

10

10th International Conference  
“Distributed Computing and Grid Technologies in  
Science and Education”

GRID2023  
3-7 July 2023



## BM@N Computing Software Architecture and its use for the mass production

Konstantin Gertsenberger  
Joint Institute for Nuclear Research, Dubna

*on behalf of the BM@N collaboration*



4 July 2023

# Nuclotron-based Ion Collider fAcility



- Beams: from  $p, d^\uparrow$  to  $Bi$
- Luminosity:  $10^{27}$  ( $Bi$ ),  $10^{32}$  ( $p$ )  $cm^{-2}s^{-1}$
- Collision energy:  $\sqrt{s_{NNAU}} = 4 - 11$  GeV  $E_{lab} = 1 - 5$  AGeV

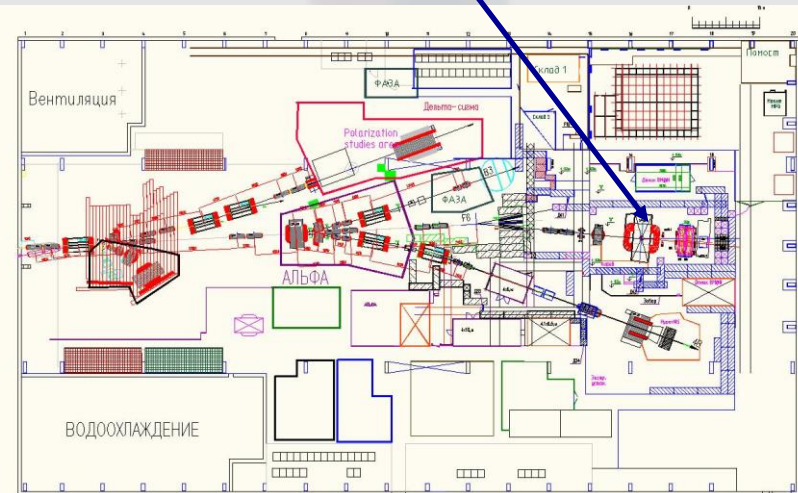
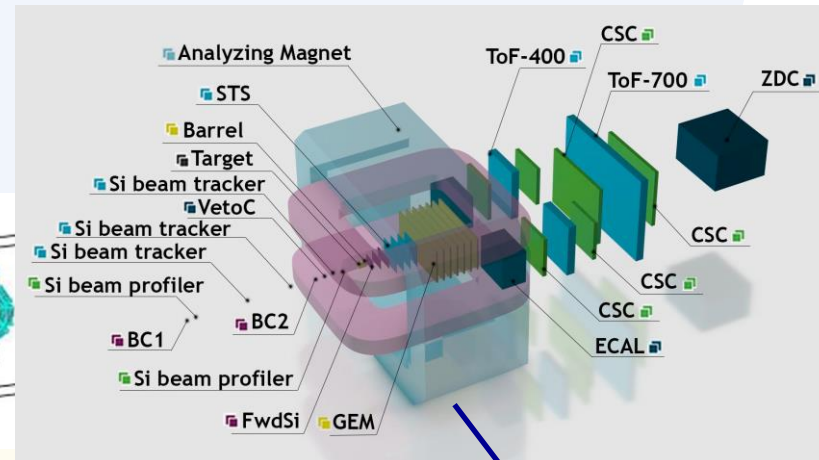
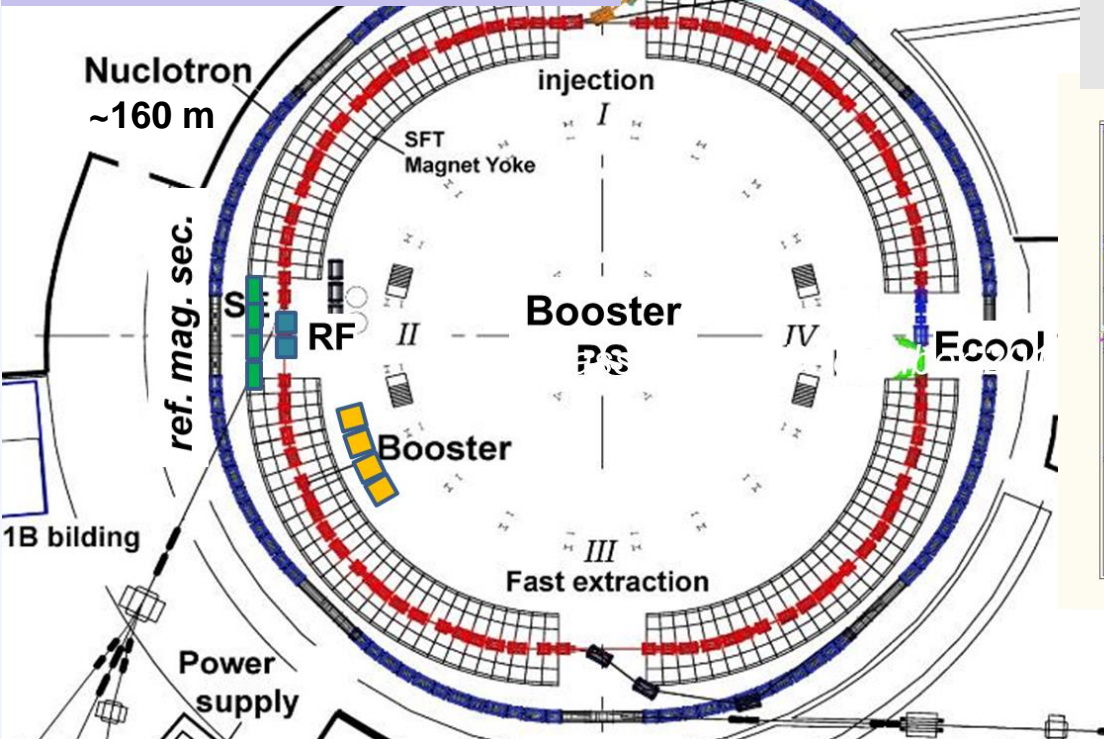
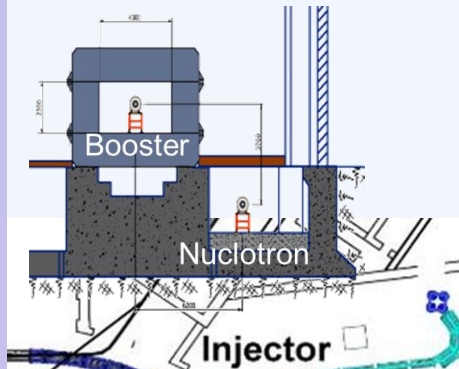
- Fixed target experiment: BM@N (2018)
- 2 interaction points: MPD (2025) & SPD (2028)
- Official site: [nica.jinr.ru](http://nica.jinr.ru), [bmj.jinr.ru](http://bmj.jinr.ru)



# Baryonic Matter @ Nuclotron

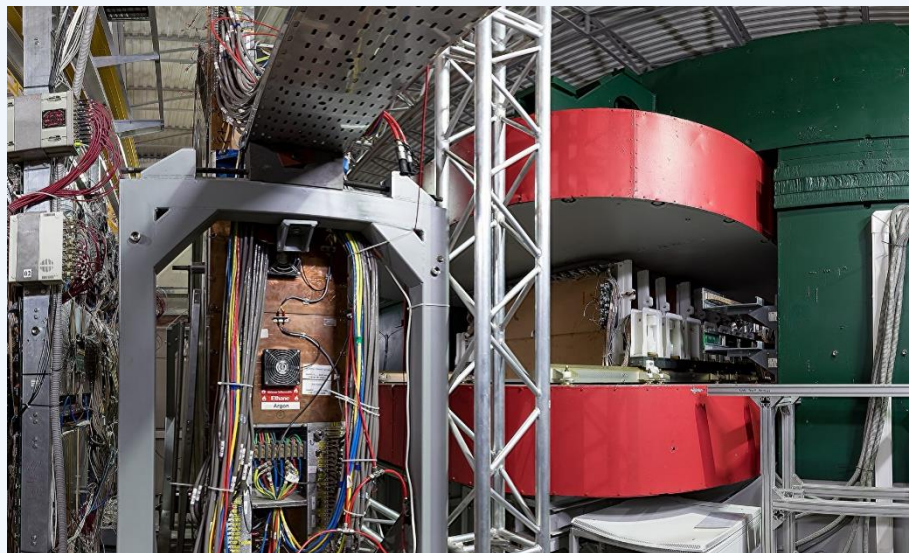
## BM@N Physics Program:

- ✓ strange / multi-strange hyperon and hypernuclei production at the threshold
- ✓ in-medium modifications of strange & vector mesons in dense nuclear matter
- ✓ hadron femtoscopy
- ✓ short range correlations
- ✓ event-by event fluctuations
- ✓ electromagnetic probes, states decaying into  $\gamma$ ,  $e$  (with ECAL)



# BM@N in Nuclotron Runs (2015 – 2023)

❖ Nuclotron Run 51 (d,C)		<i>Feb. 22 – Mar. 15, 2015</i>
❖ Nuclotron Run 52 (d)	<b>Technical</b>	<i>June 29 – June 30, 2016</i>
❖ Nuclotron Run 53 (d, d <sup>†</sup> )	<i>interaction rate: 5 kHz</i>	<i>Dec. 09 – Dec. 23, 2016</i>
❖ Nuclotron Run 54 (C)	<b>Technical+SRC Physics</b>	<i>Mar. 07 – Mar. 18, 2017</i>
❖ Nucl. Run 55 (C,Ar,Kr)	<i>interaction rate: 8 kHz</i>	<i>Mar. 03 – Apr. 05, 2018</i>
❖ Nucl. Run 56: SRC (C)	<b>Physics</b>	<i>Mar. 07 – Mar. 28, 2022</i>
❖ Nucl. Run 57: BM@N (Xe)	<i>interaction rate: 10 kHz</i>	<i>Dec. 12 – Feb. 02, 2023</i>



- Beam: **Xe** (3.8, 3.0 AGeV),  
previous runs: Kr (2.3, 2.6, 3.0 AGeV), Ar (3.2 AGeV),  
 $C^{12}$  (3.5–4.5 AGeV), d (4, 4.6 AGeV)
- Target: **Cs** or empty  
previous runs: Pb, Sn, Cu, Al,  $C_2H_4$ , C,  $H_2$
- Integrated DAQ,  $T_0$  and Trigger systems
- Detectors: FSD, GEM, CSC, ToF-400, ToF-700,  
DCH 1&2, FHCaI, ECal, LAND, profilometers...
- Detect min bias beam-target interactions to  
reconstruct hyperons, strange particles, identify  
charged particles and nucleus fragments...

