Cloud Technologies in High-Energy Physics Data Processing

Jakob Blomer
CERN, EP-SFT

ICFDT 2016
1st April 2016

Rob Joyce, NSA, Head of Tailored Access Operations

Cloud computing is a really fancy term for someone else's computer.

Rob Joyce, NSA, Head of Tailored Access Operations

Cloud computing is a really fancy term for someone else's computer.

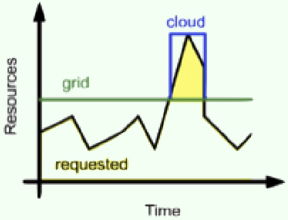
1. High-energy physic's idea of cloud computing is mostly limited to "Infrastructure-as-a-Service" (IaaS)
2. I will focus in this talk at cloud computing in the context of a means to improve our distributed systems ("the grid")

Drivers and Obstacles

1. Cost ☺ (partially)
2. Control & Trust ☹
3. Specialized applications ☹
4. New Technologies for Distributed Systems ☺
   - Virtualization
   - BLOB storage
   - NoSQL databases

Themes

1. Hybrid academic-commercial clouds



Source: Mato

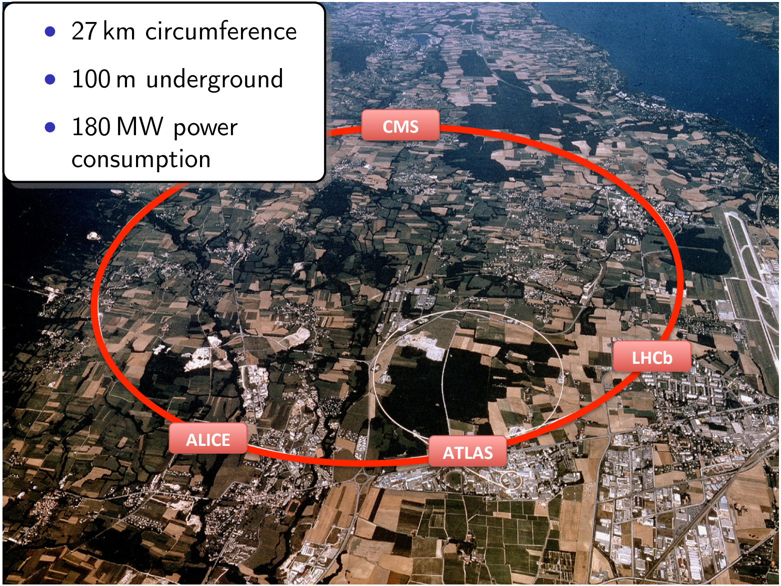
2. Offload mainly simulation (up to 50%), i. e. no data lock-in
3. "Private" adoption of cloud technology
   - OpenStack for virtualization
   - Ceph/RADOS as a BLOB store
   - "Data Mining as a Service"

1. HEP Computing Model
2. Node Virtualization
3. Scientific Application Delivery
4. Volunteer Cloud
5. Cloud Storage
6. Summary & Outlook

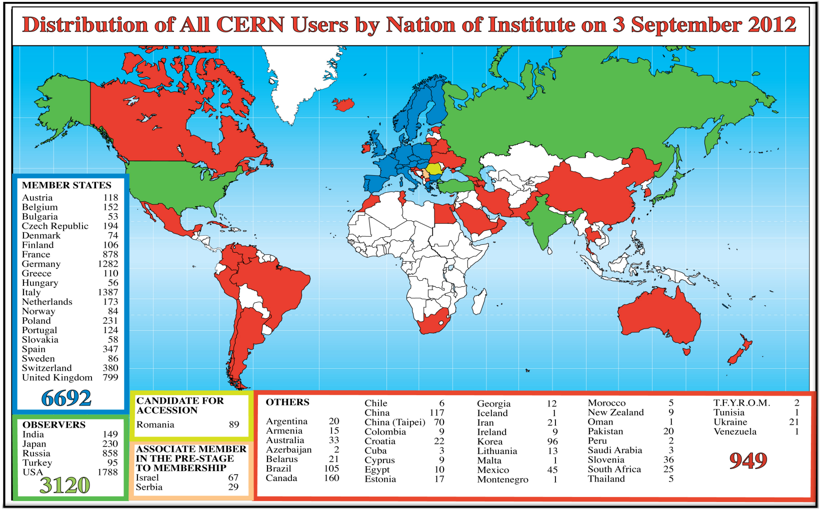
HEP Computing Model

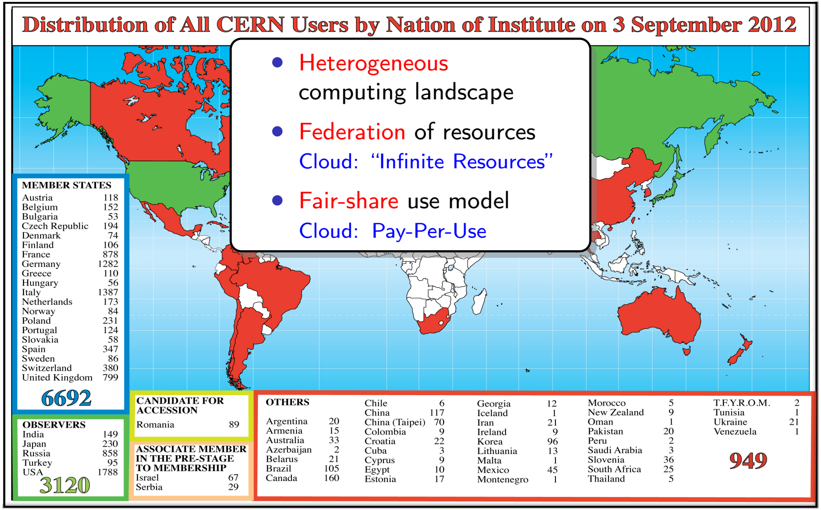
Bird's Eye View of the Large Hadron Collider

