Coffea-Casa Analysis Facility at the University of Nebraska-Lincoln

HEPiX Fall 2023

Garhan Attebury, Carl Lundstedt, John Thiltges
Oksana Shadura, Andrew Wightman, Sam Albin, Brian Bockelman

Holland Computing Center @ UNL in brief

Red (USCMS Tier2 cluster)

- ~372 nodes, 16k job slots (threads)
- 10PB Ceph storage

Swan (HPC cluster)

- ~340 nodes, 16k cores
- Mellanox IB
- 5.4PB lustre on ZFS

Anvil (Openstack cloud)

- 76x 20 core 256GB nodes
- 349TB Ceph storage (Jewel! Yikes!)

Flatiron (CMS Analysis Facility)

- 832 cores + V100 + 2x P100 GPUs
- ~100TB Ceph on NVMe

Common (shared storage)

- 1.9PB BeeGFS on ZFS

Attic (archival storage)

- 2PB ZFS

External projects

- NRP
  - 3072 cores
  - 195 GPUs
  - 14TB RAM
- PATh
  - 1.5k cores
  - 1PB storage

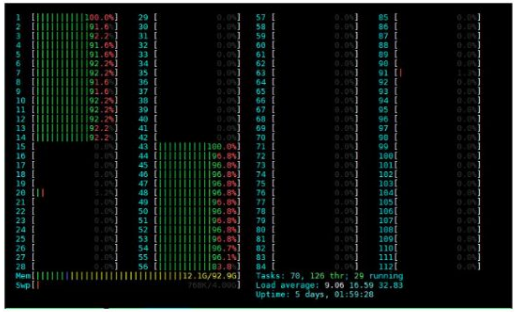
3x datacenters

600 kW in use

Analysis Facilities (from a non physicist PoV)

How the physicists see “Analysis Facilities”

Homelab (https://domalab.com)



“Analysis Facilities” could be any type of resource from laptop to Tier-2

HEP data access	Number of cores to scale	Recipe how to run code	Disk space	Favorite analysis frameworks available

Slide shamelessly stolen from Brian Bockelman
26th International Conference on Computing in High Energy & Nuclear Physics

Analysis Facilities:
What the physicist users dream about

- **Quick interactive analysis turnaround:** *"I want to get my preliminary plots to be ready over coffee break"*
- **User improvement experiences (UX):** let's help physicists focus on the physics
- **Methods for efficient data scaling, caching at AFs:** more challenges with data-intensive analysis pipeline
- **Data reusability:** AF should support extraction of user defined experiment data formats to migrate them onto other facility, laptops or workstations at home institutions or at home

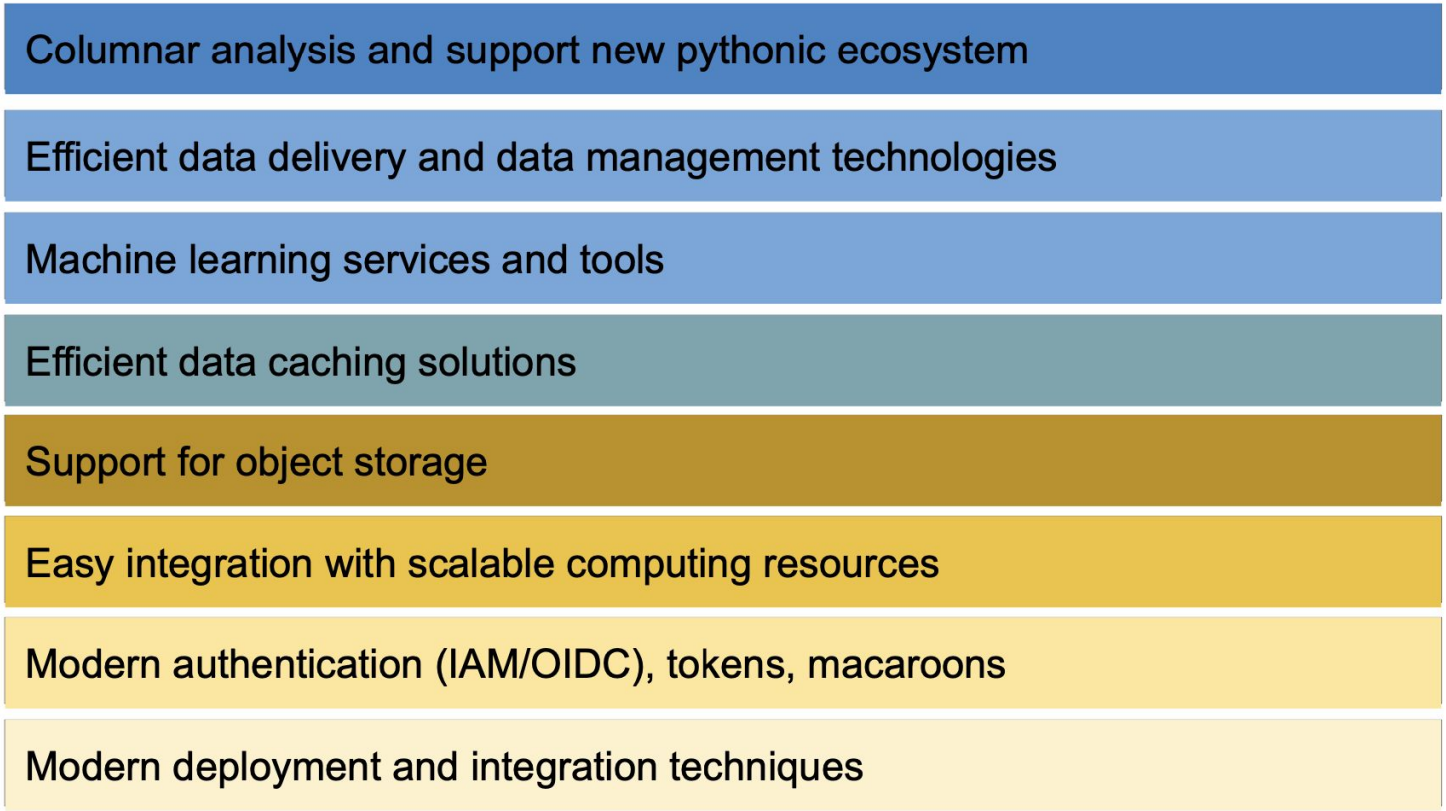
Slide shamelessly stolen from Brian Bockelman

