Intellectual Data Storage-on-Demand Systems

Pavel Lavrenko
Chief Business Development Officer
RSC Group

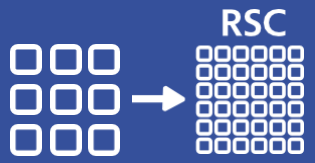
Flash Memory Summit 2020 Virtual Event

RSC BasIS Platform: Orchestration for High Performance Composable Storage Architectures

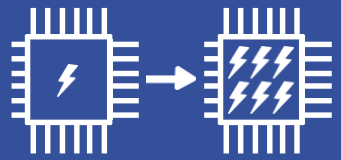
HPC innovations since 2009

Development of innovative
ultra-high dense and energy efficient HPC solutions
delivering unique features and addressing specific
end-user needs

Points of Excellence



Computing density



Power density

Energy efficiency

Ease to manage and maintain

Reliability

Market presence

Science and Education

- Leading Universities
- Russian Academy of Sciences
- CERN – JINR Cooperation
- Bioinformatics
- Astrophysics
- Medicine

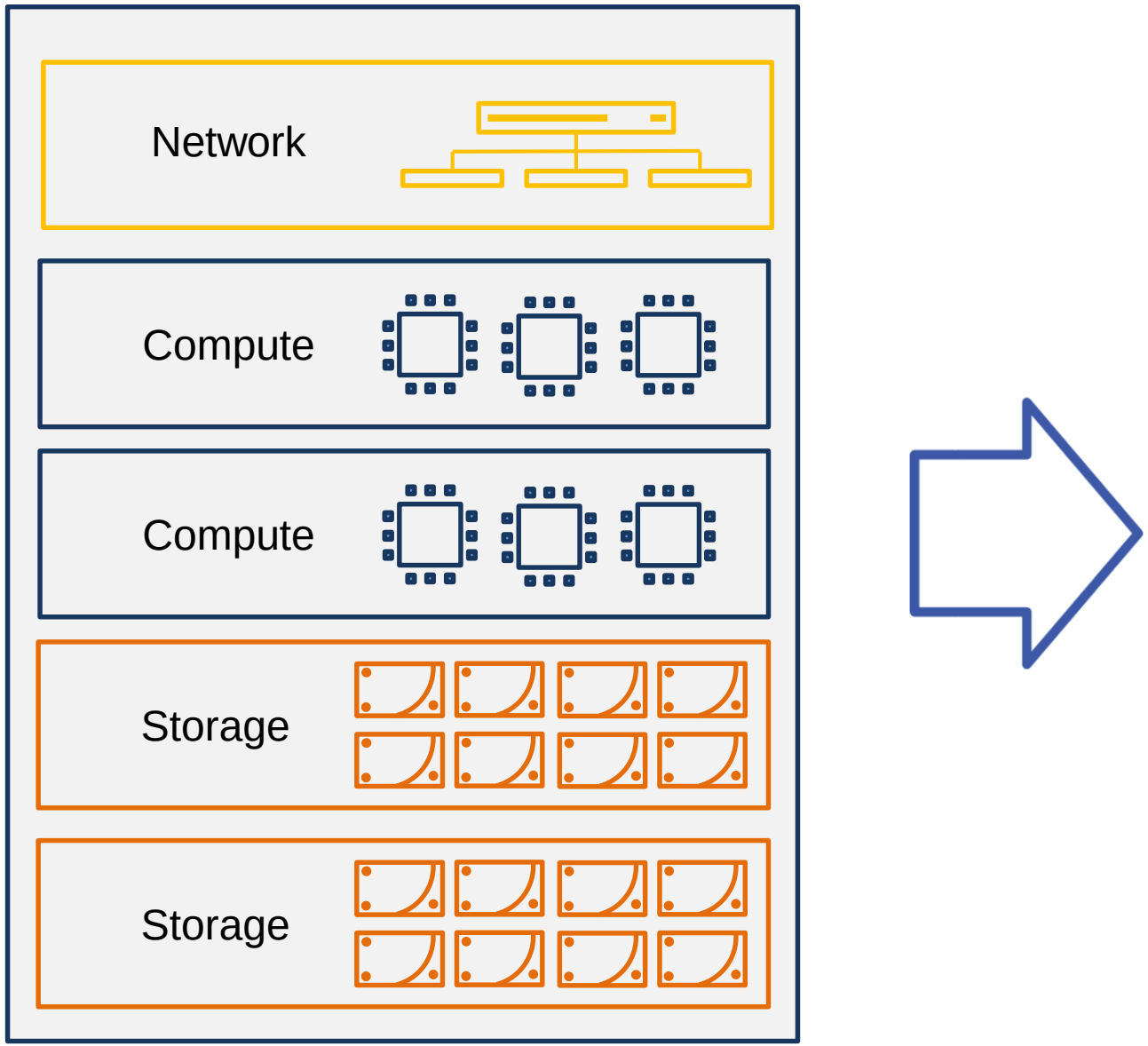
Industry

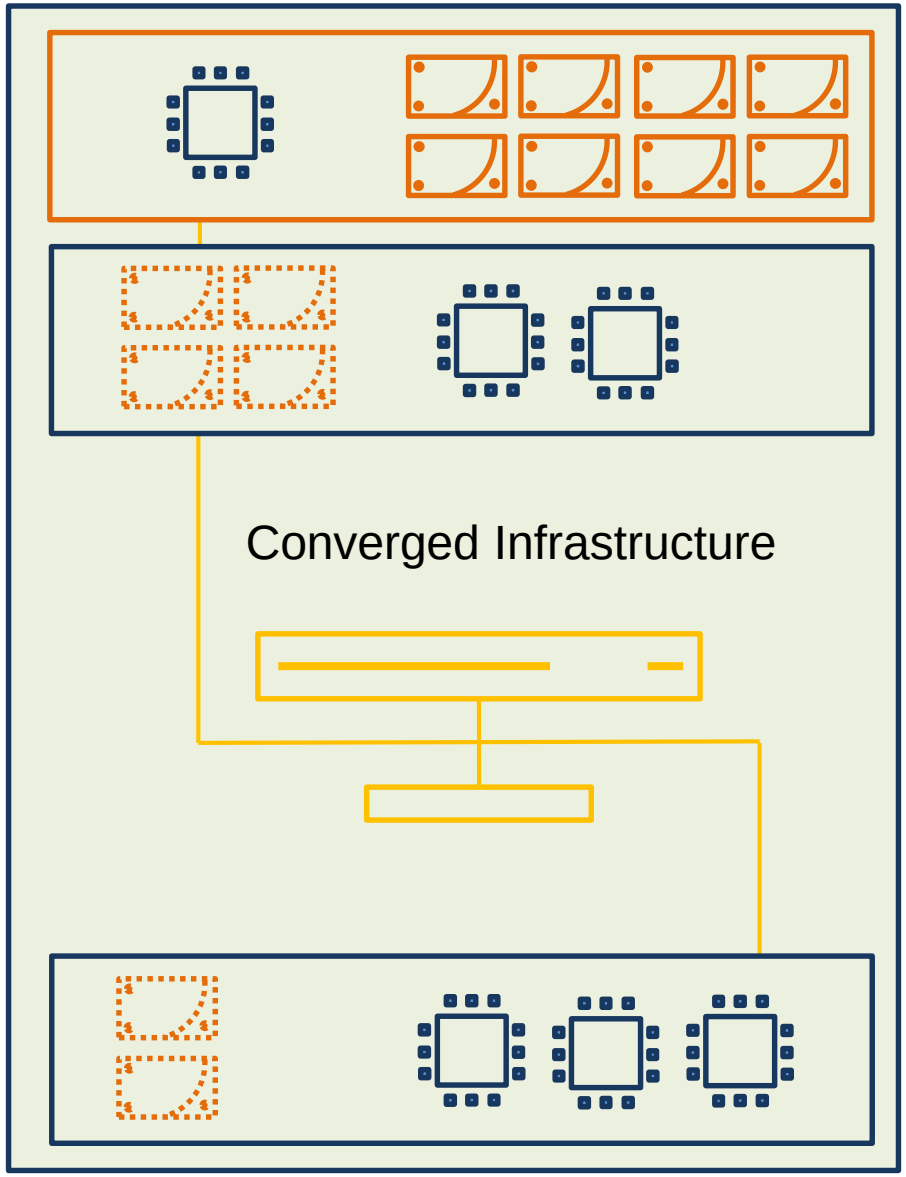
- Aviation
- Energy
- Computer Graphics
- Oil&Gas

and others

Cloud Deployments

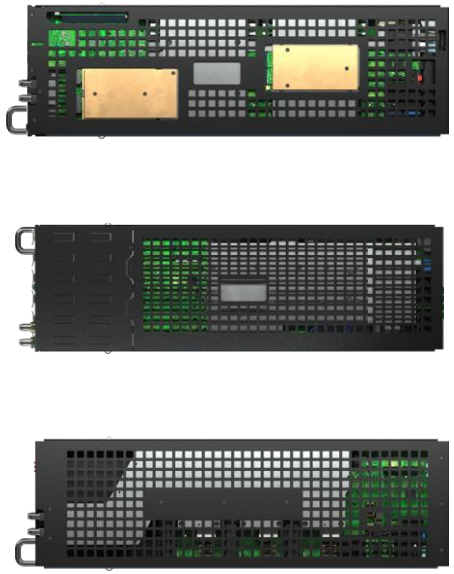
From Rackscale to Composable





RSC Composable Lineup

Unified Rack
Up to 153 nodes
0.64m², 2m height
(42U)



Compute nodes
Hyperconverged nodes
Storage nodes

RSC BasIS CDI
Orchestration Platform

Modern Approaches

1. Hyperconvergence

- Performing compute and storage tasks on unified servers
- Horizontal scalability
- Dynamic software orchestration

2. Disaggregation and dynamic composition

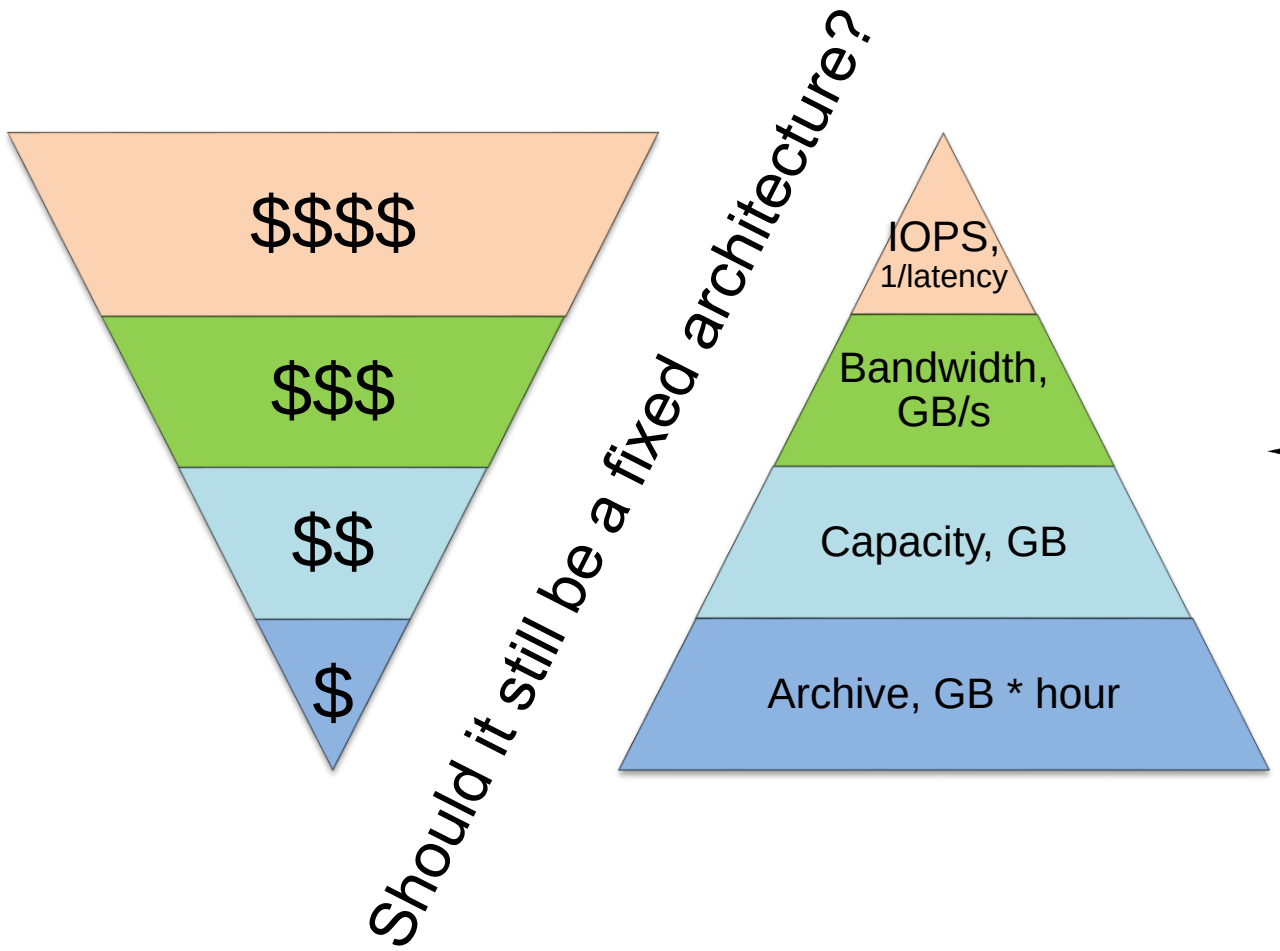
- Transforming distributed storage and compute elements into pools of devices
- Dynamically switching devices between servers with virtually zero "cost" (latency, CPU load)

