ALICE distributed data processing

14 March 2017

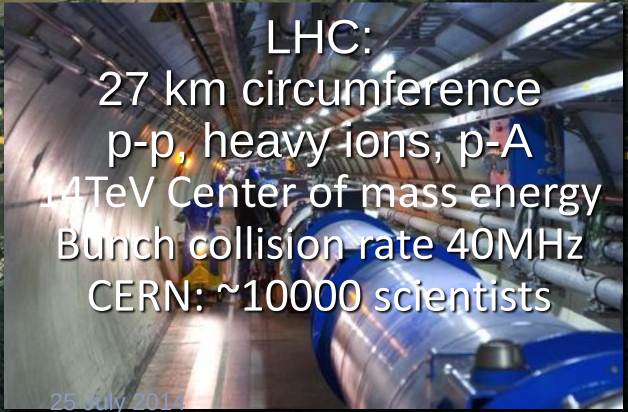
Latchezar Betev - ALICE

LHC and Experiments

p-p
B-Physics, CP Violation
(matter-antimatter symmetry)

General purpose,
p-p, heavy ions
New physics: Higgs boson,
SuperSymmetry

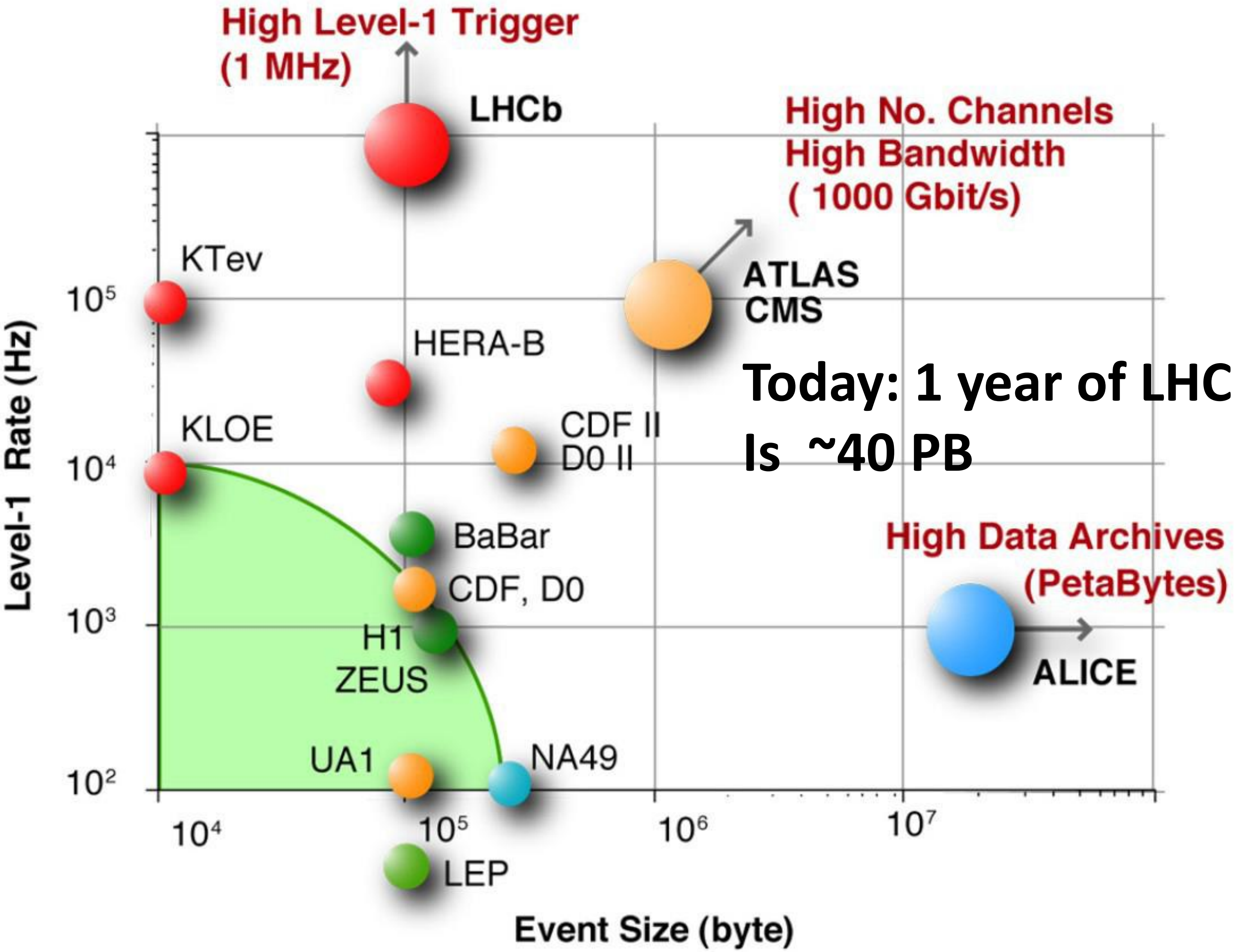
Exploration of a new energy frontier
in p-p and Pb-Pb collisions



LHC:
27 km circumference
p-p, heavy ions, p-A
14TeV Center of mass energy
Bunch collision rate 40MHz
CERN: ~10000 scientists

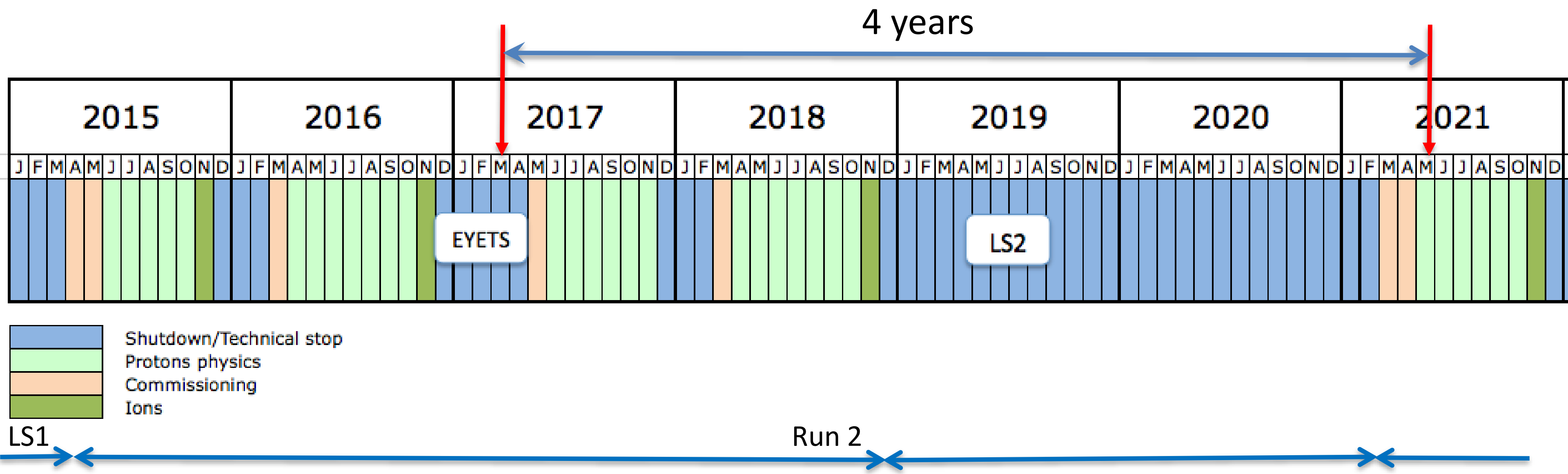
Heavy ions, pp
Quark-Gluon Plasma
(state of matter of early universe)

The data challenge in HEP



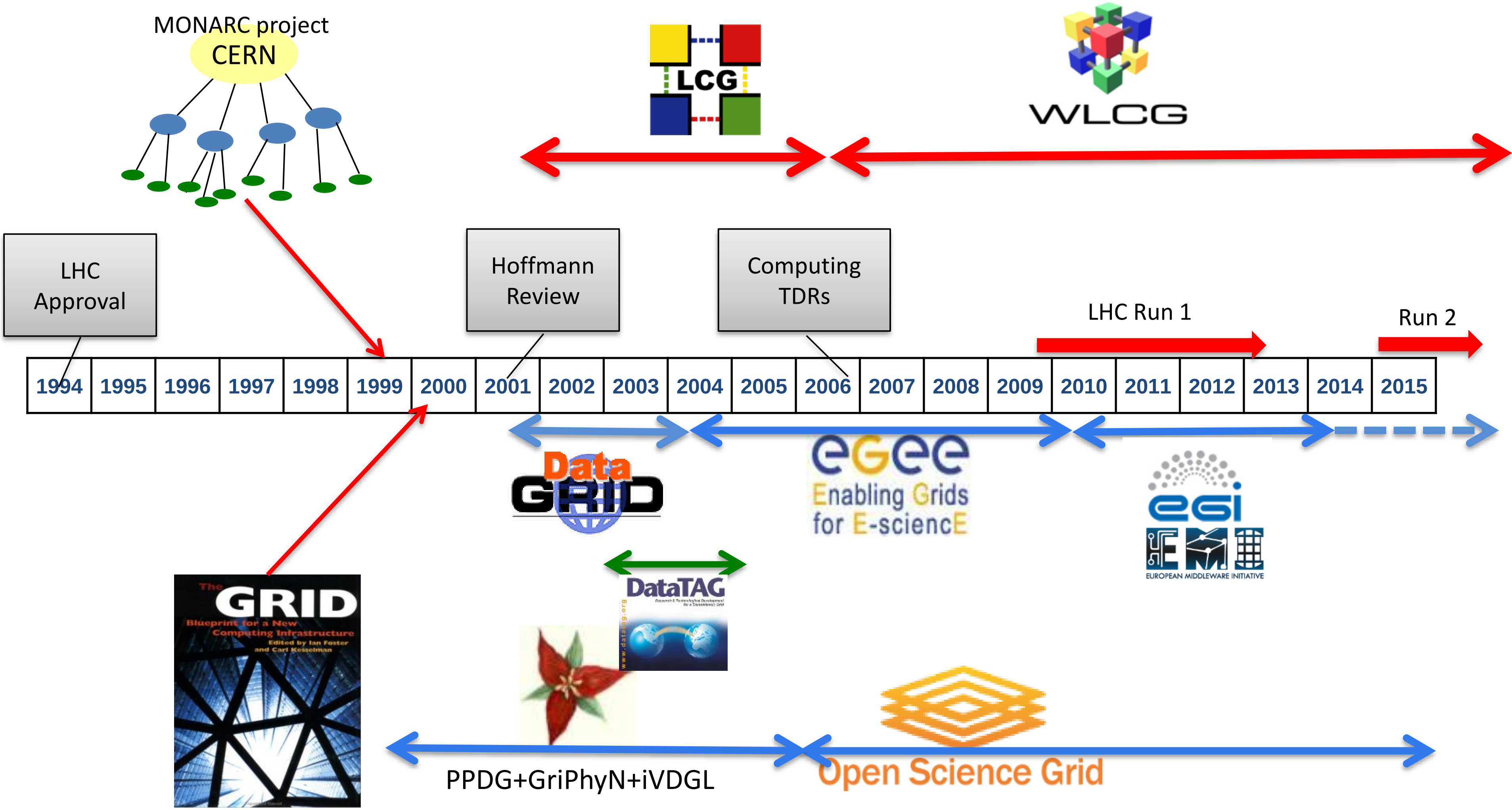
Today: 1 year of LHC data Is ~40 PB

CERN Schedule



- LS2 2019-2020
    - Upgrades of ALICE and LHCb
- LS3 2024-2026
    - Upgrades of ATLAS and CMS (HL-LHC)
- **ALICE upgrade ready in Spring 2021** – 4 years from now, fits well with the CERN openlab next project phase

