TIER-1 CMS at JINR: Past, Present and Future

N.S. Astakhov, A.S. Baginyan, S.D. Belov, A.G. Dolbilov, A.O. Golunov, I.N. Gorbunov, N.I. Gromova, I.S. Kadochnikov, I.A. Kashunin, V.V. Korenkov, V.V. Mitsyn, I.S.Pelevenyuk, S.V. Shmatov, **T.A. Strizh**, E.A. Tikhonenko, V.V. Trofimov, N.N. Voitishin, V.E. Zhiltsov

Joint Institute for Nuclear Research

The project of launching Tier-1 at JINR

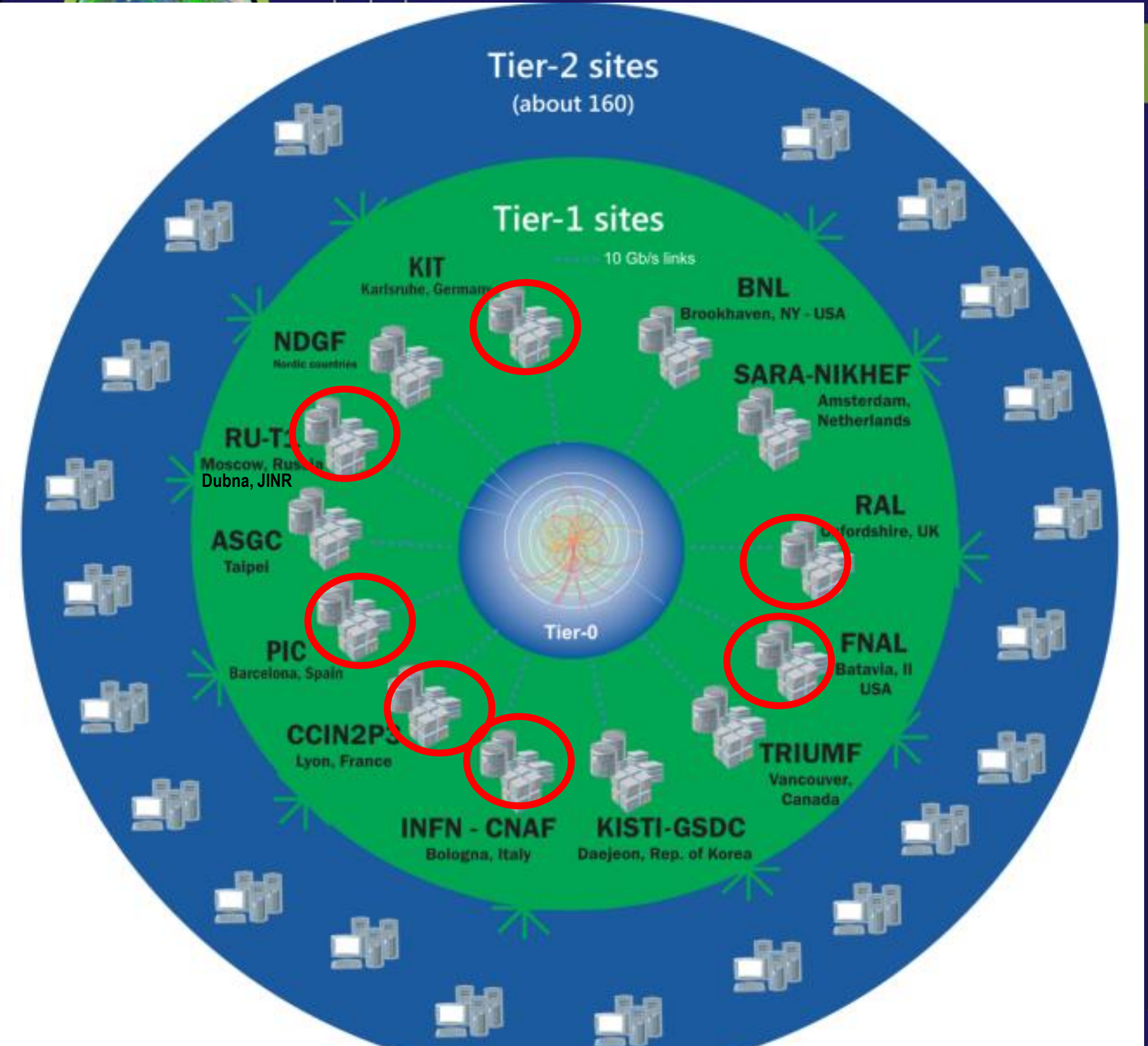
In November, 2011 at the suggestion of A.A. Fursenko at a session of the Committee on Russia-CERN cooperation, a decision was accepted on the creation in Russia of a Tier-1 center for LHC experiments on the base of NRC "Kurchatov institute" and JINR.

On September 28, 2012, a session of the Supervisory Council of WLCG project approved a work schedule on the creation of Tier-1 in Russia.

First stage (December, 2012) - creation of a prototype of Tier-1 at NRC KI and at JINR.
Second stage (November, 2013) - installation of equipment for the base Tier-1 center, its testing and finishing up to required functional characteristics.
Third stage (March, 2015) – finalization of this complex and commissioning a full-scale Tier-1 center for CMS in JINR.

In the sequel, a systematic increase of computing capacity and data storage is needed in accordance with the experiment requirements.

LHC Computing Model

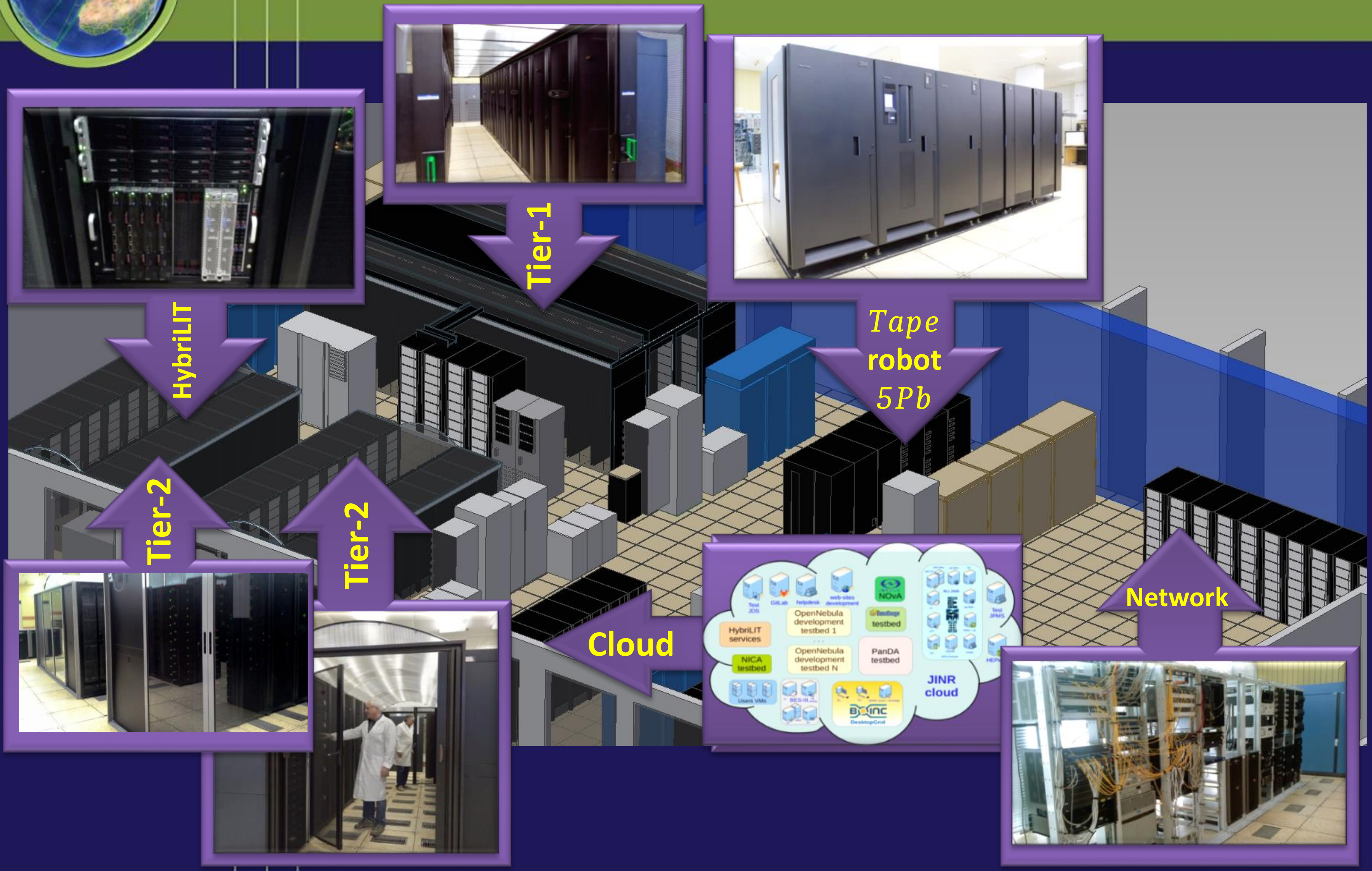


Tier-0 (CERN):
- Data recording
- Initial data reconstruction
- Data distribution

Tier-1 (11→14 centres):
- Permanent storage
- Re-processing
- Analysis
- Simulation

Tier-2 (>200 centres):
- Simulation
- End-user analysis

JINR Computing Centre for Data Storage, Processing and Analysis



Tier-1

HybriLIT

Tape **robot** *5Pb*

Tier-2

Tier-2

Cloud

Network

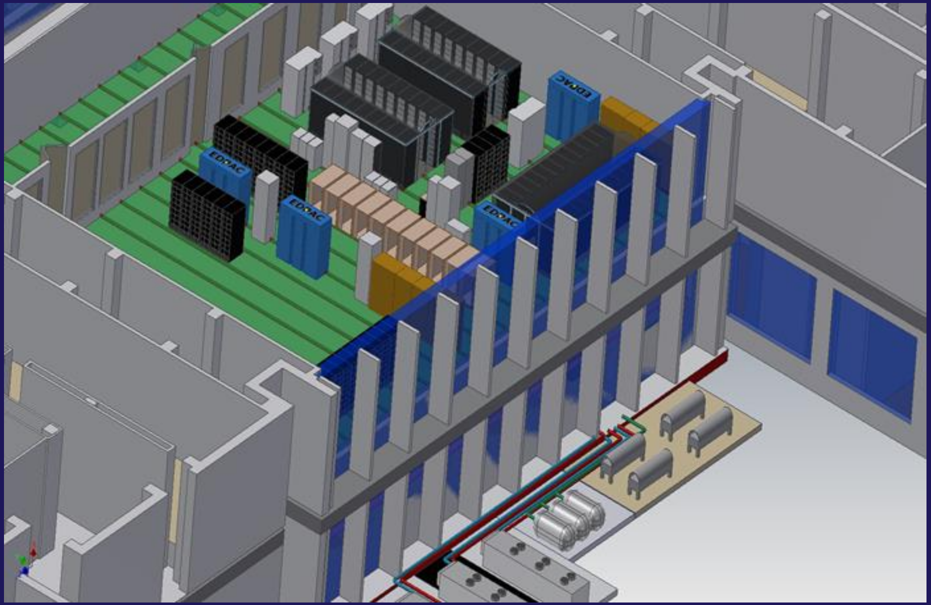
In accordance with the CMS computing model, the centers Tier-1 for CMS provide a wide range of reliable services for the whole CMS collaboration using standard grid-interfaces of the WLCG project and additional CMS services. Also required is a high-level availability of all Tier-1 services and a constant, in 7x24 mode, technical support for troubleshooting.

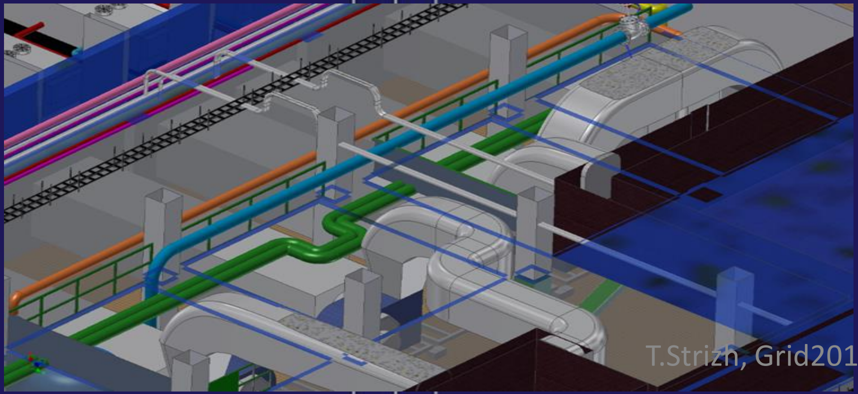
The main functions of the Tier-1 centre include:

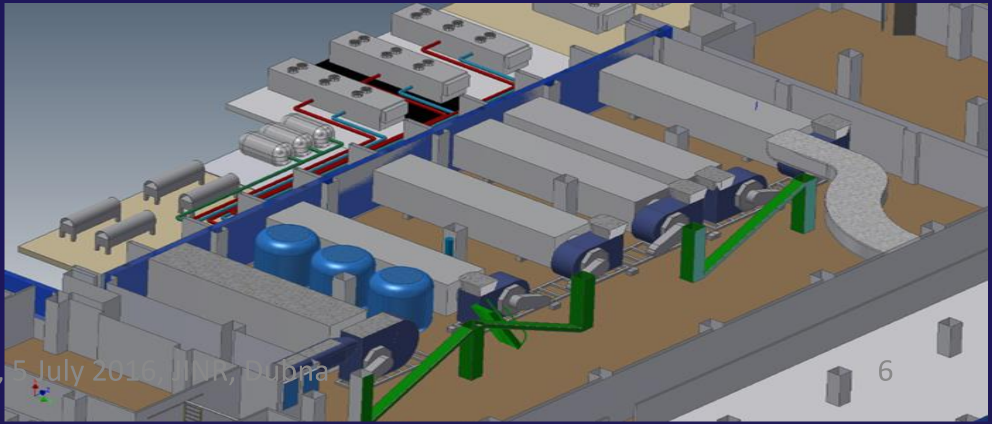
- organization of sequential data processing (data acquisition, fast browsing (skimming), reprocessing),
- storage of bulk raw experimental and simulated data
- providing data to other Tier-1-2-3 sites for copying and physical analysis,
- receiving “raw” data from Tier-0 centre for a long-time (several years) storage, replication of such data onto other Tier-1 centres.
- data analysis using algorithms of the CMS collaboration.

Creation of CMS Tier-1 in JINR

- Engineering infrastructure (a system of uninterrupted power supply, climate - control);
- High-speed reliable network infrastructure with a dedicated reserved data link to CERN (LHCOPN);
- Computing system and storage system on the basis of disk arrays and tape libraries of high capacity;
- 100% reliability and availability.







Visible to the user services of the Tier-1 centre include:

1) services of archival storage,

2) services of disk storage,

3) services of access to data and provision of data exchange,

4) services of control and recovery of the system,

5) services for data processing and analysis,

6) services to organize a user access to the WLCG resources.

CMS Tier-1

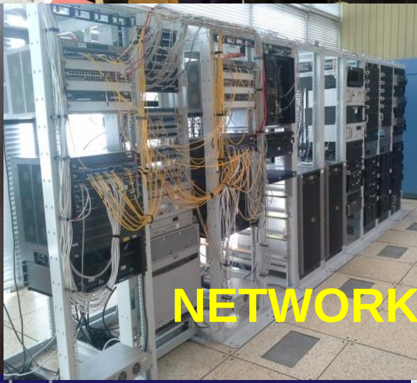
March 2015 – CMS Tier1 Inauguration

LHCOPN – 10Gbps, 2400 cores (~ 30 kHS06),
5 PB tapes (IBM TS3500), 2.4 PB disk
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

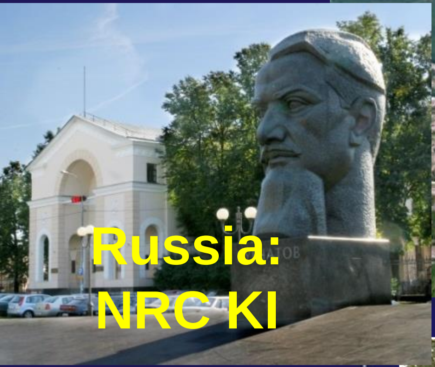
Tape Robot

Computing elements

Uninterrupted power supply


NETWORK

Cooling system


Russia:
NRC KI

US-BNL

Amsterdam/NIKHEF-SARA

Taipei/ASGC

Bologna/CNAF

Ca-
TRIUMF

JINR

NDGF

CERN

US-FNAL

UK-RAL

De-FZK

Barcelona/PIC

Lyon/CCIN2P3

Architecture and hardware resources

The computing cluster: 220 physical machines, 3400 cores /54,4 kHS06

To support the batch processing system, a special server with a cluster resource allocation system and a scheduler has been installed.

