CMS Tier-1 risk assessment

D. Bonacorsi

[deputy CMS Computing coordinator - University of Bologna]

Credits to all CMS Computing Ops colleagues, special thanks to Chris Brew

Why a T1 risk analysis?

Some experience collected already as of data taking 2009:

- Already experienced a state with one T1 not-used for custodiality
- Already experienced data loss at a T1, several major outages (of different kind)

Goals of a **T1 risk analysis**

- Identify shortcomings in operations in case of one (or more) Tier-1 downtime(s)
- Layout rough operational plans according to different scenarios
- Agree on and document the action plans
- Communicate, (test), enforce them

Work done in January-February 2010, ready for 7 TeV data taking

- Collect feedback from CMS-internal (Ops teams) brainstorming sessions
- Merge and rationalize, discuss at the mgmt level, agree on some scenarios
- Work separately on each scenario
- Open the drafts, collect feedback from CMS Computing Ops sub-project leaders
- Implement comments, freeze the work-sheets, open to WLCG
- Schedule the updates to the work-sheets

How to face potential T1 incidents (1/2)

Incidents affecting T1's are two-fold:

- Temporary (or permanent) loss of data
- Temporary loss of services / no more usability of resources

The problem is not the data loss in itself, but either:

- The bandwidth / operational effort needed to retransfer the data
- The CPU needed to re-generate the data
- The procedures which need to be well thought and digested, aiming to:
  - Avoid discussion overhead in emergency-mode periods
  - Acceptable and not-inflating effort asked to Operations teams
  - Reasonable trust we will make it in an acceptable (and quantified) time

CMS-internal brainstorming sessions gave some confidence: computing ops people agree that in 2010 still a recovery could well be possible

- Just a matter of depicting the scenarios we want to be ready for…
- … and work to get prepared to face them

How to face potential T1 incidents (2/2)

Data safety and Data accessibility

- The Computing Model envisioned 1+1 copies of RAW data on tapes
- For data *safety* it should suffice. For data *accessibility*, it does if each T1 is 'reliable' i.e. abides by its MoU obligations
  - In case of a first T1 'off', do prepare for a second Tier-1 failing
  - E.g. if another T1 fails at the same time and hosts the same data we have to use the CERN 'cold' copy

The primary action is to carefully assess the situation *(see next)*

- DataOps and FacOps agree this should be done together
- Both agree that WLCG is crucial to help dealing with the problematic site
  - Make sure that WLCG and the affected T1 communicate directly on the technical level
  - CMS does not necessarily get involved - thus relieving the Ops team which can instead work on the experiment-specific actual implementation of the crisis plan

The details of a 'crisis situation'

The details of the crisis situation will drive the reaction plan:

- Which T1 is affected
- What is the damage, which critical services are affected
  - E.g. Part of or all WAN transfers, or regional problem, or tapes-only incident, ...
- Impact on CMS workflows running at that T1 in that data-taking period
  - E.g. How many PDs reside at that T1? Was that T1 starting/running a re-processing?
- Level and quality of information flow coming out of the affected T1
  - E.g. Promptness, completeness and preciseness of info from the T1 to the outside
- Estimated duration of the downtime
  - E.g. "Few days" != "Weeks" != "Unknown" *(see later)*
- Estimated impact of the incident
  - E.g. How quickly the accumulated volume for transfers/migrations can be estimated to grow? ...

What if a T1 (not FNAL) goes down

It is the relatively-easier case.

Real data:

- Some datasets custodiality be moved to FNAL or to another "*back-up T1*"
  - For some T1's, storage resources shortage may become an issue, though
- CPU may be a no-problem at any T1
  - the possibly needed processing (re-processing, skimming) may be either done by the back-up T1 - depending on urgency and duration of the problematic T1 down - or delayed - until the affected T1 comes back

MC data:

- The region connected to the affected T1 cannot upload MC for safe storage
- Short recovery time -> MC data transfer may just be delayed
  - Risk to loose data at T2 level, but lower probability if downtime window is narrow
- Longer recovery time -> MC data needs to be moved to a back-up T1
  - Unless we accept to be vulnerable and eventually pay a cost (re-production at T2 level)
  - Anyway, a cost has to be payed, in terms of T2-T1 non-regional link commissioning
    - In good shape, performances may be not adequate though (FTS tuning, STAR channels, …)

Estimated duration of the downtime

< 2-3 days:

- Do nothing special
  - Monitor closely
  - Get regular (e.g. twice-a-day) status updates from WLCG and/or the affected T1
- Prepare for longer outage

< 1-2 weeks:

- Too long, action needed: DataOps *quickly* needs a back-up T1 (still a CMS T1)
  - Creating a new workflow to be run at the back-up T1 is not a big overhead
  - Store new data (e.g. to always have enough RAW copies for good data accessibility)
  - Store new MC (to allow T2s to be safe and dynamic)
  - (if needed) Take the ownership of running prompt-skimming
- Also, prepare for a longer outage

> 2 weeks to 'unknown':

- Action needed: DataOps needs *immediately and stably* a back-up T1
  - (same tasks as above)
  - This back-up T1 will help to restore a datasets custodiality scheme at T1s

DataOps would favor not to attempt to roll back data placement after the problematic site is back in production

- Again: this depends on the specific emergency re-arrangement...

What if the affected T1 is FNAL

Much higher impact and may jeopardize sustained CMS operations

- A large fraction of the computing resources will not be available

So far, trying to have copies of data on both side of the Atlantic

- Each T1 hosting a non-custodial copy of the currently-not-accessible custodial data at FNAL gets this data 'promoted' to custodial
- In any case the data need to be moved back to FNAL when it comes back
  - since probably such a big processing cannot be done at other T1's in a timely manner

Other interesting suggestions:

- In case, use a set of T2's as a tapeless T1 for some period ?
  - A number of T2(/T3) have "special" access to 'their' T1. Think of RALPP, CCIN2P3 T2, FNAL T3. In a case where CPUs at the remaining T1's is the problem (rather than disk or tape) these resources could be used to beef up the T1.
- Not investigate further yet.

BTW: too pessimistic to think of loosing FNAL?

- In 2011, the loss of each T1 will have an effect similar to loosing FNAL in 2010

The 4 scenarios

Reference twiki:

- https://twiki.cern.ch/twiki/bin/viewauth/CMS/T1RiskAssessment
  - NOTE: it's now open to everyone who is in the wlcg-operations e-group

We worked on:

- SCENARIO 1: Data loss at a T1
- SCENARIO 2: Partial loss of a T1
- SCENARIO 3: Procurement failure at a T1
- SCENARIO 4: Extended T1 outage

More details in the pdf's attached to the twiki.