Scaleable Computing

What about some failures?

CS7NS1/CS4400

Stephen Farrell
stephen.farrell@cs.tcd.ie

https://github.com/sftcd/cs7ns1/

Note: PRs for repo are welcome!

Contents

- 1965 NE US/Canada power outage
- 2012 Hurricane Sandy
- Facebook’s 2015 approach to failures
- General survey of Internet outages (Mar. 2018)
- Oct 2018 Github.com outage
- I broke my TV! (Oct 2018)
- Some effects of the root KSK roll (Nov 2018)
- Nov 13th Google BGP leak/hijack
- ...Your outages/failure war-stories here.

1965 NE US/Canada Power Outage

- Cascading relay trips caused major outage
  - Relay tripped => power re-routed elsewhere => another relay trip
  - https://en.wikipedia.org/wiki/Northeast_blackout_of_1965
- “ANATOMY OF POWER SYSTEM BLACKOUTS: PREVENTIVE RELAYING STRATEGIES” from 1996 considers “hidden failures” and “regions of vulnerability,” IEEE Transactions on Power Delivery, Vol. 11, No. 2, April 1996, Tamronglak et al.
  - https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=489327
- Protection systems can cause failure!

2012 Hurricane Sandy

- Hurricane caused Internet outages
  - Active measurement – pings of IPv4 /24 ranges and MaxMind geo-location
  - Post-facto analysis of outages, but near real-time is possible
- “A Preliminary Analysis of Network Outages During Hurricane Sandy”, USC/ISI Technical Report ISI-TR-685, November 2012, Heidemann et al.

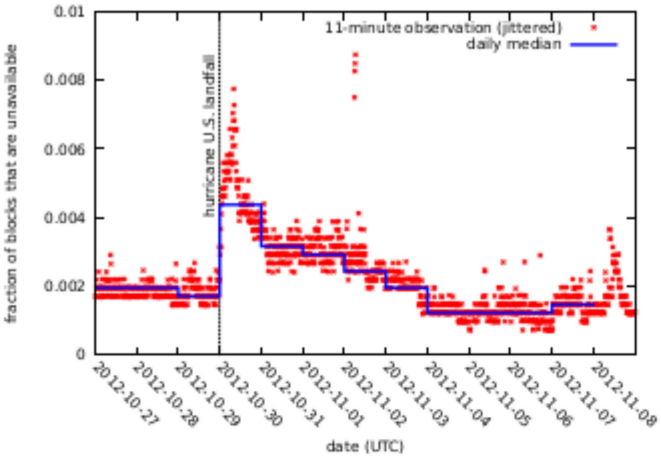

Figure 3: Median daily outages (solid line) for /24 blocks geolocated to the United States, with jittered individual readings (dots). (Dataset: [15]).

Facebook’s approach in 2015

- "Move Fast and Break Things" – dunno if they still say that:-)
- Causes: hard disk failures, workload changes, human error, fast configuration propogation
- Mitigations: A/B testing, controlled config propogation/canaries, strict validation of configs, simplify reverting, fire-drills, post-mortems, clever queues and good dashboards (that scale)
- More SLA violations when people are working on the system than when they’re not (at home, vacating, doing bureaucracy!)
- “Fail at Scale, Reliability in the face of rapid change”, ACM Queue, Oct 2015, Maurer
  - https://queue.acm.org/detail.cfm?id=2839461

Internet Outages Survey

- “A comprehensive survey on internet outages,” Journal of Network and Computer Applications, v 113, 1 July 2018, Pages 36-63, Aceto et al.
- Causes: mixed; nature (weather, earthquakes), submarine cable cuts, censorship, DDoS, BGP fun, buggy kit
- Detection Methods: passive/active/hybrid measurements
- Impacts: numbers affected, costs (to whom?), (censorship?)
- Methods for Resilience: of n/w or AS, of applications?
- Lots of good references there... graphic on next slide...

