JINR Tier1 Center Status and Plans

Andrey Baginyan, Anton Balandin, Sergey Belov,
Andrey Dolbilov, Alexey Golunov, Natalia Gromova,
Ivan Kadochnikov, Ivan Kashunin, Vladimir Korenkov, Valery Mitsyn,
Igor Pelevanyuk, Sergei Shmatov, Tatiana Strizh, Vladimir Trofimov,
Nikolay Voytishin, Victor Zhiltsov

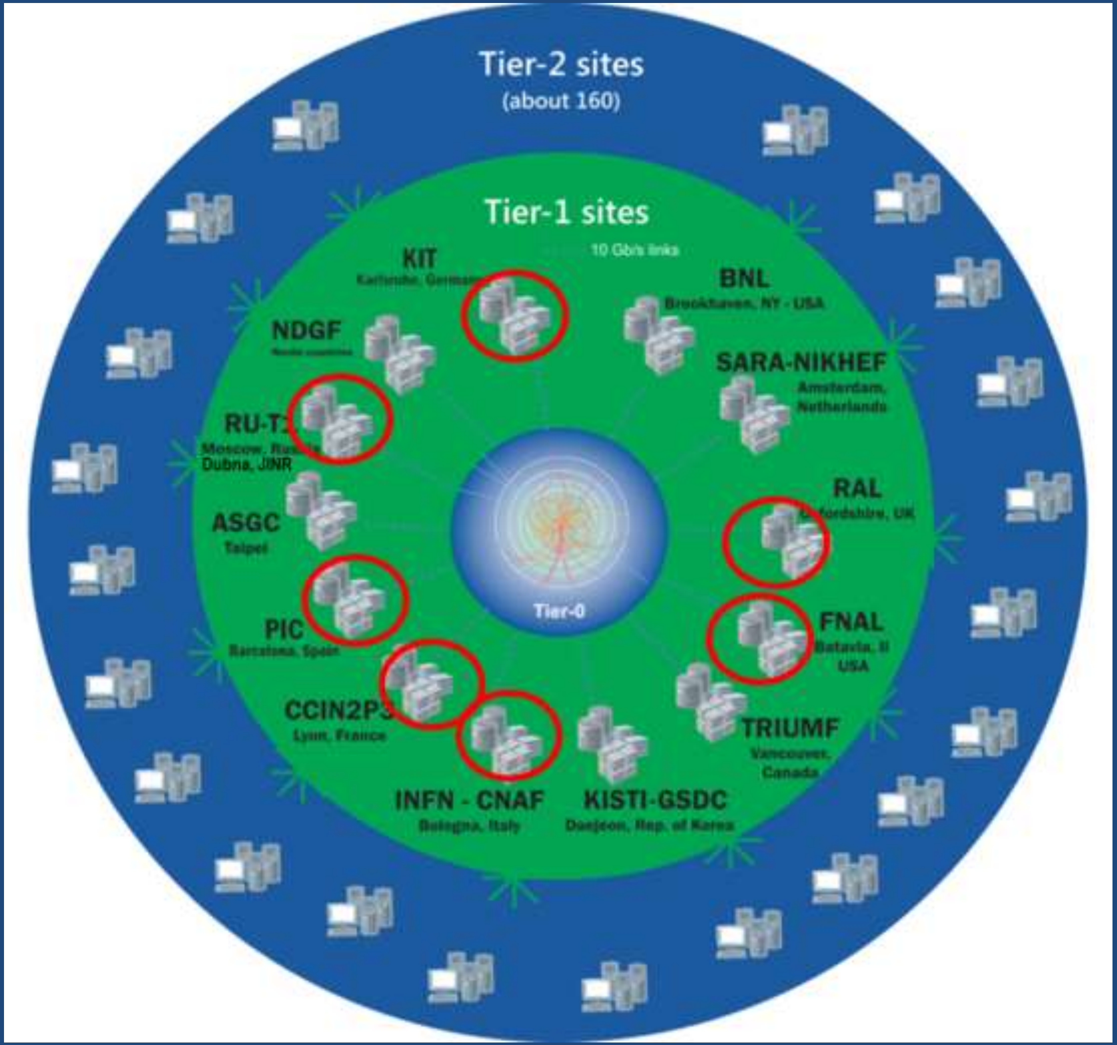
ROLCG-2018
Cluj
18-10-18

Outline

- How we started
- Main functions
- Infrastructure
- Network and telecommunication channels
- Resources
- Monitoring
- How well does it work?
- Plans for 2019

LHC Computing Model + WLCG

WLCG computing enabled physicists to announce the discovery of the Higgs Boson on 4 July 2012



42 countries
170 computing centers
2 million tasks run every day
800,000 computer cores
500 petabytes on disk and
400 petabytes on tape

Tier-0 (CERN):
Data recording
Initial data reconstruction
Data distribution

Tier-1 (14 centers):
Permanent storage
Re-processing
Analysis
Simulation

Tier-2 :
Simulation
End-user analysis

Joint NRC "Kurchatov Institute" – JINR Tier1 Computing Centre

- Proposal to create the WLCG Tier1 center in Russia: March 2011, accepted in October 2012

- ***The Federal Target Programme Project:***

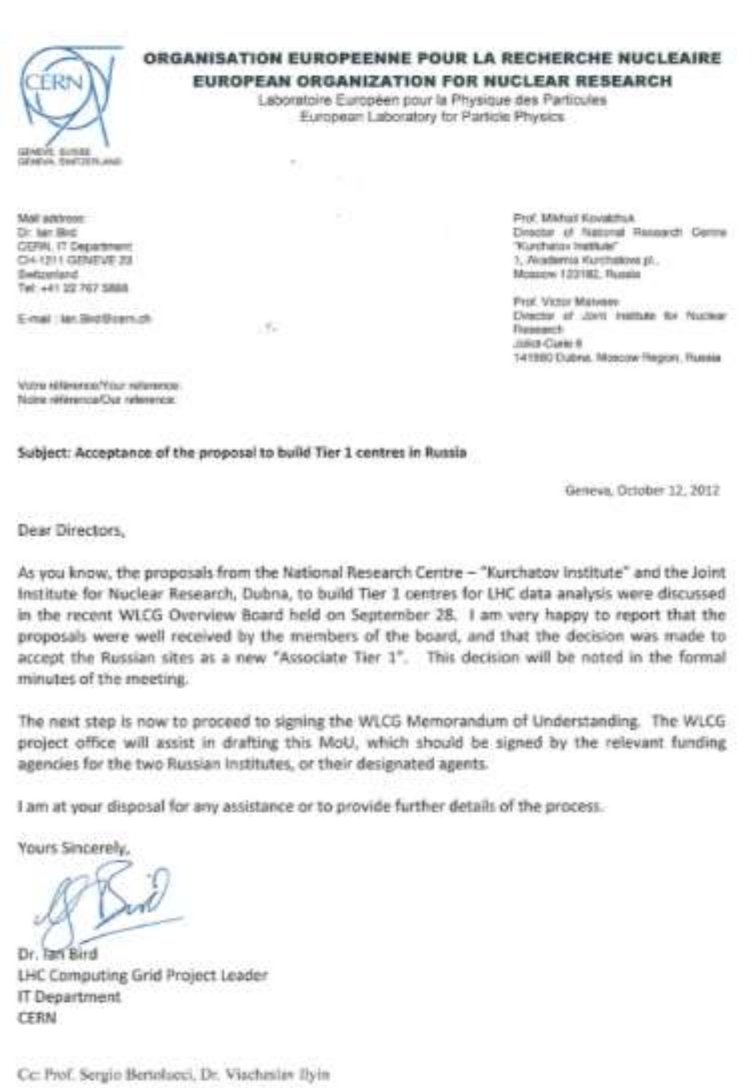
«Creation of the automated system of data processing for experiments at the LHC of Tier1 level and maintenance of Grid services for a distributed analysis of these data»

Duration: 2011 – 2013

Russia Tier1 full scope start-up in WLCG in 2014

NRC “Kurchatov Institute” supports ATLAS, ALICE and LHCb, JINR supports CMS (Compact Muon Solenoid)

Systematic increase of computing capacity and data storage is needed in accordance with the experiment requirements



CERN

GENEVE, SUISSE
GENEVA, SWITZERLAND

ORGANISATION EUROPEENNE POUR LA RECHERCHE NUCLEAIRE
EUROPEAN ORGANIZATION FOR NUCLEAR RESEARCH
Laboratoire Européen pour la Physique des Particules
European Laboratory for Particle Physics

Mail address:
Dr. Ian Bird
CERN, IT Department
CH-1211 GENEVE 23
Switzerland
Tel: +41 22 767 5888

E-mail : Ian.Bird@cern.ch

Prof. Mikhail Kovalchuk
Director of National Research Centre "Kurchatov Institute"
1, Akademia Kurchatova pl.,
Moscow 123182, Russia

Prof. Victor Matveev
Director of Joint Institute for Nuclear Research
Joliot-Curie 6
141980 Dubna, Moscow Region, Russia

Votre référence/Your reference:
Notre référence/Our reference:

Subject: Acceptance of the proposal to build Tier 1 centres in Russia

Geneva, October 12, 2012

Dear Directors,

As you know, the proposals from the National Research Centre – "Kurchatov Institute" and the Joint Institute for Nuclear Research, Dubna, to build Tier 1 centres for LHC data analysis were discussed in the recent WLCG Overview Board held on September 28. I am very happy to report that the proposals were well received by the members of the board, and that the decision was made to accept the Russian sites as a new "Associate Tier 1". This decision will be noted in the formal minutes of the meeting.

