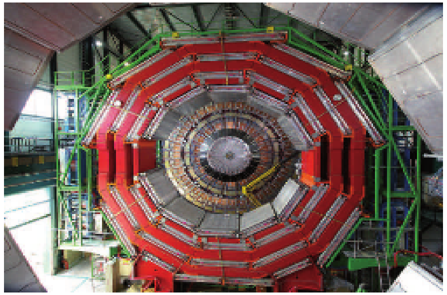
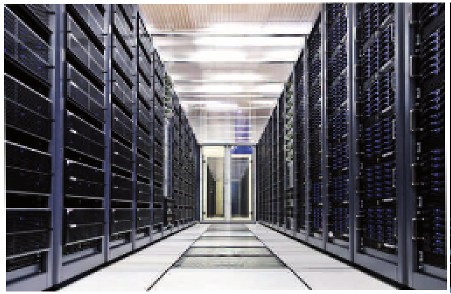
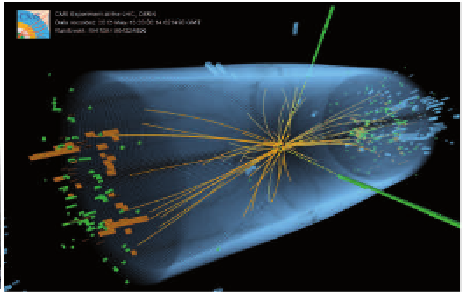
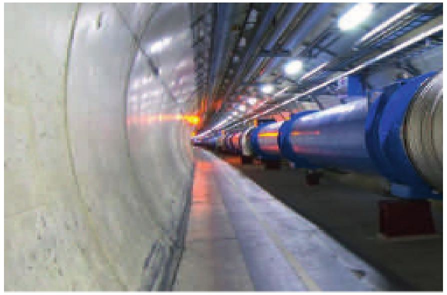
# Managing the CMS Data and Monte Carlo Processing during LHC Run 2

CHEP 2016

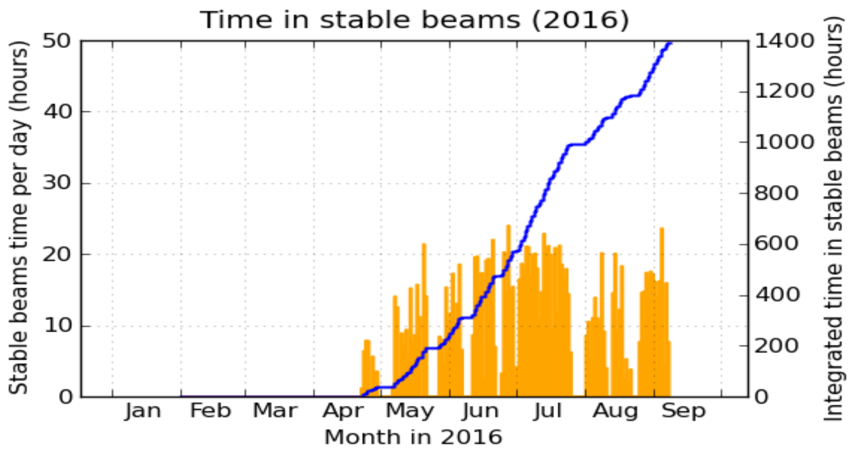
Christoph Wissing (DESY) for the CMS Collaboration
October 2016

