Bundesministerium
für Bildung
und Forschung

BMBF Forschungsschwerpunkt
ALICE Experiment
202
ALICE

Data Distribution and Load Balancing for the ALICE Online-Offline (O^2) System

Gvozden Nešković for the ALICE Collaboration

Frankfurt Institute for Advanced Studies

12.07.2018

FIAS Frankfurt Institute
for Advanced Studies

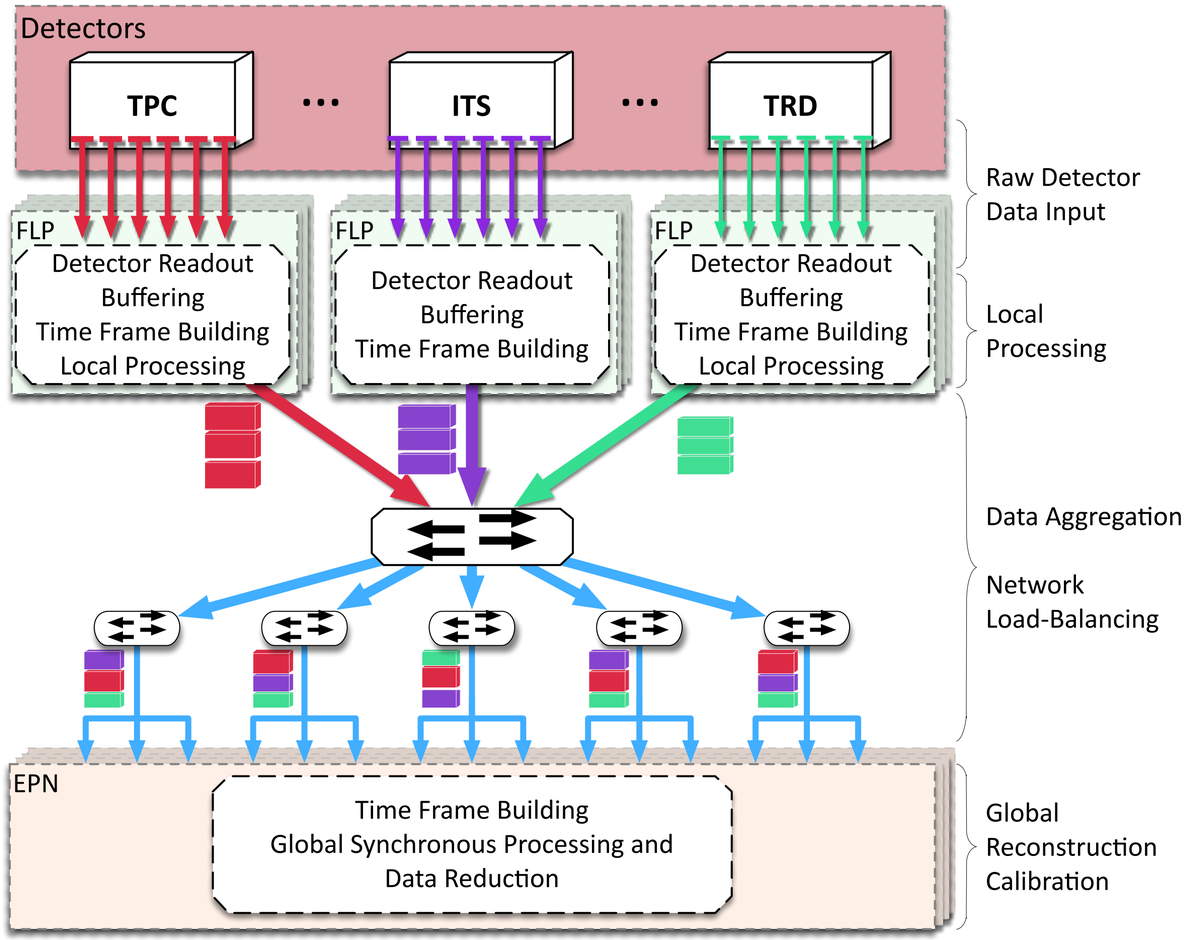
ALICE

GOETHE
UNIVERSITÄT
