



Improvements in the CMS Computing System from Run2

CHEP 2015
Ian Fisk and Maria Girone
For CMS Collaboration



Introduction

- In Run2 CMS expects to increase the HLT output to $\sim 1\text{kHz}$
 - will promptly reconstruct more than twice as many events as the final year of Run2
 - expected pile-up will require twice as much processing time per event
- The budget and evolution of technology permits less than doubling of the computing resources between 2012 and 2015
- The main focus of Long Shutdown 1 has been finding ways to do more with less and to look for efficiency improvements in every system



Computing Model Changes in LS1

- Evolution out of the MONARC model had already started in Run1. The LS1 focused on
 - Additional **Flexibility** on the way resources are accessed
 - Improved **Performance** on the way resources are used
 - Optimized **Access** to data for analysis and production
- Instrumental for these changes have been the adoption of
 - The higher level trigger for production use
 - Logical separation between the **Disk and Tape** storage systems
 - **Dynamic Data Placement**
 - **Data Federation**
 - **Distributed Tier-0**
 - **CRAB₃**



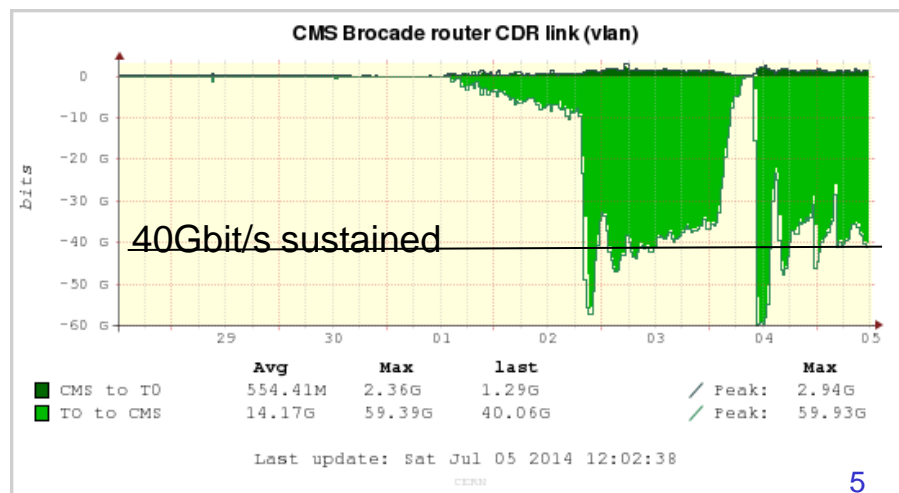
Reducing Resource Needs

- Operational improvements
 - Reducing the number of reprocessing passes expected and commissioning the use of the HLT farm for the main data processing in the winter
 - Constraining the simulation budget and developing techniques to allow simulation reconstruction to be run on more resources
 - Distributing the prompt reconstruction between CERN and the Tier-1 centers
- Technical improvements
 - Reconstruction improvements and the development of a multi-core application
 - Commissioning of the multi-core queues at Tier0 and Tier1s
 - Multi-core decreases the number of processes that need to be tracked and reduces the overall operational load



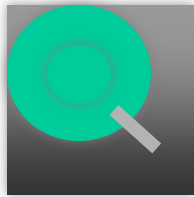
The HLT Farm

- An addition for Run II is the use of the High Level Trigger (HLT) farm for offline processing
 - It is a large computing resource (15k cores) that is similar in size to the Tier-0 in terms of number of cores, but we cannot reach this scale until March
 - Successfully interfaced using cloud computing tools. It is similar to the Tier-0 AI
- In 2014 the network link P5 to the computing center was upgraded from 20 to 60Gb/s
 - Far larger than needed for data taking but necessary to access the storage in the computing center for simulation reconstruction
 - **Will be upgraded to 120Gb/s before the 2015 run starts**
- Production workflows have been commissioned including the HI reprocessing, Gen-Sim, and Simulation reconstruction
 - All access to data is through the data federation and primarily served from CERN