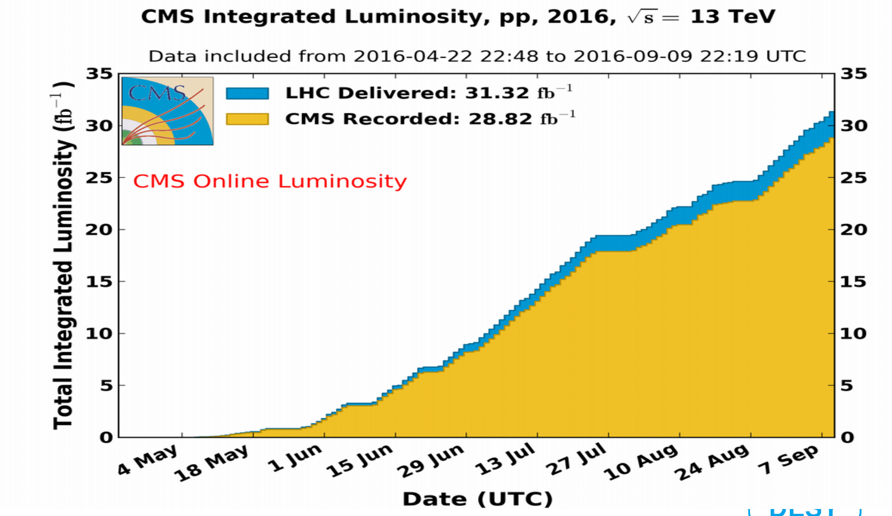
HELMHOLTZ ASSOCIATION

DESY

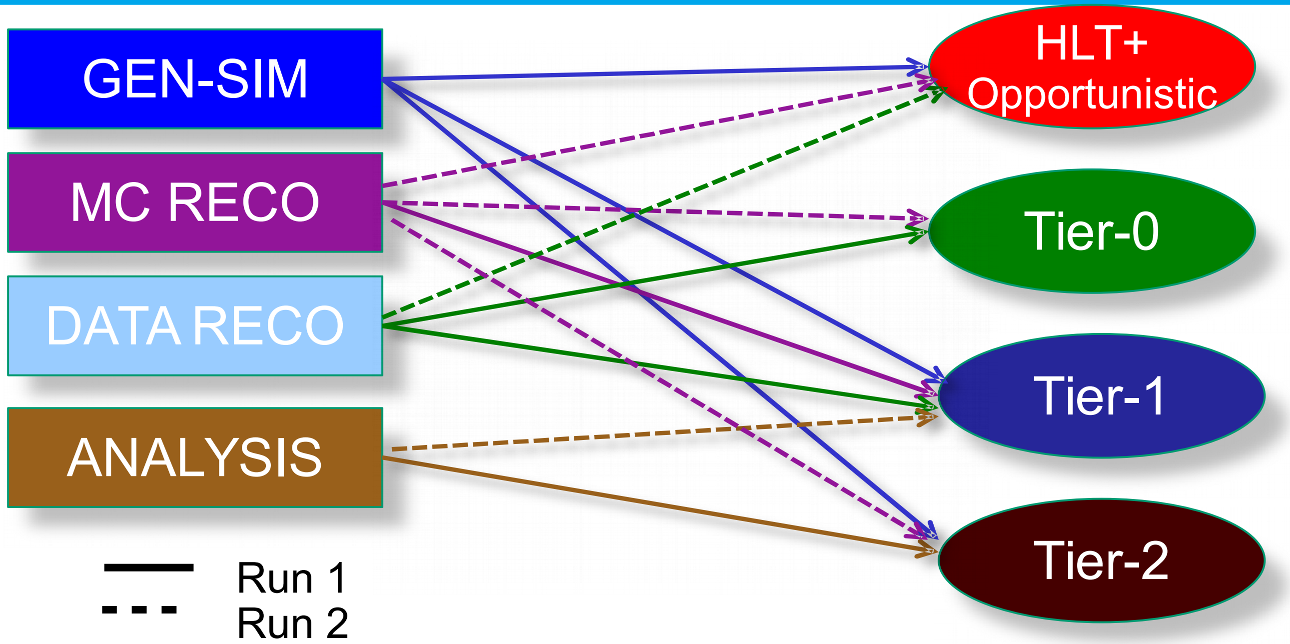
# LHC & CMS Performance

- Excellent performance of the LHC
  - Expected uptime reached in September
- Very demanding data rate taken by CMS detector
  - Data logging rate often beyond the target of 1kHz
- Increased need for corresponding Monte Carlo sets
- Computing becomes resource constrained
- Requires flexible and efficient use of all resources



