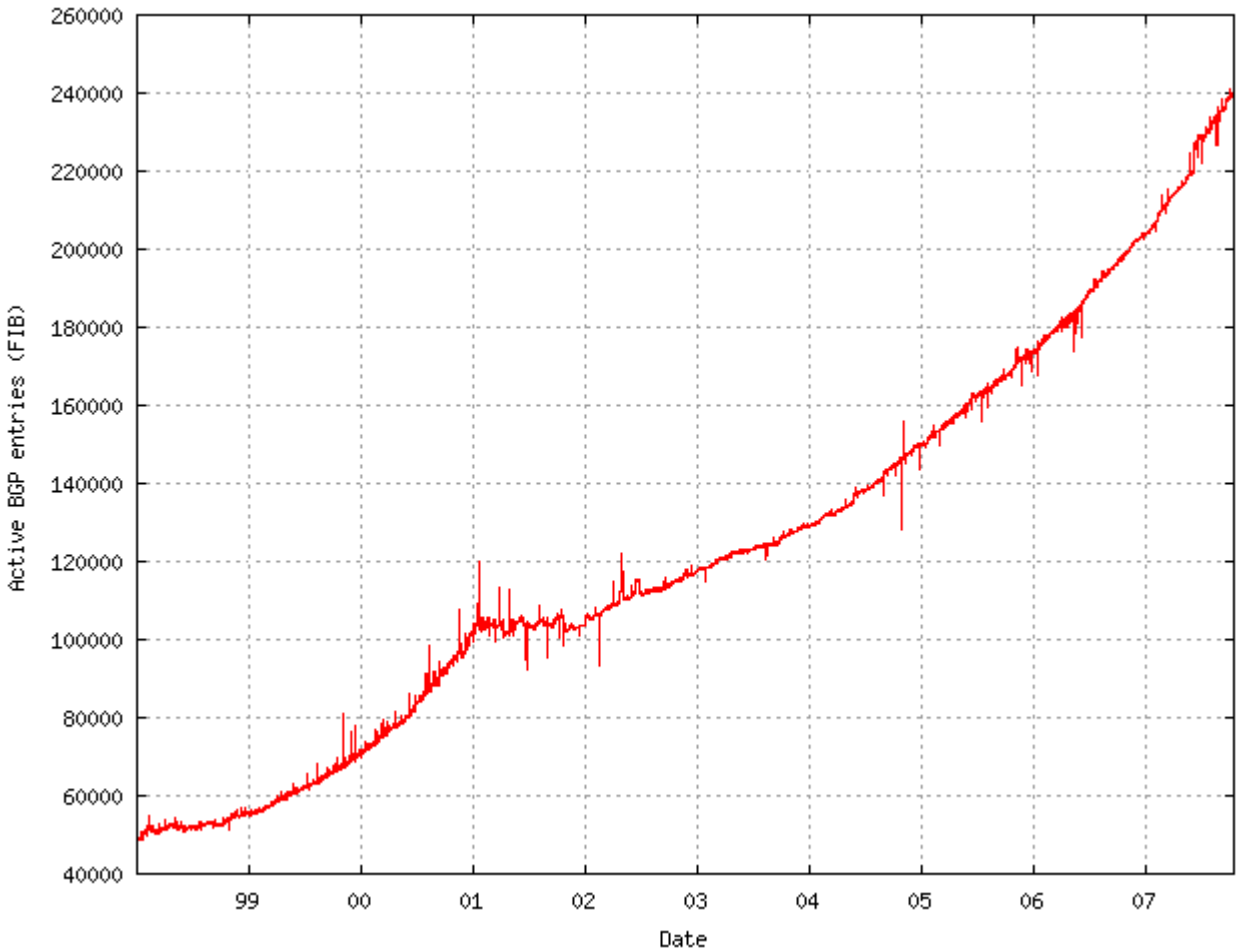
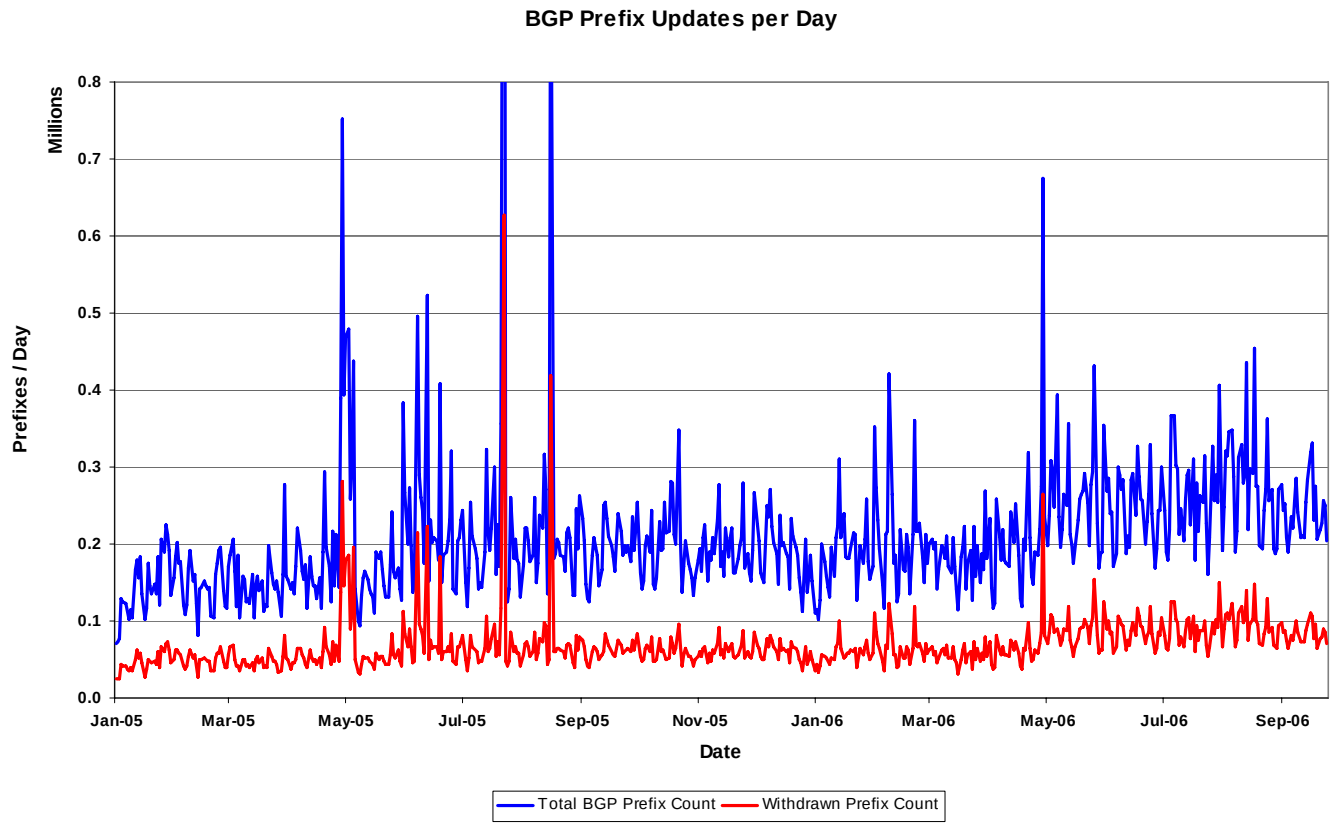
Update Damping in BGP

