

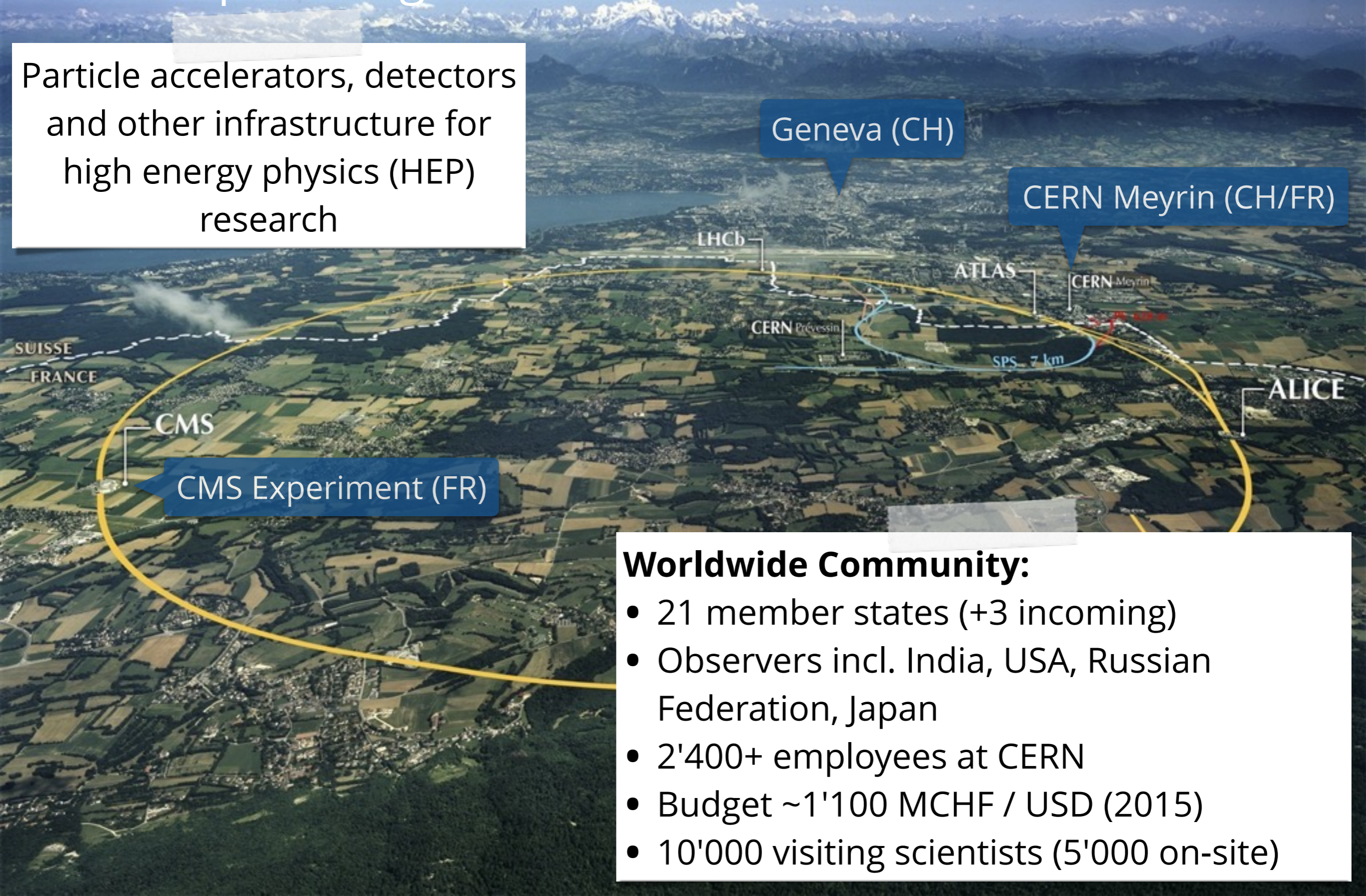
ARM64/AArch64 for Scientific Computing at the CERN CMS Particle Detector

David Abdurachmanov (FNAL)

CERN

The European Organisation for Nuclear Research

Particle accelerators, detectors and other infrastructure for high energy physics (HEP) research



Geneva (CH)

CERN Meyrin (CH/FR)

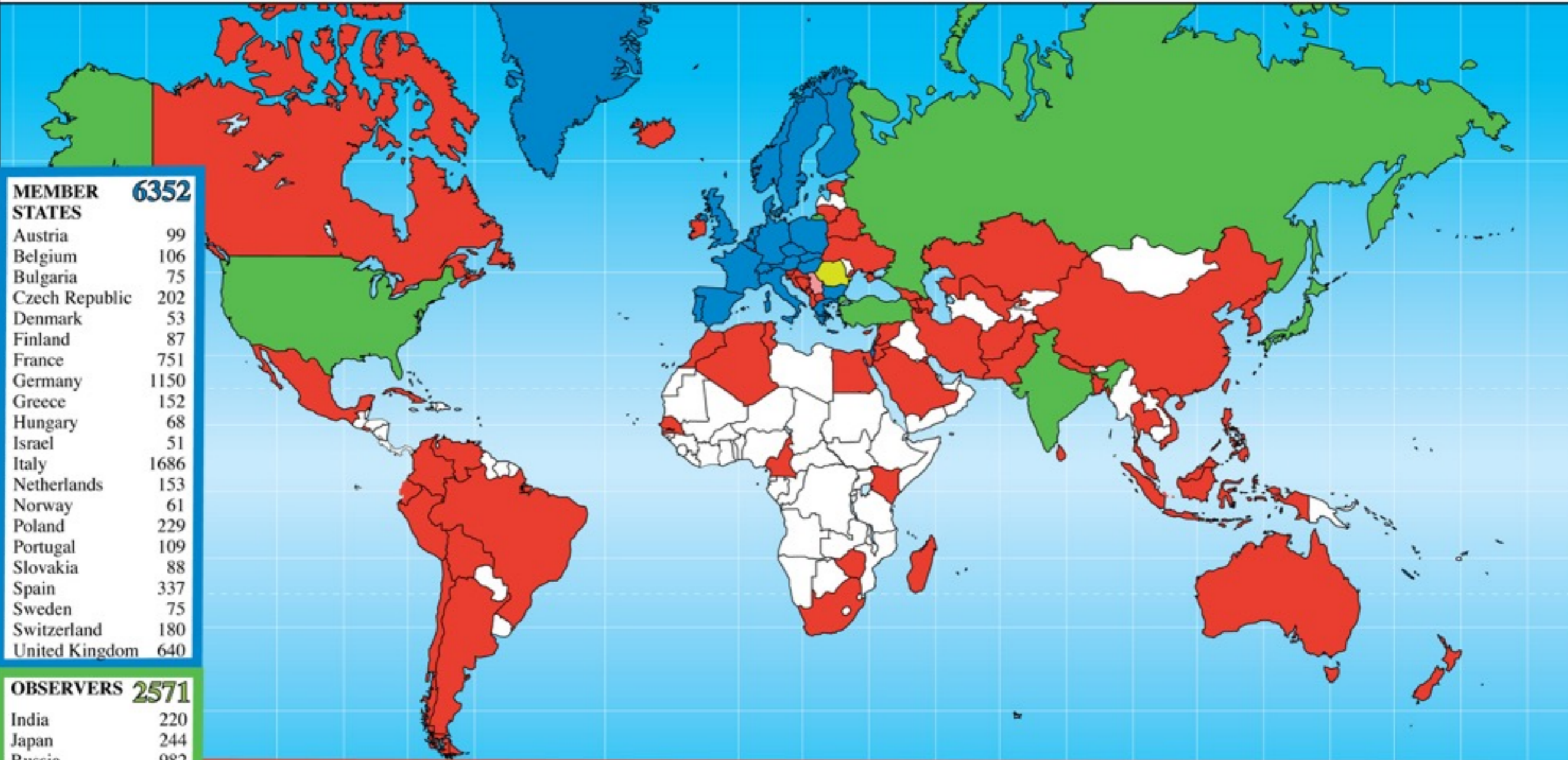
CMS Experiment (FR)

