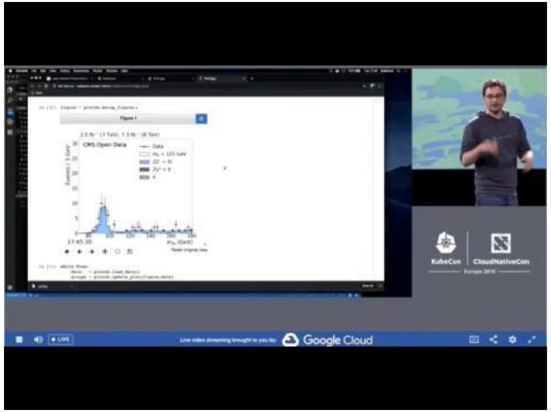
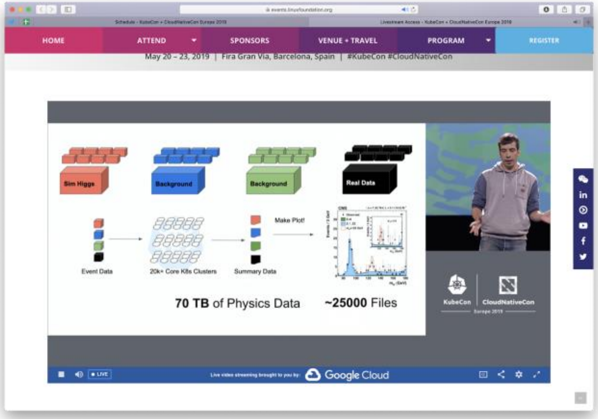
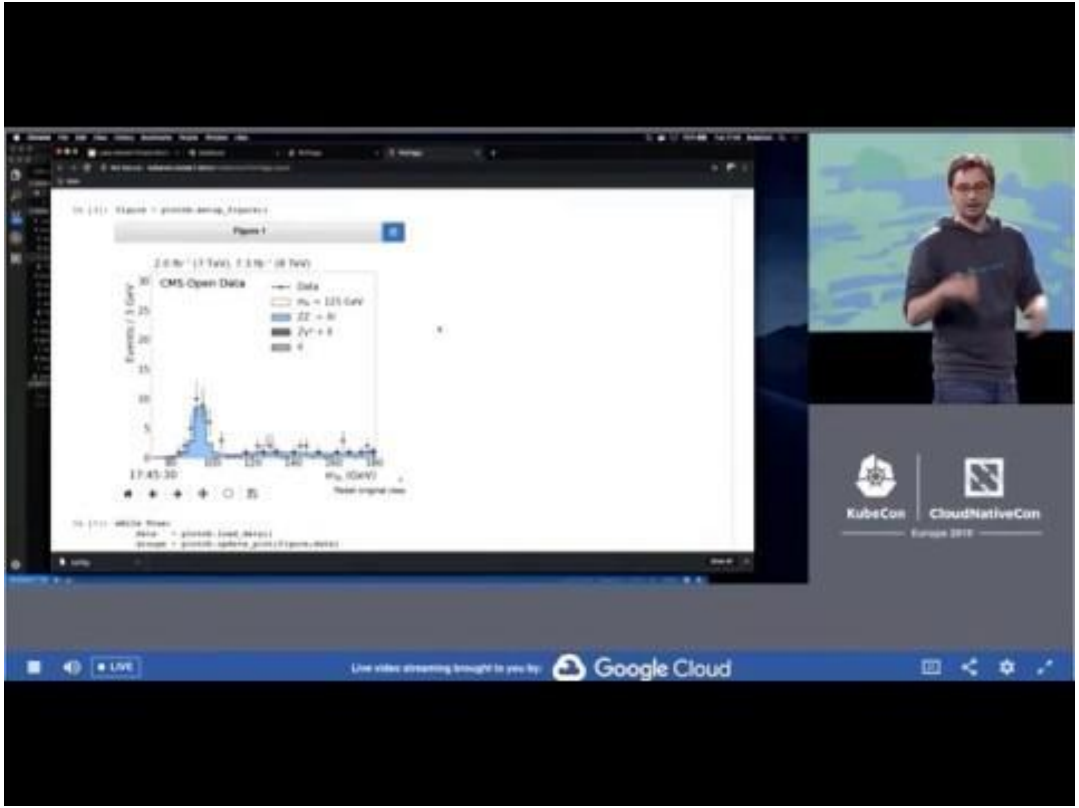
Kubecon Higgs Analysis

Demo Deep Dive

Lukas Heinrich, Ricardo Rocha





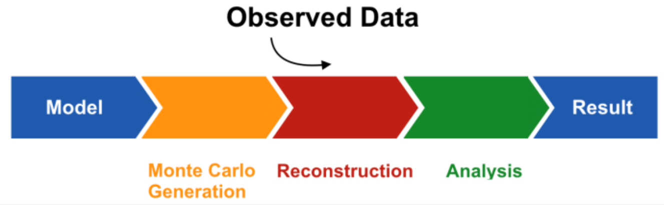


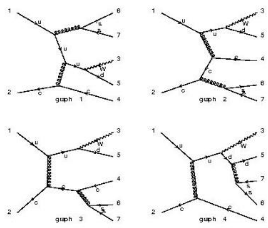
https://www.youtube.com/watch?v=CTfp2woVEkA

Our job: look at data and check against multiple theories (Higgs, SUSY, …)

Fundamental problem 1: looking for rare phenomena. Needs lots of data.
Fundamental problem 2: do not have a simple way to predict what data would look like under **different theories / assess compatibility**

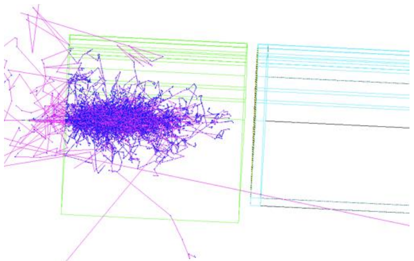
Solution:
Use **large scale compute** to process data
+ **deep stack of software** to brute-force what data looks like under theories (Monte Carlo)





