

# Quantitative and qualitative changes in the JINR cloud infrastructure

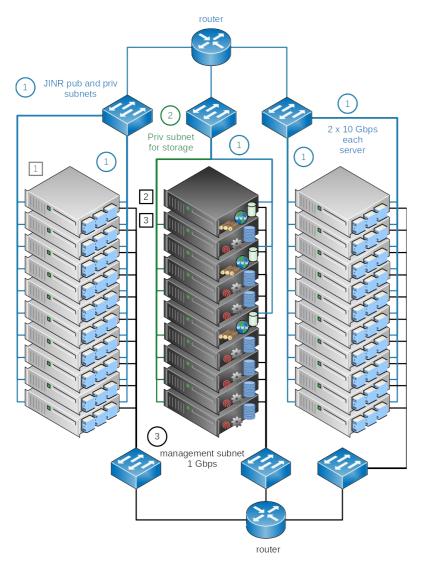
N. A. Balashov<sup>1</sup>, I.S. Kuprikov<sup>2</sup>, <u>N. A. Kutovskiy<sup>1</sup></u>, A.N. Makhalkin<sup>1</sup>, Ye. Mazhitova<sup>1,3</sup>, R. N. Semenov<sup>1,4</sup>



- <sup>1</sup> Laboratory of Information Technologies, Joint Institute for Nuclear Research
- <sup>2</sup> Dubna State University, Dubna, Russia
- <sup>3</sup> Institute of Nuclear Physics, Almaty, Kazakhstan
- <sup>4</sup> 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