- 27 km circumference
- 100 m underground
- 180 MW power consumption



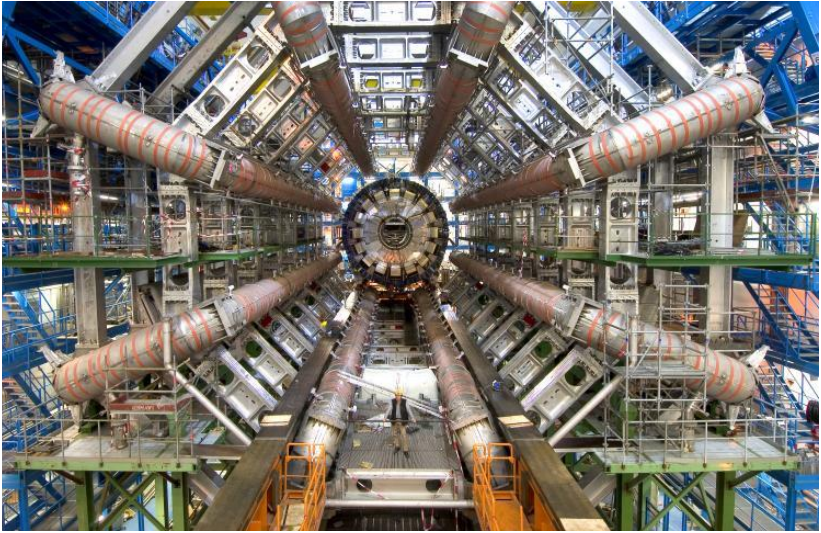
CERN Members and Users





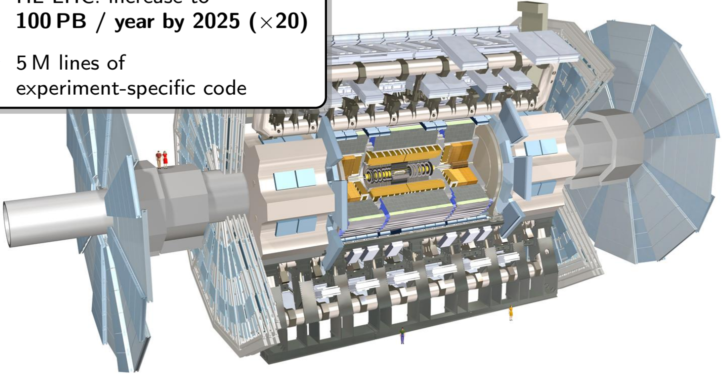
- Heterogeneous computing landscape
- Federation of resources
  Cloud: "Infinite Resources"
- Fair-share use model
  Cloud: Pay-Per-Use

The ATLAS Detector

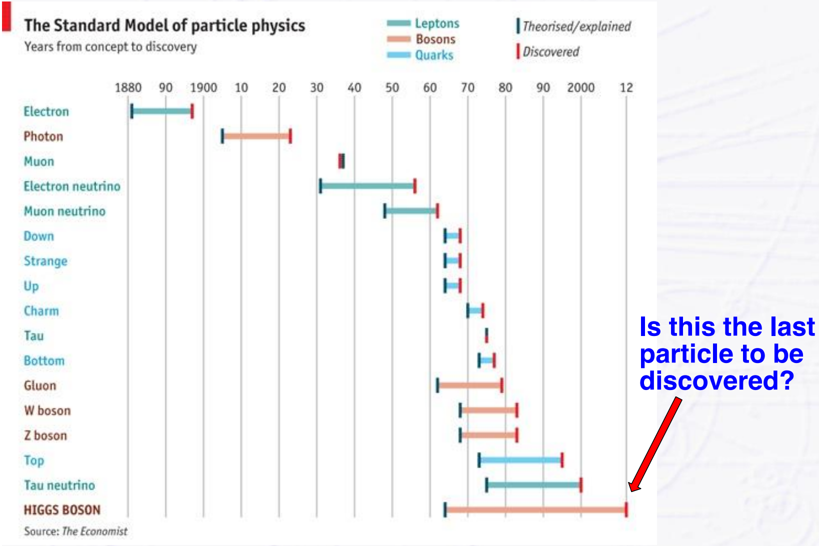


The ATLAS Detector

- 100 Million channels,
  bunch crossing every 25 ns
- 1 PB/s internal data rate
- 5 PB data / year recorded
  (without simulation)
  HL-LHC: increase to
  100 PB / year by 2025 ($\times$20)
- 5 M lines of
  experiment-specific code