The next step is now to proceed to signing the WLCG Memorandum of Understanding. The WLCG project office will assist in drafting this MoU, which should be signed by the relevant funding agencies for the two Russian Institutes, or their designated agents.

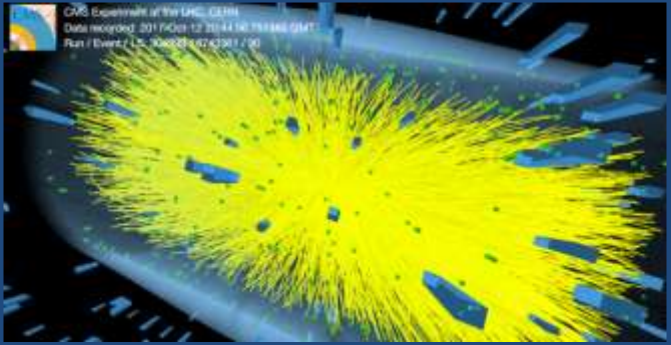
I am at your disposal for any assistance or to provide further details of the process.

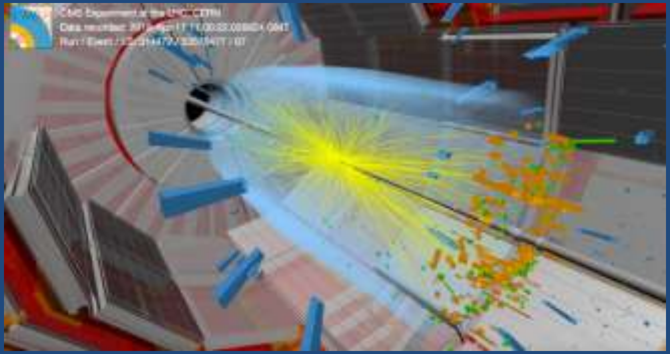
Yours Sincerely,

I Bird

Dr. Ian Bird
LHC Computing Grid Project Leader
IT Department
CERN

Cc: Prof. Sergio Bertolucci, Dr. Viacheslav Ilyin





In agreement with the CMS Computing model, the JINR Tier1 site provides:

- **acceptance of an agreed share of raw data and Monte Carlo data;**
- **provision of access to the stored data by other CMS Tier1/Tier2/Tier3 sites of the WLCG;**
- **service of FTS-channels for Russian and Dubna Member States Tier2 storage elements including monitoring of data transfers**

USER-VISIBLE SERVICES

- **Data Archiving Service**
- **Disk Storage Services**
- **Data Access Services**
- **Reconstruction Services**
- **Analysis Services**
- **User Services**

SOME SPECIALIZED SYSTEM-LEVEL SERVICES

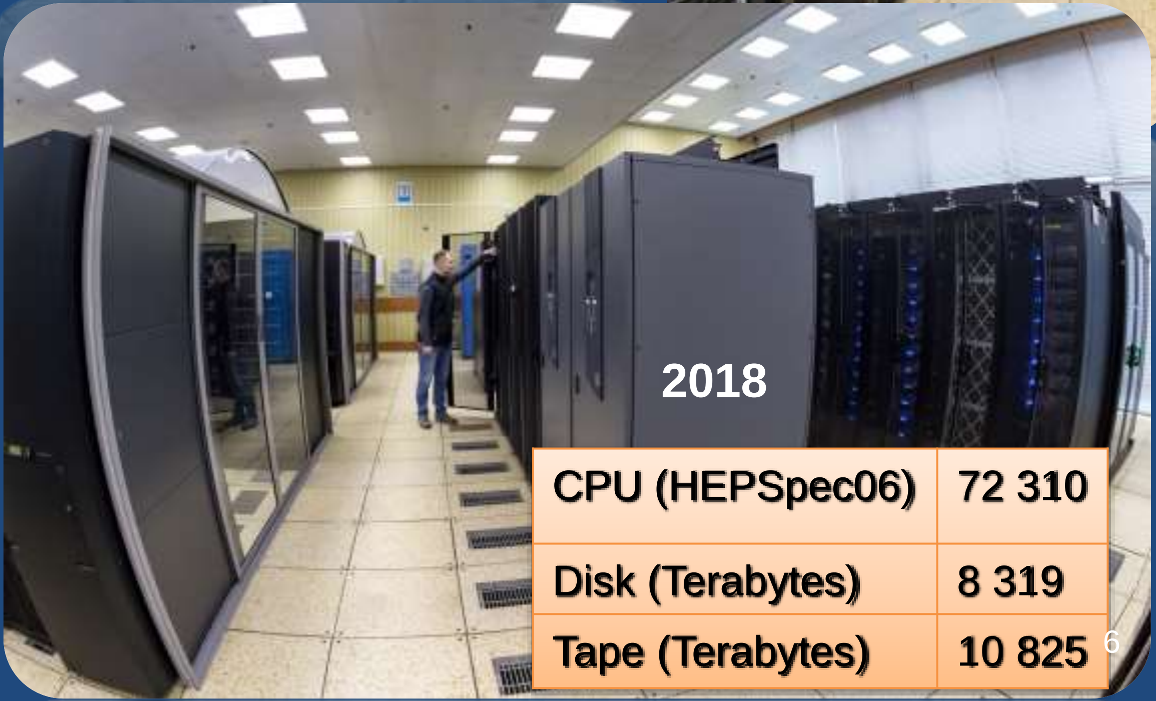
- **Mass storage system**
- **Site security**
- **Prioritization and accounting**
- **Database Services**

Tier1 prototype

2014

CPU (HEPSpec06)	14 400
Disk (Terabytes)	660
Tape (Terabytes)	72

Free space

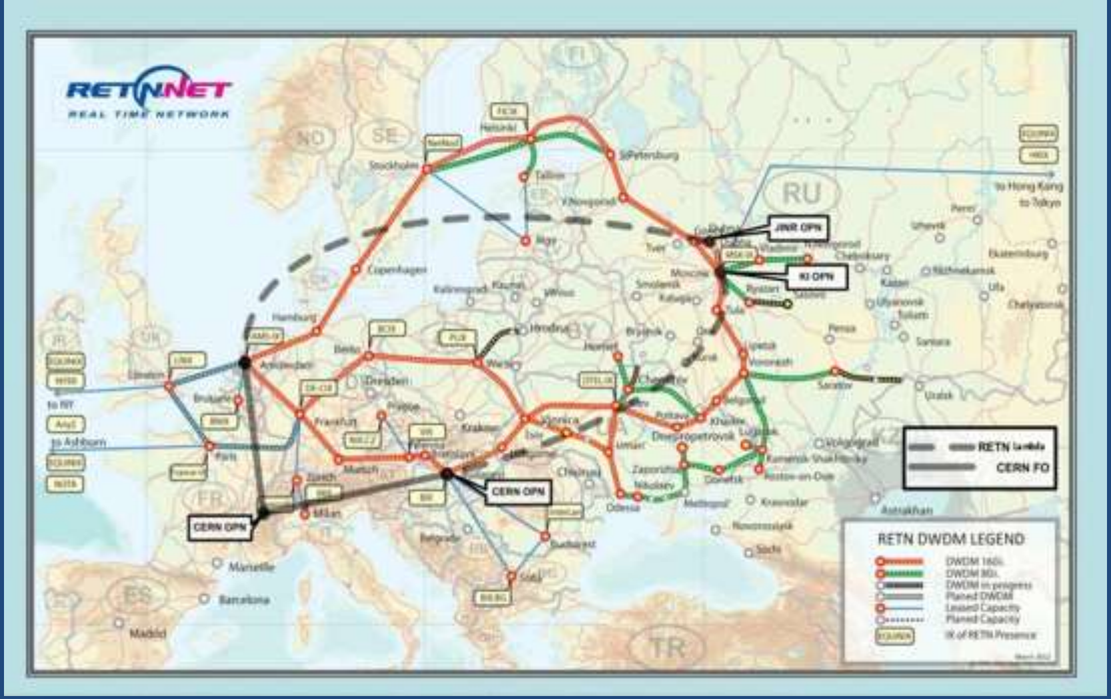


2018

CPU (HEPSpec06)	72 310
Disk (Terabytes)	8 319
Tape (Terabytes)	10 825

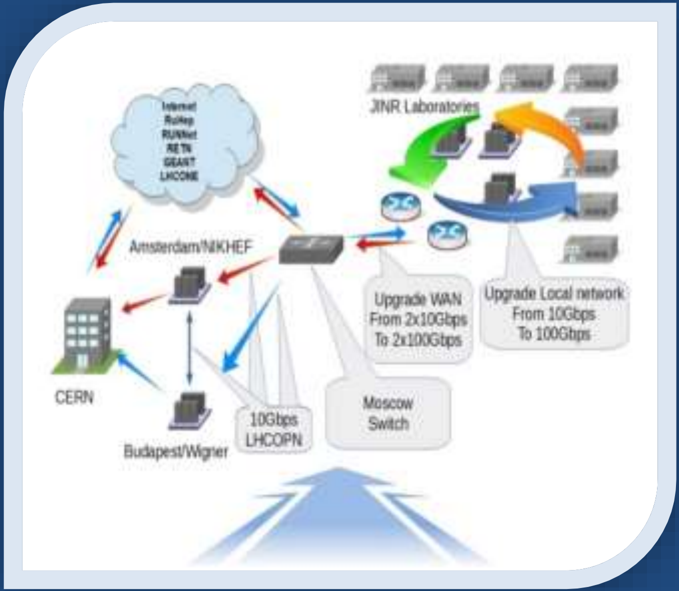
Tier1 Infrastructure

- Close-coupled, chilled water cooling InRow
- Hot and cold air containment system
- MGE Galaxy 7000 – 2x300 kW energy efficient solutions 3Ph power protection with high adaptability
- Installation of two new transformers (2.5 MW)
- Guaranteed power supply using two diesel generators



Network and telecommunication channels

The network infrastructure is meant to provide a 100% availability and reliability of the storage and computing resources of the JINR Tier-1 center.