Grid projects timeline



MONARC project

CERN

LHC Approval

Hoffmann Review

Computing TDRs

LHC Run 1

Run 2

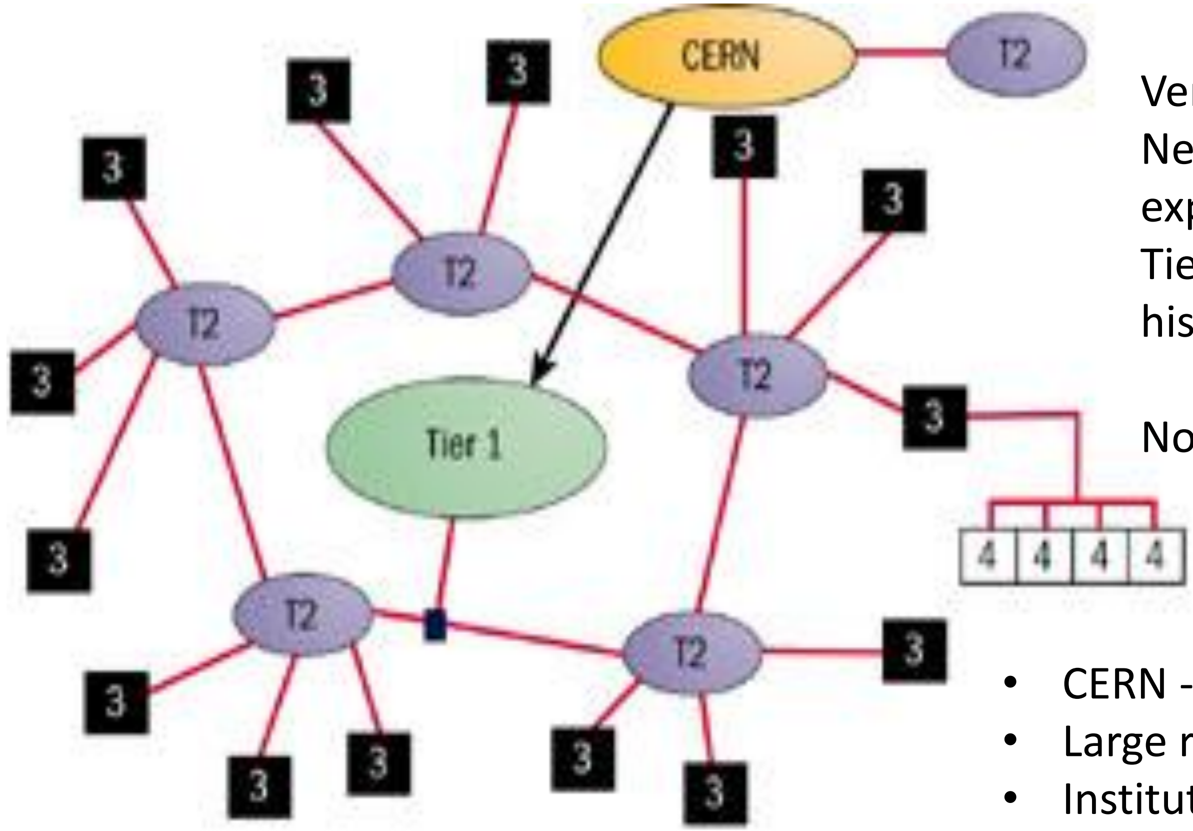
1994 | 1995 | 1996 | 1997 | 1998 | 1999 | 2000 | 2001 | 2002 | 2003 | 2004 | 2005 | 2006 | 2007 | 2008 | 2009 | 2010 | 2011 | 2012 | 2013 | 2014 | 2015

PPDG+GriPhyN+iVDGL

LCG
WLCG
Data GRID
egee
Enabling Grids for E-sciencE
egi
EUROPEAN MIDDLEWARE INITIATIVE
DataTAG
www.datatag.org
The GRID
Blueprint for a New Computing Infrastructure
Edited by Ian Foster and Carl Kesselman
Open Science Grid

MONARC model (1999)

Models of Networked Analysis at Regional Centres for LHC Experiments



Very specific data paths
Network grows ***faster*** than expected
Tier names remain for historical reasons

Now the Grid looks like a cloud

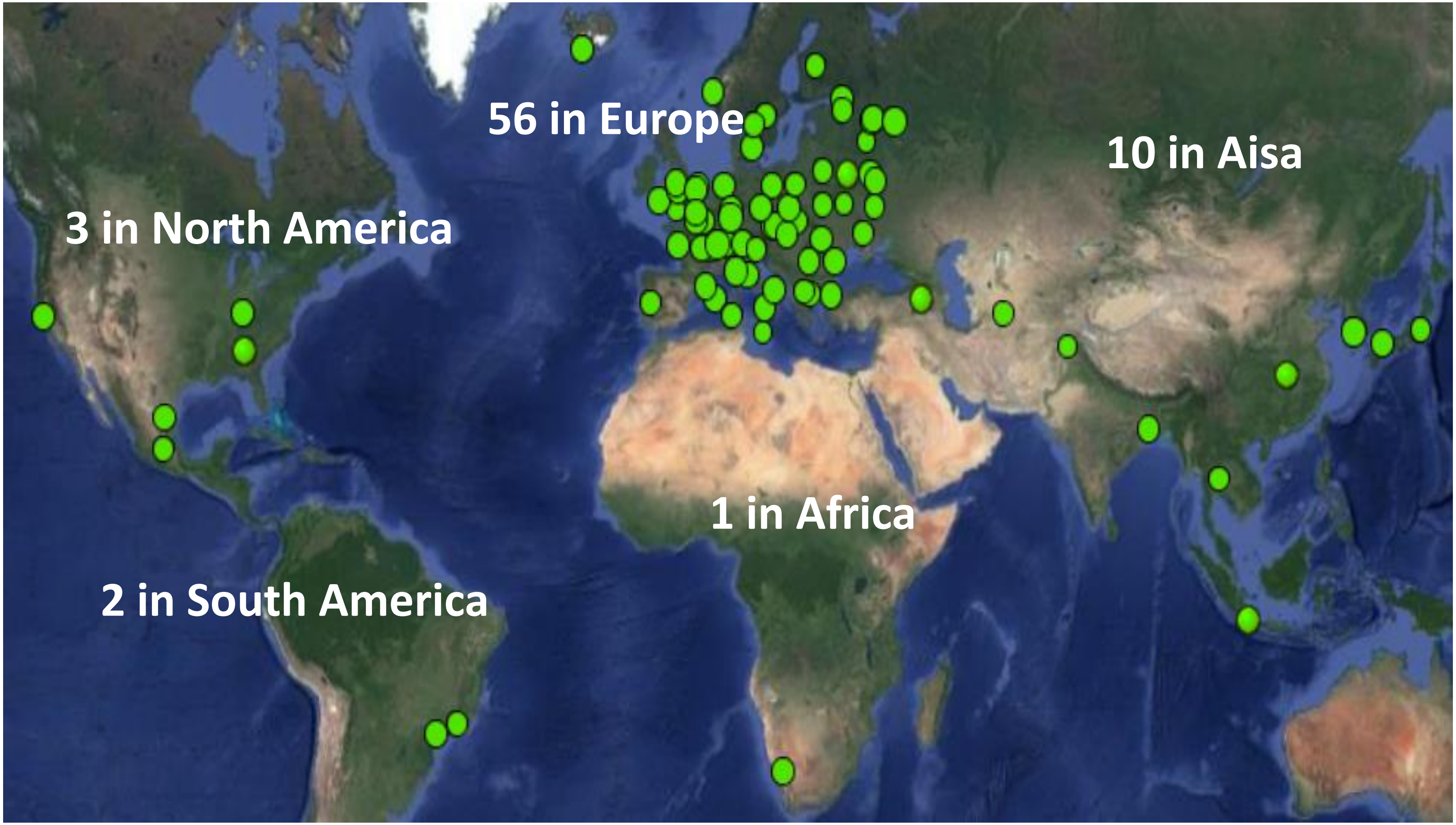
- CERN - **Tier0**
- Large regional centres - **Tier1s**
- Institute/university centres - **Tier2**
- Smaller centres - **Tier3**

Red lines – data paths

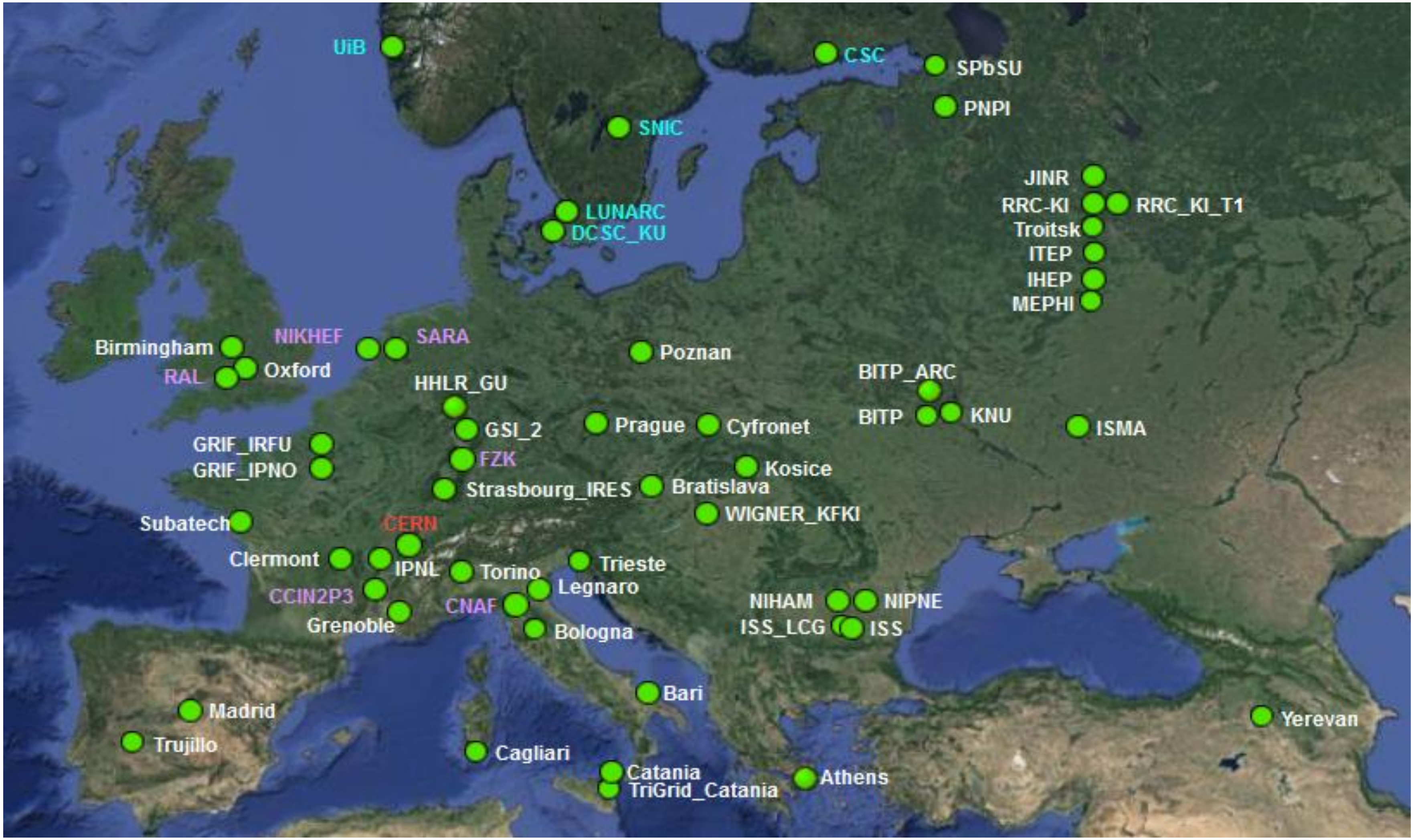
Grid building blocks (layers)

