SWAN: service for web-based analysis

D. Castro
On behalf of the SWAN team

https://cern.ch/swan

Apr 4th, 2019
Architects Forum Meeting

Introduction

SWAN in a Nutshell

- Analysis only with a web browser
  - No local installation needed
  - Based on **Jupyter Notebooks**
  - Calculations, input data and results “in the Cloud”
- Support for multiple analysis ecosystems
  - ROOT, Python, R, Octave
- Easy sharing of scientific results: plots, data, code
- Integration with CERN resources

Spark > physics_analysis_using_swan_spark_template
(autosaved)

FILE EDIT VIEW INSERT CELL KERNEL HELP

Not Trusted | Python 2

Markdown

Integration of SWAN with Spark clusters

This notebook demonstrates the functionality provided by a SWAN prototype machine that allows to offload computations to an external Spark cluster. The Spark version we are going to use is 2.1.0 and we are going to connect to the analytix cluster (as previously selected in the SWAN web form).

Step 1 - Acquire the necessary credentials to access the Spark cluster.

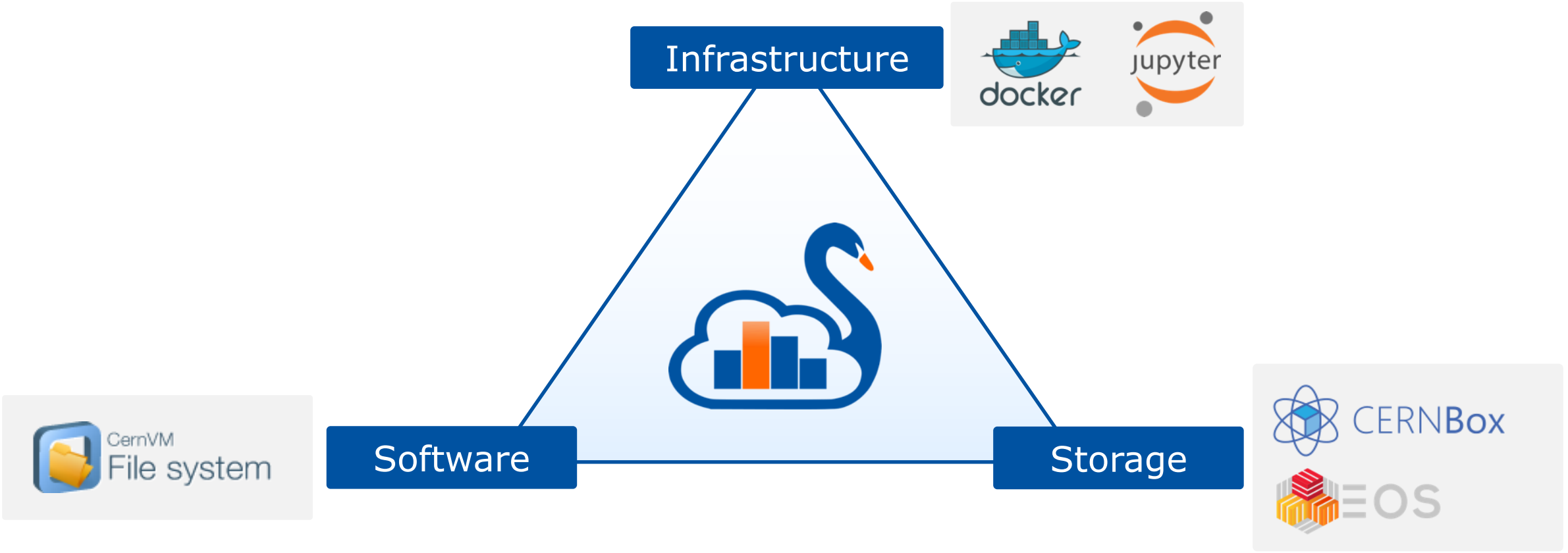
In [1]:
```
import getpass
import os, sys, re

print("Please enter your password")
ret = os.system("echo \"%s\" | kinit" % re.escape(getpass.getpass()))

if ret == 0: print("Credentials created successfully")
else:        sys.stderr.write('Error creating credentials, return code: %s\n' % ret)
```

