BGP, where are we now?

John Scudder and David Ward
IETF-68, March 22, 2007
Agenda

- Trivia
- Dynamic behavior
- Convergence properties and problems
- Convergence/stability work items
Goals and Priorities

- Goal: Maximize connectivity of Internet
- Convergence and stability are subsidiary to this
- Implication: Priorities
  - First: fastest service restoration
  - Second: minimize peak load on control plane
Focus

• This talk focuses on performance and stability
• There are other very important aspects of BGP
  • Services
  • Operations
  • Weird behaviors (wedgies, etc)
  • Security
  • ...
• But we don’t have all day
Shalt Not’s

• BGP uses ASes for loop suppression — and nothing else!
  • Speaking of “overloading things”… ASes are not locators. No topological significance.

• Auto-aggregation appears to be a non-starter
  • Even proxy aggregation is tricky, but that’s an operational consideration
MP-BGP

- BGP carries data for multiple address families (AFs)
  - Plain old IP (v4, v6)
  - VPNv4
  - Other things
- Not all AFs need to be present on all routers!
VPNs

- Often observed that VPN tables larger than Internet table
  - True, in aggregate
  - But, not true of any single VPN table
- Inherently parallelizable
  - No single PE or RR holds all VPN tables
- Operational challenges to managing
  - Some tools to do this, e.g. rt-constrain
BGP dynamic behavior

- Confusion even among routing experts
- Of course, surprising emergent behaviors are possible
- … but important to understand bounding conditions
BGP and TCP

- BGP runs over TCP
  - Flow control: important implications for dynamics
  - Intuition about TCP is usually wrong…
BGP under load

- When uncongested, BGP will pass updates as fast as they are received
- Modulo MRAI, dampening
- Degradation mode under (CPU) congestion: state compression
  - “Adaptive low-pass filter” behavior emerges
  - Things slow down, they typically do not melt
BGP under load [2]

• BGP adapts to speed of peer
  • Slow peer gets routes as slow as it wants (with state compression)
  • Fast peer gets routes as fast as it wants
  • Implication: One slow peer does not hinder overall convergence

• Update packing
  • Low prefix/update ratios when not congested… but that’s fine!
  • High ratios emerge under congestion… which is when needed
BGP convergence

- At least $O(n)$ in the size of the DFZ table
  - Fundamental to how BGP transports routes
- But full convergences don’t happen often!
  - At startup (“initial convergence”)
  - On rare occasions otherwise
- Hard to “fix” completely — but is it broke?
  - “BGP’s biggest, yet least important, problem.”
BGP convergence [2]

- Techniques to avoid full convergences
  - Graceful Restart
  - Nonstop Routing
- ... or to cover them up
  - Different flavors of fast reroute
- ... or to pre-converge by advertising extra routes
  - Best-external, multi-path and similar
Route Reflection

• RRs hide backup paths
  • Reduce RIB sizes (but less than you think)
  • Bad for convergence

• Convergence:
  • State reduction/data hiding
  • Faster convergence
  • Pick one
Known Algorithmic Deficiencies

- Path hunting
- Nonconverging policies
- At least $O(n)$ in DFZ size
Path Hunting

• Well-known amplification effect
• Approaches to reduce
  • Root cause notification
  • Propagation of backup paths
Propagation of Backup Paths

• Transit ASes seldom fully partition from each other

• However, when a single AS-AS link goes down, border router temporarily loses routes

• Due to aggressive data hiding by less-preferred border routers and RRs
Propagation of Backup Paths [2]

- Speculation: many “path disturbance” events caused by this effect
- Intra-domain backup propagation feasible today
- Cost: some additional RIB state within AS
- Benefit: faster internal convergence and global stability
Some Possible Tools

- As-pathlimit
- Aggregate withdraw
- Best-external
- Better instrumentation reusing WRD infra
- BGP free core (pick your encap)
- Dampening (with better parameters)
- Multi-path
- Root cause notification
Moving Forward

• Narrow down (or expand!) “possible tools” list

• Align costs and benefits
  • Those who pay, must benefit, or solution will never be deployed
  • Many examples of existing technically-excellent “solutions” to current problems… but problems still exist. Example: BCP-38
  • Deployment trumps all considerations!

• Focus on behavior under load (or making load go away!)
Dampening

• Misused in past (we were wrong about default parameters)

• Heavy contribution of few sites to GH data suggests very generous parameters which only penalize egregious flappers
  • Study needed to validate what constitutes “egregious”

• Given parameters, can be turned on today
  • Lower-than-low hanging fruit
  • Aligns costs and benefits
Punch Line

- BGP not in danger of falling over
  - Lots of runway
- IDR
  - Near-term improvements
- RRG
  - Fundamental changes, e.g. new routing and addressing architectures