Local Area Network – 10 Gbps, planned upgrade to 100 Gbps

Wide Area Network – 100Gbps,

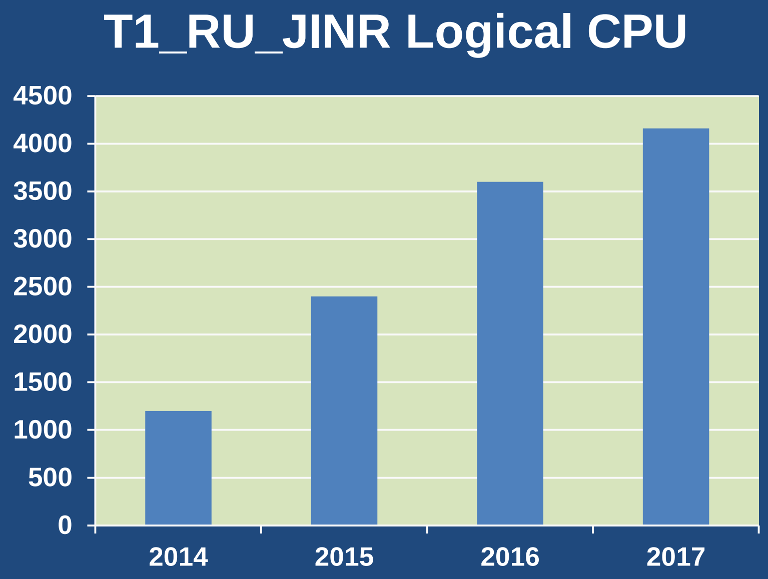
LHCOPN - 2x10Gbps

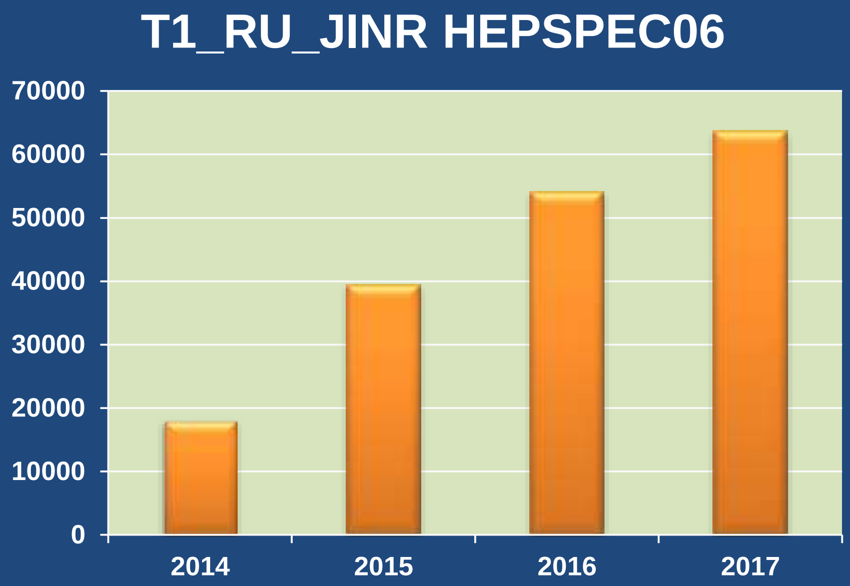
LHCONE – 10 Gbps

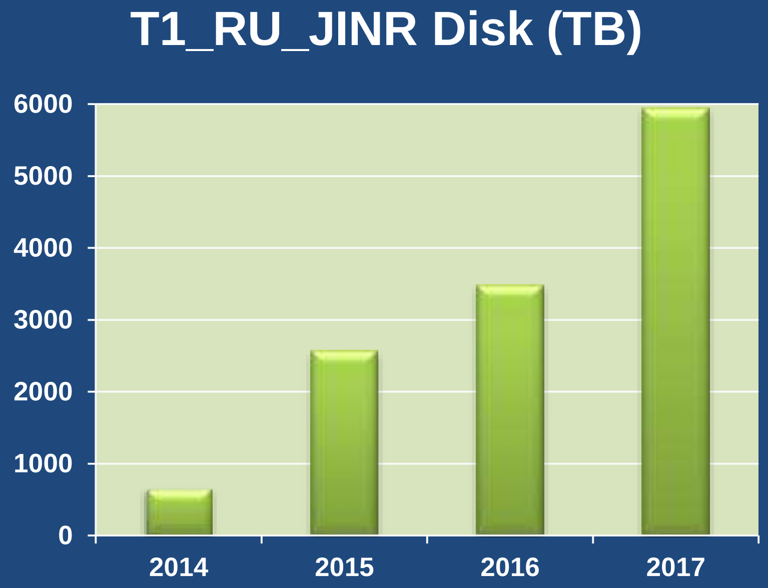
Upgrade WAN to 2x100Gbps planned

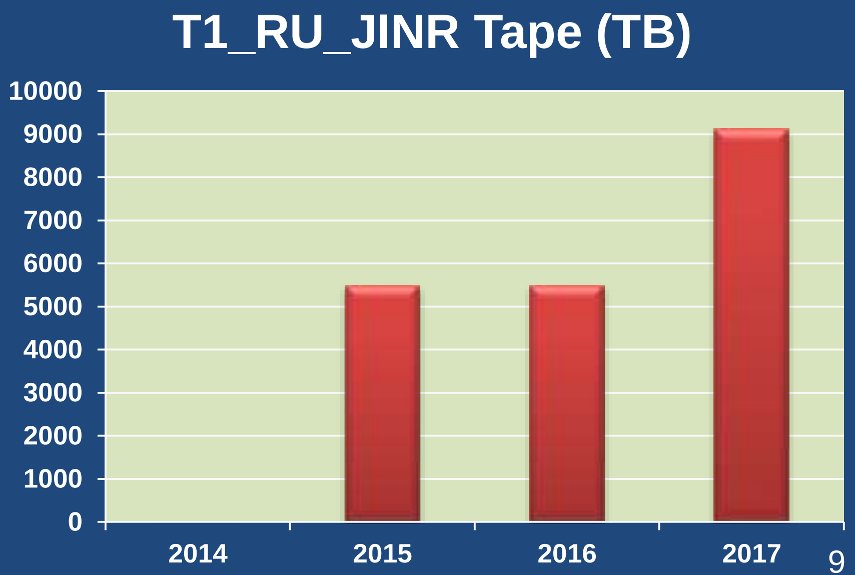
IPv6/IPv4 - enabled

Tier1 resources









Tier1 resources 2018

Computing Elements (CE)

* *Worker Node (WN)*

Typically SuperMicro Blade

100 64-bit machines: 2 x CPU (Xeon X5675 @ 3.07GHz, 6 cores per processor); 48GB RAM, 2x1000GB SATA-II; 2x1GbE.

175 64-bit machines: 2 x CPU (Xeon E5-2680 v2 @ 2.80GHz, 10 cores per processor), 64GB RAM; 2x1000GB SATA-II; 2x1GbE.

Total: 4720 core/slots for batch.

* *Software*

OS: Scientific Linux release 6 x86_64.
BATCH : Torque 4.2.10 (home made)
Maui 3.3.2 (home made)
CMS Phedex

Storage Elements(SE)
Storage System: dCache

* *Hardware*

Typically Supermicro and DELL

1st - Disk Only:

31 disk servers: 2 x CPU (Xeon E5-2650 @ 2.00GHz); 128GB RAM; 112TB h/w ZFS (24x6000GB NL SAS); 2x10G.
12 disk servers: 2 x CPU (Xeon E5-2660 @ 2.60GHz); 128GB RAM; 70TB ZFS (16x6000GB NL SAS); 2x10G.
8 disk servers: 2 x CPU (Xeon E5-2650 @ 2.29GHz) 128GB RAM; 150TB ZFS (24x8000GB NLSAS), 2x10G
12 disk servers: 2 x CPU (Xeon E5-2660 @ 2.60GHz) 128GB RAM; 150TB ZFS (24x8000GB NLSAS), 2x10G

Total space: 7.3PB

3 head node machines: 2 x CPU (Xeon E5-2683 @ 2.00GHz); 128GB RAM; 4x1000GB SAS h/w RAID10; 2x10G.
8 KVM for access protocols support.

