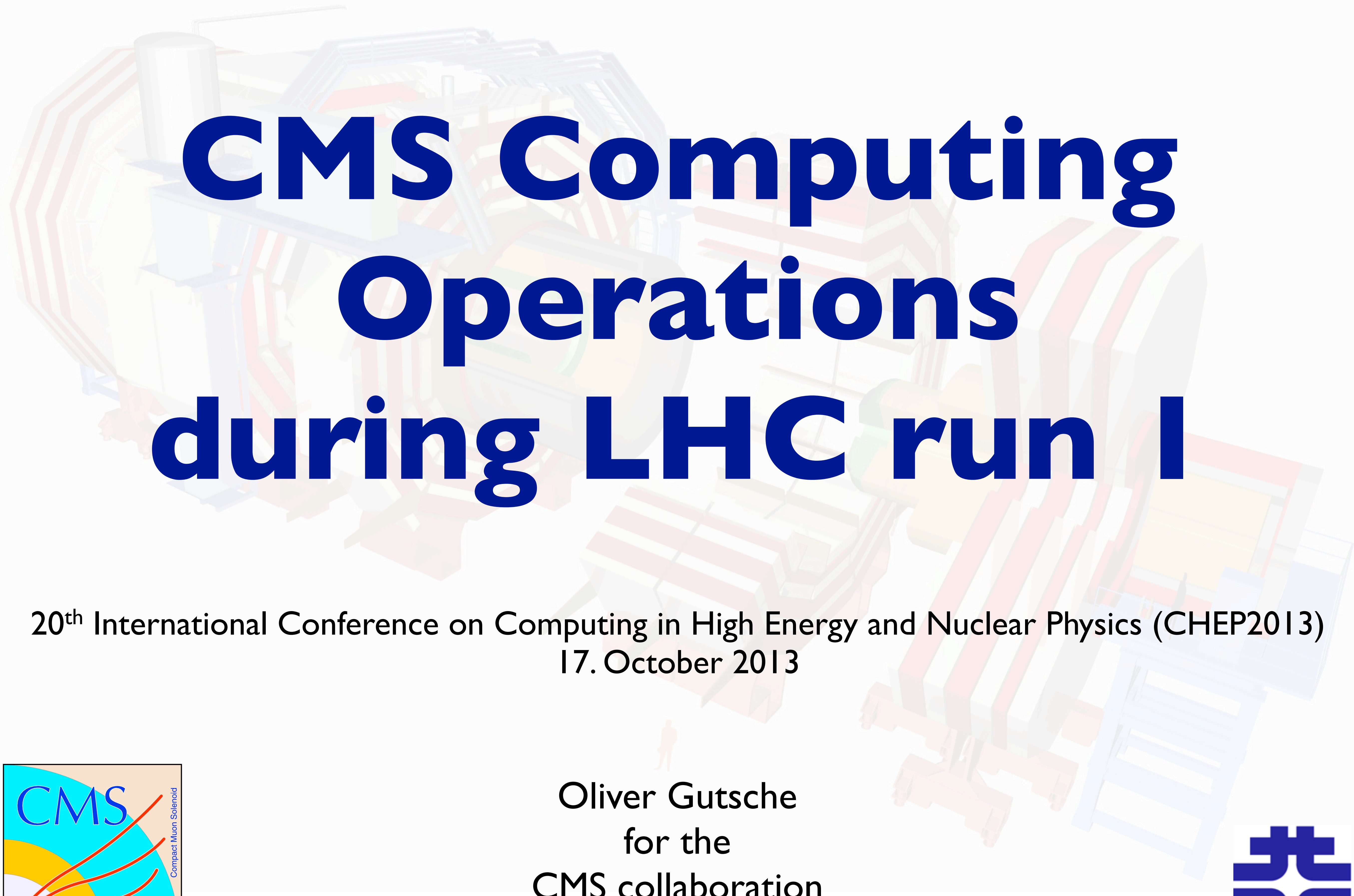
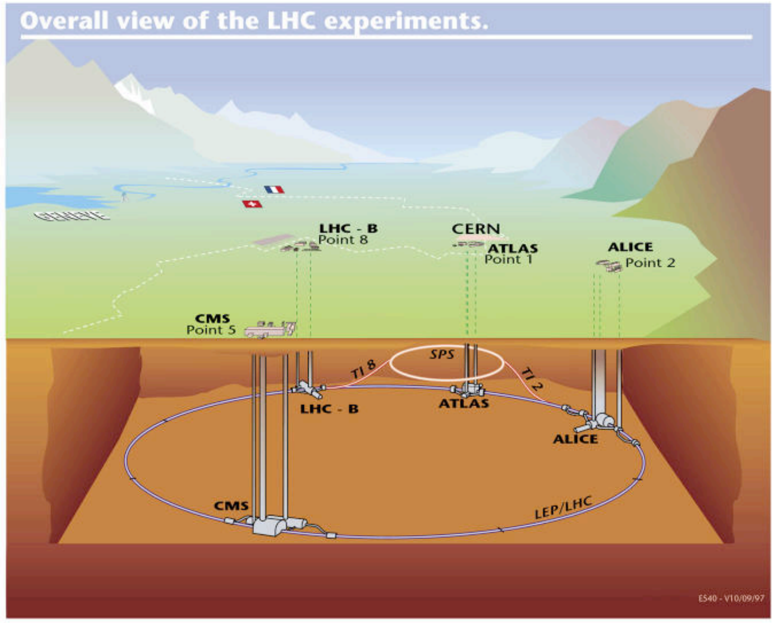
CMS Computing Operations during LHC run 1

