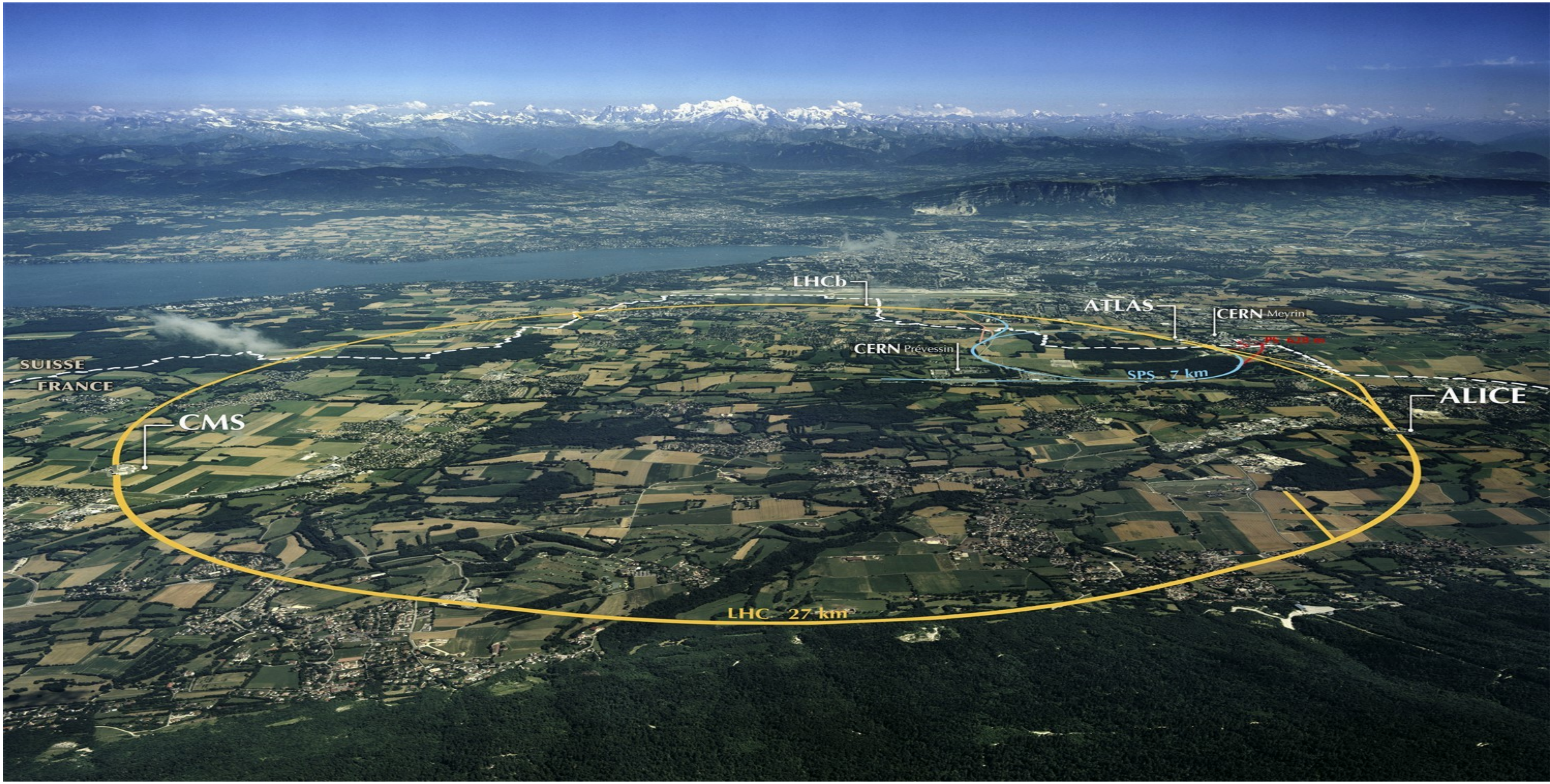
CERN

Taming Local Users and Remote Clouds with HTCondor at CERN



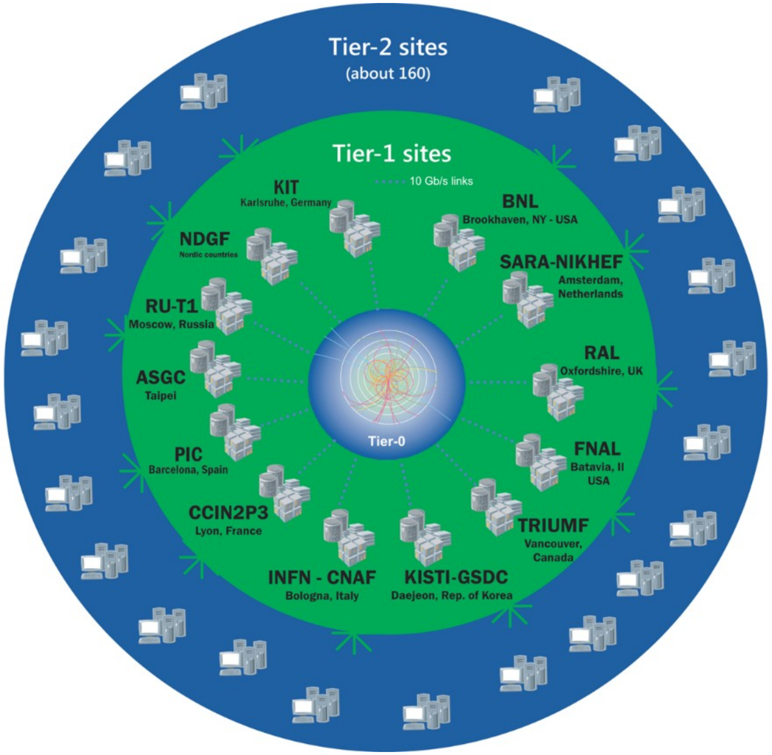
Ben Jones -- ben.dylan.jones@cern.ch

Worldwide LHC Computing Grid

TIER-0 (CERN):