2nd - support Mass Storage System:

8 disk servers: 2xCPU (Xeon X5650 @2.67GHz); 96GB RAM; 63TB h/w RAID6 (24x3000GB SATAIII); 2x10G; Qlogic Dual 8Gb FC.
8 disk servers: 2 x CPU (E5-2640 v4 @ 2.40GHz); 128GB RAM; 70TB ZFS (16x6000GB NLSAS); 2x10G; Qlogic Dual 16Gb FC.

Total disk buffer space: 1.1 PB.

1 tape robot: IBM TS3500, 10 PB

3440xLTO-6 data cartridges; 12xLTO-6 tape drives FC8.
3 head node machines: 2 x CPU (Xeon E5-2683 v3 @ 2.00GHz); 128GB RAM; 4x1000GB SAS h/w RAID10; 2x10G.
6 KVM machines for access protocols support

* *Software*

dCache-3.2
Enstore 4.2.2 for tape robot.

Tier1 data exchange 2017 - 2018

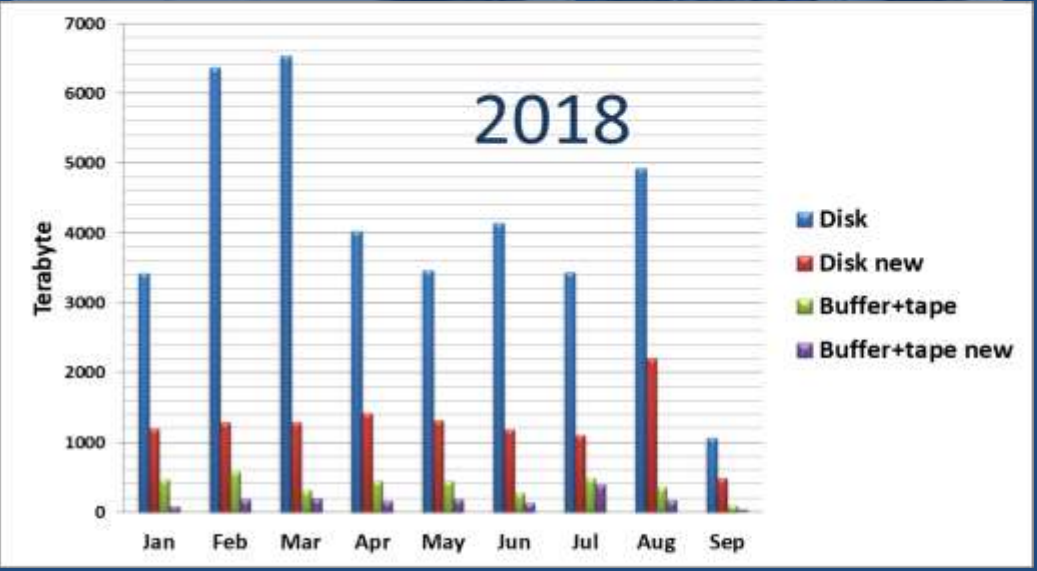
Buffer and TAPE usage:

- **157 sites worldwide transfer data FROM us;**
- **140 sites transfer data TO us;**

Leaders are CERN, KIT(Germany), RAL (UK).

Storage Element Disk only usage:

- **316 sites worldwide transfer data FROM us;**
- **150 sites transfer data TO us.**



SE disk only 2018

Total 36.50 PB

new files 11.19 PB

SE Buffer + Tape 2018

Total 6.78 PB

new files 3.38 PB

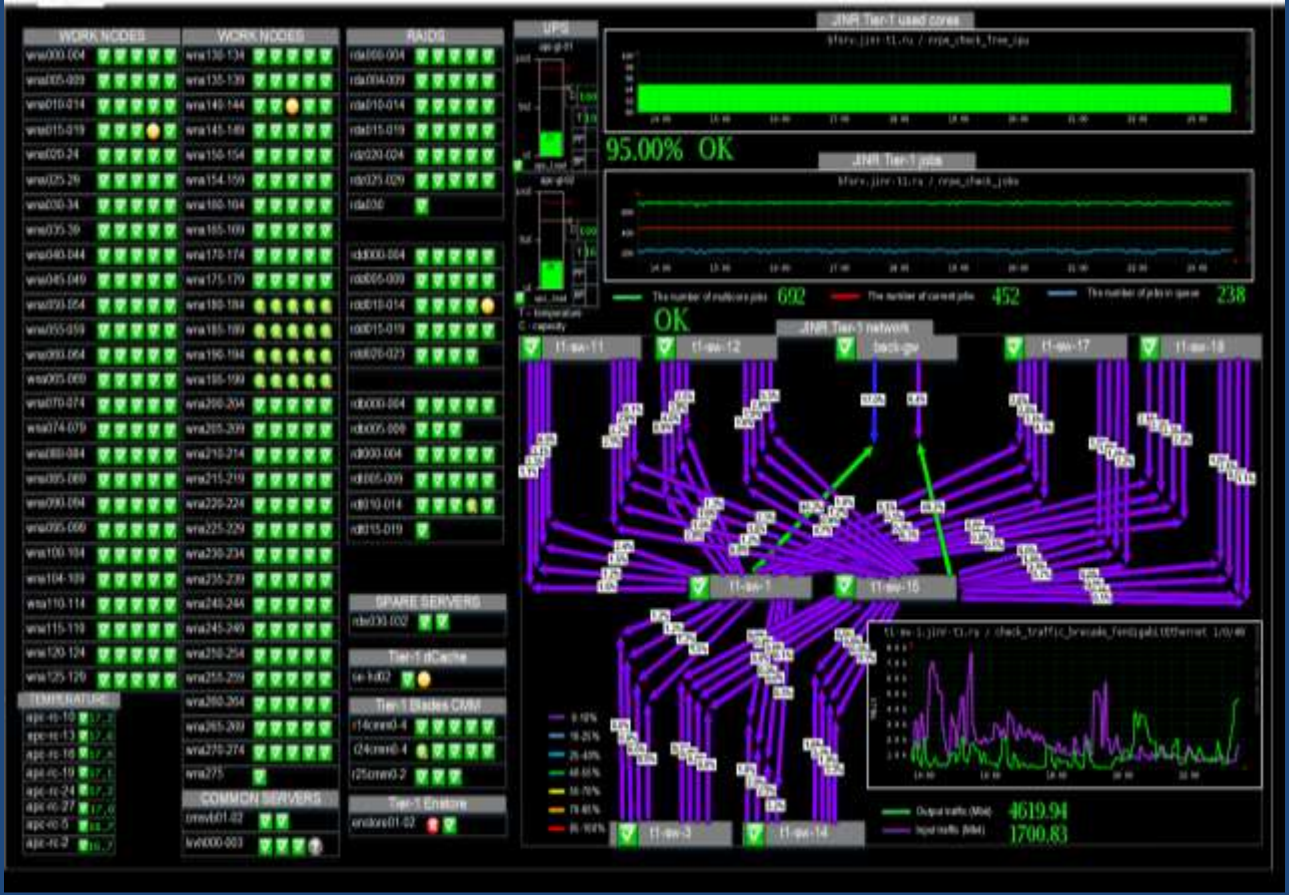
SE Buffer + Tape 2017

Total 3.64 PB

new files 1.52 PB

Tier1 hardware monitoring +

For a robust performance of the complex it is necessary to monitor the state of all nodes and services - from the supply system to the robotized tape library.



Monitoring data are collected from the wide range of hardware and software related to Tier1

- cooling systems,
- temperature sensors,
- uninterruptable power supplies (UPS),
- computing servers,
- disk arrays,
- managing services,
- L2 and L3 switches/routers
- tape robot.

~ 850 elements are under observation

~ 8000 checks in real time

~ 100 scripts

The system allows one, in a real time mode, to observe the whole computing complex state and send the system alerts to administrators and users via e-mail, sms, etc.

Tier1 services

Apart from hardware metrics, service metrics are scattered among many internal and external systems.

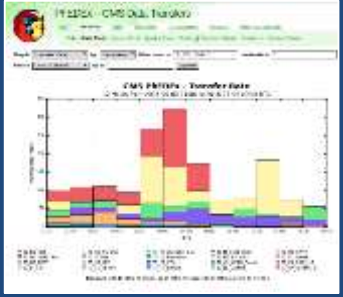
This information relates to

- data transfers,
- data storage,
- data processing.

In order to keep track of the services admin should regularly check several dozens of web pages. Interpretation of data is more complex.

