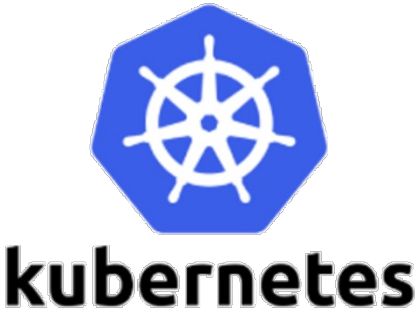


AI and Scientific Research Computing with Kubernetes

Software Stack, Applications, and Services

A tutorial at PEARC24

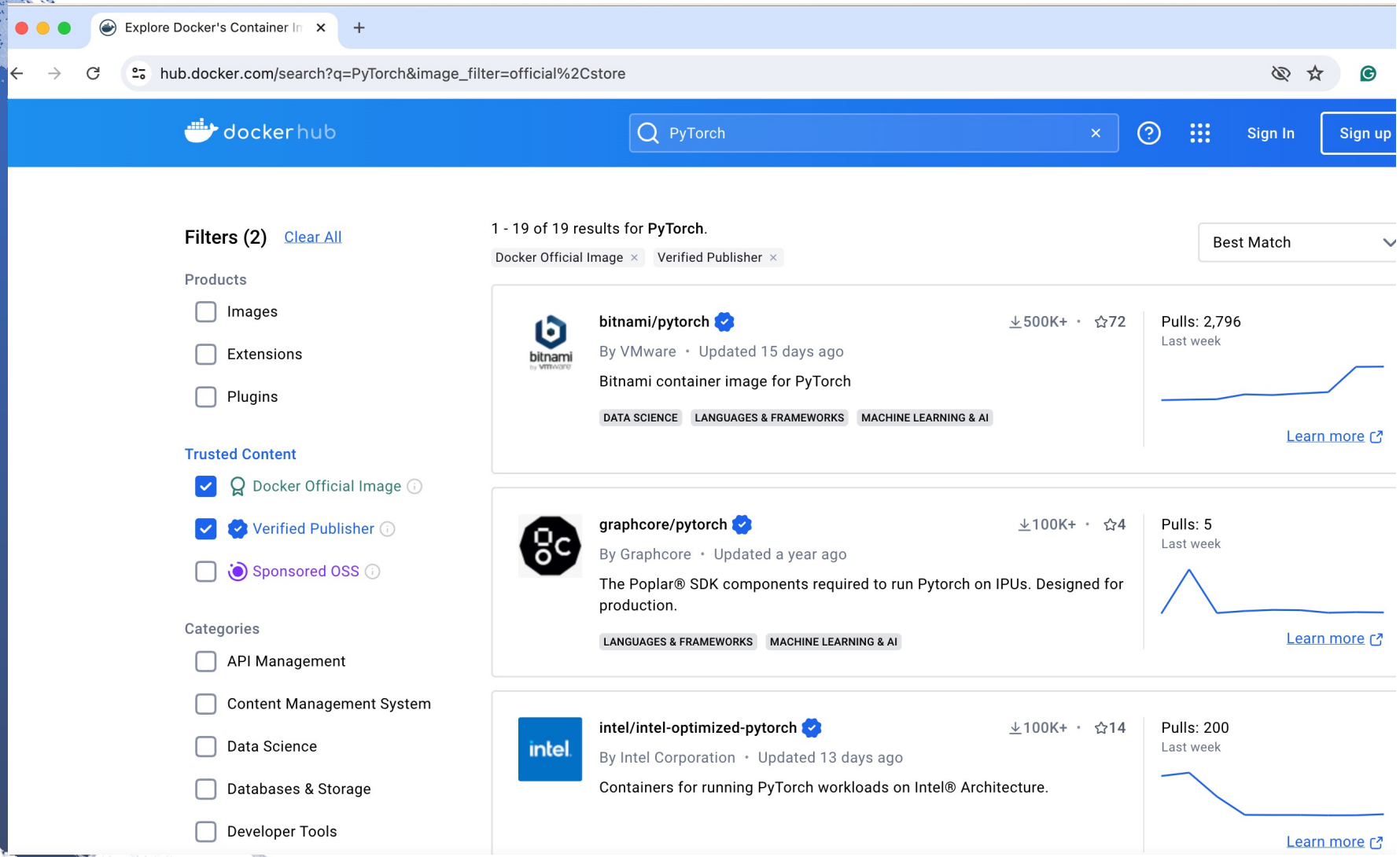
July 22, Providence, Rhode Island

Presented by Mahidhar Tatineni and Dmitry Mishin
University of California San Diego – San Diego Supercomputer Center

Ref: Tutorials at PEARC, SC, 5NRP by Igor Sfiligoi, Dmitry Mishin, and Mahidhar Tatineni

