Experience with K8s at Coffea-Casa AF@UNL

Garhan Attebury, **Carl Lundstedt**, Derek Weitzel
University of Nebraska Holland Computing Center

Mat Adamec, Ken Bloom, Oksana Shadura,
University of Nebraska, Lincoln

Mátyás Selmeci
University of Wisconsin, Madison

Brian Bockelman
Morgridge Institute

Analysis facilities: prototypes

- **Two AF facilities** with the possible outcome of adding more sites as soon as we gain experience

CMSAF @T2 Nebraska
“Coffea-Casa”
https://cmsaf-jh.unl.edu

Elastic AF @ Fermilab

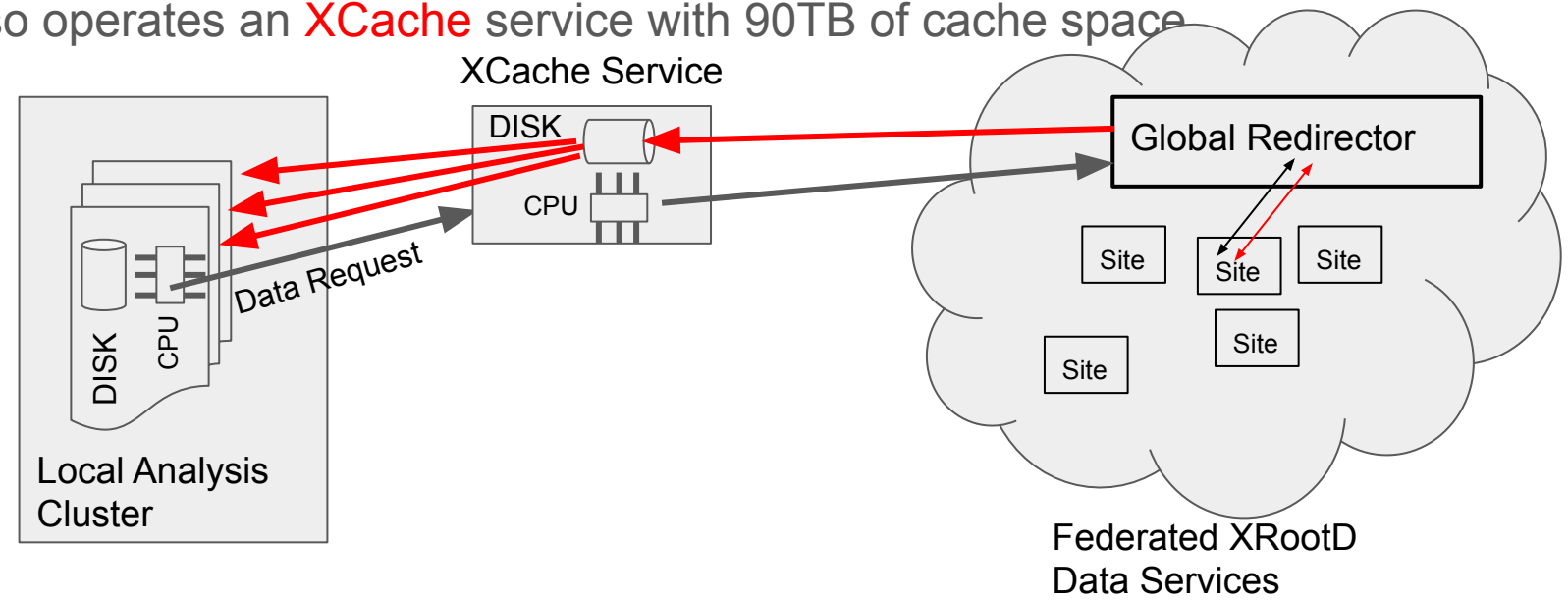
- ***Q4 2020 -*** *Invite first users to test “alpha” version of UNL AF (“coffea-casa”)*
- ***Q4 2020 -*** *Make “coffea-casa” products (Helm charts, modules) deployable in any other AF facility*
  - *Expected first test deployment of **FNAL Elastic AF** during 2021*
- ***Q4 2020 -*** *Finalize testing of ServiceX@UNL AF*
- ***Q1 2021 -*** *Deploy and test data delivery with Skyhook at UNL AF*

T2 Nebraska Site Resources

T2_US_Nebraska has over 13,000 cores available for analysis, accessible via grid interfaces, internally managed by HTCondor.

