Advances in software management tools

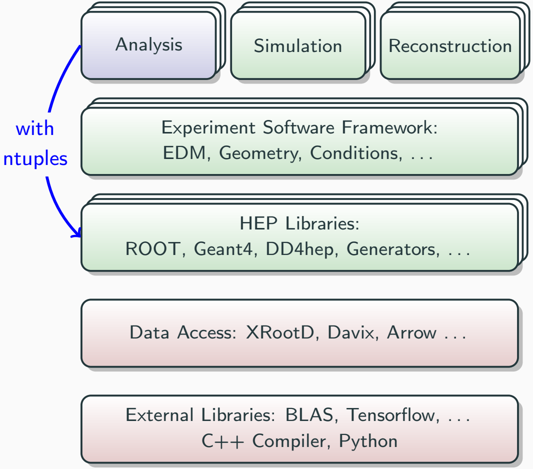
Jakob Blomer, CERN, EP-SFT

LHCP 2021
9 June 2021

LHC Experiment Software Stacks and Targets

Daily **builds** of hundreds of interdependent packages

Distribution to $\mathcal{O}$(1million) *hetereogeneous* cores



Analysis

Simulation

Reconstruction

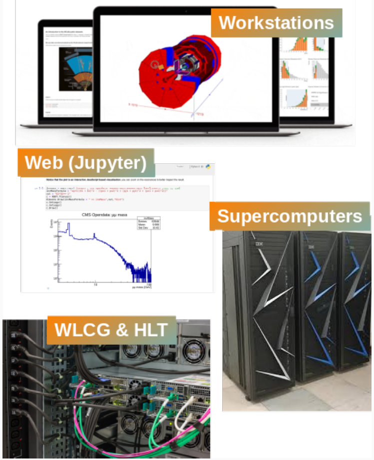
with ntuples

Experiment Software Framework:
EDM, Geometry, Conditions, . . .

HEP Libraries:
ROOT, Geant4, DD4hep, Generators, . . .

Data Access: XRootD, Davix, Arrow . . .

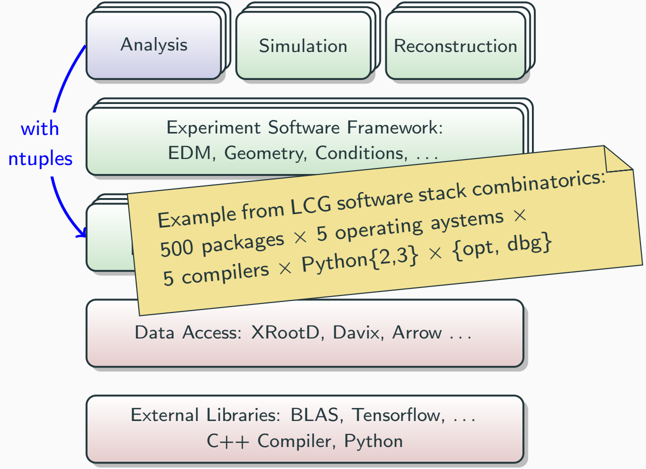
External Libraries: BLAS, Tensorflow, . . .
C++ Compiler, Python



LHC Experiment Software Stacks and Targets

Daily **builds** of hundreds of interdependent packages

Distribution to $\mathcal{O}$(1million) *hetereogeneous* cores



Analysis

Simulation

Reconstruction

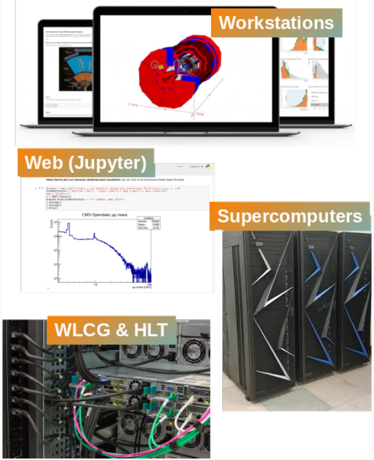
with ntuples

Experiment Software Framework:
EDM, Geometry, Conditions, . . .

Example from LCG software stack combinatorics:
500 packages × 5 operating aystems ×
5 compilers × Python{2,3} × {opt, dbg}

Data Access: XRootD, Davix, Arrow . . .

