Reperforming a Nobel Prize discovery on Kubernetes

Lukas Heinrich, Ricardo Rocha

CERN

ATLAS and CMS discovered the Higgs boson in 2012.

Since then the LHC has released some of their data publicly.

In this presentation we'll try to reproduce one of the main results from CMS Open Data using modern cloud tools.

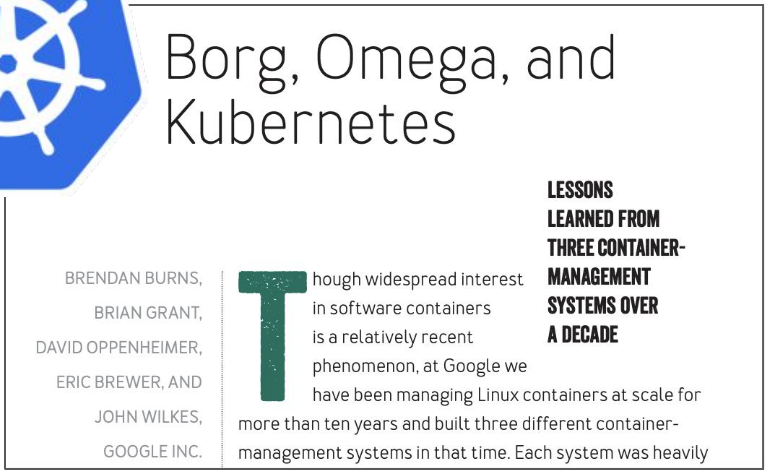

11/22/2013 5:55:18 p.m.
Running jobs: 244151
Transfer rate: 40.08 GiB/sec
WLCG
US Dept of State Geographer
© 2013 Google
Data SIO, NOAA, U.S. Navy, NGA, GEBCO
Image Landsat
Google earth
Fecha de las imágenes: 4/10/2013 66°43'28,18" N 8°52'37,10" O alt. ojo 16085.50 km

Kubernetes

Spun out of Google as an open source

container orchestration project

Built on lessons from Borg and Omega



Borg, Omega, and Kubernetes

LESSONS LEARNED FROM THREE CONTAINER-MANAGEMENT SYSTEMS OVER A DECADE

BRENDAN BURNS, BRIAN GRANT, DAVID OPPENHEIMER, ERIC BREWER, AND JOHN WILKES, GOOGLE INC.

Though widespread interest in software containers is a relatively recent phenomenon, at Google we have been managing Linux containers at scale for more than ten years and built three different container-management systems in that time. Each system was heavily

Loosely coupled collection of components to deploy, maintain and scale workloads

Declarative, Load Balancing, Self Healing, Auto Scaling

Service and Batch Workloads

Kubernetes

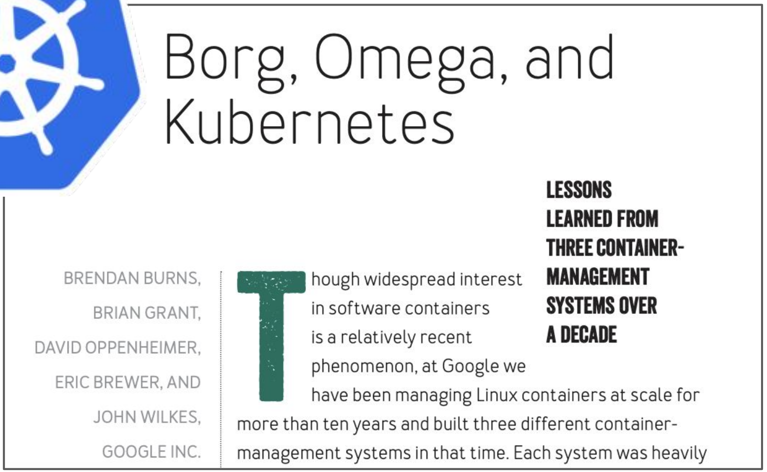
Largest open source project after kernel

- **35.000** contributors, **148.000** code commits
- **83.000** pull requests, **1.1M** contributions
- **2000+** contributing companies

Google, RedHat, VMware, Huawei, Microsoft, IBM, Fujitsu, …

Open community welcome to contributions

Special Interest Groups (SIGs) : Auto-Scaling, Multi-Cluster, Scheduling, ...



Borg, Omega, and Kubernetes

LESSONS LEARNED FROM THREE CONTAINER-MANAGEMENT SYSTEMS OVER A DECADE

BRENDAN BURNS, BRIAN GRANT, DAVID OPPENHEIMER, ERIC BREWER, AND JOHN WILKES, GOOGLE INC.