FRANKFURT AM MAIN

ALICE O²

Synchronous Processing

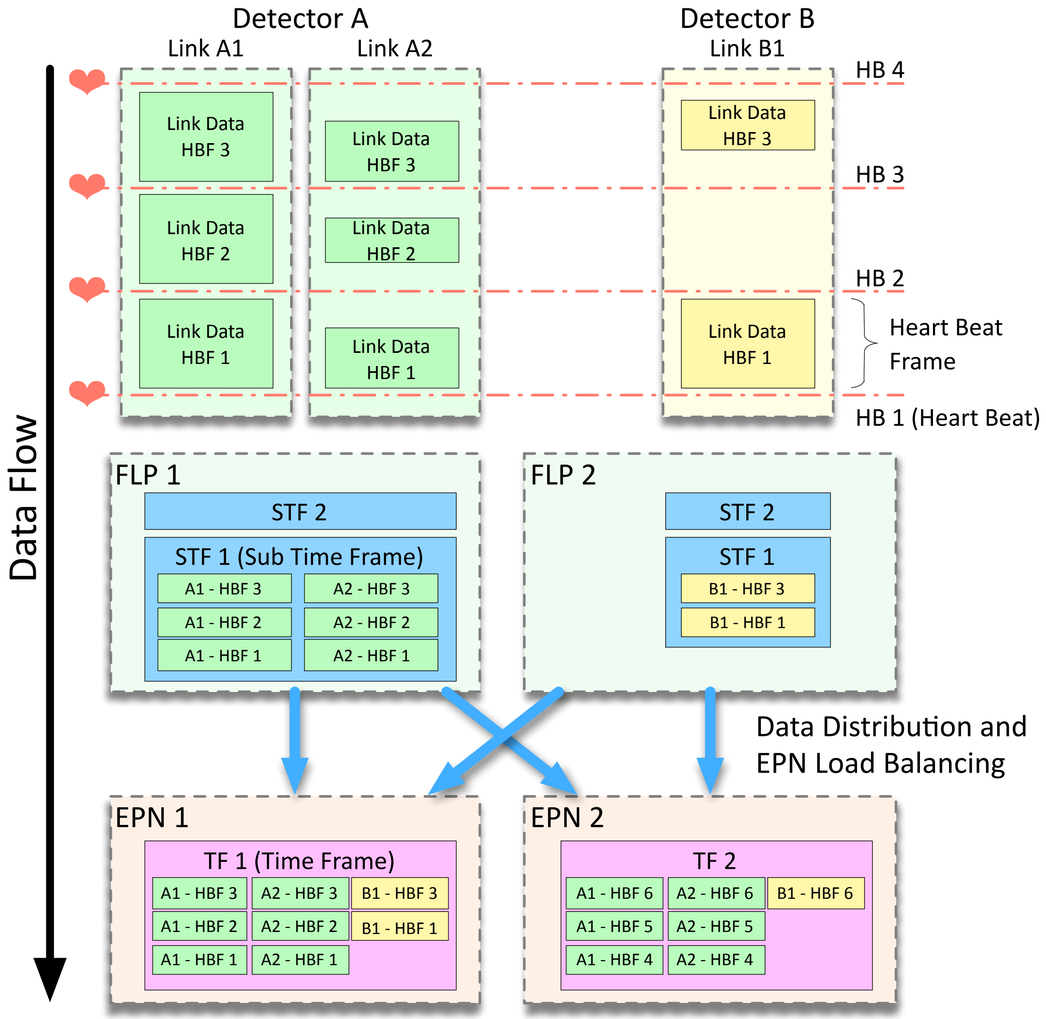
- Scope of the talk:
  - ALICE O² data flow during the synchronous processing
  - Data load balancing and network traffic shaping
- Stages of the synchronous processing:
  - Raw detector data recording
  - Local processing
  - Global data aggregation and load balancing
  - Global processing