3. End-to-end Orchestration

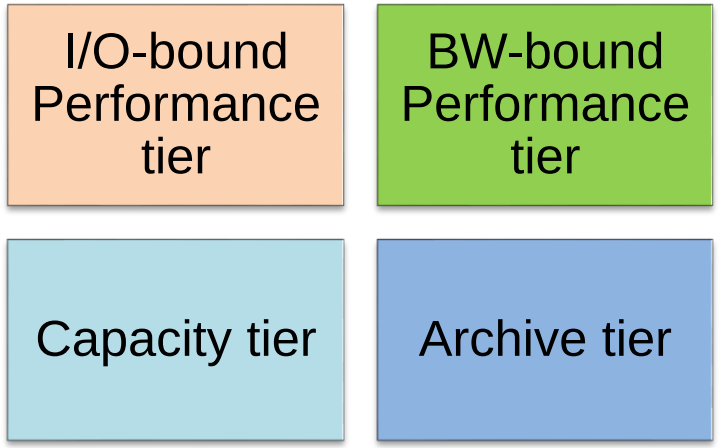
- Rebuilding configurations of the entire system "on the fly"
- Creation of several "domains" (clusters) with different configurations

RSC Vision on Multi Tier Architecture

New storage technologies open an opportunity to decompose storage into different tiers for different processing models



Composable Disaggregated Infrastructure brings on-demand resource allocation



Transformation of storage architectures

- Should sustain peak system loads
- Extremely high cost
- Should be scaled with compute growth

- Helps to sustain system overload spikes
- Still considered a huge storage investment
- Underutilized under compute-bound workloads

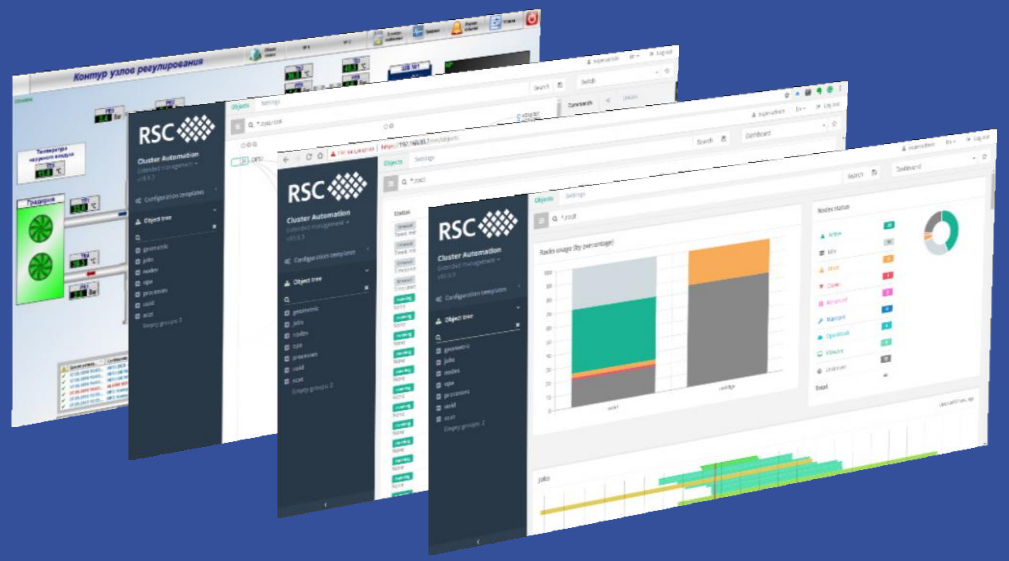
Decoupled storage tiers:
- Task-defined on-demand configuration
- Balanced Performance
- Configurable lifetime

Storage performance scales with the compute growth

Static Hardware-Centric

Dynamic Workflow-Centric

RSC BasIS Platform



RSC BasIS Orchestration

Vertical integration of
Hardware, Software
and Infrastructure components

Knowledge about all
datacenter objects and their
connections

Microagent Mesh for
Cluster Automation

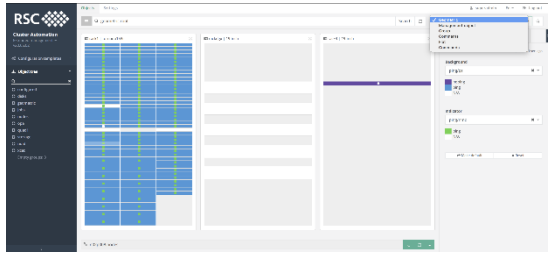
App Repository

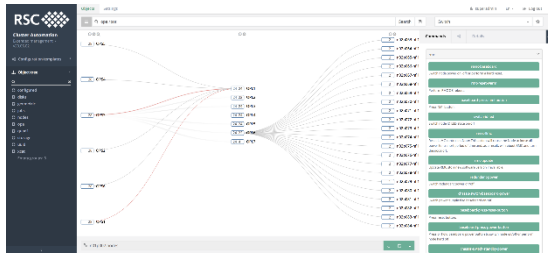
Agents

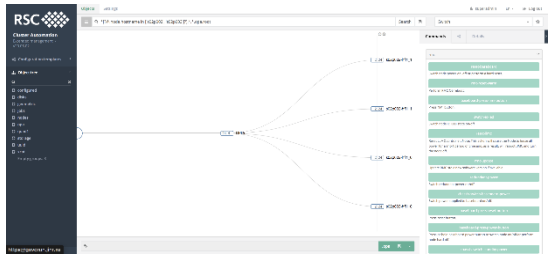
Agent Lifecycle

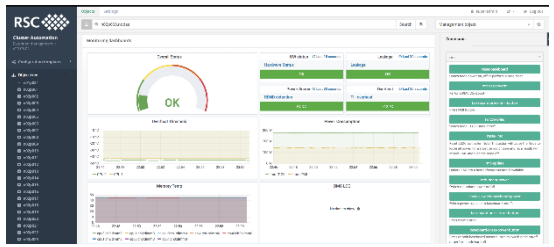
Messaging system

SDK









Knowledge of objects

- Auto-discovery
- Inventory and classification
- Knowledge of topologies
- Dynamic selection based on Query language

Continuous configuration

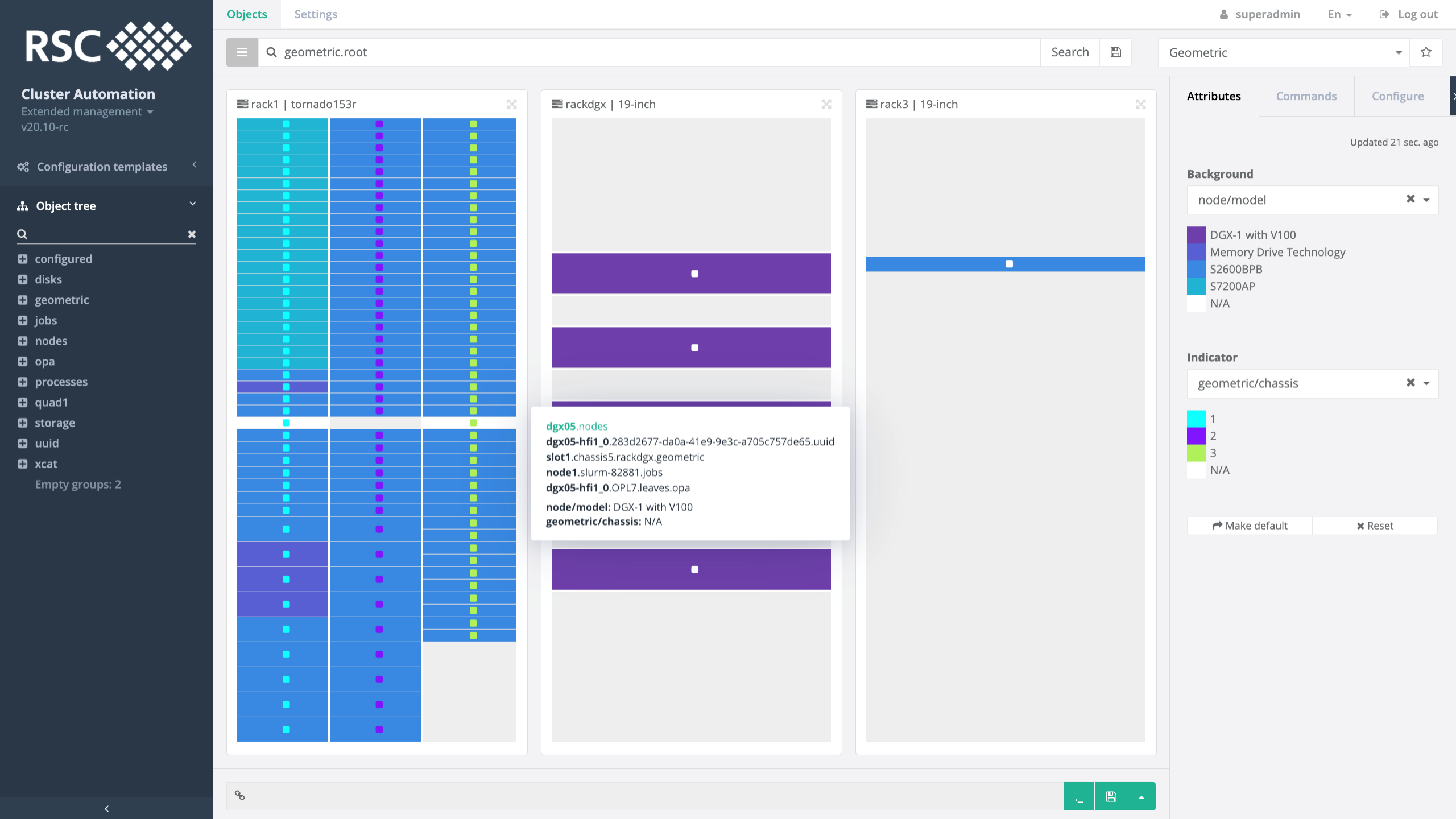
- Repository of configuration
- Maintaining consistency

Group Commands Execution

- Human operator – Platform
- Agent to agent

Monitoring

- Dynamical status representation
- GUI for drill-down analysis
- Problem-oriented dashboards


RSC
Cluster Automation
Extended management
v20.10-rc
Configuration templates
Object tree
configured
disks
geometric
jobs
nodes
opa
processes
quad1
storage
uuid
xcat
Empty groups: 2
Objects
Settings
superadmin
En
Log out
geometric.root
Search
Geometric
rack1 | tornado153r
rackdgx | 19-inch
rack3 | 19-inch
dgx05.nodes
dgx05-hfi1_0.283d2677-da0a-41e9-9e3c-a705c757de65.uuid
slot1.chassis5.rackdgx.geometric
node1.slurm-82881.jobs
dgx05-hfi1_0.OPL7.leaves.opa
node/model: DGX-1 with V100
geometric/chassis: N/A
Attributes
Commands
Configure
Updated 21 sec. ago
Background
node/model
DGX-1 with V100
Memory Drive Technology
S2600BPB
S7200AP
N/A
Indicator
geometric/chassis
1
2
3
N/A
Make default
Reset

