

|| The 8th Asia Tier Center Forum ||

HPC Project for ALICE in Korea

Hyeonjin Yu

Integrated M.S. and Ph.D student,
Chungbuk National University, South Korea

hyeonjin.yu@cern.ch


2024.09.03(Tue)

HPC Project Introduction


❖ The main activities and participating institutes

- Constructing a new ALICE grid site with HPC resources.
 - HPC resources: Nurion, 5th supercomputer in South Korea (managed by KISTI)

<https://www.ksc.re.kr/eng/resources/nurion>




Overview




Main system

- 8,305 compute nodes
- 132 CPU-only nodes (25.7PFlops total)




Storage

- 12 racks(23.88PB total)
- 3,200 8TB NL-SAS
- 768 1.2TB SSDs



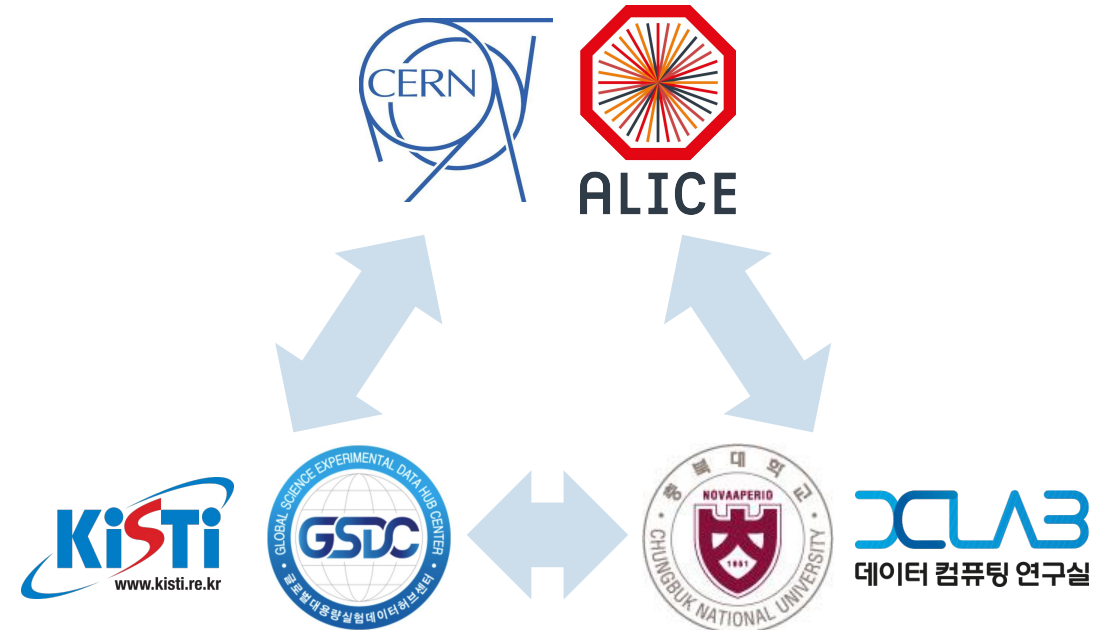
Tape Library

- 4 racks(10PB total)
- 1,700 LTO7 tapes

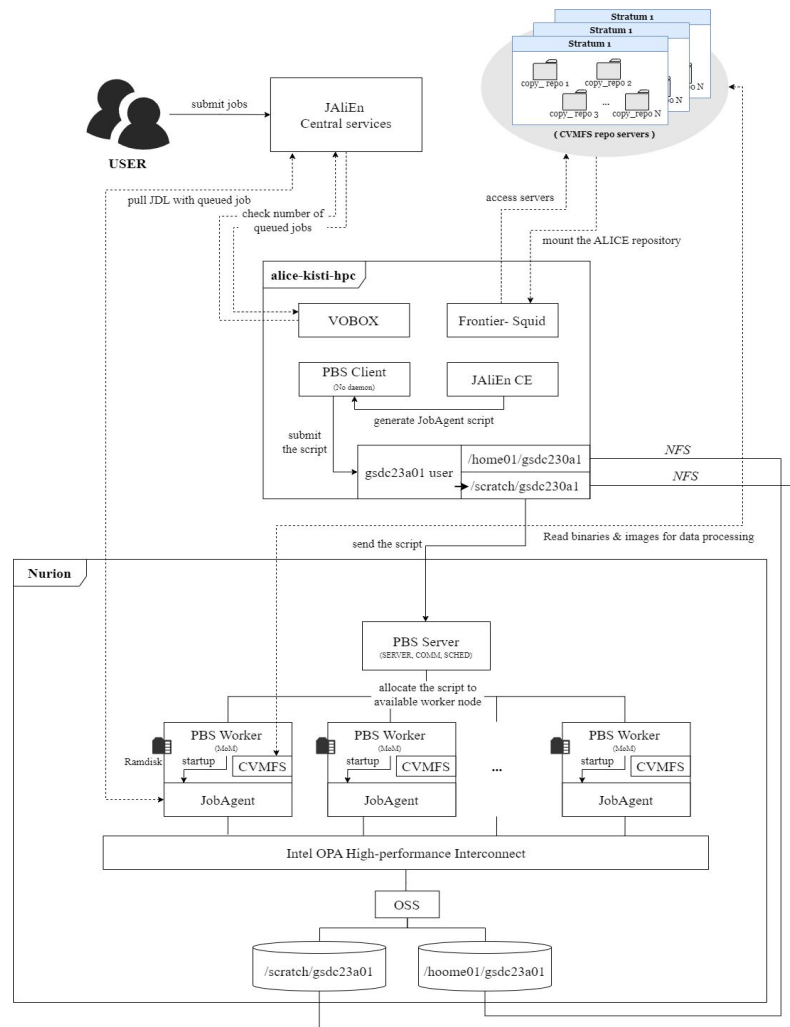


Infrastructure

- Chiller 400RT 4EA
- Free colling 250RT 2EA
- Cooling tower 450RT 4EA
- HVAC 30RT 11EA