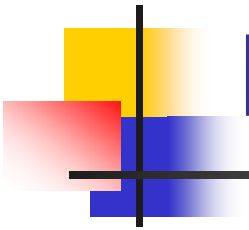
Geoff Huston
Chief Scientist, APNIC

BGP Growth: Table Size



BGP Growth: Updates (05 – 06)





Limits to Growth?

Are there practical limits to the size of the routed network ?

- limits to routing database size ?
- limits in routing update processing load ?
- practical bounds for time to reach “converged” routing states ?

Current Understandings

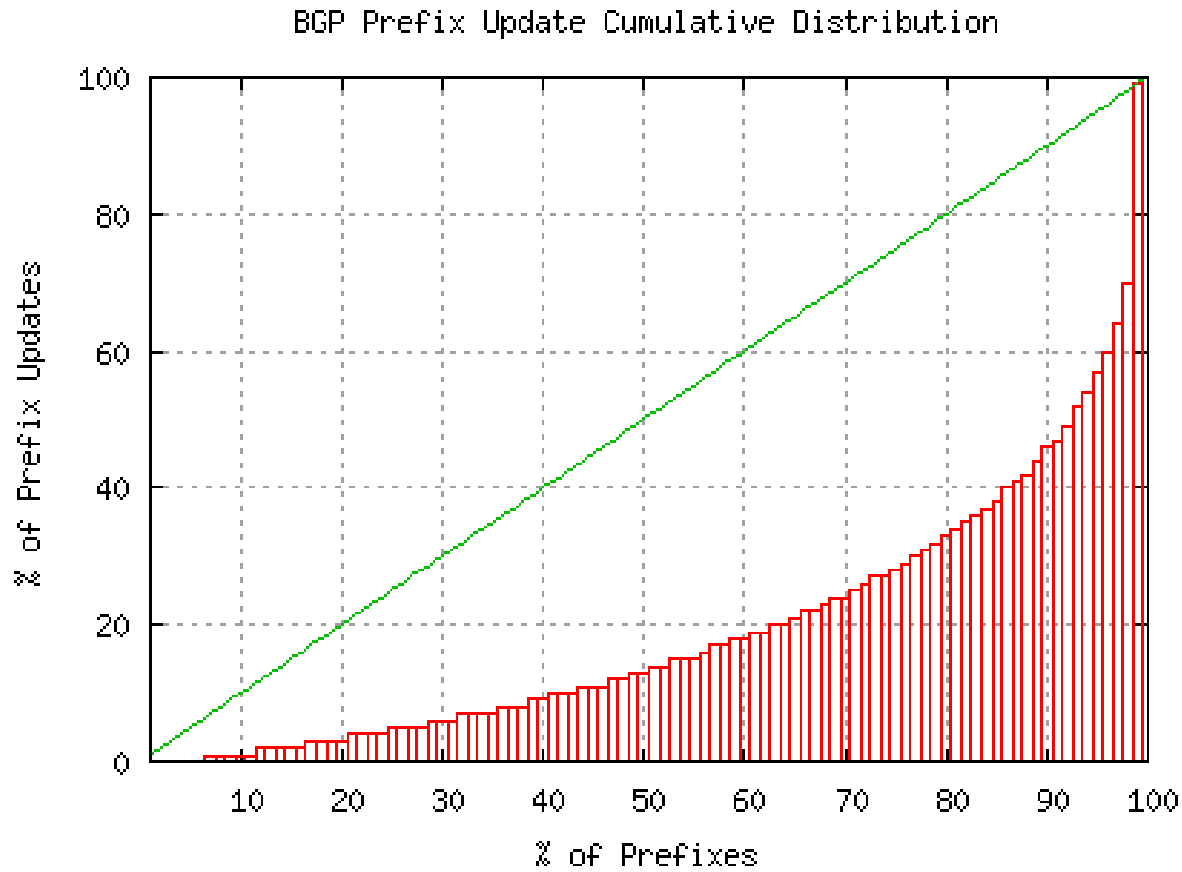
- The protocol message peak rate is increasing faster than the number of routed entries
  - BGP is a “chatty” protocol
  - Dense interconnection implies higher levels of path exploration to stabilize on best available paths
- Some concern that BGP has some practical limits in terms of size and convergence times within the bounds of currently deployed routing machinery
- Some further concern that these limits may be achieved in the near term future