Worldwide Community:

- 21 member states (+3 incoming)
- Observers incl. India, USA, Russian Federation, Japan
- 2'400+ employees at CERN
- Budget ~1'100 MCHF / USD (2015)
- 10'000 visiting scientists (5'000 on-site)

CERN: An International Laboratory

Distribution of All CERN Users by Nationality on 14 January 2014



MEMBER STATES	6352
Austria	99
Belgium	106
Bulgaria	75
Czech Republic	202
Denmark	53
Finland	87
France	751
Germany	1150
Greece	152
Hungary	68
Israel	51
Italy	1686
Netherlands	153
Norway	61
Poland	229
Portugal	109
Slovakia	88
Spain	337
Sweden	75
Switzerland	180
United Kingdom	640

OBSERVERS	2571
India	220
Japan	244
Russia	982
Turkey	146
USA	979

CANDIDATE FOR ACCESSION	
Romania	118

ASSOCIATE MEMBERS IN THE PRE-STAGE TO MEMBERSHIP	
Serbia	41

OTHERS													
Afghanistan	1	Bolivia	3	Cuba	7	Iran	28	Madagascar	4	Philippines	1	Tunisia	6
Albania	2	Bosnia & Herzegovina	1	Cyprus	16	Ireland	22	Malaysia	15	Saudi Arabia	3	Ukraine	55
Algeria	8	Brazil	108	Ecuador	3	Jordan	2	Mauritius	1	Senegal	1	Uzbekistan	4
Argentina	11	Cameroon	1	Egypt	19	Kazakhstan	1	Mexico	64	Singapore	2	Venezuela	9
Armenia	25	Canada	134	El Salvador	1	Kenya	1	Montenegro	3	Sint Maarten	2	Viet Nam	9
Australia	25	Cape Verde	1	Estonia	16	Korea, D.P.R.	1	Morocco	12	Slovenia	27	Zimbabwe	2
Azerbaijan	8	Chile	12	Georgia	36	Korea Rep.	117	Nepal	5	South Africa	16		
Bangladesh	4	China	280	Gibraltar	1	Kuwait	1	New Zealand	7	Sri Lanka	5		
Belarus	47	China (Taipei)	45	Hong Kong	1	Lebanon	12	Pakistan	41	Syria	2		
		Colombia	30	Iceland	4	Lithuania	19	Palestine (O.T.)	4	Thailand	12		
		Croatia	35	Indonesia	1	Luxembourg	4	Peru	8	T.F.Y.R.O.M.	1		

1415

CERN: Tools

The European Organisation for Nuclear Research

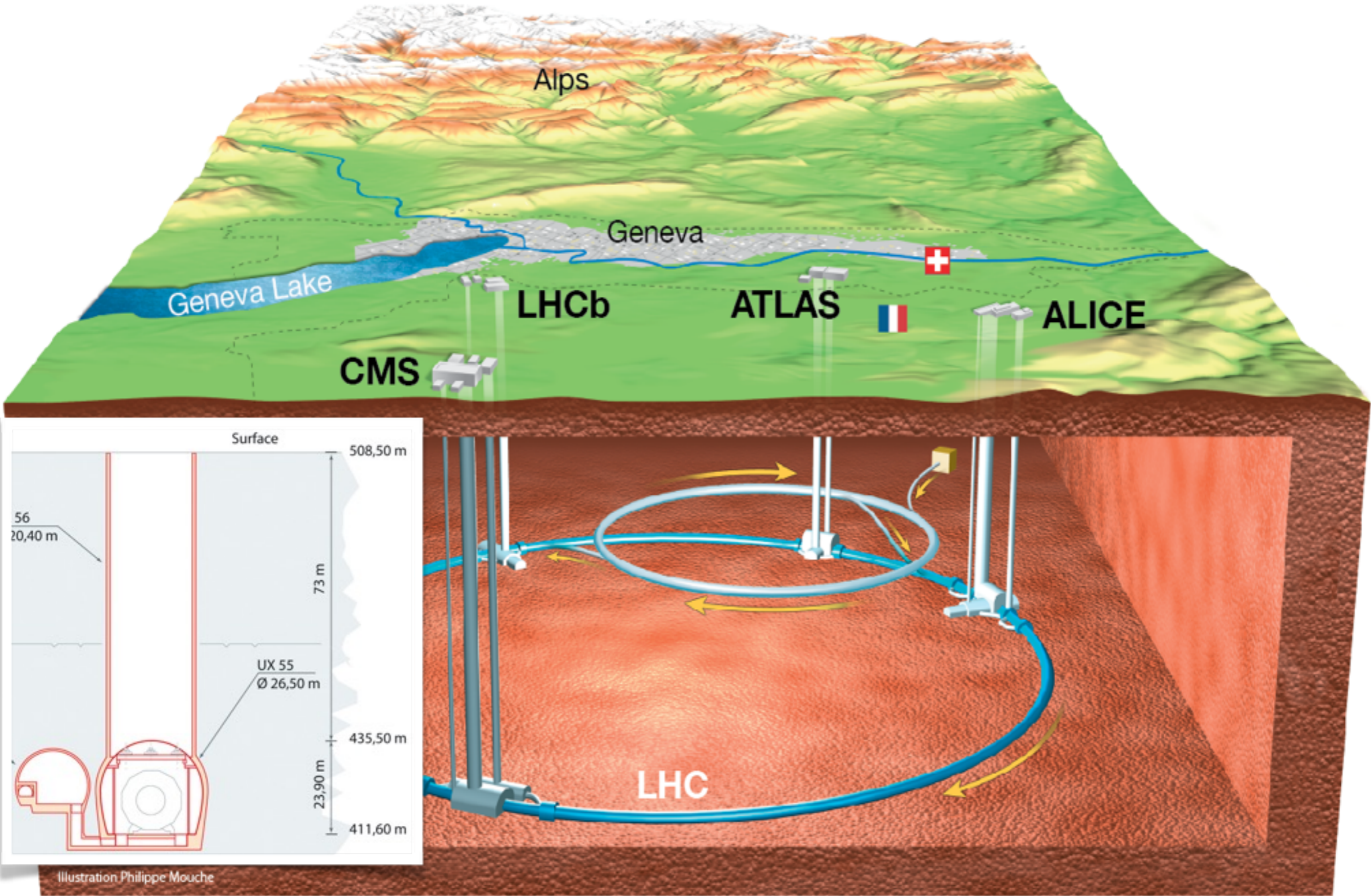


Illustration Philippe Mouche

The LHC Experiments Today

ALICE – “A Large Ion Collider Experiment”