11 PB of HDFS data storage, 12 XRootD/GridFTP data doors

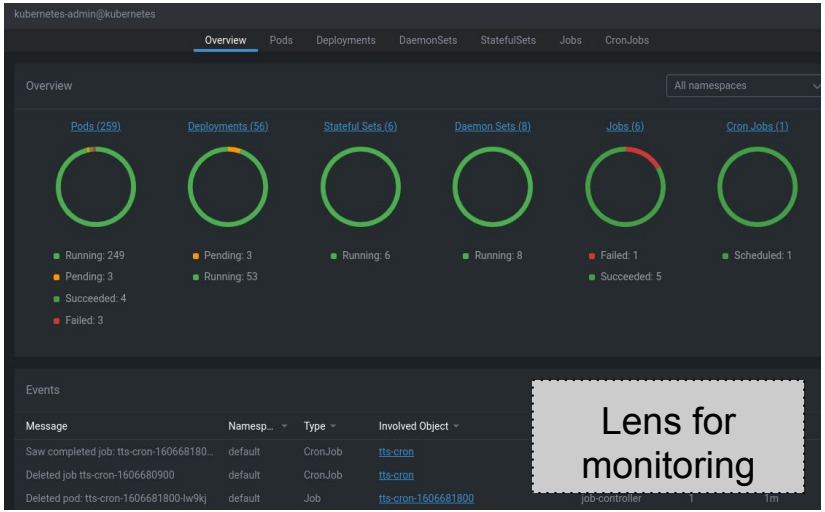
It also operates an XCache service with 90TB of cache space

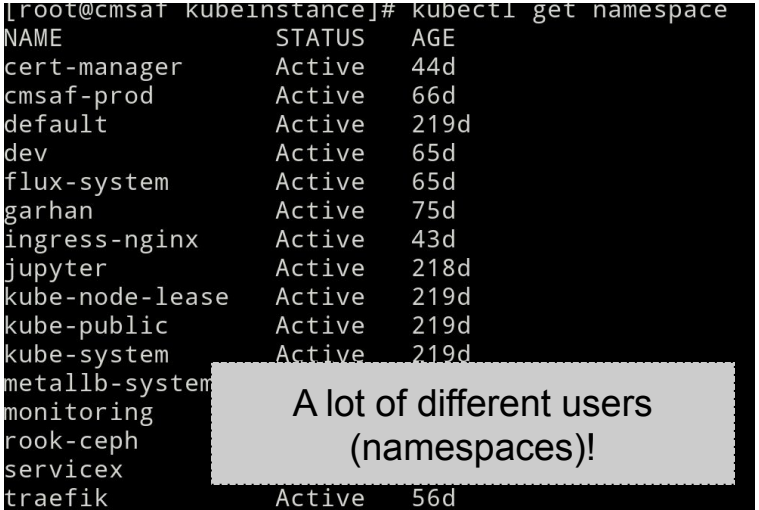


UNL K8s Cluster specs

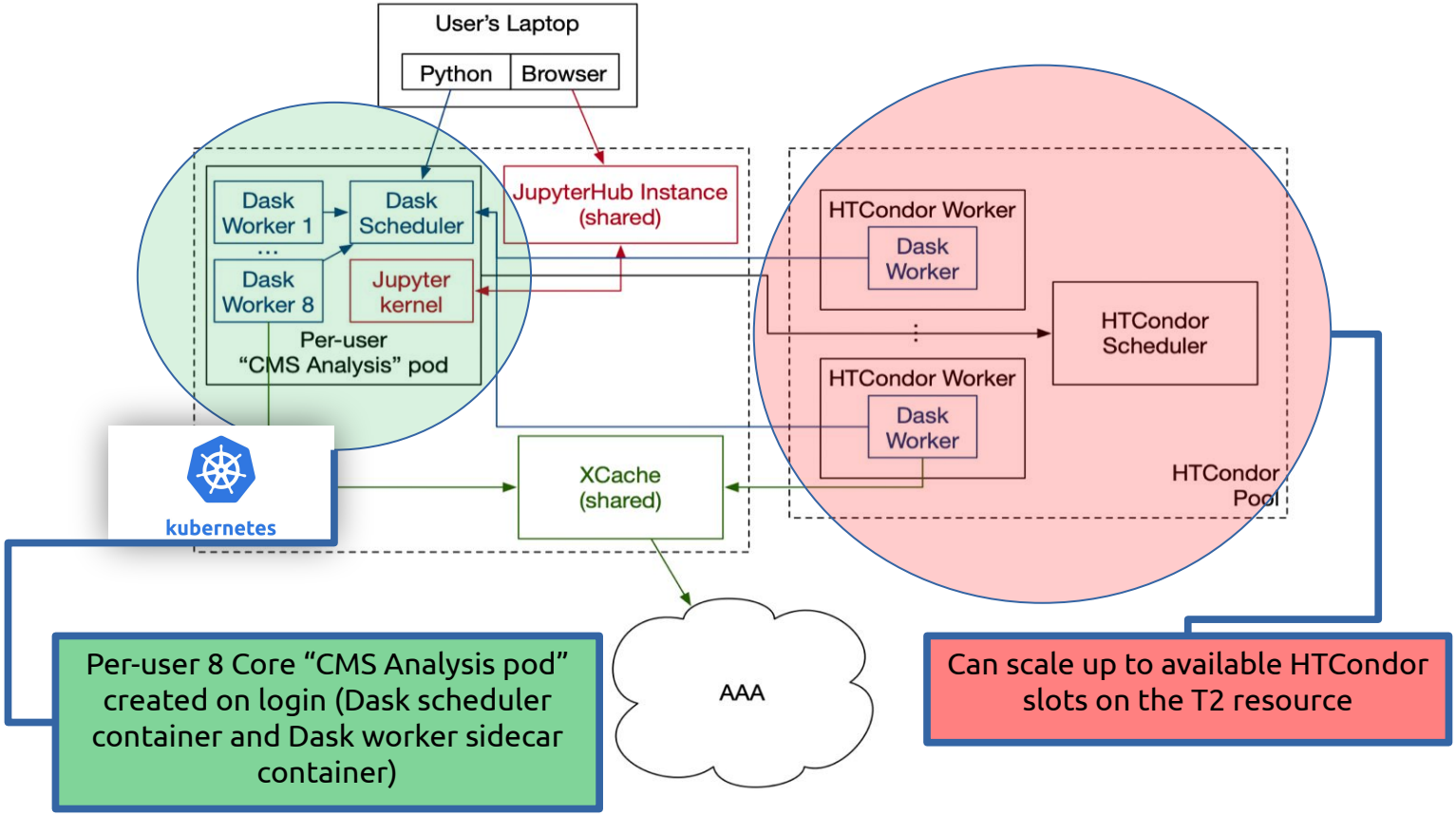
(Currently this Kubernetes setup is constructed with recycled workers and disks.)