RSC

Cluster Automation
Extended management
v20.10-rc

Configuration templates

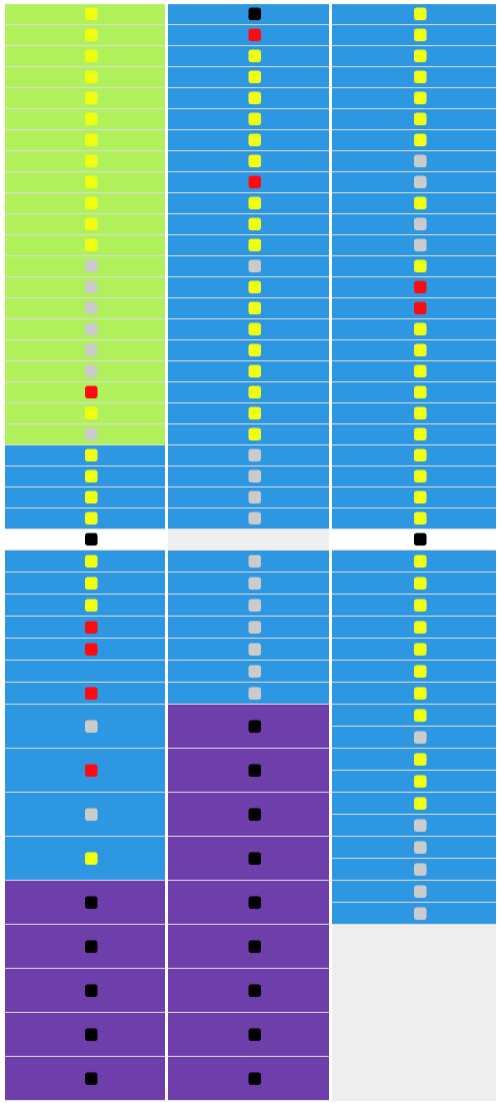
Object tree

- configured
- disks
- geometric
- jobs
- nodes
- opa
- processes
- quad1
- storage
- uuid
- xcat

Empty groups: 2

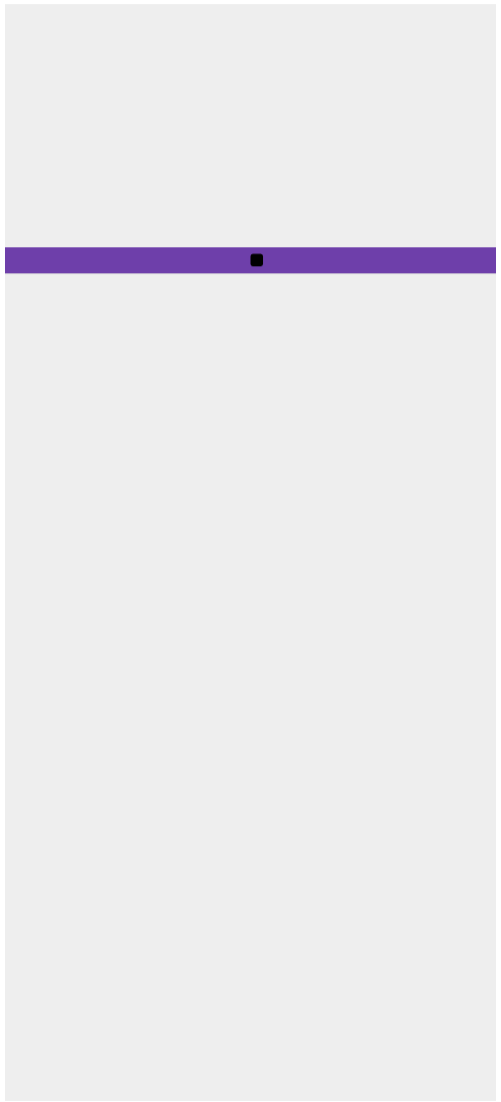
geometric.root | Search | Geometric

rack1 | tornado153r



rackdgx | 19-inch

rack3 | 19-inch



Attributes | Commands | Configure

Updated 27 sec. ago

Background

cpu/cores

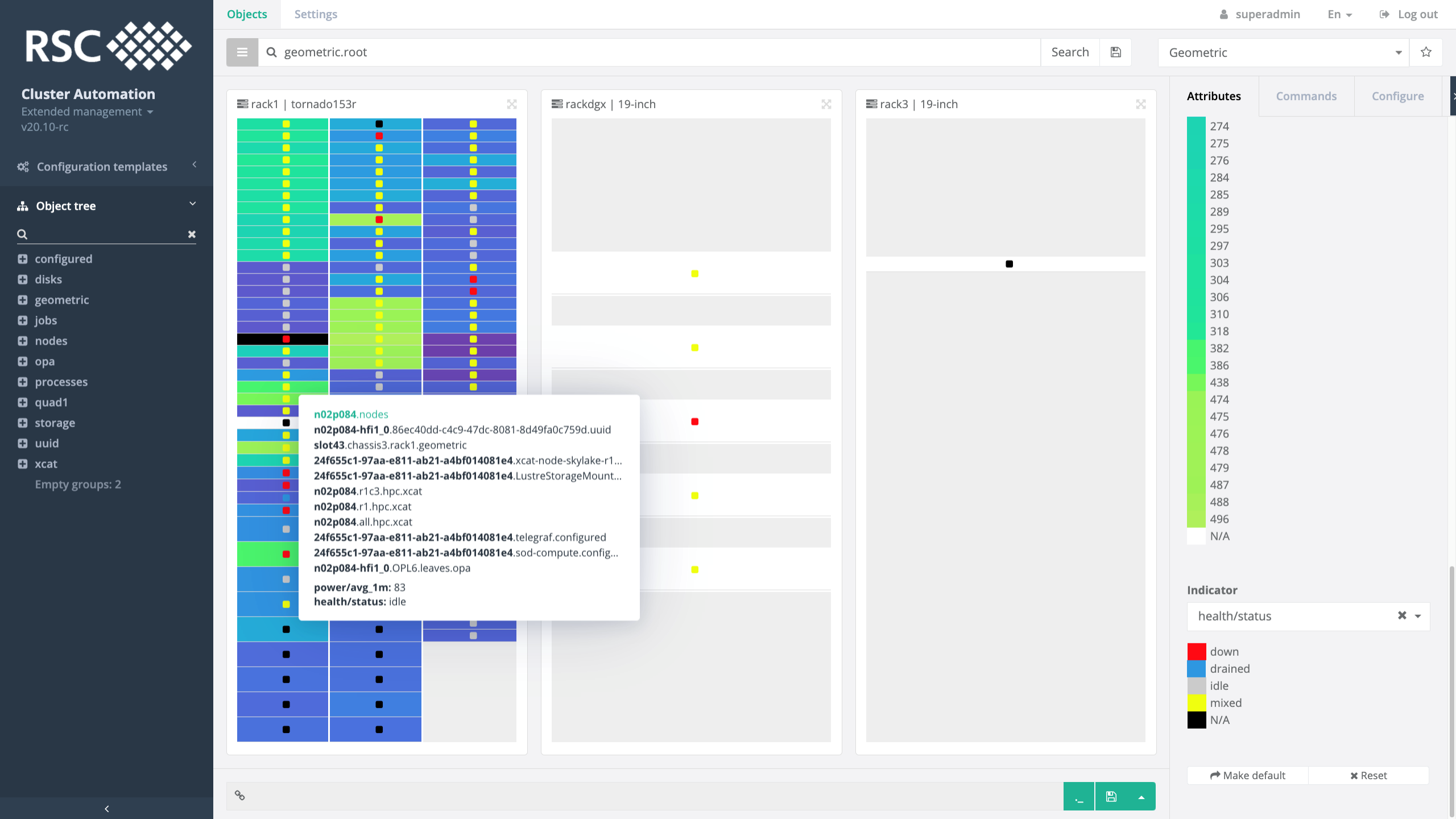
- 36
- 48
- 72
- N/A

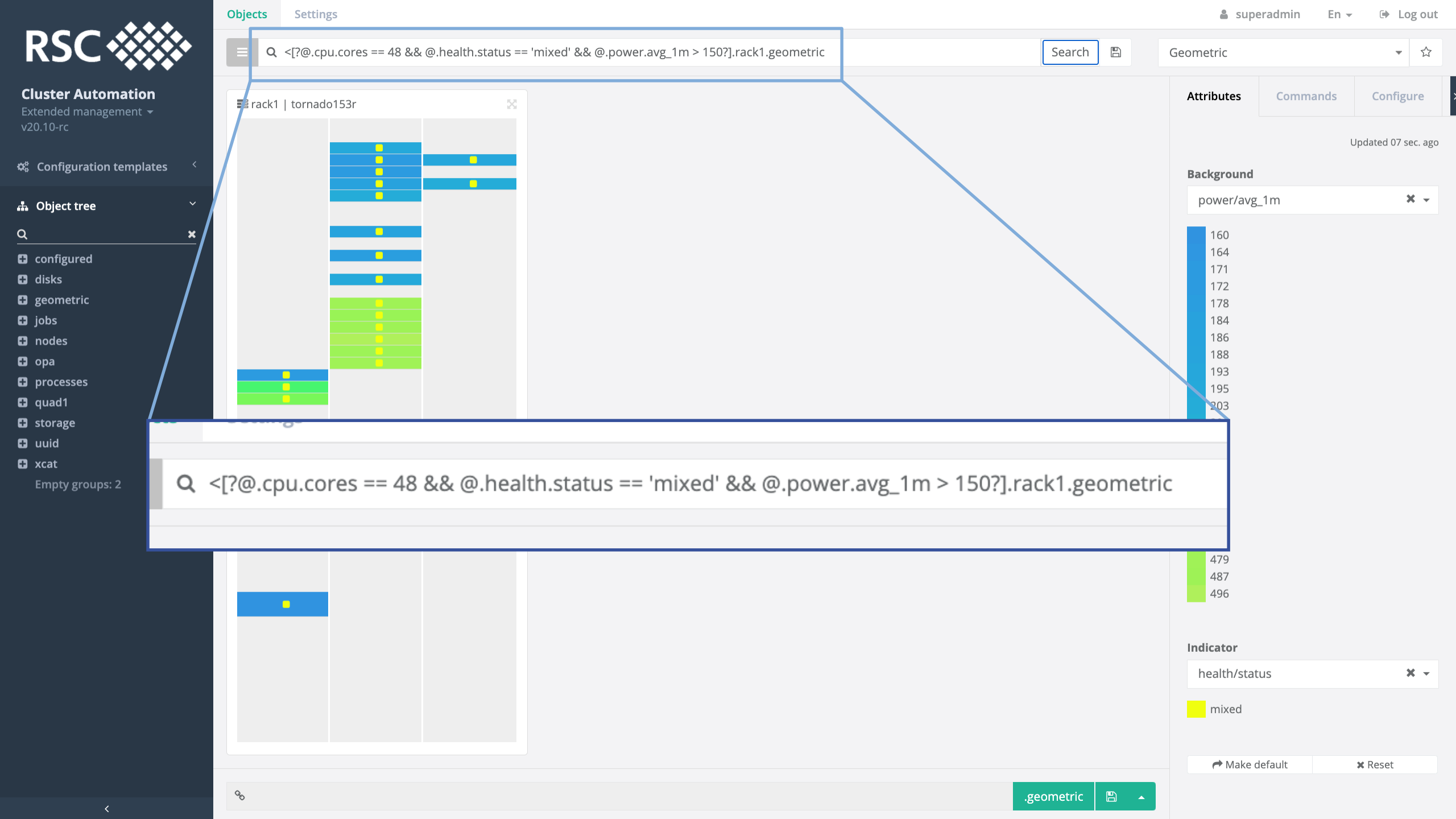
Indicator

health/status

- down
- drained
- idle
- mixed
- N/A

Make default | Reset


RSC
Cluster Automation
Extended management
v20.10-rc
Configuration templates
Object tree
configured
disks
geometric
jobs
nodes
opa
processes
quad1
storage
uuid
xcat
Empty groups: 2
Objects
Settings
superadmin
En
Log out
geometric.root
Search
Geometric
rack1 | tornado153r
rackdgx | 19-inch
rack3 | 19-inch
n02p084.nodes
n02p084-hfi1_0.86ec40dd-c4c9-47dc-8081-8d49fa0c759d.uuid
slot43.chassis3.rack1.geometric
24f655c1-97aa-e811-ab21-a4bf014081e4.xcat-node-skylake-r1...
24f655c1-97aa-e811-ab21-a4bf014081e4.LustreStorageMount...
n02p084.r1c3.hpc.xcat
n02p084.r1.hpc.xcat
n02p084.all.hpc.xcat
24f655c1-97aa-e811-ab21-a4bf014081e4.telegraf.configured
24f655c1-97aa-e811-ab21-a4bf014081e4.sod-compute.config...
n02p084-hfi1_0.OPL6.leaves.opa
power/avg_1m: 83
health/status: idle
Attributes
Commands
Configure
274
275
276
284
285
289
295
297
303
304
306
310
318
382
386
438
474
475
476
478
479
487
488
496
N/A
Indicator
health/status
down
drained
idle
mixed
N/A
Make default
Reset


