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/20200611-sessionstore+kubernetes

< Incident documentation

document status: draft

Summary

The sessionstore service suffered an outage that lead to the inability of logged-in users to submit edits. The root cause of the outage was insufficient capacity to respond to a sudden increase of requests reaching mediawiki.

TODO: These are raw numbers, to be used as input to a calculation of actual user impact that hasn't been performed yet. More to come here.

Sessionstore impact: About 32 million requests were lost between 18:36 and 19:20. Based on the pre-incident steady state of about 85% 404s from sessionstore, about 4.8 million requests were lost that *would* have returned with 2xx status. (source). It's important to point out that 404s from sessionstore are from readers and hence are totally expected. TODO: Clarify why the error rate is normally so high, and whether it's the total or 2xx requests that reflect actual impact here.

MediaWiki-reported save failures: A total of 8,953 during the same period (all edit.failures.session_loss), which is unrealistically low, given that we believe *all* logged-in edits failed during the outage. (source)

Deficit in MediaWiki-reported edits: The trough in successful edits between 18:36 and 19:00 can be seen in MediaWiki-reported stats, as well as an increase from 19:00 to about 20:30, as users—presumably both humans and bots—retried their edits that failed during the outage. During the outage, the deficit is about 24,000 edits; for the total window from 18:36 to 20:30, the deficit is about 18,000 edits (that is, about 6k edits were effectively delayed rather than dropped). (source)

Editors were affected on all wikis in all geographic regions by being unable to edit, login or logout of the sites. Readers were completely unaffected.

Contents [hide]

Timeline

Due to the outage, stashbot was intermittently unable to edit the SAL. Every !log command from #wikimedia-operations is included here for completeness.

All times in UTC. TODO: formatting

May 28

- 14:14 Start of a backlog in change propagation causing CPU and memory issues, which in turn caused iowait CPU starvation as it attempted to do garbage collection. (graph)

June 11

- 18:36 first kask processes restart (https://grafana.wikimedia.org/d/000001590/sessionstore?orgId=1&from=1591900083237&to=1591903561732) **OUTAGE BEGINS**
- 18:41 first icinga CRITICAL (Prometheus jobs reduced availability for {eventgate_analytics_cluster,eventgate_analytics_external_cluster,eventgate_main_cluster,mathoid_cluster,sessionstore} in eqiad)
- 18:42 first paging alert (LVS socket timeout for sessionstore.svc.eqiad.wmnet)
- 18:42-18:45 Icinga alerts and recoveries for a number of unrelated kubernetes services, as well as host down alerts for kubernetes[1001,1003,1005]. (Full alert history under #Detection.) It was later found the kubernetes hosts never actually rebooted, but they did stop responding to pings.
- 18:44 T255179 filed, reporting logged-in users unable to edit wikis.
- 18:47:04 <icinga-wm> PROBLEM - MediaWiki edit session loss on graphite1004 is CRITICAL: CRITICAL: 60.00% of data above the critical threshold [50.0] https://wikitech.wikimedia.org/wiki/Application_servers https://grafana.wikimedia.org/dashboard/db/edit-count?panelId=13&fullscreen&orgId=1
- 18:47 Incident opened. rzl becomes IC.

