Computing in CMS at DESY

S. Consuegra Rodríguez, L. Estévez Baños, D. Pérez Adán

Humboldt Highway II - computer cluster on renewable energies, 02-03.08.2023

Outline

> ## CMS global computing model

- Big picture
- Computing centers hierarchy (TIERs)
- Data storage and management
- Computing infrastructure at TIERs
- Event data model and software for data processing

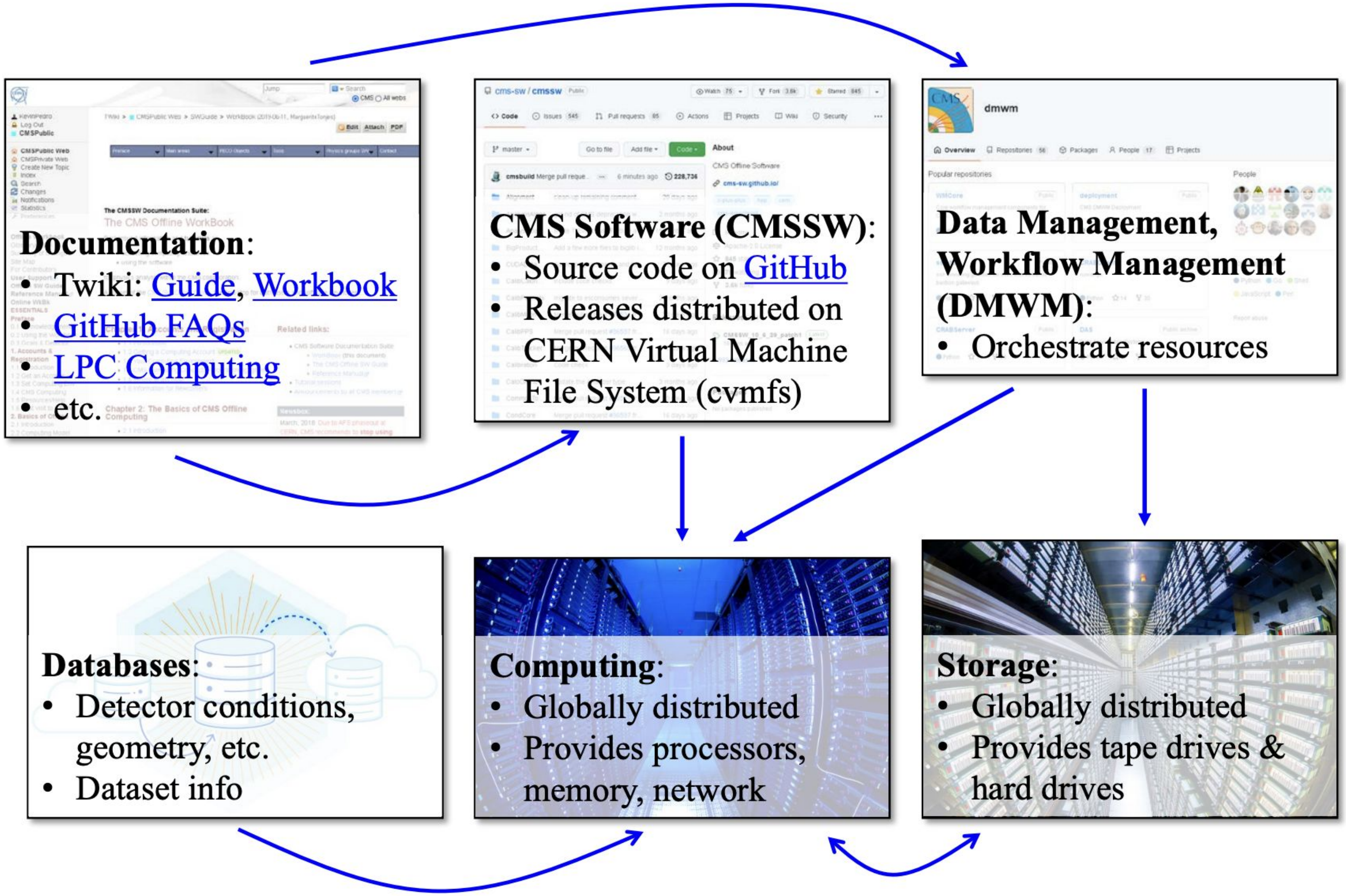
> ## CMS computing at DESY

- BIRD for batch processing
- HTCondor as resource manager
- Software access and needs

> ## Experience as user and typical needs

CMS global computing model

Big picture



Documentation:

- Twiki: Guide, Workbook
- GitHub FAQs
- LPC Computing
- etc.

CMS Software (CMSSW):

- Source code on GitHub
- Releases distributed on CERN Virtual Machine File System (cvmfs)

Data Management, Workflow Management (DMWM):

- Orchestrate resources

Databases:

- Detector conditions, geometry, etc.
- Dataset info

Computing:

- Globally distributed
- Provides processors, memory, network

Storage:

- Globally distributed
- Provides tape drives & hard drives

Source: CMS DAS 2022: Software, Computing, and Analysis Tools at CMS

CMS global computing model

Challenges

> not only in terms of

- physics reach and discovery potential of experiment
- detector to build and operate

but also in terms of data volume and necessary computing resources

> Computing and storage requirements difficult to fulfill at only one place, for both technical and funding reasons

> **Solution:** Computing environment structured as distributed system of computing services and resources that interact with each other as Grid services

> computational infrastructure intended to be **available to all CMS collaborators**, **independently of their physical locations**, and on **fair share basis**

CMS global computing model

Computing centers hierarchy (TIERs)

> Sites organized in “Tiers”, pledging CPU, disk storage, and tape resources, proportional to commitment of each funding agency within CMS Collaboration

Roles:

- **Tier-0 center**

close to experiment to execute a first calibration and reconstruction pass, the so-called “prompt reconstruction” as well as to maintain a custodial copy of all RAW data on tape; 24h/7 support guaranteed

combination of computing capacity, disk storage and tape archival space

- **Tier-1 regional centers**

(six in use by CMS), which maintain a second distributed custodial copy of the RAW data on tape, and provide CPU for re-reconstruction and MC simulation; 24h/7 support guaranteed

have adapted to support Tier-0 in the task of prompt data processing, whenever needed