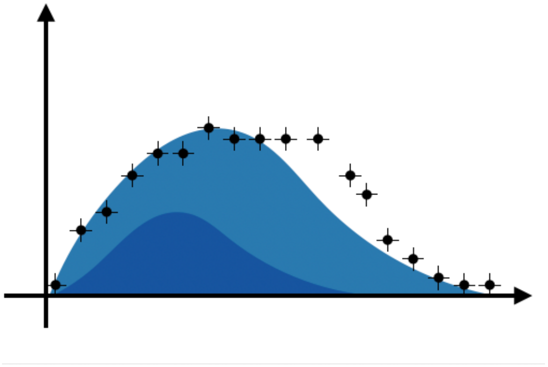
High Energy Physics

Event Evolution

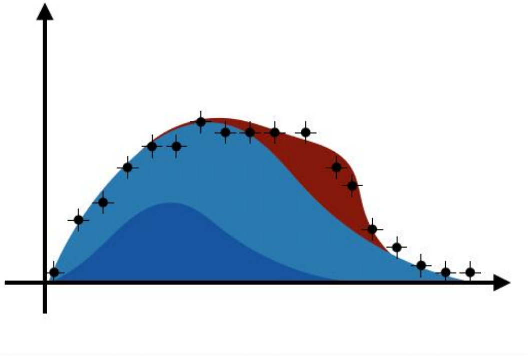


Detector Interactions

Baseline only



Baseline +
New Physics

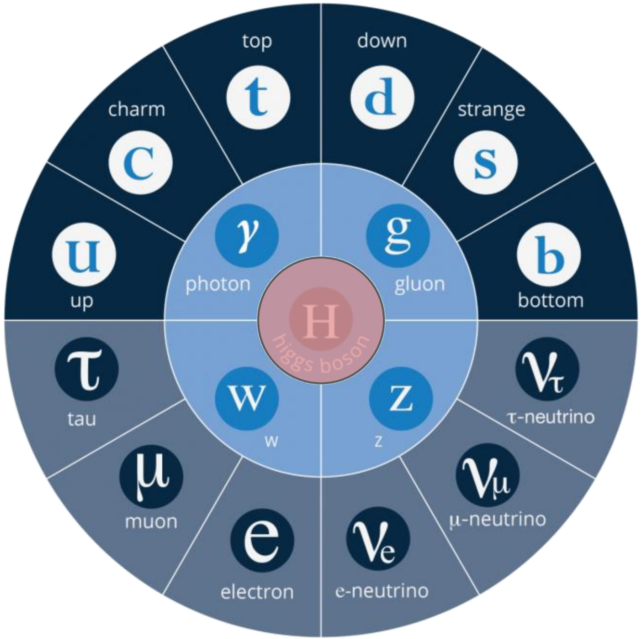


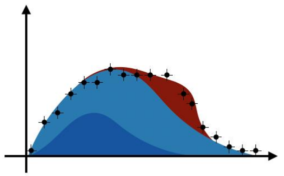
Baseline cannot describe data
… but baseline + new physics theory does -> Discovery!

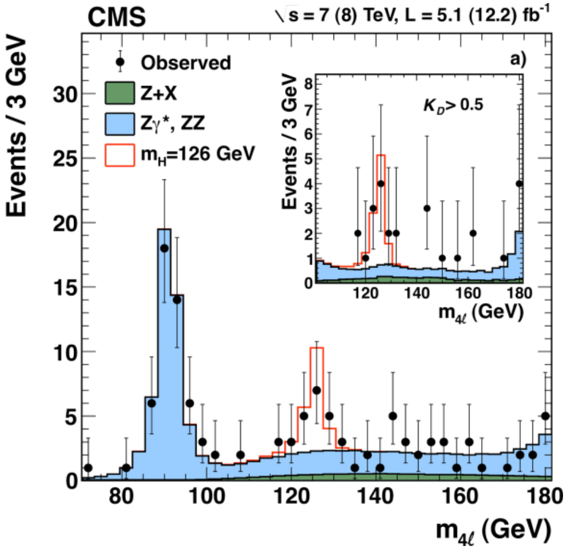
Back in 2012, CERN announced one of its most important achievements, the discovery of the Higgs boson leading to the 2013 Nobel Prize in Physics.

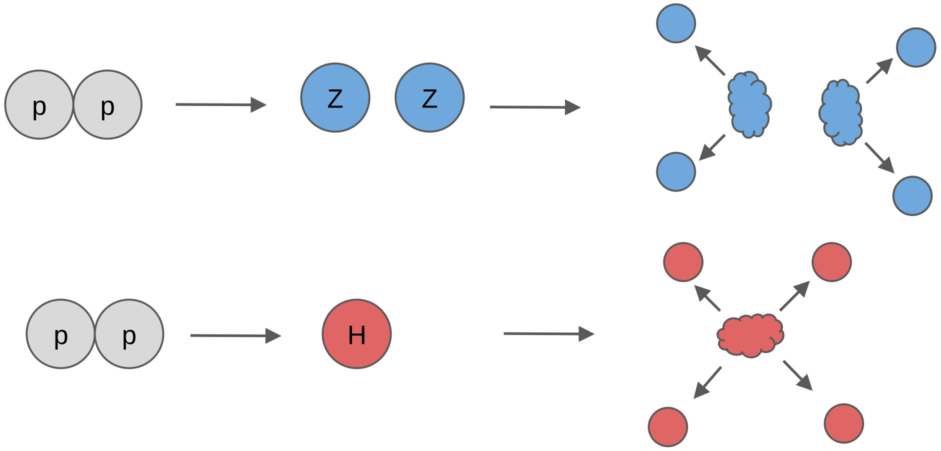
In this presentation, we will redo the data analysis that led to it, this time on top of Kubernetes, the new infrastructure stack growing in popularity in the laboratory.

