Experience and Plans for use of commercial clouds at US Tier-1s

Dirk Hufnagel (FNAL)
for BNL/ATLAS and FNAL/CMS

Many thanks to John Hover (BNL) and Burt Holzman (FNAL), who provided most of the material in this talk.

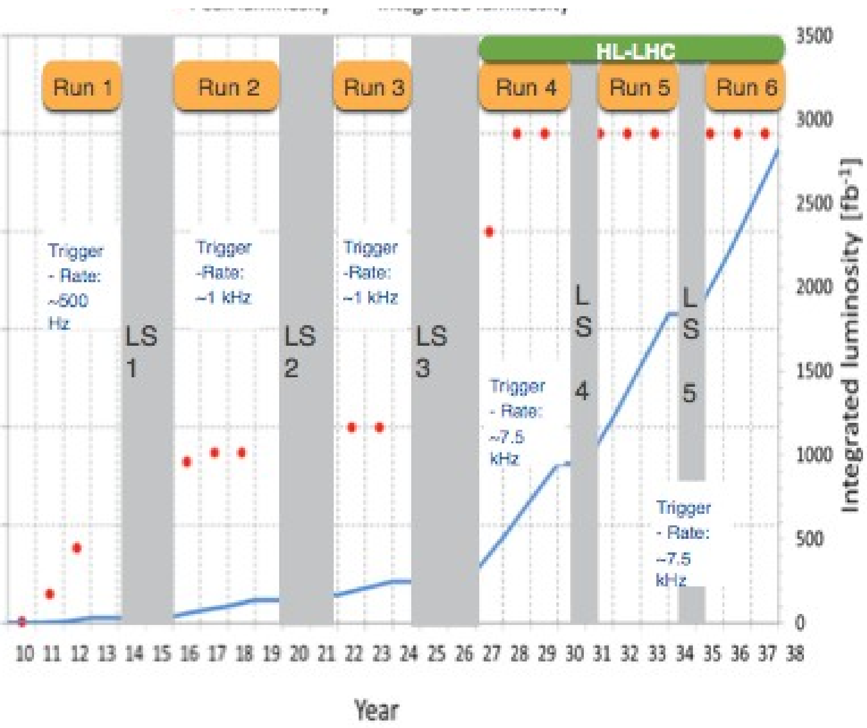
WLCG, 09.10.2016

Overview

- Motivation
- Why site extension and not standalone ?
- Experiences of BNL/ATLAS and FNAL/CMS with AWS
- Conclusion / Future