- 18:49:52 <cdanis> all the kask pods [for sessionstore] are in CrashLoopBackOff
- 18:49:53 <shdubsh> kubernetes1005 hit the oom killer
- 18:52:11 <cdanis> kask @ sessionstore is OOM-looping and killing the whole machine
- 18:56:57 <akosiaris> let me increase sessionstore capacity to make sure we aren't going down under pressure
- 18:58:55 <+logmsgbot> !log akosiaris@cumin1001 conftool action : set/pooled=false; selector: name=eqiad,dnsdisc=sessionstore
- 18:59:21 <akosiaris> !log depool eqiad, switch to codfw
- 19:00:02 <+logmsgbot> !log akosiaris@deploy1001 helmfile [CODFW] Ran 'sync' command on namespace 'sessionstore' for release 'production' .
- 19:00:25 <akosiaris> !log increase sessionstore capacity in codfw from 4 pods to 6
- 19:02:13 <_joe_> so as of now everything should be migrated and sessions should still work, because of the fact sessionstore is multi-dc
- 19:04:55 <_joe_> akosiaris: restarts happening in codfw now, so not everything is ok
- 19:05:54 <cdanis> Warning FailedScheduling 21s (x10 over 5m40s) default-scheduler 0/6 nodes are available: 2 Insufficient cpu, 4 node(s) didn't match node selector.
- 19:07:05 <rzl> if we can't stabilize on sessionstore, we should consider switching back to redis even if it logs everyone out -- I'm not saying do it right now, but let's keep it on the table
- 19:07:20 <logmsgbot> !log akosiaris@deploy1001 helmfile [EQIAD] Ran 'sync' command on namespace 'sessionstore' for release 'production' .
- 19:07:34 <akosiaris> !log increase memory limits for sessionstore in eqiad to 400Mi from 300Mi
- 19:10:07 <logmsgbot> !log akosiaris@deploy1001 helmfile [EQIAD] Ran 'sync' command on namespace 'sessionstore' for release 'production' .
- 19:10:34 <akosiaris> !log remove the podaffinity restrictions for sessionstore in eqiad
- 19:12:42 <logmsgbot> !log akosiaris@cumin1001 conftool action : set/pooled=true; selector: name=eqiad,dnsdisc=sessionstore
- 19:12:52 <akosiaris> !log repool eqiad for sessionstore
- 19:13:46 <_joe_> so, it's interesting how one service (although the most called one) failing made kube-proxy suffer so much
- 19:14:14 <akosiaris> yes, that's my main question, how on earth did this affect machines that did not run those pods?
- 19:16:19 <akosiaris> https://grafana.wikimedia.org/d/000001590/sessionstore?panelId=47&fullscreen&orgId=1&var-dc=eqiad%20prometheus%2Fk8s&var-service=sessionstore&from=1591900236776&to=1591900788867 that's the initial reason: somehow session requests to session store increased from 15k to 20k
- 19:17:02 <_joe_> so we were quite under capacity with just 4 pods, and those VMs can't host more
- 19:17:10 <akosiaris> that probably pushed the pod pretty close to the memory limit and from there malloc() failures or whatever sent it spiralling down
- 19:18:50 <rzl> two broader k8s-related AIs -- one is capacity planning (plus maybe capacity alerting) for k8s services so we don't get taken by surprise like this, the other is "why was other k8s stuff affected"
- 19:20 Sessionstore traffic recovers to pre-incident levels. TODO: graph link+embed **OUTAGE ENDS**
- 19:20:33 <_joe_> akosiaris: should we do the same in codfw btw?
- 19:21:32 <_joe_> sessions are all back to eqiad, fwiw
- 19:24:18 "MediaWiki exceptions and fatals per minute" fires again; this is determined to be a delayed alert, due to logstash ingestion backlog (https://grafana.wikimedia.org/d/000000561/logstash?orgId=1&refresh=5m)
- https://grafana.wikimedia.org/d/000000561/logstash?panelId=21&fullscreen&orgId=1&from=1591898400000&to=1591916399000
- 19:30 akosiaris merges 604844, committing his emergency changes to git for both eqiad and codfw.
- 19:32:16 <logmsgbot> !log akosiaris@deploy1001 helmfile [CODFW] Ran 'sync' command on namespace 'sessionstore' for release 'production' .
- 19:33:06 <akosiaris> !log apply emergency sessionstore fixes in codfw as well
- 19:33:38 <akosiaris> 8 pods in codfw as well, at least on the trigger front I think we are ok now
- 19:36 _joe_ observes that something has been starving kubernetes[1001-1004] CPU with iowait since May 28 or 29: https://grafana.wikimedia.org/explore?orgId=1&left=%5B%221590019200000%22,%221591912800000%22,%22eqiad%20prometheus%2Fops%22,%7B%22expr%22:%22avg%20by%20(instance)%20(irate(node_cpu_seconds_total%7Bmode%3D%5C%22iowait%5C%22,%20instance%3D~%5C%22kubernetes100.*%5C%22%7D%5B5m%5D))%22%7D,%7B%22mode%22:%22Metrics%22%7D,%7B%22ui%22:%5Btrue,true,true,%22none%22%5D%7D%5D
- 19:59:08 <rzl> I'll keep monitoring for the rest of the day just to make sure logstash catches up, since that's our only remaining thing, but otherwise I'm considering the incident resolved
- 22:12 Logstash finishes catching up; incident fully resolved.