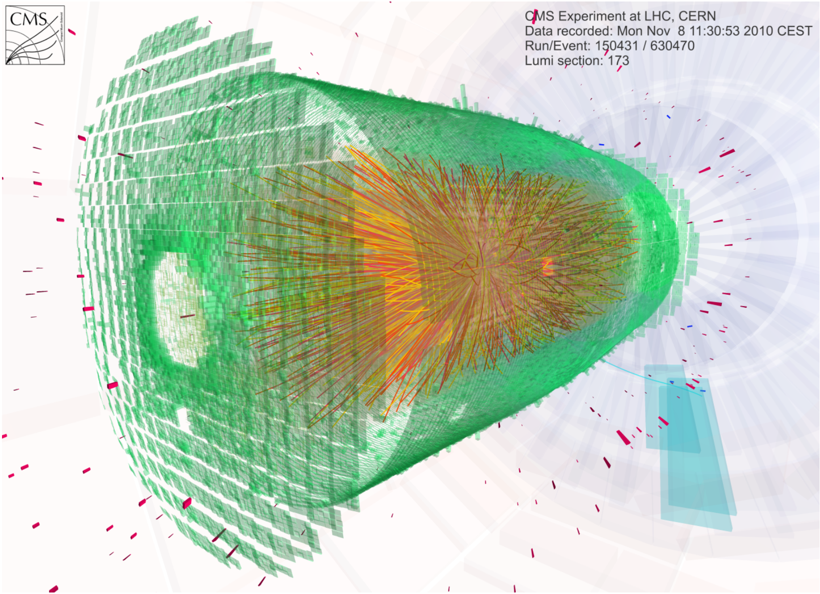


Time Line of the Standard Model



Is this the last particle to be discovered?

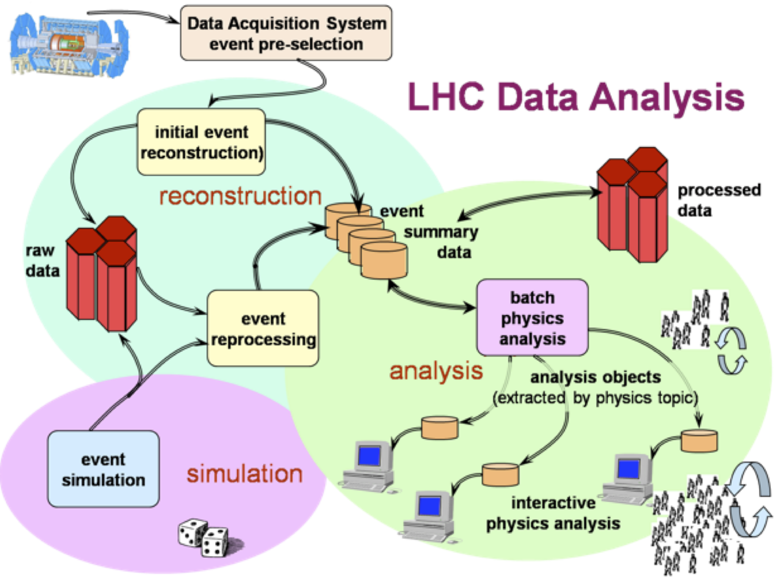
The “Event”


CMS
CMS Experiment at LHC, CERN
Data recorded: Mon Nov 8 11:30:53 2010 CEST
Run/Event: 150431 / 630470
Lumi section: 173

- **Raw Event**
  Megabytes or tens of megabytes per event
  Sensor signals corresponding to a single crossing of beams
- **Event Summary Data (ESD)**
  Hundreds of kilobytes per event
  Reconstructed the physics information from digitized sensor signals. For instance: transformation in global time and space coordinates, reconstruction of trajectories, determination of particles' charge and momentum
- **Analysis Object Data (AOD)**
  Tens of kilobytes per event
  Stripped ESD tailored to specific analysis goals

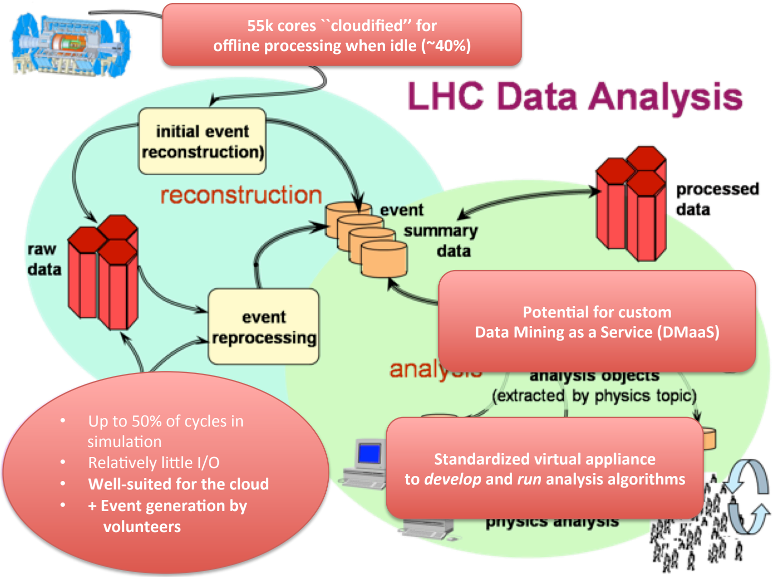
Events are independent from each other,
embarrassingly parallel workload

