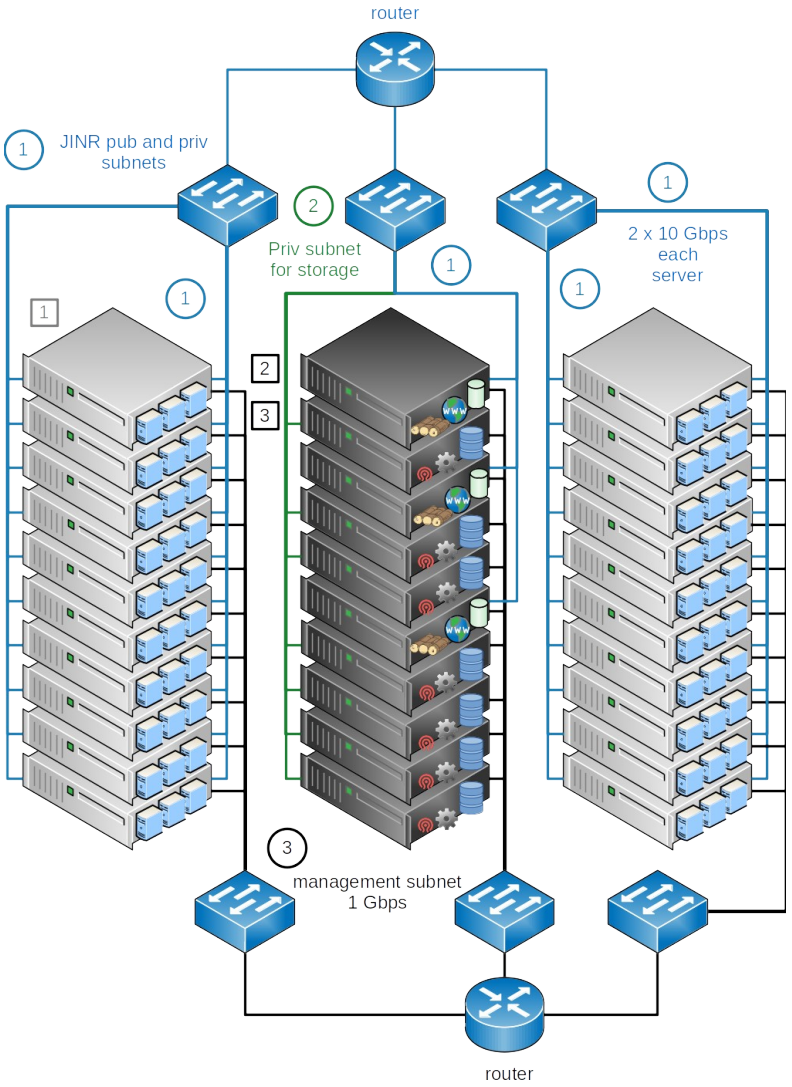
Quantitative and qualitative changes in the JINR cloud infrastructure

N. A. Balashov[1], I.S. Kuprikov[2],
N. A. Kutovskiy[1], A.N. Makhalkin[1],
Ye. Mazhitova[1,3], R. N. Semenov[1,4]

[1] Laboratory of Information Technologies, Joint Institute for Nuclear Research
[2] Dubna State University, Dubna, Russia
[3] Institute of Nuclear Physics, Almaty, Kazakhstan
[4] Plekhanov Russian University of Economics, Moscow, Russia

The 9th International Conference "Distributed Computing and Grid-technologies in Science and Education" (GRID'2021)
July 5-9, 2021, JINR, Dubna, Russia