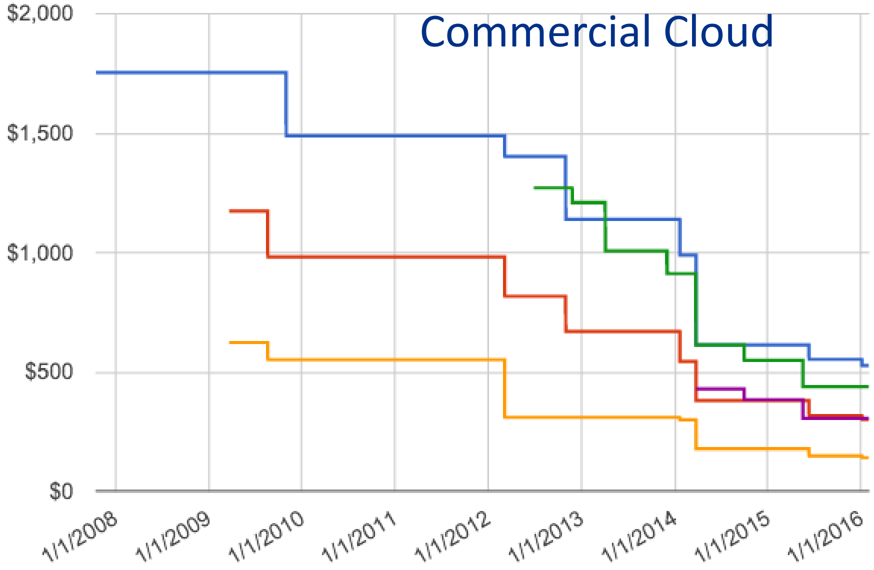
Motivation : Capacity and Cost

- HEP Computing needs will be 10x to 100x current capacity



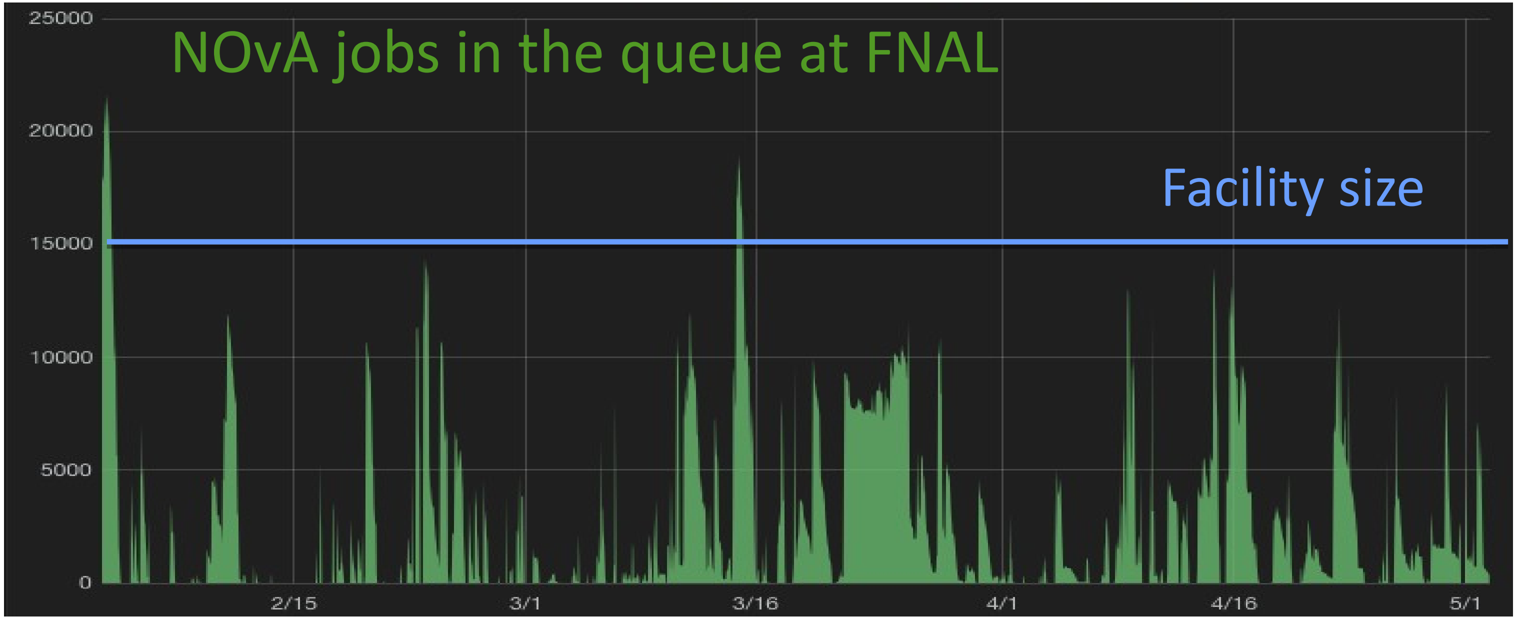
- Commercial clouds offering increased **value** for decreased **cost** compared to the past