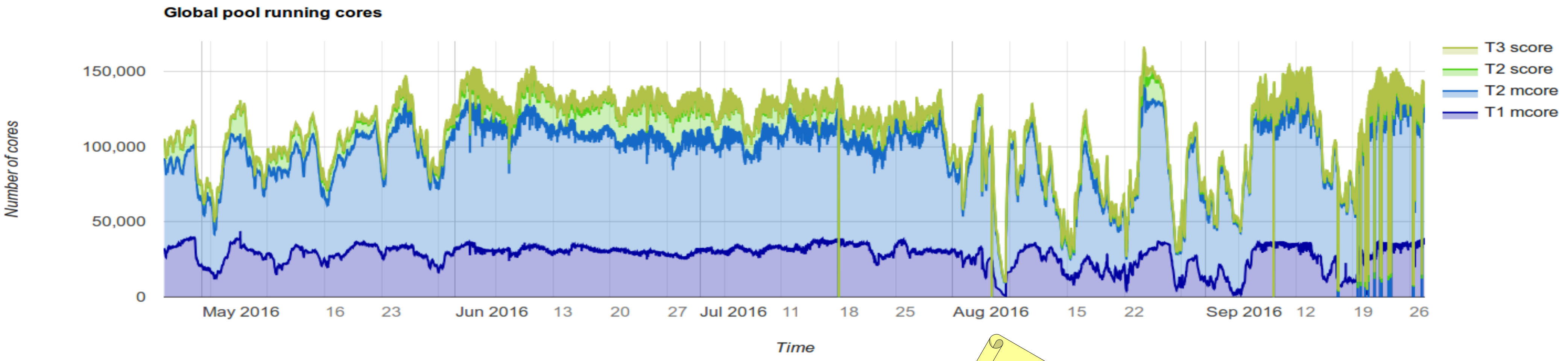



# Decoupling of Workflows and Resource Types




- Rather tight coupling of workflow types to resources in Run1
- Big gain in flexibility for Run 2
  - Almost each workflow can run anywhere
  - All CPU joined to one Global HTCondor pool + dedicated Tier-0 pool
  - (Almost) all Tier-1 & Tier-2 disk managed via Dynamic Data Management (DDM)

# Global Pool



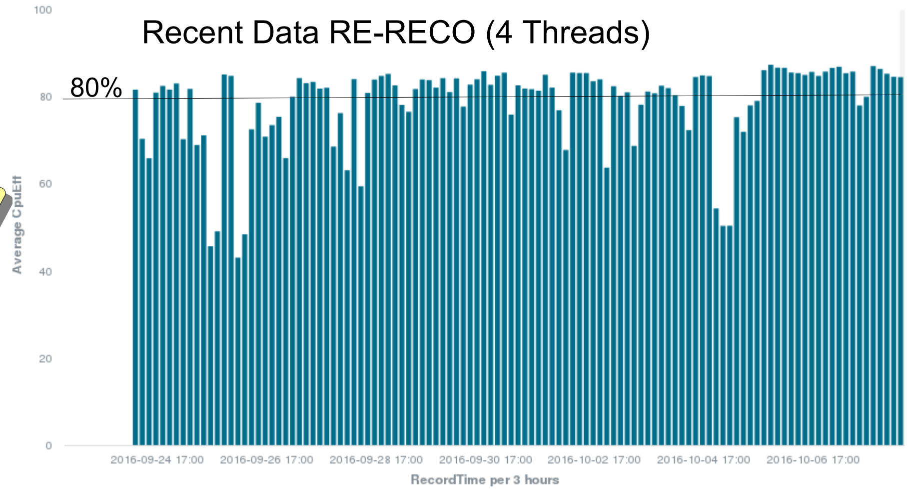

- Stably utilization of ~130.000 cores
- Multi-core pilots sent to ~90% of the resources
  - Campaign to move Tier-2 sites to multi-core pilots in Spring 2016
  - Send exclusively multi-core pilots where possible
  - “Dynamic partitioning”:
    Matching of single vs. multi-threaded application happens inside the pilot

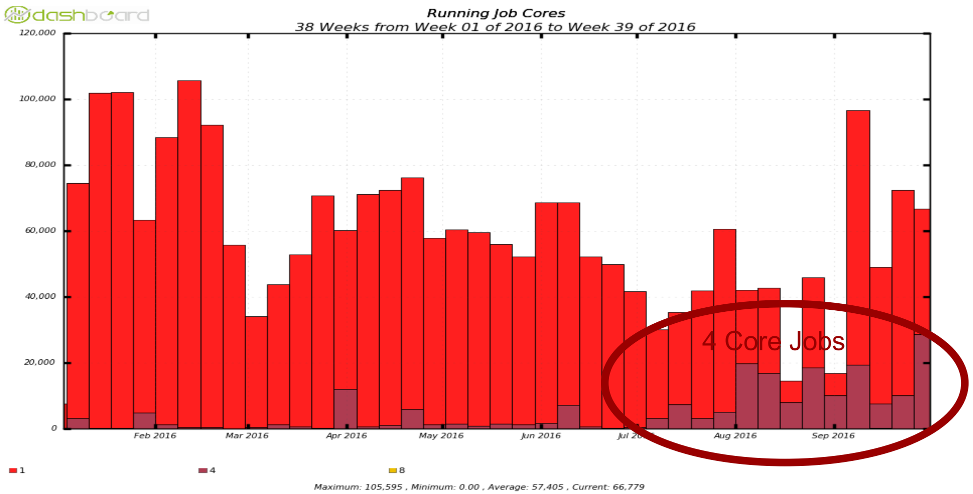
- A. Perez-Calero et al. - CMS readiness for multi-core workload scheduling
- A. Perez-Calero et al. - Stability and scalability of the CMS Global Pool Pushing HTCondor and glideinWMS to new limits