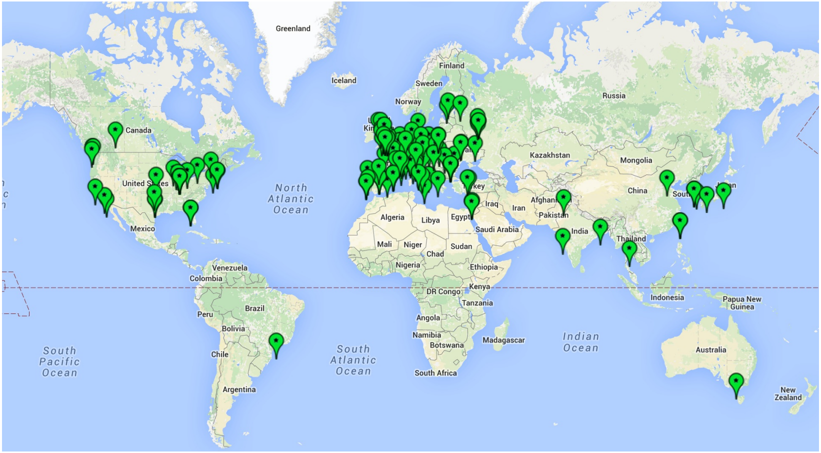
LHC Data Flow



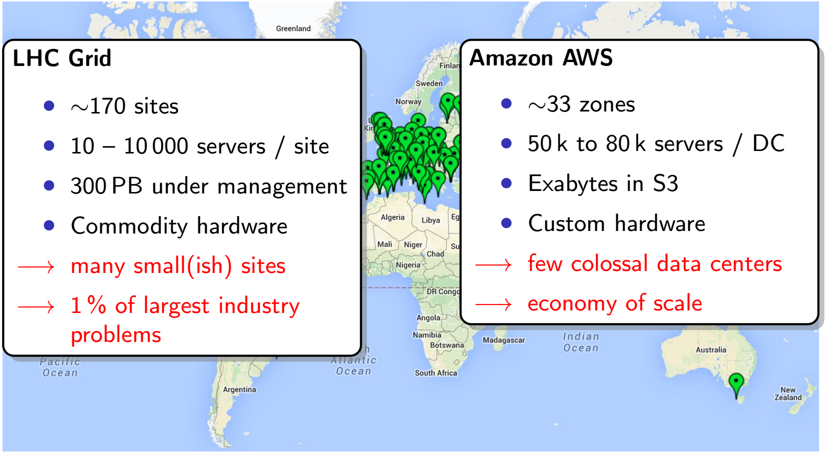
Source: Harvey et al.

LHC Data Flow




Greenland
Iceland
Finland
Sweden
Norway
Russia
Canada
Kazakhstan
Mongolia
United States
North Atlantic Ocean
China
Mexico
Algeria
Libya
Egypt
Iraq
Iran
Pakistan
India
Saudi Arabia
Thailand
Mali
Niger
Chad
Sudan
Nigeria
Ethiopia
Venezuela
Colombia
Kenya
DR Congo
Tanzania
Indonesia
Papua New Guinea
Brazil
Peru
Bolivia
Angola
Namibia
Botswana
Madagascar
Indian Ocean
Australia
South Pacific Ocean
South Atlantic Ocean
Chile
South Africa
Argentina
New Zealand

Source: WLCG



LHC Grid

- ~170 sites
- 10 – 10 000 servers / site
- 300 PB under management
- Commodity hardware

⟶ many small(ish) sites

⟶ 1 % of largest industry problems

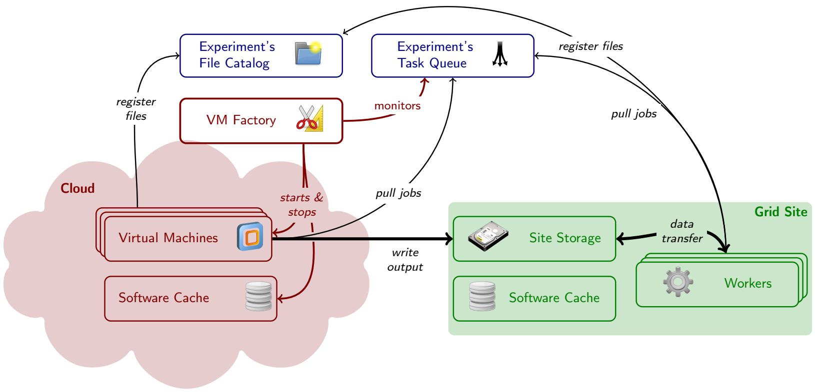
Amazon AWS

- ~33 zones
- 50 k to 80 k servers / DC
- Exabytes in S3
- Custom hardware

⟶ few colossal data centers

⟶ economy of scale

Node Virtualization



Different approaches to "VM Factory":
CloudScheduler VAC WNoDeS

Virtual Performance

- Good VM performance requires careful configuration of hypervisor and guest
- High-performance I/O devices (e. g. 10 Gbit NIC, GPGPU, Xeon Phi, . . .) only usable with direct assignment to VMs
- Just beginning to explore non x86_64 architectures (ARM64, POWER8)

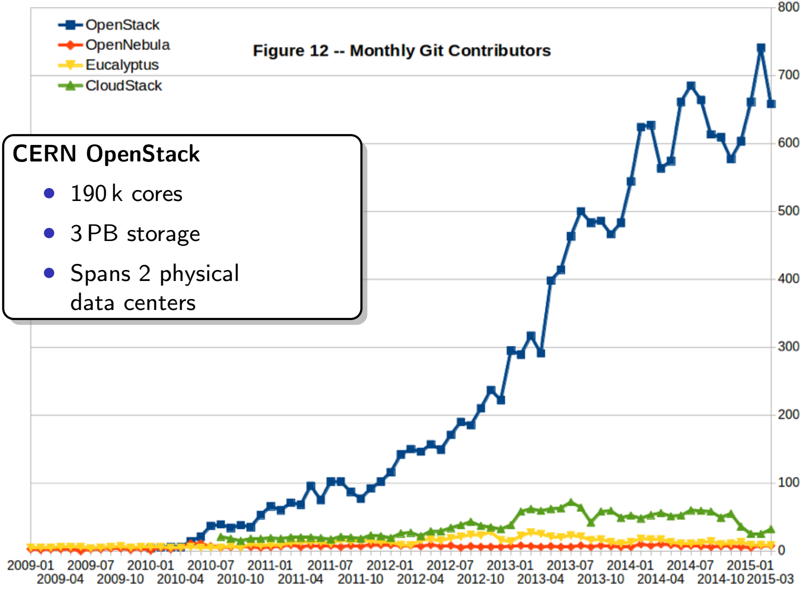
Performance engineering on CERN OpenStack:

VM sizes (cores)	Before	After
4x 8	7.8%	3.3% (batch WN)
2x 16	16%	4.6% (batch WN)
1x 24	20%	5.0% (batch WN)
1x 32	20.4%	3-6% (bare SLC6 ... batch WN)

HEPiX'15

Source: Wiebalck et al.