data recording, reconstruction and distribution

TIER-1:
permanent storage, re-processing, analysis

TIER-2:
Simulation, end-user analysis



nearly 170 sites, 40 countries

~350’000 cores

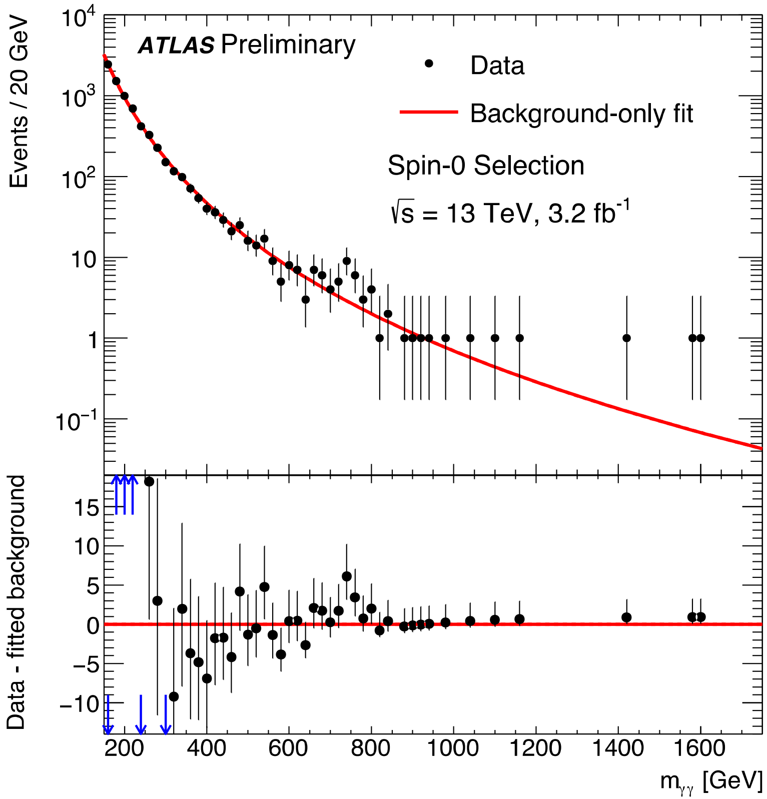
500 PB of storage