RSC
Cluster Automation
Extended management
v20.10-rc
Configuration templates
Object tree
configured
disks
geometric
jobs
nodes
opa
processes
quad1
storage
uuid
xcat
Empty groups: 2
Objects
Settings
superadmin
En
Log out
<[?@.cpu.cores == 48 && @.health.status == 'mixed' && @.power.avg_1m > 150?].rack1.geometric
Search
Geometric
rack1 | tornado153r
Attributes
Commands
Configure
Updated 07 sec. ago
Background
power/avg_1m
160
164
171
172
178
184
186
188
193
195
479
487
496
Indicator
health/status
mixed
Make default
Reset
.geometric

QDSL

DNS-like path: `slot1.chassis2.rack3.geometric, n01p001.nodes.root, switch2.access.networks`

Wildcard: `*.*.rack1.geometric, *.chassis1.*.geometric`

Ranges: `[n01p001, n02p002, n03p003].nodes, sky[08-10], n01p[001-002, 010-011]`

Filters: `*[?@._meta.type == 'head' && @.health.status in ['down', 'drain'] || @.metrics.temp > 70?]`

`<[?@.power.avg_1m > 300?].geometric` All nodes with 1 minute power consumption average is greater then 300W

NVMe Storage device location:

```
[root@gvr-head1 ~]# rsc-ba qdsl "*[?@.pcie.numa == 1 && @.block.size == '14 TiB'?].*.nodes" --name | jq .[].name
"nvme0n1"
"nvme10n1"
"nvme11n1"
"nvme12n1"
"nvme13n1"
"nvme14n1"
"nvme15n1"
"nvme1n1"
"nvme2n1"
"nvme3n1"
"nvme4n1"
"nvme5n1"
"nvme6n1"
"nvme7n1"
"nvme8n1"
"nvme9n1"
```

RSC

Cluster Automation
Extended management
v20.10-rc

Configuration templates
Object tree
n02p023
n02p024
n02p025
n02p026
n02p027
n02p028
n02p029
n02p030
n02p031
n02p032
n02p033
n02p034
n02p035
n02p036
n02p037
n02p038
n02p039
n02p040
n02p041
n02p042
n02p043
n02p044
n02p045
n02p046
n02p047
n02p048
n02p049

n02p028.nodes Search Management object

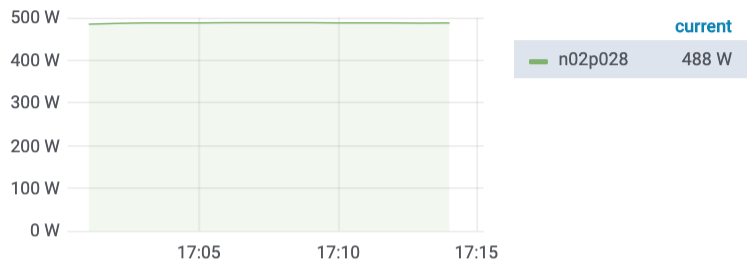
Monitoring dashboards

Power Source
RDND detection
ACDC

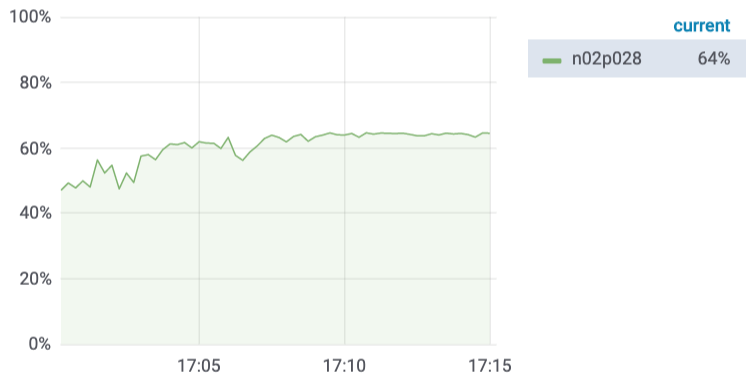
Leakage
Leakage
OK

Overheat
Till overheat
-29 °C

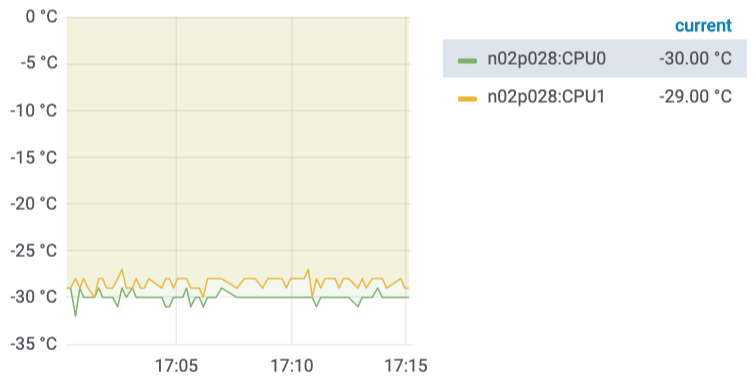
Power Consumption



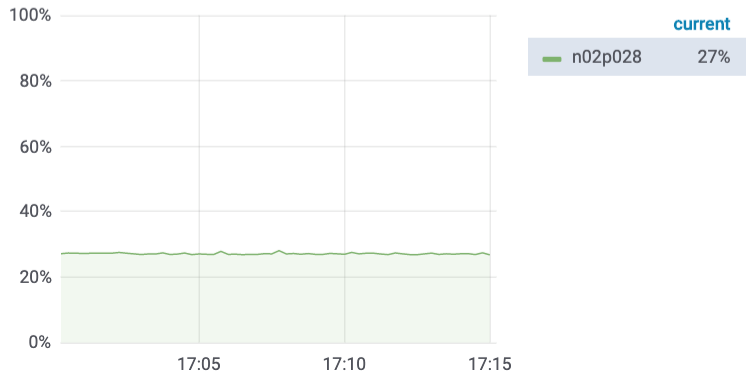
CPU Usage %



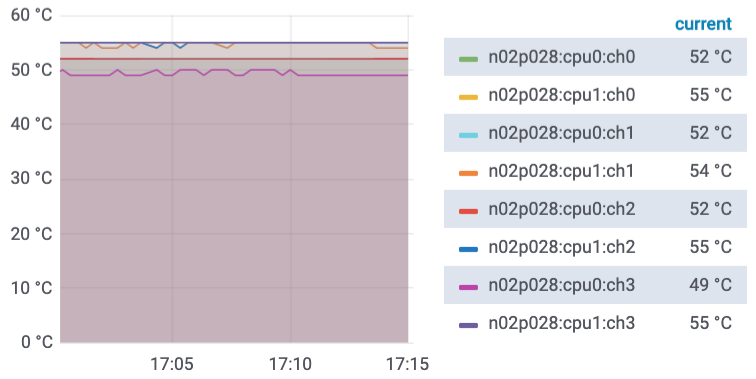
Overheat Threshold



Mem Usage %



Memory Temp



Commands

nvme-target

export
Export Disks

un-export
Un-Export Disks

export-all
Export all Disks

un-export-all
Un-Export Disks

rmc

reset-baseboard
Switch node power on, off or perform a hard reset.

rmc-reset-warm
Perform RMC OS reboot.

baseboard-press-nmi-button
Press NMI button.

switch-id-led
Switch node ID LED state on/off

reset-llmc
Reset LLMC controller. Note: This action will cause the Node to loose all power for a short period of time and, as a result, will reboot BMC and turn the node off.

rmc-update
Update RMC sto a new software version if available.

redundant-power

RSC

Cluster Automation
Extended management
v20.10-rc

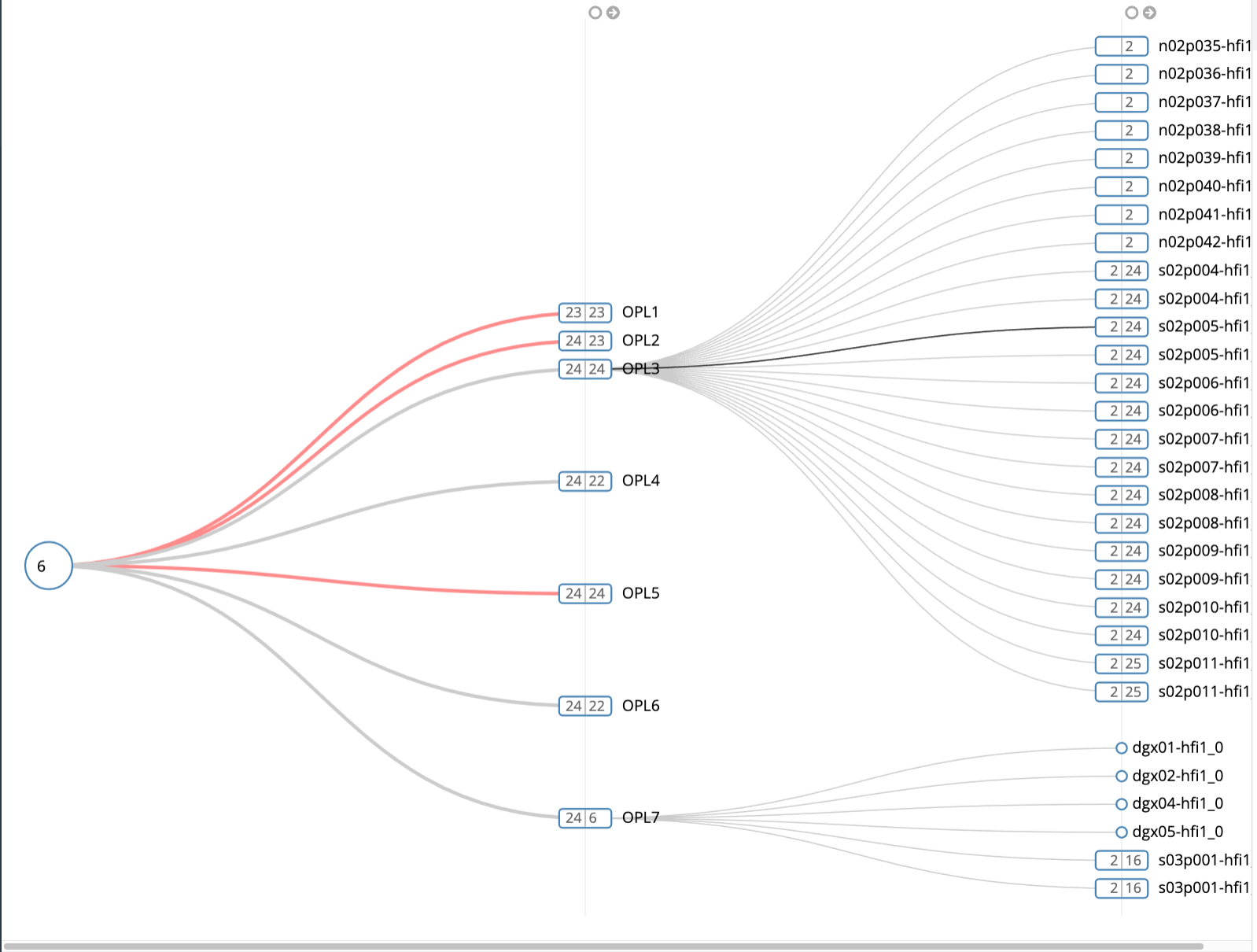
Configuration templates

Object tree

- configured
- disks
- geometric
- jobs
- nodes
- opa
- processes
- quad1
- storage
- uuid
- xcat

Empty groups: 2

opa.root | Search | Switch



Commands | Details

rmc

reset-baseboard
Switch node power on, off or perform a hard reset.

rmc-reset-warm
Perform RMC OS reboot.

baseboard-press-nmi-button
Press NMI button.

switch-id-led
Switch node ID LED state on/off

reset-llmc
Reset LLMC controller. Note: This action will cause the Node to loose all power for a short period of time and, as a result, will reboot BMC and turn the node off.

rmc-update
Update RMC sto a new software version if available.

redundant-power
Switch redundant power on/off.

chassis-switch-baseboard-power
Switch power supplied to baseboard on/off.

baseboard-press-reset-button
Press reset button.

baseboard-press-power-button
Press or hold baseboard power button to switch node on/off or perform node hard off.

chassis-switch-standby-power
Switch standby power on/off.

rmc-reset-cold
Perform full power cycle on a RMC.