Analysis Facilities:
What the resource managers dream about

- **Easy deployment and reproducible setup:** Kubernetes can help to facilitate an easy AF deployment within Tier-X facilities (e.g. co-locating next to existing computing resources)
- **Modularity:** Kubernetes is ideal for demanding applications that require complex configurations (focusing on modular orchestration)
- **“Self-healing”:** easy rollback with Kubernetes

Slide shamelessly stolen from Brian Bockelman
26th International Conference on Computing in High Energy & Nuclear Physics

Building blocks for analysis facilities



What we built: Flatiron

12x Dell R750

- 2x Xeon 6348 (56C/112T)
- 512GB RAM
- 10x 3.2TB NVMe drives
- 2x 100Gb

2x Dell S5232F-ON switches

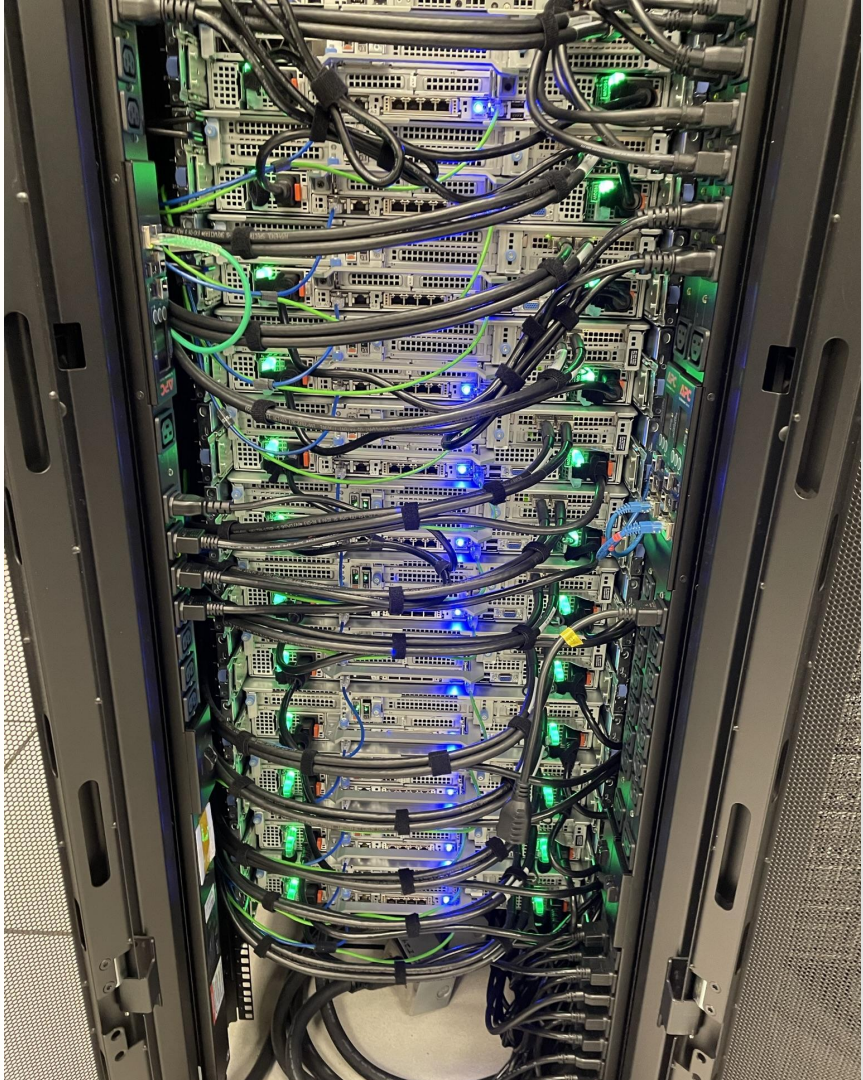
- 200+Gbps path to Internet2/ESnet

Bonus GPUs

- 1x NVIDIA V100S
- 2x NVIDIA P100

Bonus CPUs

16x 8C i7 3.0GHz desktops ~~w/GTX 980s~~

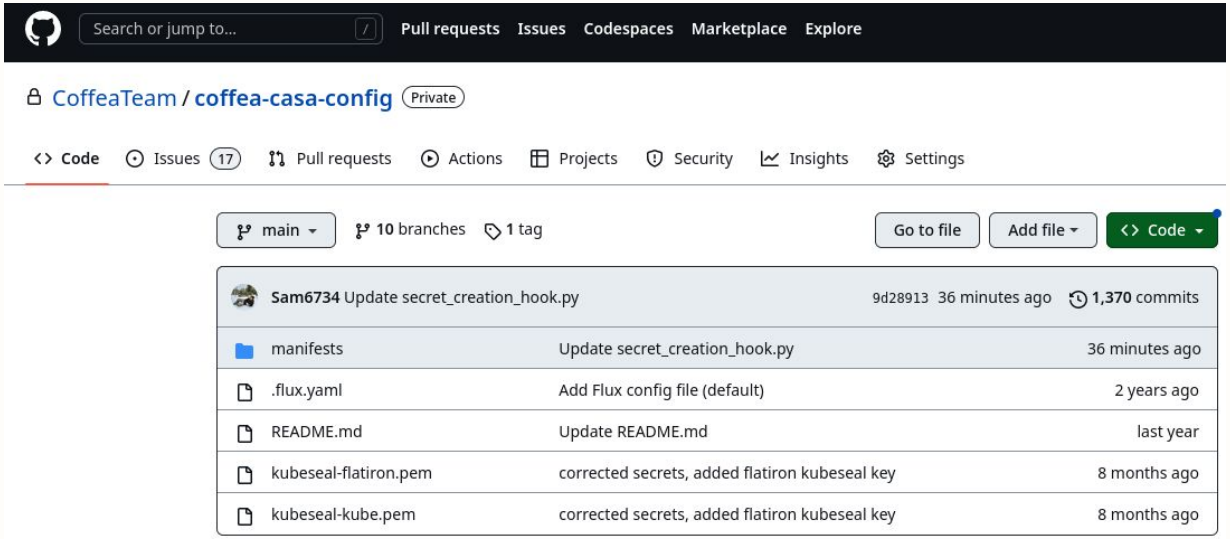


Flatiron Kubernetes Cluster

- AlmaLinux 8.7
- Kubernetes 1.27.6
- 3x master/etcd nodes, keepalived + HAProxy
- Calico + BGP and MetalLB
- Ceph via Rook.io (106TB 3x replica NVMe, 5x OSD per drive)
- Ceph via Skyhook (9TB testing)
- Dex + CILogon for API access
- Jupyter auth : OpenID Connect (OIDC)
  - CMS with CERN IAM
  - CILogon with COmanage for Opendata facility
- Traefik, Flux, Git, CVMFS, Dask, etc...

