

Changes and challenges at the JINR and its Member states cloud infrastructures

N. A. Balashov¹, I.S. Kuprikov², <u>N. A. Kutovskiy¹</u>, A.N. Makhalkin¹, Ye. Mazhitova^{1,3}, I.S. Pelevanyuk¹, R. N. Semenov¹



¹ Meshcheryakov Laboratory of Information Technologies, JINR
 ² X5 Digital, Moscow, Russia

³ Institute of Nuclear Physics, Almaty, Kazakhstan

The 10th International Conference "Distributed Computing and Grid technologies in Science and Education" (GRID'2023) July 3-7, 2023, JINR, Dubna, Russia

Plan

- JINR cloud
- Distributed informational and computing environment (DICE) based on the resources of JINR Member State organizations

JINR cloud as part of MICC



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
 - 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, Kerberos credentials, local AA
- Hardware
 - 174 servers for VMs:
 - +1 new server (since Grid2021)
 - -3 old servers (since Grid2021)
 - >5000 non-HT CPU cores (+128, -60)
 - 20 .. 32 non-HT CPU cores per physical server
 - >60 TB of RAM (+512 GB, -384 GB)

- RAM per non-HT CPU core: 5.3 GB..16 GB

- 24 servers for ceph storages with 3 PB of raw HDD disk capacity (+3 servers as reserve)
- Web-interface URL: http://cloud.jinr.ru

Ceph-based software defined storages (1/2)



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

Ceph-based software defined storages (2/2)

- SSD-based ceph storage:
 - For users and services VMs with high disk I/O requirements
 - Prometheus DB
 - COMPASS critical services
 - git@JINR
 - etc
- HDD-based ceph storage:
 - For users and services VMs with mild/low disk I/O requirements
 - Data preservation for relatively small data volumes (large data volumes are for EOS@JINR)



HDD-based ceph storage performance issue

- Thousands of jobs create sufficient load on the HDD ceph storage (VMs disks are Rados ceph block devices)
- A lack of HDD-based ceph storage performance led to decrease QoS of other services deployed in the JINR cloud
- VMs' disks migration from HDD-based ceph storage to SSD-based one was started
- Most of the VMs with high demanding disk I/O have been already migrated to SSD-based ceph storage. New users' VMs disks are created now on SSD-based ceph storage



SSD-based ceph storage performance last 90 days

Monitoring and accounting

Custom OpenNebula metrics collector

- Prometheus (TSDB + alertmanager)
- InfluxDB retired
- Grafana for visualization
- Migration from nagios to prometheus alarms is in transition phase

Resources							
Total CPU	Total RAM	Total DISK	C Total VMs	C Running VMs	Pending VMs	Failed VMs	Users count
5152	60.6 тв	470 тв	v_ 778	692	1	0	197
Hosts (2 panels)							
Clusters (3 panels)							
Virtual Networks (1 panel)							
Datastores (1 panel)							
Users (2 panels)							
Virtual Machines (6 panels)							
Resources detailed							
		CPU				RAM	
5 K				600 GB			
0 K				500 GB 400 GB			
5 K				300 GB			
				200 GB			
10:30 11:00 11:30	12:00 12:30 13:00	13:30 14:00 14:30	15:00 15:30 16:00	10:30 11:00 11:30	12:00 12:30 13:0	0 13:30 14:00 14:30	15:00 15:30 16:00
0 10:30 11:00 11:30 = 277 - 278 - 280 - 281	12:00 12:30 13:00 282 283 284 285	13:30 14:00 14:30 1 286 - 287 - 288 - 289 -	15:00 15:30 16:00 290 - 291 - 292 - 428	100 GB 10:30 11:00 11:30 - 277 - 278 - 280 - 28	12:00 12:30 13:0 1 - 282 - 283 - 284 - 3	0 13:30 14:00 14:30 185 - 286 - 287 - 288 - 289	15:00 15:3 - 290 - 291
431 - 432 - 433 447 - 448 - 449	- 434 - 435 - 436 - 437 - 450 - 455 - 371 - 372	- 438 - 439 - 440 - 441 - 373 - 374 - 375 - 451 -	442 4 43 4 44 4 45 452 4 53 4 54 2 94	- 429 - 431 - 432 - 43 - 446 - 447 - 448 - 44	3 - 434 - 435 - 436	137 = 438 = 439 = 440 = 441 172 = 373 = 374 = 375 = 451	- 442 - 443 - 444 - 445 - 452 - 453 - 454 - 294
- 297 - 305 - 321 - 322		- 386 - 387 - 392 - 393 -	394 - 395 - 396 - 400		2 - 323 - 341 - 384 - 3	85 - 386 - 387 - 392 - 393	- 394 - 395 - 396 - 400