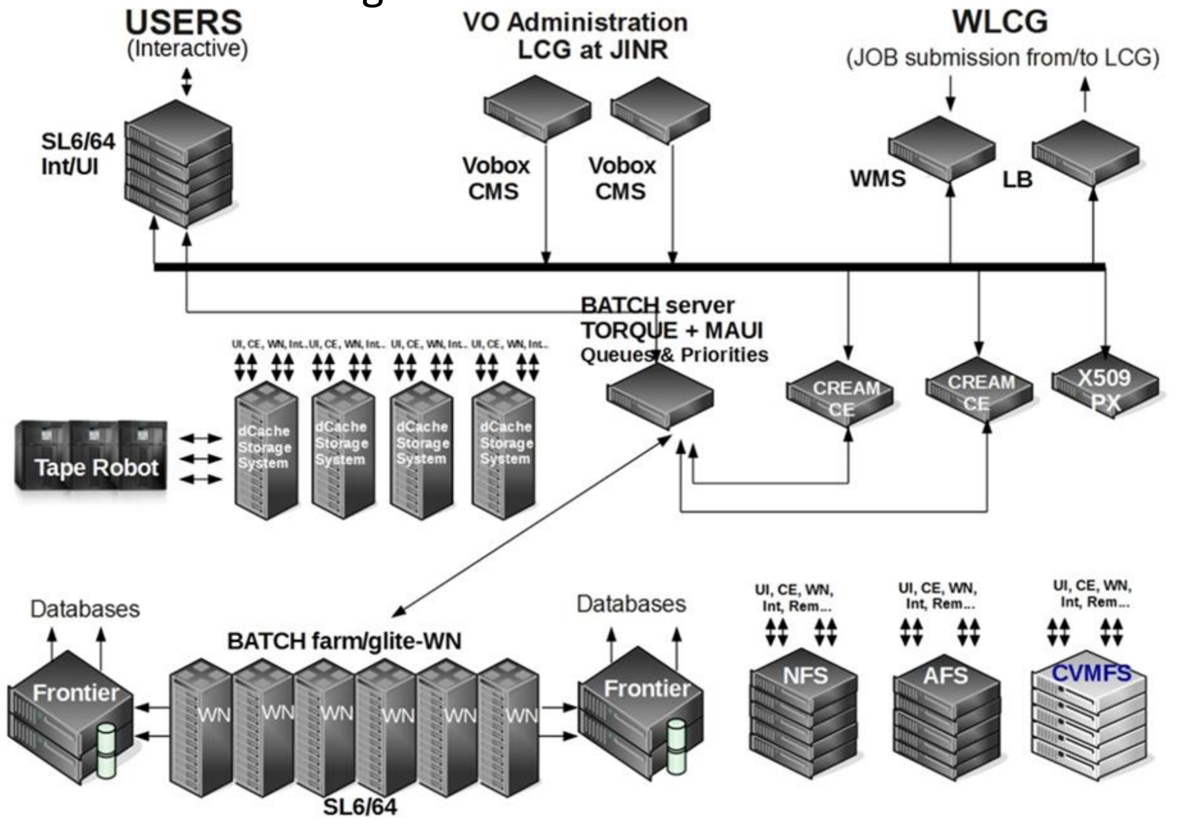
Storage systems - dCache software.

1. dCache installation - only disk servers and used for operational data storage with fast access to them.
2. dCache installation - disk servers and a tape robot. The disks serve as a buffer zone, tape robot is intended for a long-time storage of data.

Totally, 2 installations have now 3.4 PB of effective disk space, and the tape robot has a 5.4 PB of data storage capacity. To support the storage and access to data, 8 physical and 14 virtual machines have been installed.

The servers to support the grid-WLCG environment

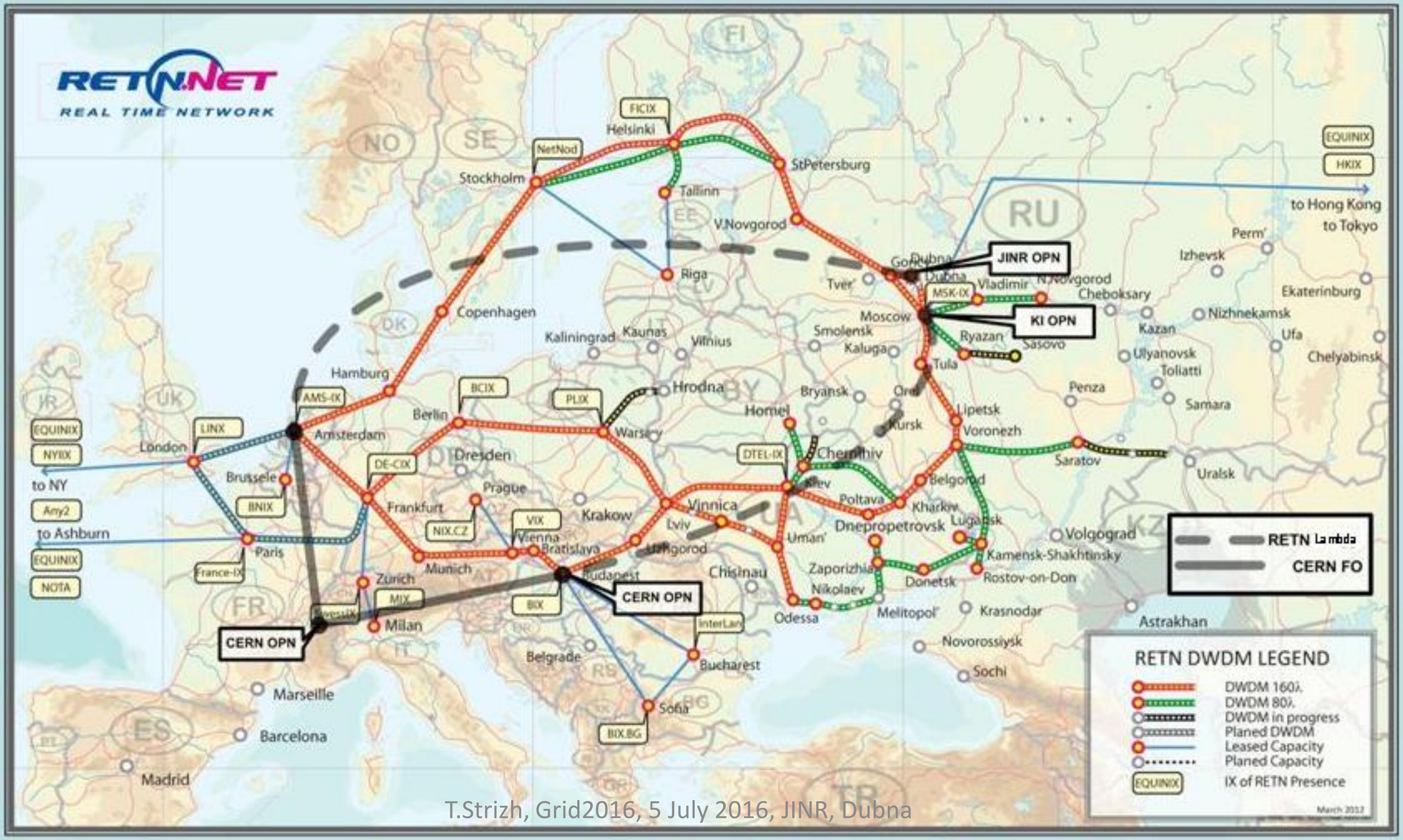
Logical structure of the Tier-1



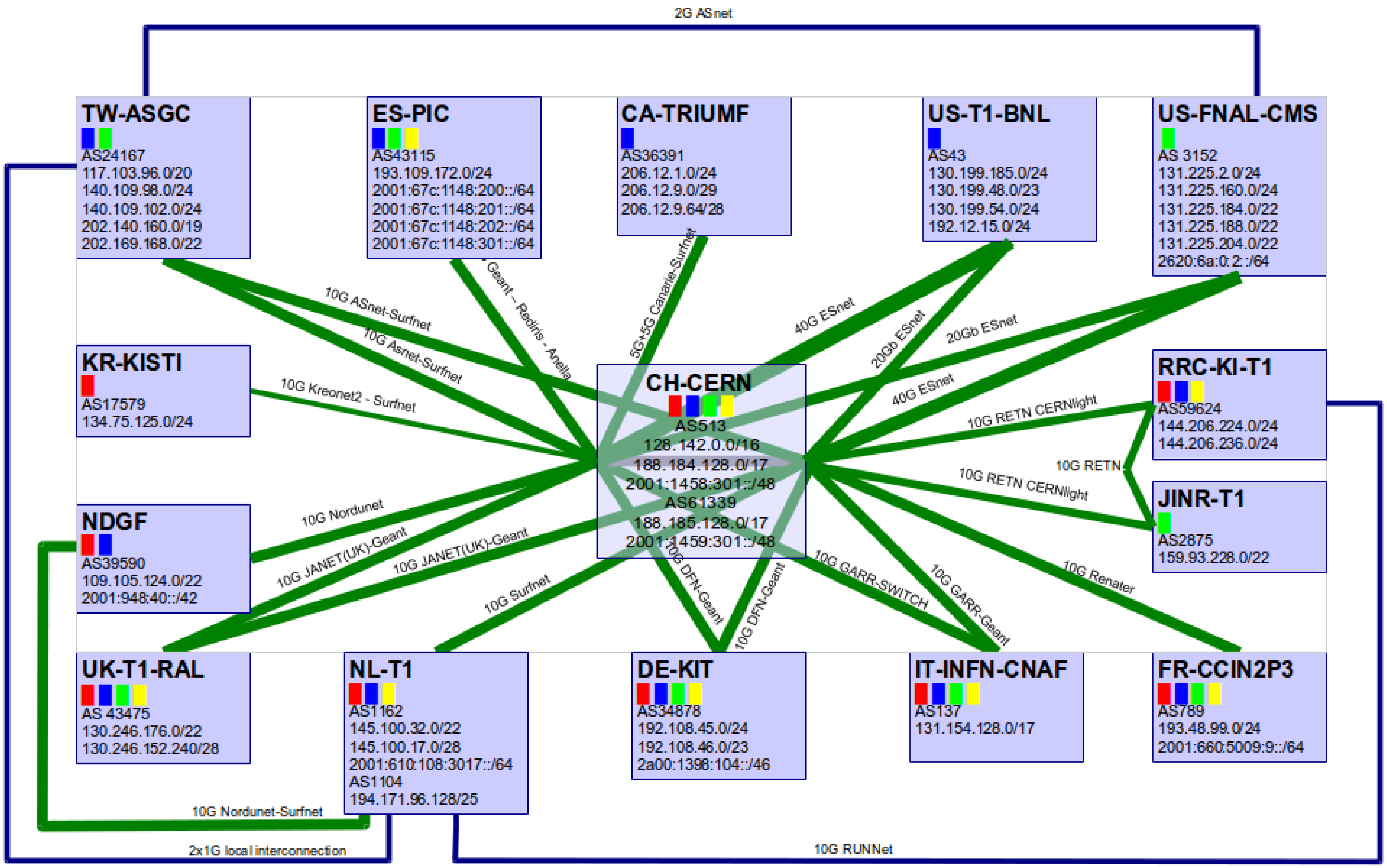
WLCG services are installed with grid middleware EMI-3. Currently, 21 services WLCG are installed to provide :

- user and virtual organizations (VO) authorization;
- task run from VO remote services.
- the WLCG information system;
- different algorithms of remote testing and verification of the service environment on local resources.

JINR Tier-1 Connectivity Scheme



LHCOPN



TW-ASGC 🟦🟩
AS24167
117.103.96.0/20
140.109.98.0/24
140.109.102.0/24
202.140.160.0/19
202.169.168.0/22

ES-PIC 🟦🟩🟨
AS43115
193.109.172.0/24
2001:67c:1148:200::/64
2001:67c:1148:201::/64
2001:67c:1148:202::/64
2001:67c:1148:301::/64

CA-TRIUMF 🟦
AS36391
206.12.1.0/24
206.12.9.0/29
206.12.9.64/28

US-T1-BNL 🟦
AS43
130.199.185.0/24
130.199.48.0/23
130.199.54.0/24
192.12.15.0/24

US-FNAL-CMS 🟩
AS 3152
131.225.2.0/24
131.225.160.0/24
131.225.184.0/22
131.225.188.0/22
131.225.204.0/22
2620:6a:0:2::/64

KR-KISTI 🟥
AS17579
134.75.125.0/24

CH-CERN 🟥🟦🟩🟨
AS513
128.142.0.0/16
188.184.128.0/17
2001:1458:301::/48
AS61339
188.185.128.0/17
2001:1459:301::/48

RRC-KI-T1 🟥🟦🟨
AS59624
144.206.224.0/24
144.206.236.0/24

NDGF 🟥🟦
AS39590
109.105.124.0/22
2001:948:40::/42

JINR-T1 🟩
AS2875
159.93.228.0/22

UK-T1-RAL 🟥🟦🟩🟨
AS 43475
130.246.176.0/22
130.246.152.240/28

NL-T1 🟥🟦🟨
AS1162
145.100.32.0/22
145.100.17.0/28
2001:610:108:3017::/64
AS1104
194.171.96.128/25

DE-KIT 🟥🟦🟩🟨
AS34878
192.108.45.0/24
192.108.46.0/23
2a00:1398:104::/46

IT-INFN-CNAF 🟥🟦🟩🟨
AS137
131.154.128.0/17

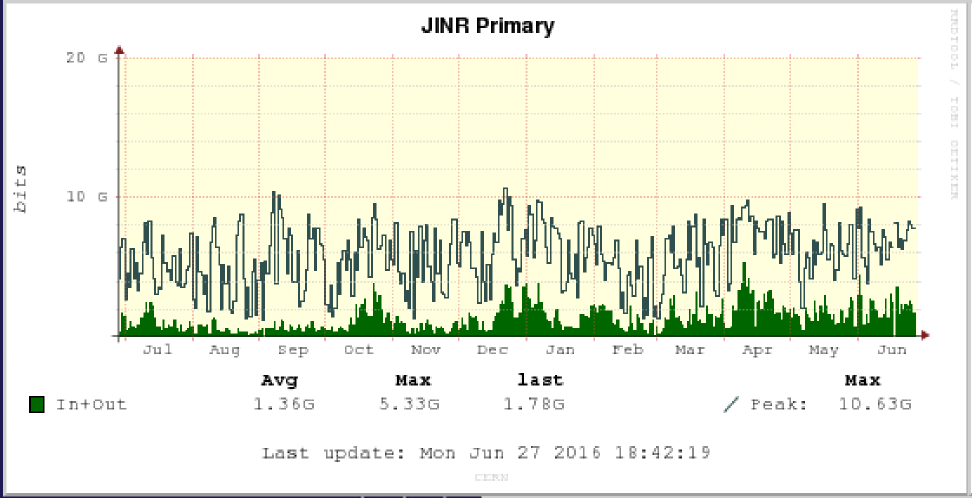
FR-CCIN2P3 🟥🟦🟩🟨
AS789
193.48.99.0/24
2001:660:5009:9::/64

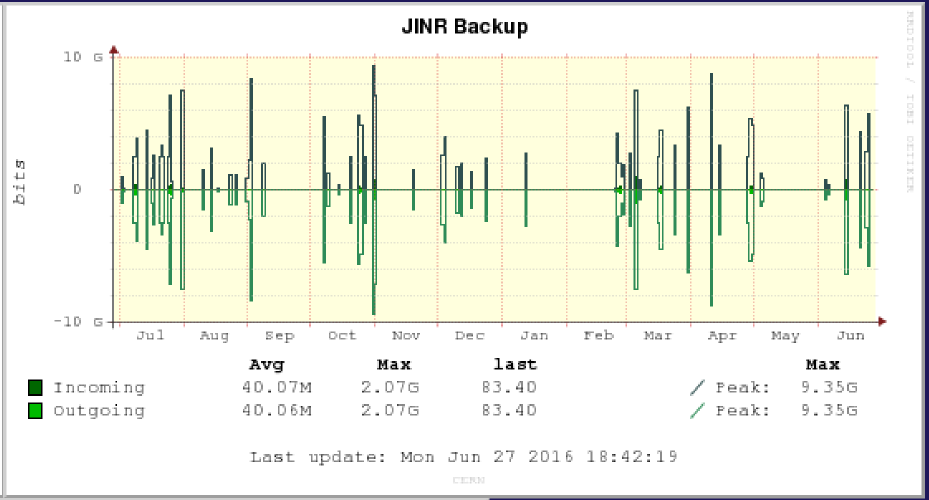
T0-T1 and T1-T1 traffic
T1-T1 traffic only
Not deployed yet
(thick) >=10Gbps
(thin) <10Gbps