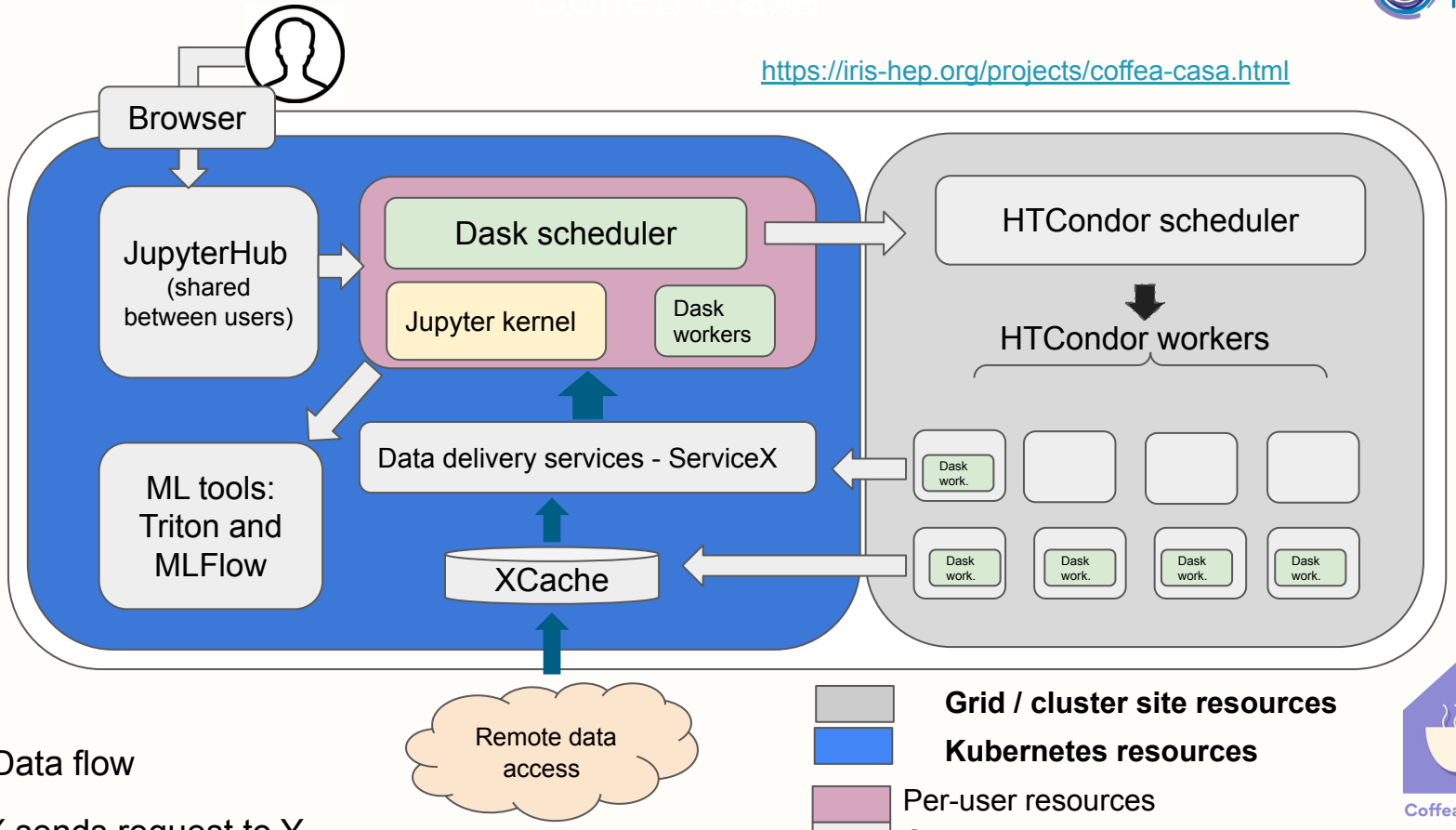
Casa Infrastructure & Management

- Configs for casa are kept in GIT
- Changes follow gitops techniques
- Changes are applied in-situ via a Flux agent



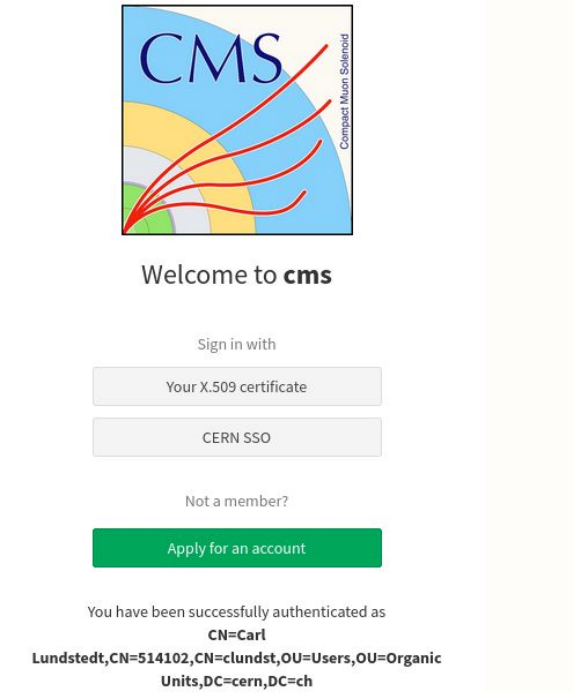
Slide shamelessly stolen from Carl Lundstedt