Profiling BGP Load

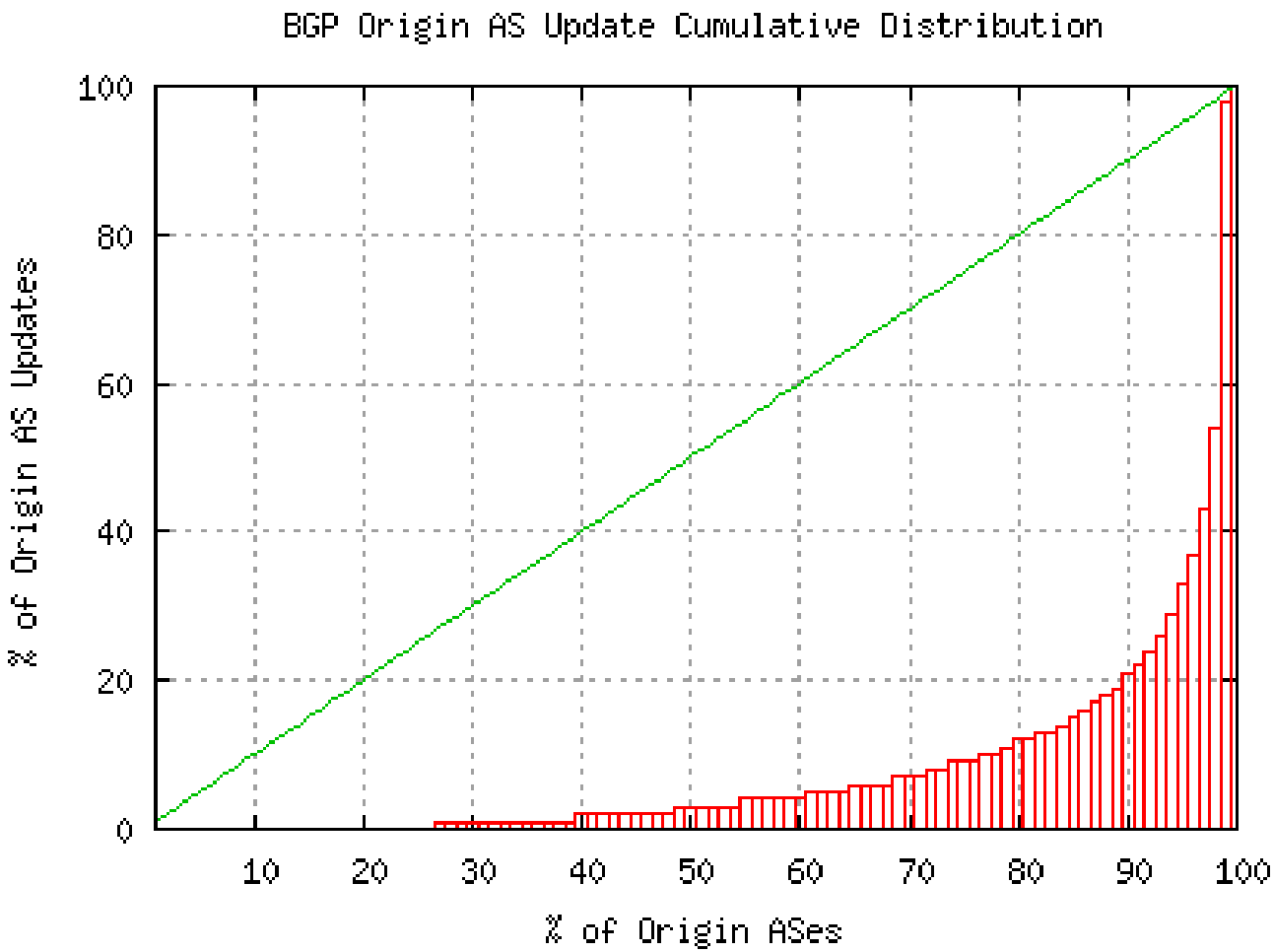
- Use a BGP monitor connected to DFZ update feeds
  - Quagga
- Log all updates
- Process logs and generate daily profile