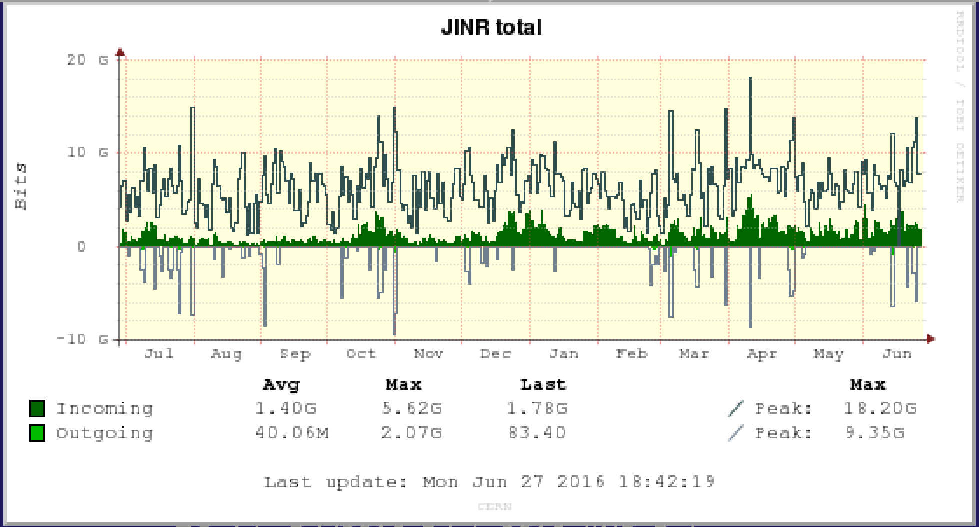
🟥 = Alice 🟦 = Atlas
🟩 = CMS 🟨 = LHCb

p2p prefix: 192.16.166.0/24 - 2001:1458:302::/48
edoardo.martelli@cern.ch 20160322

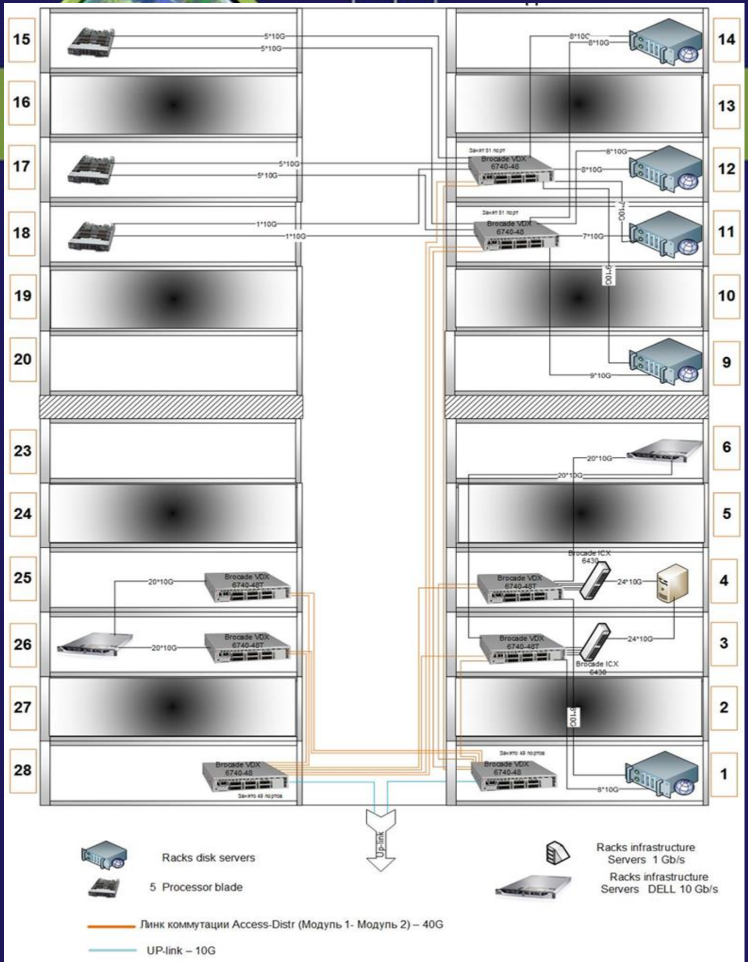
LHCOPN JINR-T1 Traffic (last year)

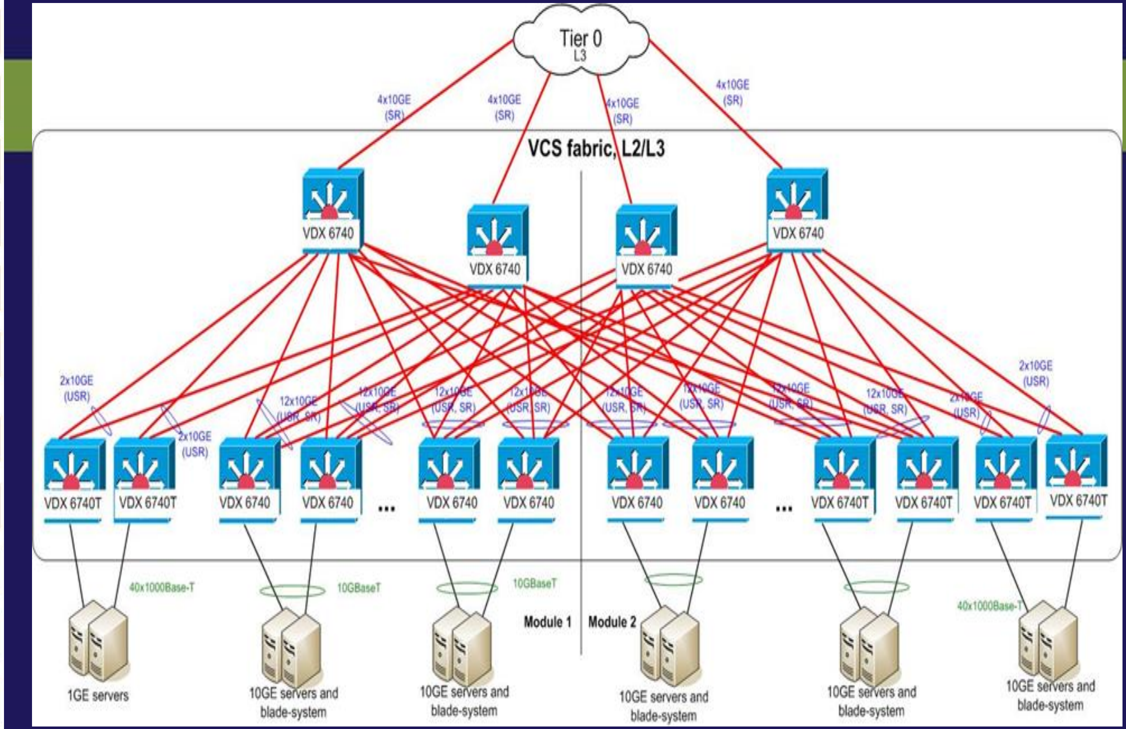






Tier-1 Network Structure





Three-layer architecture of the Tier-1 at JINR

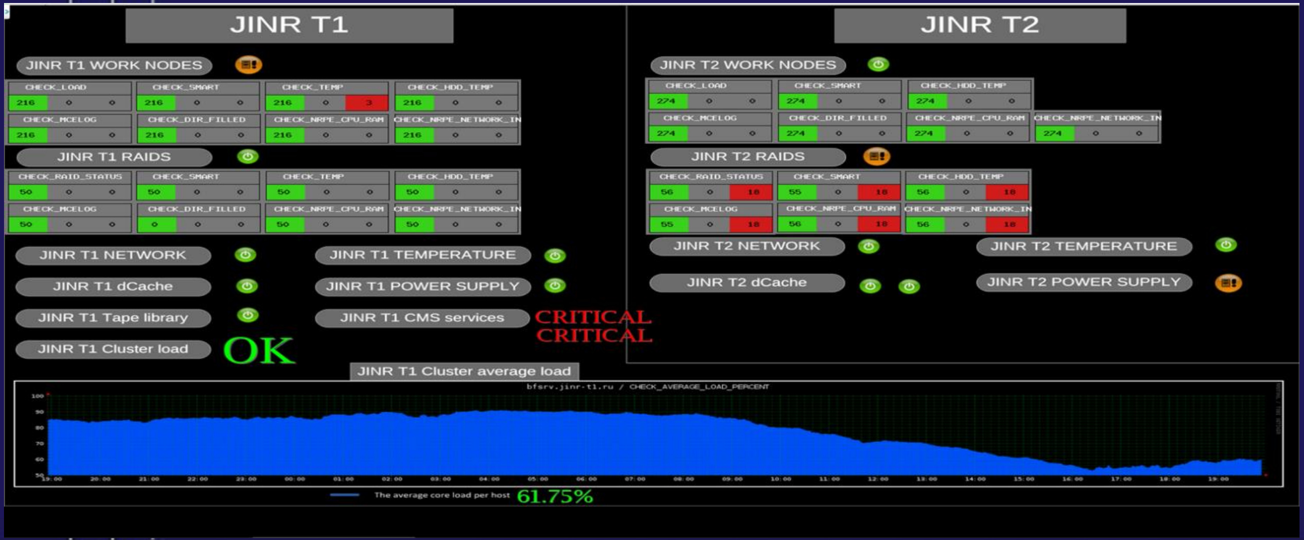
Network segment of Tier-1 module at JINR built of Brocade equipment

A diagram of a network architecture implemented Tier-1 module as of 2016. The module consists of twenty-four (24) racks, sixteen (16) of which are filled with a server equipment, and eight (8) allocated for the module cooling and to create the desired climate. Ten (10) racks are filled with disk servers. Three (3) racks are allocated by the computing blade servers. Servers that provide Grid infrastructure occupy three (3) racks.

The operation of switches assumes creation of a virtual factory, which combines up to 32 devices, capable of self-balancing traffic across all pathes.a

JINR Tier1- Tier2 Monitoring

1. Monitoring the states of all nodes and services- from the supply system to the robotized tape library
2. Global real time survey of the state of the whole computing complex
3. In case of emergency, alerts are sent to habilitated persons via e-mail, SMS, etc.
4. ~690 elements are under observation
5. ~ 3500 checks in real time



JINR T1

JINR T1 WORK NODES

CHECK_LOAD	216	0	0
CHECK_SMART	216	0	0
CHECK_TEMP	216	0	3
CHECK_HDD_TEMP	216	0	0
CHECK_MCELOG	216	0	0
CHECK_DIR_FILLED	216	0	0
CHECK_NRPE_CPU_RAM	216	0	0
CHECK_NRPE_NETWORK_IN	216	0	0

JINR T1 RAIDS

CHECK_RAID_STATUS	50	0	0
CHECK_SMART	50	0	0
CHECK_TEMP	50	0	0
CHECK_HDD_TEMP	50	0	0
CHECK_MCELOG	50	0	0
CHECK_DIR_FILLED	0	0	0
CHECK_NRPE_CPU_RAM	50	0	0
CHECK_NRPE_NETWORK_IN	50	0	0

JINR T1 NETWORK

JINR T1 TEMPERATURE

JINR T1 dCache

JINR T1 POWER SUPPLY

JINR T1 Tape library

JINR T1 CMS services CRITICAL CRITICAL

JINR T1 Cluster load OK

JINR T2

JINR T2 WORK NODES

CHECK_LOAD	274	0	0
CHECK_SMART	274	0	0
CHECK_HDD_TEMP	274	0	0
CHECK_MCELOG	274	0	0
CHECK_DIR_FILLED	274	0	0
CHECK_NRPE_CPU_RAM	274	0	0
CHECK_NRPE_NETWORK_IN	274	0	0

JINR T2 RAIDS

CHECK_RAID_STATUS	56	0	18
CHECK_SMART	55	0	18
CHECK_HDD_TEMP	56	0	18
CHECK_MCELOG	55	0	18
CHECK_NRPE_CPU_RAM	56	0	18
CHECK_NRPE_NETWORK_IN	56	0	18

JINR T2 NETWORK

JINR T2 TEMPERATURE

JINR T2 dCache

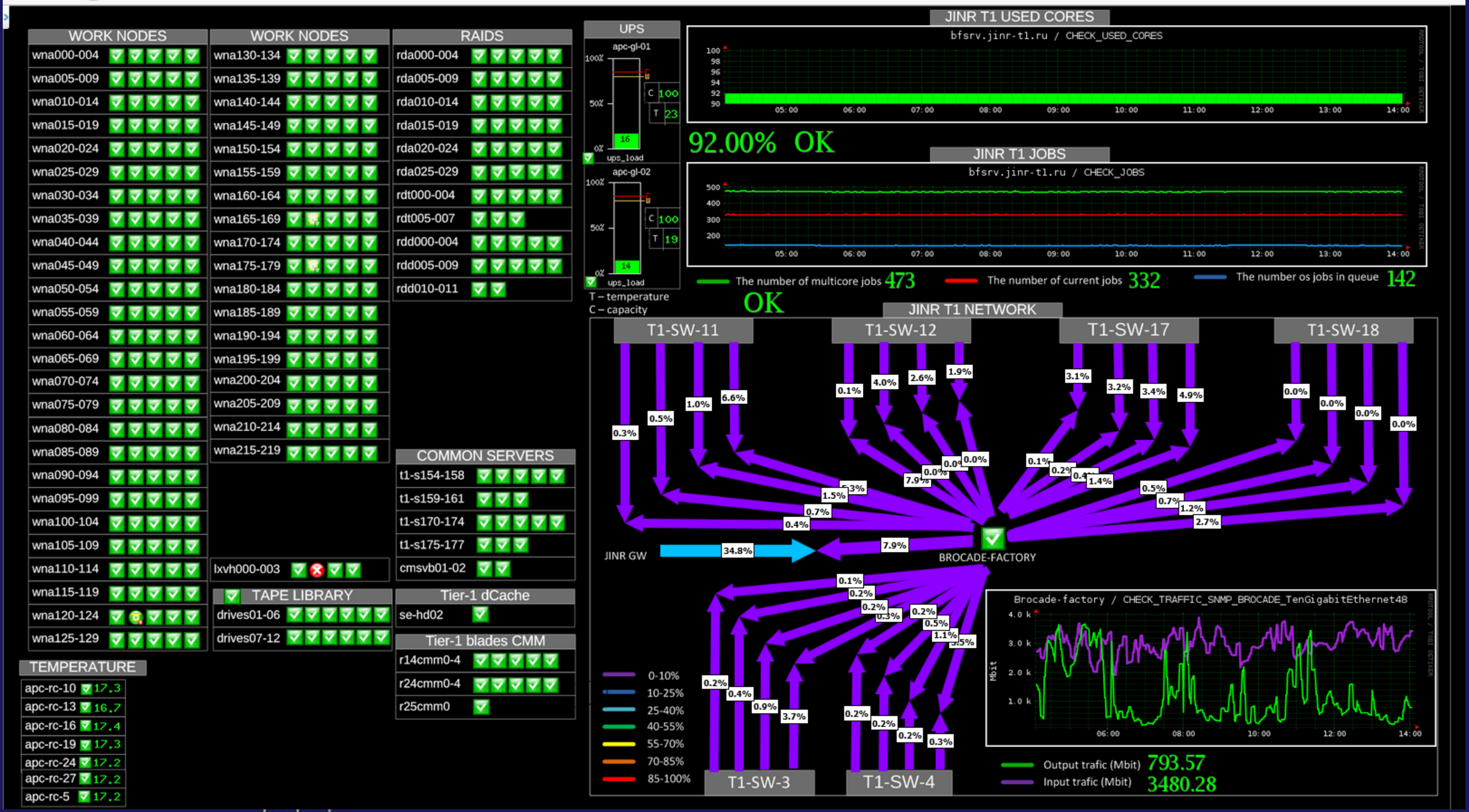
JINR T2 POWER SUPPLY

JINR T1 Cluster average load

bfsrv.jinr-t1.ru / CHECK_AVERAGE_LOAD_PERCENT

The average core load per host 61.75%

JINR Tier1 Dashboard



WORK NODES					
wna000-004	✓	✓	✓	✓	✓
wna005-009	✓	✓	✓	✓	✓
wna010-014	✓	✓	✓	✓	✓
wna015-019	✓	✓	✓	✓	✓
wna020-024	✓	✓	✓	✓	✓
wna025-029	✓	✓	✓	✓	✓
wna030-034	✓	✓	✓	✓	✓
wna035-039	✓	✓	✓	✓	✓
wna040-044	✓	✓	✓	✓	✓
wna045-049	✓	✓	✓	✓	✓
wna050-054	✓	✓	✓	✓	✓
wna055-059	✓	✓	✓	✓	✓
wna060-064	✓	✓	✓	✓	✓
wna065-069	✓	✓	✓	✓	✓
wna070-074	✓	✓	✓	✓	✓
wna075-079	✓	✓	✓	✓	✓
wna080-084	✓	✓	✓	✓	✓
wna085-089	✓	✓	✓	✓	✓
wna090-094	✓	✓	✓	✓	✓
wna095-099	✓	✓	✓	✓	✓
wna100-104	✓	✓	✓	✓	✓
wna105-109	✓	✓	✓	✓	✓
wna110-114	✓	✓	✓	✓	✓
wna115-119	✓	✓	✓	✓	✓
wna120-124	✓	⚠	✓	✓	✓
wna125-129	✓	✓	✓	✓	✓