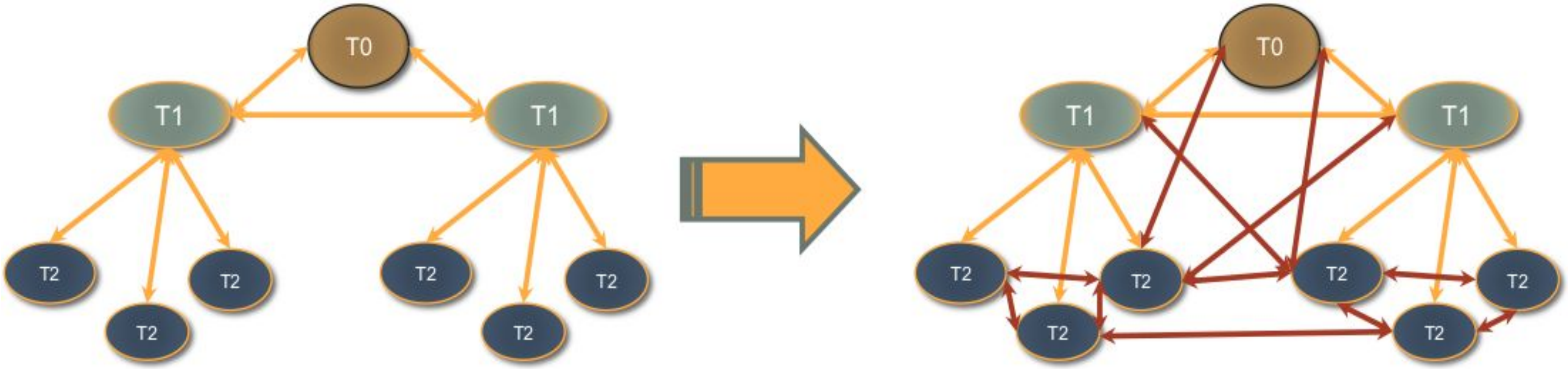
Countries: France, Germany, Italy, US, UK, and Spain

- **Tier-2 local centers**

(about 50 in use by CMS), providing support for analysis activity and MC simulation; guaranteed support only during working hours

CMS global computing model

Computing centers hierarchy (TIERs)



evolution of CMS computing model from hierarchical (left) to fully connected structure (right)

> Tier-1/2s have become more similar in order to optimize usage of resources with a high efficiency

Current computing model -> a job can now run, in principle, wherever free CPU is available while **accessing input data through high-speed wide-area network (WAN) connections**

Tier-3 sites

do not pledge resources, but nevertheless provide CPU and storage in varying amounts

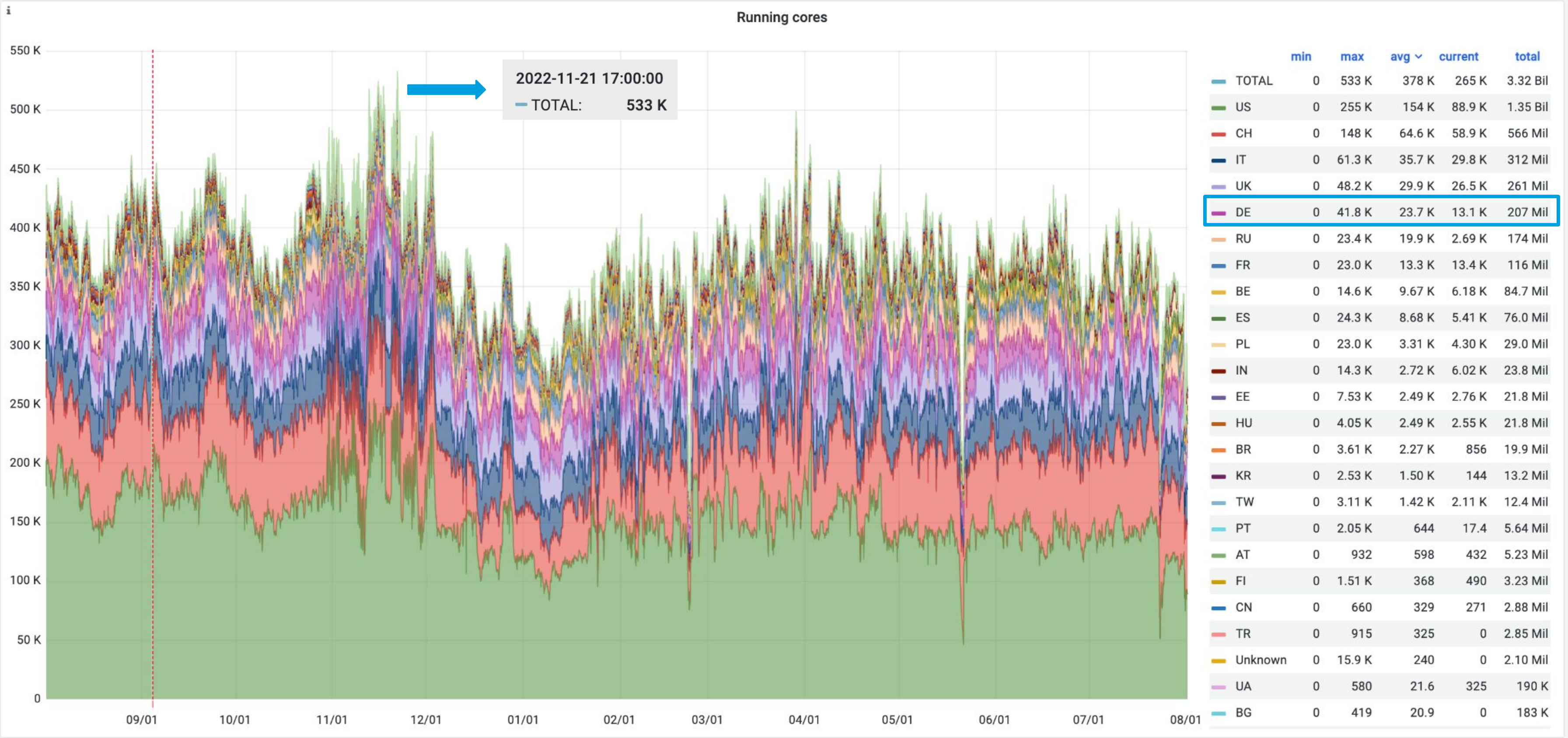
- batch farms for mainly local use
- those that predominantly support another experiment, but allow CMS to take CPU slots opportunistically when available

CMS global computing model

Resource Utilization

1 year timespan

CMS Job Monitoring



2022-11-21 17:00:00
TOTAL: 533 K

	min	max	avg	current	total
TOTAL	0	533 K	378 K	265 K	3.32 Bil
US	0	255 K	154 K	88.9 K	1.35 Bil
CH	0	148 K	64.6 K	58.9 K	566 Mil
IT	0	61.3 K	35.7 K	29.8 K	312 Mil
UK	0	48.2 K	29.9 K	26.5 K	261 Mil
DE	0	41.8 K	23.7 K	13.1 K	207 Mil
RU	0	23.4 K	19.9 K	2.69 K	174 Mil
FR	0	23.0 K	13.3 K	13.4 K	116 Mil
BE	0	14.6 K	9.67 K	6.18 K	84.7 Mil
ES	0	24.3 K	8.68 K	5.41 K	76.0 Mil
PL	0	23.0 K	3.31 K	4.30 K	29.0 Mil
IN	0	14.3 K	2.72 K	6.02 K	23.8 Mil
EE	0	7.53 K	2.49 K	2.76 K	21.8 Mil
HU	0	4.05 K	2.49 K	2.55 K	21.8 Mil
BR	0	3.61 K	2.27 K	856	19.9 Mil
KR	0	2.53 K	1.50 K	144	13.2 Mil
TW	0	3.11 K	1.42 K	2.11 K	12.4 Mil
PT	0	2.05 K	644	17.4	5.64 Mil
AT	0	932	598	432	5.23 Mil
FI	0	1.51 K	368	490	3.23 Mil
CN	0	660	329	271	2.88 Mil
TR	0	915	325	0	2.85 Mil
Unknown	0	15.9 K	240	0	2.10 Mil
UA	0	580	21.6	325	190 K
BG	0	419	20.9	0	183 K

