DATA MANAGEMENT UPGRADE IDEAS

Brainstorming

Ian Fisk

28 February 2014

Disclaimer

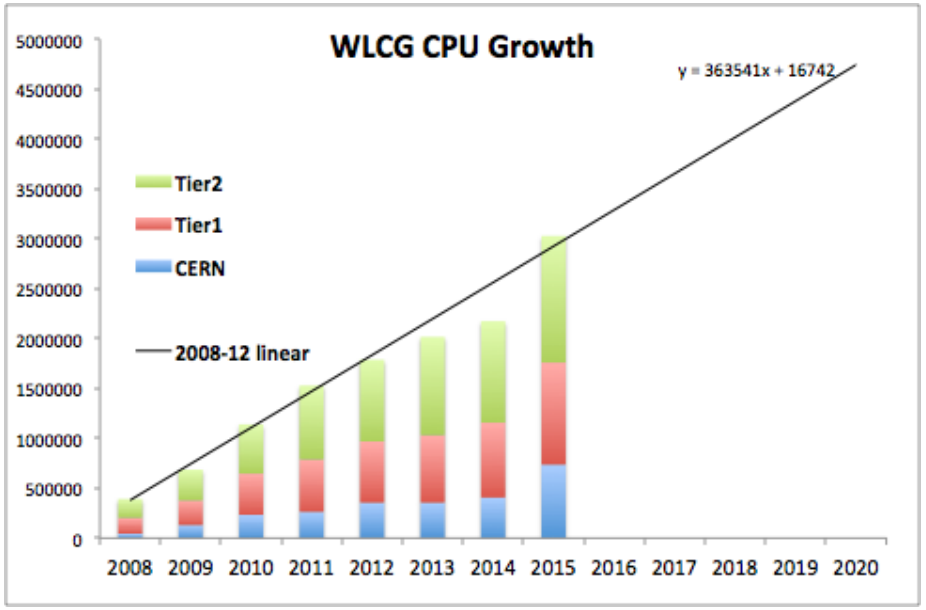
- What follows are a few personal reflections
  - They are not the choices of CMS or an indication of a program of work
  - Mostly things I thought were interesting

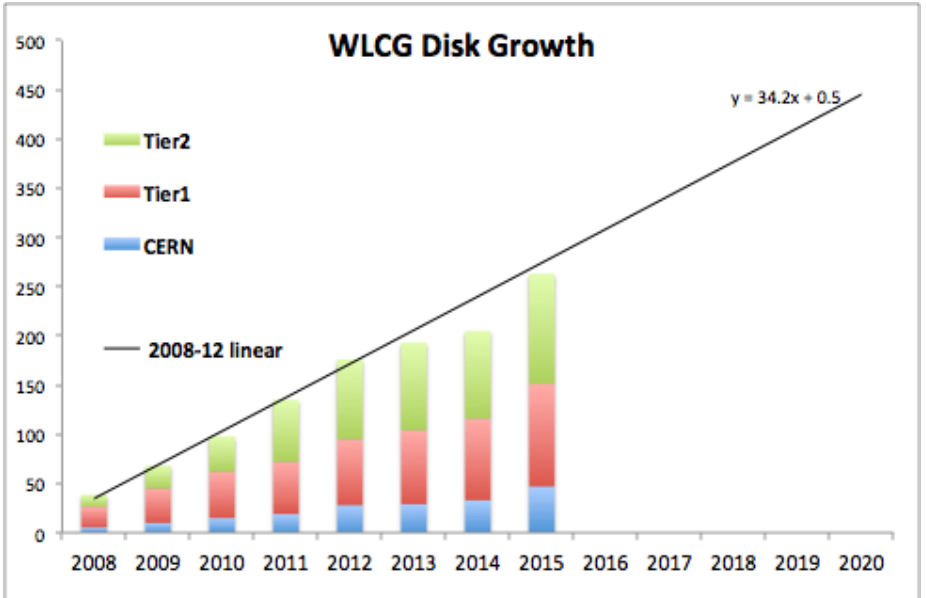
Scope of an upgrade

- When CMS speaks of upgrade we are normally speaking of after LS2 in Run3
  - Lines up well with the time scale proposed here
- Capabilities increase all the time and the computing model evolves, but people are expecting more disruptive changes at this time
- It’s a challenging environment
  - More than 100 interactions per crossing and a 10kHz of data collected
  - 50B events a year
    - A year of 2016 data every 3 weeks

Expected Growth

- Expected growth in computing doesn't really track to this load by our current extrapolation
  - Looking to 2020
    - CMS would have 150PB of disk
    - 300k CPU cores
      - Not an order of magnitude more





If we change nothing….