Containers (e. g. Docker) provide isolated namespaces on top of the same kernel and have almost native performance.

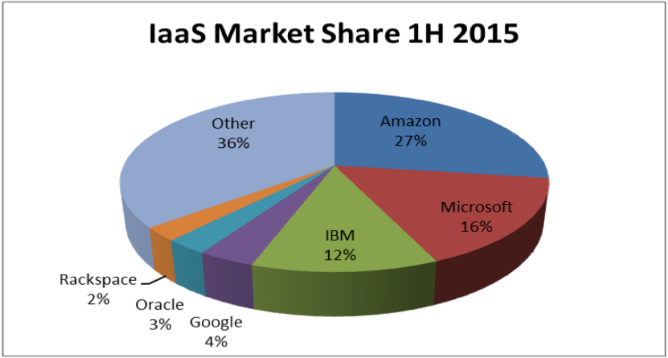


CERN OpenStack

- 190 k cores
- 3 PB storage
- Spans 2 physical data centers

Source: McNab

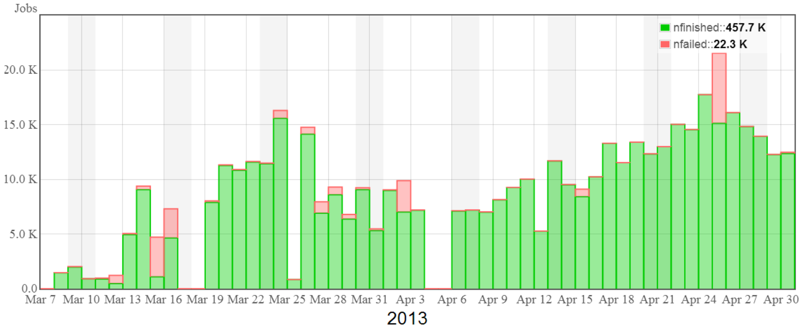
- Amazon (EC2), Microsoft (Azure), IBM (OpenStack,...), Google Cloud Platform, Rackspace (OpenStack/Azure/VMware), ...



- Ralph Finos, http://wikibon.com/public-cloud-market-shares-2014-and-2015/

Source: McNab

Failed and Finished Jobs



- Most of the job failures occurred during start up and scale up phase – as expected
- Reached throughput of 15k jobs per day

Source: Panitkin

Google I/O

Brookhaven National Lab / Amazon EC2

HEPiX'15

- 100 000 cores
- Amazon spot market biddings

CERN / Evaluation of Various Commercial Clouds

HNSciCloud

- HelixNebula Science Cloud:
  EU funded project for federated clouds
- Current tests at a few thousand cores but ramping up
- $\sim$5 k cores allocated on MS Azure
- A few thousand cores on Deutsche Börse Cloud Exchange

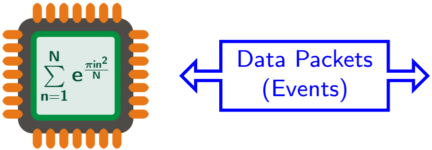
Lessons Learned

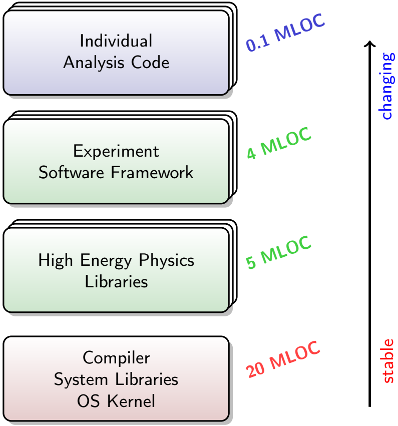
- Cloud resources require supporting services besides the VMs (caches, proxies)
- We need to figure out our cost of computing
- Cloud resources need to be continuously benchmarked
- More layers of complexity that can cause operational issues

Application Delivery

High-Energy Physics Software Stacks

```
> cmsRun DiPhoton_Analysis_cfg.py
```

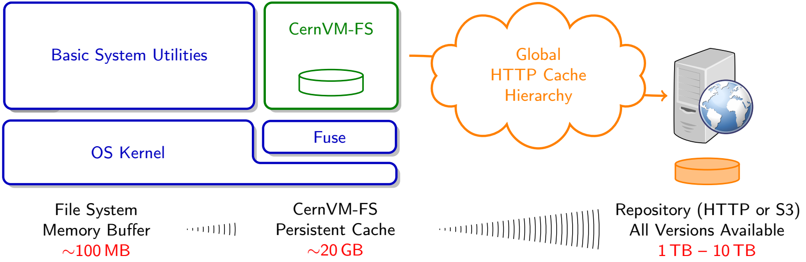




Scientific Software is Special

- Frequent Updates
- Includes its dependencies
- Released versions need to stay available
- Quick turn-around desired from source code commit to deployment