Though widespread interest in software containers is a relatively recent phenomenon, at Google we have been managing Linux containers at scale for more than ten years and built three different container-management systems in that time. Each system was heavily

Kubernetes

Lingua franca of the cloud

Managed services offered by all major public clouds

Multiple options for on-premise or self-managed deployments

Common declarative API for basic infrastructure : compute, storage, networking

Healthy ecosystem of tools offering extended functionality

Cloud Native Landscape

v0.9.9

App Definition & Development

Database & Data Analytics

BigchainDB, CarbonData, Cassandra, CockroachDB, Couchbase, Crate.io, IBM DB2, Druid, Hadoop, iguazio, MariaDB, MemSQL, mongoDB, MySQL, noms, Oracle Database, Pachyderm, PostgreSQL, Qubole, redis, RethinkDB, ScyllaDB, Spark, SQL Data Warehouse, SQL Server, snowflake, TiDB, Vertica, Vitess

Streaming

Amazon Kinesis, APEX, Dataflow, Flink, HERON, kafka, NATS, nifi, PULSAR, RabbitMQ, StreamSets, STORM

SCM

Bitbucket, GitHub, GitLab, Visual Studio Team Services

Application Definition

apache brooklyn, bitnami, habitat, Packer, OPEN API

CI/CD

AppVeyor, Bamboo, buddybuild, Buildkite, circleci, cloudbees, codefresh, codeship, Concourse, Container Ops, distelli, Drone, GitLab Runner, Jenkins, Octopus Deploy, semaphore, shippable, Solano Labs, Spinnaker, Travis, weave flux, wercker, Deploy

Orchestration & Management

Scheduling & Orchestration

kubernetes — CNCF Project

Amazon ECS, Docker Swarm, MESOS, Microsoft Azure Service Fabric, Nomad

Coordination & Service Discovery

CoreDNS — CNCF Project

Apache Zookeeper, ContainerPilot, Consul, etcd, Netflix OSS Eureka, Hyperbahn, Registrator, SkyDNS, SmartStack, vmware Haret

Service Management

linkerd — CNCF Project

GRPC — CNCF Project

envoy — CNCF Project

3SCALE, Apache Thrift, AVI Networks, backplane, DATAWIRE, F5, HA PROXY, HYSTRIX, KONG, Istio, NGINX, Netflix Ribbon, Turbine Labs, traefik, vamp, Netflix Zuul

Runtime

Cloud-Native Storage

ceph, Container Storage Interface, DATERA, DELL EMC, Diamanti, Hadoop HDFS, GlusterFS, HEDVIG, infinit, LeoFS, libStorage, Manta, MINIO, NetApp, OpenEBS, OpenSDS, portworx, PURE STORAGE, Quobyte, Sheepdog, REX-Ray, ROBIN, ROOK, springpath, StorageOS, SWIFT

Container Runtime

containerd — CNCF Project

rkt — CNCF Project

OPEN (Open Container Initiative), cri-o, Intel Clear Containers, lxd, runV, SmartOS

Cloud-Native Network

CNI — CNCF Project

Aporeto, aviatrix, Big Switch, Project Calico, canal, Contiv, cilium, CUMULUS, flannel, midokura, vmware NSX, nuage networks, OPENCONTRAIL, Open vSwitch, PLUMgrid, ROMANA, snaproute, TIGERA, weave net

Provisioning

Host Management / Tooling

ANSIBLE, CFEngine, CHEF, puppet, RUN DECK, SALTSTACK

Infrastructure Automation

BOSH, CloudFormation, Cloudify, CoreOS, HELM, INFRAKIT, JUJU, kubicorn, ManageIQ, Terraform

Container Registries

Amazon ECR, ATOMIC, Azure Registry, Docker Registry, Google Registry, HARBOR, JFrog Artifactory, Portus, QUAY

Secure Images

Notary — CNCF Project

TUF — CNCF Project

anchore, aqua, BLACK DUCK, clair, NeuVector, OpenSCAP, Sonatype Nexus, Twistlock

Key Management

Confidant, Keywhiz, Knox, Vault

Cloud

Public

aws, Microsoft Azure, Google Cloud Platform, IBM Bluemix, Alibaba Cloud, Oracle Cloud, DigitalOcean, FUJITSU, HUAWEI, Joyent, packet, Tencent Cloud

Private

DIGITAL REBAR, FOREMAN, MAAS, openstack, Scaleway, vmware

Platforms

PaaS / Container Service

APCERA, apprenda, Atomist, Cloud 66, Cloud Foundry, ContainerShip, CONVOX, DATAWIRE, DC/OS, docker, exoscale, Flynn, Galactic Fog, Giant Swarm, Gravitational, heroku, HYPER SH, Iron.io, Jelastic, KONTENA, kubermatic, kublr, Lightbend, MANTL, moby project, navops, nirmata, OPENSHIFT, PaaSTA, PHOTON, PLATFORM9, platform.sh, portainer.io, RANCHER, remind, Scalingo, Stack Point Cloud, TECTONIC, TRITON, tsuru, ubuntu

Serverless/Event-based

APEX, Azure Functions, Cloud Functions, Firebase, fission, Kubeless, AWS Lambda, NStack, nuclio, OPENFAAS, OpenWhisk, Oracle Application Container, PubNub Blocks, serverless, webtask

