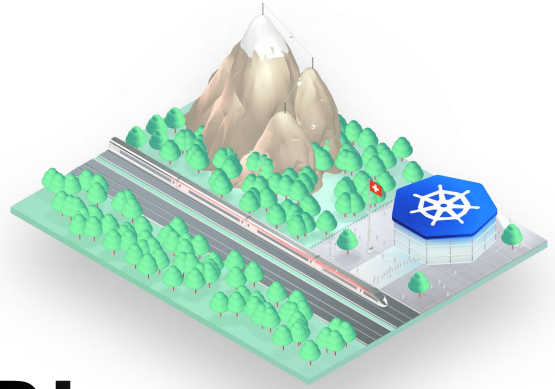


When Things Get Big

Scaling Cloud Native Workloads and Software Distribution

Ricardo Rocha, CERN @ahcorporto

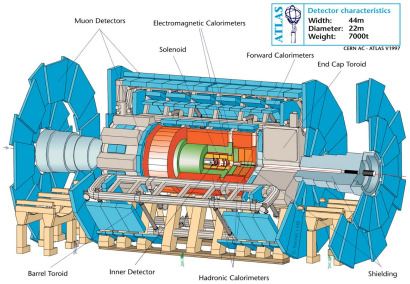
https://kcdzurich.ch/

Big **Machines**


CMS
CERN
LHC
[Large Hadron Collider]
ALICE
LHCb
ATLAS

Palazzani

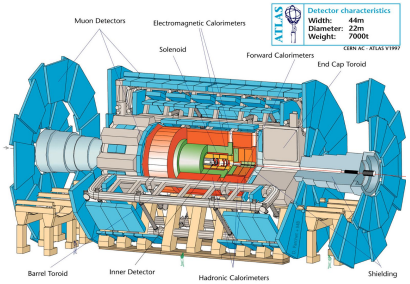
Big **Applications**



1 PB / sec

ATLAS Event Filter

< 10 GB / sec



1 PB / sec

ATLAS Event Filter

< 10 GB / sec

Typically split into Hardware and Software Filters
(this might change too)

40 million particle interactions / second

~3000 multi-core nodes

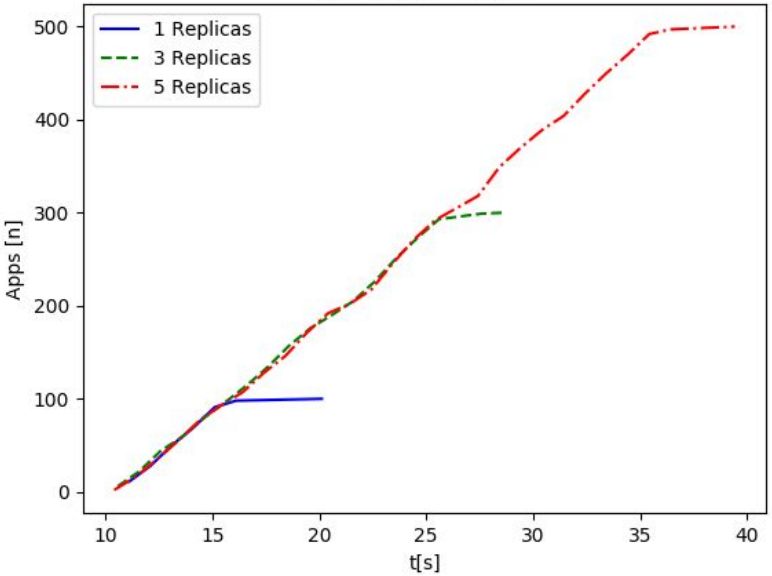
~30.000 applications to supervise

Critical system, sustained failure means data loss

Can it be improved for Run 4?

First study 2017, Mattia Cadeddu, Giuseppe Avolio

Kubernetes 1.5.x



First results… 100 node cluster

40 sec to start **500** applications on 100 nodes

Container launch rate **12Hz**

Not particularly promising ...

