Incident documentation/20130419-Parsoid

< Incident documentation

From Roan's email to the Ops mailing list

We had the Parsoid LVS services showing as down in Icinga from 23:46 UTC on April 18th until 00:20 UTC on April 19th. I say "showing down" because the actual downtime was much shorter, but a separate issue caused Icinga to believe they were still down.

We deployed a large Parsoid update today. Because we were deploying config (really compiled code) and code changes that were incompatible with each other, we knew a short period of downtime would happen, and that's exactly what happened.

However, we neglected to notify ops about this, or to silence monitoring, so when Parsoid went down, it paged everyone. I noticed this because phones started going off around the office, but I was surprised that mine didn't, because I asked to receive Parsoid pages. After chatting with Daniel it turns out that he hasn't made Parsoid outage page me yet because he hasn't found a clean way to separate "people who should get all pages" from "people who should only get Parsoid pages" yet. The relevant ticket is https://rt.wikimedia.org/Ticket/Display.html?id=4318 . I did get Watchmouse pages 5 minutes after the fact, so that's working.

I'm very sorry for paging everyone, I should have realized that this would happen because as I said, brief downtime was expected behavior here. As I was told on IRC, I should have silenced monitoring before the deployment, or at least notified someone in ops so they could've told me I needed to silence monitoring. But I didn't realize I was about to cause pages, so I didn't think to do these things, and that's completely my fault, for which I apologize.

After being told this, I tried to actually go and silence monitoring for Parsoid LVS. I was kind of baffled by the UI first (as we'd migrated to Icinga recently), and I couldn't find the silencing feature because Icinga unhelpfully changed its icon from the crossed-out megaphone to a generic red X icon that's shared with six other actions. Leslie kindly came over to my desk to show me how to silence the checks, but then I got a permission error. Leslie found out I didn't actually have sufficient permissions to silence things, and fixed that in https://gerrit.wikimedia.org/r/#/c/59976/ .

Once Parsoid came back up, both pybal and Icinga kept reporting it was down. The backends were working fine though, and because of the depool limit in pybal some were still pooled and serving requests. It turned out there was a bug in the new Parsoid code that caused the connection to never be closed, but only if you requested / (the root). The monitoring (both pybal and Icinga) hit that URL, then timed out after 5s and reported failure, but our actual usage of Parsoid hit different URLs (for actual articles, e.g. /en/Main_Page) which worked fine. Through a combination of me trying the request locally on the box itself, me running tcpdump on the box, and Gabriel remembering a recent change related to Connection: Close, we figured out what the issue was and deployed a fix. Once we did that, Icinga and pybal instantly showed everything as up again. A few Parsoid backends are still down but those are expected, and I've acknowledged those in Nagios.

Of course /ideally/, we wouldn't have to have downtime even with config changes. The reason we have it now is because the config and the code are in separate repos, and git-deploy restarts the service after every deployment. That means that between the two deployments, the system is in an inconsistent and unstartable state, so it goes down for a few minutes until the second deployment brings it back up again. It would generally be nice not to restart things automatically, because that would also allow us to restart one box first, test the change there, then restart the others. I've filed a bug against git-deploy for this: https://bugzilla.wikimedia.org/show_bug.cgi?id=47393 .

Sorry for the pages!

Roan

---

fwiw, Roan did not even page all of ups, watchmouse just paged him and mailed the rest of us and we didn't add

him to Icinga paging yet. 21:33, 19 April 2013 (UTC)

Category: Incident documentation