Application Delivery with the CernVM File System

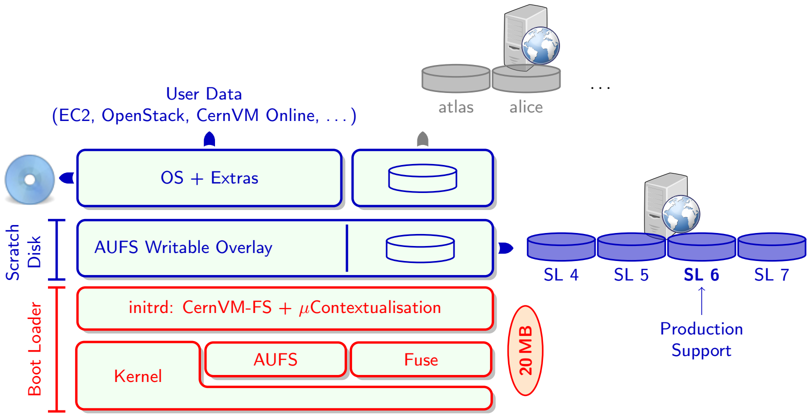


CiSE'15

- Independent *repositories*, e. g. /cvmfs/atlas.cern.ch
- Single point of publishing
- HTTP Transport, access and caching on demand
- $>$ 100 Million files, $\sim$100,000 deployed clients (Grid, Cloud, HPC)

Twofold system: μCernVM boot loader + OS delivered by CernVM-FS

CHEP'13



$\sim$ 30 000 CernVM booted per day

CernVM Hypervisor Support Status

The success of CernVM is largely based on the fact that it runs in practically **all** cloud environments.

Hypervisor / Cloud Controller	Status
VirtualBox	✓
VMware	✓
KVM	✓
Xen	✓
Microsoft Hyper-V	✓
Vagrant	✓
OpenStack	✓
OpenNebula	✓
CloudStack	✓
Amazon EC2	✓
Google Compute Engine	✓
Microsoft Azure	✓
Docker	✓[1]

[1] Release Candidate

CernVM: complete and portable environment
for developing and running HEP data processing tasks

CernVM Appliance Use Cases

CernVM: complete and portable environment for developing and running HEP data processing tasks

Use Cases

1. IaaS Clouds
2. Development Environment
3. Volunteer Computing
4. Long-Term Analysis Preservation
5. Outreach & Education

Infrastructure-as-a-Service Cloud

Various clouds:

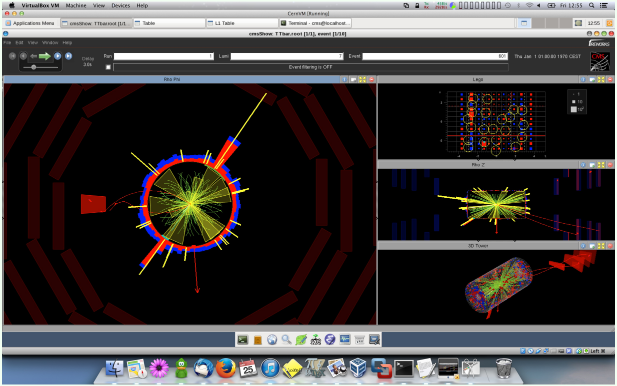
- ATLAS high-level trigger farm
- Cloud resources seamlessly integrated with experiment task queues (e. g. ATLAS CloudScheduler, LHCb VAC)
- Commercial providers
- ALICE software release testing on CERN OpenStack
- ...

CernVM: complete and portable environment for developing and running HEP data processing tasks

Use Cases

1. IaaS Clouds
2. Development Environment
3. Volunteer Computing
4. Long-Term Analysis Preservation
5. Outreach & Education

Interactive Users: Linux Event Viewer on Mac



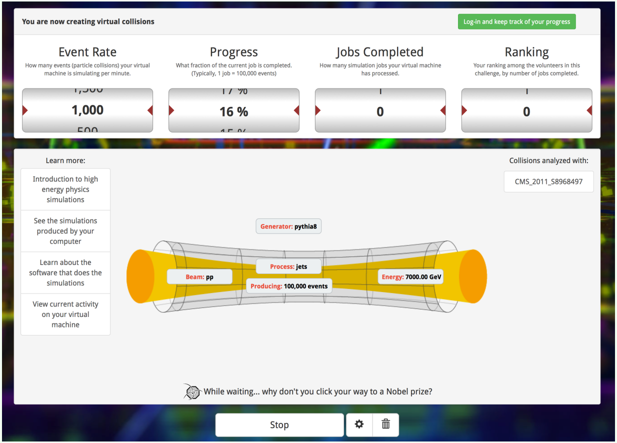
~300 users in 2015, out of which ~65 external users

CernVM: complete and portable environment for developing and running HEP data processing tasks

Use Cases

1. IaaS Clouds
2. Development Environment
3. Volunteer Computing
4. Long-Term Analysis Preservation
5. Outreach & Education

Test4Theory and Computing Challenge



CernVM Appliance Use Cases

CernVM: complete and portable environment
for developing and running HEP data processing tasks

Use Cases

1. IaaS Clouds
2. Development Environment
3. Volunteer Computing
4. Long-Term Analysis Preservation
5. Outreach & Education

ALEPH (1989-2000) software in CernVM

jakob — aleph@cernvm-aleph01:~/test/ALPHA — ssh — 66×18

```
pb-d-128-141-134-74:~ jakob$ ssh -X aleph@cernvm-aleph01
aleph@cernvm-aleph01's password:
[aleph@cernvm-aleph01 ~]$ source setaleph.sh
[aleph@cernvm-aleph01 ~]$ cd test/ALPHA/
[aleph@cernvm-aleph01 ALPHA]$ sh alpha.sh
     ***********************************************
     **********        ALPHA RUN      **** 11.6 ***
     ***********************************************
     ***********************************************

Wed Mar 19 16:10:27 CET 2014

   *****************************************************************
   ***    Compilation  and creation of the makefile 6lep.mk
   *****************************************************************
gmake -f /home/aleph/test/ALPHA/6lep.mk
gmake: `6lep' is up to date.
```