- **Network** connects Grid resources
- **Resource layer** is the actual grid resources: computers and storage
- **Middleware** provides the tools that enable the network and resources layers to participate in a Grid
- **Application software** scientific/engineering programs running on the Grid + portals and development toolkits to support the applications

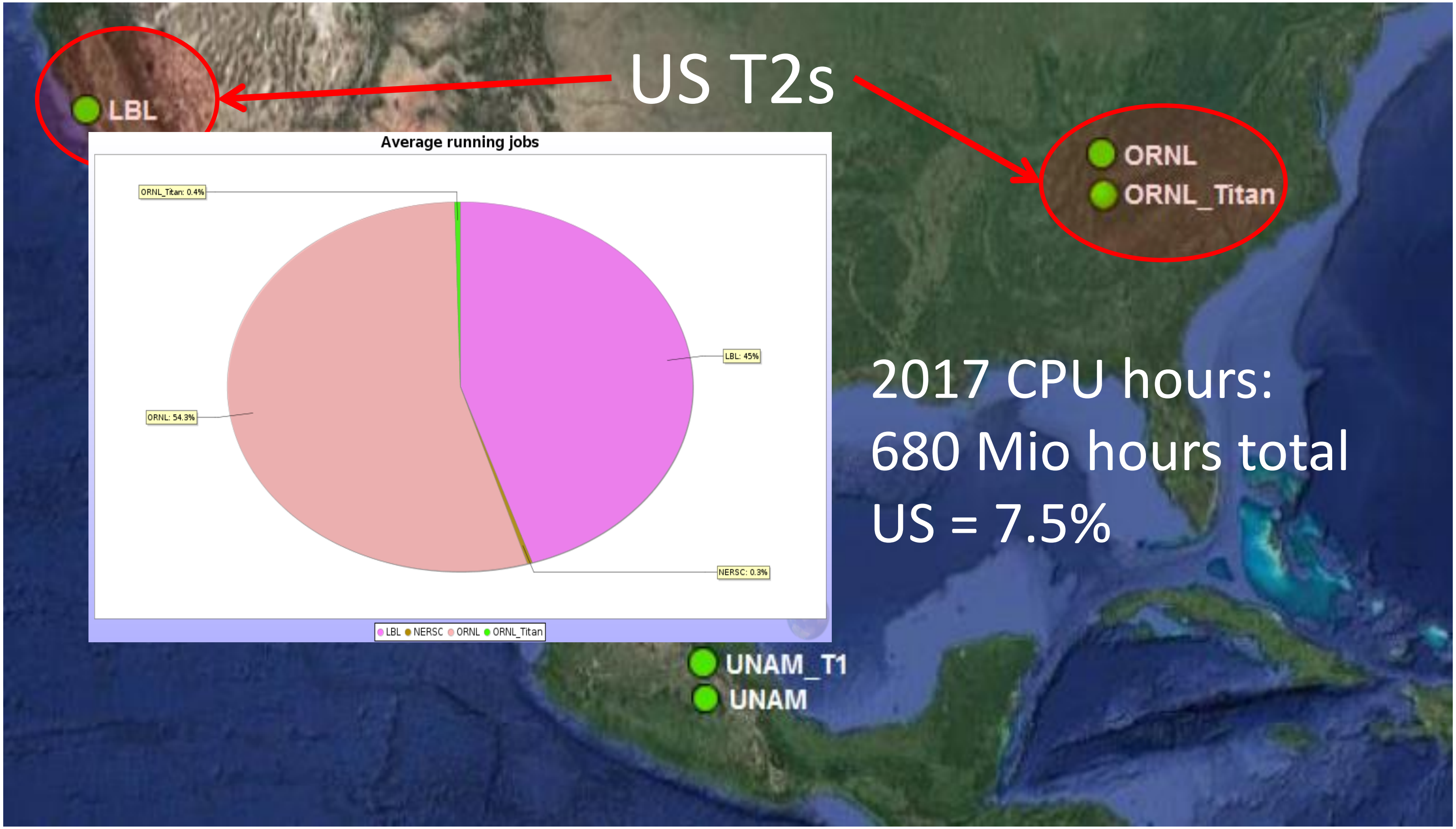
The ALICE Grid sites



Zoom on Europe



Zoom on North America



US T2s

2017 CPU hours:

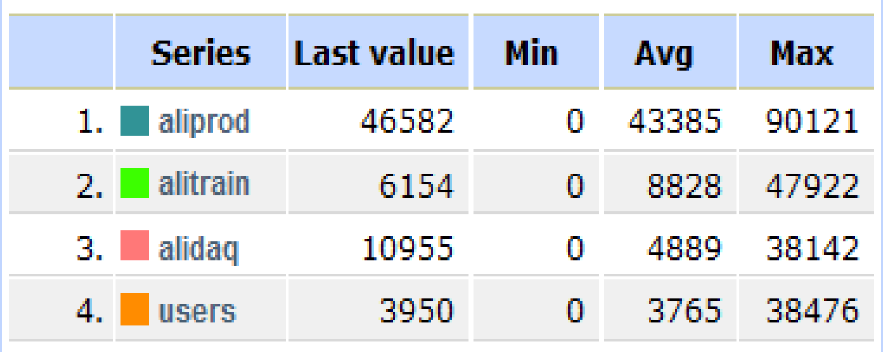
680 Mio hours total

US = 7.5%

Use cases - Offline data processing

- RAW data collection and distribution
  - Unprocessed events from the detectors
- Data processing
  - Calibration, tracking, simulation (physics and detector)
- Analysis objects
  - The data containers for physics analysis
- Analysis
  - The analysis process, resulting in publications

Resources share



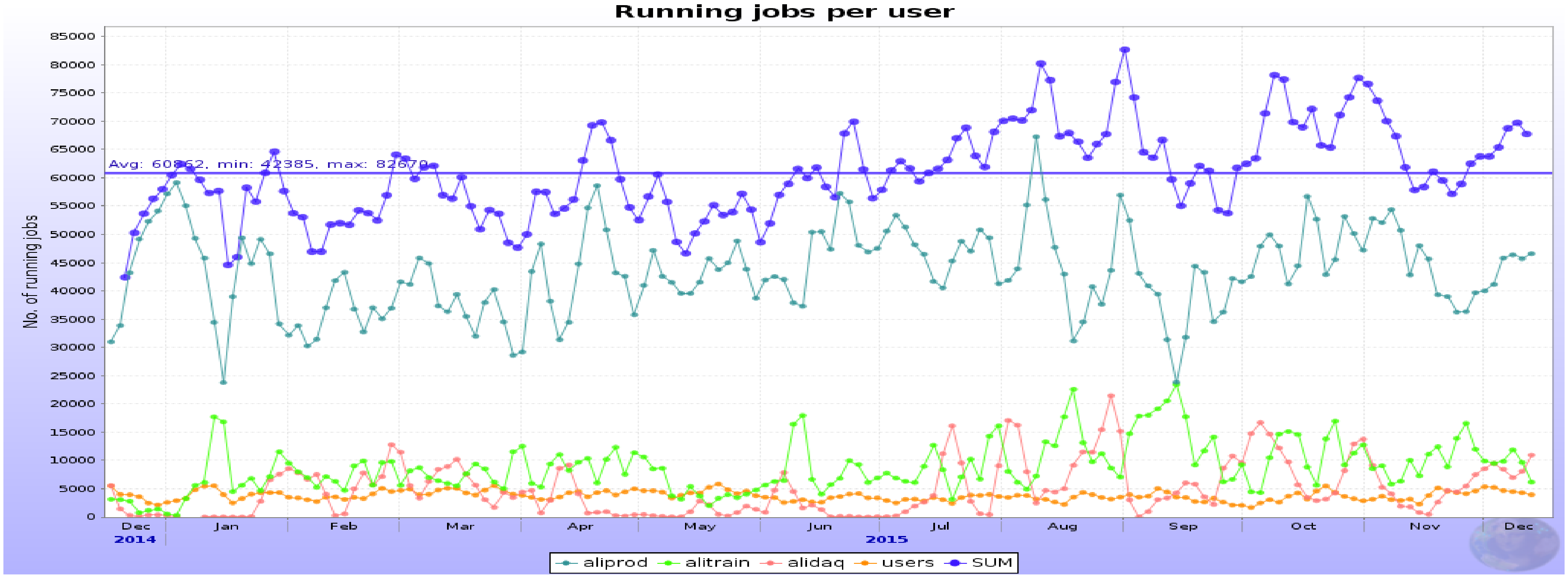
	Series	Last value	Min	Avg	Max
1.	🟦 aliprod	46582	0	43385	90121
2.	🟩 alitrain	6154	0	8828	47922
3.	🟥 alidaq	10955	0	4889	38142
4.	🟧 users	3950	0	3765	38476