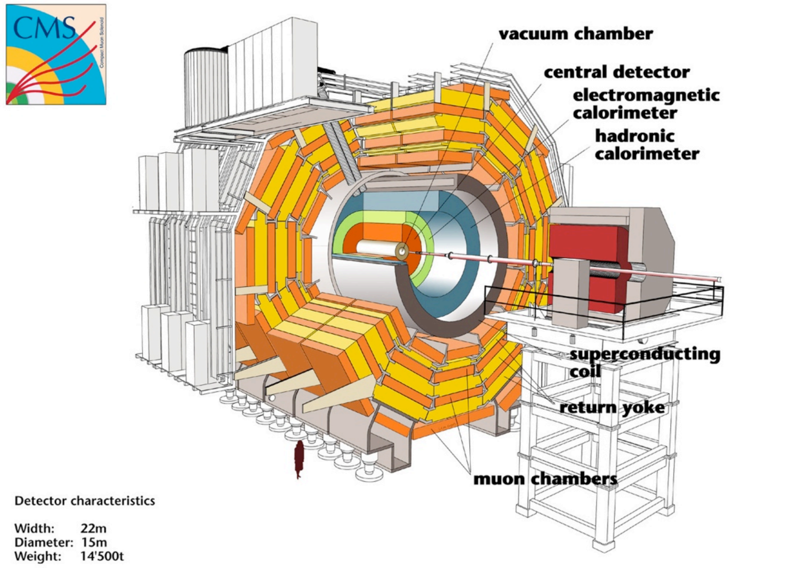


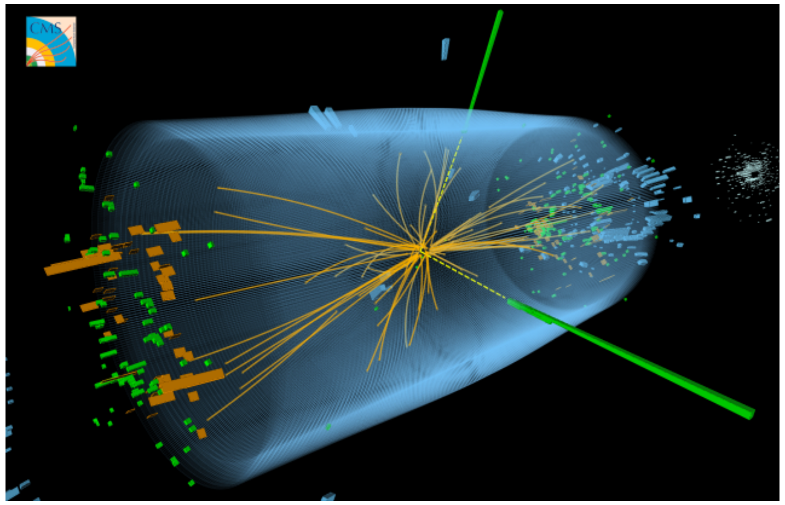
20th International Conference on Computing in High Energy and Nuclear Physics (CHEP2013)
17. October 2013