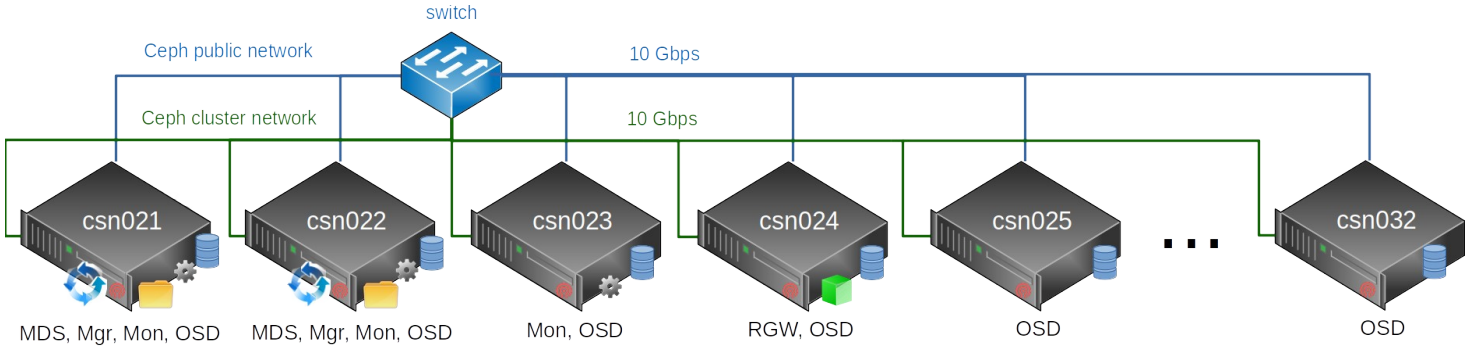
JINR cloud highlights



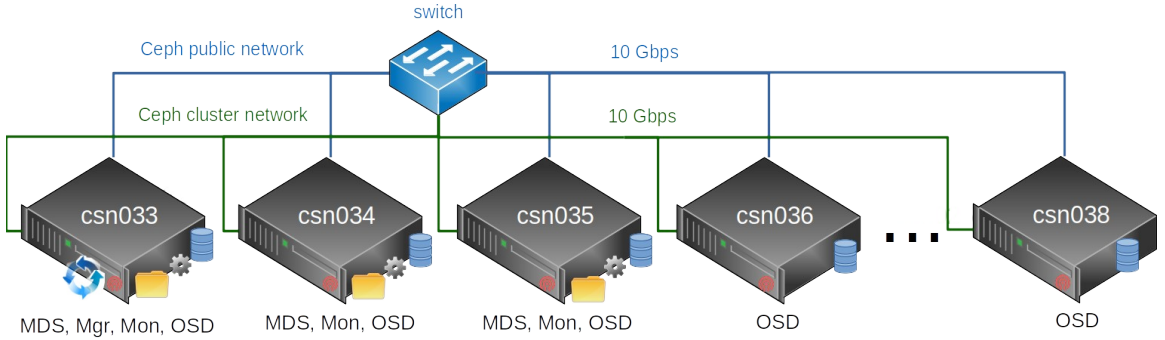
HA setup: 3 FNs, leader elections based on raft consensus algorithm
Distributed storage: ceph, 3x replicas

- Purpose
  - increase the efficiency of hardware and proprietary software utilization
  - improve IT-services management
- Implementation:
  - Cloud platform: OpenNebula (v5.12.0.4 CE)
  - Virtualization: KVM (dropped OpenVZ support)
  - Storage back-end for KVM VM images: ceph block-device
  - user interfaces: web GUI and command line interface
  - Authentication in the cloud web-GUI : JINR central user database (LDAP+Kerberos)
  - VM access: rsa/dsa-key or Kerberos credentials
- Hardware
  - 176 servers for VMs (+96 servers since Grid2018)
    - >5000 non-HT CPU cores (+3400)
      - 20 .. 32 non-HT CPU cores per physical server
    - >60 TB of RAM (+52 TB)
      - RAM per non-HT CPU core: 5.3 GB..16 GB
  - 21 servers for ceph storages with 3 PB of raw disk capacity (+2.1 PB)
- Web-interface URL: http://cloud.jinr.ru

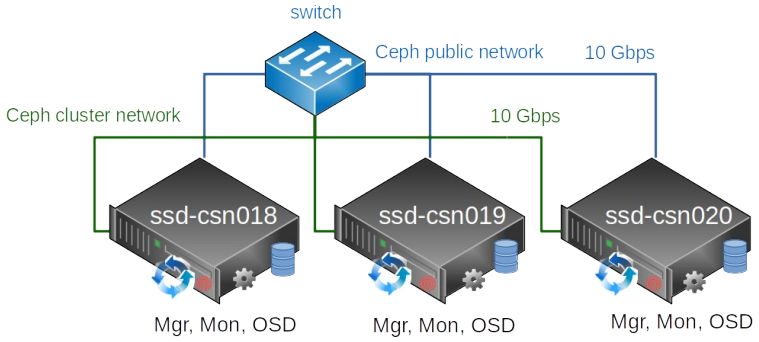
Ceph-based software defined storages



General purpose storage
ceph version: 14.2.21
Total raw capacity: 1.1 PiB
Replication: 3x
Connectivity: 2x10GBase-T



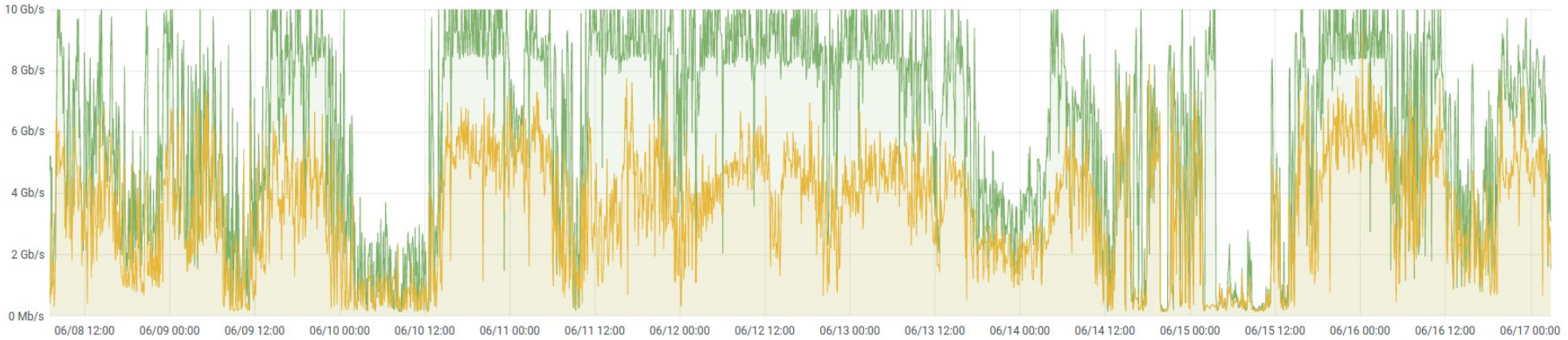
NOvA storage
ceph version: 15.2.11
Total raw capacity: 1.5 PiB
Replication: 3x
Connectivity: 2x10GBase-T



Pure SSD storage
ceph version: 15.2.13
Total raw capacity: 419 TiB
Replication: 3x
Connectivity: 4x10GBase-T (bonding)
+ 2x100Gbps – to be connected

Works on the storage for the NOvA experiment were supported by the grant of the Russian Science Foundation (project № 18-12-00271).