Disk Tape Separation

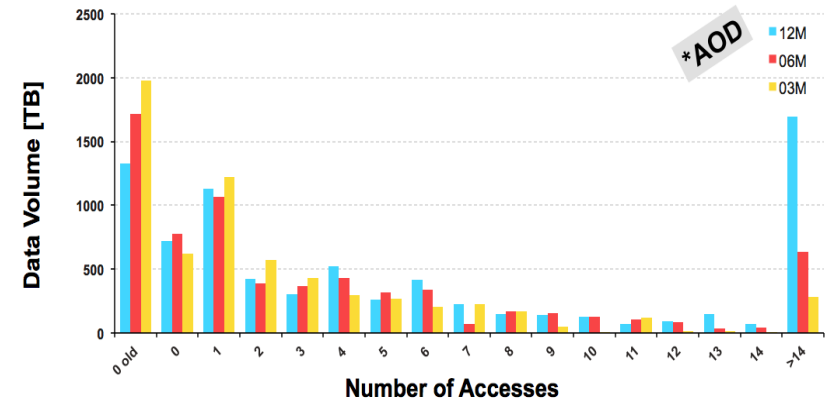


- CMS logically separated disk and tape.
Creating two sites: one disk and one tape
 - **Simple change with far reaching consequences**
 - Improved disk management
 - Ability to treat the archival services independently
 - Reduces functional differences between sites
 - Reduced differences between sites
 - Eliminates the need for storage classes



Improvements in Data Management

- In addition to work on data federation we have tried to improve our traditional data placement and access
- The use of samples is continuously monitored through the data popularity
 - Number of replicas depends on the popularity of the datasets
- Samples will be replicated dynamically if the load is high and replicas removed if they are not accessed for a period of time
- Samples are only deleted when there is new data to replicate, disks are kept full
- “0 old”, shows the volume data that were last accessed prior to the period covered by the plot; the “0” bin stands for no access in the period selected
- The zero bin includes un-accessed replicas and datasets that have only **one** copy on disk
- As of today, the system has triggered the deletion of roughly 1.5 PB of the least popular samples residing at Tier-2 sites, and it is being enabled in a few Tier-1s