Price of one core-year on Commercial Cloud



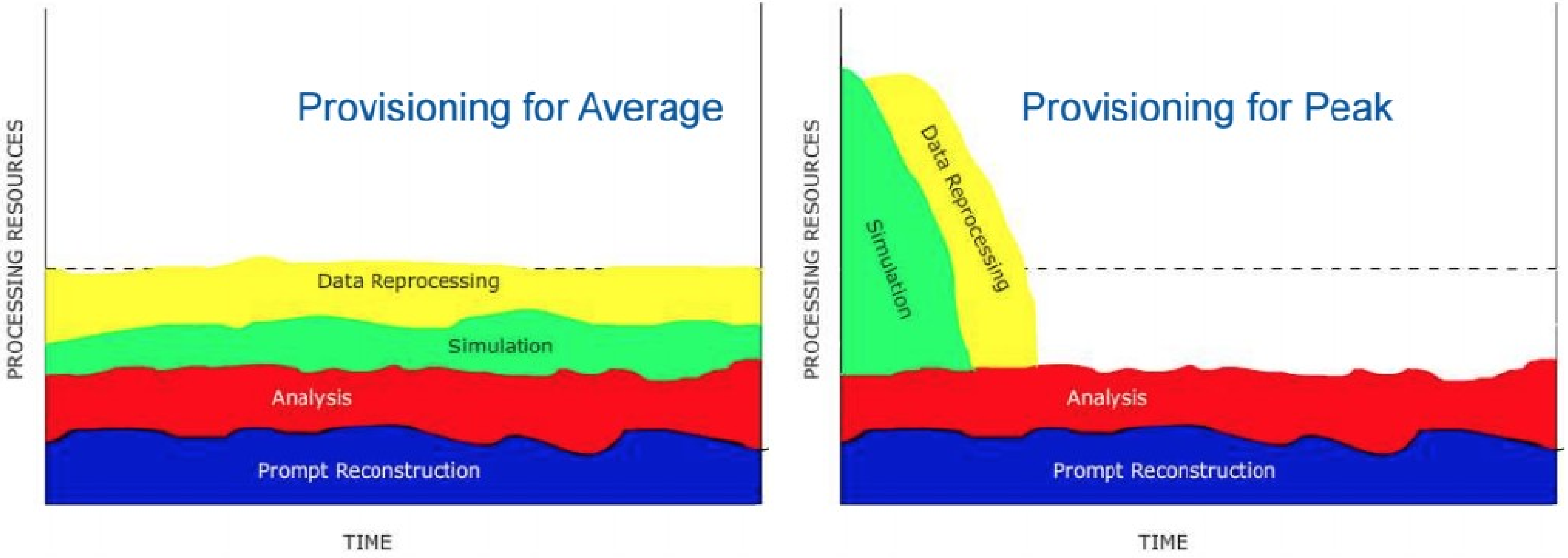
Motivation : Elasticity

- Computing schedules driven by real-world considerations (detector, accelerator, conferences …) but also ingenuity:
  - – this is research and development of cutting-edge science
- Resource use is not steady-state and worse, not always predictable



Motivation : Elasticity

- Provision base capacity at site and keep overflow capacity in cloud that can be rented as needed to deal with demand spikes (paradigm shift for our users: exposure to cost of computing)



Why site extension ?

- HEP hardware funding is not to experiments, but to facilities (that pledge resources to experiments).
- Need 'glue' to fit external cloud resources into the existing HEP computing infrastructure (longterm storage, interface to experiments workflow management etc).
- Both imply that a site extension scheme works better than treatment as standalone resources (the larger the cloud procurement the larger the 'host' site needs to be).

=> HEP Cloud scheme for FNAL/BNL (more later)