In 2022: peaked at 533,000 cores! with increasing usage of high performance computing (HPC) centers

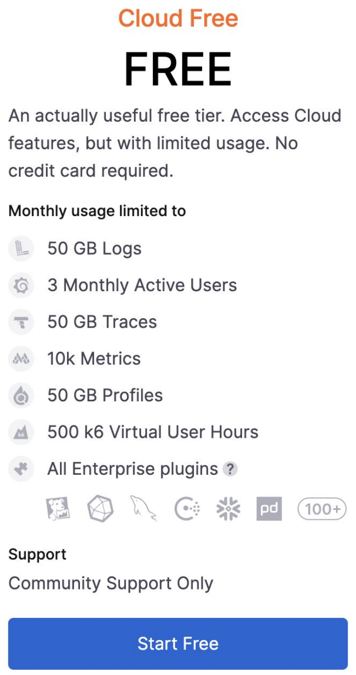
CMS global computing model

CMS Job Monitoring

Resource Utilization

Last week

24.07-31.07. 2023



Cloud Free

FREE

An actually useful free tier. Access Cloud features, but with limited usage. No credit card required.

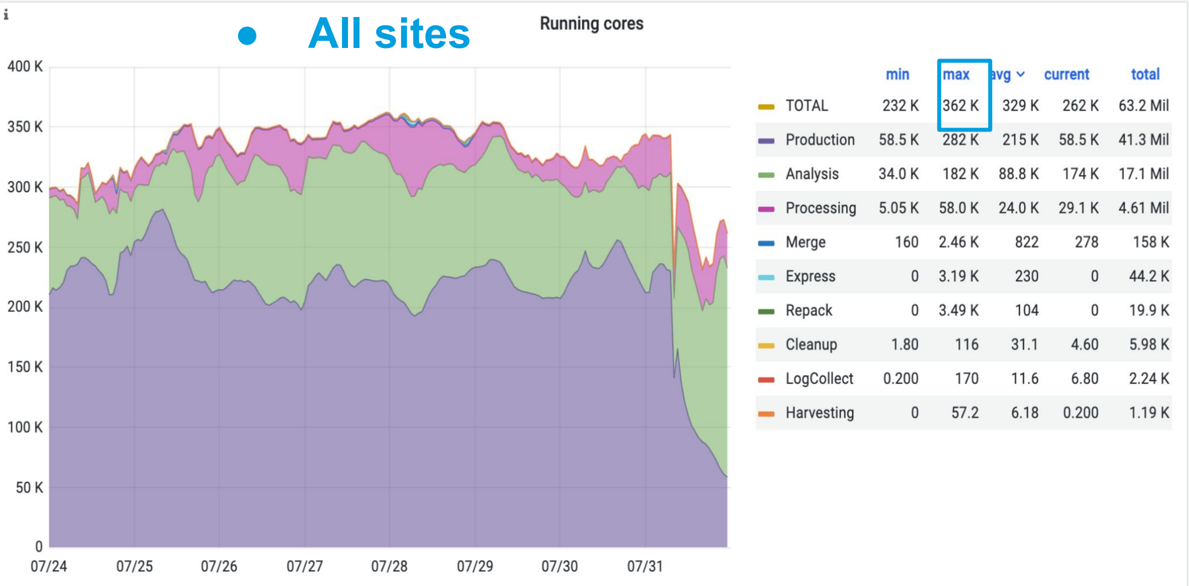
Monthly usage limited to

- 50 GB Logs
- 3 Monthly Active Users
- 50 GB Traces
- 10k Metrics
- 50 GB Profiles
- 500 k6 Virtual User Hours
- All Enterprise plugins ?

Support

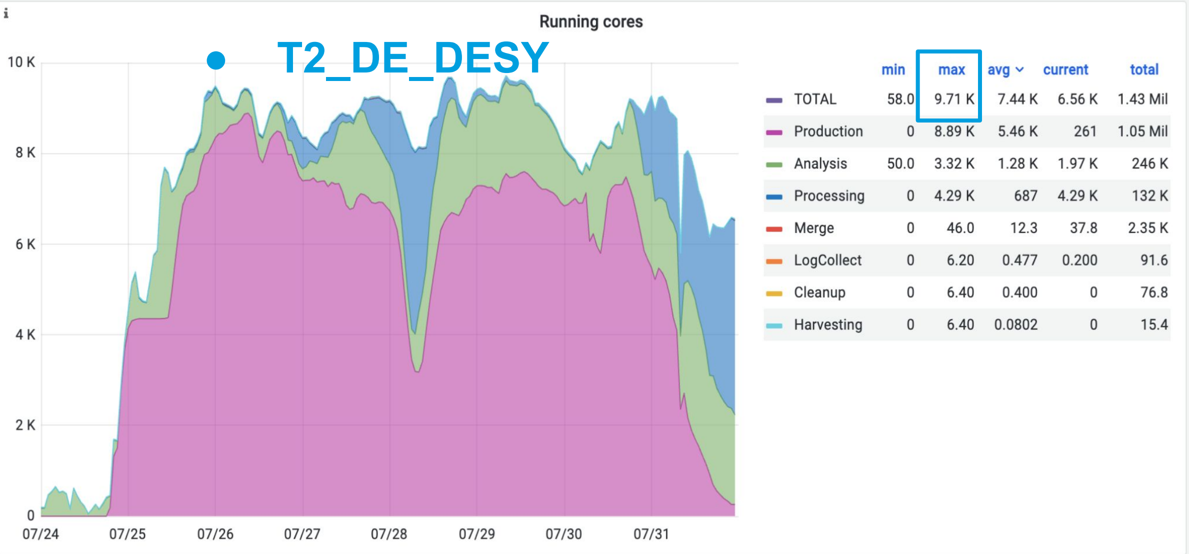
Community Support Only

Basic subscription free, price plans according to needs



- All sites

	min	max	avg	current	total
TOTAL	232 K	362 K	329 K	262 K	63.2 Mil
Production	58.5 K	282 K	215 K	58.5 K	41.3 Mil
Analysis	34.0 K	182 K	88.8 K	174 K	17.1 Mil
Processing	5.05 K	58.0 K	24.0 K	29.1 K	4.61 Mil
Merge	160	2.46 K	822	278	158 K
Express	0	3.19 K	230	0	44.2 K
Repack	0	3.49 K	104	0	19.9 K
Cleanup	1.80	116	31.1	4.60	5.98 K
LogCollect	0.200	170	11.6	6.80	2.24 K
Harvesting	0	57.2	6.18	0.200	1.19 K



