CMS Computing:
Recent Experiences and Future Challenges

Christoph Wissing (DESY)
for the CMS Collaboration

NEC 2013
September 9th-14th, 2013
Proton Collisions at the Energy Frontier

Access to rare processes

- High center of mass energy
- High luminosity
- Very distinct event selection
Large Hadron Collider (LHC)

Beam Energy 7 TeV on 7 TeV *
Bunch Crossing 40 Mhz *
pp Collisions ~1 GHz
Event rate (e.g. Higgs) ~10^{-2} Hz

* Design value, 2011: 7 TeV, 2012: 8 TeV, 20 MHz Bunch Crossing
## CMS Detector

<table>
<thead>
<tr>
<th>Name</th>
<th>Channels</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pixel</td>
<td>66 M</td>
</tr>
<tr>
<td>SiStrip</td>
<td>9.3 M</td>
</tr>
<tr>
<td>Muons</td>
<td>890 k</td>
</tr>
<tr>
<td>Preshower</td>
<td>140 k</td>
</tr>
<tr>
<td>ECAL</td>
<td>76 k</td>
</tr>
<tr>
<td>HCAL</td>
<td>9 k</td>
</tr>
</tbody>
</table>

Data from: 2008 JINST 3 S08004, Table 9.1

- **RAW data**
  - 0.3 MB/event
  - 1 MB/event (MC)

- **RECO data**
  - 500 kB/event

- **AOD data**
  - 100 kB/event

- 2 stage Trigger
  - some 100 ev/s
Tier 0
- prompt reconstruction
- store RAW data and export to T1s

Tier 1
- re-reconstruction
- long term storage of RAW and MC data

Tier 2
- MC production
- User analysis

Tier 3
- Mainly user analysis

More than 50 CMS centers, in more then 20 countries

Flags taken from Wikipedia:
http://de.wikipedia.org/wiki/Liste_der_Nationalflaggen
Run I: Luminosity, Collected Data and MC Events

CMS Integrated Luminosity, pp

Data included from 2010-03-30 11:21 to 2012-12-16 20:49 UTC

~35B Events
~36B in 2012-13

~58B Events
Run I – Event Sizes for Data and MC Events

Observations

> Data Processing for Run I
  - Full legacy reprocessing of 2011 pending
    > 2011 data reprocessed several times already
  - Low activity shortly after Run end

> MC production
  - Activity slightly reduced at Run end

> Data tiers
  - Switch to AOD visible
  - Clear reduction in disk usage
  - RECO might become transient in Run2

~14PB

30PB
21PB in 2012-13

Christoph Wissing
Events with high pile up end of 2011
- (Too) High memory usage
- Could not use all job slots in Tier-0!
- (Too) Long processing time

Enormous code optimization campaign
- Focus on track reconstruction
- Gain of factor in memory footprint
- Gain of factor 3 for high pile up events

Efforts ongoing to further optimize existing code
Challenges for Computing in 2015 and beyond

- Event Pile-up expected to rise with luminosity
  - More complex events demanding more CPU for prompt reconstruction
  - Expect \( \sim 2.5 \) times CPU time per event

- Logging rate of 800Hz – 1200Hz expected [assuming trigger thresholds similar to 2012]
  - Increase by factor 2

- LHC bunch spacing from 50ns to 25ns
  - Another factor 2 in reconstruction time needed with 2012 software
  - Efforts ongoing to regain by some code improvements

- We see up to a factor of 10 increase in CPU demands

- Storage demands would increase by \( \sim 2-4 \)
Growth of WLCG Resources

- WLCG resources growing roughly linearly with time
  - Mainly due to “~constant budget” for the sites
- Have to use existing resources more efficiently to be able to cope with 2015 data taking conditions
Multi Core Usage

> Forked Mode

- Improved memory usage: share libraries, geometry etc in memory
- Already supported in recent CMSSW
- Tested with present production tools
  > Can use multi-core queues at sites
- Slight disadvantage:
  > Inefficiencies in “merge” step

![Memory usage per core vs number of cores](image_url)

![Diagram of multi core usage](image_url)
Towards a multi-threaded Application

- Staged approach with increasing complexity
  - Phase 1: Process individual events in parallel
  - Phase 2: Process modules within one event in parallel
  - Phase 3: Process runs and lumi sections in parallel

- Need to have certain parts thread safe
  - Mixture of thread-safe and unsafe modules allowed

- Benefit of multi cores on when vast majority can run multi-threaded (Amdahl's Law)
ARM Architecture

➢ Hitting the power wall
  ▪ No gain in CPU clock speed
  ▪ Limit: Power consumption

➢ Operational costs
  ▪ 25%-50% for power
  ▪ Consider performance per watt rather than performance per purchase

➢ ARM architecture
  ▪ Mainly used for mobile devices
  ▪ Huge quantities
  ▪ Very power effective
  ▪ 64bit ARMv8 expected this year
  ▪ Commercial ARM servers in 2014
Relaxing the strict Tier Computing Model