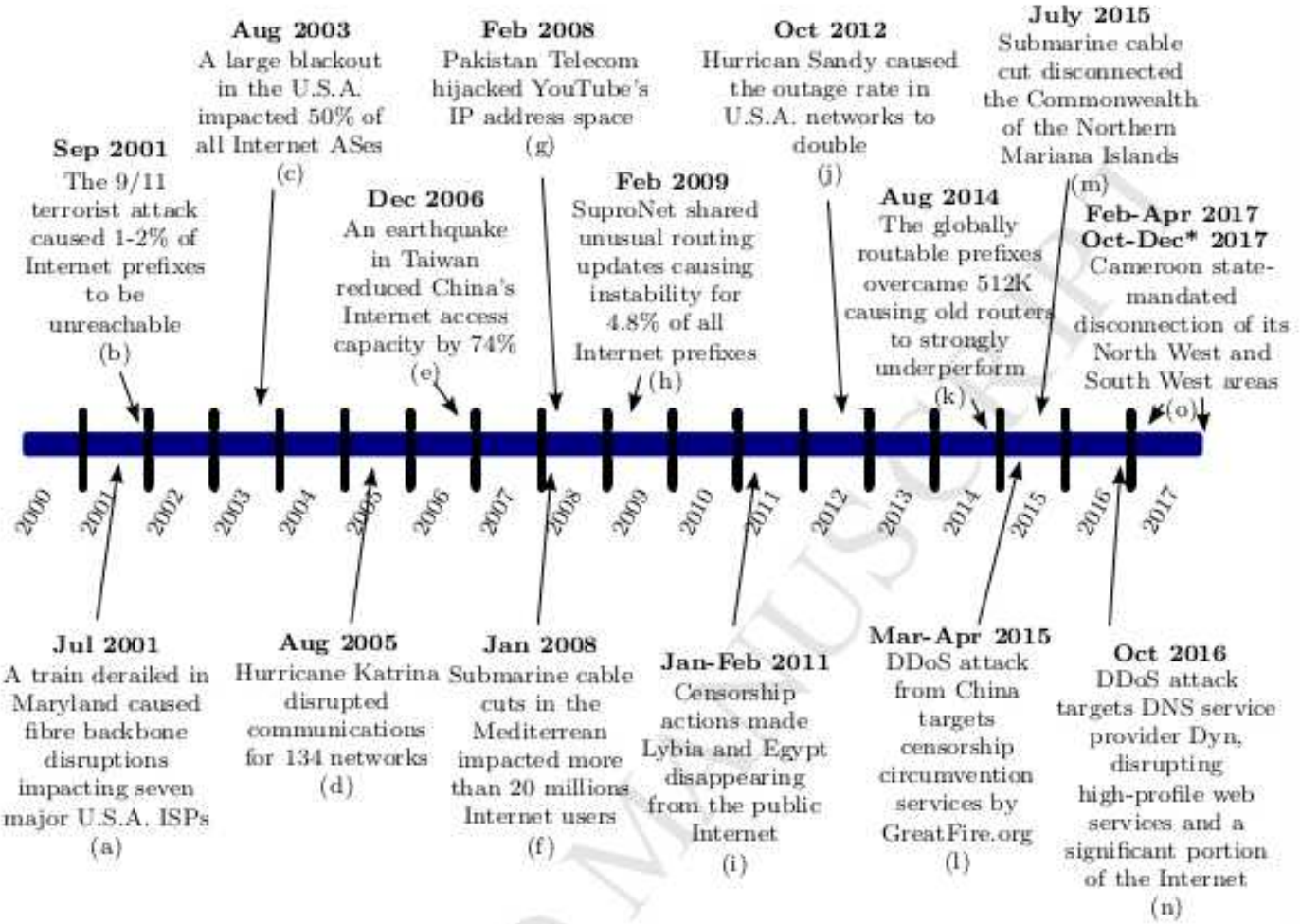

Figure 2. A timeline reporting concrete examples of some of the main and well known Internet outages. References (in chronological order): (a) [225], (b) [178], (c) [73], (d) [11], (e) [9], (f) [59], (g) [52], (h) [256], (i) [72, 71], (j) [74], (k) [221], (l) [165], (m) [27], (n) [159], (o) [29].
* Still ongoing at the time of writing.

“Internet outage”

- Aceto et al’s, definition of “Internet outage” doesn’t seem so good to me:
  - “...the particular condition in which the network lies when one or multiple network elements located in a specific geographic area either do not work properly or are not reachable due to intentional or accidental events.”
- What’d be a better definition?

Github.com outage

- Oct 21 2018 outage
- Cause: 43 seconds n/w partition
- Result: ~24 hours of downtime for github.com web site features
  - US coast-coast latency (60ms) and DB failover/recovery problems
- Notable: git CLI kept working; lack of OOB status reporting; growth => complexity, lack of history
- Write up: https://blog.github.com/2018-10-30-oct21-post-incident-analysis/

I broke my TV!

- <Draws on the blackboard>
- In turning on DNS/TLS (DoT) at my main home router, I got too enthusiastic and set set top boxes (STB) to use that recursive, which won’t work (STB’s on WAN side of DoT recursive).
- ~72 hours later STB’s needed new DHCP leases, and they somehow also needed DNS at that time (WTF? but whatever, some multicast IPTV stuff I guess) which didn’t work of course
  - Confusion abounded given the latency between the bad action and the breakage

DNS Root KSK Roll

- October 13 2018: Eir had a problem:
  - https://www.rte.ie/news/2018/1013/1002966-eir-outage/
  - Ultimate (speculated) cause: new 2017 key signing key for DNS roots!
- <Draws on blackboard again>

  then
- Re-uses Geoff Huston's slides from last week
  - https://iepg.org/2018-11-04-ietf103/2018-11-04-kskroll-iepg.pdf
- Background on what/how they measure:
    - https://www.potaroo.net/presentations/2017-09-13-how-labs-measures.pdf

Google BGP hijack/leak

- Tuesday Nov 13th:-)
  - https://www.theregister.co.uk/2018/11/13/google_russia_routing/
- Nigerian ISP advertises google ranges via BGP
  - Apparently accidentally, though not all BGP hijacks are accidents
- Result: for 74 minutes, routes to google were via TransTelekom (RU) and China Telecom, who blackholed the trafffic
- Lesson: BGP is fragile, RPKI may help

Interesting failures you've seen?

- Yes, Rosettahub:-) But how specifically?
- What else...

Contents - revisited

- **Cascading protection system fail**: 1965 NE US/Canada power outage
- **Nature**: 2012 Hurricane Sandy
- **Processes**: Facebook's 2015 approach to failures
- **Surveys**: General survey of Internet outages (Mar. 2018)
- **Backend orchestration flaws**: Oct 2018 Github.com outage
- **Too smart by half**: I broke my TV! (Oct 2018)
- **Complex systemic effects**: Some effects of the root KSK roll (Nov 2018)
- **Legacy Infrastructure:** Nov 13th Google BGP leak/hijack
- **Various:-) :** ...Your outages/failure war-stories here.

Conclusion

- Failure is real and inevitable
  - Fail to plan to fail => fail :-)
- We learn a **lot** from considering failures
  - Especially our own
- Lots of code is concerned with handling failures
  - If that’s not true for your system, then you likely have trouble ahead!