Hostname	Role	Description	CPU	RAM
red-kube-vm00[1,2,3]	masters	VMs	2	8GB
red-kube-c07[24,26,28,30]	workers	R710s	24	96GB
red-kube-c10[35,36,37]	workers	Sun X2200	8	32GB
red-kube-c69[21-26]	workers	Sun X2200	8	24GB
red-kube-c69[27-30]	workers	2x4U Supermicro	16	64GB
red-kube-c6931	workers	1U Supermicro	8	32GB





Analysis Facility @ T2 Nebraska



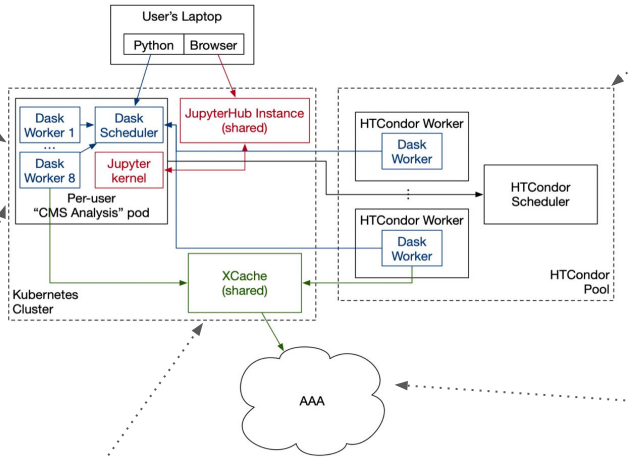
Current status of Analysis Facility @ T2 Nebraska

Enabled token authentication in HTCondor infrastructure

Security - TLS enabled communication between workers and scheduler

CoffeaCasaCluster: HTCondorCluster integration for Dask to allow auto-scaling out to the local HTCondor pool

We are using **a highly customized "CMS Analysis" container** with all the necessary dependencies



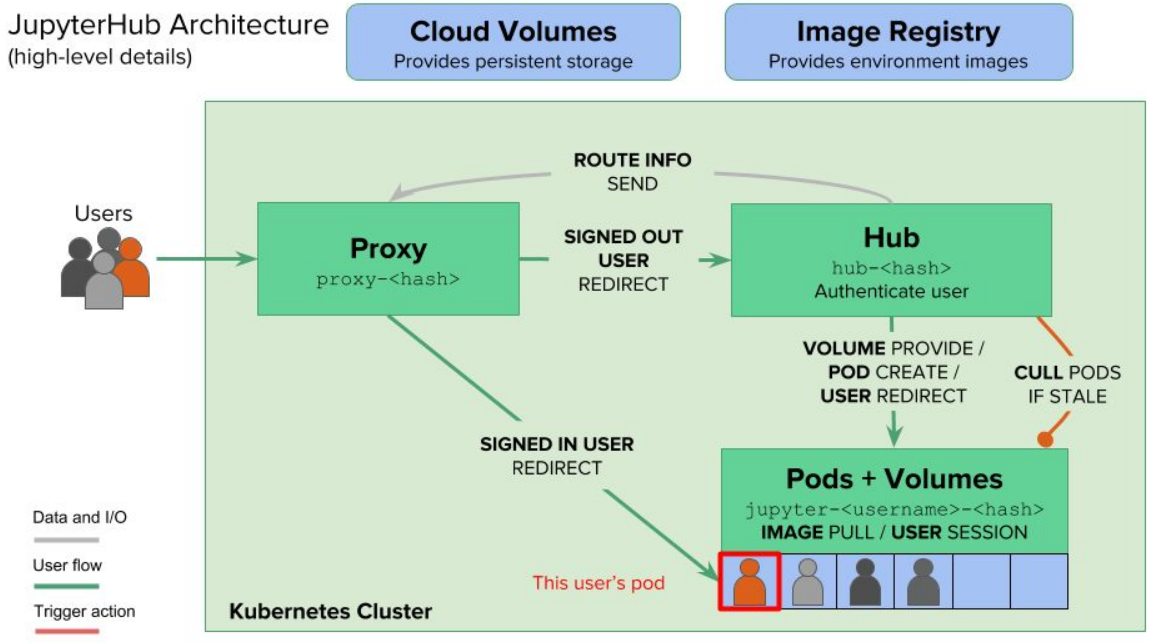
All of this is being *incorporated into a Helm chart* - many rough edges, but it will be portable to other sites