71% - MC

14% - Organized analysis

9% - RAW data reconstruction

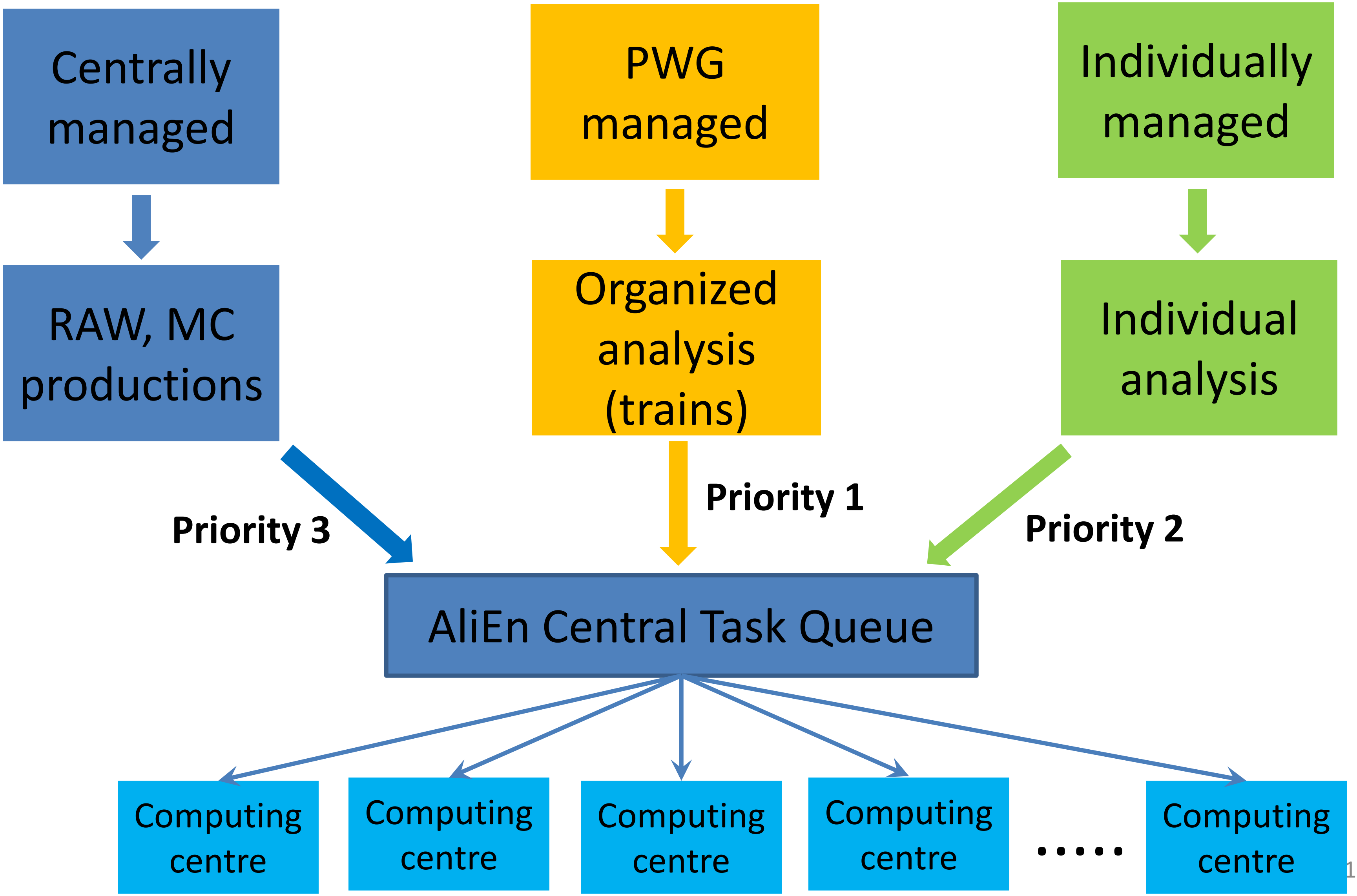
6% - user analysis



ALICE data model

- All ALICE data are annotated in the AliEn catalogue
  - Including the location on site SEs
- Data files are accessed directly
  - Jobs go to the data, in case of local failure reads from closest replica
  - User access to data is managed through a shell, which connects to the catalogue and downloads/uploads data to the site SEs
- Exclusive use of xrootd protocol
  - Also supporting http, ftp, torrent for downloading other input files
  - At the end of the job N replicas are uploaded from the job itself (2x ESDs, 2xAODs, 1x logs and other service files)

Computing tasks and workflow



Centrally managed

RAW, MC productions

PWG managed

Organized analysis (trains)

Individually managed

Individual analysis

Priority 3

Priority 1

Priority 2

AliEn Central Task Queue

Computing centre

Computing centre

Computing centre

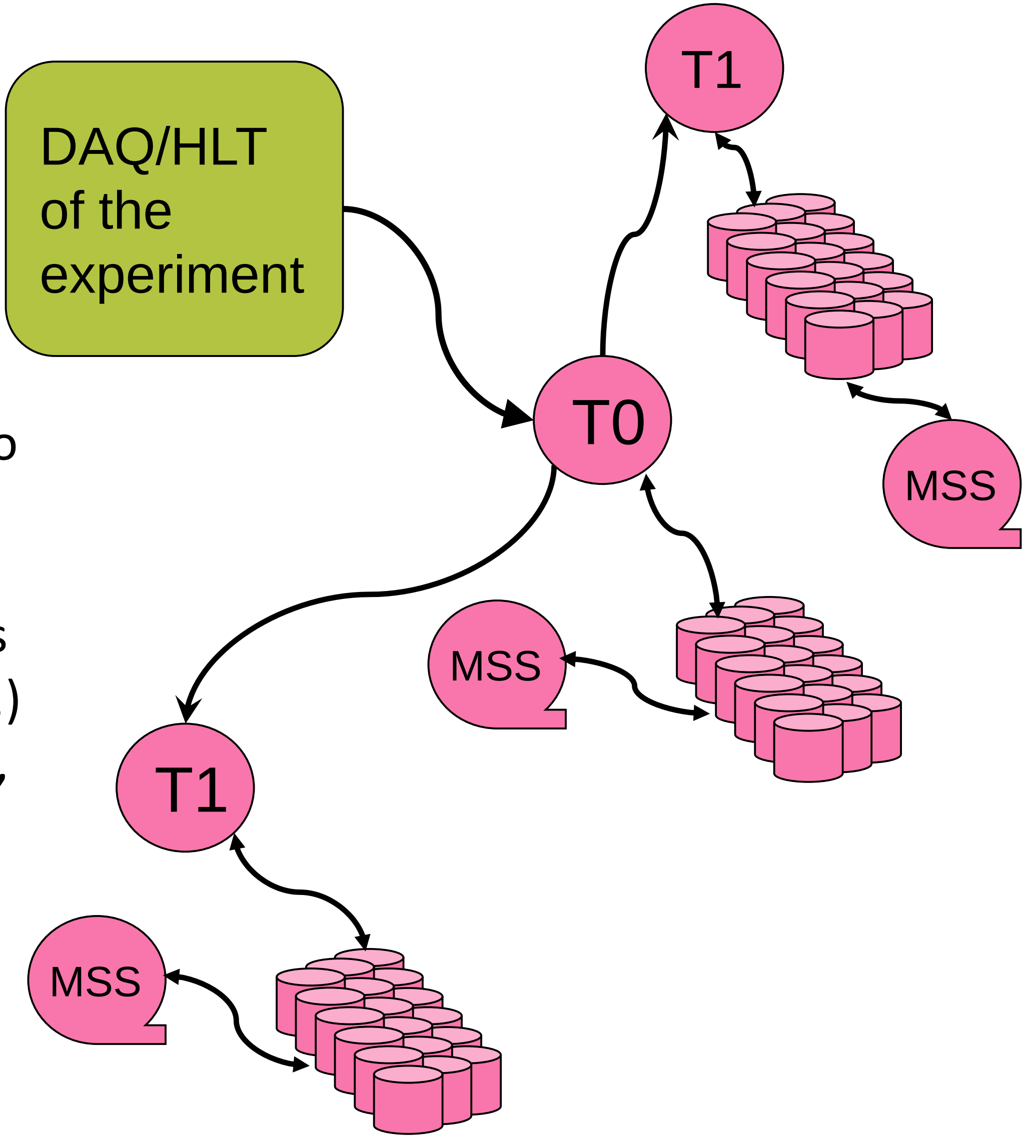
Computing centre

.....

Computing centre

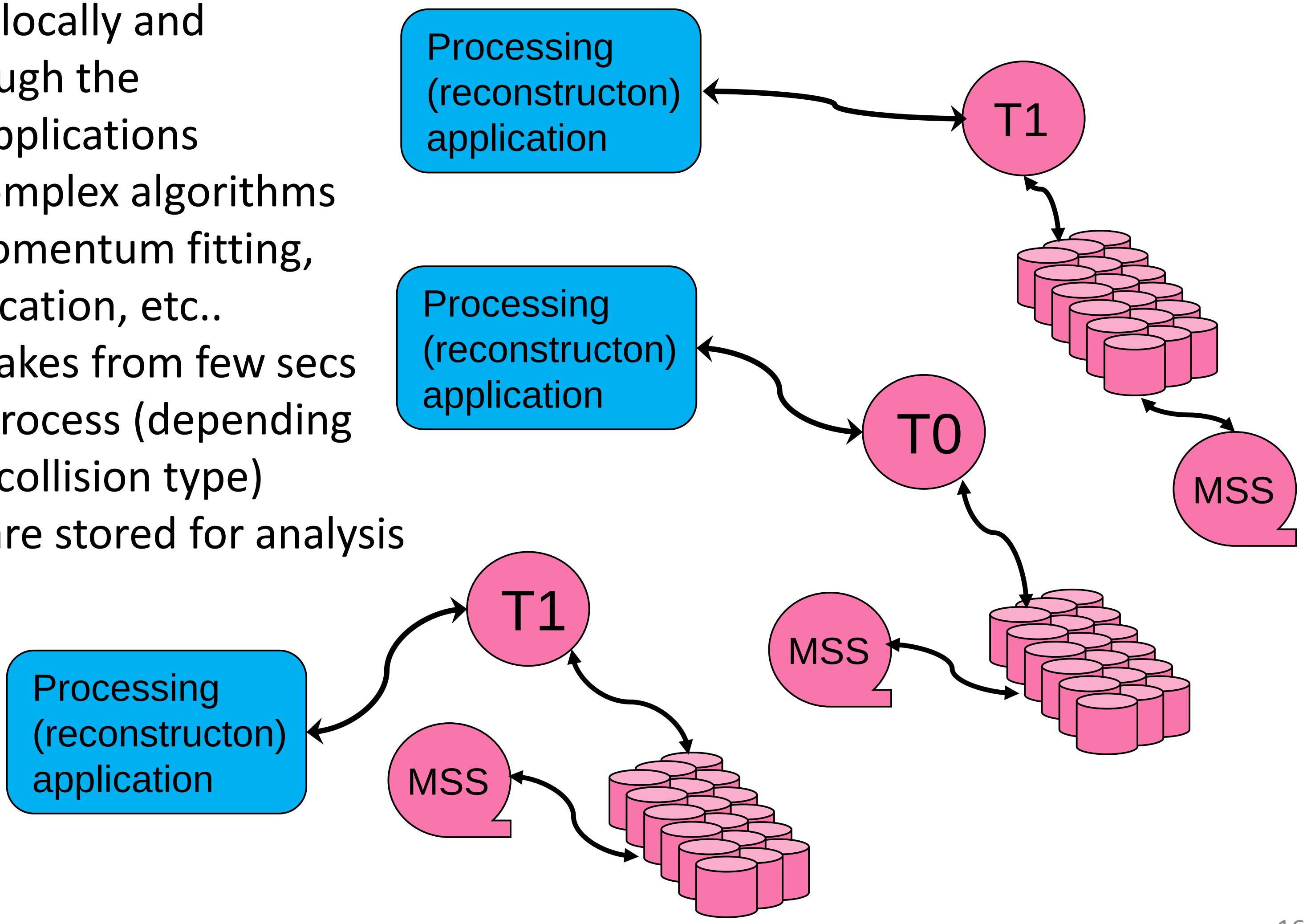
RAW Data distribution

- RAW data is first collected at the T0 centre (CERN)
- One or two copies are made to the remote T1s with custodial storage capabilities
- Custodial (MSS) usually means tape system (but not necessarily!)
- The RAW data is irreplaceable, hence multiple copies



RAW data processing

- RAW data is read from the T0/T1s storage locally and processed through the experiment’s applications
- These are complex algorithms for tracking, momentum fitting, particle identification, etc..
- Each event takes from few secs to minutes to process (depending on complexity, collision type)
- The results are stored for analysis