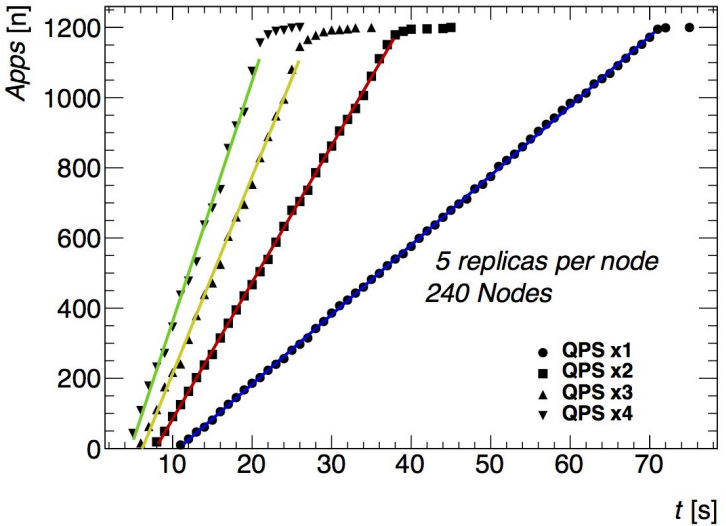
Issues Found

QPS defaults were very conservative

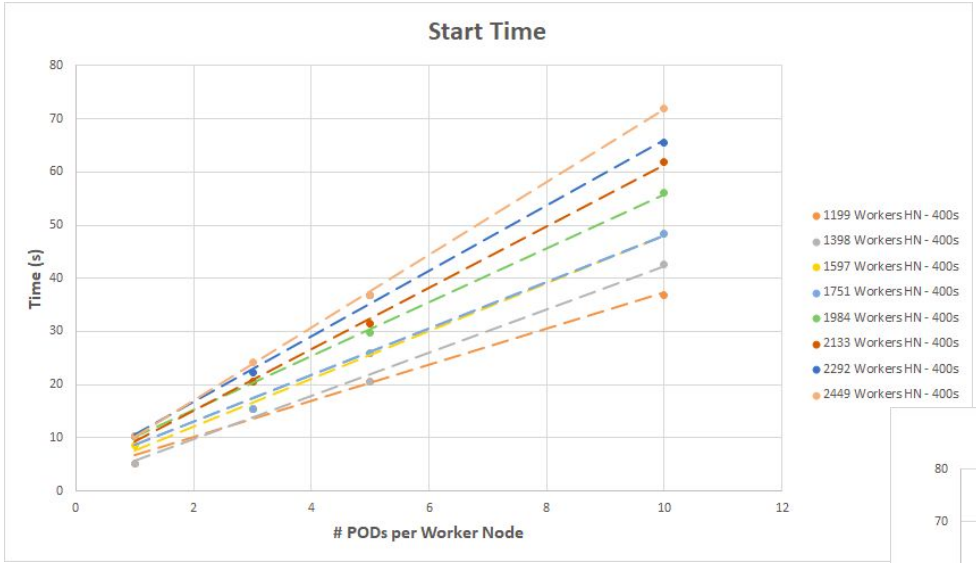
Tested x1, x2, x3, x4 values

QPS x4 gives 57Hz: **> 330 % improvement**

Does Kubernetes prefer larger clusters ?



2022



Kubernetes 1.21.2

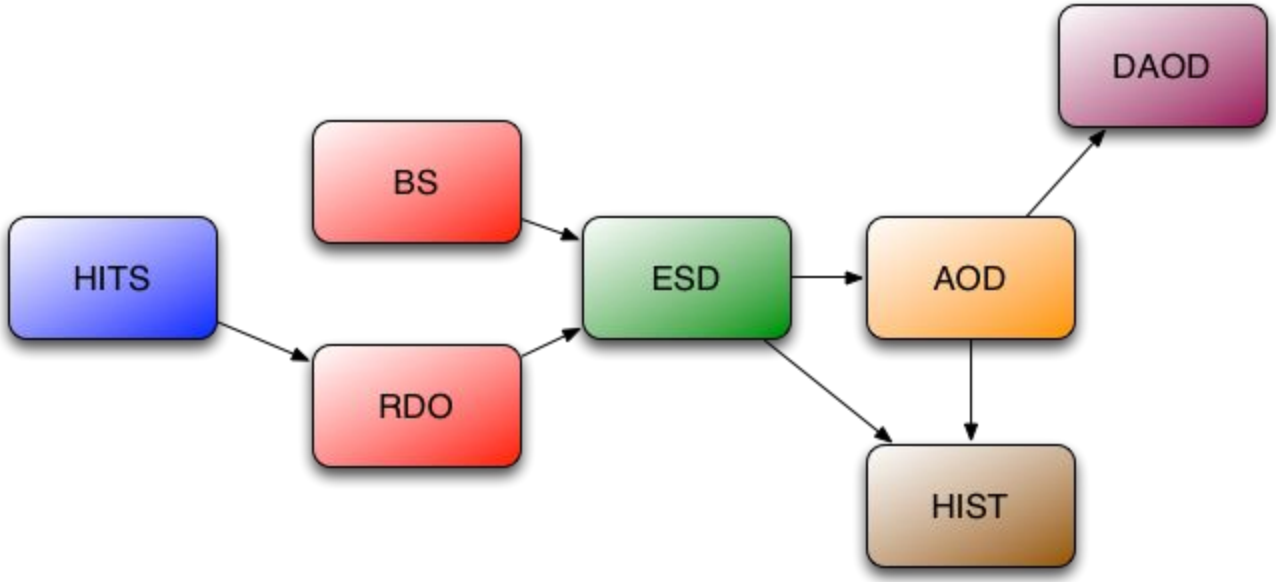
2500 Nodes

Prediction 5000 Node Cluster

PODs per Host	Start Time (s) [95% CL]	Stop Time (s) [95% CL]
1	[13.8 – 29.1]	[18.7 – 26.2]
3	[39.0 – 48.4]	[44.4 – 51.8]
5	[75.1 – 80.3]	[81.1 – 82.9]

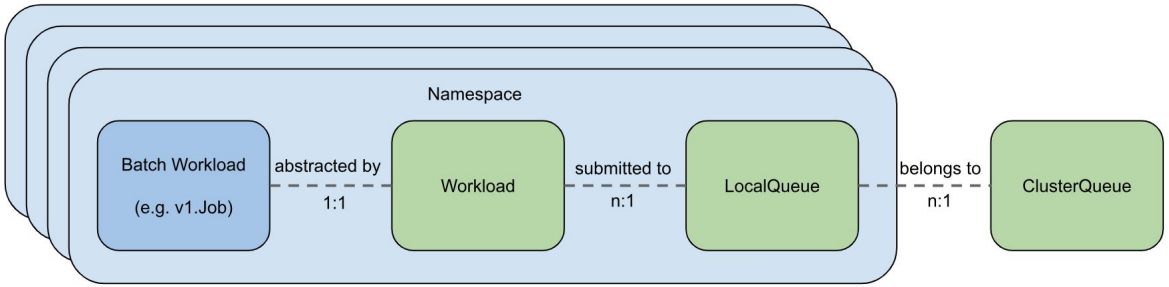
Credit: Giuseppe Avolio, ATLAS

Big **Workloads**


DAOD
BS
HITS
ESD
AOD
RDO
HIST


11/22/2013 5:55:18 p.m.
Running jobs: 244151
Transfer rate: 40.08 GiB/sec
WLCG
US Dept of State Geographer
© 2013 Google
Data SIO, NOAA, U.S. Navy, NGA, GEBCO
Image Landsat
Google earth
Fecha de las imágenes: 4/10/2013 66°43'28,18" N 8°52'37,10" O alt. ojo 16085.50 km