Oliver Gutsche
for the
CMS collaboration

Outline





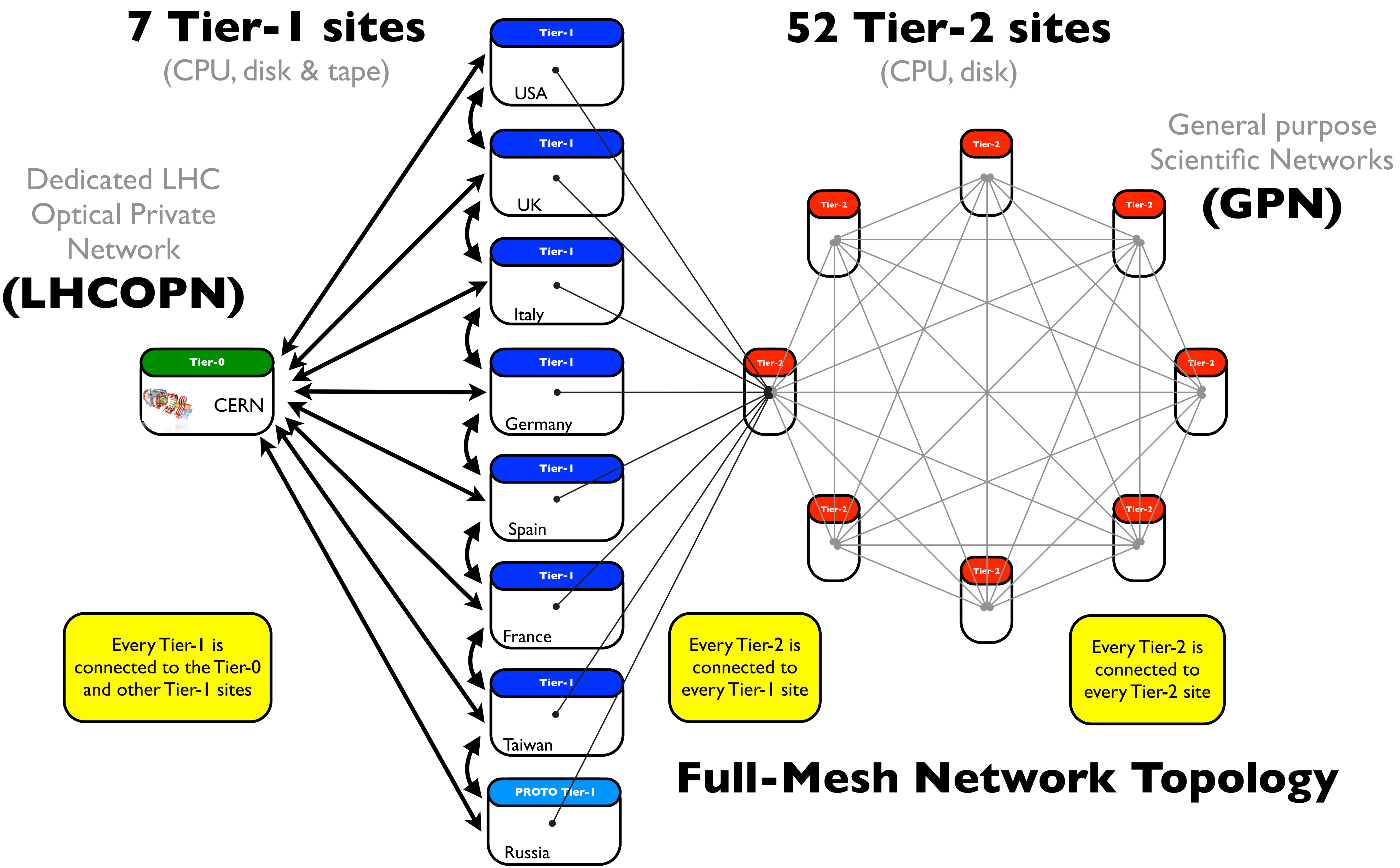


- Focus of this talk: CMS computing operations in LHC run 1 (2010-2012)
  - Introduction of CMS computing infrastructure, services and workflows
  - Lessons learned and plans for improvements for LHC run 2 (2015-2017)

- This talk will focus on the techniques and lessons learned and not on how the resources were used.

