Noteworthy post-mortems

Tomasz Nowak (my slides)

https://github.com/danluu/post-mortems

Allegro, 31.08.2018

https://shorturl.at/dMY01

- Discount: 100x Honor 7C phones, price 850 zł → 1 zł.
- Allegro was prepared. They manually scaled up resources.
- “due to a misconfiguration, some services reserved much more resources than they actually needed.”
- “bad interactions between the autoscaler scaling services up and the cluster watchdog killing off unresponsive instances.”
- connection pools and file descriptors were saturated for incoming and outgoing network connections
- “one of the circuit breakers between our services was not correctly configured” – too high threshold for triggering
- “we plan to introduce a mechanism which will be able to tell backend services to generate simplified, cacheable responses”
- “the traffic which brought us down, was in large part bots rather than human users”

CloudFlare, 03.03.2013

https://shorturl.at/qrDM2

- DDoS against a costumer.
- Usual response: profile the attack, find pattern, block it.
- Added a rule to all routers.
- "What should have happened is that no packet should have matched that rule because no packet was actually that large. What happened instead is that the routers encountered the rule and then proceeded to consume all their RAM until they crashed."
- "many of the routers crashed in such a way that they did not reboot automatically and we were not able to access the routers' management ports. Even though some data centers came back online initially, they fell back over again because all the traffic across our entire network hit them and overloaded their resources."
- "ask them to physically access the routers and perform a hard reboot"

CloudFlare, 21.06.2022

https://shorturl.at/aELM1

- “a Change Request ticket was created, which includes a dry-run of the change, as well as a stepped rollout procedure. Before it was allowed to go out, it was also peer reviewed by multiple engineers.”
- The changes looked harmless, but in the routers’ configs, some lines were reoredered. `term REJECT-THE-REST` was put before other rules.
- Deployed to prod only on some locations, no harm as they didn’t use this part of their new architecture.
- Then deployed to busiest locations, but still not with locations with this architecture.
- Changes reached all locations, took 19 locations offline, 50% of requests failed.
- Fix was “delayed as network engineers walked over each other’s changes, reverting the previous reverts, causing the problem to re-appear sporadically.”

Facebook, 5.10.2021

https://shorturl.at/fhvGH

- Maintenance of servers / routers. Often need to take part of the backbone offline.
- A command was ran to assess the availability of global backbone capacity, which unintentionally took down all the connections in the network. The systems are designed to prevent mistakes like this, but a bug in that audit tool prevented it from stopping the command.
- For reliable operation, their DNS servers disable BGP advertisements when they can't speak to data centers, since this is an indication of a bad network connection. The backbone was shut down so it disabled BGP advertisements.
- They couldn't access their data centers and the loss of DNS broke internal tools they normally use to resolve such outages.
- They sent engineers to the data centers to debug and restart the systems. This took time, because these facilities are designed with high levels of physical and system security.
- Dips in power usage in the range of tens of megawatts.

Mozilla, 2.02.2022

https://shorturl.at/tDZ02

- Firefox has some services on Google Cloud Platform, which some day changed load balancers to use HTTP/3 by default.
- All Firefox browsers were suddenly hanging. They saw that they didn't change anything, but their data collection system started using HTTP/3.
- Explicit disable of HTTP/3 in GCP fixed the issue, but they didn't understand why at the time.
- Their data collection system was their only Rust code with their network stack, which used another library for network access, which lower-cases the `Content-Length` header, but their HTTP/3 code was case-sensitive.
- It looped the code, which disabled network communication, as all communication goes through one socket thread.

Google, 24.07.2014

https://shorturl.at/dprY0

- Gmail, Google+, Calendar, Documents, etc. were shut down for 25 minutes.
- Internal system that generates configurations encountered a bug, it generated incorrect ones.
- Config was sent to live services, it made the users' requests be ignored, and services generated errors.
- The same system later automatically generated a new correct configuration and began sending it, issue resolved.

Google, 31.07.2009

https://shorturl.at/rHL16

- Manually and automatically updated list of malicious sites.
- Human error: / added to the rules.
- Every google result was labelled with “This site may harm your computer”.
- Issue resolved after an hour.

