Root Cause Analysis for PBI000000001020:

on May, 2012

Report submitted Tim Doody April, 2012

CS-doc-4727

Summary of Incident

On November 28, 2011 the NIS master of the CDF worker nodes was switched by the Grid and Cloud Computing Department as part of a service migration to new hardware: "We will deploy the two new CDF servers fcdfsrv13 and fcdfsrv14 and replace servers fcdfsrv0, fcdfsrv5." This work was performed inside the change management process #CHG 3218. Impact was classified as moderate/limited.

Email was sent to members of the Fermilab Experiments Facilities Department, FEF, informing them of the YP/NIS change. There was no acknowledgement/confirmation of receipt of the said email from FEF back to Grid and Cloud Computing.

On December 7th the change request was marked as completed and closed with a rating of successful with the following email message:

> "fcdfsrv0 and fcdfsrv5 were retired as scheduled and fcdfsrv13 and fcdfsrv14 were deployed as scheduled, work was completed about 13:00 on 11/28/2011.
> We are holding this change open because the work of the migration of the rest of the worker nodes and gatekeepers continues incrementally with the next batch of activity slated to be on December 13, 2011."

On Dec. 12th the Grid and Cloud Computing Department notifies FEF again, via email, of the YP/NIS change and that updates to the work nodes and slave server. Again, there was no acknowledgement/confirmation by FEF that they received the email and are proceeding with necessary follow-up actions.

During December 2011 and January 2012 CDF worker nodes were re-installed in batches to correct OSG scratch space oversubscription and to collapse the two OSG clusters into one. At the end of January, the last NIS slave server that was running with stale YP/NIS information since the November change was shutdown – the YP/NIS clients on the CDF worker nodes became dysfunctional and were rebooted at which point the YP/NIS clients on the CDF worker nodes stopped functioning.

People from FEF realized the issue quickly and Grid & Cloud people diagnosed it as out-of-date NIS configuration.

About 1100 running jobs of CDF were lost before the configuration was corrected. The running jobs that were lost are automatically restarted by the batch system. However, data handling jobs are unable to re-process files that were processed previously (by a lost job), thus causing incomplete results. Special recovery projects need to be made by the end user to recover the missing data.

*It should also be noted: that I was advised that NIS management of the CDF worker nodes has caused incidents previously.

List of related tickets:

- CHG 3218 (Change ticket of the November OSG head node replacement)
- RITM 15340 (Request ticket to re-install the last batch of worker nodes)
- INC 197630
- INC 197581

Background Concepts

The following points are useful in understanding the nature and impact of this problem:

- YP/NIS (Yellow Pages or Network Information Service) - information of the CDF worker nodes changes rarely.
- System management of worker/slave nodes including NIS clients and NIS slave servers is done by Fermilab Experiments Facilities Department, FEF.
- System management of the OSG head nodes and NIS master of the worker nodes is done by Grid and Cloud Computing Department.
- Both departments are under the Scientific Computing Facilities Quadrant.
- The analysis farm is used by the CDF collaboration (plus other OSG users opportunistically) and Running Experiments department is coordinating use and management of the farm in the Computing Sector.
- The CDF experiment was not in no-interrupt-conference-preparation.

Timeline

This is a partial timeline for the change, service disruption and restoration.

Monday, November 28th

15:54 an email from the Grid and Cloud Computing Department is sent to the Fermilab Experiments Facilities to inform about the NIS master change.

Monday, December 12, 2011

Another email from the Grid and Cloud Computing Department to the Fermilab Experiments Facilities is sent to inform about the NIS master change.

Friday, January 20th

15:14 request to Fermilab Experiments Facilities to re-install last batch of CDF worker nodes.

Monday, January 23rd

09:59 Fermilab Experiments Facilities starts draining jobs on last batch of CDF worker nodes.

Thursday, January 26th

09:33 Fermilab Experiments Facilities starts re-installation of last batch of CDF worker nodes.

10:16 network switch connecting about 150 CDF worker nodes fails. This causes the batch system to loose connection with jobs and them to be restarted later.

11:34 network switch is back up but emergency maintenance to replace failed component scheduled for 13:30.

12:40 Running Experiments black-lists worker nodes on failed switch to avoid new jobs from starting until after switch repair.

14:15 network switch is repaired. Some jobs are lost during the work and scheduled to be restarted later.

14:33 Fermilab Experiments Facilities checks worker nodes on failed switch and finds NIS master is unreachable.

14:37 Grid and Cloud Computing diagnoses binding to old NIS master.

15:37 Fermilab Experiments Facilities has updated NIS configuration on all CDF worker nodes for new master.

Analysis

The lack of comprehensive communication (before and after) between the two departments; Grid and Cloud Computing Department and the Fermilab Experiments Facilities Department led to necessary changes/updates to YP/NIS not being implemented on the CDF slaves and/or slave servers. This is main reason the CDF Farms became unusable and jobs were lost. Simple email messages advising of changes that were made and needed follow-up action were and are not sufficient.

Direct Cause

Because the worker nodes/slave servers were not updated to point to the two new CDF YP/NIS master servers - fcdfsrv13 and fcdfsrv14 after they were installed, the worker nodes continued to use their shared, cached, and out-dated YP/NIS information. The farm would continue to work until the last node in the farm with out-dated information was re-installed (wiping out cached information). Once the

outdated cached information was completely cleared, the CDF farms became dysfunctional.

Contributing Factors

A number of factors contributed to the failure of the CDF farms.

- The Change Request identified the possible impact of the change as *moderate/limited*. In hindsight, it was not.
- System management of the OSG head nodes and NIS master of the worker nodes, and system management of worker/slave nodes including NIS clients and NIS slave servers are done by two separate departments increasing the importance of coordination and communications with additional follow-up.

Root Cause

A root cause of the CDF farms going down was that the worker nodes and slave servers were running in December and January with mis-configured YP/NIS information. The only reason the farms continued to run in those two months was that the worker nodes and slave servers had cached information that was shared.

Once the last node(s) in the farm with the incorrect old cached YP/NIS information were reinstalled and the cache cleared, the incorrect master server configuration became an issue as the systems tried to get updated information from systems that were no longer there. The CDF farms became dysfunctional.

Observations & Comments

It appears that the Grid & Cloud Computing Department and the Fermilab Experiment Facilities Department have a good working relationship. They are comfortable working with each other. Sending emails between the two departments to coordinate tasks is common.

Recommendations

Based on the Problem investigation and root cause analysis, the following list of recommendations are made for preventing future occurrences of CDF farms becoming unusable.

1. The two Departments use Service Now ticket/request when requesting required specific tasks be done by another department.
2. When making changes to any services that require follow on tasks to be done by another group/department, the other group/department should be given advance warning of the upcoming change and possibly should participate in the preparation of the Change Request. Participation and communication between all participating groups are required for change(s) to be completed successfully.

3. I am not a Linux expert, but could an alias of some kind be used on the YP/NIS master servers so the changes in physical hardware to not require updates to workers and slave servers?

Root Cause Analysis Committee

The following people served on the RCA committee (?? February 2012):

Tim Doody (lead)

Stu Fuess (previous CDF worker node NIS RCA)

Michael Kaiser (Change and Release Manager)

Mike Kirby, CDF experiment

Tanya Levshine, Grid and Cloud Computing

Ed Simmonds, Fermilab Experiments Facilities