Application of OMAT in HT-Condor Resource Management

Qingbao Hu(huqb@ihep.ac.cn),

Wei Zheng, Xiaowei Jiang, Jingyan Shi

On behalf of Computing Center, IHEP

ISGC 2021

Outline

- IHEP Computing Platform
- HT-Condor Status @IHEP
- OMAT @IHEP
- HT-Condor resource management
- Summary

IHEP Computing Platform

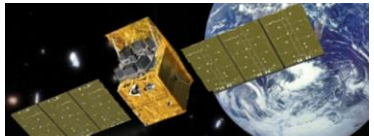
- Multiple experiments supported

BESIII

BESIII (Beijing Spectrometer III at BECPII)

JUNO

JUNO (Jiangmen Underground Neutrino Observatory)

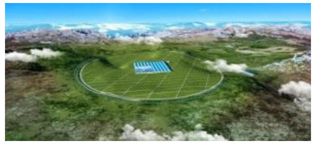


HXMT(Hard X-Ray Moderate Telescope)

CSNS

中国散裂中子源

China Spallation Neutron Source



LHAASO (Large High Altitude Air Shower Observatory)

HEPS (High Energy Photon Source)

CMS

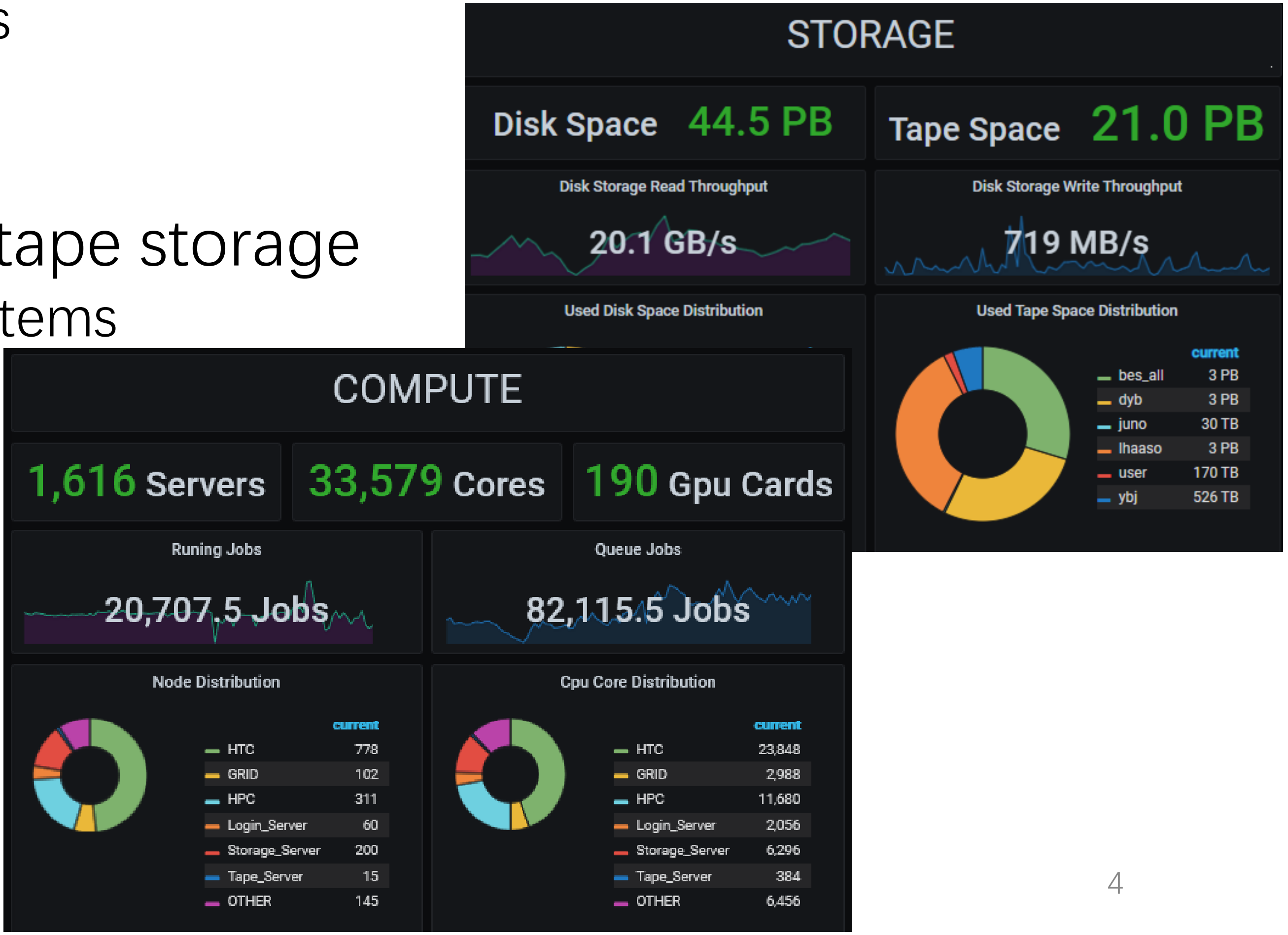
ATLAS EXPERIMENT

LHCb

Belle II

Computing Resources

- 33K CPU cores, 190 GPU cards to support more than 10 experiments
  - HTCondor cluster runs for HTC jobs
  - Slurm cluster runs for HPC jobs
  - WLCG tier2 sites & DIRAC sites
- About 44 PB disk storage, 21 PB tape storage
  - Luster and EOS as two main file systems
  - Castor for tape storage,
  - EOS CTA coming soon
- Network
  - IPv4 / IPv6 dual-stack
  - Ethernet / IB protocols supported
  - LHCOne member



STORAGE

Disk Space 44.5 PB

Tape Space 21.0 PB

Disk Storage Read Throughput: 20.1 GB/s

Disk Storage Write Throughput: 719 MB/s

Used Disk Space Distribution

Used Tape Space Distribution

	current
bes_all	3 PB
dyb	3 PB
juno	30 TB
lhaaso	3 PB
user	170 TB
ybj	526 TB

COMPUTE

1,616 Servers

33,579 Cores

190 Gpu Cards

Runing Jobs: 20,707.5 Jobs

Queue Jobs: 82,115.5 Jobs

Node Distribution

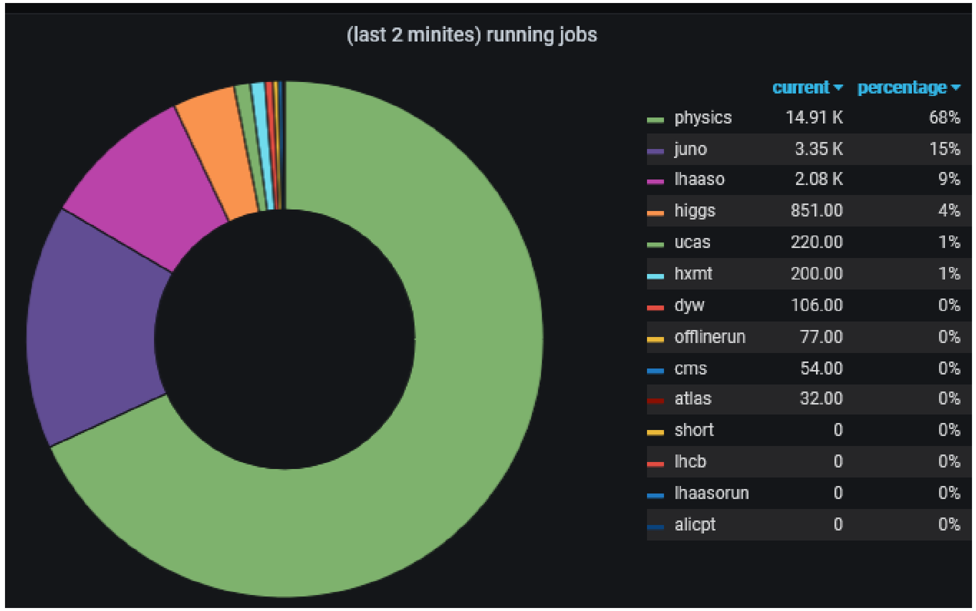
	current
HTC	778
GRID	102
HPC	311
Login_Server	60
Storage_Server	200
Tape_Server	15
OTHER	145

