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/20170718-JobQueue

< Incident documentation

Summary

Deploying a change to the Jobrunner service using Scap3 resulted in the wrongful starting of job runners in Codfw. This is a problem given Jobrunner is not (yet) an active-active service, and Codfw is currently the passive/read-only DC.

Jobrunners in Codfw were active for approximately 30 minutes.

Timeline

- 21:40 Krinkle attempts to deploy a change to mediawiki/services/jobrunner using Trebuchet from `tin:/srv/deployment/jobrunner/jobrunner`.
- 21:43 Command `git deploy sync` fails at the fetch stage ("0/44 minions completed fetch").
- 21:47 thcipriani confirmed nothing happened.
- 21:54 Krinkle attemts to deploy the change with Scap3 instead. Also from `tin:/srv/deployment/jobrunner/jobrunner`.
- 21:57 Command `scap deploy -v` failed. Sync and restart succeeded for jobrunner-canaries (mw1299.eqiad.wmnet, mw2247.codfw.wmnet), group1, and group2. But restart failed for a server in group3 (mw1260.eqiad.wmnet). There are 9 groups in total. Deployment paused at this point.
- 22:00 It seems this server was disabled intentionally.
- 21:02 Krinkle notices some of the servers where restart reportedly succeeded in group1 and group2, are in Codfw. Krinkle aborts deployment.
- 22:07 Jobrunner error spike in Grafana.
  Image: https://phabricator.wikimedia.org/F8795232 / View in Grafana
- 22:15 Start of Codfw jobrunner confirmed.
  Image: https://phabricator.wikimedia.org/F8795242 / View in Grafana
- 22:25 thcipriani manually stops the jobrunner service on Codfw nodes.
- 22:34 Krinkle starts rollback to ensure cluster is in a consistent state (re-deploy of previous version, with restarts disabled)
- 22:58 thcipriani manually re-restarts Eqiad jobrunners that were previously updated with the new code

Detailed log

Log of `#wikimedia-operations` on IRC:

- 21:43 <Krinkle> **!log Attempt to deploy** https://gerrit.wikimedia.org/r/349364 (mediawiki/services/jobrunner) failed.
- 21:44 <Krinkle> hashar: Command line output at https://phabricator.wikimedia.org/P5759
- 21:44 <hashar> Krinkle: jobrunner should be cleaned from the deployment server. It is no more deployed by Trebuchet but using scap
- 21:44 Krinkle assumed Trebuchet because Jobrunner documents it, and T129148 (Deploy jobrunner with scap3) was still open and has an unresolved subtask.
- 21:46 <Krinkle> thcipriani: What did (if anything) 'sync' do just now?
- 21:47 <thcipriani> I think it probably didn't do anything but I'm checking...