Size: 26m long, 16m wide, 16m high; weight: 10'000 t

35 countries, 118 Institutes

Material costs: 110 MCHF / USD

ATLAS – “A Toroidal LHC ApparatuS”

Size: 46m long, 25m wide, 25m high; weight: 7'000 t

38 countries, 174 institutes

Material costs: 540 MCHF / USD

CMS – “Compact Muon Solenoid”

Size: 22m long, 15m wide, 15m high; weight: 14'000 t

40 countries, 172 institutes

Material costs: 500 MCHF / USD

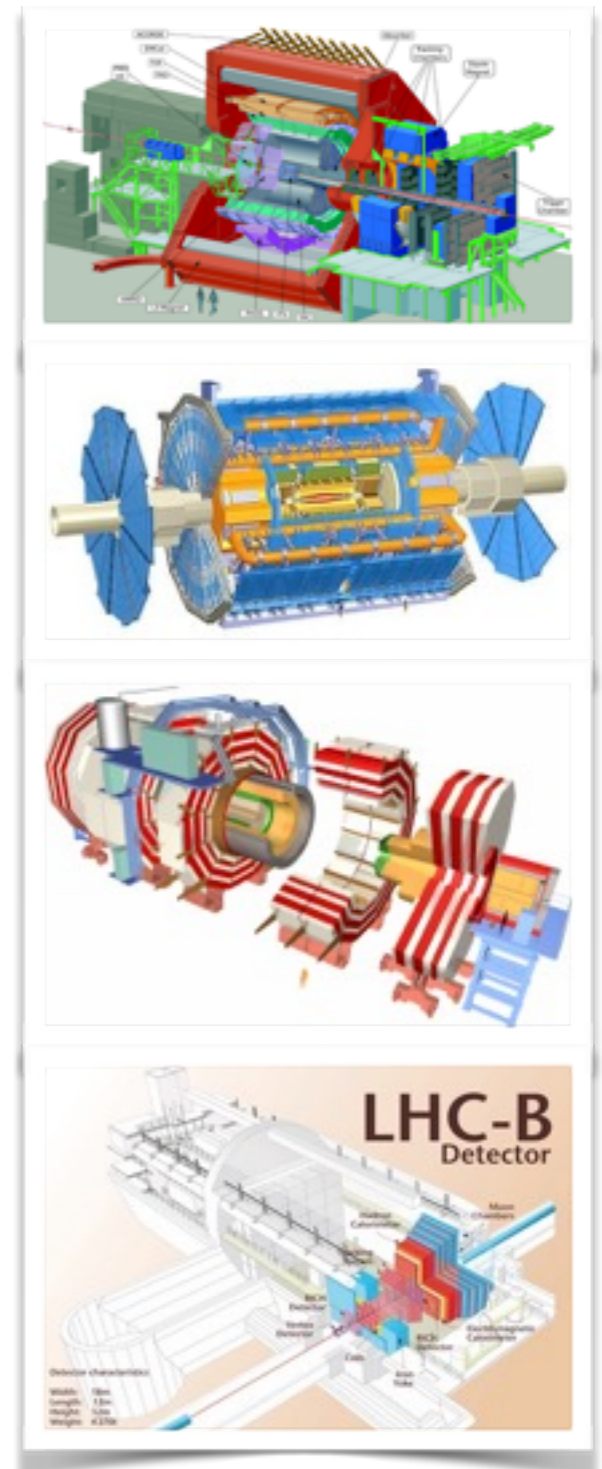
LHCb – “LHC beauty” (b-quark is called “beauty” quark)

Size: 21m long, 13m wide, 10m high; weight: 5'600 t

15 countries, 52 Institutes

Material costs: 75 MCHF / USD

Regular upgrades, first was in 2013/14 (Long Shutdown 1)



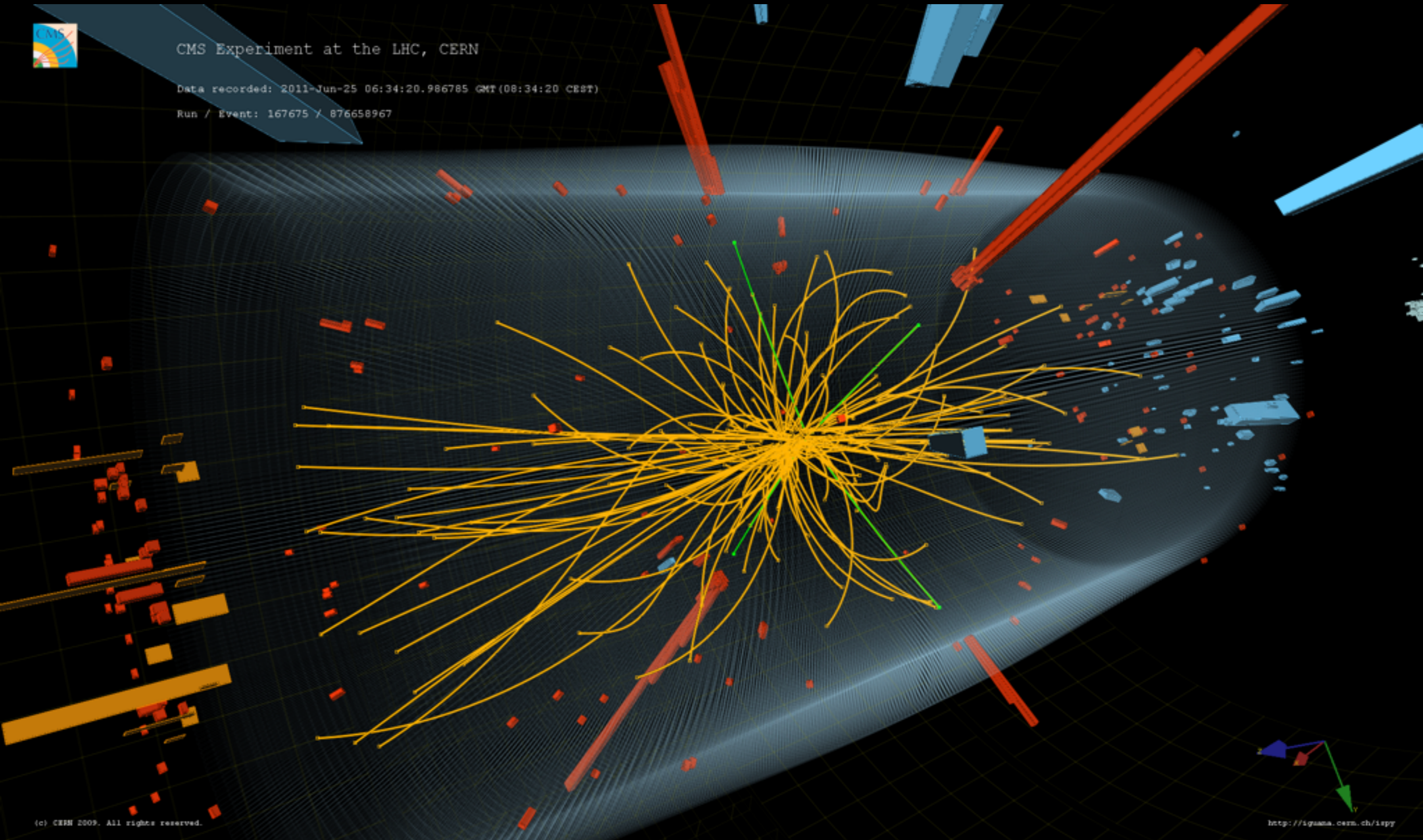
Higgs-boson @ CMS



CMS Experiment at the LHC, CERN

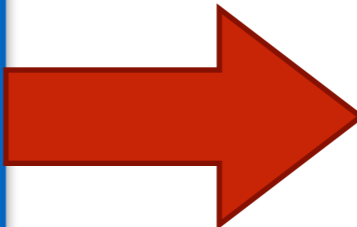
Data recorded: 2011-Jun-25 06:34:20.986785 GMT (08:34:20 CEST)