https://iris-hep.org/projects/coffea-casa.html



Slide shamelessly stolen from Carl Lundstedt

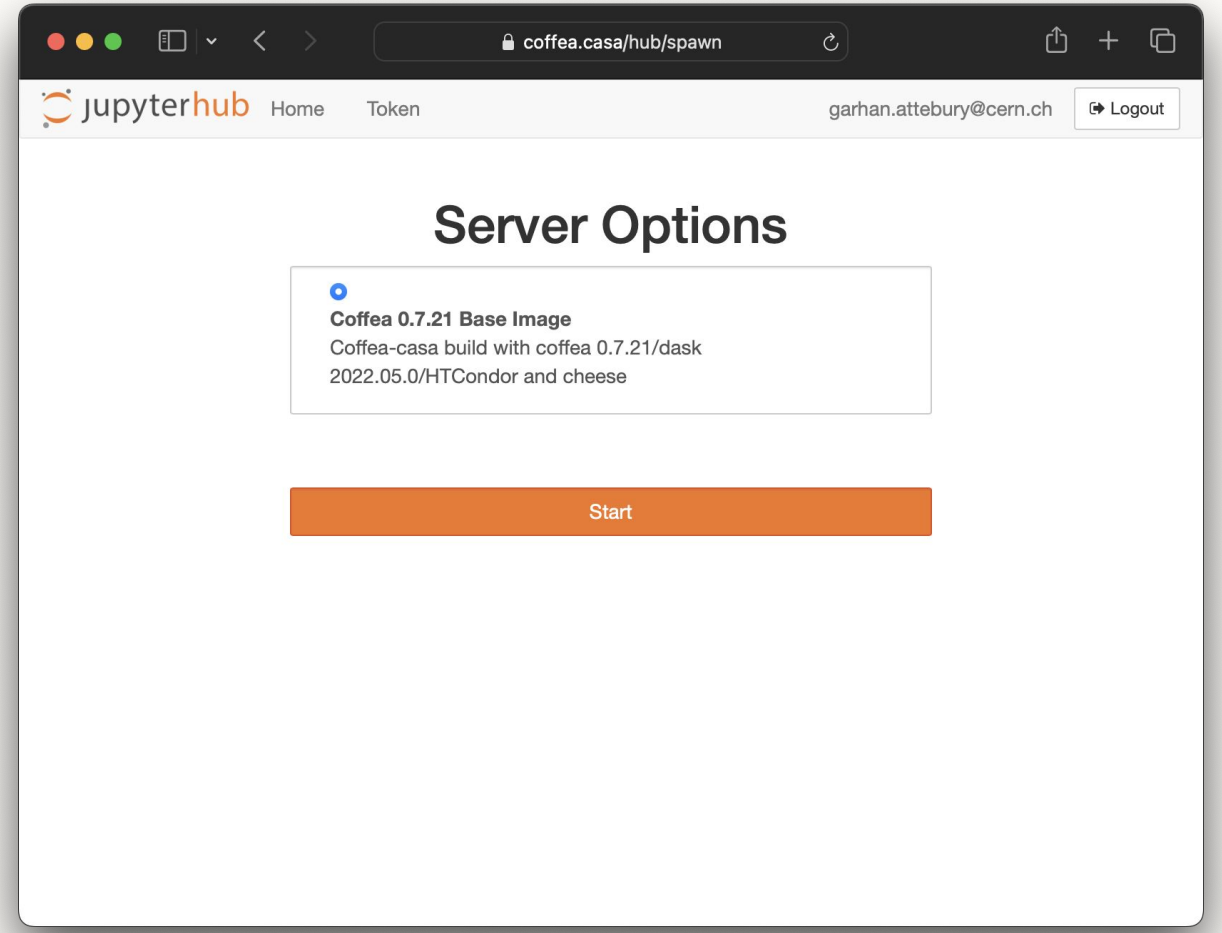