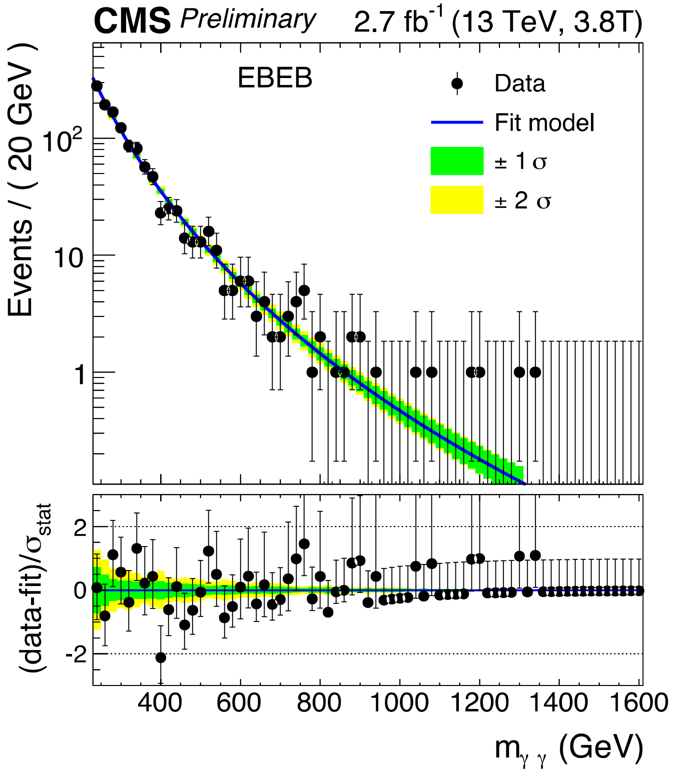
> 2 million jobs/day

10-100 Gb links

Run 2 has only just started

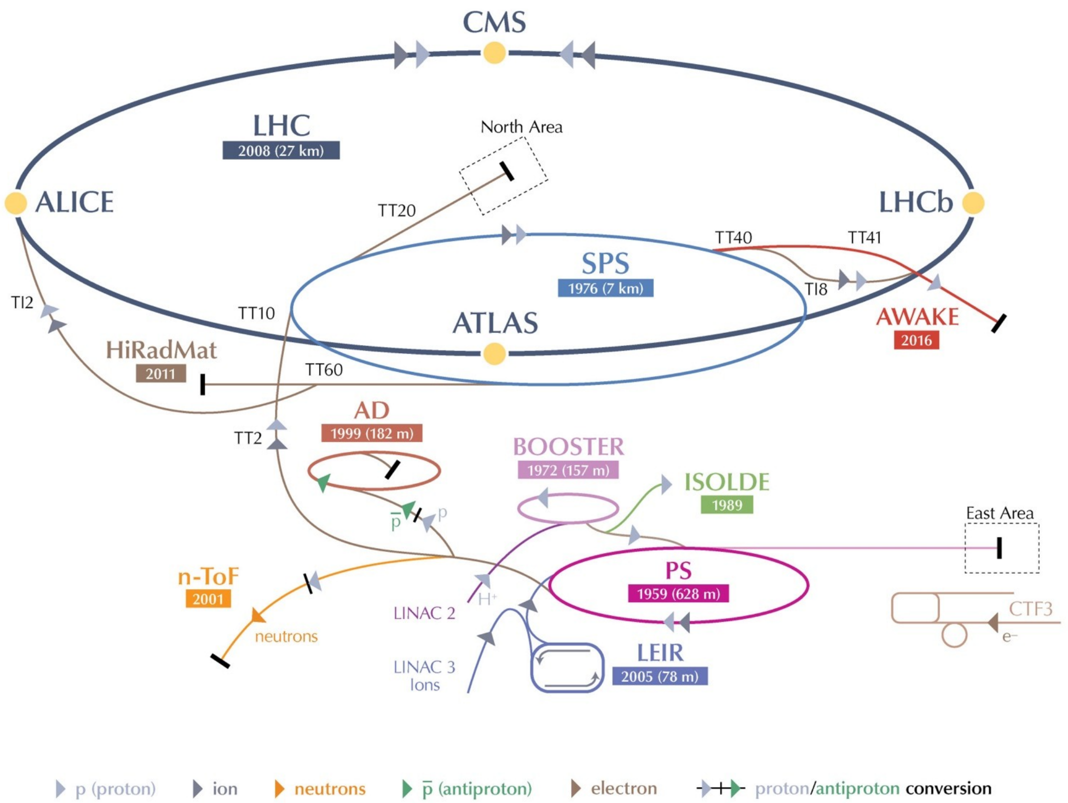
- Hint of an excess with diphoton mass of 750 GeV
  - Seen by ATLAS and CMS – coincidence or a new signal?