WORK NODES					
wna130-134	✓	✓	✓	✓	✓
wna135-139	✓	✓	✓	✓	✓
wna140-144	✓	✓	✓	✓	✓
wna145-149	✓	✓	✓	✓	✓
wna150-154	✓	✓	✓	✓	✓
wna155-159	✓	✓	✓	✓	✓
wna160-164	✓	✓	✓	✓	✓
wna165-169	✓	⚠	✓	✓	✓
wna170-174	✓	✓	✓	✓	✓
wna175-179	✓	⚠	✓	✓	✓
wna180-184	✓	✓	✓	✓	✓
wna185-189	✓	✓	✓	✓	✓
wna190-194	✓	✓	✓	✓	✓
wna195-199	✓	✓	✓	✓	✓
wna200-204	✓	✓	✓	✓	✓
wna205-209	✓	✓	✓	✓	✓
wna210-214	✓	✓	✓	✓	✓
wna215-219	✓	✓	✓	✓	✓

lxvh000-003	✓	☒	✓	✓

RAIDS					
rda000-004	✓	✓	✓	✓	✓
rda005-009	✓	✓	✓	✓	✓
rda010-014	✓	✓	✓	✓	✓
rda015-019	✓	✓	✓	✓	✓
rda020-024	✓	✓	✓	✓	✓
rda025-029	✓	✓	✓	✓	✓
rdt000-004	✓	✓	✓	✓	✓
rdt005-007	✓	✓	✓		
rdd000-004	✓	✓	✓	✓	✓
rdd005-009	✓	✓	✓	✓	✓
rdd010-011	✓	✓			

COMMON SERVERS					
t1-s154-158	✓	✓	✓	✓	✓
t1-s159-161	✓	✓	✓		
t1-s170-174	✓	✓	✓	✓	✓
t1-s175-177	✓	✓	✓		
cmsvb01-02	✓	✓			

✓ TAPE LIBRARY						
drives01-06	✓	✓	✓	✓	✓	✓
drives07-12	✓	✓	✓	✓	✓	✓

Tier-1 dCache	
se-hd02	✓

Tier-1 blades CMM					
r14cmm0-4	✓	✓	✓	✓	✓
r24cmm0-4	✓	✓	✓	✓	✓
r25cmm0	✓				

TEMPERATURE		
apc-rc-10	✓	17.3
apc-rc-13	✓	16.7
apc-rc-16	✓	17.4
apc-rc-19	✓	17.3
apc-rc-24	✓	17.2
apc-rc-27	✓	17.2
apc-rc-5	✓	17.2

UPS

T – temperature
C – capacity

JINR T1 USED CORES

92.00% OK

JINR T1 JOBS

🟩 The number of multicore jobs 473 🟥 The number of current jobs 332 🟦 The number os jobs in queue 142

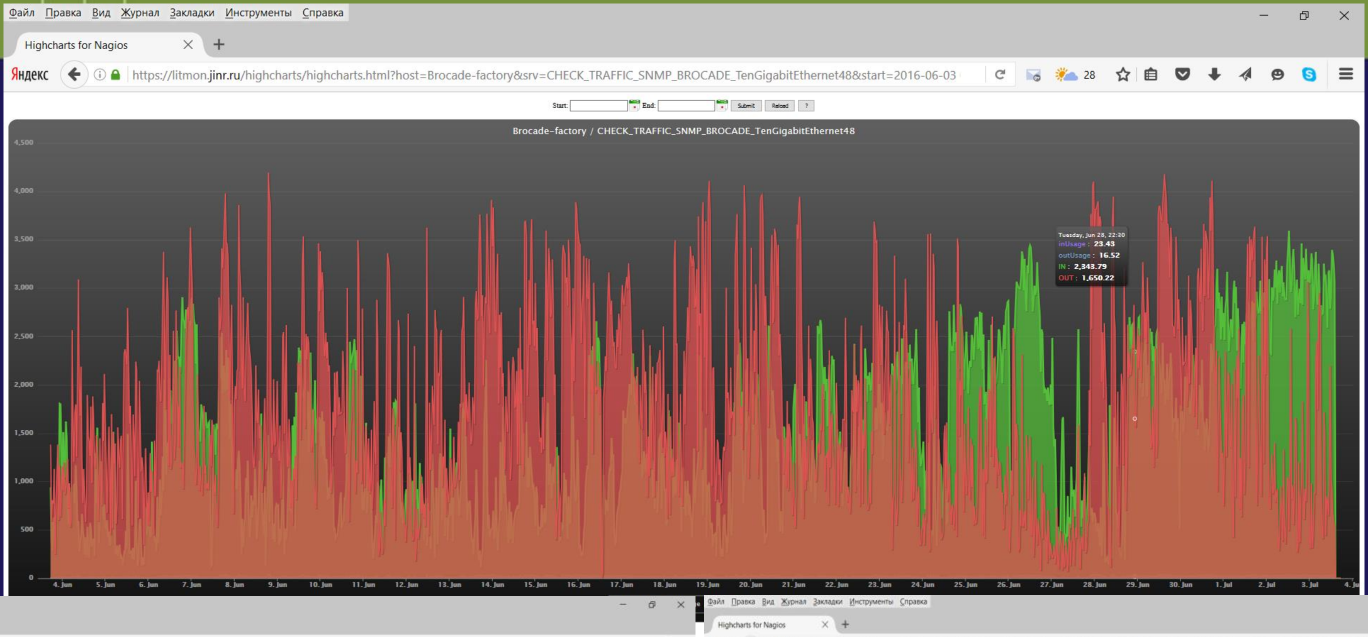
OK

JINR T1 NETWORK

Brocade-factory / CHECK_TRAFFIC_SNMP_BROCADE_TenGigabitEthernet48
Mbit
4.0 k
3.0 k
2.0 k
1.0 k
06:00
08:00
10:00
12:00
14:00

🟩 Output trafic (Mbit) 793.57
🟪 Input trafic (Mbit) 3480.28

Tier-1 Brocade Factory Traffic



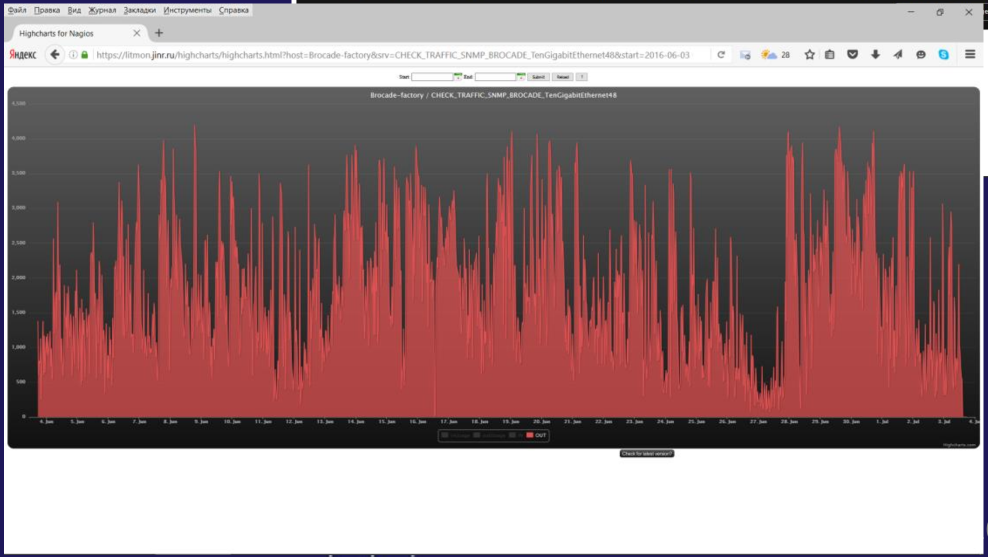
https://litmon.jinr.ru/highcharts/highcharts.html?host=Brocade-factory&srv=CHECK_TRAFFIC_SNMP_BROCADE_TenGigabitEthernet48&start=2016-06-03

Brocade-factory / CHECK_TRAFFIC_SNMP_BROCADE_TenGigabitEthernet48

Tuesday, Jun 28, 22:30
inUsage : 23.43
outUsage : 16.52
IN : 2,343.79
OUT : 1,650.22

Output traffic

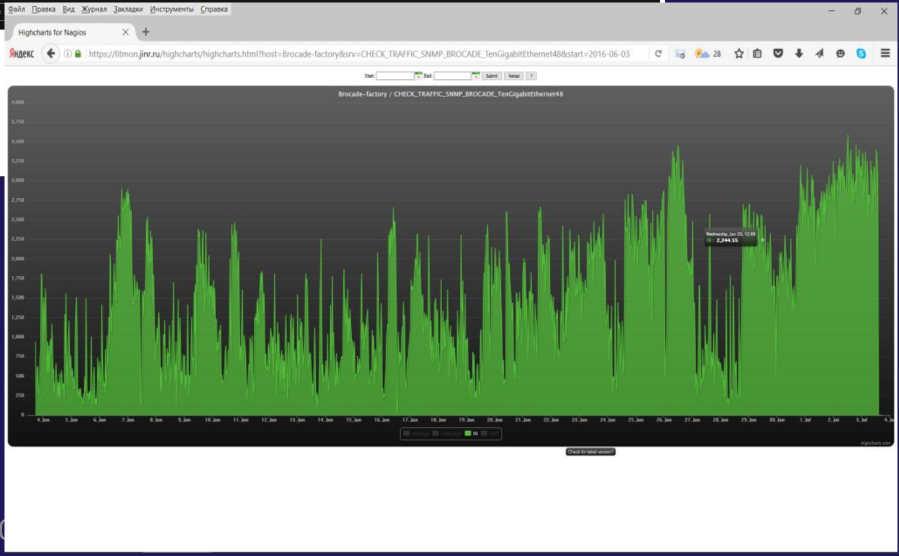
Input traffic



Файл Правка Вид Журнал Закладки Инструменты Справка
Highcharts for Nagios
Яндекс
Start:
End:
Submit
Reload
?
OUT
Check for latest version?

https://litmon.jinr.ru/highcharts/highcharts.html?host=Brocade-factory&srv=CHECK_TRAFFIC_SNMP_BROCADE_TenGigabitEthernet48&start=2016-06-03

Brocade-factory / CHECK_TRAFFIC_SNMP_BROCADE_TenGigabitEthernet48

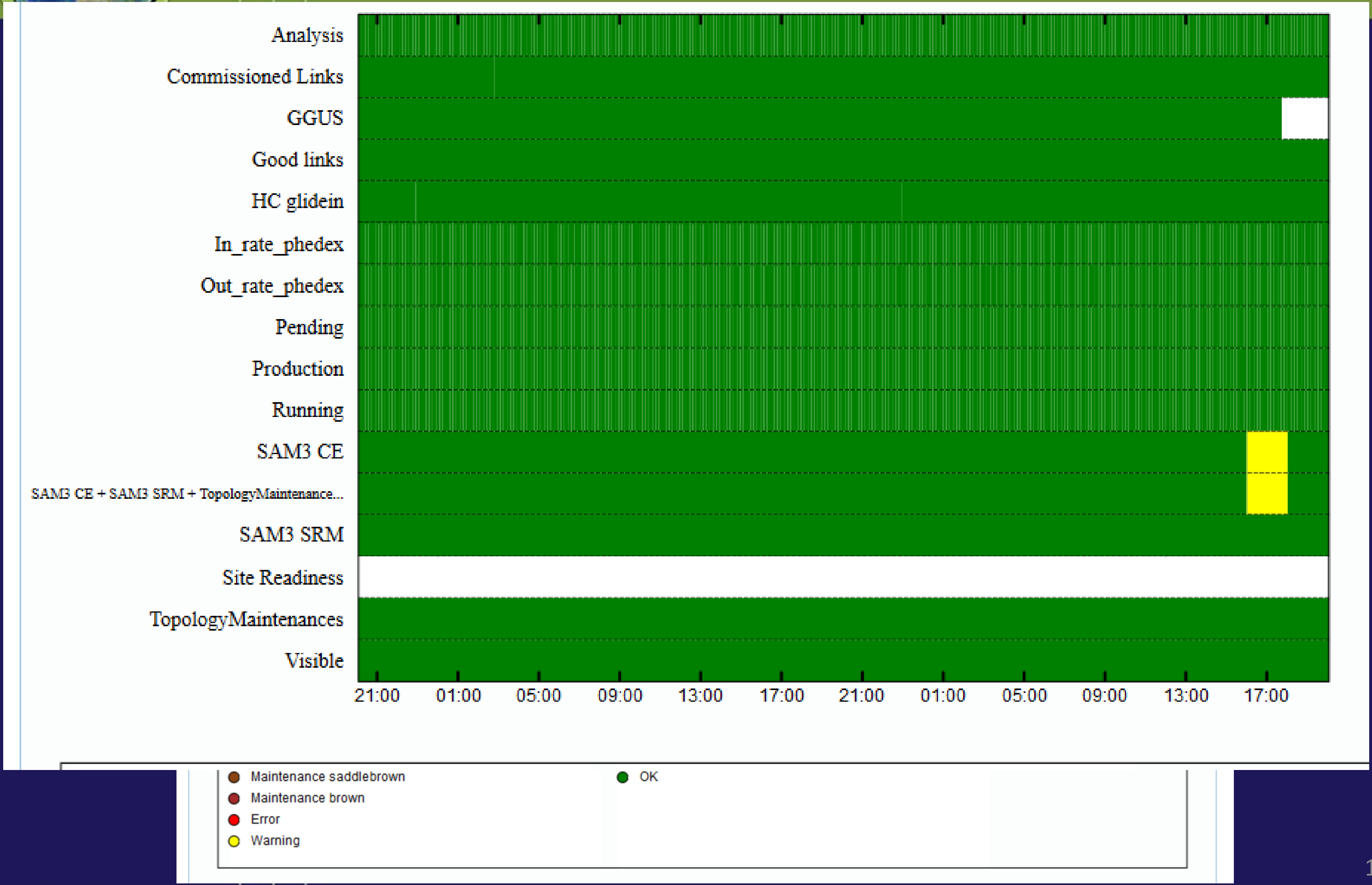


Файл Правка Вид Журнал Закладки Инструменты Справка
Highcharts for Nagios
Яндекс
Start:
End:
Submit
Reload
?
IN
Check for latest version?

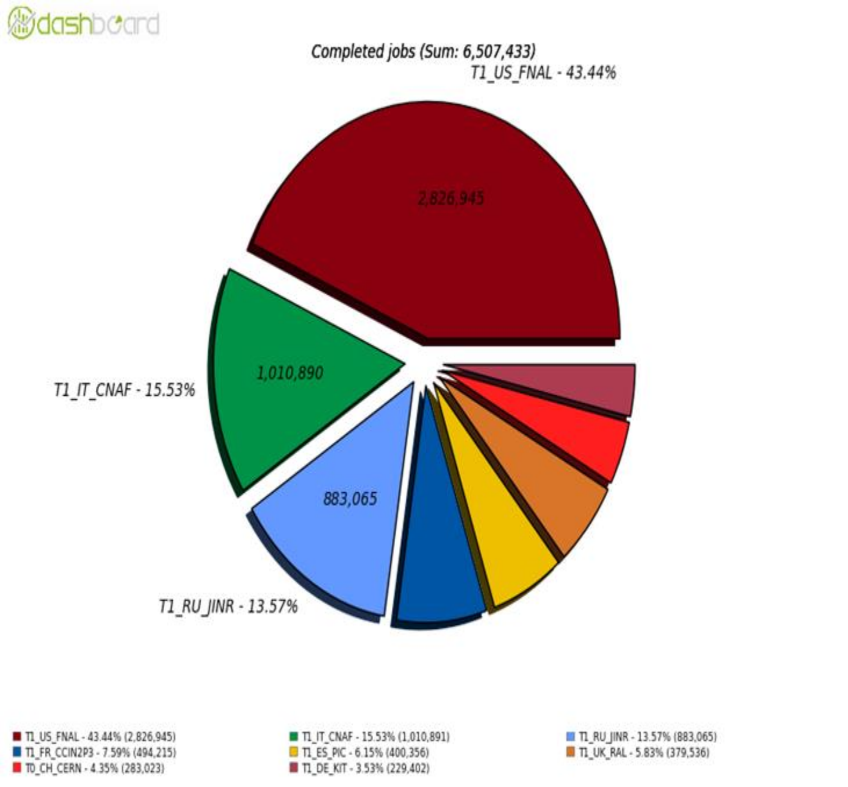
https://litmon.jinr.ru/highcharts/highcharts.html?host=Brocade-factory&srv=CHECK_TRAFFIC_SNMP_BROCADE_TenGigabitEthernet48&start=2016-06-03