ALICE O² Synchronous processing

Data Flow

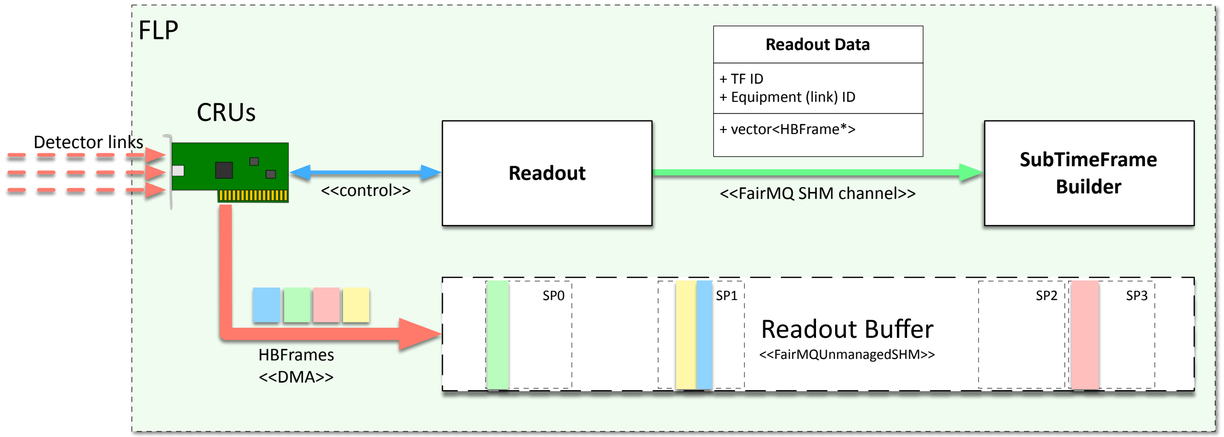
- Heart-Beat Frame (HBF, ~90 μs):
  - Detector data recorded in between two heart beats
  - Both contiguous and triggered readout detectors
- Sub Time Frame (STF, ~20 ms):
  - Subset of detector data recorded on a single First Level Processor (FLP)
  - The size depends on the detector and geographical region of the links
- Time Frame (TF, ~20 ms):
  - Complete data sets of all the detectors
  - Input for the global synchronous reconstruction



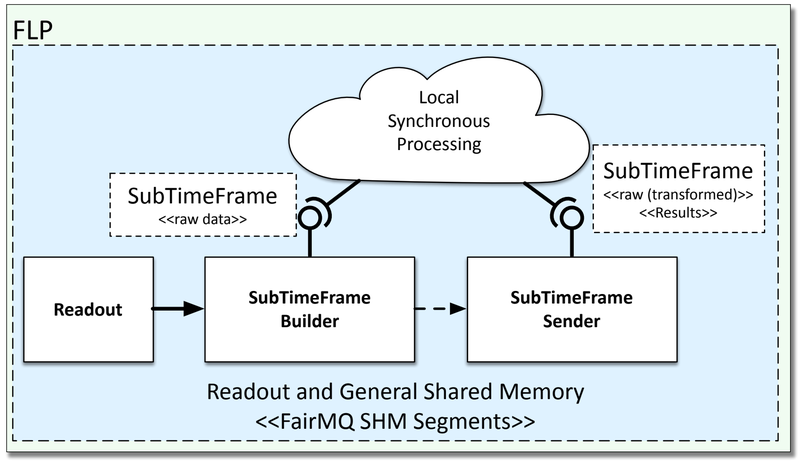
ALICE O² Data Distribution

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 (SHM)
  - Data blocks are never copied by the CPU



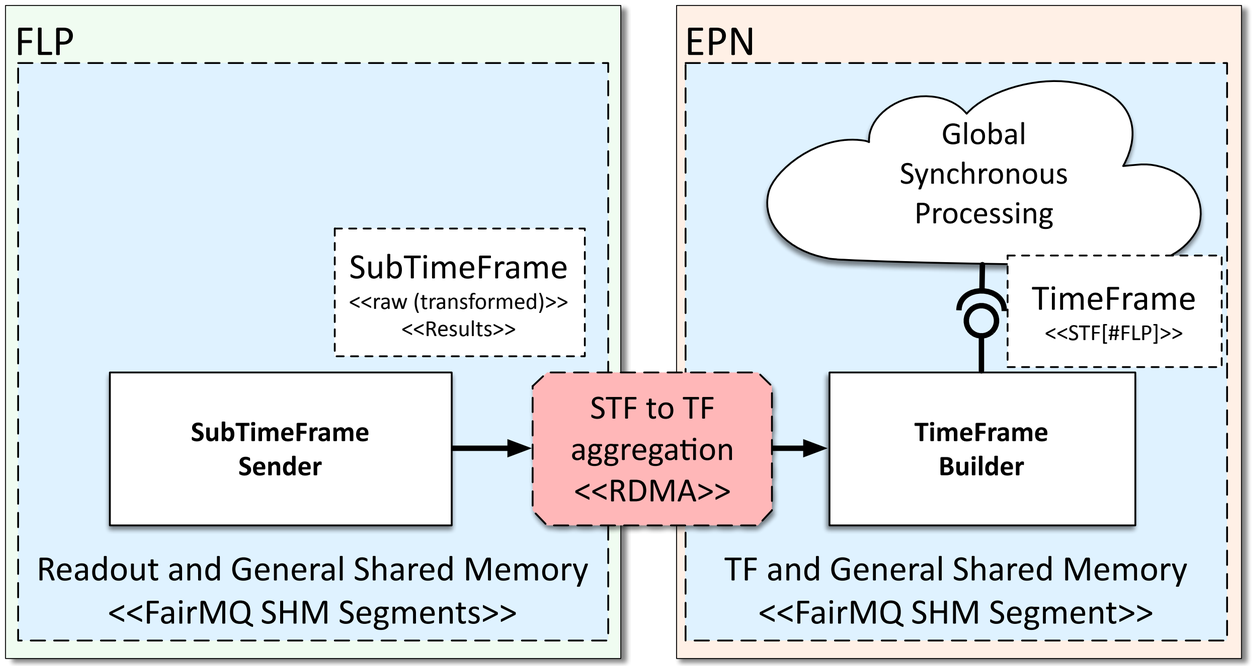
- Multi-process approach:
  - Provided by the ALFA Framework
  - New shared memory transport in FairMQ for intra-node communication
  - Data flows from process to process via exchange of SHM messages