# BM@N Run 8 Data Production

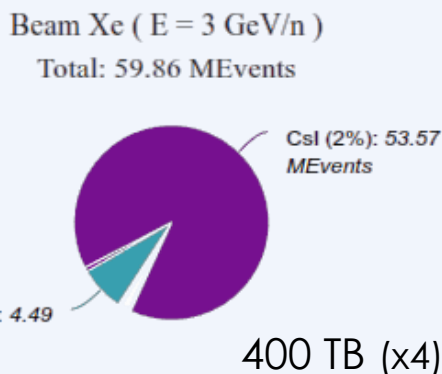
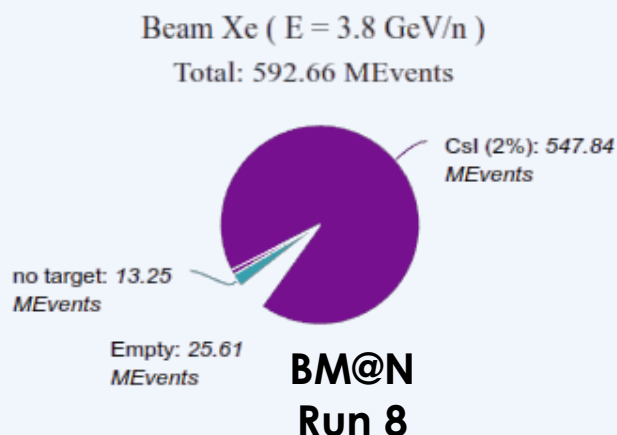
Description	Value	Unit	Symbol	Comment
<b>Data acquisition time</b>	<b>720</b>	<b>hour</b>	<i>T</i>	<i>Eacc</i> = 32%, 62%
run duration	20	min	<i>Trun</i>	
run time break	2.5	min	<i>Tbr</i>	
Beam intensity (3.8 AGeV)	up to 900k / 2.2 up to 900k / 12	Xe <sup>+</sup> /sec	<i>Ibeam_spill</i> <i>Ibeam_period</i>	409k 75k
<b>Trigger rate</b>	<b>8k / 2.2</b>	<b>event/sec</b>	<i>Itrigger</i>	3 636
<b>Event size</b>	<b>0,57</b>	<b>MB</b>	<i>Vevent</i>	
<b>Data rate</b>	<b>2</b>	<b>GB/sec</b>	<i>Idata</i>	= <i>Itrigger</i> * <i>Vevent</i>
Avg event/sec per all data	280	event/sec	<i>Revent</i>	50% empty
<b>Raw file size</b>	<b>15</b>	<b>GB</b>	<i>Vraw</i>	5-10% spill data
Event count per file	25 000		<i>Ievent</i>	= <i>Vraw</i> / <i>Vevent</i>
<b>Total event count</b>	<b>645 M</b>		<i>Nevent</i>	= <i>T</i> * <i>Revent</i> * <i>Trun</i> / ( <i>Trun</i> + <i>Tbr</i> )
Total file count	25 800		<i>Nfile</i>	= <i>Nevent</i> / <i>Ievent</i>
Total run count	1 920		<i>Nrun</i>	= <i>T</i> / ( <i>Trun</i> + <i>Tbr</i> )
<b>Total raw data size</b>	<b>378</b>	<b>TB</b>	<i>Nraw</i>	= <i>Vevent</i> * <i>Nevent</i>
Total replicated raw data (x4)	1512	TB	<i>Nraw_repl</i>	LHEP EOS x2 + MLIT EOS x2
<b>Avg digit file size</b>	<b>870</b>	<b>MB</b>	<i>Vdigit</i>	
Total digit file size (x3 software version)	64	TB	<i>Ndigit</i>	= <i>Nraw</i> * <i>Vdigit</i> / <i>Vraw</i> * 3
<b>Avg DST file size</b>	<b>2 000</b>	<b>MB</b>	<i>Vdst</i>	
Total DST file size <i>1 version for each digit file</i>	150	TB	<i>Ndst</i>	= <i>Nraw</i> * <i>Vdst</i> / <i>Vraw</i> * 3



# Data Collected in BM@N Run 8 (comparing with Run 7)

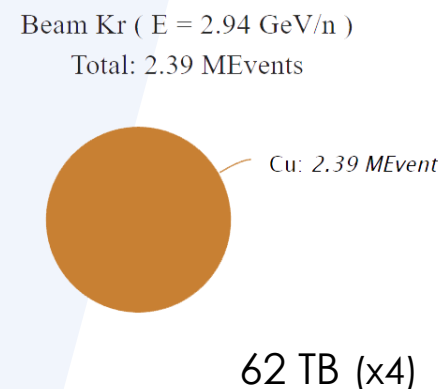
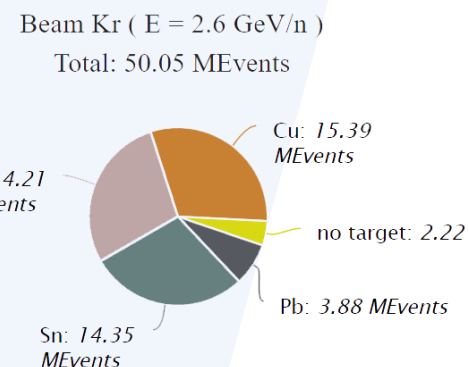
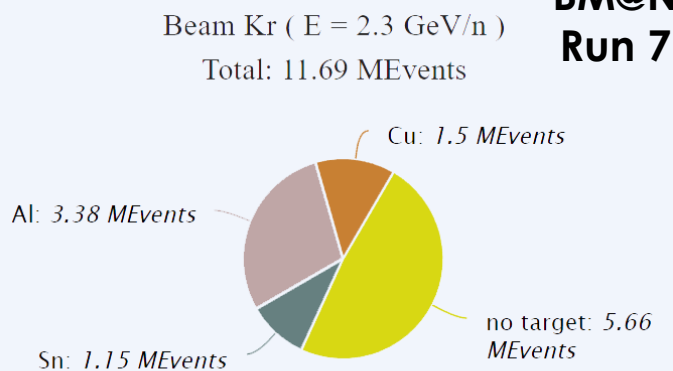
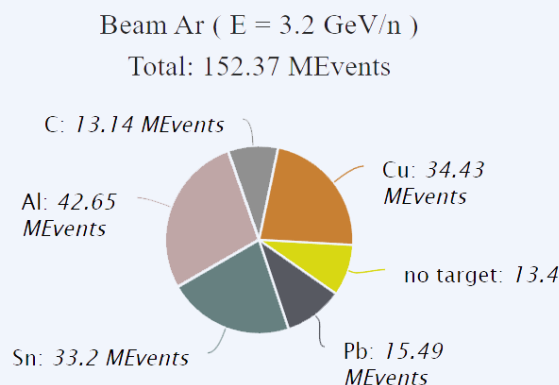
## 1<sup>st</sup> Physics BM@N Run

Two beam energy available for Xe-beam  
CsI target is used as more similar to Xe  
More than 600M events were collected



## Technical BM@N Run 7

One beam energy available for Ar-beam and  
three for Kr-beam  
Wide set of targets used: (C, Al, Cu, Sn, Pb)



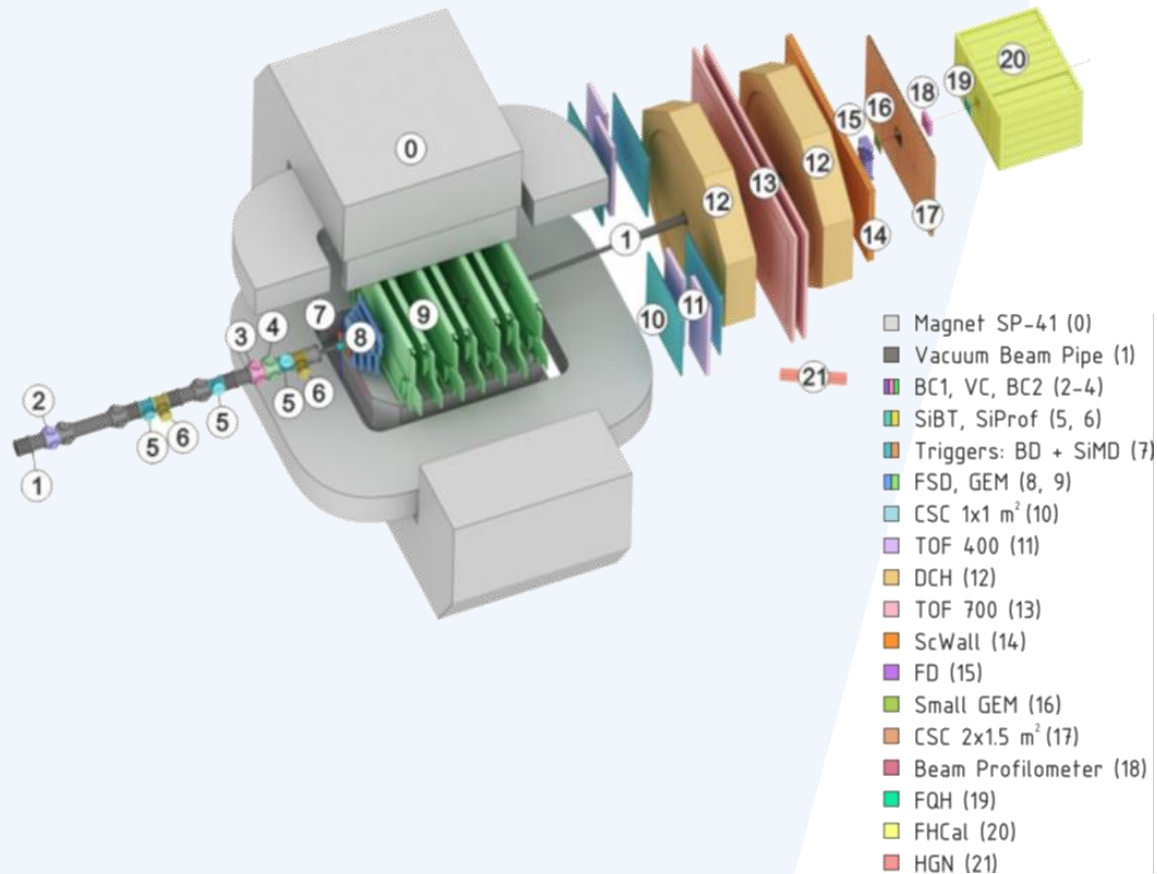
# BmnRoot Framework

The **BmnRoot** framework is developed for realistic event simulation, reconstruction of experimental or simulated data and following physics analysis of ion collisions with a fixed target at the BM@N facility.

has been actively developing since 2014

C++ classes, Linux/macOS,

based on  ROOT and FairRoot



The BmnRoot software is available in GitLab@JINR: <https://git.jinr.ru/nica/bmnroot>

# BmnRoot. Event Data Model

## DAQ Storage

raw data in a binary format

**raw\_run.data**  
≈ 600 KB/event

RAW  
binary  
format

RAW  
ROOT  
format

digitizer

BmnDataToRoot.C  
converter + decoder

**digi\_exp.root**  
≈ 35 KB/event

DIGIT  
ROOT  
format

Storage Levels

**persistent**

transient

reconstruction

run\_reco\_bmn.C

**dst\_reco.root**  
≈ 90 KB/event

physics  
analysis

macro/physics/

Geant 4, Fluka

simulation

run\_sim\_bmn.C

**digi\_sim.root**

SIM  
ROOT  
format

DST  
ROOT  
format

miniDST  
for PhA

## Event Generators

(DCM-)SMM, QGSM, UrQMD...