- T2_DE_DESY

	min	max	avg	current	total
TOTAL	58.0	9.71 K	7.44 K	6.56 K	1.43 Mil
Production	0	8.89 K	5.46 K	261	1.05 Mil
Analysis	50.0	3.32 K	1.28 K	1.97 K	246 K
Processing	0	4.29 K	687	4.29 K	132 K
Merge	0	46.0	12.3	37.8	2.35 K
LogCollect	0	6.20	0.477	0.200	91.6
Cleanup	0	6.40	0.400	0	76.8
Harvesting	0	6.40	0.0802	0	15.4

CMS global computing model

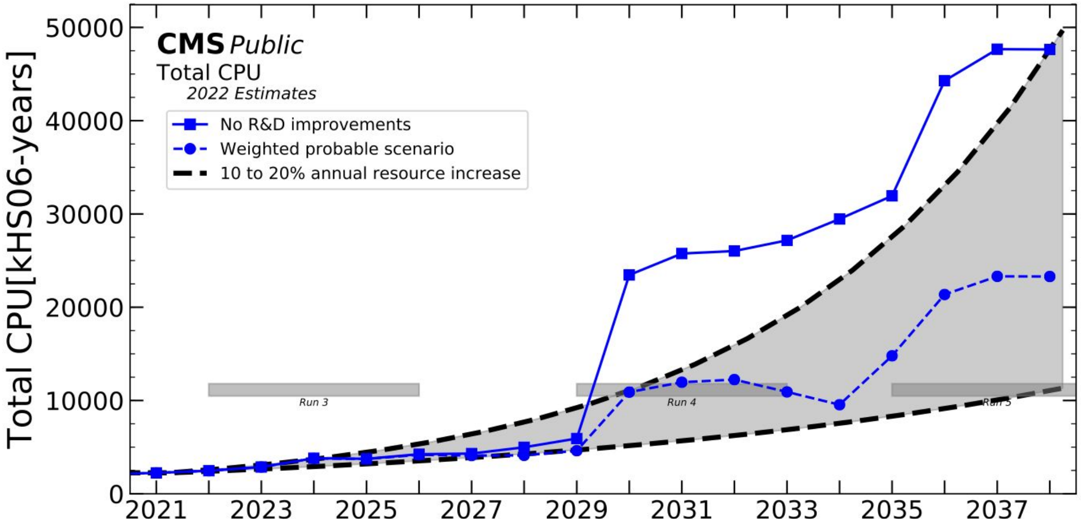
Projections for High luminosity LHC (HL-LHC)

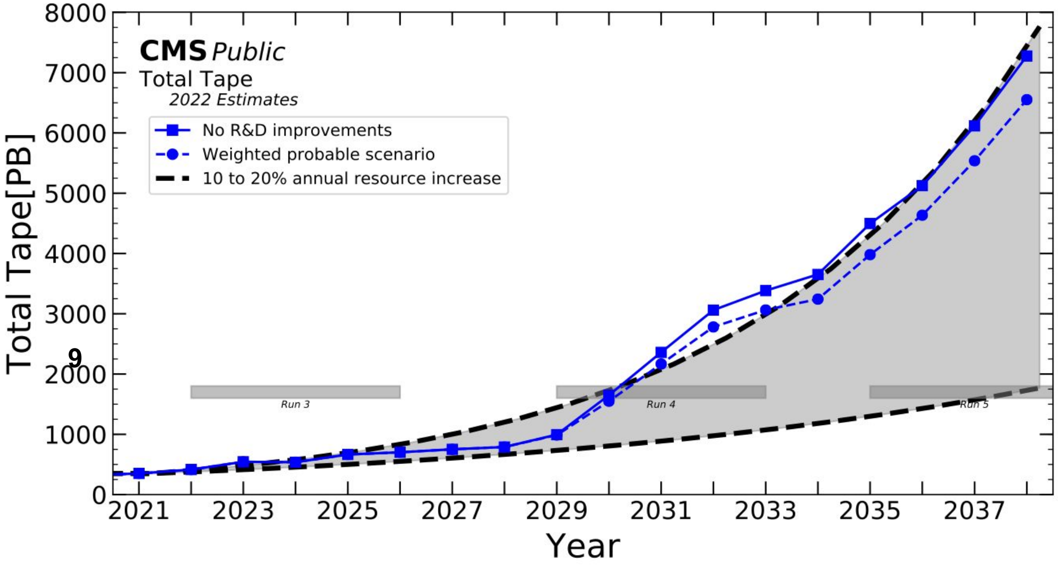
CMSOfflineComputingResults

> Approximate breakdown of CPU time, disk, and tape requirements into primary processing and analysis activities during a typical HL-LHC year

> baseline scenario is considered

i.e. projected effects from on-going R&D to reduce the computing resources needed by CMS not considered





CMS global computing model

Data storage

> Magnetic tape:

- long-term archiving
- thousands of cartridges at T0 and T1 sites

CERN-OBJ-IT-137

held inside tape libraries with robotic arms that load them into tape drives where they can be read and written

> Hard disk drives:

- distributed access at T1, T2 and T3 sites
- Random access: job can read any data
- Needs to be staged from tape first

CMS global computing model

Data management

> During Run 1 and Run 2, CMS used PhEDEx and Dynamo as data management tools

For Run 3 and beyond

> need to adopt a more scalable, flexible and powerful system to

- increasingly automate the data management
- include possibility to scale up transfers to around 100 petabytes per day by late 2020s for start of the High Luminosity LHC

> Rucio system

Data management project for scientific communities adopted by CMS at end of 2020

> Incorporated higher-level decisions about data placement

CMS global computing model

Data management

> All Rucio services built into single Kubernetes cluster which can be brought up from scratch in under an hour

> Rucio removes data as additional space is needed at a site

> Only data that is not held in place by one or more rules eligible to be removed

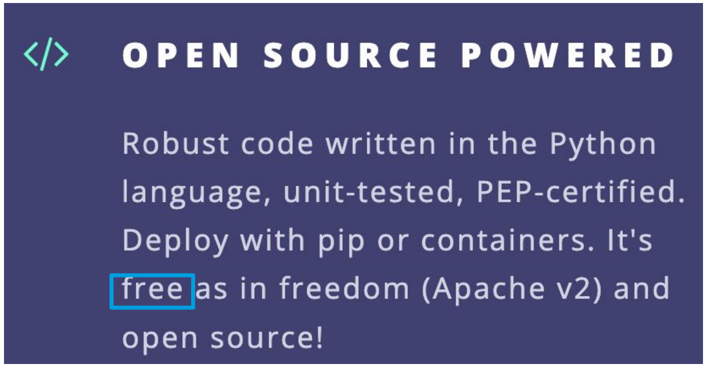
> To make this decision, Rucio uses last access time of data.

> This information is taken from job reports, CMSSW file reads, and monitoring of AAA (Any data, Anywhere, Any time) system

SCIENTIFIC DATA MANAGEMENT

RUCIO

https://github.com/rucio



</> OPEN SOURCE POWERED

Robust code written in the Python language, unit-tested, PEP-certified. Deploy with pip or containers. It's free as in freedom (Apache v2) and open source!

