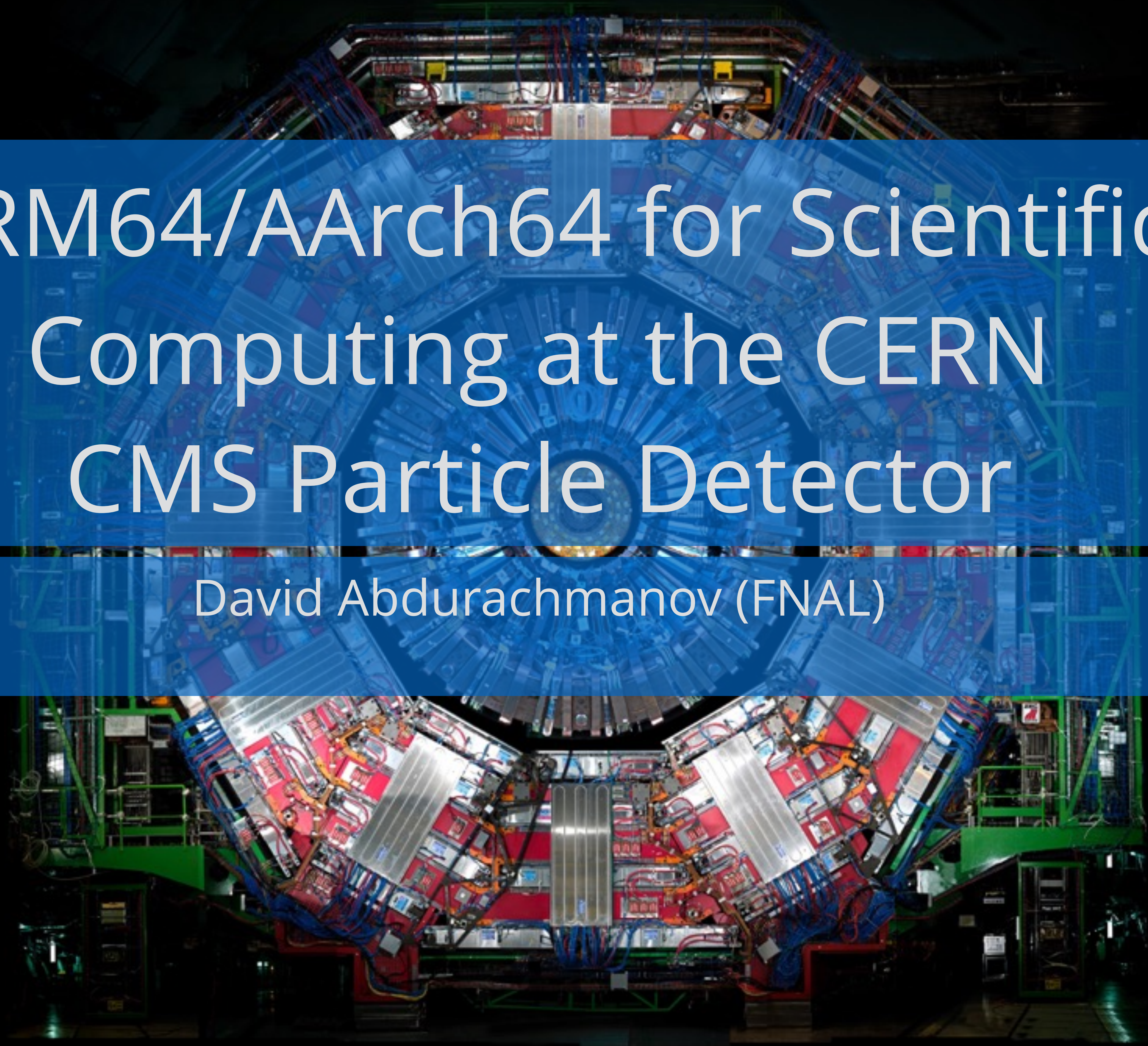
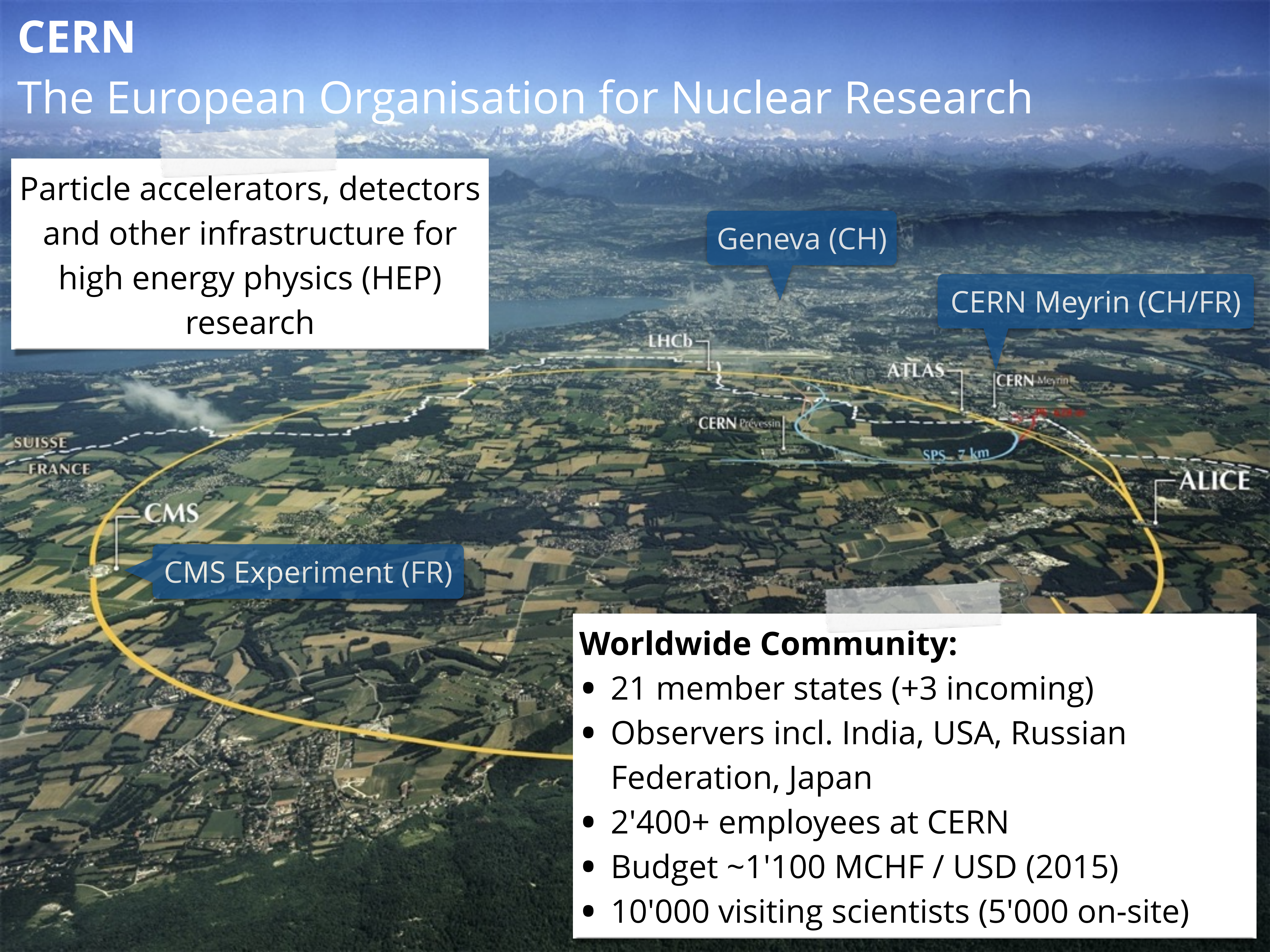