**generator.dat**  
≈ 10 KB/event

GEN  
format

**RAW** → **DIGIT** → **DSTexp** → PhA

**RAW**: raw (binary) event data collected by the  
DAQ system after the Event Builder

**DIGIT**: detector readings (event digits) after the  
digitizer macro

**DSTexp**: reconstructed data of experimental events

**GEN** → SIM → **DSTsim** → PhA

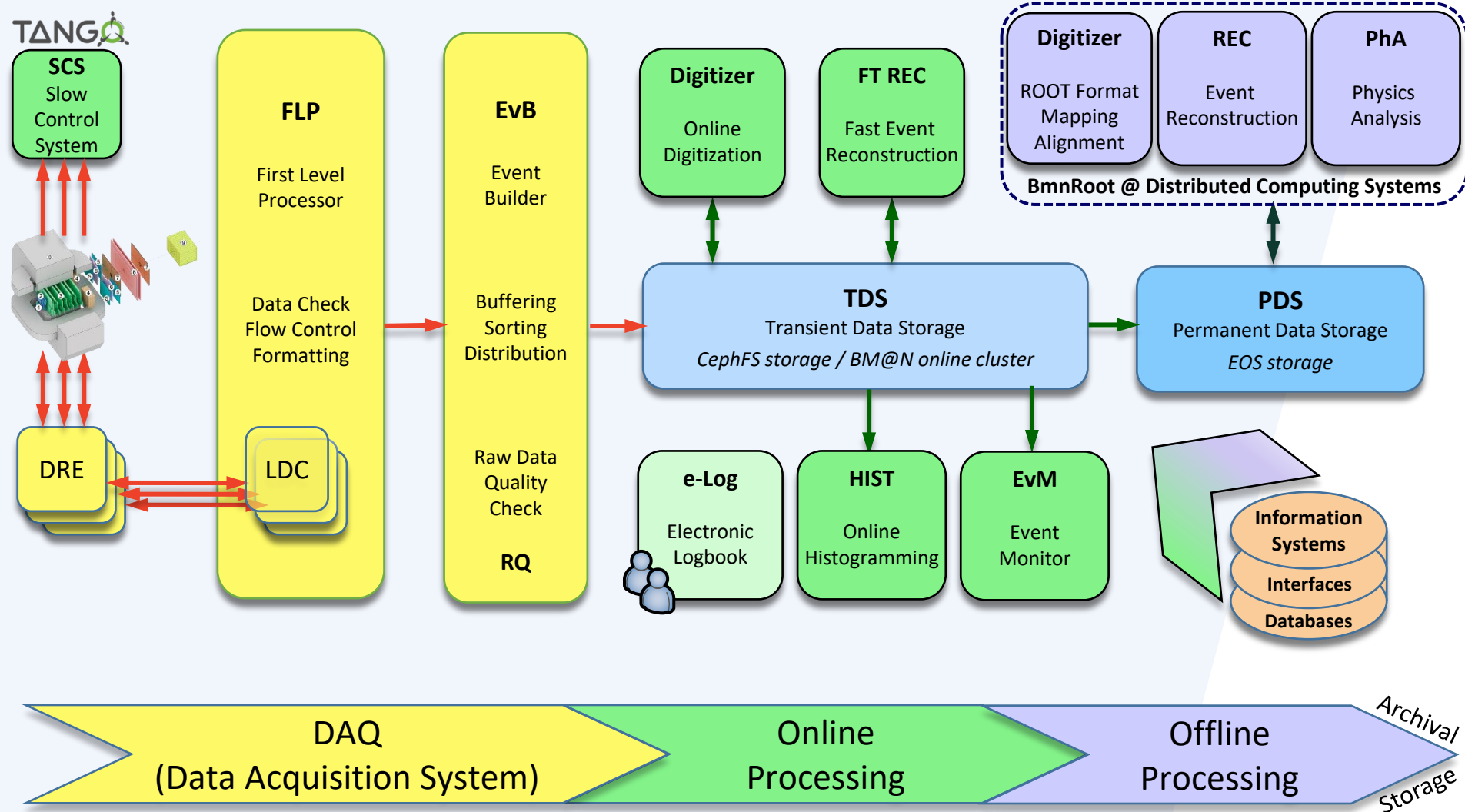
**GEN**: particle collisions description received by  
an event generator

**DSTsim**: reconstructed data of simulated events





# BM@N Data Processing Flow



# Prerequisites of BM@N distributed computing

- ✦ high interaction (trigger) rate up to **15 kHz**
- ✦ high particle multiplicity up to **hundreds of reconstructible particles** for the fixed target collisions at the BM@N energies
- ✦ large BM@N data stream:
  - is estimated up to **10 PB** of raw data per year
  - 500m simulated events ~ 0.5 PB
- ✦ long sequential event digitizing and reconstruction of hundreds of millions of events takes **decades**
- ✦ NICA computing platforms can be used to successfully process BM@N events concurrently

# Components of BM@N distributed complex

- ❖ **computing platforms** for the BM@N experiment
- ❖ **data storages** on distributed FS for experimental and simulated files
- ❖ **software distribution system** as a central repository of the experiment software
- ❖ **workload management system** for parallel task/job distribution
- ❖ **file and event catalogues** organizing smart namespaces with metadata
- ❖ **data transfer services** enabling the transfer of large amounts of data between users and storages within the federal administration
- ❖ **workflow management service** orchestrating task flows on data processing
- ❖ **information systems** based on databases providing necessary information for offline and online processing
- ❖ **user interfaces** (Web, API, CLI) to manage databases and distributed data processing
- ❖ central **authentication and authorization system** to regulate access rights
- ❖ **monitoring system** to control state of server nodes, databases and interfaces



# Computing Platforms for BM@N

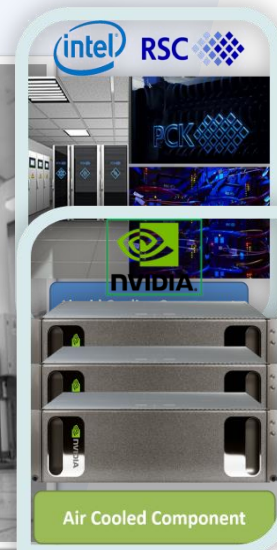
NICA Cluster  
*ncx[101-106].jinr.ru*  
(LHEP, b.216)



GRID Tier1&2 Centres  
*lxui.jinr.ru* (CICC)  
(MLIT, b.134)



HybriLIT platform (SC «Govorun»)  
*hydra.jinr.ru*  
(MLIT, b.134)



OS: CentOS / Scientific Linux 7.9

Central Software Repository based on **CVMFS** for the experiments

**EOS: 1 PB** (*replicated*)  
**GlusterFS: 300 TB** (*for NICA*)  
**Sun Grid Engine: 300** cores/user

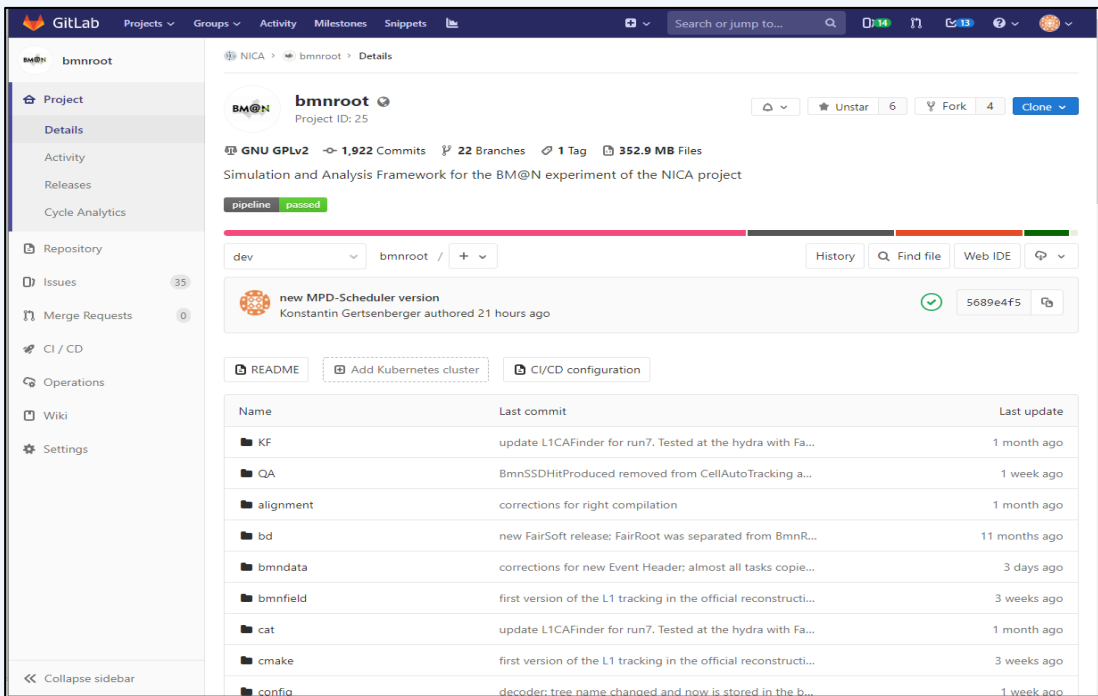
**EOS: 1 PB** (*replicated*)  
**SLURM: 0 – 2500** cores  
(*for NICA*)