RSC

Cluster Automation
Extended management
v20.10-rc

Configuration templates

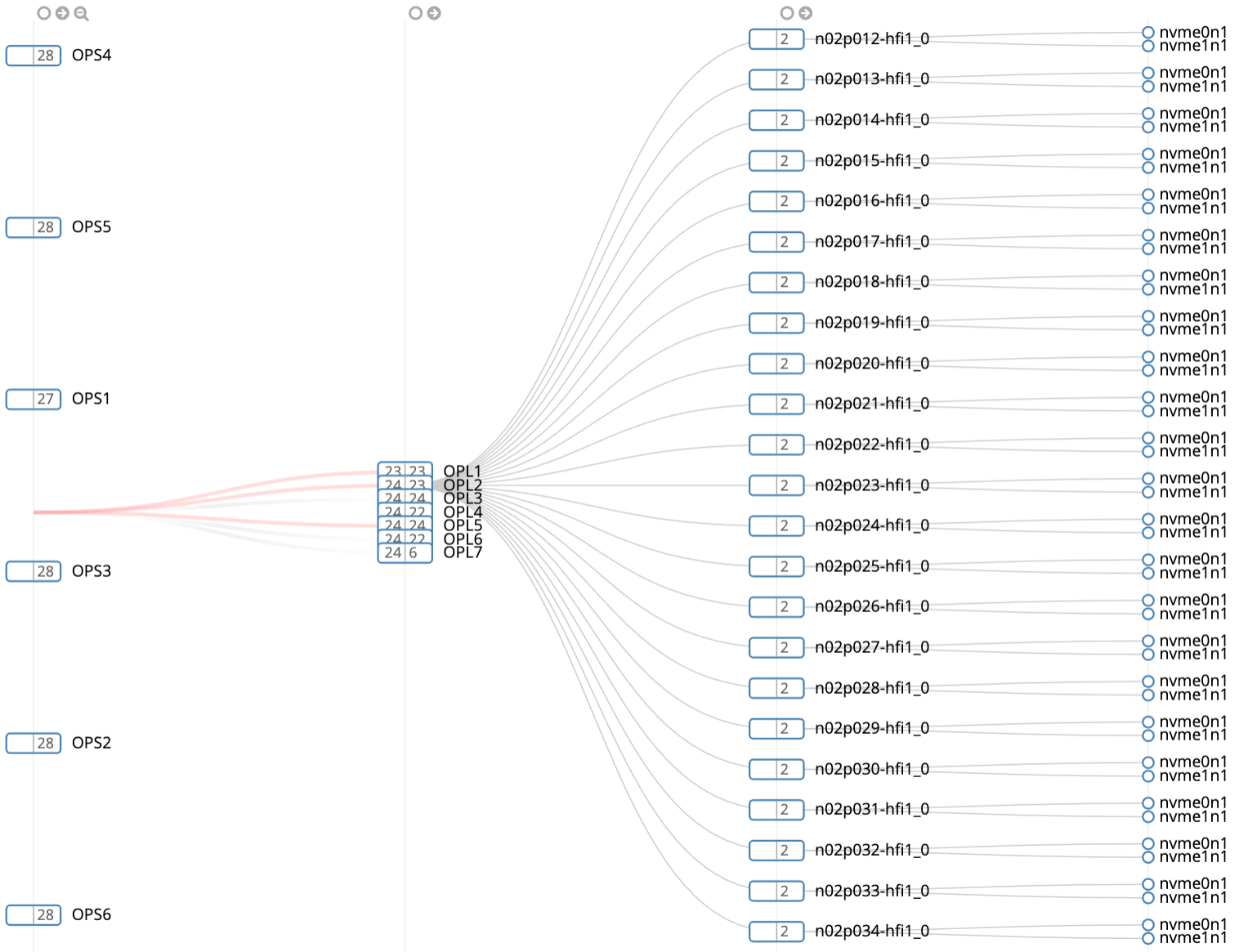
Object tree

- configured
- disks
- geometric
- jobs
- nodes
- opa
- processes
- quad1
- storage
- uuid
- xcat

Empty groups: 2

opa.root Search

Switch



Commands | Details

rmc

reset-baseboard
Switch node power on, off or perform a hard reset.

rmc-reset-warm
Perform RMC OS reboot.

baseboard-press-nmi-button
Press NMI button.

switch-id-led
Switch node ID LED state on/off

reset-llmc
Reset LLMC controller. Note: This action will cause the Node to loose all power for a short period of time and, as a result, will reboot BMC and turn the node off.

rmc-update
Update RMC sto a new software version if available.

redundant-power
Switch redundant power on/off.

chassis-switch-baseboard-power
Switch power supplied to baseboard on/off.

baseboard-press-reset-button
Press reset button.

baseboard-press-power-button
Press or hold baseboard power button to switch node on/off or perform node hard off.

chassis-switch-standby-power
Switch standby power on/off.

rmc-reset-cold
Perform full power cycle on a RMC.

RSC

Cluster Automation
Extended management
v20.10-rc

Configuration templates

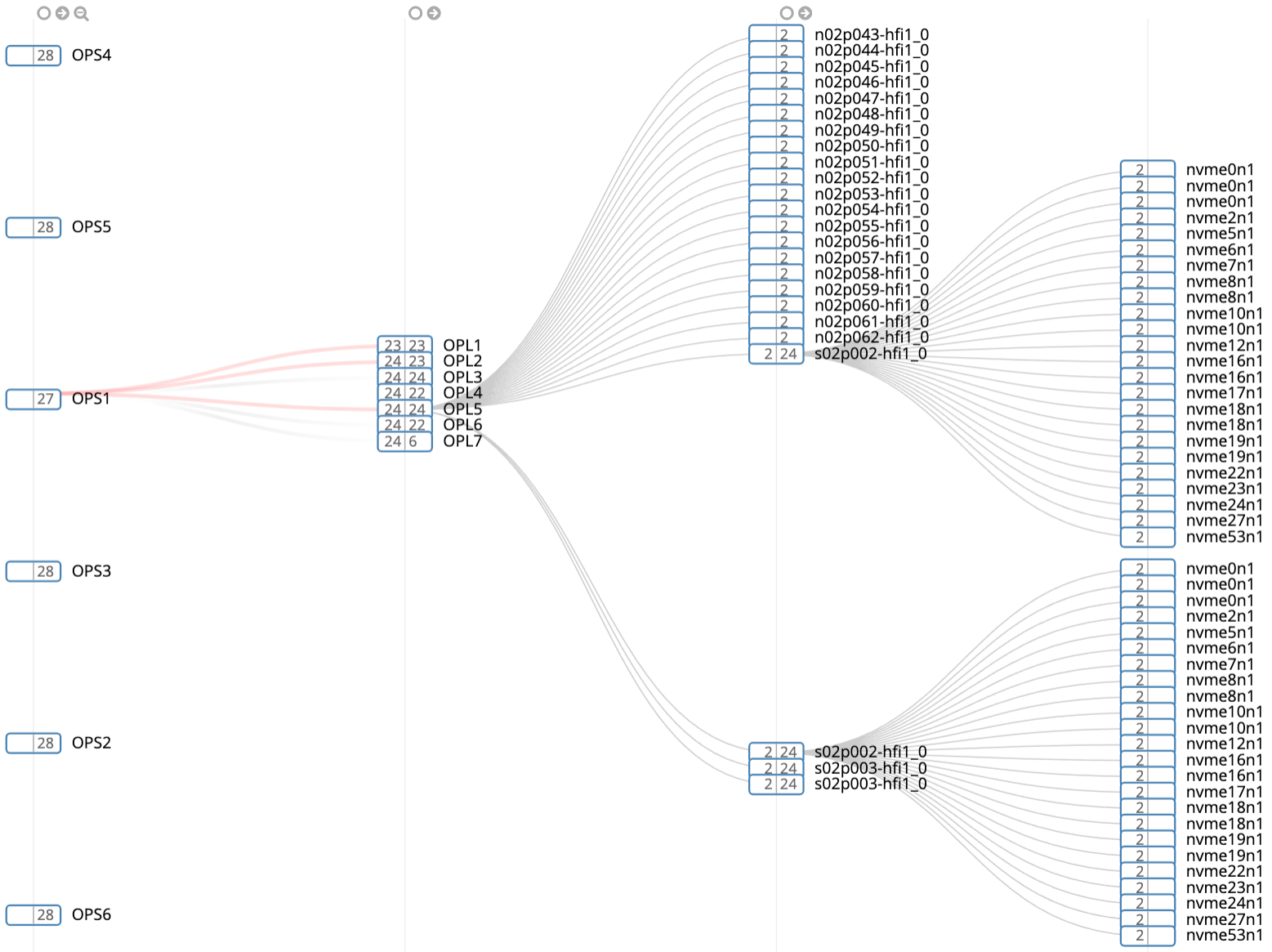
Object tree

- configured
- disks
- geometric
- jobs
- nodes
- opa
- processes
- quad1
- storage
- uuid
- xcat

Empty groups: 2

opa.root Search

Switch



Commands | Details

rmc

reset-baseboard
Switch node power on, off or perform a hard reset.

rmc-reset-warm
Perform RMC OS reboot.

baseboard-press-nmi-button
Press NMI button.

switch-id-led
Switch node ID LED state on/off

reset-llmc
Reset LLMC controller. Note: This action will cause the Node to loose all power for a short period of time and, as a result, will reboot BMC and turn the node off.

rmc-update
Update RMC sto a new software version if available.

redundant-power
Switch redundant power on/off.

chassis-switch-baseboard-power
Switch power supplied to baseboard on/off.

baseboard-press-reset-button
Press reset button.

baseboard-press-power-button
Press or hold baseboard power button to switch node on/off or perform node hard off.

chassis-switch-standby-power
Switch standby power on/off.

rmc-reset-cold
Perform full power cycle on a RMC.