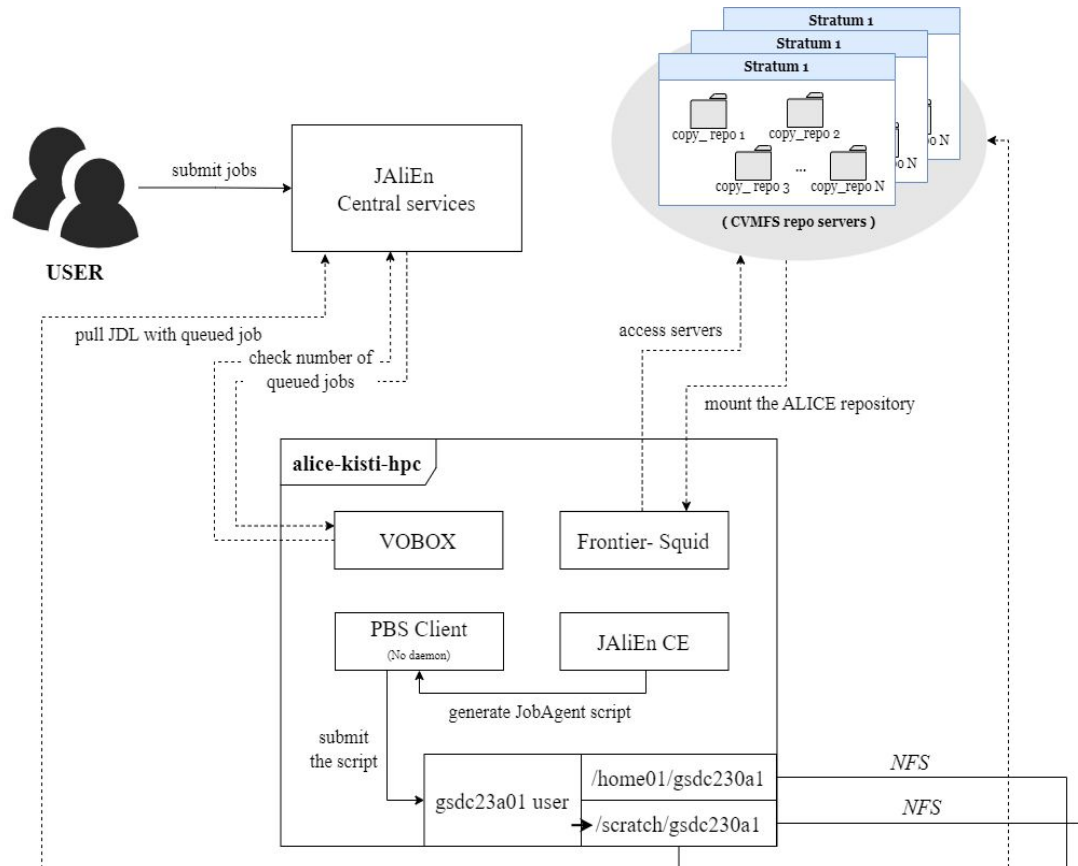


❖ HPC Grid Site: 'KISTI_GSDC_Nurion'



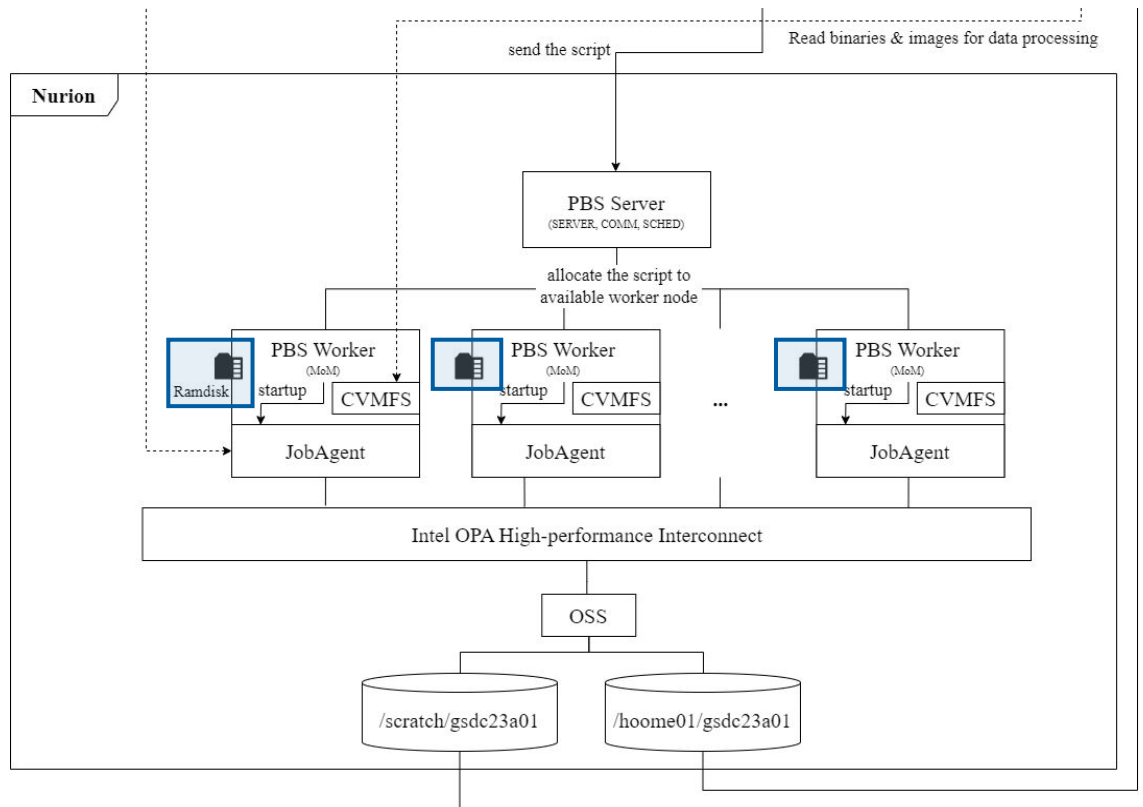
1. Acquire submitted ALICE grid jobs from ALICE users via VOBOX
2. Create JobAgent startup scripts to search for available computational nodes on alice-kisit-hpc node
3. Transit to the designated job submission user, gsdc23a01
4. Move to the designated submission workspace that gsdc23a01 owns, /scratch/gsdc23a01
5. Submit jobs executing the JobAgent script to the local batch queue by gsdc23a01 on /scratch/gsdc23a01
6. Allocate the jobs on one of HPC worker nodes by the PBS server
7. Process ALICE grid jobs for the JobAgent's lifetime