ARM64/AArch64 for Scientific Computing at the CERN CMS Particle Detector

David Abdurachmanov (FNAL)

CERN
The European Organisation for Nuclear Research

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

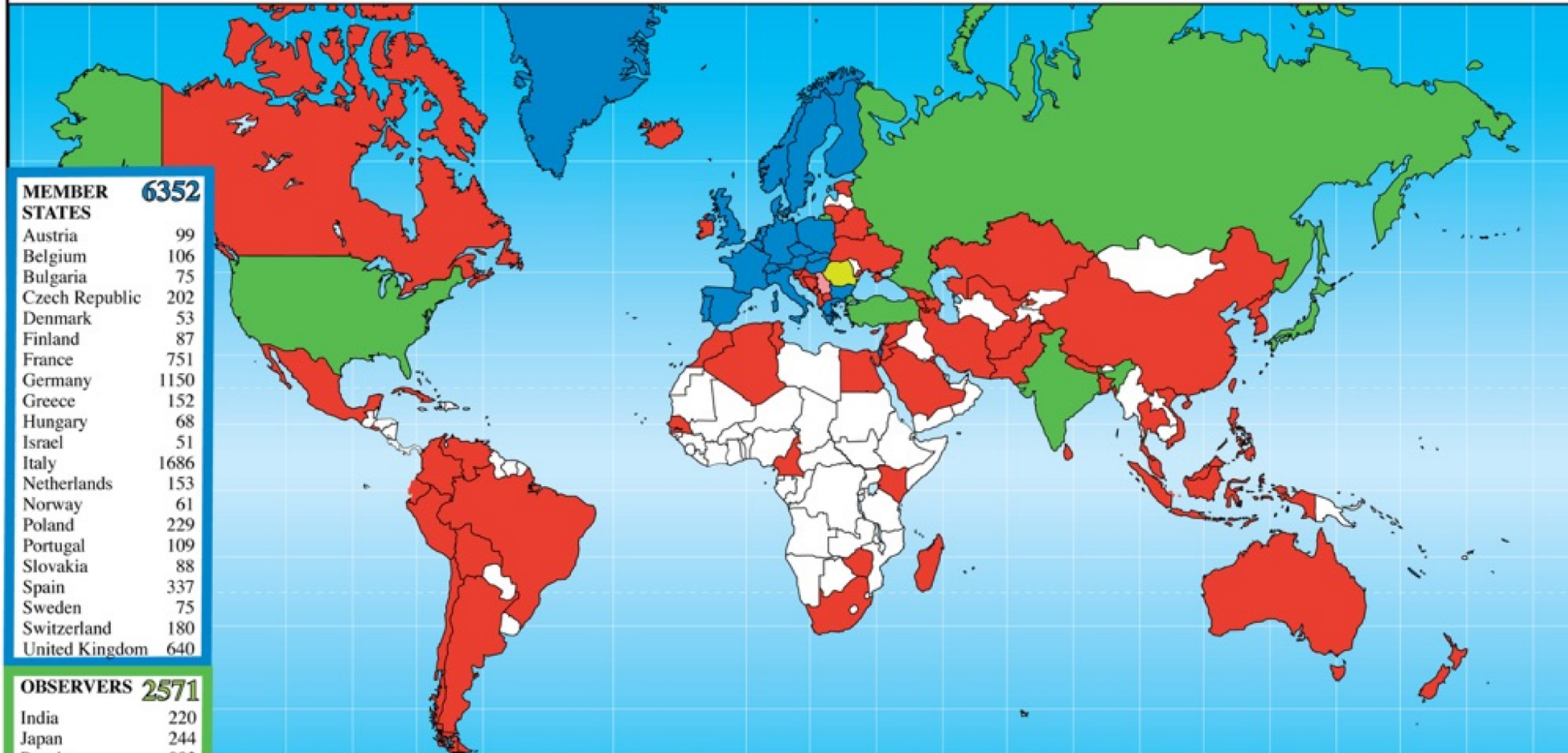


Worldwide Community:

- 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

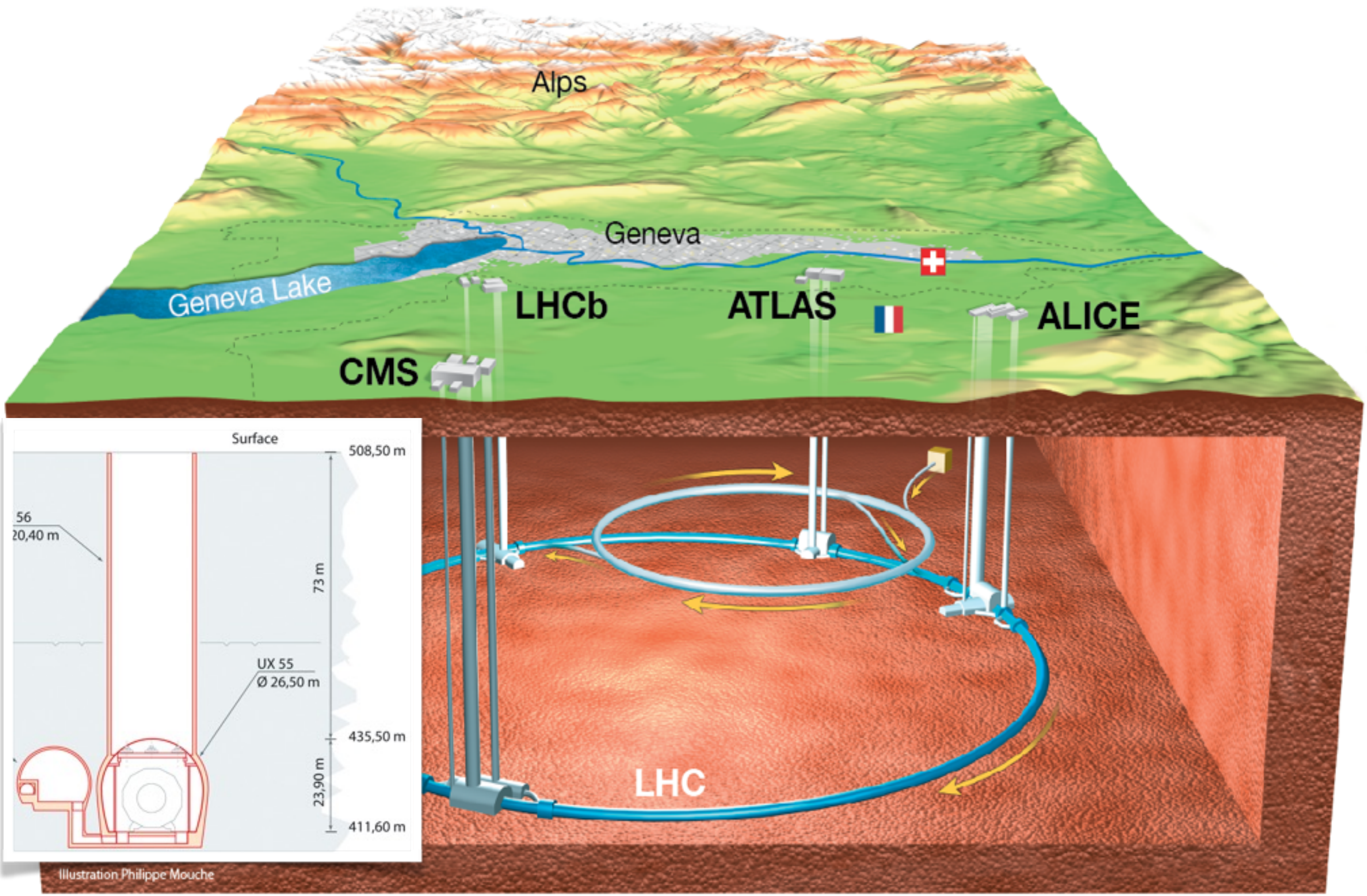
ASSOCIATE MEMBERS IN THE PRE-STAGE TO MEMBERSHIP	
Serbia	41

