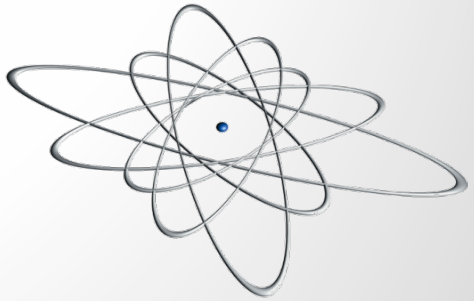
ALFA:
A common concurrency, message based framework for ALICE and FAIR experiments

Mohammad Al-Turany
GSI scientific computing

This talk

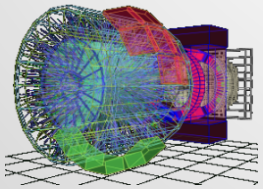
- Introduction: GSI / FAIR
- FairRoot
- From FairRoot to ALFA
- Basic features and components of ALFA
- Examples and Prototypes

Research Areas at GSI

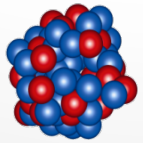


Nuclear Physics

Nuclear reactions
Superheavy elements
Hot dense nuclear matter



Alice



Atomic Physics

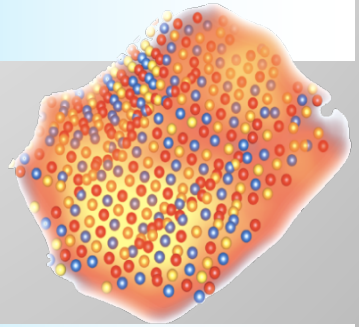
Atomic Reactions
Precision spectroscopy of highly charged ions

Biophysics and radiation medicine

Radiobiological effect of ions
Cancer therapy with ion beams

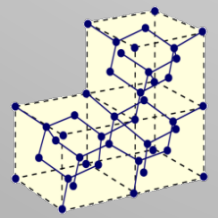
Plasma Physics

Hot dense plasma
Ion-plasma-interaction



Materials Research

Ion-Solid-Interactions
Structuring of materials with ion beams



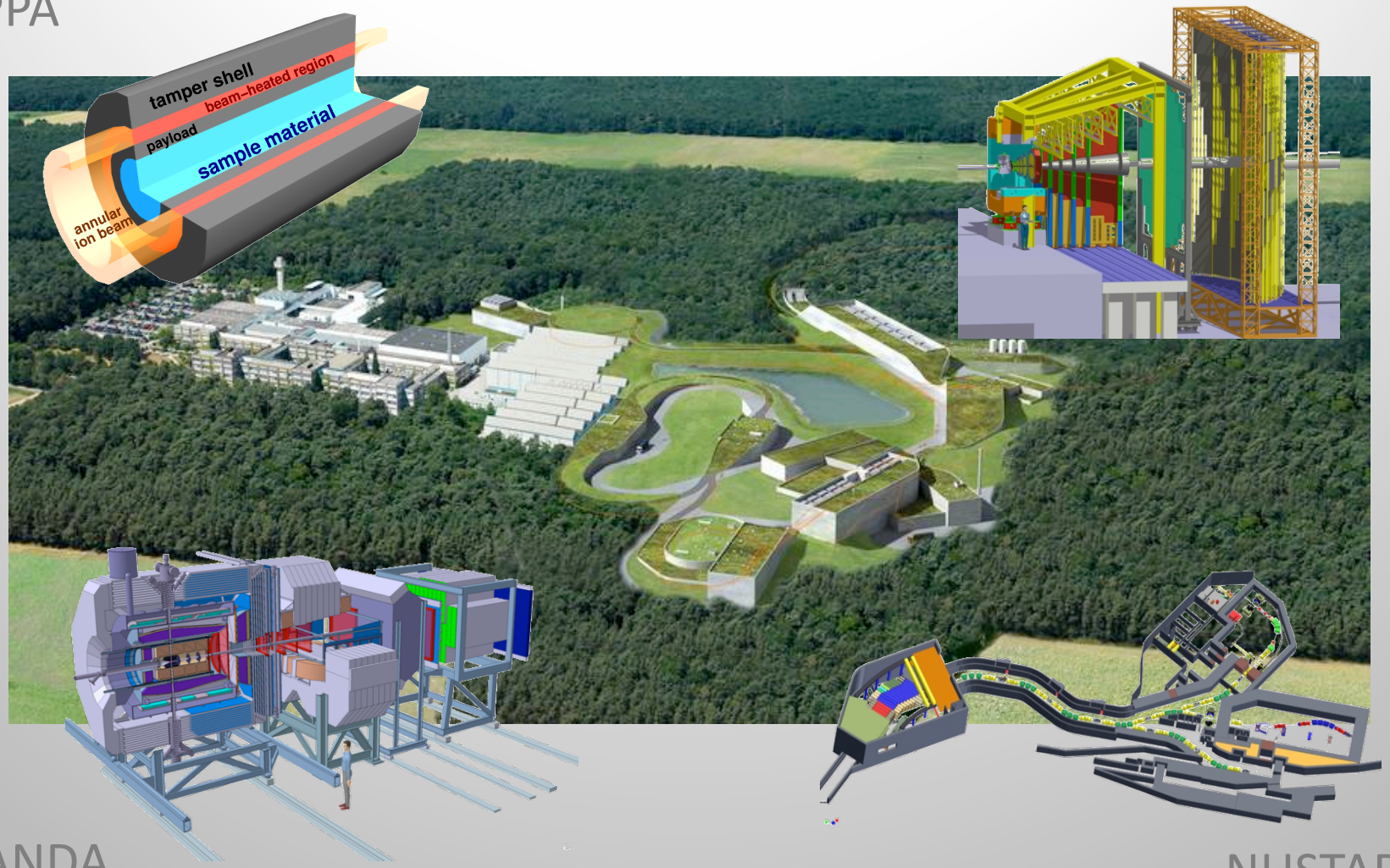
Accelerator Technology

Linear accelerator
Synchrotrons and storage rings

FAIR – four research pillars

APPA

CBM



PANDA

NUSTAR

FAIR – four research pillars

APPA
- Atomic Physics and Fundamental Symmetries,
- Plasma Physics,
- Materials Research,
- Radiation Biology,
- Cancer Therapy with Ion Beams / Space Res.

CBM
- Dense and Hot Nuclear Matter

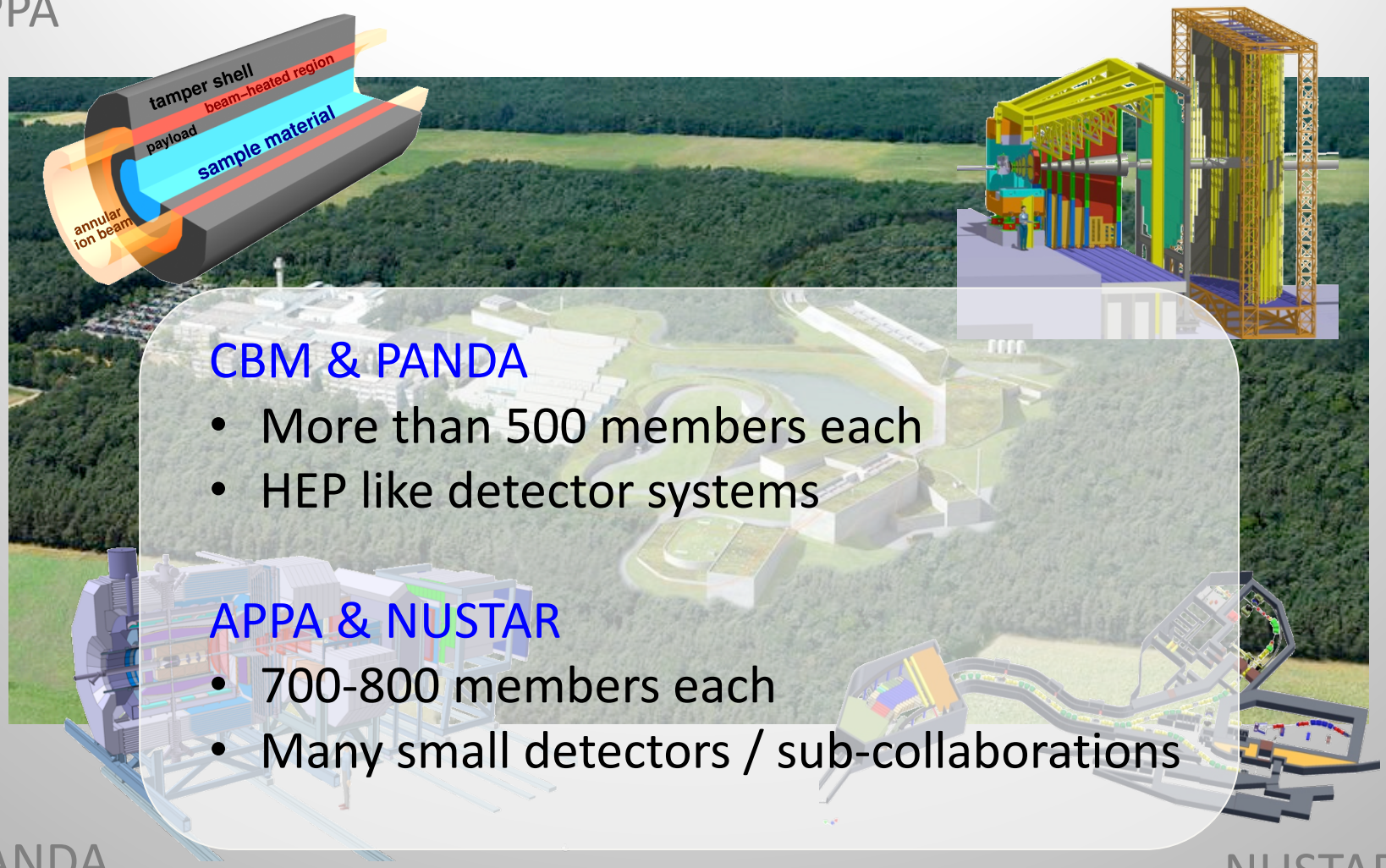
NUSTAR
- Nuclear Structure far off stability,
  Physics of Explosive Nucleosynthesis (r process)