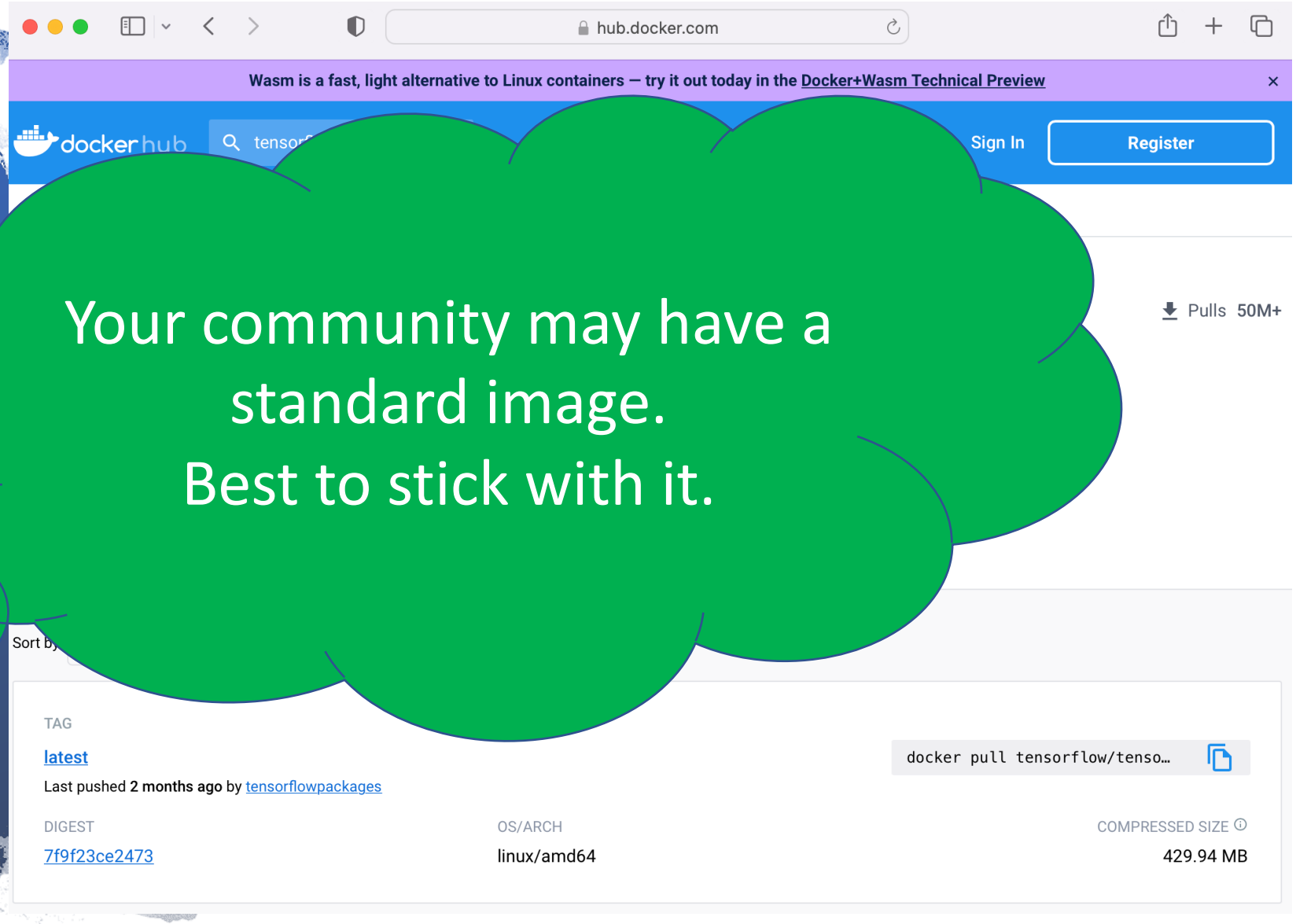
https://hub.docker.com/

Find your container images

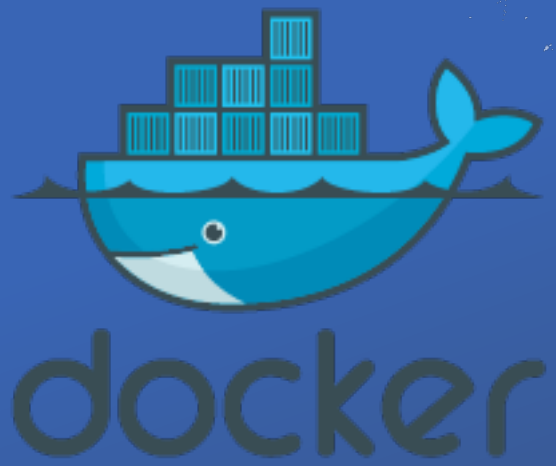


https://hub.docker.com/

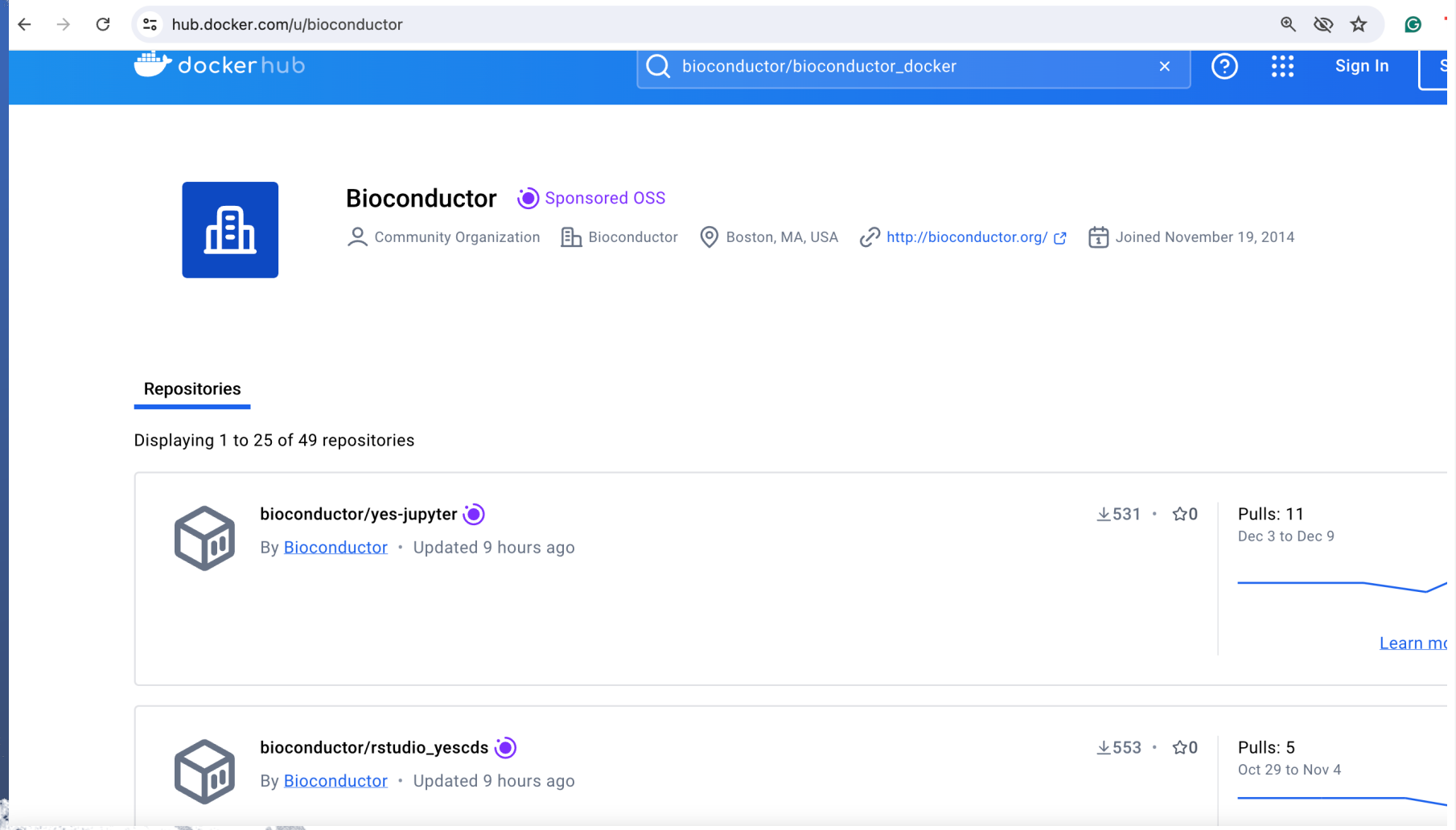