**ZFS: 200 TB**  
**Lustre (Hot Storage): 300 TB<sub>ssd</sub>** (*for NICA*)  
**SLURM: bmn – 192** cores

**BM@N software have been installed & configured on JINR CVMFS**

**Automatic software deployment of the BmnRoot package on CVMFS with GIT CI**

# BmnRoot. Automatic Deployment to CVMFS



## GIT: Version Control System

Repository branch protection

Role-based access control to projects

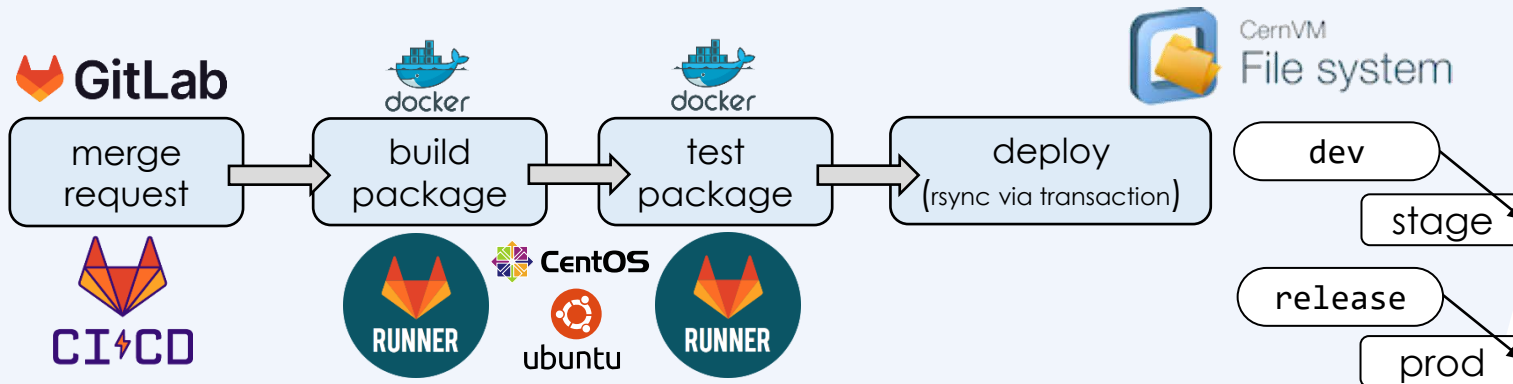
Issue Tracker

Automated Tests & Deployment

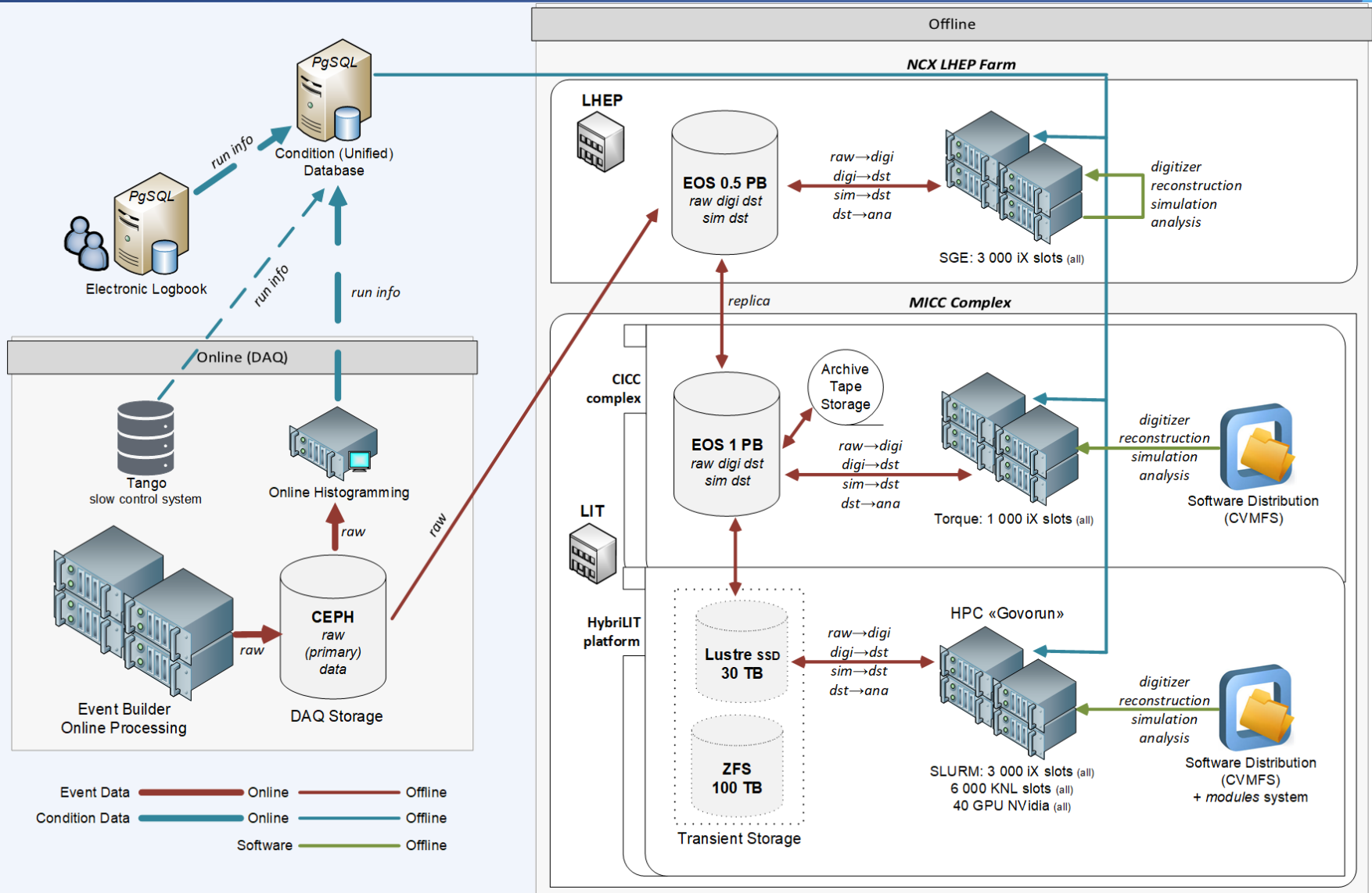
## Software Distribution via

## CernVM File System

Read-only network file system  
with aggressive caching, optimized  
for software distribution via HTTP  
in a fast, scalable and reliable way

