Main page
Recent changes
Server admin log (Prod)
Server admin log (RelEng)
Deployments
SRE/Operations Help
Incident status

Cloud VPS & Toolforge
Cloud VPS documentation
Toolforge documentation
Request Cloud VPS project
Server admin log (Cloud VPS)

Tools
What links here
Related changes
Special pages
Permanent link
Page information
Cite this page

Print/export
Create a book
Download as PDF
Printable version

Toolforge webservices are in the final stages of migrating to the toolforge.org domain.
Please help us clean up older documentation referring to tools.wmflabs.org!

Incident documentation/20180628-dumps

< Incident documentation

https://dumps.wikimedia.org and related NFS content were offline or in a degraded state for several hours

Summary

During an upgrade of the storage capacity on the dumps NFS/web servers (labstore1006 and labstore1007), six disks were marked offline by the controller, causing the volume to mount as read-only. Actions taken to correct this required a temporary shutdown of the web service and fail over of the NFS services.

Timeline

- 14:40 or so: New shelves were plugged into labstore1006 and labstore1007 per Task T196651, these were subsequently shut down
- 14:48:42 errors show up in the nginx logs for labstore1007 but we do not know this:
  - `[crit] 3989#3989: *32764157 pread() "/srv/dumps/xmldatadumps/public/other/pagecounts-ez/merged/2013/2013-02/pagecounts-2013-02-05.bz2" failed (5: Input/output error) while sending response to client, client: XXXXX, server: dumps.wikimedia.org, request: "GET /other/pagecounts-ez/merged//2013/2013-02//pagecounts-2013-02-05.bz2 HTTP/1.0", host: "dumps.wikimedia.org"`
- 14:52 cron job spam shows we have a problem with rsyncs to the labstores, via messages such as

  `rsync: mkstemp "/srv/dumps/xmldatadumps/public/other/pageviews/2018/2018-06/.pageviews-20180628-130000.gz.vriD77" failed: Read-only file system (30)`
- 15:07 Icinga alert for labstore1007 for failed Puppet resources; these failures are real though we don't realize it.
- 15:10 Alert is triggered for degraded RAID on labstore1007 as both systems show 6 disks failed in the main RAID 10 block which serves the dumps Task T198407
  - At this point, the filesystem is mounted read-only, but there were no obvious signs this was so initially.
- 15:31 Chris says the new shelves have been disconnected.
- 15:31 various failures registered on labstore1006 in the logs (similar are available from labstore1007):
  - 15:31:37 `labstore1006 kernel: [9105061.802485] hpsa 0000:08:00.0: hpsa_update_device_info: LV failed, device will be skipped.`
  - 15:32:04 `labstore1006 kernel: [9105088.811707] Aborting journal on device dm-1-8.`
  - 15:41:03 `labstore1006 kernel: [9105628.421645] EXT4-fs (dm-1): previous I/O error to superblock detected`
  - 15:41:03 `labstore1006 kernel: [9105628.525987] EXT4-fs (dm-1): Remounting filesystem read-only`
- 15:32 Icinga alert for degraded RAID for labstore1006.
- 15:50 Because there is a network maintenance window coming up in ten minutes which will take the attention of the cloud team, the labstore boxes are put on hold.
- 16:00 Maintenance window for cloud; SRE services subteam meeting for other folks.

- 16:20 Examination of physical and logical volumes on labstore1006/7 shows that one pv is missing; the device (would be /dev/sdc) is undetected. However, data appears to still be served from these hosts, though we don't test that too closely, rather leave what seems to be served running.
- 17:20 or so: network maintenance window for cloud team ends
- 17:30 In an effort to bring labstore1007 back up, the group decided to reboot the server in hopes that pvscan would pick up the missing physical volume. Labstore1006 was left for later since Toolforge was reading dumps from it. Labstore1007 came back with the following message, which no one knew anything about, and then dropped to 'enter root for maintenance':
  - Slot 3 Port 1 : Smart Array P441 Controller - Configuration Required - 1779-Slot 3 Drive Array - Replacement drive(s) detected OR previously failed drive(s) now appear to be operational : Port: 1E, box:1, bay: 2 (SATA) bay: 4 (SATA) bay: 6 (SATA) bay: 8 (SATA) bay: 10 (SATA) bay: 12 (SATA) Logical drive(s) disabled due to possible data loss.Action: Resolve any issues that disabled drive. Restore data from backup if drive(s) replaced.
  - This is the point when https://dumps.wikimedia.org became unavailable.
- Faidon is pinged and brought up to speed
- 18:46 Bryan emails various lists to announce the dumps webserver outage.
- We verify with Chris that the cabling is back to how it was (yes); Faidon decides to cold reboot, as the previous reboot had been without power off.
- The same screen is presented; we decide to 'repair and accept data loss'.
- 18:46 The host comes back up and the file system mounts cleanly with all disks intact.
- Faidon sees complaints about no redundant cable path; we have Chris try adding a second cable for a loop; this fails.
- Chris discovers that the cabling for adding the second shelves was wrong, based on the docs Brooke found.
- 19:16 The cabling is set back the way it was and labstore1007 is rebooted and comes up clean.
- Labstore1006 is rebooted; we agree to 'repair' there too; it comes up clean.
- Chase moves nfs dumps service to labstore1007 so that the new shelf can be added to labstore1006 the following day.
- 20:00 Both servers and their storage are now back in operation, but labstore1006 is not providing any services.

Conclusions

- We don't have a lot of expertise re: HP controller configuration/error handling.
- Puppet resource warnings are 99% of the time false alarms, so we ignore them.
- There is no monitoring of dump web server errors of the sort that would indicate system problems. Only today did it become clear (many hours after service restoration) that files were not in fact all being served as normal, though many were served just fine (how the *&^%$ were they?).
- We discovered that the failover procedures on the labstore1006/7 systems could use a little work to ensure they are cleaner. More importantly, this underlined the need to always make changes to the systems one at a time so that we never have both in a failure mode at the same time.

Links to relevant documentation

Where is the documentation that someone responding to this alert should have (cookbook / runbook). If that documentation does not exist, there should be an action item to create it.

Actionables

- Improve processes around failover of the dumps cluster (phab:T198420)
- Better coordination for any work to labstore 1006/7, even work that should have no impact, so only one system at a time is touched
- Double-check that storage arrays are not turned on when they are newly connected to a host, and that cabling configurations are correct
- Document HP controller behavior in the case that it detects new disks without changes to the configuration of arrays/logical drives
- Document the procedure to recover from errors of the above sort
- Make sure at least one person from cloud and one SRE are available to handle these issues independently of other maintenance or meetings?

Category: Incident documentation

This page was last edited on 19 October 2018, at 12:06.