Cpu Core Distribution

	current
HTC	23,848
GRID	2,988
HPC	11,680
Login_Server	2,056
Storage_Server	6,296
Tape_Server	384
OTHER	6,456

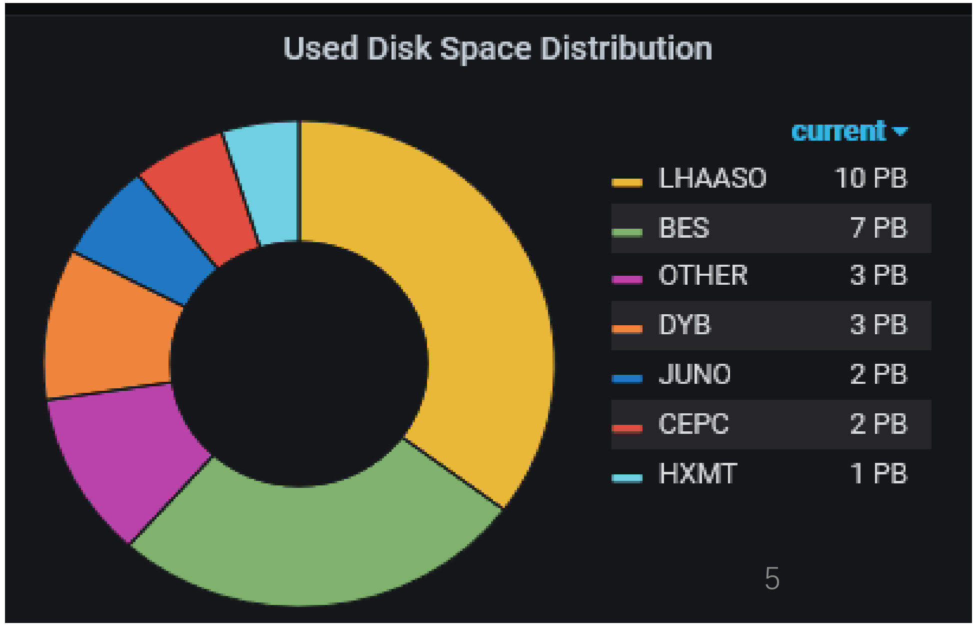
HT-Condor Status @IHEP

- IHEP HTCondor cluster
  - 22,000+ job slots
  - There are more than 100,000,000 jobs finished each year.
  - Support multiple HEP experiments including BES, JUNO, LHAASO, HXMT etc.
  - Various file systems such as Lustre, EOS, NFS, AFS are the dedicated storage for the experiments separately.
  - Work node supports a variety of experimental HTC job running environments, in order to improve resource utilization. The fairness among all the jobs is guaranteed by the experiment quota configured at the scheduler.



(last 2 minites) running jobs

	current ▾	percentage ▾
physics	14.91 K	68%
juno	3.35 K	15%
lhaaso	2.08 K	9%
higgs	851.00	4%
ucas	220.00	1%
hxmt	200.00	1%
dyw	106.00	0%
offlinerun	77.00	0%
cms	54.00	0%
atlas	32.00	0%
short	0	0%
lhcb	0	0%
lhaasorun	0	0%
alicpt	0	0%

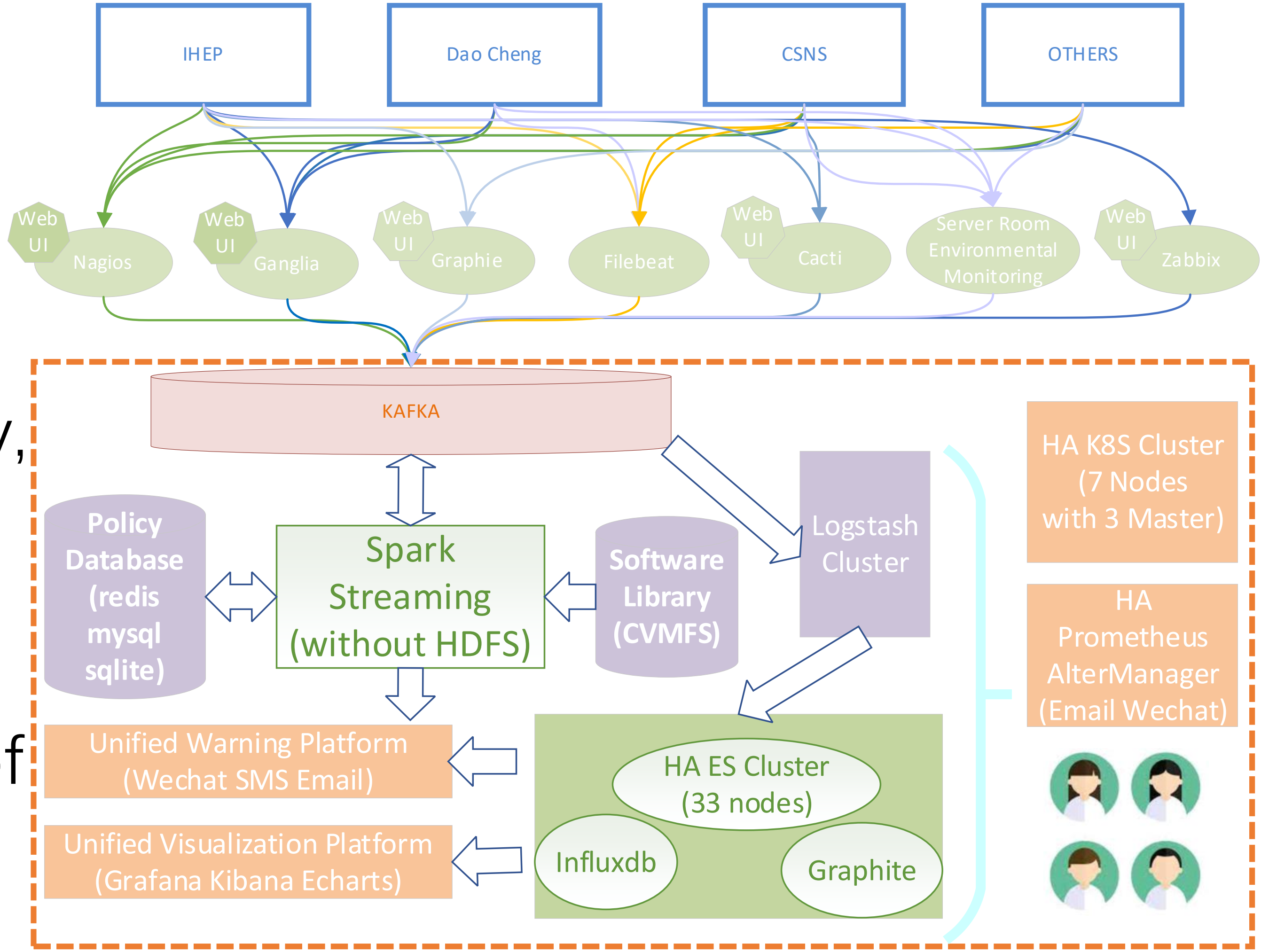


Used Disk Space Distribution

	current ▾
LHAASO	10 PB
BES	7 PB
OTHER	3 PB
DYB	3 PB
JUNO	2 PB
CEPC	2 PB
HXMT	1 PB

Open Maintain Analyze Tools @IHEP

- OMAT is an integrated framework based on a variety of open-source tools, supporting data aggregation, real-time analysis, index query, alarm and visualization.
- It is currently applied to support the daily monitoring operation, and maintenance of IHEP computing cluster.