# CMSSW became multi-threaded

- **Advantages**
  - Memory is shared between threads
  - Need less jobs for the same amount of events
- **Efficiency of multi-core applications**
  - Big fraction of code needs to be thread-safe [Amdahl's law]
  - Achieved good CPU efficiency in recent software releases
- **Usage in Production**
  - PromptRECO (Tier-0) since 2015
  - Re-reco since end of 2015
  - Digi-reco since Summer 2016

C. Jones - CMS Event Processing Multi-core Efficiency Status



# Dynamic Data Management (DDM)

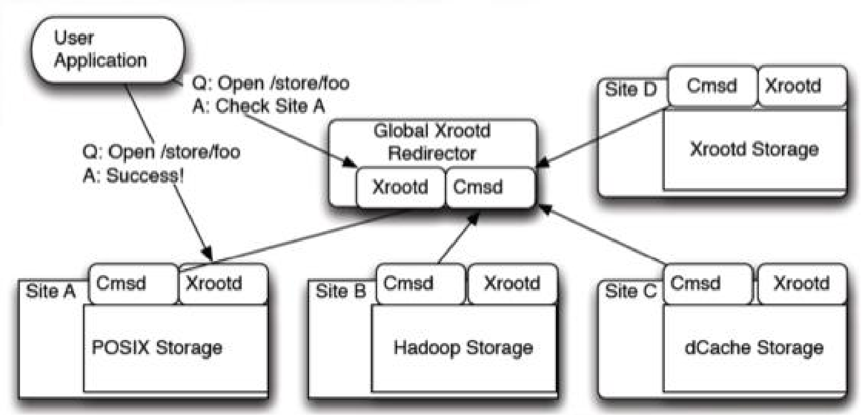
- DDM manages today about 54 PB of disk space
  - All Grid sites (Tier-0, Tier-1s and Tier2s) contribute to the DDM pool
- DDM controls the Phedex groups AnalysisOps and DataOps
- DDM creates new subscriptions or removes subscriptions based on
  - Data popularity
    - Access of data is recorded
    - Create more replicas for 'popular' datasets, lower the replication for less popular datasets
  - Disk usage level on a given site
    - Keep sites filled at a 'safe' level and always use available disk space
  - A set of DDM policy rules (examples, actual config my be different!)
    - Keep at least 2 copies of 2016 AOD data
    - Keep at least 3 copies of MINIAODSIM from main 2016 MC production campaign
    - Delete RECO datasets from disk after 3 months of lifetime

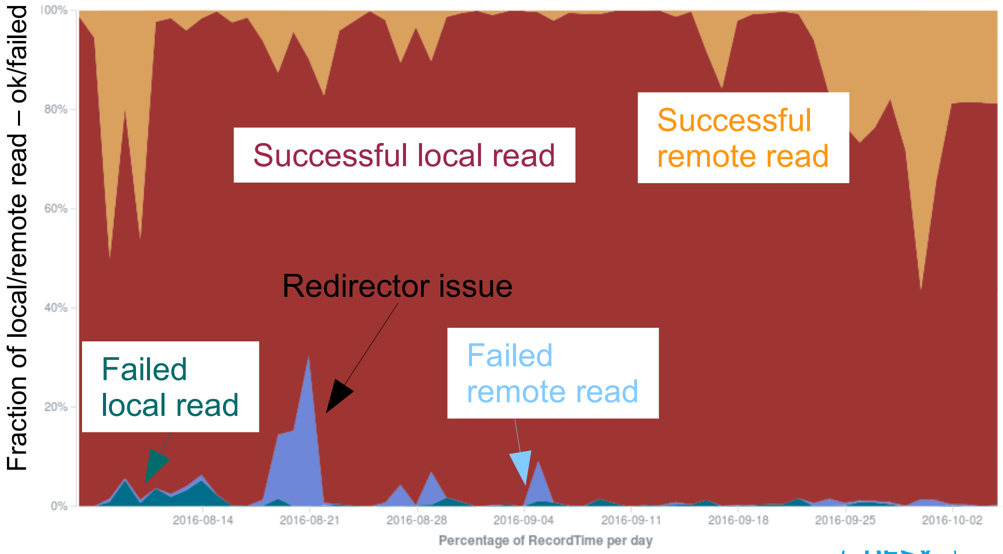
Y. Iiyama - Dynamo - The dynamic data management system for the distributed CMS computing system