Processing (reconstructon) application

T1

Processing (reconstructon) application

T0

MSS

MSS

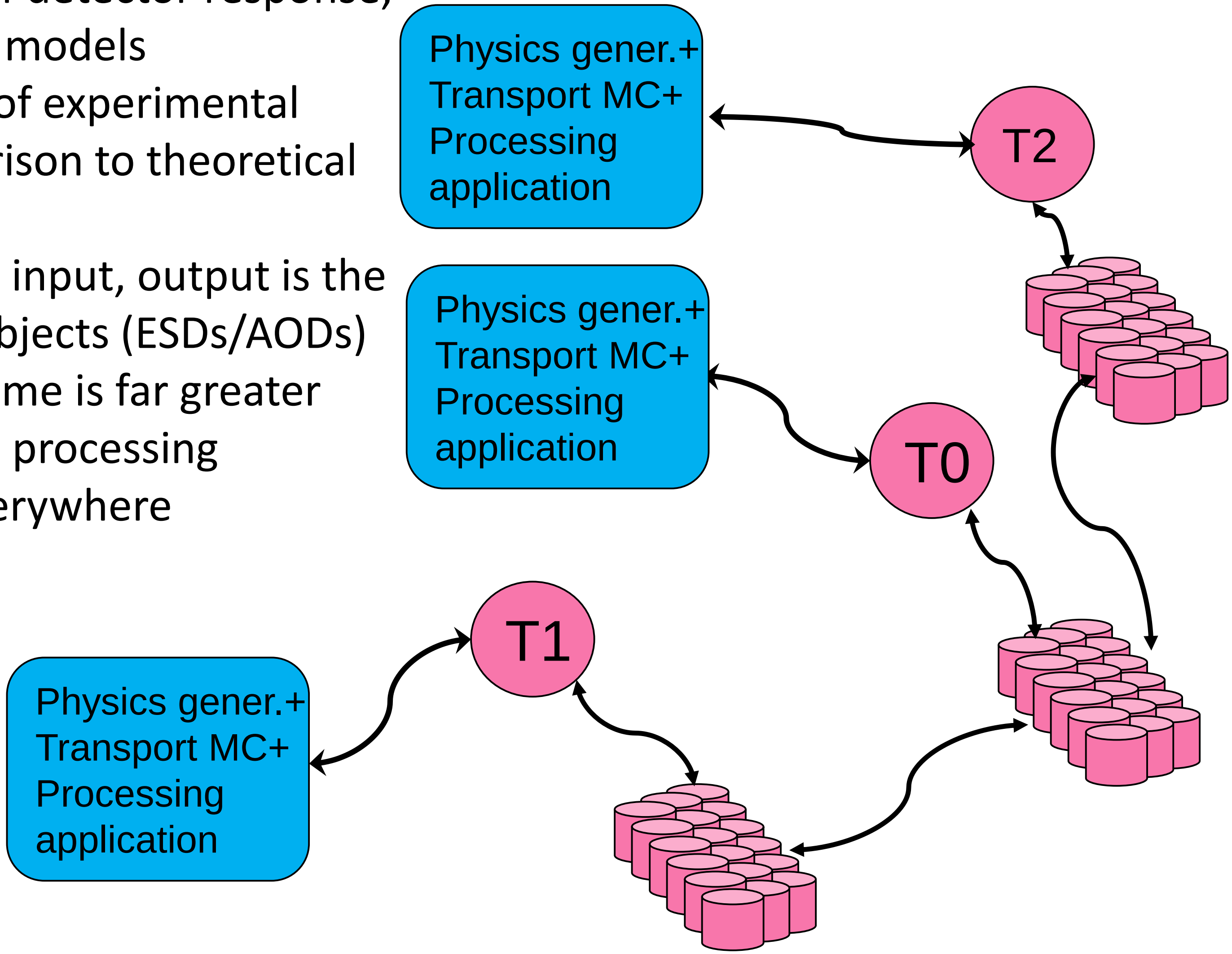
T1

Processing (reconstructon) application

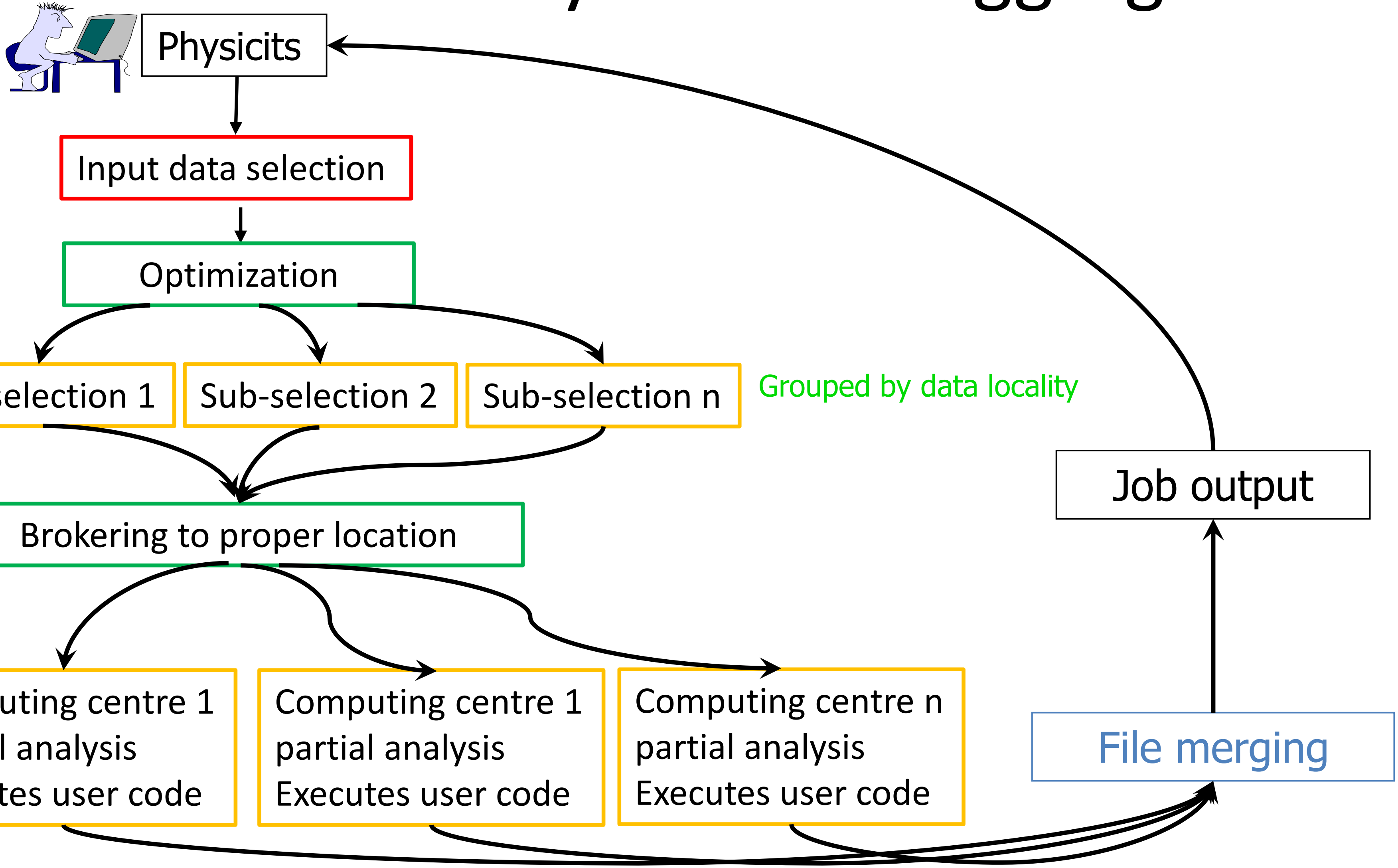
MSS

Monte-Carlo production

- Simulation of detector response, various physics models
- Corrections of experimental results, comparison to theoretical predictions
- MC has little input, output is the Same type of objects (ESDs/AODs)
- Processing time is far greater Than RAW data processing
- MC runs everywhere



Distributed analysis – data aggregation



Physicits

Input data selection

Optimization

Sub-selection 1

Sub-selection 2

Sub-selection n

Grouped by data locality

Brokering to proper location

Computing centre 1
partial analysis
Executes user code

Computing centre 1
partial analysis
Executes user code

Computing centre n
partial analysis
Executes user code

File merging

Job output

Size and evolution of the Grid

- Cores per site vary from hundreds (few) to tens of thousands
  - Average site is 1000 cores
- ~200K CPU cores in the WLCG Grid
- Storage capacity per site – hundred of TBs to tens of PBs
- In a “Flat budget world” Grid growth is assured by technology advances: Moore’s law (or whatever is left of it) and Kryder’s law (with modifications)
- In general, the Grid

resources grow at about

20% per year (Grid’s law)

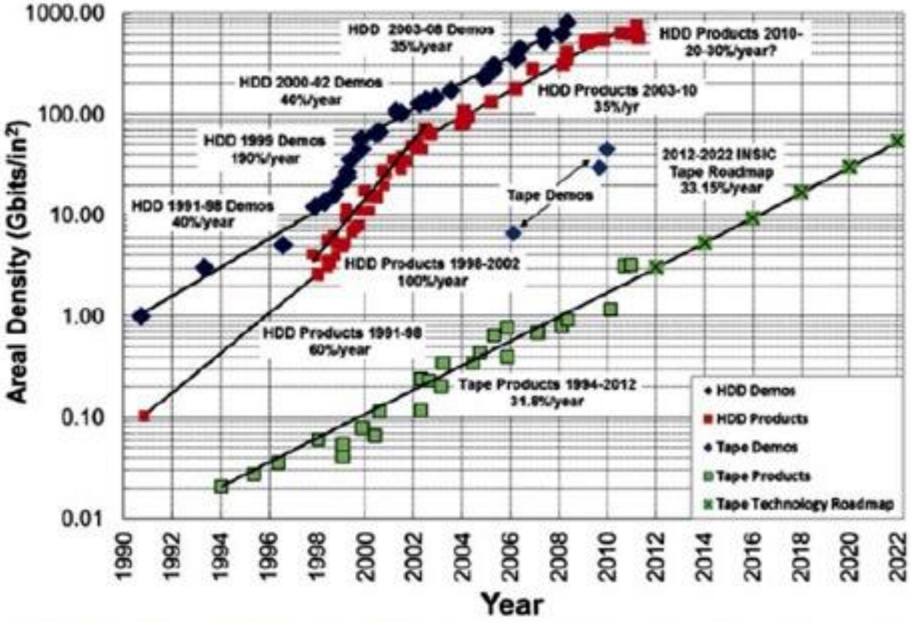
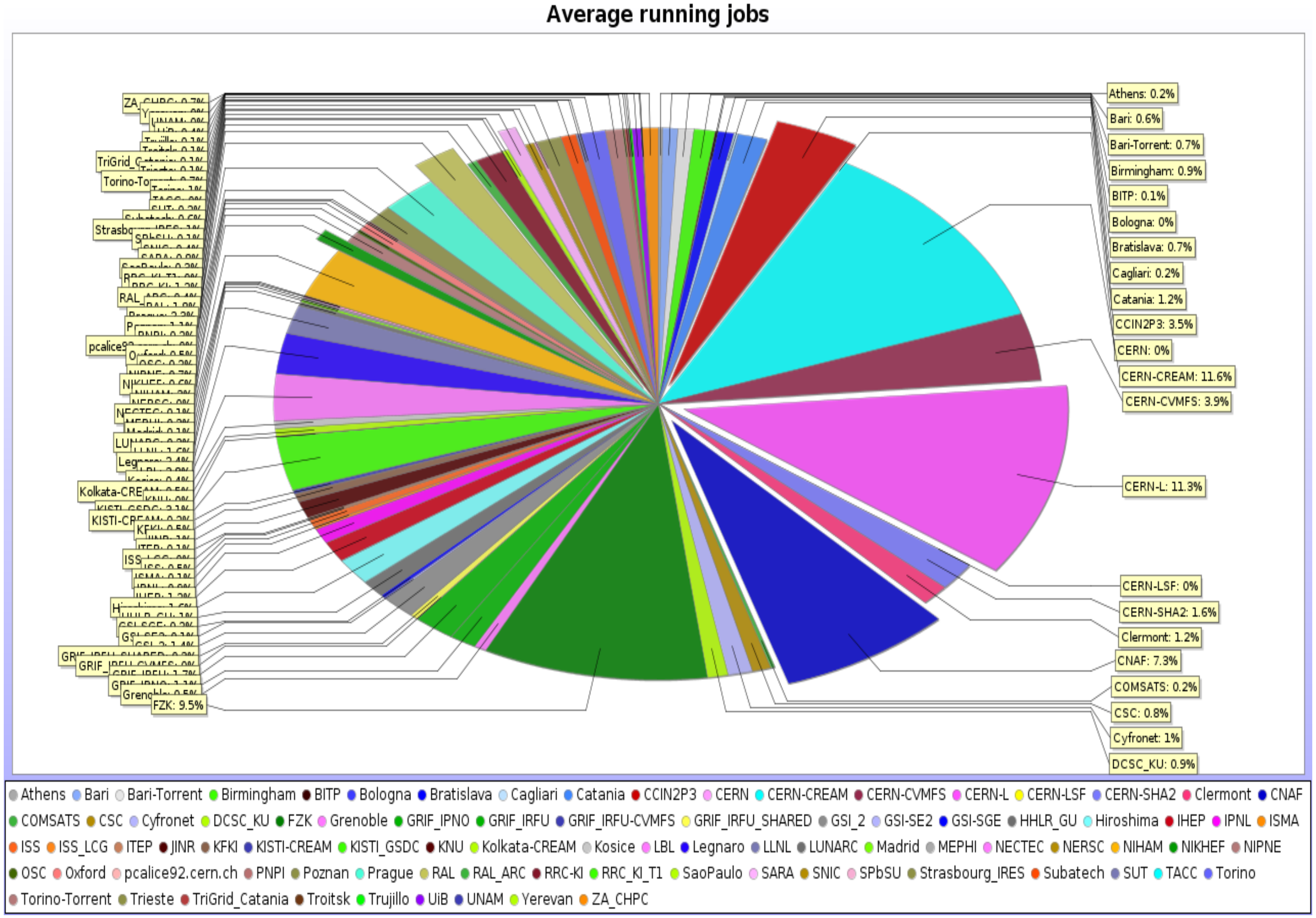