Demonstrates that VMs + Containers can bridge 10+ years

▸ DPHEP Status Report

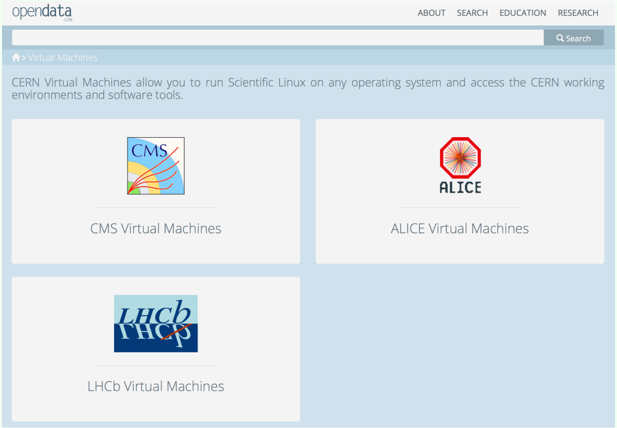
CernVM Appliance Use Cases

CernVM: complete and portable environment
for developing and running HEP data processing tasks

Use Cases

1. IaaS Clouds
2. Development Environment
3. Volunteer Computing
4. Long-Term Analysis Preservation
5. Outreach & Education

CERN OpenData Portal, CERN@School



Volunteer Computing

- Commonly known as …@Home projects
- Computing on *spare* resources of "interested citizens"
- Opportunistic, volatile resources
- Big projects comparable to a TOP500 supercomputer

- Outreach program
- Requires **curation of volunteers**

SETI@Home

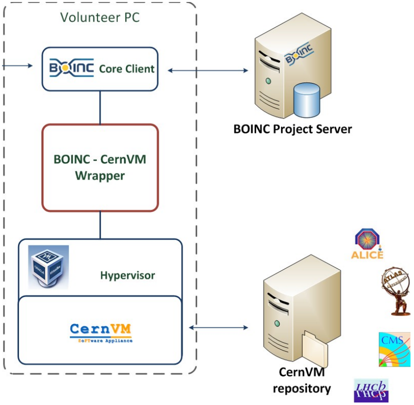
- Search for extraterrestrial life
- 130 000 active participants
- 630 Tera-FLOPS

Einstein@Home

- Search for gravitational waves
- 34 000 active participants
- 470 Tera-FLOPS

LHC@Home 2.0 ACAT'12

- Monte-Carlo simulations, parameter tuning
- First BOINC project using virtual machines, avoids application porting



Source: Harutyunyan

LHC@Home 2.0

- 600–700 VMs connected
- $>$ 2 trillion events created

ATLAS@Home

- 10 k running jobs
- 2nd biggest simulation site

CMS@Home and Beauty@Home under development

Cloud Storage

Thoughts behind Key-Value and BLOB Stores

~2000–2010

SQL DBs and file systems do not scale to the needs of large web services

1. Can we scale out storage (“horizontal scaling”)
   if we give up on the relational data model and the file system hierarchy?
   - Dictionary as a simple, easy to distribute
     yet useful data structure

Scaling

Thoughts behind Key-Value and BLOB Stores

~2000–2010

SQL DBs and file systems do not scale to the needs of large web services

1. Can we scale out storage ("horizontal scaling")
   if we give up on the relational data model and the file system hierarchy?
   - Dictionary as a simple, easy to distribute
     yet useful data structure

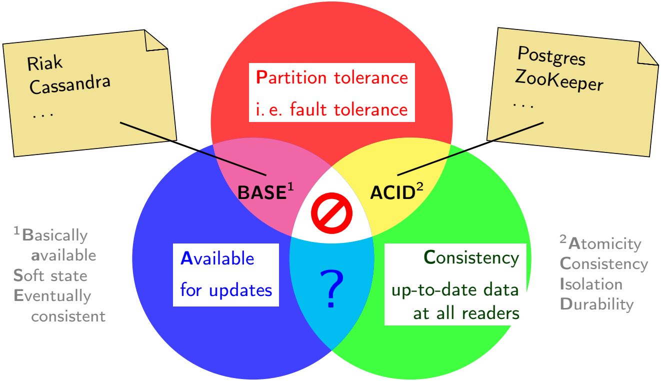
   Scaling

2. In the presence of inevitable faults, can we still make progress ("availability")?
   - Boosted by Amazon Dynamo paper SOSP'07
   - Idea of "eventual consistency":
     heal data inconsistency once the system recovers from faults

   Fault-Tolerance

Even though the *interface* is simple, the *implementation* of a distributed key-value or BLOB store is highly non-trivial.