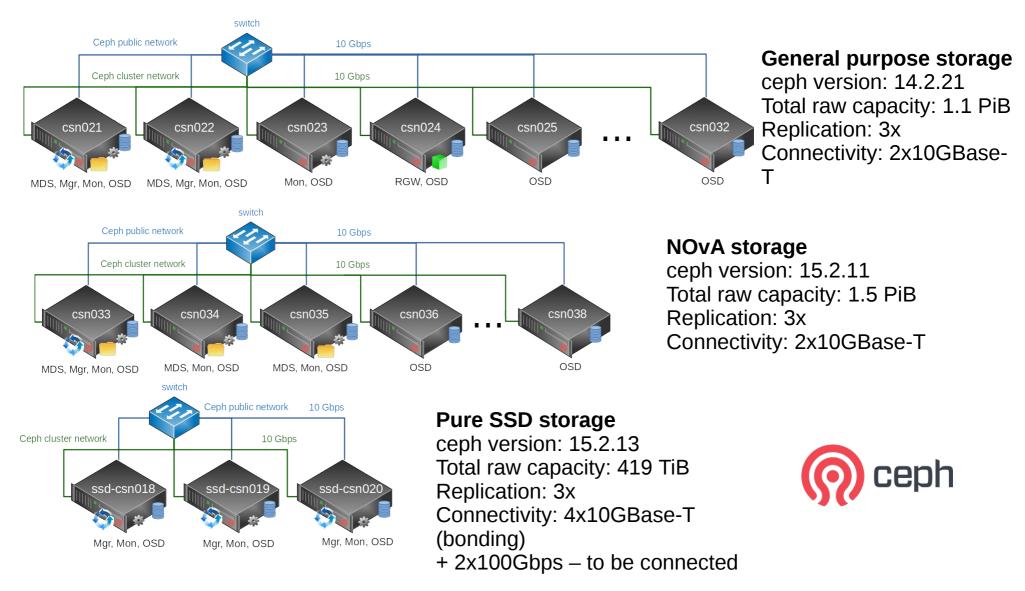
## JINR cloud highlights



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

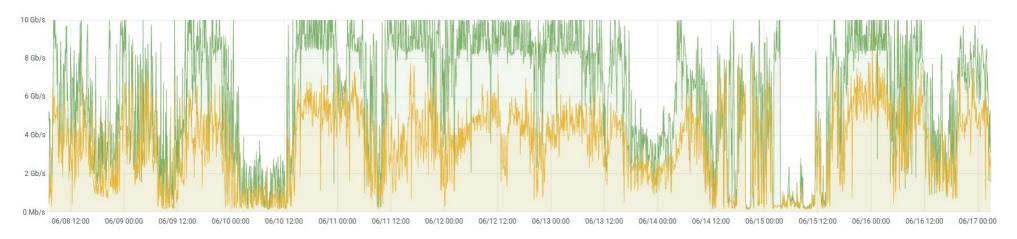
- Purpose
  - increase the efficiency of hardware and proprietary software utilization
  - improve IT-services management
- Implementation:
  - Cloud platform: OpenNebula (v5.12.0.4 CE)
  - Virtualization: KVM (dropped OpenVZ support)
  - Storage back-end for KVM VM images: ceph blockdevice
  - user 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



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

#### Network bottleneck

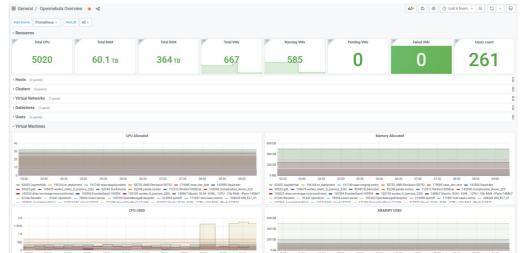


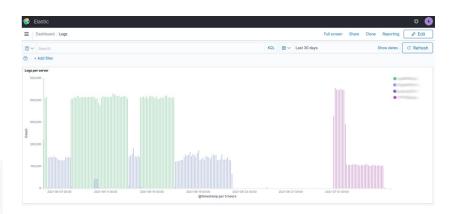
- Thousands of jobs create sufficient load on the network
- Faced with network bottleneck what led to services misbehavior
- Network update is scheduled for the next week

## 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

CLUSTER STATE																		
Status		Monitors In	Monitors In Quorum		Pools			Cluster Capacity			Used Capacity				Available Capacity			
HEALTHY		3				1.09 PiB			492 TiB			55.79%						
OSD STATE																		
OSDs IN	OSDs OUT	OSDs UP	OSDs DOWN		Agerage PGs per	OSD		Average OSD App	oly Latency		Aw	erage OSD Com	mit Latency		Average	Monitor Lat	ency	
148	0	148	0		59.0			16.81	ns	_	~	16.8 r	ns		1	A/A		
CLUSTER																		
	C	apacity					IOPS						Throu	ighput ~				
.24 PIB				1500							600 MB							
07 Pi8 09 Ti8				1000						-	400 MB		2021-07-05 12:23:00 - Write: 470 MB - Read: 46.2 kB	_		_		
28 7/8				500							200 MB		- HOSO. 40.2 KD					
546 T/B 12:23	12:24	12:25 12:26	12:27	0	12:23	12:24	12:25	12:26	1	2.27	100 MB							
- Available		min ma 621 TiB 621		- Write				min 753		g current 89 753		12:23	12:24	12:25	12 min	26 max	12.2	7 CUITE
Used		492 TIB 492		- Read							- Write				417 MB			518
<ul> <li>Total Capacity</li> </ul>		1.09 Pi8 1.09	PIB 1.09 PIB 1.09 PIB	- Total				997	1173 10	80 1094	- Read				41.2 kB	12.1 M8	3.88 MB	5.95
LATENCY (2 panels)																		
OBJECTS (3 parets)																		
RECOVERY																		