Google, 05.04.2017

https://shorturl.at/wxzC8

- Replica of master server for config propagation lost access to the file system.
- Replica was changed to be the master server, system detected something was wrong, all load balancers started using oldest correct config from master server (which was outdated).
- Updating the master server triggered garbage collector to remove outdated configs.
- Server failed health check, automatic reboot. Clients got errors.

OWASA water treatment plant, 10.02.2017

https://shorturl.at/ptvW9

- “unintentional [water treatment plant] operator key stroke”
- prompted a water shortage in Orange County that closed businesses, placed towns under states of emergency. Customers were unable to use or drink their water for more than twenty-four hours.
- accidentally instructed the plant’s fluoride feed pump to change feed rate 10% → 80%.
- Fluoride is added to water to prevent tooth decay.
- “The operator tried to change the command about twelve seconds later, but according to the report, the change didn’t register.”
- A lead operator at the plant noticed the extra high levels but did not take corrective action.
- “The water main break may have been the result of the pipe bending because of pressure and settling.”
- “The break leaked about 1.2 million gallons of water, causing pressure in the system to plummet. As a result, customers were ordered not to use or drink their water because low flow can allow bacteria to grow in pipes.”

Rust crates.io, 21.06.2023

https://shorturl.at/cjlnF

- 15 minutes of crates.io failure, 3M failed requests.
- Pull request for migrating to AWS S3, refactored how endpoints create URLs.
- Missing slash when generating URLs, e.g. `static.crates.iocrates` instead of `static.crates.io/crates`.
- There were tests, but on prod there were different env vars.
- Easy one minute fix, just rollback deployment.

Amazon AWS, 02.06.2012

https://shorturl.at/cdpNZ

- Weather warnings of a storm over their datacenter. Changes in datacenter cancelled, more personnel overnight.
- Power fluctuations, a large voltage spike, electrical switches initiated transfer to generator power.
- Generators failed to provide stable voltage. Servers were running on UPS, but not for long. 20 minutes of downtime.
- But generators were regorously tested! They always passed 8h of load testing, months ago they successfully generated power, weekly tests, they were new.
- Clients' servers were offline. Booting them was slow, as data volumes could be in an inconsistent state (e.g. poweroff mid-write). It requires manual checks by clients to see if everything is fine.
- A bug appeared, AWS load balancing filled up event queue, clients also did a lot of requests for new instances. Events fell behind.
- Another bug appeared and some "Multi Availability Zones" didn't change the main databases from the faulthy datacenter to the backup ones. It triggered a failsafe, required manual intervention.

ARPANET, 27.10.1980

https://shorturl.at/eBFM3

- The network was unusable for several hours. All the “routers” (Interface Message Processor) didn’t communicate. Restarting the IMPs was OK, until they connected to the net, and then they failed.
- One IMP flipped some bits in routing data.
- There were checksum, but they were only checked periodically for software efficiency.
- Incorrect sequence numbers caused packets to be sent in a loop and router buffers began to fill.
- Full buffers caused loss of keepalive packets and nodes took themselves off the network.

FirstEnergy / General Electric, 14.08.2003

https://shorturl.at/mxzRY

- Hot day, high demand for electricity (cooling). Tree branches touched power lines in Ohio.
- Bug with race condition appeared in the energy management system. It stalled FirstEnergy's control room alarm system for over an hour. System operators were unaware of the malfunction. The failure deprived them of both audio and visual alerts for important changes in system state.
- The lack of alarms led operators to dismiss a call from American Electric Power about the tripping and reclosure of a 345 kV shared line in northeast Ohio.
- Overload of a power line → more heat in the power line → metal conductors expand → line sags too low → a flash over to nearby objects (e.g. trees) occurs → transient increase in current → power line disconnected for safety.
- Big changes of voltage, generators go off for safety, cascading effect, 256 power plants offline.
- Blackout, 55 million people affected for 2h-4 days, including NYC.

JaneStreet, REDACTED

https://example.com

- REDACTED

EVE Online, 5.12.2007

https://shorturl.at/cehoU

- Game update. A script changed CWD and deleted some old files, e.g. the game's `boot.ini`.
- The script command ignored the CWD and assumed that it was from the root directory.
- Windows users got their `/boot.ini` file removed.
- It passed tests, as Windows recovers when it's on the first partition of the boot drive (and fails to recover otherwise), the testing env was always the same.