http://bgpupdates.potaroo.net

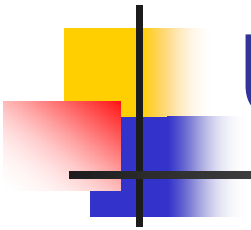
Update Distribution by Prefix



Update Distribution by Origin AS

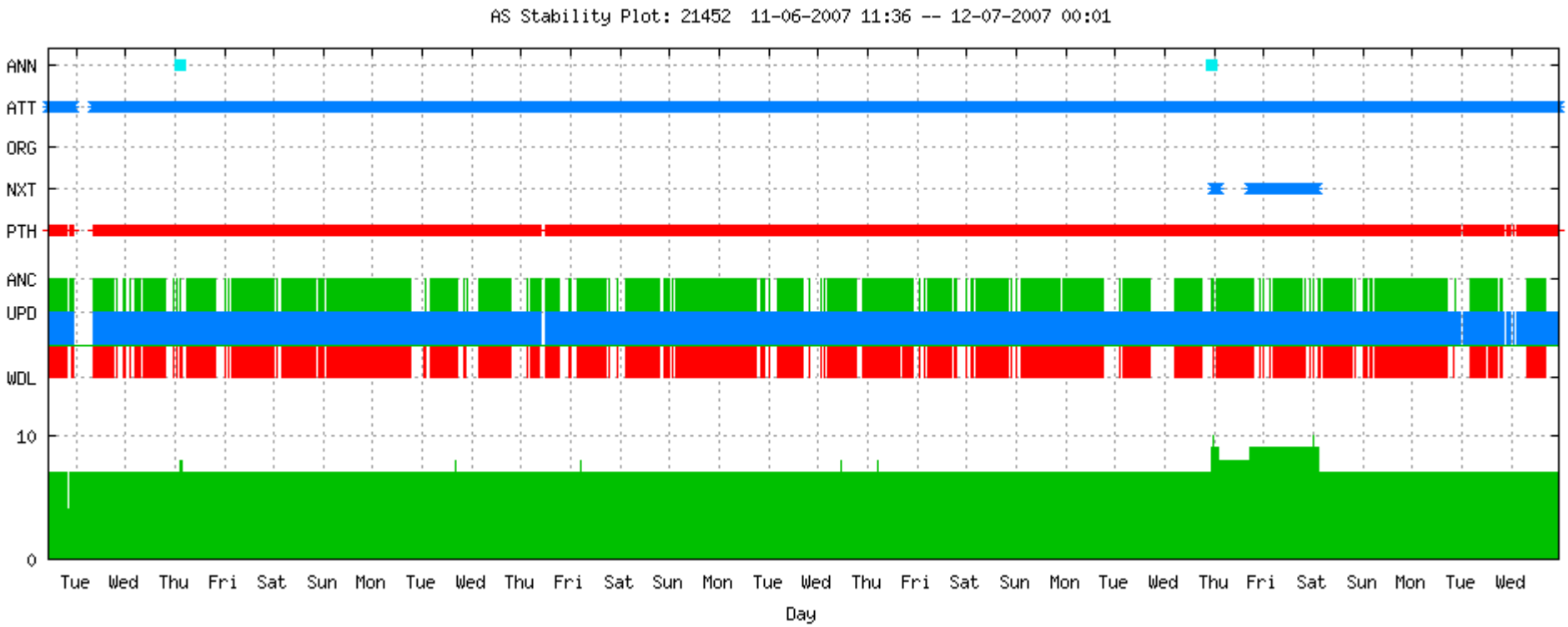


Previous Analysis of BGP Update Profile



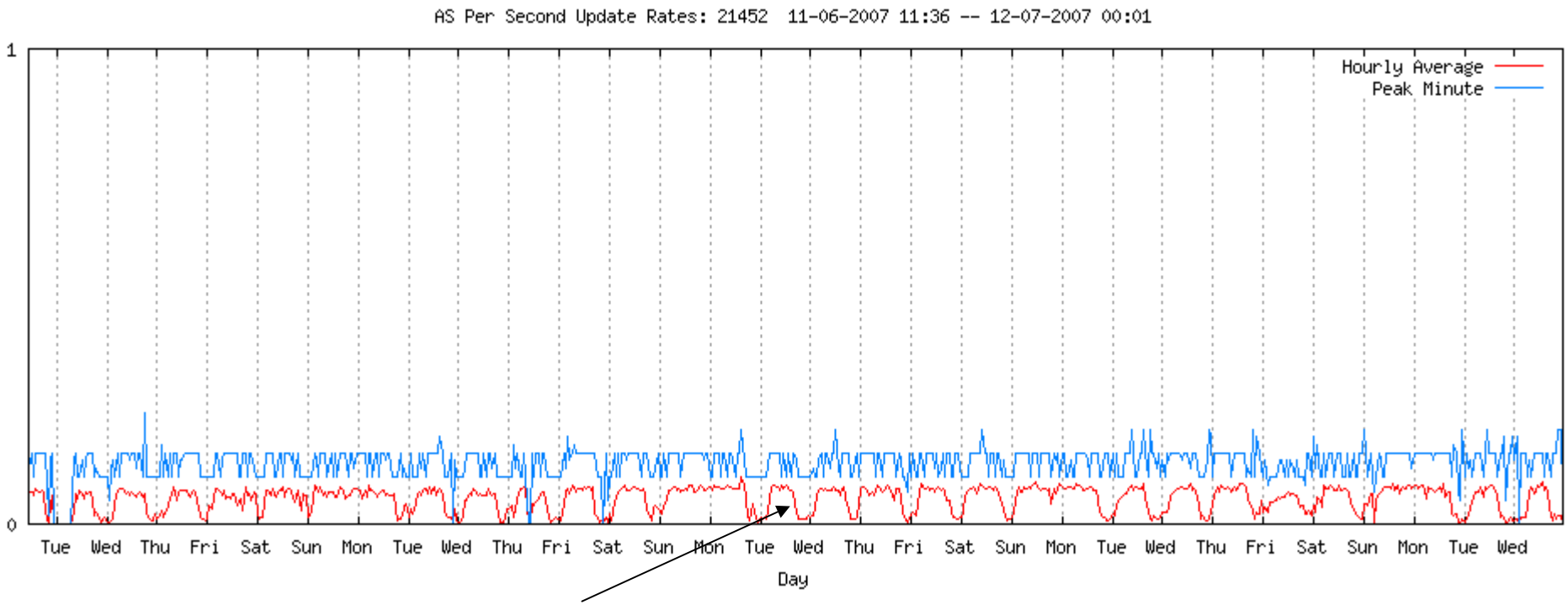
- Update load profile and convergence times do not appear to be precisely aligned to routing table size
- The BGP load profile is heavily skewed, with a small number of route objects, and a small number of origin AS's, contributing a disproportionate amount to the routing update load
  - Background load appears to be heavily related to close-to-collector routing events that affect large numbers of routed objects
  - Intense load appears to be related to close-to-origin routing events that affect small numbers of routed objects with each event
- As the network grows the highly active component of route load does not appear to grow proportionally

What's the cause here?



