Service Incident Report for ASGC

Atlas T1, T2 and CMS T2 file loss due to XFS file system metadata crashed to one of our disk servers.

Description:

Since Feb 2013, we spotted that some of our DPM disk servers performed poorly, while disk servers are supposed to have around 1GB/s throughput they only have 600MB/s. So, we started to trace this issue, but, we didn't see anything wrong, so, we invited our hardware vendor to get involved in this investigation.

After we tried several analyses, we still have no clue. In Oct, our vendor insist to do I/O meter benchmark and also guaranteed that there is no impact on existing filesystem, so, we started with testbed and saw no problem there, then we put one disk server into DPM readonly mode and ran the benchmark program on it. The benchmark seemed fine, but it got I/O error immediately after the benchmark was done, and that caused the entire controller and disks frozen at that time.

We tried to recover storage when things went bad, but we found that the XFS file system metadata had been damaged, so that, we couldn't locate any file on the filesystem. After we tried several data rescue tools with vendor, we could only manage to recover a few files (about 0.07%) .

Impact

There were 1,007,444 files lost. They were part of T0D1 files which belong to Atlas calibdisk, datadisk, groupdisk, hotdisk, localgroupdisk, scratchdisk, and several files belong to CMS T2. There is huge impact on ATLAS production and user analysis.

Time line of the incident

- **15:12 28th Oct, 2013 UTC.** Doing I/O benchmark and got it finished. It''s only ran on one disk server . The process was enduring for 10 mins.

- **15:38 28th Oct, 2013 UTC.** We got alarm regarding read/write error to that disk server. Confirm that the disk directories couldn't be accessible. We immediately disabled all partitions on this disk (f-dpmp28.grid.sinica.edu.tw) and started the recovery process.

- **29th Oct, 2013 UTC.** Deploy another server and attached storage to that server, use dd to make image from original partitions in order to try some data rescue tools without bothering original source. But, after several trials, only xfs_irecover could work.

- **1st Nov, 2013 UTC.** Enable FAX to make Atlas job to be able to use xrootd redirector in order to reduce job failure rate. Confirm that the tool was functional and started to scan all partitions. Use checksum to identify the real filename between DPM database and recovered files.

- **5th Nov, 2013 UTC.** The recovery process had been finished. Recovery rate was only 0.07% , had confirmed 140TB data loss. List the lost files to DDM OPS and ATLAS users. Submit file invalidation request for CMS T2 lost files.

- **11th Nov, 2013 UTC.** CMS T2 file invalidation has been done.

Analysis & Improvement

- The vendor never does double check to confirm if this benchmark program will impact on file system with different circumstances (such as multipath, dual controllers..), and still, they don't think they are also responsible for this incident in any term, we still have some arguments with them...

- We didn't backup xfs journal regularly. In this case, we would only lose inodes, if we have journal backup, it should be able to increase the recovery rate. We will have regular xfs journal backup

afterward.

- The vendor will prepare another independent machine with the same spec as our production system. We will go ahead to trace the performance issues to improve our system efficient.