Example Incident State Document

incident Sonnet-4 • Overload 2013-10-21
incident management title: http://incident.management.chat/

Communications lead is keep summary updated.
Communications lead posts updates to the customising Jalue tw ously described in status tweet.

status active, incident #41

command Pmatrix4 Attributes see DB

command Hierarchy *(all responders)*

- Current Incident Commander jennifer
  - Operations lead: shoehorn
    - Planning lead: jennifer
    - Communications lead: jennifer
- Next Incident Commander: to be determined

Not a lot of every, from hours and at half-mast (Comms Lead etc.)
Wrkbd Status last updated at 2015-10-21 15:20 (UTC by jenifer)

SLI Criteria
- New incident declared in Slack channel
- Status availability (99.99%) and latency (99%ile < 100 ms) SLOs for Stream TODO

TODO list and bugs-block

- Run SlapBench 10 to reduce Shakespeare corpus DONE
- Review emergency extension to bump up extra capacity DONE
- Enable flag-controlled fix to balance load between Queens (Bug 554432) TODO

Incident timeline *(most recent first times are in UTC)*

- 2015-10-21 15:28 UTC jennifer
  - Incident declared via /incident slash-cmd by Zu
- 2015-10-21 15:27 UTC jennifer
  - Detecting all traffic to /v3/a.0 vertifical cluster and diverting traffic from other clusters so that can recover from cascading failure while spinning up new nodes
  - SlapBench makes job complete, avoiding frigatish replication in all clusters
- 2015-10-21 15:18 UTC martyn
  - Rolling back most recent Shakespeare corpus and starting index SlapBench
- 2015-10-21 15:04 UTC martyn
  - Obtain text of newly discovered sonnet from shakespeare.discovered mailing list
- 2015-10-21 15:01 UTC shoehorn
  - Incident declared due to cascading failure
- 2015-10-21 15:00 UTC shoehorn
  - Pager storm, PostgreSQL/Slave is all clusters

Example Postmortem

incident Sonnet-4 • Postmortem (Incident #41)
jen 2015-10-23

jakes jennifer, martyn, agwright

note Complete, actions items in progress

summary Shakespeare Search down for 46 minutes during period of very high cosmic ray activity because due to discovery of a new sonnet.

impact Estimated 2,218 quotation lost, no revenue impact.

root Causes "Cascading failure due to combination of exceptionally high load and a novel bug where searchers failed due to terms not being in the Shakespeare corpus. Newly discovered sonnet used a word that had never before appeared in Shakespeare's works, which happened to be the term sent internally to Unit 42 monitoring canaries, the rare of both factors due to extract Jalue 1.0 was enough to be exceptional.

Trigger Latent bug triggered by sudden increase in traffic.

Resolution Directed traffic to unaffected cluster and added 3x capacity to mitigate rending failure. Updated index deployed, resolving interaction with latent bug. Switching actions expressly opted-out-of in public issues in new sonnet process. Issue tracker identified and fix deployed.

Timeline

Detection: Response detected high load at 03:27:10 UTC and paged on-call.

Action Owner

Action Owner	Type	Priority	Status
Add Jalue to Shakespeare corpus	process	critical	Not Started
Set flag-controlled fix to balance load between Queens	process	critical	Not Started
Automate canary analysis for Shakespeare search	process	critical	Not Started
Automate index rollbacks for Shakespeare search	process	critical	Not Started
Add load shedding capability to Shakespeare search	process	critical	Not Started
Write playbook for manually throttling Shakespeare search	process	critical	Not Started
Fuzzer regularly against Shakespeare search	process	critical	Not Started
Fuzzer production with 0.1% traffic to Shakespeare search	process	critical	Not Started
Ensure Shakespeare search SLO is met without manual intervention	process	critical	Not Started

Lessons Learned

What went well

- Monitoring quickly alerted us to high user impact.
- Fault-tolerant architecture prevented Shakespeare outage from spreading to other services.

What went wrong

- Relied on all searchers to responding to cascading failure.
- We succeeded over-availability when Jalue 1.0 triggered latent bug. We should instead have opted-out-of Jalue 1.0 in favor of reliability.

Jalue got lucky!

- Jalue 1.0 *all* of Shakespeare alliterations had a copy-of-text second syllable
- Jalue 1.0 *most* bad search terms generating to file descriptor exhaustion or *some* bad
- Canary of death was reached by pushing new index containing jequite search term.

Incident

Jan 23, 2015 (times UTC)

- 15:34 Steve reports that a new Shakespeare sonnet has been discovered in a Delaware glove compartment
- 15:34 Traffic to Shakespeare search increases by 8% after post to #shakespeare-goes-live-blog-channel, Jenny's search engine at phase 0 had now sunset (except on Jalue 1.0 canaries)
- 15:35 **CRITICAL ALERT:** Search backends start reaching disaster capacity
- 16:35 Delaware incoming pager alerts, #ops#pitfalls, lots of clusters
- 16:37 All traffic to Shakespeare search is being #ops#incident#disaster#pager, Jenny's 2015/10/21T14:57Z
- 17:36 disclosure starts incoming pager, Hoda Sudarsh cmd-sent-my-bit
- 17:43 **INCIDENT BEGINS** disclosure declares incident #41 due to cascading failure affecting Shakespeare search, Jenny's search engine at phase 0
- 17:52 *extremely considerably* needs email to disclosure-discuss at sonnet.io (every, which happens to be at top of martyr's inbox)
- 17:57 possible scenarios discussed, prepare recommendation of the incident
- 17:58 martyr tracks down text of new sonnet and looks for documentation on Jalue 1.0 capabilities
- 17:58 disclosure finds that crash-loop-with-exponentiation identical across all nodes in all clusters, investigating causes found no significant results