Ceph prometheus module + prometheus + grafana

#### Hardware inventory

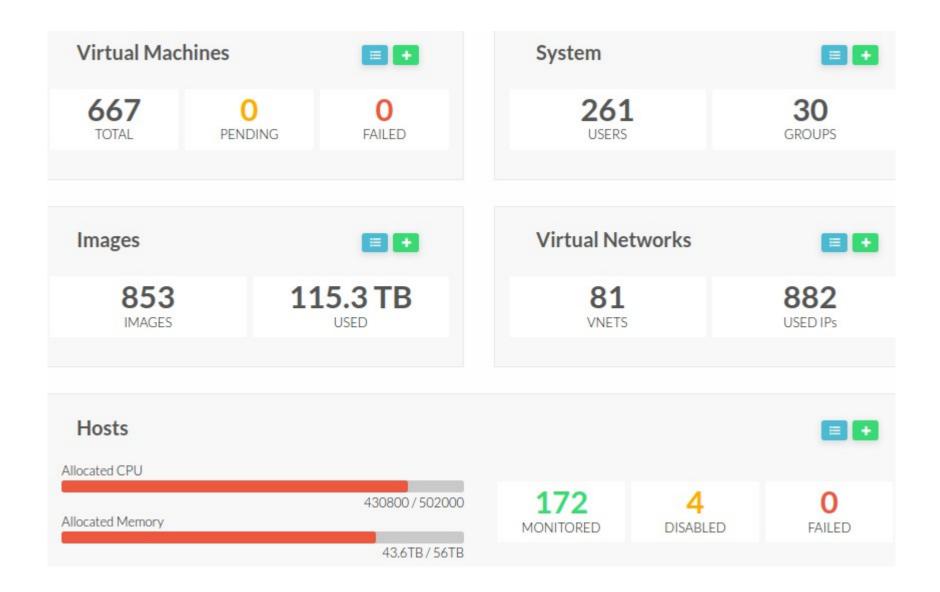
v 📜 🔻													
	Infrast	ructure											
ome				-		-							
guration Management	Rack:	10 💌	Enclosure: 2	2 Server:	205 Network Device: 1-	43 😻 Stora	ge System: 0	SAN SAN S	Switch: 0	NAS: 0	Tape L	ibrary: 0 🛕 Power Co	nnecti
Overview Contacts New contact	Create a new Rack Search for Rack obj	Create a n ects Search for	ew Enclosure Enclosure objec	Create a new Server ts Search for Server obj	Create a new Network Device ects Search for Network Device objects	Create a new Stor Search for Storage	age System e System objects	Create a new SAN Search for SAN Sw	Switch Create a ritch objects Search f	new NAS or NAS object	Create a new Tape L s Search for Tape Libr	ibrary Create a new Power Col ary objects Search for Power Conne	nection ction ob
Search for contacts Locations New Cl	Uirtual Virtual	ization											
Search for CIs Documents Software catalog Groups of CIs	Farm	🗮 Виртуалы	ная машина 🗦	Raft HA VM 3 test	tbed $>$ 🏠 Добро пожаловать $>$ 🧃	cfn012 > 🗐 112	> 🎤 Preferenc	es > 🕒 Oven	riew 🗦 🧮 Server				
	Create a new Far Search for Farm (	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	
pdesk		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	
dent Management	End I	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	
blem Management		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	
ange management	PC: (	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	28
vice Management	Create a new PC Search for PC ob	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	cts
s serverum us seri		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	
a administration	Softv	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	
nin tools		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	
	Midd	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	n: (
	Create a new Mic	csn033	HP	ProLiant XL420 Gen10	Intel(R) Xeon(R) Gold 6226 CPU @ 2.70GHz	384(12x32), 2933MHz	high	production	(022222222222)	414	192.168.220.33	NOvA ceph cloud storage node	
	Search for Middle	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	ects
Secondo 😚	Patcl	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	