393
AD
ANTIMATTER
FACTORY
ELENA

CERN's Accelerator Complex



LHC Large Hadron Collider **SPS** Super Proton Synchrotron **PS** Proton Synchrotron

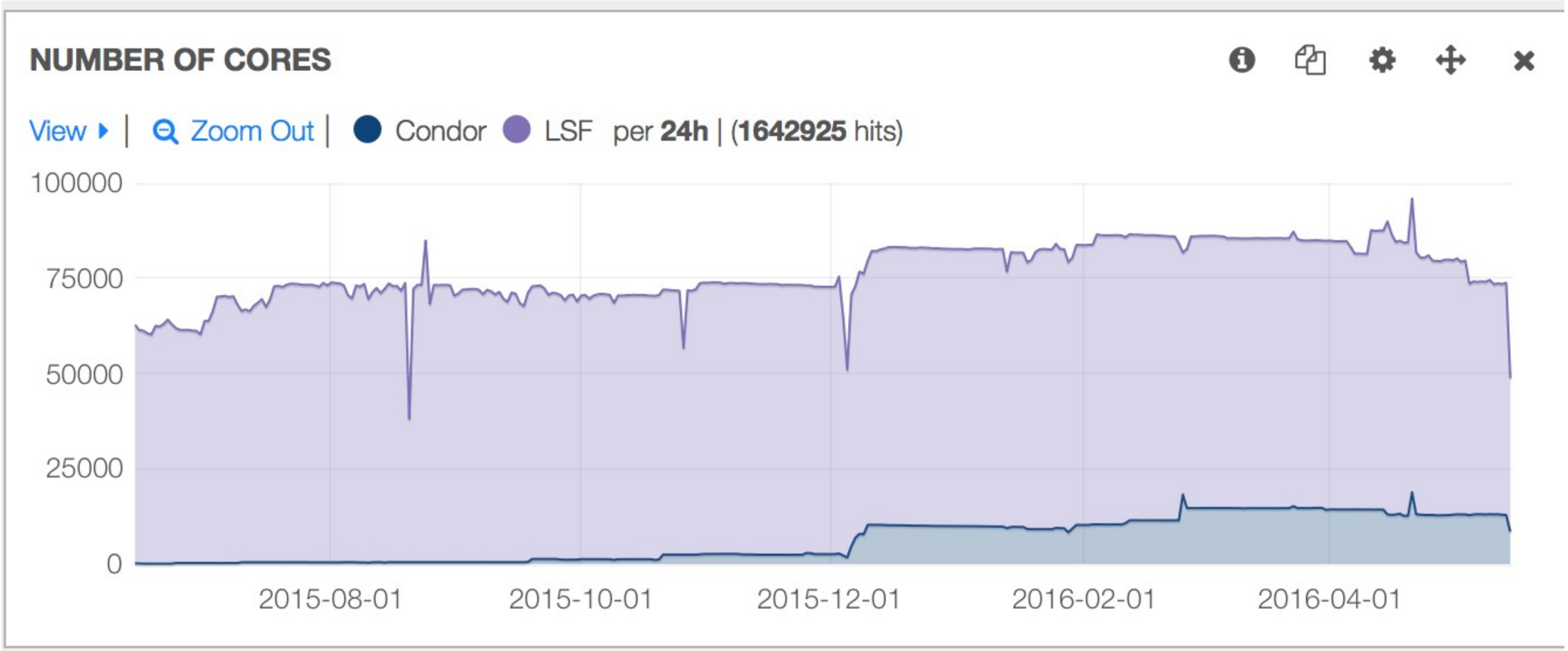
AD Antiproton Decelerator **CTF3** Clic Test Facility **AWAKE** Advanced WAKefield Experiment **ISOLDE** Isotope Separator OnLine DEvice

LEIR Low Energy Ion Ring **LINAC** LINear ACcelerator **n-ToF** Neutrons Time Of Flight **HiRadMat** High-Radiation to Materials

Batch Service at CERN

- Service used for both grid and local submissions (roughly equal %)
- Local public share open to all CERN users
- 400-600K job submissions / day
- ~60K Running jobs
- Migration to HTCondor from LSF underway
  - Some grid already moved, local imminent
  - Vanilla, but planning on Docker