Run / Event: 167675 / 876658967

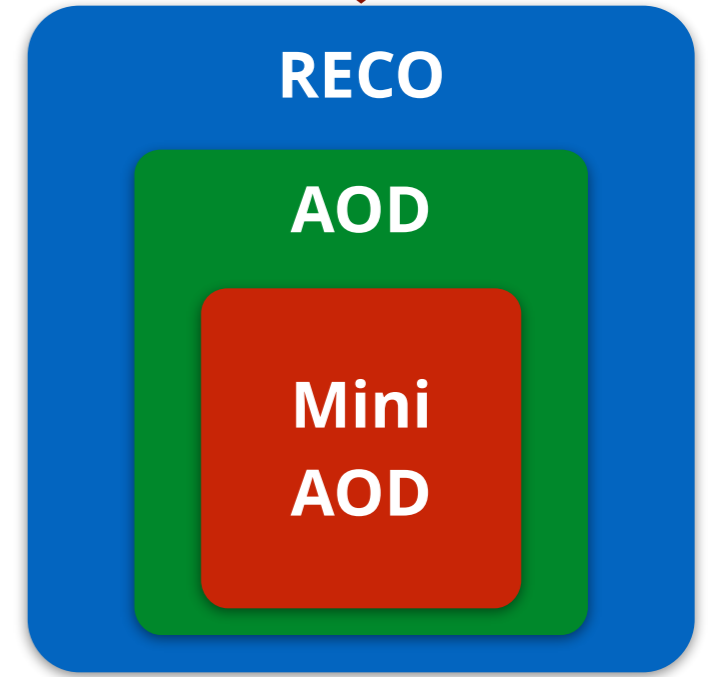
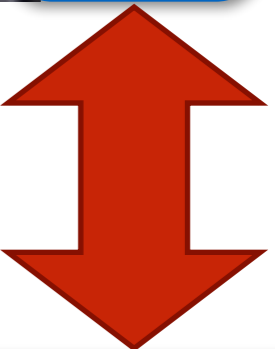


Data and Algorithms

CMS Detector



Worldwide LHC Computing Grid (WLCG)



HEP data is organised as **Events** (particle collisions)

Simulation, Reconstruction and Analysis workloads processes "one event at a time"

Events are independent of each other, thus trivial (embarrassing) parallel processing

Event processing workloads are composed of a number of algorithms selecting and transforming **RAW** event data into reconstructed event data and statistics

RAW - 300KB (ZLIB), **RECO** - 1MB/event (ZLIB), **AOD** - 300KB (LZMA), **Mini AOD** - 30KB/event (ZLIB)

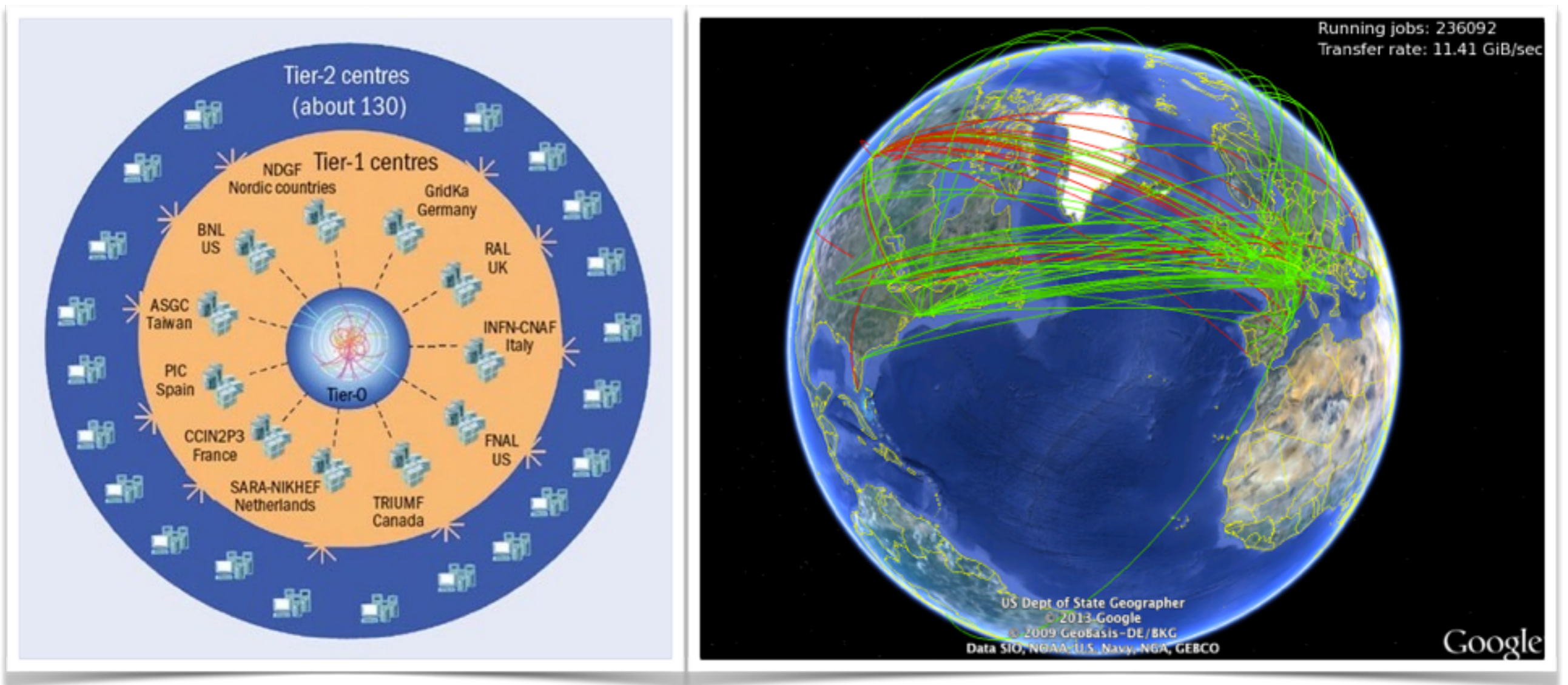
Worldwide LHC Computing Grid

A virtual super computing to store, distribute, reconstruct and analyse LHC data

Based on more than 170 computing centres in 42 countries

Distribute and analyse ~30PB of data annually generated by LHC

Experiments produce >15PB of new data annually



WLCG Software

No single job batch submission system, incl. **LSF**, **HTCondor**, Slurm, SGE, Torque/Pbs