CMS global computing model

Computing infrastructure at TIERs

Minimal Tier-2 Setup

- **1000 CPU cores** (i.e. 10 machines with dual-socket and 48-core CPUs) compute servers, i.e. worker nodes, should have 2 GB of memory per core
  - ❖ 64-bit Linux (preferentially Alma or another RedHat Enterprise Linux compatible distribution) with CVMFS running on the worker nodes
  - ❖ worker nodes also need scratch space for running jobs, about 20 GB per core
- **500 TBytes of disk space** (i.e. one 36-disk server with 18 TB hard drives)
- a dual-stack, IPv4 and v6 accessible, WebDAV and XRootD storage endpoint
- **2 Gbps Internet connectivity**
- a batch system with 8-core/48-hour queues and a server for either an ARC or HTCondor compute element interface/service
- a small server for HTTP caching
- support during work days/hours

CMS global computing model

Event data model

- centered around concept of an ***Event***

- Physically, a collection of information about a single beam crossing ("collision") stored in memory for processing then written to file (tape/disk)

- In software terms, it starts as a collection of RAW data from a detector or MC event, stored as a single entity in memory, a C++ type-safe container called `edm::Event`

- An Event is a C++ object container for all RAW and reconstructed data related to a particular collision

- During processing, data are passed from one module to next via the Event, and are accessed only through the Event

CMS global computing model

Data tiers

GEN: Intermediate and outgoing stable particles from the collision simulation

SIM: Detailed description of energy deposits left by stable outgoing particles in the detector material

Two options available:

Full MC: highly accurate GEANT4-based application

Fast MC: parametric fast simulation application

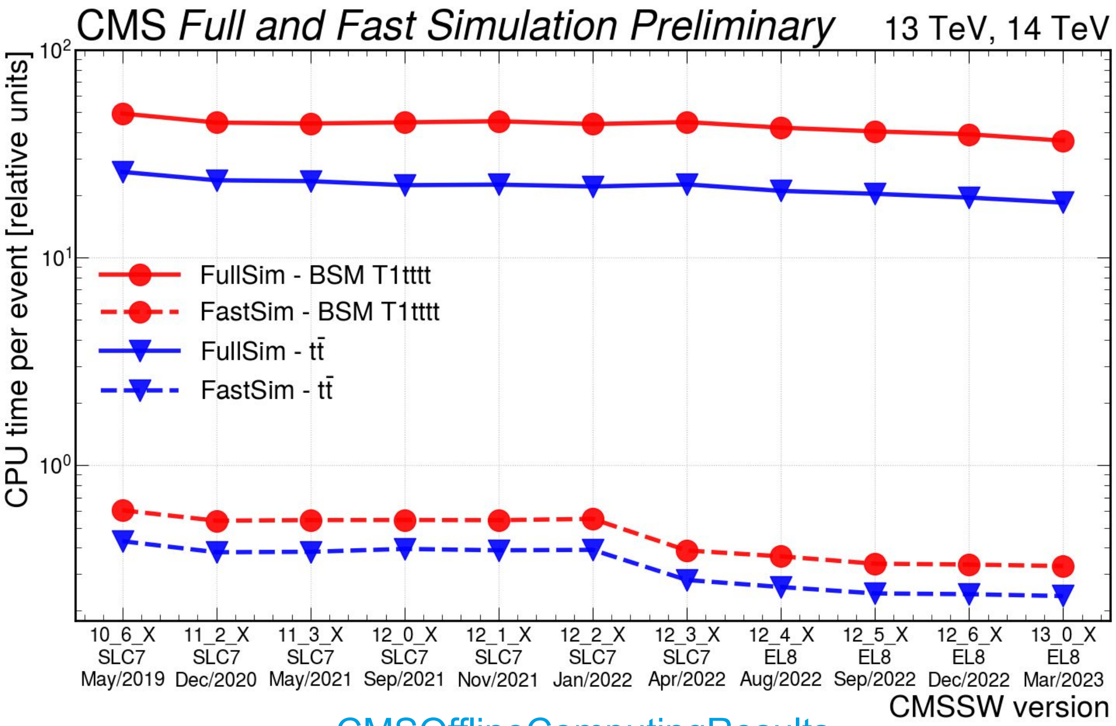
DIGI: Digitized detector readout or simulation

CMS global computing model

Full vs Fast Simulation comparison

Fast simulation:

- trades accuracy for a 100-fold decrease in detector simulation time or 10-fold decrease in total CPU time per simulated event
- level of inaccuracy introduced typically less than 10% wrt to Full simulation in final analysis observables



CMSOfflineComputingResults

- With a small cluster prioritise fast simulation could be a good approach

CMS global computing model

Data tiers

RAW: Packed detector readout data

RECO: Detailed description of calibrated detector hits and low-level physics objects

AOD: Reduced description of calibrated detector hits and low-level physics objects, uncalibrated high-level physics objects

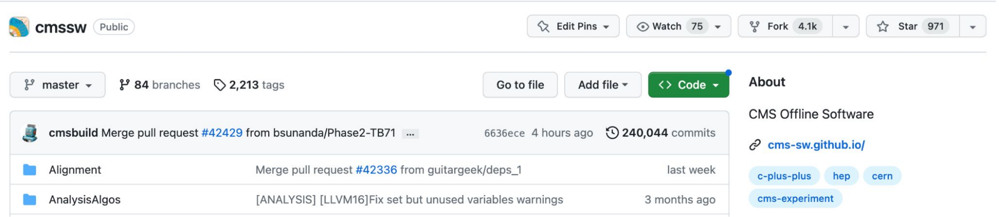
MiniAOD: Reduced low-level physics objects and calibrated high-level physics objects

truncated floating-point representation used for most object attributes. Introduced for Run 2

NanoAOD: Compact data format containing only high-level physics object attributes stored as (arrays of) primitive data types, introduced during Run 2

CMS global computing model

Software for data processing



cmsbuild Merge pull request #42429 from bsunanda/Phase2-TB71 …	6636ece 4 hours ago	240,044 commits
Alignment	Merge pull request #42336 from guitargeek/deps_1	last week
AnalysisAlgos	[ANALYSIS] [LLVM16]Fix set but unused variables warnings	3 months ago

About

CMS Offline Software

cms-sw.github.io/

c-plus-plus hep cern cms-experiment

- ~6M lines of code organized into directories: Subsystem/Packages
- + associated tools, build system (SCRAM), etc
- ~500 external packages (outside software distributed w/ CMSSW)

CMSSW executable (one), **cmsRun** and many plug-in modules

configured at run time by the user's job-specific configuration file. This file tells cmsRun

- which data to use
- which modules to execute
- which parameter settings to use for each module
- what is the order or the executions of modules, called *path*
- how the events are filtered within each path, and
- how the paths are connected to the output files

CMS computing at DESY

> DESY hosts one of the **Tier-2** centers - operated by the DESY CMS group

- CPU capacity is divided
  - central production of MC events
  - used by individual users
- Storage capacity is also split
  - space centrally managed by CMS
  - local and national user community