# Remote Data Access via AAA Storage Federation

- Efficient remote data access important for flexibility
- CMS application I/O got improved for remote reads
- Present technology choice
  - Xrootd based storage federation
  - Sites “publish” storage inventory to regional re-director
  - Hierarchy of re-directors
    - Two redundant regional re-directors in Europe and US
    - One redundant global re-director at CERN
- Central production uses AAA routinely
  - RE-RECO of data
  - Classical Pileup-Mixing for MC DIGI-RECO
    - I/O intense read for pileup local
    - GEN-SIM of “primary physics process” via AAA
  - MC DIGI-RECO with premixed pileup

•D. Lange - CMS Full Simulation Status

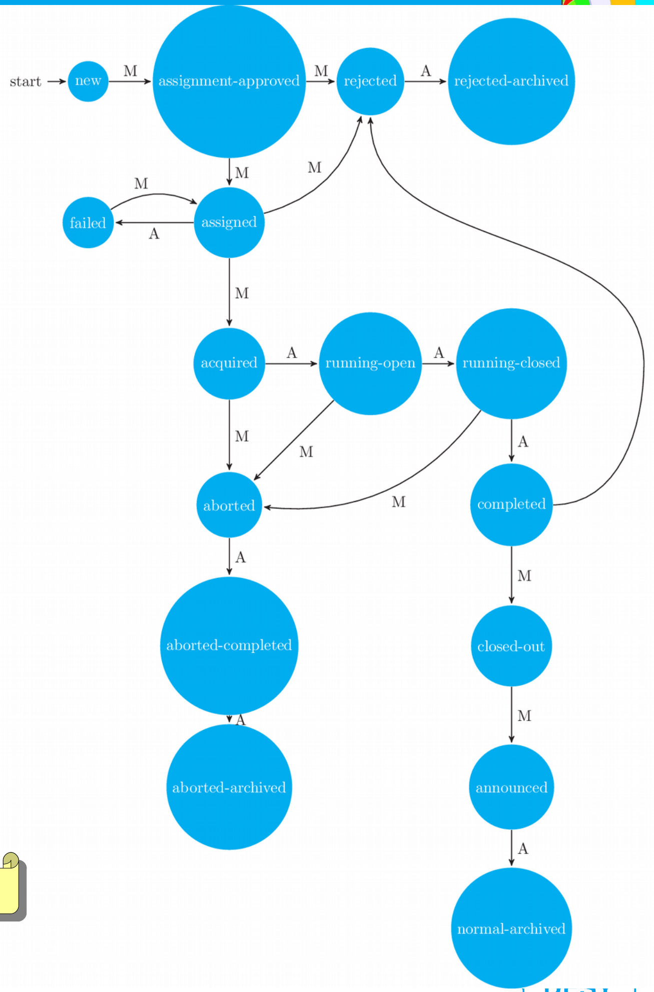


# Workflow Management

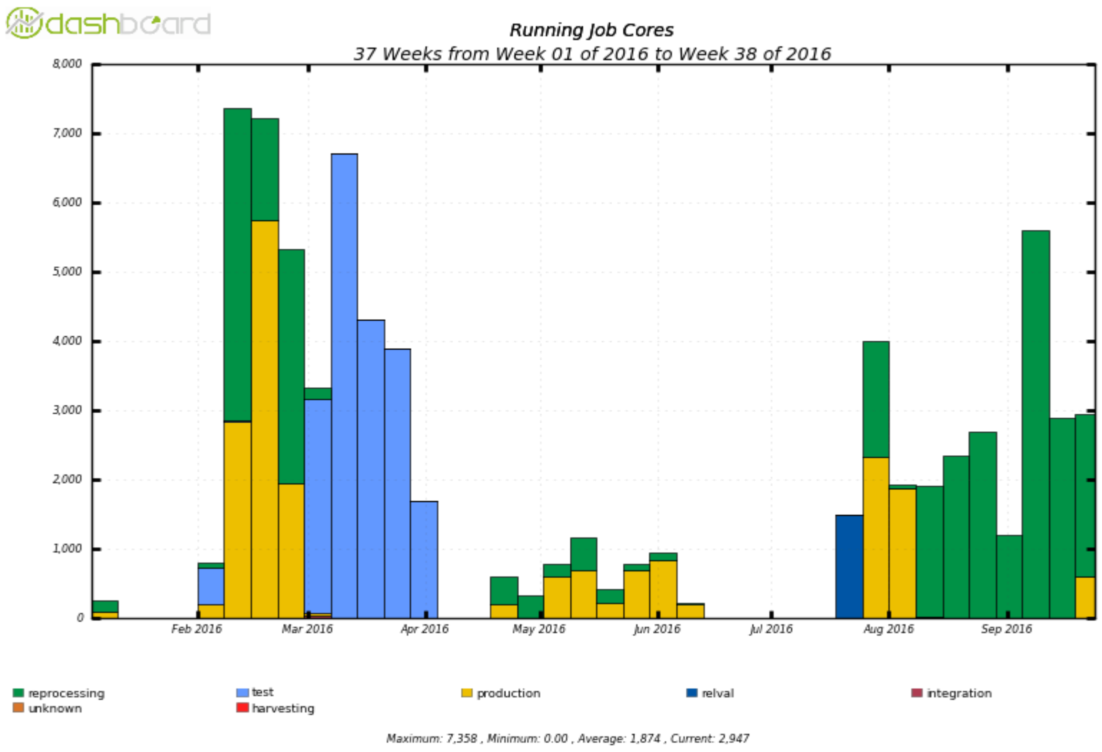
- One central Request Manager
- Unified Tool
  - Handling of input data staging and placement
  - Assignment of request to resources