IHEP

Dao Cheng

CSNS

OTHERS

Web UI

Nagios

Web UI

Ganglia

Web UI

Graphie

Filebeat

Web UI

Cacti

Server Room Environmental Monitoring

Web UI

Zabbix

KAFKA

Policy Database (redis mysql sqlite)

Spark Streaming (without HDFS)

Software Library (CVMFS)

Logstash Cluster

Unified Warning Platform (Wechat SMS Email)

Unified Visualization Platform (Grafana Kibana Echarts)

HA ES Cluster (33 nodes)

Influxdb

Graphite

HA K8S Cluster (7 Nodes with 3 Master)

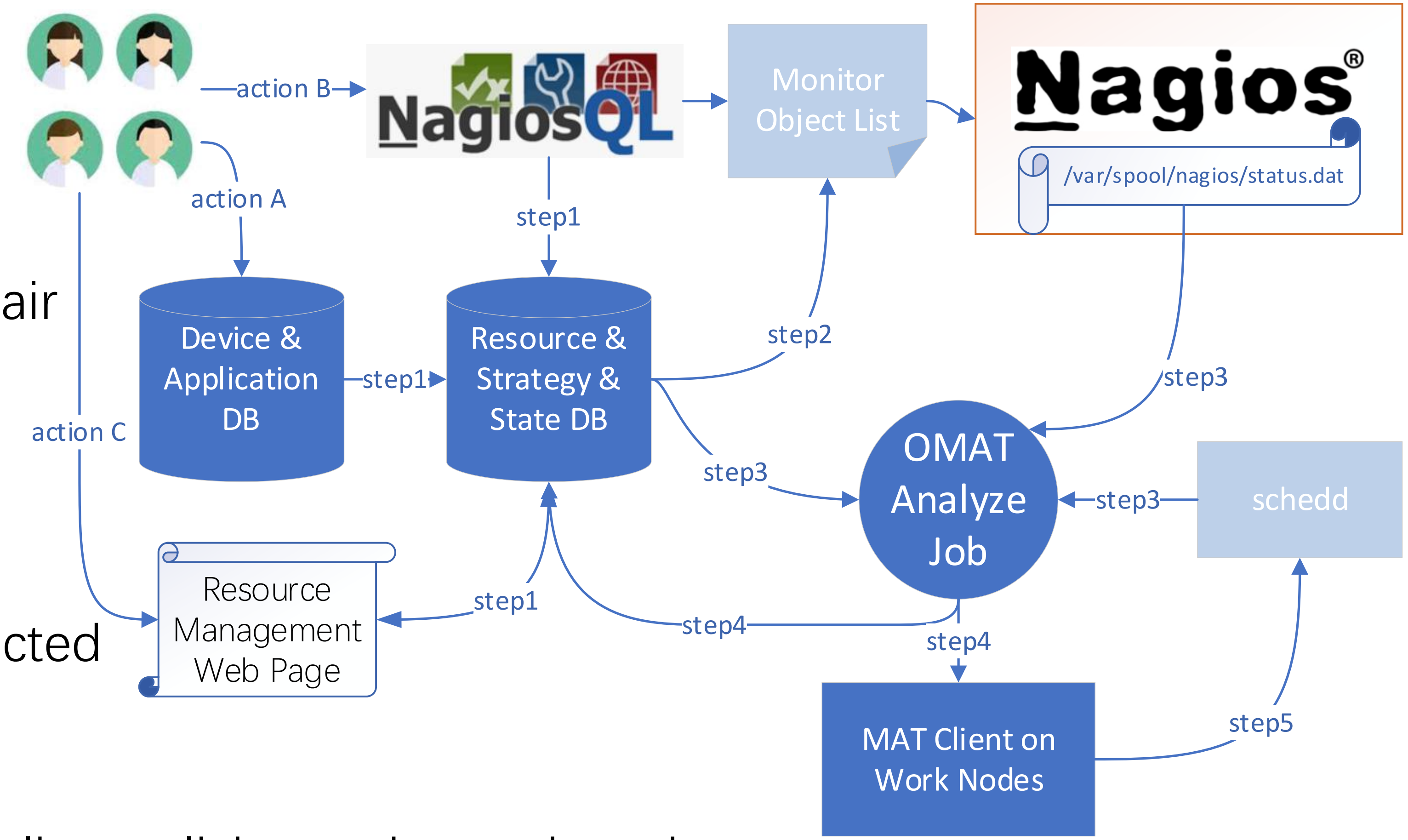
HA Prometheus AlterManager (Email Wechat)

Resource Management Requirements

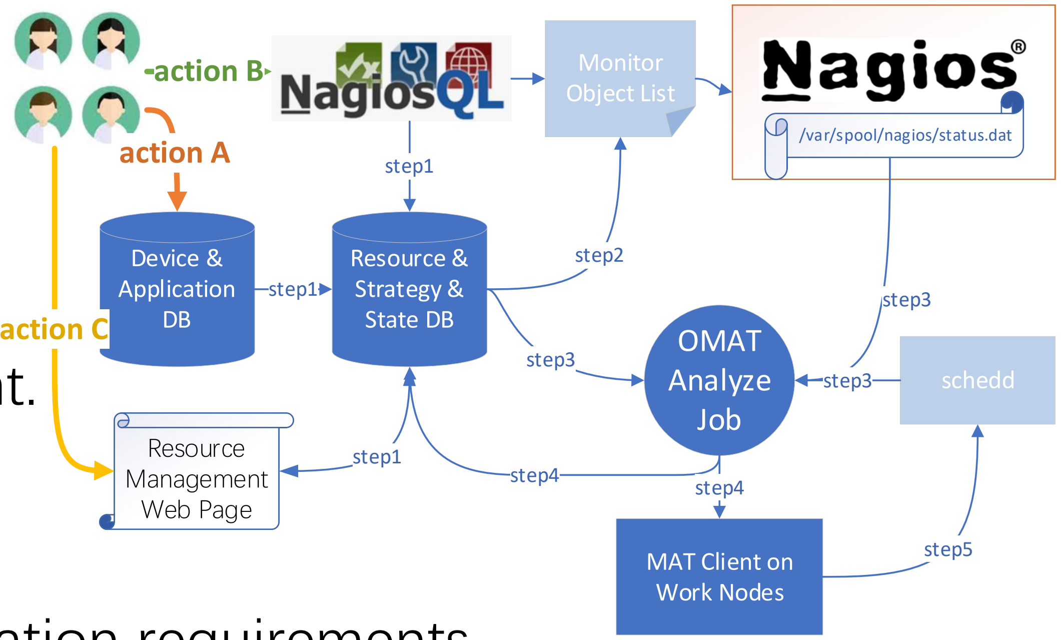
- As the cluster has been expanding, there are more and more special scheduling requirements from the experiments, which need to adjust work node configuration frequently.
- Requirements for resource management
  - Computing resources added and decreased automatically through convenient policy configuration.
  - Application and monitoring services need to be associated flexibly.
  - Sensitive detection node abnormal situation and real-time adjustment of scheduling strategy, to prevent the occurrence of job “black hole”.
  - Provide accessible resource adjustment records to help administrators gain insight into resource usage.