Demo Idea: reproduce Higgs discovery



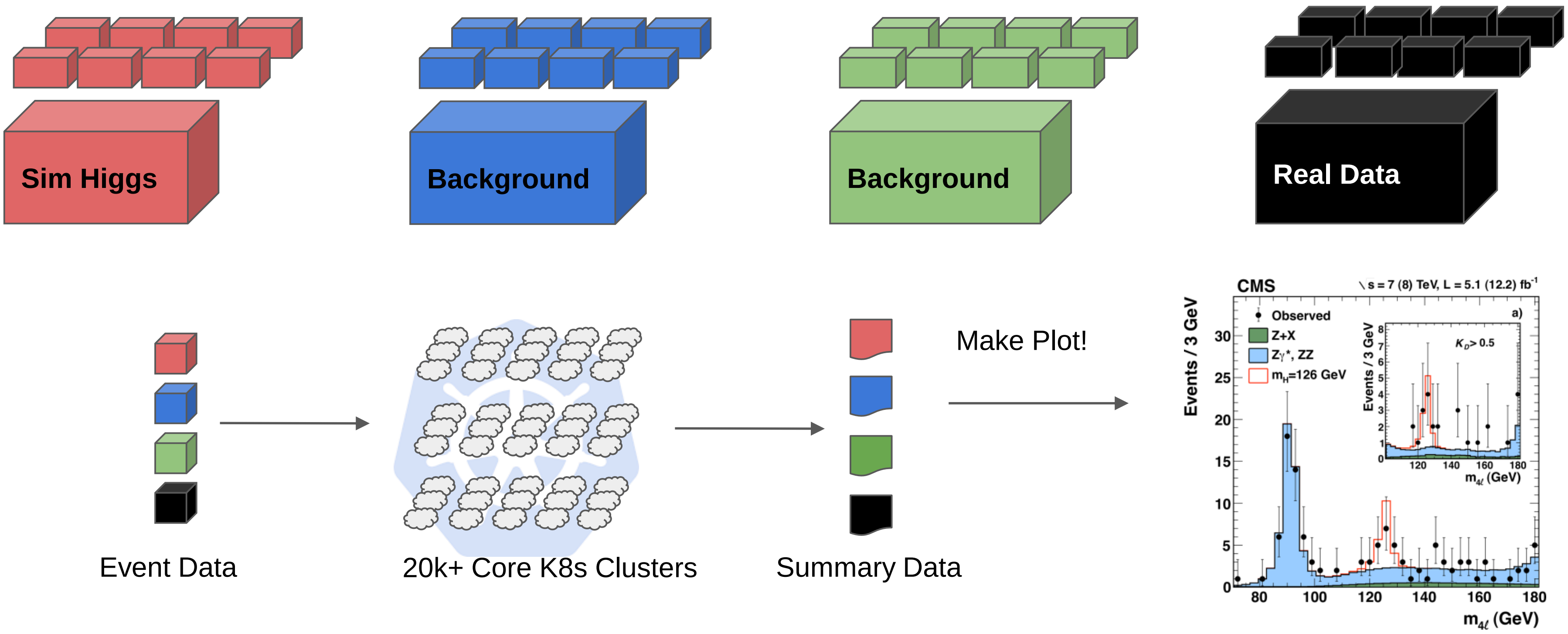






Demo

https://github.com/cernops/higgs-demo



70 TB of Physics Data **~25000** Files

70 TB Dataset

OpenStack Magnum

25000 **Kubernetes Jobs**

Job Results

Interactive Visualization

Aggregation

Moving data from EOS to CERN S3

Initial dataset (opendata) available on /eos

S3 is more cloudy, we wanted to test with that to ease transition to GCP

https://gitlab.cern.ch/rbritoda/eos2s3

(a Kubernetes backed dummy file transfer service)

16 parallel transfer processes

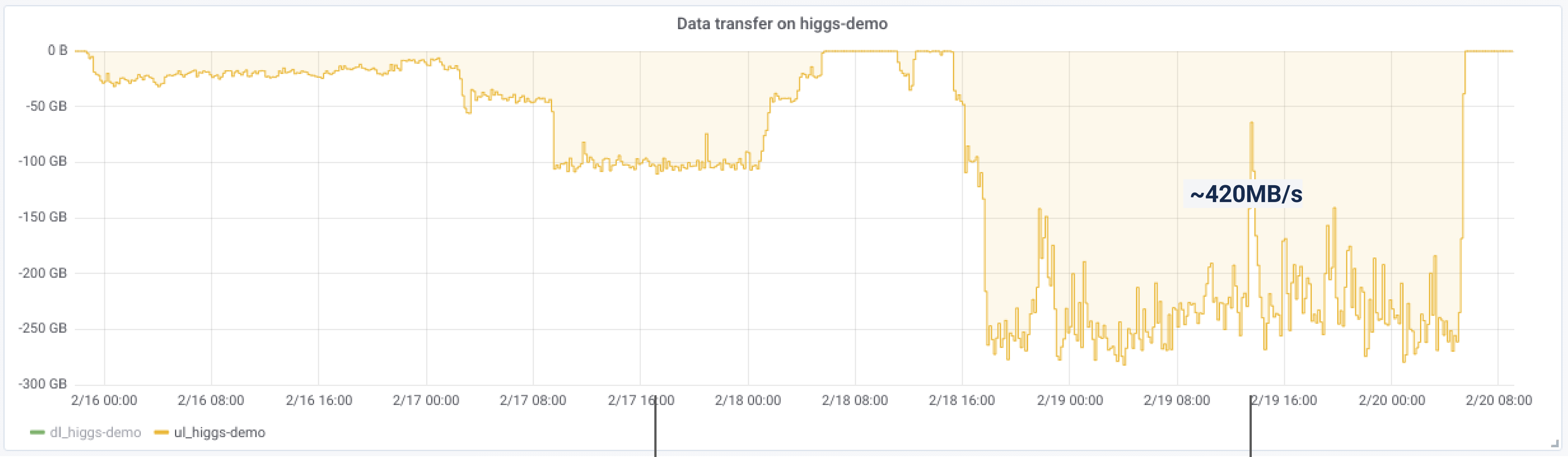
Moving data from EOS to CEPH/S3

s3cmd
(single transfer)

s3cmd
(parallel)

aws cli
(multipart)

Moving data from EOS to CEPH/S3



CEPH / S3
(optimization #1)

CEPH / S3
(optimization #2)

Big thanks to the CEPH team

CERN Analysis Run

Kubernetes 1.12

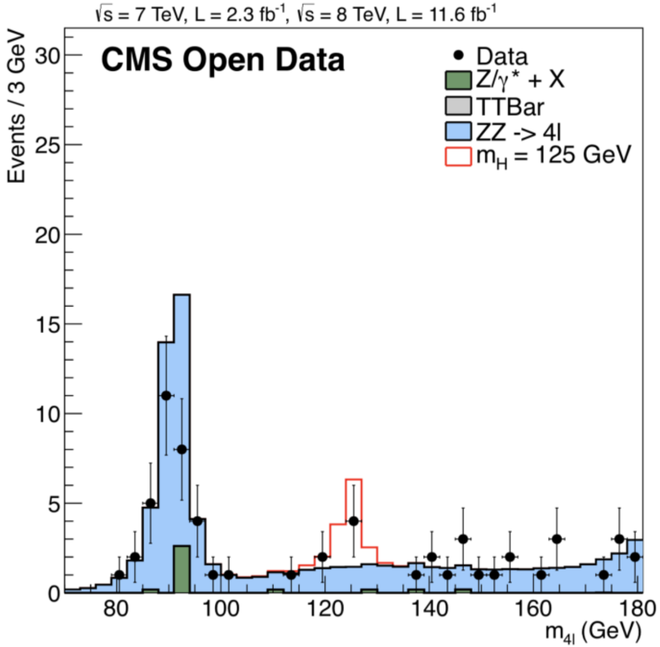
61 Nodes (VMs)

