CMS Networking Requirements

WLCG Workshop – Manchester (UK)

Christoph Wissing (DESY) for CMS Offline & Computing

June 2017

HELMHOLTZ ASSOCIATION

DESY

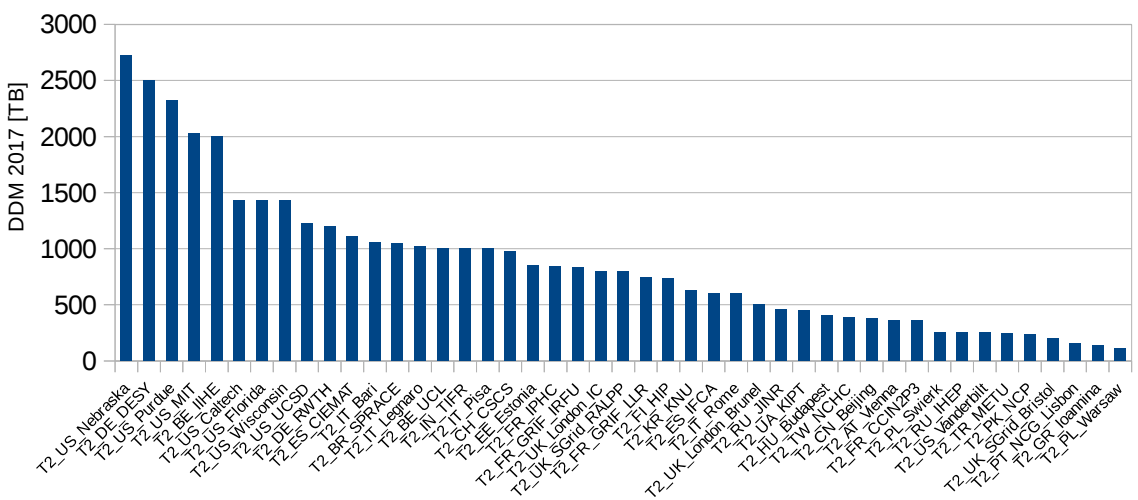
I/O Demands of CMS Workflow Types

Workflow Type	GEN-SIM	DIGI (ClassicPU)	DIGI (PreMixing)	RECO	Analysis
Typical Input Rate	negligible	5MB/s/core	0.5MB/s/core	0.1MB/s/core	~0.5MB/s/core

- GEN-SIM: Completely CPU limited
- DIGI
  - Classical mixing: Pile-up events (PU) mixed in as required
    - Very I/O demanding, particularly for high PU conditions
    - Flexible: Can produce 'any' PU distribution
  - PreMixing: Prepared library with already premixed pile-up events
    - Much lighter I/O requirements
    - Needs one storage intense library (O(100s TB) per PU scenario
- RECO: Moderate I/O demands
- Analysis: Can vary by orders of magnitude
- Output rate is typically very moderate O(0.05MB/s/core)

Tier-2 Network Infrastructure and Capacity

- Some observations from an incomplete survey among the CMS Tier-2 early 2017
- Typical US Tier-2
  - 100Gbit/s WAN connectivity, similar LAN connectivity
  - ~6-10k (HT) CPU cores, ~3PB storage
  - CMS-only sites allowing OSG VO opportunistically
- European & Asian Tier-2
  - 10-20Gbit/s WAN connectivity, similar LAN connectivity
    - Several have less
  - Large spread in size
  - Some (larger) sites are shared among bigger VOs



Remote Data Access

- All centrally managed workflows except Classical mixing can be executed with remote access
  - Typically 15% of core-hours are spent in production jobs with non-local data
- Fraction of Analysis jobs with remote data access varies between ~10% and 30%
- File opening sequence
  - Always attempt “local” open (for some special places “local” is not strictly local)
  - If local open failed attempt to open via Federation in Region
    - In France and Italy an open via a National Federation is attempted before the Region
    - UK and Spain also planning for a National Federation
  - If still failed attempt to open via Global Federation
- Presently CMS has two Regions: US region and Europe/Asia
- Xrootd client in CMSSW tries to distribute file opens according to measured speed
- Actual amount of remote reads at a site difficult to predict and to steer
- Scheduling prefers sites with having data locally
- On average 10% CPU efficiency decrease for remote read compared to local read