The Four Casa Instances

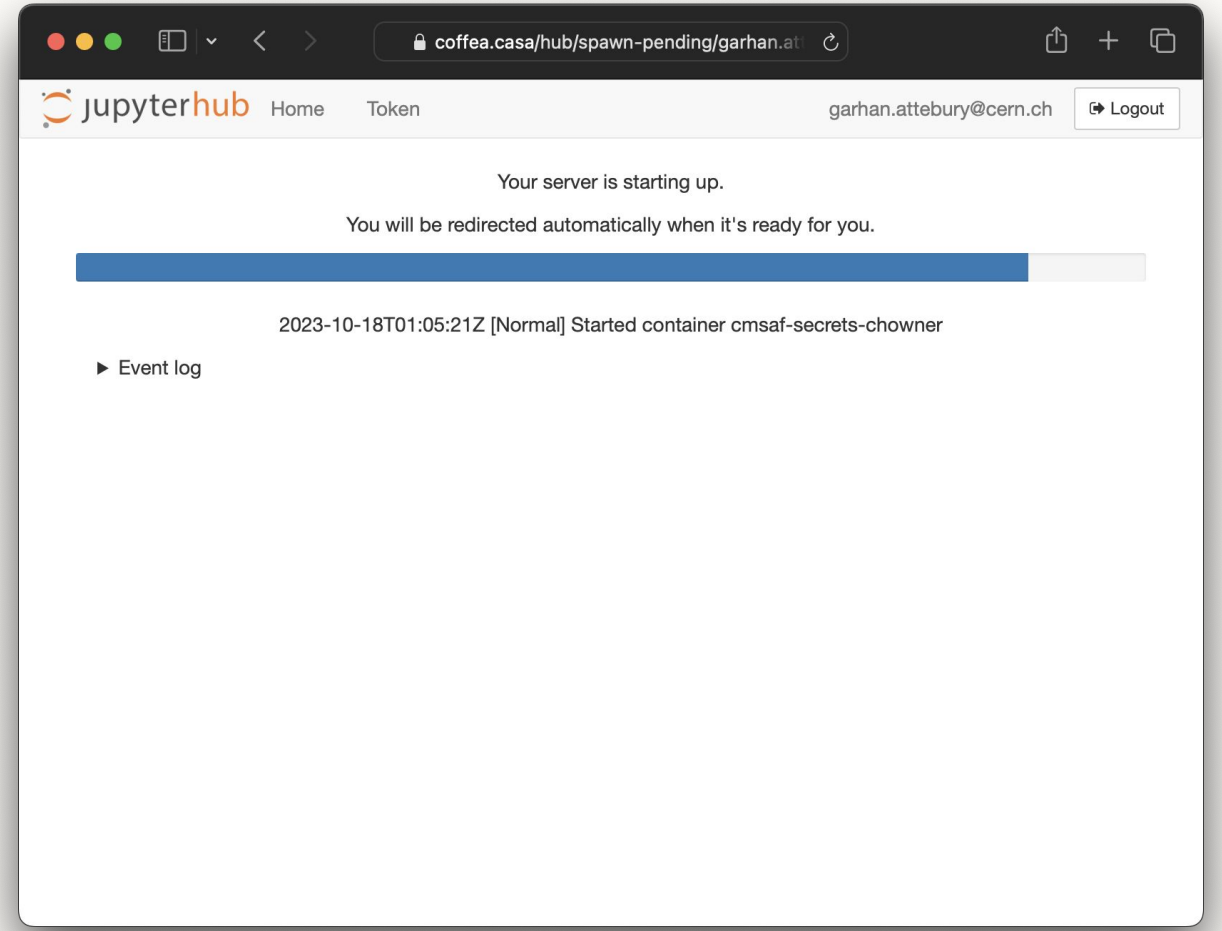


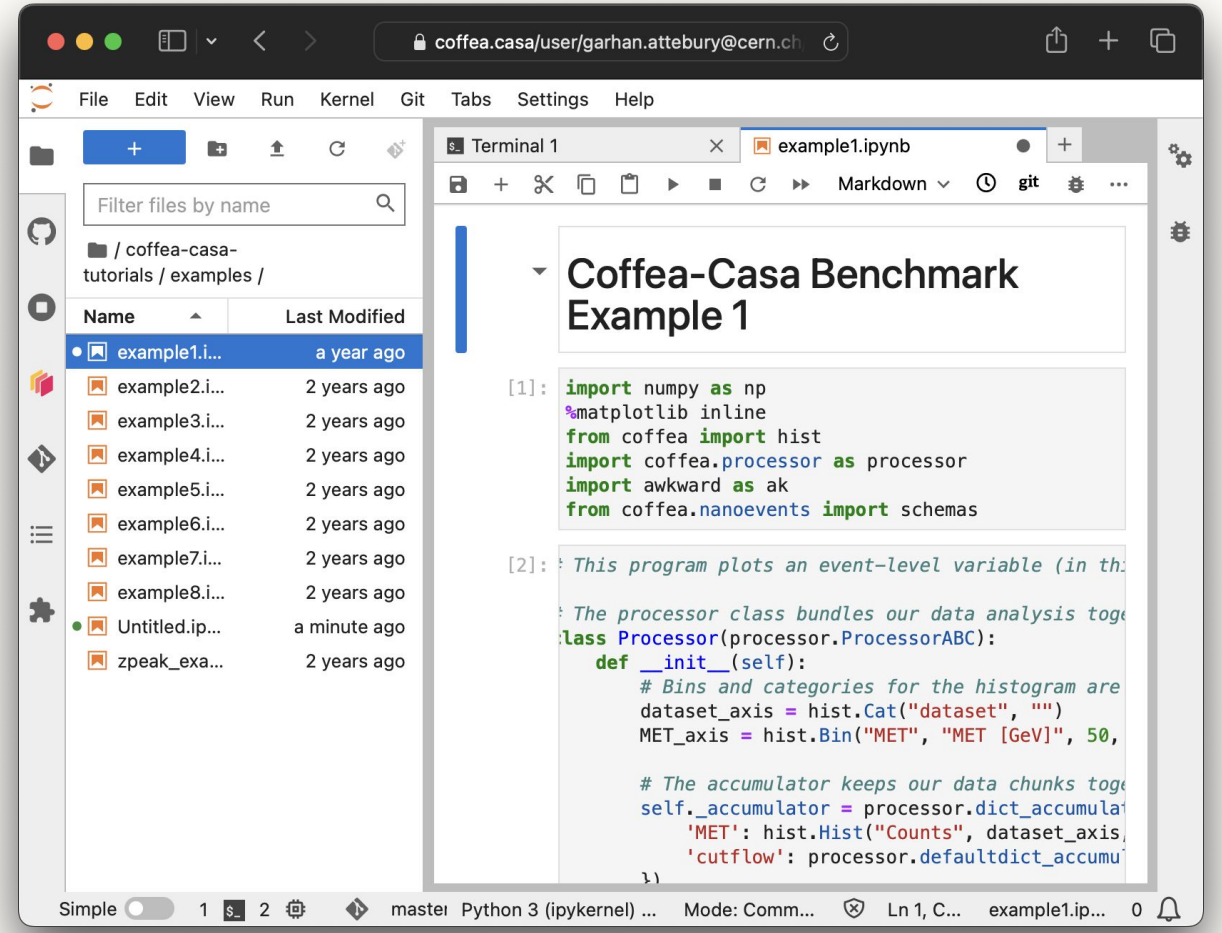
- CMS-Prod (https://coffea.casa)
- CMS-Dev