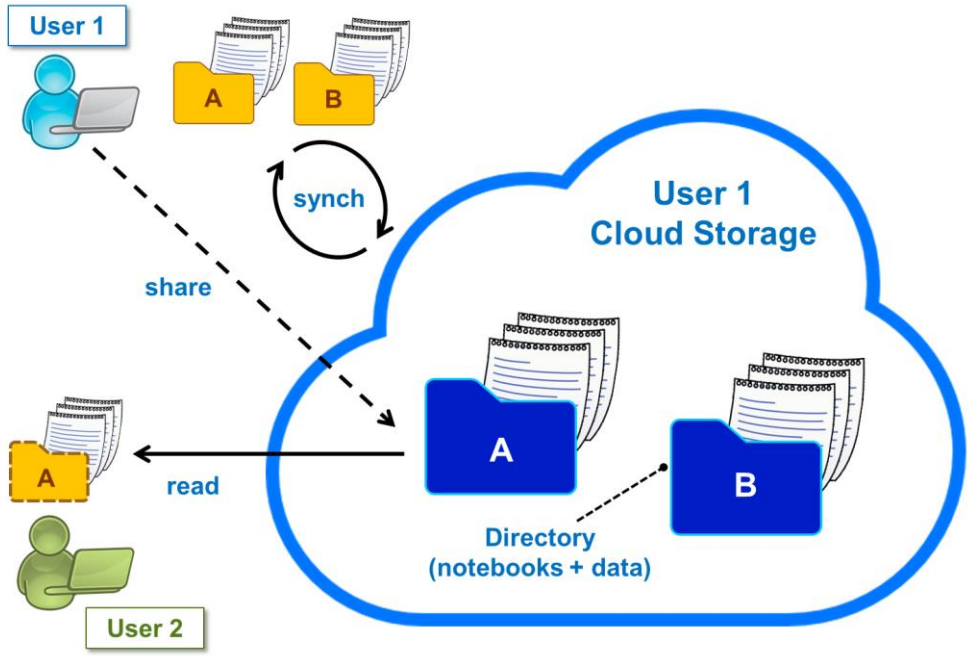

CERN

Integrating services



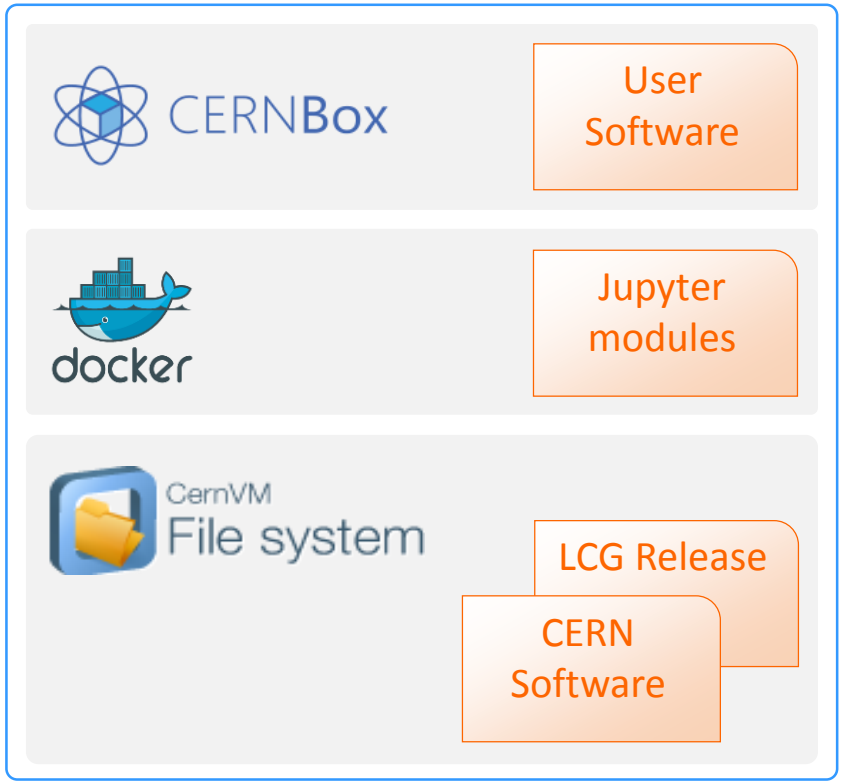
Storage

- Uses EOS disk storage system
  - Same mountpoints as LXPlus
- CERNBox is SWAN's home directory
  - Storage for your notebooks and data
- Sync&Share
  - Files synced across devices and the Cloud
  - Collaborative analysis



Software

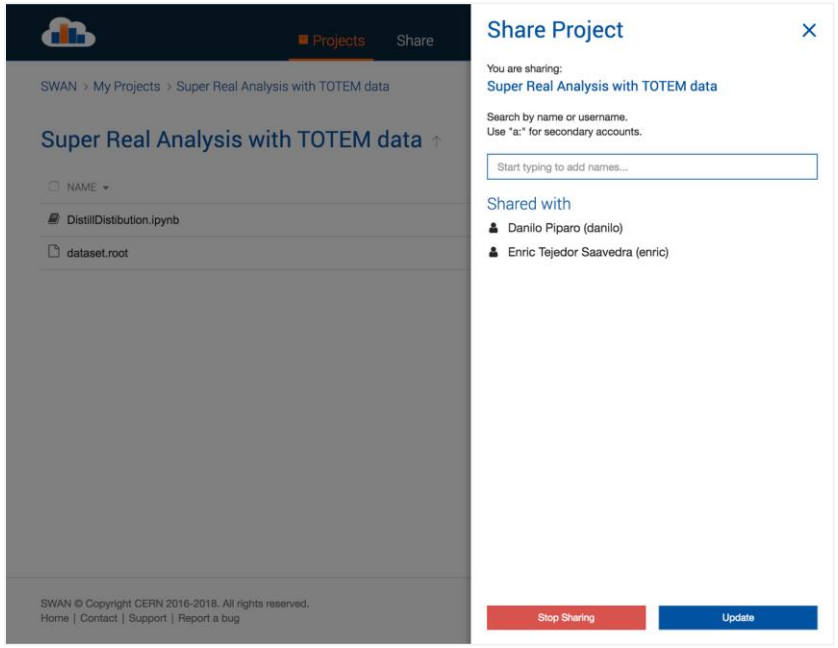
- Software distributed through CVMFS
  - LCG Releases
  - Same mountpoints as LXPlus
  - Step towards reproducibility (across time and people)
- Possibility to install libraries in user cloud storage
  - Good way to use custom/not mainstream packages
  - Configurable environment