Workloads, not just Pods

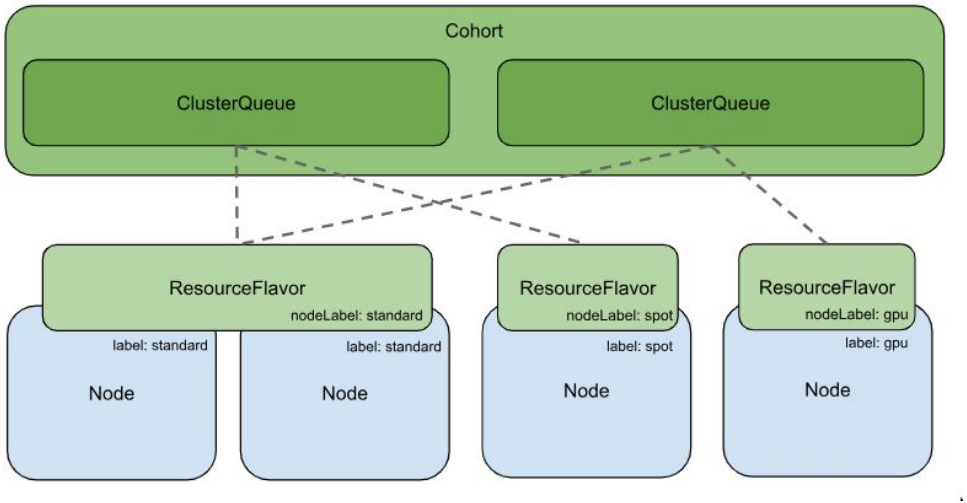


https://kueue.sigs.k8s.io/

Workloads, not just Pods

Queues and Priorities, not all workloads are equal

Fair Sharing, resource usage optimization



https://kueue.sigs.k8s.io/

Workloads, not just Pods

Queues and Priorities, not all workloads are equal

Fair Sharing, resource usage optimization

Gang Scheduling, Array Jobs

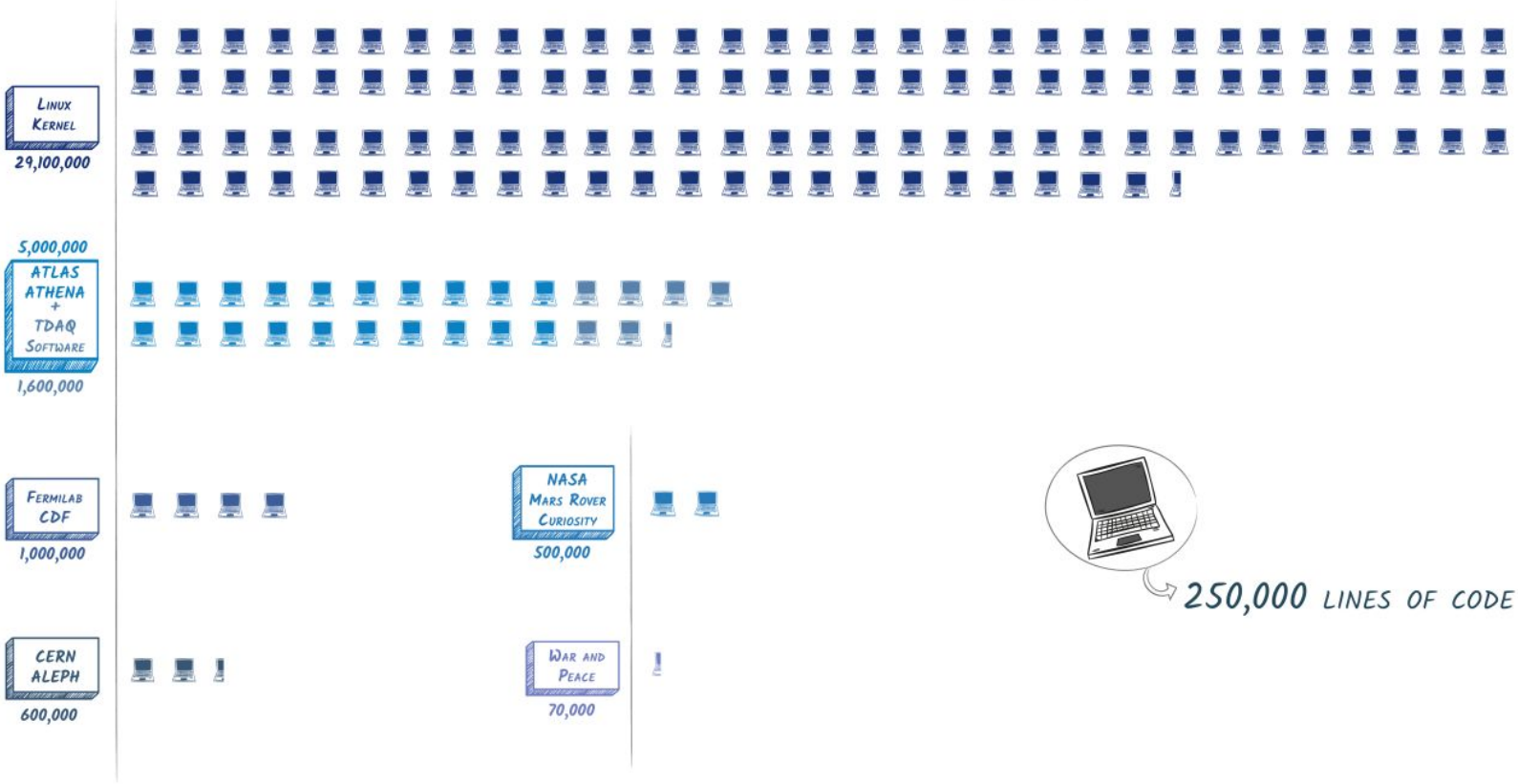
Kubernetes Batch WG
https://github.com/kubernetes/community/tree/master/wg-batch

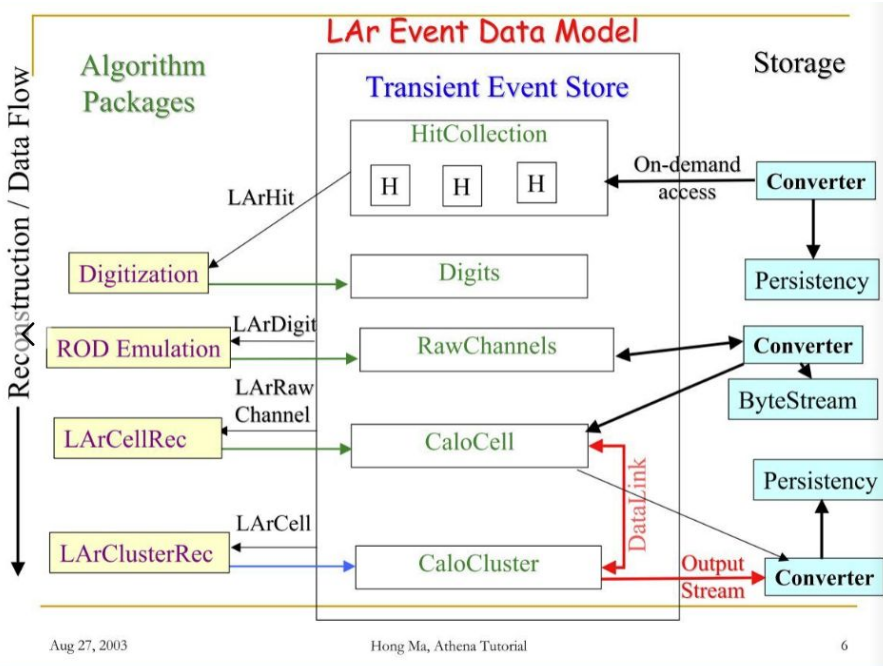
CNCF Batch System Initiative
https://github.com/cncf/tag-runtime/issues/38