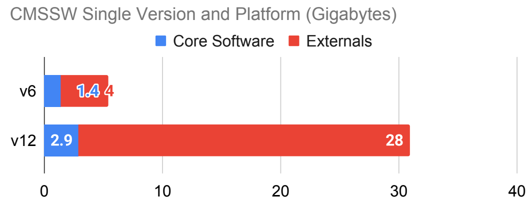
External Libraries: BLAS, Tensorflow, . . .
C++ Compiler, Python



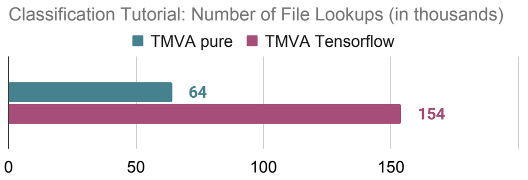
Software Management for HL-LHC

Compared to run 1-2, we now find

- Multiple target architectures: x86_64 micro-architectures (e. g. AVX512), AArch64, Power, GPUs
- A growing Python software ecosystem, in particular for machine learning tasks
- More agile software development: automated integration builds, nightly builds
- Generally we tend to add code and externals more often than removing software



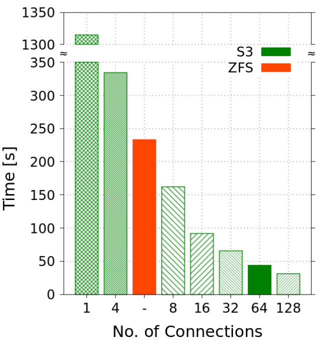
CMSSW Single Version and Platform (Gigabytes)



Classification Tutorial: Number of File Lookups (in thousands)

My estimate: the software management problem for HL-LHC grows by a factor of 3-5.

Software Management for HL-LHC



Performance engineering example: improving CernVM-FS write performance with Ceph/S3

▸ vCHEP'21

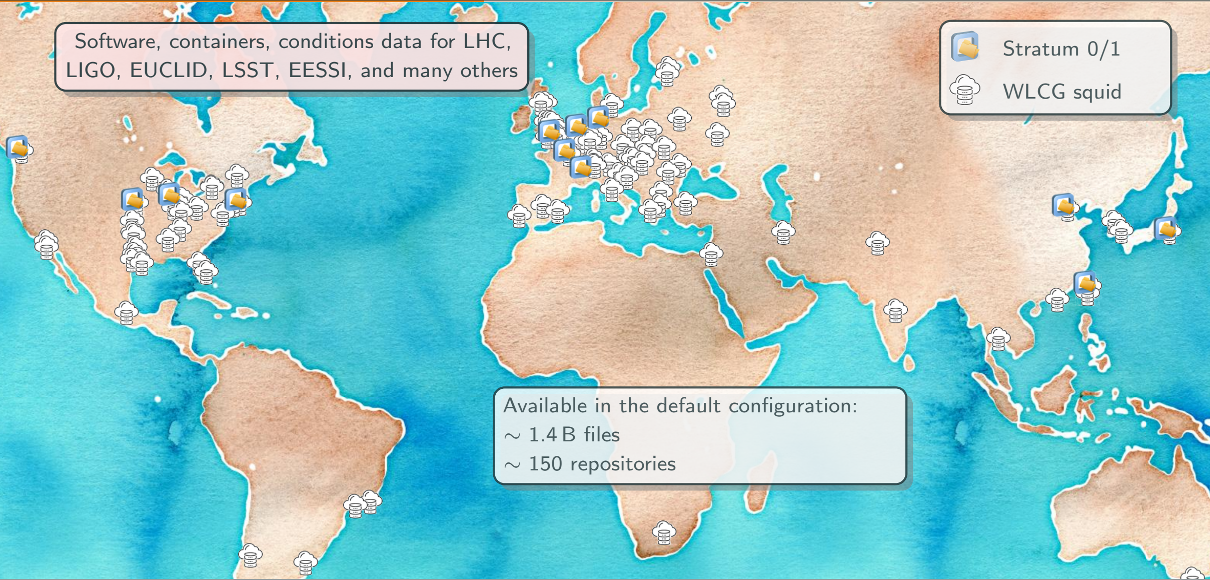
- Investment needed in the **performance** of software build, test, and distribution tools
- Constructing software stacks becomes an **engineering discipline** ("software librarian")
  - Tendency to build complete stack including OS layer → independent from target node OS
  - Dedicated projects for turnkey stacks
    ▸ LCG ▸ Key4HEP
  - Significant effort is going into adopting Spack package manager for maintaining HEP software stacks
- Preserving stacks gets harder: preservation tools are available but we risk to "capture the mess"