Pod customization hook to create secrets and services **-** pod can expose the Dask scheduler to the outside world and can authenticate with services like HTCondor and XRootD

Integration of XRootD - each pod's unique secret includes and **auto-generated macaroon *authorizing the pod to access files at the site XCache server***

Developed a custom XRootD client plugin enabling whenever the prefix *root://xcache/* is used, hostname is replaced with the correct one for the local site (using environment variables) and token authorization is automatically used & embedded in the URL

Jupyterhub Helm charts (forked from Z2JH)



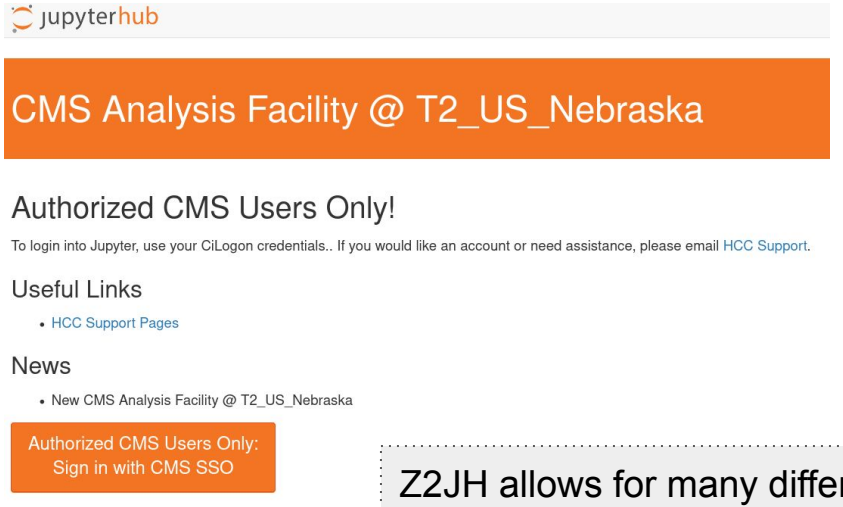

```
[root@cmsaf kubeinstance]# kubectl get pods --namespace cmsaf-prod
NAME                                    READY   STATUS    RESTARTS   AGE
continuous-image-puller-4dsfd           1/1     Running   0          99m
continuous-image-puller-5djdr           1/1     Running   0          99m
continuous-image-puller-6dc64           1/1     Running   0          99m
continuous-image-puller-8gcfl           1/1     Running   0          99m
continuous-image-puller-8pkg4           1/1     Running   0          99m
continuous-image-puller-bmmfx           1/1     Running   0          99m
continuous-image-puller-dvt4h           1/1     Running   0          99m
continuous-image-puller-h99jz           1/1     Running   0          99m
continuous-image-puller-h9g7v           1/1     Running   0          99m
continuous-image-puller-hjp1n           1/1     Running   0          99m
continuous-image-puller-jrqh7           1/1     Running   0          99m
continuous-image-puller-lgnzc           1/1     Running   0          99m
continuous-image-puller-mbfzb           1/1     Running   0          99m
continuous-image-puller-n77lz           1/1     Running   0          99m
continuous-image-puller-rn4jx           1/1     Running   0          99m
continuous-image-puller-tnqlt           1/1     Running   0          99m
continuous-image-puller-v2ghm           1/1     Running   0          99m
continuous-image-puller-vvtlv           1/1     Running   0          99m
continuous-image-puller-xbdwj           1/1     Running   0          99m
continuous-image-puller-zdn7l           1/1     Running   0          99m
hub-79f76c79bc-ndm5t                    1/1     Running   0          98m
jupyter-matousadamec-40gmail-2ecom      2/2     Running   0          11m
jupyter-oksana-2eshadura-40cern-2ech    2/2     Running   0          81m
proxy-7f5ddf7576-8t2vv                  1/1     Running   1          55d
user-scheduler-5954b4765d-4ds6w         1/1     Running   3          55d
user-scheduler-5954b4765d-v7bn7         1/1     Running   9          55d
```