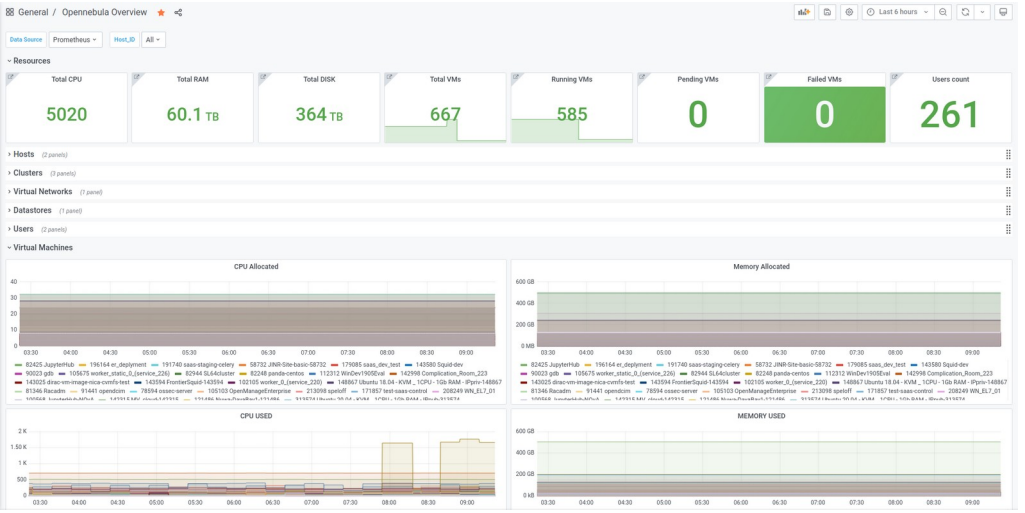
Network bottleneck



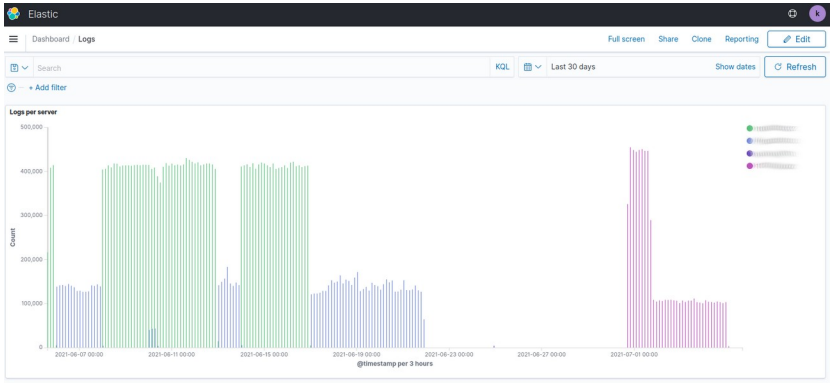
- Thousands of jobs create sufficient load on the network
- Faced with network bottleneck what led to services misbehavior
- Network update is scheduled for the next week

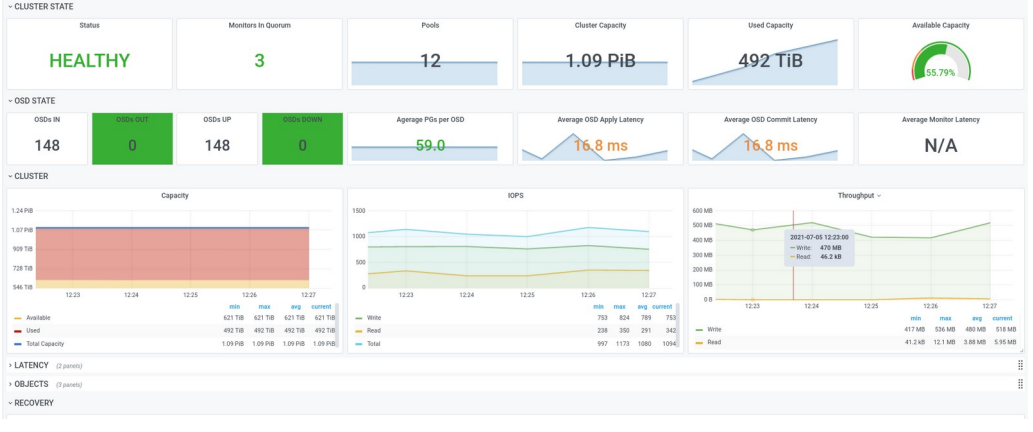
Monitoring and accounting

- Custom OpenNebula metrics collector
  - Prometheus (TSDB + alertmanager)
  - InfluxDB (for backward compatibility)
  - Grafana for visualization