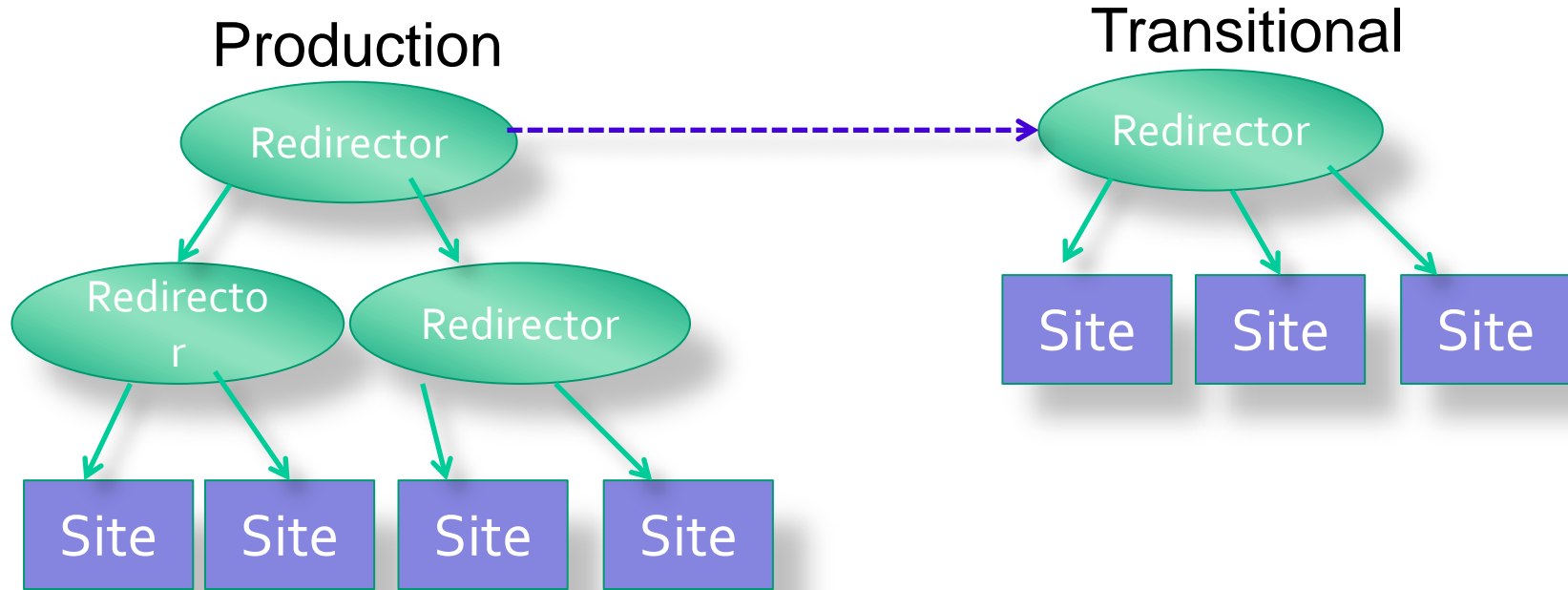
Improvements in Data Access

- Any Data, Anytime, Anywhere (AAA) has been a primary focus area in 2014
 - CERN, all Tier-1s, most of the Tier-2 sites serve data in the federation
 - More than 90% of all current CMS data should be accessible
 - Sufficient IO capacity to provide 20% of the total access (~100TB/day)
 - Enable to system of hierarchical redirectors to maintain access within geographic regions when possible
 - Nearly all sites are configured to use the federation to access samples, if they aren't available locally
 - Optimization of the IO has been an ongoing activity for several years, which has paid off in high CPU efficiency over the wide area
 - Big push in 2014 to commission sites to measure IO and file open rates and to understand the sustainable load and to deploy and use advanced monitoring
- CMS has developed a new user analysis data format that is **5-10x smaller than the Run 1** format
 - Design and content based on Run 1 experience across analysis groups
 - Targets most analysis needs (80-90%)
 - Potential for big analysis improvements in Run 2:
 - Increased analysis agility in limited computing resources: Recreating miniAOD is much faster than rerunning the reconstruction for a vast range of performance improvements



Creation of the Transitional Federation

- The addition of the production workflow puts additional constraints on the required reliability of the Federation



- Validated sites are in the production federation, sites being commissioned are in an independent federation and only when a sample cannot be found in production are they used
- This new concept has been a result of a collaborative effort between ATLAS and CMS federation developers



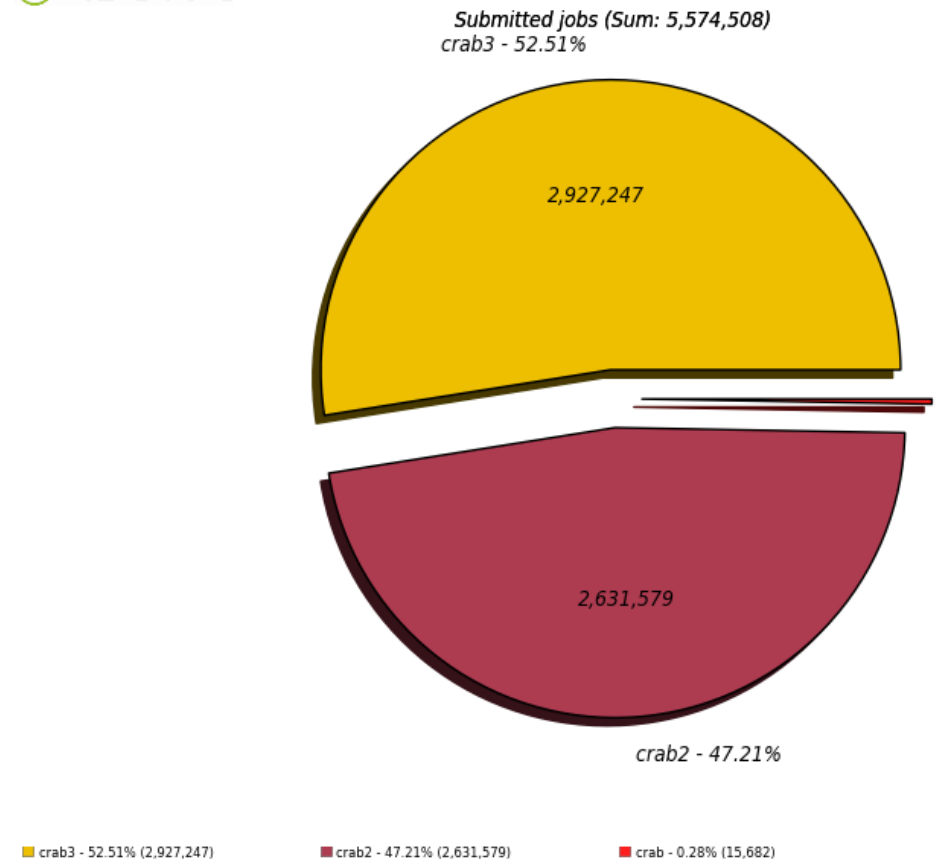
Tier-0 Commissioning

- The Tier-0 submission infrastructure was reworked over LS₁
 - The capability of distributing Prompt Reco was added to allow the Tier-1 centers to contribute
 - The Tier-0 processing predominately come from the CERN Agile Infrastructure with direct resources provisioning through Condor
 - CMS was an early adopter of CERN AI
 - The Tier-0 infrastructure was implemented using the components from the rest of the computing reprocessing system
 - This reduces the software maintenance of the system and reduces the overall operations load