❖ 1) VObox node: 'alice-kisti-hpc'



- **installed packages**
 - vbox
 - cvmfs
 - frontier-squid
 - pbs-pro
- **roles**
 - authentication (via vbox)
 - proxy server (via frontier-squid)
 - pbs client (via pbs-pro)
- **for job submission, it needs:**
 - **/scratch/gsd23a01** (as a job submission path)
 - **gsdc23a01** (as a job submission user)

❖ 2) PBS cluster: 'Nurion' as a supercomputer



NODE: PBS server

- **installed package**
 - pbs-pro
- **roles**
 - allocating jobs to worker nodes

NODE: PBS worker

- **installed packages**
 - cvmfs
 - pbs-pro
- **roles**
 - processing jobs (via cvmfs, pbs-pro)
- **why ramdisk used?**
 - instead of disks

NURION

<https://www.ksc.re.kr/eng/resources/nurion>

5th Supercomputer Summary

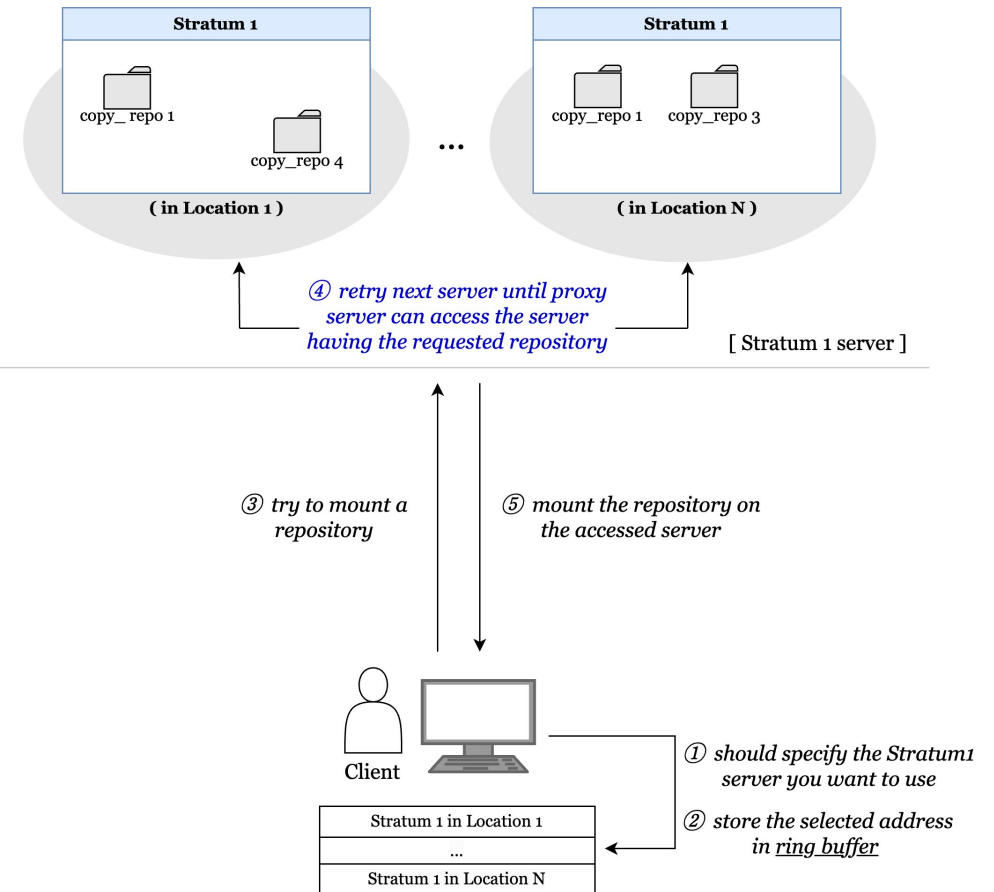
Nurion is a system consisting of compute nodes, CPU-only nodes, Omni-Path interconnect networks, Burst Buffer high-speed storage, Luster-based parallel file system, and water cooling device based on Rear Door Heat Exchanger (RDHx). Nurion's compute nodes are 8,305 Intel Xeon Phi processors (named "Knight Landing") nodes and CPU-only nodes are 132 Intel Xeon processors (named "Skylake") nodes. Total theoretical performance is 25.7 petaflops, which was ranked 11th in the world in June 2018 (<http://www.top500.org>).

	Compute Node	CPU-only Node
Model	Cray CS500	
Architecture	Xeon Phi cluster	CPU cluster
Processor	Intel Xeon Phi 7250 1.4GHz	Intel Xeon 6148 2.4GHz
Total nodes	8,305	132
CPU each node	1	2
Cores per CPU	68	20
Cores each node	564,740	5,280
Memory per core	1.4GB	2.4GB
Memory each node	96GB	192GB
Total memory	797.3TB	25.3TB
Compute power(peak)	25.3PF	0.4PF
Storage capacity		21PB
Interconnect		OPA

❖ The key softwares to be used in the site

- What are their respective roles?