Figure 40: Areal Density of Hard Disk and Tape Laboratory Demonstrations and Products [Reference 71].

Resources distribution

Remarkably stable 50/50 share between T1s and T2s over 10 years



Central services

AliEn FC

Central catalogue of logical file names (LFN)

- With owner:group and unix-style permissions
- Size, MD5 of files, metadata
- A GUID is associated to each LFN
- Multiple physical file names (PFN) can be associated to a LFN
- root://<redirector>//<HH>/<hhhhh>/<GUID>

  HH and hhhhh are hashes of the GUID

Task queue

Central queue for all jobs executed on the Grid

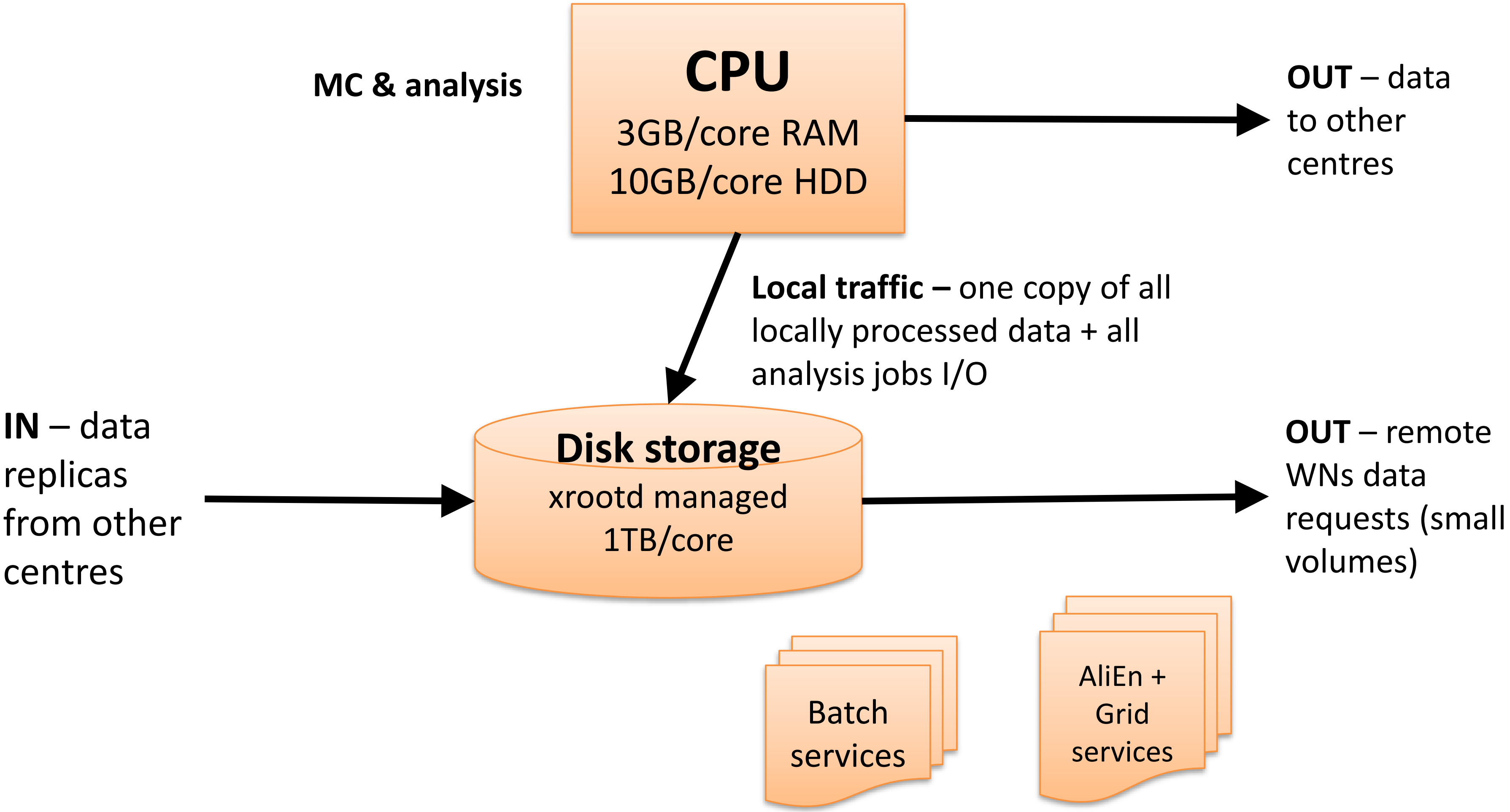
- Master Jobs are using JDL and submitted to the queue
- A service splits the jobs into sub-jobs matching sites capabilities
- Job broker assigns jobs to sites
- Job traces are kept of each sub- and master job from start to completion
- Quotas and priorities are assigned to each user and job

Splitter,
Optimizer,
LDAP,
Authen,
Broker

API services

All running on a set of servers at CERN

GRID node



MC & analysis

CPU

3GB/core RAM

10GB/core HDD

OUT – data to other centres

Local traffic – one copy of all locally processed data + all analysis jobs I/O

IN – data replicas from other centres

Disk storage

xrootd managed

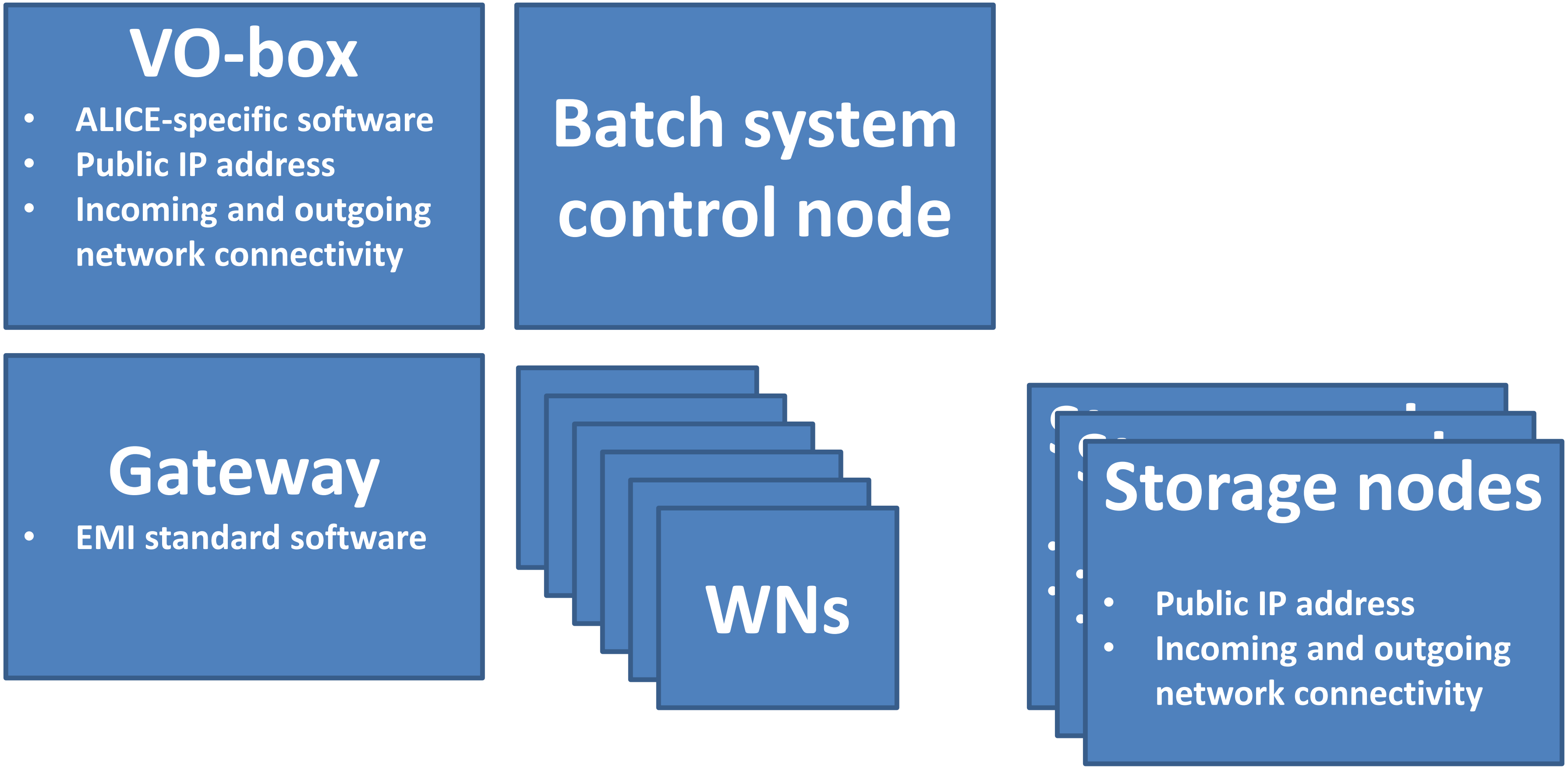
1TB/core

OUT – remote WNs data requests (small volumes)