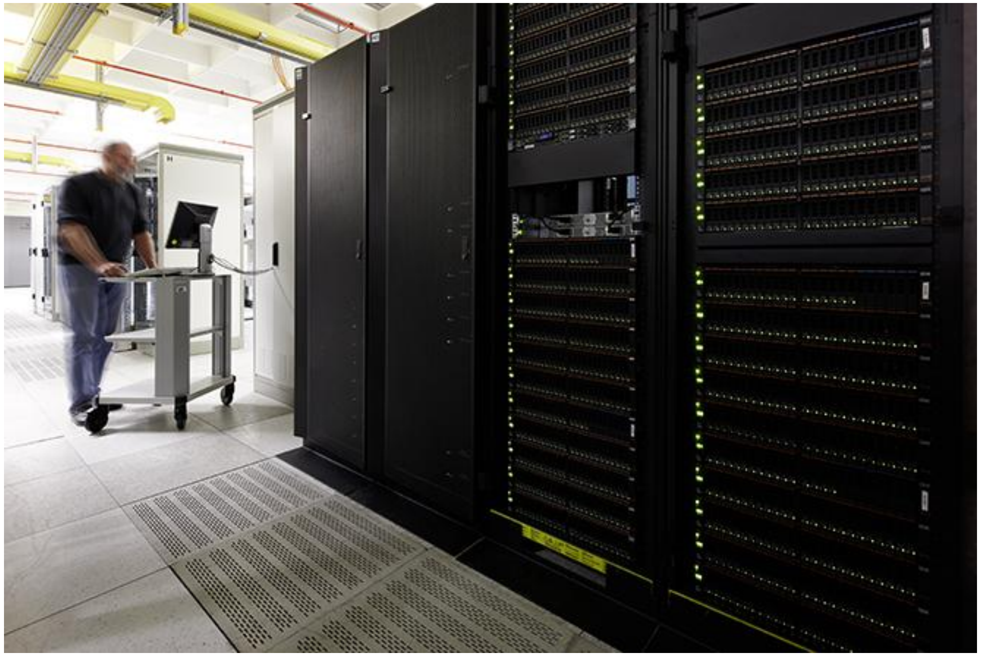


Foto from FH@DESY

Aided Complemented

Large batch facility -> National Analysis Facility (NAF)

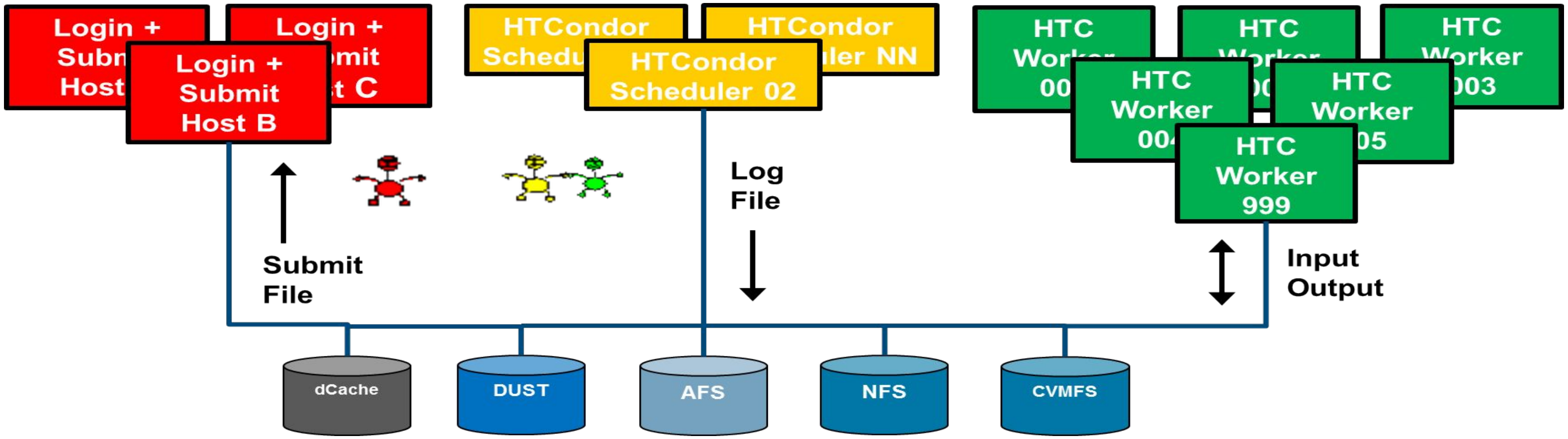
- Interactive Workgroup Servers: users can access the services through them
- Direct access to the Tier-2 resources located at DESY
- Fast cluster file system ideal for reading and writing large files by means of a high bandwidth

BIRD for batch processing

> Large computing infrastructure for batch processing at DESY

- Technology deployed and relaying on mounted filesystems
  - **Data should be directly accessible on the workernodes** for a ‘per-job’ basis execution
  - Example of integration of various subsystems: AFS, DUST, PNFS/DCACHE, CVMFS
  - This data can include:
    - Individual user data (e.g. personal work-space & code, your BSM MC simulation)
    - Central experiment-specific software and data (e.g. CMSSW, MC ‘grid-packs’)
    - Externally provided software and data (e.g. ROOT, LHAPDF)
- Powered by Scientific Linux (SL)
  - Singularity containers available for older versions of SL -> *reliable functionality w/o requirements on software backward compatibility*
  - Easy access to **free software** and possibility to host a number of licensed software with time-dynamic usage by several users -> *lower number of licenses for more users*

