US sites - operations and resources planning
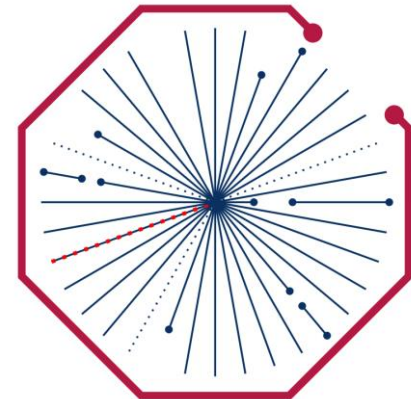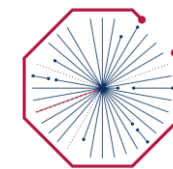
ALICE T1/T2 Workshop

September 28, 2022

Irakli Chakaberia

US Department of Energy Supported ALICE-USA Group

- ALICE-USA Computing project provides and maintains compute and storage resources for ALICE

- Fulfills DOE funded MoU-based ALICE USA obligations for compute and storage resources to ALICE

- Operates ALICE grid facilities at 2 DOE labs

- Project was proposed in 2009

- Operational changes 2017-2021
  - ORNL T2 relocated within ORNL, personnel retained
  - New LBNL/ITD cluster - HPCS, PDSF retired in 2019
  - Report all resources to the WLCG under the US_LBNL_ALICE federation
  - Funded R&D use of HPC resources
  - Opportunistic use of Lawrencium supercomputer at LBNL

Average running jobs
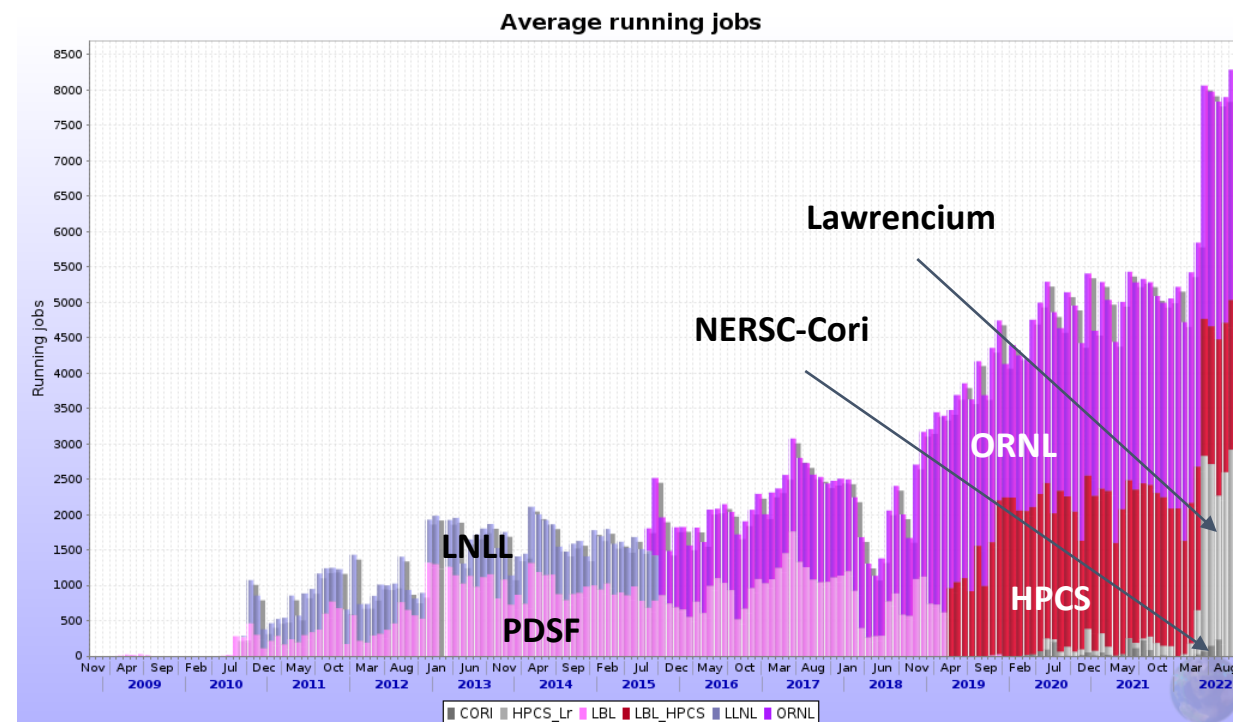
US-ALICE Review Meetings

- External project review in July 2021

- Received favorable report

- US Operations review in May 2022

2021 RRB CPU Delivery vs. Obligations

- For the accounting reasons obligations are evenly split between sites, in reality during the reporting period ORNL had slightly less CPU

- ORNL only reported, to OSG, 4 hours of downtime during RRB | 2021 year

- LBNL has reported about 3 weeks of downtime

CPU Obligations	kHS06
ALICE-USA	70.25
LBNL HPCS T2	35.125
ORNL T2	35.125

USA T2 Site	Per/core CPU (HS06/Core)	CPU Delivered (MHS06*hr) ALICE-Monitored	CPU Delivered (MHS06*hr) WLCG-Monitored	ALICE-USA Obligation (MHS06*hr)	Delivered per WLCG-Monitored Obligation (%)
LBNL	16.7	303.3	318.8	290.1	109.9
ORNL	11.5	302.2	303.2	307.6	98.6
Total		605.5	621.9	597.6	104.1
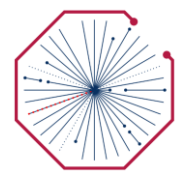
WLCG Summary | RRB 2021

Target Availability for each site is 97.0%.