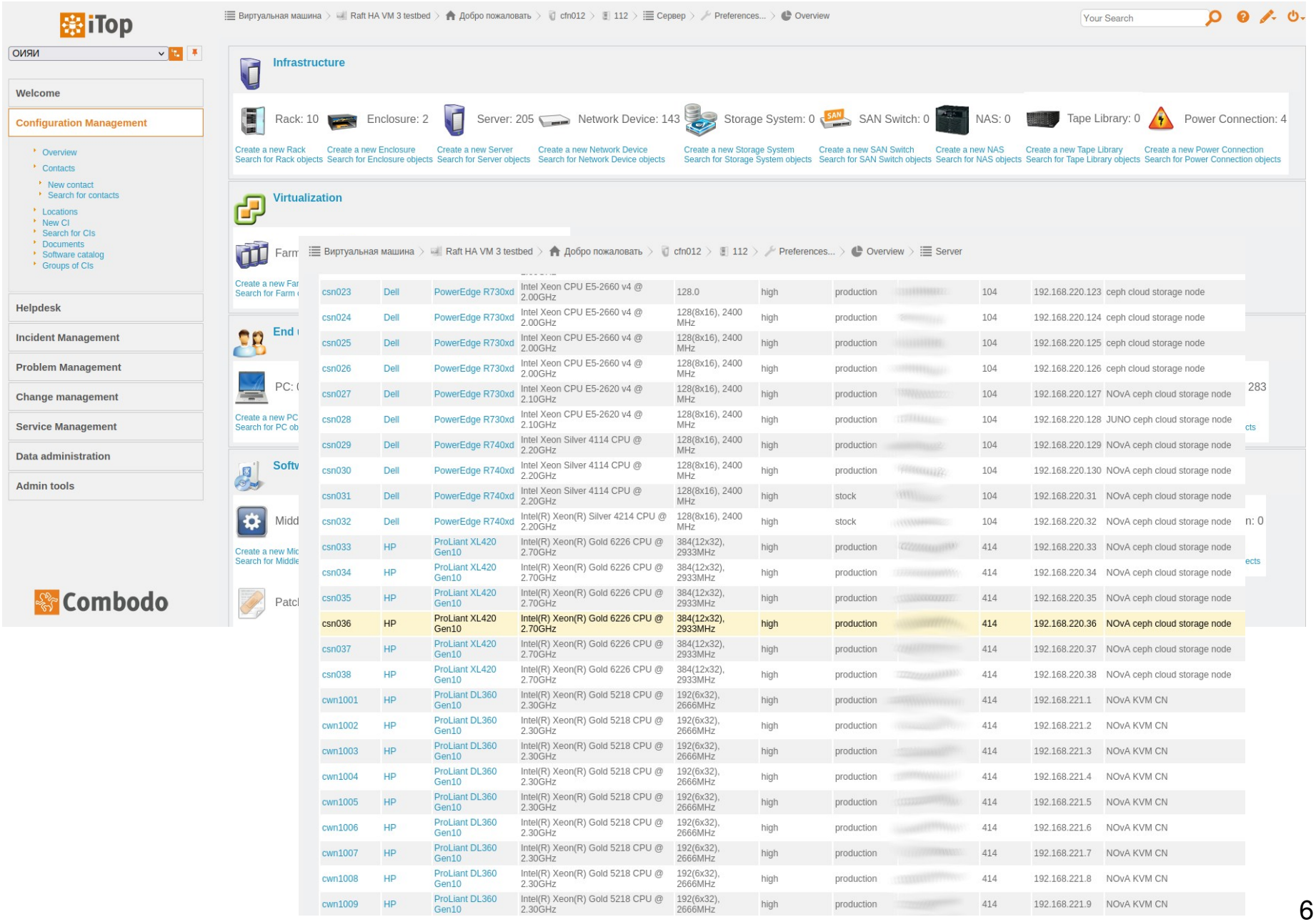
- OpenDistro for ElasticSearch
  - OpenNebula logs
  - Kibana for visualization





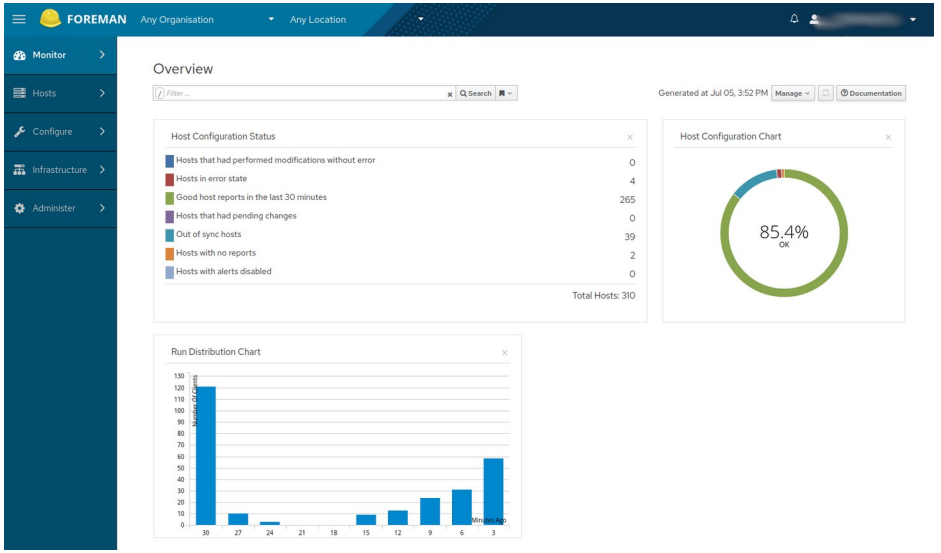
Ceph prometheus module
+ prometheus + grafana