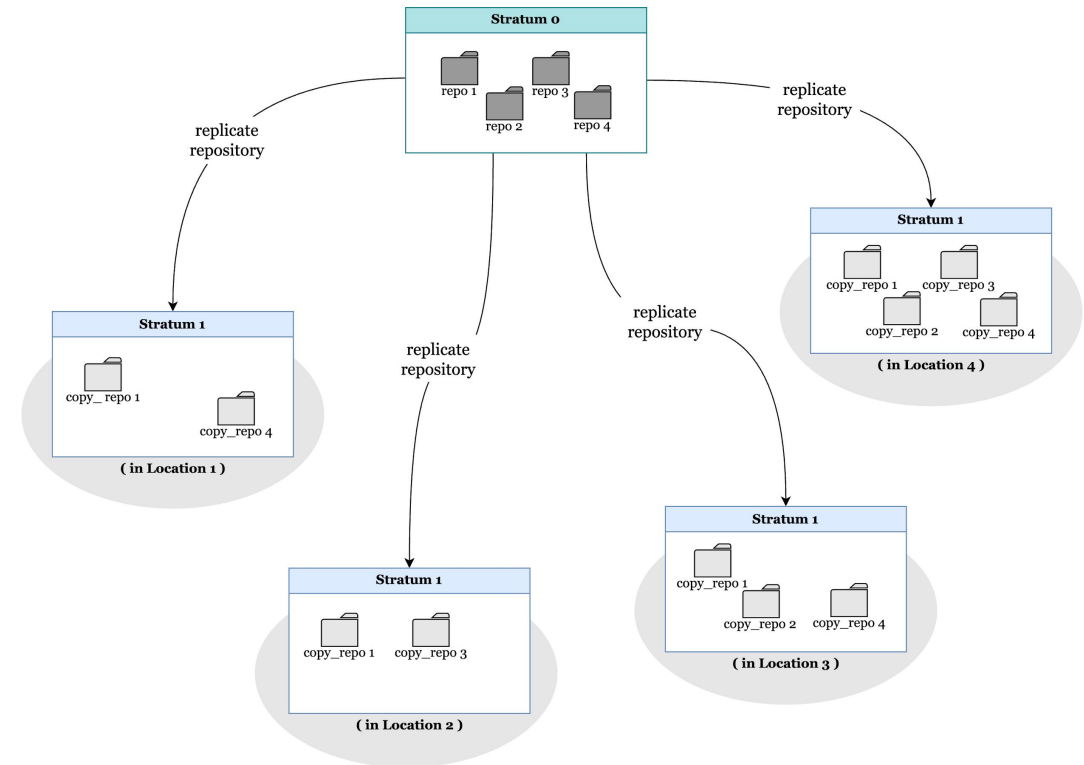
Frontier-squid	<ul style="list-style-type: none"> • A proxy server utilized to mount the ALICE cvmfs repositories necessary for ALICE Grid jobs.
CVMFS	<ul style="list-style-type: none"> • A file system that stores repositories including packages, experimental data and so on.
NFS	<ul style="list-style-type: none"> • A networking protocol used for sharing /home/gsd23a01 and /scratch/gsd23a01 directories.
PBS	<ul style="list-style-type: none"> • A distributed workload management system for managing and monitoring your computational workload.
VOBOX	<ul style="list-style-type: none"> • A system which supports ALICE VO services, authorizing users, defining site information.



❖ The key softwares to be used in the site

- What are their respective roles?

Frontier-squid	<ul style="list-style-type: none"> • A proxy server utilized to mount the ALICE cvmfs repositories necessary for ALICE Grid jobs.
CVMFS	<ul style="list-style-type: none"> • A file system that stores repositories including packages, experimental data and so on.
NFS	<ul style="list-style-type: none"> • A networking protocol used for sharing /home/gsd23a01 and /scratch/gsd23a01 directories.
PBS	<ul style="list-style-type: none"> • A distributed workload management system for managing and monitoring your computational workload.
VOBOX	<ul style="list-style-type: none"> • A system which supports ALICE VO services, authorizing users, defining site information.



❖ The key softwares to be used in the site

- What are their respective roles?

Frontier-squid	<ul style="list-style-type: none"> • A proxy server utilized to mount the ALICE cvmfs repositories necessary for ALICE Grid jobs.
CVMFS	<ul style="list-style-type: none"> • A file system that stores repositories including packages, experimental data and so on.
NFS	<ul style="list-style-type: none"> • A networking protocol used for sharing /home/gsd23a01 and /scratch/gsd23a01 directories.
PBS	<ul style="list-style-type: none"> • A distributed workload management system for managing and monitoring your computational workload.
VOBOX	<ul style="list-style-type: none"> • A system which supports ALICE VO services, authorizing users, defining site information.

```
[root@alice-kisti-hpc ~]# mount -t nfs      :/home01/gsd23a01 /home01/gsd23a01
[root@alice-kisti-hpc ~]# mount -t nfs      :/scratch/gsd23a01 /scratch/gsd23a01
```

[on gsd23a01 user]

```
[gsdc23a01@alice-kisti-hpc gsd23a01]$ ll /home01/gsd23a01/certs
total 12
-r--r-----. 1 gsd23a01 in0138 1281 Oct  5 23:04 ca.pem
-r--r-----. 1 gsd23a01 in0138 1704 Oct  5 23:04 ldap.key
-r--r-----. 1 gsd23a01 in0138 1383 Oct  5 23:04 ldap.pem
```