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



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

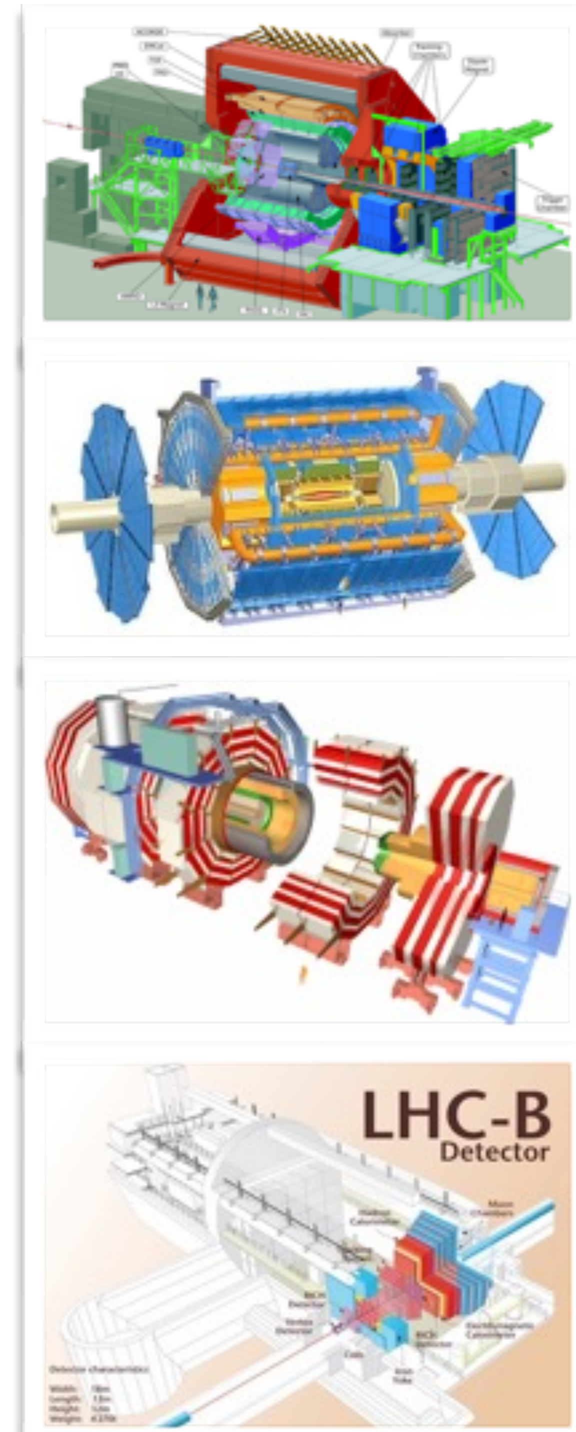
LHCb – "LHC beauty" (b-quark is called "beauty" quark)