- 21:52 <thcipriani> Krinkle: it seems like nothing has changed, I don't see any new tags on any of the servers I spot-checked. It doesn't look like it fetched anything afaict.
- 21:51 <Krinkle> thcipriani: Walk me through the new workflow and I'll document it on Wikitech? (I still want to deploy this change)
- 21:53 <thcipriani> so the new process is: get the repo on tin the way it should look in `/srv/deployment/jobrunner/jobrunner`, and then run `scap deploy -v`
- 21:54 <logmsgbot> **!log krinkle@tin Started deploy [jobrunner**/jobrunner@5f6099f]: (no justification provided)
- 21:57 <Krinkle> third group failed to restart one server: `21:57:00 ['/usr/bin/scap', 'deploy-local', '-v', '--repo', 'jobrunner/jobrunner', '-g', 'default', 'promote', '--refresh-config'] on mw1260.eqiad.wmnet returned [70]: Failed to restart jobrunner.service: Unit jobrunner.service is masked.`
- 21:58 <RainbowSprinkles> It was set to masked on purpose, iirc
- 21:59 <mutante> masked would survive reboots, so really disabled
- 22:01 <thcipriani> IIRC there was something about how we didn't want to start one of the jobrunner services on a particular node...something like that.
- 22:02 <logmsgbot> !log krinkle@tin Finished deploy [jobrunner/jobrunner@5f6099f]: (no justification provided) (duration: 07m 58s)
- 22:02 <Krinkle> OK. I won't rollback in that case.
- 22:03 <Krinkle> Does this mean I just started jobrunners on the codfw servers?
- 22:05 <Krinkle> It looks like JobRunner is active on mw2161.codfw.wmnet (random spot check using `ps aux | grep`).
- 22:05 <Krinkle> No jobchron though, maybe that's fine?
- 22:05 <icinga-wm> **PROBLEM - Check systemd state on mw2153** is CRITICAL: CRITICAL - degraded: The system is operational but one or more units failed.
- 22:06 <Krinkle> thcipriani: Can we find out if jobrunner was already active on those codfw nodes?
- 22:07 <Krinkle> jobrunner error count in Grafana just went from 1K to 7K.
  Image: https://phabricator.wikimedia.org/F8795232.
  Dashboard: https://grafana.wikimedia.org/dashboard/db/job-queue-health?refresh=1m&orgId=1&from=1500412247458&to=1500420052382
- 22:09 <thcipriani> Krinkle: I just spot-checked mw2159 and it looks like it's still running the old version of the code...so what happens with scap is it'll stop deploying once it hits the failure limit (which is, I suppose, 1 in this instance) I think we need to redeploy with a higher failure limit to account for the service masking. This would explain why you only hit 1 server that had the problem: it was just the first one you hit. I think you've deployed 17 of 36 servers so far.
- 22:10 <Krinkle> Yes, it reached server 3 of group default3 and stopped when the 4th one in that group failed. But group default1 has 3 codfw servers and also 1 codfw server in group default2
- 22:12 <icinga-wm> **PROBLEM - Check systemd state on mw2247** is CRITICAL: CRITICAL - degraded: The system is operational but one or more units failed.
- 22:11 <Krinkle> Did these codfw wrongly start a jobrunner where previously they were not?
- 22:15 <thcipriani> well for mw2243 it seems that it was started 18 mins ago. "active (running) since Tue 2017-07-18 21:56:30 UTC;" 18min ago
- 22:15 <Krinkle> https://grafana.wikimedia.org/dashboard/file/server-board.json?refresh=1m&orgId=1&var-server=mw2243&var-network=bond0&from=1500410677549&to=1500420743298 / https://phabricator.wikimedia.org/F8795242
- 22:15 <Krinkle> `bash: salt: command not found`
- 22:18 [private message] <thcipriani> I think you could probably stop them with dsh using keyholder
- 22:22 [private message] <thcipriani> Failed to **stop jobrunner.service**: Access denied as mwdeploy
- 22:22 [private message] <Krinkle> Reaching out in `_sec`
- 22:25 [#mediawiki_security] <thcipriani> Krinkle: I think I got it
- 22:25 [#mediawiki_security] <thcipriani> seems to have done the trick: `SSH_AUTH_SOCK=/run/keyholder/proxy.sock dsh -f codfwlist -M -r ssh -o -oUser=mwdeploy -- sudo /usr/sbin/service jobrunner stop`
- 22:28 <thcipriani> Krinkle: if you want to move forward with the deploy, you can remove service_name from the scap.cfg and it will not try to restart any services.
- 22:28 <Krinkle> thcipriani: I'd prefer to revert for now.
- 22:32 <thcipriani> lemme try something first...
- 22:33 <thcipriani> Krinkle: yes, all looks correct in the current state