Foto from Confluence@DESY

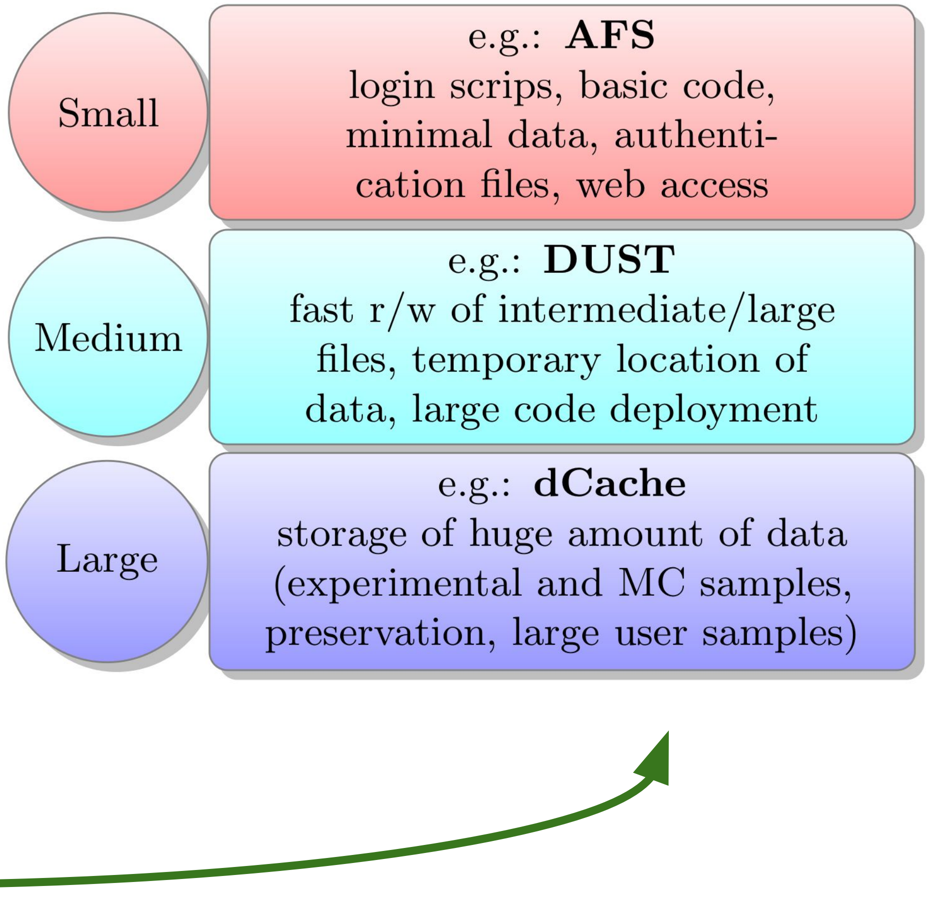


Different technology for different data

> Experimental collaboration and user needs varies depending on the stage of data processing

> Event sample production workflow might vary significantly

- Initially large amounts of space is required to store as much information as possible from a collision event
  - MC generated events typically not the biggest problem (provided that advanced compression formats such as ROOT are used)
  - Detector simulation and event reconstruction usually takes up most of the CPU and storage capacity
- Final stages require fast accessing and processing over large datasets but produce smaller outputs



Small

e.g.: **AFS**
login scrips, basic code, minimal data, authentication files, web access

Medium

e.g.: **DUST**
fast r/w of intermediate/large files, temporary location of data, large code deployment

Large

e.g.: **dCache**
storage of huge amount of data (experimental and MC samples, preservation, large user samples)

Alternative: Differentiated Storage

Multi-purpose batch-cluster for data processing

- Data access through shared filesystems makes it easier for collaboration and resource saving
  - Any Data, Anytime, Anywhere (AAA): provided by **XROOTD** in CMS
    - Remote access to huge databases using scalable architecture and communication protocol
    - Logical File Name (LFN) identifies files anywhere within all (disk) storage
  - Global Grid resources access through Grid user certificate (authentication) and Virtual Organizations (authorization)

<- need a powerful batch processing infrastructure ->

HTCondor
Software Suite

Automating and managing of jobs via HTCondor

- ★ Ability to manage shared resources with distributed ownership of various kind (single/set of clusters, cloud resources, temporary)
- ★ Ideal for splitting processing over large datasets
- ★ Powerful match-making between owners and consumers of resources, on request and by priority
- ★ Meta-scheduler realization for inter-dependent jobs via DAGman
- ★ Easy command interface for users with extensive documentation and wide use across many CERN associated institutes and CERN itself

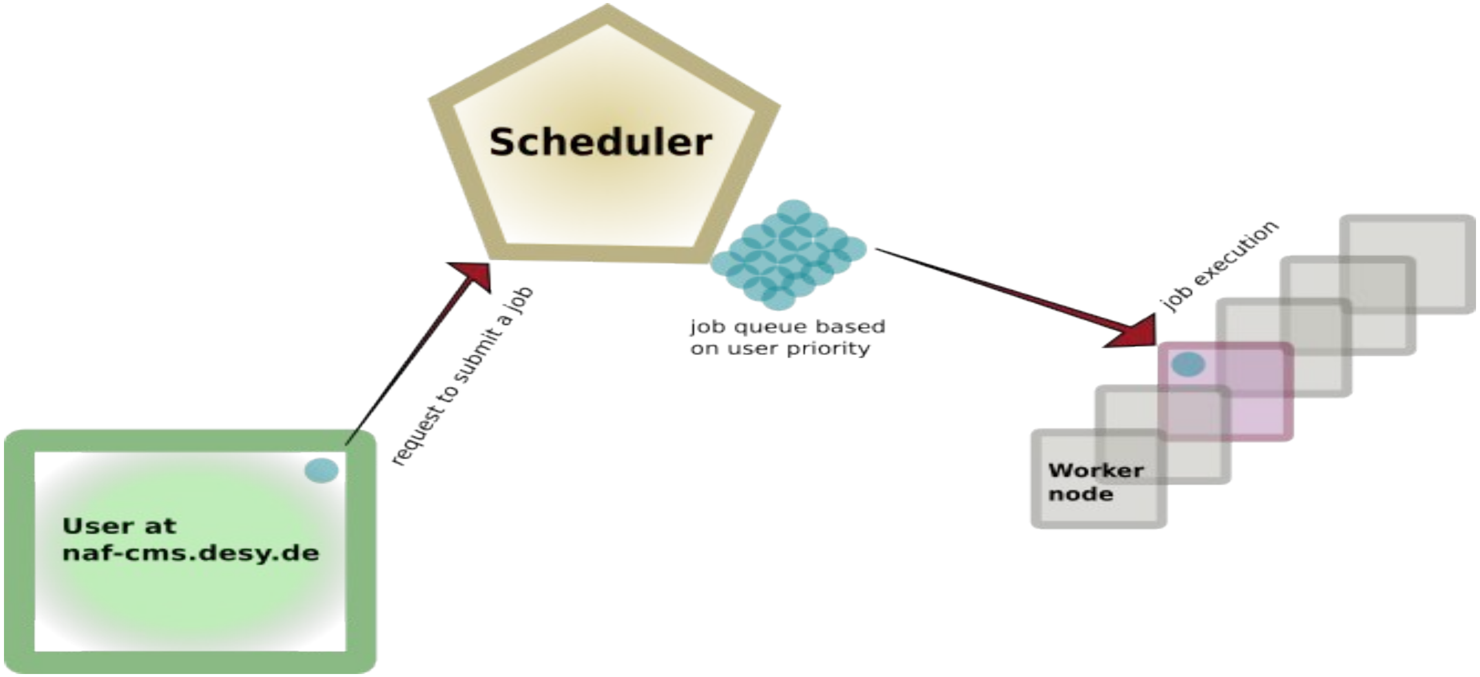
Multi-purpose batch-cluster for data processing

Easy remote access to WGS via SSH

HTCondor Software Suite

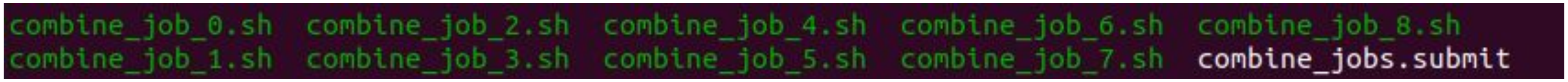
Typical generic jobs can be run with 1 core, 2GB of memory and 3h allocated time