- 4 Cores / 8 GB nodes
- 40GB disks (SSDs)

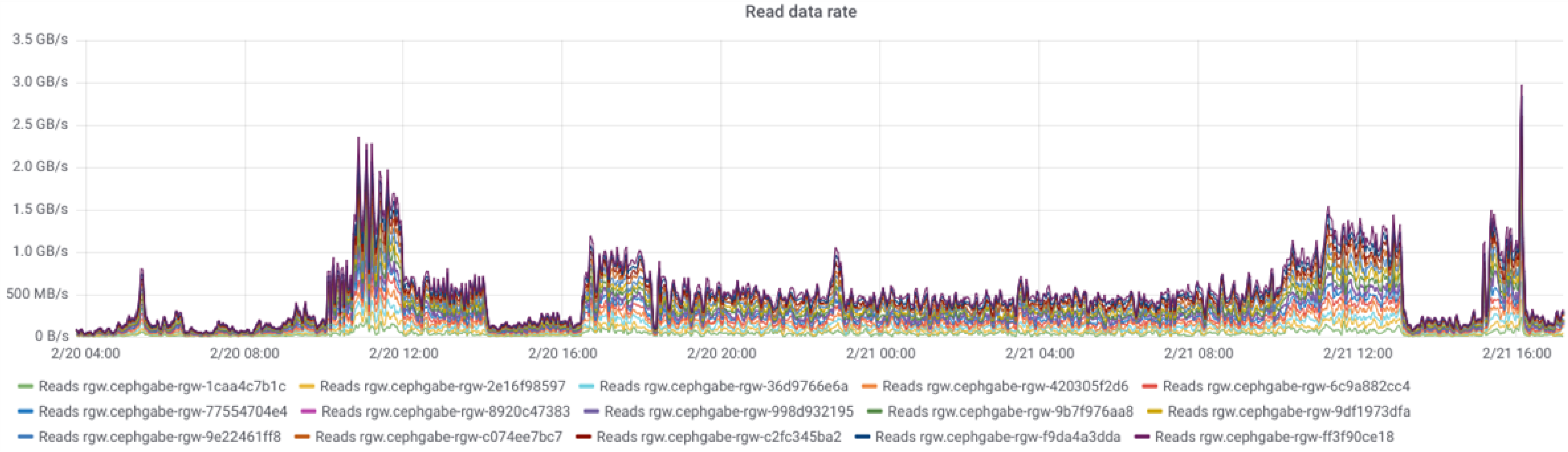
Running on 36TB (half the dataset)

Total time: 19h with 244 cores (~1GB/s)

Goal for GCP: 250x speedup to run it in <10min



CERN Analysis Run



CERN openlab

Google Cloud

Google Cloud Storage

70 TB Dataset

Cluster on GKE

Max **25000 Cores**

Single Region, 3 Zones

25000 **Kubernetes Jobs**

Job Results

Interactive Visualization

Aggregation

First Transfer to Zurich Region

Added GCS support to eos2s3 (using gsutil)

Zurich because closer is better?

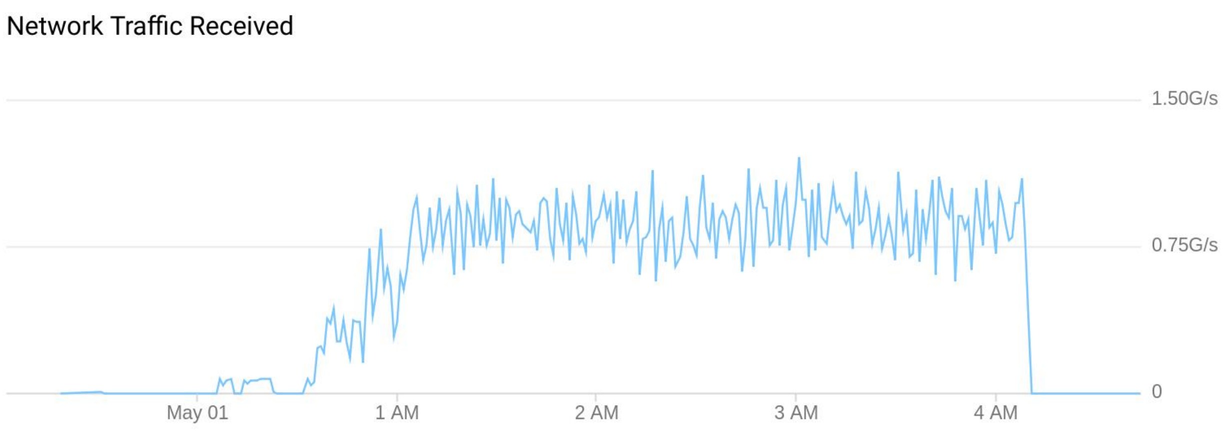
But using the Internet 70TB will take a long time… or?

First Transfer to Zurich Region

Added GCS support to eos2s3 (using gsutil)

Zurich because closer is better?

But using the Internet 70TB will take a long time… or?



Late night good news!

First Transfer to Zurich Region

Added GCS support to eos2s3 (using gsutil)

Zurich because closer is better?

But using the Internet 70TB will take a long time… or?

Sample destination: zrh04s15-in-f10.1e100.net

```
Traceroute from kubecon-demo-012-2elgq755lsas-minion-77 (i692316327109xx.cern.ch):
traceroute to zrh04s15-in-f10.1e100.net (172.217.168.74), 30 hops max, 46 byte packets
 1  10.100.104.1 (10.100.104.1)  0.005 ms  0.005 ms  0.003 ms
 2  l513-v-rbrmx-1-xxx.cern.ch (10.42.xx.x)  1.351 ms  0.147 ms  0.123 ms
...
 8  e773-e-rbrxl-2-xxx.cern.ch (192.65.xx.xx)  1.152 ms  1.172 ms  0.996 ms
 9  google-zurich-10g.cern.ch (192.65.184.202)  6.102 ms  6.176 ms  6.339 ms
10  74.125.243.113 (74.125.243.113)  7.166 ms  7.001 ms  6.996 ms
11  172.253.50.5 (172.253.50.5)  15.032 ms  64.233.175.167 (64.233.175.167)  7.797 ms
8.035 ms
12  zrh04s15-in-f10.1e100.net (172.217.168.74)  7.012 ms  6.924 ms  7.109 ms
```

Late night good news!

First Transfer to Zurich Region

Added GCS support to eos2s3 (using gsutil)

Zurich because closer is better?

But using the Internet 70TB will take a long time… or?