SWAN features

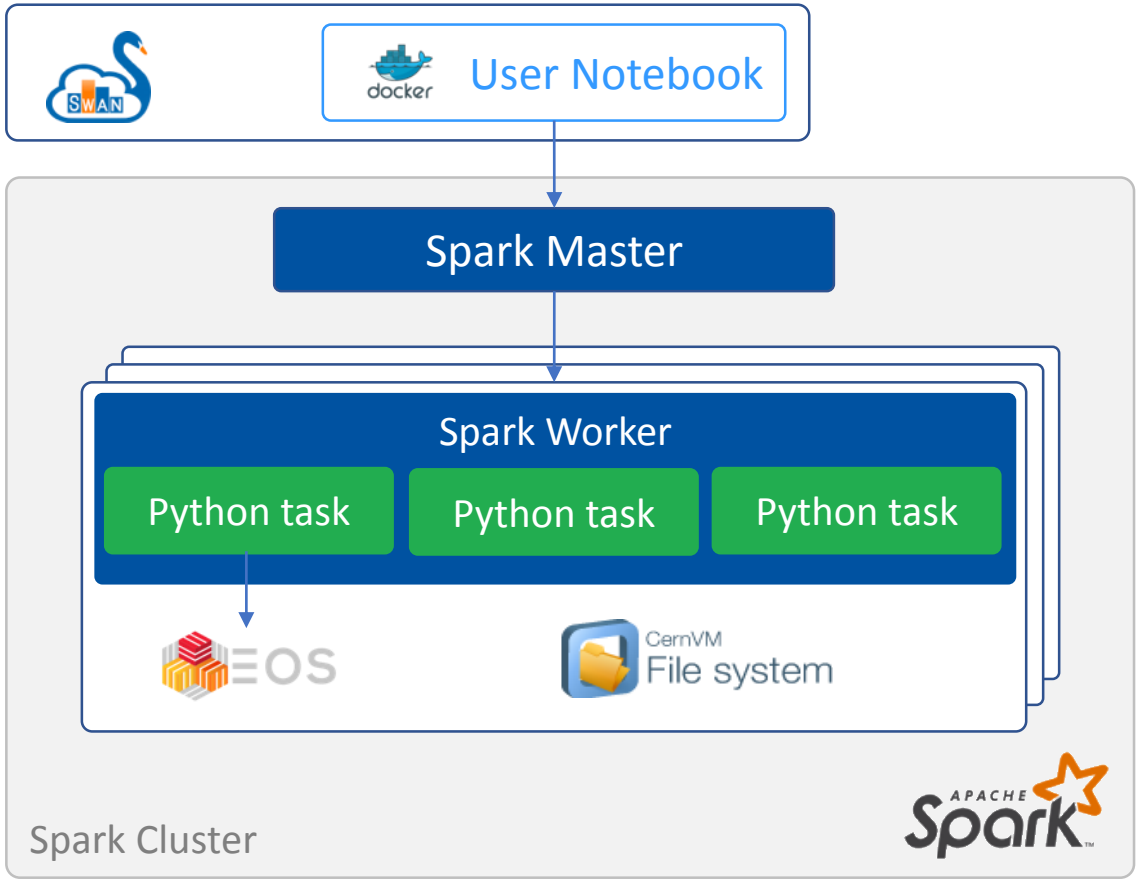
Sharing made easy

- Sharing from inside SWAN interface
  - Integration with CERNBox
  - List shares from other users
- Users can share “Projects”
  - Special kind of folder that contains notebooks and other files, like input data
  - Self contained
- Concurrent editing not supported *yet* by Jupyter
  - Safer to clone
  - Will be available with Jupyterlab