Batch services

AliEn + Grid services

Structure of a T2



VO-box

- ALICE-specific software
- Public IP address
- Incoming and outgoing network connectivity

Batch system control node

Gateway

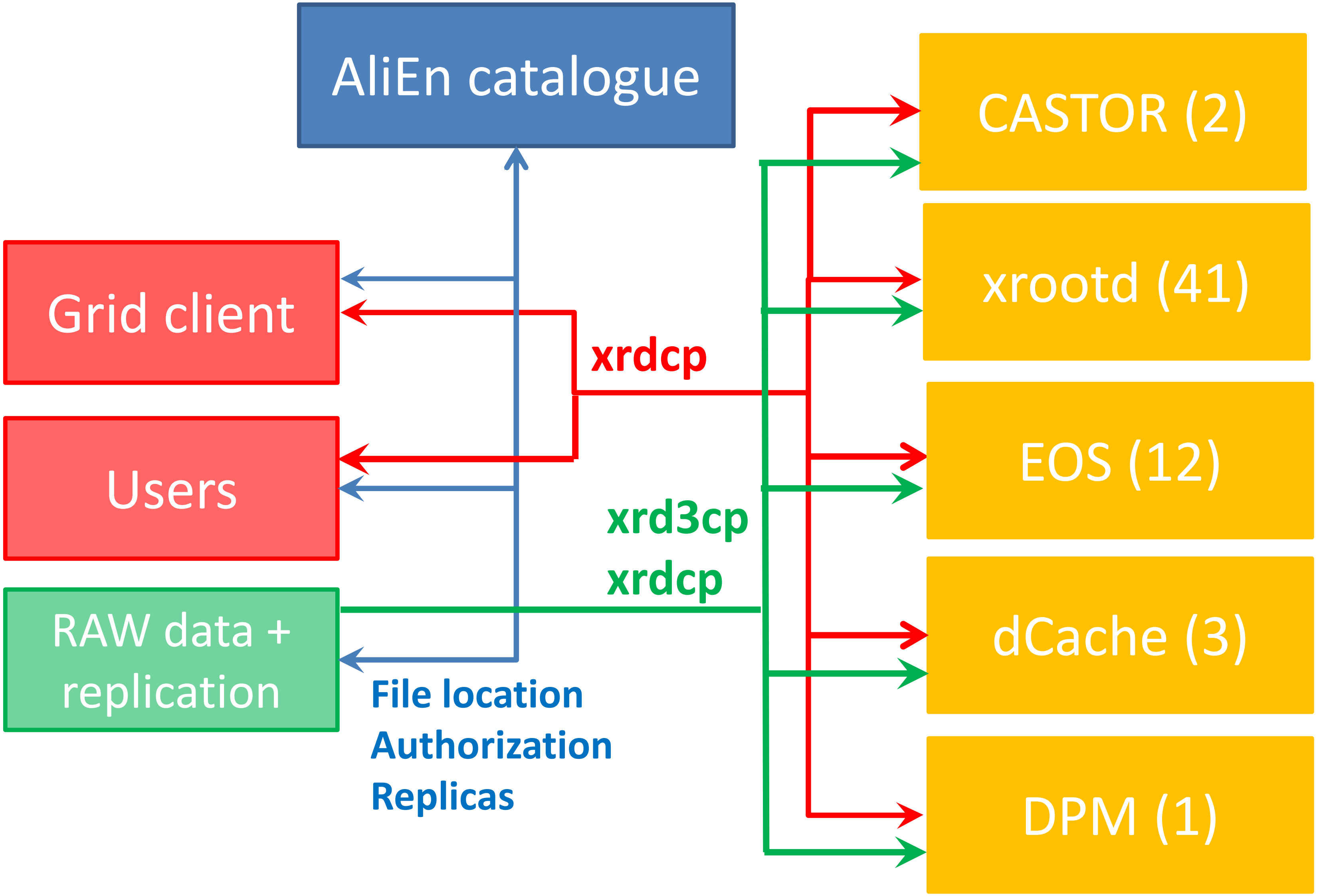
- EMI standard software

WNs

Storage nodes

- Public IP address
- Incoming and outgoing network connectivity

Storage types, protocol and interactions



Contribution - USA

- Based on the present contribution to M&O-A
  - 44 US M&O payers for a total of 615 - 54 (CERN)
  - => 7.84%, from October 2016 RRB and based on the requirements document for the CRSG

Year	CPU KHS06 (cores*)	Disk PB
2017	48.8 (3250)	4.4
2018	58.3 (3890)	5.7
2019	89.1 (5940)	7.1

*assumes 15HS06/core

  - 2020 either the same as in 2019, or the same as in 2019 and less in 2019 to have a smooth growth, or +20% as compared to 2019 to prepare for RUN3.

ALICE O^2 in a nutshell

Requirements

1. LHC min bias Pb-Pb at 50 kHz
   ~100 x more data than during Run 1
2. Physics topics addressed by ALICE upgrade
   - Rare processes
   - Very small signal over background ratio
   - Needs large statistics of reconstructed events
   - Triggering techniques very inefficient if not impossible
3. 50 kHz > TPC inherent rate (drift time ~100 µs)
   Support for continuous read-out (TPC)
   - Detector read-out triggered or continuous

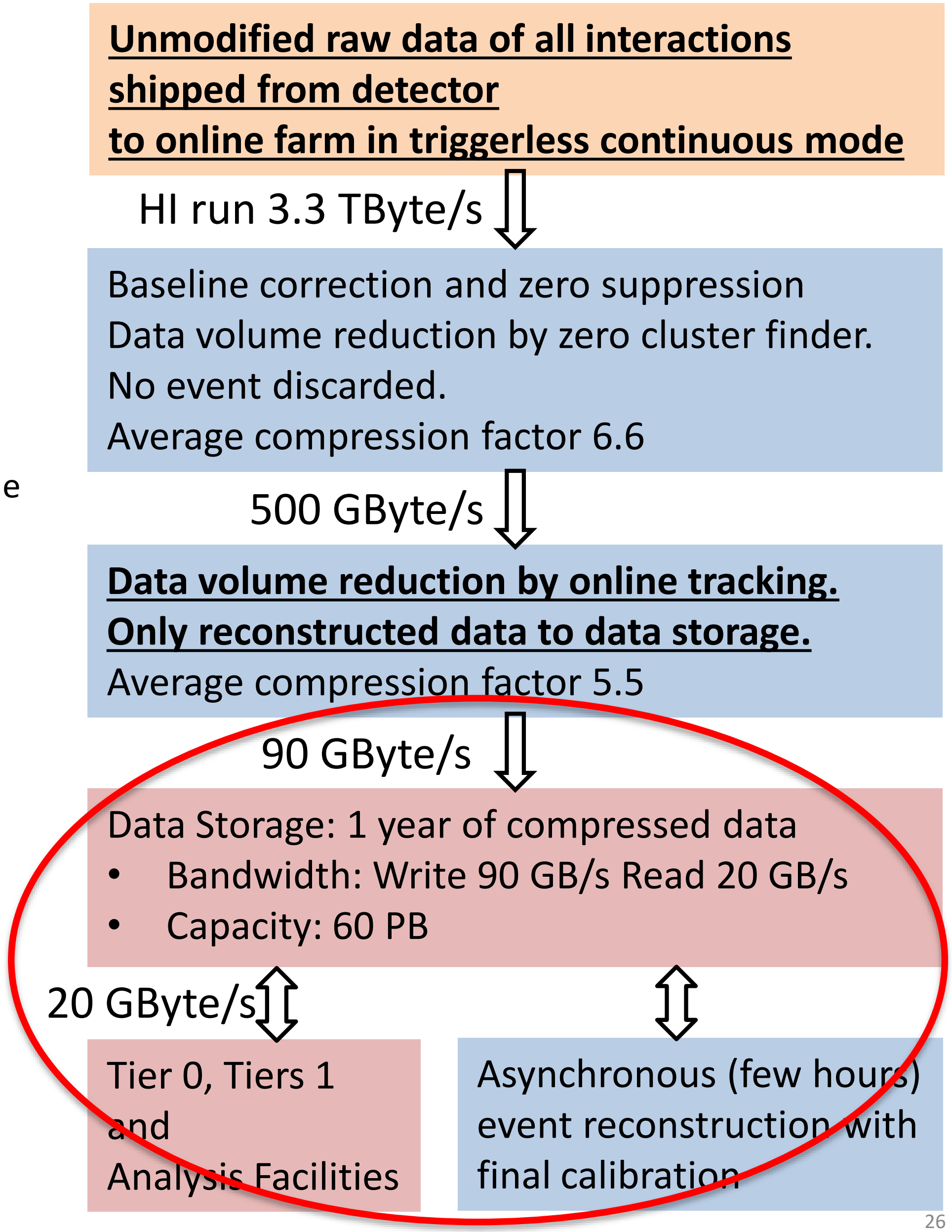
New computing system

- Read-out the data of all interactions

➔ Compress these data intelligently
by online reconstruction

➔ One common online-offline
computing system: O^2

- Paradigm shift compared to approach
for Run 1 and 2



Unmodified raw data of all interactions shipped from detector to online farm in triggerless continuous mode

HI run 3.3 TByte/s

Baseline correction and zero suppression
Data volume reduction by zero cluster finder.
No event discarded.
Average compression factor 6.6

500 GByte/s

Data volume reduction by online tracking.
Only reconstructed data to data storage.
Average compression factor 5.5

90 GByte/s

Data Storage: 1 year of compressed data

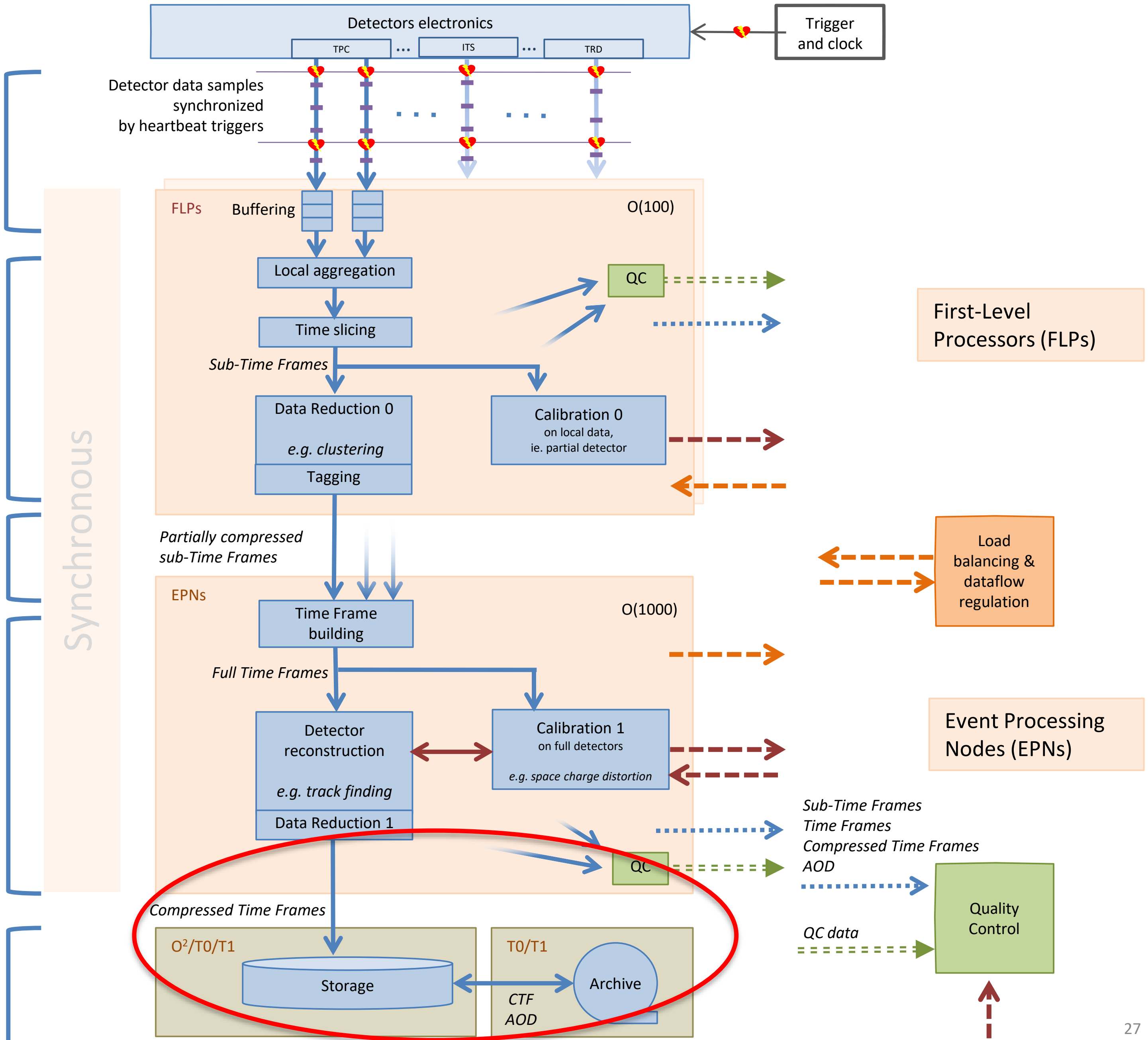
- Bandwidth: Write 90 GB/s Read 20 GB/s
- Capacity: 60 PB