CNCF Research User Group
https://community.cncf.io/research-end-user-group/

Big **Software**

Counting Lines of Code





Releases

- Release: A complete set of code compiled, built and frozen for use
  - Defined by a set of tags for each package,
    - Tag collector
  - Nightly builds
    - Used by developers to work towards a release.
    - Not guarranteed to work
  - Developer Release: approximately every 3 weeks(currently 6.6.0)
    - Mostly working
  - Major release: approximately every 6 months (currently
    - Major milestones, such as Data Challenges
- Release plan and Release Status can be found on the "Software Development" page
- USATLAS builds same releases as CERN
  - /afs/usatlas.bnl.gov/software/dist/

	compressed	uncompressed
alpine	3MB	6MB
ubuntu:22.04	30MB	78MB

	compressed	uncompressed
alpine	3MB	6MB
ubuntu:22.04	30MB	78MB
admutils	870MB	1.2GB

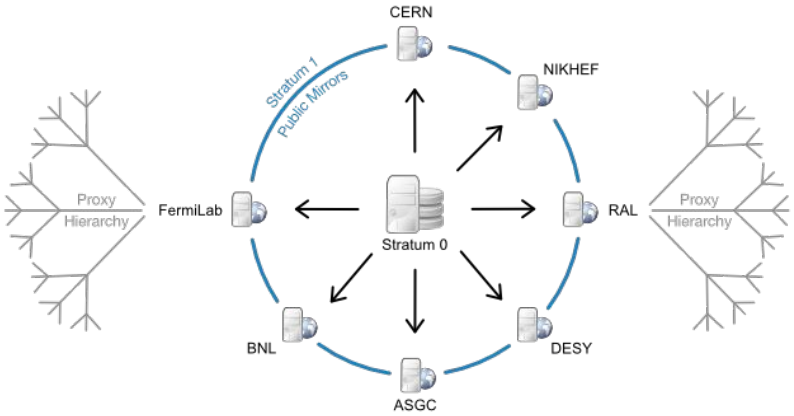
	compressed	uncompressed
alpine	3MB	6MB
ubuntu:22.04	30MB	78MB
admutils	870MB	1.2GB
atlas/athena	18.7GB	61GB

Why the heck is this thing 24 GB?!?!

```
docker images us-docker.pkg.dev/colab-images/public/runtime:latest
REPOSITORY                                        TAG       IMAGE ID       CREATED       SIZE
us-docker.pkg.dev/colab-images/public/runtime   latest    01b91234d65c   2 weeks ago   24.1GB
```

@Ricardo Rocha ^ here's another one for your "large images running in production" list. 😛

	compressed	uncompressed
alpine	3MB	6MB
ubuntu:22.04	30MB	78MB
admutils	870MB	1.2GB
atlas/athena	18.7GB	61GB
<???>	76GB	124GB



Container SW Distribution

16 May 2019, 12:00 → 17 May 2019, 23:10 Europe/Zurich

513/R-070 - Openlab Space (CERN)

Ricardo Brito Da Rocha (CERN)

Registration — *You are registered for this event.* Check details

Thursday, 16 May

14:00 → 15:00 **SM18 Visit** 1h 513/R-070 - Openlab Space

Friday, 17 May

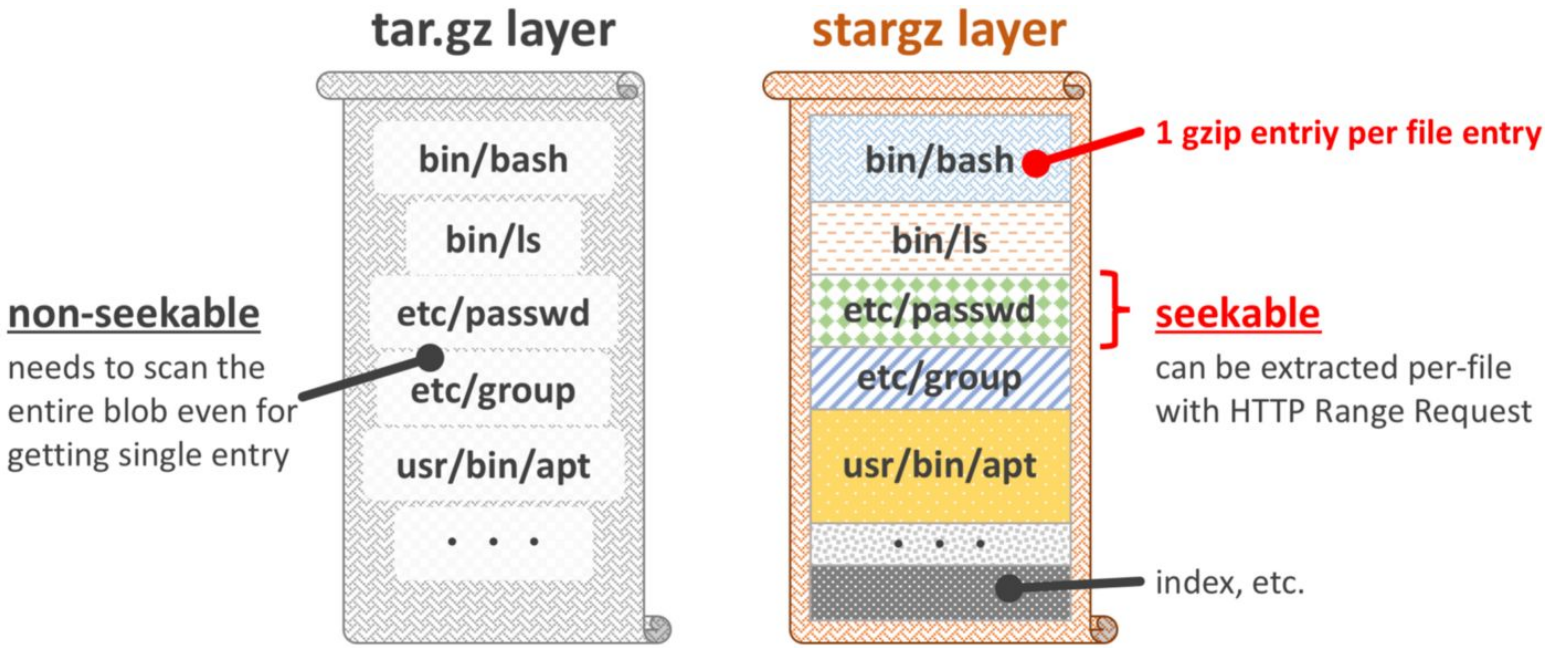
09:00 → 09:25 **CVMFS and Software Distribution on the Grid** 25m 513/R-070 - Openlab Space
Speaker: Jakob Blomer (CERN)
cvmfs-containers.pdf