No single storage solution, incl. NFS, GlusterFS, **Hadoop** (popular in US)

Contains **10-years** worth of CPUs (100+ SKUs)

Common operating system: **RHEL/CentOS/Scientific Linux (SL)**

Dominated by **SL 6** co-developed by CERN and Fermilab

CentOS 7 + Special Interest Group to follow **SL 6**

Software and essential precomputed data (e.g. LUT) distributed via **CernVM File System (CVMFS)**

HEP SPEC '06 benchmark is used for accounting in WLCG and by experiments

Designed to represent worker node activity under full load

CMS Software Bundle

CMSSW is **open-source** and available at GitHub

Mostly written in **C++14**, C, **Python** and Fortran

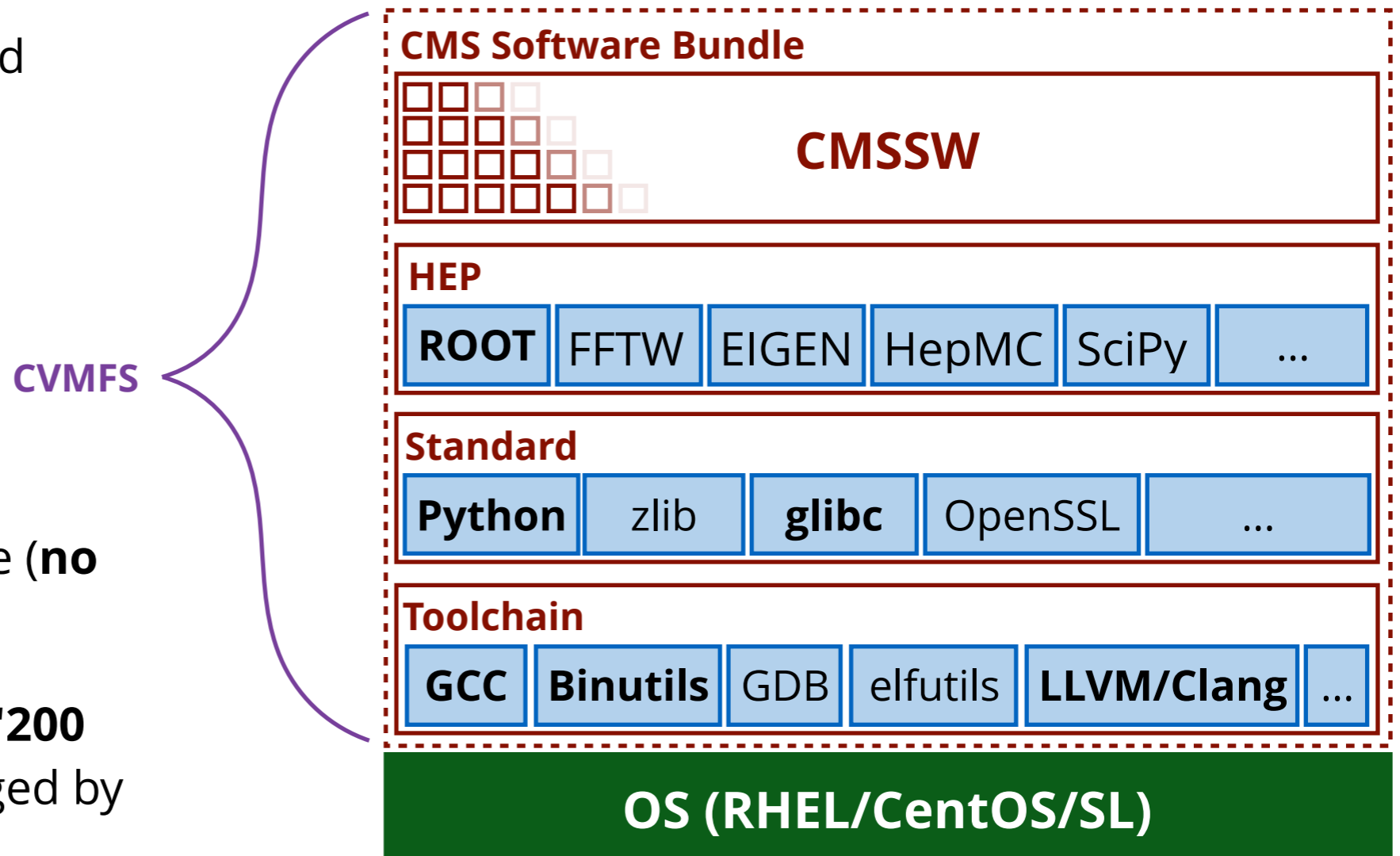
Packaged as **RPMs** and via custom build infrastructure (**no mock and Koji**)

CMSSW is composed of **~1'200** individual packages managed by **SCRAM**

CMSSW is like **Software Collection** package or **Linux Container** without actually being any of them

SCRAM is **S**ource **C**onfiguration, **R**elease, **A**nd **M**anagement tool

Think **scl** + **CMake** + **make**



Why new architectures?

Distributed computing in **HEP before ~2000** had multiple vendors involved, and incl. special workstations and heterogeneous computing

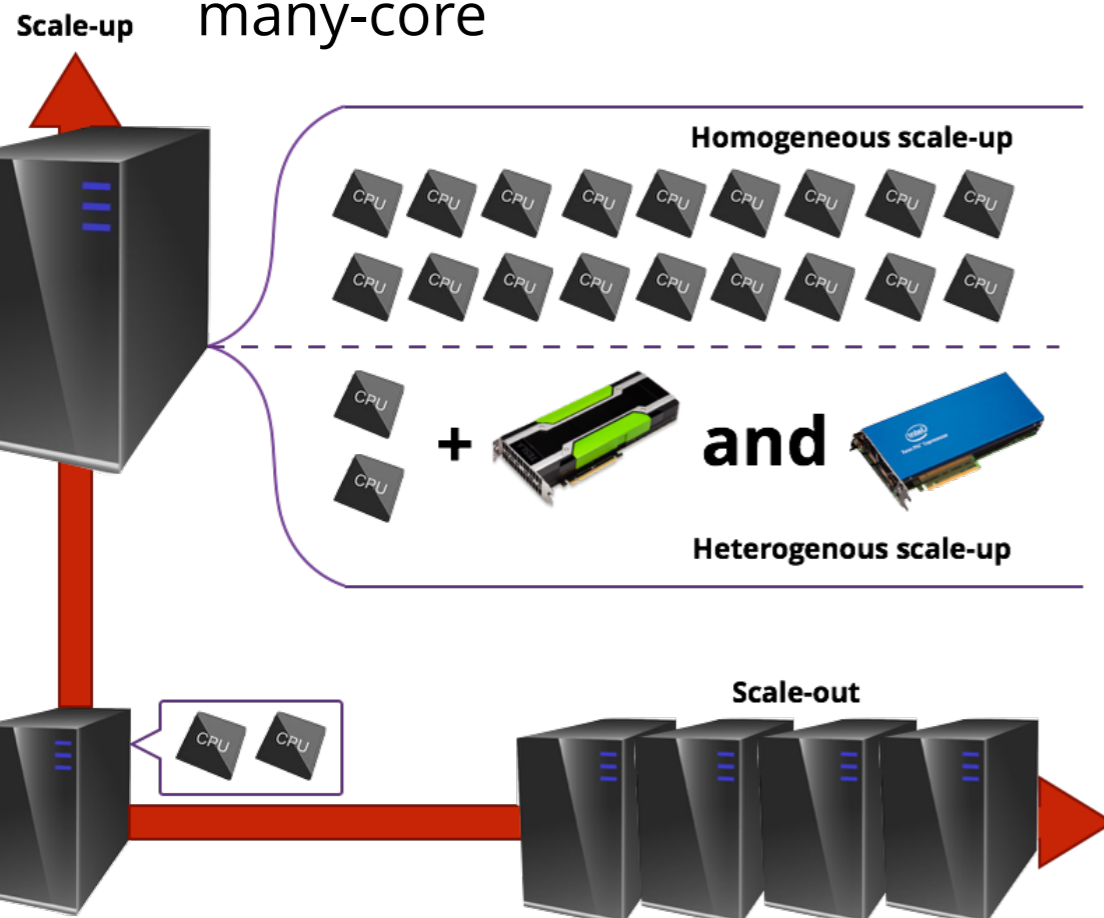
High Throughput Computing (HTC) converged on x86/Linux **at ~2000**