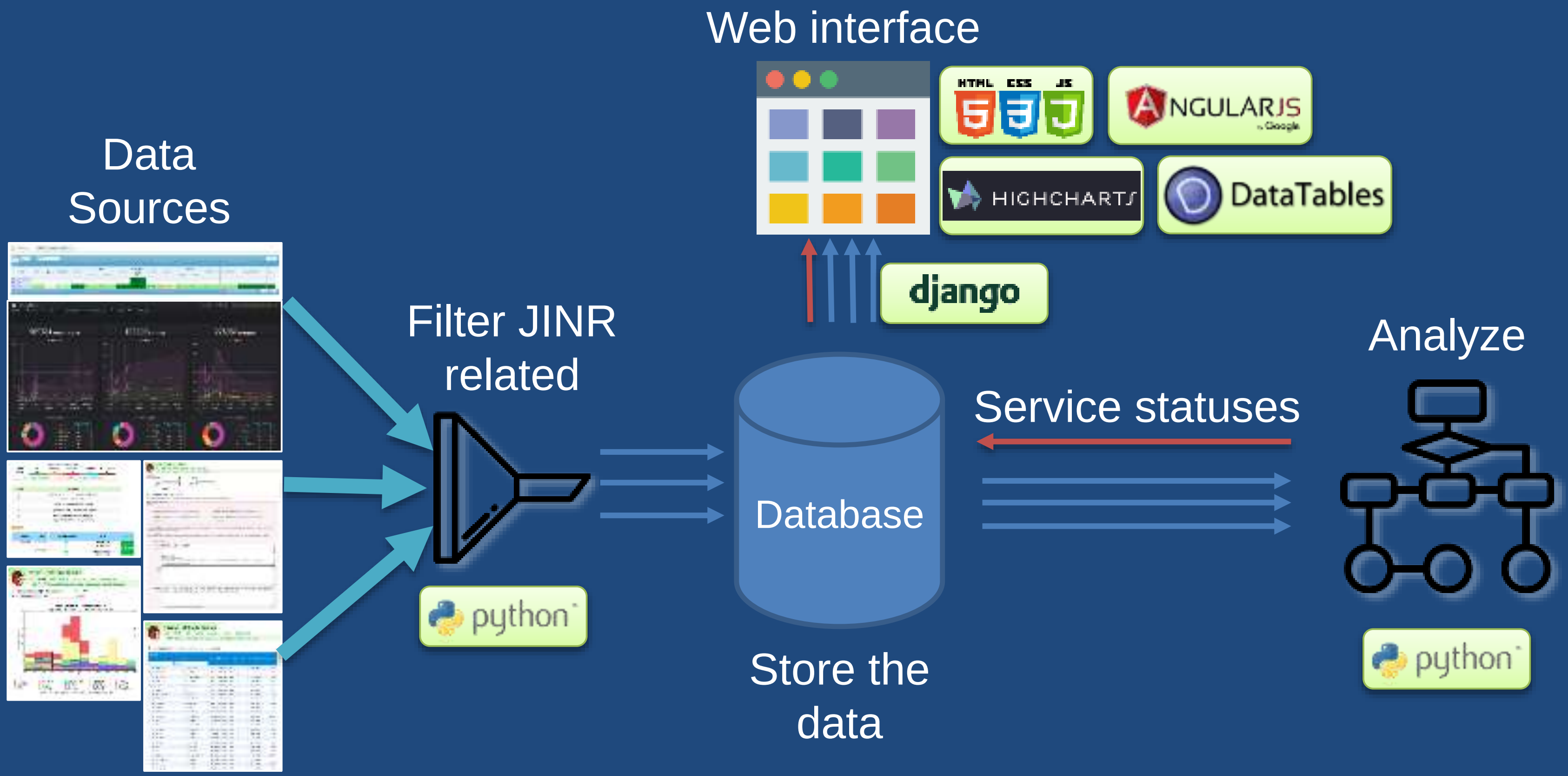
Aim is:

- Provide a single source of aggregated monitoring information.
- Perform basic analysis of data and provide status of the system.



Tier1 services monitoring system (I)

Idea is to collect, aggregate and analyze data from different sources. Then provide in comprehensive form on the web page. In case of critical failures – inform administrators.



Data Sources

Filter JINR related

Web interface

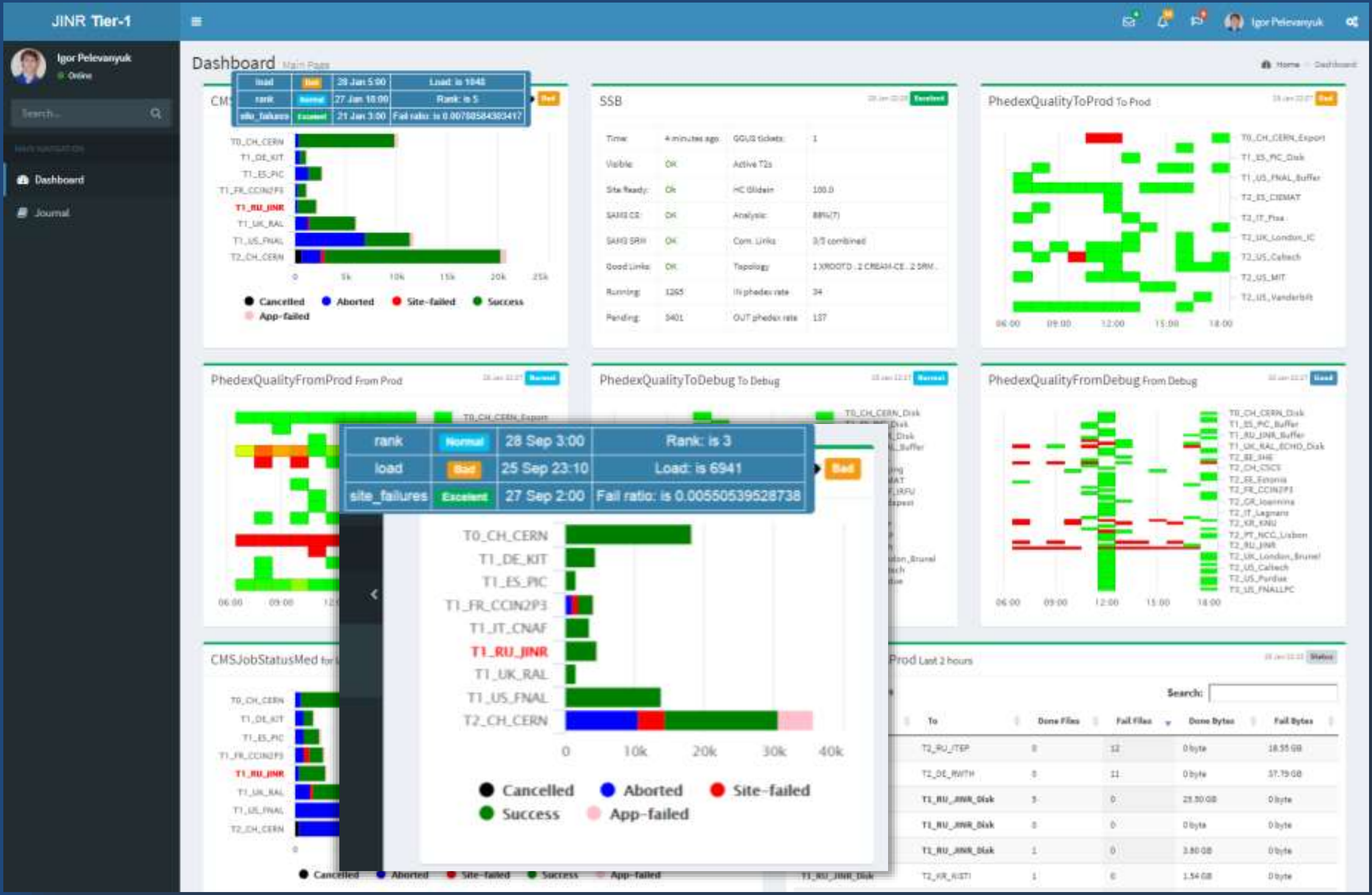
Database

Store the data

Service statuses

Analyze

Tier1 services monitoring system (II)



load	Bad	28 Jan 5:00	Load: is 1048
rank	Normal	27 Jan 10:00	Rank: is 5
site_failures	Excellent	21 Jan 3:00	Fail ratio: is 0.00780584383417

SSB 28 Jan 12:27 Excellent

Time:	4 minutes ago	GGUS tickets:	1
Visible:	OK	Active T2s	
Site Ready:	Ok	HC Glidein	100.0
SAM3 CE:	OK	Analysis:	88%(7)
SAM3 SRM:	OK	Com. Links:	3/5 combined
Good Links:	OK	Topology:	1 XROOTD , 2 CREAM-CE , 2 SRM ...
Running:	1265	IN phedex rate	34
Pending:	3401	OUT phedex rate	137

rank	Normal	28 Sep 3:00	Rank: is 3
load	Bad	25 Sep 23:10	Load: is 6941
site_failures	Excellent	27 Sep 2:00	Fail ratio: is 0.00550539528738

Prod Last 2 hours 28 Jan 12:23 Status

Search:

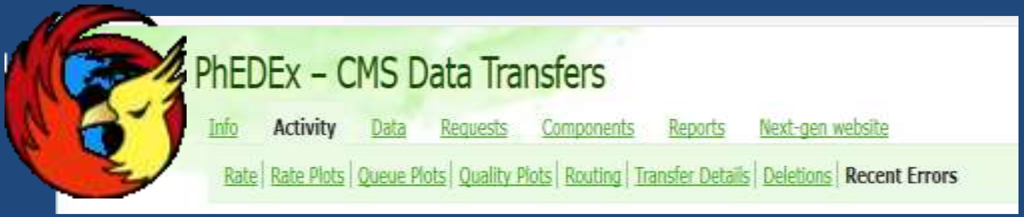
To	Done Files	Fail Files	Done Bytes	Fail Bytes
T2_RU_ITEP	0	12	0 byte	18.55 GB
T2_DE_RWTH	0	11	0 byte	37.79 GB
T1_RU_JINR_Disk	5	0	23.50 GB	0 byte
T1_RU_JINR_Disk	0	0	0 byte	0 byte
T1_RU_JINR_Disk	1	0	3.80 GB	0 byte
T2_KR_KISTI	1	0	1.54 GB	0 byte

PhEDEx operations: issues

PhEDEx system was designed to operate mostly automatically. But sometimes, due to different reasons it requires intervention to fix errors manually.