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

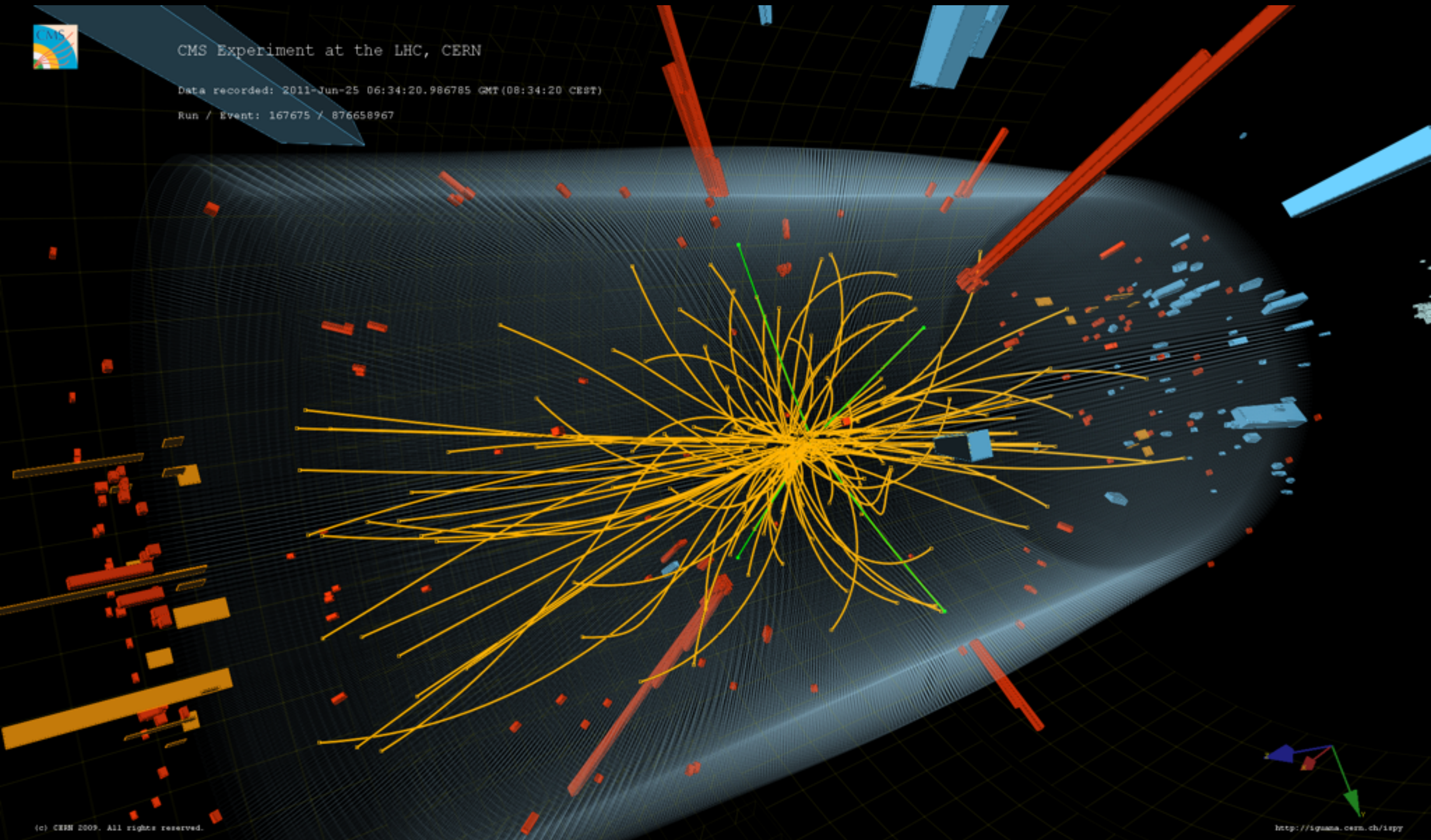
15 countries, 52 Institutes

Material costs: 75 MCHF / USD

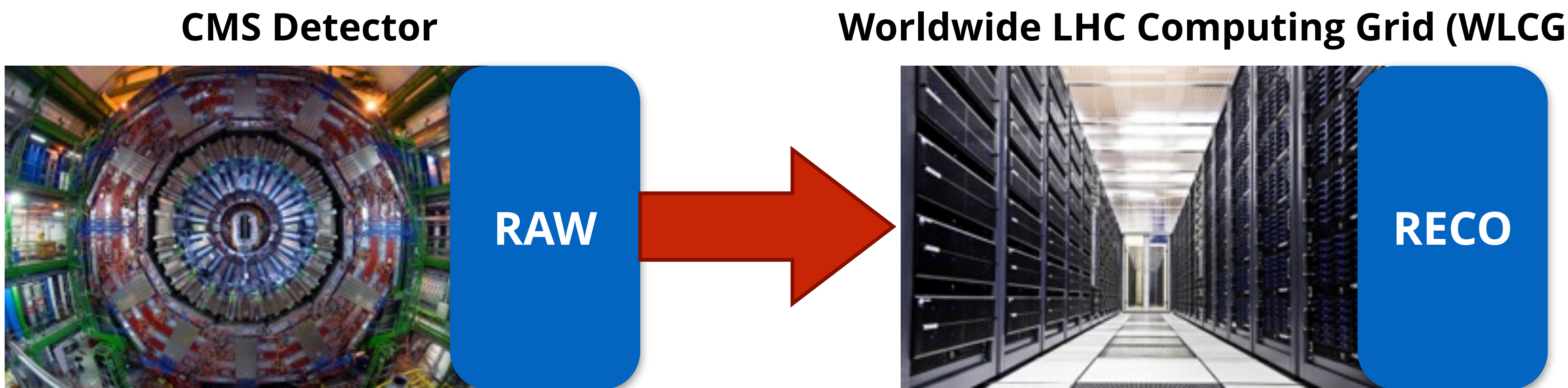
Regular upgrades, first was in 2013/14 (Long Shutdown 1)



Higgs-boson @ CMS



Data and Algorithms



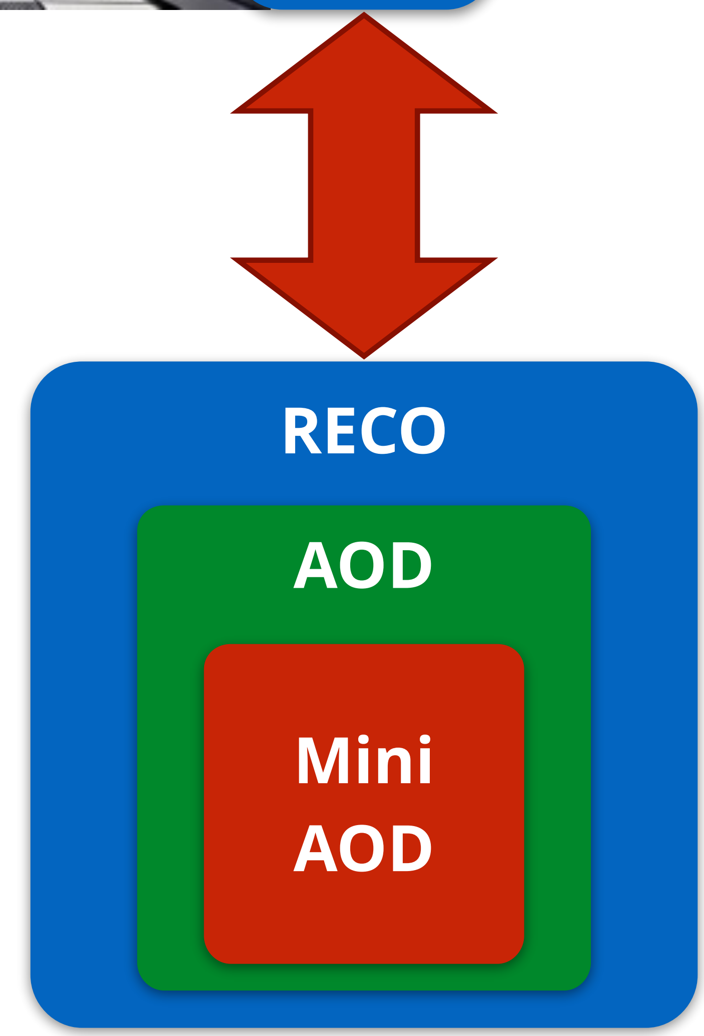
HEP data is organised as **Events** (particle collisions)