RSC

Cluster Automation
Extended management
v20.10-rc

Configuration templates

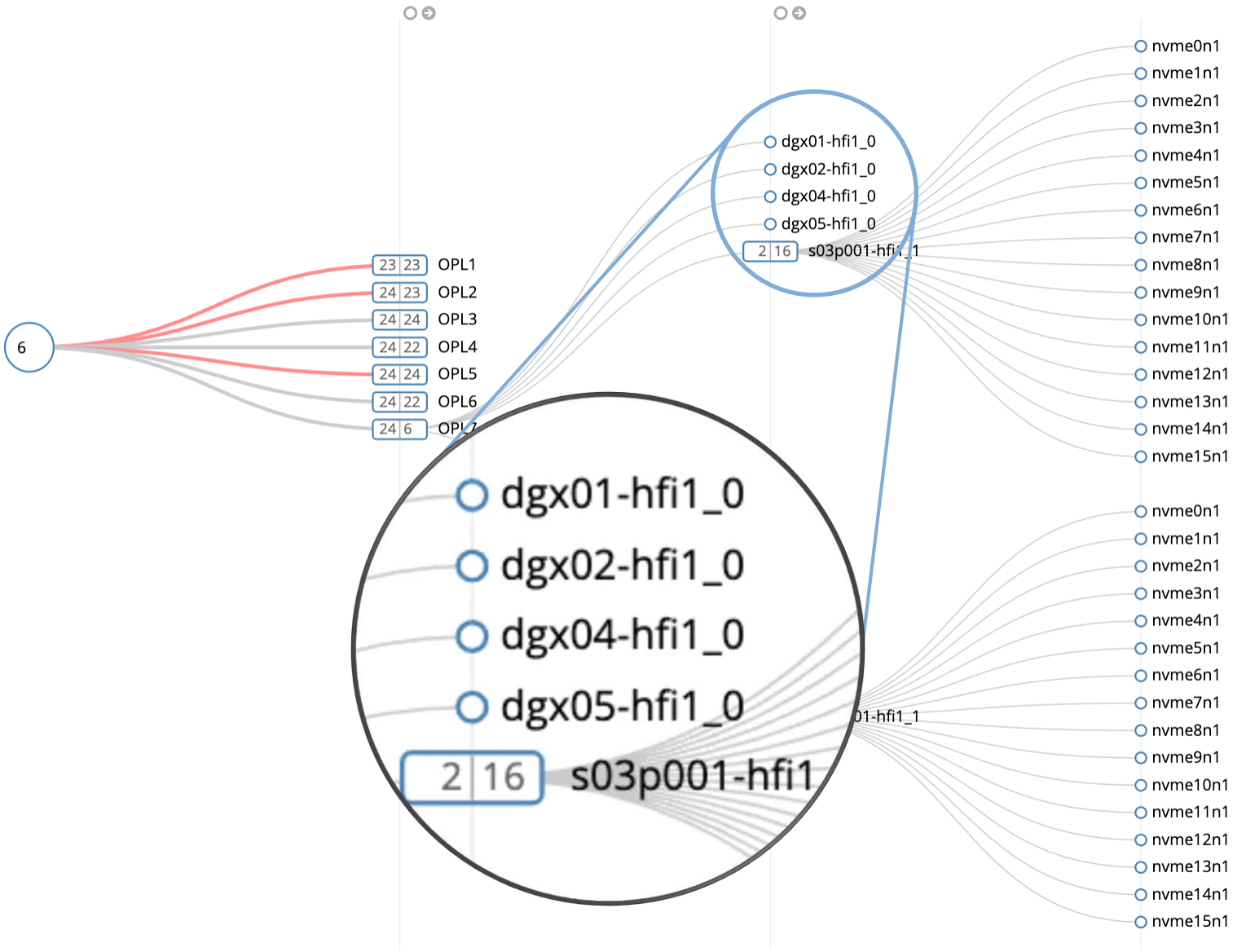
Object tree

- configured
- disks
- geometric
- jobs
- nodes
- opa
- processes
- quad1
- storage
  - control
  - disk-pool
  - lit
  - metastor
  - node-pool
  - objstor
- uuid
- xcat

Empty groups: 2

opa.root Search

Switch



Commands | Details

rmc

reset-baseboard
Switch node power on, off or perform a hard reset.

rmc-reset-warm
Perform RMC OS reboot.

baseboard-press-nmi-button
Press NMI button.

switch-id-led
Switch node ID LED state on/off

reset-llmc
Reset LLMC controller. Note: This action will cause the Node to loose all power for a short period of time and, as a result, will reboot BMC and turn the node off.

rmc-update
Update RMC sto a new software version if available.

redundant-power
Switch redundant power on/off.

chassis-switch-baseboard-power
Switch power supplied to baseboard on/off.

baseboard-press-reset-button
Press reset button.

baseboard-press-power-button
Press or hold baseboard power button to switch node on/off or perform node hard off.

chassis-switch-standby-power
Switch standby power on/off.

rmc-reset-cold
Perform full power cycle on a RMC.

RSC

Cluster Automation
Extended management
v20.10-rc

Configuration templates

Object tree

- configured
- disks
- geometric
- jobs
- nodes
- opa
- processes
- quad1
- storage
- uuid
- xcat

Empty groups: 2

*[?@.pcie.numa == 1?].s02p001.nodes Search Flat

Items: 0 / 9 Sort by: fqdn Update freq.: 30 sec.

nvme6n1
nvme6n1.s02p001.nodes
block (block)
device: /dev/nvme6n1
disk (nvme)
model: INTEL SSDPELKX020T8
pcie (pcie)
numa: 1

nvme7n1
nvme7n1.s02p001.nodes
block (block)
device: /dev/nvme7n1
disk (nvme)
model: INTEL SSDPELKX020T8
pcie (pcie)
numa: 1

nvme8n1
nvme8n1.s02p001.nodes
block (block)
device: /dev/nvme8n1
disk (nvme)
model: INTEL SSDPELKX020T8
pcie (pcie)
numa: 1

nvme9n1
nvme9n1.s02p001.nodes
block (block)
device: /dev/nvme9n1
disk (nvme)
model: INTEL SSDPELKX020T8
pcie (pcie)
numa: 1

nvme10n1
nvme10n1.s02p001.nodes
block (block)
device: /dev/nvme10n1
disk (nvme)
model: INTEL SSDPELKX020T8
pcie (pcie)
numa: 1

nvme11n1
nvme11n1.s02p001.nodes
block (block)
device: /dev/nvme11n1
disk (nvme)
model: INTEL SSDPELKX020T8
pcie (pcie)
numa: 1

nvme16n1
nvme16n1.s02p001.nodes
block (block)
device: /dev/nvme6n1
disk (nvme)
model: INTEL SSDPELKX020T8
pcie (pcie)
numa: 1

nvme19n1
nvme19n1.s02p001.nodes
block (block)
device: /dev/nvme9n1
disk (nvme)
model: INTEL SSDPELKX020T8
pcie (pcie)
numa: 1

nvme23n1
nvme23n1.s02p001.nodes
block (block)
device: /dev/nvme11n1
disk (nvme)
model: INTEL SSDPELKX020T8
pcie (pcie)
numa: 1

.nodes

Attributes Commands

nvme
- [ ] firmware-rev
- [x] model
- [ ] serial
- [ ] state
- [ ] vendor

nvmf
- [ ] allow_any_host

partition
- [ ] device
- [ ] fs-label
- [ ] fs-type
- [ ] id
- [ ] num-blocks
- [ ] offset
- [ ] size
- [ ] table-type

pcie
- [ ] address
- [ ] driver
- [ ] enabled
- [x] numa
- [ ] slot
- [ ] speed
- [ ] width

rdma
- [ ] host-traddr
- [ ] nqn

Usage scenarios and lifecycle

Current focus is to provide the best experience with the well known Lustre file system

Standalone configuration

by admin request

“Prolonged” deployments

Integration with other schedulers

- High Performance Computing (SLURM)
- Cloud (OpenStack, K8S) can be easily supported

On-demand during application run

Other usage models are possible

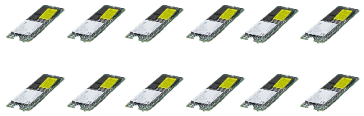
Hardware platform

Compute node

Highest density
2 CPU
384 GB RAM / 4 NVMe SSD

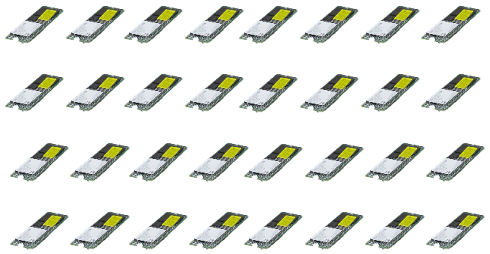
«Fat» compute node

2 CPU
12 NVMe SSD



Storage node

1 Petabyte
32 EDSFF.L NVMe SSD



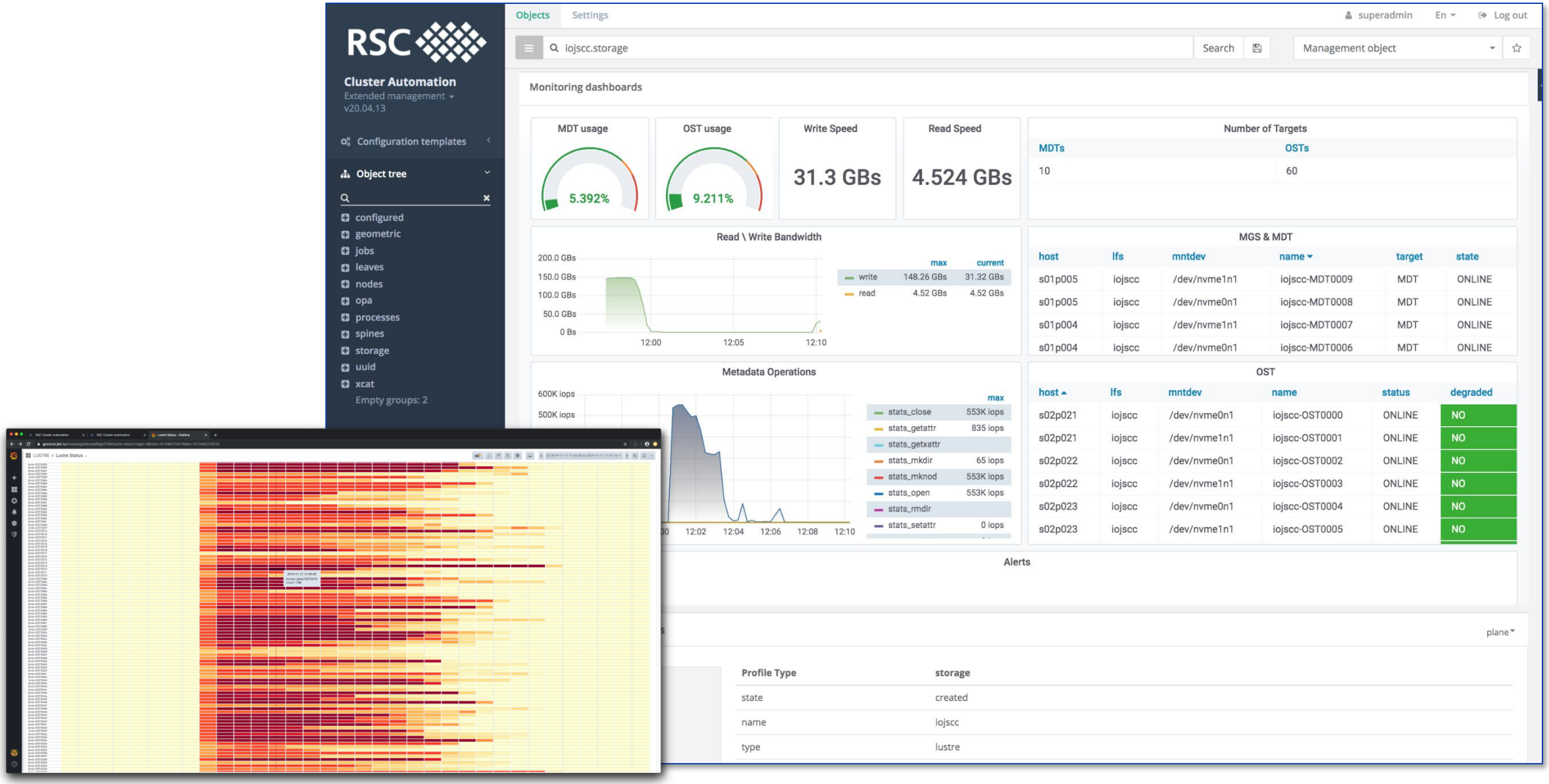
Lifecycle

- disk pools
- storage servers pools

- storage clusters
- storage services

```
API:
Select storage nodes
            create-node-pool        n01p[001-020].nodes
Select drives
            create-disk-pool        *[?@.pcie.numa == 0 && @.block.size == '1.8 TiB'?].disks
Storage initiation
            create-storage          lustre1.storage
Storage cluster initiation
            create-cluster          lustre1.storage
Storage services initiation
            create-service          serviceType (MGS, MDS, OSS), numOfDrivers, raidEngine, raidLevel
Map storage topology to disk and server pools
            distribute
Run storage
            storage-launch
Run high-availability services
            pacemaker-launch
```