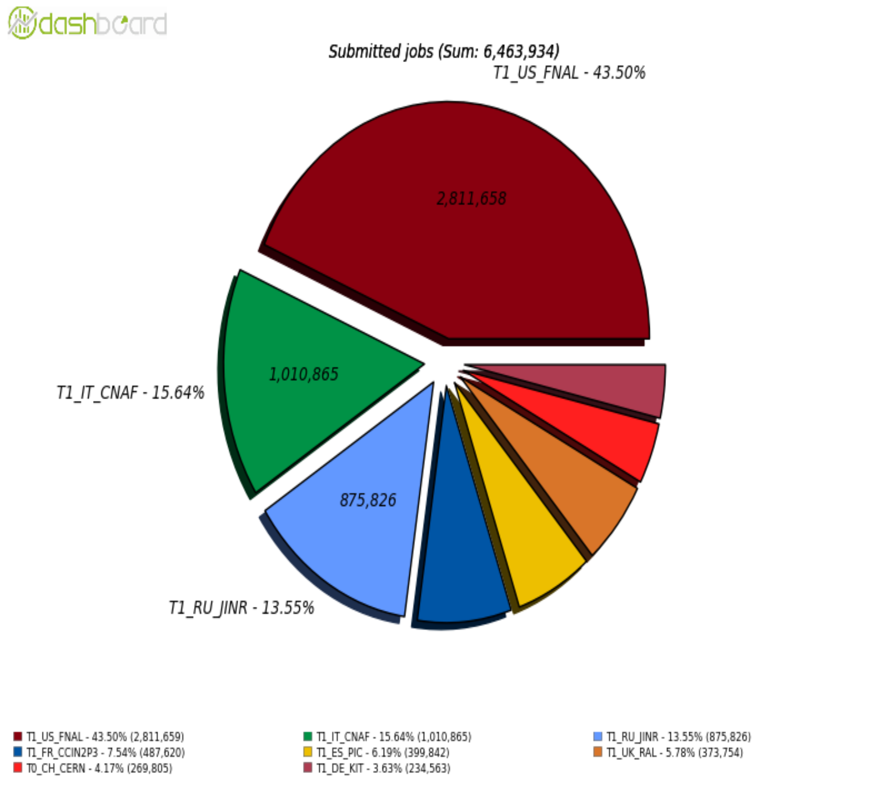
Brocade-factory / CHECK_TRAFFIC_SNMP_BROCADE_TenGigabitEthernet48

Default Metrics for the site T1_RU_JINR

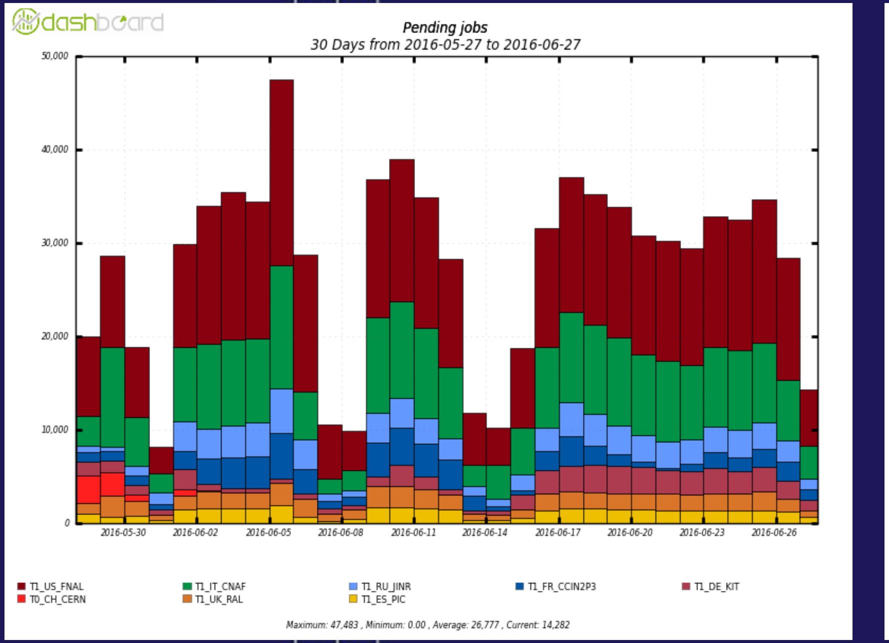


Last month jobs



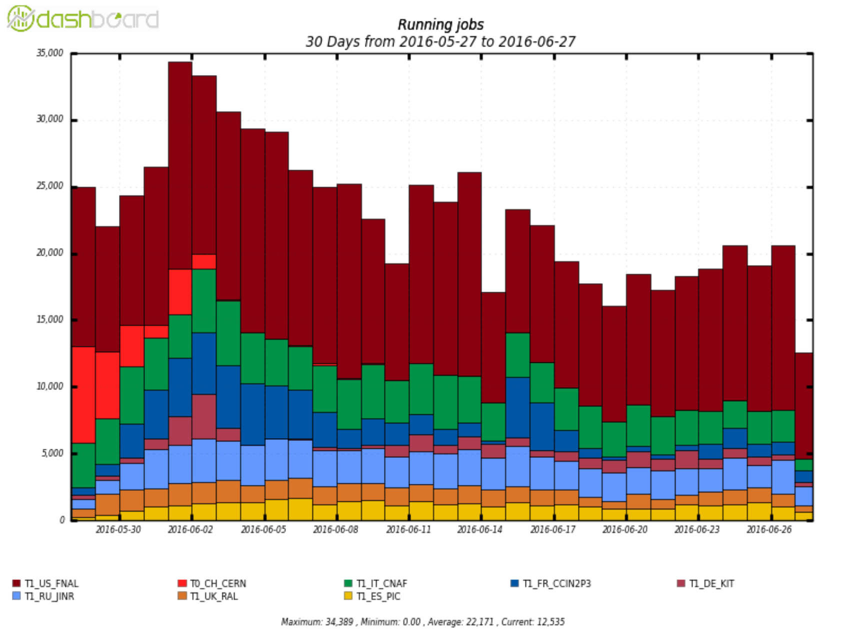


Last month pending and running jobs


Pending jobs

30 Days from 2016-05-27 to 2016-06-27

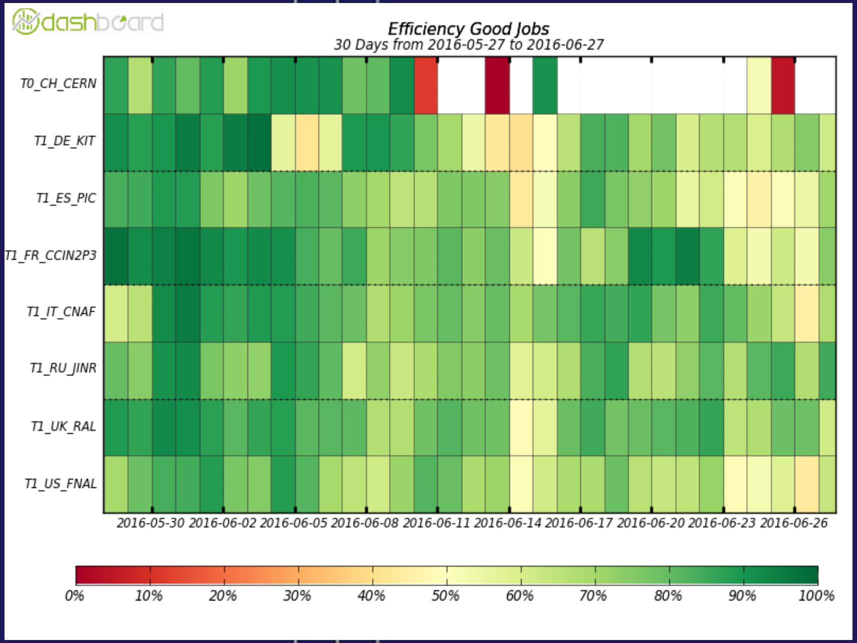
Maximum: 47,483 , Minimum: 0.00 , Average: 26,777 , Current: 14,282

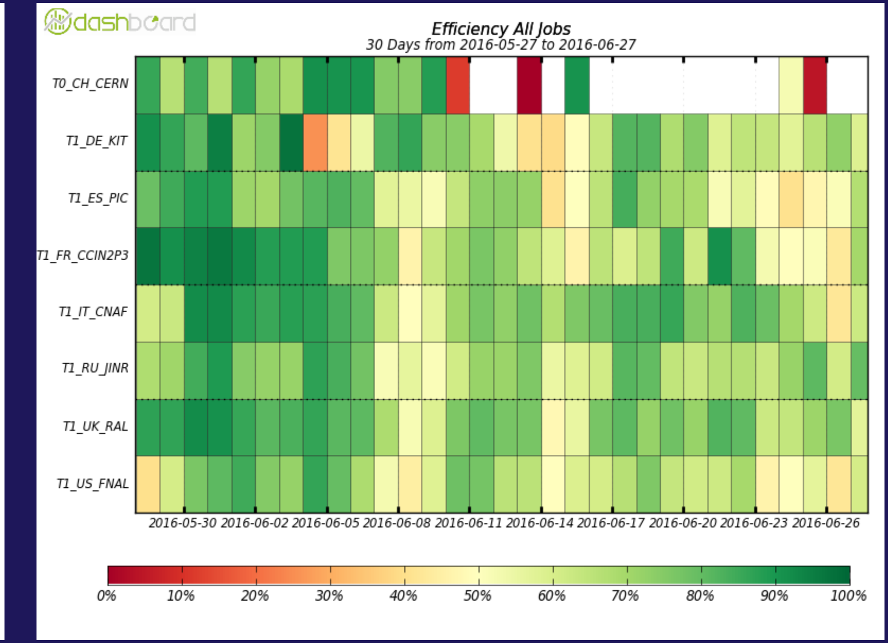

Running jobs

30 Days from 2016-05-27 to 2016-06-27

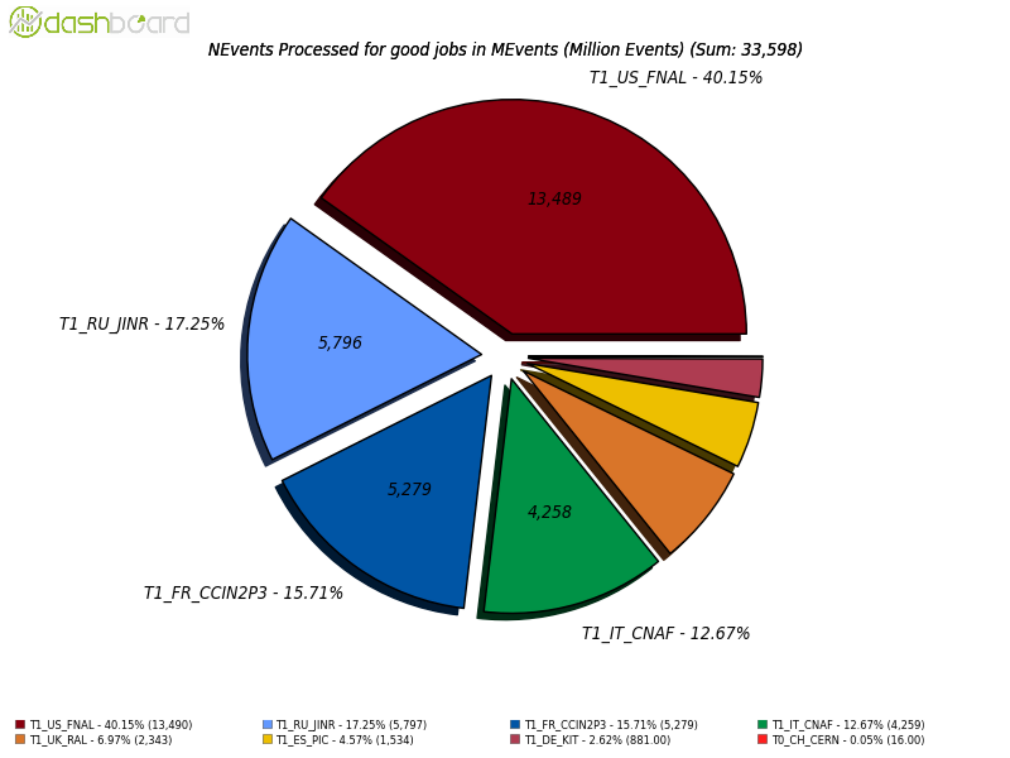
Maximum: 34,389 , Minimum: 0.00 , Average: 22,171 , Current: 12,535