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

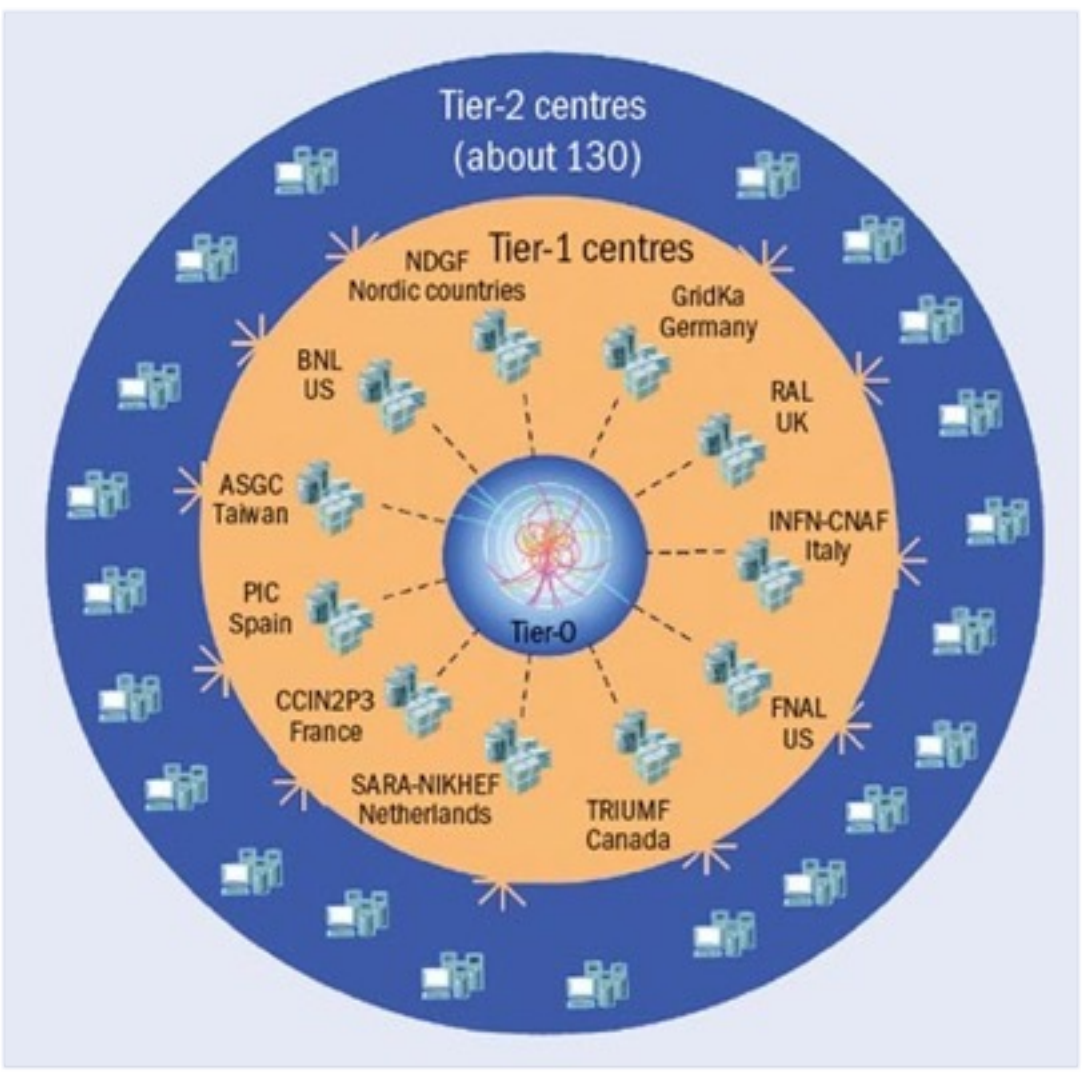
RAW - 300KB (ZLIB), **RECO** - 1MB/event (ZLIB), **AOD** - 300KB (LZMA), **Mini AOD** - 30KB/event (ZLIB)