Turnkey Stacks: Example

```
[jblomer@lxplus]$ source \
  /cvmfs/sft.cern.ch/lcg/views/LCG_100/x86_64-centos7-gcc10-opt/setup.sh
[jblomer@lxplus]$ root --version
ROOT Version: 6.24/00
Built for linuxx8664gcc on Apr 14 2021, 14:33:50
From tags/v6-24-00@v6-24-00
[jblomer@lxplus]$ python -V
Python 3.8.6
[jblomer@lxplus]$ python -c "import tensorflow"
...
```

Software Distribution: Status of CernVM-FS

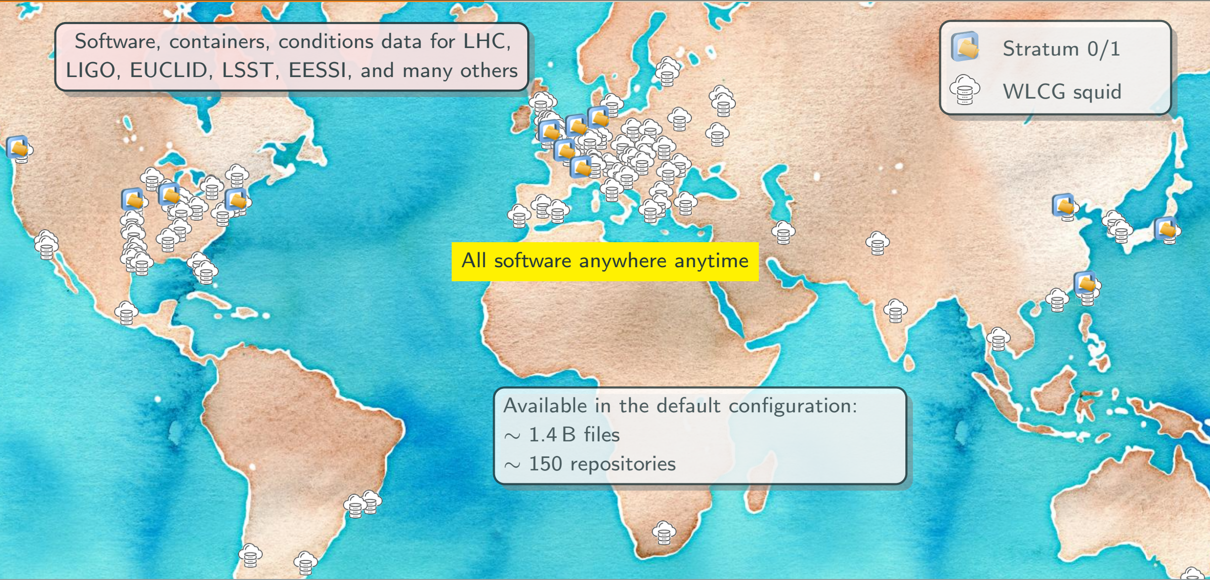
Software, containers, conditions data for LHC, LIGO, EUCLID, LSST, EESSI, and many others



Available in the default configuration:
∼ 1.4 B files
∼ 150 repositories

Software Distribution: Status of CernVM-FS

Software, containers, conditions data for LHC, LIGO, EUCLID, LSST, EESSI, and many others



Stratum 0/1

WLCG squid

All software anywhere anytime

Available in the default configuration:

$\sim$ 1.4 B files

$\sim$ 150 repositories

Containers for LHC Experiment Applications

SINGULARITY

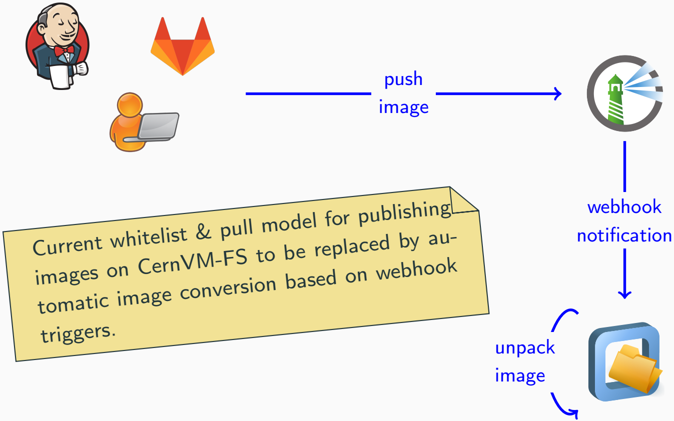
kubernetes

- Decouples application from node operating system
- Solves the performance overhead problem of VMs
- Convenient to compose new container images based on existing layers → excellent tool to explore new software
- Allows for capturing a reproducible software environment ▸ HEP Benchmarks

- Poorly addresses the distribution problem at scale
- Facilitates a “black box” approach that impedes a proper understanding of the software stack inside

A smart approach to containers combines their isolation capabilities
with the distribution efficiency and well-maintained content of CernVM-FS

CernVM-FS as a Container Hub



Current whitelist & pull model for publishing images on CernVM-FS to be replaced by automatic image conversion based on webhook triggers.

