	FLP/S1 - CORE [4]	CORE-EPN/S1 [2]	EPN/S1 - **EPN 3**
FLP 4 - FLP/S1	FLP/S1 - CORE [1]	CORE-EPN/S1 [1]	EPN/S1 - **EPN 4**
FLP 4 - FLP/S1	FLP/S1 - CORE [2]	CORE-EPN/S1 [3]	EPN/S1 - **EPN 5**

ALICE O^2 Data Distribution
Summary

- Load Balancing:
  - Readout system: CPT-CRU control loop
    - Maximizes likelihood of global data recording
    - Discards incomplete TFs
  - EPN Load Balancing:
    - Evenly utilize available processing resources
  - Network:
    - Perform traffic shaping and congestion avoidance

- Data Distribution Data Flow:
  - Efficiency enabled by the new SHM and RDMA Transports in FairMQ
  - Flexible deployment with DDS on a cluster or dedicated test setups

- Backups

Source Based Routing

- Ability to determine the data path through the network on sender side
  - Also referred to as “path addressing”
- InfiniBand:
  - Subnet Manager (SM) creates optimal paths for each pair of nodes (N^2 number of paths)
  - Each HCA port can be assigned k LIDs ($k=2^m$), leading to the total $2^k \times N^2$ paths to choose from
- Ethernet:
  - Source based routing not supported
  - BGP-ECMP uses hash vales to select the paths (and performs load distribution when multiple equally suitable paths exist)
  - Can lead to unpredictable congestion in the network core

ALICE O² Data Distribution
TimeFrame Scheduling

- Schedule contains a list of (TF,EPN) pairs:
  - Distributed to all FLPs in advance
  - Longer list avoids latency issues for reliable schedule distribution
- Each FLP performs transfers from the schedule in an unique permutation:
  - Avoid hot-spots in the network core
    - Unexpected congestion control of the underlying network fabric
  - Prevent the "in-cast" traffic patterns at the receiver
  - Investigate optimizing transfer schedules of high data rate FLPs (requires more buffer space at FLPs)