09:25 → 09:50 **WLCG and the case for userspace containers** 25m 513/R-070 - Openlab Space
Speakers: Alessandra Forti (University of Manchester (GB)), Lukas Alexander Heinrich (CERN)
ContainerSWDistr.pdf

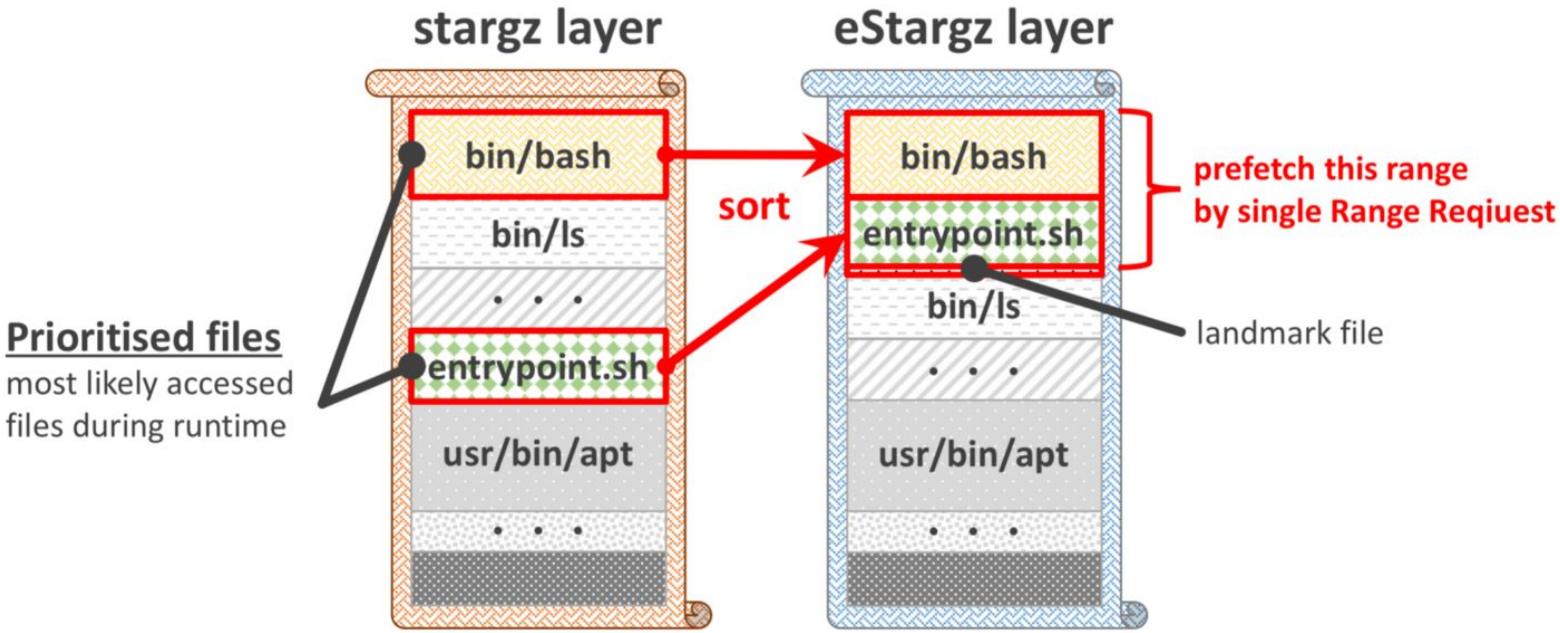
09:50 → 10:15 **Overview of Containerd** 25m 513/R-070 - Openlab Space
Speaker: Phil Estes (IBM)
FOSDEM 2019_ con...