Hardware inventory



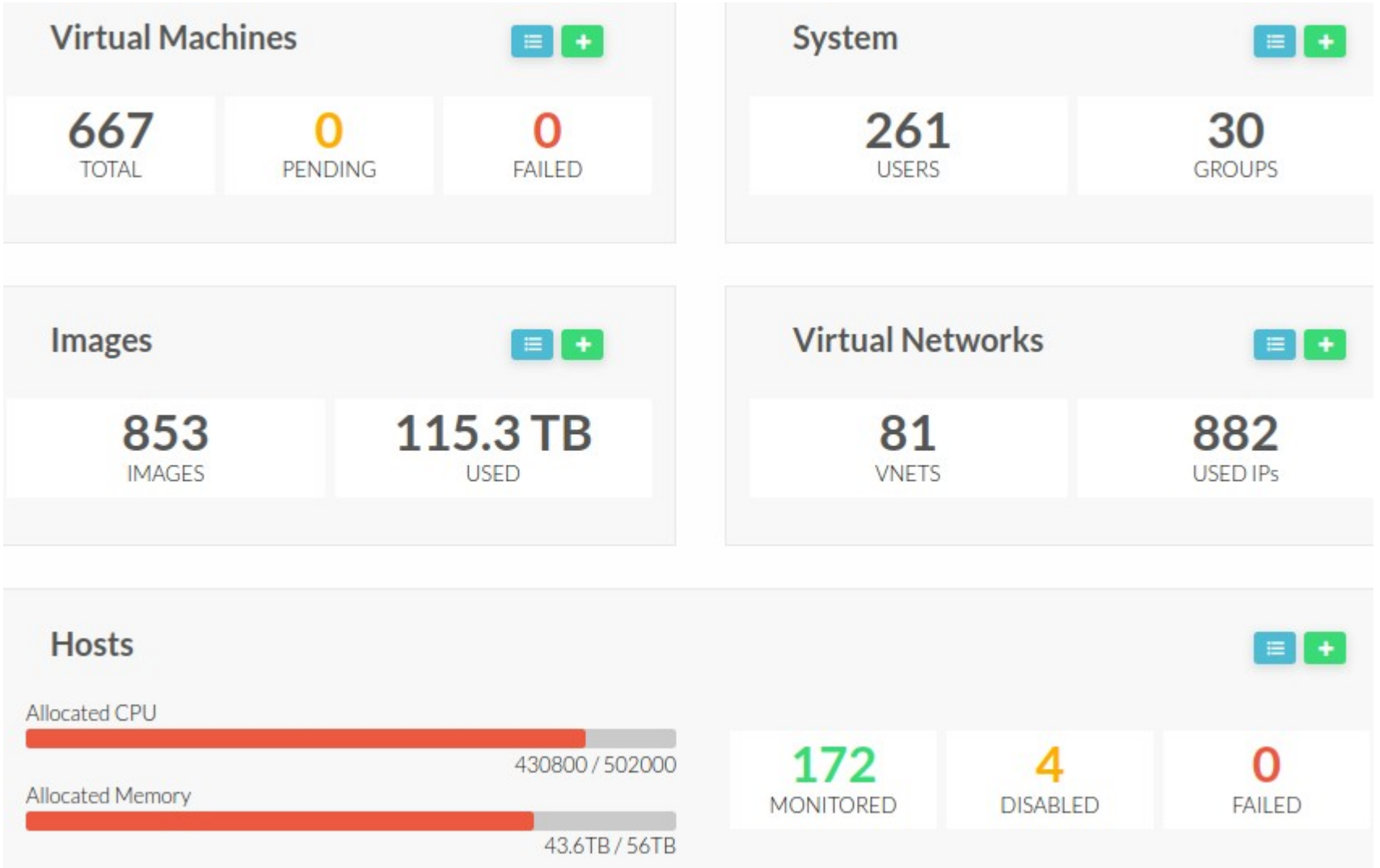
csn023	Dell	PowerEdge R730xd	Intel Xeon CPU E5-2660 v4 @ 2.00GHz	128.0	high	production		104	192.168.220.123	ceph cloud storage node
csn024	Dell	PowerEdge R730xd	Intel Xeon CPU E5-2660 v4 @ 2.00GHz	128(8x16), 2400 MHz	high	production		104	192.168.220.124	ceph cloud storage node
csn025	Dell	PowerEdge R730xd	Intel Xeon CPU E5-2660 v4 @ 2.00GHz	128(8x16), 2400 MHz	high	production		104	192.168.220.125	ceph cloud storage node
csn026	Dell	PowerEdge R730xd	Intel Xeon CPU E5-2660 v4 @ 2.00GHz	128(8x16), 2400 MHz	high	production		104	192.168.220.126	ceph cloud storage node
csn027	Dell	PowerEdge R730xd	Intel Xeon CPU E5-2620 v4 @ 2.10GHz	128(8x16), 2400 MHz	high	production		104	192.168.220.127	NOvA ceph cloud storage node
csn028	Dell	PowerEdge R730xd	Intel Xeon CPU E5-2620 v4 @ 2.10GHz	128(8x16), 2400 MHz	high	production		104	192.168.220.128	JUNO ceph cloud storage node
csn029	Dell	PowerEdge R740xd	Intel Xeon Silver 4114 CPU @ 2.20GHz	128(8x16), 2400 MHz	high	production		104	192.168.220.129	NOvA ceph cloud storage node
csn030	Dell	PowerEdge R740xd	Intel Xeon Silver 4114 CPU @ 2.20GHz	128(8x16), 2400 MHz	high	production		104	192.168.220.130	NOvA ceph cloud storage node
csn031	Dell	PowerEdge R740xd	Intel Xeon Silver 4114 CPU @ 2.20GHz	128(8x16), 2400 MHz	high	stock		104	192.168.220.31	NOvA ceph cloud storage node
csn032	Dell	PowerEdge R740xd	Intel(R) Xeon(R) Silver 4214 CPU @ 2.20GHz	128(8x16), 2400 MHz	high	stock		104	192.168.220.32	NOvA ceph cloud storage node
csn033	HP	ProLiant XL420 Gen10	Intel(R) Xeon(R) Gold 6226 CPU @ 2.70GHz	384(12x32), 2933MHz	high	production		414	192.168.220.33	NOvA ceph cloud storage node
csn034	HP	ProLiant XL420 Gen10	Intel(R) Xeon(R) Gold 6226 CPU @ 2.70GHz	384(12x32), 2933MHz	high	production		414	192.168.220.34	NOvA ceph cloud storage node
