Incident Report – Accidental Restarting of Disk Array 10/11/11

Neil Brown, Craig Strachan – Version 1.0 22/11/11

Overview

On 10th November 2011, the Services Unit scheduled downtime to power down the disk array kbevo1 located at Kings Buildings, attach a new JBOD and perform various other tasks such as updating the firmware on the array. Operator error resulted in the array located in the forum, ifevo1 being shut down instead while still in service. Although ifevo1 was quickly brought back on-line, the need to ensure that no corruption of the data on ifevo1 had occurred meant that the data stored of this array was unavailable for up to 15 hours.

Sequence of events

As part of ongoing work in the JCMB server room, IS wished to repatch the network connection between KB and the central area resulting in a short loss of network connectivity between KB and the central area. The Services Unit decided to use this period of down time to attach a new JBOD to the disk array kbevo1, a task requiring the shutting down of the various servers (mostly AFS file servers) attached to the array, followed by the array itself. It was anticipated that this work would take no more than an hour. The Inf unit also decided to reboot some network equipment and the server hati which has the dual role of slave KDC and AFS database server at this time. Since very few RW volumes are located on kbevo1, it was anticipated that the disruption to users would be minimal. A weeks warning of the intended work was circulated to the School's users.

At 7:50am on the 10th of November, the process of shutting down the servers attached to kbevo1 began. This was completed by 8:10 whereupon the process of powering down kbevo1 began. This is accomplished by connecting to the array's internal web server, navigating to the appropriate page and shutting down the array's two controllers. Unfortunately, the address for ifevo1 was entered into the web browser in error this array's controllers shut down instead. Note that no special warning or error message is generated if an array of this type is shut down while actively serving data. The new JBOD was attached to kbevo1 and the array powered back on. The array came back on-line at 8:25.

At this point, conversation with colleagues in the central area revealed that the downtime was having a far greater effect than had been anticipated. Logging into machines took far longer than normal or failed completely and even when logged in, there were problems accessing the AFS file system. At first this was put down to a combination of the effects of the KB slave KDC being down, the possibility that some machines might have been using the KB AFS database machine and were taking some time to switch over (this behaviour had been observed previously) and that some machines might have been making use of the AFS file infrastructure volumes stored on the KB servers. Work therefore focused on bringing the KB servers back up in the expectation that this would resolve the issues.

At this point, another problem became apparent, namely that when booting, the KB servers were getting stuck when running updaterpms. Since the network connection with the central area had been restored, there was no obvious reason why this should be the case. In the belief that all the issues being observed would be resolved once the halted hosts at KB were back up, it was decided to alter the reboot precess of these machines to skip updaterpms. This was done and all halted hosts

were back up by 9:51. This did not resolve the issues being seen.

Since 9:30, nagios had been sending out messages about a fibre channel path being down. In the rush to get services restored as quickly as possible, these had been assumed to apply to the KB servers and disk array but now on closer inspection, they proved to be associated with one of the Forum arrays, ifevo1. A check of ifevo1 showed that both controllers had been shut down at 8:10, the time at which the KB array had supposedly been shut down. Examination of the logs of the servers connected to ifevo1, squonk, crocotta, bunyip and cameleopard, showed that they contained messages about the fibre channel connections failing but that the affected file systems had not been remounted read-only. Ifevo1 contains only AFS data and so attempts were made to shutdown the AFS file server process on the servers. These attempts hung, clearly attempting to complete I/O to the array. Given a choice between shutting the entire server done leaving this I/O incomplete or restarting the array and possibly allowing the I/O to complete successfully, the latter seemed preferable and so the controllers on the array were restarted via the web interface. When this was done, the AFS file server processes on the servers terminated normally

Recovery

The affected array contained 16TB of AFS file space, a mixture of user and group data, all of it RW volumes. After discussion it was felt that although there was a possibility that the multipath daemon might have cached the pending I/O until access to the array was restored thus preventing corruption of the file system, the only prudent course of action was check each affected file system using fsck.

Since all of the affected servers were also serving data from other disk arrays, the ifevo1 partitions were unmounted on each server and the AFS file server process restarted. Work then began on checking the partitions on ifevo1. To avoid overloading the array, only one partition per server was checked at a time. The -f flag was used with fsck to force a full check. As partitions passed the fsck check, they were remounted on the server and the file server process on the server restarted, thus allowing users to access their data as soon as it became available. All user data had been restored by 7:50pm and all group data by 11:00pm. Of the 45 partitions checked, only one showed definite signs of corruption (which fsck fixed) though in several other cases fsck optimised the directory structure of the partition. This may have been a sign of underlying problems or it may have been coincidence.

Communication with users

An initial message to sys-announce announcing “wide spread problems with the computing systems” was sent out at 09:36am. At around this time, the Head of School was informed of the situation, a flip chart announcing the problems was set up at the main entrance of the Forum and computing staff started touring the Forum informing users of the problem. The admin staff on each floor were also recruited to help spread the word. An article giving more details of the issue was put on the blog at around 02:30pm. Users were updated on progress made via updates to this article, further postings to sys-announce and updates to the flip chart. Due to an oversight, the status page on the DICE web site was not updated to advertise the problem until after 12:15pm.