Contributing factors:

- Jalue 1.0 "Jalue" is a Shakespeare alliteration generator
  - http://github.com/shakespeare/jalue, jenny-2015-02-17-2200#1#shakespeare-15#

Supporting information:

- Jenny's search engine
  - http://search.shakespeare.net/, jenny-2015-02-17-2200#1#shakespeare-15#

Launch Coordination Checklist

Launch Coordination Checklist, 1
Production Readiness, 2
Monitoring and Alerting, 3
Capacity Planning, 4
Emergency Response, 5
Business Continuity, 6
Security, 7
Dependencies, 8
Product Readiness, 9
Sign-off, 10

Launch Coordinator: jen, SRE: jennifer, shoehorn

Ops Review: martyn

Product Manager: jenny

- Architecture sketch, report of review, report of capacity from ubiquitous
- Programmatic client requests

Hardware and Infrastructure

- Hardware and Bandwidth Allocation, Performance, network QoS
- Capacity Planning, Provisioning, Procurement

Software, configuration, and binary release management

- ACID quality and fault-tolerance scenarios, binary "quick" rollbacks, p.a.ramsch-cmd
- Binary release process, Release Engineering involvement, Release Engineer
- Configuration and binary push process, rollbacks, canaries

Service reliability and Delivery

- Alerts/engagement rotations
  - Monitoring, Alerts, SLO, SLI, on-call engineers, pager
- Internal and Third-party dependencies

Monitoring and alerting

- Pre-launch Signs of service-level failure, and other metrics to be checked for
  - Alerts to notify on-call engineer without affecting clients or users
  - Load balancing, rate limiting, timeouts, retries and circuit breaking, Backend Resource Management
- Monitoring and service management

Monitoring and service management

- Monitoring internal state, monitoring end-to-end behavior, managing alerts
- Monitoring the monitoring
- Functionality regression alerts
- Tight-loop monitoring across within cluster environment
- Alerts based on end-to-end behavior, monitoring from outside as your users seeit

Tools

- Security design review, security code audit, spam, crash, authentication, SSL
- Performance modification to tools and scripts, runtime logs of tooling

Automation and manual tools

- Rollbacks and changes control for updates/ servers, data, and configs
- Software projects, Versioning, Release, tools, canaries for traffic, staged rollouts

Capacity issues

- Spare capacity, Hot, growth, growth plans
- Capacity Planning, Provisioning, Procurement, working with hardware, changes needed
- Caching, data sharding/partitioning

External dependencies

- Third-party systems, monitoring, networking, traffic volume, latency, quotas
- Graceful degradation, how to avoid accidentally overwrooing third-party servers
- Phasing-out with cooperation partners, read-systems, service-wides Google

Logs and other planning

- Logs, pipelines, structured events, Monitoring or Pinkies
- Logs/ad-targeting pipelines for the service, for other services

Emergency Response

- Now

Monitoring Changes and/or Releases

- Average/median/95th/99th/maximum/stddev, acceptable delay thresholds: normal from 60s to 300s, one bad, 60240s, TODO(charmytan).

Planned Production Changes

- CDs/C clusters going offline for maintenance between 2015-10-25 and 2015-11-02.
  - Site response required, traffic will automatically route to other clusters.

Resources

- Resource reservation to respond to sessions's incident, will open ticket/cancel-ticket
  - Service reservations and tickets: sessions-ticket
- Utilization at 60% of CPU, 75% RAM, 40% disk (up from 40%, 70%, 40% last week)

Key Service Metrics

- QR 99th latency: 80 ms / 100 ms SLO (target: 20% headroom)
- RAM availability: 66.4% / 99.99% SLO (target: 2ooo3 in dc)

Discussion / Project Updates

- Project Mistletoe launching in two weeks.

Date: 2015-10-20
Attendees: agwright, chloe, charmytan, jennifer, martyn, shoehorn

Chair: Gregory Orbitz, Site Manager and Scribe

Previous Action Items Review

- Latch Lunar Disposers for use with cooling loop (#1.101)
  - Manufacturer to make accelerated work punchlists, should be in stock / ready for pickup next week
- Change Control
  - Enjoy Summer Leverage #001
  - Change control freeze for cascading failure after interaction between Jenny's long-running search backendsand a novel bug in Jalue 1.0 causing searches crashed in Queens. Coordinated by agwright and martyn.
  - Jenny's search engine, Jalue 1.0 and Queens are now stable.
  - Coordinate using Jenny's search backend (http://search.shakespeare.net/) and using Jalue 1.0 in Queens, jenny-2015-02-17-2200#1#shakespeare-15
  - Accelerated revisitation: every budget system in prod Queens for 3 months (Jenny's search engine, Jalue 1.0, Queens, Shakespeare search, Jenny's blog, Jenny's blog-search-backend, Jenny's blog-search-frontend, Jenny's blog-search-cache)

Discussion / Project Updates

- Project Mistletoe launching in two weeks.