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

Incident documentation/20150217-LabsOutage

< Incident documentation

Contents [hide]

Summary

The disk controller on virt1005 failed at about 17:00 UTC. All instances on virt1005 failed -- this included instances vital to the operation of many labs projects, including Tools and Deployment-prep.

A reboot of virt1005 restored disk service, but in order to avoid future calamity all instances were migrated to virt1012. Most tools services were restored by 18:45 UTC, but the complete migration process took several more hours, prolonging the outage for some labs instances. All services were restored (on new hardware) by 23:00 UTC.

A complete list of affected instances is below.

Timeline

- [17:00] Shinken reports that several tools hosts are down. Nova shell commands on virt1000, specifically "$ nova show <hostid>" reveal that all the failed VMs are hosted on virt1005. Failed VMs include the Tools web proxy and the labs-wide web proxy, so many labs and tools users notice the outage all at once.
- [17:15] Andrew notices that basic commands like 'free' and 'df' are failing on virt1005, calls for help on IRC. Brandon immediately diagnosis this as a disk failure, looks at dmesg and determines that the problem is in the disk controller.
- [17:30] Andrew sends an outage notice to Labs-l which includes a list of all downed instances.
- [17:30] After a bit of nervous discussion, Brandon, Faidon and Andrew agree to risk a reboot. Andrew reboots virt1005 from the mgmt console, at which point virt1005 starts up without complaint.
- [17:40] At this point virt1005 is up but all VMs are in a 'shutoff' state. It's agreed that the safest option is to evacuate VMs from virt1005 in anticipation of another disk failure. Because the instances are already off, Andrew opts to cold-migrate the instances rather than boot them and attempt to evacuate running VMs.
- [17:45-18:00] Andrew retools the cold-migrate script to handle a bulk evacuation. In brief, those changes allowed the script to take a single argument, an instance ID. It also left the migrated instances in a shutoff state after transport. For posterity, here's what that retooling looked like:

```
root@virt1000:~# diff cold-migrate cold-migrate-virt1005
12,16c12,15
< INSTANCENAME=$1
< TOHOST=$2
< PROJECT=$3
< FROMHOST=`OS_TENANT_NAME=${PROJECT} nova show ${INSTANCENAME} | grep 'OS-EXT-SRV-ATTR:host' | awk '{ print $4 }'`
< INSTANCE=`OS_TENANT_NAME=${PROJECT} nova show ${INSTANCENAME} | grep ' id ' | awk '{ print $4 }'`
---
> INSTANCE=$1
> TOHOST='virt1012'
> FROMHOST=`nova show ${INSTANCE} | grep 'OS-EXT-SRV-ATTR:host' | awk '{ print $4 }'`
> INSTANCENAME=`nova show ${INSTANCE} | grep ' name ' | awk '{ print $4 }'`
41,43c40,42
< echo "Rebooting the instance"
< echo ""
< OS_TENANT_NAME=${PROJECT} nova reboot ${INSTANCENAME}
---
> #echo "Rebooting the instance"
> #echo ""
> #OS_TENANT_NAME=${PROJECT} nova reboot ${INSTANCENAME}
```

- [18:00] Before starting a bulk migration, a few critical instances were chosen for special treatment: tools-webproxy and tools-submit.
- [18:40] tools-webproxy and tools-submit are up and running on virt1012. Marc tidies up the tools grid and most tools are able to resume normal operation.
- [18:45-22:45] The remaining instances are migrated to virt1012 in a batch job. Andrew watches the migration and selectively starts each instance as it migrates, depending on its previous run state on virt1005. Again, Marc cleans up grid issues as tools exec nodes restart.
- [20:00] The labs web proxy finishes migration and restarts, thus ending the labs-wide outage.
- [23:15] Alex Monk notes that beta is still suffering and starts MySQL on deployment-db1 (PID file exists, but MySQL had not started). At this point Beta resumes more-or-less normal operations.

Conclusions

Due to the one-off nature of most labs projects, Labs will always be especially vulnerable to catastrophic hardware failures. Nonetheless, there are several things we could do to limit the effect of such disasters; some of these options are practical and some of them grandiose.