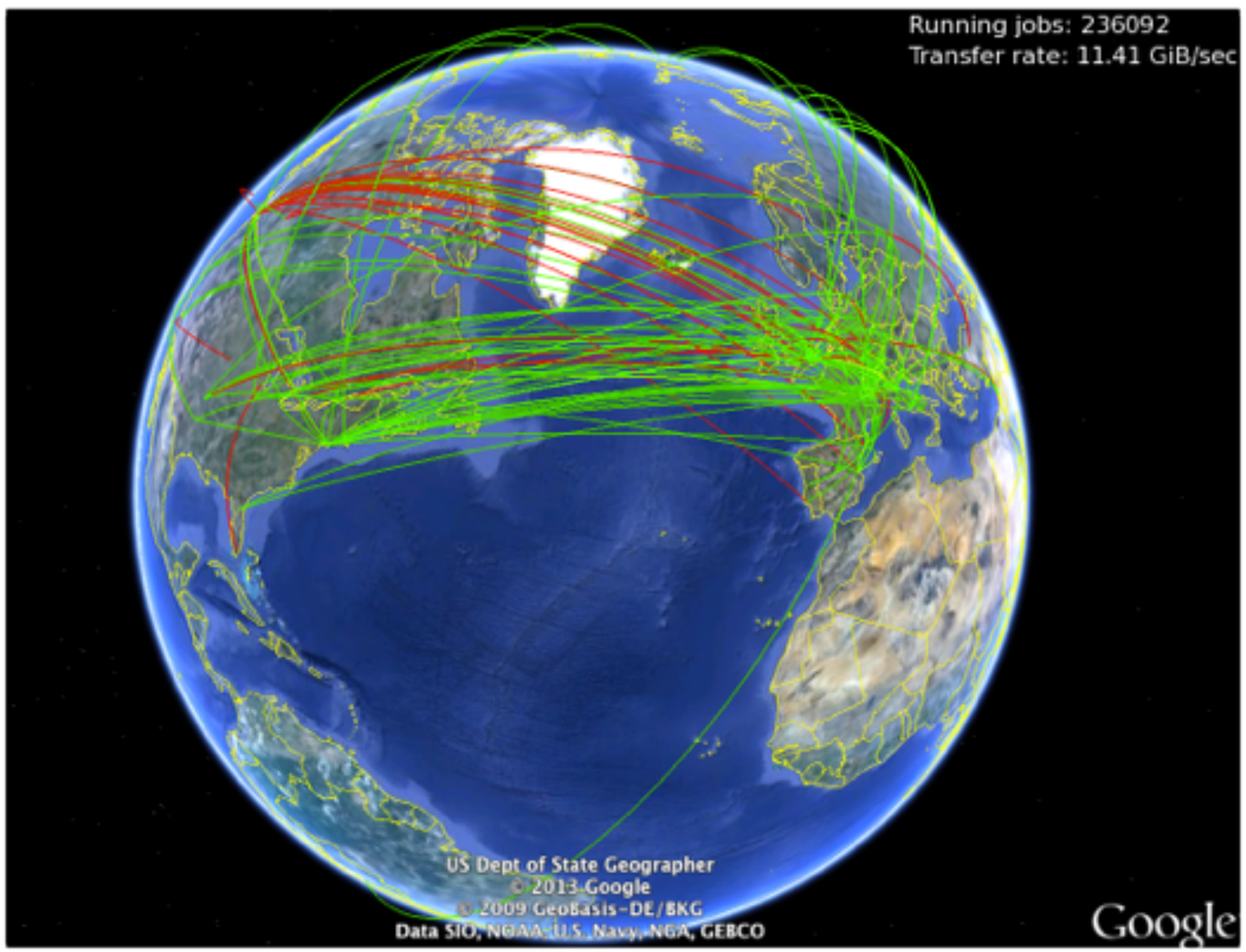


Worldwide LHC Computing Grid

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)**

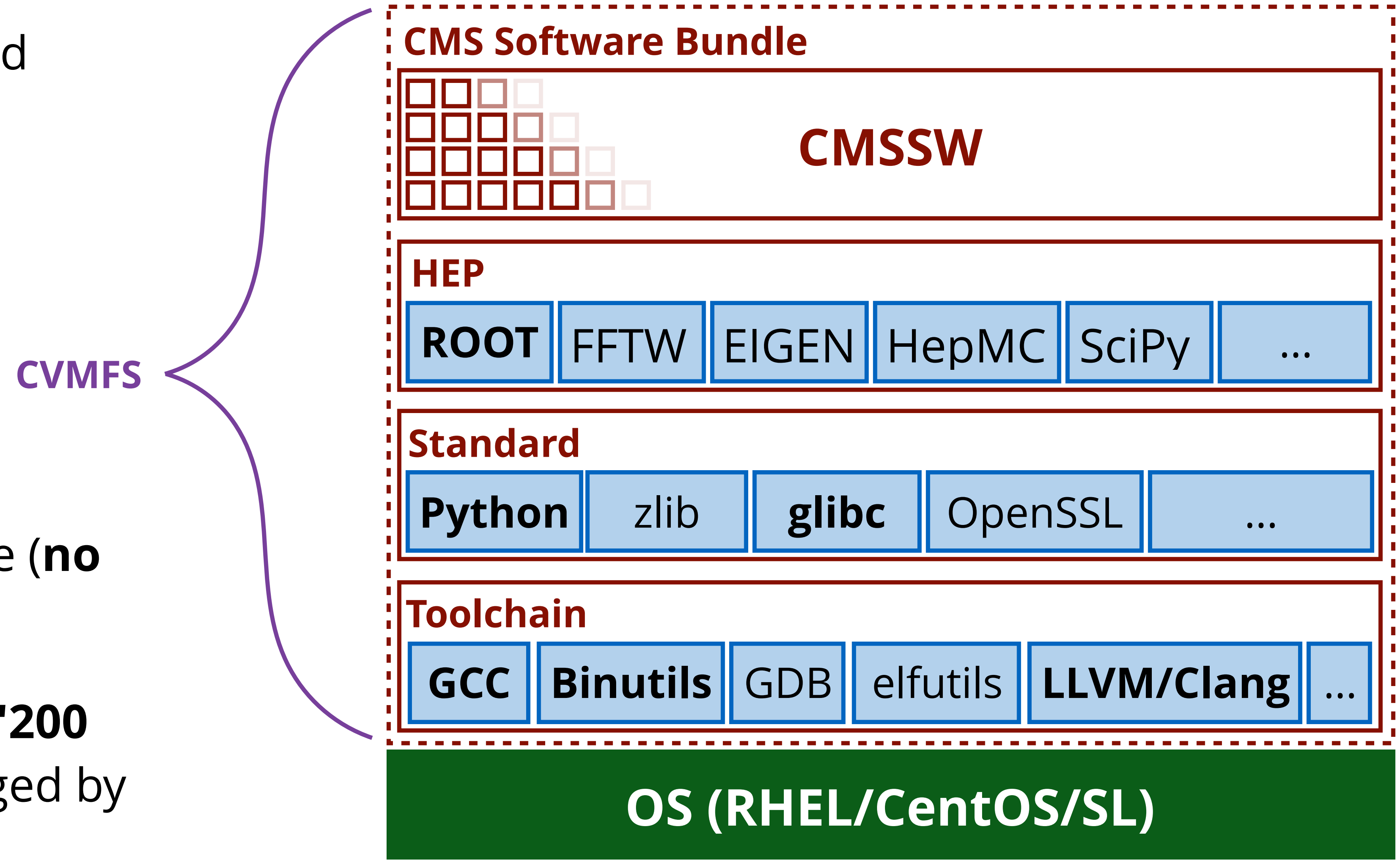
HEP SPEC '06 benchmark is used for accounting in WLCG and by experiments

- 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 **RPM**s 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

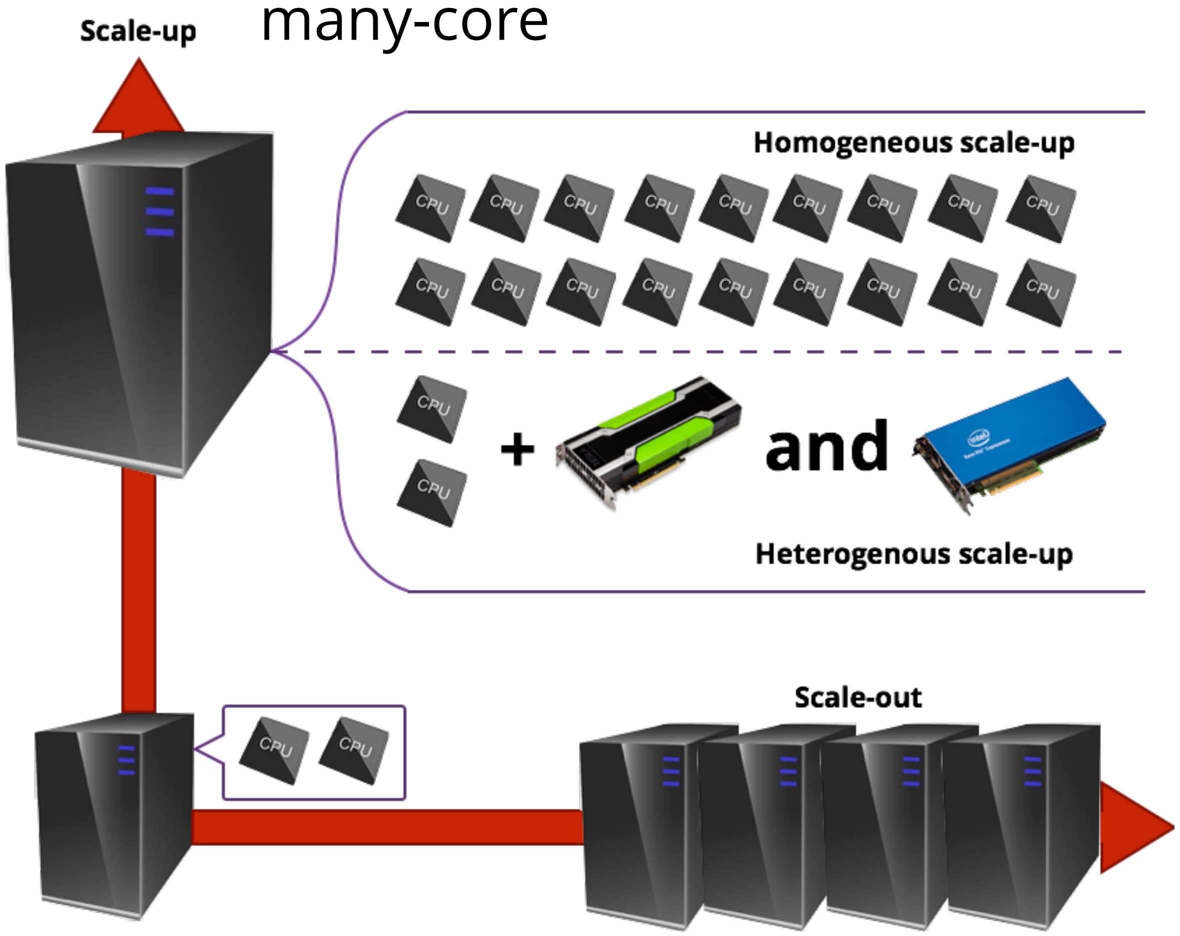
Hight Throughput Computing (HTC) converged on x86/Linux **at ~2000**

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 Phis (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