11:00 → 11:30 **Containerd Remote Snapshotter** 30m 513/R-070 - Openlab Space

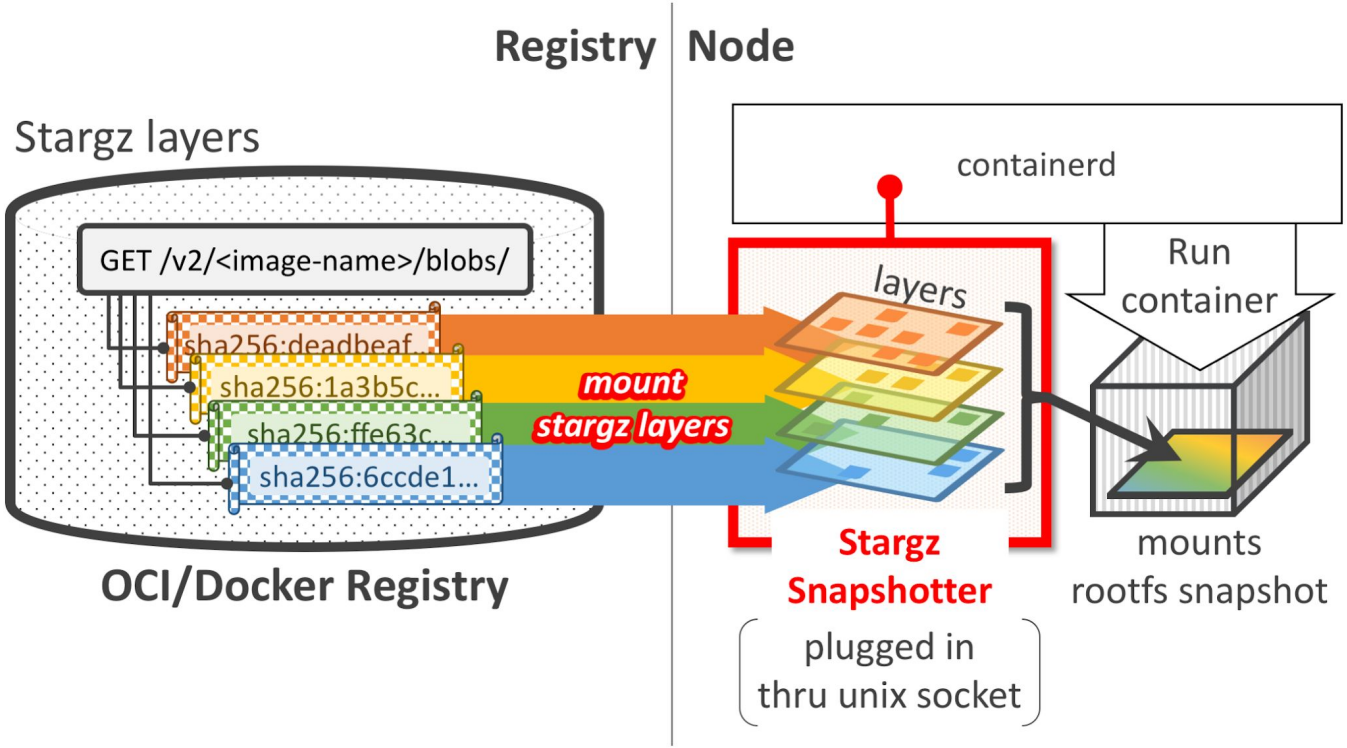
11:30 → 12:15 **Discussion: Containerd Remote Snapshotter** 45m 513/R-070 - Openlab Space



https://github.com/containerd/stargz-snapshotter



https://github.com/containerd/stargz-snapshotter



https://github.com/containerd/stargz-snapshotter

Containers & Kubernetes

Introducing GKE image streaming for fast application startup and autoscaling

November 4, 2021

Image streaming is a method of pulling container images in which GKE streams data from eligible images as requested by your applications. You can use Image streaming to allow your workloads to initialize without waiting for the entire image to download, which leads to significant improvements in initialization times. The shortened pull time provides you with benefits including the following:

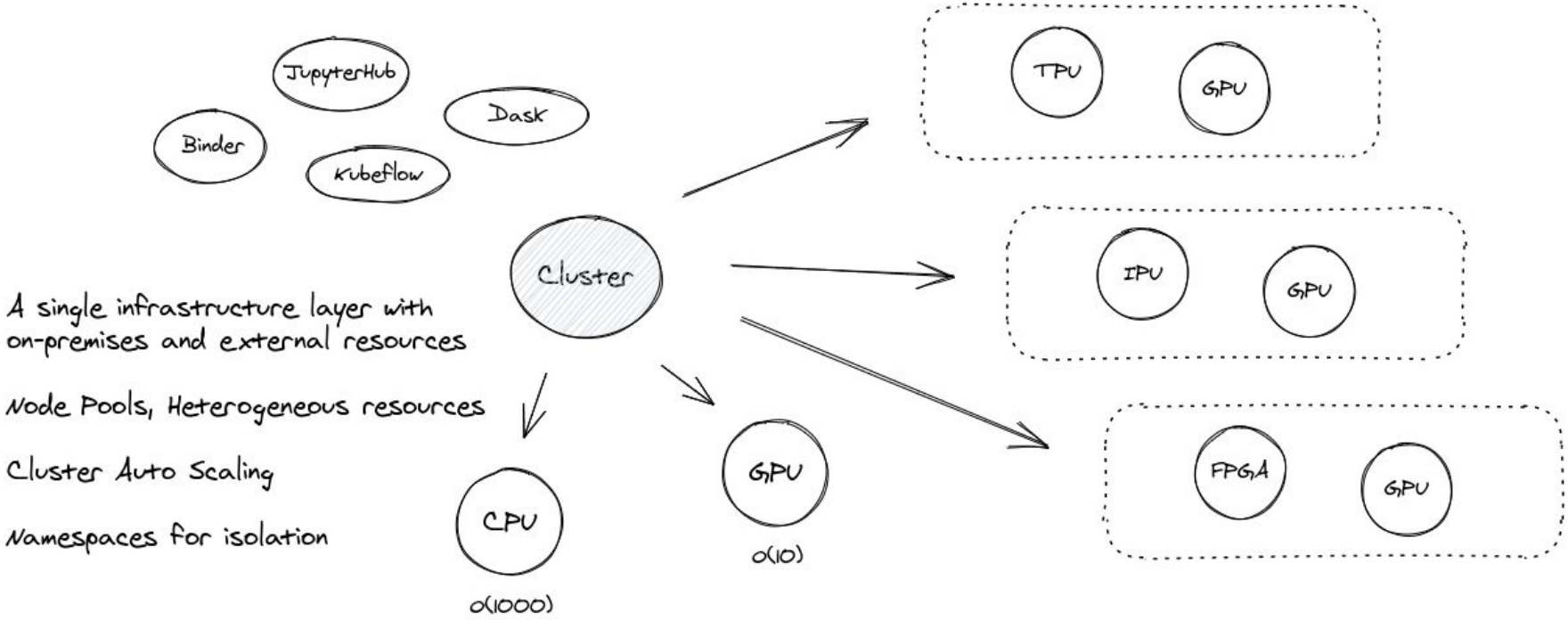
- Faster autoscaling
- Reduced latency when pulling large images
- Faster Pod startup