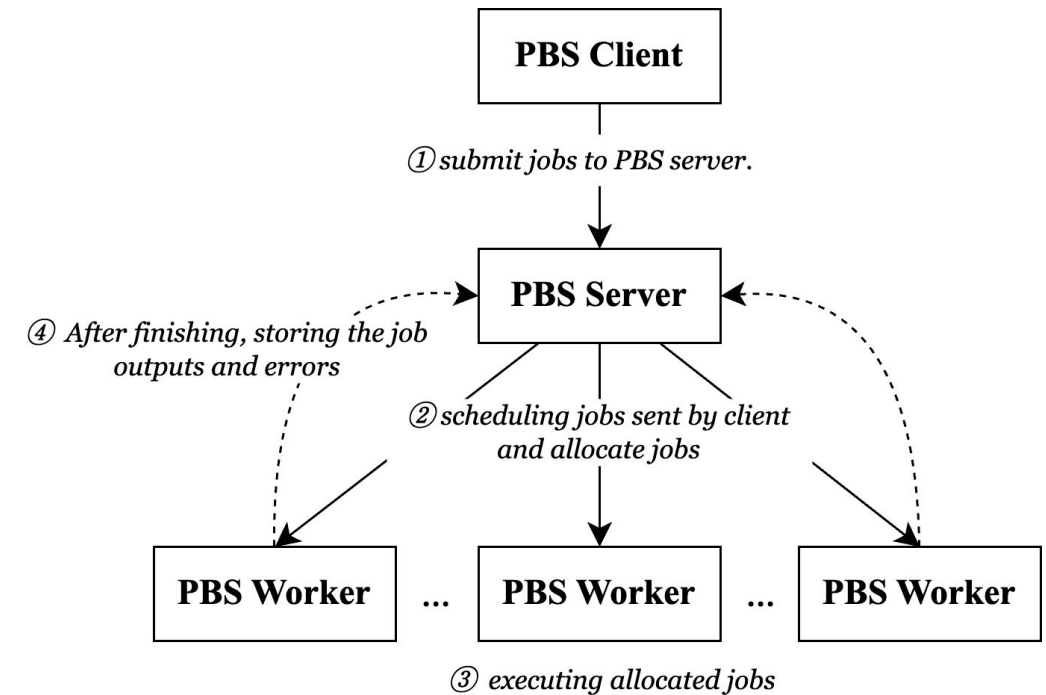
Access succeeded!

```
[gsdc23a01@alice-kisti-hpc ~]$ ll /home01/gsd23a01
total 56900
drwxr-xr-x. 2 gsd23a01 in0138   4096 Oct  5 23:04 certs
drwxr-xr-x. 2 gsd23a01 in0188   4096 May 18 14:06 job_examples
-rw-r--r--. 1 gsd23a01 in0188 58254308 Sep 26 15:52 pbspro-execution-2020.1.
3.20210315160738-0.el7.x86_64.rpm
[gsdc23a01@alice-kisti-hpc ~]$
[gsdc23a01@alice-kisti-hpc ~]$ ll /scratch/gsd23a01/
-rwxr-x---. 1 gsd23a01 in0188   328 Oct 30 11:34 cvmfs.sh
-rw-r-----. 1 gsd23a01 in0188    81 Oct 19 13:09 test.c
-rwxr-x---. 1 gsd23a01 in0188  8360 Oct 23 14:57 test.exe
```

❖ The key softwares to be used in the site

- What are their respective roles?

Frontier-squid	<ul style="list-style-type: none">• A proxy server utilized to mount the ALICE cvmfs repositories necessary for ALICE Grid jobs.
CVMFS	<ul style="list-style-type: none">• A file system that stores repositories including packages, experimental data and so on.
NFS	<ul style="list-style-type: none">• A networking protocol used for sharing /home/gsd23a01 and /scratch/gsd23a01 directories.
PBS	<ul style="list-style-type: none">• A distributed workload management system for managing and monitoring your computational workload.
VOBOX	<ul style="list-style-type: none">• A system which supports ALICE VO services, authorizing users, defining site information.



❖ The key softwares to be used in the site

- What are their respective roles?

Frontier-squid	<ul style="list-style-type: none"> • A proxy server utilized to mount the ALICE cvmfs repositories necessary for ALICE Grid jobs.
CVMFS	<ul style="list-style-type: none"> • A file system that stores repositories including packages, experimental data and so on.
NFS	<ul style="list-style-type: none"> • A networking protocol used for sharing /home/gsd23a01 and /scratch/gsd23a01 directories.
PBS	<ul style="list-style-type: none"> • A distributed workload management system for managing and monitoring your computational workload.
VOBOX	<ul style="list-style-type: none"> • A system which supports ALICE VO services, authorizing users, defining site information.

```
[root@alice-kisti-hpc ~]# cat site-info.def
GROUPS_CONF=/opt/glite/yaim/etc/groups.conf
USERS_CONF=/opt/glite/yaim/etc/users.conf

SITE_NAME=KR-KISTI-GSDC-01

VOBOX_HOST=`hostname -f`
WMS_HOST=rocwms01.grid.sinica.edu.tw
PX_HOST=myproxy.cern.ch
BDII_HOST=lcg-bdii.cern.ch

#SE_LIST=alice-t1-se.sdfarm.kr
SE_LIST=my-se.my-domain

#VOS="alice dteam ops"
VOS="alice"

VO_ALICE_SW_DIR=
VO_ALICE_DEFAULT_SE=my-se.my-domain
VO_ALICE_VOMS_SERVERS="'voms://voms2.cern.ch:8443/voms/alice?/alice/' 'voms://lcg-voms2.cern.ch:8443/voms/alice?/alice/' "
VO_ALICE_VOMSES="'alice lcg-voms2.cern.ch 15000 /DC=ch/DC=cern /OU=computers/CN=lcg-voms2.cern.ch alice 24' 'alice voms2.cern.ch 15000 /DC=ch/DC=cern/OU=computers/CN=voms2.cern.ch alice 24'"
VO_ALICE_VOMS_CA_DN="'/DC=ch/DC=cern/CN=CERN Grid Certification Authority' '/DC=ch/DC=cern/CN=CERN Grid Certification Authority'"

[root@alice-kisti-hpc ~]# tail /opt/glite/yaim/etc/users.conf
14320:ali1_120:14200:alicet1:alice::
14321:ali1_121:14200:alicet1:alice::
14322:ali1_122:14200:alicet1:alice::
14323:ali1_123:14200:alicet1:alice::
14324:ali1_124:14200:alicet1:alice::
14325:ali1_125:14200:alicet1:alice::
14326:ali1_126:14200:alicet1:alice::
14327:ali1_127:14200:alicet1:alice::
14328:ali1_128:14200:alicet1:alice::
100018801:gsdc23a01:1000188:in0188:alice:sgm:
```

```
[root@alice-kisti-hpc ~]# cat /opt/glite/yaim/etc/groups.conf
"/alice/ROLE=lcgadmin":::sgm:
"/alice/ROLE=production":::prd:
"/alice/ROLE=pilot":::pilot:
"/alice":::
```

❖ 8 nodes are added in the HPC cluster

- In previous ATCF, the site had 2 nodes for testing whether it operates without problems.
- Currently, the HPC site has a total 10 KNL nodes by adding 8 nodes on June 24.
 - As it, the maximum CPU cores that the site supports is 680 (=68*10).