Architecture

- Device DB
  - device status (In use, in repair or scrapped)
- Resource DB
  - The set of experiments supported by each node
  - The set of experiments affected by each monitoring service
- MAT Client
  - Maps all experiment scheduling policies to the work node configuration and re-configs "startd" after receiving the new policy.

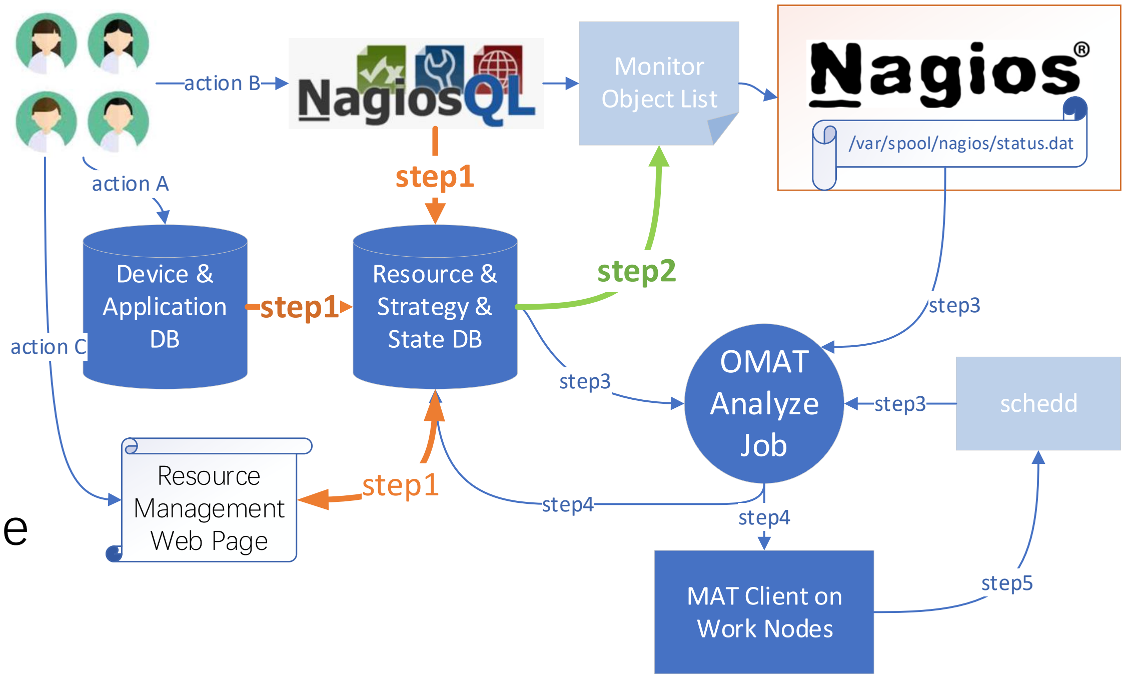


Key Action



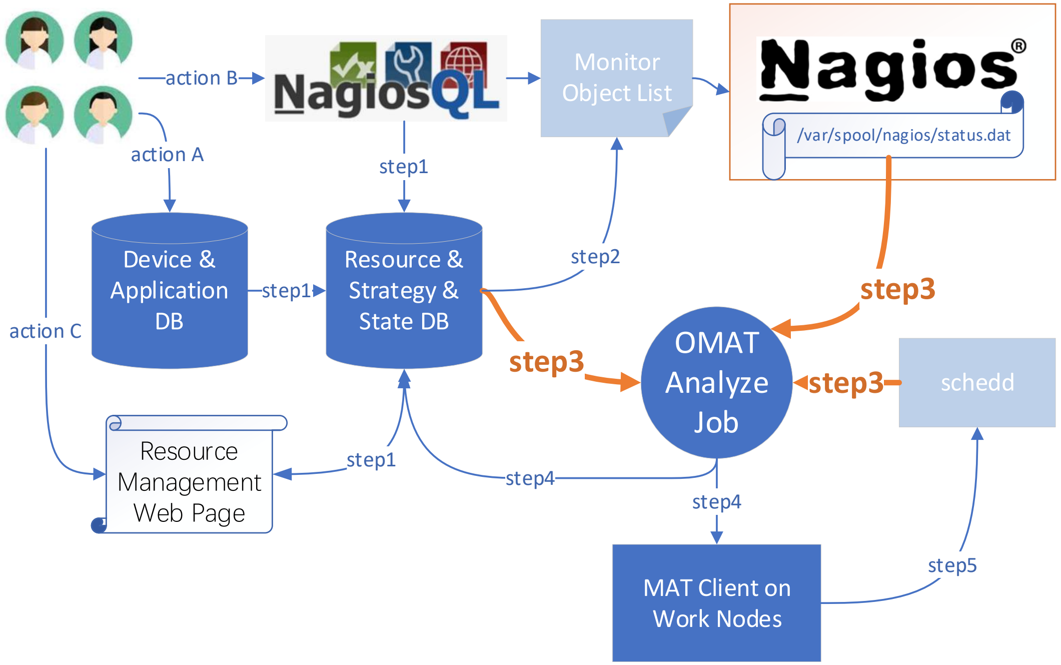
- Action A:
  - The administrator deploys the work node operating system and updates work node’s stat.
  - Add new experiment application.
- Action B:
  - Add monitoring services according to application requirements.
- Action C:
  - Configure the relationship between the experiments and the monitoring servers.
  - Enable or disable the specific monitoring service or experiment.
  - Adjust, remove, or restore the shared scope of work node for specific requirements.

Key Process (1)



- Step1:
  - Synchronize the computing resources, delete the obsolete resources, increase the available resources and provide computing resources for all open experiments.
  - Synchronize the monitoring service, increase the new monitoring service and disable the correlation analysis policy.
  - Synchronize the experiment application, increase the new experiment and disable sharing policy.
  - Wait for administrators to configure the experiment and monitoring service mapping and enable sharing policies. (Action C)
- Step2:
  - Synchronize the work node list and monitoring service to update monitor objects and re-config Nagios.