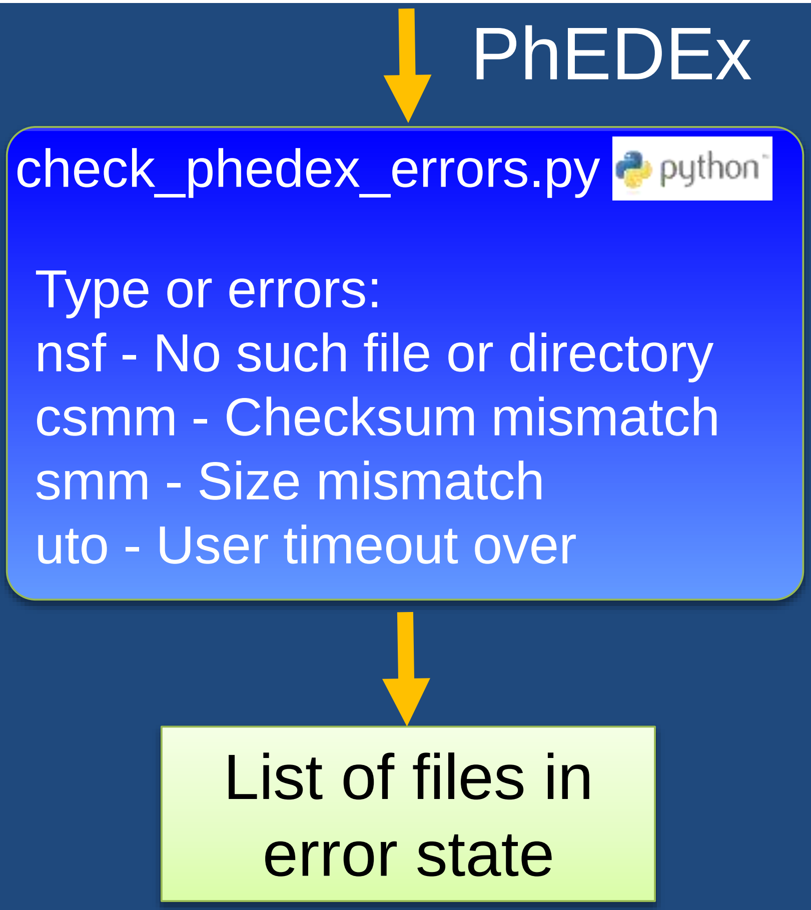
Source of information about errors is a corresponding PhEDEx webpage. Every error is a big form with source/destination site, time of assigned/start/done, PFNs to/from, transfer/detail/validate logs.

In order to simplify operation python script was written to list important errors and provide relevant information about them.



PhEDEx – CMS Data Transfers

Info
Activity
Data
Requests
Components
Reports
Next-gen website
Rate | Rate Plots | Queue Plots | Quality Plots | Routing | Transfer Details | Deletions | Recent Errors



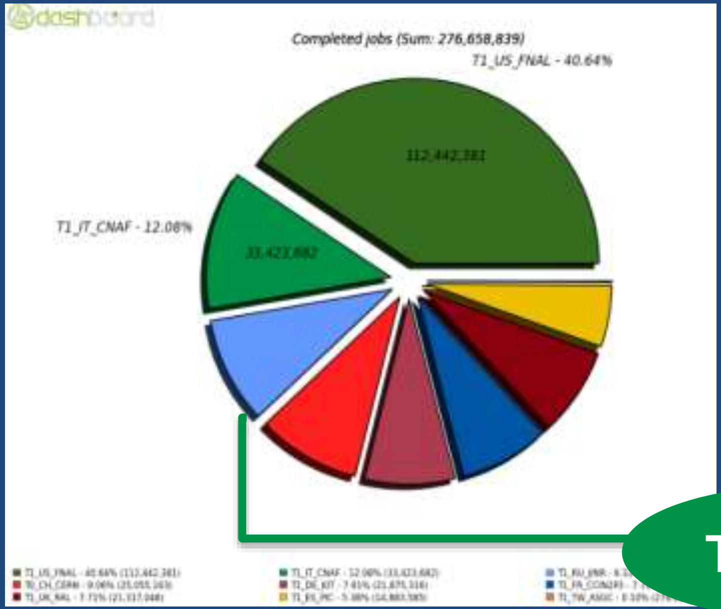
PhEDEx

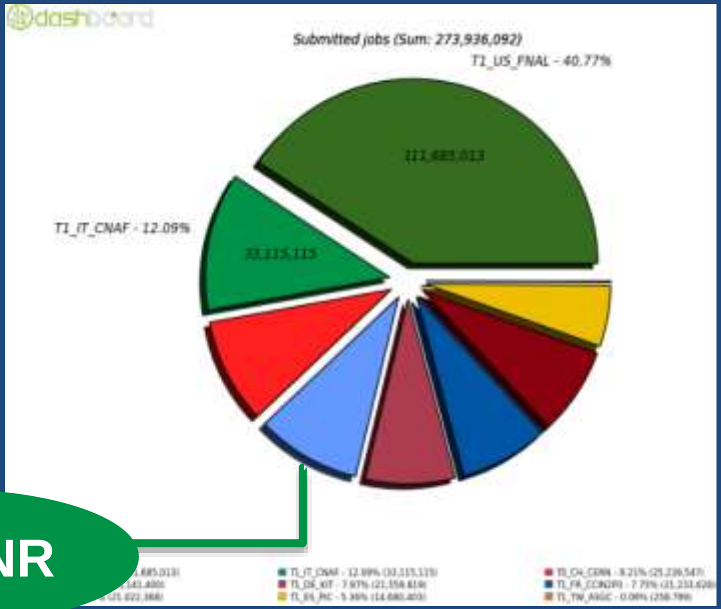
check_phedex_errors.py

Type or errors:
nsf - No such file or directory
csmm - Checksum mismatch
smm - Size mismatch
uto - User timeout over