Observability & Analysis

Monitoring

Prometheus — CNCF Project

AppDynamics, AppNeta, AXIBASE, catchpoint, centreon, Cloud Watch, COBE, COSCALE, DATADOG, dynatrace, Grafana, graphite, influxdata, INSTANA, Iconic, Librato, LIGHTSTEP, meros, Nagios, NETSIL, New Relic, NODESOURCE, opsclarity, opbeat, Outlyer, OpenTSDB, RIEMANN, Rocana, sensu, SENTRY, Server Density, signal fx, StackDriver, StackRox, StackState, StatsD, sysdig, WAVEFRONT, weave cloud, ZABBIX

Logging

fluentd — CNCF Project

elastic, graylog, loggly, logz.io, Loom, papertrail, sematext, splunk, sumologic

Tracing

OPENTRACING — CNCF Project

JAEGER — CNCF Project

Appdash, Cloud Sleuth, Dynatrace APM, ZIPKIN

---

github.com/cncf/landscape

This landscape is intended as a map through the previously uncharted terrain of cloud native technologies. There are many routes to deploying a cloud native application, with CNCF Projects representing a particularly well-traveled path.

CLOUD NATIVE COMPUTING FOUNDATION

Redpoint

Amplify PARTNERS

Greyed logos are not open source

Kubernetes and containers at CERN

Started offering Mesos/DCOS, Swarm and Kubernetes

Now only Swarm and Kubernetes

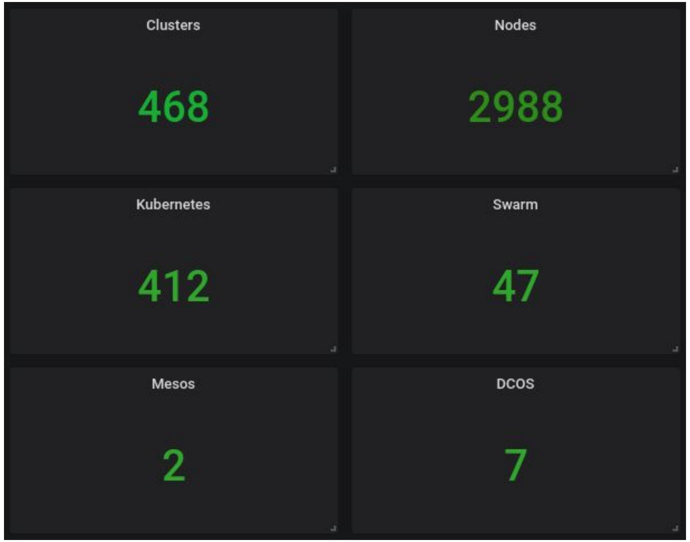
Kubernetes by far the most popular solution