Availability Algorithm:

@ALICE_CE * @ALICE_VOBOX * all AliEn-SE
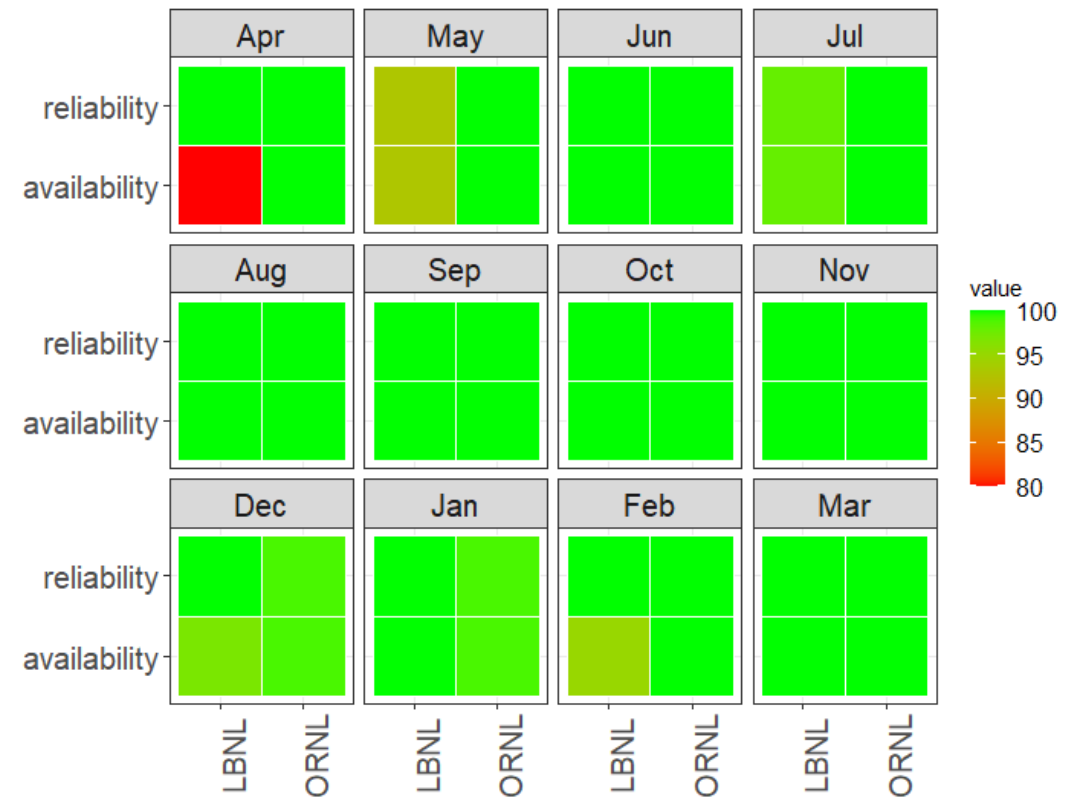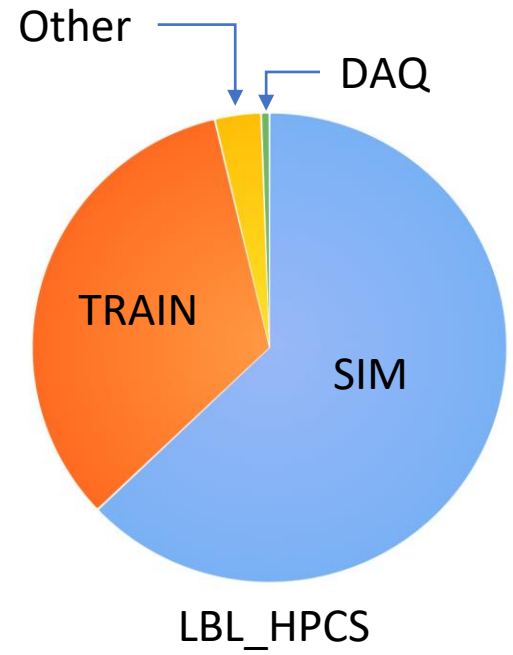
MonAlisa Summary | RRB 2021