ALICE O² Data Distribution

Inter-Node Data Transport: **R**emote **D**irect **M**emory **A**ccess

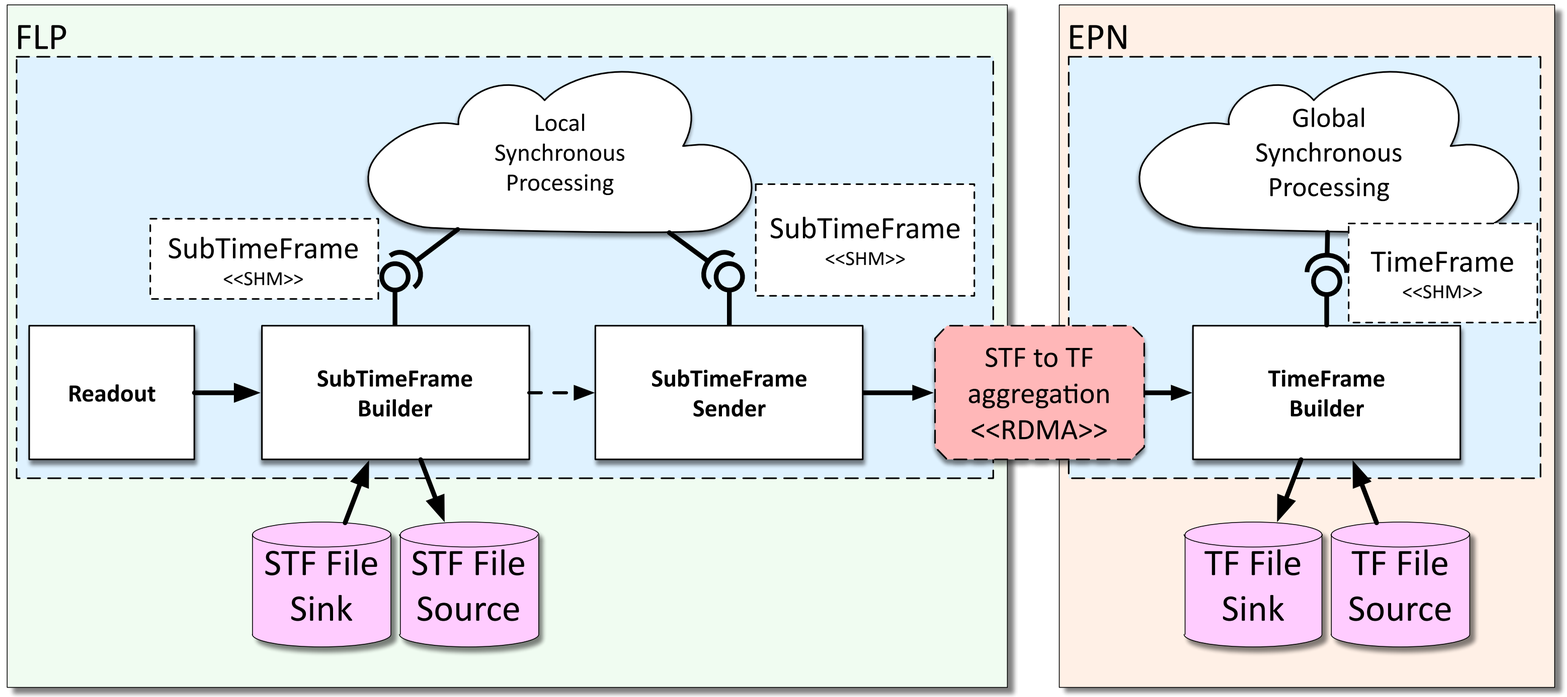
- Efficient inter-node data transport:
  - Supported by modern HPC interconnects
  - Use network hardware to move data out of the node (RDMA, RoCEv2)
  - Higher bandwidth and lower latencies with minimal CPU overhead
- New FairMQ transport for RDMA:
  - Data transport offloaded the to the network interface cards
  - TFs placed into SHM segment of EPNs
    - No explicit CPU data copies, end-to-end!