TODO: Clearly indicate when the user-visible outage began and ended.

Detection

Detection was automated, with the first IRC alert about five minutes after the kask pods initially crashed, and the first (and only) page about 30 seconds later. Icinga's full transcript from #wikimedia-operations is below, comprising 39 total alerts.

Note that the "MediaWiki exceptions and fatals per minute" alerts persisted, spuriously, for some hours after the underlying problem was solved. This was due to Logstash's delay in processing MediaWiki log entries: the alert reacts to the rate of log entry *ingestion*, not log entry *production*, so when Logstash is behind, the alert is behind too.

```
18:41:46 <icinga-wm> PROBLEM - Prometheus jobs reduced availability on icinga1001 is
CRITICAL: job=
{swagger_check_eventgate_analytics_cluster_eqiad,swagger_check_eventgate_analytics_extern
_cluster_eqiad,swagger_check_eventgate_main_cluster_eqiad,swagger_check_mathoid_cluster_e
ad,swagger_check_sessionstore_eqiad} site=eqiad
https://wikitech.wikimedia.org/wiki/Prometheus%23Prometheus_job_unavailable
https://grafana.wikimedia.org/d/NEJu05xZz/prometh
18:42:08 <icinga-wm> PROBLEM - PyBal backends health check on lvs1015 is CRITICAL:
PYBAL CRITICAL - CRITICAL - sessionstore_8081: Servers kubernetes1001.eqiad.wmnet,
kubernetes1003.eqiad.wmnet, kubernetes1005.eqiad.wmnet, kubernetes1006.eqiad.wmnet are
marked down but pooled https://wikitech.wikimedia.org/wiki/PyBal
18:42:19 <icinga-wm> PROBLEM - LVS sessionstore eqiad port 8081/tcp - Session store-
sessionstore.svc.eqiad.wmnet IPv4 #page on sessionstore.svc.eqiad.wmnet is CRITICAL:
CRITICAL - Socket timeout after 10 seconds
https://wikitech.wikimedia.org/wiki/LVS%23Diagnosing_problems
18:42:20 <icinga-wm> PROBLEM - Cxserver LVS eqiad on cxserver.svc.eqiad.wmnet is
CRITICAL: /v1/mt/{from}/{to}{/provider} (Machine translate an HTML fragment using
TestClient.) timed out before a response was received: / (root with wrong query param)
timed out before a response was received: /v1/dictionary/{word}/{from}/{to}{/provider}
(Fetch dictionary meaning without specifying a provider) timed out before a response
was received: /v2/suggest/source
18:42:20 <icinga-wm> ggest a source title to use for translation) timed out before a
response was received: /v1/list/pair/{from}/{to} (Get the tools between two language
pairs) timed out before a response was received: /_info/name (retrieve service name)
timed out before a response was received: /v1/page/{language}/{title}{/revision}
(Fetch enwiki protected page) timed out before a response was received
https://wikitech.wikimedia.org/wiki/CX
18:42:22 <icinga-wm> PROBLEM - restbase endpoints health on restbase1021 is CRITICAL:
/en.wikipedia.org/v1/feed/featured/{yyyy}/{mm}/{dd} (Retrieve aggregated feed content
for April 29, 2016) timed out before a response was received
https://wikitech.wikimedia.org/wiki/Services/Monitoring/restbase
18:42:26 <icinga-wm> PROBLEM - restbase endpoints health on restbase1027 is CRITICAL:
/en.wikipedia.org/v1/feed/featured/{yyyy}/{mm}/{dd} (Retrieve aggregated feed content
for April 29, 2016) timed out before a response was received
https://wikitech.wikimedia.org/wiki/Services/Monitoring/restbase
18:42:26 <icinga-wm> PROBLEM - restbase endpoints health on restbase-dev1005 is
CRITICAL: /en.wikipedia.org/v1/feed/featured/{yyyy}/{mm}/{dd} (Retrieve aggregated feed
content for April 29, 2016) timed out before a response was received
https://wikitech.wikimedia.org/wiki/Services/Monitoring/restbase
18:42:28 <icinga-wm> PROBLEM - PyBal backends health check on lvs1016 is CRITICAL:
PYBAL CRITICAL - CRITICAL - sessionstore_8081: Servers kubernetes1003.eqiad.wmnet,
kubernetes1004.eqiad.wmnet, kubernetes1005.eqiad.wmnet are marked down but pooled
https://wikitech.wikimedia.org/wiki/PyBal
18:42:28 <icinga-wm> PROBLEM - MediaWiki exceptions and fatals per minute on
icinga1001 is CRITICAL: cluster=logstash job=statsd_exporter level=ERROR site=eqiad
https://wikitech.wikimedia.org/wiki/Application_servers
https://grafana.wikimedia.org/d/000000438/mediawiki-alerts?
panelId=2&fullscreen&orgId=1&var-datasource=eqiad+prometheus/ops
18:42:38 <icinga-wm> PROBLEM - LVS wikifeeds eqiad port 8889/tcp - A node webservice
supporting featured wiki content feeds. termbox.svc.eqiad.wmnet IPv4 on
wikifeeds.svc.eqiad.wmnet is CRITICAL: CRITICAL - Socket timeout after 10 seconds
https://wikitech.wikimedia.org/wiki/LVS%23Diagnosing_problems
18:42:38 <icinga-wm> PROBLEM - Citoid LVS eqiad on citoid.svc.eqiad.wmnet is CRITICAL:
/api (bad URL) timed out before a response was received: /api (Zotero and citoid
alive) timed out before a response was received
https://wikitech.wikimedia.org/wiki/Citoid
18:42:42 <icinga-wm> PROBLEM - LVS echostore eqiad port 8082/tcp - Echo store-
echostore.svc.eqiad.wmnet IPv4 on echostore.svc.eqiad.wmnet is CRITICAL: CRITICAL -
Socket timeout after 10 seconds
https://wikitech.wikimedia.org/wiki/LVS%23Diagnosing_problems
18:42:44 <icinga-wm> PROBLEM - restbase endpoints health on restbase1018 is CRITICAL:
/en.wikipedia.org/v1/feed/featured/{yyyy}/{mm}/{dd} (Retrieve aggregated feed content
for April 29, 2016) timed out before a response was received
https://wikitech.wikimedia.org/wiki/Services/Monitoring/restbase
18:42:54 <icinga-wm> PROBLEM - eventgate-main LVS eqiad on eventgate-
main.svc.eqiad.wmnet is CRITICAL: / (root with no query params) timed out before a
response was received: / (root with wrong query param) timed out before a response was
received https://wikitech.wikimedia.org/wiki/Event_Platform/EventGate
18:42:56 <icinga-wm> PROBLEM - eventgate-logging-external LVS eqiad on eventgate-
```

```
logging-external.svc.eqiad.wmnet is CRITICAL: / (root with no query params) timed out
before a response was received: / (root with wrong query param) timed out before a
response was received: /robots.txt (robots.txt check) timed out before a response was
received https://wikitech.wikimedia.org/wiki/Event_Platform/EventGate
18:43:02 <icinga-wm> PROBLEM - restbase endpoints health on restbase-dev1004 is
CRITICAL: /en.wikipedia.org/v1/feed/featured/{yyyy}/{mm}/{dd} (Retrieve aggregated feed
content for April 29, 2016) timed out before a response was received
https://wikitech.wikimedia.org/wiki/Services/Monitoring/restbase
18:43:14 <icinga-wm> PROBLEM - Host kubernetes1003 is DOWN: PING CRITICAL - Packet
loss = 100%
18:43:18 <icinga-wm> PROBLEM - eventgate-analytics-external LVS eqiad on eventgate-
analytics-external.svc.eqiad.wmnet is CRITICAL: WARNING:urllib3.connectionpool:Retrying
(Retry(total=2, connect=None, read=None, redirect=None)) after connection broken by
ConnectTimeoutError(urllib3.connection.VerifiedHTTPSConnection object at
0x7efce6308518, Connection to eventgate-analytics-external.svc.eqiad.wmnet timed out.
(connect timeout=15)): /?spec https://wi
18:43:18 <icinga-wm> org/wiki/Event_Platform/EventGate
18:43:48 <icinga-wm> PROBLEM - Host kubernetes1005 is DOWN: PING CRITICAL - Packet
loss = 100%
18:43:56 <icinga-wm> PROBLEM - Host kubernetes1001 is DOWN: PING CRITICAL - Packet
loss = 100%
18:44:08 <icinga-wm> RECOVERY - restbase endpoints health on restbase1021 is OK: All
endpoints are healthy https://wikitech.wikimedia.org/wiki/Services/Monitoring/restbase