PANDA
- Hadron Structure & Dynamics with cooled antiproton beams

FAIR – four research pillars

APPA

CBM



CBM & PANDA

- More than 500 members each
- HEP like detector systems

APPA & NUSTAR

- 700-800 members each
- Many small detectors / sub-collaborations

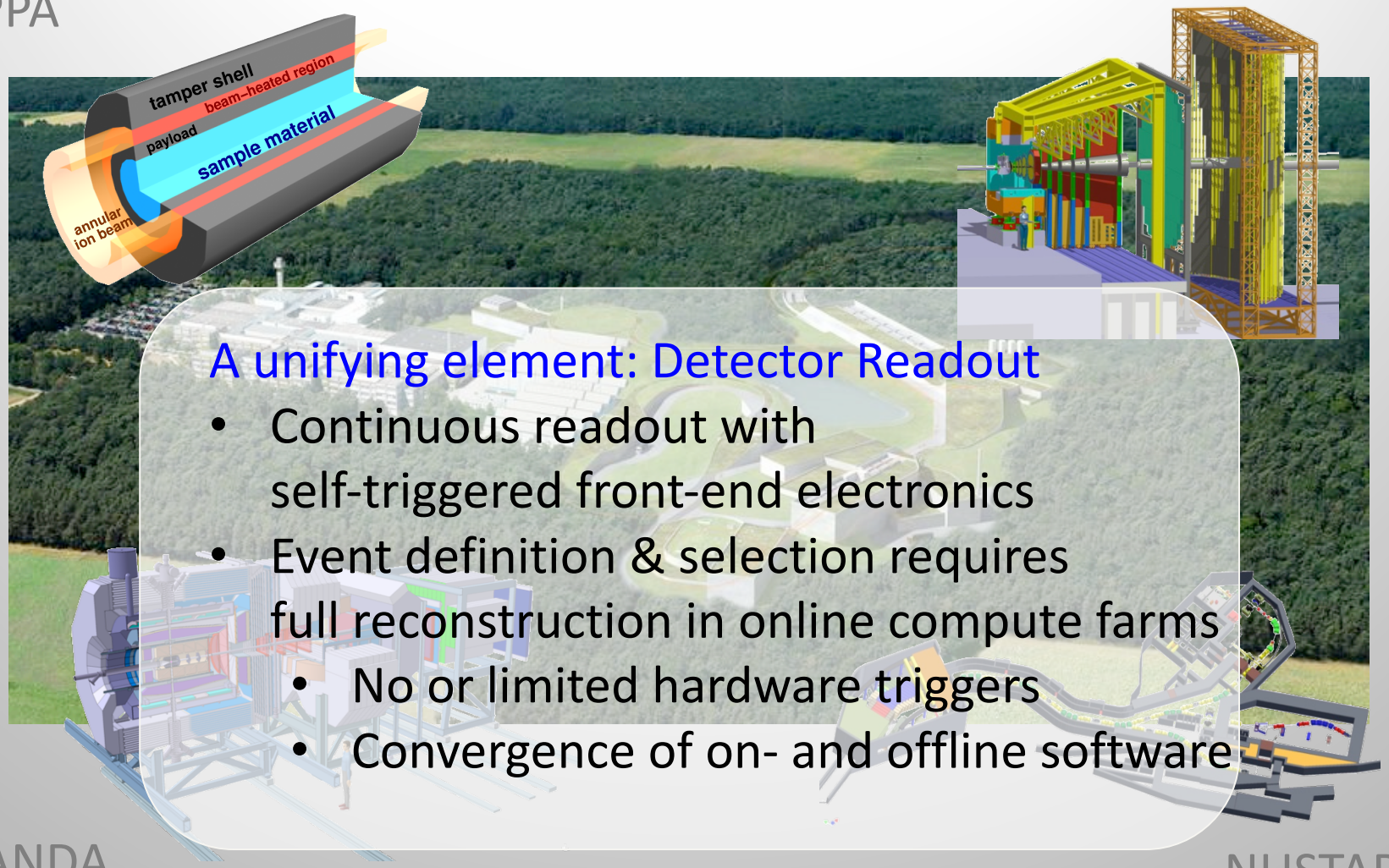
PANDA

NUSTAR

FAIR – four research pillars

APPA

CBM



A unifying element: Detector Readout

- Continuous readout with self-triggered front-end electronics
- Event definition & selection requires full reconstruction in online compute farms
  - No or limited hardware triggers
  - Convergence of on- and offline software

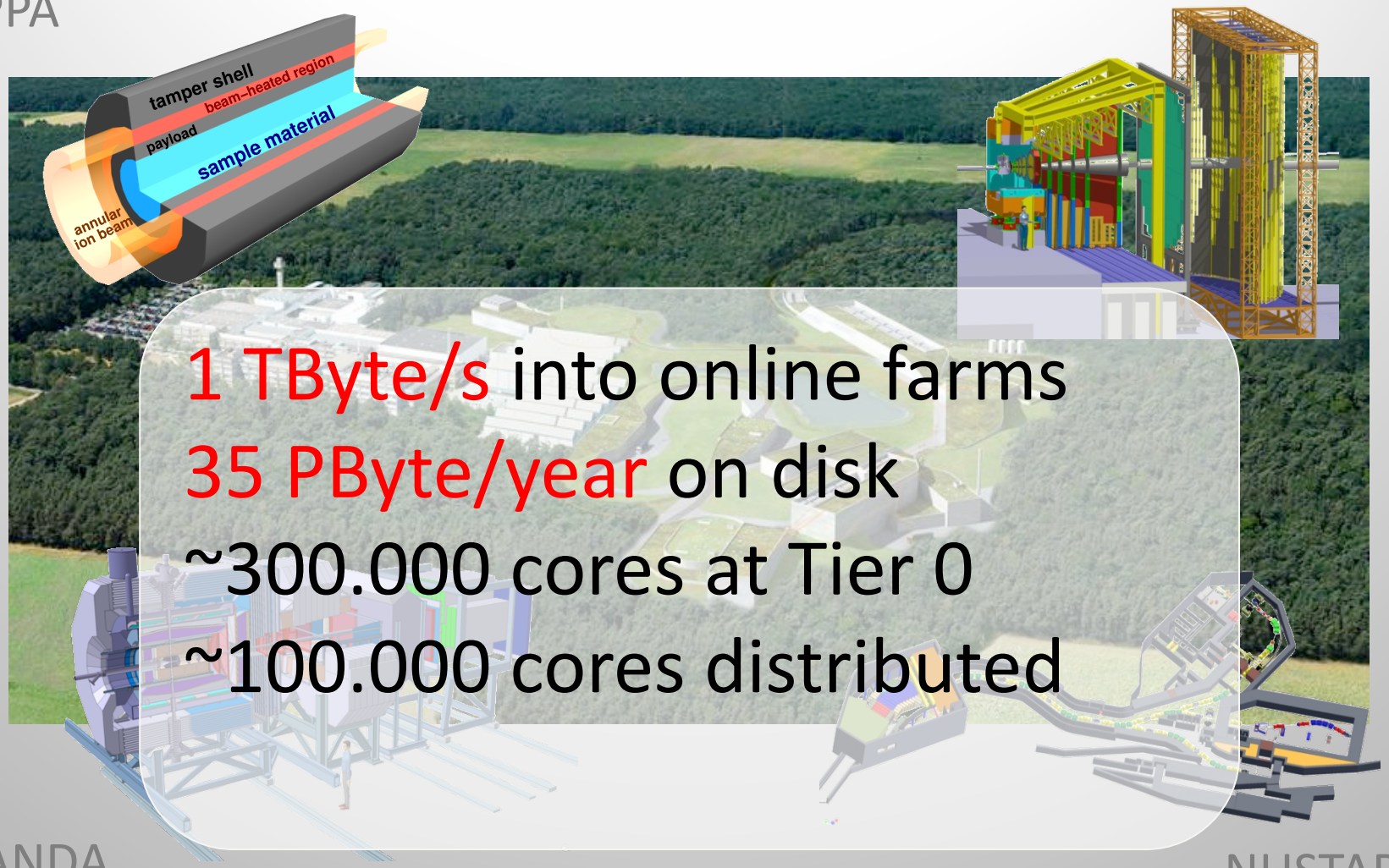
PANDA

NUSTAR

Computing at FAIR

APPA

CBM



1 TByte/s into online farms

35 PByte/year on disk

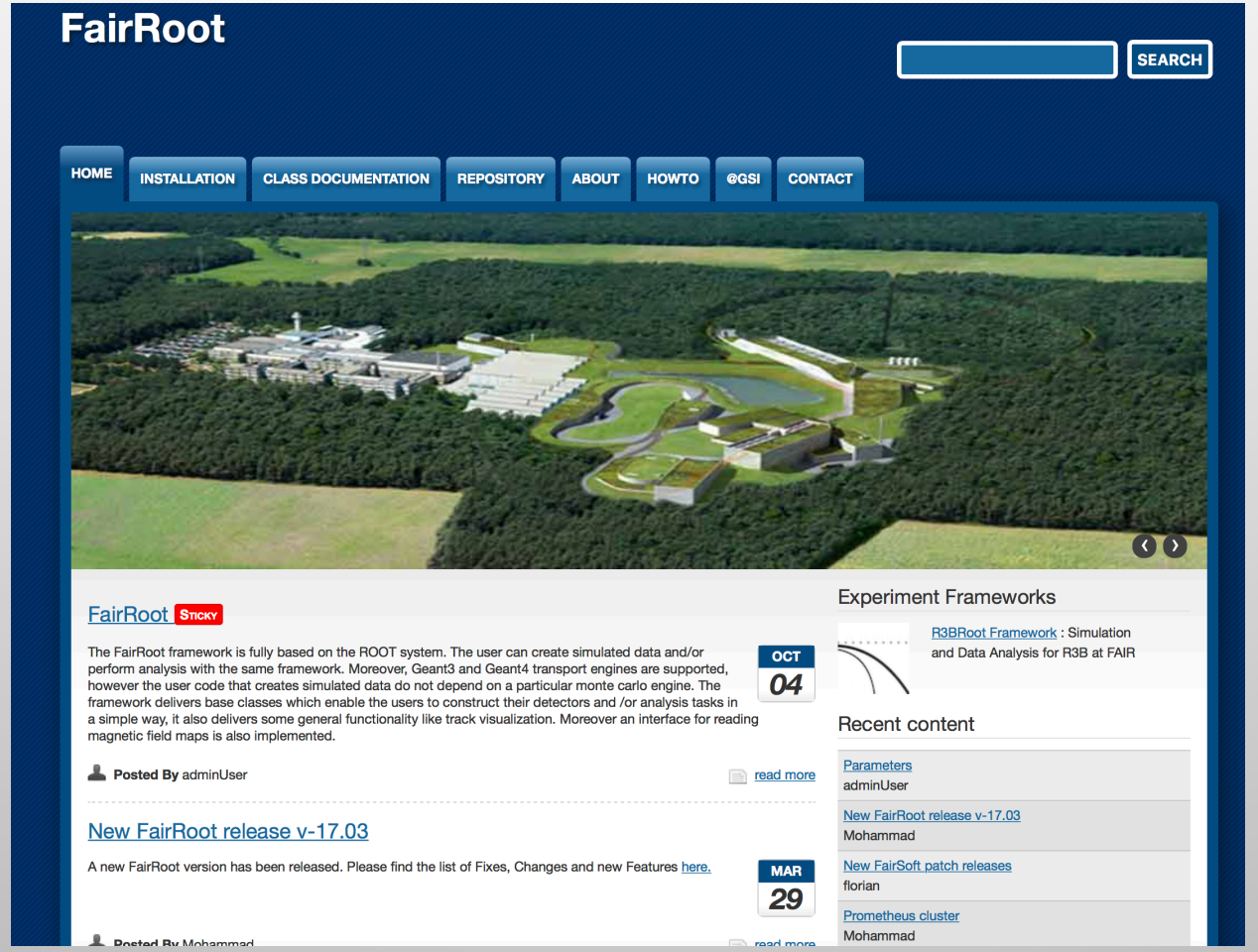
~300.000 cores at Tier 0

~100.000 cores distributed

PANDA

NUSTAR

FairRoot



fairroot.gsi.de

How it started?

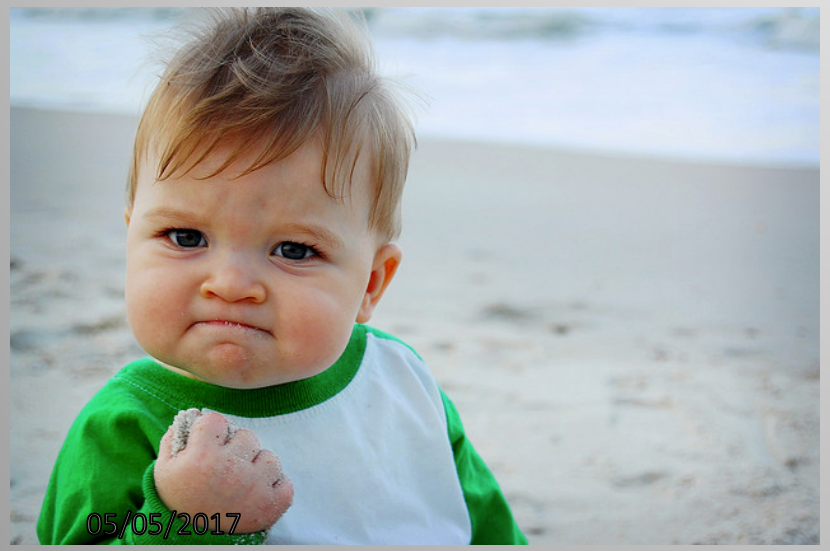
CBM collaboration 2003

- We need simulations for the LOI
- It has to be easy, fast, reliable, ..etc
- We have no manpower for software
- We need it yesterday

FLUKA
ROOT
PAW
VMC
Pythia
Geant4
AliRoot
Urqmd
Geant3
Hydra

Software for FAIR Experiments (FairRoot)

Agreement between GSI-IT management and the experiments to create a core team in the IT with participation of the experiment



After decision by Panda collaboration to use CbmRoot, the common part was called FairRoot

Software for FAIR Experiments (FairRoot)

FAIR and non-FAIR experiments join the effort to build one platform for simulation and reconstruction software

What about

Heterogeneous architectures

- Accelerator cards (GPUs, Xeon Phi, etc)

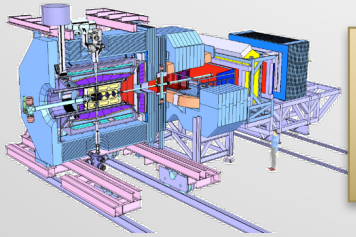
Concurrency?