ALICE O² Data Distribution

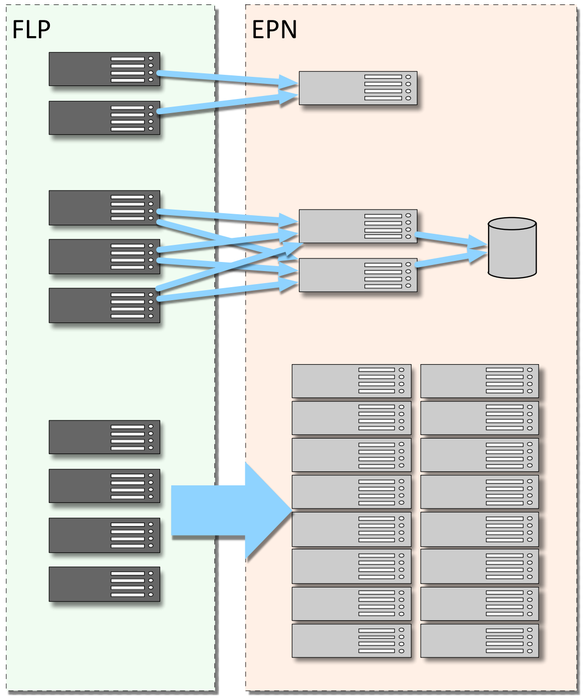
Complete Data Distribution Chain

- Components:
  - SubTimeFrame Builder
  - SubTimeFrame Sender
  - TimeFrame Builder



Components of the Data Distribution Chain

- Flexible Deployment:
  - Small readout test
  - Detector data-taking
  - O² commissioning