[The cluster 1 year ago]

vnode	state	njobs	run	susp	mem f/t	ncpus f/t	nmics f/t	ngpus f/t	jobs
...									
node8304	free	0	0	0	94gb/94gb	68/68	0/0	0/0	--
node8305	free	0	0	0	94gb/94gb	68/68	0/0	0/0	--



[The current cluster]

```
[gsvc23a01@alice-kisti-hpc ~]$ pbsnodes -aSj | grep node63
```

node6309	free	1	1	0	93gb/93gb	4/68	0/0	0/0	15502772.pbs
node6310	free	1	1	0	93gb/93gb	4/68	0/0	0/0	15502772.pbs
node6311	free	1	1	0	93gb/93gb	4/68	0/0	0/0	15502772.pbs
node6312	free	1	1	0	93gb/93gb	4/68	0/0	0/0	15502772.pbs
node6313	free	1	1	0	93gb/93gb	4/68	0/0	0/0	15502772.pbs
node6314	free	1	1	0	93gb/93gb	4/68	0/0	0/0	15502772.pbs
node6315	free	1	1	0	93gb/93gb	4/68	0/0	0/0	15502772.pbs
node6316	free	1	1	0	93gb/93gb	4/68	0/0	0/0	15502772.pbs
node6317	free	1	1	0	93gb/93gb	4/68	0/0	0/0	15502772.pbs
node6318	free	1	1	0	93gb/93gb	4/68	0/0	0/0	15502772.pbs

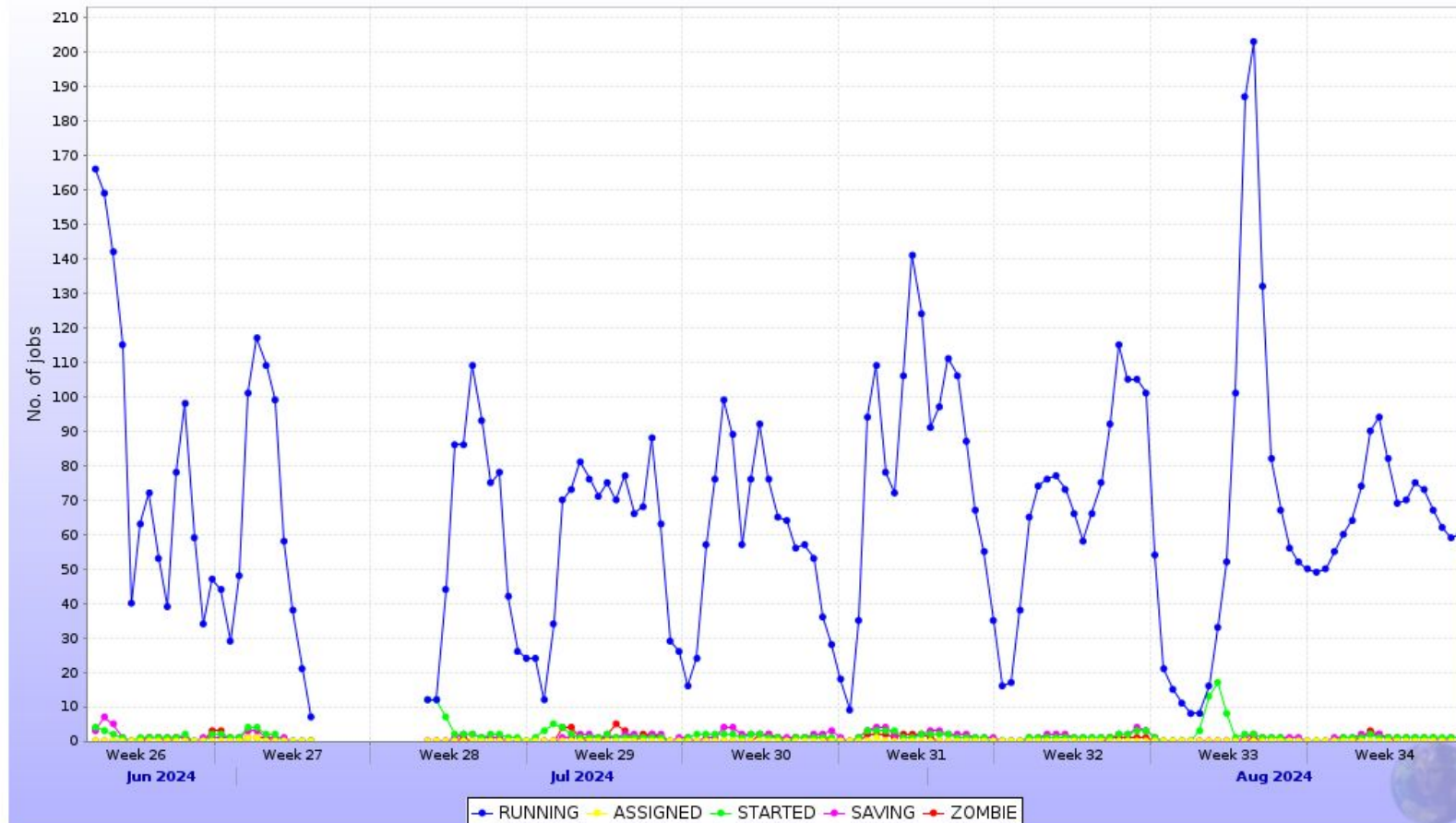
the maximum cores of KISTI_GSDC_Nurion = 15% of KISTI Tier 1's cores

Current State of The Site

❖ 1) The plot of active jobs in KISTI_GSDC_Nurion (Jun 24 - Aug 24)