Networking expectations from 1996

- Event Summary Data (ESD) live to one regional centre
- Analysis Object Data (AOD) live to one regional centre

Reached in 2007

Computing models from today based on too pessimistic network evolution

- Strict hierarchical model: Rather static data distribution T0->T1->T2
- Job execution always local to data storage (LAN access)

Today's networks can support different approaches

- Mesh-like data distribution establish rather early in CMS
- First experiences with WAN access to data

Much more dynamic Computing model in preparation
Wide Area Data Access

1. Attempt local file open
2. Ask a redirector, if 1.) fails

- Join site Storage Elements into Storage Federations
  - Almost 40 Tier-2 sites publish their data into a federation
- Want to have access to all CMS data remotely
  (Any data Anytime Anywhere)

More than 40 Tier-2s have configured this “Fallback”
Another ~10 still need to do it
Disk Tape Separation

Presently: Disk-Tape coupling at Tier-1

- Uncontrolled access to files on tape forbids user analysis at Tier-1
- Produced files go to tape “immediately”
  - Cannot use site without writing to tape → inefficient and inflexible

Solution: Separate disk and tape

- Large disk pool and small tape read/write pool
- Staging and flushing from/to tape is a subscription in CMS data management tool
- Files on disk can be read via WAN xrootd (never trigger tape operations)
- Tier-1s can be open for 'chaotic' analysis jobs
- Intermediate data products can be easily deleted
- Production and (tape) archiving location now independent
HLT Cloud

► High Level Trigger Farm
  ▪ Sizable resource available during LHC “no beam” time
    ► Shut downs (several months)
    ► Technical stops (time scale week(s))
    ► Between LHC fills (time scale hours)
  ▪ Installed close to CMS detector at LHC point 5

► CPU Resources
  ▪ More then 1,200 Machines with over 13k cores
  ▪ Sums up to about 195k HS06 [compared to 120k for Tier-0 and 150k for Tier-1]

► Good network connectivity from CMS cavern to CERN Computer Center
  ▪ Presently 2x10Gbit

► Setup based on OpenStack Cloud tool
  ▪ Deployment of VMs
HLT Cloud: First Experiences

Running jobs
81 Days from Week 21 of 2013 to Week 32 of 2013

- Data accessed over “WAN” using xrootd
  - ~50kB/s per RECO job
- Could achieve 6000 concurrent jobs
- Bandwidth limitations under investigation
- Upgrade of HLT-CERN link to 100Gb/s
GOAL: Transparent access to classical Grid sites and Cloud resources

Recent GlideinWMS supports most use cases

- Integration with CMS tools required
Opportunistic Sites

- Main physics program needs to fit with pledged resources
  - Budget is tight – no room for 'extras'
- CMS might have access to additional not-pledged resources in an 'opportunistic' way
  - Grid sites mainly supporting other VOs – have them already “Tier-3s”
  - Super Computers after successful proposal
  - Test infrastructures for Clouds (or other technologies)
  - Commercial resources during off-working hours
- Gained 1st experiences in 2013
  - RE-RECO campaign at SDSC
  - Rather successful, but lot's of things done 'by hand'
  - More transparent integration in progress
Setup of Opportunistic Sites

➢ Software distribution for CMSSW and Grid client
  ▪ Access software via CVMFS
    ➢ Usually requires local installation/configuration by root used
    ➢ Solved by using Parrot

➢ Parrot
  ▪ Tool for attaching programs to remote I/O systems through the filesystem interface
  ▪ Provides CVMFS repositories in user space during run time of a job
  ▪ Only a small performance penalty

➢ Work load management
  ▪ Ongoing integration with existing GlideinWMS production infrastructure

➢ Squid serves for DB constants
  ▪ Evaluating “Public squids” accessed remotely
  ▪ Might require “on site squids” for larger opportunistic resources
Analysis Infrastructure

Present Tool – CRAB 2
- Typically 20k jobs per day
- Peaks to 40k jobs
- Mainly GlideinWMS submission
- Few gLite based submissions

Future Tool – CRAB3
- Supports different submission backends
  - GlideinWMS – similar to CMS production infrastructure
  - PANDA – based on ALTAS pilot factory
- Asynchronous stageout of user files
  - Write file to local site first
  - Transfer to “home site” of user by FTS
- Presently in integration phase
Summary

CMS Computing during 1st LHC run quite busy but successful

- Relevant data and MC production got delivered for discoveries
- Huge effort to maintain performance of CMS software with more demanding data taking conditions

Next LHC running period challenging for CMS computing

- Running conditions call for almost a factor of 10 in CPU capacity
- Several improvements underway to use realistically available resources much more efficient
  - Exploit multi-core capabilities of modern processors
  - Evaluate power efficient hardware architectures such as ARM
  - Strong wide area networks allow more dynamic data placement and job scheduling
  - Access temporarily available resources, e.g. HLT farm

LHC running resumes in beginning 2015

- A bit more then a year sounds much, but is ambitious given the amount of tasks