- Average job CPU efficiency and total wall time
- JobMix defines the average CPU efficiency delivered
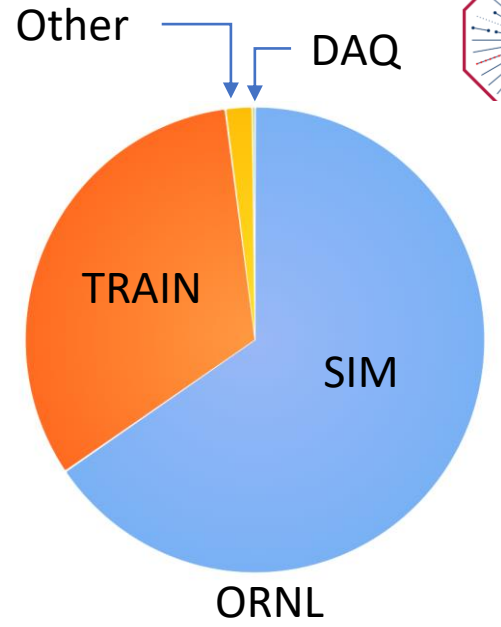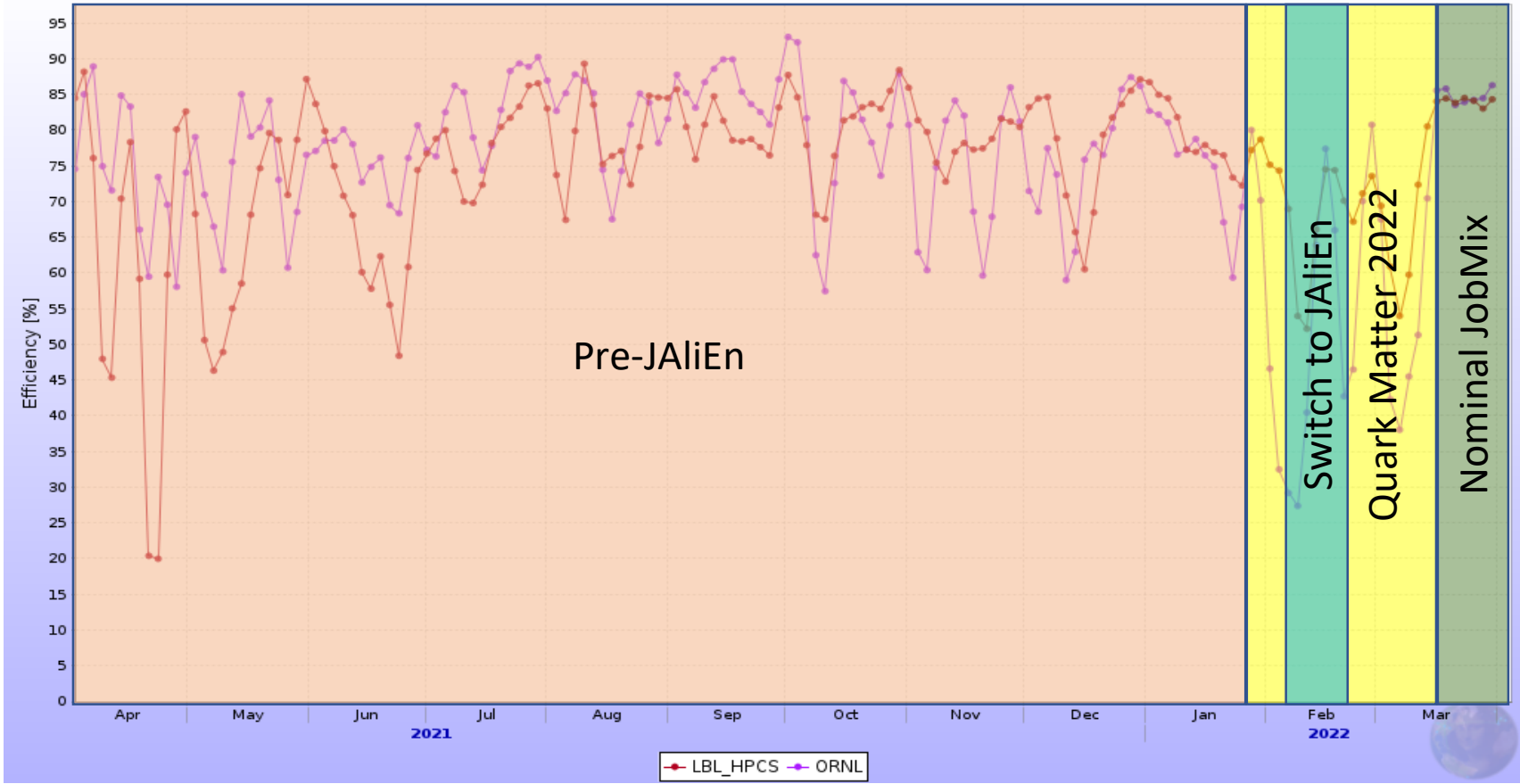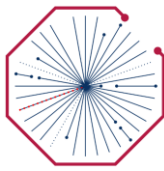
CPU Performance | RRB 2021

Cori and Lawrencium were b[...]
and tu[...]
period[...]

Running Jobs

ORNL added another

Both sites switched to JAliEn

LBL-HPSC
- SLURM issues even before JAliEn
- Mix-up with env variables
- Problem with ulimit –u setting
- DNS services got blocked by the firewall after replacing the master node cluster
- We ended up about 200 jobs below what we had before the switch

ORNL
- Changes needed for configuration
- After the initial issues with the switch, it looks like a steady job load but there were some efficiency problems

Lawrencium & CORI – no activity

ORNL Pledge

LBNL Pledge

Both sites switched to JAliEn
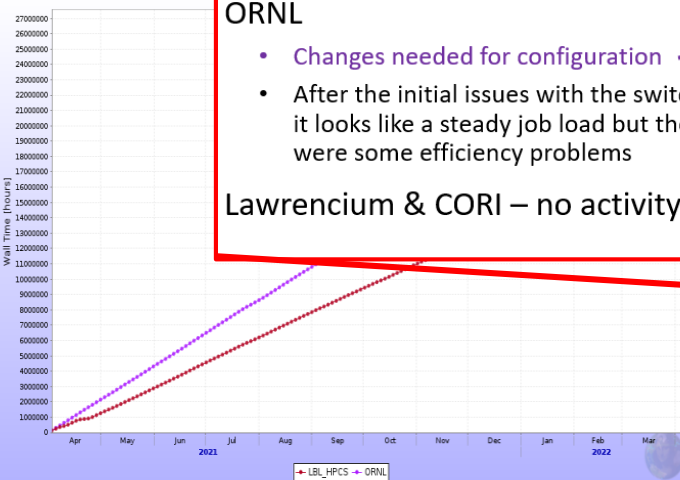
Storage Performance | RRB 2021

- SE availability has been very high – over 95% on average

- IPv6 ~~the R~~

- IPv6 ~~avail~~ ~~read~~ ~~T2~~

AliEn SEs availability for writing
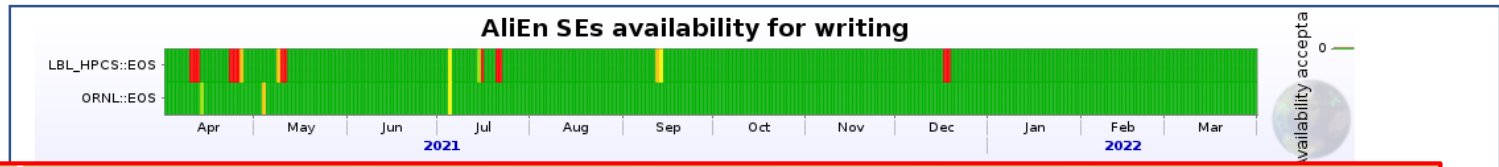
Site name	VOBox and SE problems
LBL_HPCS alice-vobox.lbl.gov	**Networking**: low buffer size kernel parameters: tcp_wmem_max=4 MB only **Networking**: No IPv6 public address **CVMFS**: old CVMFS version detected (2.9.2), please upgrade to 2.9.4.

ORNL::EOS   01 Apr 2021 00:42   31 Mar 2022 23:47   8755   27   99.69%   99.68%

Do we need IPv6 on WNs? How important is it?