- Multi-/Many-Core
- SIMD

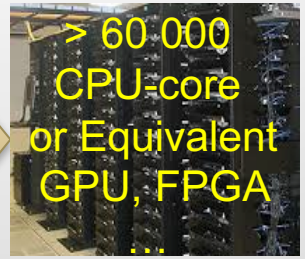
Online computing?

Handling 1 TByte/s data transport in the online systems

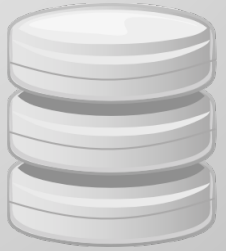
PANDA



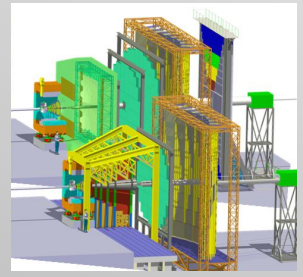
300 GB/s
20M Evt/s



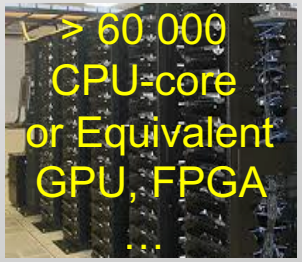
< 1 GB/s



CBM



1 TB/s



1 GB/s

ALICE LS2 Upgrade - Strategy

About 3 TByte/s detector readout

Storage bandwidth 90 GByte/s

Many physics probes have low S/B:
classical trigger/event filter approach not efficient

Store only reconstruction results, discard raw data

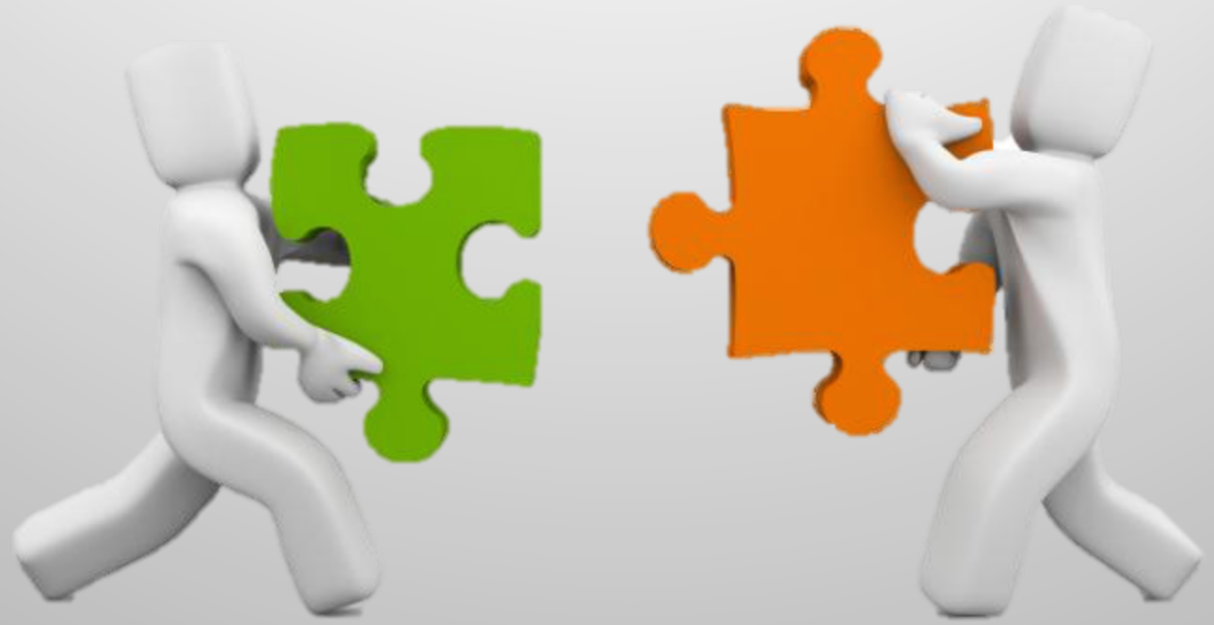
Data reduction by (partial) online reconstruction and compression

>100.000 cores + GPUs + FPGAs

Implies much tighter coupling between online and offline reconstruction software

Two projects – same requirements

- Massive data volume reduction
- Data reduction by (partial) online reconstruction
- Online reconstruction and event selection



ALICE – FAIR Framework: ALFA

- Developed in common by FairRoot Group (GSI), FAIR experiments and ALICE
- Has data-flow based model (Message Queues based multi-processing)
- Provides configuration, process management and monitoring tools
- Provides unified access to configuration parameters and databases

Scalability through multi-processing with message queues?

Each process assumes limited communication and reliance on other processes.

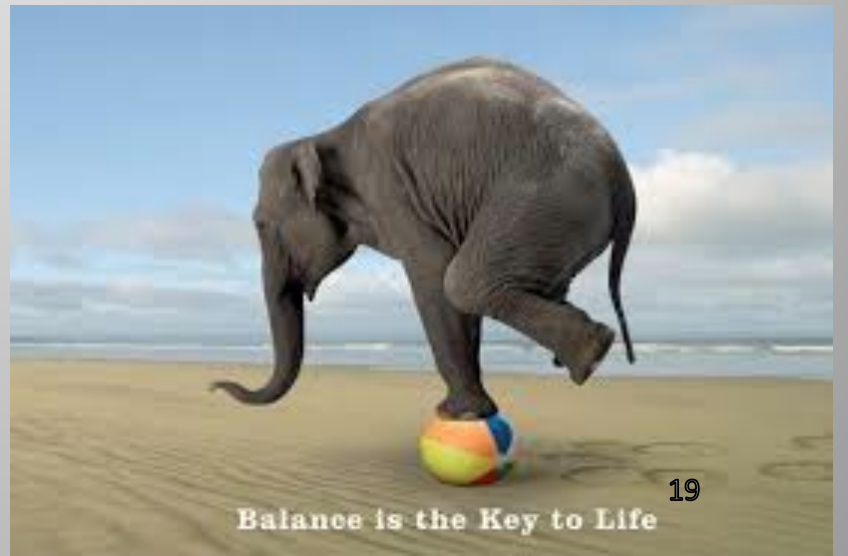
- No locking, each process runs with full speed
- Easier to scale horizontally to meet computing and throughput demands (starting new instances) than applications that exclusively rely on multiple threads which can only scale vertically.

Correct balance between reliability and performance

Each "Task" is a separate process, which:

- Can be multithreaded, SIMDized, …etc.
- runs on different hardware (CPU, GPU, …, etc.)
- Be written in an any supported language (Bindings for 30+ languages)

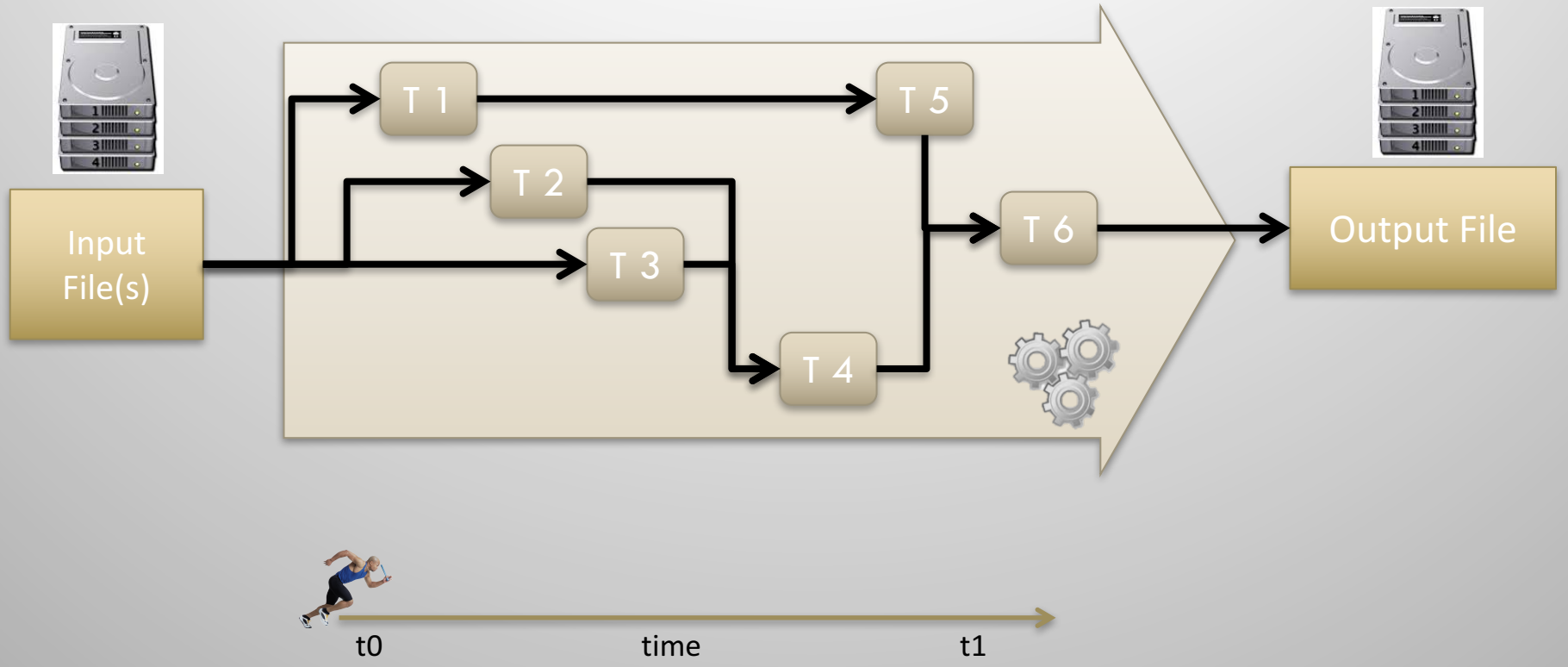
Different topologies of tasks can be adapted to the problem itself, and the hardware capabilities



ALFA uses FairMQ to connect different pieces together

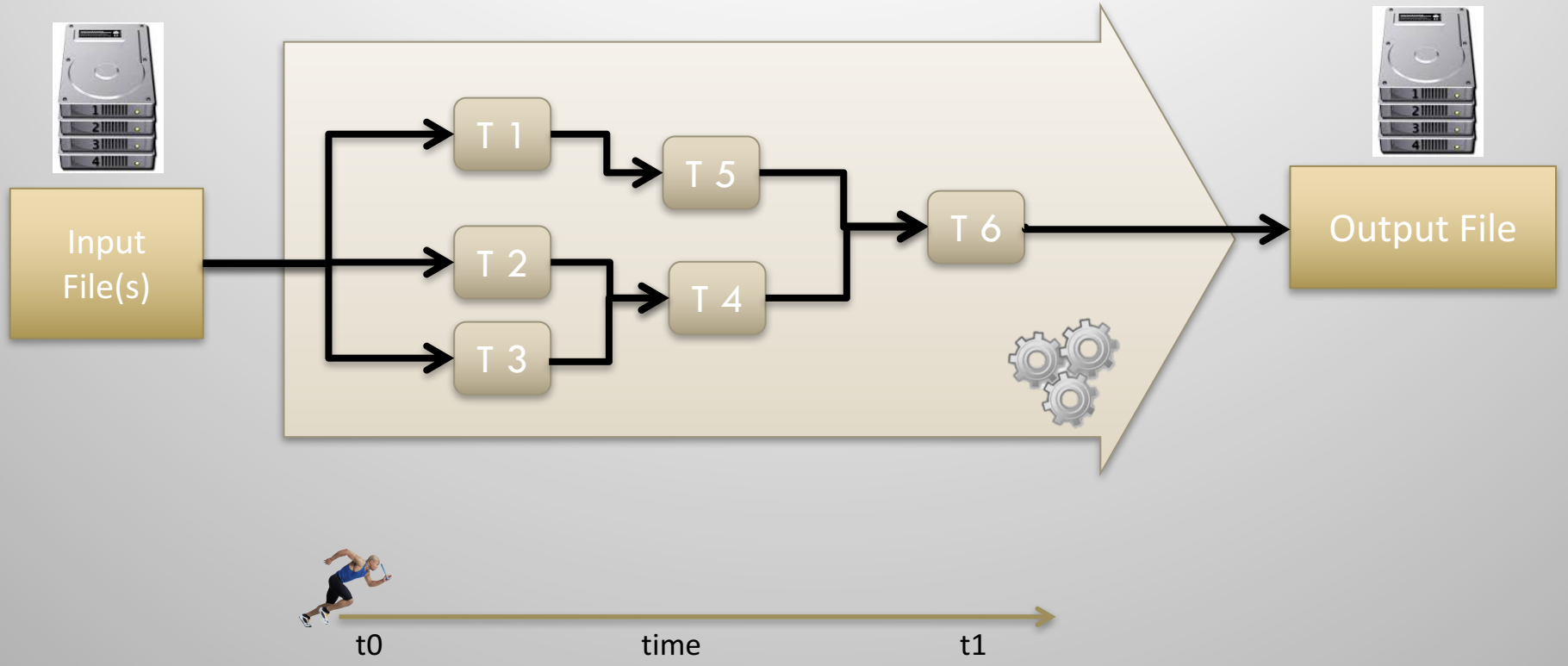
FairRoot: Where we are now?