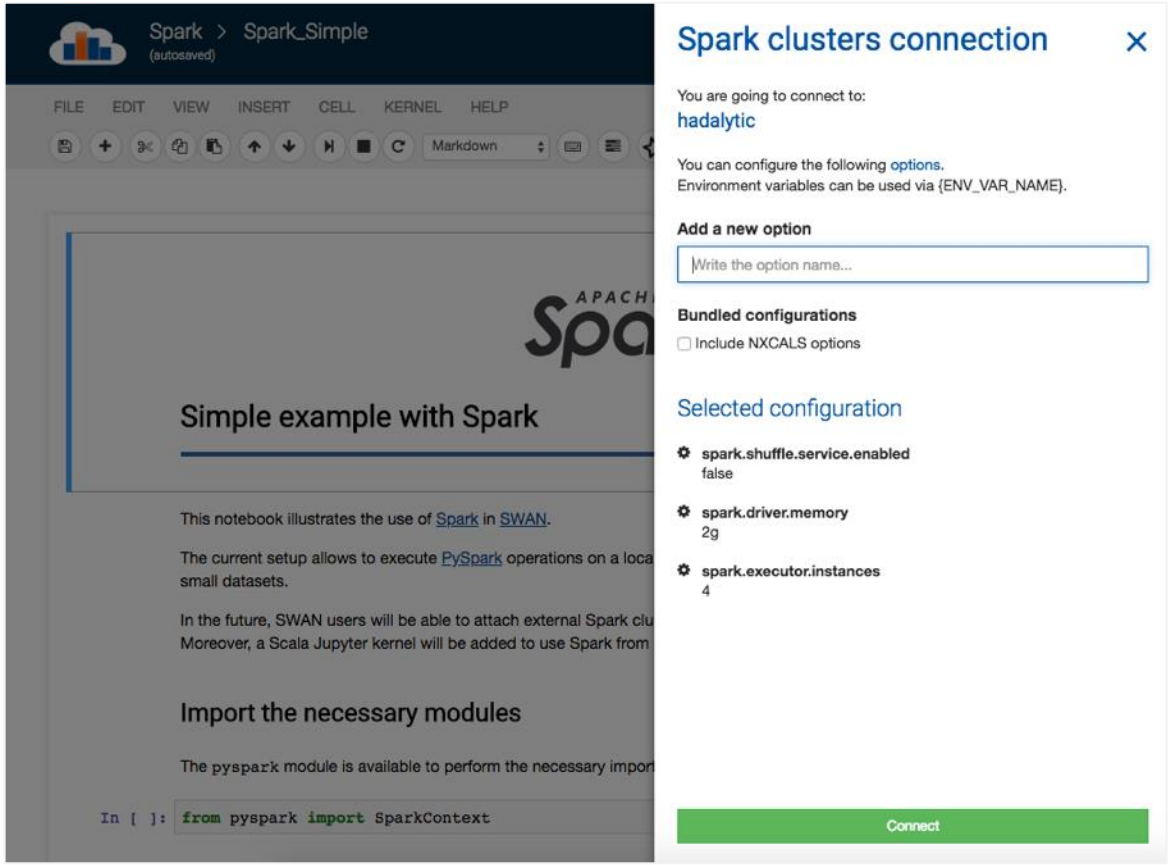


Integration with Spark

- Connection to CERN Spark Clusters
- Same environment across platforms
  - User data – EOS (xrootd)
  - Software – CVMFS
- Graphical Jupyter extensions developed
  - Spark Connector
  - Spark Monitor



Spark Connector/Monitor



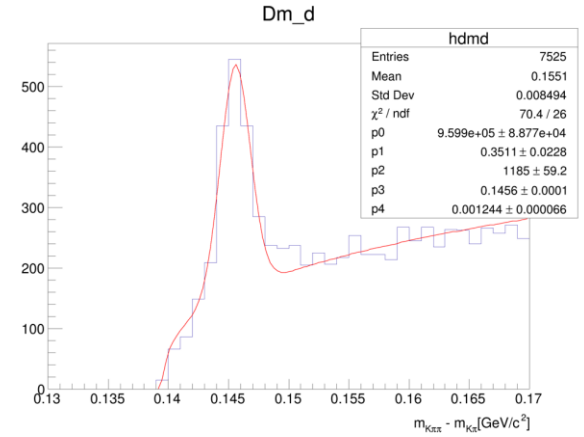
Google Summer of Code

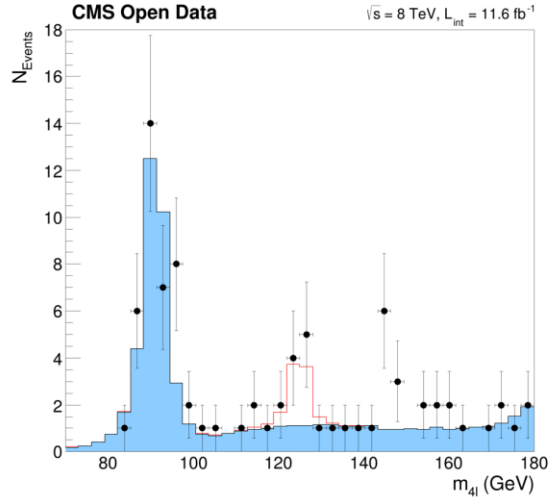
```
In [ ]: def f(x):
            global a
            a+=x
        RDD9.foreach(f)
        RDD9.foreach(f)
        print(a.value)
        #Display should appear automatically
```

Use cases

Use cases

- Final steps of analysis
  - NanoAOD processing with RDataFrame
  - Reconstruction of Higgs boson decaying to two Z bosons from events with four leptons
- Exploration
  - H1 analysis with RDataFrame
- Teaching
  - CERN Summer student courses: ~150 students, data analysis with ROOT
  - CERN School of Computing exercises: ~70 students, parallelism
  - CERN ATLAS PhD student courses: ~50 students, declarative data analysis
  - Machine learning tutorials at CERN