- Opendata-Prod (https://coffea-opendata.casa)
- Opendata-Dev

Slide shamelessly stolen from Carl Lundstedt


coffea.casa/hub/spawn
jupyterhub
Home
Token
garhan.attebury@cern.ch
Logout
Server Options
Coffea 0.7.21 Base Image
Coffea-casa build with coffea 0.7.21/dask
2022.05.0/HTCondor and cheese
Start


coffea.casa/hub/spawn-pending/garhan.att
jupyterhub
Home
Token
garhan.attebury@cern.ch
Logout
Your server is starting up.
You will be redirected automatically when it's ready for you.
2023-10-18T01:05:21Z [Normal] Started container cmsaf-secrets-chowner
Event log


coffea.casa/user/garhan.attebury@cern.ch
File Edit View Run Kernel Git Tabs Settings Help
Filter files by name
/ coffea-casa-tutorials / examples /
Name
Last Modified
example1.i... a year ago
example2.i... 2 years ago
example3.i... 2 years ago
example4.i... 2 years ago
example5.i... 2 years ago
example6.i... 2 years ago
example7.i... 2 years ago
example8.i... 2 years ago
Untitled.ip... a minute ago
zpeak_exa... 2 years ago
Terminal 1
example1.ipynb
Markdown
git
Coffea-Casa Benchmark Example 1
[1]:
import numpy as np
%matplotlib inline
from coffea import hist
import coffea.processor as processor
import awkward as ak
from coffea.nanoevents import schemas
[2]:
# This program plots an event-level variable (in th
# The processor class bundles our data analysis toge
class Processor(processor.ProcessorABC):
def __init__(self):
# Bins and categories for the histogram are
dataset_axis = hist.Cat("dataset", "")
MET_axis = hist.Bin("MET", "MET [GeV]", 50,
# The accumulator keeps our data chunks toge
self._accumulator = processor.dict_accumula
'MET': hist.Hist("Counts", dataset_axis,
'cutflow': processor.defaultdict_accumu
})
Simple 1 2 master Python 3 (ipykernel) ... Mode: Comm... Ln 1, C... example1.ip... 0

Workflow Scale Out

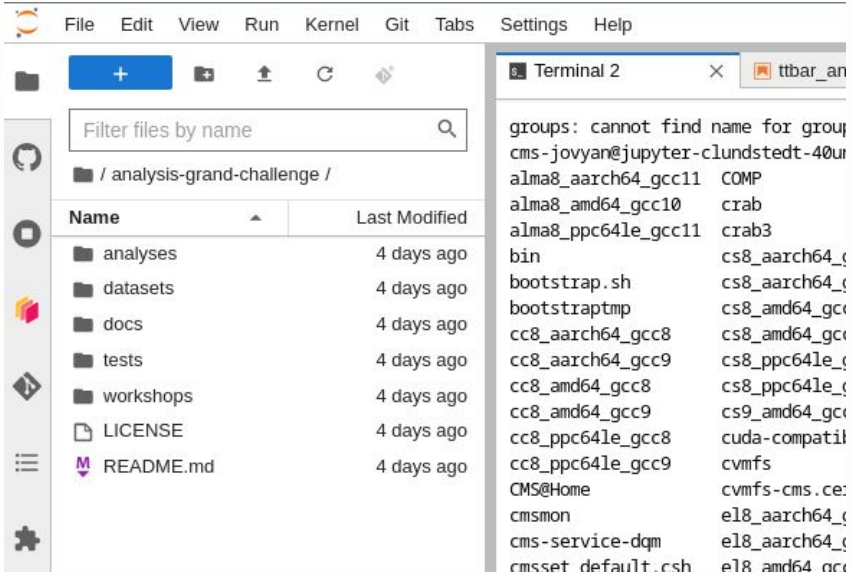
Scale out is accomplished with a custom Dask-Jobqueue Class that deploys Dask worker nodes in either our T2 resource or in an condor cluster running inside the Flatiron kubernetes.