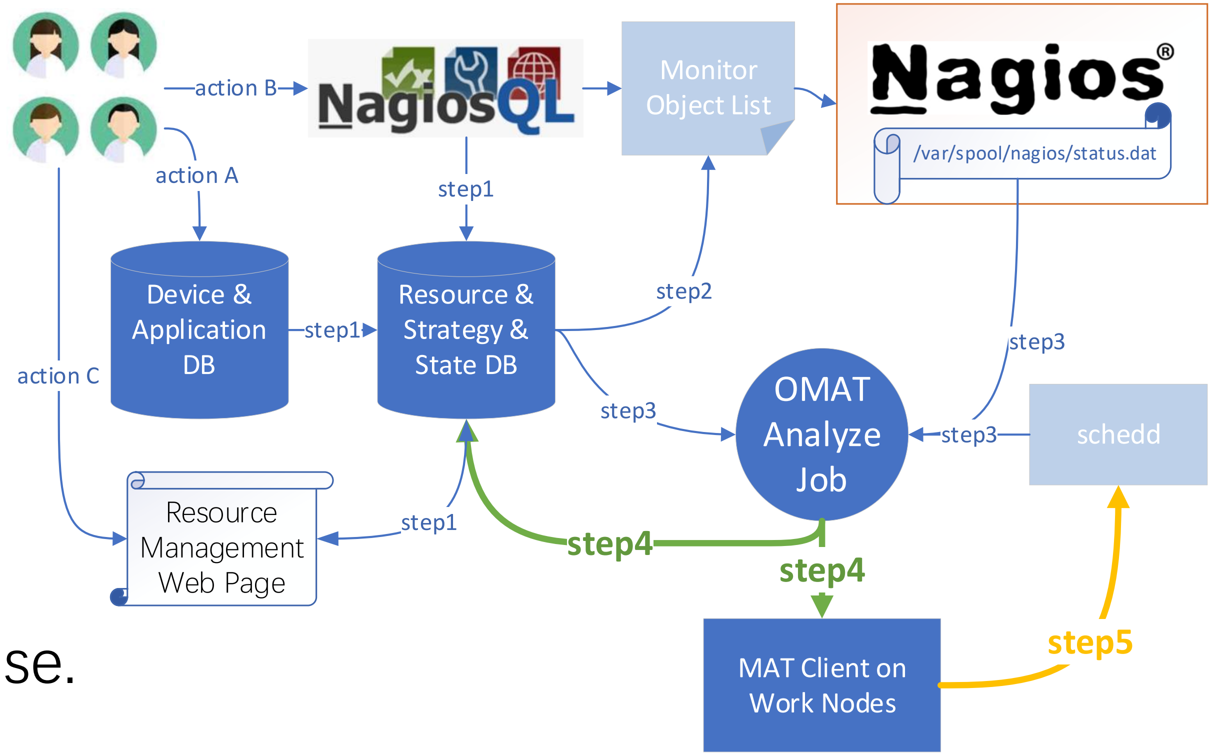
Key Process(2)



- Step3:
  - Compare the pre-allocation experiment list of a node with the experiment which affected by the node's current abnormal service to analyze the list of experiments that nodes should assign.
  - To avoid the job “black hole” status, instantly change the application sharing policy for nodes affected by the abnormal service.
  - To reduce the impact of monitoring service jitter, the sharing policy of the corresponding node is changed after the abnormal service is restored for some time.
  - Compare the active nodes retrieved from the scheduler with all useful HTC nodes from the resource database, to detect the missing work node.
  - Compare the active nodes which are monitored by the Nagios with all useful HTC nodes from the resource database and disable sharing policy of the work node which are not monitored.

Key Process(3)

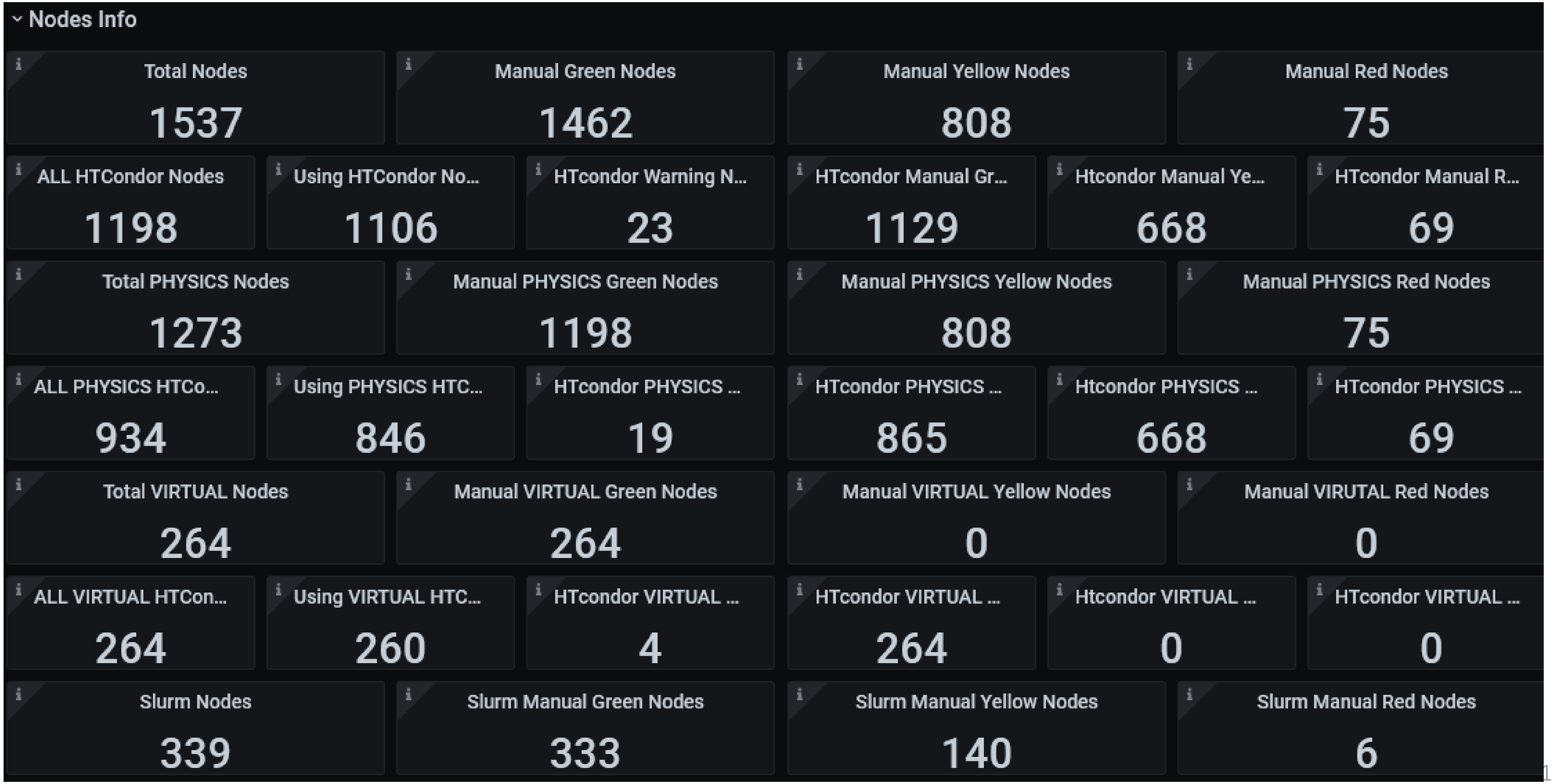
- Step4:
  - Check_nrpe command is used to actively push the new policy file to the target node.
  - Record synchronization status and change history into the state database.
- Step5:
  - The work node receives and analyze the new share policy to create scheduler configuration and re-configs “startd” service.



Architectural Features

- Real-time capture of the abnormal service information of nodes by the /var/spool/nagios/status.dat of Nagios server.
- Quickly analyze data from multiple data sources and obtain analysis results, based on OMAT’s data streaming processing technology.
- The policy files are concurrent actively pushed and deployed based on the “check_nrpe” command to improve the update efficiency.
- The average delay between abnormal service detected by Nagios and correction policy created is less than 10 seconds.
- The delay between a policy taking effect on the node with policy modified by abnormal service detected or adjust manually by the administrator is less than 1 minutes.