18:44:12 <icinga-wm> RECOVERY - restbase endpoints health on restbase1027 is OK: All
endpoints are healthy https://wikitech.wikimedia.org/wiki/Services/Monitoring/restbase
18:44:12 <icinga-wm> RECOVERY - restbase endpoints health on restbase-dev1005 is OK:
All endpoints are healthy
https://wikitech.wikimedia.org/wiki/Services/Monitoring/restbase
18:44:14 <icinga-wm> RECOVERY - Host kubernetes1003 is UP: PING WARNING - Packet loss
= 33%, RTA = 47.03 ms
18:44:18 <icinga-wm> RECOVERY - LVS wikifeeds eqiad port 8889/tcp - A node webservice
supporting featured wiki content feeds. termbox.svc.eqiad.wmnet IPv4 on
wikifeeds.svc.eqiad.wmnet is OK: HTTP OK: HTTP/1.1 200 OK - 945 bytes in 0.003 second
response time https://wikitech.wikimedia.org/wiki/LVS%23Diagnosing_problems
18:44:20 <icinga-wm> RECOVERY - Host kubernetes1005 is UP: PING OK - Packet loss = 0%,
RTA = 0.19 ms
18:44:20 <icinga-wm> RECOVERY - LVS echostore eqiad port 8082/tcp - Echo store-
echostore.svc.eqiad.wmnet IPv4 on echostore.svc.eqiad.wmnet is OK: HTTP OK: Status
line output matched 200 - 258 bytes in 0.016 second response time
https://wikitech.wikimedia.org/wiki/LVS%23Diagnosing_problems
18:44:22 <icinga-wm> RECOVERY - Citoid LVS eqiad on citoid.svc.eqiad.wmnet is OK: All
endpoints are healthy https://wikitech.wikimedia.org/wiki/Citoid
18:44:28 <icinga-wm> RECOVERY - restbase endpoints health on restbase1018 is OK: All
endpoints are healthy https://wikitech.wikimedia.org/wiki/Services/Monitoring/restbase
18:44:28 <icinga-wm> RECOVERY - eventgate-analytics-external LVS eqiad on eventgate-
analytics-external.svc.eqiad.wmnet is OK: All endpoints are healthy
https://wikitech.wikimedia.org/wiki/Event_Platform/EventGate
18:44:34 <icinga-wm> RECOVERY - Host kubernetes1001 is UP: PING OK - Packet loss = 0%,
RTA = 0.18 ms
18:44:34 <icinga-wm> RECOVERY - eventgate-main LVS eqiad on eventgate-
main.svc.eqiad.wmnet is OK: All endpoints are healthy
https://wikitech.wikimedia.org/wiki/Event_Platform/EventGate
18:44:38 <icinga-wm> RECOVERY - eventgate-logging-external LVS eqiad on eventgate-
logging-external.svc.eqiad.wmnet is OK: All endpoints are healthy
https://wikitech.wikimedia.org/wiki/Event_Platform/EventGate
18:44:44 <icinga-wm> RECOVERY - restbase endpoints health on restbase-dev1004 is OK:
All endpoints are healthy
https://wikitech.wikimedia.org/wiki/Services/Monitoring/restbase
18:45:04 <icinga-wm> PROBLEM - Host kubernetes1003 is DOWN: PING CRITICAL - Packet
loss = 100%
18:46:10 <icinga-wm> RECOVERY - Host kubernetes1003 is UP: PING OK - Packet loss = 0%,
RTA = 54.98 ms
18:47:04 <icinga-wm> PROBLEM - MediaWiki edit session loss on graphite1004 is CRITICAL:
CRITICAL: 60.00% of data above the critical threshold [50.0]
https://wikitech.wikimedia.org/wiki/Application_servers
https://grafana.wikimedia.org/dashboard/db/edit-count?panelId=13&fullscreen&orgId=1
18:47:28 <icinga-wm> PROBLEM - k8s API server requests latencies on argon is CRITICAL:
instance=10.64.32.133:6443 verb=LIST https://wikitech.wikimedia.org/wiki/Kubernetes
https://grafana.wikimedia.org/dashboard/db/kubernetes-api
18:49:18 <icinga-wm> RECOVERY - k8s API server requests latencies on argon is OK: All
metrics within thresholds. https://wikitech.wikimedia.org/wiki/Kubernetes
https://grafana.wikimedia.org/dashboard/db/kubernetes-api
18:52:10 <icinga-wm> PROBLEM - Host kubernetes1003 is DOWN: PING CRITICAL - Packet
loss = 100%
18:54:16 <icinga-wm> PROBLEM - Host kubernetes1001 is DOWN: PING CRITICAL - Packet
loss = 100%
18:54:46 <icinga-wm> PROBLEM - Host kubernetes1005 is DOWN: PING CRITICAL - Packet
loss = 100%
```