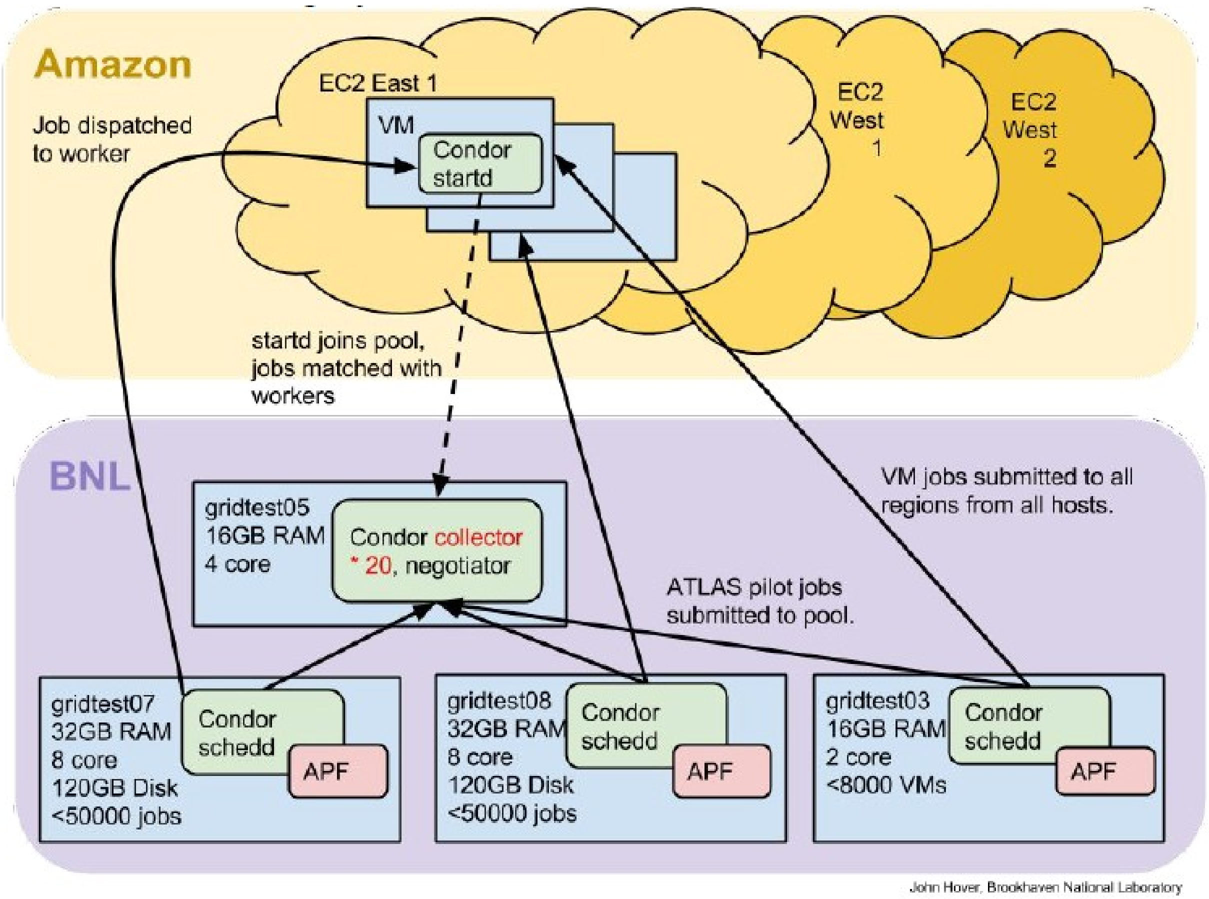
ESnet/AWS peering

- Big thanks to ESnet and Brookhaven for establishing the 100Gbit peering with AWS
- Without that the FNAL/CMS test would not have been possible at the scale we were running it
- If we want to use other cloud providers at the same scales, we need similar peering agreements (AFAIK there are ongoing efforts in this direction)