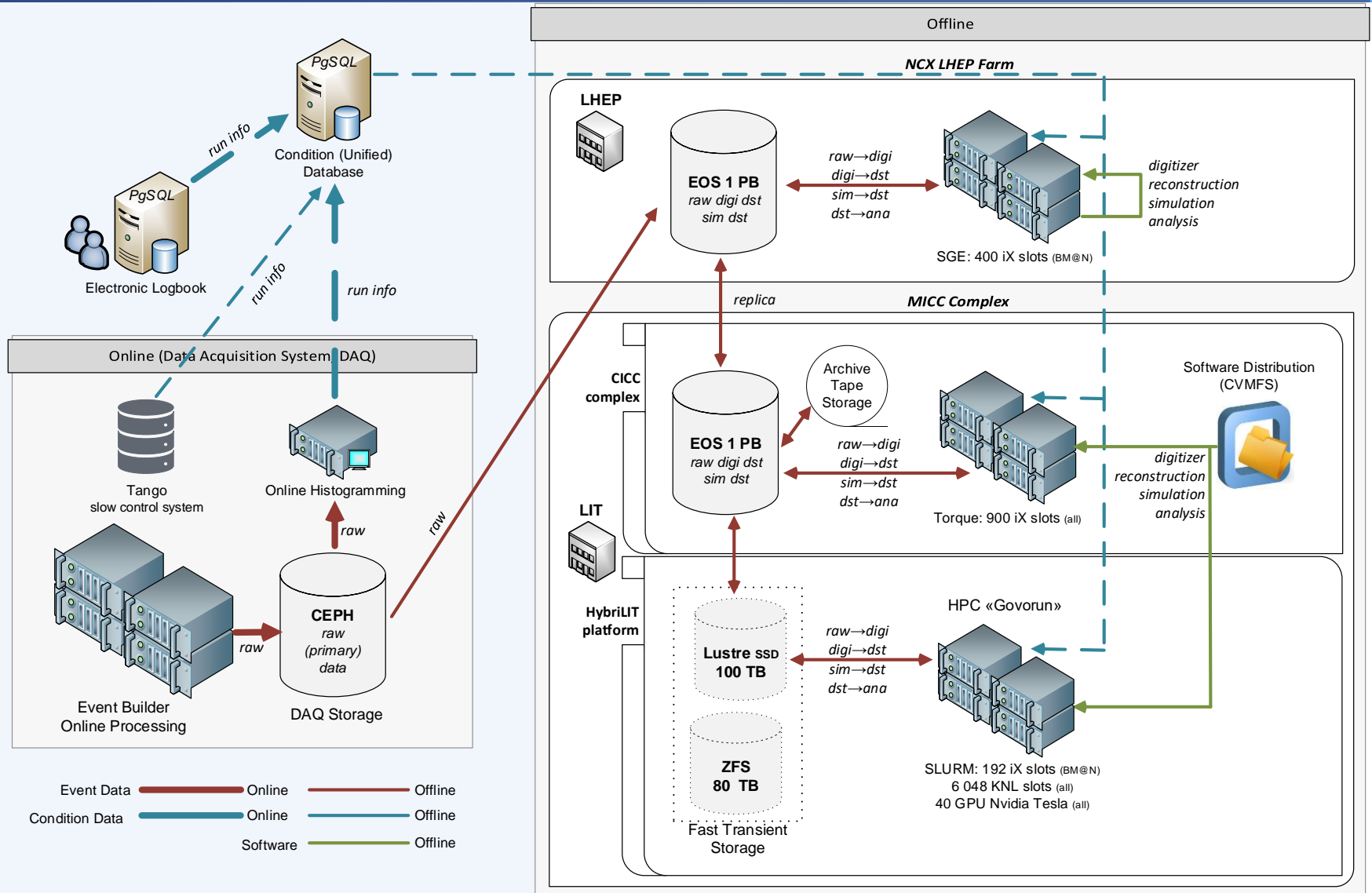


# BM@N WorkFlow. Run 7 Status (2018)

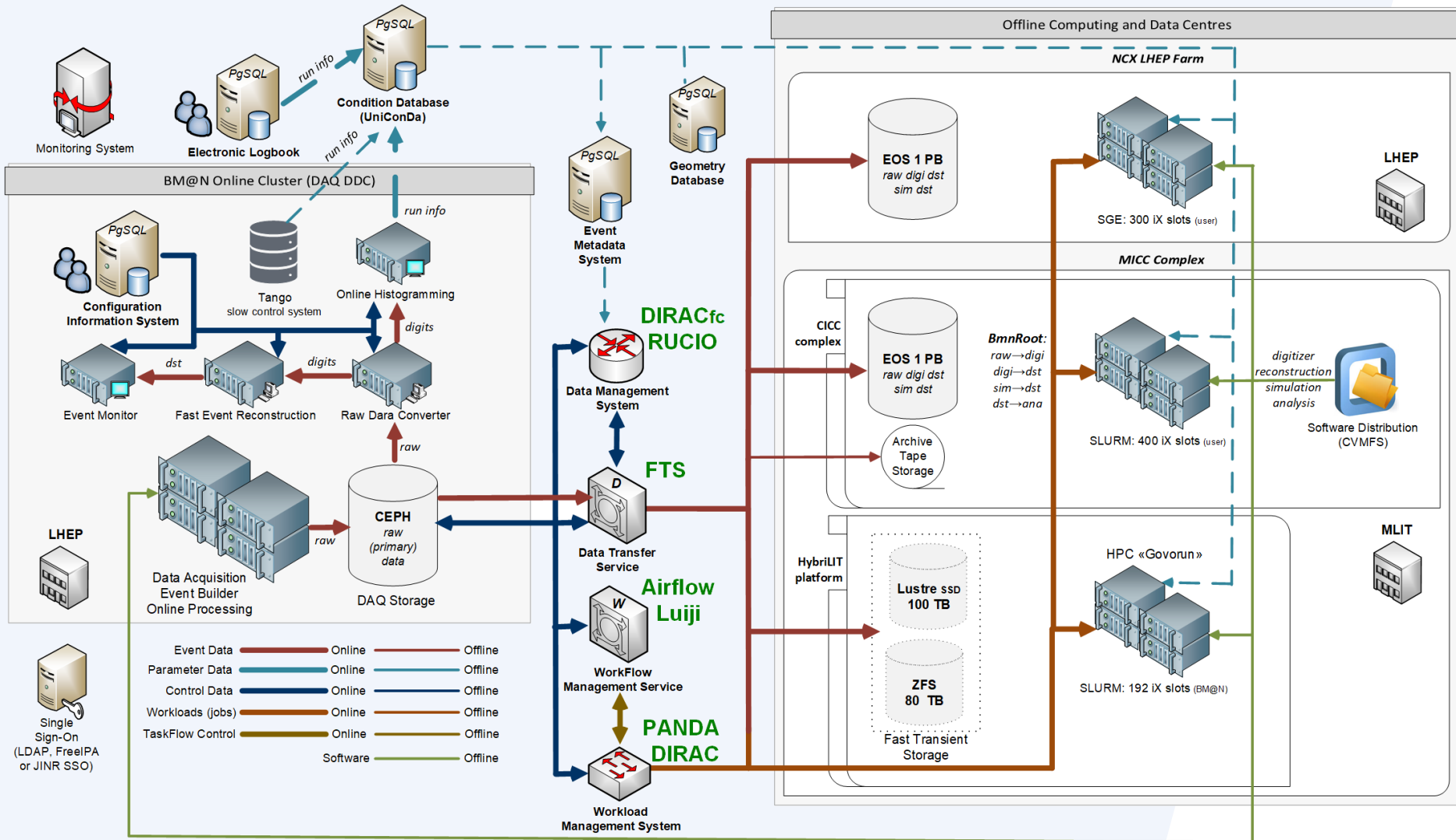




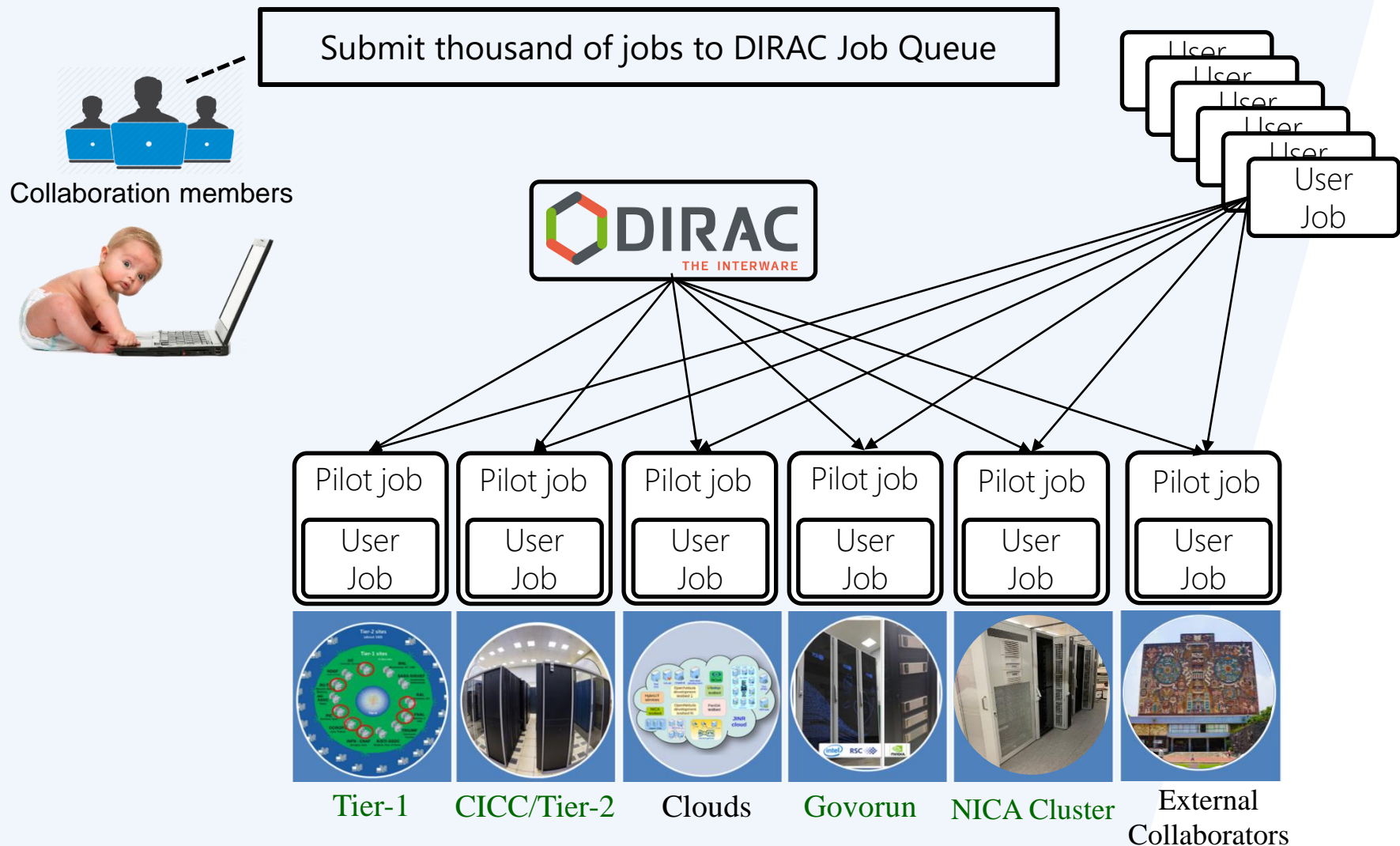
# BM@N WorkFlow. Status 2020



# BM@N Computing Software Architecture



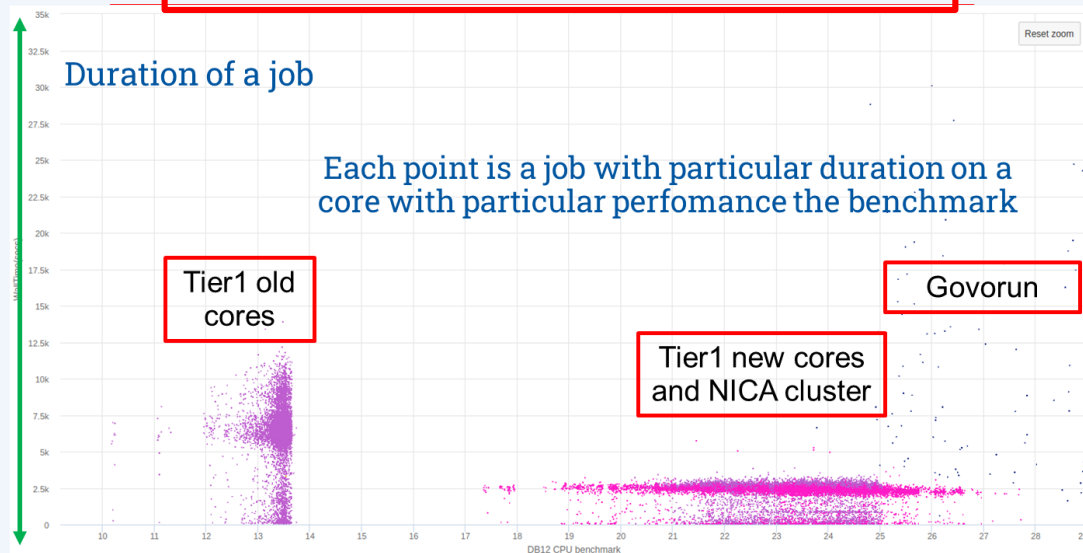
# DIRAC Workload Manager for BM@N





# BM@N Mass Production via DIRAC (Run 8)

Total duration of Raw2Digi campaign – 18 hours



CPU core performance on benchmarks

Total files: **30 741**    Total raw size: **393 TB**  
Average transfer speed (20 streams): **1.92 GB/s**  
Total transfer duration: **2d 15h**  
Max transfer speed (R+W) EOS@MLIT: **7.5 GB/s**

Disk usage: tmp file: **8 GB** result file: **800 MB**  
Total disk usage per job (15 GB): **25 GB**  
RAM usage: **2 GB**

Total wall time: **10 years**

Quotas (cores):

Tier1: 1500 (for NICA)

Tier2: 1000 (for NICA)

Govorun: 192 (BM@N)

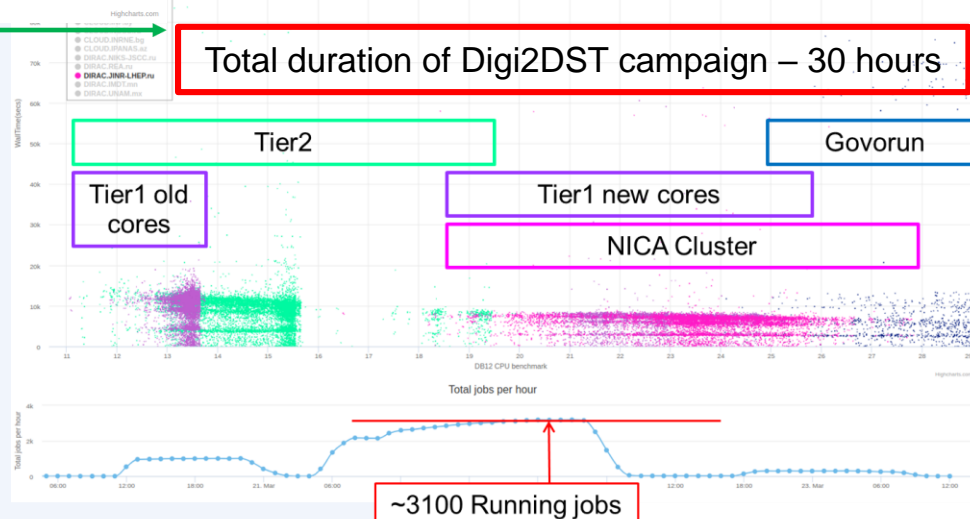
NICA cluster: 300 (per user)

NICA cluster: max 100 slots (**big files**)

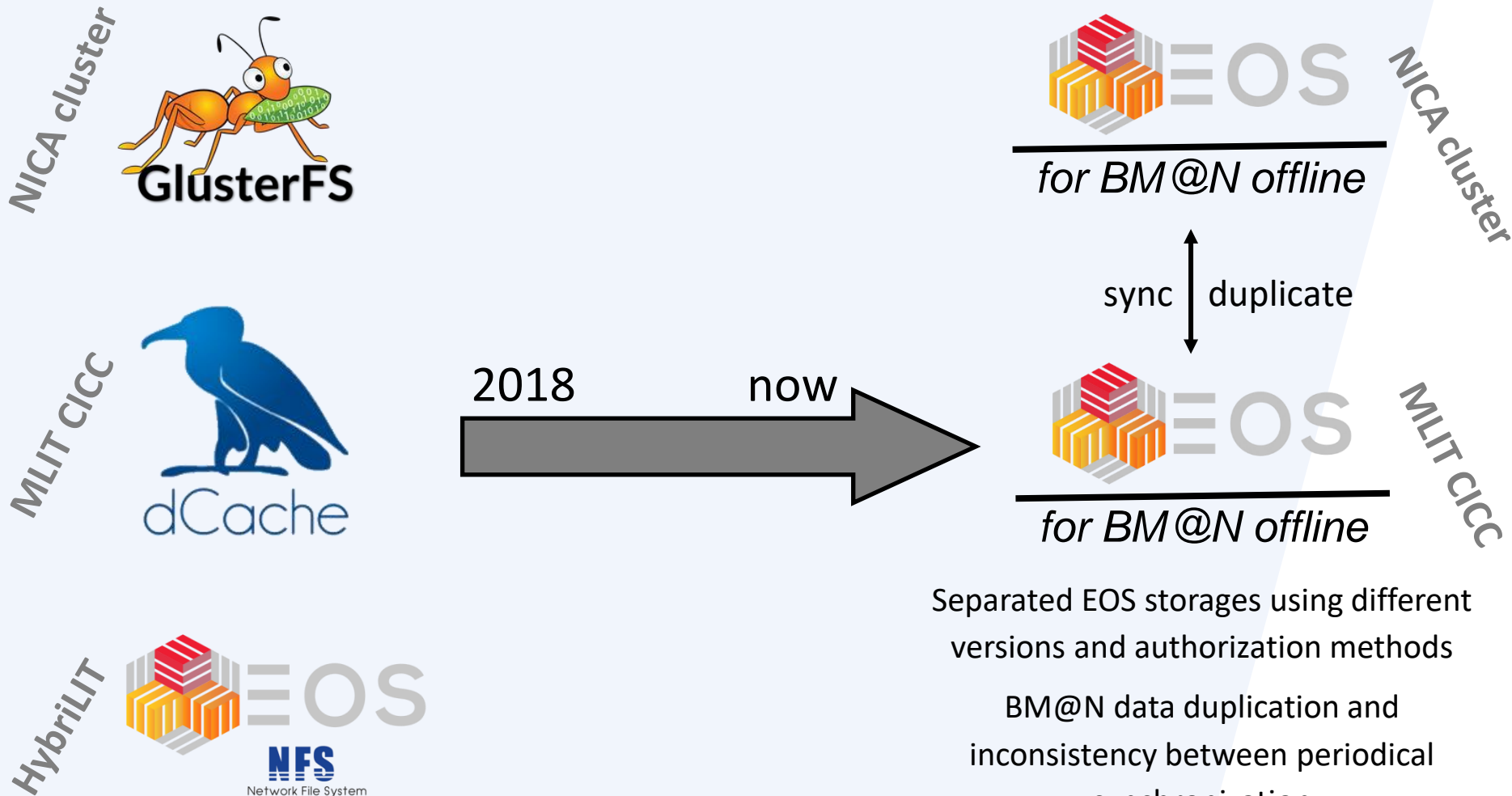
Igor Pelevanyuk  
(4 July 15:00)

