FNAL-USCMS Tier1 Service Incident Report

Description

Checksum errors were being reported for files that were being read from dCache disk instance. Further investigation revealed that many files were having this issue and all the files were coming from 2 dCache storage nodes, which had been deployed the previous week. These two storage nodes share the same Nexsan storage device, which is supposed to be configured so that the two servers each only see half of the LUNS that are available from the Nexsan. The Nexsan was configured incorrectly though so in reality the servers had the same two LUNS mounted and were writing into them simultaneously.

Impact

A total of 672 files were corrupted. Of those 417 did not have replicas in any other location and were permanently lost. The loss was spread across several individual datasets. All of those were Monte Carlo simulation and could be remade if needed. The largest impact to any individual datasets reached about 20% loss of statistics due to a low total number of files in those datasets. Although the impact to CMS appears minimal, FNAL in its response has treated this very seriously, as if unique precious data was lost.

Timeline

2015-04-14 10:15:57 Added dCache pools for new cmsstor401-410 to dCache disk instance pool servers, each having two dCache pools

2015-04-22 10:42:03 Incident INC000000535889 Opened by David Mason about 'Transfers FNAL Disk->Buffer fail due to wrong checksum'

2015-04-22 11:33:15 Initial analysis of one of the impacted files for the issue by local dCache admin pointing to the checksum verification mechanism

2015-04-22 14:29:10 Second analysis by local dCache admin points to local file corruption

2015-04-22 15:27:39 Service owners start investigating checksum mechanism on dCache pool

2015-04-22 15:41:50 Service owners confirm the file was properly transferred and corrupted afterwards on disk

2015-04-22 16:03:58 Service owners randomly checks two more files on the dCache pool where the analyzed corrupted file lives (w-cmsstor409-disk-disk2), both are

OK (CRC matches the database).

2015-04-22 16:29:24 Service owners analyze the 28 files that are failing transfers; they are all stored on dCache pools from servers cmsstor409 and cmsstor410

2015-04-22 17:11:04 Service owners set the 4 dCache pools on cmsstor409 and cmsstor410 as read only; Whole dCache pool checksum check started for all the FY15 dCache pools purchase (cmsstor4xx)

2015-04-22 18:10:55 dCache pool checksum verification show significant amount of mismatches on some pools; Issue is escalated

2015-04-22 19:54:08 List of potentially corrupted files is available

2015-04-22 20:59:24 A potential root cause is detected: the SAN backend is misconfigured and shows all devices to both servers with corrupted data (cmsstor409 and cmsstor410); This means that both servers are writing data to the same block devices, thus the corruption; All corrupted files are located in 4 pools that belong to two servers that share the same SAN backend.

2015-04-23 08:28:04 All checksum checks are done, final results confirm analysis from 2015-04-22 20:59:24.

2015-04-23 11:35:40 Migration tasks for the non-corrupted data on the affected dCache pools (cmsstor409 and cmsstor410) starts

2015-04-23 18:10:08 All migration tasks are done, all non-corrupted files copied to sane dCache pools

2015-04-24 13:35:53 Final list of corrupted files provided to CMS

2015-04-28 18:49:35 Corrupted files removed from the namespace (and the pool)

2015-04-29 09:53:10 cmsstor409 and cmsstor410 are completely removed from the production dCache and ready for reconfiguration and re-commissioning

Follow Up Actions

- The infrastructure services team that installs and configures the hardware is adding monitoring to ensure that no two nodes have the same UUIDs mounted. This check will be run via Check_MK regularly and will have alarms enabled.
- The infrastructure services team is also reviewing burn-in and acceptance procedures to see if they can be modified to include tests that would catch this type of configuration issue and other sorts of configuration mistakes.

- The distributed computing services team that supports the dCache service is also going to modify the procedures it uses when adding additional storage. The storage will be attached to a test stand initially and dCache writes/reads will be performed to ensure the software/hardware integration is working properly.
- In both teams we are evaluating the shared Nexsan configuration to see if a one server/one Nexsan configuration can handle the load.
- We have put a moratorium on storage installations until all these processes have been tested and implemented. The storage that was involved in this incident will be used for the tests and will not be re-deployed to production dCache until we have tested the new procedures.

Summary

In summary, files were corrupted on write due to the accidental misconfiguration of a newly deployed storage unit. Steps are now being taken to do more checks during installation to prevent misconfigurations from happening again, and also to ensure that if a misconfiguration like this happens again, it will be detected before the unit goes into production.