96.79	97.33
99.92	99.68

98 → 100%   100%

IPv6 Only
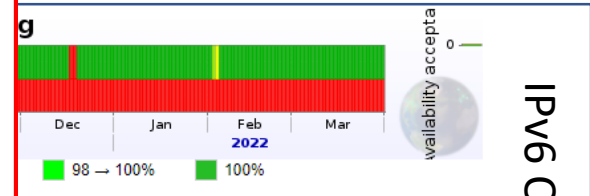
Statistics

Link name	Data		Individual results of writing tests			Overall
	Starts	Ends	Successful	Failed	Success ratio	Availability
LBL_HPCS::EOS	31 Mar 2021 23:42	31 Mar 2022 23:47	8420	341	96.11%	96.09%
ORNL::EOS	31 Mar 2021 23:42	31 Mar 2022 23:50	0	8756	0%	0%
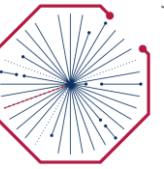
Lessons Learned - LBNL

- AliEn to JAliEn switch needed accurate attention to ENV and LDAP settings
    - `ulimit` setting became important with much more java and singularity processes spawn by JAliEn
    - Initial TTL at LBL was set for 48 hours that highlighted the issue with the jobAgents sitting idle
    - Setting was changed to match one at ORNL (25 hours) that seems optimal with the length of jobs, thus SLURM jobs expire and not let jobAgents sit idle for long
    - This change recovered about 200 jobs for LBNL and later the issue with the idle jobAgent was fixed

Lessons Learned - ORNL

- Unusual number of jobs terminating simultaneously on certain hosts

- No obvious reasons (nothing logged, no anomalous system behavior)

- SLURM reported that the host was down for one cycle, then it was back up again - marked the jobs dead, killed after reconnection

- Mean run time across all jobs on cray-012 varied between 4 hours and 15.3 hours. Mean run time across all jobs on 109 varied between 12 minutes and 15.1 hours.

- Diagnosing: host, network, switches – no issues

- Log files report no problems until they stop

- Solution: a couple of nodes had stale /etc/host files - they were down on last configuration push. Updating host files resolved the issue.

- Would a configuration mismatch monitoring and logging be a good idea to avoid such problems in the future?!

Active jobs in ORNL

Zombie grass

Slide from: Pete Eby & Steve Moulton
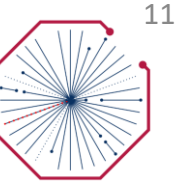
- HTTPS support was enabled on both ORNL and LBNL sites in Mid 2021
- NERSC security scans discovered a vulnerability in the implementation.
- John reported a security vulnerability to EOS team in August.
- EOS replied that they will fix the issue.
- ALICE-USA sites disabled the HTTPS support
- Fix was implemented pretty promptly in September 2021
- But we only learned about it by explicitly inquiring about the subject in June 2022
- This raises questions:
  - Should we have a convenient way of communicating which EOS/XRootD/CVMFS/etc. versions have the necessary developments/fixes and need to be updated?
  - JAliEn handles updates pretty seamlessly and transparently for the sites but the rest need manual involvement from the system admins, however they need a timely cue to act
- For the moment HTTPS is enabled with the new implementation and by explicitly firewalling the access for the library with vulnerable implementation (XrdHttp is is use and libmicrohttpd is firewalled)
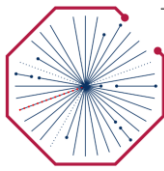
ALICE-USA Obligations and Acquisition Strategy

- ALICE-USA share of resources in respect to WLCG / C-RCG recommendations for the next few years and planned hardware acquisition strategy table

Year	FY2021	FY2022	FY2023	FY2024	FY2025
ALICE Requirements					
CPU (kHS06)	852	1013	1164	1338	1579
Disk (PB)	91	104	121	138	163
ALICE-USA Participation					
ALICE M&O-A	561	571	566	566	566
ALICE-USA M&O-A	46	46	46	46	46
ALICE-USA/ALICE (%)	8.20%	8.06%	8.13%	8.13%	8.13%
ALICE-USA Obligations					
CPU (kHS06)	69.9	81.6	94.6	108.7	128.3
Disk (PB)	7.5	8.4	9.8	11.2	13.2

| Resource	Installed	FY2022	FY2023	FY2024	FY2025
LBNL HW & Costs					
CPU change (+/- kHS06)		-2.0+10.0	-7.0+11.0	-3.0+10.0	0.0+10.0
CPU Installed (kHS06)	35	43	47	54	64
Disk change (+/- PB)		0.0+2.0	0.0+0.0	-1.6+2.1	0.0+0.9
Disk installed (PB)	2.75	4.75	4.75	5.25	6.15
ORNL HW & Costs					
CPU change (+/- kHS06)		-10.0+16.0	-11.0+17.0	0.0+7.0	0.0+10.0
CPU Installed (kHS06)	35	41	47	54	64
Disk change (+/- PB)		-2.75+3.0	0.0+2.0	0.0+1.2	0.0+0.9
Disk installed (PB)	2.75	3.00	5.00	6.20	7.10