csn035	HP	ProLiant XL420 Gen10	Intel(R) Xeon(R) Gold 6226 CPU @ 2.70GHz	384(12x32), 2933MHz	high	production		414	192.168.220.35	NOvA ceph cloud storage node
csn036	HP	ProLiant XL420 Gen10	Intel(R) Xeon(R) Gold 6226 CPU @ 2.70GHz	384(12x32), 2933MHz	high	production		414	192.168.220.36	NOvA ceph cloud storage node
csn037	HP	ProLiant XL420 Gen10	Intel(R) Xeon(R) Gold 6226 CPU @ 2.70GHz	384(12x32), 2933MHz	high	production		414	192.168.220.37	NOvA ceph cloud storage node
csn038	HP	ProLiant XL420 Gen10	Intel(R) Xeon(R) Gold 6226 CPU @ 2.70GHz	384(12x32), 2933MHz	high	production		414	192.168.220.38	NOvA ceph cloud storage node
cwn1001	HP	ProLiant DL360 Gen10	Intel(R) Xeon(R) Gold 5218 CPU @ 2.30GHz	192(6x32), 2666MHz	high	production		414	192.168.221.1	NOvA KVM CN
cwn1002	HP	ProLiant DL360 Gen10	Intel(R) Xeon(R) Gold 5218 CPU @ 2.30GHz	192(6x32), 2666MHz	high	production		414	192.168.221.2	NOvA KVM CN
cwn1003	HP	ProLiant DL360 Gen10	Intel(R) Xeon(R) Gold 5218 CPU @ 2.30GHz	192(6x32), 2666MHz	high	production		414	192.168.221.3	NOvA KVM CN
cwn1004	HP	ProLiant DL360 Gen10	Intel(R) Xeon(R) Gold 5218 CPU @ 2.30GHz	192(6x32), 2666MHz	high	production		414	192.168.221.4	NOvA KVM CN
cwn1005	HP	ProLiant DL360 Gen10	Intel(R) Xeon(R) Gold 5218 CPU @ 2.30GHz	192(6x32), 2666MHz	high	production		414	192.168.221.5	NOvA KVM CN
cwn1006	HP	ProLiant DL360 Gen10	Intel(R) Xeon(R) Gold 5218 CPU @ 2.30GHz	192(6x32), 2666MHz	high	production		414	192.168.221.6	NOvA KVM CN
cwn1007	HP	ProLiant DL360 Gen10	Intel(R) Xeon(R) Gold 5218 CPU @ 2.30GHz	192(6x32), 2666MHz	high	production		414	192.168.221.7	NOvA KVM CN
cwn1008	HP	ProLiant DL360 Gen10	Intel(R) Xeon(R) Gold 5218 CPU @ 2.30GHz	192(6x32), 2666MHz	high	production		414	192.168.221.8	NOvA KVM CN
cwn1009	HP	ProLiant DL360 Gen10	Intel(R) Xeon(R) Gold 5218 CPU @ 2.30GHz	192(6x32), 2666MHz	high	production		414	192.168.221.9	NOvA KVM CN