- 50B events is 25PB of RAW
  - ~15PB of AOD per reco pass
- Assuming the same for as now for MC
  - 1.25B Complete simulation events a month
  - 17PB of AOD MC
- Assuming the same storage placement, Tier-2s will have 60PB and 1 copy of the data will fit, plus user space and some physics controlled space
- Imagining 60s per event (very optimistic) a reprocessing pass using half the CPUs takes more than 6 months
- Indicates that we could be within factors of 2-3 of where we would need to be in terms of computing resources

Options

- For upgrades we could
  - Reduce the amount of data we need to deal with
  - Increase the number of resources we have (preferably with large resources we don’t pay for)
  - Try new techniques to boost our efficiency

REDUCING DATA

Tensions

- In every modern accelerator the volume of events you can collect is limited by your offline computing capabilities, not the physics desires
- The trigger is typically limited by how many events it can reject and not how many it can accept.
  - CMS DAQ could take 10kHz of data in 2015
- Bandwidth out is more than the processing capacity and we need to make hard choices about what can be kept long term

Not all events are created equal

- Currently every event we accept is treated in largely the same way
  - Most events are uninteresting background. And even when we know much more about one after detailed reconstruction and analysis, we still treat each event like it has the same physics potential.
- We make a decision in 100ms and live with that through long term data preservation
  - Reconstructing and carrying events along

Archiving

- We should be able to write data to tape
- Current estimates are $0.04 per GB for tape. We should be able to write ~600 CMS-like RAW events per $0.01. Tape charges for a nominal LHC year at 10kHz would be ~$800k
- Affording to reprocess and to analyze all the data is much harder. This is the active data sample

Defining the active dataset

- It is not obvious the active dataset needs to be defined by the online trigger where 100ms of thought was given
- As events are understood they may be
  - Eliminated from the active dataset
  - Immediately analyzed and not reprocessed
  - Put into background distributions
  - Kept in a reduced active dataset for further processing
- Progressively pair down the active dataset with understanding and time

Gains

- This is not a revolutionary idea, and it has some traction in the experiment since it's easy to understand
  - Also a natural evolution from the skimming and slimming all the experiments currently do
- Also very low risk. Original data is still on tape if there is a mistake, but defining an dynamic active dataset will give a lot of flexibility to the amount of data we collect initially without exploding the offline computing requirements

INCREASING RESOURCES

Increasing Resources

- CMS has been trying to increase the pool of resources we can use to Compute
  - Opportunistic is most attractive because it has a low cost and can be large
- We define opportunistic as any resource we don’t pay for
  - Cloud allowances, super computers, university clusters at night, HLT farms, etc.

Problem is Storage

- The problem is while opportunistic CPU is achievable and transient,
  - Storage is a longer term commitment, so to open up the most diversity of resources you need to solve the storage problem.
- How to deliver data to diverse processing resources when there is not local storage
  - Or only very transient cache