GitHub, 21.10.2018

https://shorturl.at/hioBN

- Maintenance of network between data centers. Mistake, 43 seconds of outage.
- US West and US East got disconnected, algorithm changed a read replica to a write primary in US West.
- Connection restored, but now there are two primaries, each had `writes` that the other didn't have; unable to revert to one primary.
- Decision to stop writes, e.g. pushes, webhook delivery, GitHub Pages builds. GitHub partially unusable.
- Plan: restore backups, synchronize replicas, fall back to one primary. Backups are made often, but reverting terabytes takes hours.
- Data on the sites was in weird state, inconsistent. No data loss though.
- Restoring backups took longer than expected, as there was additional load from active GitHub pages and lot of synchronization attempts.
- 24 hours of degradation.

Knight Capital, 1.08.2012

https://shorturl.at/gAEKY

- Bankruptcy of Knight Capital, 440 million dollar loss.
- new market, one month announcement, Knight wanted to prepare. They updated software (which was 8 years old) that split big orders into many smaller ones. They reused a boolean for activating unused code for a new functionality.
- Mistake: deployed to seven out of eight servers, no review, no alert system in case of discrepancy, no written procedures for supervision.
- 97 automatic email messages about unusual state, but not on high-priority channel, and they weren't read.
- Seven servers worked correctly, the eight ran the unused code that looped splitting of orders.
- No kill switch, 20 minutes of looking at what is wrong, they reverted to old code, now 8 out of 8 servers were "bad".
- 45 minutes running, long on 80 stocks, 3.5 billion dollars, short 3.15 billion dollars.

Microsoft, 29.02.2012

https://shorturl.at/rBTY7

- Code for creating certificates set the valid-to date for current time plus one year. Failed on February 29th.
- Any new Virtual Machines in Windows Azure didn't initialize, automatically after 25 minutes the machines reboots, after three times it reports a hardware error and moves the VM to another server. When a certain number of servers fails, manual intervention is needed.
- Admins disabled service management. Created a fix, tested it, pushed to prod.
- Some clusters weren't fully updated and it required time for them to update, admins pushed for faster deployment, but the ones that weren't updated didn't have fully compatible networking with the fix, which shut down those clusters from the network.
- Outage that lasted for most of a day.

Linux, 7.01.2012

https://shorturl.at/gxyC5

- Leap second occurred, `CLOCK_REALTIME` in Linux was rewound by one second.
- Not done through `hrtimer bsae.offset`, thus `TIMER_ABSTIME` `CLOCK_REALTIME` timers got expired one second early, including timers set for less than one second.
- This caused applications that used sleep for less than one second in a loop to spinwait without sleeping, causing high load on many systems. This caused a large number of web services to go down in 2012.

European Space Agency, 4.06.1996

https://shorturl.at/adlm7

- Unsuccessful test flight of Ariane 5 expendable launch system. The rocket veered off its flight path after launch and was destroyed. Loss: 4x Cluster mission spacecraft, 370 million dollars.
- Software reused from Ariane 4, as then everything worked.
- Ariane 5's flight path was different. It had greater acceleration. Pre-flight tests had never been performed on this code under Ariane 5 flight conditions.
- A data conversion from 64-bit float to 16-bit int overflowed. For efficiency they removed software handler for this error trap (though there were some for other errors).
- This led to a cascade of problems, culminating in destruction of the entire flight.

Valve, 14.01.2015

https://shorturl.at/beuG7

-
  ```
  STEAMROOT="$(cd "$0%/*"&& echo $PWD)"
  # Scary!
  rm -rf "$STEAMROOT/"*
  ```
- Removing or moving `/.local/steam` caused `STEAMROOT` to be empty, which did `rm -rf /`
- After this blew up on social media, there were widespread reports that this was reported to Valve months earlier. But Valve doesn't triage most bugs, resulting in an extremely long time-to-mitigate, despite having multiple bug reports on this issue.

Thanks!

Tomasz Nowak

https://github.com/danluu/post-mortems