- Task hierarchy (User code) runs sequentially in one process
- Tasks implement only algorithms (can be exchanged/replaced)



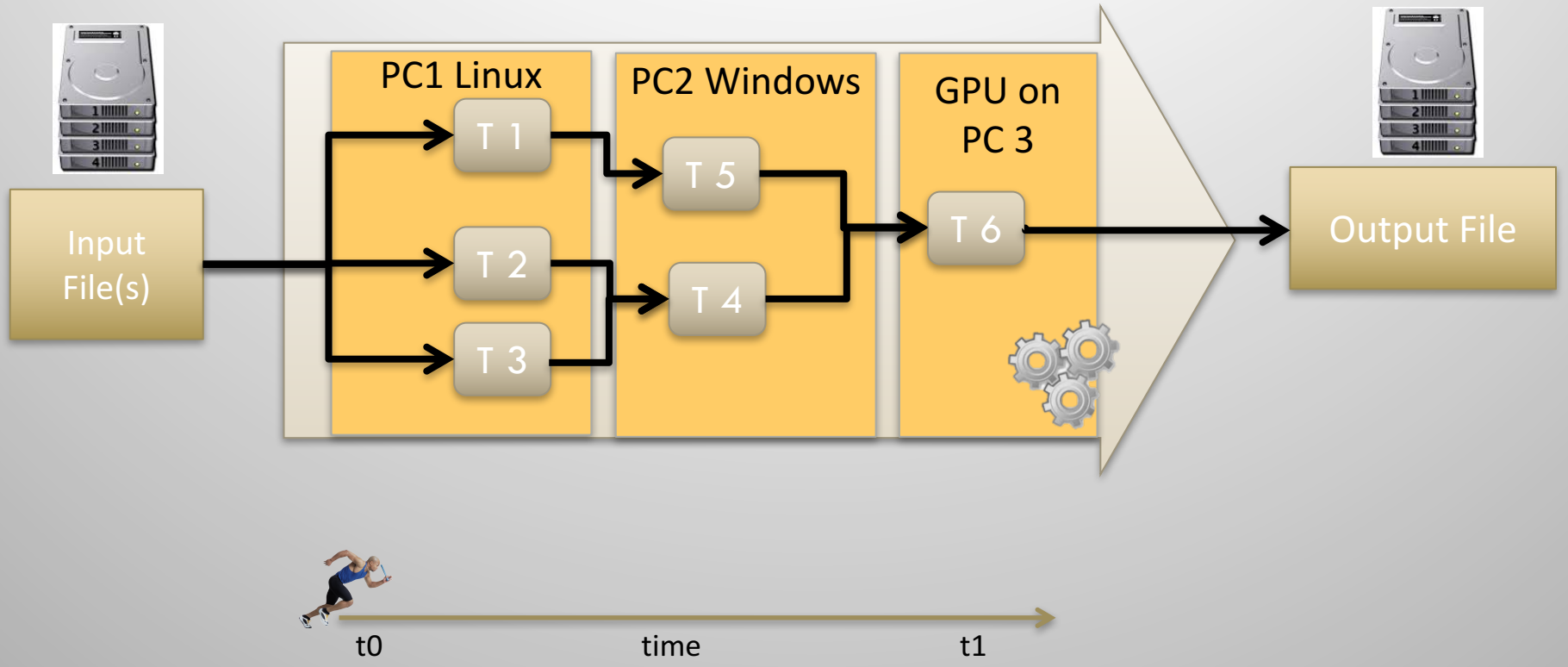
With FairMQ (ALFA)

- Each Task is a process (can be Multi-threaded)
- Message Queues for data exchange
- Support multi-core and multi node



With FairMQ (ALFA)

- Each Task is a process (can be Multi-threaded)
- Message Queues for data exchange
- Support multi-core and multi node



Message format ?

The framework does not impose any format on messages.

It supports different serialization standards

- BOOST C++ serialization
- Google's protocol buffers
- ROOT
- Flatbuffers
- MessagePack
- User defined

How to deploy ALFA on a laptop, few PCs or a cluster?

DDS: Dynamic Deployment System

Users describe desired tasks and their dependencies using topology (graph) files

Users are provided with a WEB GUI to create topology (Can be created manually as well).

The system takes so called “topology file” as the input.

DDS

One of the key challenges of the FairMQ approach:
Process Management for 10.000 to 100.000 devices

Control

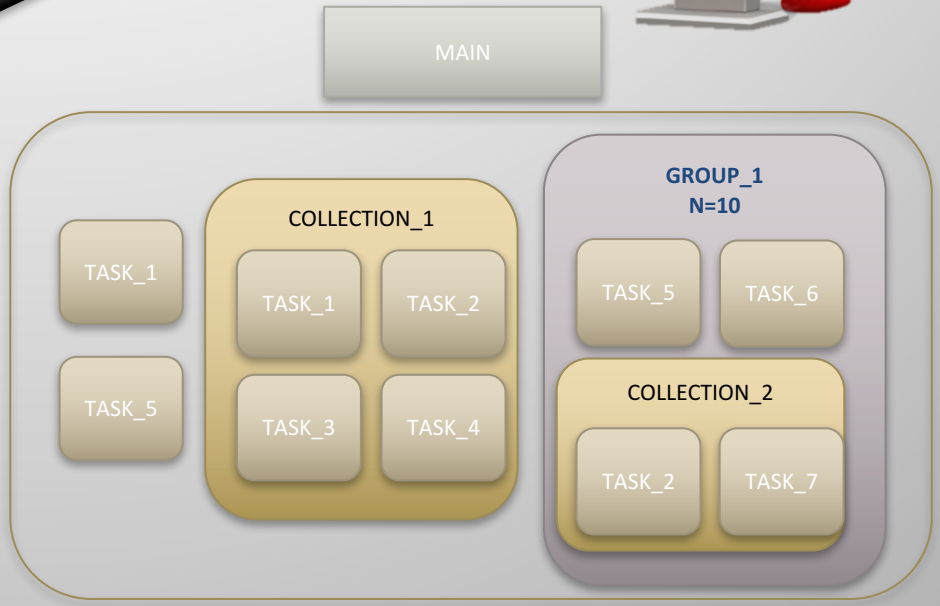
Monitoring

Configuring

Dynamic Deployment System

- Separate module in FairRoot / ALFA
- Xml description of process topology

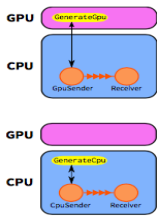
http://dds.gsi.de/

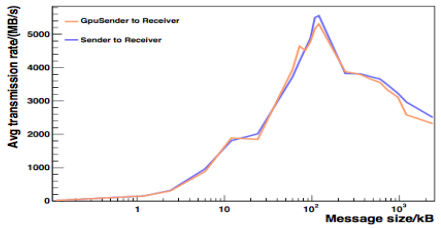


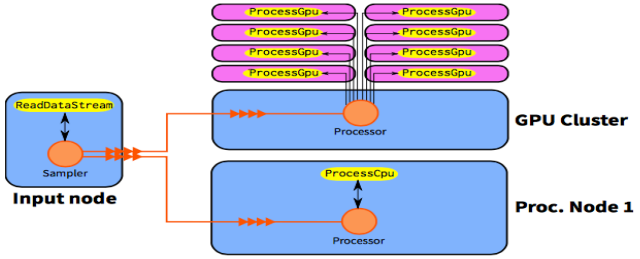
GPUs in ALFA

GPUs and Message Queues

- Explore communication/data transfer to GPUs
- FairMQ: implementation of Message Queues in the FairRoot framework (☞Apr 14: M. Al-Turany, A. Rybalchenko, F. Uhlig)
- Test system with implementation of Circle Hough algorithm
  - Modular structure
  - CPU and GPU version of processing task
  - FairMQ: stream input data to CPU/GPU processing tasks
  - Maximum flexibility of architecture and data transfer interface







L. Bianchi | Online Tracking with GPUs at PANDA | CHEP2015

13 / 14

Mitglied der Helmholtz-Gemeinschaft

Ludovico BIANCHI

http://indico.cern.ch/event/304944/session/1/contribution/363

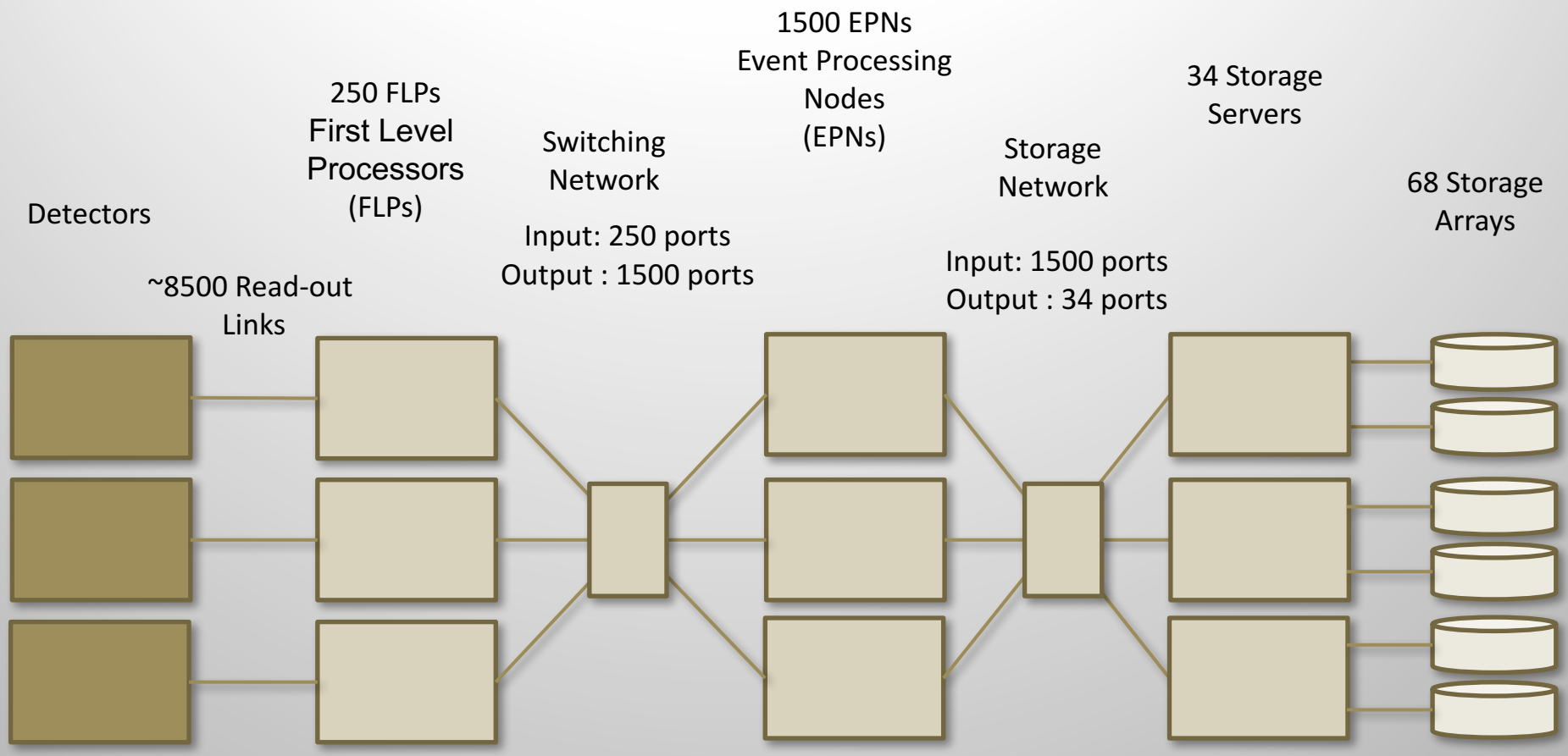
ALICE

FAIR

Is the data processing strategy feasible?