- **Device Name:** **CIXP-GOOGLE** [Last Operation]
- Location: 0000 0-0000 (Zone: EQUINIX ZURICH)
- Manufacturer: UNKNOWN
- Model/Type: UNKNOWN
- Generic Type: COMPUTER
- Description: CONNECTION TO EQUINIX ZURICH VIA 10GBPS BY SWITCH
- Tag: IT/CS
- Serial Number:
- Operating System: UNKNOWN Version: UNKNOWN
- CERN Inventory number:
- Network Interface Card(s): ---- PRIVATE INFO ---- [Why?]
- Responsible for the device: CIXP-SUPPORT E-GROUP IT CS
  CIXP-SUPPORT@CERN.CH / Tlf: 72613
- Main User of the device:
- HCP Response: This system **CAN** obtain an IP address automatically [more info]
- IPv6 Ready: This system **IS NOT** IPv6 ready
- Last changed: 01-08-2018 (08:10)

https://network.cern.ch/sc/fcgi/sc.fcgi?Action=SearchForDisplay&DeviceName=cixp-google

Interface(s) Information

Interface Name	IP Address	Service Name
GOOGLE-ZURICH-10G.CERN.CH	192.65.184.202 2001:1458:0:33::2	S513-E-XE4

Late night good news!

Retransfer to NL Region

On Google’s request

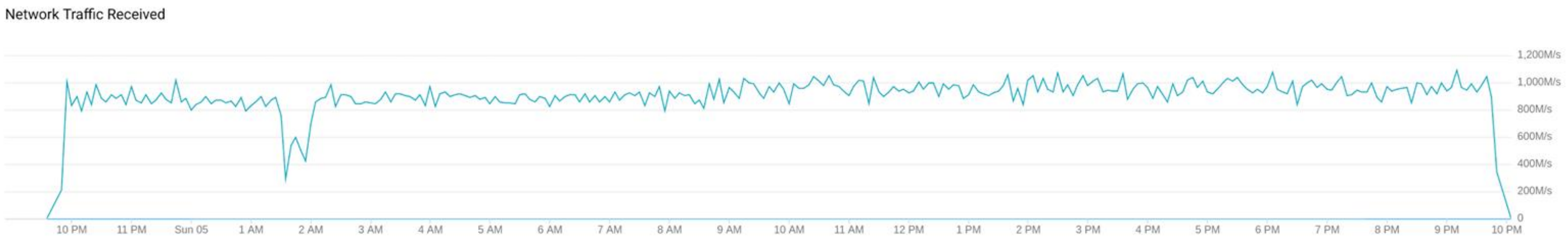
Higher flexibility in terms of available capacity

Why retransfer? Can’t GCP replicate cross regions?

Ingress is free, Ingress is free

(even when running on credits, this counts)