CPU Acquisition

- We are gradually retiring out of warranty hardware

- HS06 for the planned CPU acquisition is roughly estimated based on the measurements we made on similar nodes
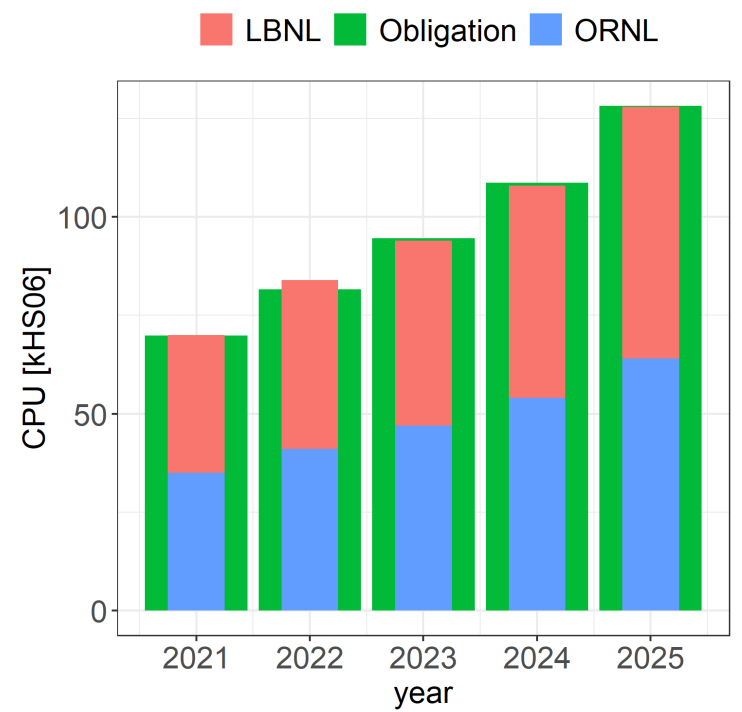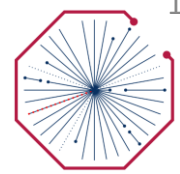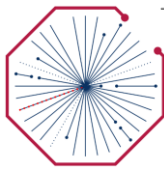
Storage Acquisition

- Our storage utilization has been on a lighter side
- We keep evaluating the utilization and purchase the new storage based on it and optimized for the available budget
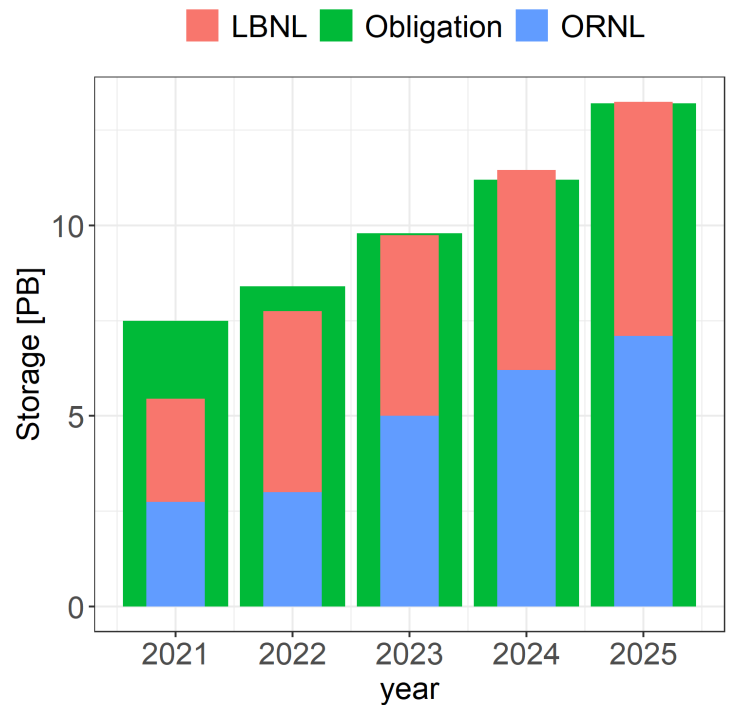
HPCS Lawrencium

- R&D project for using opportunistic resources on HPC clusters is now fully operational

- We make steady use of Lawrencium for the past half a year month

- Started with only MC jobs to minimize frustration if preemption was Performance was too disruptive, but it proved so reliable that we mixed in train jobs for the past month

CPU Efficiency per Job Type

All Jobs

But quickly increased the limit

Started with a low allocation
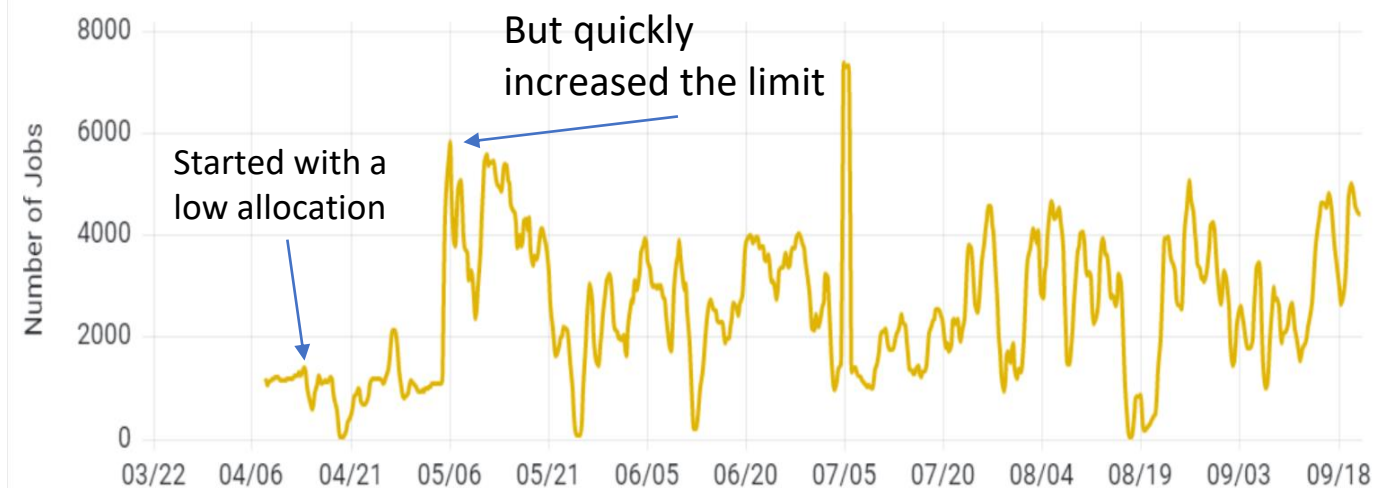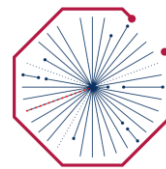
2022 RRB CPU Delivery vs. Obligations

- Due to some budget uncertainties the FY2021-2022 hardware acquisition was postponed and is in progress right now