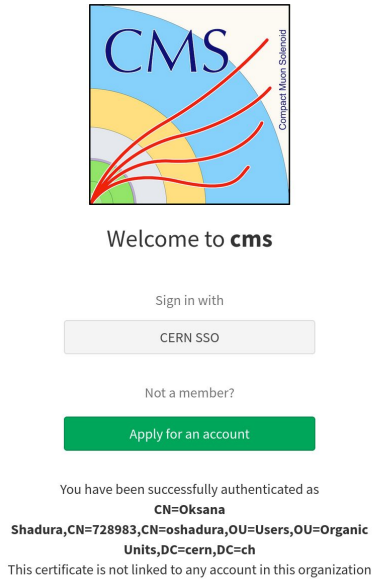
A lot of customisations (CMS-related and not only)!

```
NAME                                    STATUS   VOLUME                                     CAPACITY   ACCESS MODES   STORAGECLASS      AGE
claim-brian-2ebockelman-40cern-2ech     Bound    pvc-dfd8dd0a-2f5f-4a53-a075-8e1e9662840d   10Gi       RWO            rook-ceph-block   42d
claim-clundstedt-40unl-2eedu            Bound    pvc-5e4565cf-c019-4d86-bdd9-e88daa5ff8e6   10Gi       RWO            rook-ceph-block   58d
claim-garhan-2eattebury-40cern-2ech     Bound    pvc-f25ad566-162b-40db-9f54-5486fec0dff0   10Gi       RWO            rook-ceph-block   56d
claim-john-2ethiltges-40cern-2ech       Bound    pvc-dc42b1d6-93d6-45af-9224-dc72b8e57fcb   10Gi       RWO            rook-ceph-block   56d
claim-kenbloom-40unl-2eedu              Bound    pvc-3eef8c17-e9b7-44b1-bd23-403fcefab801   10Gi       RWO            rook-ceph-block   7d22h
claim-matousadamec-40gmail-2ecom        Bound    pvc-e060c809-3153-4384-b62b-70a3f4a2bc8e   10Gi       RWO            rook-ceph-block   55d
claim-oksana-2eshadura-40cern-2ech      Bound    pvc-0a63e8fe-5f77-47e8-833a-bc7da8c34747   10Gi       RWO            rook-ceph-block   44d
hub-db-dir                              Bound    pvc-b65d0477-e72a-4529-a549-d8d3e98e194e   1Gi        RWO            rook-ceph-block   55d
```

Authentication@Coffea-Casa

```
auth:
  type: custom
  custom:
    className: oauthenticator.generic.GenericOAuthenticator
    ……
```



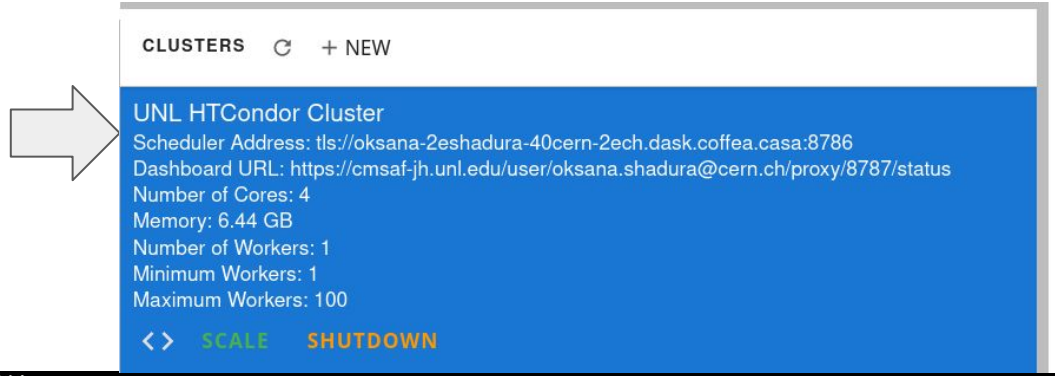


Z2JH allows for many different, standard SSO solutions and it should be fairly easy for any experiment to plugin their SSO solution (or do user/password management if they desire).

Traefik@Coffea-Casa - modern HTTP reverse proxy and load balancer that makes deploying microservices easy

For *Coffea-Casa*:

- Allow hub to add DNS entries to traefik service in traefik namespace for each user:
  - For Dask scheduler;
  - For Dask workers.