Can we create a small scale but yet realistic processing topology ?

O2 Facility


Detectors
~8500 Read-out Links
250 FLPs
First Level Processors
(FLPs)
Switching Network
Input: 250 ports
Output : 1500 ports
1500 EPNs
Event Processing Nodes
(EPNs)
Storage Network
Input: 1500 ports
Output : 34 ports
34 Storage Servers
68 Storage Arrays
3.1 TB/s
500 GB/s
90 GB/s

The prototype:

In ALICE 92.5% of the data is generated by the TPC

focus on TPC processing

The data from the TPC front-end will arrive via multiple links in the FLP nodes

use present readout layout with 216 links

Local cluster reconstruction is running on hardware accelerator cards in real-time on the input streams

Prototype start with clusters (space points) in the main memory of FLP nodes

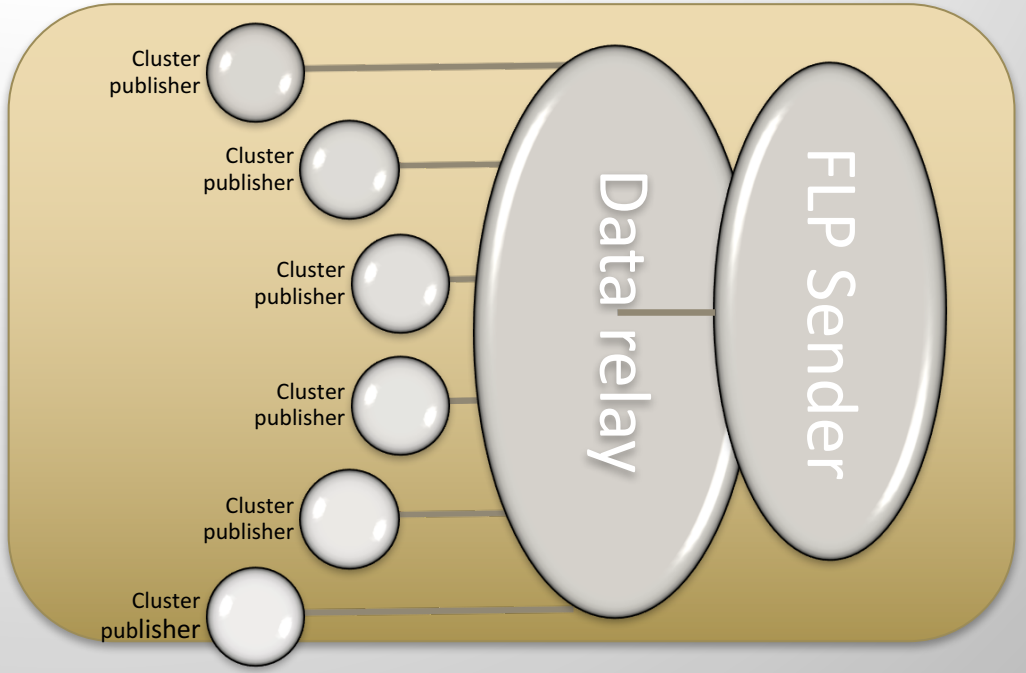
PROTOTYPE

Matthias Richter

FLP devices

- 36 Data sources
  - 36 x 6 cluster publisher
  - 36 Merger (Data relay)
  - 36 FLP Sender

216+36+36 = 288 processes



use present readout layout with 216 links

FLP: First level data processer

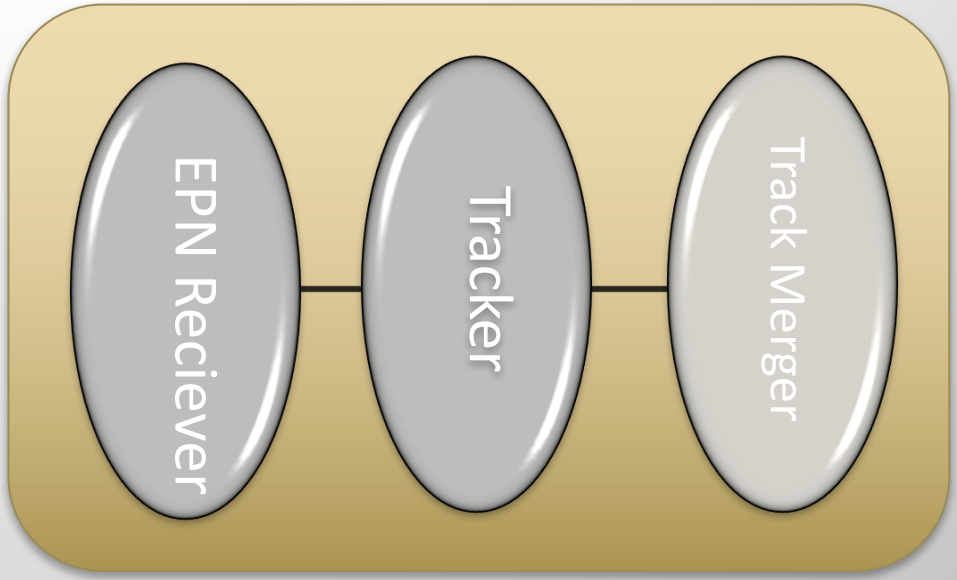
PROTOTYPE

Matthias Richter

EPN devices

- 28 Data consumers
  - 28 recievers
  - 28 Trackers (GPU)
  - 28 Track mergers

28+28+28 = 84 processes

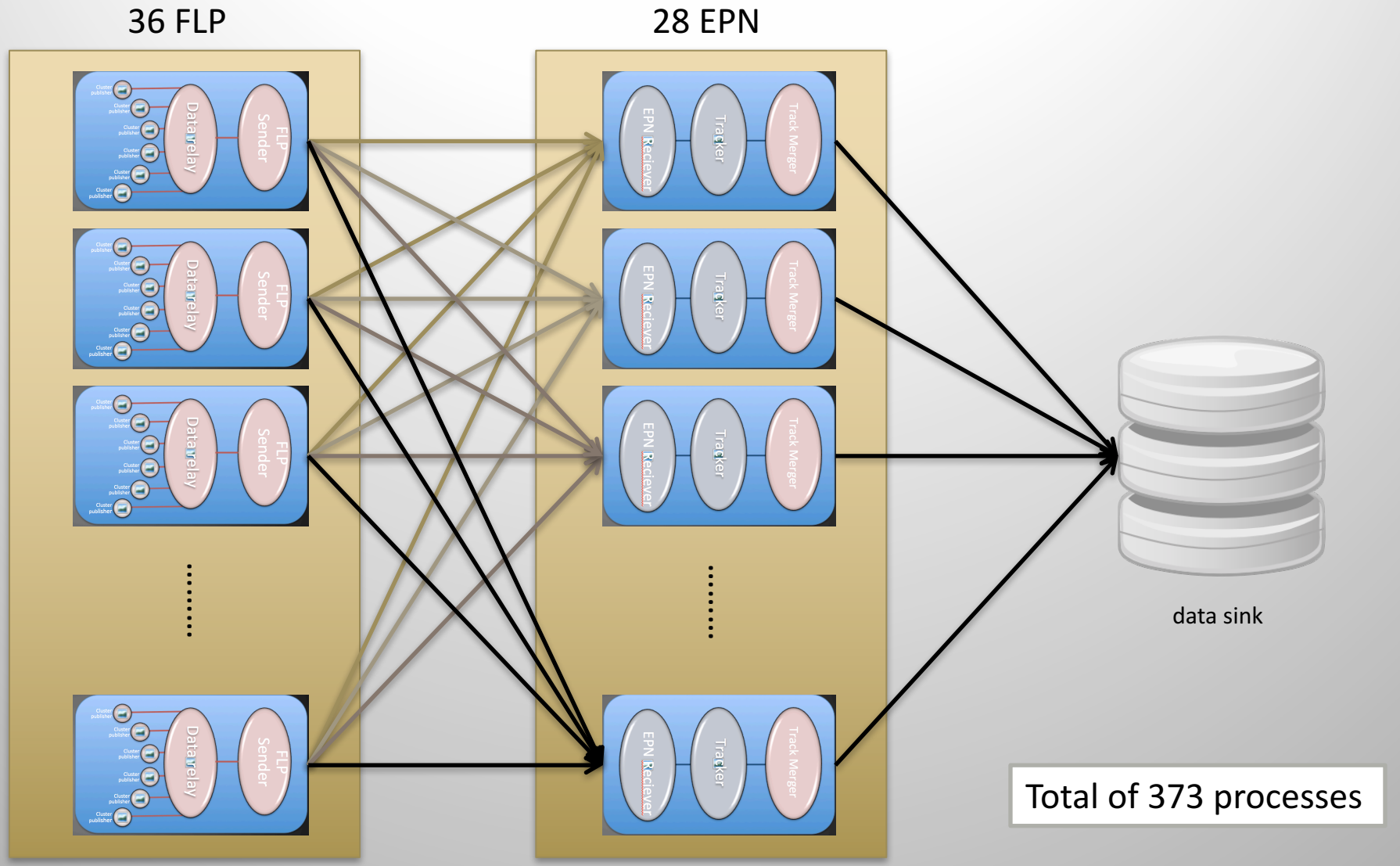


EPN: Event Processing Node

Implementing the TPC reconstruction

PROTOTYPE

Matthias Richter



Total of 373 processes

FAIR

Hardware

- Small scale test environment (40 nodes) using parts of existing ALICE HLT development cluster :
  - 16 core Intel Xeon 2.26 GHz
  - 24 core AMD Opteron 2.1 GHz
  - GPU used as accelerator card for particle track finding
- Network protocol IP over InfiniBand

Matthias Richter

Results

- The topology is processing aggregated size of 1.6 GByte/s (limited by the cluster publishers)
- FLP to EPN data transportation prove to fulfill the requirement
- Efficient process scheduling and deployment system tested with the prototype
- System is ready for larger test

How to switch form root single-core processing (FairRoot Tasks) to FairMQ multi-core pipeline processing

FairRoot/Examples/MQ/9-PixelDetector

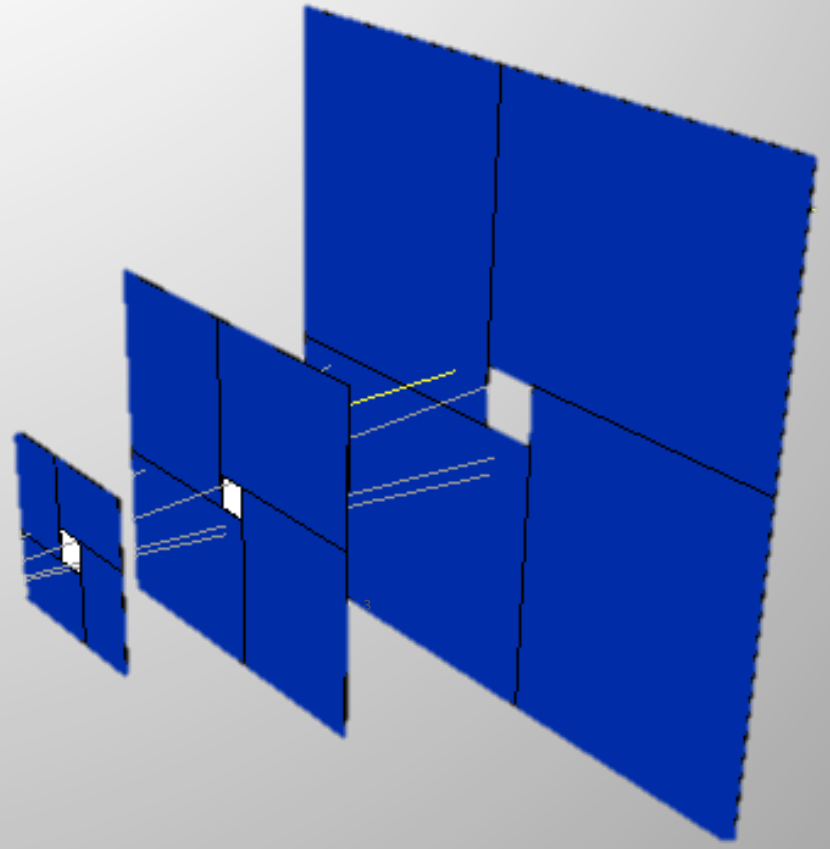
Radoslaw Karabowicz, GSI

- Detector simulation,
- Digitization,
- reconstruction (hit finding, tracking, track fitting),
- Shows how to switch from root single-core processing to FairMQ multi-core pipeline processing.