loss = 100%
18:55:06 <icinga-wm> RECOVERY - Host kubernetes1003 is UP: PING OK - Packet loss = 0%, RTA = 0.25 ms
18:55:06 <icinga-wm> RECOVERY - Host kubernetes1001 is UP: PING OK - Packet loss = 0%, RTA = 0.23 ms
18:55:14 <icinga-wm> PROBLEM - Cxserver LVS eqiad on cxserver.svc.eqiad.wmnet is CRITICAL: /v2/suggest/source/{title}/{to} (Suggest a source title to use for translation) timed out before a response was received https://wikitech.wikimedia.org/wiki/CX
18:55:16 <icinga-wm> PROBLEM - Too many messages in kafka logging-eqiad on icinga1001 is CRITICAL: cluster=misc exported_cluster=logging-eqiad group={logstash,logstash-codfw,logstash7-codfw,logstash7-eqiad} instance=kafkamon1001:9501 job=burrow partition={0,1,2,3,4,5} site=eqiad topic=udp_localhost-err https://wikitech.wikimedia.org/wiki/Logstash%23Kafka_consumer_lag https://grafana.wikimedia.org/d/000000484/kafka-consumer-lag?from=now-3h&to=now&orgI
18:55:16 <icinga-wm> e=eqiad+prometheus/ops&var-cluster=logging-eqiad&var-topic=All&var-consumer_group=All
18:55:18 <icinga-wm> PROBLEM - MediaWiki exceptions and fatals per minute on icinga1001 is CRITICAL: cluster=logstash job=statsd_exporter level=ERROR site=eqiad https://wikitech.wikimedia.org/wiki/Application_servers https://grafana.wikimedia.org/d/000000438/mediawiki-alerts?panelId=2&fullscreen&orgId=1&var-datasource=eqiad+prometheus/ops
18:55:20 <icinga-wm> RECOVERY - Host kubernetes1005 is UP: PING OK - Packet loss = 0%, RTA = 0.18 ms
18:58:34 <icinga-wm> PROBLEM - Host kubernetes1001 is DOWN: PING CRITICAL - Packet loss = 100%
18:58:38 <icinga-wm> PROBLEM - Host kubernetes1006 is DOWN: PING CRITICAL - Packet loss = 100%
18:58:48 <icinga-wm> RECOVERY - Cxserver LVS eqiad on cxserver.svc.eqiad.wmnet is OK: All endpoints are healthy https://wikitech.wikimedia.org/wiki/CX
18:59:02 <icinga-wm> RECOVERY - Host kubernetes1006 is UP: PING OK - Packet loss = 0%, RTA = 0.23 ms
18:59:04 <icinga-wm> RECOVERY - Host kubernetes1001 is UP: PING OK - Packet loss = 0%, RTA = 0.21 ms
19:00:48 <icinga-wm> RECOVERY - MediaWiki exceptions and fatals per minute on icinga1001 is OK: All metrics within thresholds. https://wikitech.wikimedia.org/wiki/Application_servers https://grafana.wikimedia.org/d/000000438/mediawiki-alerts?panelId=2&fullscreen&orgId=1&var-datasource=eqiad+prometheus/ops
19:02:18 <icinga-wm> RECOVERY - PyBal backends health check on lvs1015 is OK: PYBAL OK - All pools are healthy https://wikitech.wikimedia.org/wiki/PyBal
19:02:21 <icinga-wm> RECOVERY - LVS sessionstore eqiad port 8081/tcp - Session store-sessionstore.svc.eqiad.wmnet IPv4 #page on sessionstore.svc.eqiad.wmnet is OK: HTTP OK: Status line output matched 200 - 258 bytes in 0.012 second response time https://wikitech.wikimedia.org/wiki/LVS%23Diagnosing_problems
19:02:36 <icinga-wm> RECOVERY - PyBal backends health check on lvs1016 is OK: PYBAL OK - All pools are healthy https://wikitech.wikimedia.org/wiki/PyBal
19:03:44 <icinga-wm> RECOVERY - Prometheus jobs reduced availability on icinga1001 is OK: All metrics within thresholds. https://wikitech.wikimedia.org/wiki/Prometheus%23Prometheus_job_unavailable https://grafana.wikimedia.org/d/NEJu05xZz/prometheus-targets
19:04:26 <icinga-wm> PROBLEM - MediaWiki exceptions and fatals per minute on icinga1001 is CRITICAL: cluster=logstash job=statsd_exporter level=ERROR site=eqiad https://wikitech.wikimedia.org/wiki/Application_servers https://grafana.wikimedia.org/d/000000438/mediawiki-alerts?panelId=2&fullscreen&orgId=1&var-datasource=eqiad+prometheus/ops
19:06:02 <icinga-wm> PROBLEM - High average GET latency for mw requests on api_appserver in codfw on icinga1001 is CRITICAL: cluster=api_appserver code=200 handler=proxy:unix:/run/php/fpm-www.sock https://wikitech.wikimedia.org/wiki/Monitoring/Missing_notes_link https://grafana.wikimedia.org/d/RIA1lzDZk/application-servers-red-dashboard?panelId=9&fullscreen&orgId=1&from=now-3h&to=now&var-datasource=codfw+prometheus/ops&var-cluster=api_appserver&var-m