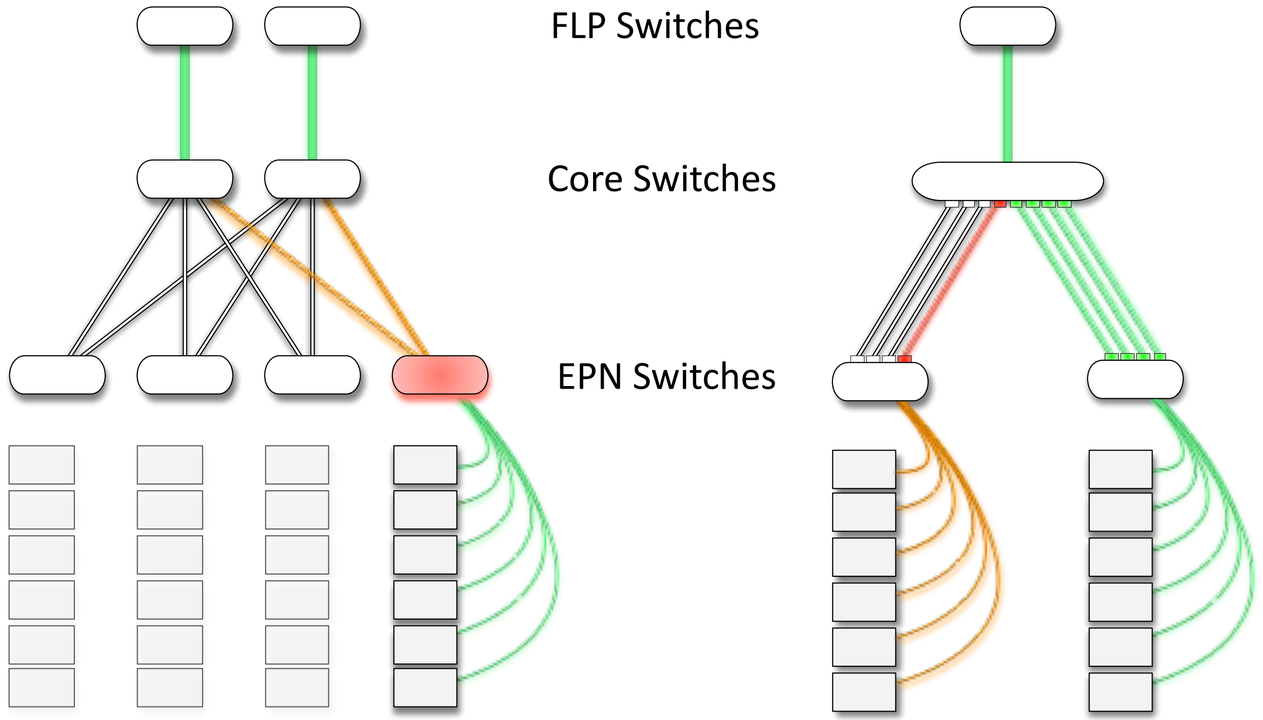


Deployment Setups

ALICE O² Load Balancing

Network traffic shaping

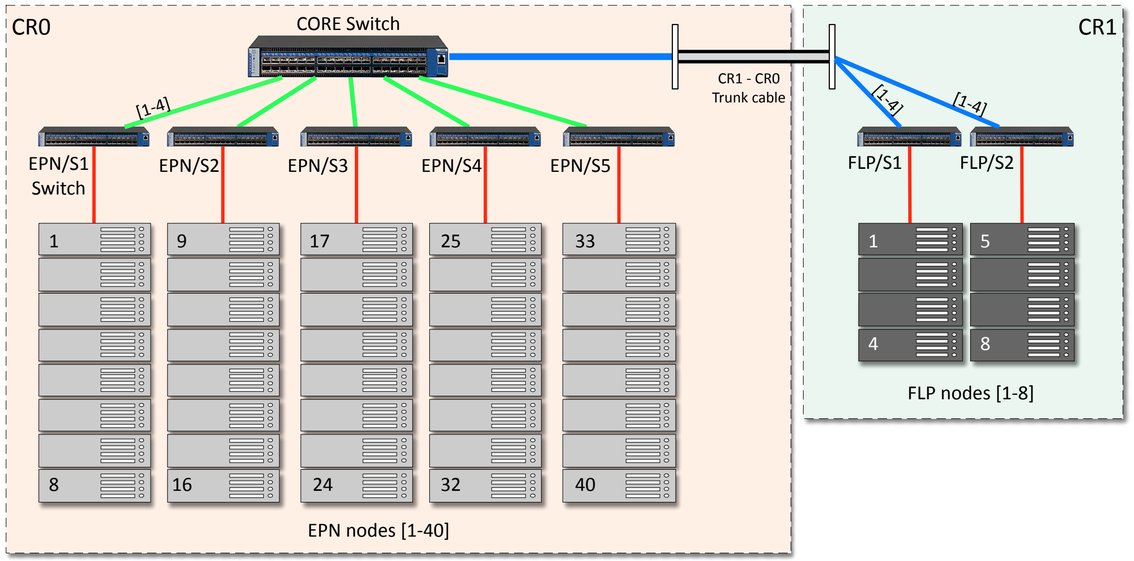
- Network requirements:
  - Aggregate FLP data rate of 4 Tb/s
  - Uneven data rates between FLPs
  - Several TFs aggregated at the same time
- Network congestion decreases effective throughput
  - Congestion can occur on the level of switches, links, and ports
- Network traffic shaping
  - Avoid congestion by spreading traffic evenly across the network fabric
  - Prevent simultaneous data streams to a single EPN (“In-cast”)



ALICE O² Network traffic shaping

TF Schedules (ongoing research)

- Ensure efficient use of EPN processing resources
  - Evenly distribute TFs to EPNs with free compute and memory resources
- Implement resource accounting scheme
  - Maintain information about EPNs ready to process TFs
  - Distribution according to realistic model of network link utilization
  - Result is a schedule for distribution of upcoming TFs for all FLPs
- TF Schedule properties:
  - Contains sufficient number of TFs (EPNs) to absorb aggregate FLP data rate
  - Ensures schedule arrival is in time
  - Singular occurrence of any given EPN in a schedule
  - Links in the network evenly utilized



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

Example of a network topology and potential TF Schedule

ALICE O^2 Data Distribution and Load Balancing

Summary

- Data distribution for synchronous processing:
  - Zero-copy intra-node data transport using SHM
  - CPU offloaded inter-node data transport with RDMA
  - O^2 data distribution chain ready for detector tests

- Ongoing research in load balancing of processing and network resources:
  - Efficient use of EPN processing resources
  - Network traffic shaping for congestion avoidance