Infrastructure management

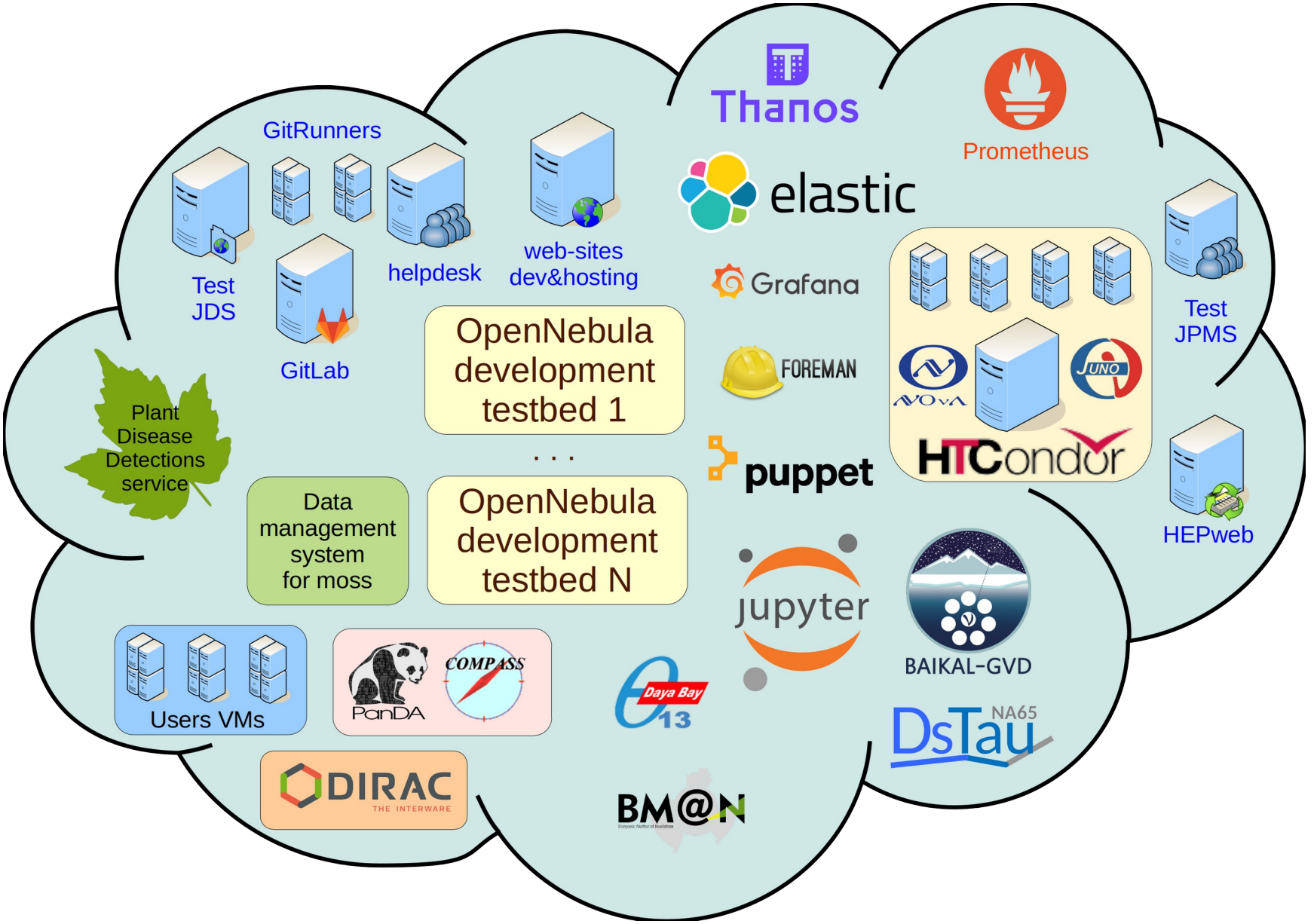


- Infrastructure as a Code (IaaC)
- Foreman + puppet
  - Profile + role model
- Physical servers and virtual machines
- Hosts autodiscovery feature
- Puppet manifests management is done via git
- Sensitive information is kept in HashiCorp Vault

Usage (1/2)


Virtual Machines
667
TOTAL
0
PENDING
0
FAILED
System
261
USERS
30
GROUPS
Images
853
IMAGES
115.3 TB
USED
Virtual Networks
81
VNETS
882
USED IPs
Hosts
Allocated CPU
430800 / 502000
Allocated Memory
43.6TB / 56TB
172
MONITORED
4
DISABLED
0
FAILED

Usage (2/2)