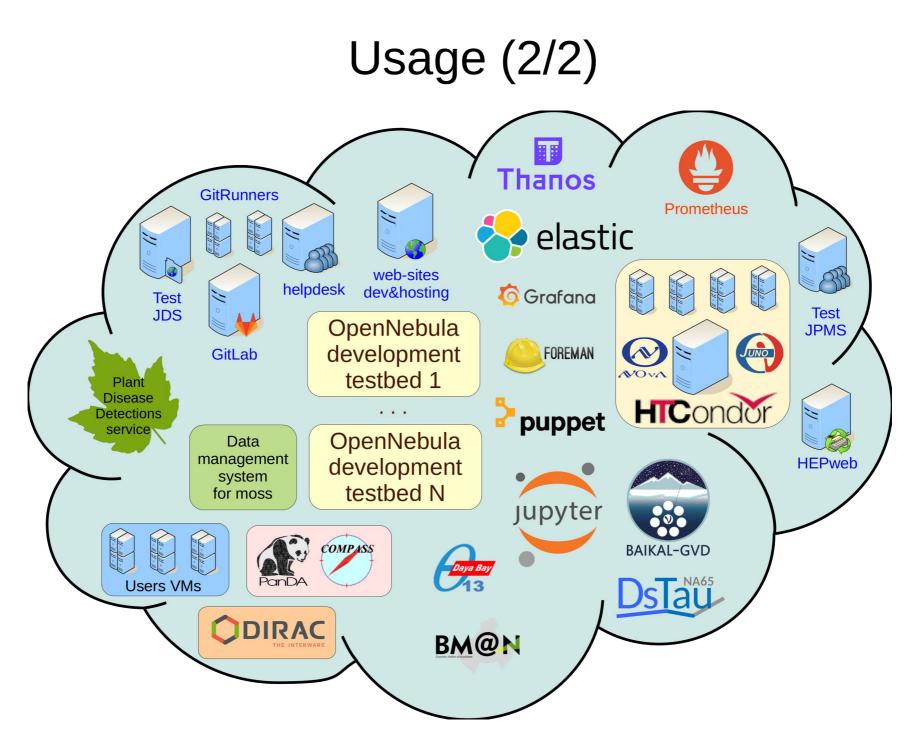
#### Infrastructure management

$\equiv$ $\bigcirc$ for	EMAN		4 <b>1</b> -
🚯 Monitor	>	Overview	
😫 Hosts	>	$\fbox{\sc product}{\sc product} \label{eq:linear} \fbox{\sc product}{\sc product} \label{eq:linear} \begin{tabular}{ c c c c c c c c c c c c c c c c c c c$	Generated at Jul 05, 3:52 PM Manage v Documentation
🗲 Configure	>	Host Configuration Status ×	Host Configuration Chart ×
👬 Infrastructure	>	Hosts that had performed modifications without error O Hosts in error state 4	
🔅 Administer	>	Good host reports in the last 30 minutes 265 Hosts that had pending changes 0	05.44
		Out of sync hosts         39           Hosts with no responts         2           Hosts with no responts         0	85.4% «
		Total Hosts: 310	
		Run Distribution Chart ×	

- Infrastructure as a Code (laaC)
- 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)



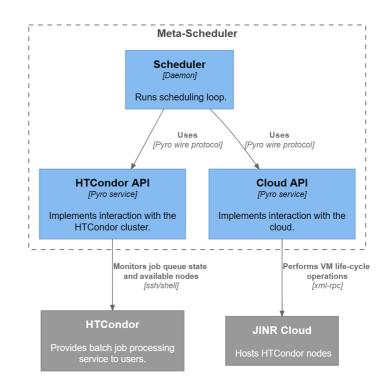


### Cooperation with DLNP

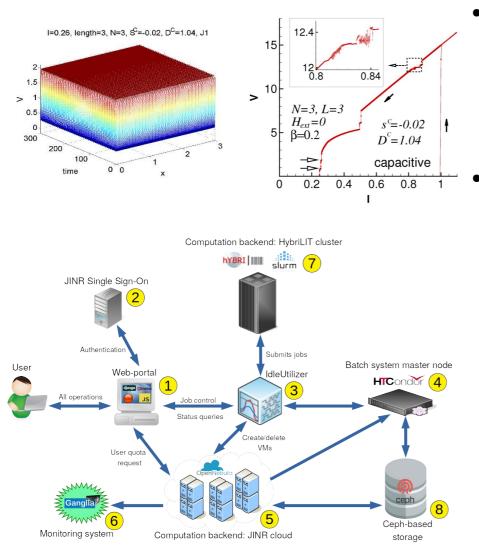
#### DLNP neutrino experiments contribution into JINR cloud components

	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 ondemand
- 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