REFRESH Disabled

CMS dashboard TASK MONITORING

CMS Dashboard Task Monitoring

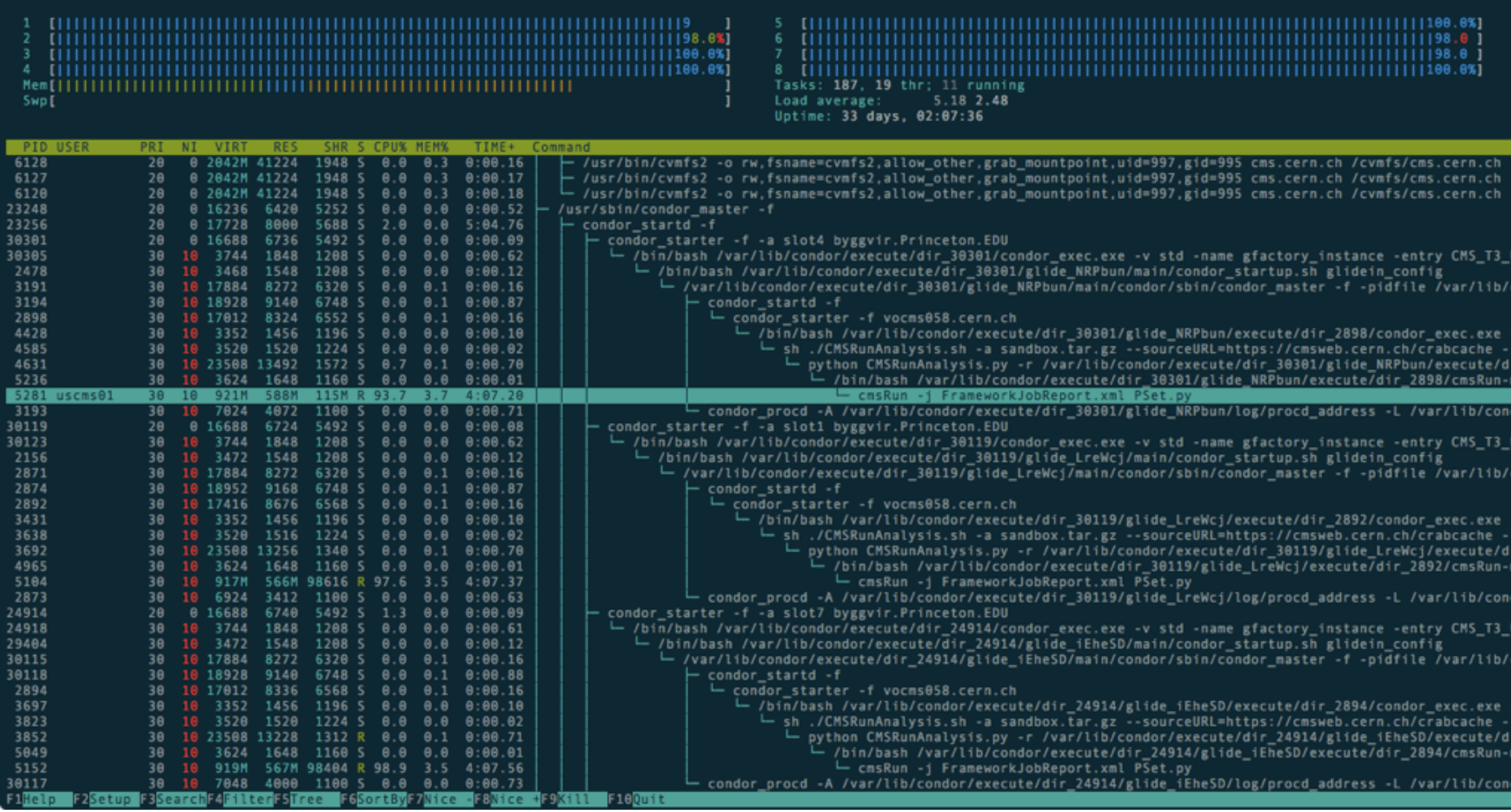
Start » [Justas Balcas] » Tasks » Jobs

Data | Charts | Show 25 entries

Task: 150608_200051:jbalcas_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	[illegible]	2015-06-08T20:01:22	2015-06-08T20:05:37	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)



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

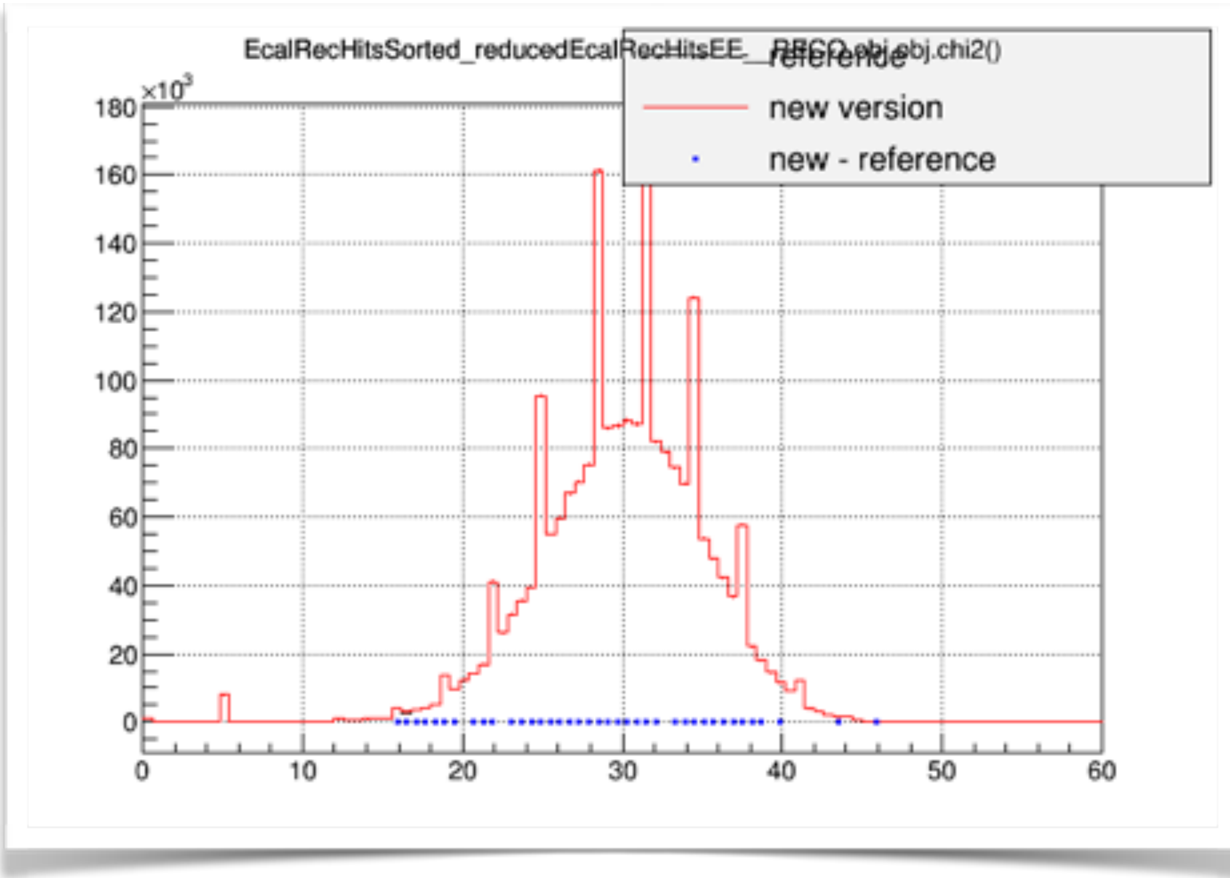
Moving to **RHELSA 7** on **HP Moonshot** + 6 x **m400** (production-level system)