Lessons to be learned and points to consider

This document was discussed at the Computing Operation Meeting 23/11/2011. Notes, comments, points made at that meeting are marked up

thus.

1. Apart from the delay in updating the system information page, a good job was done of informing users of what was happening. It would perhaps be preferable to have a formal checklist of steps to be taken when dealing with an incident of this nature.

The checklist is a good idea. USU should drive. **Action** USU

Some (Alastair, Alison) couldn't update the web alert message because they didn't have a home directory. Having the page in e.g. Drupal would help (assuming it wasn't affected by whatever incident is occurring)

People could (if they know where) update the file on the live filesystem of www.inf. However, it currently has networked home directories, so this should be changed to localhomes. **Action** Services

Everyone should remember to fill out the planned maintenance web page http://www.inf.ed.ac.uk/systems/support/interruptions.html

Still some confusion over whether the sys-announce@inf list is catching all the people we think it should. Should all active accounts ie "the passwd file" be the same list of users as sys-announce? At the moment there are a lot more people in the passwd file, as we're not doing automatic account deletions - yet.

It is assumed that staff that have "left" drop off sys-announce immediately and aren't given the same grace period on that list, as their actual account has. This should be checked at rectified **Action** RAT

2. This incident had a major impact on users because it affected their home directories. We should perhaps revisit the possibility of allowing users to have their home directories on their local machines if they prefer.

Are local home directories the thin end of wedge? People could store dot files locally and main data in AFS. Disconnected mode in AFS 1.6 could help. Various points, for and against, were revisited. See 24/2/2010 meeting for last time we (inconclusively) argued about this. We should at least record arguments, for and against, properly. **Action** take to CEG

3. This loss of the jabber server during this incident handicapped efforts to identify and resolve the problem. We should make efforts to determine why the jabber service became unavailable and make it more robust. In addition, the provision of a text based jabber client would have circumvented issues caused by X being unable to access home directories.

Processes on server were hung, Stephen had to kill -9 bash to get it going again. Possibly the last time the daemon was started it was done so manually by someone still in their (affected) home directory. **Action** Neil and Ian to see if they can replicate this scenario.

Should more servers be set up for computing staff with local home directories (but still with access to AFS). **Action** All units to consider which servers should have local home directories.

There is a terminal based jabber client. Graham uses it, it's called `finch`.

Also could people have not used X's failsafe mode?

4. The immediate cause of this incident was user error by a member of staff while working in the early hours of the morning in a less than ideal environment. We should look at ways of minimising the likelihood of errors such as this on being repeated. Possible ideas include:
    - Changing the user name and password used for accessing the disk arrays to something more closely related to each array

Action Services Unit will do this.

    - Not doing maintenance out of hours
    - Not attempting to do too much during the maintenance period
    - Scheduling maintenance at the end of the working day thus removing the imperative to have things working again by 9:00am

This doesn't suit (or is even possible) for some computing staff

    - Always having two people involved with one monitoring the activities of the other

Good idea for important stuff such as shutting down disk arrays

5. Would it have been safe to simply remount the partitions on ifevo1 and trust to the multipath daemon to have avoided corruption? It might be worth conducting some experimentation to investigate this.

Until we can investigate and work out what's going on, then we should always fsck. Though Simon suggests that ext3 filesystems should be fine to "just bring back".

6. The DICE desktop in the KB server room arizona needs to be configured to give computing staff a local home directory.

And luse in AT. **Action** USU to reconfigure server room desktops.

7. Less data would have been affected and the time time taken to restore this data diminished if our data was spread over a larger number of smaller disk arrays. However this would have serious cost implications.
8. The initial rebooting of the KB servers (and the consequent determination of the true cause of the problems) was delayed by the time taken for updaterpms to fail when it couldn't access the package directories. MPU is currently investigating how this can be improved.

This would certainly be welcomed

9. Why did nagios not report the problems with the IF fibre until 09:30am, well over an hour after ifevo1 was shut down? This should be investigated.

Action Ian will investigate – and did see below

10. What should ssh's behaviour be when trying to access a host which doesn't have access to home directories? Some delay is to be expected but users (or at least computing staff) should still be able to gain access to the machine.

Simon observed : The logins hung because the file server (process) was still up, but their storage was down. The fileserver accepts the request and then blocks on I/O. The transport layer sees that the fileserver is running so never times out the request.

Logins hung as openssh does a krb5_kuserok check. This looks for various files in the user's home directory (including .k5login), to see if the user can login. Hence if the I/O is blocking for the reason above, ssh logins also block.

Is there anything that could be done at the ssh end to timeout these problems?

Other comments:

Is there some way for COs to do a quick check of the evo status? Maybe via SNMP?