    - Largely automated based on configurations for different campaigns
  - Recover failed parts of a request
  - Close-out and announce finished requests
- WMAgent(s)
  - Several instances
  - Job splitting and submission to Global Pool
  - Job tracking

J.-R. Vlimant - Software and Experience with Managing Workflows for the Computing Operation of the CMS Experiment



# High Level Trigger (HLT) Farm as Processing Resource

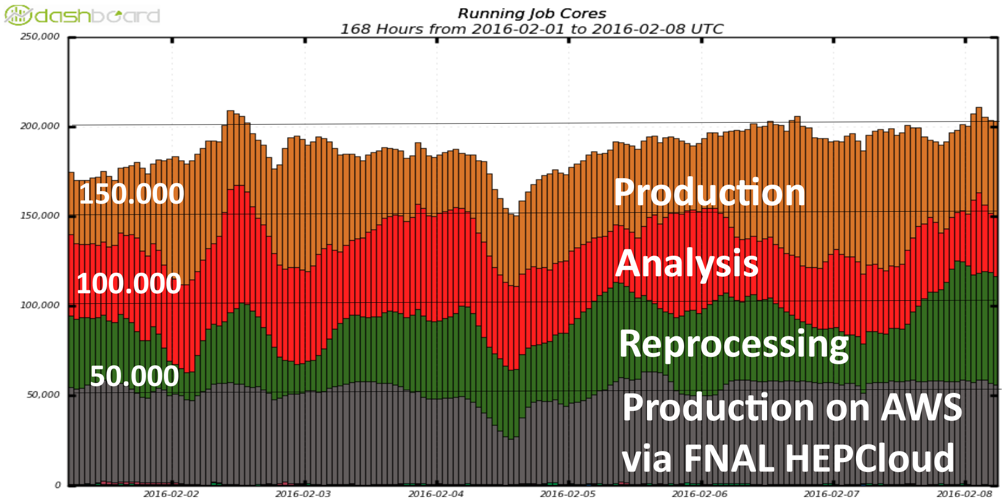
- HLT is a significant resource: ~15k cores
- Routinely used during longer breaks
  - HTL 'converted' to Openstack cloud
  - VMs join the Global HTCondor Pool
- Inter-fill mode
  - “Old” mode:
    - Start cloud and launch 2h jobs
    - Kill, if beam comes back
  - New mode:
    - Suitable for present LHC performance
    - Suspend running VM to disk, during beam
    - Resume VMs when luminosity gets lower
    - Successfully commissioned