CRAB3

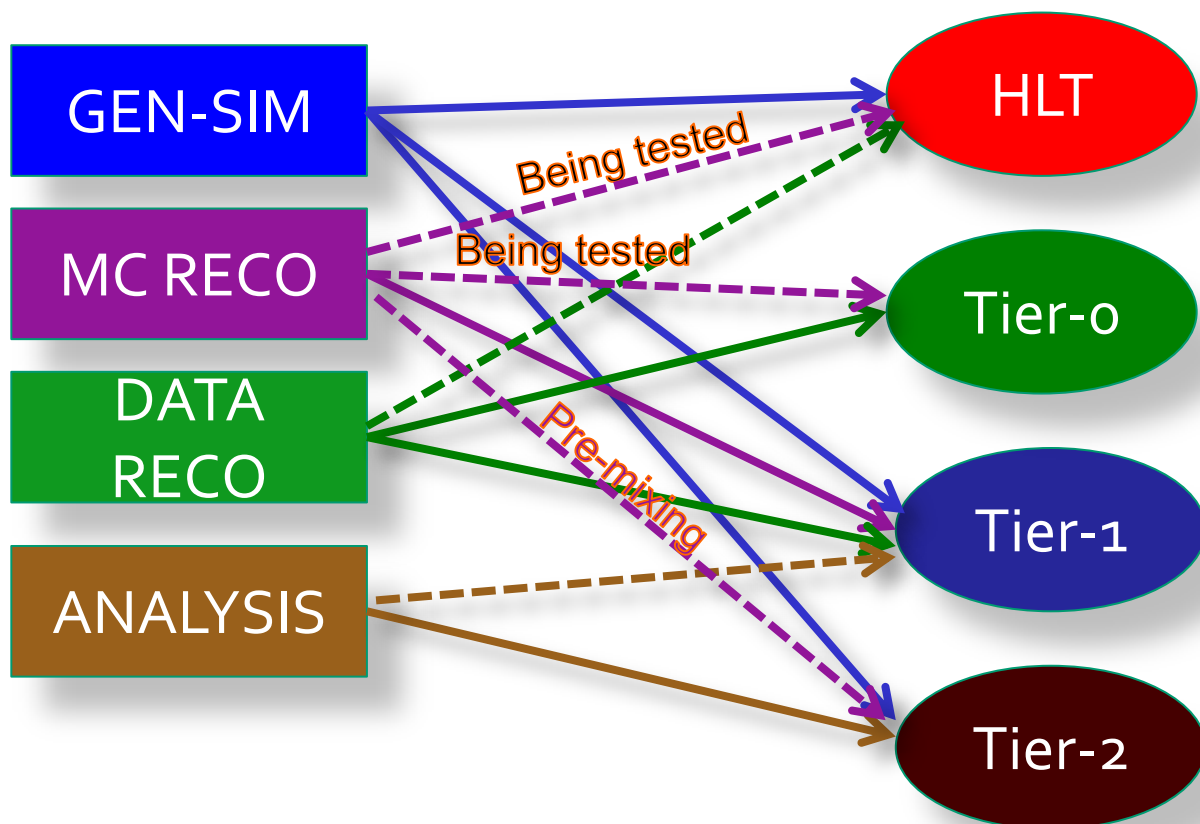
- The CMS Remote Analysis Building (CRAB) has been the user interface to the distributed computing resources since the beginning
- Completed the development of the next generation of CRAB during LS1 (CRAB3)
 - Is a client-server model
 - A light client uploads to a server
 - Given much better resubmission and allows the experiment more central prioritization
 - Output files are handed asynchronously





Blurring the Site Boundaries

- In Run2 CMS computing resources are intended to work more like a coherent system than a collection of sites with specialized functions
 - Improved networks have been key to this
- Data Federation will make CMS datasets available transparently across the Grid
- One central queue for all resources and all workflows
- The HLT farm is now an integrated resource when we are not in data-taking mode





- A lot of work has been done in LS₁
 - We have a functional data federation, which we expect to declare production ready before the start of the run
 - We have much better flexibility in where workflows can run
 - We have reworked the Tier-0 infrastructure for distributed prompt-reconstruction

- We needed to make big increases in operational efficiency to survive the conditions in Run2 with the resources that could be afforded