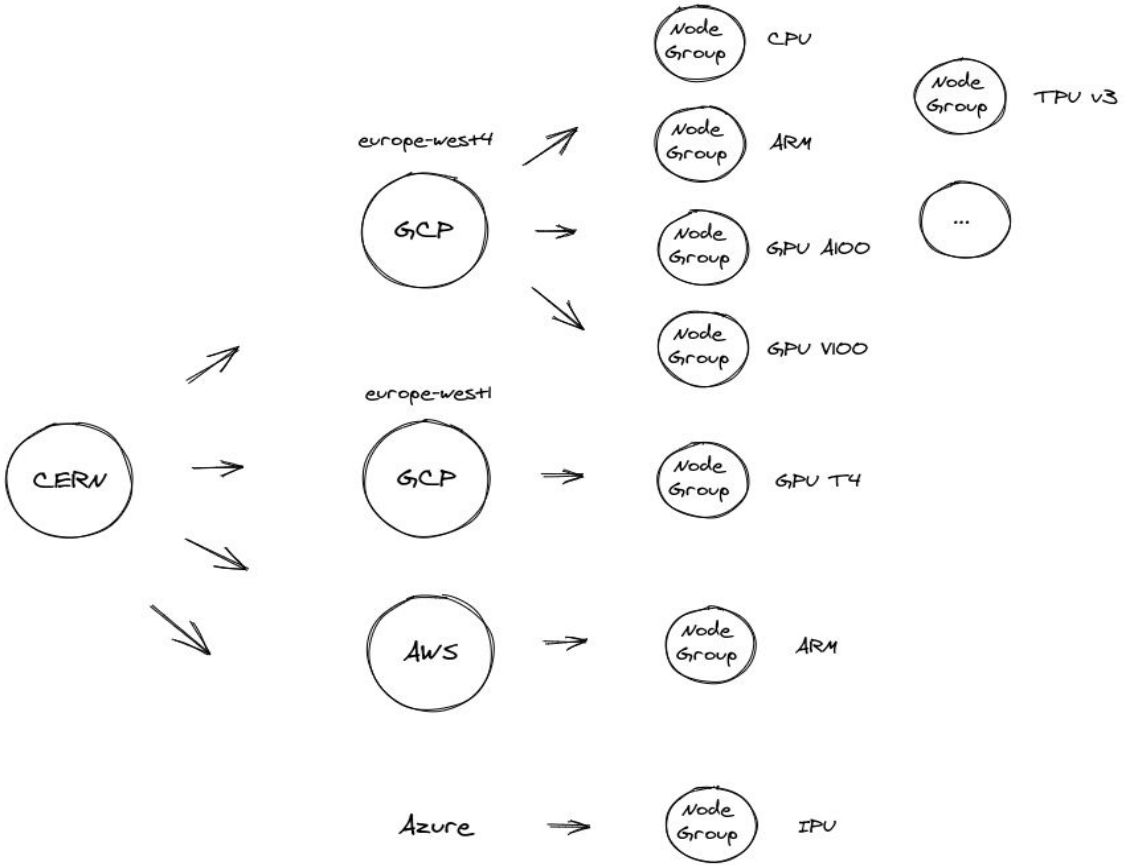
- You might not notice the benefits of Image streaming during the first pull of an eligible image. However, after Image streaming caches the image, future image pulls on any cluster benefit from Image streaming.

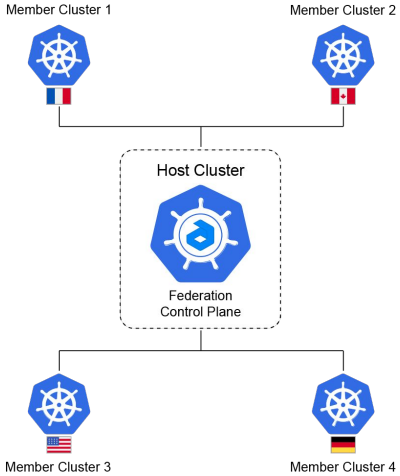
https://cloud.google.com/blog/products/containers-kubernetes/introducing-container-image-streaming-in-gke

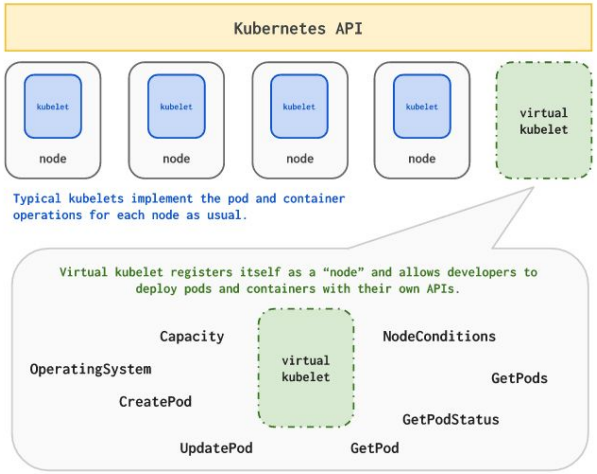
Quick Demo

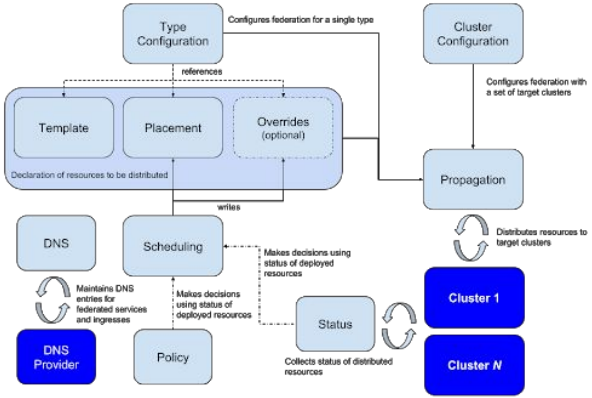
Big **Needs**

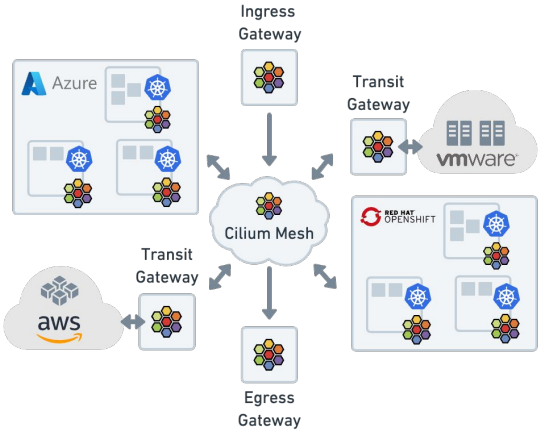

JupyterHub
Binder
Dask
Kubeflow
Cluster
A single infrastructure layer with
on-premises and external resources
Node Pools, Heterogeneous resources
Cluster Auto Scaling
Namespaces for isolation
CPU
o(1000)
GPU
o(10)
TPU
GPU
IPU
GPU
FPGA
GPU