Commodity hardware enabled the current model of WLCG:

Build Once, Run Everywhere

Two vendors: Intel (dominating) and AMD

Commodity hardware itself is limited by power wall with stop-gap solution as many-core



Specialised processors and heterogeneous computing rise up

Lightweight general-purpose low-power high-density, vector units, GPUs, Xeon Phi (highly-parallel long-vector), etc

The focus is shifting to **performance/watt**, not just **performance/price**

Why ARM?

ARM dominates mobile and embedded market instead of Intel and AMD

The focus is on **low-power** and **high-efficiency** SoCs

ARMv8 provides **64-bit** ISA, **LE** and **LP64** data model

With the help of various partners (e.g., APM, AMD, Cavium) enters datacenter market

The business model of licensing Intellectual Property (IP) to partners enables market competition and heterogeneous specialisation described earlier

We see ARM playing a role in the strategies of other big players (e.g., AMD, Nvidia)

CUDA 6.5 added support ARMv8 64-bit (Aug 20, 2014)

Porting to ARMv8

CMSSW was originally ported to **ARMv7 (32-bit)** few years ago

High-end mobile SoC based development boards were used

ODROID-U2 (Exynos 4412 Prime), ODROID-XU2 (Exynos 541), Arndale Octa (Exynos 5420), Jetson TK1 (Tegra K1)

Resolved majority of porting issues and found numerous issues in CMSSW (even affecting x86_64)

CMSSW for **ARMv8 (64-bit)** port was started early

Step1: ARM Foundation Model

Step2: QEMU + binfmt_misc + user mode emulation

Step3: APM Mustang

Step4: HP Moonshot + m400

For ARMv8 we wanted full stack, CMSSW itself was not enough

We needed the port of **Open Science Grid (OSG)** repositories

CVMFS was required for software distribution

T3_US_Princeton_ARM

Building a Tier-3 on ARMv8 Server-on-Chip for WLCG

Hardware: x86_64 master node, APM Mustang and HP Moonshot + 6 x m400 worker nodes

No local storage element, only remote data access and stage out

APM Mustang

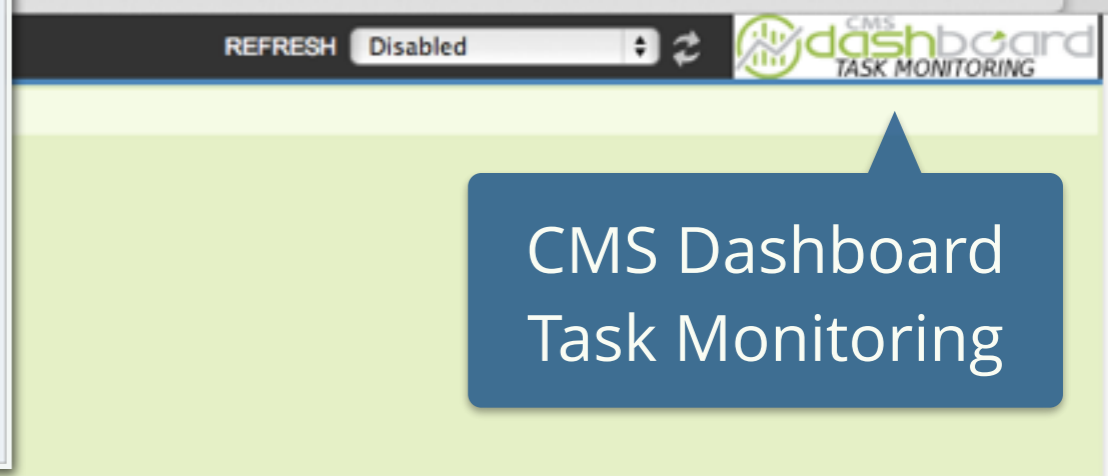


HP Moonshot + m400





On **June 26, 2015** CMS **successfully** executed CMSSW based job on **AArch64 worker node** via standard job injection pipeline and received output files



Start » [Justas Balcas] » Tasks » Jobs

Data Charts Show 25 entries Task: 150608_200051:jbalkas_crab_ARM_TEST_2-output2 NJobTotal: 1000 Pending: 822 Running: 0 Unknown: 0 Cancelled: 0 Success: 168 Failed: 2 WNPostProc: 8 ToRetry: 0

Id	Status	AppExitCode	Site	Retries	Submitted	Started	Finished	Wall Time	Job Log	File Access	FTS File Status
1	finished	0	T3_US_Princeton_ARM	1	2015-06-08T20:01:22	2015-06-08T20:05:35	2015-06-08T20:15:16	00:09:41	Job Log, Job Log JSON, Post Job Log	File Info	N/A
2	finished	0	T3_US_Princeton_ARM	1	2015-06-08T20:01:22	2015-06-08T20:05:35	2015-06-08T20:15:15	00:09:38	Job Log, Job Log JSON, Post Job Log	File Info	N/A
3	finished	0	T3_US_Princeton_ARM	1	2015-06-08T20:01:22	2015-06-08T20:05:37	2015-06-08T20:15:25	00:09:48	Job Log, Job Log JSON, Post Job Log	File Info	N/A
4	finished	0	T3_US_Princeton_ARM	1	2015-06-08T20:01:22	2015-06-08T20:05:37	2015-06-08T20:15:32	00:09:55	Job Log, Job Log JSON, Post Job Log	File Info	N/A
5	finished	0	T3_US_Princeton_ARM	1	2015-06-08T20:01:22	2015-06-08T20:05:37	2015-06-08T20:15:34	00:09:57	Job Log, Job Log JSON, Post Job Log	File Info	N/A
6	finished	0	T3_US_Princeton_ARM	1	2015-06-08T20:01:22	2015-06-08T20:09:29	2015-06-08T20:16:00	00:06:31	Job Log, Job Log JSON, Post Job Log	File Info	N/A
7	finished	0	T3_US_Princeton_ARM	1	2015-06-08T20:01:22	2015-06-08T20:24:29	2015-06-08T20:28:52	00:04:23	Job Log, Job Log JSON, Post Job Log	File Info	N/A
8	finished	0	T3_US_Princeton_ARM	1	2015-06-08T20:01:22	2015-06-08T20:24:29	2015-06-08T20:29:03	00:04:34	Job Log, Job Log JSON, Post Job Log	File Info	N/A
9	finished	0	T3_US_Princeton_ARM	1	2015-06-08T20:01:22	2015-06-08T20:24:30	2015-06-08T20:29:32	00:05:02	Job Log, Job Log JSON, Post Job Log	File Info	N/A
10	finished	0	T3_US_Princeton_ARM	1	2015-06-08T20:01:22	2015-06-08T20:24:31	2015-06-08T20:28:10	00:03:39	Job Log, Job Log JSON, Post Job Log	File Info	N/A