Spark as a Service, WebLogic, JIRA

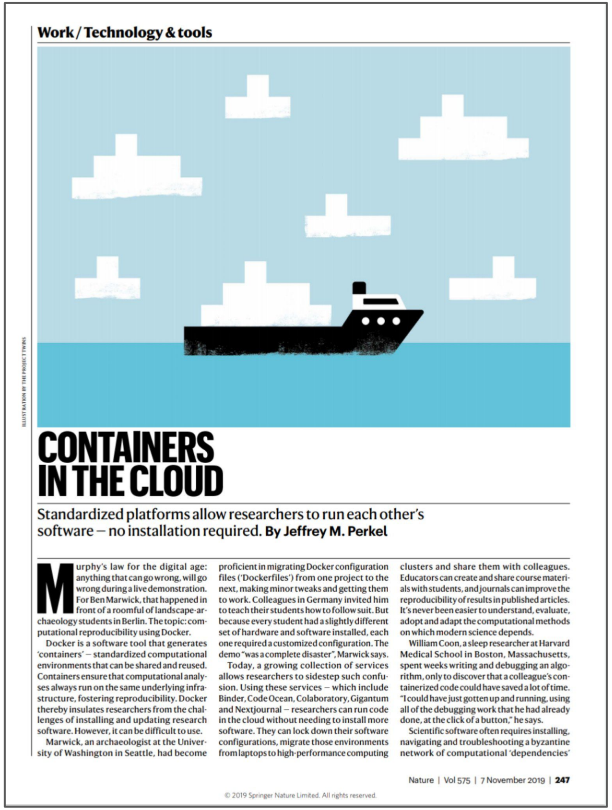
INSPIRE-HEP, REANA/RECAST, Jupyter

OpenStack, Batch / Condor

And many others



Containers & Open Data in Science



nature

Containers & (Open) Science

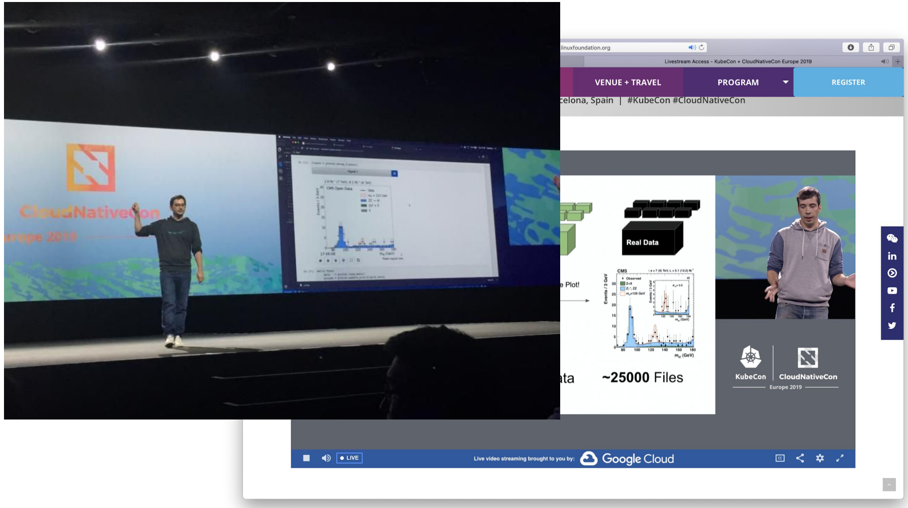
Challenge: H→4l re-discovery on CMS Open Data

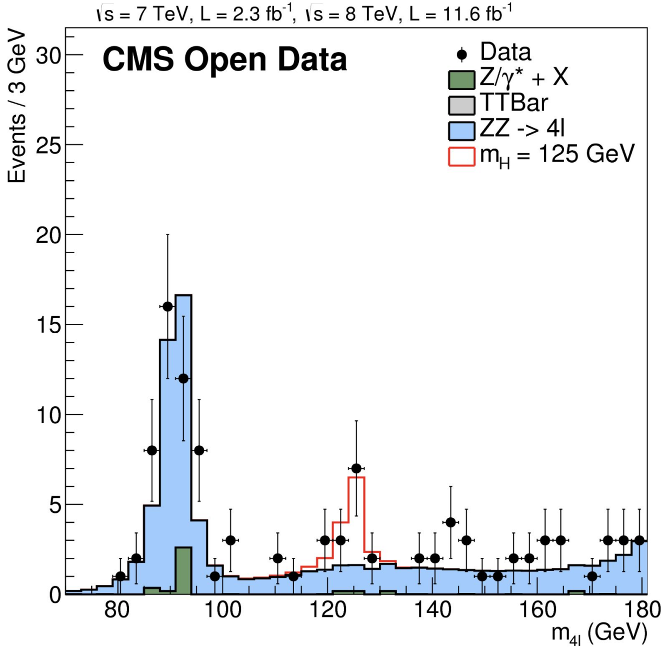
Benchmark analysis based on Open LHC Data.

Goal: Fit it within a live demo for 20-minute Keynote at KubeCon EU 2019

Learn something about cloud-native analysis, reproducibility, Open Data.

Have some Fun.

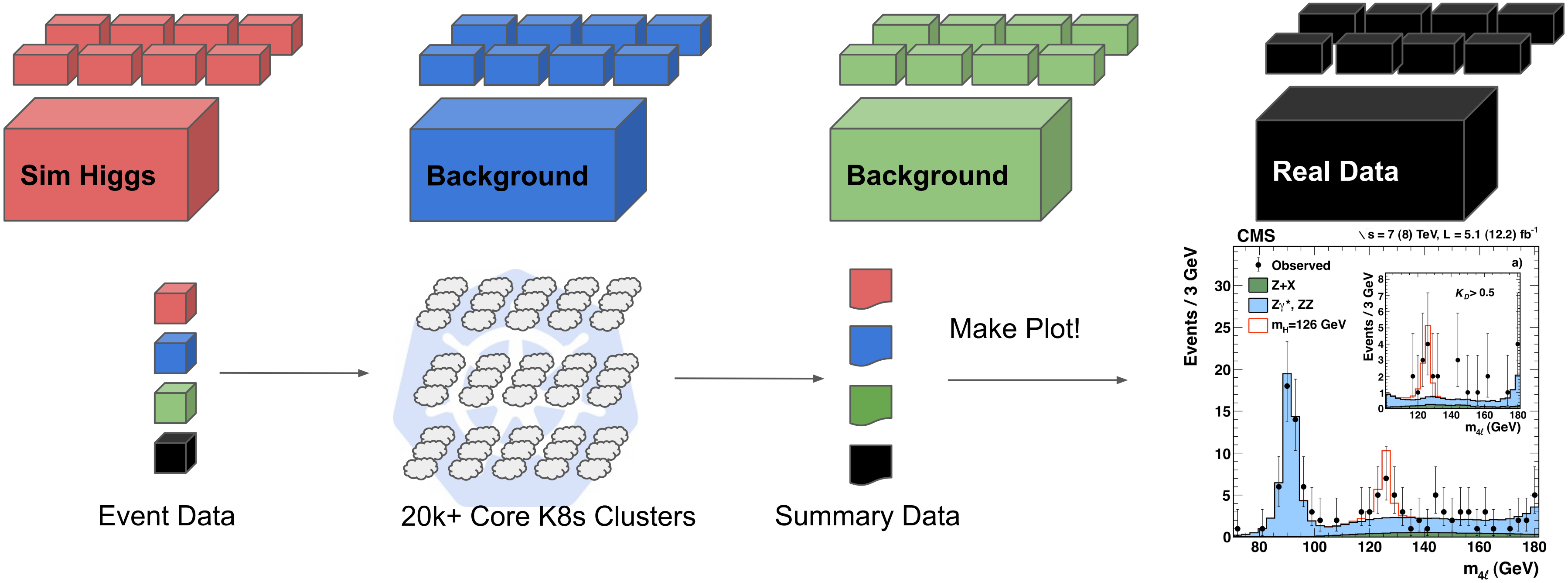




Demo

Challenge: H→4l re-discovery on CMS Open Data

what would this look like in a cloud-native approach?