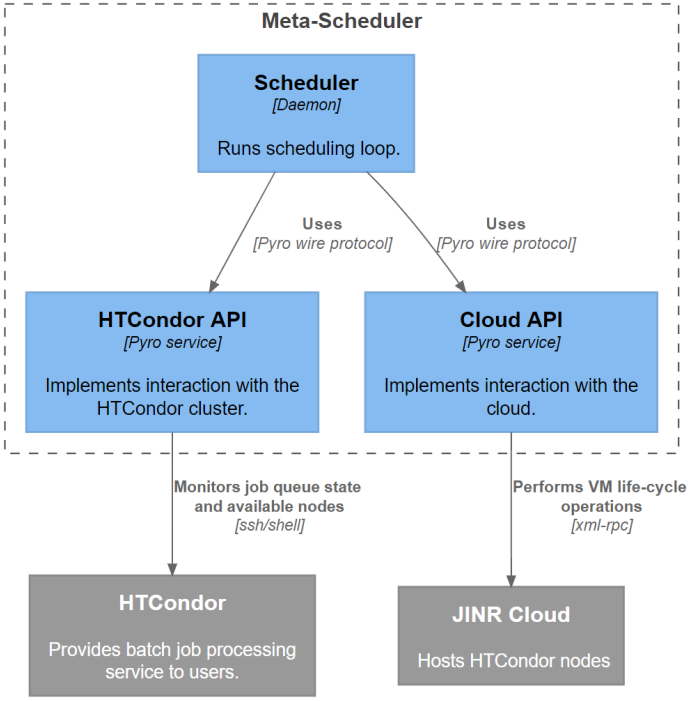

Thanos
Prometheus
elastic
GitRunners
Test
JDS
helpdesk
web-sites
dev&hosting
Grafana
Test
JPMS
GitLab
OpenNebula
development
testbed 1
FOREMAN
Plant
Disease
Detections
service
. . .
puppet
HTCondor
Data
management
system
for moss
OpenNebula
development
testbed N
HEPweb
jupyter
BAIKAL-GVD
Users VMs
PanDA
COMPASS
Daya Bay
DsTau
NA65
DIRAC
THE INTERWARE
BM@N

Cooperation with DLNP

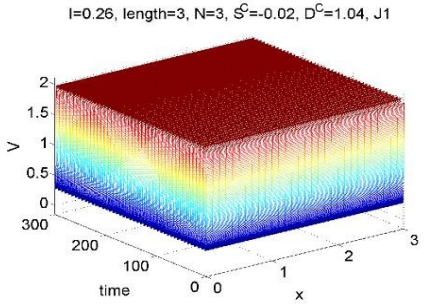
DLNP neutrino experiments contribution into JINR cloud components

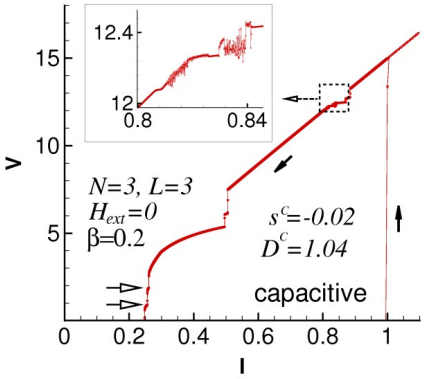
	Total number of CPU cores, items	Total amount of RAM, TB	Total amount of storage, TB
Baikal-GVD	84	0.768	0
JUNO	2976	35.97	128
NOvA/DUNE	1020	5.79	2144

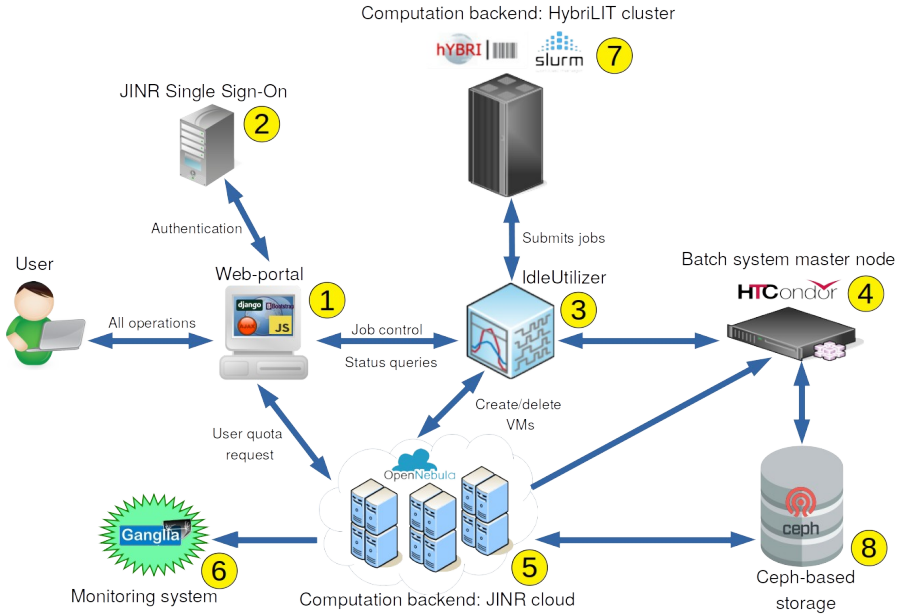
- One of the way to increase Neutrino Computing Platform (NCP) resources utilization efficiency is to organize resources sharing across NCP participants
- Cloud Meta-Scheduler is intended to implement such sharing by dynamic scaling of the HTCondor cluster on-demand
- Meta-Scheduler prototype is deployed and testing



Cooperation with BLTP







- **Purpose:**
  - to simplify the usage of the JINR MICC resources by providing intuitive web-interface for scientists to run computational jobs
- **Available apps:**
  - Long Josephson junctions stack simulation
  - Superconductor-Ferromagnetic-Superconductor Josephson junction simulation
  - Annular Array of JJs average
  - Long Josephson junction coupled with the ferromagnetic thin film
  - Stack of short JJ
  - Stack of short JJ with LC shunting

The work is supported by the Russian Science Foundation under grant #18-71-10095

Conclusion&Plans

- The JINR cloud resources are growing as well as a number of its users
- Most HW contribution is done by neutrino experiments
- Quantitative change requires changes in architecture: splitting ceph storage into serveral instances, ceph with SSD disks for VMs sensitive to disk I/O, network upgrade
- Keep OpenNebula up to date
- Finish migration from nagios/icinga-based monitoring to prometheus-based one
- Increase a degree of automation by adding more profiles and roles in foreman/puppet
- Put the cloud meta-scheduler in production mode