OpenDistro for ElasticSearch

- OpenNebula logs
- Kibana for visualization







Ceph prometheus module + prometheus + grafana

Hardware inventory

🚼 iTop	🧮 Виртуальная маши	ина > 🗐 Raft	HA VM 3 testbe	d > 🏫 Добро пожалов	ать $>$ 🧃 cfn012 $>$ 🗐 112 $>$ 🧮 Се	рвер 〉 🎤 Preferenc	es > Cvervie	w			Your	Search	01.
т	Infrastr	ructure											
Welcome	U					-							
Configuration Management	Rack: 1	10 💌	Enclosure: 2	2 Server: 2	205 Network Device: 1	43 😻 Stora	.ge System: 0 🧃	SAN SAN S	Switch: 0	NAS: 0	Tape Lik	orary: 0 🛕 Power C	onnection: 4
Overview Contacts New contact	Create a new Rack Search for Rack obje	Create a ne	ew Enclosure Enclosure objec	Create a new Server ts Search for Server obje	Create a new Network Device ects Search for Network Device objects	Create a new Stor Search for Storage	rage System C e System objects S	reate a new SAN earch for SAN Sv	Switch Create a vitch objects Search fr	new NAS or NAS objects	Create a new Tape Lii S Search for Tape Libra	orary Create a new Power C ry objects Search for Power Con	onnection nection objects
 Search for contacts Locations New CI 	Uirtuali	ization											
Search for CIs Documents Software catalog Groups of CIs	Farm	🗏 Виртуальн	ная машина $>$	Raft HA VM 3 test	bed > 🏫 Добро пожаловать > 🧃	cfn012 > 🗐 112	> 🌽 Preference	s > 🔮 Over	view $> egin{array}{c} & & \\ $				
	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	
elpdesk		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	
ncident 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	
roblem 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	
hange 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 nod	e 283
arvice Management	Create a new PC	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 nod	e
	Search for PC 00	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 nod	e
ata 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 nod	e
dmin 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 nod	e
	Midd	csn032	Dell	PowerEdge R740xd	Intel(R) Xeon(R) Silver 4214 CPU @ 2 20GHz	128(8x16), 2400	high	stock		104	192.168.220.32	NOvA ceph cloud storage nod	e n: 0
	Create a new Mir	csn033	HP	ProLiant XL420	Intel(R) Xeon(R) Gold 6226 CPU @	384(12x32),	high	production	(222200000180)	414	192.168.220.33	NOvA ceph cloud storage nod	e
	Search for Middle	csn034	HP	ProLiant XL420	Intel(R) Xeon(R) Gold 6226 CPU @	384(12x32),	high	production		414	192.168.220.34	NOvA ceph cloud storage nod	ects
Combodo	Date!	csn035	HP	ProLiant XL420	Intel(R) Xeon(R) Gold 6226 CPU @	384(12x32),	high	production		414	192.168.220.35	NOvA ceph cloud storage nod	e
	Faici	csn036	HP	ProLiant XL420	Intel(R) Xeon(R) Gold 6226 CPU @	384(12x32),	high	production		414	192.168.220.36	NOvA ceph cloud storage nod	e
		csn037	HP	ProLiant XL420	Intel(R) Xeon(R) Gold 6226 CPU @	2933MHZ 384(12x32),	high	production		414	192.168.220.37	NOvA ceph cloud storage nod	e
		csn038	HP	Gen10 ProLiant XL420	2.70GHz Intel(R) Xeon(R) Gold 6226 CPU @	2933MHz 384(12x32),	high	production		414	192 168 220 38	NOvA ceph cloud storage nod	e
		cwn1001	HP	ProLiant DL360	2.70GHZ Intel(R) Xeon(R) Gold 5218 CPU @	2933MHz 192(6x32),	hiah	production		414	192.168.221.1	NOVA KVM CN	
		cwn1002	HP	Gen10 ProLiant DL360	2.30GHz Intel(R) Xeon(R) Gold 5218 CPU @	2666MHz 192(6x32),	high	production		414	192,168,221,2	NOVA KVM CN	
		cwn1003	нр	Gen10 ProLiant DL360	2.30GHz Intel(R) Xeon(R) Gold 5218 CPU @	2666MHz 192(6x32),	high	production		414	192 168 221 3	NOVA KVM CN	
		cwn1004	HP	Gen10 ProLiant DL360	2.30GHz Intel(R) Xeon(R) Gold 5218 CPU @	2666MHz 192(6x32),	high	production		414	192 168 221 4	NOVA KVM CN	
		cwn1005	нр	Gen10 ProLiant DL360	2.30GHz Intel(R) Xeon(R) Gold 5218 CPU @	2666MHz 192(6x32),	high	production		414	192 168 221 5	NOVA KVM CN	
		CWD1006	нр	Gen10 ProLiant DL360	2.30GHz Intel(R) Xeon(R) Gold 5218 CPU @	2666MHz 192(6x32),	high	production		414	102 160 221 6	NOVA KVM CN	
		CWI1000	UD	Gen10 ProLiant DL360	2.30GHz Intel(R) Xeon(R) Gold 5218 CPU @	2666MHz 192(6x32),	high	production		414	102 160 221 7		
		cwn1007	HP	Gen10 ProLiant DL360	2.30GHz Intel(R) Xeon(R) Gold 5218 CPU @	2666MHz 192(6x32),	high	production		414	192.108.221.7		
		CWITOOR	THP	Gen10 ProLiant DL360	2.30GHz Intel(R) Xeon(R) Gold 5218 CPU @	2666MHz 192(6x32).	nign	production		414	195.108.221.8		
		cwn1009	HP				nian	production		414	192.168.221.9	NUVA KVM CN	