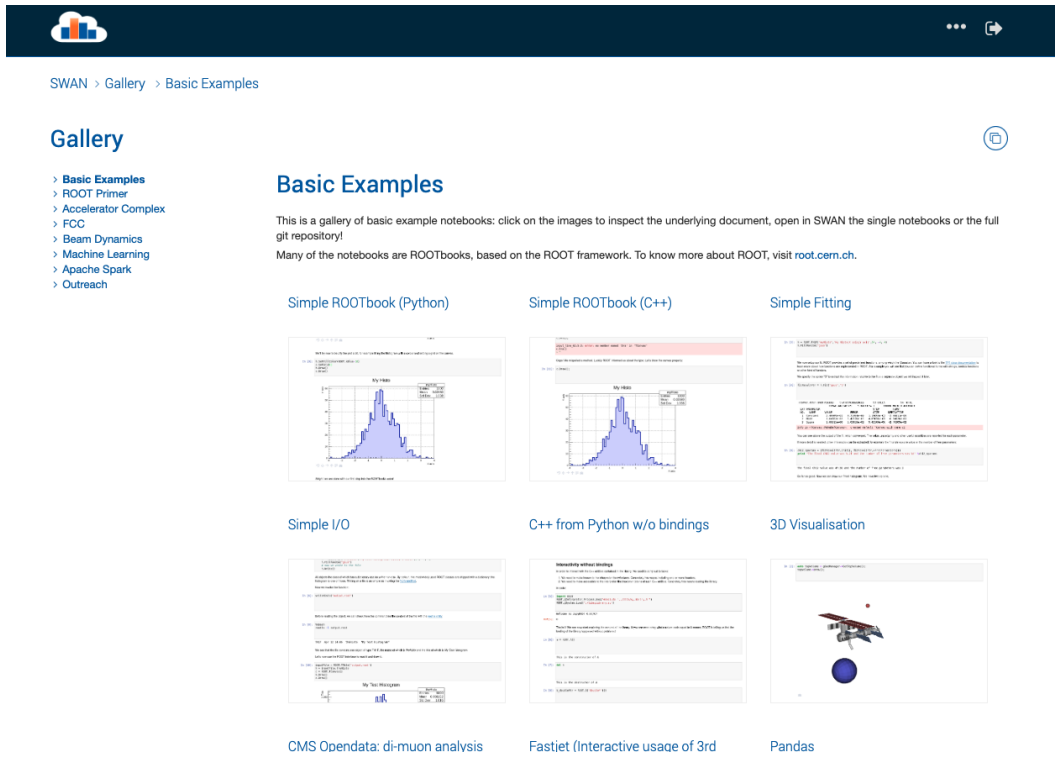


https://root.cern/doc/master/group__tutorial__dataframe.html

User community

- SWAN development is guided by our user community
  - New features (libs, kernels, ...) are requested by users from their real usage needs
- Gallery of examples
  - Made in collaboration with our users
  - More than 50 notebooks in 8 categories

Example notebooks at **cern.ch/swan** or inside the service



Access with only a click

Science Box: SWAN on Premises

- Packaged deployment of SWAN
  - Includes all SWAN components: CERNBox/EOS, CVMFS, JupyterHub
  - Deployable through Kubernetes or docker-compose
- Some successful external/community installations
  - CERN UP2U Pilot
    - Used by students to learn physics and other sciences
  - PSNC (UP2U EU project)
    - SWAN and CERNBox for students and teachers
  - Open Telekom Cloud (Helix Nebula)
    - TOTEM analysis
  - AARNet
    - Australia's Academic and Research Network

UP2U

Physiscope UNIGE

HELIX NEBULA THESCIENCECLOUD

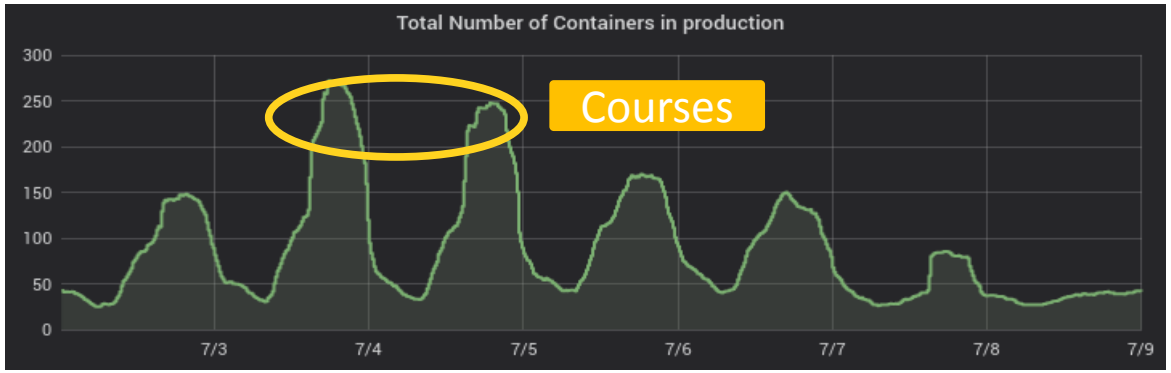
aarnet Australia's Academic and Research Network