BM@N Run 8 data reconstruction on  
a distributed infrastructure with DIRAC



Total duration of Digi2DST campaign – 30 hours



# Cold Data Storages for the BM@N experiment



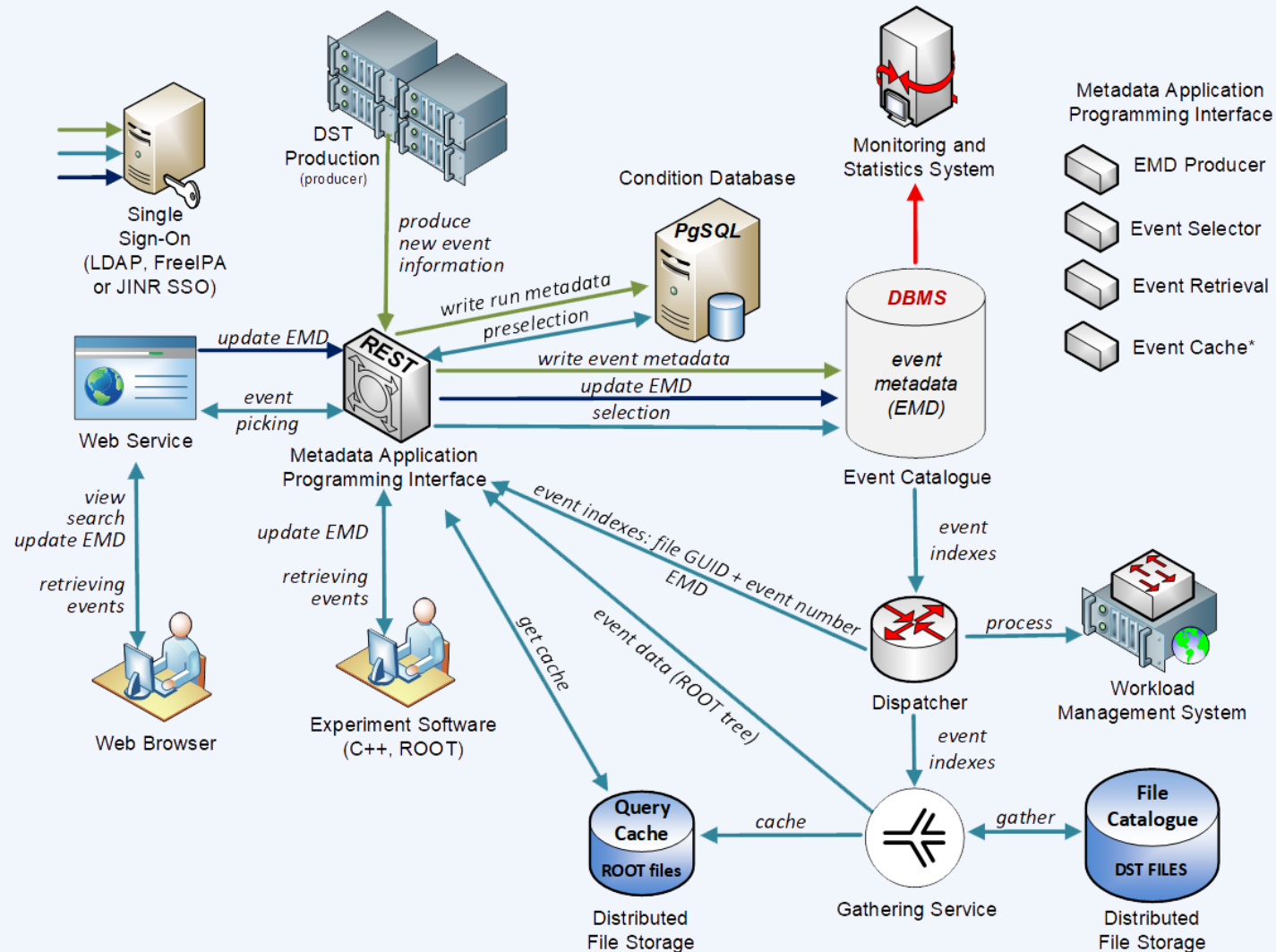
# File Catalogue Choice for BM@N

- File Catalogues map a Logical File Name (LFN) to the Physical File Name (PFN) at distributed computing platforms
- The native  File Catalog (DFC) combines both replica and metadata functionality. In the DFC metadata can be associated with any directory, and subdirectories inherit the metadata of their parents
-  is a Distributed Data Management System initially developed for the ATLAS experiment in 2014 providing file and dataset catalogue and transfers between sites and staging capabilities, policy engines, caching, bad file identification and recovery, and many other features.





# BM@N Event Catalogue



Event Catalogue is based on PostgreSQL

Integrated with the Condition Database

REST API and Web UI developed on Kotlin multiplatform

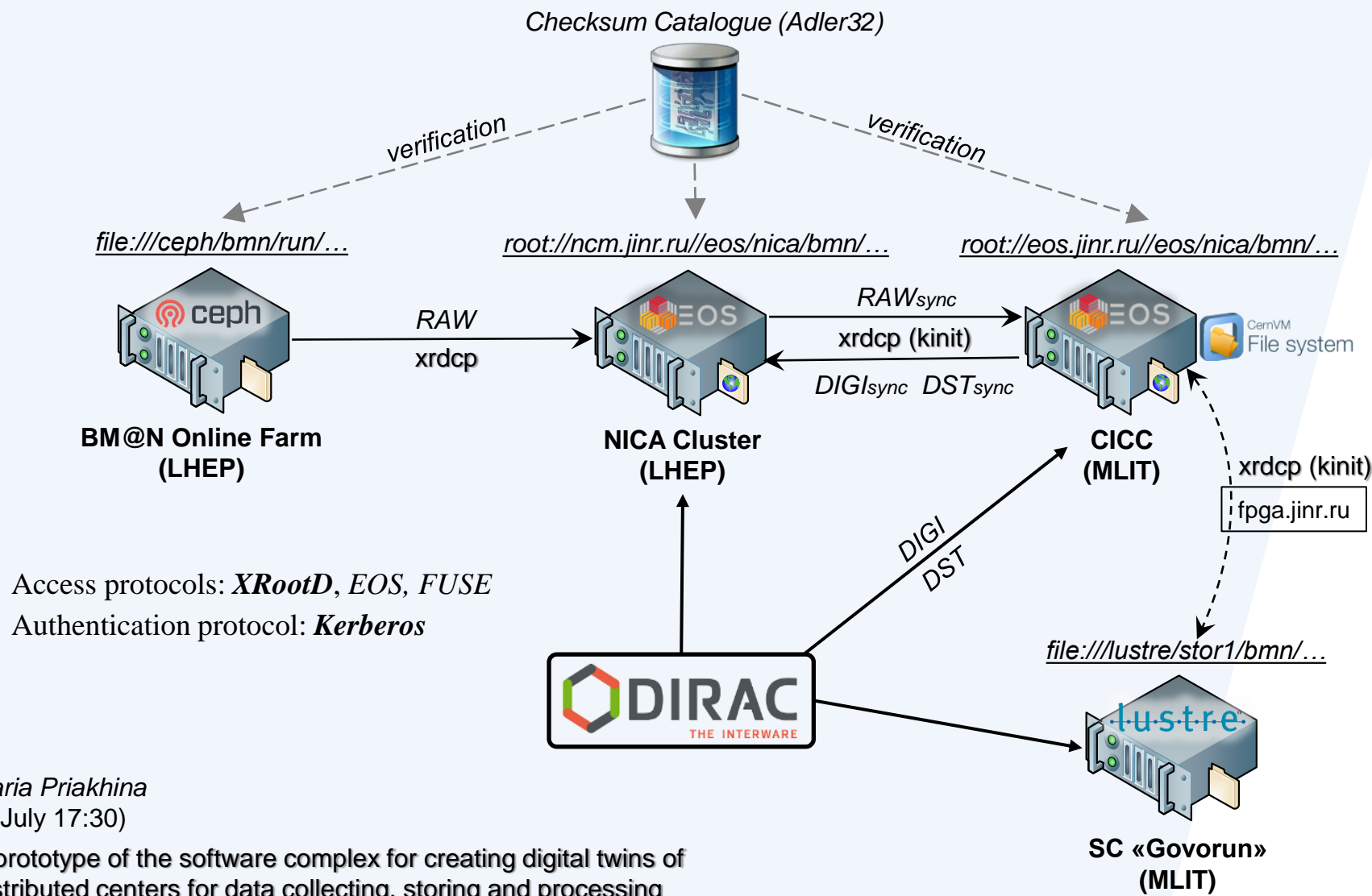
Configurable to support any metadata

ROOT macro to write new event metadata to the Catalogue

Role-based access control

Monitoring

# Current BM@N Data Transfer



Daria Priakhina  
 (4 July 17:30)

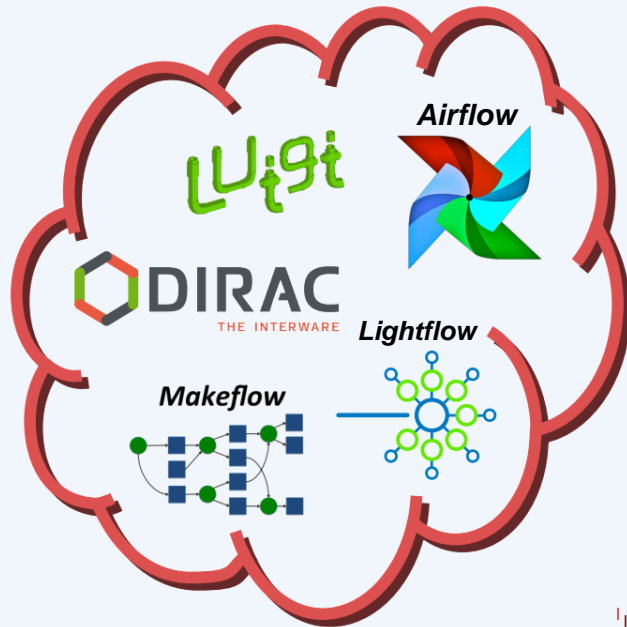
A prototype of the software complex for creating digital twins of distributed centers for data collecting, storing and processing

# Data Transfer Service for the NICA experiments



- Selected and validated by ATLAS, CMS & LHCb: Rucio and DIRAC run on top
- Transfer scheduling with real-time optimisation
- Web-based and messaging based monitoring
- Clients – CLI, REST (e.g. curl), python
- Protocol support: SRM, GridFTP, WEBDAV/HTTP(S), XRootD.
- Horizontally scalable multi-threaded server with “zero config”
- WebFTS portal for simplifying user's experience

# Workflow Management Services



Ability to define dependencies between different tasks and scheduling them is a key

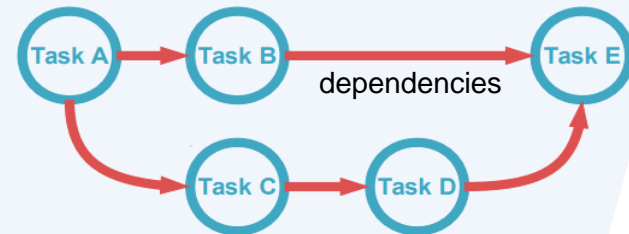
Gaining popularity as businesses increasingly have many complex Extract, Transform, Load (ETL) operations to schedule e.g. search, indexing/ranking, monitoring, statistics creation.