- Next time, triage failed instances: If greater care had been taken with the migration and restoration of the downed instances, this outage could have been much less painful. Specifically, if the labs web proxy and the deployment db server had moved to the head of the line, the outage would have been at least an hour shorter for most users.
- Replace old hardware: Virt1001-1009 are Cisco servers that are out of warranty. Replacing them with newer hardware may reduce hardware failures. Maybe.
- Web-proxies: it may be possible to maintain multiple web proxies (both for tools and for general labs use) such that there's redundancy if one of them is lost. There would need to be some sort of failover/load-balancer between them, the implementation details of which are beyond this author.
- Tools grid: It's already the case that the tools grid engine is fairly well distributed and resistant to failures. It may be possible to add an additional submit host in order to improve this resistance.
- Shared filesystem for instance storage: If instance storage wasn't local to the virt hosts, evacuation from a failed server would be trivial, and outages such as this would be drastically reduced.

Actionables

- Investigate strategies to reduce the SPOF nature of the web proxies. https://phabricator.wikimedia.org/T89995
- Continue ongoing efforts to harden the tool grid against individual VM failure.
- Consider replacing all Cisco servers with new hardware.

- Experiment with using Ceph as a backend for instance storage. (This has been on the long-term list for a while, but won't happen this quarter.)

Affected Instances

Here is a complete list of instances affected, as reported by "nova list --all-tenants --host virt1005" early in the outage:

```
| 684e5f6f-3fbf-42a1-a44e-2953448448fd | accounts-mwoauth | ACTIVE | - | Running
| public=10.68.17.44 |
| 6f147938-fe0c-4868-abd6-da7d0667c07f | bastion2 | ACTIVE | - | Running |
public=10.68.16.66, 208.80.155.153 |
| b141a571-5fb0-4c62-90b2-e76268511a6b | catalogcompiler | ACTIVE | - | Running
| public=10.68.16.24 |
| eded2579-f205-4d34-b935-d8fa0361c237 | cephticon3 | SHUTOFF | - | Shutdown |
public=10.68.16.132 |
| d33dba53-45b6-4525-943a-ad24b100cd48 | cg-puppetmaster | ACTIVE | - | Running
| public=10.68.17.133 |
| e97213bf-a56e-4d79-a6e3-508e6157d19c | cvresearch-web | ACTIVE | - | Running |
public=10.68.16.91 |
| f8d3db4c-6fd3-4ea7-9b80-ef1c50e4b0f8 | dannyb | ACTIVE | - | Running |
public=10.68.16.139 |
| d0da5ac8-34ca-43a9-b63d-64ee438c29cc | deployment-cxserver03 | ACTIVE | - |
Running | public=10.68.16.150 |
| aec2245c-e965-4f06-8496-a0d9abf519ee | deployment-db1 | ACTIVE | - | Running |
public=10.68.16.193 |
| 0294f3af-eba5-4ac8-9205-7c1aba8808d9 | deployment-logstash1 | ACTIVE | - |
Running | public=10.68.16.134 |
| 09bbebac-631d-4126-ad3e-48ef64e72eb8 | deployment-restbase02 | ACTIVE | - |
Running | public=10.68.16.234 |
| fcaa135e-3fce-4fae-afe0-ded789fe6f6a | deployment-sca01 | ACTIVE | - | Running
| public=10.68.17.54 |
| 07976e9b-6676-46d8-afc2-ee1cfd3a25b1 | dns-test-dzahn | ACTIVE | - | Running |
public=10.68.16.232 |
| 4c354f1e-c733-4e01-ab25-a7954e3d8d67 | dynamicproxy-gateway | ACTIVE | - |