Wide Area Data Access & AAA

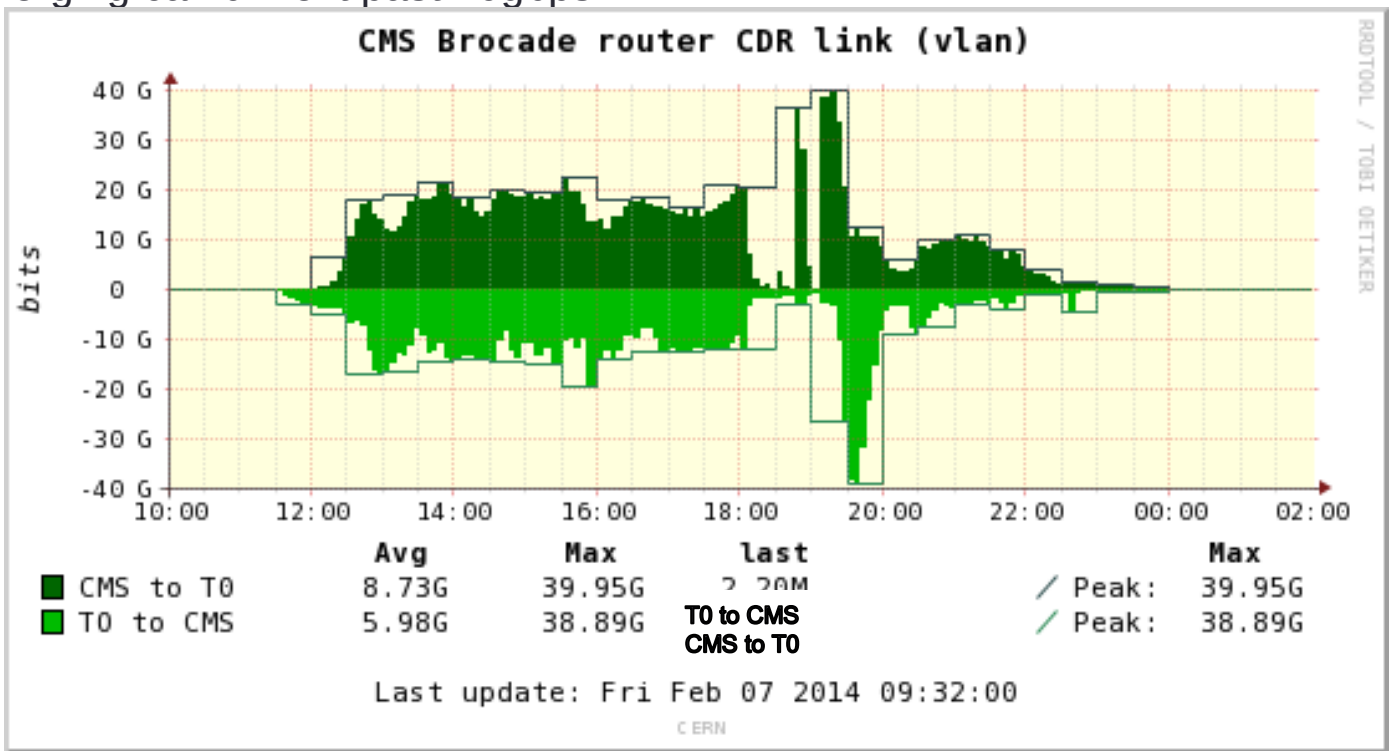
- US-CMS pioneered this effort, with the “Any data, anywhere, anytime (AAA)” project
  - Federation should allow the sharing of data serving to processing resources across sites

- By Summer 2014, CMS will complete the deployment and testing of the data federation in preparation for Run2
  - All Tier1s and 90% of Tier2s serving data
  - Nearly all files from data collected or derived in 2015 should be accessible interactively
  - Scale tests currently ongoing: file opening at 250Hz! Now moving to file reading scale tests

Goal: access 20% of data across wide area; 200k jobs/day, 60k files/day, O(100TB)/day. This is really modest and is really a first step

Proof of Concept

- The CMS HLT farm is 40% the size of the Tier-1s combined. We can built it up from an empty cloud to fully populated with VMs and pilots within 30 minutes
  - We can keep the farm busy with 20gbps for data reprocessing
    - Merging can drive it past 40gbps



T0 to CMS
CMS to T0

What we need

- Data federation is a good first step, but what we need is a data intensive scientific content delivery network
  - Delivering to super computers, clouds, clusters, people, etc

Improvements

- Content delivery networks are different from data federations
  - Intelligent and dynamic placement
  - Intelligent replication and pre-placement
  - Intelligent data selection and global balancing of capacity
- Many of the commercial solutions rely on more replication than we could ever afford, so we will need to do some smarter development

IMPROVING TECHNIQUES

Change the access

- Looking at the processing and analysis model used in CMS it would have a lot in common with the one when we used the first computers
  - Events are processed from files
  - Users make selections on the files and calculate their own quantities
  - Users make distributions
    - And maybe discoveries

Maybe time for big data tools

- Not all communities with large data volumes don’t access them this way
  - Another option is to use something like map reduce
  - Map is just a function that calculates something
  - A grouping function that groups recurring results
  - Reduce just summarizes how often that thing happened from the groups

Similar to us

- Map functions are designed to run on widely distributed systems and can be completely parallelized
- Grouping and Reduce functions group and store the statistics persistently
- We do similar things
  - We calculate some quantity on a high parallelized system, but each user does the grouping and reducing themselves
  - Potentially there is a big efficiency gain by keeping the statistics and sharing the output

Potential Benefits

- You need storage to keep the groups and results of the mapping functions, but you don’t need as many replicas of the data
- You need a lot of CPU to calculate and update the maps, but it is more structured IO scales more linearly
- Combining cuts is just taking subsets from groups
- Maps are validated and released code

Potential Architectures

- Large scale data reduction centers could be done as map reduce facilities
  - Chewing through potentially tens of petabytes of data in short period of time that could result in analysis samples to move to local clusters to perform the final steps
- Pushes to a model with large scale potentially distributed centers for a specialized task and analysis centers close to users

Outlook

- The needs for upgrade at least for Run3 are large and challenging but not so far from what we could achieve even with the current system
  - Probably some combination of reducing the load, increasing the resources, and improving the techniques are all needed
- Should be a place we can bring in new effort if we aim high enough