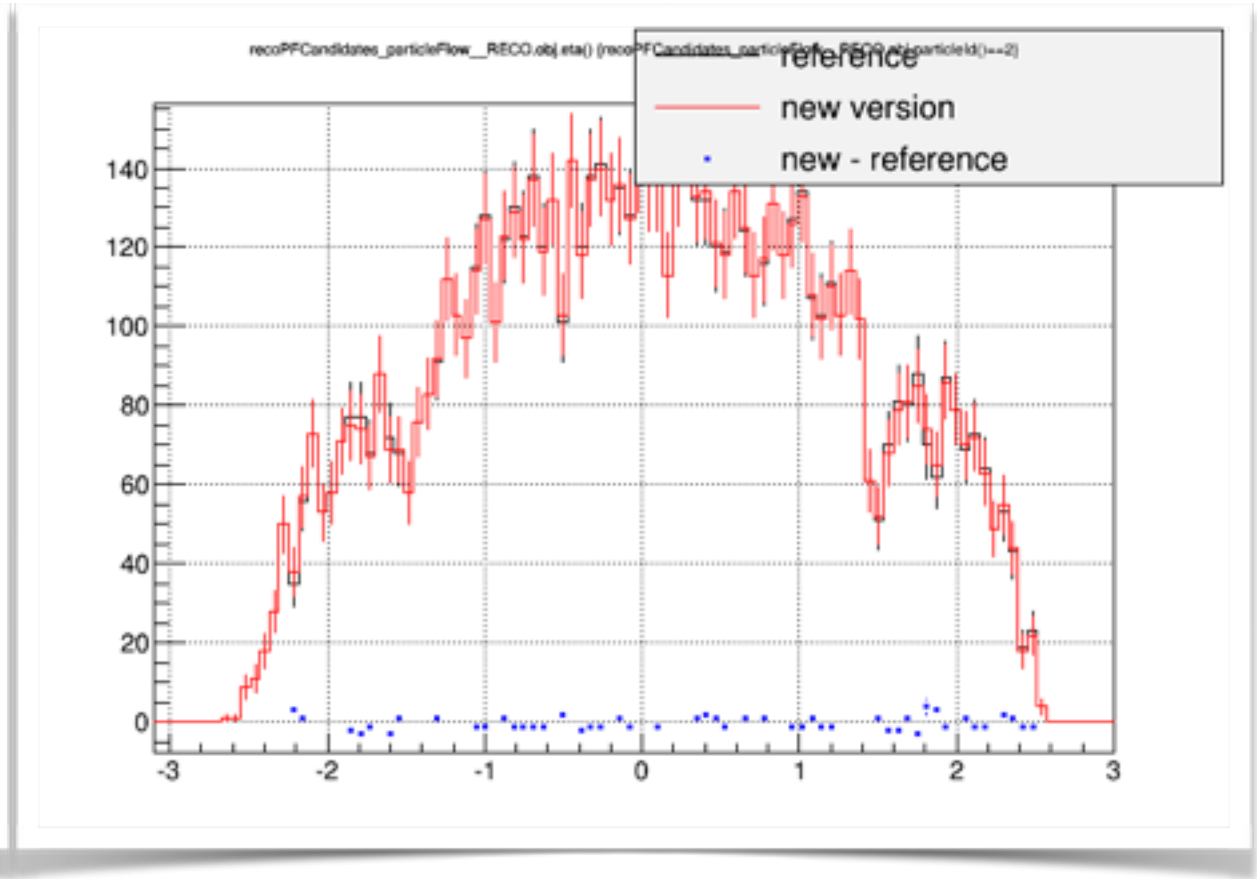
Initial Physics Validation

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

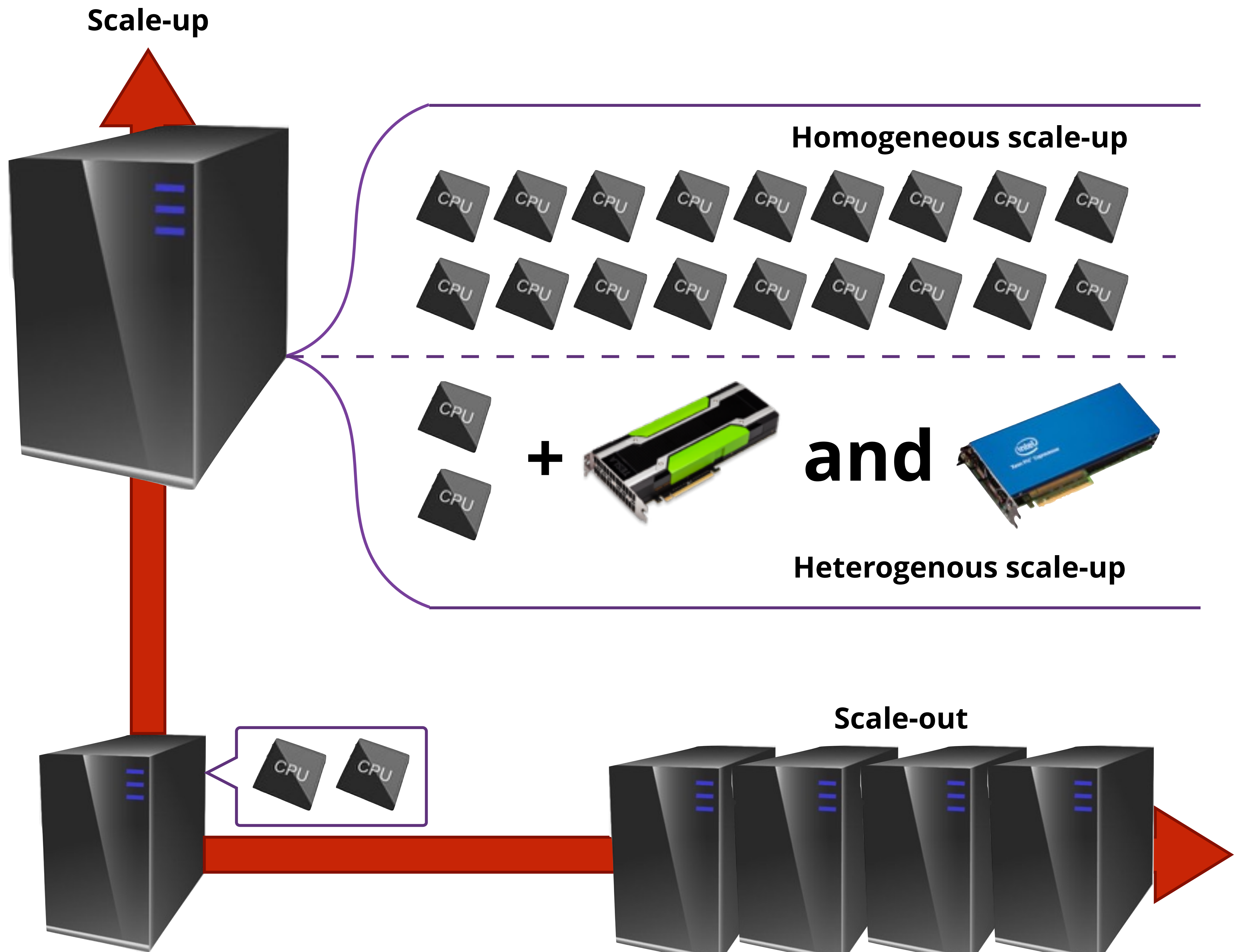
Successful demonstration of remote read of input files and remote stage out of results

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

davidlt at cern dot ch


Scale-up
Homogeneous scale-up
CPU
+
and
Heterogenous scale-up
Scale-out