Find your container images



https://hub.docker.com/

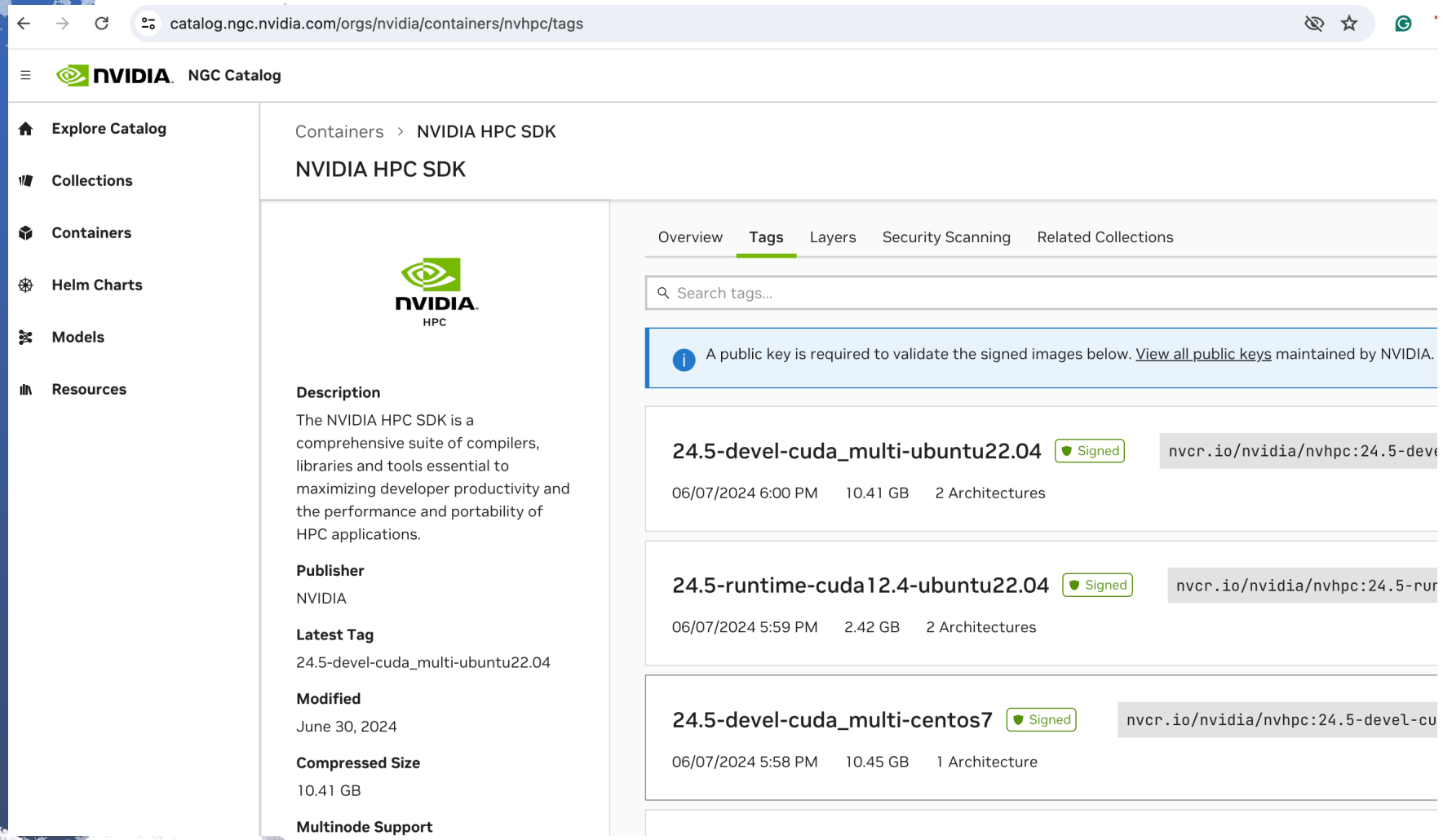


Community maintained images



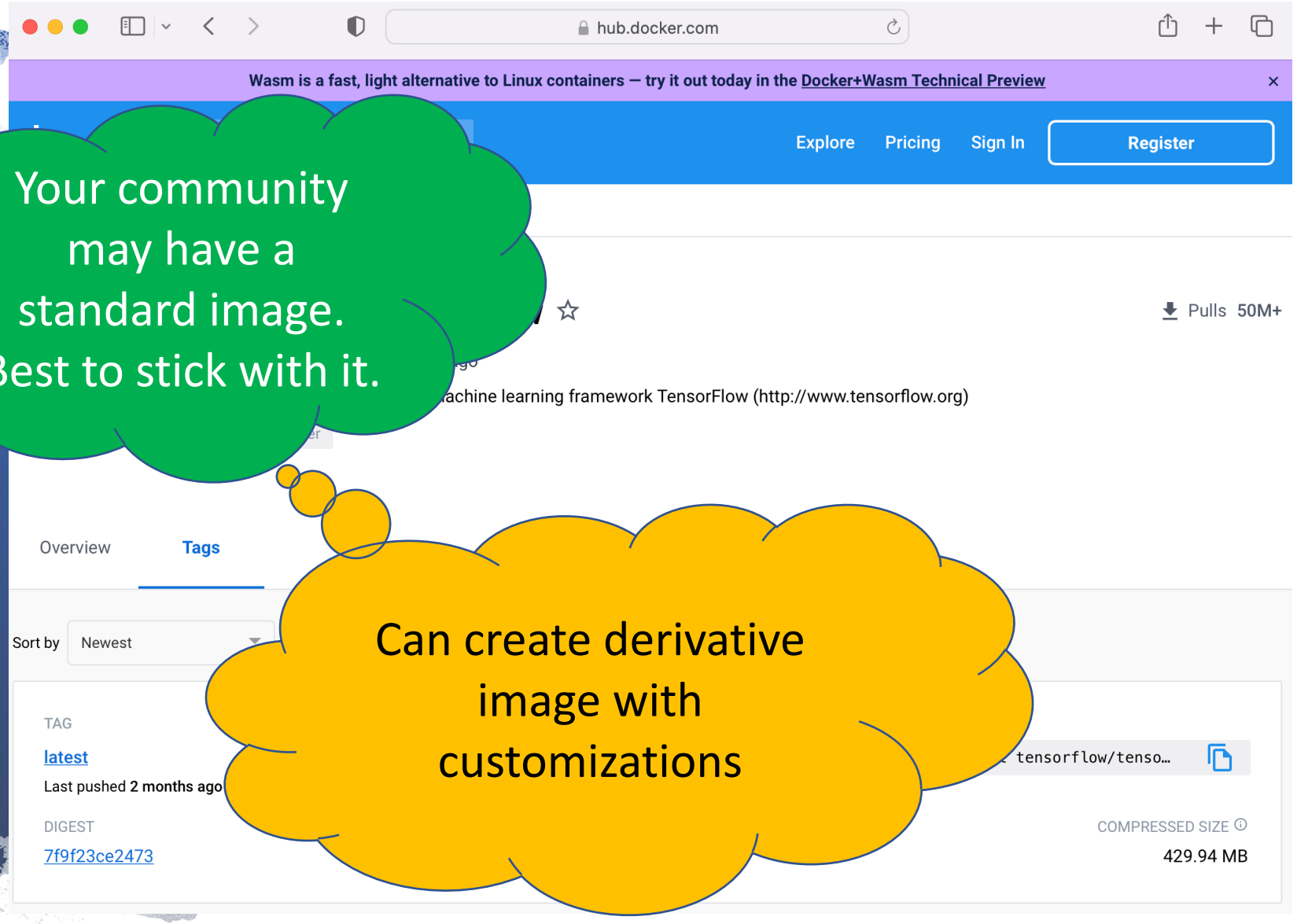
Vendor optimized and signed images

https://catalog.ngc.nvidia.com/
(Look for public images)



https://hub.docker.com/

Find your container images



Your community may have a standard image. Best to stick with it.

Can create derivative image with customizations

Explore the rich ecosystem

Kubernetes is quite powerful by itself

- But it is the surrounding ecosystem that makes it so valuable!