Introduction

CMS Computing Infrastructure



7 Tier-1 sites

(CPU, disk & tape)

52 Tier-2 sites

(CPU, disk)

Dedicated LHC Optical Private Network

(LHCOPN)

General purpose Scientific Networks

(GPN)

Every Tier-1 is connected to the Tier-0 and other Tier-1 sites

Every Tier-2 is connected to every Tier-1 site

Every Tier-2 is connected to every Tier-2 site

Full-Mesh Network Topology

CMS Computing Services

Building blocks needed to support the execution of CMS workflows on a distributed infrastructure through GRID technologies

Transfer system:

- Files are organized in datasets with similar physics content
- Transfer system replicates datasets at CMS sites
- Name: *PhEDEx*

Bookkeeping system:

- Metadata catalogue of all CMS datasets and files
- Name: *DBS*

Constants system:

- Distributed access to constants for all jobs at all sites based on HTTP SQUID caches
- Name: *Frontier*

Software distribution system:

- CMS software releases are installed locally or accessed through shared GRID filesystem based on HTTP SQUID caches
- Name: *CVMFS*

Submission infrastructure:

- Used by production and analysis system to execute jobs on distributed infrastructure
- Names: *gLite WMS, HTCondor_G, GlideIn WMS*

Monitoring system:

- All jobs are instrumented to report monitoring information to central instance
- Name: *DashBoard*

Site monitoring system:

- Service at sites (CE, SE, etc) are probed periodically
- Name: *SAM (Site Availability Monitoring)*

Site stress test system:

- Sites are probed periodically with complete workflows
- Name: *HammerCloud*

Production system:

- State machine to execute production workflows
- No custom code outside a software release
- Name: *WMAgent*

Analysis system:

- User system to discover, split and execute user analysis projects
- Move user-specific code to GRID workernode
- Name: *CRAB*

Main data flows

Archiving

Tape based: grouping by physics content of related files on separate sets of tapes (tape families) have to be defined before files start arriving at a Tier-1 site

T0 → T1

- **RAW data**
  - RAW data is recorded at Tier-0
    - Stored on tape at CERN as "cold" backup copy
  - RAW data is distributed across the Tier-1s via the LHCOPN network links
    - Archived on tape and current year kept on disk for quick access

T2 → T1

- **Simulation output**
  - Output of Monte Carlo generation and simulation is produced mainly on Tier-2s (CPU intensive workflows)
  - Archived on tape distributed across the Tier-1s via the LHCOPN and GPN network links

T1 → T1

- **Analysis Object Data (AOD) production**
  - Output of reconstructions of RAW data and simulation is produced on Tier-1s (strong I/O capabilities required)
  - Archived on tape at Tier-1 site that has archival copy of input and ran reconstruction workflow

Serving

T1 → T2

- **Analysis Object Data (AOD) access**
  - Access to reconstructed data and simulated events on Tier-2 sites
    - Samples are distributed to the Tier-2s via the GPN network links
    - Tier-2 disk is logically separated into managed portions for common samples and unmanaged areas for user files (ntuples)physics samples

Main workflows

Production workflows

T0

- **Data recording**
  - Collisions are recorded by the detector, selected by the trigger and sent from the detector pit to the CERN computing center in binary format
  - 10% of the selected collisions are express repacked into the CMS ROOT format and reconstructed
    - Latency of express reconstruction is 1 hour to allow for prompt alignment and calibration workflows and data quality monitoring
  - 100% of the selected collisions are repacked into the CMS ROOT format and promptly reconstructed
    - Prompt reconstruction starts 48 hours after events have been recorded to incorporate updated calibrations

T1/2

- **Simulation**
  - Collisions are generated using theory software packages and the detector response is simulated using GEANT4
    - CPU intensive workflow
    - Needs little or no input

T1

- **Data and simulated event reconstruction**
  - Collisions are reconstructed with different software versions and/or calibrations
    - RAW data and simulated events archived on Tier-1s need to be pre-staged to disk for efficient access
    - Reconstruction of simulated events needs to access additional datasets as input to simulate additional interactions in the detector (PileUp)

Analysis

T2

- **Data and simulated event analysis**
  - Collisions are accessed by physicists using officially released and self-written code
    - Input is distributed beforehand through transfer system and analysis jobs are sent to data location
    - Output is stored on Tier-2s outside the transfer system and catalogues but can be elevated
  - Multi-user access instead of single user production mode