Storage-on-Demand Dashboard



Resilience

High availability

Storage-on-demand HA configuration is based on disk RAIDs and node active-passive pairs

The Redundancy topology is based on the following knowledge:

- Power domains
- Cooling domains
- Disk distribution
- Network fabric topology
- Required redundancy on disk and node levels

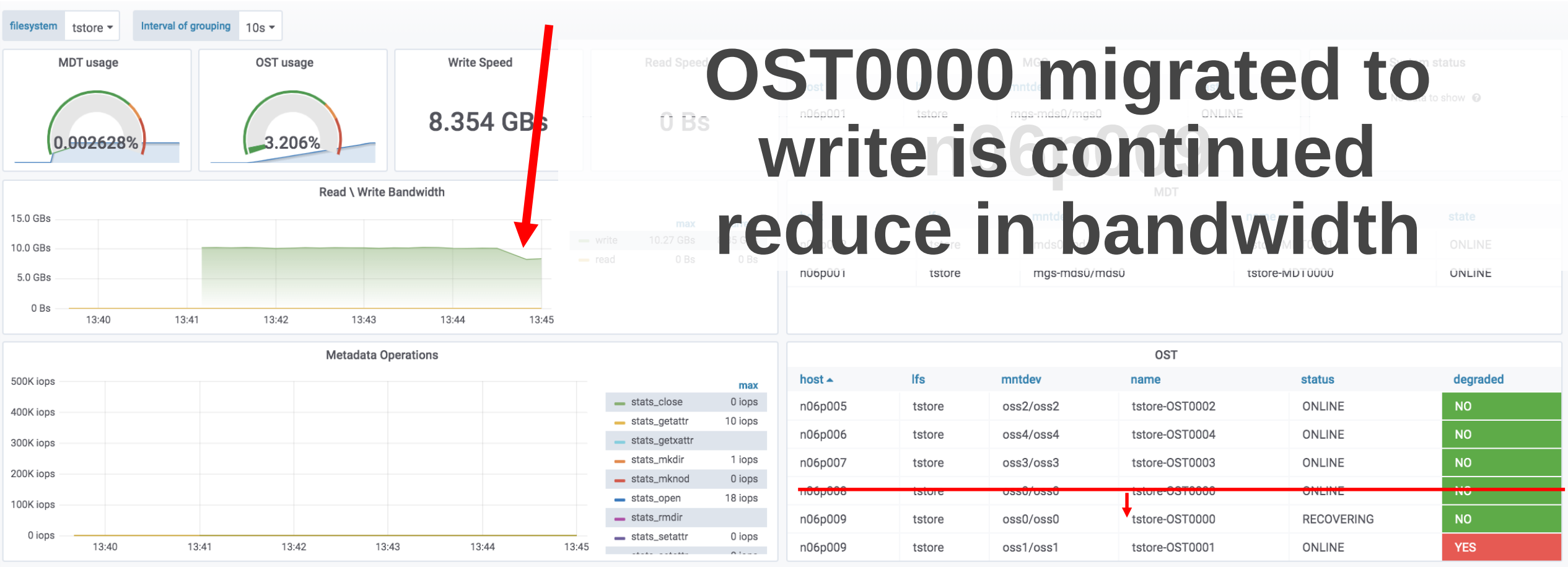
Optimal fault tolerant topology will be selected automatically or could be provided manually

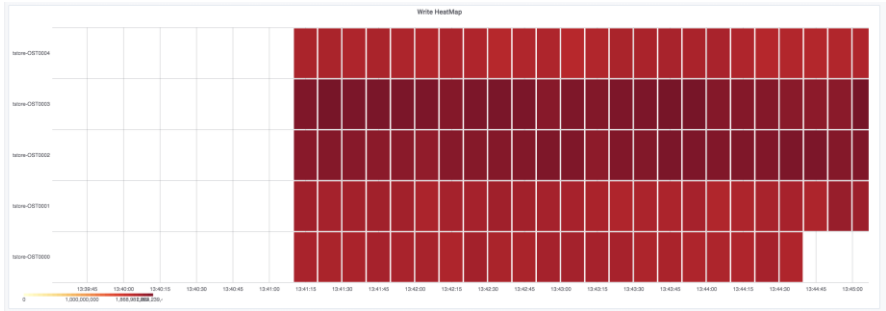
HA Realtime failover

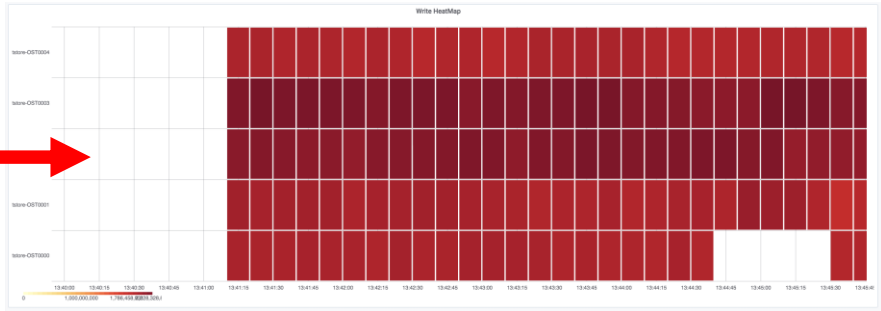
node n06p008 goes down
write to OST0000 is paused

HA Realtime failover

OST0000 migrated to write is continued reduce in bandwidth

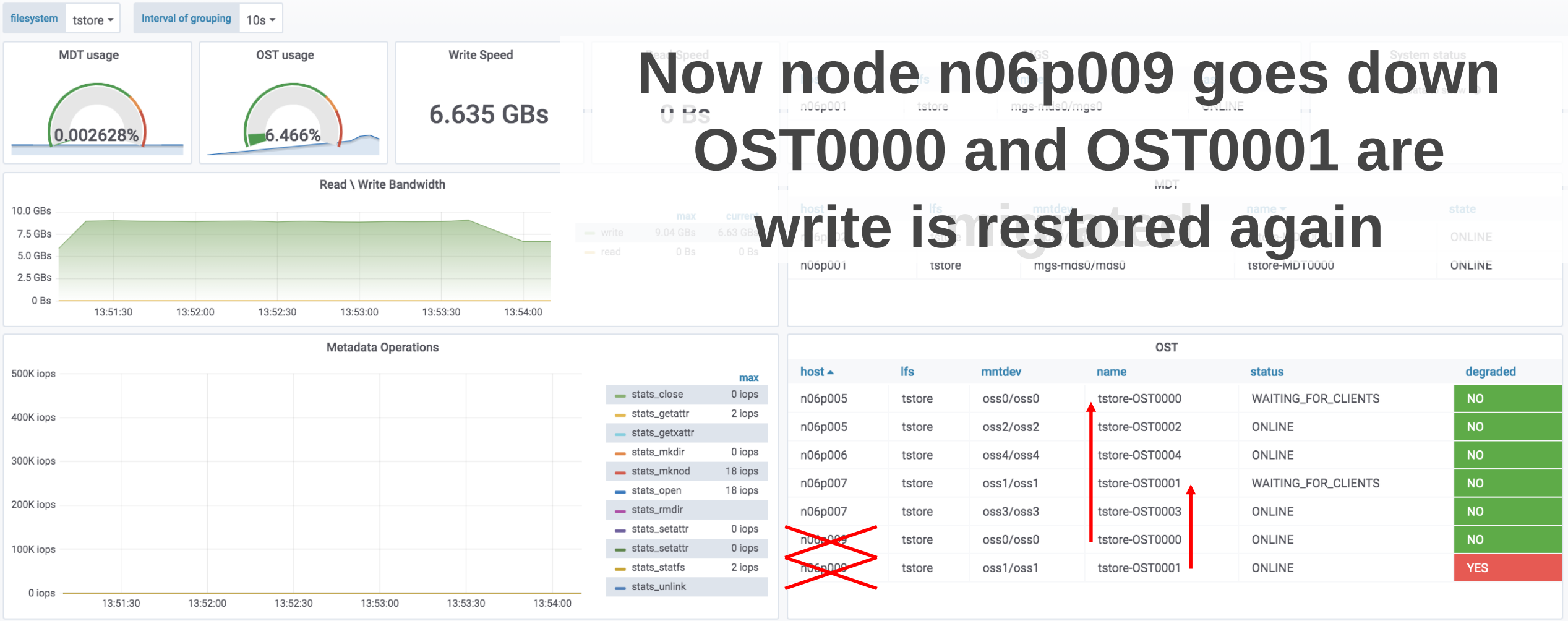




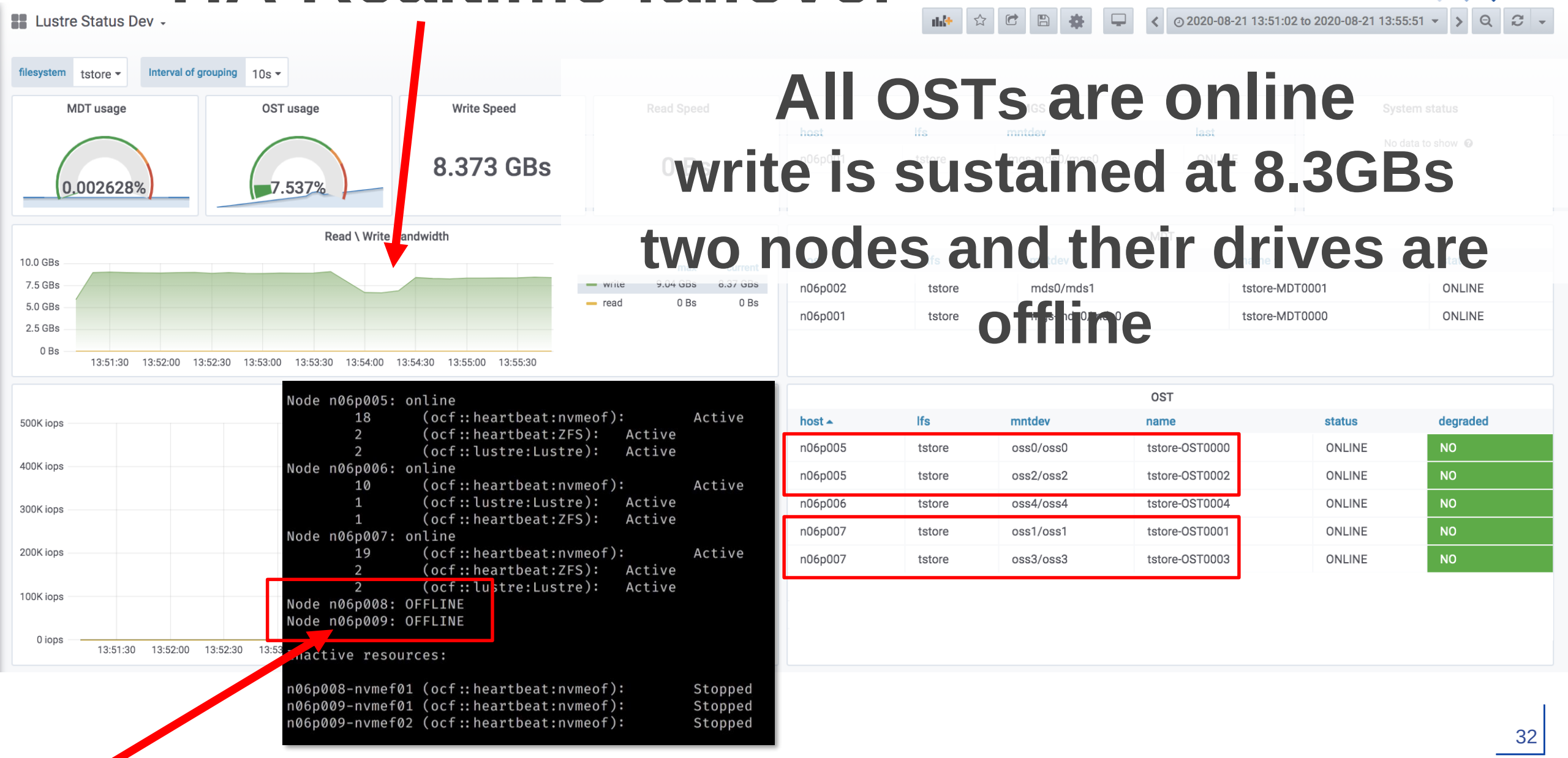


HA Realtime failover

Now node n06p009 goes down
OST0000 and OST0001 are
write is restored again



HA Realtime failover



All OSTs are online

write is sustained at 8.3GBs

two nodes and their drives are offline

Results

IO500 Benchmark 2020

Lustre on Demand in POLYTECH:

4 MDS (Total 8 MDT) Intel Optane NVMe
55 OSS (109 OST)* Intel NVMe
59 clients nodes with Intel Xeon Platinum 8268

* One OST NVMe became faulty during the benchmark

* * *

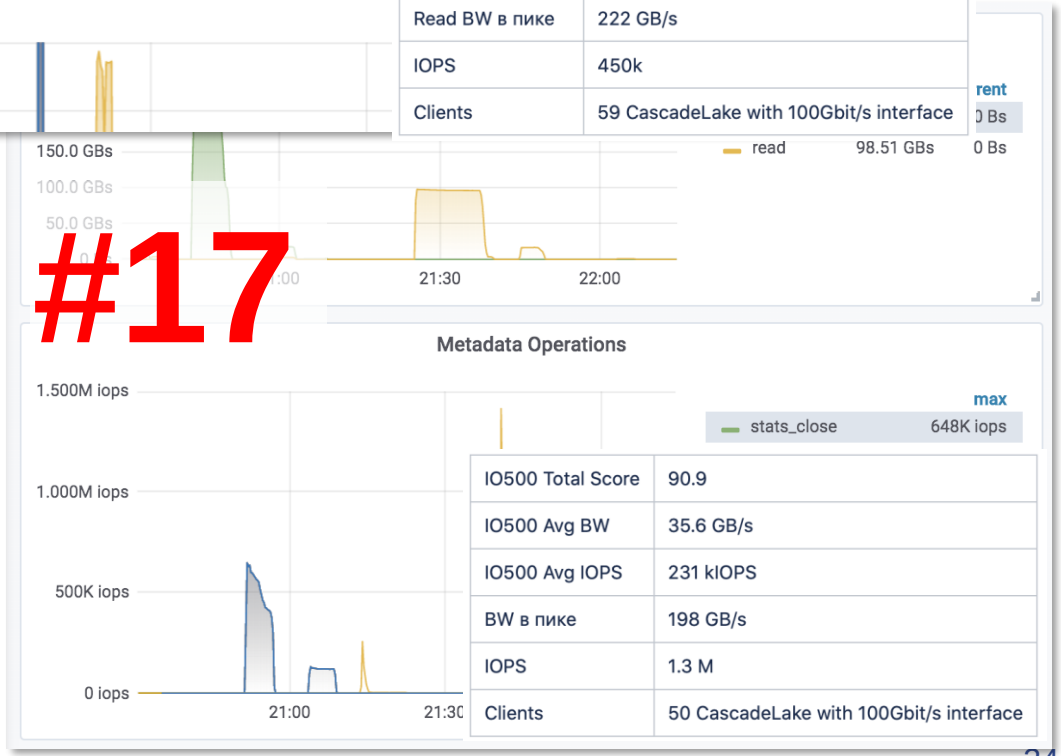
Lustre on Demand in JINR:

12 MDS (Total 24 MDT): Intel Optane NVMe
14 OSS (144 OST) Intel NVMe
50 clients nodes with Intel Xeon Platinum 8268

#22

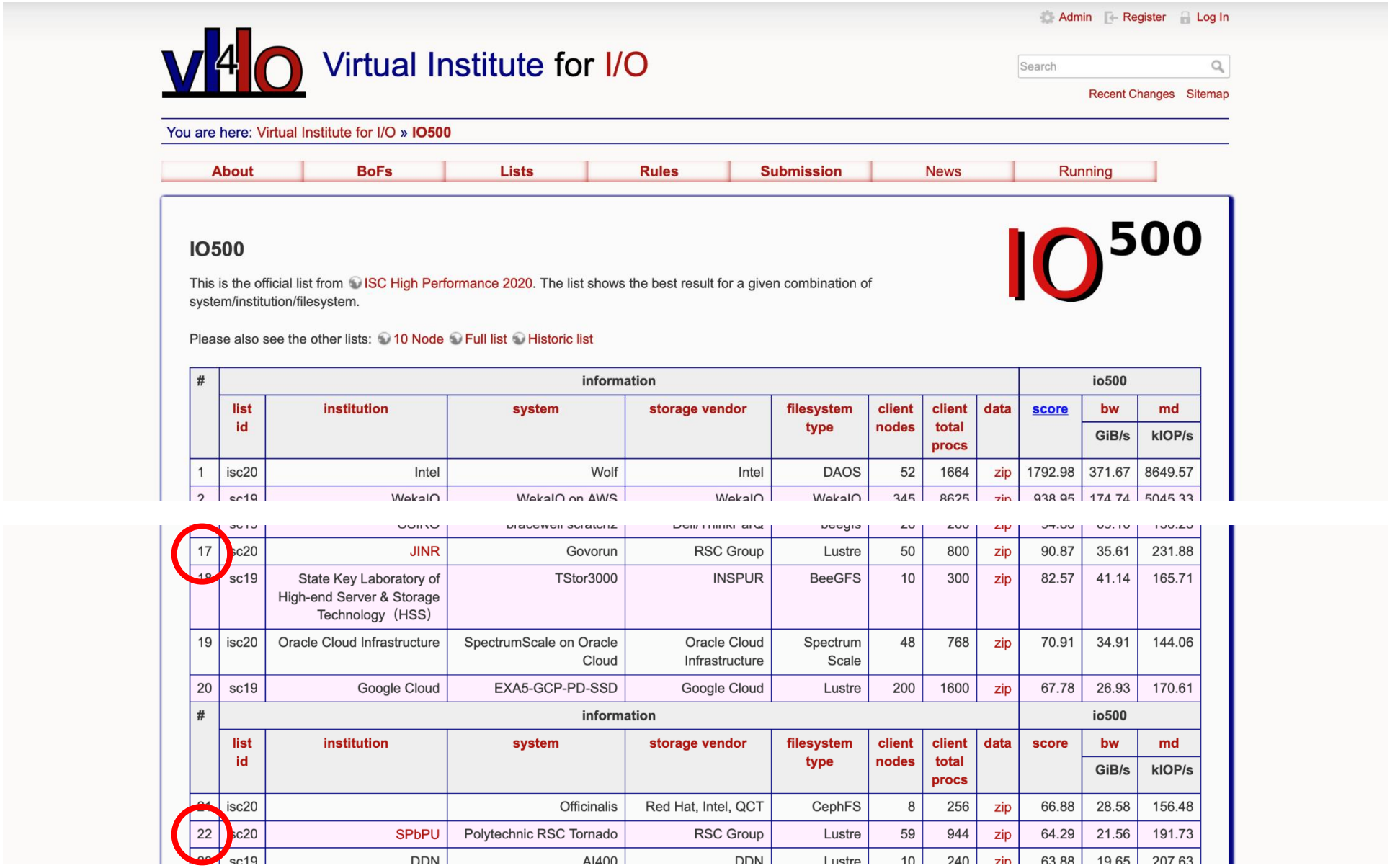
IO500 Total Score	64.1
IO500 Avg BW	21.0 GB/s
IO500 Avg IOPS	195 kIOPS
Write BW в пике	144 GB/s
Read BW в пике	222 GB/s
IOPS	450k
Clients	59 CascadeLake with 100Gbit/s interface

#17



IO500 Total Score	90.9
IO500 Avg BW	35.6 GB/s
IO500 Avg IOPS	231 kIOPS
BW в пике	198 GB/s
IOPS	1.3 M
Clients	50 CascadeLake with 100Gbit/s interface

IO500 Benchmark 2020



Admin Register Log In

Virtual Institute for I/O

Search

Recent Changes Sitemap

You are here: Virtual Institute for I/O » **IO500**

About | BoFs | Lists | Rules | Submission | News | Running

IO500

This is the official list from ISC High Performance 2020. The list shows the best result for a given combination of system/institution/filesystem.

Please also see the other lists: 10 Node Full list Historic list

#	information								io500		
	list id	institution	system	storage vendor	filesystem type	client nodes	client total procs	data	score	bw GiB/s	md kIOP/s
1	isc20	Intel	Wolf	Intel	DAOS	52	1664	zip	1792.98	371.67	8649.57
2	sc19	WekaIO	WekaIO on AWS	WekaIO	WekaIO	345	8625	zip	938.95	174.74	5045.33
17	isc20	JINR	Govorun	RSC Group	Lustre	50	800	zip	90.87	35.61	231.88
18	sc19	State Key Laboratory of High-end Server & Storage Technology（HSS）	TStor3000	INSPUR	BeeGFS	10	300	zip	82.57	41.14	165.71
19	isc20	Oracle Cloud Infrastructure	SpectrumScale on Oracle Cloud	Oracle Cloud Infrastructure	Spectrum Scale	48	768	zip	70.91	34.91	144.06
20	sc19	Google Cloud	EXA5-GCP-PD-SSD	Google Cloud	Lustre	200	1600	zip	67.78	26.93	170.61
#	information								io500		
	list id	institution	system	storage vendor	filesystem type	client nodes	client total procs	data	score	bw GiB/s	md kIOP/s
21	isc20		Officinalis	Red Hat, Intel, QCT	CephFS	8	256	zip	66.88	28.58	156.48
22	sc20	SPbPU	Polytechnic RSC Tornado	RSC Group	Lustre	59	944	zip	64.29	21.56	191.73
23	sc19	DDN	AI400	DDN	Lustre	10	240	zip	63.88	19.65	207.63

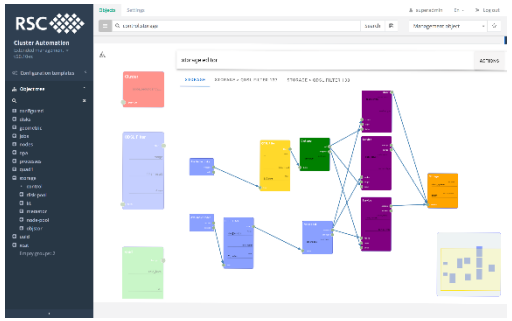
Roadmap

Upcoming features

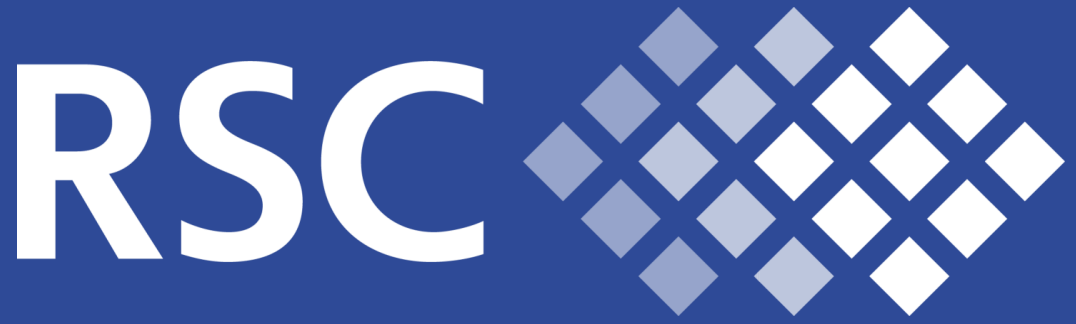
Improving Lustre – substituting slow zfs with faster NVMeoF-aware RAID engine:

+20% in bandwidth in IO500
+100% IOPS in IP500

Intel DAOS support (fastest clustered storage, 3 out of 5 systems in top 5 of IO500 list (2020)



New IO500 submission should be published by the time you are watching this presentation ☺



www.rscgroup.ru
hq@rsc-tech.ru