CERN

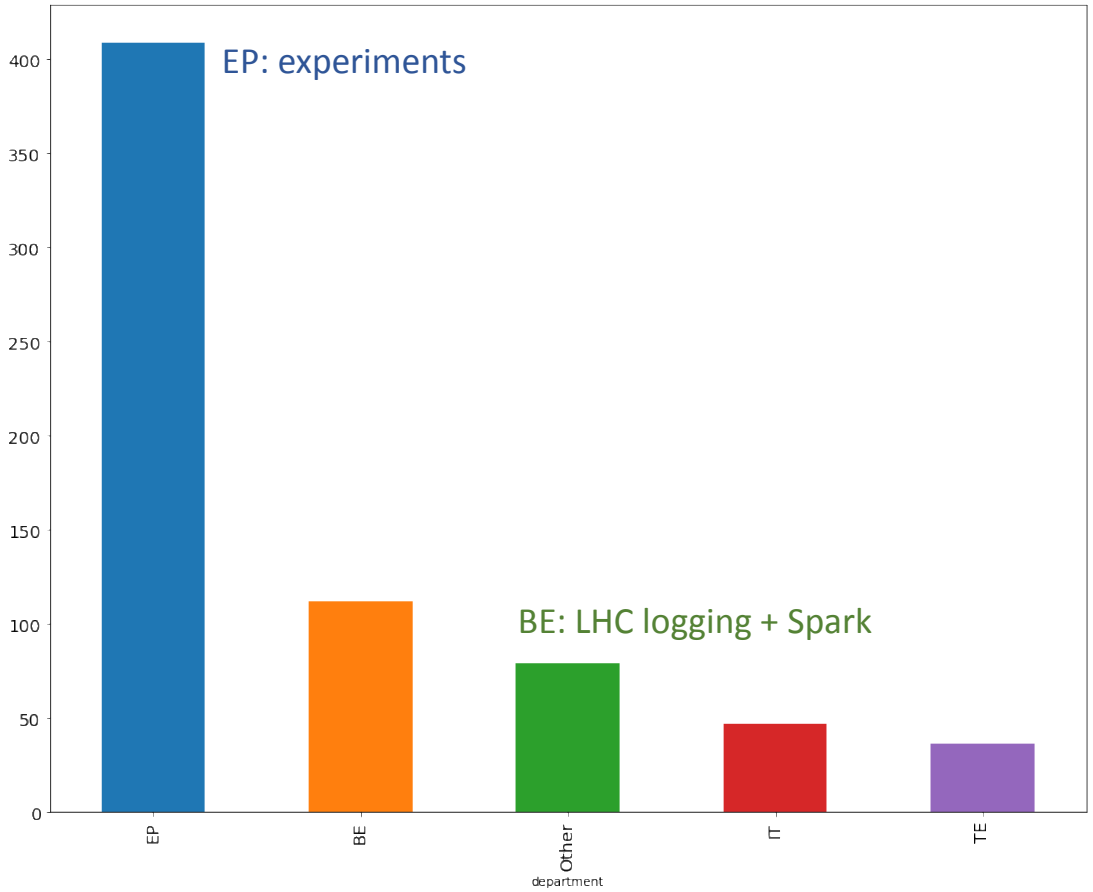
Usage numbers

Usage numbers

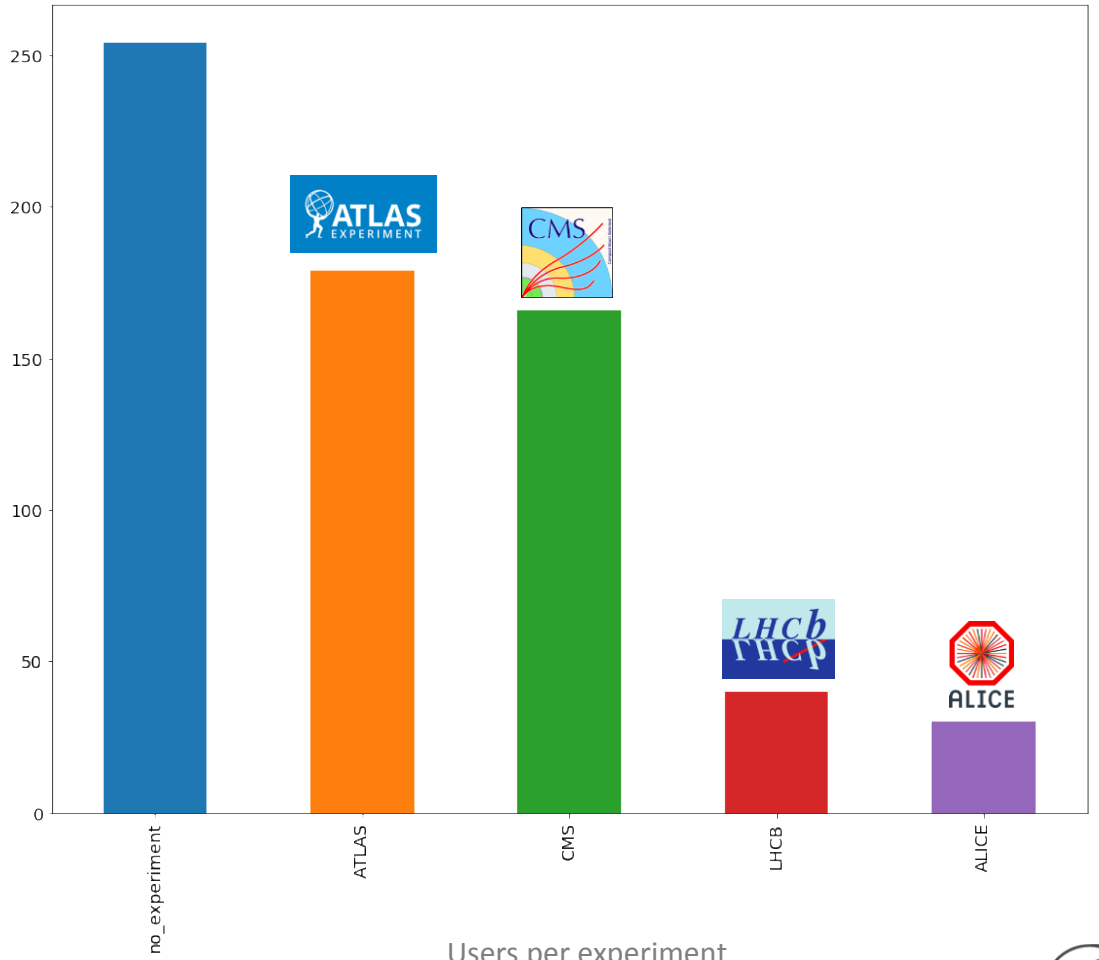
- **~200 user sessions** a day on average
  - Users doubled last year with new SWAN interface
- **~1700 unique users** in 7 months
  - 728 in March
- Spark cluster connection: 15 – 20 % of users
  - SWAN as entry point for accessing computational resources
  - Used for monitoring LHC accelerator hardware devices (NXCals)