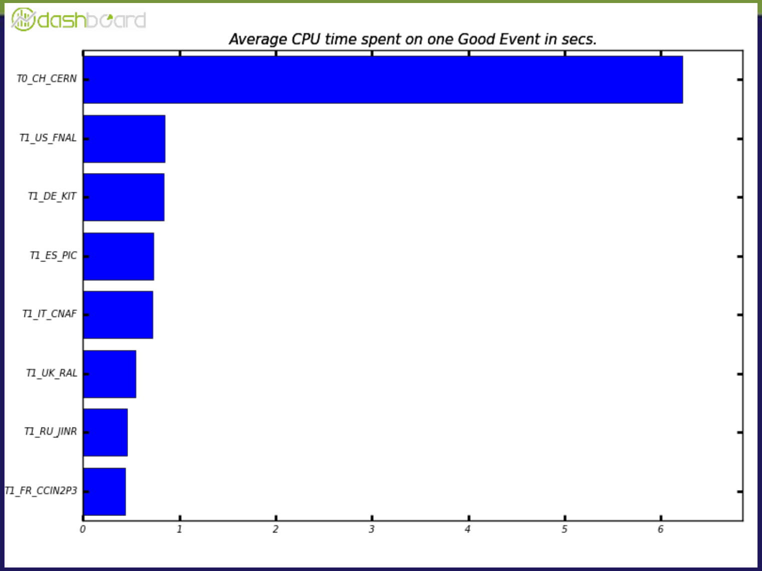
Efficiency good&all jobs

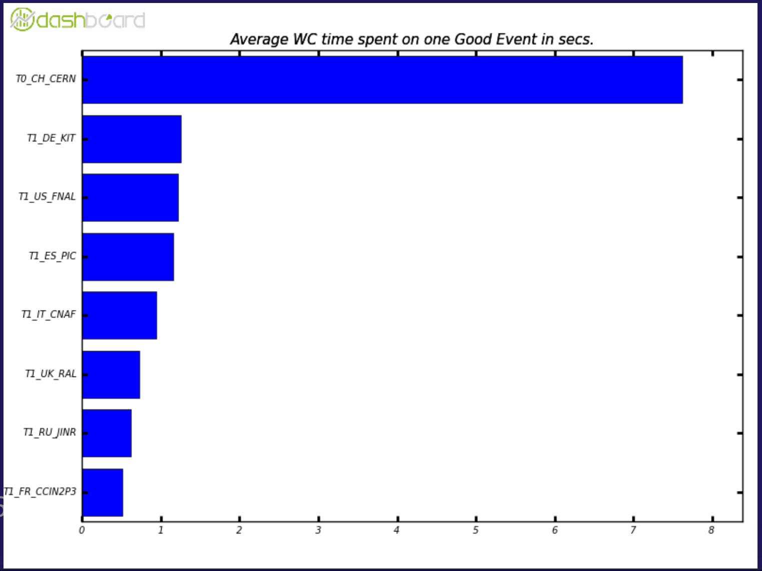




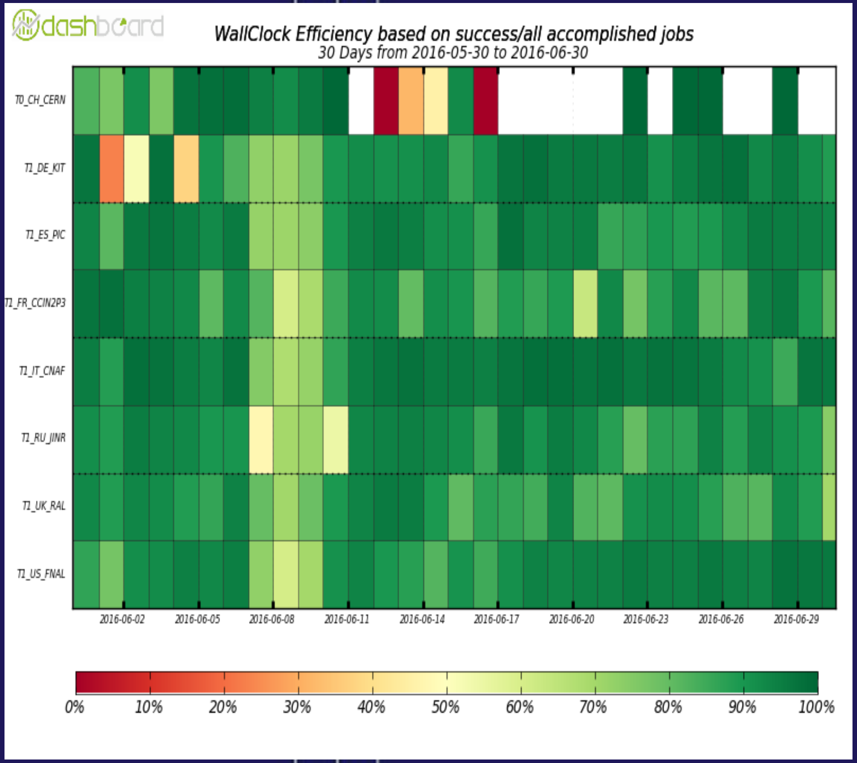
Last month events processed

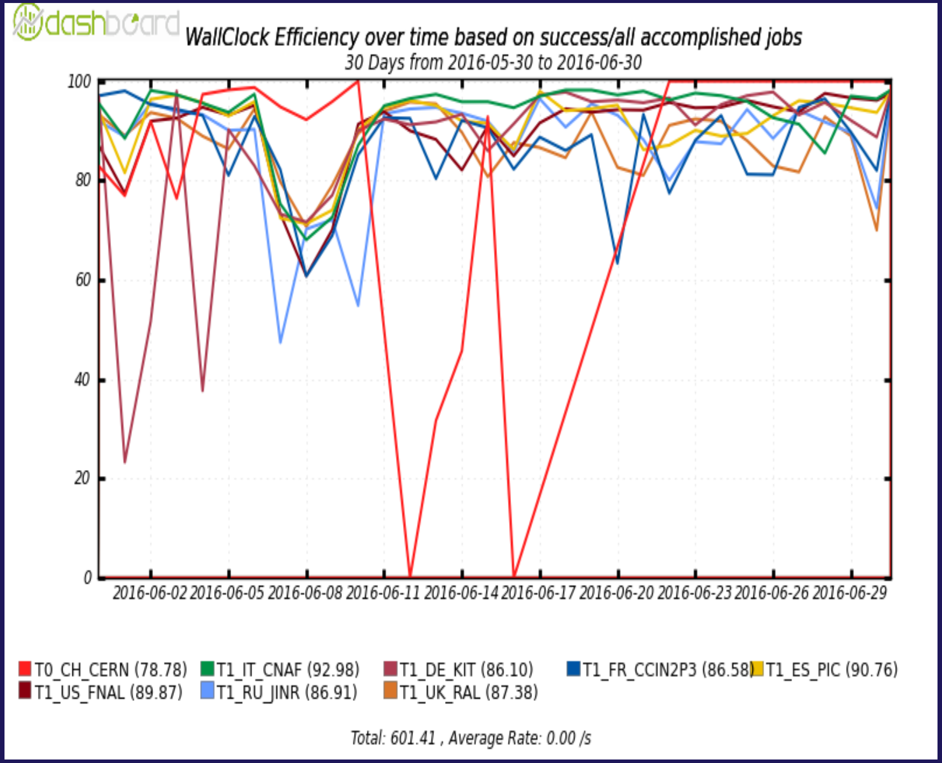






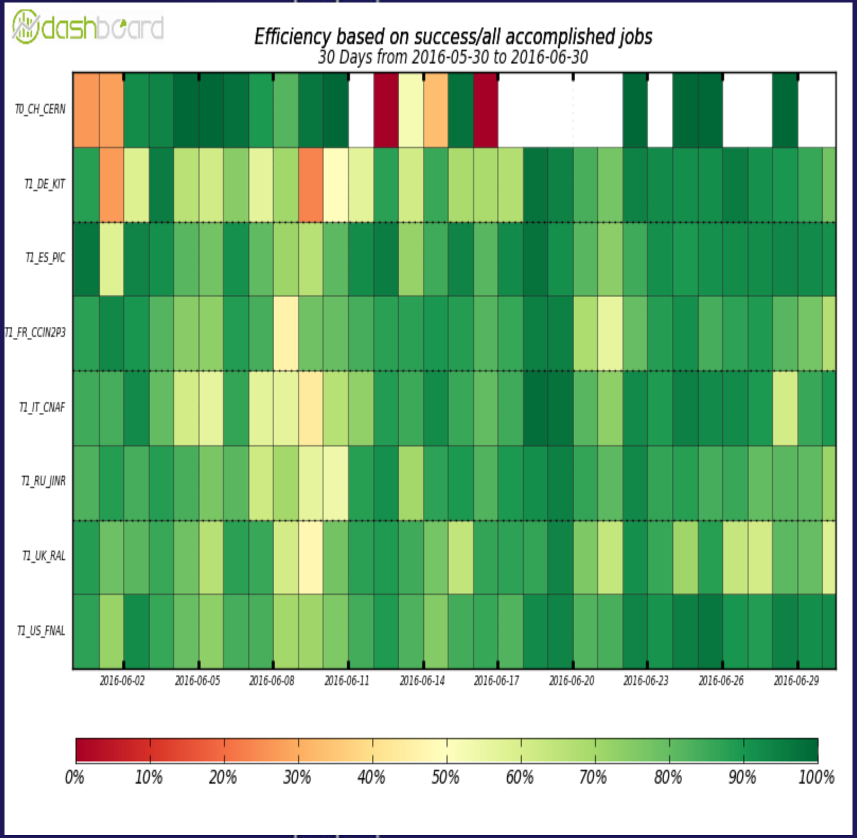
WallClock Efficiency based on success/all accomplished jobs

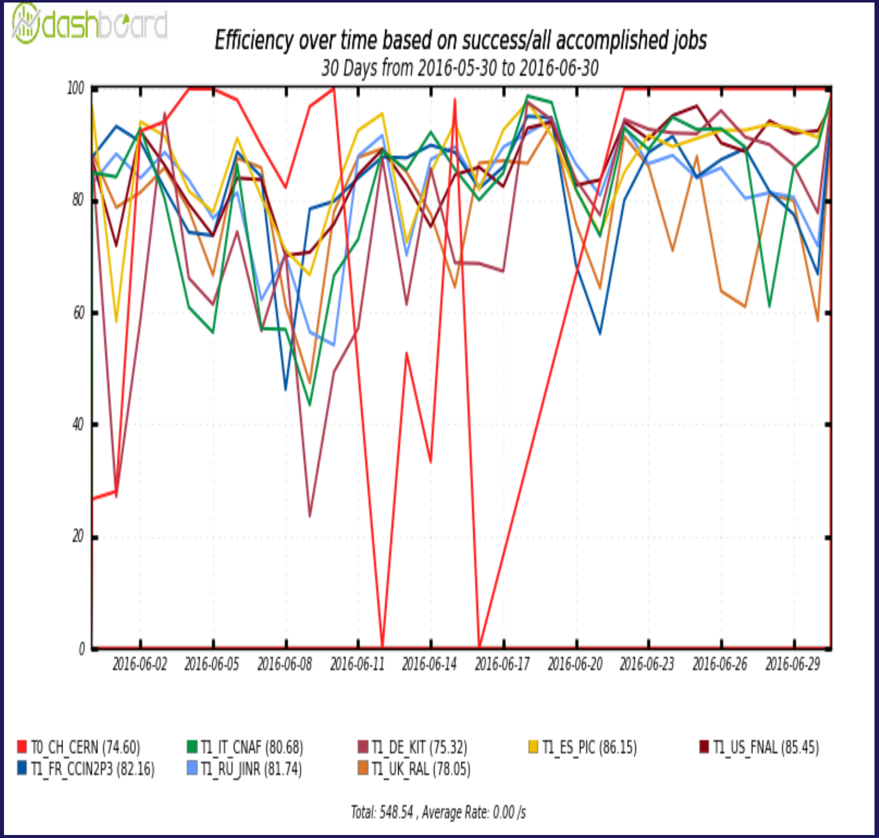

WallClock Efficiency based on success/all accomplished jobs
30 Days from 2016-05-30 to 2016-06-30


WallClock Efficiency over time based on success/all accomplished jobs
30 Days from 2016-05-30 to 2016-06-30

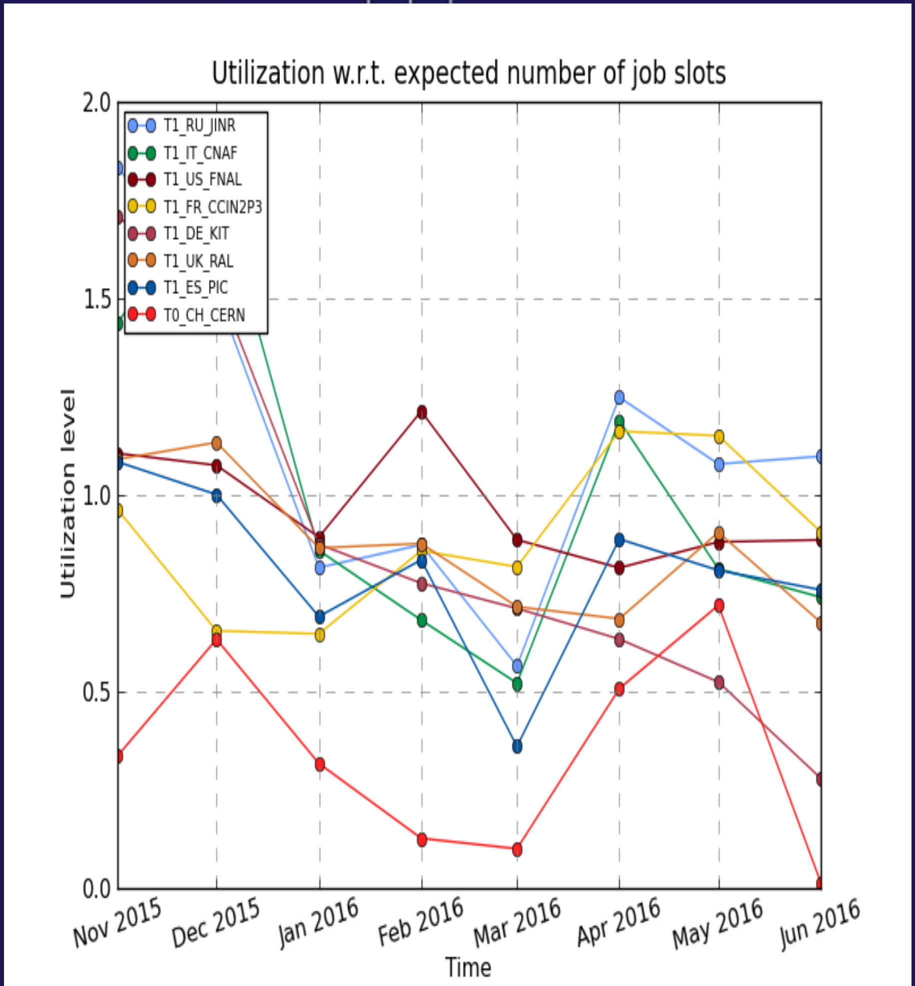
Total: 601.41 , Average Rate: 0.00 /s

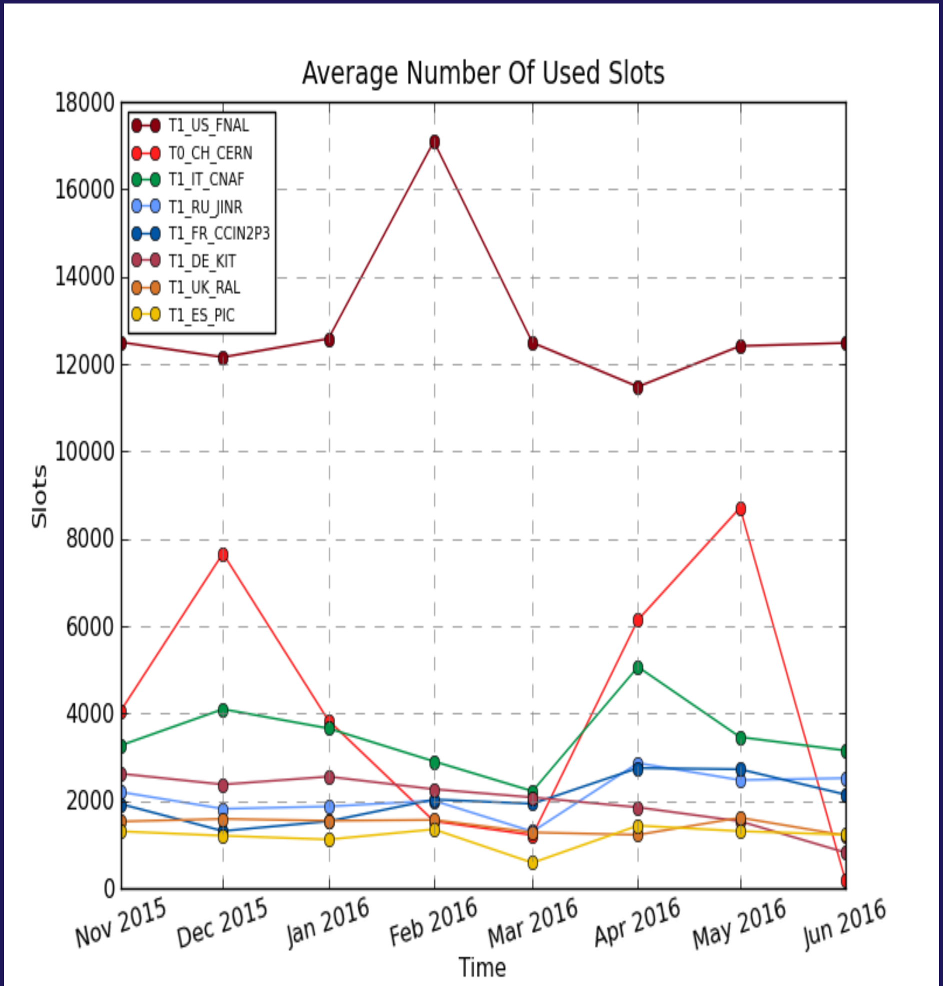
Efficiency based on success/all accomplished jobs