Kubernetes has become an industry standard

- Many projects build on top of it to deliver additional capabilities

Helm package manager

Helm has become the main package manager

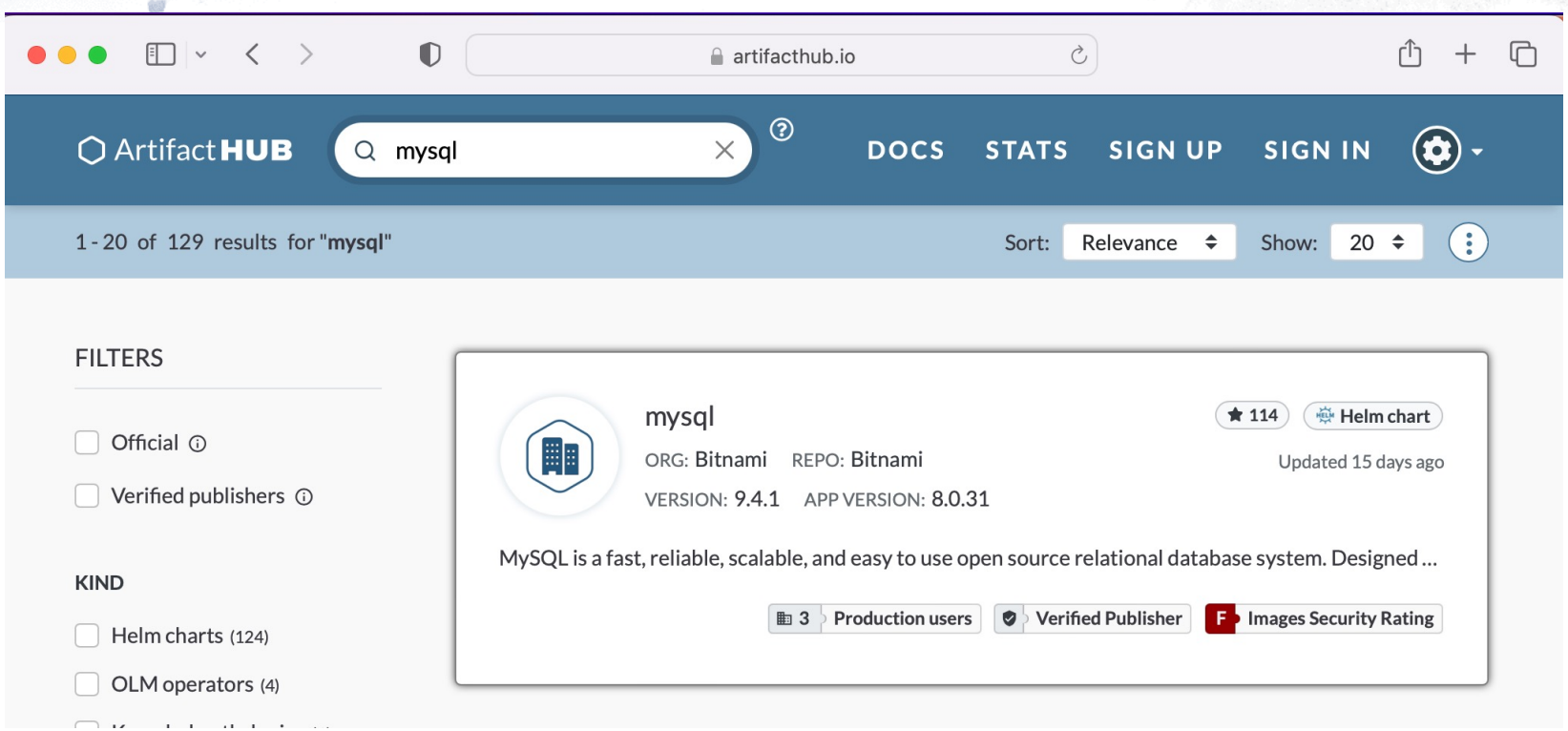
- Many pods can be bundled in a coherent package
- Known as “helm charts”