19:10:02 <icinga-wm> PROBLEM - High average GET latency for mw requests on appserver in codfw on icinga1001 is CRITICAL: cluster=appserver code=200 handler=proxy:unix:/run/php/fpm-www.sock https://wikitech.wikimedia.org/wiki/Monitoring/Missing_notes_link https://grafana.wikimedia.org/d/RIA1lzDZk/application-servers-red-dashboard?panelId=9&fullscreen&orgId=1&from=now-3h&to=now&var-datasource=codfw+prometheus/ops&var-cluster=appserver&var-method=GET
19:11:28 <icinga-wm> RECOVERY - High average GET latency for mw requests on api_appserver in codfw on icinga1001 is OK: All metrics within thresholds. https://wikitech.wikimedia.org/wiki/Monitoring/Missing_notes_link https://grafana.wikimedia.org/d/RIA1lzDZk/application-servers-red-dashboard?panelId=9&fullscreen&orgId=1&from=now-3h&to=now&var-datasource=codfw+prometheus/ops&var-cluster=api_appserver&var-method=GET
19:11:50 <icinga-wm> RECOVERY - High average GET latency for mw requests on appserver in codfw on icinga1001 is OK: All metrics within thresholds.

```
https://wikitech.wikimedia.org/wiki/Monitoring/Missing_notes_link
https://grafana.wikimedia.org/d/RIA1lzDZk/application-servers-red-dashboard?
panelId=9&fullscreen&orgId=1&from=now-3h&to=now&var-
datasource=codfw+prometheus/ops&var-cluster=appserver&var-method=GET
19:18:54 <icinga-wm> RECOVERY - MediaWiki exceptions and fatals per minute on
icinga1001 is OK: All metrics within thresholds.
https://wikitech.wikimedia.org/wiki/Application_servers
https://grafana.wikimedia.org/d/000000438/mediawiki-alerts?
panelId=2&fullscreen&orgId=1&var-datasource=eqiad+prometheus/ops
19:23:22 <icinga-wm> RECOVERY - MediaWiki edit session loss on graphite1004 is OK: OK:
Less than 30.00% above the threshold [10.0]
https://wikitech.wikimedia.org/wiki/Application_servers
https://grafana.wikimedia.org/dashboard/db/edit-count?panelId=13&fullscreen&orgId=1
19:24:18 <icinga-wm> PROBLEM - MediaWiki exceptions and fatals per minute on
icinga1001 is CRITICAL: cluster=logstash job=statsd_exporter level=ERROR site=eqiad
https://wikitech.wikimedia.org/wiki/Application_servers
https://grafana.wikimedia.org/d/000000438/mediawiki-alerts?
panelId=2&fullscreen&orgId=1&var-datasource=eqiad+prometheus/ops
19:42:26 <icinga-wm> RECOVERY - MediaWiki exceptions and fatals per minute on
icinga1001 is OK: All metrics within thresholds.
https://wikitech.wikimedia.org/wiki/Application_servers
https://grafana.wikimedia.org/d/000000438/mediawiki-alerts?
panelId=2&fullscreen&orgId=1&var-datasource=eqiad+prometheus/ops
19:43:14 <icinga-wm> PROBLEM - Logstash rate of ingestion percent change compared to
yesterday on icinga1001 is CRITICAL: 413.5 ge 210
https://phabricator.wikimedia.org/T202307
https://grafana.wikimedia.org/dashboard/db/logstash?orgId=1&panelId=2&fullscreen
19:46:02 <icinga-wm> PROBLEM - MediaWiki exceptions and fatals per minute on
icinga1001 is CRITICAL: cluster=logstash job=statsd_exporter level=ERROR site=eqiad
https://wikitech.wikimedia.org/wiki/Application_servers
https://grafana.wikimedia.org/d/000000438/mediawiki-alerts?
panelId=2&fullscreen&orgId=1&var-datasource=eqiad+prometheus/ops
19:55:07 <icinga-wm> PROBLEM - Citoid LVS eqiad on citoid.svc.eqiad.wmnet is CRITICAL:
/api (bad URL) timed out before a response was received: /api (Zotero and citoid
alive) timed out before a response was received
https://wikitech.wikimedia.org/wiki/Citoid
19:56:11 <icinga-wm> RECOVERY - Citoid LVS eqiad on citoid.svc.eqiad.wmnet is OK: All
endpoints are healthy https://wikitech.wikimedia.org/wiki/Citoid
20:04:21 <icinga-wm> PROBLEM - MediaWiki exceptions and fatals per minute on
icinga1001 is CRITICAL: cluster=logstash job=statsd_exporter level=ERROR site=eqiad
https://wikitech.wikimedia.org/wiki/Application_servers
https://grafana.wikimedia.org/d/000000438/mediawiki-alerts?
panelId=2&fullscreen&orgId=1&var-datasource=eqiad+prometheus/ops
```