The first AArch64 based WLCG site (demonstrator)

```

1 [|||||] 9 ]
2 [|||||] 98.0% ]
3 [|||||] 100.0% ]
4 [|||||] 100.0% ]
Mem[|||||]
Swp[|||||]
]
Tasks: 187, 19 thr; 11 running
Load average: 5.18 2.48
Uptime: 33 days, 02:07:36

PID USER PRI NI VIRT RES SHR S CPU% MEM% TIME+ Command
6128 20 0 2042M 41224 1948 S 0.0 0.3 0:00.16 /usr/bin/cvmfs2 -o rw,fsname=cvmfs2,allow_other,grab_mountpoint,uid=997,gid=995 cms.cern.ch /cvmfs/cms.cern.ch
6127 20 0 2042M 41224 1948 S 0.0 0.3 0:00.17 /usr/bin/cvmfs2 -o rw,fsname=cvmfs2,allow_other,grab_mountpoint,uid=997,gid=995 cms.cern.ch /cvmfs/cms.cern.ch
6120 20 0 2042M 41224 1948 S 0.0 0.3 0:00.18 /usr/bin/cvmfs2 -o rw,fsname=cvmfs2,allow_other,grab_mountpoint,uid=997,gid=995 cms.cern.ch /cvmfs/cms.cern.ch
23248 20 0 16236 6420 5252 S 0.0 0.0 0:00.52 /usr/sbin/condor_master -f
23256 20 0 17728 8000 5688 S 2.0 0.0 5:04.76 condor_startd -f
30301 20 0 16688 6736 5492 S 0.0 0.0 0:00.09 condor_starter -f -a slot4 byggvir.Princeton.EDU
30305 30 10 3744 1848 1208 S 0.0 0.0 0:00.62 /bin/bash /var/lib/condor/execute/dir_30301/condor_exec.exe -v std -name gfactory_instance -entry CMS_T3_U
2478 30 10 3468 1548 1208 S 0.0 0.0 0:00.12 /bin/bash /var/lib/condor/execute/dir_30301/glide_NRPbun/main/condor_startup.sh glidein_config
3191 30 10 17884 8272 6320 S 0.0 0.1 0:00.16 /var/lib/condor/execute/dir_30301/glide_NRPbun/main/condor/sbin/condor_master -f -pidfile /var/lib/c
3194 30 10 18928 9140 6748 S 0.0 0.1 0:00.87 condor_startd -f
2898 30 10 17012 8324 6552 S 0.0 0.1 0:00.16 condor_starter -f vocns058.cern.ch
4428 30 10 3352 1456 1196 S 0.0 0.0 0:00.10 /bin/bash /var/lib/condor/execute/dir_30301/glide_NRPbun/execute/dir_2898/condor_exec.exe -
4585 30 10 3520 1520 1224 S 0.0 0.0 0:00.02 sh ./CMSRunAnalysis.sh -a sandbox.tar.gz --sourceURL=https://cmsweb.cern.ch/crabcache --
4631 30 10 23508 13492 1572 S 0.7 0.1 0:00.70 python CMSRunAnalysis.py -r /var/lib/condor/execute/dir_30301/glide_NRPbun/execute/di
5236 30 10 3624 1648 1160 S 0.0 0.0 0:00.01 /bin/bash /var/lib/condor/execute/dir_30301/glide_NRPbun/execute/dir_2898/cmsRun-m
5281 uscms01 30 10 921M 588M 115M R 93.7 3.7 4:07.20 cmsRun -j FrameworkJobReport.xml PSet.py
3193 30 10 7024 4072 1100 S 0.0 0.0 0:00.71 condor_procd -A /var/lib/condor/execute/dir_30301/glide_NRPbun/log/procd_address -L /var/lib/cond
30119 20 0 16688 6724 5492 S 0.0 0.0 0:00.08 condor_starter -f -a slot1 byggvir.Princeton.EDU
30123 30 10 3744 1848 1208 S 0.0 0.0 0:00.62 /bin/bash /var/lib/condor/execute/dir_30119/condor_exec.exe -v std -name gfactory_instance -entry CMS_T3_U
2156 30 10 3472 1548 1208 S 0.0 0.0 0:00.12 /bin/bash /var/lib/condor/execute/dir_30119/glide_LreWcj/main/condor_startup.sh glidein_config
2871 30 10 17884 8272 6320 S 0.0 0.1 0:00.16 /var/lib/condor/execute/dir_30119/glide_LreWcj/main/condor/sbin/condor_master -f -pidfile /var/lib/c
2874 30 10 18952 9168 6748 S 0.0 0.1 0:00.87 condor_startd -f
2892 30 10 17416 8676 6568 S 0.0 0.1 0:00.16 condor_starter -f vocns058.cern.ch
3431 30 10 3352 1456 1196 S 0.0 0.0 0:00.10 /bin/bash /var/lib/condor/execute/dir_30119/glide_LreWcj/execute/dir_2892/condor_exec.exe -
3638 30 10 3520 1516 1224 S 0.0 0.0 0:00.02 sh ./CMSRunAnalysis.sh -a sandbox.tar.gz --sourceURL=https://cmsweb.cern.ch/crabcache --
3692 30 10 23508 13256 1340 S 0.0 0.1 0:00.70 python CMSRunAnalysis.py -r /var/lib/condor/execute/dir_30119/glide_LreWcj/execute/di
4965 30 10 3624 1648 1160 S 0.0 0.0 0:00.01 /bin/bash /var/lib/condor/execute/dir_30119/glide_LreWcj/execute/dir_2892/cmsRun-m
5104 30 10 917M 566M 98616 R 97.6 3.5 4:07.37 cmsRun -j FrameworkJobReport.xml PSet.py
2873 30 10 6924 3412 1100 S 0.0 0.0 0:00.63 condor_procd -A /var/lib/condor/execute/dir_30119/glide_LreWcj/log/procd_address -L /var/lib/cond
24914 20 0 16688 6740 5492 S 1.3 0.0 0:00.09 condor_starter -f -a slot7 byggvir.Princeton.EDU
24918 30 10 3744 1848 1208 S 0.0 0.0 0:00.61 /bin/bash /var/lib/condor/execute/dir_24914/condor_exec.exe -v std -name gfactory_instance -entry CMS_T3_U
29404 30 10 3472 1548 1208 S 0.0 0.0 0:00.12 /bin/bash /var/lib/condor/execute/dir_24914/glide_iEheSD/main/condor_startup.sh glidein_config
30115 30 10 17884 8272 6320 S 0.0 0.1 0:00.16 /var/lib/condor/execute/dir_24914/glide_iEheSD/main/condor/sbin/condor_master -f -pidfile /var/lib/c
30118 30 10 18928 9140 6748 S 0.0 0.1 0:00.88 condor_startd -f
2894 30 10 17012 8336 6568 S 0.0 0.1 0:00.16 condor_starter -f vocns058.cern.ch
3697 30 10 3352 1456 1196 S 0.0 0.0 0:00.10 /bin/bash /var/lib/condor/execute/dir_24914/glide_iEheSD/execute/dir_2894/condor_exec.exe -
3823 30 10 3520 1520 1224 S 0.0 0.0 0:00.02 sh ./CMSRunAnalysis.sh -a sandbox.tar.gz --sourceURL=https://cmsweb.cern.ch/crabcache --
3852 30 10 23508 13228 1312 R 0.0 0.1 0:00.71 python CMSRunAnalysis.py -r /var/lib/condor/execute/dir_24914/glide_iEheSD/execute/di
5049 30 10 3624 1648 1160 S 0.0 0.0 0:00.01 /bin/bash /var/lib/condor/execute/dir_24914/glide_iEheSD/execute/dir_2894/cmsRun-m
5152 30 10 919M 567M 98404 R 98.9 3.5 4:07.56 cmsRun -j FrameworkJobReport.xml PSet.py
30117 30 10 7048 4000 1100 S 0.0 0.0 0:00.73 condor_procd -A /var/lib/condor/execute/dir_24914/glide_iEheSD/log/procd_address -L /var/lib/cond

F1 Help F2 Setup F3 Search F4 Filter F5 Tree F6 SortBy F7 Nice F8 Nice + F9 Kill F10 Quit