Infrastructure management

\equiv \bigcirc fore	MAN	Any Organisation Any Location	۰ ۹
Monitor	>	Overview	
Hosts	>	[] Filter	Generated at Jul 05, 3:52 PM Manage ~ 💭 🕲 Documentation
🗲 Configure	>	Host Configuration Status	Host Configuration Chart ×
👬 Infrastructure	>	Hosts that had performed modifications without error Hosts in error state	0
🏟 Administer	>	Good host reports in the last 30 minutes 26 Hosts that had pending changes	5
		Out of sync hosts 2 Hosts with no reports Hosts with alerts disabled	9 85,4% ск ок
		Ran Distribution Chart ×	0
		20 30 30 30 30 30 30 30 30 30 30 30 30 30	

- 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)

Virtual Machi	nes		System		= +		
778 TOTAL	2 O PENDING FAILED		197 USERS	7	34 GROUPS		
Images		= +	Virtual Ne	tworks	= +		
600 IMAGES	177.3 TB USED		99 VNETS		990 USED IPs		
Hosts					= +		
Allocated Memory		493700 / 515200 49.5TB / 56.5TB	170 MONITORED	6 DISABLED	O FAILED		



Number of active cloud users



Inefficient resources utilization

- Some users don't care about an efficiency of resources utilization
 - Users ask more resources they use in fact → query to cut amount of allocated resource to used ones
 - CPU load on some VMs is around zero pretty long time → query to either delete VM or to undeploy it
 - Regular clean up of unused VM images

User support

User support is done on basis of Helpdesk@JINR service (running on iTop software)

🔠 i Ton

Search for User Request Objects											
Ref:		Organization:	* Any * 🔻	۲itle Title	s	Description:]			
Start date:		Resolution date:		Close date	e	Status:	* Any * 🔻				
Operational status: *	Any * 🗸	Caller:		🔍 Origin	:: * Any * 🔻	Request Type:	* Any * 🔻				
Impact: *	Any *	Urgency:	* Any * 🔻	Priority	* Any * 🔻	Service:	* Any *				
Service subcategory: *	Any *	Team:	* Any * 👻	Agent		🔍 Hot Flag:	* Any * 🔻				
Resolution code: *	Any *	User satisfaction:	* Any * 🔻	SLA tto passed		SLA ttr passed:		J			
								Search			
				Search							

Number of user requests on cloud resoures during last 365 days 15 Cloud Storage

Number of incidents on cloud resoures during last 365 days



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	1000	5.72	2144
TAIGA	0	0	120

One of the way to increase Neutrino Computing Platform (NCP) resources utilization efficiency is to organize resources sharing across NCP participants

HTCondor-CEs:

- v9.0.11
- IAM-based token authentication support
- ~3400 CPU cores in total



Conclusion&Plans

- Most efforts on JINR cloud are focused on increasing QoS now
 - VMs disks migration from HDD-based ceph storage to SSD-based one
- Migration from nagios/icinga-based monitoring to prometheusbased is still in progress
- Increase a degree of automation by adding more profiles and roles in foreman/puppet

Training and testing JINR cloud (t-cloud)

- For user and admin trainings as well as for development and testing
- Servers for VMs:
 - 32 x Supermicro X8DTT-F: Intel(R) Xeon(R) CPU X5650@2.67GHz (12 CPU cores, 24 GB of RAM)
- Ceph (v17.2.5)
 - 3x Supermicro X9DR3-F
 - ~400 TB of raw disk space
 - Triple replication
 - 2x 10GBase-T
- OpenNebula
 - 6.0.0.2
 - Web-interface: http://t-cloud.jinr.ru