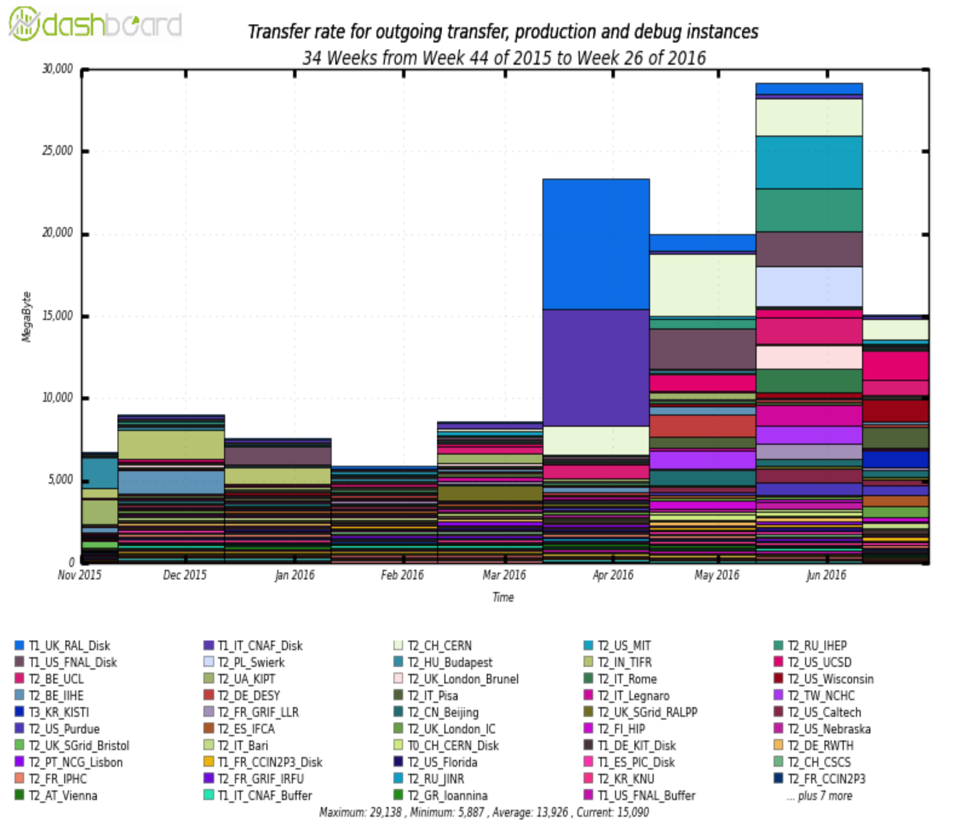


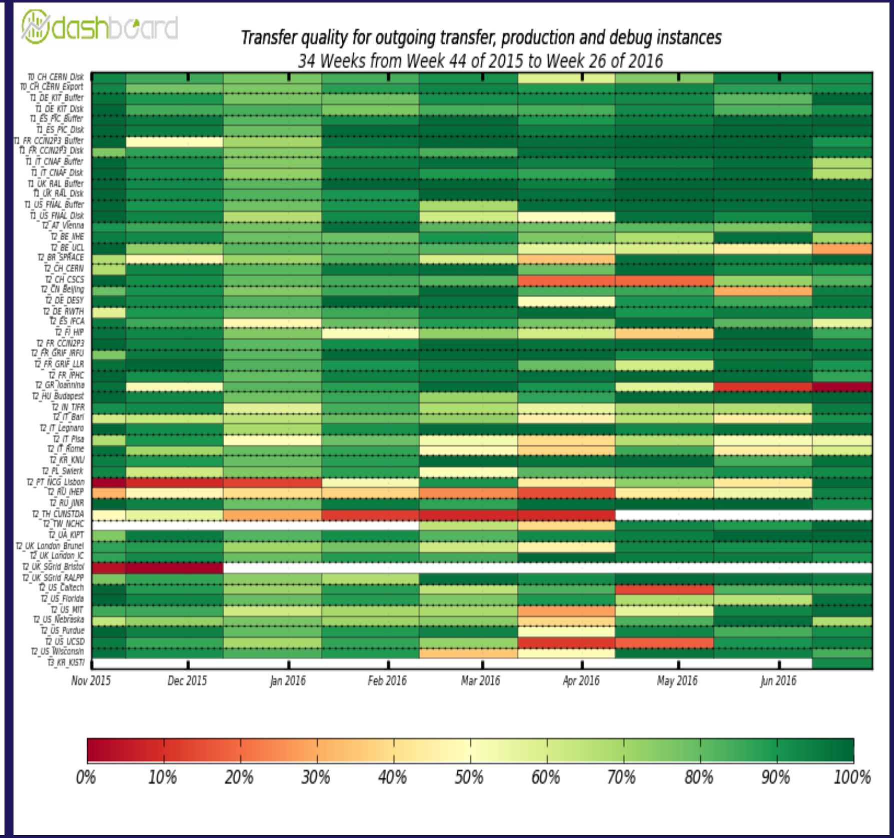
Resource Utilization



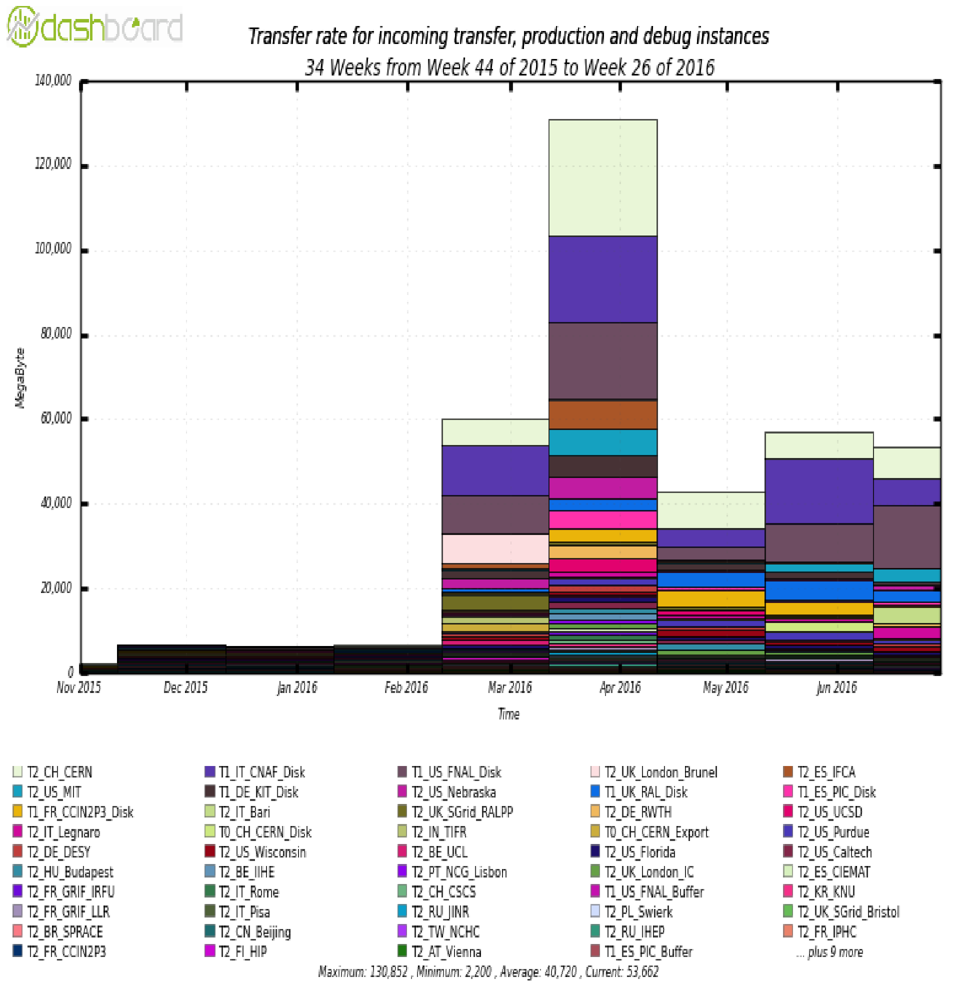


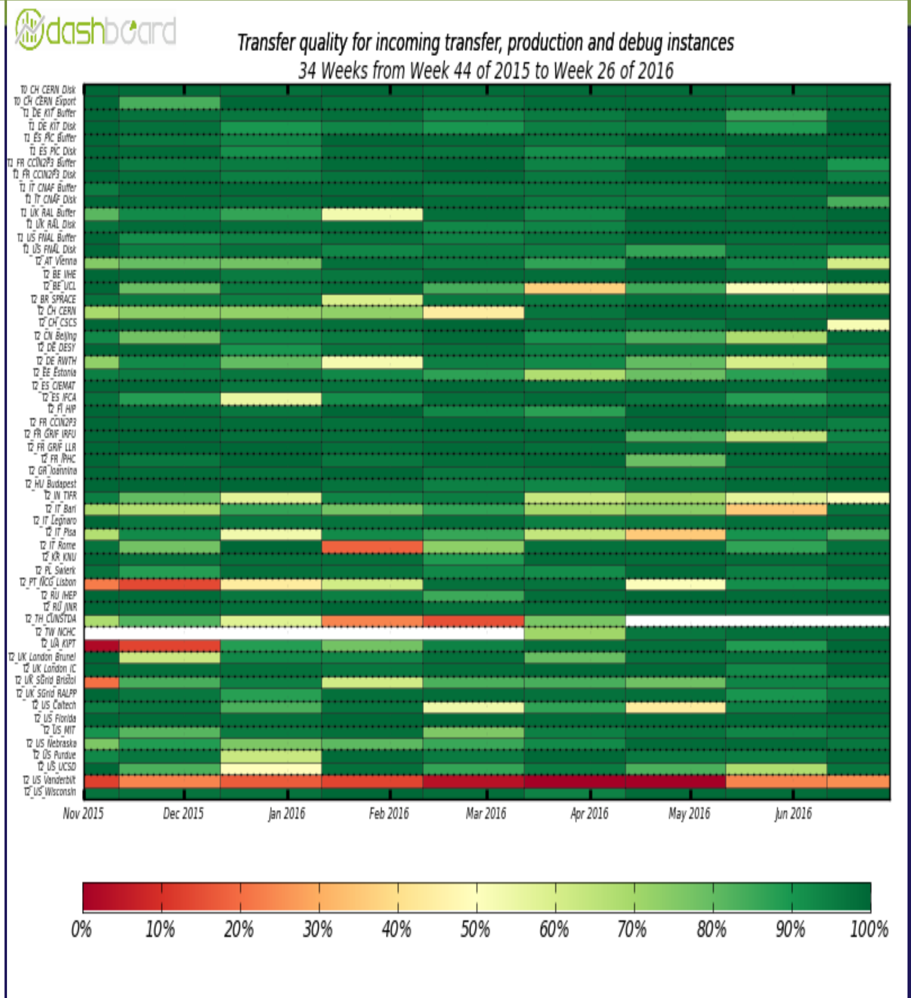
Data transfer from the JINR T1 via Production and Debug Phedex instances



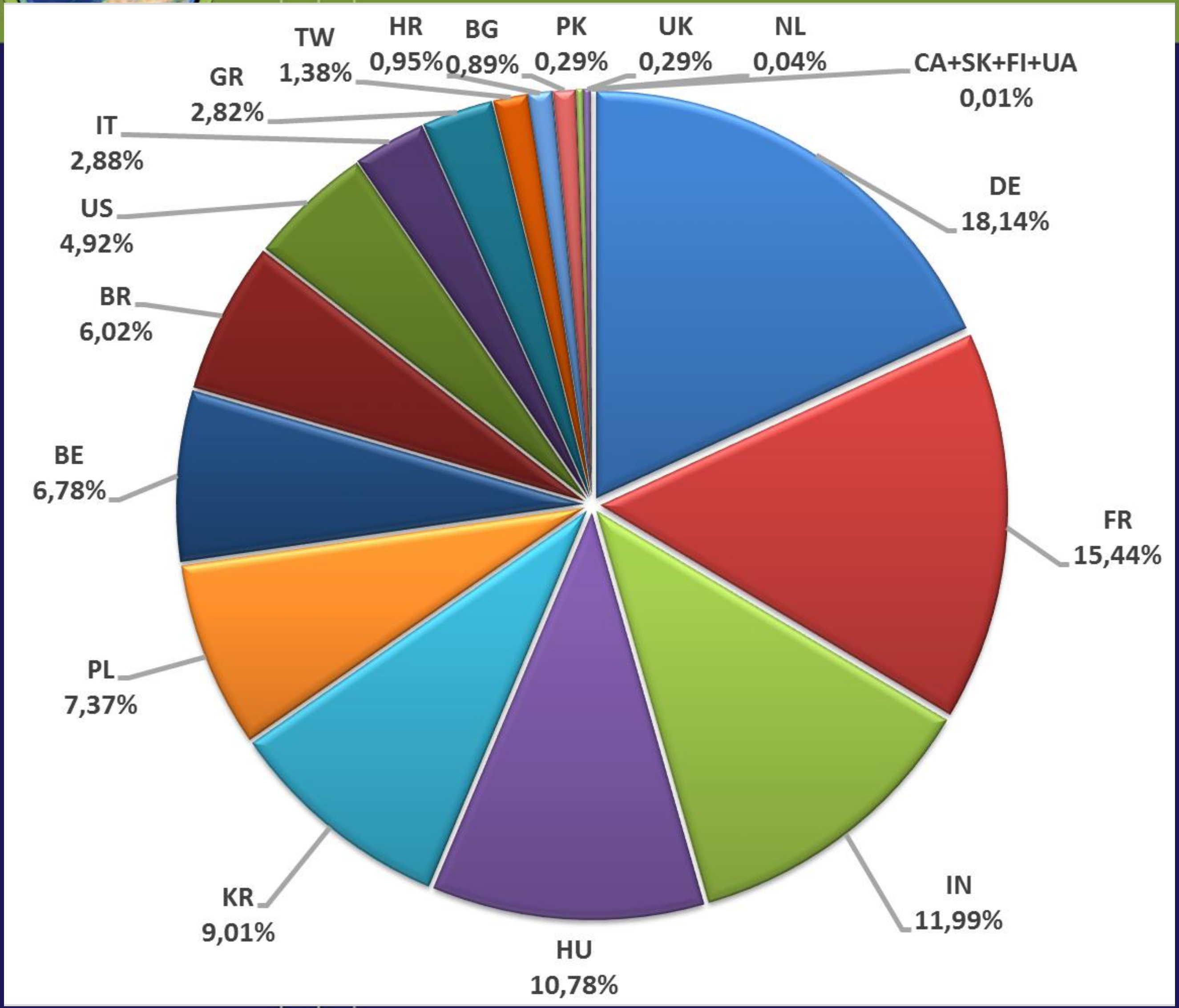


Data transfer to the JINR T1 via Production and Debug Phedex instances



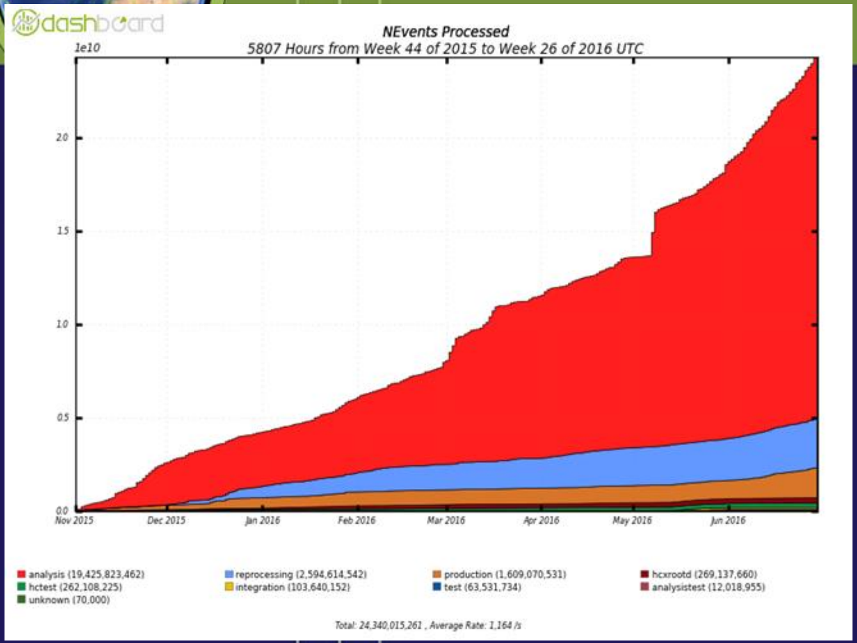


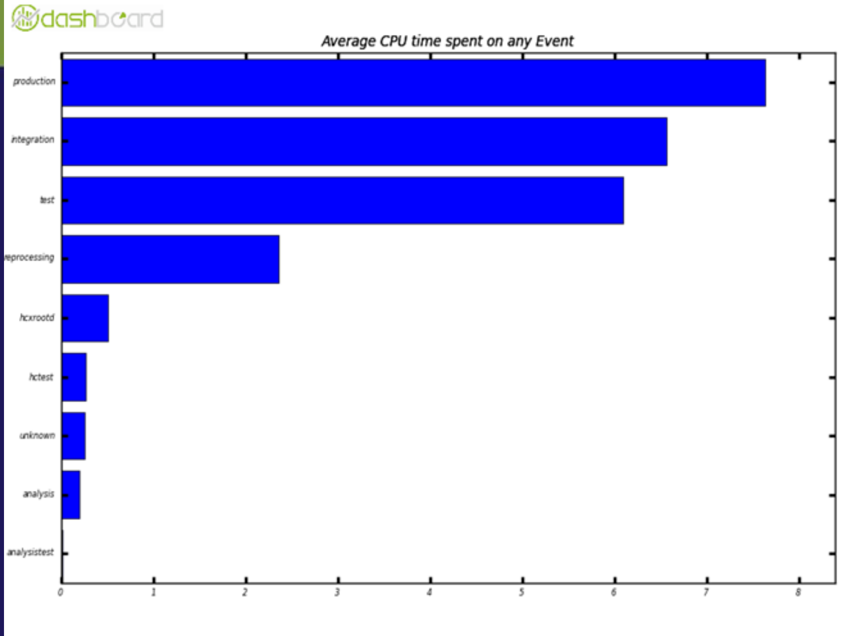
Last month requests to JINR-T1

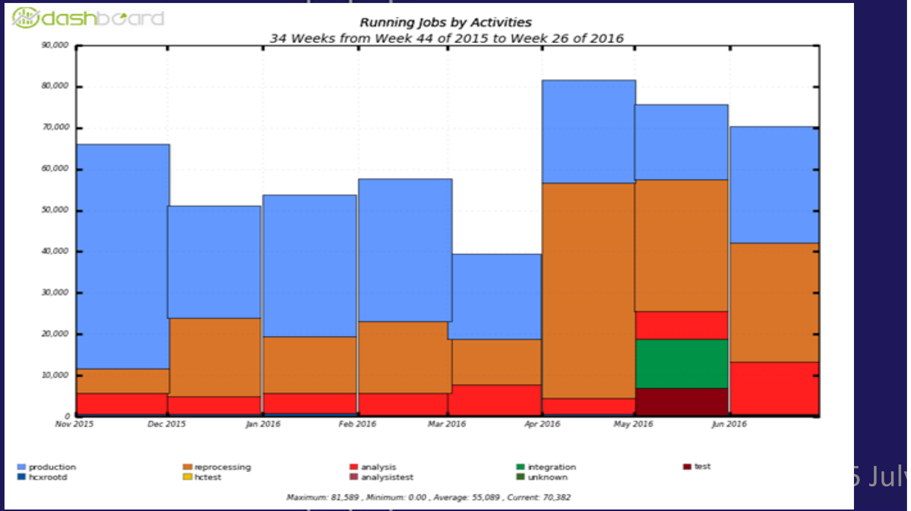


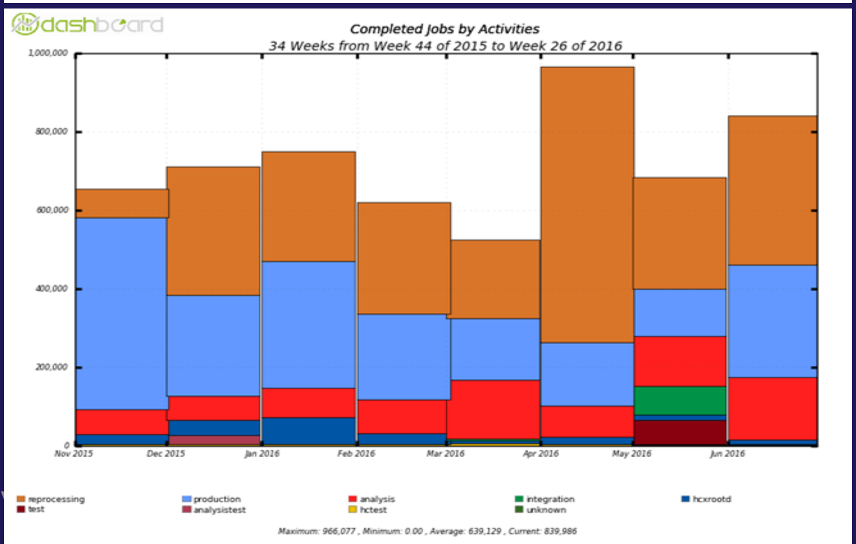
Raw data transfers to JINR-T1:
250-300 MBps
>1 TB/hour
~30TB/day