```

Heterogenous computing

Batch job submitted from **x86_64** machine at CERN to **aarch64** worker node at Princeton University (**T3_US_Princeton_ARM**)

Showcased on **Fedora 19** on **APM Mustang** development board

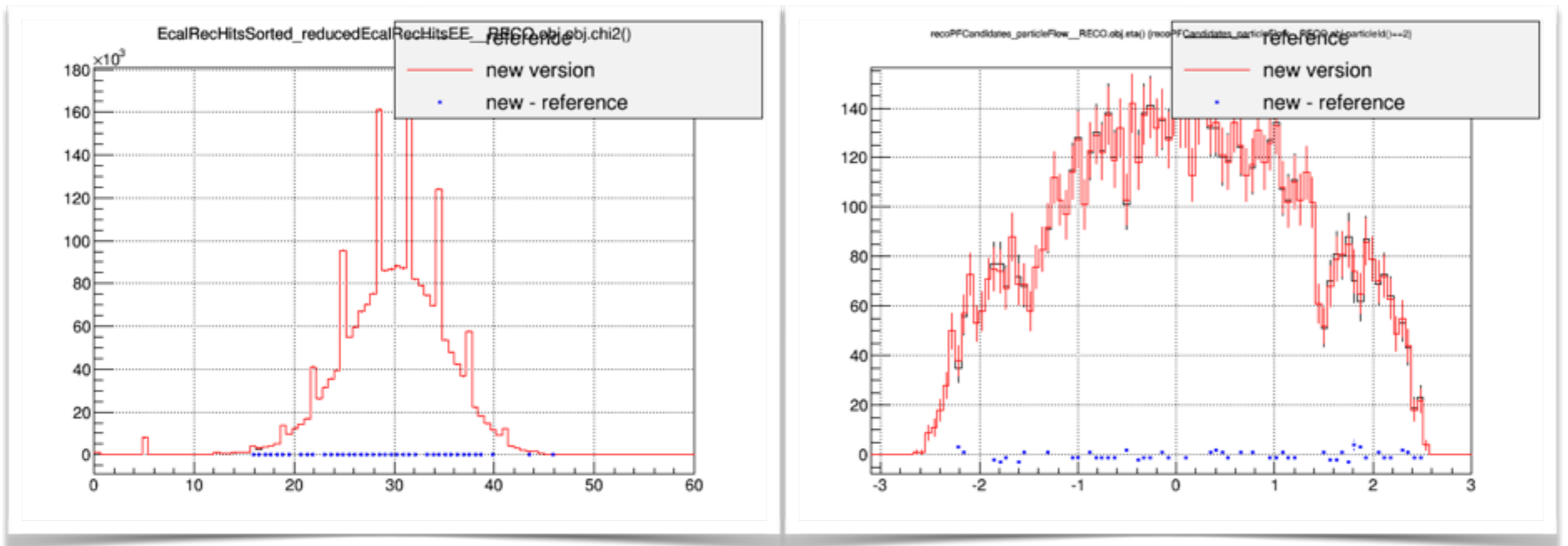
Moving to **RHELSA 7** on **HP Moonshot** + 6 x **m400** (production-level system)

Initial Physics Validation

For CMSSW on AArch64 to become a production-ready architecture a number of physics validation steps must be passed

We performed high-statistics (9000 events) reconstruction comparison between x86_64 and aarch64

~950 differences detected, but majority minimal, i.e. non-significant



Demo

CMSSW (**cmsRun**) processing **LHC** data on **AArch64**

If you are **silicon/hardware** vendor and have amazing product for **HEP + HTC**, do not hesitate to **contact!**

We are listening

Summary

We successfully ported CMSSW, essential parts of OSG, CVMFS to ARMv8 64-bit/AArch64

We demonstrated heterogeneous successful job submission and execution from x86_64 machine to aarch64 in a different continent using WLCG and CMS Computing infrastructure

Successful demonstration of remote read of input files and remote stage out of results

Initial look into high-statistics (9000 events reconstruction) comparison against x86_64 reference showed minimal differences

CMSSW for AArch64 is available on CVMFS now for any site

Contact

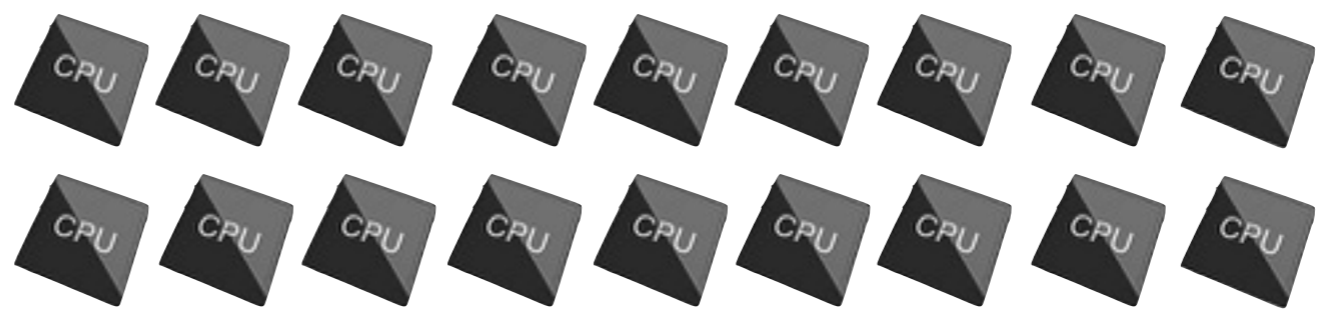


dauidt at cern dot ch

Scale-up



Homogeneous scale-up



Heterogenous scale-up

Scale-out