https://github.com/FairRootGroup/FairRoot/tree/master/examples/MQ/9-PixelDetector

fairroot/examples/MQ/9-PixelDetector

- 3 stations with 4 rectangular sensor each:
  - size: 5x 5cm^2, inner hole: 1x1cm^2, at z = 5cm;
  - size: 10x10cm^2, inner hole: 1x1cm^2, at z = 10cm;
  - size: 20x20cm^2, inner hole: 2x2cm^2, at z = 20cm;
- each sensor divided into pixels (0.01x0.01cm^2), that are grouped into FE modules (110 pixels x 116 pixels)

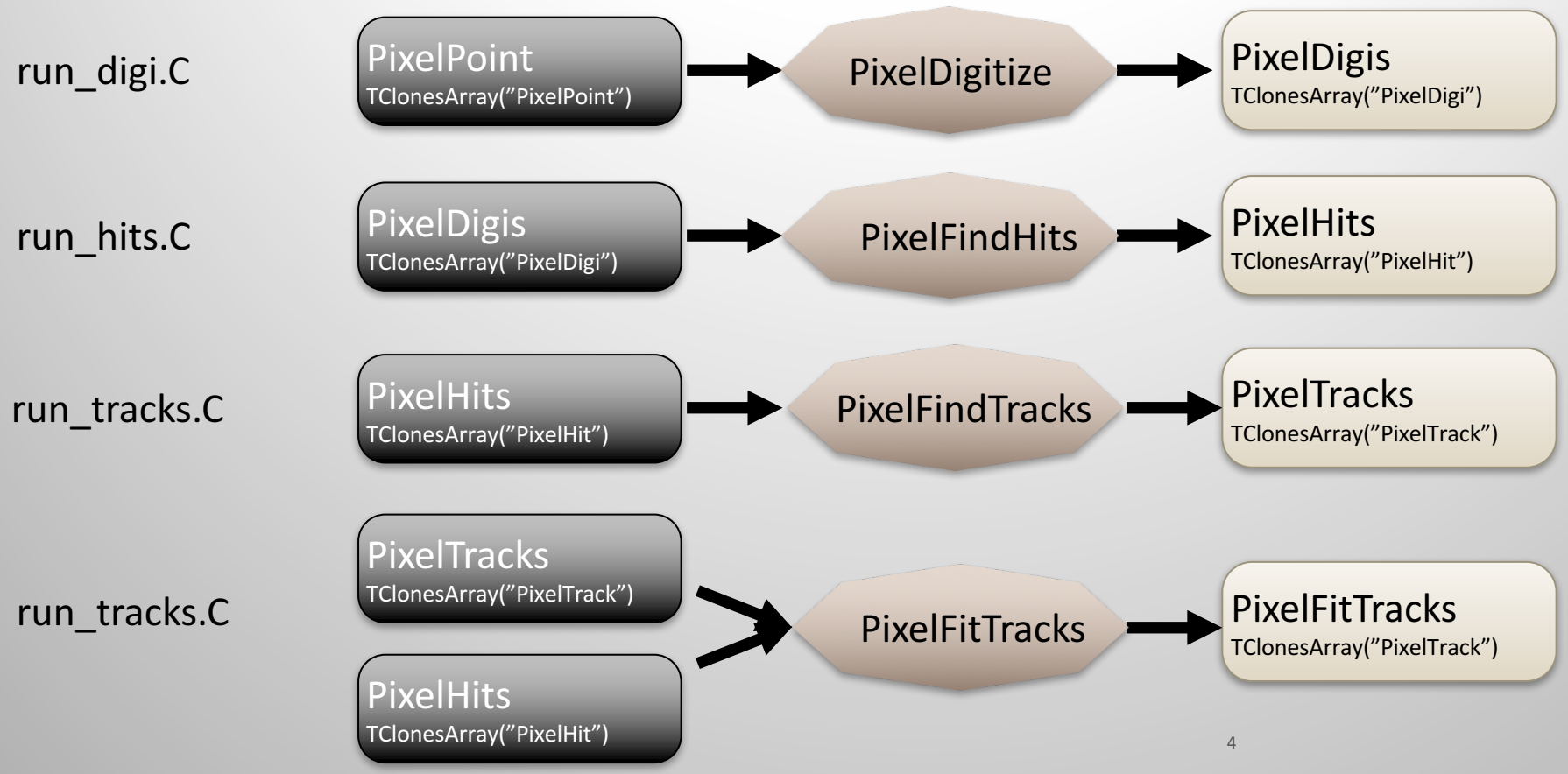


FEs numbering on one sensor

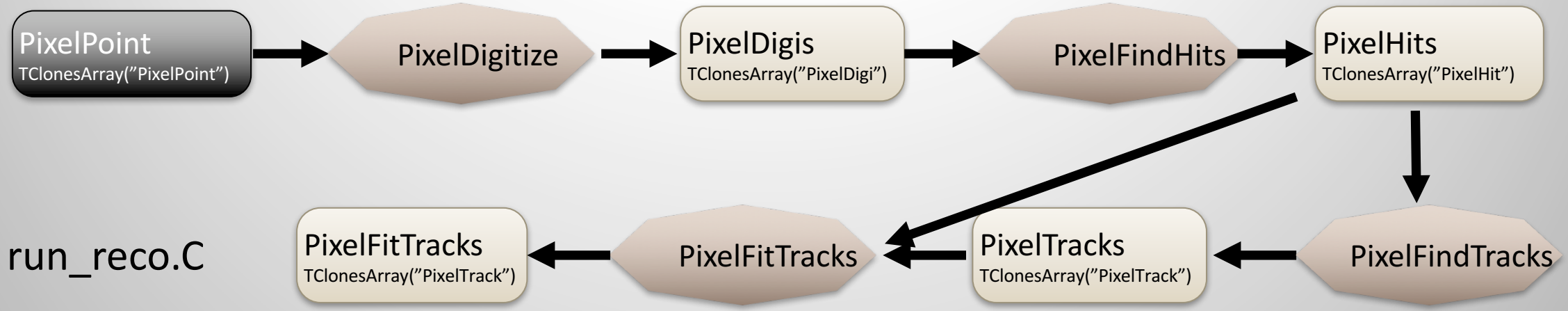
…					
FE 5	…				
FE 4	FE 68	…			
FE 3	FE 67	FE 131	…		
FE 2	FE 66	FE 130	FE 194	…	
FE 1	FE 65	FE 129	FE 193	FE 257	…