- 22:34 <logmsgbot> **!log krinkle@tin Started deploy [jobrunner**/jobrunner@5f6099f]: (no justification provided)
- 22:34 <Krinkle> thcipriani: "jobrunner/jobrunner: promote and restart_service stage(s): 100% (ok: 2; fail: 0; left: 0)"
- 22:34 <Krinkle> restart_service is mentioned but not running, or..
- 22:38 <thcipriani> Krinkle: definitely didn't restart/reload service
- 22:39 <Krinkle> thcipriani: OK. Minor bug/enhancement to avoid scaring log messages :)
- 22:39 <thcipriani> :)
- 22:42 <logmsgbot> !log krinkle@tin Finished deploy [jobrunner/jobrunner@5f6099f]: (no justification provided) (duration: 08m 18s)
- 22:44 <thcipriani> Krinkle: ok, do we need to restart jobrunner on the machines in eqiad that were restarted previously?
- 22:45 <Krinkle> thcipriani: Yes.
- ...
- 22:54 <Krinkle> Let's do the restarts first.
- 22:54 <thcipriani> ok, doing restarts
- 22:58 <thcipriani> **!log restarted jobrunner** on mw1299.eqiad.wmnet mw1168.eqiad.wmnet mw1164.eqiad.wmnet mw1305.eqiad.wmnet mw1304.eqiad.wmnet mw1301.eqiad.wmnet mw1259.eqiad.wmnet mw1166.eqiad.wmnet mw1300.eqiad.wmnet
- 22:58 <thcipriani> ^ Krinkle should be done
- 22:59 <Krinkle> thcipriani: Thanks

Conclusions

What weakness did we learn about and how can we address them?

- Trebuchet was not cleared out for Jobrunner.
  - git deploy doesn't check to see if a repo has ever been deployed with scap/should be deployed with scap
- Documentation was outdated for jobrunner, dsh, and salt
- Jobrunner should have been reverted back to Trebuchet, or at least not Scap3 left enabled without known issue T167104 resolved.
- ~~Salt no longer exists (since when?) with no documented alternative.~~

  This is not correct, Salt is in the process of being replaced as part of this quarter TechOps goals, see bug T164780 but is currently still working as it was before, no changes have been yet made to the production Salt infrastructure. –-Volans

  > I tried on tin and got `salt: command not found` > I see now that Salt is only available on the salt-master and root-only. This used to not be the same (long ago?), which his why using Salt to restart job runners is (was) part of the jobrunner deployment process at Jobrunner#Deployment. Given it is now root-only, Salt is effectively unavailable to deployers. – Krinkle

- DSH
  - Command on the Dsh page worked, but needed some tweaking. Actual command was `SSH_AUTH_SOCK=/run/keyholder/proxy.sock dsh -f codfwlist -M -r ssh -o -oUser=mwdeploy -- sudo /usr/sbin/service jobrunner stop`
  - Created a list of codfw jobrunners via `cat /etc/dsh/group/jobrunner /etc/dsh/group/jobrunner-canaries | grep -v eqiad > codfwlist`
  - the `mwdeploy` user does, thankfully, have permissions to restart jobrunner:

```
mwdeploy@mw2153:~$ sudo -l
Matching Defaults entries for mwdeploy on mw2153:
    env_reset, env_keep+=HOME,
secure_path=/usr/local/sbin\:/usr/local/bin\:/usr/sbin\:/usr/bin\:/sbin\:/bin


User mwdeploy may run the following commands on mw2153:
    (www-data, mwdeploy, l10nupdate) NOPASSWD: ALL
    (root) NOPASSWD: /usr/sbin/service hhvm restart
    (root) NOPASSWD: /usr/sbin/service apache2 start
    (root) NOPASSWD: /usr/sbin/service hhvm start
    (root) NOPASSWD: /usr/sbin/apache2ctl graceful-stop
    (mwdeploy) NOPASSWD: ALL
    (root) NOPASSWD: /usr/sbin/service jobrunner *
```

Actionables

- bug T129148 Disable Trebuchet for Jobrunner.
- bug T129148 Update Deployment documentation to Scap3 on wikitech:Jobrunner.
- bug T129148 **Important** Disable restart command for Jobrunner Scap3 deployments and document how to manually restart eqiad job runners.
- bug T167104 Figure out how to make Scap3 safely restart Jobrunner.
- scap: Don't log `restart_service stage(s): 100%` when no restart was issued.

Categories: Pages using deprecated source tags | Incident documentation