<https://airflow.apache.org>

Platform to author, schedule and monitor workflows (as code)

Python scripting to define Directed Acyclic Graphs (DAGs) as collections of tasks (Operators) with dependencies between them



DAGs can be given a regular schedule, triggered manually, or even trigger each other

Includes Flask-based web monitoring to visualize pipelines, monitor progress and troubleshoot issues

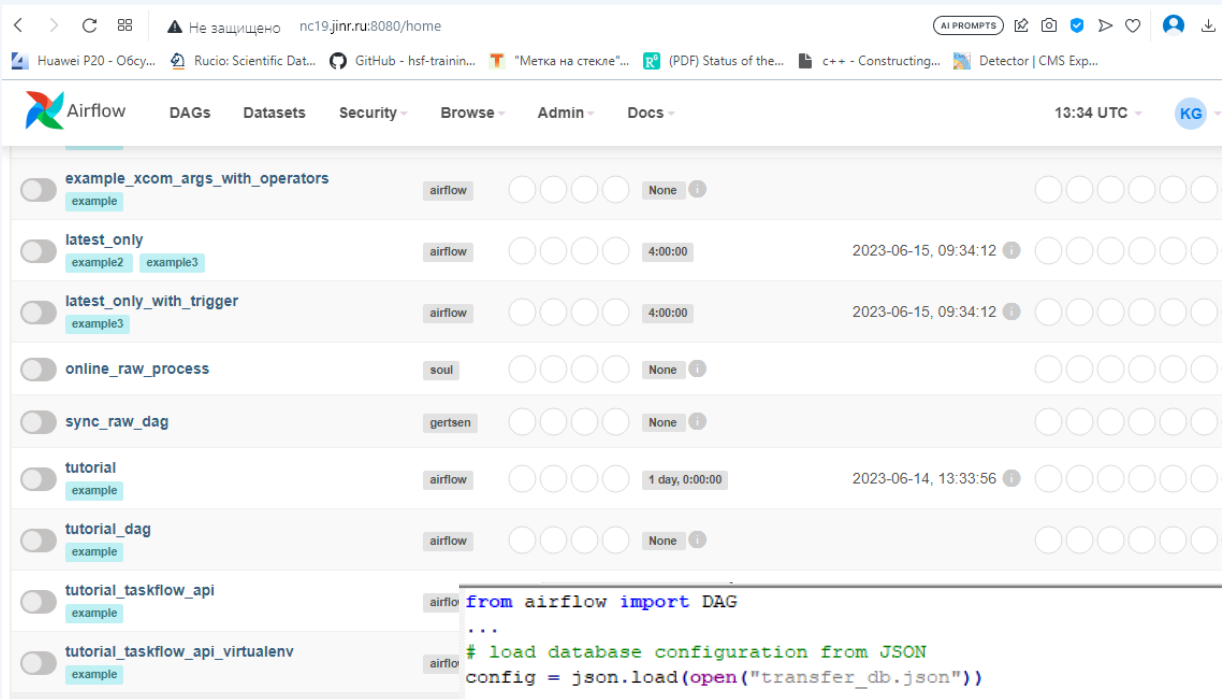
Uses Jinja2 templating and SQLAlchemy to write tasks that render templated strings e.g. in database query strings or bash scripts

Operators	Sensors	Hooks
Defines the task in a DAG Different operator types, e.g. Python, Bash, SQL	Operator that periodically executes a query Won't complete until conditions are met	Defines the interface to some external system Retrieves authentication stored in Airflow DB

\*D. Dossett, M. Sevir. "Automating Calibration at the Belle II Detector"



# First steps in BM@N Workflow Management



The screenshot shows the Apache Airflow web interface. The top navigation bar includes links for DAGs, Datasets, Security, Browse, Admin, and Docs. The main content area displays a list of DAGs with their names, owners, and execution status. The DAGs listed are:

- example\_xcom\_args\_with\_operators (owner: airflow, status: None)
- latest\_only (owner: airflow, status: 4:00:00, last run: 2023-06-15, 09:34:12)
- latest\_only\_with\_trigger (owner: airflow, status: 4:00:00, last run: 2023-06-15, 09:34:12)
- online\_raw\_process (owner: soul, status: None)
- sync\_raw\_dag (owner: gertsen, status: None)
- tutorial (owner: airflow, status: 1 day, 0:00:00, last run: 2023-06-14, 13:33:56)
- tutorial\_dag (owner: airflow, status: None)
- tutorial\_taskflow\_api (owner: airflow, status: None)
- tutorial\_taskflow\_api\_virtualenv (owner: airflow, status: None)

Airflow **deployed** on the NC-farm

Used in BM@N Run 8 to **transfer raw data** emerging on the NICA-cluster to the LIT EOS storage and **to check the integrity** of the source and destination files

To be employed for **managing online** (for emerging raw data files) **and offline data production** via DIRAC



MC simulation pipeline  
event filtering      digitizing  
reconstruction      analysis

...

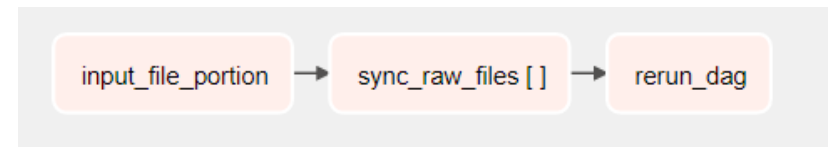
```
from airflow import DAG
...
# load database configuration from JSON
config = json.load(open("transfer_db.json"))
...
with DAG('sync_raw_dag', description='This DAG is for copying new raw data files from an input directory to LIT EOS',
        default_args=default_args, schedule_interval=None, catchup=False, max_active_runs=1) as dag:

    @task
    def input_file_portion():
        ...
        return process_list

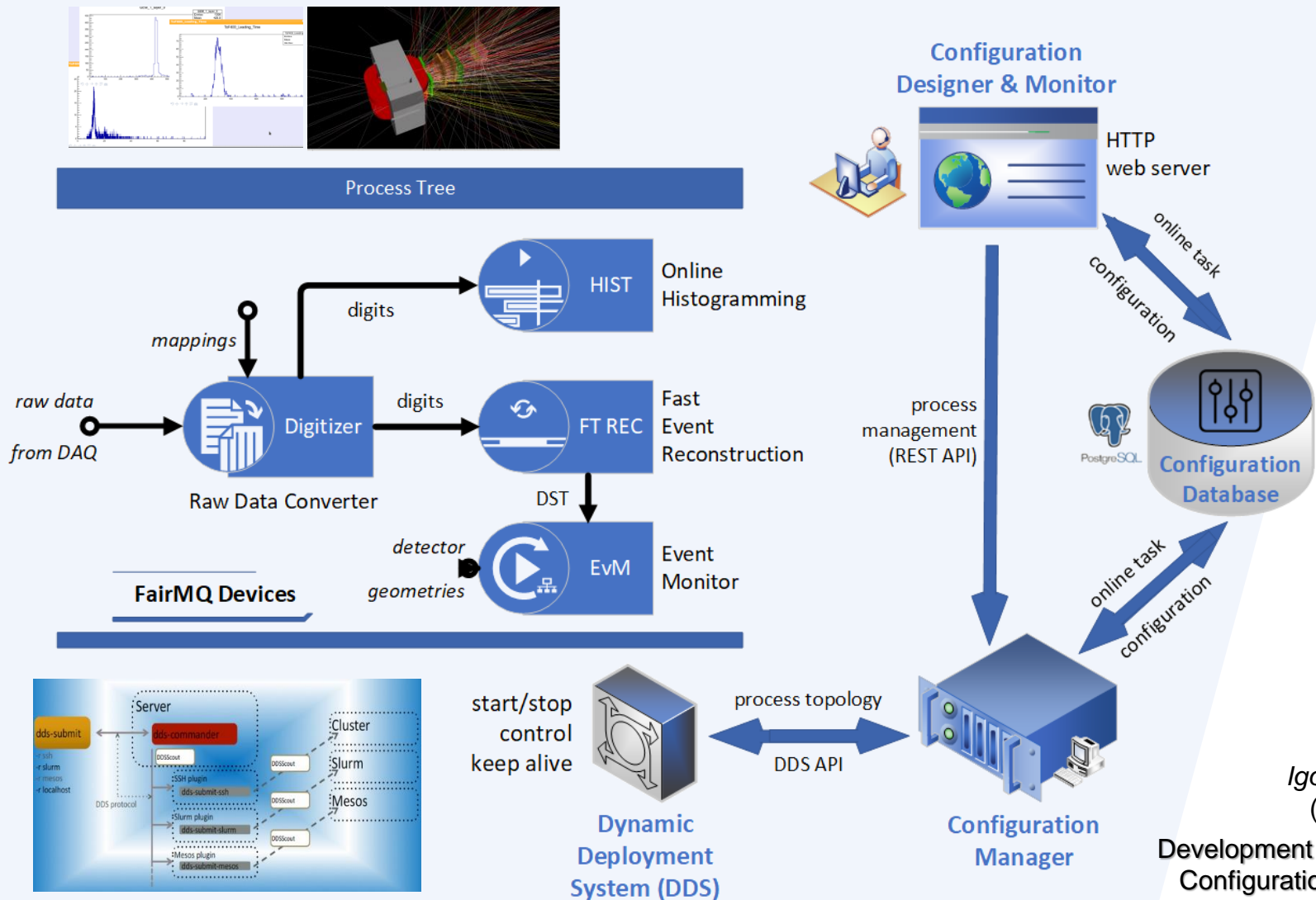
    @task(max_active_tis_per_dag=8)
    def sync_raw_files(input_file_path):
        ...

    trigger = TriggerDagRunOperator(task_id='rerun_dag',
                                    trigger_dag_id="sync_raw_dag")

    sync_raw_files.expand(input_file_path=input_file_portion()) >> trigger
```



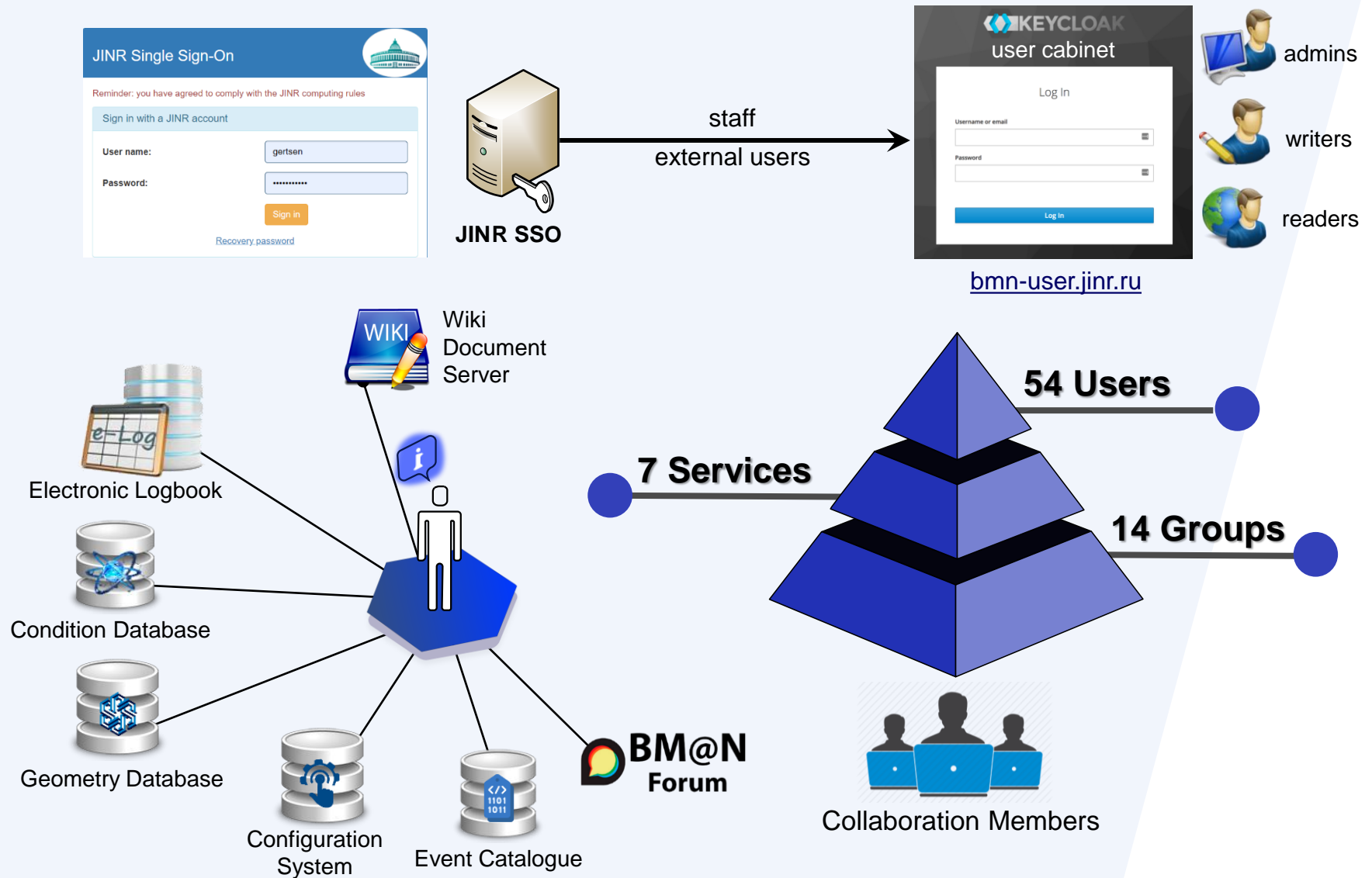
# Online Distributed Processing (OCS)