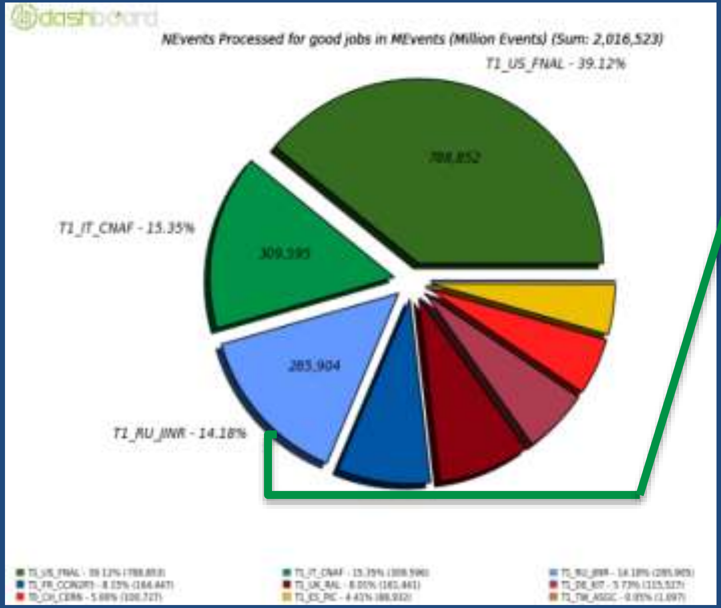
List of files in error state

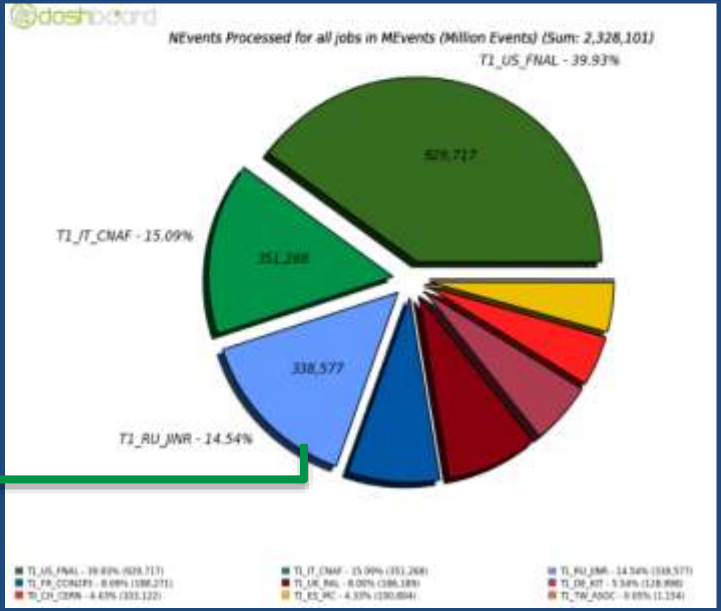
From: 2014-01-01 to: 2018-10-17

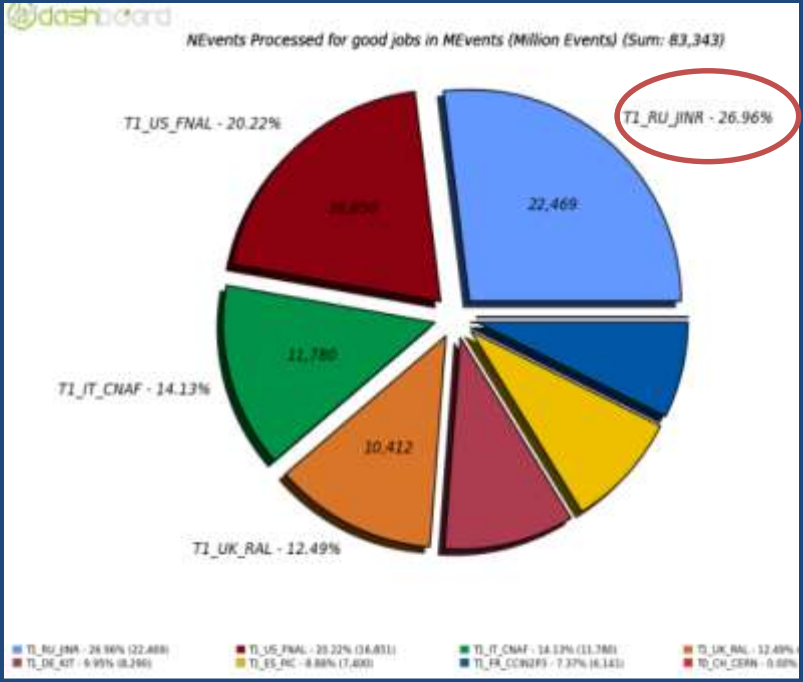


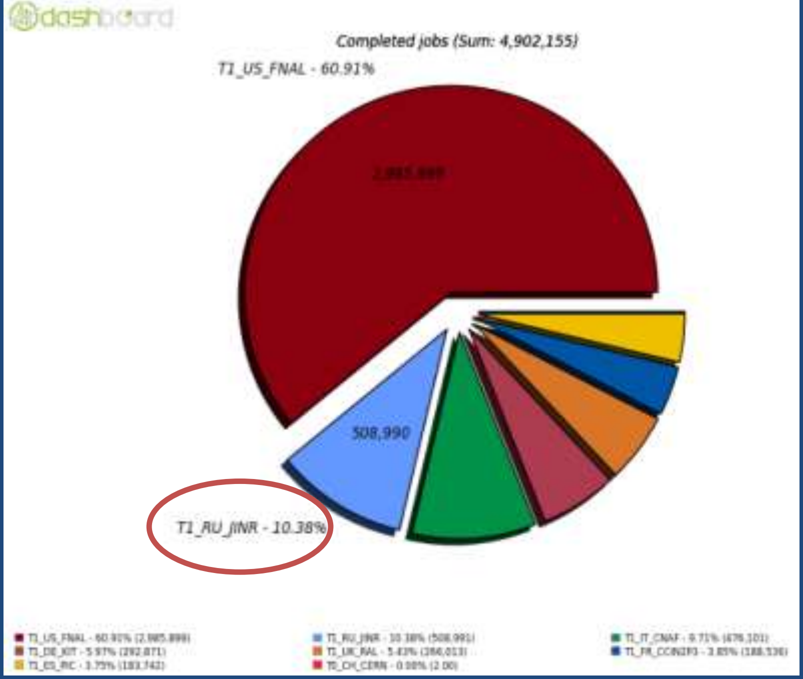


T1_JINR

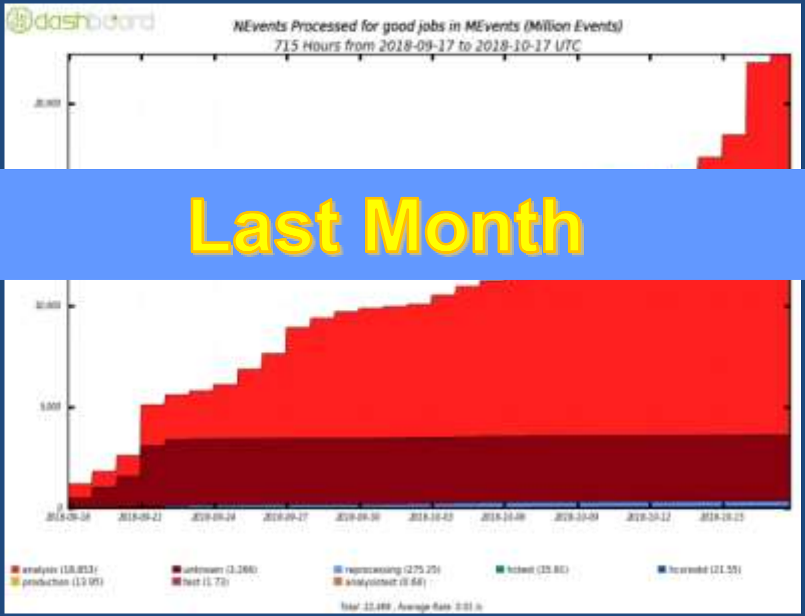


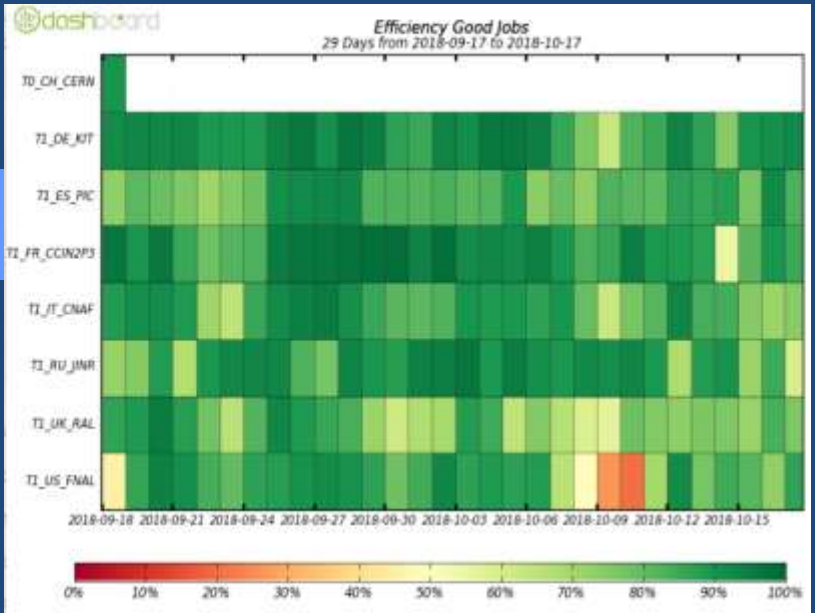



dashboard
NEvents Processed for good jobs in MEvents (Million Events) (Sum: 83,343)
T1_RU_JINR - 26.96%
T1_US_FNAL - 20.22%
T1_IT_CNAF - 14.13%
T1_UK_RAL - 12.49%
22,469
16,850
11,780
10,412
T1_RU_JINR - 26.96% (22,469)
T1_US_FNAL - 20.22% (16,851)
T1_IT_CNAF - 14.13% (11,780)
T1_UK_RAL - 12.49%
T1_DE_KIT - 9.95% (8,290)
T1_ES_PIC - 8.88% (7,400)
T1_FR_CCIN2P3 - 7.37% (6,141)
T0_CH_CERN - 0.00%

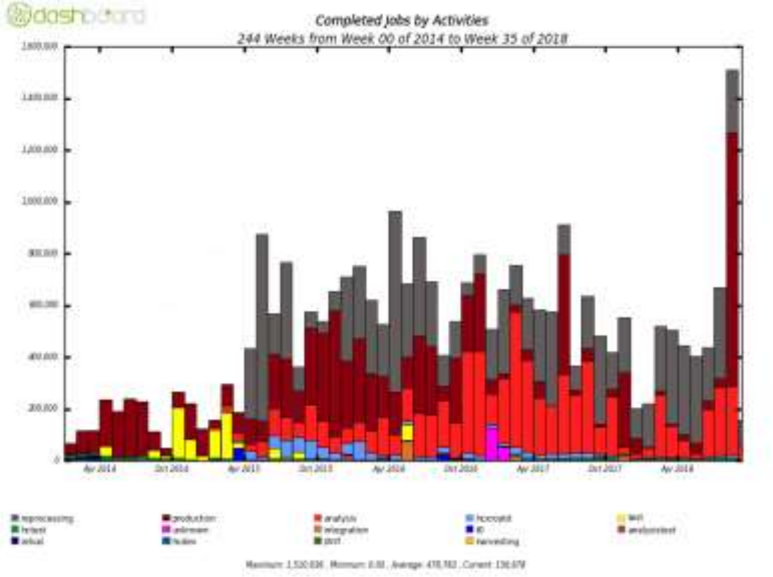

dashboard
Completed jobs (Sum: 4,902,155)
T1_US_FNAL - 60.91%
2,985,899
508,990
T1_RU_JINR - 10.38%
T1_US_FNAL - 60.91% (2,985,899)
T1_DE_KIT - 5.97% (292,871)
T1_ES_PIC - 3.75% (183,742)
T1_RU_JINR - 10.38% (508,991)
T1_UK_RAL - 5.43% (266,013)
T0_CH_CERN - 0.00% (2.00)
T1_IT_CNAF - 9.71% (476,101)
T1_FR_CCIN2P3 - 3.85% (188,536)