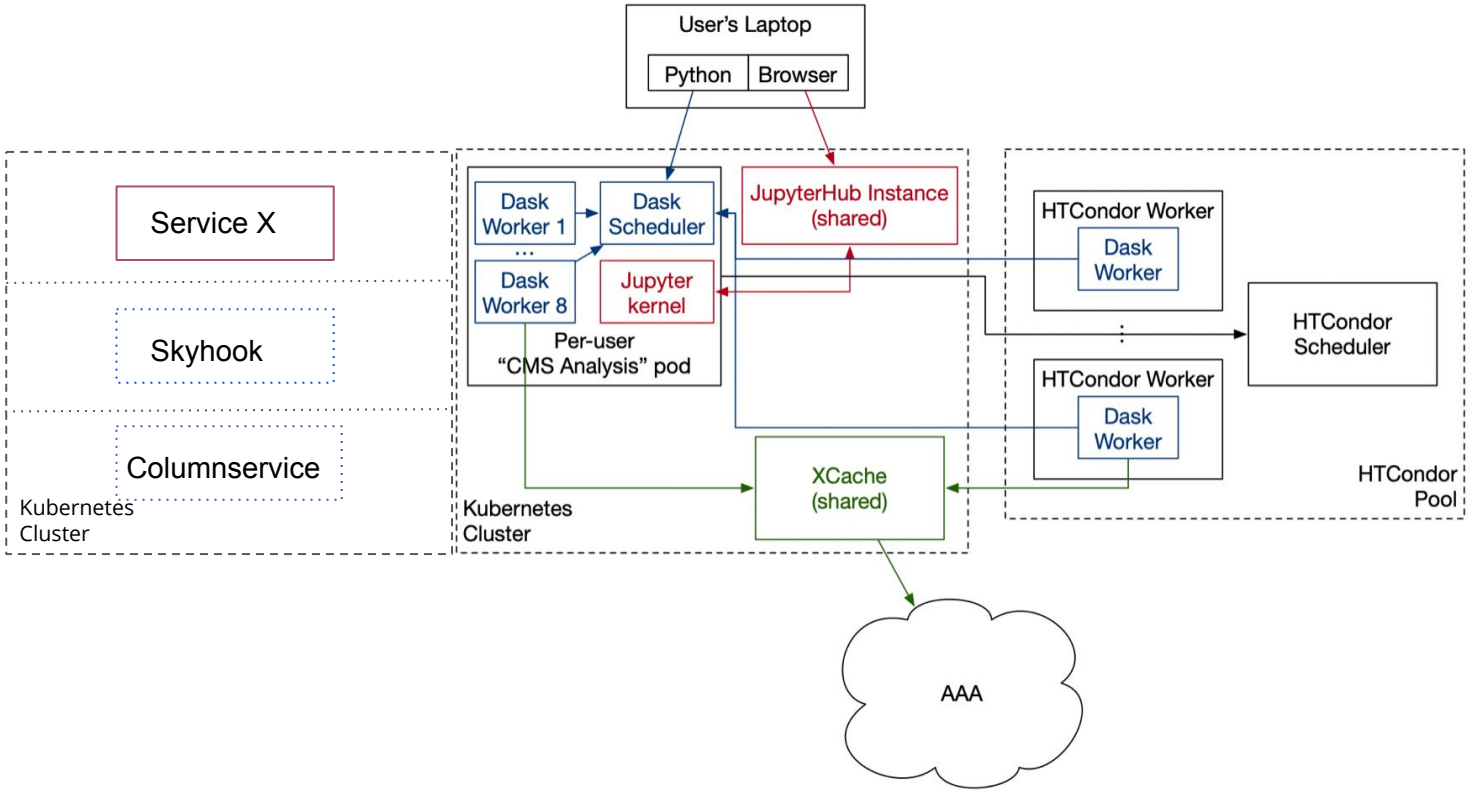

```
[root@cmsaf kubeinstance]# kubectl get svc -n traefik
NAME      TYPE           CLUSTER-IP      EXTERNAL-IP     PORT(S)                                                    AGE
traefik   LoadBalancer   10.105.250.20   129.93.183.29   80:31125/TCP,443:30886/TCP,8786:31237/TCP,8788:31039/TCP   56d
```

```
[root@cmsaf kubeinstance]# kubectl get svc -n cmsaf-prod
NAME                                        TYPE           CLUSTER-IP       EXTERNAL-IP     PORT(S)                      AGE
brian-2ebockelman-40cern-2ech-dask-service  ClusterIP      10.111.6.103     <none>          8786/TCP,8788/TCP            42d
clundstedt-40unl-2eedu-dask-service         ClusterIP      10.102.180.145   <none>          8786/TCP,8788/TCP            58d
garhan-2eattebury-40cern-2ech-dask-service  ClusterIP      10.101.225.186   <none>                                       56d
hub                                         ClusterIP      10.110.64.41     <none>                                       55d
john-2ethiltges-40cern-2ech-dask-service    ClusterIP      10.111.43.146    <none>                                       55d
kenbloom-40unl-2eedu-dask-service           ClusterIP      10.105.180.87    <none>                                       7d22h
matousadamec-40gmail-2ecom-dask-service     ClusterIP      10.96.160.93     <none>          8786/TCP,8788/TCP            55d
oksana-2eshadura-40cern-2ech-dask-service   ClusterIP      10.101.192.253   <none>          8786/TCP,8788/TCP            57d
proxy-api                                   ClusterIP      10.104.84.33     <none>          8001/TCP                     55d
proxy-public                                LoadBalancer   10.104.108.146   129.93.183.32   443:30388/TCP,80:30837/TCP   55d
```