Radoslaw Karabowicz, GSI

data classes, tasks and macros

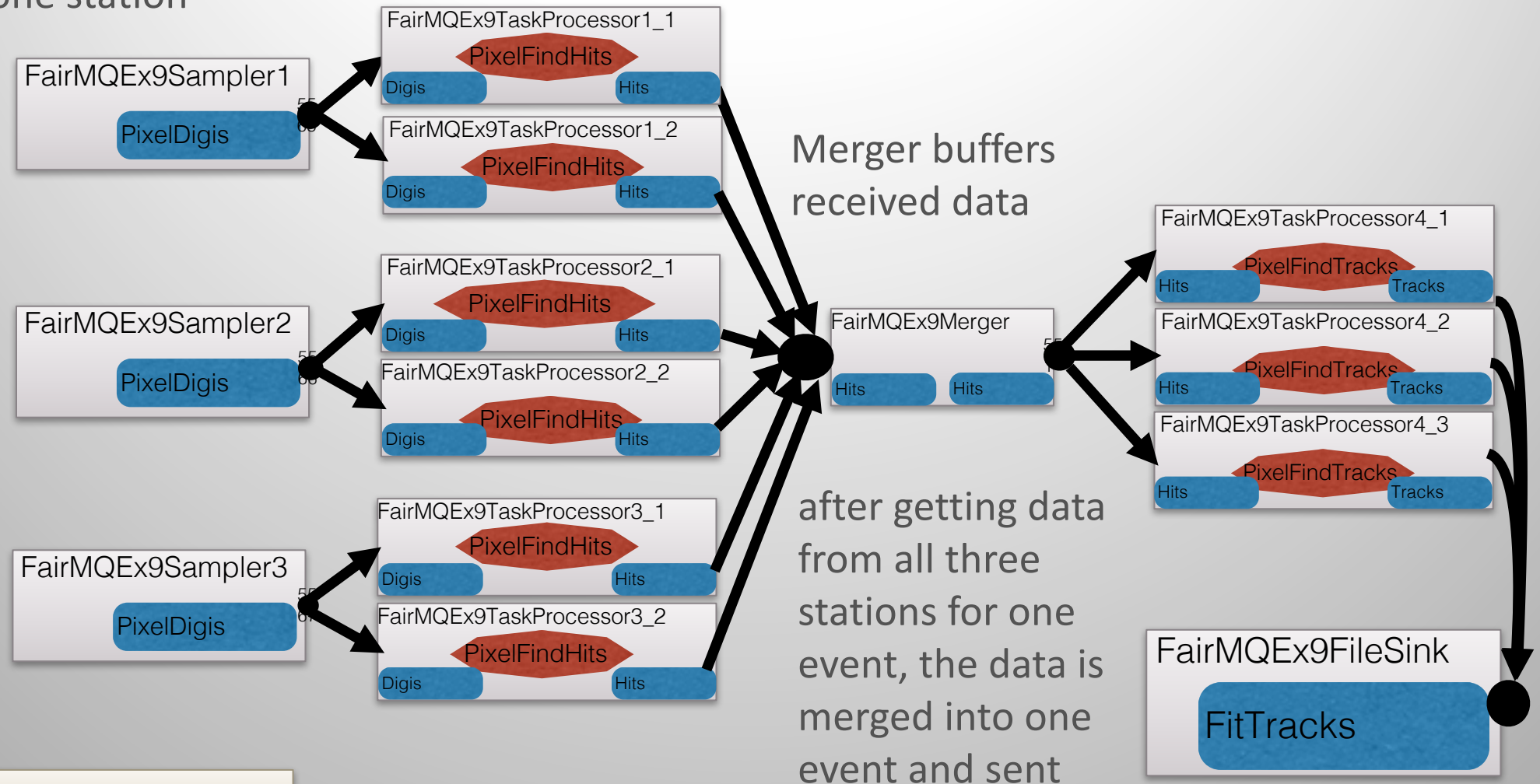


data classes, tasks and macros



Example topology

Each sampler reads from file with data from one station



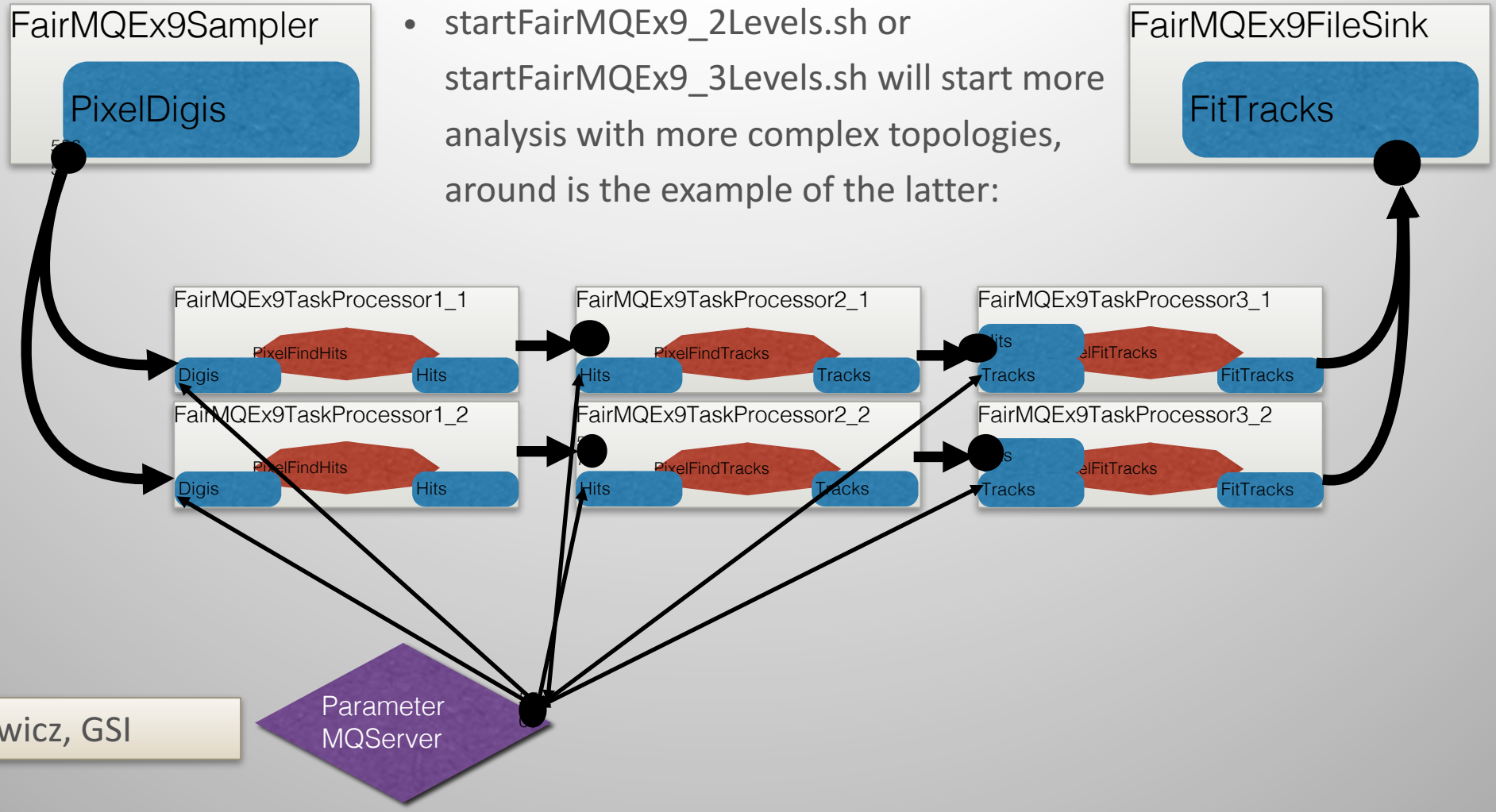
Merger buffers received data

after getting data from all three stations for one event, the data is merged into one event and sent

Radoslaw Karabowicz, GSI

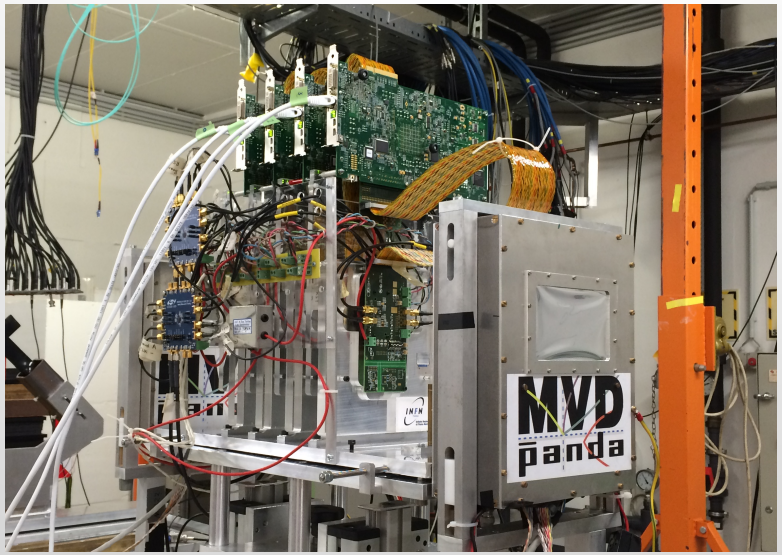
Other topologies:

- startFairMQEx9_2Levels.sh or startFairMQEx9_3Levels.sh will start more analysis with more complex topologies, around is the example of the latter:



Radoslaw Karabowicz, GSI

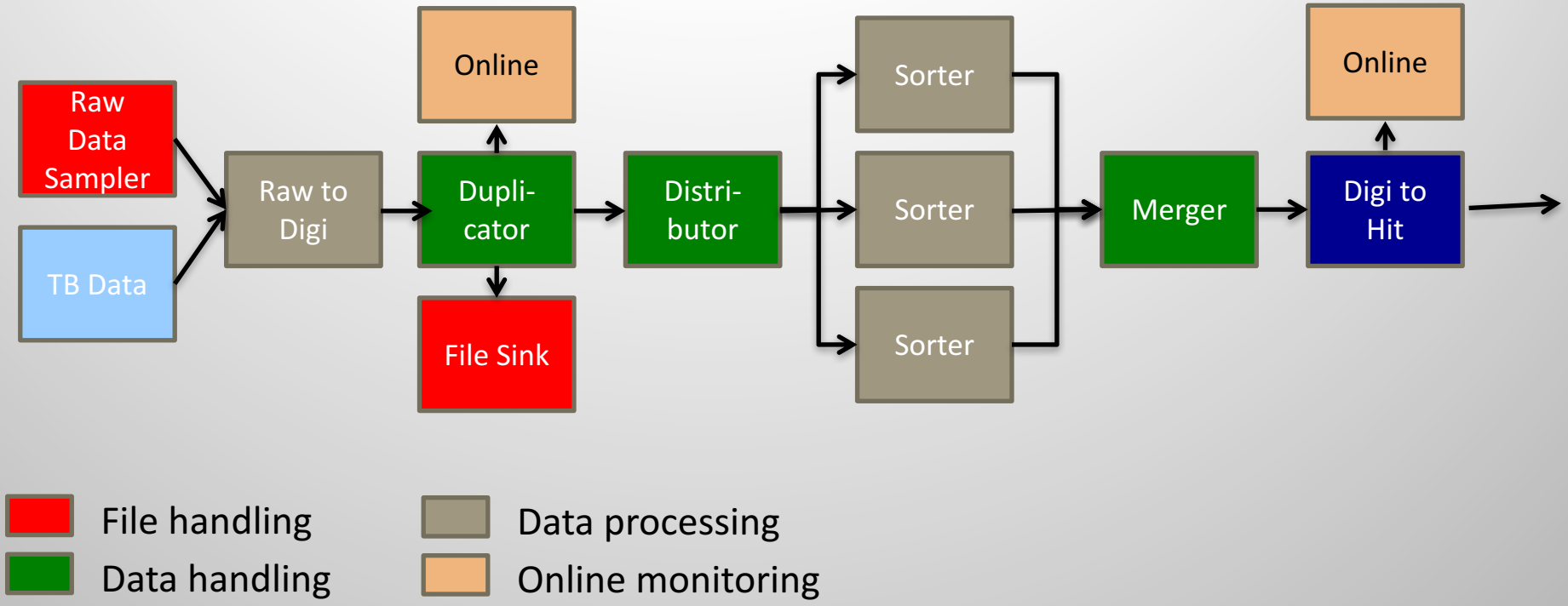
Test of FairMQ with Real Data



- Parallel readout of 4 pixel detectors
- Readout done by 4 FPGA boards sending their data to two PCs
- On the PC bitstream converted into raw data
- Raw data send via FairMQ to a FileSink

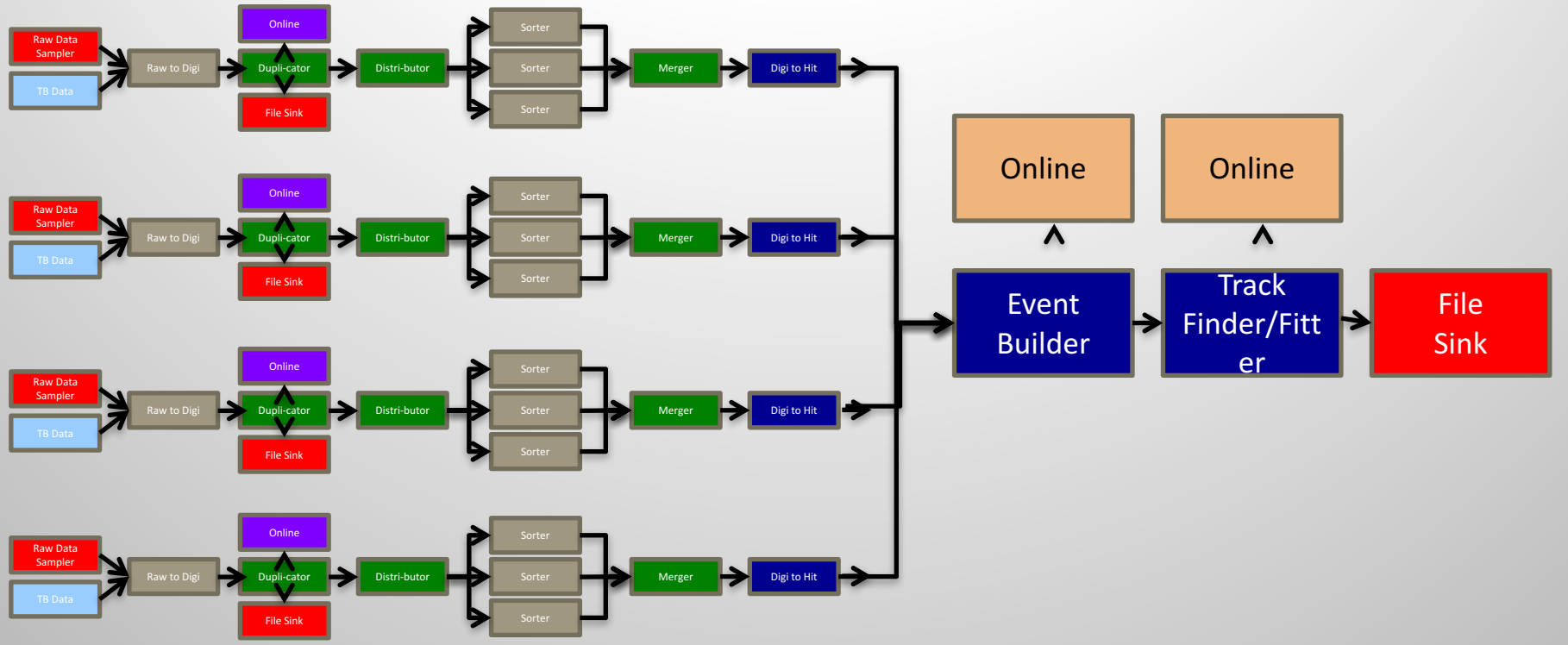
Tobias Stockmanns
https://indico.cern.ch/event/505613/contributions/2227258/