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

70 TB Dataset

OpenStack Magnum

25000 **Kubernetes Jobs**

Job Results

Interactive Visualization

Aggregation

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

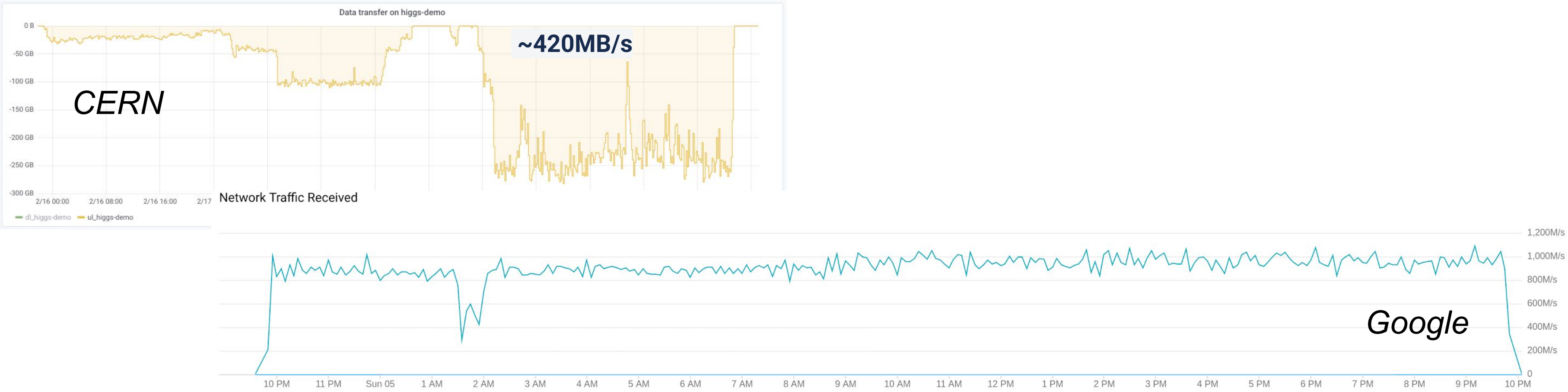
→ jupyter

Interactive Visualization

Aggregation

Data Upload

Initial dataset (opendata) available on /eos

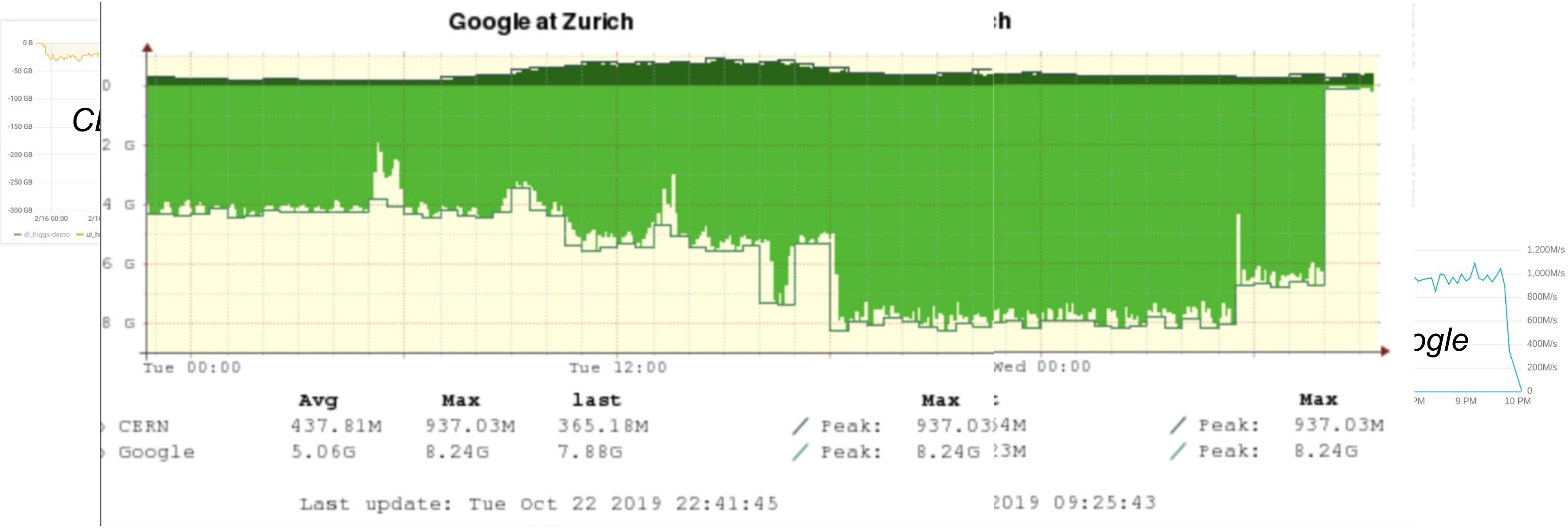


~1 day for full dataset transfer, done first to Zurich then to NL

Ingress is free, Ingress is free...