Work Node Resource Overview



Nodes Info

Total Nodes
1537

Manual Green Nodes
1462

Manual Yellow Nodes
808

Manual Red Nodes
75

ALL HTCondor Nodes
1198

Using HTCondor No...
1106

HTcondor Warning N...
23

HTcondor Manual Gr...
1129

Htcondor Manual Ye...
668

HTcondor Manual R...
69

Total PHYSICS Nodes
1273

Manual PHYSICS Green Nodes
1198

Manual PHYSICS Yellow Nodes
808

Manual PHYSICS Red Nodes
75

ALL PHYSICS HTCo...
934

Using PHYSICS HTC...
846

HTcondor PHYSICS ...
19

HTcondor PHYSICS ...
865

Htcondor PHYSICS ...
668

HTcondor PHYSICS ...
69

Total VIRTUAL Nodes
264

Manual VIRTUAL Green Nodes
264

Manual VIRTUAL Yellow Nodes
0

Manual VIRUTAL Red Nodes
0

ALL VIRTUAL HTCon...
264

Using VIRTUAL HTC...
260

HTcondor VIRTUAL ...
4

HTcondor VIRTUAL ...
264

Htcondor VIRTUAL ...
0

HTcondor VIRTUAL ...
0

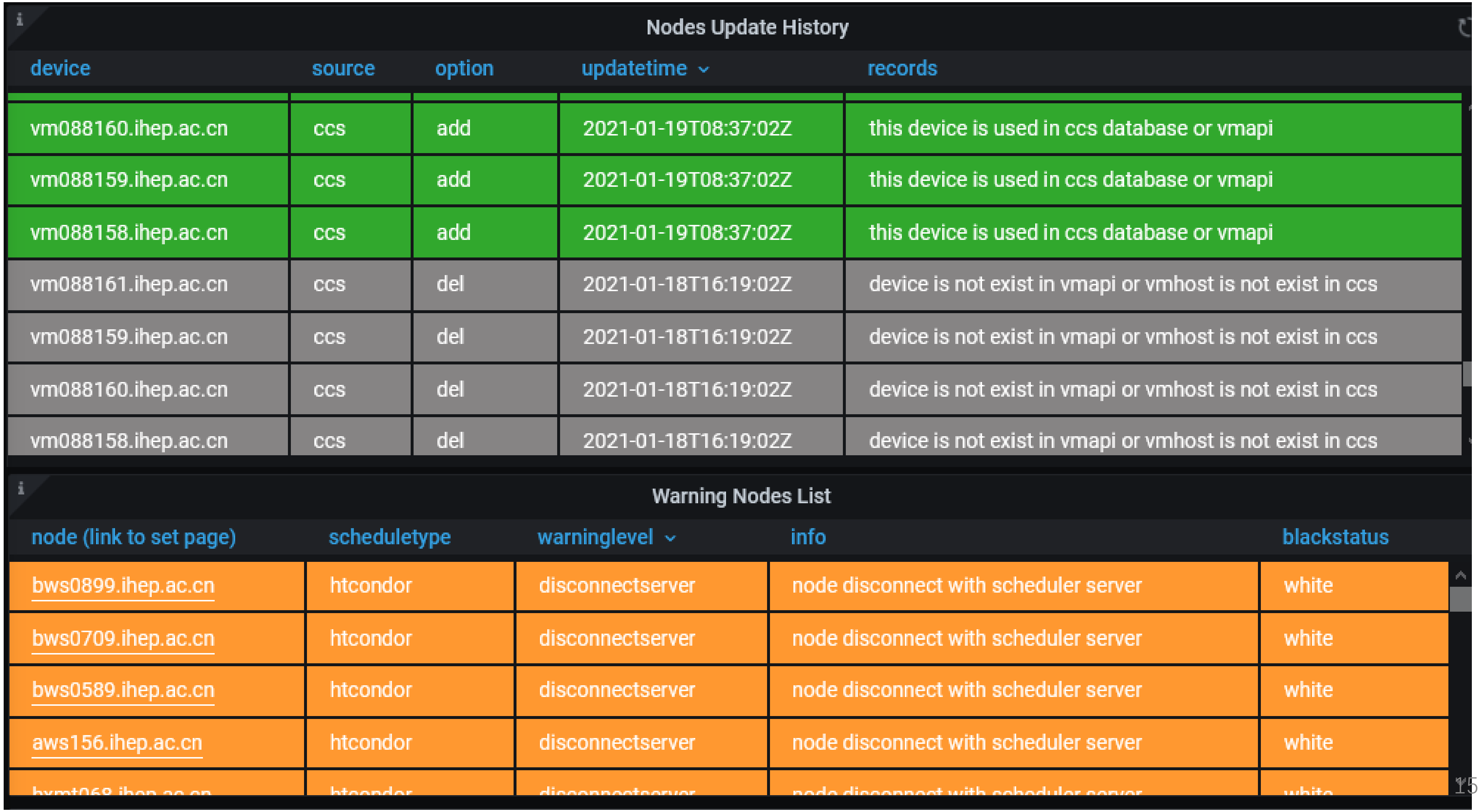
Slurm Nodes
339

Slurm Manual Green Nodes
333

Slurm Manual Yellow Nodes
140

Slurm Manual Red Nodes
6

Work Node Resource Monitoring



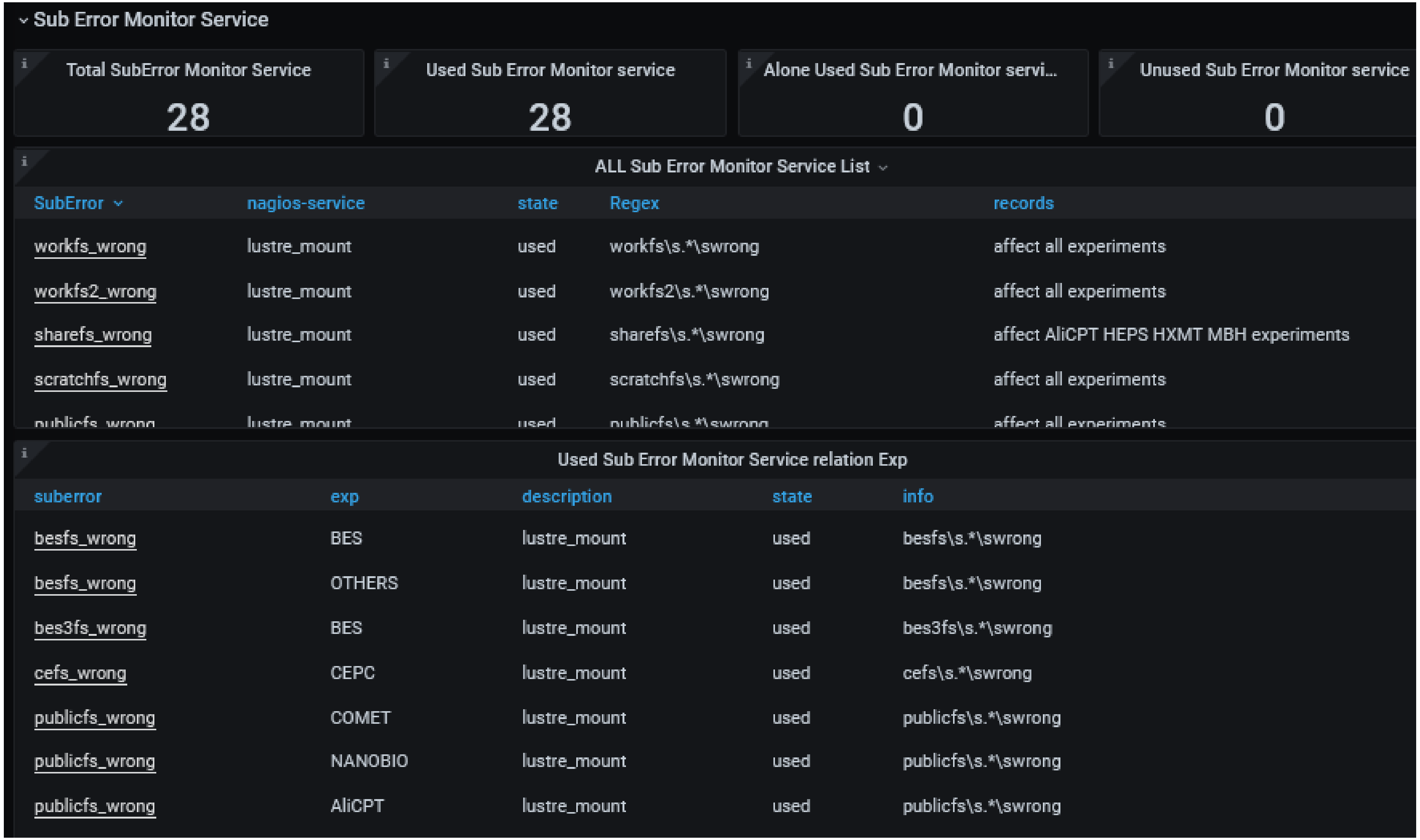
Nodes Update History

device	source	option	updatetime ⌄	records
vm088160.ihep.ac.cn	ccs	add	2021-01-19T08:37:02Z	this device is used in ccs database or vmapi
vm088159.ihep.ac.cn	ccs	add	2021-01-19T08:37:02Z	this device is used in ccs database or vmapi
vm088158.ihep.ac.cn	ccs	add	2021-01-19T08:37:02Z	this device is used in ccs database or vmapi
vm088161.ihep.ac.cn	ccs	del	2021-01-18T16:19:02Z	device is not exist in vmapi or vmhost is not exist in ccs
vm088159.ihep.ac.cn	ccs	del	2021-01-18T16:19:02Z	device is not exist in vmapi or vmhost is not exist in ccs
vm088160.ihep.ac.cn	ccs	del	2021-01-18T16:19:02Z	device is not exist in vmapi or vmhost is not exist in ccs
vm088158.ihep.ac.cn	ccs	del	2021-01-18T16:19:02Z	device is not exist in vmapi or vmhost is not exist in ccs