BNL/ATLAS AWS September 2015

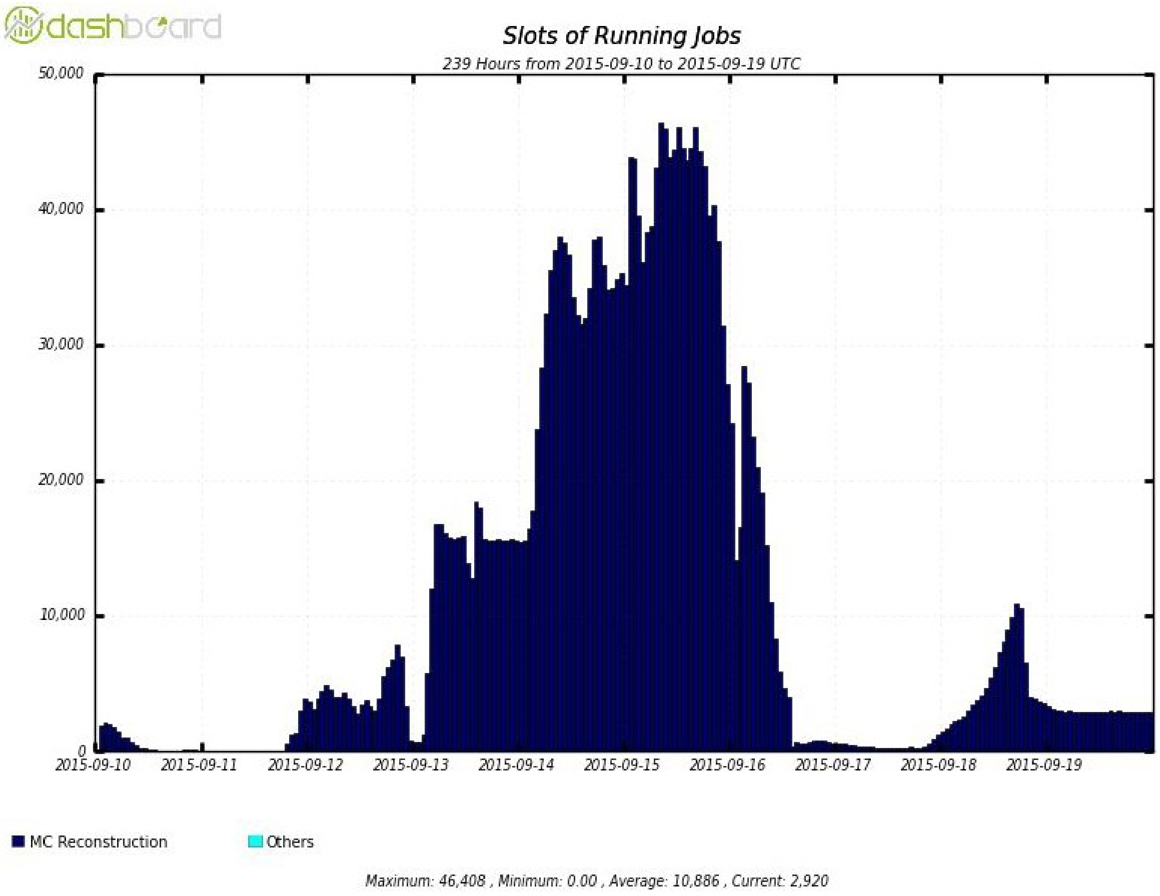
- ~45k cores
- US East region only (networking not ready on west coast)
- 10Gb dedicated bandwidth between AWS and BNL
- Input data automatically pre-subscribed (copied to S3)
- Output pulled from S3 asynchronously after the run
- ~6000 jobs (8-core multicore)
- ~4000 simultaneous VMs: mix of 8-,16-, and 32-core types (8 >> 16 > 32)
- Ran ~5 days. 437,000 jobs completed
- 3.2 million CPUhrs
- Compute cost approx $57K, Data+storage around $500