BGP Updates recorded at AS2.0, June 28 – July 12
AS21452

What's the cause here?

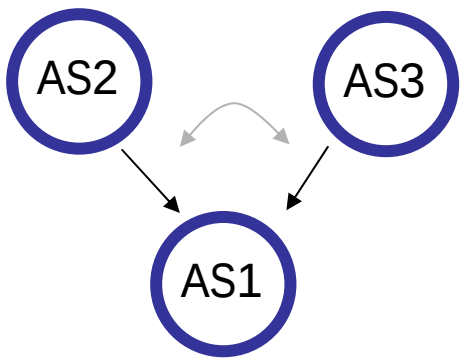


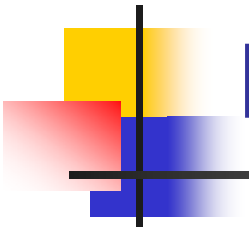
This daily cycle of updates with a weekend profile is a characteristic signature of the origin AS performing some form of load-based routing

BGP Updates recorded at AS2.0, June 28 – July 12
AS21452

Poor Traffic Engineering?

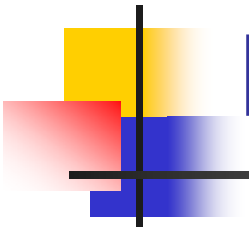
- An increasing trend to “multi-home” an AS with multiple transit providers
- Spread traffic across the multiple transit paths by selectively altering advertisements
- The use of load monitors and BGP control systems to automate the process
- Poor tuning (or no tuning!) of the automated traffic engineering process produces extremely unstable BGP outcomes!





BGP Update Load Profile

- It appears that the majority of the BGP load is caused by a very small number of unstable origination configurations, possibly driven by automated systems with limited or no feedback control
- This problem is getting larger over time
- The related protocol update load consumes routing resources, but does not change the base information state – it generally oscillates across a small set of states that do not imply local forwarding change



Mitigating BGP Update Loads

Current set of deployed “tools” to mitigate BGP update overheads:

1. Minimum Route Advertisement Interval Timer (MRAI)
2. Withdrawal MRAI Timer
3. Route Flap Damping
4. Output Queue Compression

1. MRAI Timer

- Optional timer in BGP
  - ON in Ciscos (30 seconds)
  - OFF in Junipers (0 secconds)
- Suppress the advertisement of successive updates to a peer for a given prefix until the timer expires
- Commonly implemented as suppress ALL updates to a peer until a per-peer MRAI timer expires

2. Withdrawal MRAI TIMER

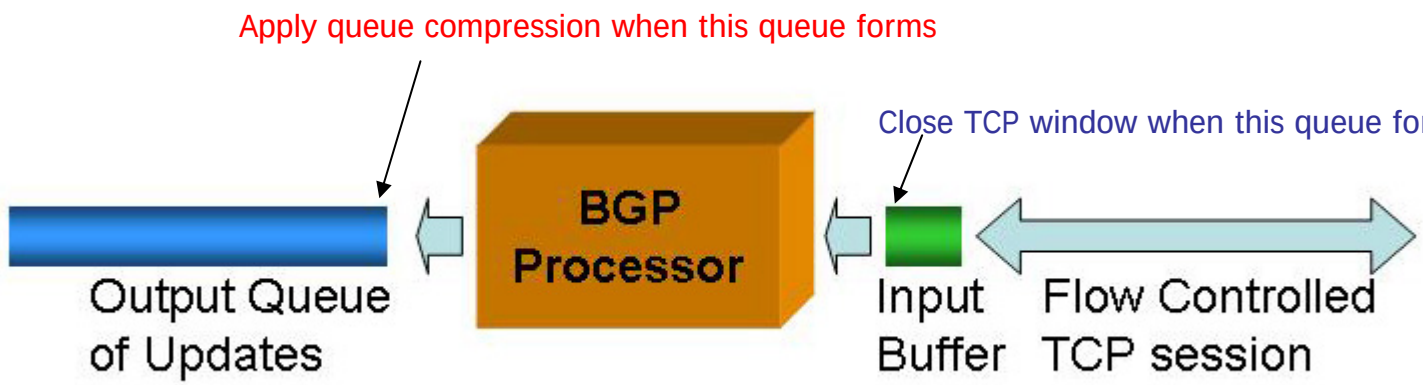
- Variant on MRAI where withdrawals are also time limited in the same way as updates

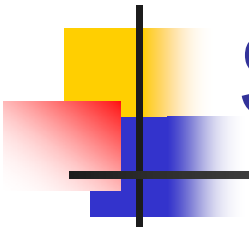
3. Route Flap Damping

- RFD attempts to apply a heuristic to identify noisy prefixes and apply a longer term suppression to update propagation
- Uses the concept of a “penalty” score applied to a prefix learned from a peer
  - Each update and withdrawal adds to the score
  - The score decays exponentially over time
  - If the score exceeds a suppress threshold the route is damped
  - Damping remains in place until the score drops below the release threshold
  - Damping is applied to the adj-rib-in

4. Output Queue Compression

- BGP is a rate-throttled protocol (due to TCP transport)
  - A process-loaded BGP peer applies back pressure to the 'other' side of the BGP session by shutting down the advertised TCP recv window
  - The local BGP process may then perform queue compression on the output queue for that peer, removing queued updates that refer to the same prefix





Some Observations

- RFD – long term suppression
  - Route Flap damping extends convergence times by hours with no real benefit offset
- MRAI – short term suppression
  - MRAI variations in the network make path exploration noisier
  - Even with piecemeal MRAI deployment we still have a significant routing load attributable to Path Exploration
- Output Queue Compression
  - Rarely triggered in today's network!