No need to maintain external IPs

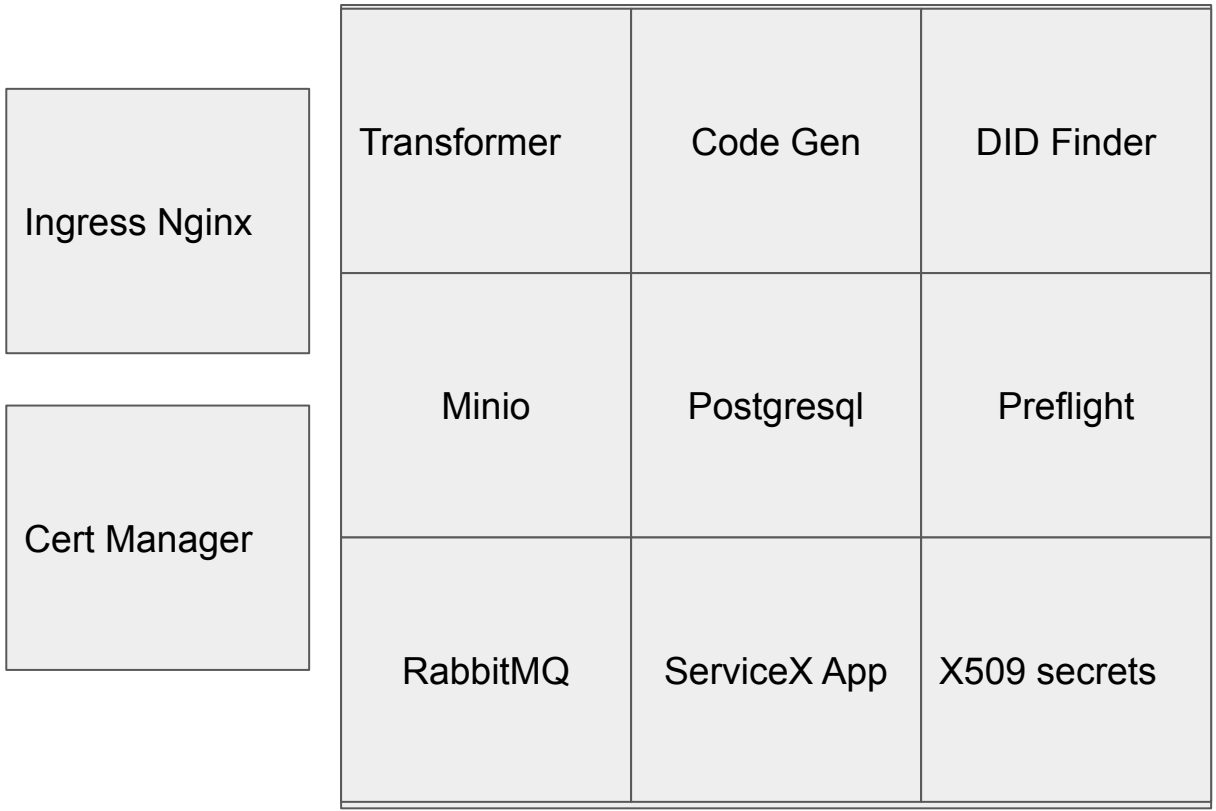
Next step is to add a generic DNS management at Coffea-Casa Helm Charts (easily adaptable to any cluster/grid site perspectives)