https://helm.sh

Typically used to deploy new capabilities

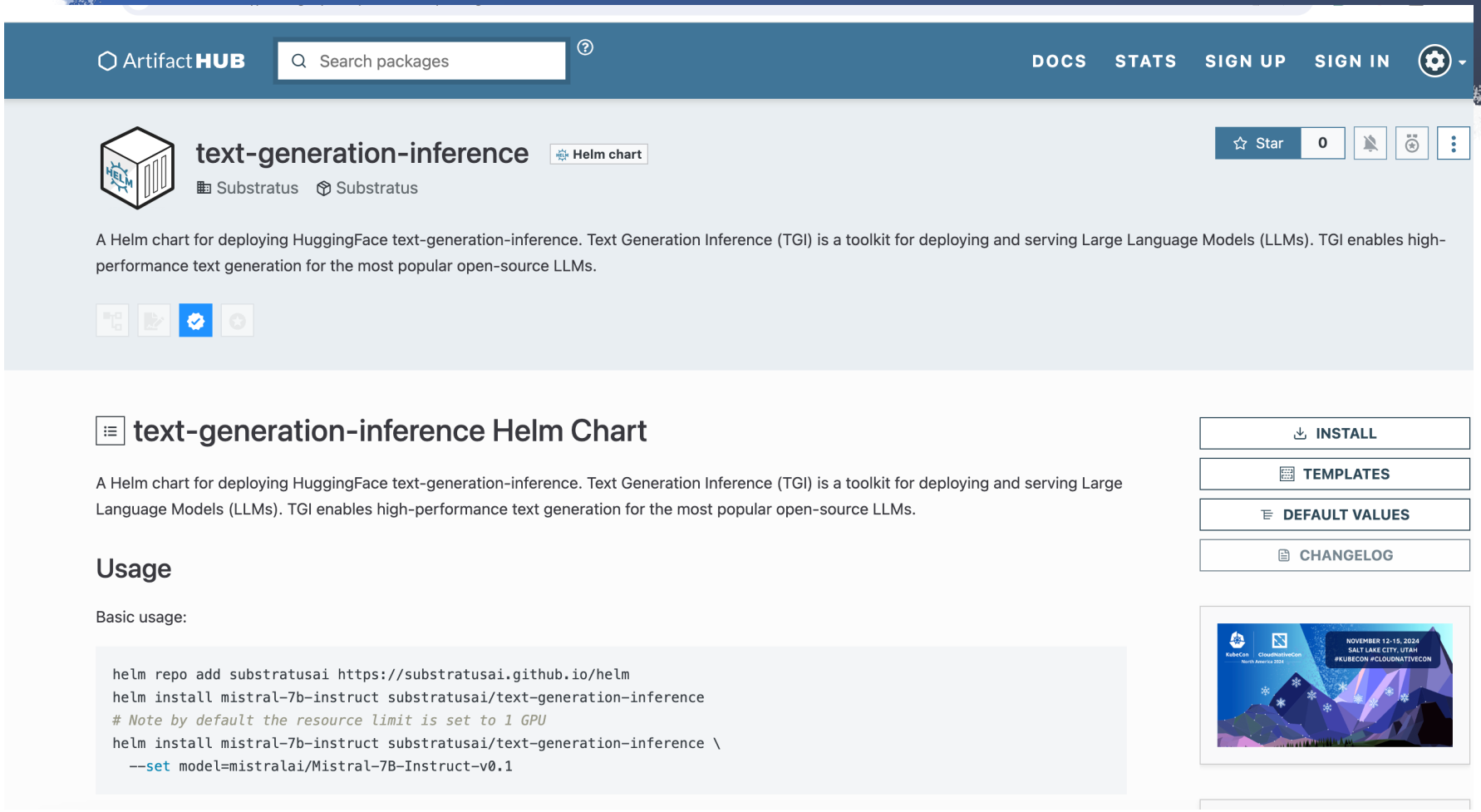
- e.g. multi-pod services
- But can be used for complex science workflows, too
- We will have a LLM as a service deployment example in hands-on session

Large catalogue of services



https://artifacthub.io/

Large catalogue of services



https://artifacthub.io/

And if K8S still does not look appealing

- If you do not like YAML and Pods, it is easy to create an **overlay batch system** on top of Kubernetes
  - Kubernetes capabilities are really a superset of most batch systems (if we ignore scheduling capabilities)
- We have plenty of experience using **HTCondor** as an overlay batch system for Kubernetes (in collaboration with OSG)
  - Users never see Kubernetes, just HTCondor
  - Resources provisioned by Kubernetes given to HTCondor for management (the pilot concept)

https://doi.org/10.48550/arXiv.2205.01004

MPI jobs

- Kubernetes does not natively support MPI jobs
  - Or gang-scheduling in general
- Several external projects help getting a cluster to support them, e.g.
  - kubeflow: https://www.kubeflow.org/docs/components/training/mpi/
  - kube-openmpi: https://github.com/everpeace/kube-openmpi

MPI jobs

- Kubeflow MPI allows for specs on both launcher (where mpirun command runs) and worker nodes (where the actual MPI tasks land)
- There are pods on the launcher node and each of the worker nodes
- stdout/stderr from the MPI job is in the launcher logs as that is where the mpirun is executed
- Example job showing the launcher and worker pods below

```
mahidhar@vgr-1-20:~$ kubectl get pods -n default | grep mpiexample
mpiexample-mahi-launcher-nqbth          1/1     Running             0          35s
mpiexample-mahi-worker-0                1/1     Running             0          35s
mpiexample-mahi-worker-1                1/1     Running             0          35s
mahidhar@vgr-1-20:~$
```

Acknowledgements

This work was partially funded by
US National Science Foundation (NSF) awards
OAC-1541349, OAC-1826967, OAC-2112167,
OAC-2030508, OAC-1841530 and CNS-1730158.

The end