Data Upload

Initial dataset (opendata) available on /eos



Ingress is free, Ingress is free…

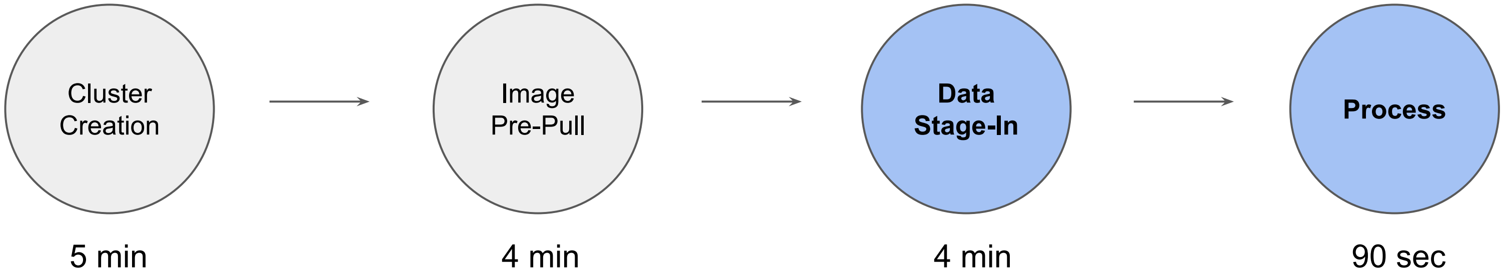
GCP Analysis Run

Kubernetes clusters on GKE (Managed Kubernetes service on GCP)

Today’s run included

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

10560 cores, 69 TB RAM



GCP Analysis Run

Network guarantees 2Gb/core up to 16 core nodes (**32 Gbit per VM** !)

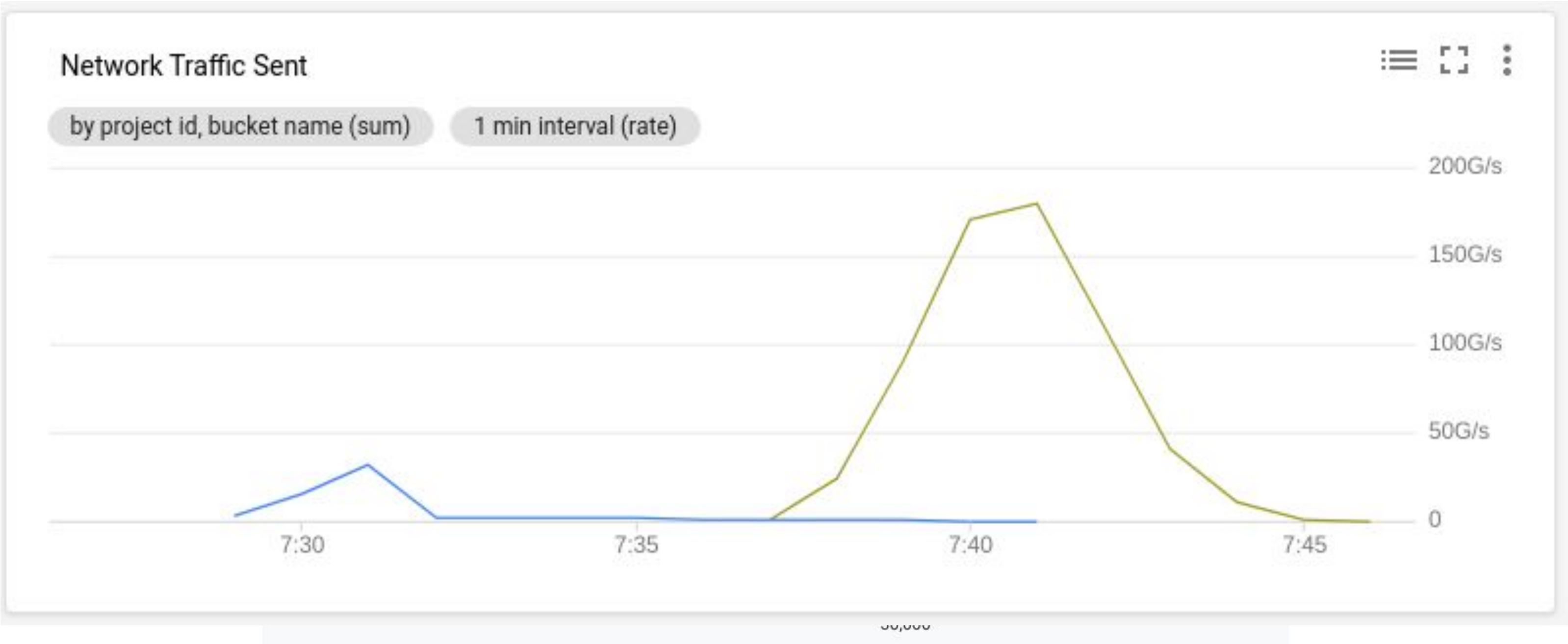
GCS can handle these rates somehow, and we end up bound by local i/o

Ended up using in-memory filesystems to go around this

	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

GCP Analysis Run

Network guarantees 2Gb/core up to 16 core nodes (**32 Gbit per VM** !)