Analysis Facility @ T2 Nebraska



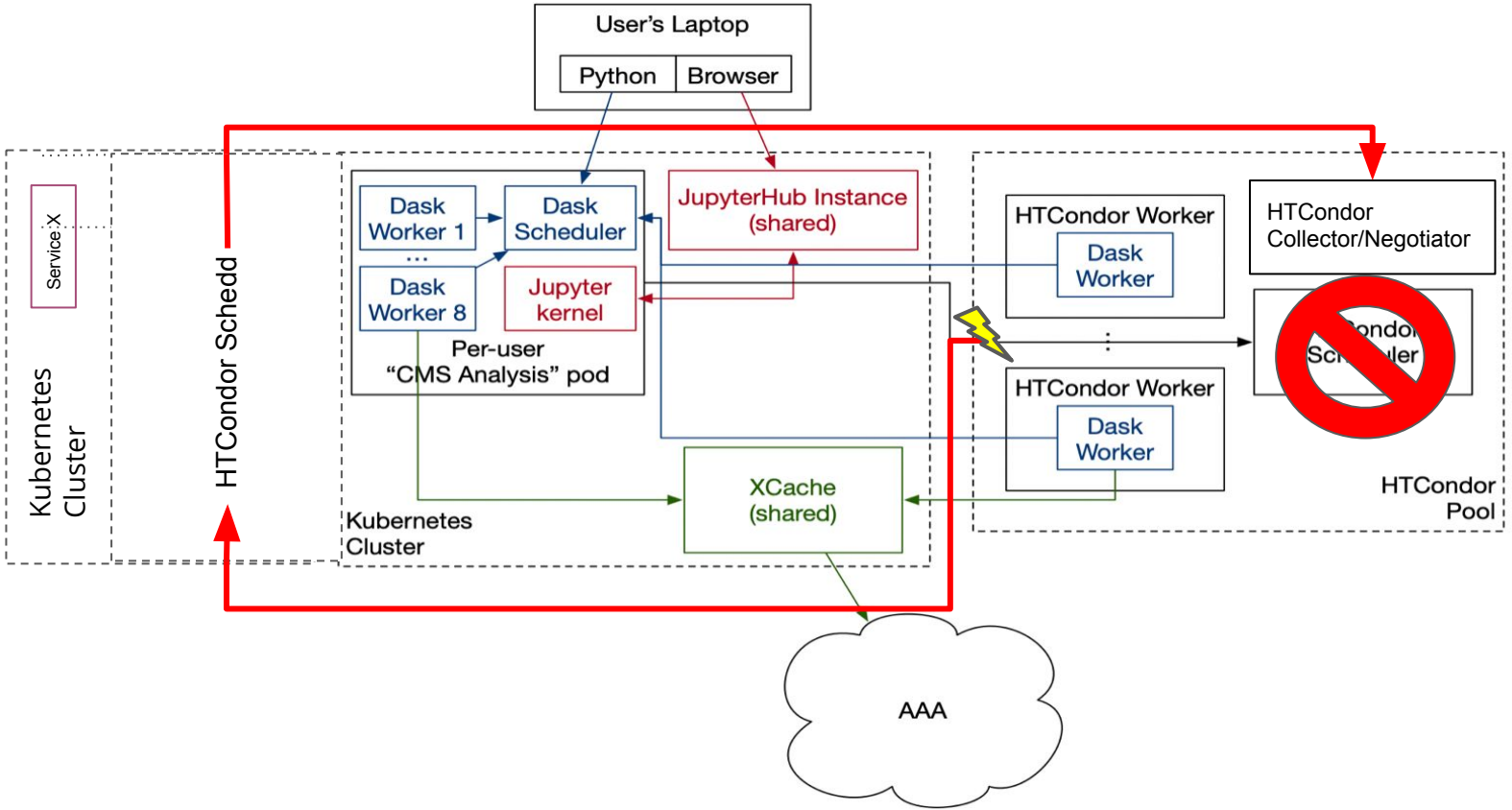
We are looking for the volunteers (*other sites*) to try our developments!

ServiceX (RC3)



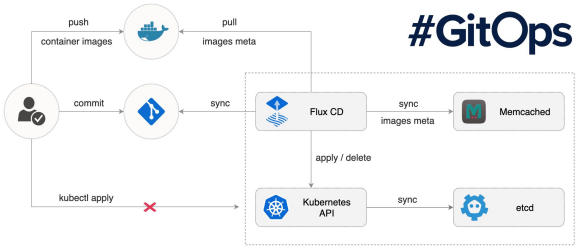
Next step is to test opt-out of ServiceX user management system for Coffea-Casa...

WIP: HTCondor Dedicated Schedd Integration

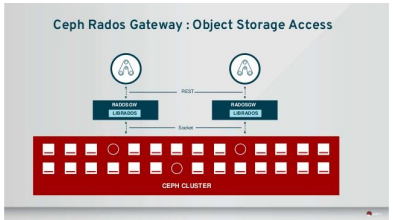


Other K8s Coffea-Casa Plans

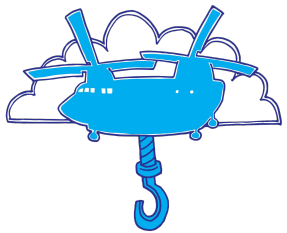
- Add Flux CD for Coffea-casa Helm Charts at UNL



- Adopt ServiceX to use Rados Gateway (RGW) instead of default Minio



- Integrate SkyHook at Coffea-Casa (via Rook.io)



Coffea-Casa Timeline

- ***Q4 2020 -*** *Invite first users to test “alpha” version of UNL AF (“coffea-casa”)*
- ***Q4 2020 -*** *Make “coffea-casa” products (Helm charts, modules) deployable in any other AF facility*
    - *Expected first test deployment of **FNAL Elastic AF** during 2021*
- ***Q4 2020 -*** *Finalize testing of ServiceX@UNL AF*
- ***Q1 2021 -*** *Deploy and test data delivery with Skyhook at UNL AF*