- This causes slight under delivery for the ALICE-USA sites starting RRB2022

- The use of Lawrencium however picks up the slack

USA T2 Sites	Per/cpre CPU (HS06/Core)	CPU Delivered (MHS06*hr) ALICE-Monitored	CPU Delivered (MHS06*ht) WLCG-Monitored	ALICE-USA Obligation (MHS06*hr)	Delivered (%) WLCG-Monitored	Delivered (%) ALICE-Monitored	Delivered ALICE-Monitored Excluding Lawrencium
LBNL	16.7	148.9	191.8	89.544	2.141883543	166.3%	82.4%
ORNL	11.5	82.7	84.8	89.544	0.946723399	92.4%	92.4%
Total	28.2	231.6	276.6	179.088	1.544303471	**129.3%**	**87.4%**
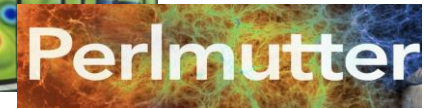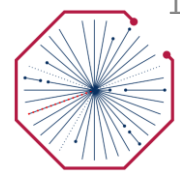
NERSC Allocation

- As part of the project, we request allocation on NERSC HPCs

- Cori is generally used as out R&D platform for HPCs

- Cori was announced to retire in January 2023, need to start planning to move the workflow to the new Perlmutter machine
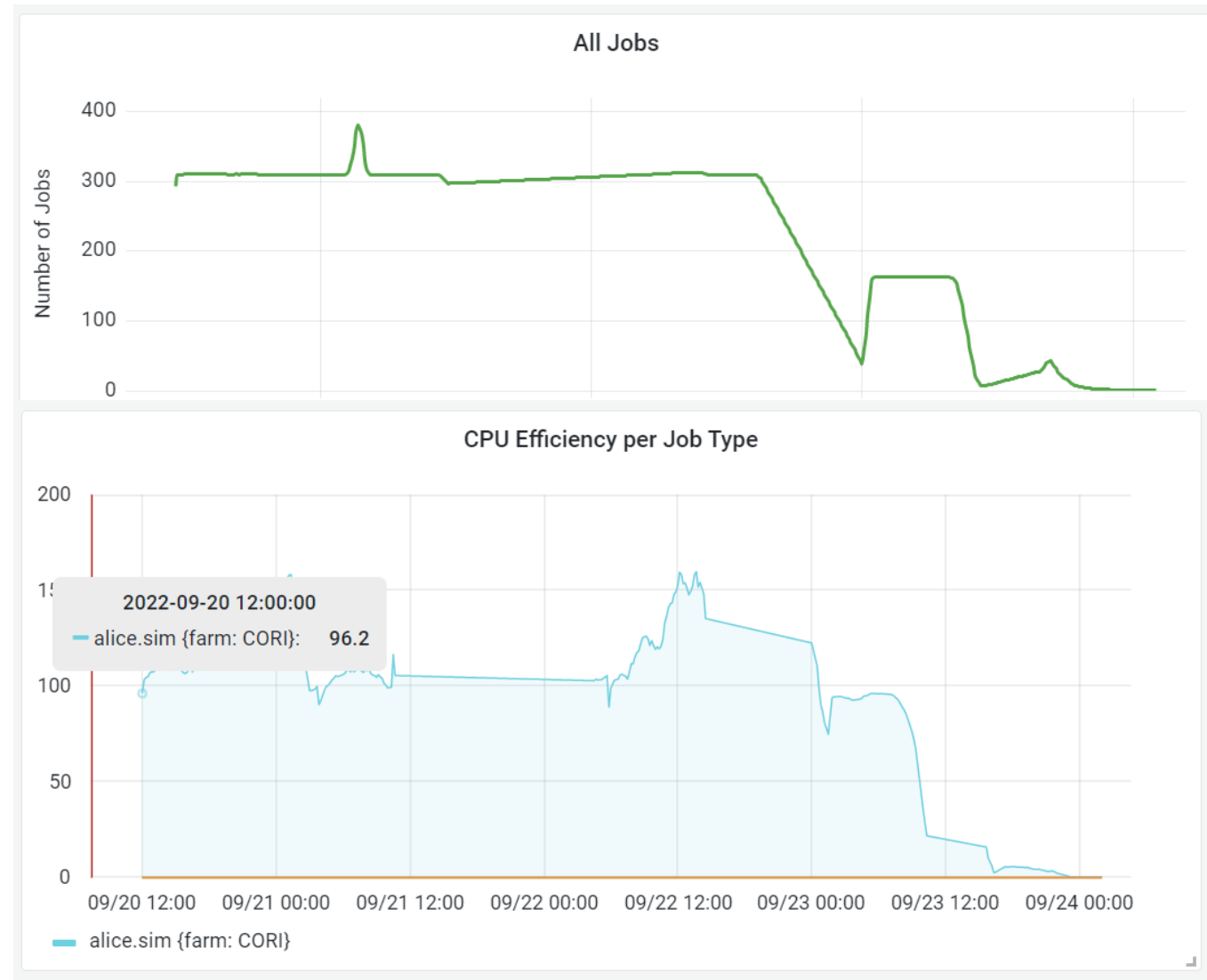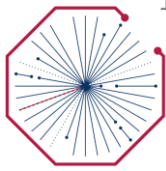
Cori → Perlmutter

Cori

Partition	# of nodes	Processor	Clock Rate	Cores Per Node	Memory Per Node
Login	20	Intel Xeon Processor E5-2698 v3	2.3 GHz	32	515 GB
Haswell	2,388	Intel Xeon Processor E5-2698 v3	2.3 GHz	32	128 GB
KNL	9,688	Intel Xeon Phi Processor 7250	1.4 GHz	68	96 GB (DDR4), 16 GB (MCDRAM)
Large Memory	20	AMD EPYC 7302	3.0 GHz	32	2 TB