name	address	nthreads	cpu	memory	limit	memory %	managed	unmanaged o	unmanaged	d	# fds
Total (26)		52	11 %	9.9 GiB	150.1 GiB	6.6 %	70.7 KiB	1.8 GiB	8.2 GiB	0.0	974
htcondor--187ç	tls://red-c7123.unl.edu:33963	2	12 %	395.1 MiB	5.7 GiB	6.7 %	2.6 KiB				
htcondor--187ç	tls://red-c7124.unl.edu:37153	2	12 %	393.1 MiB	5.7 GiB	6.7 %	2.6 KiB				
htcondor--187ç	tls://red-c7125.unl.edu:34933	2	10 %	402.2 MiB	5.7 GiB	6.9 %	2.6 KiB				
htcondor--187ç	tls://red-c7122.unl.edu:45901	2	14 %	359.6 MiB	5.7 GiB	6.1 %	2.6 KiB	55.0 MiB	304.6 MiB	0.0	37
htcondor--187ç	tls://red-c7127.unl.edu:45173	2	8 %	399.3 MiB	5.7 GiB	6.8 %	2.6 KiB	55.6 MiB	343.7 MiB	0.0	37
htcondor--187ç	tls://red-c7126.unl.edu:36439	2	14 %	362.5 MiB	5.7 GiB	6.2 %	2.6 KiB	55.7 MiB	306.8 MiB	0.0	37
htcondor--187ç	tls://red-c7123.unl.edu:38615	2	14 %	400.6 MiB	5.7 GiB	6.8 %	2.6 KiB	57.8 MiB	342.8 MiB	0.0	37
htcondor--187ç	tls://red-c7124.unl.edu:40741	2	12 %	398.5 MiB	5.7 GiB	6.8 %	2.6 KiB	55.7 MiB	342.8 MiB	0.0	36

Slide shamelessly stolen from Carl Lundstedt

Storage & Data Access

- Each user give 10GB of persistent storage on login
- XCache via Tokens issued at login
- cern.ch CVMFS mounted in the user pods
- dasgoclient / rucio / EOS access
- User's T2 */store/user* mounted in the user pod



Slide shamelessly stolen from Carl Lundstedt

Triton Inference Service

- To leverage the few GPUs we have an inference service is deployed
- Training sets are able to be stored in an S3 bucket deployed for just this task.

s3://rook-ceph-rgw-my-store.rook-ceph.svc:80/triton-c9adf042-ffb8-4221-bd42-e385efb1d0e2

```
I0426 14:13:43.918751 1 metrics.cc:650] Collecting metrics for GPU 0: Tesla V100S-PCIE-32GB
I0426 14:13:43.918929 1 tritonserver.cc:2214]
+----------------------------------+------------------------------------------------------------------------------------
| Option                           | Value
+----------------------------------+------------------------------------------------------------------------------------
| server_id                        | triton
| server_version                   | 2.25.0
| server_extensions                | classification sequence model_repository model_repository(unload_dependents) schedule_policy model
| model_repository_path[0]         | s3://rook-ceph-rgw-my-store.rook-ceph.svc:80/triton-c9adf042-ffb8-4221-bd42-e385efb1d0e2
| model_control_mode               | MODE_EXPLICIT
| startup_models_0                 | *
| strict_model_config              | 0
| rate_limit                       | OFF
| pinned_memory_pool_byte_size     | 268435456
| cuda_memory_pool_byte_size{0}    | 67108864
| response_cache_byte_size         | 0
| min_supported_compute_capability | 6.0
| strict_readiness                 | 0
| exit_timeout                     | 30
+----------------------------------+------------------------------------------------------------------------------------
```


Slide shamelessly stolen from Carl Lundstedt

How it's going...

"We have a Tier-2 cluster!"

"Great! Lets do kubernetes too."

"We now have a Tier-2 and kubernetes!"

"Great! Lets integrate the two."

"We now have a Tier-2, k8s, and a boatload of technical debt..."

Ok, not that bad. We have other kubernetes clusters for other projects. Enough expertise locally that we're not *not* going to do it.

Wishlists...

- Dask workers as HTCondor jobs on our Tier-2 (done)
- Dask workers with native access to the Ceph storage on our Tier-2 (technically possible, but *no*)
- Skyhook on the Tier-2 Ceph storage (custom images, technically possible, but again *no*)

- All roads lead to → k8s it seems
- USCMS Tier2 native in Kubernetes? Why not?

Future plans

- Continue iterating to make user experience more pleasant
  - CVMFS available within notebooks (exists)
  - EOS access and dasgoclient/rucio clients (exists)
  - Custom image support via Binderhub (not yet)
- Performance, especially latency
  - Ideally not on 13 year old hardware this time.
  - GPUs (more, and MIG-able ones would be nice)
- Reproducibility at other sites
- Expanded / refreshed / new kubernetes clusters (Cilium?)
- IPv6 (sorry Dave!)
  - Tier-2 Ceph and k8s are v4 only at present

- Rather than integration of Flatiron with Tier-2, recreate Tier-2 natively within Flatiron

… and now to tear it all down

Questions?

Thanks to the others involved: Carl Lundstedt, John Thiltges
Oksana Shadura, Andrew Wightman, Sam Albin, Brian Bockelman

https://iris-hep.org/projects/coffea-casa.html

Coffea-Casa: building composable analysis facilities for the HL-LHC
Presentation from Brian Bockelman at CHEP 2023

AGC Analysis Facility Summary Presentation
by Carl Lundstedt at AGC Workshop, May 4, 2023