LSF v HTCondor

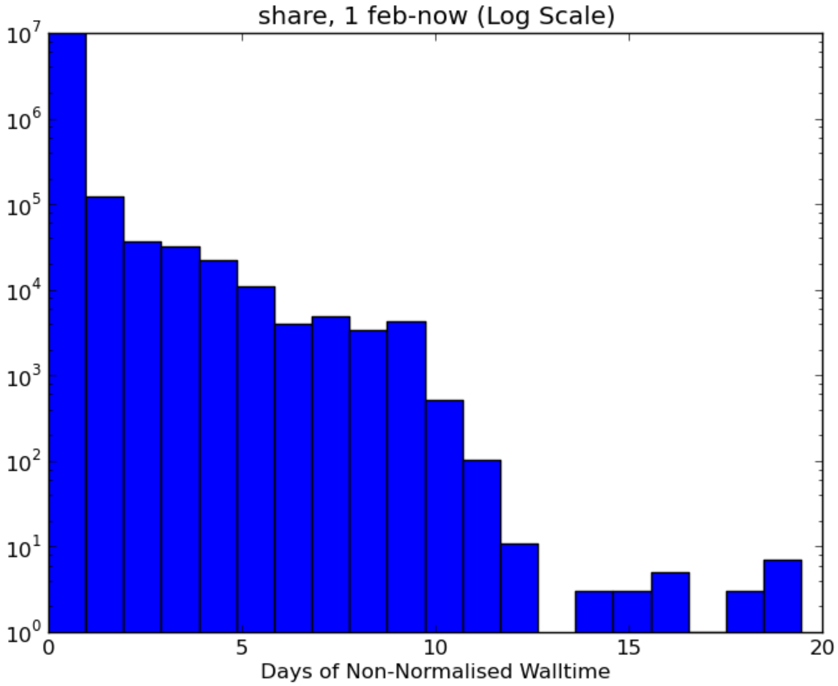


Requirements

- Users need:
  - Auth tokens for internal services (yes, including AFS $HOME)
  - A schedd to use, and to query
  - An accounting group to be charged to
  - Some nice defaults for things like job timing
- We need:
  - Defined units of compute: 1 core means 2gb RAM, 20GB scratch

Token expiry

- Standard expiry time is 1 day, renewable to 7 days
- Our jobs can last for > 7 days
- Mechanism required to keep authentication tokens refreshed (and in our case, beyond standard)



Integration with Kerberos

- Controlled by config vars:
  - SEC_CREDENTIAL_PRODUCER
  - SEC_CREDENTIAL_MONITOR
  - SEC_CREDENTIAL_DIRECTORY
- Ability to add scripts to generate initial kerberos tickets, and to renew as necessary
- Condor will monitor, copy to sandbox and refresh as necessary

Submit Side

- condor_submit calls SEC_CREDENTIAL_PRODUCER which acquires a kerberos AS-REQ
- The AP-REQ is handed to the schedd’s condor_credd by condor_submit, and is written to the schedd’s SEC_CREDENTIAL_DIRECTORY
- The SEC_CREDENTIAL_MONITOR monitors the directory, and acquires/renews TGT

Execute Side

- condor_starter copies credentials into the SEC_CREDENTIAL_DIRECTORY when a user has jobs scheduled to execute, and removes when there are no jobs left
- The SEC_CREDENTIAL_MONITOR will acquire/renew TGTs with the stored AP-REQ
- The TGT will be copied to the job sandbox when the job runs, with KRB5_CCNAME set

Schedds

- Number of user jobs means we need multiple schedds
- Want to make it easy & cheap to query, so needs to be static assignment
- Currently using zookeeper as the k/v store
- znode contains schedd
  - /htcondor/users/$username/{current,old}
- LOCAL_CONFIG_FILE piped script to contact zk on submit / query for schedd