Job processing by activities

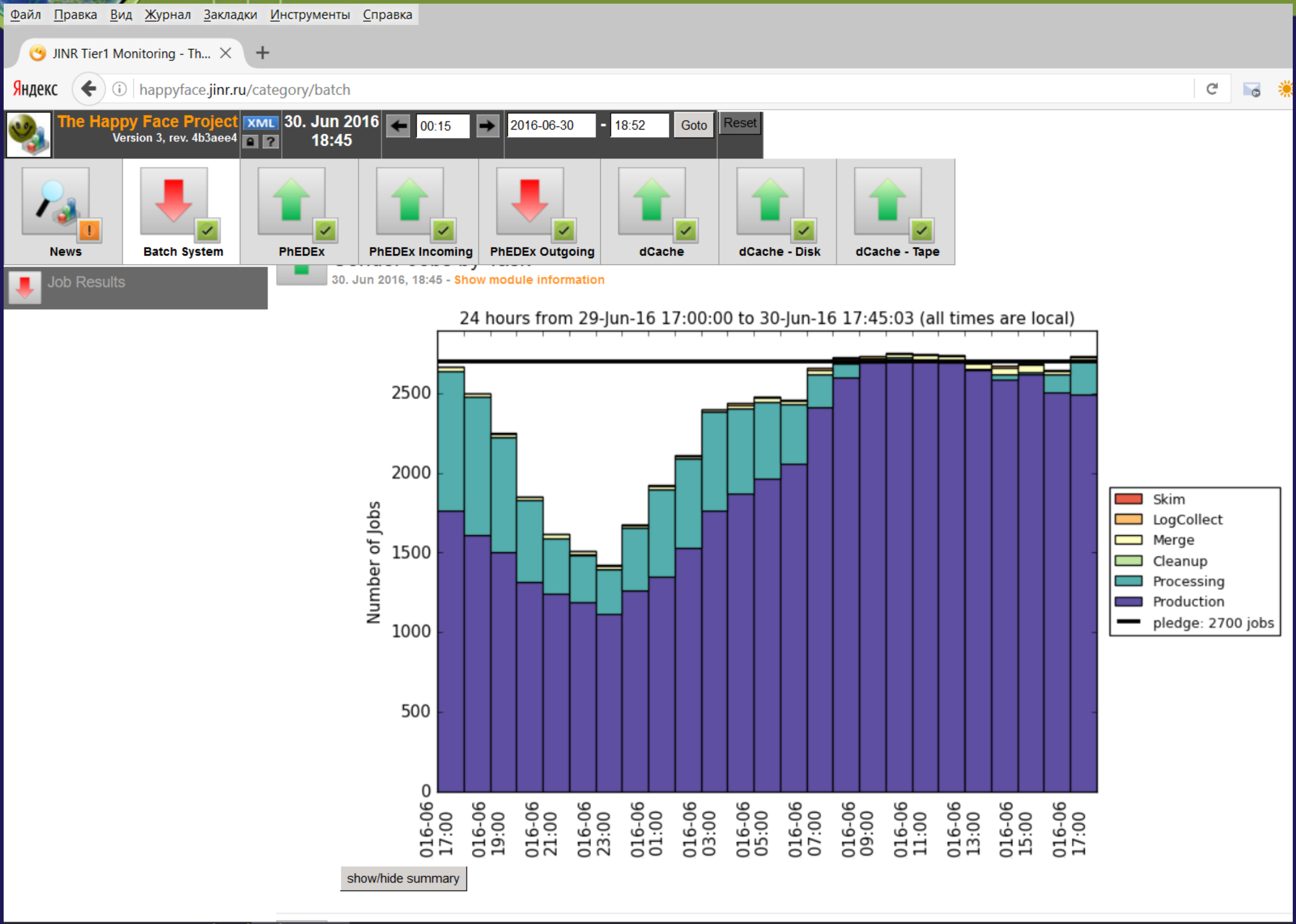








Happyface installation at JINR



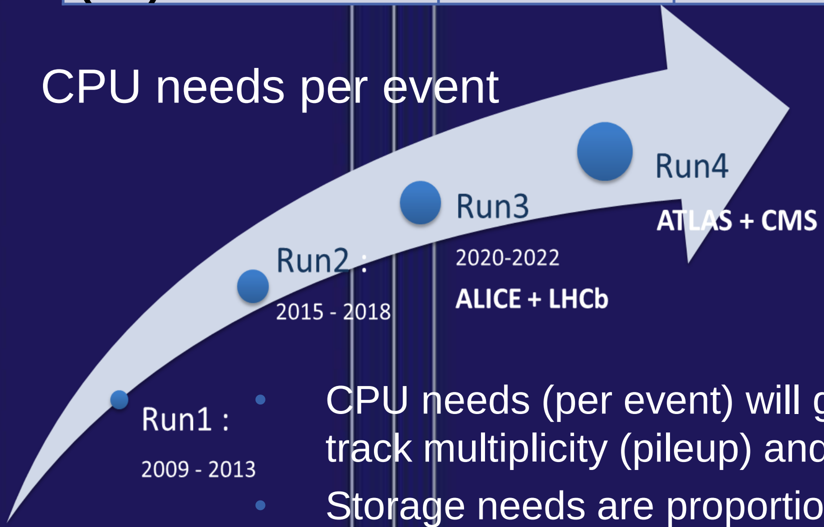
30. Jun 2016, 18:45 - Show module information

Development plans for the Tier-1 centre

Planned yearly growth of Tier-1 resources – absolute values and percentage growth over the previous year.

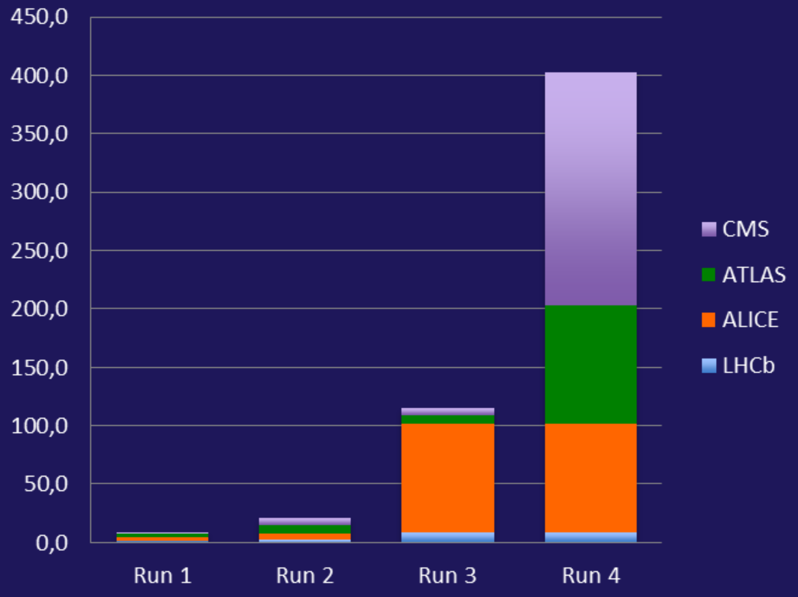
	2016	**2017**	**2018**	**2019**
Processor capacity of the core/kHS06	3400/54,4	4200/67,2 (24%)	5200/83,2 (23%)	10000/160 (52%)
Disk storage (TB)	3390	5070 (49%)	6100 (20%)	8000 (80%)
Tape storage (TB)	10000	20000	20000	20000

CPU needs per event



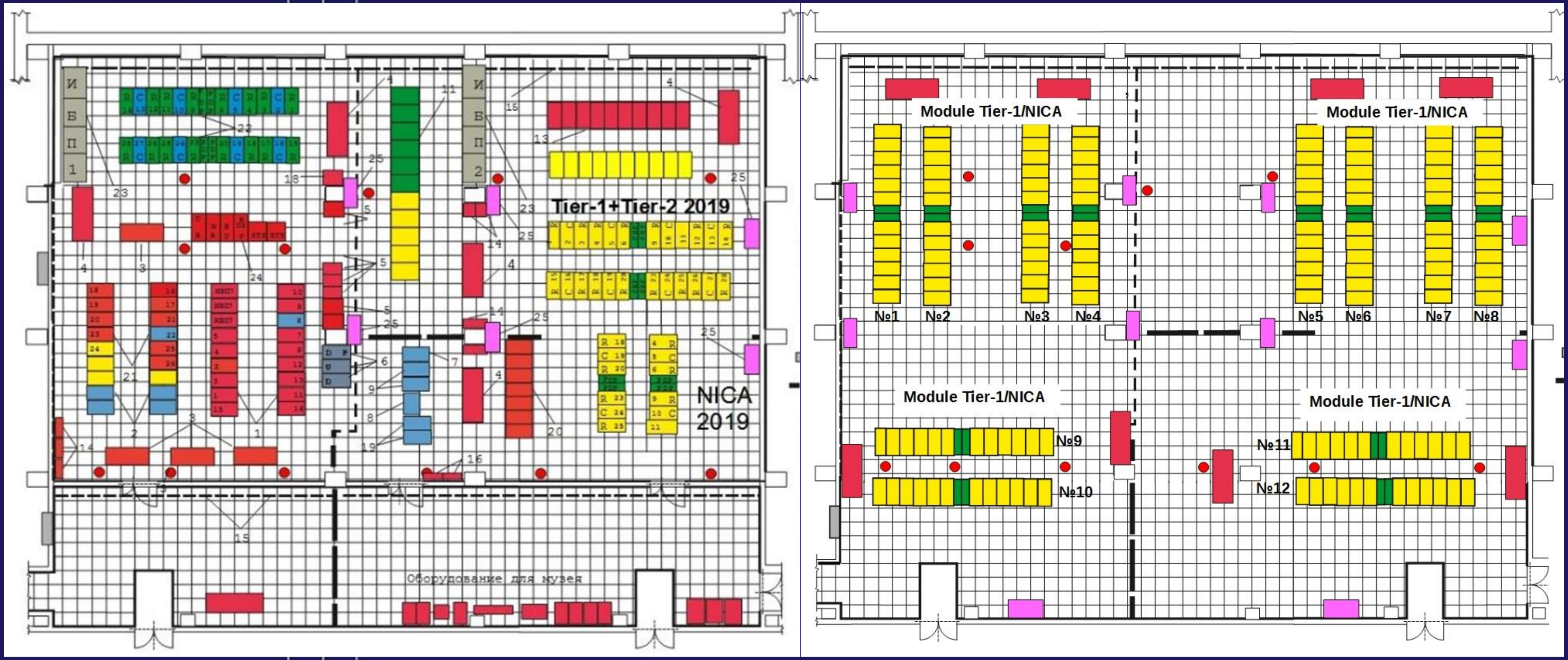
- CPU needs (per event) will grow with track multiplicity (pileup) and energy
- Storage needs are proportional to accumulated luminosity
- Grid resources are limited by funding and fully utilized

LHC Upgrade 2019-2021. Computing Needs



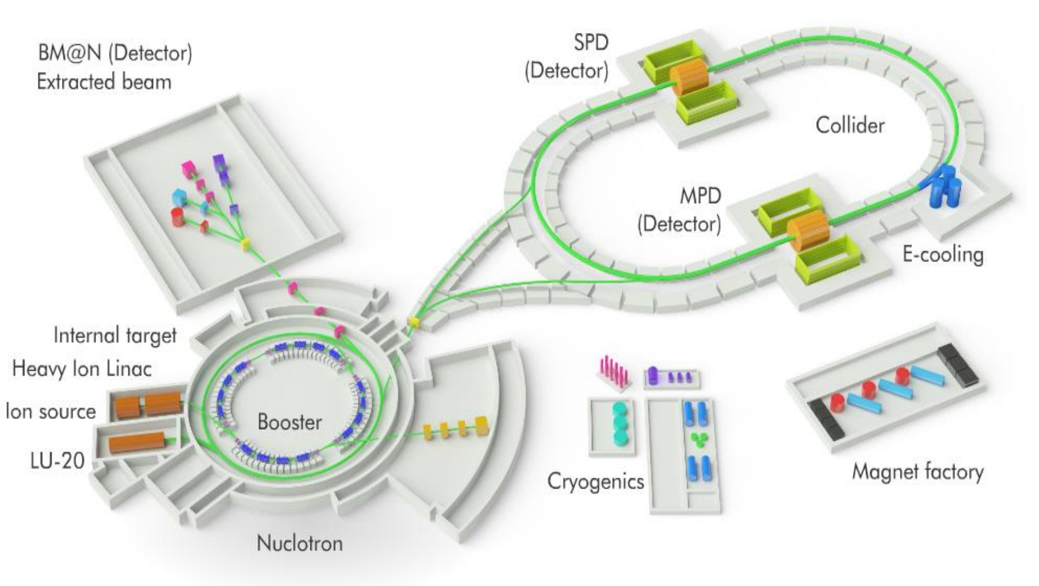
Goals and tasks of the MICC engineering infrastructure:

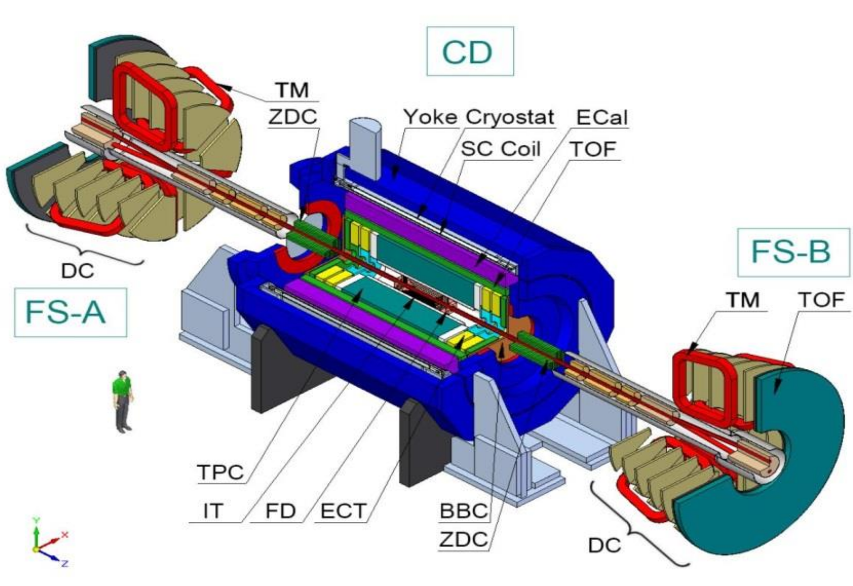
The engineering infrastructure should provide a reliable functioning of the Complex 24 hours a day, 7 days a week round-the-year



A scheme of arranging the equipment in the computer hall on the 2-nd and 4-th floors

NICA Accelerator Complex





For the NICA project the data stream has the following parameters:

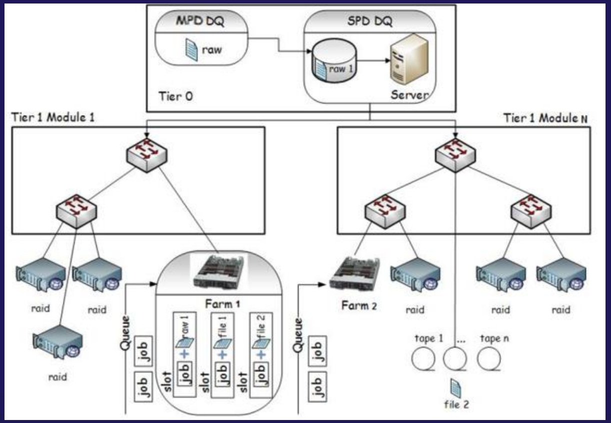
- high speed of the event set (up to 6 kHz),
- in central Au-Au collision at the NICA energies, about 1000 charged particles are generated,
- predicted event quantity - 19 billion;
- the total amount of initial data can be valued as 30 PB annually or 8.4 PB after compression.

(*MPD Conceptual Design Report v. 1.4*)

Simulation of the distributed computer infrastructure

GRID
SaaS
PaaS
IaaS

A model for studying processes has been created:
✓Tape robot,
✓Disk array,
✓CPU Cluster.



Importance of the Tier-1 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 Tier-1 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!