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 Questions

A stunt… or could we come up with a usable model?

Technically feasible. What do these technologies imply for LHC computing?

Analysis Models,
Infrastructure,
Funding Models,
….

Opportunities for Infrastructure

Simplified deployments (Federation), common APIs

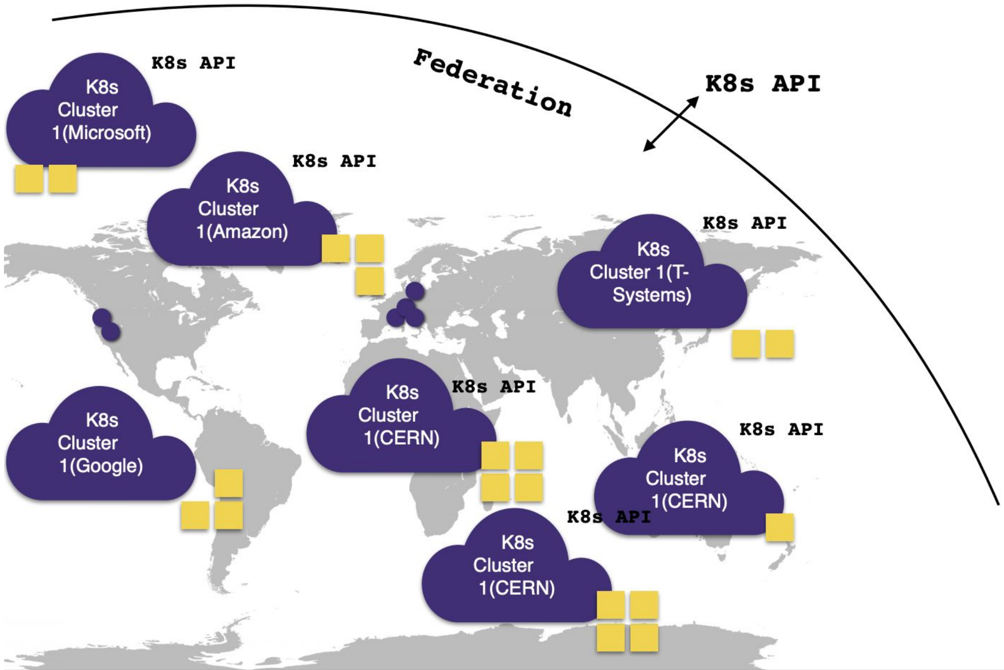
Bursting Scale-out to near-arbitrary scales

Auto-scaling

Access to special hardware

(FPGAs, TPUs, …) easily

Integrated into LHC computing



Opportunities for Analysis Models

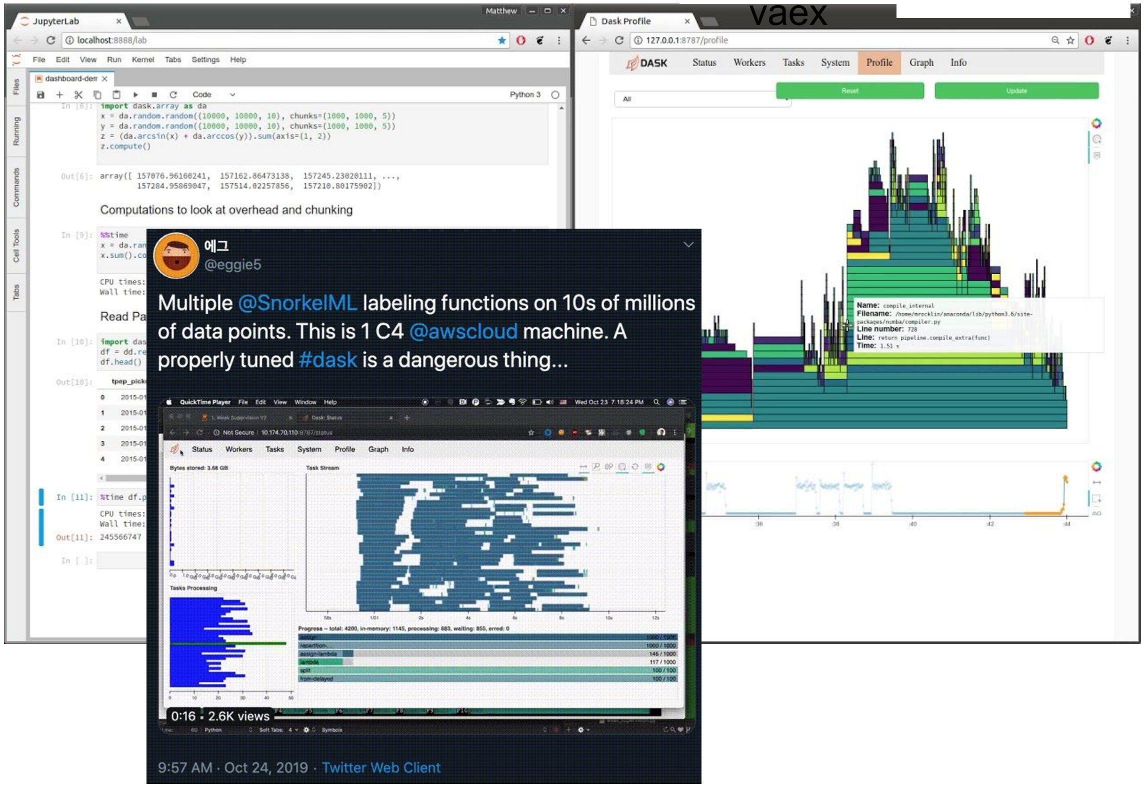
Rich gateway to scale-out, adaptive, on-demand computing