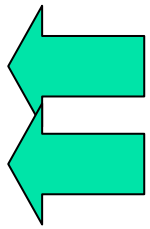
BGP Update Types

	Code	Description
Announced-to-Announced Updates	AA+	Announcement of an already announced prefix with a longer AS Path (update to longer path)
	AA-	Announcement of an announced prefix with a shorter AS Path (update to shorter path)
	AA0	Announcement of an announced prefix with a different path of the same length (update to a different AS Path of same length)
	AA*	Announcement of an announced prefix with the same path but different attributes (update of attributes)
	AA	Announcement of an announced prefix with no change in path or attributes (possible BGP error or data collection error)
Withdrawn-to-Announced Updates	WA+	Announcement of a withdrawn prefix, with longer AS Path
	WA-	Announcement of a withdrawn prefix, with shorter AS Path
	WA0	Announcement of a withdrawn prefix, with different AS Path of the same length
	WA*	Announcement of a withdrawn prefix with the same AS Path, but different attributes
	WA	Announcement of a withdrawn prefix with the same AS Path and same attributes
Announced-to-Withdrawn	AW	Withdrawal of an announced prefix
Withdrawn-to-Withdrawn	W W	Withdrawal of a withdrawn prefix (possible BGP error or a data collection error)

April 2007 BGP Update Profile

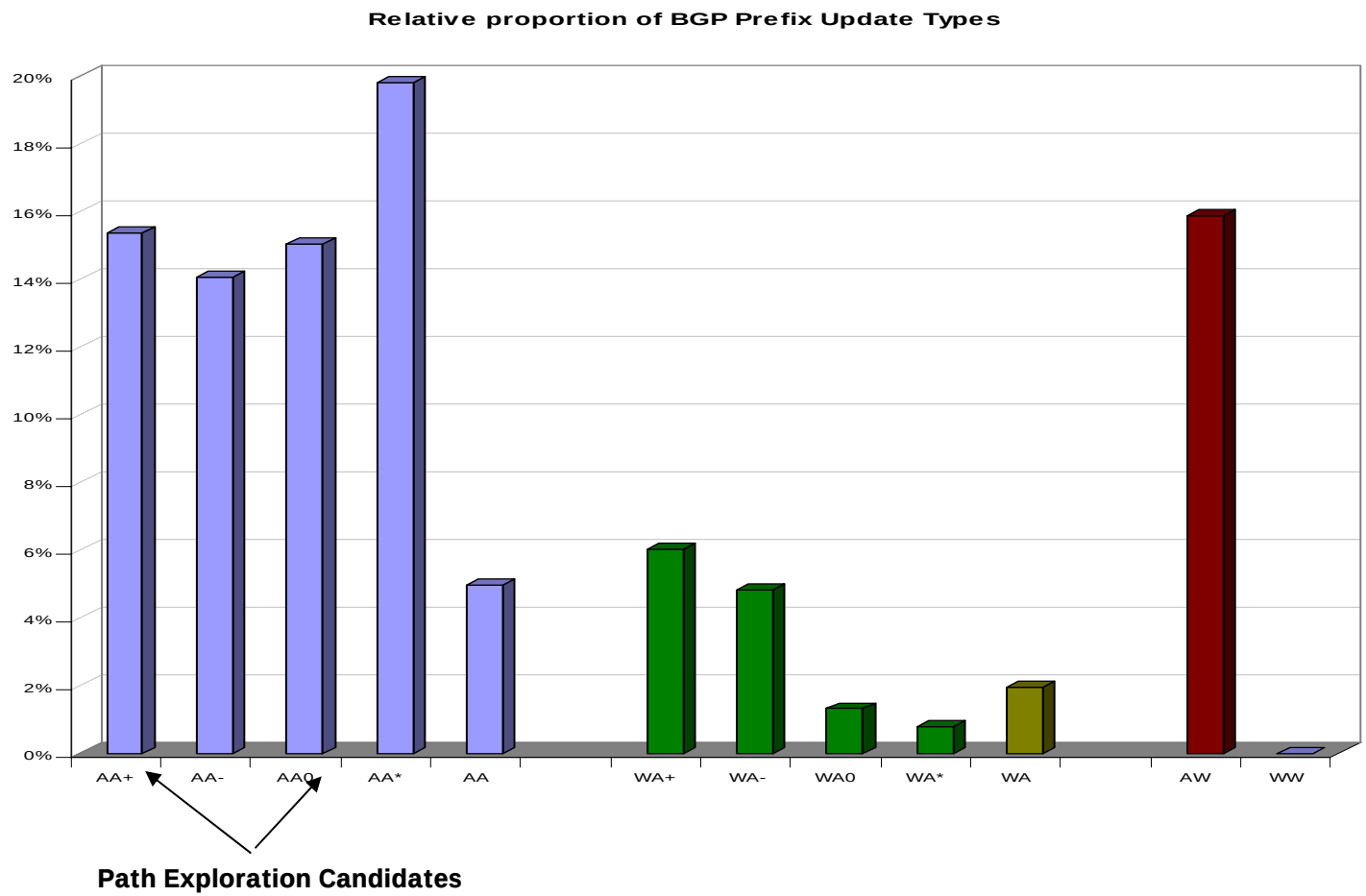
Totals of each type of prefix updates, using a recording of all BGP updates as heard by AS2.0 for the month of April 2007