Retransfer to NL Region



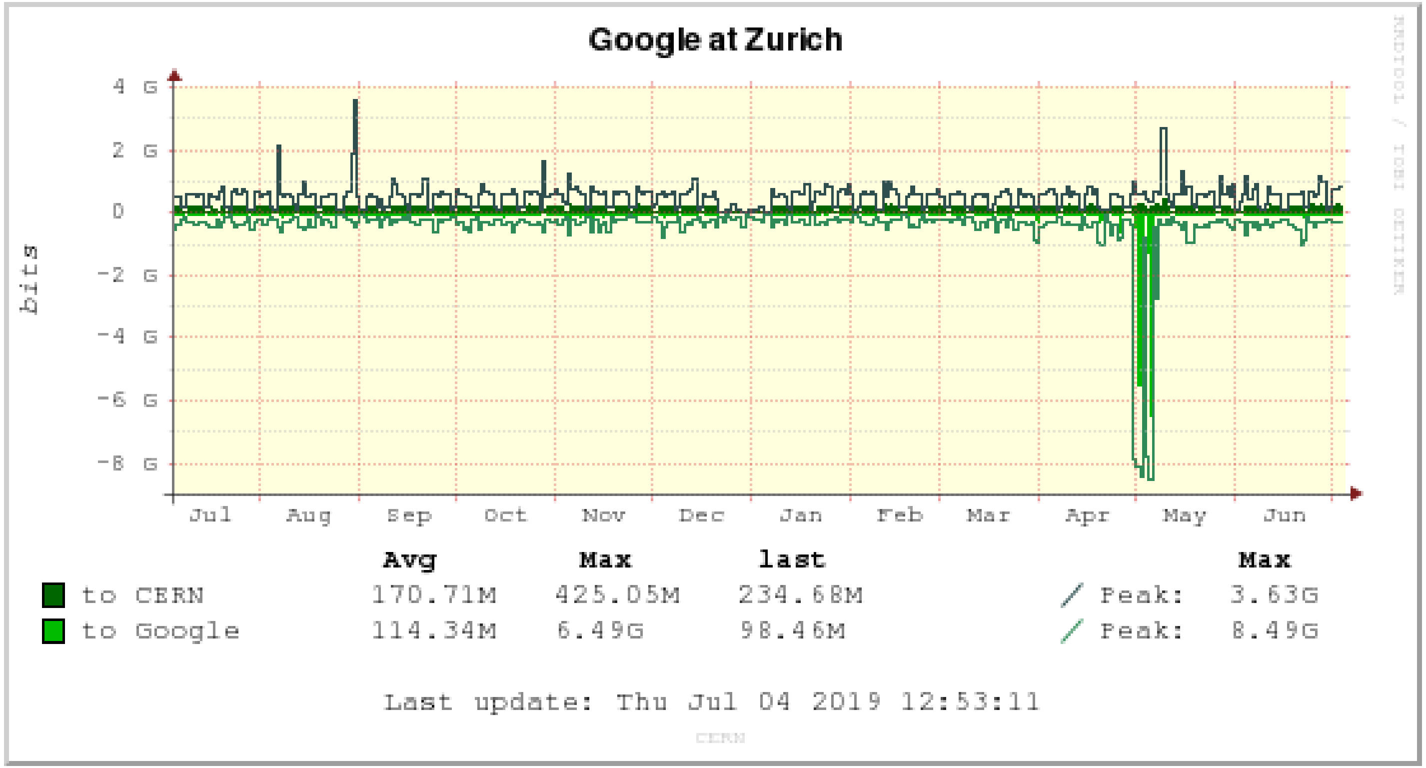
1 day to transfer the full 70TB dataset

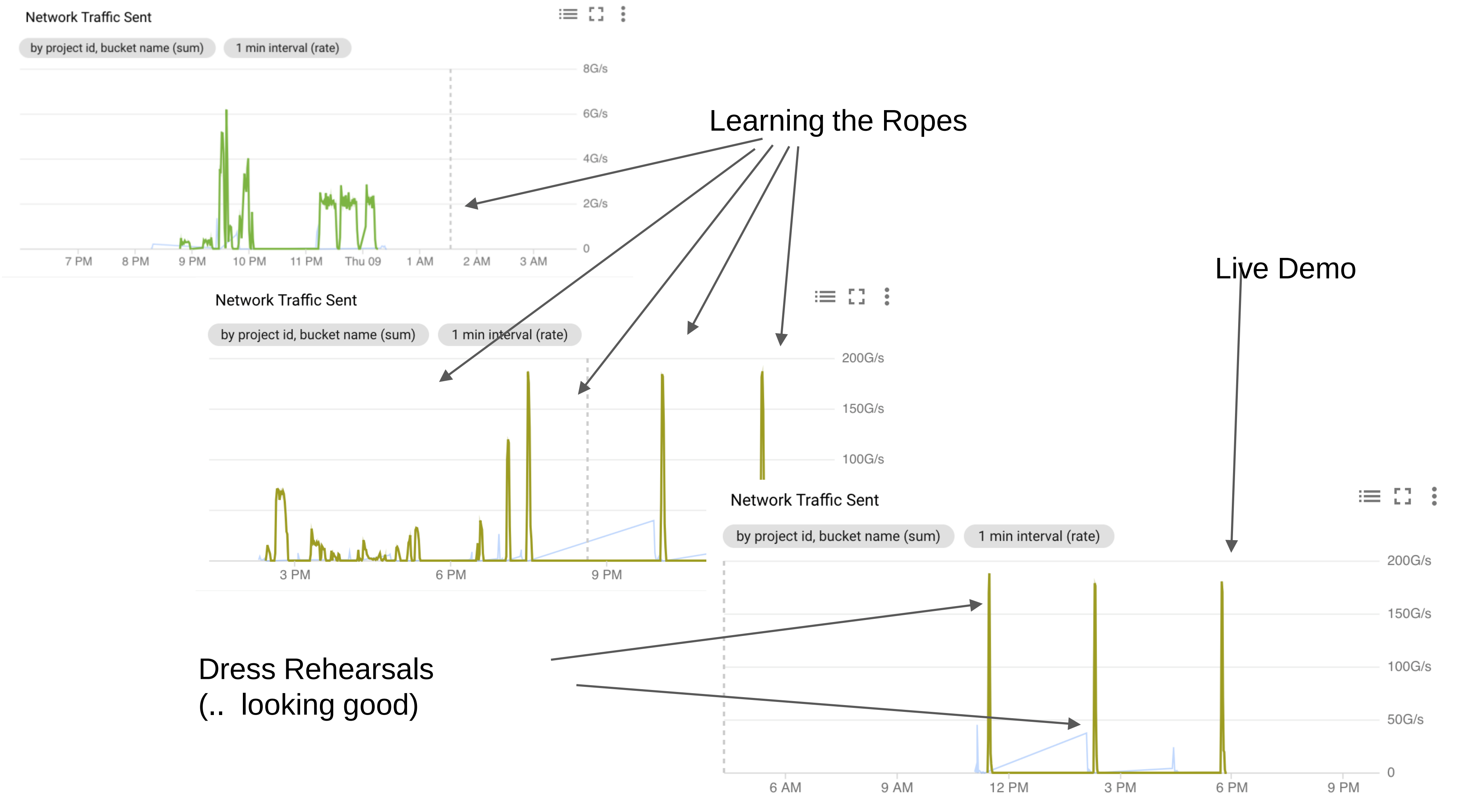
Still going through Google Zurich first

Direct NL/100Gb possible?

Similar rate as for Zurich

Retransfer to NL Region




Network Traffic Sent
by project id, bucket name (sum)
1 min interval (rate)
8G/s
6G/s
4G/s
2G/s
0
7 PM
8 PM
9 PM
10 PM
11 PM
Thu 09
1 AM
2 AM
3 AM
Learning the Ropes
Network Traffic Sent
by project id, bucket name (sum)
1 min interval (rate)
200G/s
150G/s
100G/s
3 PM
6 PM
9 PM
Live Demo
Network Traffic Sent
by project id, bucket name (sum)
1 min interval (rate)
200G/s
150G/s
100G/s
50G/s
0
6 AM
9 AM
12 PM
3 PM
6 PM
9 PM
Dress Rehearsals
(.. looking good)

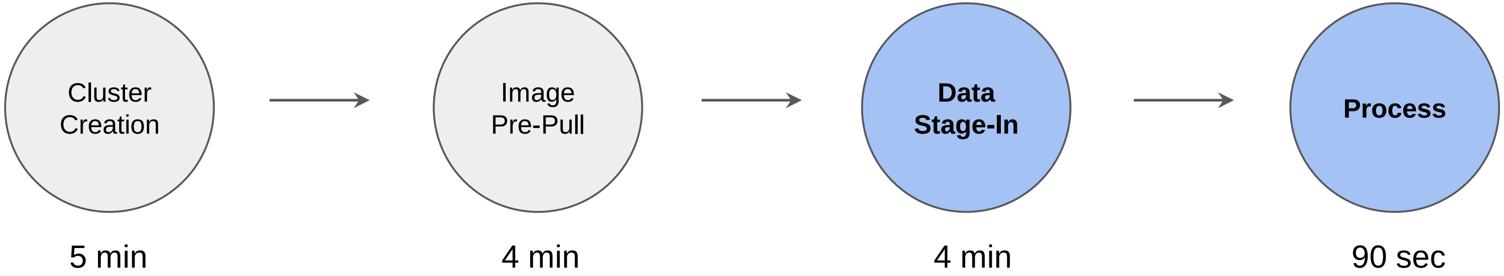
GCP Analysis Run

Kubernetes clusters on GKE (Managed Kubernetes service on GCP)

Today’s run included (real demo run was ~2x that)

660 nodes: n1-highmem-16, 104 GB RAM

10560 cores, 69 TB RAM



Single cluster, slow scheduling

Kubernetes components throttle queries to the api-server

Defaults are very conservative (20 QPS)

Would mean a very slow job scheduling rate (5 pods/sec ?)

- We knew this from previous scale tests we’ve done at CERN
  - Currently we cannot tune this in GKE, coming soon

Decision: split the load into multiple clusters

Storage choices

	Zonal standard persistent disks	Regional persistent disks	Zonal SSD persistent disks	Regional SSD persistent disks	Local SSD (SCSI)	Local SSD (NVMe)
Maximum sustained IOPS						
Read IOPS per GB	0.75	0.75	30	30	266.7	453.3
Write IOPS per GB	1.5	1.5	30	30	186.7	240
Read IOPS per instance	3,000	3,000	15,000 - 60,000*	15,000 - 60,000*	400,000	680,000
Write IOPS per instance	15,000	15,000	15,000 - 30,000*	15,000 - 30,000*	280,000	360,000

Storage choices

We first chose persistent SSDs (~15000 IOPS)

And saw huge amount of io wait, nodes evicted

As we were network and storage service bound at CERN, this was not obvious…

Storage choices

GCP network guarantees 2Gb/core up to 16 core nodes

With 16 core nodes, we get **32 Gb per VM** !