Warning Nodes List

node (link to set page)	scheduletype	warninglevel ⌄	info	blackstatus
bws0899.ihep.ac.cn	htcondor	disconnectserver	node disconnect with scheduler server	white
bws0709.ihep.ac.cn	htcondor	disconnectserver	node disconnect with scheduler server	white
bws0589.ihep.ac.cn	htcondor	disconnectserver	node disconnect with scheduler server	white
aws156.ihep.ac.cn	htcondor	disconnectserver	node disconnect with scheduler server	white
bxmt068.ihep.ac.cn	htcondor	disconnectserver	node disconnect with scheduler server	white

Service impact experiments



Sub Error Monitor Service

Total SubError Monitor Service	Used Sub Error Monitor service	Alone Used Sub Error Monitor servi...	Unused Sub Error Monitor service
28	28	0	0

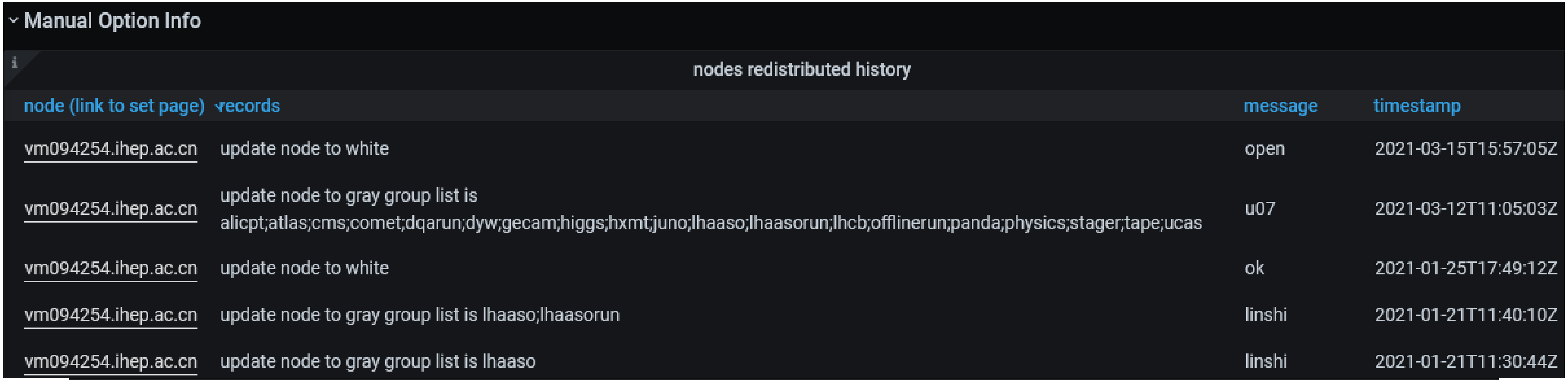
ALL Sub Error Monitor Service List

SubError	nagios-service	state	Regex	records
workfs_wrong	lustre_mount	used	workfs\s.*\swrong	affect all experiments
workfs2_wrong	lustre_mount	used	workfs2\s.*\swrong	affect all experiments
sharefs_wrong	lustre_mount	used	sharefs\s.*\swrong	affect AliCPT HEPS HXMT MBH experiments
scratchfs_wrong	lustre_mount	used	scratchfs\s.*\swrong	affect all experiments
publicfs_wrong	lustre_mount	used	publicfs\s.*\swrong	affect all experiments

Used Sub Error Monitor Service relation Exp

suberror	exp	description	state	info
besfs_wrong	BES	lustre_mount	used	besfs\s.*\swrong
besfs_wrong	OTHERS	lustre_mount	used	besfs\s.*\swrong
bes3fs_wrong	BES	lustre_mount	used	bes3fs\s.*\swrong
cefs_wrong	CEPC	lustre_mount	used	cefs\s.*\swrong
publicfs_wrong	COMET	lustre_mount	used	publicfs\s.*\swrong
publicfs_wrong	NANOBIO	lustre_mount	used	publicfs\s.*\swrong
publicfs_wrong	AliCPT	lustre_mount	used	publicfs\s.*\swrong

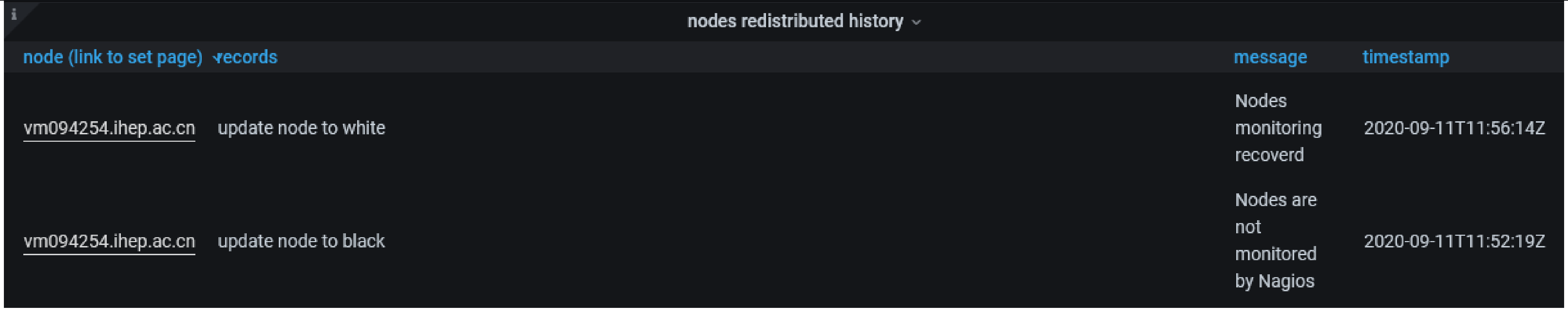
Pre-allocation node's experimental group