CERN
europe-west4
GCP
europe-west1
GCP
AWS
Azure
Node Group
CPU
Node Group
ARM
Node Group
GPU A100
Node Group
GPU V100
Node Group
GPU T4
Node Group
ARM
Node Group
IPU
Node Group
TPU v3
...

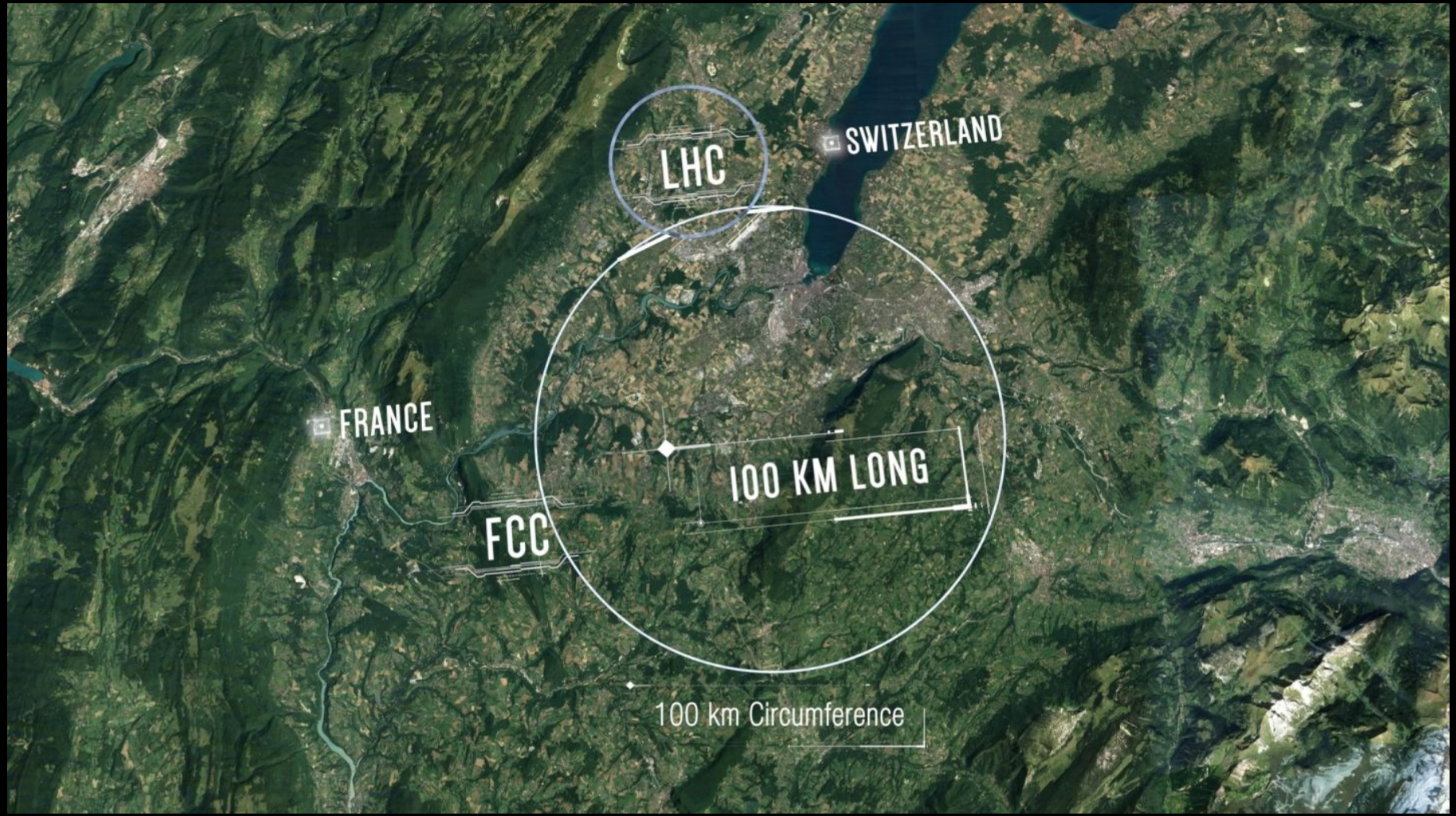

Member Cluster 1
Member Cluster 2
Host Cluster
Federation
Control Plane
Member Cluster 3
Member Cluster 4


Kubernetes API
kubelet
node
kubelet
node
kubelet
node
kubelet
node
virtual
kubelet
Typical kubelets implement the pod and container operations for each node as usual.
Virtual kubelet registers itself as a "node" and allows developers to deploy pods and containers with their own APIs.
Capacity
NodeConditions
OperatingSystem
virtual
kubelet
GetPods
CreatePod
GetPodStatus
UpdatePod
GetPod


Type
Configuration
Configures federation for a single type
Cluster
Configuration
references
Configures federation with
a set of target clusters
Template
Placement
Overrides
(optional)
Declaration of resources to be distributed
Propagation
writes
Distributes resources to
target clusters
DNS
Scheduling
Makes decisions using
status of deployed
resources
Maintains DNS
entries for
federated services
and ingresses
Makes decisions
using status of
deployed resources
Cluster 1
Status
DNS
Provider
Policy
Collects status of distributed
resources
Cluster N


Ingress
Gateway
Azure
Transit
Gateway
vmware
Cilium Mesh
RED HAT
OPENSHIFT
Transit
Gateway
aws
Egress
Gateway

2017 ---------- 2023


LHC
SWITZERLAND
FRANCE
FCC
100 KM LONG
100 km Circumference

Questions?