GCS can handle these rates somehow, and we end up bound by local i/o

Network throttling?

```
trickle -s -d $DOWNLOAD_MAX_KB -u $UPLOAD_MAX_KB gsutil cp ... - | cat > $DESTFILE
```

Helps, but not to stay under 10min… local SSDs? That’s 280000 IOPS, enough!

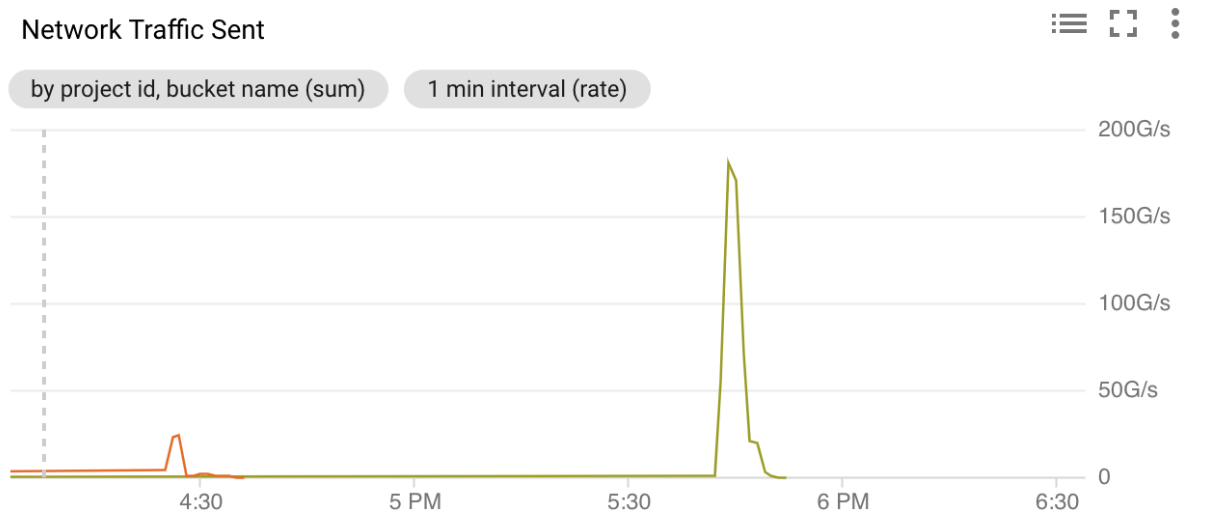
But one only gets full disks (375GB) and it’s a scarce resource

Storage choices

In the end we relied on shared memory for storage

Exploring high memory instances

medium: Memory



GCP Pricing

Billing is updated daily, though there are APIs to query for details

Considering a ~10 minutes run it implies (compute table prices, NL region)

$1.043 * 1530 / 6 = **$260** (~5x cheaper if using pre-emptibles)

Parking storage cost for the dataset (monthly cost, lots of room for creativity)

$0.020 * 70000 = $1400

Total under $300 usd

Running on credits, no Committed Use or Sustained Compute discounts

Open Data, Reproducibility, Reusability

The LHC is a unique machine. Likely that no other machine will probe the same physics regime. Preserving our work for future use-cases is crucial.

Two approaches:

The “Museum”: preserve by archiving / documenting

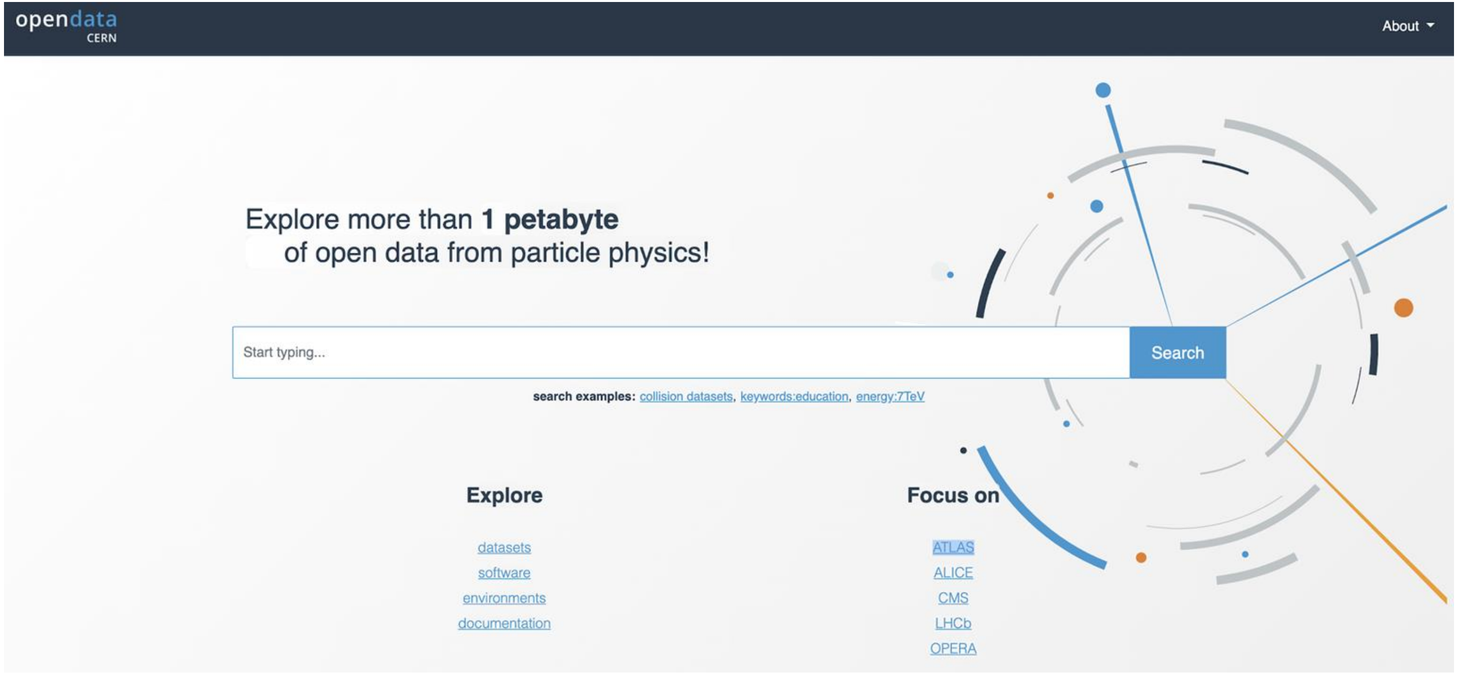
The “Hangar”: preserve to be reused / stay operational