- the large gap between the maximum and minimum

Active jobs in KISTI_GSDC_Nurion



max	median	min
218	67	6

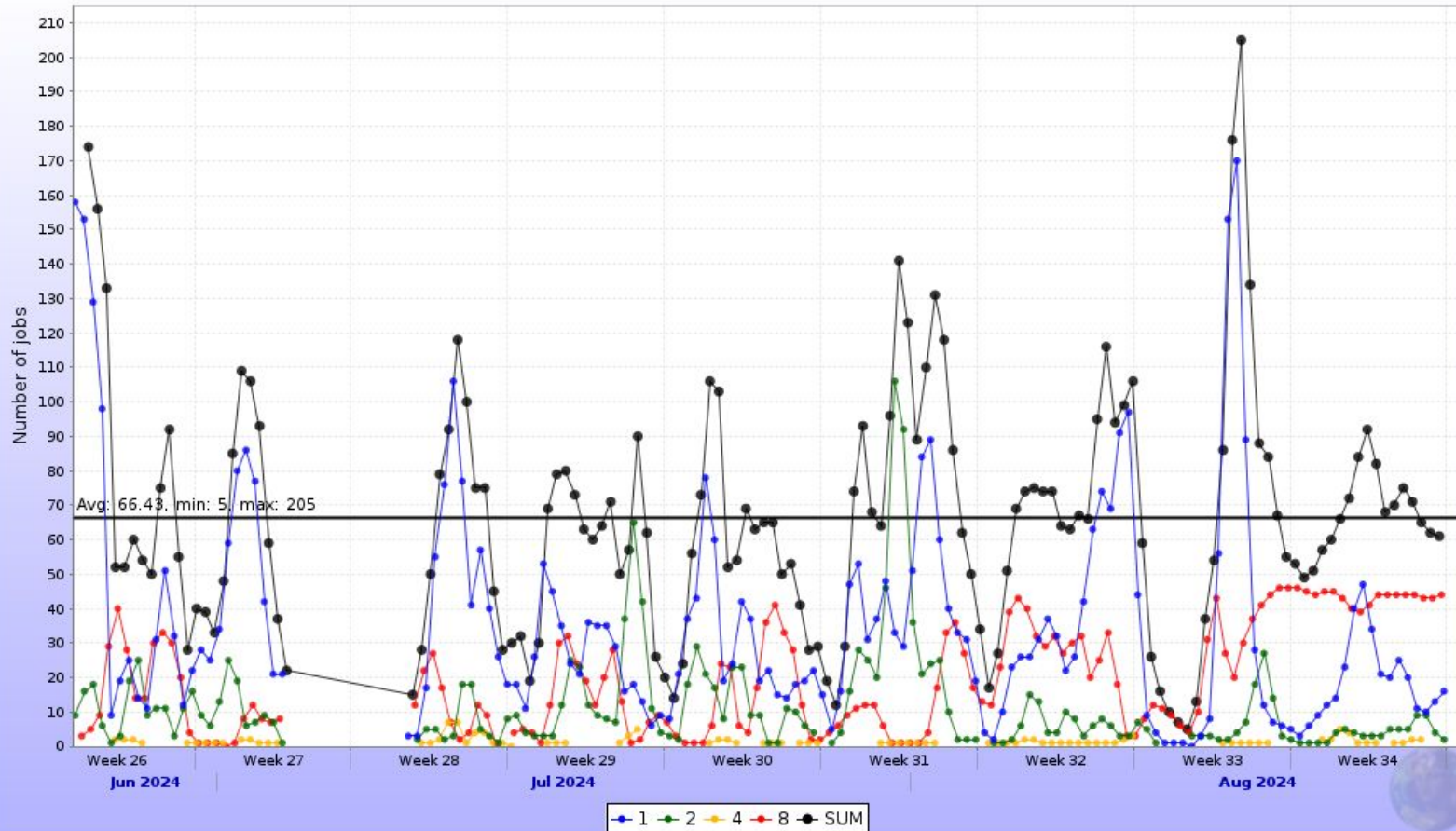
Average jobs for 2 months	Average jobs on Jun 24 - Jul 24	Average jobs on Jul 25 - Aug 24
67	61	70

- submission time and date that affect the shape of plots
(ex. Fewer jobs are submitted on weekends than on weekdays)

❖ 2) The plot of Number of jobs at KISTI_GSDC_Nurion (Jun 24 - Aug 24)

- Mostly 1-core / 2-core / 8-core jobs among the executed since adding nodes

Number of jobs at KISTI_GSDC_Nurion



	Average jobs for 2 months	Average jobs on Jun 24 - Jul 24	Average jobs on Jul 25 - Aug 24
1-core	35	38	32
2-core	11	12	10
4-core	1	1	1
8-core	19	12	25

Future Plans - OS upgrade

❖ The OS of CE node will be upgraded: CentOS 7 → Alma 9

- CentOS 7 end of life : June 30, 2024
 - Now over than the date of EOL
- Necessary to install the newer operating system

```
[gsdc23a01@alice-kisti-hpc ~]$ lsb_release -a
LSB Version:      :core-4.1-amd64:core-4.1-noarch
Distributor ID:  CentOS
Description:     CentOS Linux release 7.9.2009 (Core)
Release:         7.9.2009
Codename:        Core
[gsdc23a01@alice-kisti-hpc ~]$
```



CentOS



AlmaLinux

❖ We are going to transit VOBox support type from bare-metal to ‘Container’

- **Why we change it?**
 - because of fast recovery
 - because of easy service deployment
- **Easy to say...**
 - If VOBox nodes are suddenly shutdown or VOBox service is not operated normally,
 - We just need to the new node to replace the error nodes,
 - We create a VOBox container with container images uploaded in docker registry on the new one.
- **Benefits of introducing VOBox container**
 - for administrator : can take fast and easy actions for recovering VOBox service shutdown
 - for grid user : can utilize the computing environment in the site without long waiting time for recovery

❖ To apply the VOBox container..

- The activity of develop a new container images for PBS-Pro version is required,
 - As not including the schedules to support the images.