Virtual Machines			System		= +	
25 TOTAL	O PENDING	O FAILED	29 USERS		3 GROUPS	
Images		= +	Virtual Ne	tworks	= +	
5 IMAGES		20 GB	2 VNETS		26 USED IPS	
Hosts					= +	
Allocated CPU						
Allocated Memory		9700/38200	32 MONITORED	O DISABLED	O FAILED	
		212GB / 745.5GB				

JINR DICE: participants

 To join resources for solving common tasks as well as to distribute a peak load across resources of partner organizations



JINR DICE: resources

Organization	Country	Status	non-HT CPU cores	RAM, GB	Storage, TB
Plekhanov Russian Economic University	RU	integrated	132	608	51.1
Institute of Nuclear Physics	KZ	integrated	84	840	6.8 (SSD)
Institute of Physics of the National Academy of Sciences of Azerbaijan	AZ	maintenance	16	96	56
North Ossetian State University	RU	integrated	84	672	17
Academy of Scientific Research & Technology - Egyptian National STI Network	EG	maintenance	98	704	13.8
Institute for Nuclear Research and Nuclear Energy	BG	integrated	20	64	4
St. Sophia University «St. Kliment Ohridski»	BG	integrated	48	250	4.7
Scientific Research Institute of Nuclear Problems of the Belarusian State University	BY	integrated	132	290	127
Institute of Nuclear Physics	UZ	integrated	98	890	6.6 (SSD)
Georgian Technical University	GE	in progress	50	308	20
Total			762	4722	

Hardware inventory



PerfSONAR

- To monitor network connectivity of participants
 - http://cloud-perfsonar.jinr.ru



There is a challenge to deploy PS instance at some sites because all cloud VMs are behind NAT

Low external network **bandwidth** (e.g. 100 Mbps shared with the whole organization) is the main contributor into high CPU wall time of jobs Most **suitable** type of jobs for such kind of resources is **MC with negligible input data** 22/30

Experiments software distribution model



Metrics aggregation



Grafana World Map plugin



Usage: SARS-CoV-2 research via F@H



Usage: BM@N, SPD and Baikal-GVD



BM@N workflow with simulation jobs was tested successfully.



Simulation and reconstruction SPD jobs was tested successfully as well but not yet on production storage (EOS)



Baikal-GVD tried to use JINR DICE resources (apart from the JINR cloud) but results were poor due to insufficient network bandwidth at some clouds

Issues

- RU-NOSU:
 - often power cuts
 - Additional public IP address for perfSONAR instance
- KZ-INP:
 - slow response on hardware interventions requests
 - Too strict firewall
 - Additional public IP address for perfSONAR instance
- UZ-INP:
 - Too strict firewall
 - Additional public IP address for perfSONAR instance
- EG-ASRT
 - Not accessible from the JINR subnet
 - No reply from technical specialists
- AZ-IP:
 - Under maintenance pretty long time
- GE-GTU:
 - Due to COVID-related restrictions a hardware has not been delivered yet to GTU

JINR DICE web-portal

DICE

DISTRIBUTED INFORMATION AND COMPUTING ENVIRONMENT

DICE

INFRASTRUCTURE REGISTRATION CONTACTS

Infrastructure

A distributed infrastructure is a powerful com telecommunication channels and special softw different organizations into a single informatic tasks.

The integration of computing power of the Joir organizations into a unified distributed inform topical task, the solution of which would signif scientific results.

The relevance of creating such an environmer and resources in solving fundamental and app impossible without the use of new approache of systems for distributed storage of large am



http://dice.jinr.ru dice@jinr.ru

Communication with the experiments

INFORMATION AND COMPUTING

DISTRIBUTED

ENVIRONMENT

The JINR DICE cloud segment resources are configured to support the following virtual organizations (VOs) representing a scientific experiment and/or collaboration: BM@N, MPD, and Baikal-GVD. More than 15,000 Monte-Carlo simulation jobs for the BM@N VO were successfully completed as well as about 9000 jobs for Baikal-GVD VO. Testing jobs for MPD were performed on the distributed cloud infrastructure.







CONTACTS

Participants of the DICE Technical details Communication with the experiments Research on the SARS-CoV-2 Tutorials Publications JINR Cloud Infrastructure



INFRASTRUCTURE REGISTRATION

Q RU

Contact person

Yelena Mazhitova Junior Researcher Phone.: +7 (49621) 6-59-36 Email: dice@jinr.ru

Conclusion

- Number of JINR DICE participants and an amount of its resources is dynamically changed
- Only MC jobs with negligible input data are suitable for resources with low external network bandwidth
- Technical implementation of OpenNebula metrics aggregation is done. Its dissemination over JINR DICE clouds is in progress
- Migration from hand-drawn JINR DICE map to grafana World Map plugin is in progress
- JINR DICE web-portal is running:
 - Technical information about JINR DICE infrastructure
 - tutorials, publications
 - contacts