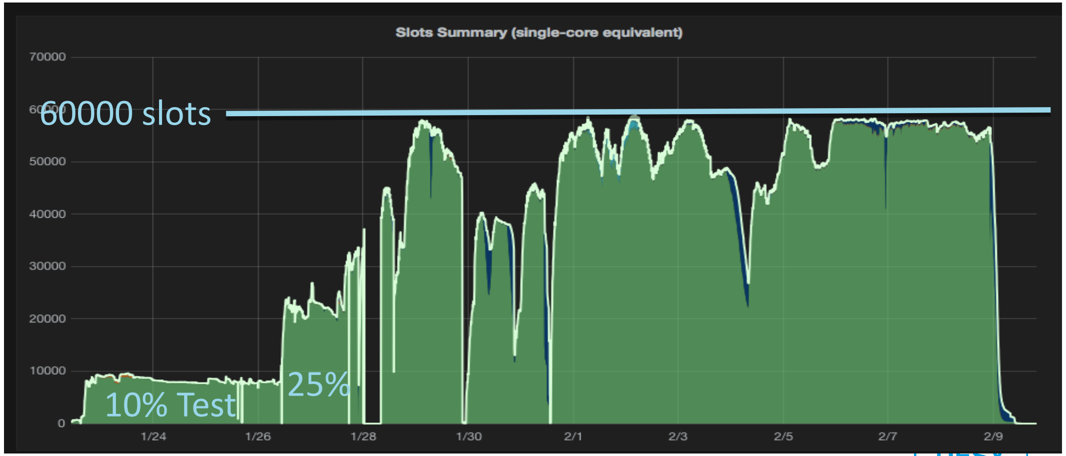


# Adding new Types of Resources: Clouds

- Dynamic extension of FNAL
  - Send jobs transparently to AWS cloud
- Reached 50k cores in AWS
- Quite some lessons learned
  - Pricing on Spot market
  - Costs for data handling
  - Suitable workflows
- Contribution to official MC production
  - ~0.5 billion events
- Important experience for other Cloud projects
  - Cloud procurement by CERN
  - Academic & commercial clouds being evaluated in various countries

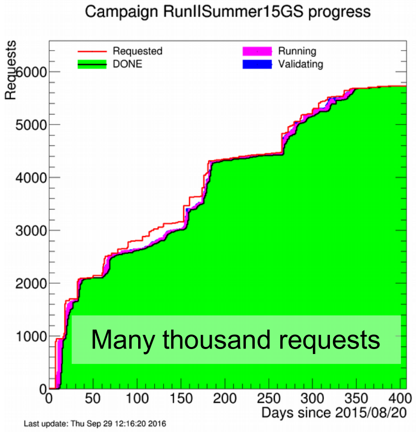
M. Girone - Experience in using commercial clouds in CMS