Create Container

1. Clone the repository containing the desired preconfigured setup:

Scheduler	Command
HTCondor	<code># git clone https://gitlab.cern.ch/mstoretv/dockervobox.git</code>
ARC/Generic	<code># git clone https://gitlab.cern.ch/mstoretv/dockervobox_arc.git</code>

- There is only support the images of HTCondor and ARC/Generic scheduler.

❖ Information of Dockerfile (HTCondor ver.)

- The base image : [gitlab-registry.cern.ch/linuxsupport/alma9-base](https://gitlab.cern.ch/linuxsupport/alma9-base) (alma 9)

```
1 FROM gitlab-registry.cern.ch/linuxsupport/alma9-base
2
3 # Add safeguards, repos and packages
4 # hadolint ignore=DL3033
5 RUN sed -i '$ d' /etc/dnf/dnf.conf
6
7 ##Packages Installation##
8 > RUN curl https://repository.egi.eu/sw/production/cas/1/current/repo-
28
29 # Setup ssh access and add user(s)
30 # hadolint ignore=DL4006
31 > RUN /usr/bin/ssh-keygen -A && \ ...
35 > RUN /usr/bin/gsissh-keygen -q -t ed25519 -f /etc/gsissh/ssh_host_ed2
38
39 ###CONFIG###
40 RUN echo "alias ll='ls -la'" >> /root/.bash_profile
41 RUN echo -e "export LC_ALL=C\nexport LANG=C\nexport LANGUAGE=C" >> /
42
43 COPY ./configs/vobox/etc/vobox-proxy.conf /var/lib/vobox/alice/etc/
44 > RUN mkdir /var/lib/vobox/alice/stop && \ ...
47
48 #supervisord
49 COPY ./configs/supervisord/supervisord.conf /etc/
50
51 #alicesgm user
52 COPY ./configs/alicesgm/ /home/alicesgm
53 # hadolint ignore=SC2039
54 > RUN mkdir -p /home/alicesgm/bin && ln -s /cvmfs/alice.cern.ch/bin/al
58
59 #nftables
60 COPY ./configs/nftables/vobox_ipv4.nft /etc/nftables/
61 COPY ./configs/nftables/vobox_ipv6.nft /etc/nftables/
62 RUN echo "include \"/etc/nftables/vobox_ipv4.nft\"" >> /etc/sysconfi
63 RUN echo "include \"/etc/nftables/vobox_ipv6.nft\"" >> /etc/sysconfi
64
65 #voms
66 RUN dnf -y install wlcg-iam-lsc-alice && \
```

```
67 | dnf -y install wlcg-iam-vomses-alice
68
69 #gsisshd
70 RUN rm /etc/gsissh/ssh_config
71 RUN rm /etc/gsissh/ssh_config.d/50-redhat.conf
72 COPY ./configs/gsissh/ssh_config /etc/gsissh/
73 RUN update-crypto-policies --set DEFAULT:SHA1 || exit
74
75 #rsyslog
76 RUN rm /etc/rsyslog.conf
77 COPY ./configs/rsyslog/rsyslog.conf /etc/rsyslog.conf
78
79 #myproxy (need to change the environment variable = alice-kisti-hpc.s
80 RUN echo "export MYPROXY_SERVER=myproxy.cern.ch" >> /etc/profile.d/g
81
82 #cron
83 COPY ./configs/cron/alice-box-proxyrenewal /etc/cron.d/
84 COPY ./configs/cron/edg-mkgridmap /etc/cron.d/
85 COPY ./configs/cron/edg-mkgridmap.conf /etc/
86 COPY ./configs/cron/grid-mapfile-local /etc/
87 COPY ./configs/cron/alicesgm /var/spool/cron/
88 RUN chown -R alicesgm.alicesgm /var/spool/cron/alicesgm && chmod 600
89 | chmod 644 /etc/cron.d/[ae]*
90
91 #condor -> pbs-pro (need to change below code suit for pbs-pro sched
92 COPY ./configs/condor/00-minicondor.vobox /etc/condor/config.d/
93 COPY ./configs/condor/02_container_extra.config /etc/condor/config.d/
94 COPY ./configs/condor/99-alice-vobox.conf /etc/condor/config.d/
95 RUN chmod 644 /etc/condor/config.d/*
96 COPY ./configs/condor/ce-usage.sh /home/alicesgm/
97 RUN chown -R alicesgm.alicesgm /home/alicesgm/ce-usage.sh && \
98 | chmod a+x /home/alicesgm/ce-usage.sh
99
100 #more manpages
101 COPY ./extras/manpage_bundle.tar.gz /tmp
102 RUN mkdir -p /usr/share/man/overrides/ && tar -xf /tmp/manpage_bundl
103
```

```
104 #use a "last" without nologout bug
105 RUN rm /usr/bin/last
106 COPY ./extras/last /usr/bin/
107 RUN chmod 755 /usr/bin/last
108
109 #JALien-V0Box (CVMFS shortcut function)
110 RUN echo 'jalien-vobox () { /cvmfs/alice.cern.ch/scripts/vobox/jalie
111
112 ###INIT###
113
114 #Add init and service scripts
115 COPY ./init.sh /init.sh
116 COPY ./services/* /services/
117 RUN mkdir -p /etc/init.d/ && \
118 | mkdir -p /etc/rc.d/init.d/ && \
119 | chmod u+x /init.sh && \
120 | chmod -R u+x /services && \
121 | ln -s /services/alice-box-proxyrenewal /etc/init.d/ && \
122 | chmod 755 /services/alice-box-proxyrenewal
123 # mv /etc/rc.d/init.d/functions /etc/rc.d/init.d/functions2 && \
124 # ln -s /services/functions /etc/rc.d/init.d/
125
126 #Add function to give users/scripts systemd-like call for supervisor
127 RUN echo 'systemctl () { supervisorctl "$@"; }' >> /etc/bashrc
128
129 #Run init script upon container start
130 CMD [ "/init.sh" ]
131
```

develop the container image(PBS-Pro ver.) based on its HTCondor version.

Thank you!

Hyeonjin Yu

hyeonjin.yu@cern.ch