Perlmutter

| Partition	# of nodes	CPU	Cores Per Node	GPU	Aggregated Memory [TB]
GPU	1536	1x AMD EPYC 7763	64 for CPU	4x NVIDIA A100 (40GB)	240
CPU	3072	2x AMD EPYC 7763	64	-	384
Login	40	1x AMD EPYC 7713	64	1x NVIDIA A100 (40GB)	
Large Memory	4	1x AMD EPYC 7713	64	1x NVIDIA A100 (40GB)	
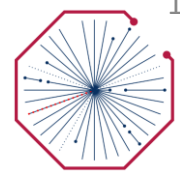
ALICE-USA Local cluster monitoring
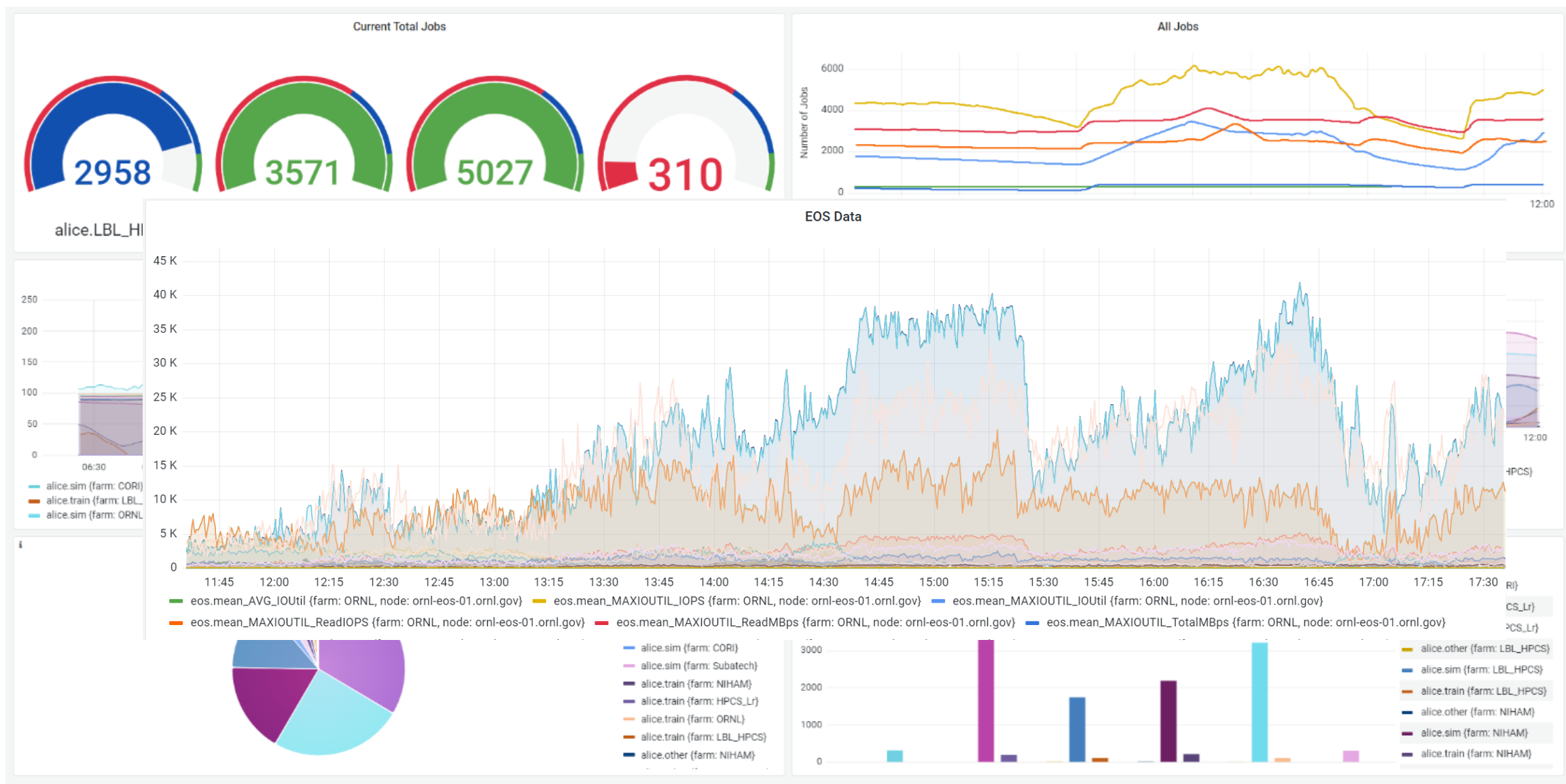
ALICE-USA Analysis Facility

- ALICE Collaboration has introduced the concept of Analysis Facility (AF) in its Run 3 computing model
  - ALICE AF is meant to be a Grid Facility like T2s
  - AFs will provide resources dedicated for the AOD level physics analysis

- The proposal to host such an AF by the ALICE-USA was presented to the DOE
  - ~2000 cores for analysis jobs
  - ~2.5 PB of usable disk storage

- AF is planned to be hosted at LBL
  - In parallel with our current T2
  - The data would be accessible from NERSC by ALICE-USA physicists (small latency access)

- DOE showed interest in the project

- Project is pending additional consideration by the DOE
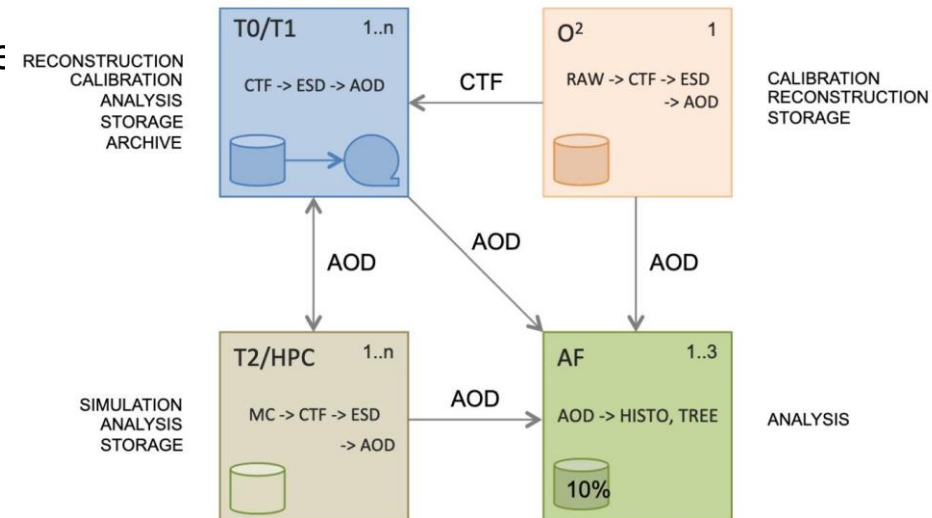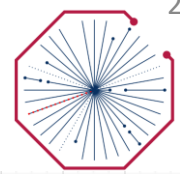
ALICE-USA Analysis Facility Prototype