Code	Count
AA+	607,093
AA-	555,609
AA0	594,029
AA*	782,404
AA	195,707
WA+	238,141
WA-	190,328
WA0	51,780
WA*	30,797
WA	77,440
AW	627,538
W W	0

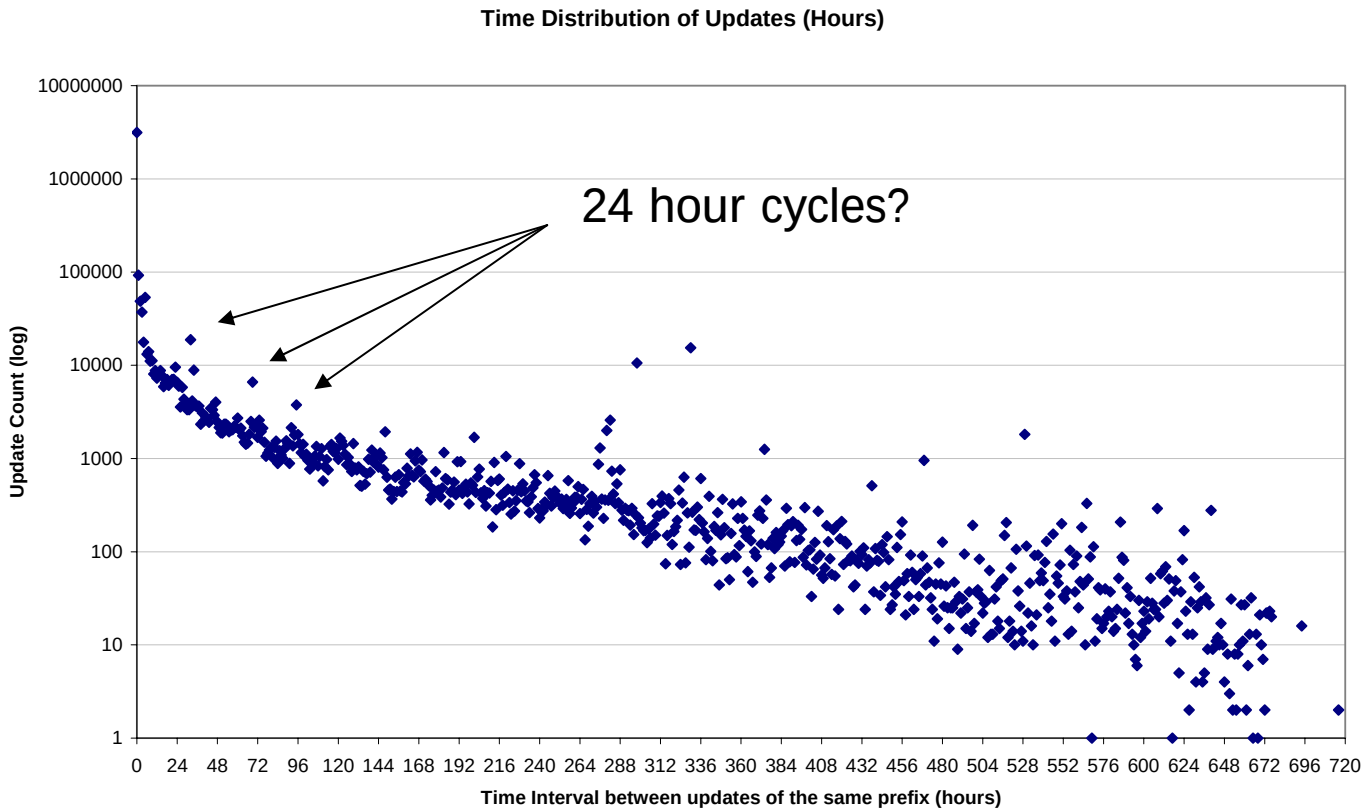


BGP Path Exploration?