March (top 5)

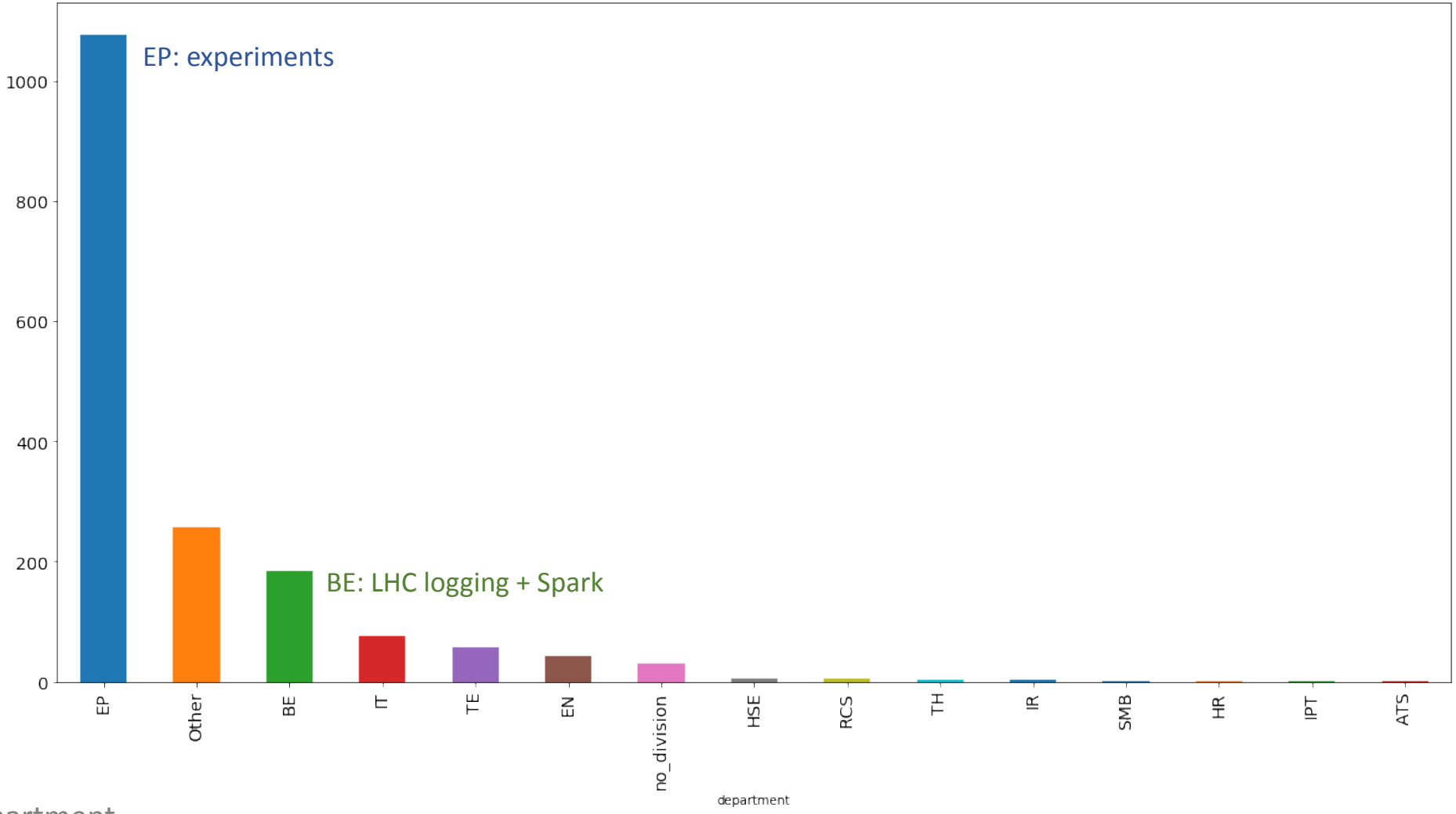


Users per department



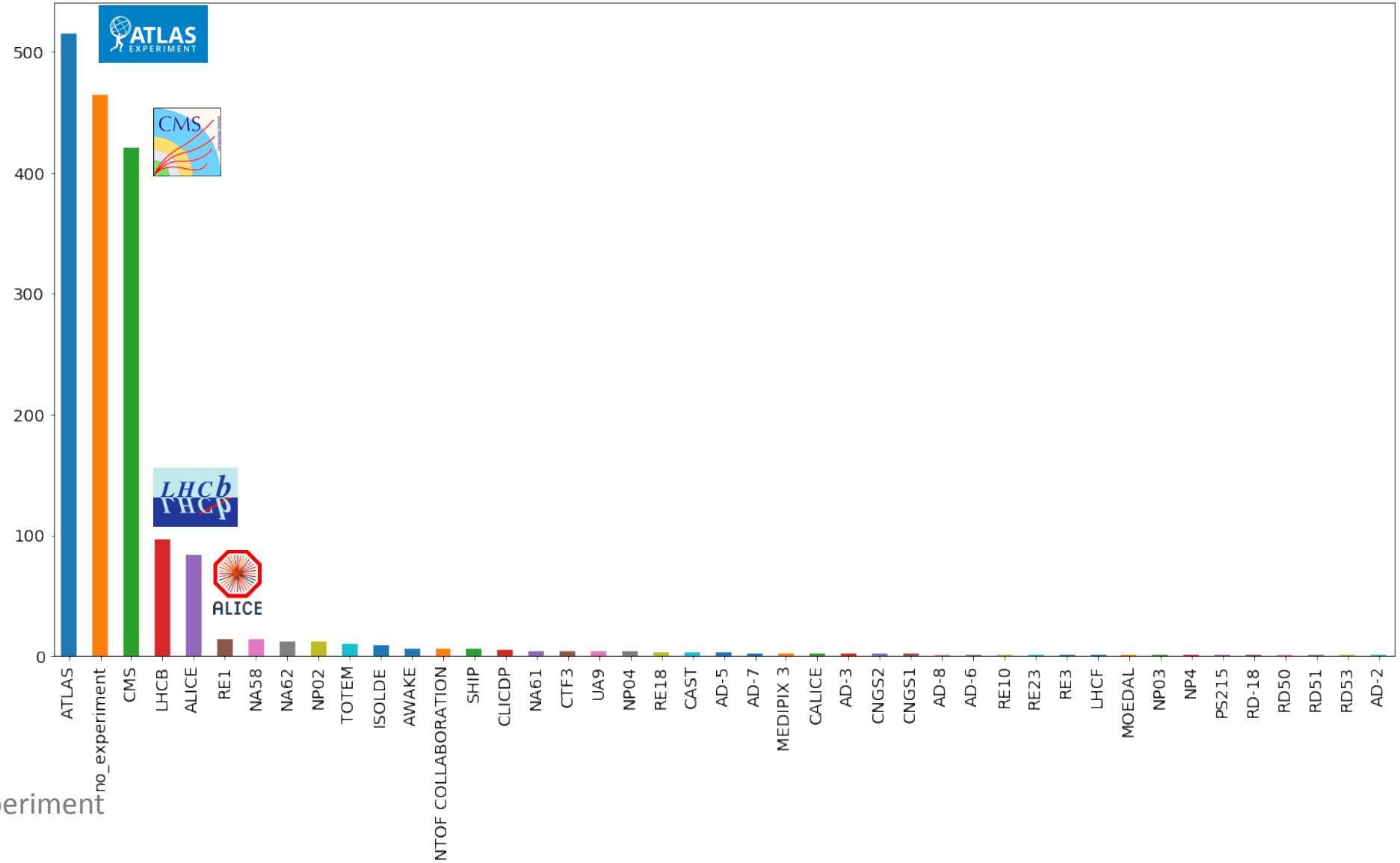
Users per experiment

7 months period



Users per department

7 months period



Users per experiment

Looking ahead

Future work/challenges

- Move to Jupyterlab
  - Porting the current extensions
  - Concurrent editing
- New architecture
  - Based on Kubernetes
- Exploitation of GPUs
  - Speed up computation of GPU-ready libraries (e.g. TensorFlow)
  - Interactive access to GPU resources (complementary to the batch access to GPU)

- Ongoing effort:
  - Submit batch jobs from the notebook

Conclusion

Conclusions

- SWAN is a CERN service that provides Jupyter Notebooks on demand
  - Promotes a cloud-based analysis model
  - Federates CERN services for software, storage and infrastructure
  - Deployable on premises
- SWAN fosters collaboration and results sharing between scientists
- The new Jupyterlab interface will bring new possibilities for collaborative analysis
  - With the introduction of concurrent editing of notebooks
- SWAN became a fundamental Interface for Mass Processing Resources (Spark)
  - Not only for Physics analysis but also for monitoring the LHC hardware
- Usage is growing
  - Missing dedicated manpower to ensure its sustainability

SWAN: service for web-based analysis

Thank you

Diogo Castro
diogo.castro@cern.ch