Some Lessons from History

• Ross Callon (retired)
  (& John Scudder, Juniper)
• IETF 102, Montreal
• July 2018
Lessons from History

• There have been major Internet issues
  • “Interesting” events in 1980’s, 90’s, 200x’s
  • We didn’t always know what we were doing

• Some knowledge is in the mind of old folks

• I thought it would be wise to write some of these down

• Examples to follow
  • I have tried to be fully vendor-neutral (eg, have not considered issues with proprietary protocols)
Arpanet Collapse (early 1980's)

• A switch crashed and restarted
  • Forwarded old packets in output queue

• Result: Old Route Update was propagated...
  • While another update was in progress
  • Old update was exactly 1/2 way around circular sequence space (a>b; b>c; c>a)
  • Update A replaced Update B
  • Update B replaced Update C
  • Update C replaced Update A...
Arpanet Collapse (early 1980's)

• Problem: Three updates chased each other around the Arpanet for hours
• Solution: All but two packet switches had to be manually shut down
• (today, with hundreds of routers per network, this could be quite unpleasant)
OSPF Flooding Issue (early 1990’s)

- Stable Network, Well-connected core with single-homed stubs

- S.P. thought: I really care about reliability. Let's multi-home stubs...
OSPF Flooding Issue...

- Redundancy added:

- Result: Collapse
- What happened?
OSPF Flooding Issue...

• Core router had LSA to send out
  • Transmitted to all adjacent routers

• Stub routers all forwarded the LSA to their neighbors...
OSPF Issue...

• Result: Other core router was overwhelmed with LSAs forwarded by stub routers

• Lesson: Buffering and discarding duplicate LSAs is a difficult part of OSPF/IS-IS
  • No one predicted this
Flooding Issue with IP over ATM

- Similar issue occur with full mesh of circuits over an ATM core
- Mesh groups added to deal with this
Lost Hellos (~1992)

- Network stable for long periods of time
- Multiple random changes in short order cause processor to fall behind
  - Processor drops Hellos, adjacencies dropped
  - More routing updates transmitted
  - Widespread CPU congestion, more Hellos dropped
  - Entire network disconnects
  - Problem stabilizes, network recovers
  - ~20 minutes later, many LSAs are refreshed, problem repeats
Lost Hellos...

• **Solution**
  • Optimize protocol processing
  • Prioritize Hello processing
  • Randomize timers
  • (Apply to all routing protocols)

• **This was known in early 90's**

• **But...**
ATM Switches, mid 1990's

- ATM Network partitions, re-connects
- New updates flooded between partitions
- CPUs congest, drop Hellos
- Adjacencies dropped, Network Disconnects
- ('Prioritize Hellos' wasn't well enough known)
IP Nets: DDoS Attacks

- Attacker compromises many hosts, uses them to launch a coordinated attack
- Result: Link Congestion
Slammer, January 2003

• Slammer worm
  • Very rapid propagation (doubles in ~8sec)
  • Widespread congestion in IP networks worldwide

• Result
  • Routers drop Hellos, Adjacencies dropped
  • Network disconnects
  • (not clear if result of link or CPU congestion)
  • Issue getting management plane to respond
Solution: Prioritize Hellos + ...

• Give priority, guaranteed resources for real time protocol functions
• Prioritized queues
  • Inside router, and on egress
Invalid Update Issue

• IS-IS (and OSPF) defined in mid 1980's
  • Smaller CPUs, which also forwarded packets
  • ⇒ Original spec minimizes CPU strain
  • In forwarding IS-IS updates: Check outer wrapper, forward, then check internals

• IS-IS & OSPF were widely deployed, interworked well
  • IS-IS was solid for several years
Invalid Update Issue,…

• Bad interface trashes update
  • One in ~65,000 have checksum which passes
  • Check outer wrapper (OK)
  • Forward (OK)
  • Check internals: Field out of range, Crash

• Result: Entire area crashes
  • Many rtrs, multiple vendors
Distance Vector (RIP) Count to Infinity

- Distance vector count to infinity is fairly well known
Many fixes have been proposed, some deployed
There was an interesting “augmenting” to this problem in the early NSFnet...
Delay Based Routing

• There have been multiple “interesting” experiments with routing based on real time (queuing) delay

• Early NSFnet “Fuzzball” routers had delay based “Hello” protocol in the core, mapped to RIP around the edges
  • Delays vary dynamically, feeding unstable metrics into RIP
  • This was not pretty

• An Arpanet variant used a linear combination of hop count and real time delay, carefully overdamped
  • When delays grew (congestion), it became under-damped
Non-deterministic routing

• ”BGP wedgies” are a well-documented example.

• One set of policy configurations can result in multiple different stable forwarding topologies (“multistable”), depending on timing.
  • Because policies are local, but forwarding is global.

• Much more detail in RFC 4264.
BGP “wedgies” simple example

- C uses BGP community to tell X “use this link as a last resort only”
- When primary fails, all is well.
- But when primary is restored, forwarding topology has a new stable state. (And not what C intended.)
BGP MED Oscillation

• Actually, BGP isn’t even always multistable.
• The BGP MED path attribute can cause persistent oscillations (see RFC 3345).

• How did this happen?
  • BGP route selection assumes total order.
  • MED gives only a partial order (MED is only comparable if source AS is the same).

• Protocol was designed to be correct with a flat IBGP
  • MED wasn’t considered when designing route reflection, which does data hiding.
  • Even if it had been, not clear there would have been a solution.
Optional Transitive BGP Attributes

• Some BGP data is opaque to routers handling it, and can transit across them.  
  • Optional Transitive Path Attributes, most famously.

• When the data is handled by a router that does understand it, the router says “oh my goodness my peer has sent me a bad update it must be insane” and resets the session.
  • But the peer didn’t misbehave. Some router far across the Internet did.
  • This means one naughty router can cause a very large number of sessions to reset.

• Best intentions by protocol designers, but a terrible outcome.

• Fixed by RFC 7606 (keep the session up but delete the malformed routes, don’t assume the peer has gone insane).
BGP – a few lessons

• Simple protocols have complex behaviors when assembled into large systems.

• Extensible protocols lead to small extensions that have surprising consequences when they interact.

• If you serve several masters (protocol correctness, business reality) something has to give.

• Data that is sometimes opaque leads to results that are sometimes surprising.

• The worse-is-better design philosophy is powerful.
Other examples...

• Operator errors
• Distribution of full BGP routes into IGP (IS-IS, OSPF, ...)
• Scaling
• Signaling System 7 (SS7) failure
• Rumors of other issues
• And note, I have not mentioned multicast...
  • Eg, “multicast grenades” are in principle possible
What To Do With This Information?

• I had been intending to write an Internet Draft (to RFC)
  • This isn’t going to happen (I am retired, and like it that way)
  • Adding more detail and additional examples would be useful

• “Those who cannot remember the past arecondemned to repeat it”
  • Old saying (possibly originally by George Santayana)

• Today, repeating these failures is not acceptable
  • We all depend upon a stable and reliable Internet

• Hopefully, this presentation can be helpful