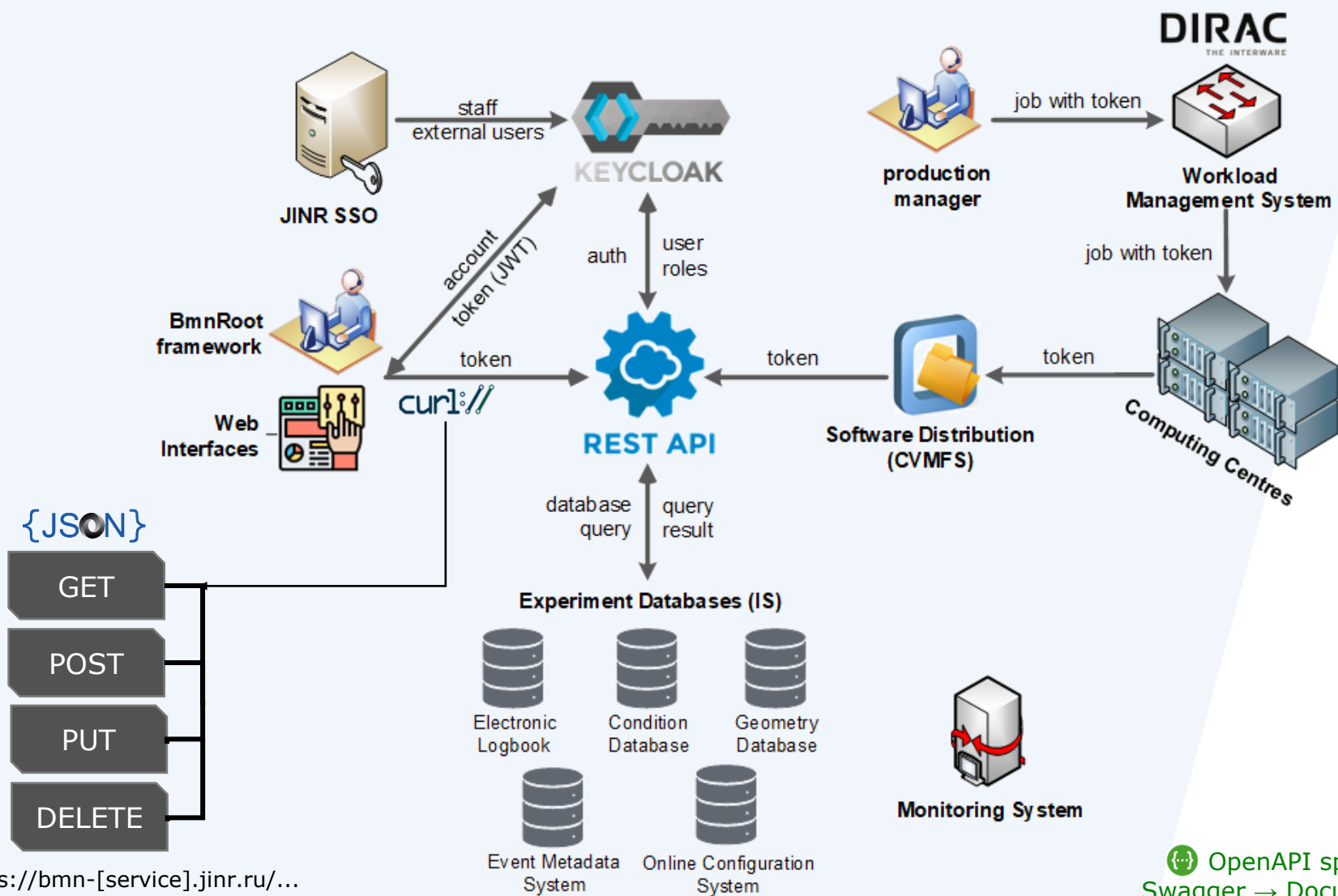
Igor Alexandrov  
(4 July 15:15)


Development of the Online  
Configuration System for  
the BM@N experiment

# Migration from FreeIPA to JINR Single Sign-On



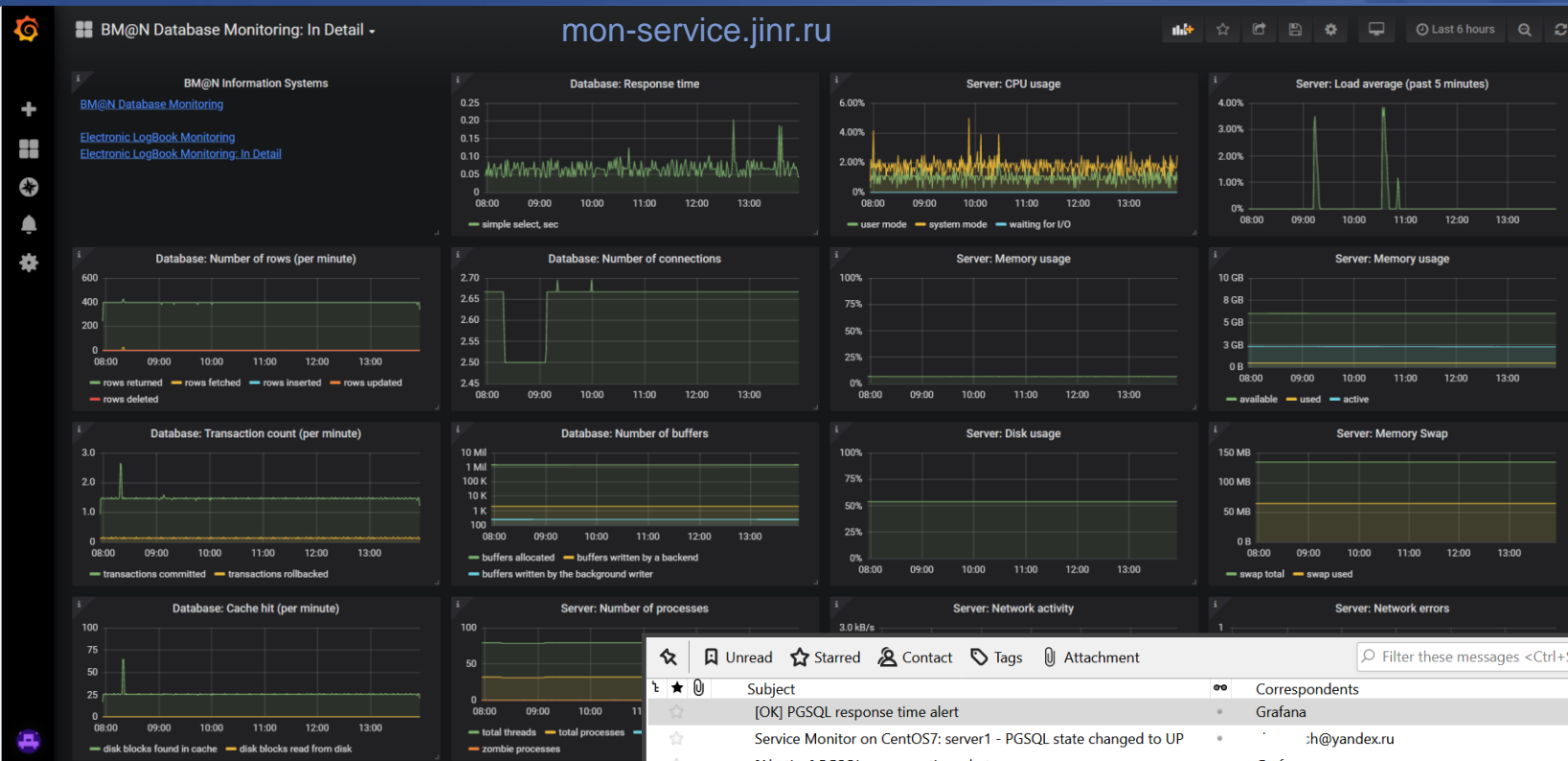
# REST APIs for BM@N Information Systems



 OpenAPI specification  
Swagger → Documentation



# Monitoring System for BM@N software complex



- hosts
- databases
- web-sites

- **Condition Database**  
simple or detailed visualization

- **Electronic Logbook**  
simple or detailed visualization

...

*Email Alerting*

The email alert interface shows a list of messages with columns for Subject, Correspondents, and Date. The selected message is:

From: Grafana <h@yandex.ru>  
Subject: [OK] PGSQL response time alert  
To: Me

**[OK] PGSQL response time alert**

Grafana: Database monitoring warning!

0.12

# Conclusions

- More than **600 millions of collision event** ( $\approx 400$  TB) were obtained in the first BM@N physics run, and it is expected that the amount of data will be increased by an order of magnitude in future runs.
- BM@N data mass production is being successfully performed via the **DIRAC** workload management system. Integration of **DFC and Airflow** is in progress.
- Many **information and software systems** of the experiment have been adapted and now are being involved in BM@N mass production for Run 8 to reduce the time of obtaining physics results. The Monitoring System has been implemented to track and visualize their states, and send notifications in case of any malfunction.
- Migration from existing FreeIPA (LDAP) single authentication/authorization system of the BM@N experiment to the Keycloak system using **JINR SSO** accounts is in progress now.
- A lot of efforts have been invested to implement the designed **BM@N software – computing architecture**, but a set of necessary services still to be developed or completed for the full automation of the BM@N distributed data processing.

*New participants are welcomed.*

# Thank you for your attention!



**Director: S. V. SHMATOV. Scientific Leader: V. V. KORENKOV**

JINR MLIT  
Contribution  
to BM@N

***BM@N is open for  
cooperation and  
young people!***

*Igor ALEXANDROV, Evgeniy ALEXANDROV, Irina FILOZOVA, et alia*

***Development of the Geometry Database and Online Configuration Systems***

*Nikita BALASHOV:*

***CVMFS Deployment, GitLab Services, Docker Containers***

*Igor PELEVANYUK:*

***DIRAC workload management system and BM@N mass production***

*Dmitriy PODGAYNY, Oksana STRELTSOVA, Maksim ZUEV*

***HybriLIT and SC Govorun support***

*Daria PRIAKHINA, Vladimir TROFIMOV*

***Modelling System for BM@N computing infrastructure***

*Zarif SHARIPOV, Zafar TUKHLIEV*

***Automation of BM@N Alignment***



thanks to the DDC,  
CICC, NCX &  
HybriLIT teams for  
computing support

contact email: [gertsen@jinr.ru](mailto:gertsen@jinr.ru)