Last Month


dashboard
NEvents Processed for good jobs in MEvents (Million Events)
715 Hours from 2018-09-17 to 2018-10-17 UTC
0
2018-09-18
2018-09-21
2018-09-24
2018-09-27
2018-09-30
2018-10-03
2018-10-06
2018-10-09
2018-10-12
2018-10-15
analysis (18,853)
production (13.95)
unknown (3,266)
test (1.73)
reprocessing (275.25)
analysistest (0.68)
hctest (35.81)
hcxrootd (21.55)
Total: 22,469 , Average Rate: 0.01 /s

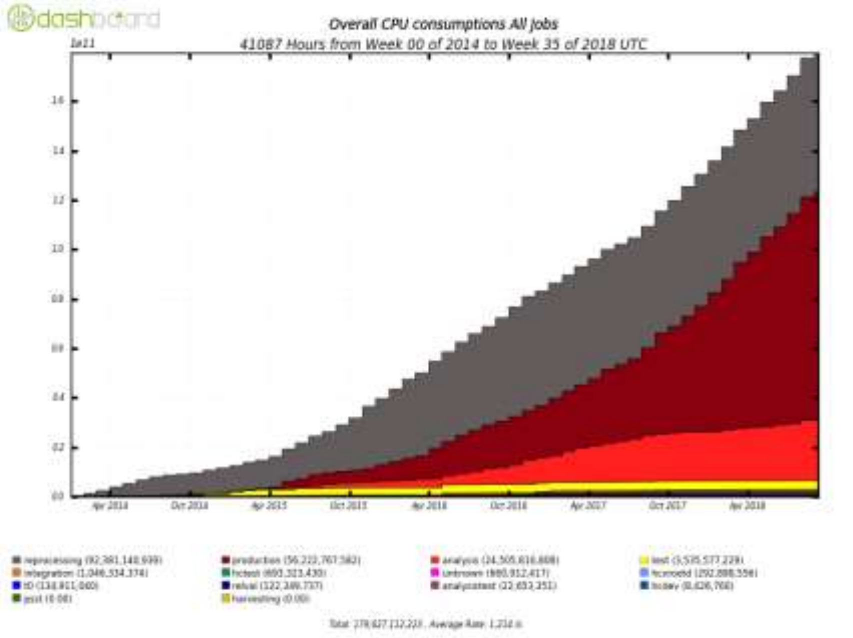

dashboard
Efficiency Good Jobs
29 Days from 2018-09-17 to 2018-10-17
T0_CH_CERN
T1_DE_KIT
T1_ES_PIC
T1_FR_CCIN2P3
T1_IT_CNAF
T1_RU_JINR
T1_UK_RAL
T1_US_FNAL
2018-09-18
2018-09-21
2018-09-24
2018-09-27
2018-09-30
2018-10-03
2018-10-06
2018-10-09
2018-10-12
2018-10-15
0%
10%
20%
30%
40%
50%
60%
70%
80%
90%
100%

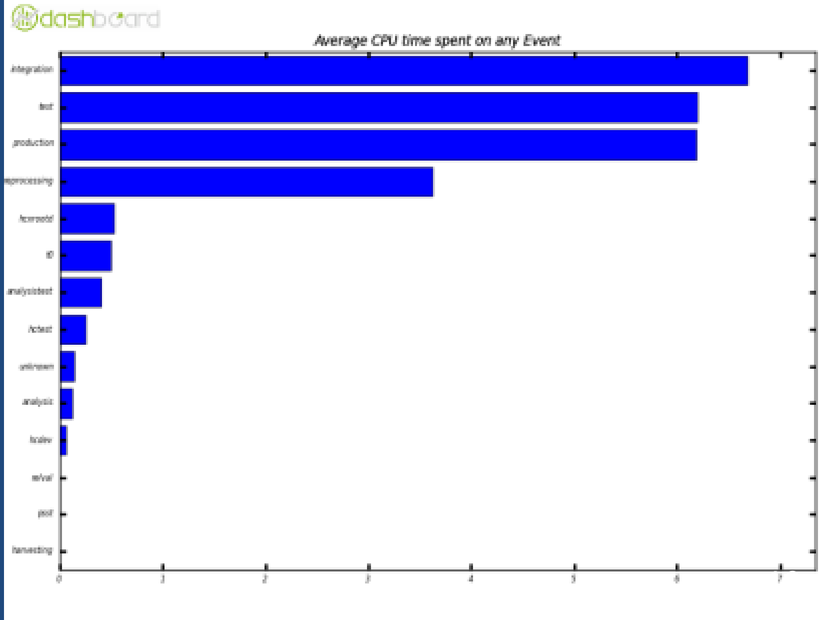
Jobs Processing by Activities



Activities	Nevents
Analysis	281 555 814 018
Reprocessing	10 042 089 961
Production	5 569 434 071
Test	428 136 099
Etc.	…..

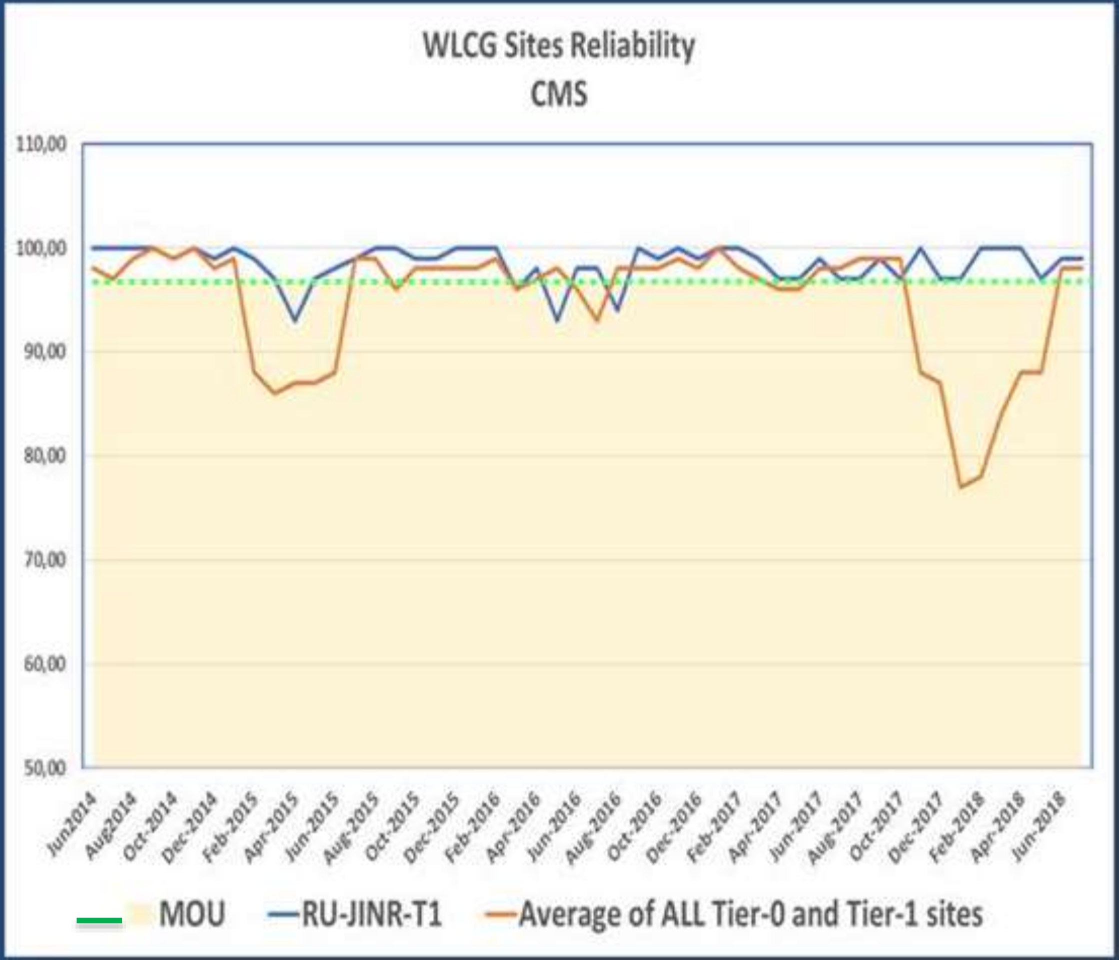
Total: 308 560 399 956 events
Average Rate: 2.086/s





Tier1 site reliability

JINR Tier1 (blue) monthly results compared to the average Tier1 reliabilities (orange) as well as with the WLCG target for site reliability (green dotted line) which is set to 97% since 2009, according to the WLCG MoU



JINR Tier1 site upgrade plans for 2019

CPU (HEPSpec06):

72310 → 90000;

Disk storage volume:

7.3 PB → 8PB;

Tape robot volume:

10 PB → 20 PB.

Importance of the Tier1 center at JINR

- Creation of conditions for JINR physicists, JINR Member States, RDMS-CMS collaboration for a full-scale participation in processing and analysis of data of the CMS experiment on the Large Hadron Collider.
- The invaluable experience of launching the Tier1 center will be used for creating a system of storage and data processing of megaproject NICA and other scale projects of the JINR-participating countries.
- The studies in the field of Big Data analytics assume significance for the development of the perspective directions of science and economy as well as analysis and forecasting of processes in various fields.

Thank you for your attention!