HPC-Cloud workshop BoF

K8s

- Nomad: happz with heat templates, mainly since they allow a K8s deployment without needing to know much about openstack. K8s is the platform they deploy on.
- In general it seems projects prefer file based deployments (yml files) and cli. Rather than a GUI. More understanding of what is being deployed. KISS approach for small clusters.
- Guidelines about how to update K8s clusters. Frank has ideas about pragmatic improvements that can be made, but still likely to be some hands on work.
- Some discussion about HEAT vs Terraform for the recipe. HEAT works well and is native to openstack (autoscaling possible extension). Terraform possible, but probably not a priority. But not ruling out.
- Step by Step guide is seen as very useful.
- pre-installed k8s images (nvidia drivers etc)?
- Re-arrange storage layout imagefs and nodefs for images and user data. Seems to have a +1 from other projects.
- Discussion about providing an image registry. Harbor, could be used for cahce of images etc. Jorge says they went for a docker hub license. Issues with using gitlab registry as a cahce may be solved in update coming soon *watch this space*
- Filesystem mounts into k8s. Nexus-POSIX vs Manila. Nexus-Posix needed if you want to mount on Raven etc. But this means strict user mappings.
- Alternatives to Nexus-POSIX. often issues with inode limits. Will take up again in storage BoF.

Automation/Infrastucture as Code

- Use cases of cloud-init/user-data: Usually simpler/lower-level configs
- Container pipelines:
  - Docker-in-Docker vs. Kaniko
  - Methods to avoid caching to ensure reproducibility
  - Ability to build on top of kubernetes
- Terraform already in use
- CI/CD:
  - Github Actions

  - Contrast between MPCDF/GWDG gitlabs vs. github, institutional vs. public access
- Automation as recovery plan
  - Roadmap item: Cinder backup for block volumes

Storage

- dcor using a cron job on the archive server (archive.mpcdf.mpg.de) to pull s3 as a backup.
- bagit can be used for archiving data. provides a way to bake in metadata and checksums. Check out bits and bytes articles for extiension from mpcdf and also available via modules (needs checking becuase this is prob dated).
- CEPH-FS as an alternative to nexus-posix.
  - would like to mount on raven and cloud VMs
  - manila limitation due to single nfs server
- Object storage.
  - multiple users, can apply for free TiB.
  - possible for users to get more than 1TiB if they pay (group wide quota)
  - CEPH roadmap includes a possible solution for more fine grained user access management within a project.
- Discussion about staging data to/from S3 into /ptmp.
  - can dowonload data without issue but uploading seen as a possible issue. Cannot stream data into S3, so caching somewhere and then doing a PUT. Can possibly use /ptmp but sometimes slow. Can consider using shm
- Accessing data in s3 (HDF5 based data?)
  - This is for sciserver so the issues relating to access management may be solved by the application.
  - IP based access policy per bucket may also help
  - Side note that the filesystem (fuse) like wrappers for S3 are nice for small scale. But not advised for production at scale.

Slurm/clusters

- GPU clusters: not there yet
  - People using workstations with GPUs with not so much concurrency
- HTCondor: look into that
- Q about module system being available on every virtual cluster
  - Ubuntu support?

    - See how we can make it available to more users
    - Apptainer/docker containers with all the environment vs. normal HPC workflow with modules et al
        - Modules allow for the same environment as RAVEN
- Wishlist for SLURM offerings: No additional points
- Wishlist for JADE
    - Concurrency problem writing to the slurm config over NFS on many nodes at the same time
    - Mount the module system? (see Ubuntu support or lack thereof)
- Other clusters?
    - HTCondor would go here