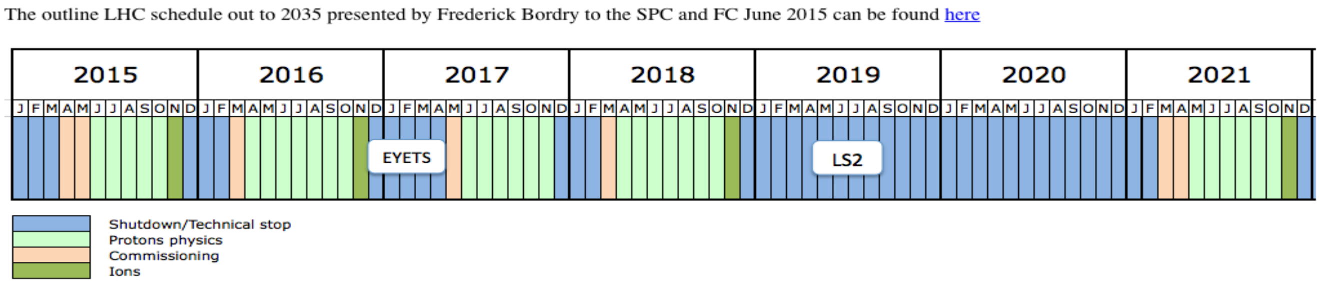
Jobflavours

- Current LSF service has defined queues for local
  - Defined as “normalised” minutes/hours/weeks
  - Slot limits for individual queues (ie limit long queues)
- Use mixture of startd and schedd policy to enforce
- Default to short time to ensure users think about it, restrict number of long slots

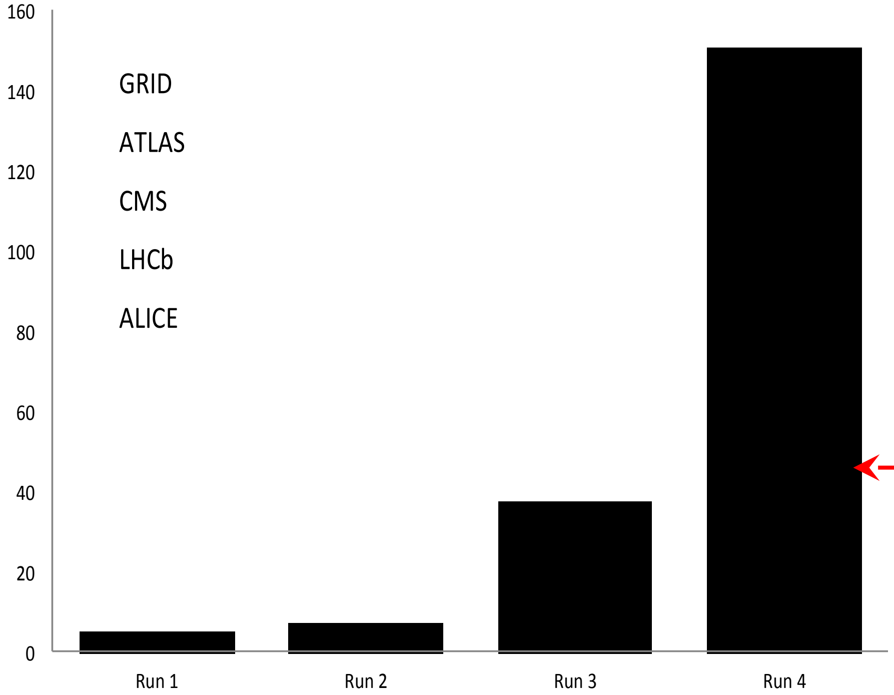
Accounting groups

- Assigning users to accounting groups and delegating control to VO admin
  - (hopefully using BNL’s solution)
- Use SUBMIT_REQUIREMENTS and PROTECTED attribute to make Account group required and immutable
- Also use SUBMIT_REQUIREMENTS to ensure users don’t request > 2gb / core

Compute Growth Outlook



The outline LHC schedule out to 2035 presented by Frederick Bordry to the SPC and FC June 2015 can be found here



Compute: Growth > x50
Moore’s law only x16

What we can afford

… and 400PB/year in Run 4

Cloudy with a chance of jobs…

- “There is no cloud: it’s just someone elses computer”
- Current public cloud investigations around adding flat capacity (cheaper), not elasticity
- Treat additions to htcondor pool as just more computers, somewhere else
- We maybe trust them slightly less

HTCondor cloud details

- Most of plant mapped to worker-node, cloudy nodes mapped to xcloud-worker
  - Differences in how we add in mapfile
- We use a less trusted CA (no perms beyond joining pool)
- For most VOs, common submission point using routes
  - Jobs: +remote_queue="externalcloud"
  - Machines: XBatch =?= True

Questions

ergency Procedure
Overpressure in Oxford magnet
XHE-2 and/or OXHE-1 releasing helium
ecovery (on recovery line)
magnetic field (if on)
Nitrogen level - refill if needed
OID CONTACT
ryogenic fluids
STAND at
harge point.
arlsberg
OXFORD
STRUMENTS