Single Front-End



Tobias Stockmanns
https://indico.cern.ch/event/505613/contributions/2227258/

Multiple Front-Ends

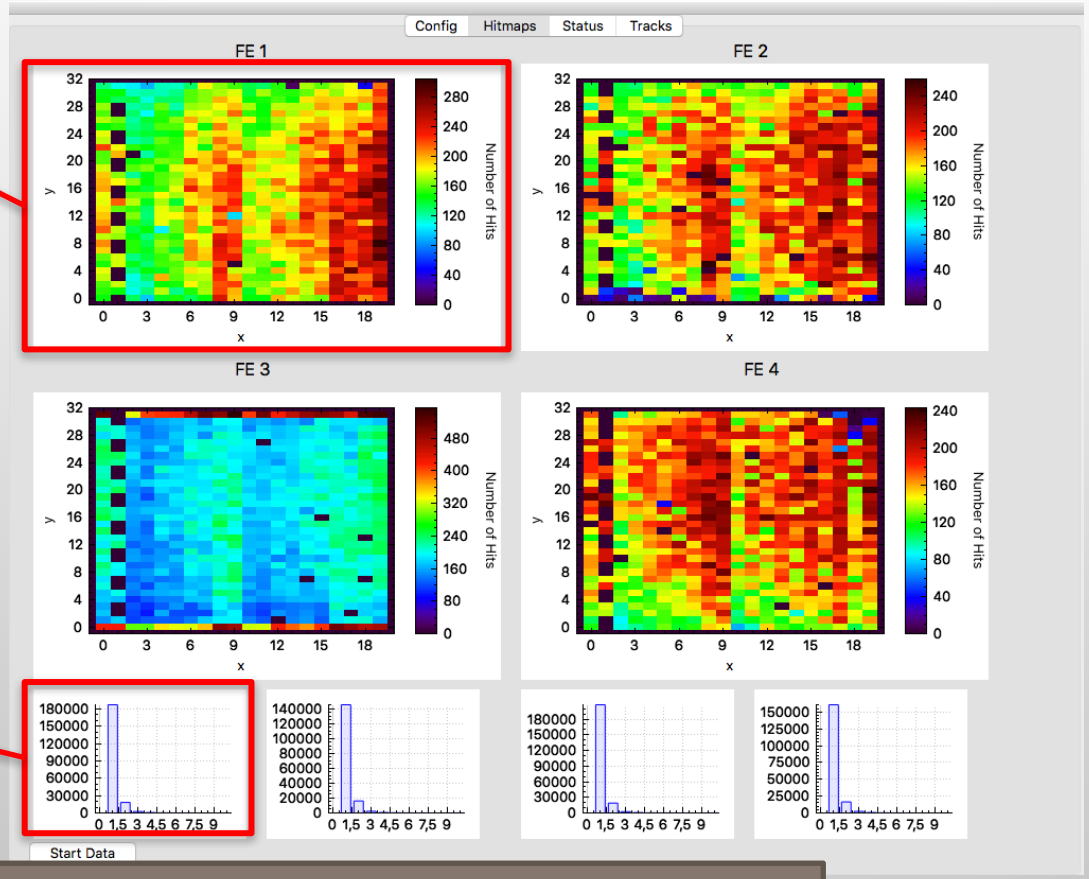


Tobias Stockmanns
https://indico.cern.ch/event/505613/contributions/2227258/

Online Monitoring

Hitmap

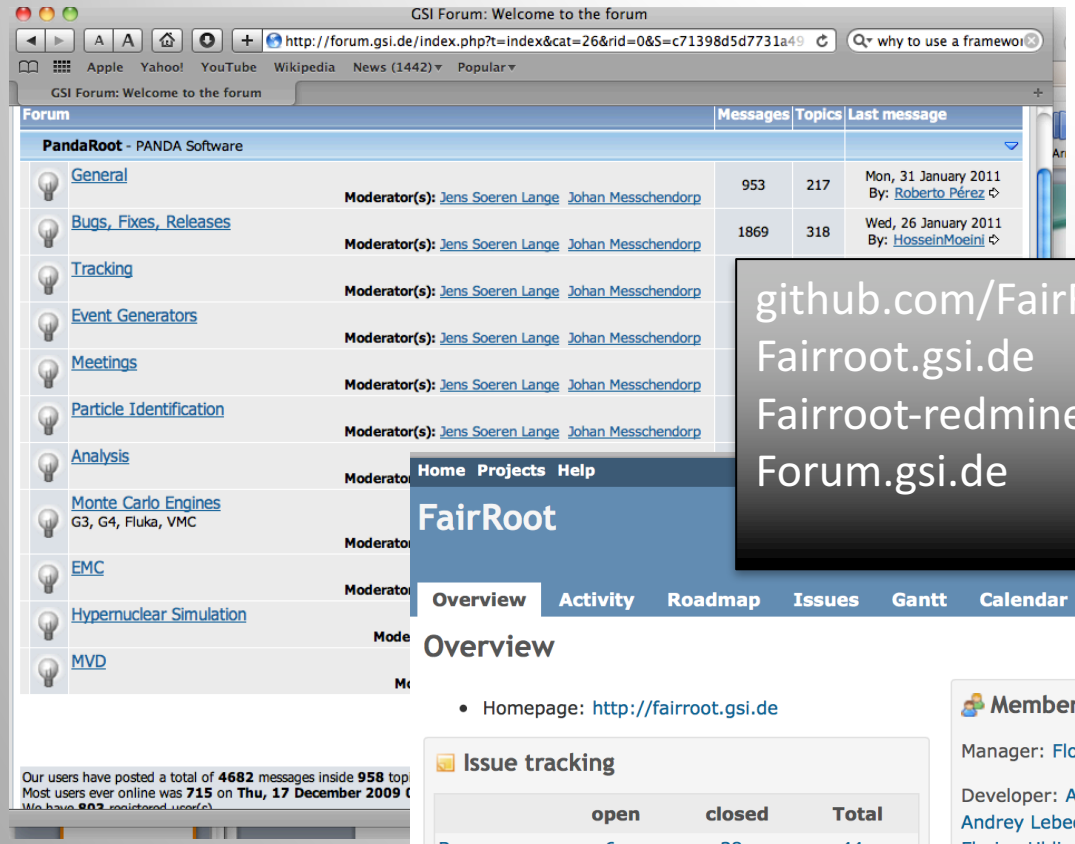
Cluster finder

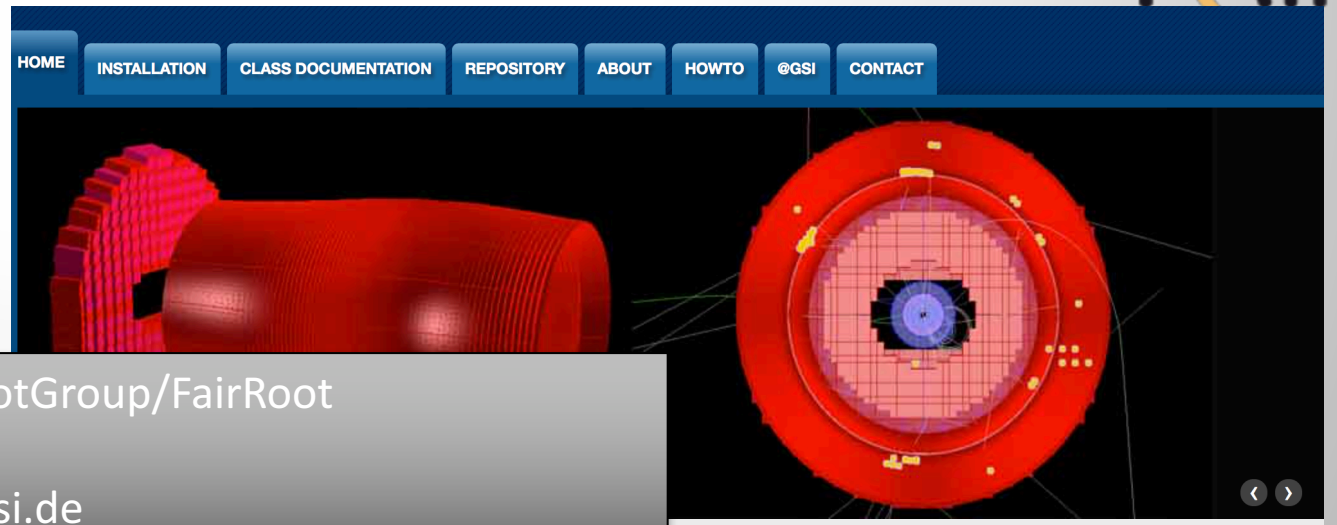


- Running on 6 PC
  - 4 for each FE
  - 1 for tracking
  - 1 for control

Tobias Stockmanns
https://indico.cern.ch/event/505613/contributions/2227258/

More about FairRoot/ALFA





github.com/FairRootGroup/FairRoot
Fairroot.gsi.de
Fairroot-redmine.gsi.de
Forum.gsi.de

Home Projects Help

FairRoot

Overview Activity Roadmap Issues Gantt Calendar News Documents Wiki Files Repository

Overview

- Homepage: http://fairroot.gsi.de

Issue tracking

	open	closed	Total
Bug	6	38	44
Feature	19	20	39
Support	6	4	10

View all issues | Calendar | Gantt

Members

Manager: Florian Uhlig, Mohammad Al-Turany

Developer: Alexey Rybalchenko, Anar Manafov, Andrey Lebedev, Dennis Klein, Dmytro Kresan, Florian Uhlig, Mohammad Al-Turany, Nicolas Winckler, Radoslaw Karabowicz

Reporter: Alexey Rybalchenko, Anar Manafov, Andrey Lebedev, Dennis Klein, Dmytro Kresan, Erik Krebs, Fons Rademakers, Klaus Goetzen, Laurent Aphecetche, Marcin Chrzaszcz, Martin Kohn, Mohammad Al-Turany, Nicolas Winckler, Prometeusz Jasinski, Radoslaw Karabowicz, Redmine Admin, Stefano Spataro, Thomas Ruf, Volker Friese

Latest news

Spent time

88.30 hours

Details | Report

OCT 04

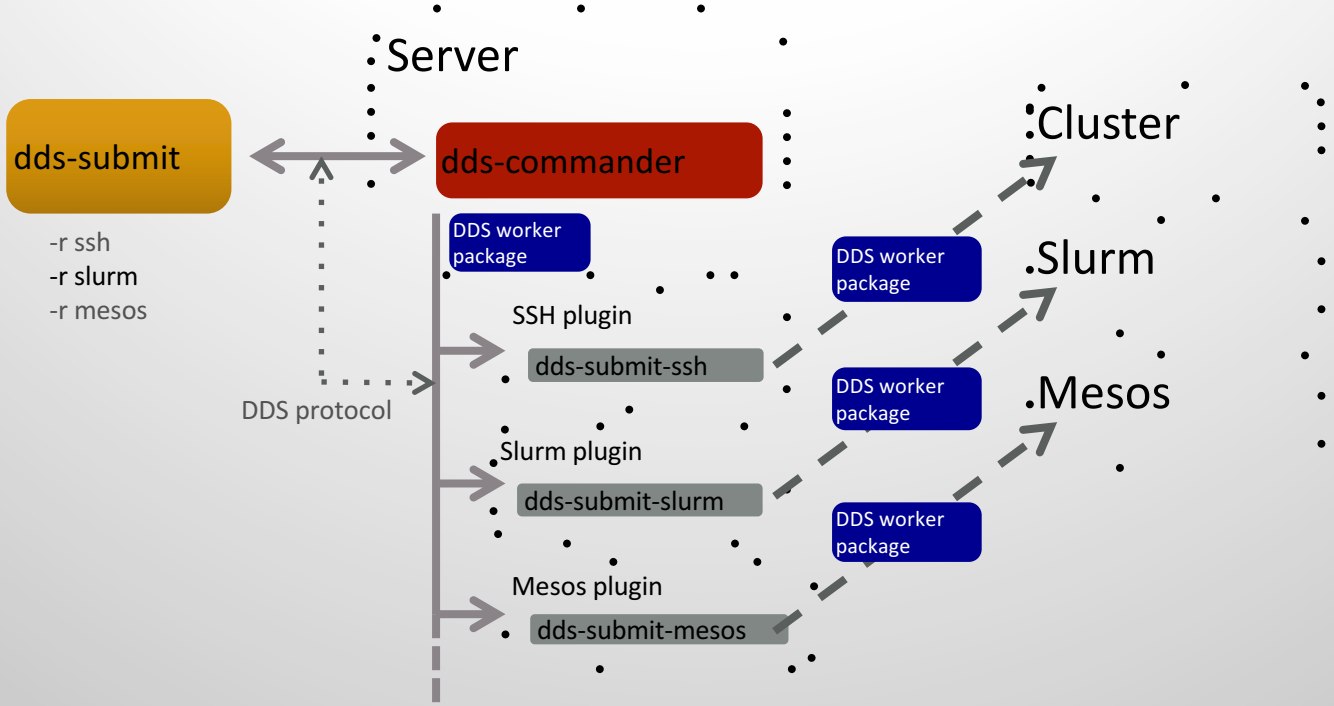
read more

Recent blog posts

- New stable version of R3BRoot is available "feb14"
- New tag release of FairRoot is available v-13.12
- New external packages (dec 13) available
- Service Interruption of the Wikis, Subversion and Forum
- New external packages (apr 13) available
- New structure of FairRoot repository
- New external packages (sep12) available
- PROOF available

Backup

DDS Plugins



1. dds-commander starts a plugin based on the dds-submit parameter,
2. plugin connects back to dds-commander,
3. plugin receives submission details,
4. plugins takes WN package and deploys it to WNs.