- With the retired out-of-warranty LBNL WN and storage we plan to build ALICE AF prototype facility

- In existing setup ALICE shares storage and CPU resources with STAR experiment

- Plan is to move STAR storage and CE to ITD Lr cluster

- Rearrange storage into 3 alicefsts and 1 aliceAFfst
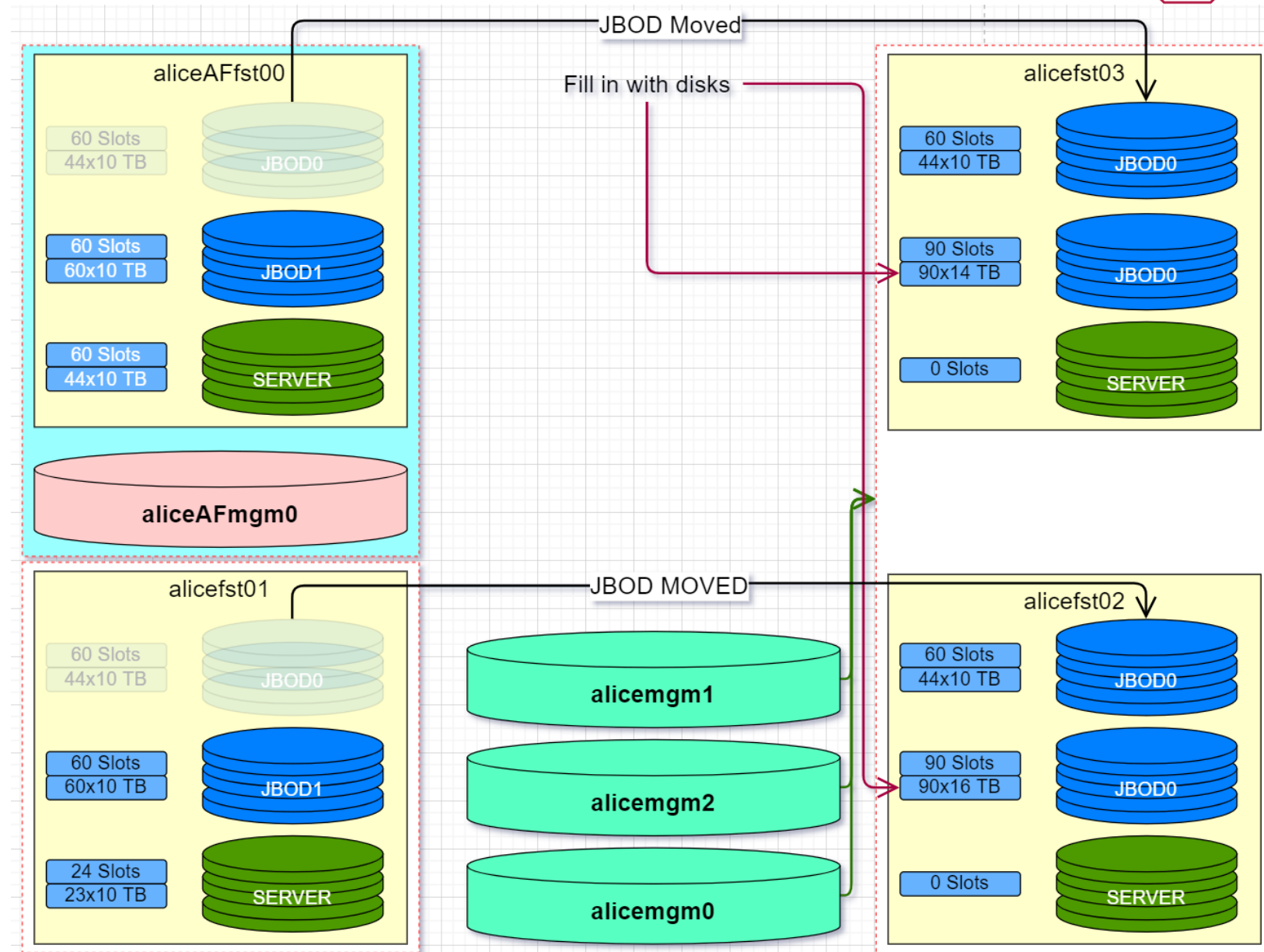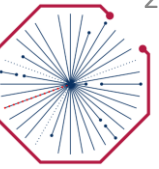
Potential use of ORNL out-of-warranty hardware

- ORNL will also retire some hardware that could be used for:
  - Second prototype AF
    - Feedback: Perhaps a dedicated site where developers can test fast data access type of jobs
  - US-based CCDB storage element
    - Feedback: This may have value to have db replica of a particular type

- Perhaps both?!

Summary

- ALICE-USA Computing project grows along with ALICE to meet it's ALICE-USA obligations

- We currently operate two T2 sites, one at LBL and one at ORNL

- Sites continue to deliver pledged resources with very high availability and reliability

- Both sites have successfully switched to JAliEn in the middle of February

- Local R&D projects have good results in utilizing the HPC recourses – Lawrencium is now a well-respected member of ALICE grid family

- Hosting of an AF, as part of the ALICE-USA computing project, was evaluated, proposed, and plans for the prototype facility are in place