- Centralized image management
- Vulnerability scan
- OCI artifact signing
- Aims at service deployment

- Efficient storage & distribution
- Enables large-scale image deployement
- Aims at scientific apps

CernVM-FS as a Container Hub

/cvmfs/unpacked.cern.ch

- $>$ 1000 images
- $>$ 6.5 TB
- $>$ 95 M files

/cvmfs/singularity.opensciencegrid.org

- $>$ 630 images
- $>$ 2.5 TB
- $>$ 60 M files

Images are readily available to run with singularity, including **base operating systems**, **experiment software stacks**, **explorative tools (ML etc.)**, **user analyses**, and special-purpose containers such as **folding@home**

```
[jblomer@lxplus.cern.ch]$ singularity exec \
  '/cvmfs/unpacked.cern.ch/registry.hub.docker.com/library/debian:stable' \
  cat /etc/issue
Debian GNU/Linux 10 \n \l
```

Container Runtime Integration

Runtime	Type	CernVM-FS Support
Singularity	flat (+ layers)	**native**
podman	layers	**native (use image storage from /cvmfs)**
containerd[1] / k8s	layers	**plugin** ▸ remote snapshotter
docker	layers	***"graph driver"* image storage plugin**[2]

Documentation chapter on containers & CernVM-FS:

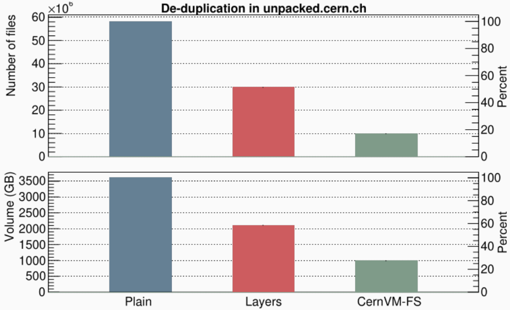
→ `https://cvmfs.readthedocs.io/en/latest/cpt-containers.html`

[1] Requires containerd version >= 1.4

[2] Expected to be replaced by containerd as foundation of docker

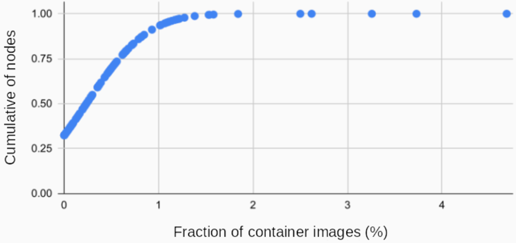
Container Distribution Efficiency

De-duplication works properly only on file-level granularity



CernVM-FS exploits that only a tiny fraction ($\sim$ 2 %) of images are used at runtime

▸ vCHEP'21

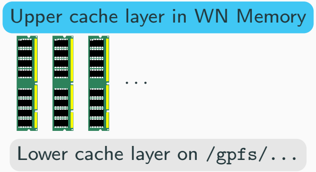


Integration of Supercomputers (HPC Sites)

During the past years, a number of CernVM-FS HPC add-ons facilitated HPC deployments
▸ CernVM-FS on HPC

- Preloader: work around compute nodes without Internet connectivity
- Tiered cache: work around diskless compute nodes

- Unprivileged mounting of /cvmfs with ▸ cvmfsexec
- Integration with HPC container workflows through singularity ▸ fusemount plugin



Upper cache layer in WN Memory

. . .

Lower cache layer on `/gpfs/...`

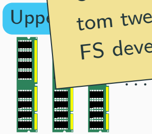
Integration of Supercomputers (HPC Sites)

During the past years, a number of CernVM-FS HPC add-ons facilitated HPC deployments
▸ CernVM-FS on HPC

- Preloader: work around compute nodes without Internet connectivity
- Tiered cache: work around diskless compute nodes

- Unprivileged mounting of /cvmfs with ▸ cvmfsexec
- Integration with HPC container workflows through singularity ▸ fusemount plugin

Onboarding a new HPC site may need some custom tweaks. There is experience in the CernVM-FS development team. Please get in touch.