20 GByte/s

Tier 0, Tiers 1 and Analysis Facilities

Asynchronous (few hours) event reconstruction with final calibration

Data flow & processing (1)



Detectors electronics

TPC … ITS … TRD

Trigger and clock

Detector data samples synchronized by heartbeat triggers

Raw data input

Synchronous

FLPs

Buffering

O(100)

Local aggregation

QC

Time slicing

Sub-Time Frames

Local processing

Data Reduction 0

e.g. clustering

Tagging

Calibration 0

on local data,
ie. partial detector

First-Level Processors (FLPs)

Partially compressed sub-Time Frames

Frame dispatch

Load balancing & dataflow regulation

EPNs

O(1000)

Time Frame building

Full Time Frames

Detector reconstruction

e.g. track finding

Data Reduction 1

Calibration 1

on full detectors

e.g. space charge distortion

Event Processing Nodes (EPNs)

Global processing

QC

Sub-Time Frames
Time Frames
Compressed Time Frames
AOD

QC data

Quality Control

Compressed Time Frames

Storage

O^2/T0/T1

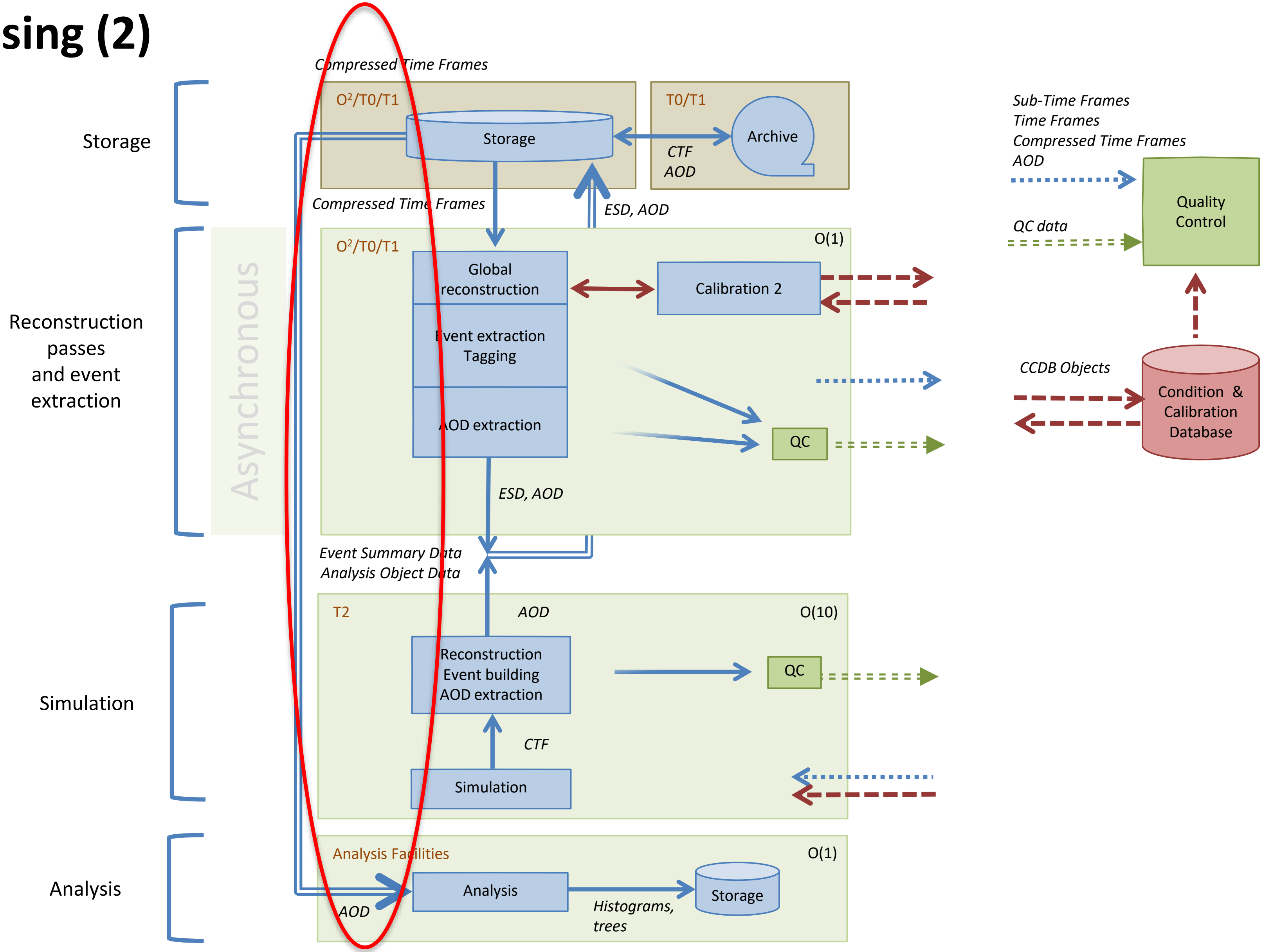
Storage

T0/T1

Archive

CTF
AOD

Data flow & processing (2)



Storage

Reconstruction passes and event extraction

Simulation

Analysis

Asynchronous

Compressed Time Frames

O²/T0/T1

Storage

T0/T1

Archive

CTF
AOD

Compressed Time Frames

ESD, AOD

O²/T0/T1

O(1)

Global reconstruction

Event extraction Tagging

AOD extraction

Calibration 2

QC

ESD, AOD

Event Summary Data
Analysis Object Data

T2

AOD

O(10)

Reconstruction
Event building
AOD extraction

QC

CTF

Simulation

Analysis Facilities

O(1)

AOD

Analysis

Histograms, trees

Storage

Sub-Time Frames
Time Frames
Compressed Time Frames
AOD

QC data

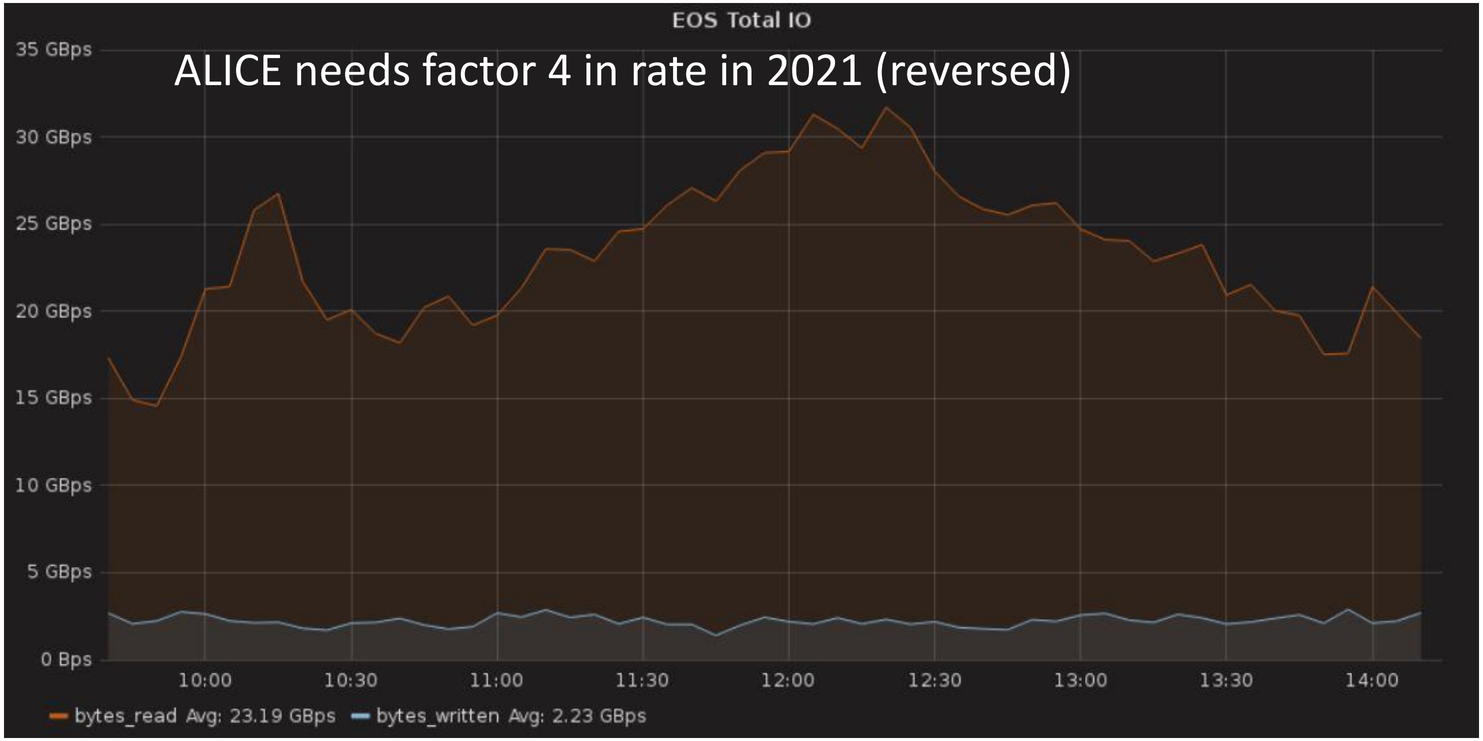
Quality Control

CCDB Objects

Condition & Calibration Database

Challenges

- Rates to storage – write 90GB/sec , read 20GB/sec out (+ delta)
- Capacity – 60PB in a single instance (first year)
- High availability – on the critical path for data taking
- Complex interactions with various systems – experiment/Grid/analysis
- Current experience (borrowed from EOS)



ALICE needs factor 4 in rate in 2021 (reversed)