Running | public=10.68.16.65, 208.80.155.156 |
| 5d3003d5-2d9a-4c88-80a5-b9a3db8415ea | etcd02 | ACTIVE | - | Running |
public=10.68.17.157 |
| 33da85da-eea6-4bf5-977c-5a75cc32215f | incident-test | ACTIVE | - | Running |
public=10.68.17.109 |
| b029ff1b-0643-42ed-9c3c-f8d112e98a0d | integration-slave1009 | ACTIVE | - |
Running | public=10.68.17.17 |
| 42cdf096-eedf-4f64-8995-ab9874976a5f | jawiki-echo | SHUTOFF | - | Shutdown |
public=10.68.16.174 |
| a25019a1-fa5f-4c5a-be57-ff975c2e4552 | legoktm | ACTIVE | - | Running |
public=10.68.17.84 |
| 3cc2200a-5432-4fe0-8db6-33ec35732b12 | map | ACTIVE | - | Running |
public=10.68.16.181 |
| e381861f-aa7e-4d7a-894e-c919e86c73e3 | mathoid2 | ACTIVE | - | Running |
public=10.68.16.194 |
| b7076a0a-008f-4282-88b2-c5c4ea0b3ace | megacron-two | ACTIVE | - | Running |
public=10.68.16.49, 208.80.155.149 |
| 799c226f-1377-410c-b590-3e07a748693e | mlp | SHUTOFF | - | Shutdown |
public=10.68.16.3 |
| d05d5ec9-48d1-42cc-96de-b40c485bed51 | mwui | ACTIVE | - | Running |
public=10.68.16.61 |
| 20853100-12c6-480d-9eb7-a3d9e1864280 | nemobis | ACTIVE | - | Running |
public=10.68.17.131 |
| 0153e94a-2f43-4b87-ac06-d6be88cb0c6c | opengrok-web | ERROR | - | Shutdown |
public=10.68.16.184 |
| d65b95f9-d28f-484c-a98f-7119aa085b89 | openid-wiki | SHUTOFF | - | Shutdown |
public=10.68.16.185 |
| 1e4668ee-d3aa-4ff7-9dbe-819267f6ec84 | otto-cass2 | ACTIVE | - | Running |
public=10.68.17.60 |
| 6dbee509-88fb-43ac-990c-f8e47f16d7e2 | pirsquared-dev | ACTIVE | - | Running |
public=10.68.17.57 |
| 9fcfcd5f-7467-43b3-bc54-bfd02e5fbfc3 | pubsubhubbub | ACTIVE | - | Running |
public=10.68.16.69 |
| 89040a4e-6eeb-466b-8a22-484fe37f2eb0 | rcstream | ACTIVE | - | Running |
public=10.68.17.114, 208.80.155.180 |
| 120cc401-ed7a-44c5-b905-2d0eae23b6af | tools-exec-03 | ACTIVE | - | Running |
```

```
public=10.68.16.32, 208.80.155.142 |
| 30b98f1d-1c5a-49c1-b800-f4c535addc12 | tools-exec-07 | ACTIVE | - | Running |
public=10.68.16.36, 208.80.155.146 |
| 5cd684db-d0a6-4241-a11f-daf4c1b2f717 | tools-exec-09 | ACTIVE | - | Running |
public=10.68.17.64, 208.80.155.152 |
| 523df61c-07f0-41ba-924d-e2b8e474b4d7 | tools-exec-cyberbot | ACTIVE | - |
Running | public=10.68.16.39 |
| 96c37c36-970b-4cc7-a7ba-d1ee90a225b5 | tools-submit | ACTIVE | - | Running |
public=10.68.17.1 |
| cdce426b-ef6f-47e7-96e4-bcb3647f4709 | tools-webgrid-04 | ACTIVE | - | Running
| public=10.68.17.174 |
| 79aeb31c-a1c1-41af-9e00-df2c7e248924 | tools-webgrid-tomcat | ACTIVE | - |
Running | public=10.68.16.29 |
| 8d92c507-d253-425d-b7f4-2af3678a39ae | tools-webproxy | ACTIVE | - | Running |
public=10.68.16.4, 208.80.155.131 |
| 31e8206d-fa5c-4e62-a805-8cfb7def1f46 | toolsbeta-puppetmaster3 | ACTIVE | - |
Running | public=10.68.16.92 |
| a0a49d0b-0fae-45bf-938b-7c942567fa8c | upload-wizard | SHUTOFF | - | Shutdown
| public=10.68.16.228 |
| 65f97389-7f55-40bd-b441-bae3c9b91737 | wdq-titan2 | ACTIVE | - | Running |
public=10.68.16.169 |
| 09733bff-d485-40d8-9a7f-4a3670509741 | wikidata-reports | SHUTOFF | - |
Shutdown | public=10.68.17.34 |
| 303fcd47-12dd-4c00-8961-13dd82d03ee7 | wikimetrics-staging1 | ACTIVE | - |
Running | public=10.68.16.77 |
```

Category: Incident documentation