Manual Option Info

nodes redistributed history

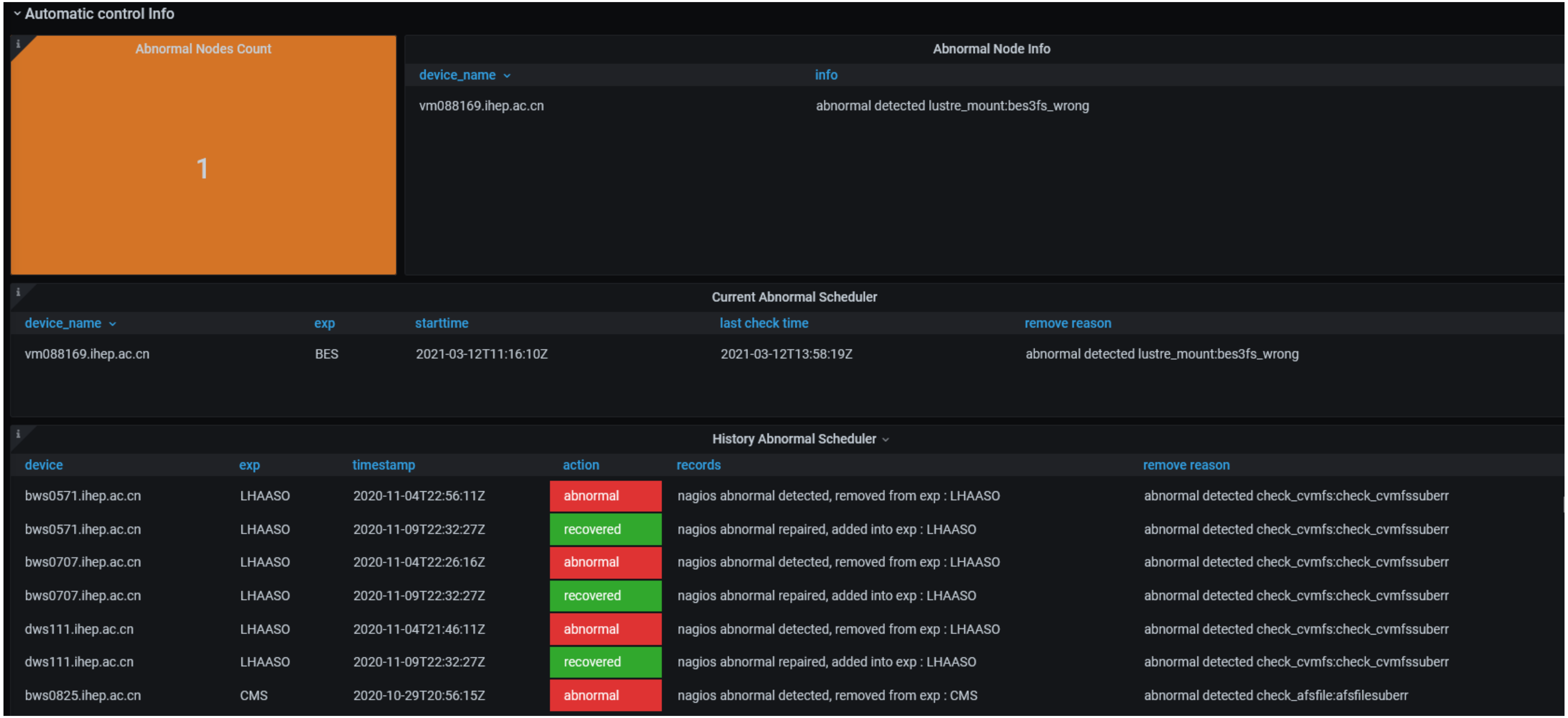
node (link to set page)	records	message	timestamp
vm094254.ihep.ac.cn	update node to white	open	2021-03-15T15:57:05Z
vm094254.ihep.ac.cn	update node to gray group list is alicpt;atlas;cms;comet;dqarun;dyw;gecam;higgs;hxmt;juno;lhaaso;lhaasorun;lhcb;offlinerun;panda;physics;stager;tape;ucas	u07	2021-03-12T11:05:03Z
vm094254.ihep.ac.cn	update node to white	ok	2021-01-25T17:49:12Z
vm094254.ihep.ac.cn	update node to gray group list is lhaaso;lhaasorun	linshi	2021-01-21T11:40:10Z
vm094254.ihep.ac.cn	update node to gray group list is lhaaso	linshi	2021-01-21T11:30:44Z



nodes redistributed history

node (link to set page)	records	message	timestamp
vm094254.ihep.ac.cn	update node to white	Nodes monitoring recoverd	2020-09-11T11:56:14Z
vm094254.ihep.ac.cn	update node to black	Nodes are not monitored by Nagios	2020-09-11T11:52:19Z

Adjust node's experimental group



Automatic control Info

Abnormal Nodes Count

1

Abnormal Node Info

device_name	info
vm088169.ihep.ac.cn	abnormal detected lustre_mount:bes3fs_wrong

Current Abnormal Scheduler

device_name	exp	starttime	last check time	remove reason
vm088169.ihep.ac.cn	BES	2021-03-12T11:16:10Z	2021-03-12T13:58:19Z	abnormal detected lustre_mount:bes3fs_wrong

History Abnormal Scheduler

device	exp	timestamp	action	records	remove reason
bws0571.ihep.ac.cn	LHAASO	2020-11-04T22:56:11Z	abnormal	nagios abnormal detected, removed from exp : LHAASO	abnormal detected check_cvmfs:check_cvmfssuberr
bws0571.ihep.ac.cn	LHAASO	2020-11-09T22:32:27Z	recovered	nagios abnormal repaired, added into exp : LHAASO	abnormal detected check_cvmfs:check_cvmfssuberr
bws0707.ihep.ac.cn	LHAASO	2020-11-04T22:26:16Z	abnormal	nagios abnormal detected, removed from exp : LHAASO	abnormal detected check_cvmfs:check_cvmfssuberr
bws0707.ihep.ac.cn	LHAASO	2020-11-09T22:32:27Z	recovered	nagios abnormal repaired, added into exp : LHAASO	abnormal detected check_cvmfs:check_cvmfssuberr
dws111.ihep.ac.cn	LHAASO	2020-11-04T21:46:11Z	abnormal	nagios abnormal detected, removed from exp : LHAASO	abnormal detected check_cvmfs:check_cvmfssuberr
dws111.ihep.ac.cn	LHAASO	2020-11-09T22:32:27Z	recovered	nagios abnormal repaired, added into exp : LHAASO	abnormal detected check_cvmfs:check_cvmfssuberr
bws0825.ihep.ac.cn	CMS	2020-10-29T20:56:15Z	abnormal	nagios abnormal detected, removed from exp : CMS	abnormal detected check_afsfile:afsfilesuberr

Summary

- It greatly shortens the time between the occurrence of node abnormal service and the effectiveness of the corrected sharing policy and reduces the impact of abnormal service on the experimental application. Enhance the efficiency of work node sharing.
- Provide great convenience to the administrator to adjust the allocation of resources.
- Provides a resource monitoring panel to help administrators more intuitive insight into the use of resources.

THANKS