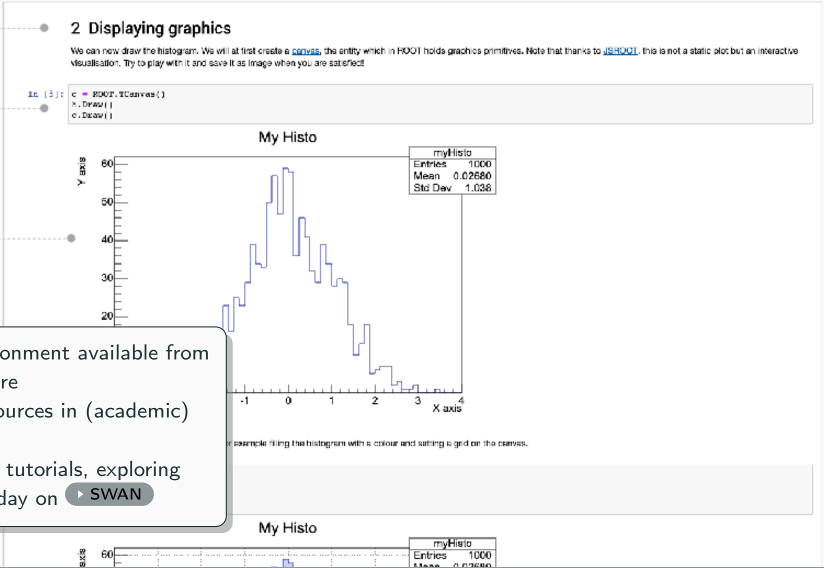
Lower cache layer on `/gpfs/...`

SWAN: Jupyter Notebook Service for HEP

Text

Code

Graphics



2 Displaying graphics

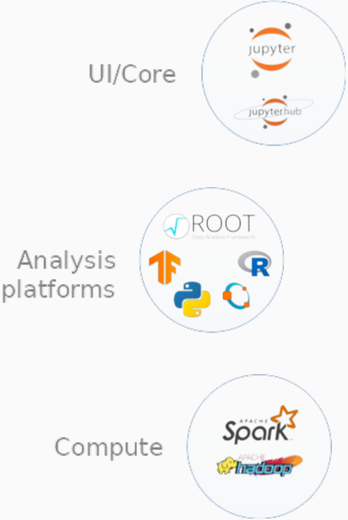
We can now draw the histogram. We will at first create a canvas, the entity which in ROOT holds graphics primitives. Note that thanks to JSROOT, this is not a static plot but an interactive visualisation. Try to play with it and save it as image when you are satisfied!

```
In [5]: c = ROOT.TCanvas()
        h.Draw()
        c.Draw()
```

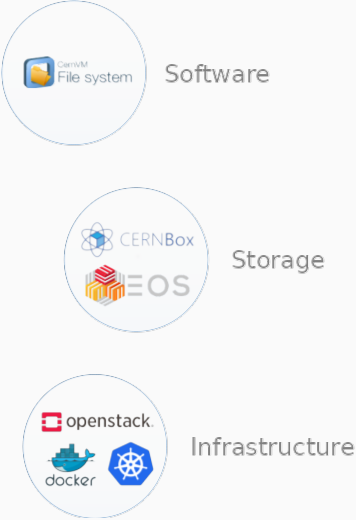
r example filling the histogram with a colour and setting a grid on the canvas.

- Data analysis environment available from all devices, anywhere
- Computational resources in (academic) clouds
- Great for teaching, tutorials, exploring
- 200-250 users per day on ▸ SWAN

SWAN: Jupyter Notebook Service for HEP



- SWAN presents the familiar environment as a notebook
- A selection of software stacks readily available
- Data access via EOS home folder
- Can scale out computation to Spark
- Can be deployed on-premise with ▸ Science Box



Summary

- Like our other LHC computing problems, the problem of software stack construction and distribution is growing in scale
  - We need to invest in the engineering of software construction
  - We need to invest in the performance of the distribution (both publishing and reading)
- We should be smart about combining industry standard tools with HEP tailor-made tools (e. g. CernVM-FS and Docker or Jupyter and EOS)
- After several years of effort, harnessing HPC environments is now a reality; still some case-by-base adjustments needed
- Jupyter notebooks—integrated with the HEP ecossytem—have matured as very useful tools for teaching and outreach

Backup Slides

Container Image Sizes

Distribution of container images sizes in
`/cvmfs/unpacked.cern.ch` and `/cvmfs/singularity.opensciencegrid.org`