Operation Teams

- CMS supports the infrastructure and the sites through
  - Central teams of operator and experts
  - Site contacts that bridge CMS to site administrators
- Team members are both **physicists and engineers**, physicists and institutes get awarded service credit to fulfill author list requirements. Totals for 2012:
  - Central: 40 FTE, Site contacts: 60 FTE

Computing Operations

- Handles **central services, site support, production workflows**
- Formed from experts with management functions and teams of operators
  - Tier-0 team, Workflow team, Transfer team, Site support team, Submission infrastructure team, Monitoring team

Physics support

- Handles **support of analysis users** and their interaction with the analysis system
- Because of the large user base of CMS (up to 500 active users at a given time), most effort is spent to support analysis users on an individual or small group basis

Details and lessons learned

Computing Shifts

- CMS established **computing shifts to monitor all running systems, alarm sites and services before problems become critical and triage problems to experts** for fast problem resolution especially during data taking
  - **CSP**: Computing Shift Person
    - 24/7 coverage by 3 shifts of non-expert collaborators (preferably in 3 time zones: Europe, Asia, US) who follow shift-instructions every 2 hours and alarm via tickets or phone/chat
  - **CRC**: Computing Run Coordinator
    - Computing expert shift for 1 week, during data taking based at CERN: on-call interface between CMS detector operation and CMS data taking coordination and computing, main contact for CSP shifters, ability to intervene out of business hours in case of problems for central services

- **Lessons learned:**
  - Vital for operations during LHC run 1, most of the interactions with sites are being initiated by tickets from CSP shifters
  - Many CSP checks and tasks could be automated. Process will be lengthy and work intensive. CMS didn't have yet the time and manpower but plans to convert most of the CSP checks into automated systems.

Tier-0 operation

- The Tier-0 processing infrastructure and its operation are of very high importance to CMS (in case of problems, data taking is impacted, data quality monitoring is limited, buffer space at CERN is filling up)
  - CMS operated the Tier-0 with two full time operators on both sides of the atlantic to increase coverage during the business day
    - Weekend and emergencies were covered by the CRC and the operators if available

- **Lessons learned:**
  - Coverage on both sides of the Atlantic crucial especially during startup phases of detector operations
  - Operators need to have intimate knowledge of the processing infrastructure (close to be a full time developer) to be able to react to changing data taking conditions and problems resulting thereof

Web services

- Many of the central services (like the transfer service and the metadata service) rely on web interfaces.
- CMS provides access to all services through a common load-balanced web platform called CMSWEB.
  - CMSWEB also provides specialized functionalities like NoSQL databases.
  - All installed services go through a structured deployment procedure with testbed deployment and testing.
    - The release schedule is regular and enforced (every month).

- **<u>Lessons learned:</u>**
  - The regular deployments successfully kept the systems stable while rolling out fixes and new developments.
    - At times painful, but it was necessary in the end to formalize the development process and introduce predictability for upgrades and updates and increase the chances for a problem-free rollout.
  - In LHC run 1, not all services were under the umbrella of the CMSWEB deployment. The goal is to transition all services (if possible) to this common testing and deployment scheme.

Processing infrastructure operation

- Manpower intensive due to the large variety of different workflow types and the reliance on many different systems, central and at the sites
  - Especially the last 5% to completion of workflows is time and manpower intensive
- Tier-1 Storage model introduces significant latency in completing work:
  - Datasets can only be processed at sites that archive the input on tape and will save the output on tape directly (workflows running at a Tier-1 immediately write to tape)
  - Introduces latency if workflows can only be processed at one site while other sites are idle
    - Happens if the variety of workflows to be run in parallel is limited

- **Lessons learned:**
  - Manpower needs are difficult to reduce further because of the magnitude of different failure modes, but continued improvement is a goal of CMS before and during LHC run 2. (With experience and more stable workflow variety, manpower needs are expected to go down)
  - The processing inefficiency due to the Tier-1 storage model is planned to be solved by
    - Separating disk/tape at Tier-1s into a big read/write disk pool from and a small tape reading/writing disk pool. The transfer system will be used to bring samples online. Tape writing can be handled selectively and also independent of processing location (process at Tier-1 A and archive at Tier-1 B).
    - Xrootd remote access through the CMS data federation (AAA project) will increase the flexibility further and reduce the latency significantly by exploiting the strong LHCOPN network between the Tier-1s