N-X-211
UNITED STATES


Open Data, Reproducibility, Reusability

When Preserving for reusability two choices

- Open Ended New Research (Open Data)
- Reinterpretation of analyzed data (RECAST)

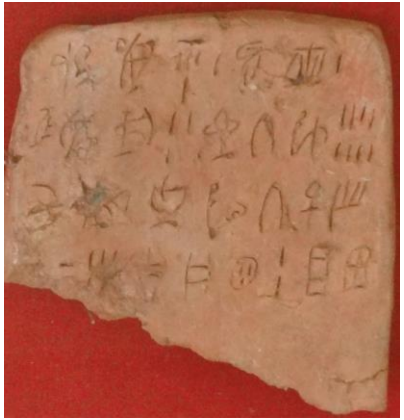


Lot of Love for CERN Open Data. But common reservation: Is it realistic to even analyze PB scale data as a individual without much infra (e.g. non-CERN physicists)?

Demo proved that you can get the necessary scale easily on public cloud. Expect prices to drop.

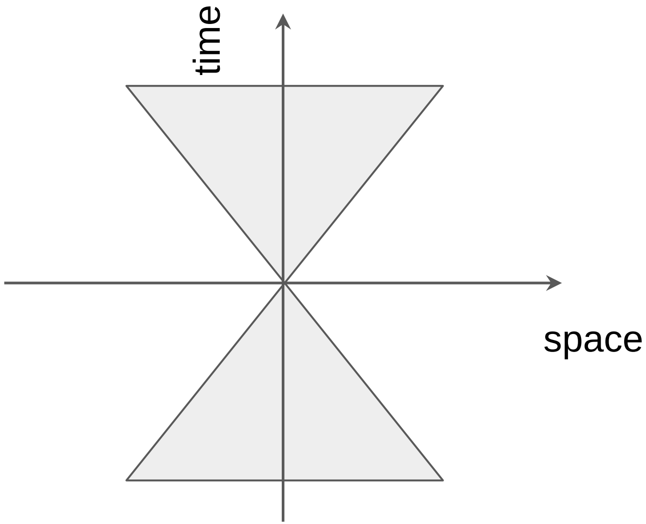
Software Preservation

But data is not enough! Need to have software too!



Preserved, but undecipherable data (Linear A)

What does it mean? Spacetime invariance.

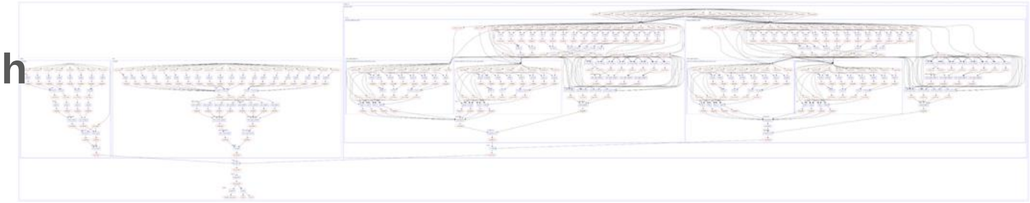


Preserved if you can run it
in the future
in a different data center

Preservation == Reproducible Deployment…

Containerized Workflows

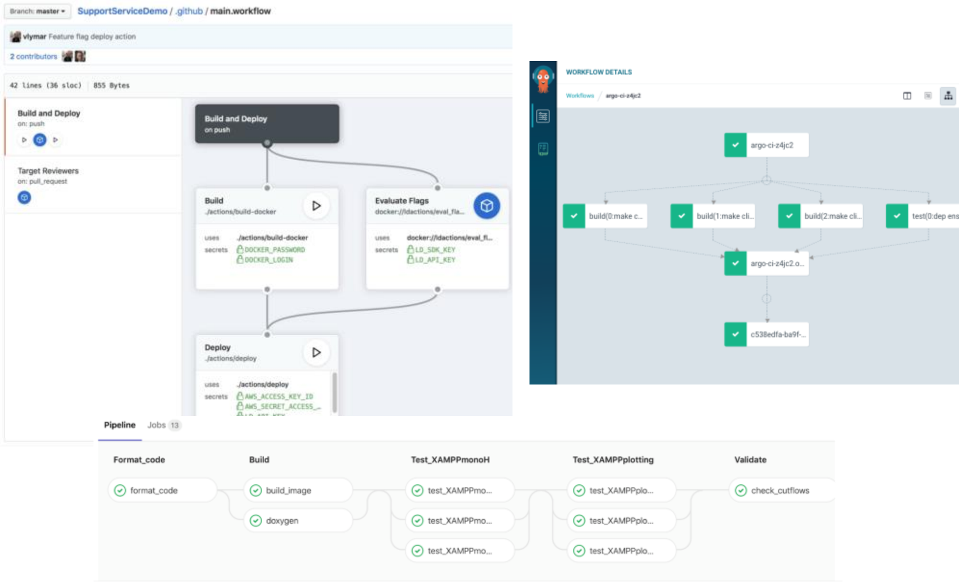
Reproducible research data analysis platform



But Software Preservation is not enough
Need to know what to do with it?

Declarative, cloud-native pipelines are portable way to preserve analyses

Used in CERN Analysis Preservation and REANA to re-execute old analyses



CERN
ANALYSIS PRESERVATION

Credits

Clemens Lange

Thomas Hartland

Google and CERN openlab for the credits and support

And of course the Kubernetes community

Questions?

Jul 1 – 4, 2019	Compute Engine N1 Predefined Instance Ram running in Netherlands: 24588.679 Gibibyte-hours [Currency conversion: USD to CHF using rate 0.977] (Source:Kubecon Demo [nimble-valve-236407])	CHF 112.07
Jul 1 – 4, 2019	Compute Engine N1 Predefined Instance Core running in Netherlands: 3845.211 Hours [Currency conversion: USD to CHF using rate 0.977] (Source:Kubecon Demo [nimble-valve-236407])	CHF 130.78
Jul 1 – 3, 2019	Cloud Storage Regional Storage Netherlands: 6984.399 Gibibyte-months [Currency conversion: USD to CHF using rate 0.977] (Source:Kubecon Demo [nimble-valve-236407])	CHF 136.51