BNL/ATLAS AWS September 2015



BNL/ATLAS AWS September 2015

Cores Sept 10-19, 2015

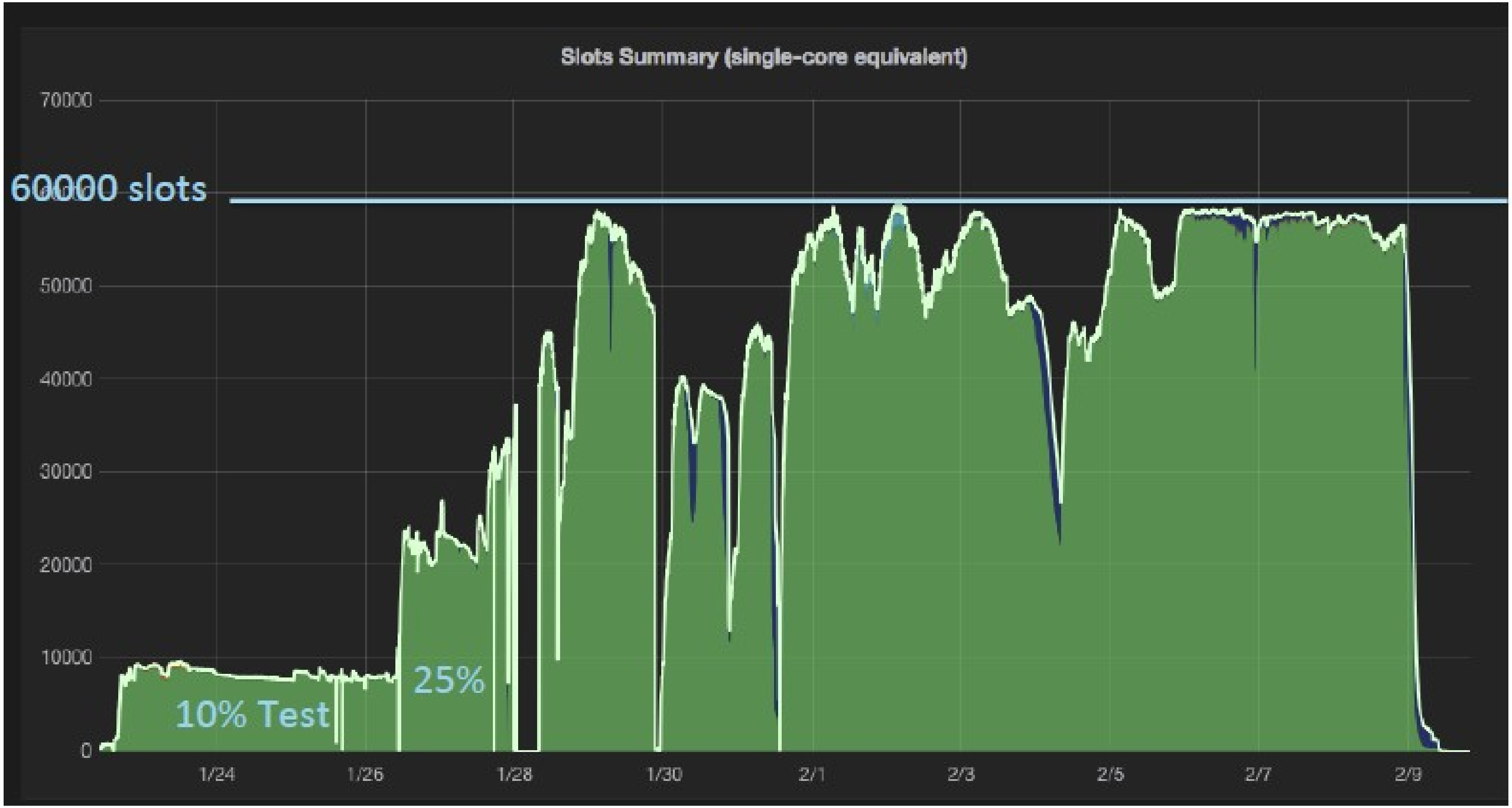


FNAL/CMS AWS January/February 2016

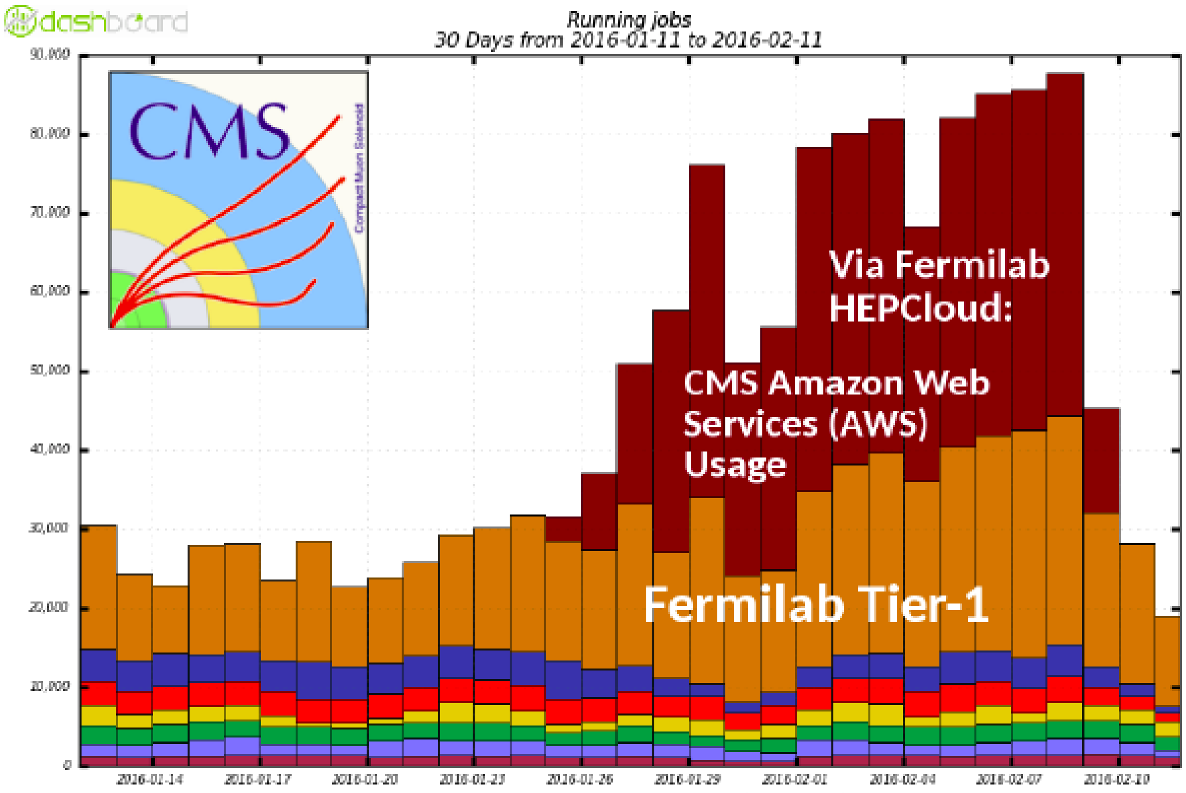
- ~60k cores
- US East 1, US West 1 and US West 2 regions
- Pileup samples pre-staged to S3
- Output staged directly back to FNAL from jobs
  (knew we needed more than 10Gbit,
  so the 100Gbit peering was essential)
- 2.9 million jobs, 15.1 million wall hours
- 9.5% badput – includes preemption from spot pricing
- 87% CPU efficiency
- 518 million events generated