Transfer and Site Support

- The CMS transfer system is functioning very well and only few last-percentage problems need to be solved actively by transfer support.
  - Main problems are site and infrastructure related (slow network links, instabilities in SRM endpoints at sites, authentication problems, etc.)
- The site issues are further being worked on by dedicated site support operators who work with the site admins to solve problems and improve longterm stability.
  - SAM tests and HammerCloud results are key for a reliable site infrastructure and crucial for the operation of all CMS sites

- **Lessons learned:**
  - Low level network problems need to be better monitored, the deployment of PerfSonar and its integration in standard monitoring procedures is planned to solve this problem.
  - CMS has a large group of reliable sites but a small group of more problematic ones. CMS is intensifying the site support effort to work with the few unreliable sites to enable them to become as reliable as the others. Site reliability will be very important for LHC run 2 (see disk resource utilization).

Physics support

- Difficult and manpower intensive task
  - Many different user-generated workflows with diverse requirements, more challenging than single-user central processing/production → users are in control of definition and validation
    - A much harder problem than central production workflows which are all run by the same user
- CMS' dedicated support team of experts is helping users and supporting the analysis GRID tools (very busy!)
  - Large improvements have been made by moving to a pilot based submission infrastructure

- **Lessons learned:**
  - Inefficiencies through strictly sending jobs to Tier-2s where data is stored locally
    - CMS is planning to dynamically re-broker analysis jobs exploiting the pilot based submission infrastructure in combination with remote access to files/datasets through xrootd (AAA)
  - Failure rates dominated by remote stage out problems of user-generated content (ntuples, not registered in transfer and metadata systems)
    - CMS is planning to move to an asynchronous stage out implementation that first stores the output locally and then moves it to a destination Tier-2 site using transfer system techniques

Disk resource utilization

- In LHC run 1, disk resources at Tier-1s were managed by the individual mass storage systems (MSS) with limited or no control over what is kept on disk.
- T2 disk was separated in managed space allocations (managed by the transfer system) maintained by individual physics groups and central allocations for commonly used datasets (like physics background simulations, etc.), as well as unmanaged space for user-generated output

- **Lessons learned:**
  - Setup prevented analysis at Tier-1s because of the lack of predictability of user access to datasets that could overwhelm the tape systems
    - The disk/tape separation at the Tier-1 sites will allow CMS to open the sites for analysis access without endangering the tape systems
  - Tier-2 disk group space setup introduced inefficiencies in which datasets are available on the Tier-2 sites for analysis (some physics groups are not able to use their space allocations very efficiently and tend to store a fraction of datasets that are not accessed at all).
    - CMS wants to simplify the setup and resolve the physics group allocations difficulties:
      - Dynamic data placement based on popularity information of datasets (gained through tracking of access patterns of datasets and queue depth of analysis jobs) is planned to take over populating the managed disk space.
      - An automatic cache release in several stages will take care of releasing disk space that can be used better for dataset in higher demand.

Submission infrastructure

- At the beginning of LHC run 1, CMS supported mainly two direct GRID submission infrastructures: gLite WMS and HTCondor_G
- During the run, CMS transitioned to the pilot based GlideIn WMS system which provides higher efficiencies for job completion.
- While supporting all three submission modes, CMS relied on different setups for production and analysis and used prioritization on site level to balance between the different activities (analysis/production) and different analysis users

- **Lessons learned:**
  - The mix of submission modes lead to inefficiencies in prioritizing workflows (production vs. analysis and also within production and analysis)
    - CMS plans to move to a global GlideIn WMS pool (see Poster P3.29) and handle all prioritization within a single system, removing the need to prioritize on site level

Final thoughts

Summary & Outlook

- LHC run 1 provided a lot of manpower intensive operational challenges for CMS computing under rapidly changing conditions and requirements
  - This is not unusual and is expected to continue during LHC run 2 to some extent
- During the successful LHC run 1 operation, CMS gained a lot of operational experience gained and identified areas of improvement:
  - Disk/tape separation at the Tier-1s
  - Global GlideIn WMS pool
  - Dynamic data placement and automatic cache release
  - Global data federation (AAA, reference to separate talk)