T1 risk assessment

Scenario 1 – “Data loss at a T1”

Scenario 2 – “Partial loss of a T1” (assumption[1]: storage outage)

Scenario 3 – “Procurement failure at a T1”

Scenario 4 – “Extended T1 outage”

PRELIMINARY – Daniele Bonacorsi – Last update: 22 Feb 2010

Actions:

1. Assess the incident details (no need to quantify the damage to CMS, yet). Talk to the affected T1 staff people. Be sure to have WLCG in the loop, and share the info and check you are in-sync. Acnowledge the emergency state, schedule ad-hoc communication flow among {Data,Fac}Ops and CMS representatives at WLCG daily calls.

2. Profit of the WLCG support and infrastructure to follow-up with the T1 staff for a prompt solution of the problem. CMS Ops should focus instead on the following bullets.

3. The central Transfer Operator(s) centrally suspend all transfers in progress.

4. Quantify the current damage to CMS. Identify the list of currently unavailable files/datasets, and the overall size.

5. If and only if the affected T1 is FNAL: (same as Scenario 1)

   ✦ 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 will 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.

   ✦ Depending on the details of the crisis assessment, the Computing Coordination will consider the option to use (a set of) T2's as a tapeless T1 for some period

6. Depending on the downtime duration, take different actions:

[1] This assumption is done based on the fact that a farm massive failure is somehow better depicted in Scenario 3. Additionally, the feedback from {Data,Fac}Ops clearly indicate a higher level of trust on the farming sector than on the storage sector. This feedback is briefly summarized as follows. The CPU’s may be a no-problem at any T1, since the possibly needed processing (re-processing, skimming) may be 1) either done by the back-up T1 - depending on urgency and duration of the problematic T1 down, or b) delayed - until the affected T1 comes back. But, T1 sites should be normally receiving more jobs and running more activities. The underutilization of T1 farms are a risk for resource provisioning processes at T1’s, and hence may influence the T1 capability to absorb peaks in a crisis situation at another T1.

- ✦ "**SHORT**, i.e.< **2-3 days**": do nothing special
  - Monitor closely
  - Get regular (e.g. twice-a-day) status updates from the affected T1
  - Also MC data upload from regional T2's to the affected T1 may just be delayed
  - Always: prepare for longer outage

- ✦ "**MEDIUM**, i.e. < **1-2 weeks**": DataOps needs a backup T1
  - Must be chosen among the CMS ones
  - Computing Coordination and {Data,Fac}Ops propose which one(s)
  - The relevant CMS T1 contacts give green light to:
    - Store new data (to always have the 1+2 RAW copies)
    - Store new MC (to allow T2s to be safe and dynamic)
    - (if needed) Take the ownership of running prompt-skimming
  - DataOps creates new workflows to be run at the back-up T1
  - Closely monitor the situation
  - Always: prepare for a longer outage

- ✦ "**LONG**, i.e. > **2 weeks to unknown**": DataOps **immediately** needs a backup T1
  - Must be chosen among the CMS ones
  - Computing Coordination and {Data,Fac}Ops propose which one(s)
  - The relevant CMS T1 contacts give green light to
    - Store new data (to always have the 1+2 RAW copies)
    - Store new MC (to allow T2s to be safe and dynamic)
    - (if needed) Take the ownership of running prompt-skimming
    - Not rolling back to the situation before the incident (i.e. this T1 will stably help to restore a datasets custodiality scheme at T1's)

7. In case of "MEDIUM" and "LONG" downtimes, and depending on the actual data-taking time in which the incident occurred, DataOps worries about 'old' data as well, and starts to discuss with Physics and assess priorities on what needs to be reproduced/reprocessed/retransferred at/to the eventual back-up T1.

8. In case you need to treat some data as "lost" to proceed, follow guidelines as from the "data loss" scenario:

   ➡ See the "Scenario 1" worksheet (bullets 9, 10, 11)

9. When the problem is fixed and the affected T1 site is back to operations:

   - ✦ The central Transfer Operator(s) runs a global `BlockDownloadVerify` to verify the overall status and accessibility of CMS data.
   - ✦ The affected T1 site could negotiate with the Computing Coordination to eventually roll back data placement (but consider DataOps would favour not to attempt it)