- Gen+Sim+Digi+Reco all done in the same job, only keep AODSIM and MINIAODSIM, **NOT** the GEN-SIM

FNAL/CMS AWS January/February 2016



FNAL/CMS AWS January/February 2016



Via Fermilab HEPCloud:

CMS Amazon Web Services (AWS) Usage

Fermilab Tier-1

Tier-1 (Italy)
Tier-1 (Germany)
Tier-1 (France)
Tier-1 (Russia)
Tier-1 (UK)
Tier-1 (Spain)

Fermilab/AWS cost comparison

- Average cost per core-hour
  - On-premises resource: **.9** cents per core-hour
    - Includes power, cooling, staff
  - Off-premises at AWS: **1.4** cents per core-hour
    - Ranged up to 3 cents per core-hour at smaller scale
- Benchmarks
  - Specialized (“ttbar”) benchmark focused on HEP workflows
    - On-premises: **0.163** events/second
    - Off-premises: **0.158** events/second

- Raw compute performance roughly equivalent
- Cloud costs larger – but approaching equivalence

Conclusions / Random thoughts

- Price we got on spot market based on economic factors (how much we bid vs. resource utilization). No way to reliable predict how long we could have kept our tests running at the price we were bidding, but nothing to indicate it couldn't have been a long time given sufficient funding and work.

- Caveat: You need to use many regions/zones/flavors. Any given one you can run out of affordable resources, overall there always seems to be capacity somewhere.

Conclusions / Random thoughts

- Longer term outlook gets you into crystal ball gazing
  - how aggressive will AWS keep expanding capacity ?
  - utilization of AWS vs. time ?
  - business model for dealing with idle resources ?

- Nothing to indicate AWS is slowing down expanding capacity.
- Safe to say daily/weekly/yearly utilization will continue to vary.

  But need to keep provisioning for peak use !

- AWS is the biggest player in the commercial cloud market, but there are large competitors (Google, Microsoft). Competition should keep prices low, but also could disrupt/change business practices (in the long run).

Conclusions / Random thoughts

- For right now, there is no indication that we couldn't make very good use of AWS spot market at large (for us) scales.

Future / HEP Cloud

- HEP Cloud is a way to extend the existing Tier 1 facilities to be able to transparently (for the user) execute jobs on external resources like HPC and commercial clouds for instance.
- Joint collaboration between FNAL and BNL. High level goals to provide a single concept/solution have been agreed on, still a lot to be worked out technically though.
- For more there are two HEP Cloud talks during CHEP.