out-of-core dataframes

- Many systems natively integrate w/ k8s
- Rich real-time monitoring

How do you move smoothly Between real-time analysis and batch/scheduled work?

- k8s supports both well



Open Data accessible to everyone at scale

LHC experiments part of growing list of experiments with complex open data problems: data complexity, data volume. large collaborations.

Open Data only useful, if it is feasible for external researchers to analyze it .

Demo goal: Show that public cloud can provide required scale on-demand.

```
[16:01:21] cmsusr@e6f7bea2253e /Users/lukasheinrich/Code/awesomedemo/higgs-demo/CMSSW_5_3_32/src $ \root -b
  *******************************************
  *                                         *
  *        W E L C O M E  to  R O O T       *
  *                                         *
  *   Version   5.32/00   2 December 2011   *
  *                                         *
  *  You are welcome to visit our Web site  *
  *          http://root.cern.ch            *
  *                                         *
  *******************************************

ROOT 5.32/00 (branches/v5-32-00-patches@42372, Jun 10 2014, 18:26:00 on linuxx8664gcc)

CINT/ROOT C/C++ Interpreter version 5.18.00, July 2, 2010
```

Moving to Cloud-based technologies: analysis preservation as a by-product

Beyond a VM: Containerized CMSSW
~decade old software to reproduce results

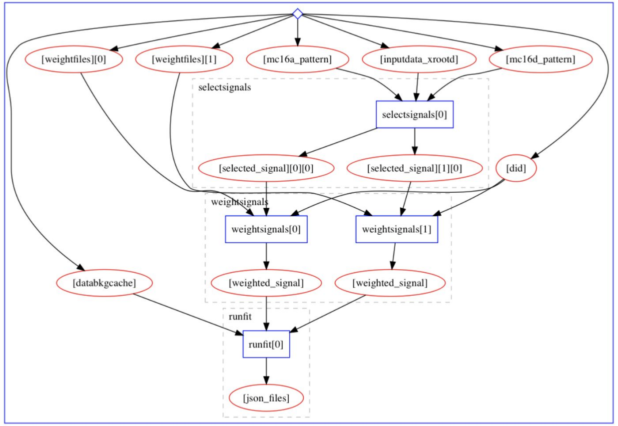
cmsopendata/cmssw_5_3_32 ☆

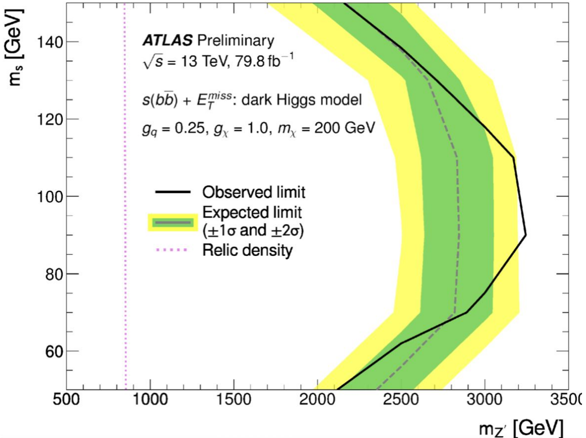
By cmsopendata • Updated 4 months ago

Container

Effective **re-use** of HEP analysis to generate new science results based on archived software.

Only possible through container-based workflows exposed to the user





Conclusions

Demonstrated Tbps analysis of CMS Open Data.

Modern cloud computing paradigms can give individuals scale to realistically analyze LHC data, foster reproducibility & reusability of LHC analyses.

Opens up exciting opportunities to evolve the LHC computing landscape as we look towards Run-4 / the HL-LHC etra. Cross-team collaborations are crucial for R&D[*].

Thought expt: if we started today, what would our infrastructure look like?

We did learn a lot & had some fun.

[*] ATLAS (LH) IT (RR) CMS (Clemens Lange) for R&D towards Run-4