Adapting some resource-intensive (e.g. full simulation) jobs to this constraints can be difficult



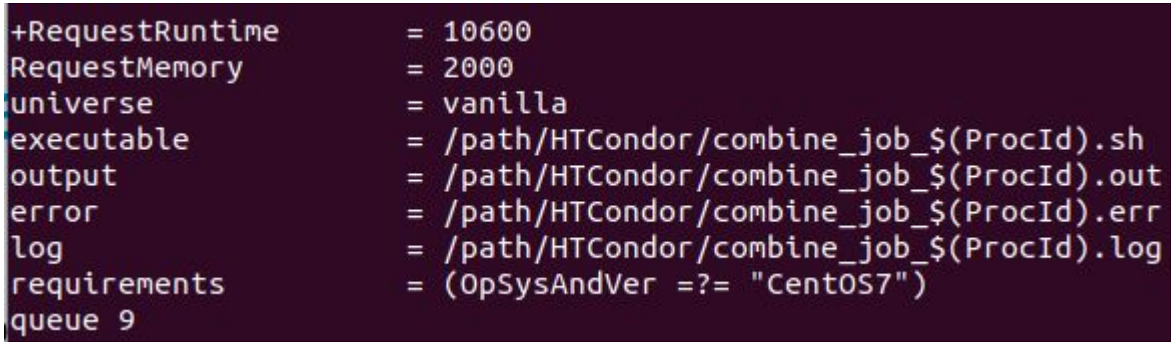
Multi-purpose batch-cluster for data processing

Simple HTCondor example



```
combine_job_0.sh  combine_job_2.sh  combine_job_4.sh  combine_job_6.sh  combine_job_8.sh
combine_job_1.sh  combine_job_3.sh  combine_job_5.sh  combine_job_7.sh  combine_jobs.submit
```

A set of bash scripts increasingly numbered



```
+RequestRuntime       = 10600
RequestMemory         = 2000
universe              = vanilla
executable            = /path/HTCondor/combine_job_$(ProcId).sh
output                = /path/HTCondor/combine_job_$(ProcId).out
error                 = /path/HTCondor/combine_job_$(ProcId).err
log                   = /path/HTCondor/combine_job_$(ProcId).log
requirements          = (OpSysAndVer =?= "CentOS7")
queue 9
```

seconds
MB
environment
executable
print output
print errors
HTC log stat
OS
Number of jobs

A HTCondor configuration file (**condor_submit**) for submission of all jobs in one single task

Jobs/machines can be easily checked and managed with a suite of different command tools

Experience as user and typical needs

> Particle-level MC event generation can be handled with O(100) cores for an average of O(100K) total events per sample running during a few hours

- Including full detector simulation changes the picture dramatically

> Storage at particle-level (e.g. HEPMC) is on the range of O(10) GBs for such a sample

- feasible for phenomenological studies w/o full detector information

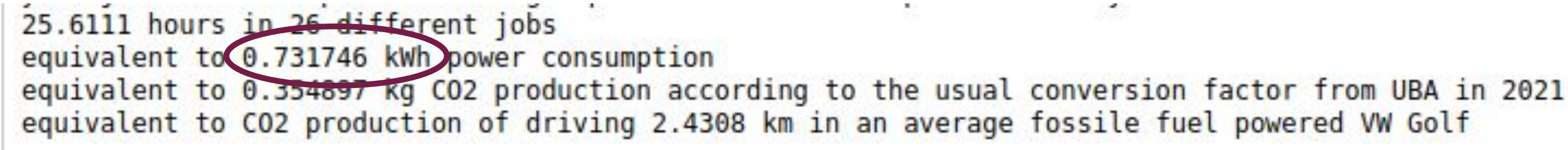
> Software sharing and accessibility through mounted filesystems is of great advantage for job operation and environment homogeneity

> Realistic user fair-share should always be taken into consideration during the high-demanded working hours of the cluster

Experience as user and typical needs

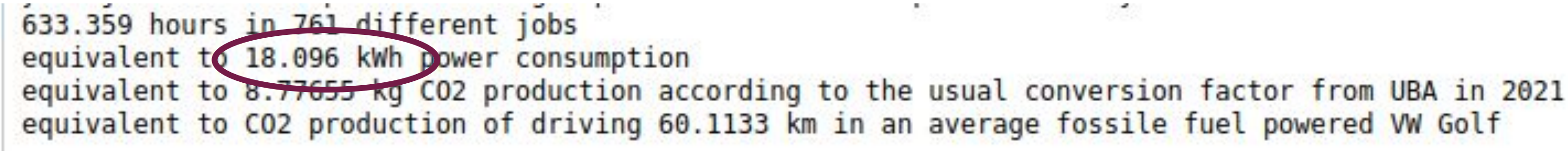
Example of weekly consumption for a single user (NAF)

“Relaxed” week of job submission: statistics analysis, plotting scripts



```
25.6111 hours in 26 different jobs
equivalent to 0.731746 kWh power consumption
equivalent to 0.354897 kg CO2 production according to the usual conversion factor from UBA in 2021
equivalent to CO2 production of driving 2.4308 km in an average fossile fuel powered VW Golf
```

Intensive week of job submission: MC generation



```
633.359 hours in 761 different jobs
equivalent to 18.096 kWh power consumption
equivalent to 8.77655 kg CO2 production according to the usual conversion factor from UBA in 2021
equivalent to CO2 production of driving 60.1133 km in an average fossile fuel powered VW Golf
```

Summary

> Scale of CMS computing model is commensurate with the mission of the experimental facility

> Distributed computing environment required by data volume and spread of resources across different territories

> DESY hosts one important CMS Tier-2 center and provides computing infrastructure for a national-wide facility

> A small batch-cluster handled with free-software alternatives such as HTCondor seems robust, uncomplicated, and adaptable

Thank you!

Contact

DESY. Deutsches
Elektronen-Synchrotron

www.desy.de

Sandra Consuegra Rodríguez (DESY)
Luis Ignacio Estévez Baños (DESY)
Danyer Pérez Adán (DESY)

0000–0002–1383–1837
0000-0001-6195-3102
0000-0003-3416-0726

sandra.consuegra.rodriguez@desy.de
luis.estevez.banos@desy.de
danyer.perez.adan@desy.de

Backup

> **Additional material**

CMS global computing model

Computing infrastructure at TIERs

Small to Moderate Tier-1 Setup

- Tier-1 pledges usually aligned with number of authors of country
- countries with the most authors provide one

Specification for country with modest to large number of authors:

- **2000 to 8000 CPU Cores** (i.e. about 20 to 80 dual socket machines) as worker nodes with 10 Gbps connection to the storage
- **2.5 to 10 PBytes of disk space** (performance of spinning magnetic disks in a RAID or Erasure Coding is sufficient)
- **8 to 32 PBytes of archival storage** (usually tape in an automated library)
- two servers for a redundant Compute Element setup with access to a local batch system spanning the worker nodes
- two servers for a redundant HTTP cache setup
- **100 Gbps network connectivity** to CERN
- 24x7 support