A lot of alerts that fired were unrelated to sessionstore. However the total volume (39) was manageable. The first alerts pinpointed the problematic service accurately.

Conclusions

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

What went well?

- Outage was root caused quickly
- We were able to switchover temporarily to codfw (and back to eqiad) within 5mins each.
- It was during a timeframe many SREs were able to respond.
- Automated monitoring detected and informed of the incident quickly.
- We already had very good graphs and statistics built in our infrastructure

What went poorly?

- This could have been caught beforehand and extra capacity could have been added to the system, manually or automatically
- The dedicated nodes for sessionstore were already filled with instances and could not afford.

Where did we get lucky?

- Nowhere.

How many people were involved in the remediation?

- At least 4 SREs

Links to relevant documentation

Add links to information that someone responding to this alert should have (runbook, plus supporting docs). If that documentation does not exist, add an action item to create it.

Actionables

- Increase Logstash ingestion capacity / handle logspam situations better https://phabricator.wikimedia.org/T255243
- Increase capacity of the sessionstore dedicated kubernetes nodes https://phabricator.wikimedia.org/T256236 ✓ **Done**
- Increase kubernetes capacity overall https://phabricator.wikimedia.org/T252185 (codfw) and https://phabricator.wikimedia.org/T241850 (eqiad) ✓ **Done**
- Investigate the iowait issues plaguing kubernetes nodes since 2020-05-29. https://phabricator.wikimedia.org/T255975 ✓ **Done**
- Investigate the apparent network connectivity loss for kubernetes100[1,3,5] during the incident
- Investigate adding resource utilization alerts to services hosted on kubernetes
- Adopt SLIs/SLOs for sessionstore https://phabricator.wikimedia.org/T256629

TODO: Create missing tasks and add the #Wikimedia-Incident-Prevention Phabricator tag

Categories: Incident documentation | Incident documentation drafts