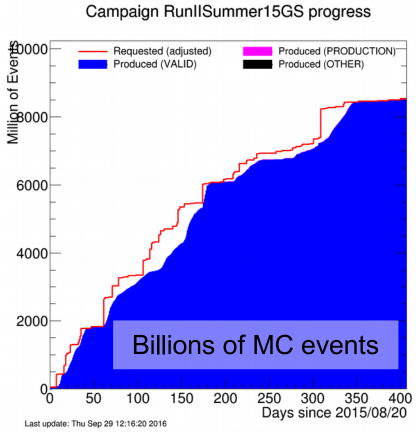


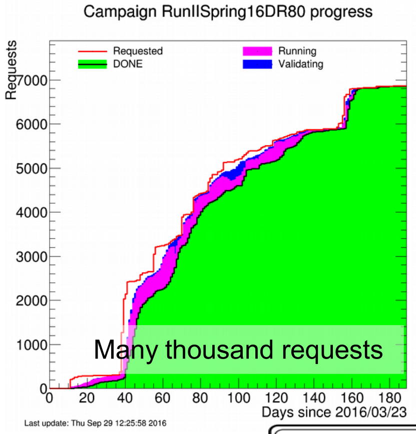
# Production and Processing in 2016

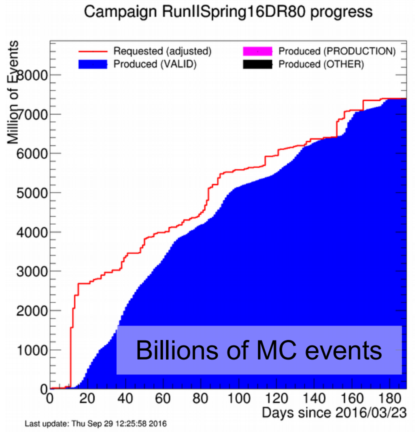
## MC Generation



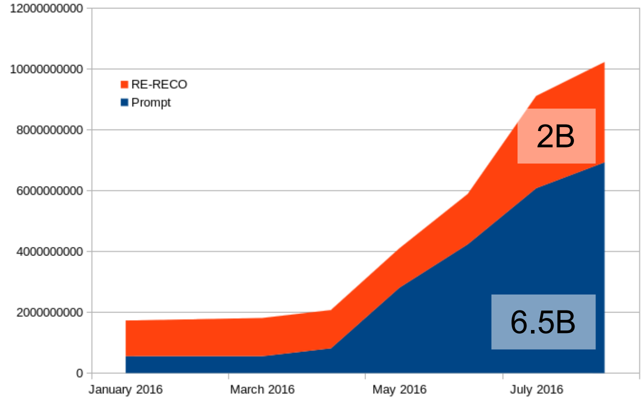



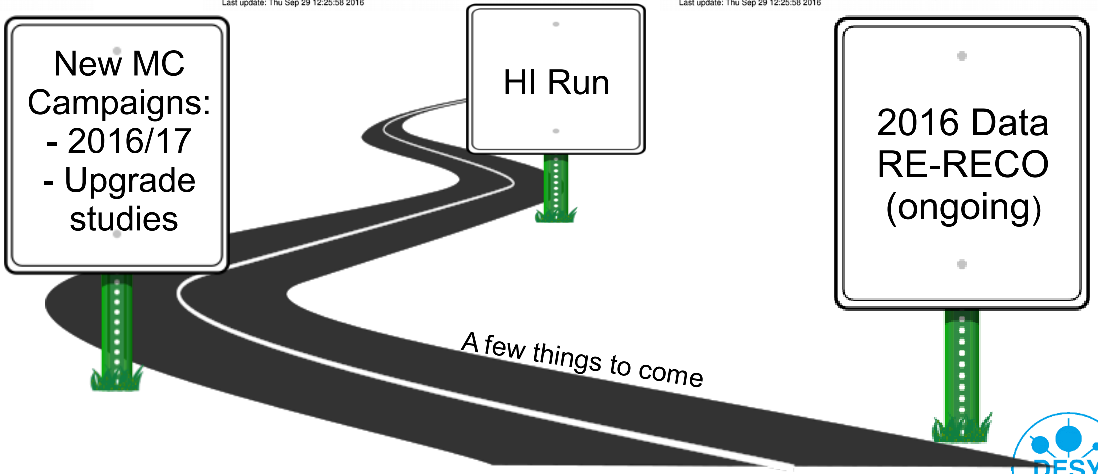

## MC DIGI-RECO




## Data: Prompt + RE-RECO



# Summary

- The LHC is performing even beyond planned performance
- Data taking, data processing and corresponding MC production became resource constrained
- A number of recent developments enable CMS to cope with the situation
  - Pooling of resources
  - Agile utilization
  - Tools for automation
  - Provisioning of resources beyond classical Grid sites
- Need to pay close attention to computing resource availability vs experiment plans

*"That's the kind of problems you want to have" (J. Butler – CMS Spokesperson)*

# Related CHEP Contributions

- A. Perez-Calero et al. - CMS readiness for multi-core workload scheduling
- A. Perez-Calero et al. - Stability and scalability of the CMS Global Pool: Pushing HTCondor and glideinWMS to new limits
- C. Jones - CMS Event Processing Multi-core Efficiency Status
- Y. Iiyama - Dynamo - The dynamic data management system for the distributed CMS computing system
- D. Lange - CMS Full Simulation Status
- J.-R. Vlimant - Software and Experience with Managing Workflows for the Computing Operation of the CMS Experiment
- HLT Poster?
- M. Girone - Experience in using commercial clouds in CMS