Requirements

- Network requirements scale (to first order) with CPU availability
  - 5MB/s/core would satisfy most demanding workflow – but appears not affordable for large sites
  - Commissioning for 1-2MB/s/core should be the target for LAN capacity
  - A few thousand remote connections of ~0.5MB/s can be expected (for sites with sizable storage)
- “Full” Tier-2: Many CPUs and large disk capacity
  - Some 10Gbit/s or 100Gbit/s for both LAN and WAN are advisable for sites with several 1000 cores
- “CPU-rich” Tier-2: Disk storage for caching only
  - CMS has no experiences with such a site (yet)
  - Mainly the same targets like for the “full” T2 apply
    - 1-2MB/s/core for LAN (reading from cache)
    - “Good” WAN connectivity (Some 10Gbit/s or more) for filling the cache or reading in (directly) from remote
- “Disk-rich” Tier-2: More storage than average, perhaps hosting disk for co-located CPU only site
  - Good WAN connection even more important
- “Disk-poor” Tier-2: Rather similar to “CPU-rich”

Disk-less Tier-2

- The “extreme”: T2_CH_CERN_HLT
  - (up to) ~12000 CPU cores, any data access from CMS EOS instance at CERN
  - 60 Gbit/s dedicated link between CMS cavern (at Cessy) and CERN computer center
  - Used routinely
- Do not know of firm plans to make a larger Tier-2 resource disk-less
- Have a number of disk-less resources and expect more
  - Cloud extensions
    - Well known HEPCloud extending FNAL
    - Recently used an Indian Tier-3 Cloud (loosely) coupled to the existing Tier-2 at TIFR
  - Examples of co-located CPU resources
    - Tier-3 in Omaha (co-located to T2 in Nebraska), HPC resources at CSCS (which is a T2), ...
  - CMS UK community working towards disk-less Tier-3s
    - Note: UK CMS Tier-3s are typically ATLAS Tier-2s
- Success relies on powerful connectivity

(Towards) IO Aware Scheduling

- Network and IO demand are not (yet) part of the CMS job scheduling
  - Managed to run IO-intense workflows into complete failure, when thousands were running at one site
    - Mitigation: Avoid certain types of workflows at some large sides
  - Heavy remote access reached (safe self-protecting) limits of storage systems at some sites
    - Difficult to mitigate: Run less workflows with remote access, create additional replicas
    - SE protection mechanisms employing better metrics than just number of TCP connections
- Submission Infrastructure team is investigating options with HTCondor developers
  - Dedicated high-IO slot(s) per multi-core pilot
    - In use already for IO-intense merging in the Tier-0 and Nebraska
  - Side wide limits for high-IO slots
  - Most challenging: Consideration of remote IO in scheduling

Network Monitoring

- Currently very limited usage of network monitoring service by CMS
  - Phedex: Routing decision based on internal success and throughput measurements
  - Limited development capacity in CMS to employ existing monitoring systems
- Wish list
  - Some feature (might) exist already
  - Capacity of a network link
  - Present utilization of a network link
  - Load reporting of storage systems
    - Experiences show limitations are often not the network work links, but the IO capability of a storage system
    - Could profit from consistent metrics across storage technologies

Summary

- LAN and WAN connectivity at the same level at many sites
  - US sites have typically 100Gbit/s
  - European sites (several) 10Gbit/s with a large spread
- CMS schedules typically 10-15% of jobs with remote data access
  - The “pentaly” in CPU efficiency is a drop of ~10% for remote access
  - Large gain in flexibility
- Computing model likely to evolve towards scenarios that require fast interconnets
- Need to improve CMS transfer system and scheduling system to better exploid network metrics