A distributed storage system can have at most two out of three desirable properties ▸ Brewer '97 ▸ Brewer '12



[1] **B**asically **a**vailable **S**oft state **E**ventually consistent

[2] **A**tomicity **C**onsistency **I**solation **D**urability

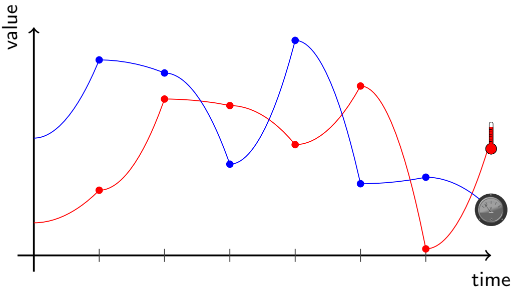
The tradeoffs between availability and consistency can be granular and subtle
For instance: a disconnected ATM might still allow small withdrawals

Examples of NoSQL Databases in HEP

- CMS evaluated Cassandra, MongoDB, Riak
  CHEP'15
- ALICE evaluated RAMCloud and Riak
  Note

Conditions Data

- State of the detector at the time of data taking
- Small (terabytes / year) but critical to reconstruct data
- Traditionally kept in SQL databases



Workflows *Application Logic*

- Virtual machines and volumes
- CERNbox (like Dropbox) replacing some use cases of OpenAFS ▸ CHEP'15
- Zenodo (research data sets) ▸ Link
- Internal Storage for CernVM-FS on S3

Building Blocks

Data Provisioning *Object Stores*

- Very good scalability on local networks
- Flat namespace
- Microsoft Azure: 70 PB (2011), Facebook f4: 65 PB (2014)
- Open Source: Ceph / RADOS CERN: 3 PB, tested up to 30 PB

Workflows on Top of BLOB Storage

Workflows *Application Logic*

- Virtual machines and volumes
- CERNbox (like Dropbox) replacing some use cases of OpenAFS CHEP'15
- Zenodo (research data sets)
- Internal S ... M-FS on S3

Traditionally: workflows on top of distributed file systems

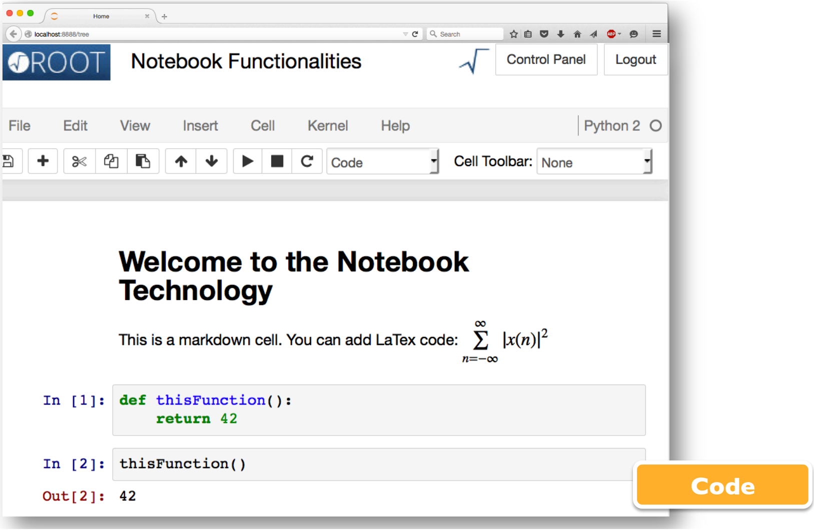
Building Blocks

Data Provisioning *Object Stores*

- Very good scalability on local networks
- Flat namespace
- Microsoft Azure: 70 PB (2011), Facebook f4: 65 PB (2014)
- Open Source: Ceph / RADOS CERN: 3 PB, tested up to 30 PB

Summary & Outlook

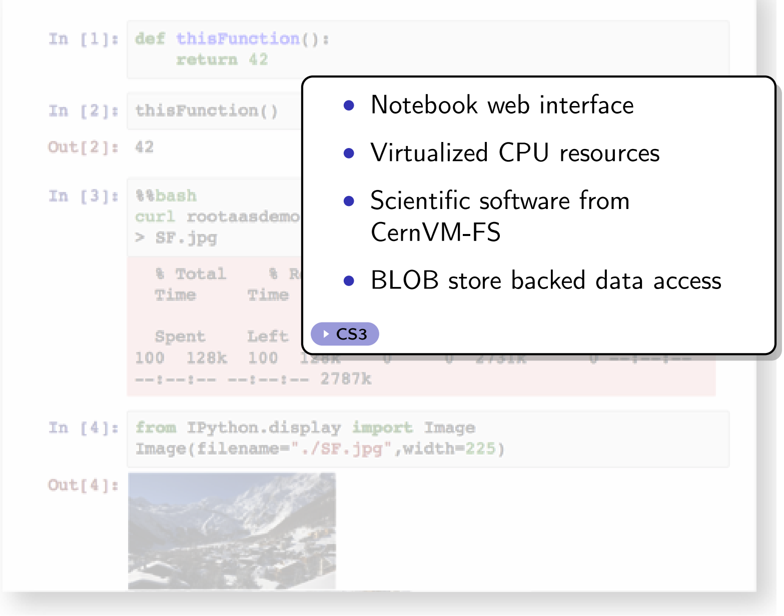
The Future: Custom Cloud Apps?



Source: Piparo

The Future: Custom Cloud Apps?

Source: Piparo



- Notebook web interface
- Virtualized CPU resources
- Scientific software from CernVM-FS
- BLOB store backed data access

CS3

Summary I

Cloud Computing Resources in HEP

- $\sim$5 % of resources provided by virtual machines, mostly for simulation
- Cost of commercial VMs almost on par with in-house cost
- Volunteer clouds: great opportunity for outreach

Cloud Storage Resources in HEP

- Storage: more expensive commercially because of data transfer costs
- Scalable key-value interface requires more application logic
- Very promising open source components: Ceph, Riak, RAMCloud, ...

Did utility computing arrive?

1. On the infrastructure level:
   x86_64 CPUs and *unstructured* storage are commoditiy
2. HEP still has unique and specialised needs and applications:
   - Wide-area data management
   - Custom, optimized I/O format
   - Complex scientific software stacks
   - Legacy software
3. Our custom distributed systems must support
   our custom application workflows
4. We have today much better building blocks available
   than at the time when we designed the grid