Backup

References

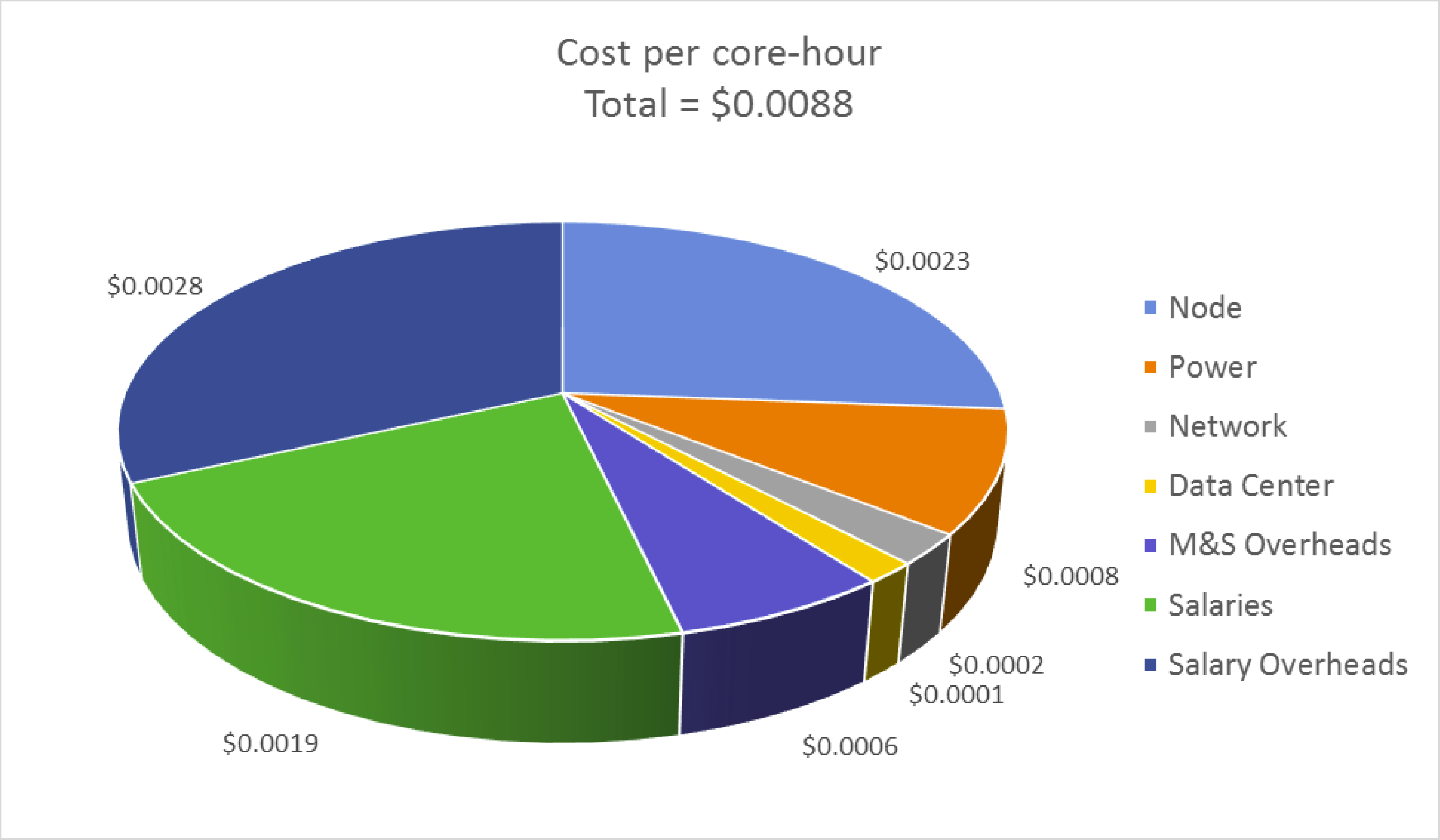
- OSG AHM 2016 : BNL/ATLAS

  https://indico.fnal.gov/contributionDisplay.py?contribId=25&sessionId=5&confId=10571

- ICHEP 2016 : FNAL/CMS

  https://indico.cern.ch/event/432527/contributions/1072465/

FNAL cost estimate


Cost per core-hour
Total = $0.0088
$0.0023
$0.0028
$0.0008
$0.0002
$0.0001
$0.0006
$0.0019
Node
Power
Network
Data Center
M&S Overheads
Salaries
Salary Overheads