BGP Update Profile

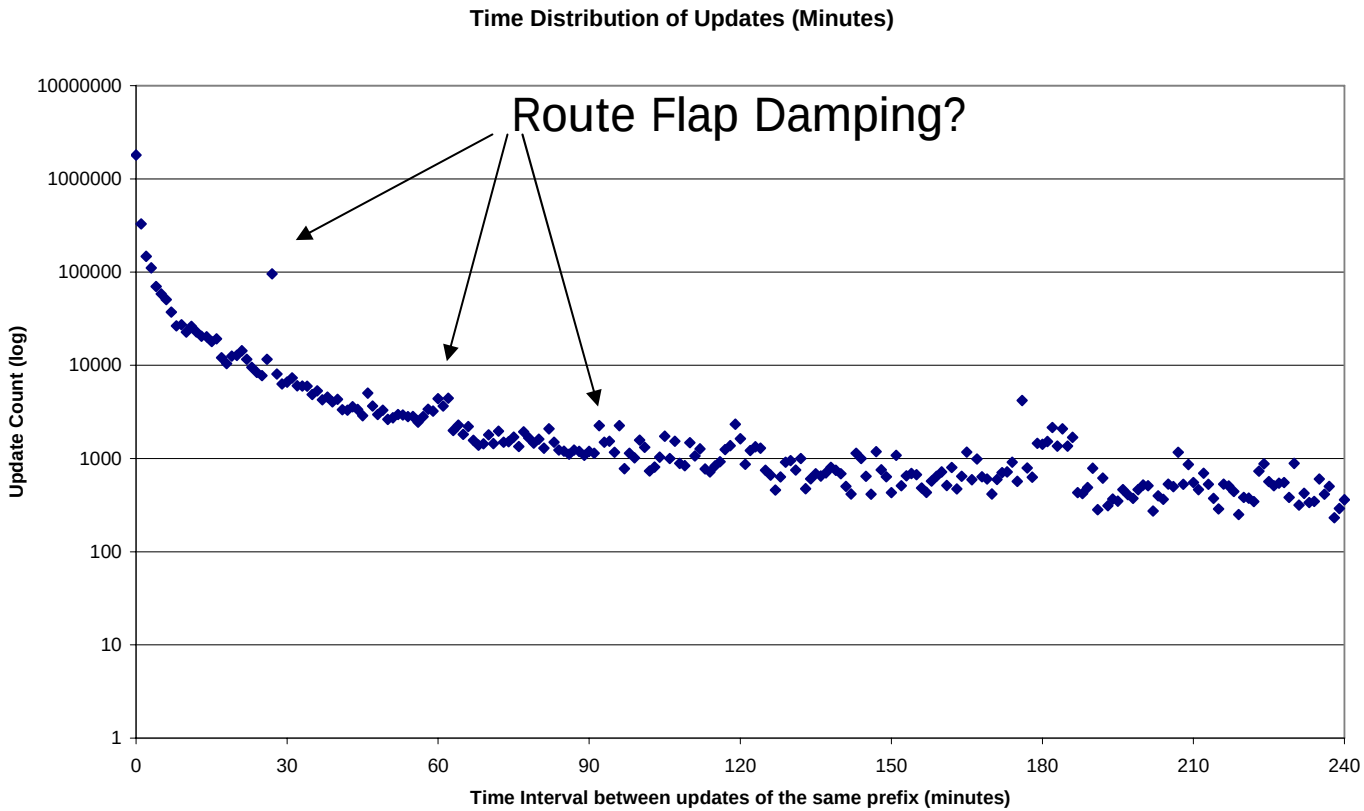


Time Distribution of Updates



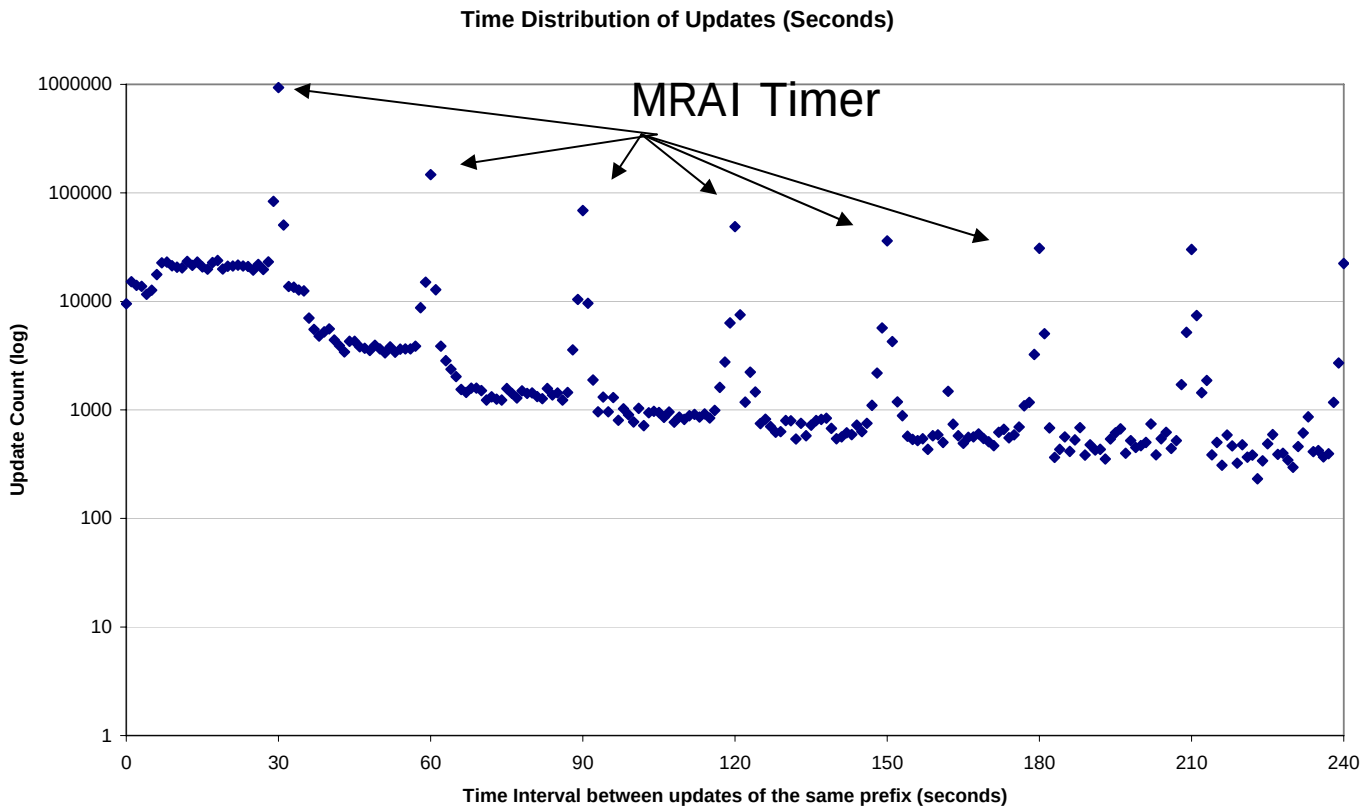
Elapsed time between received updates for the same prefix - hours

Time Distribution of Updates



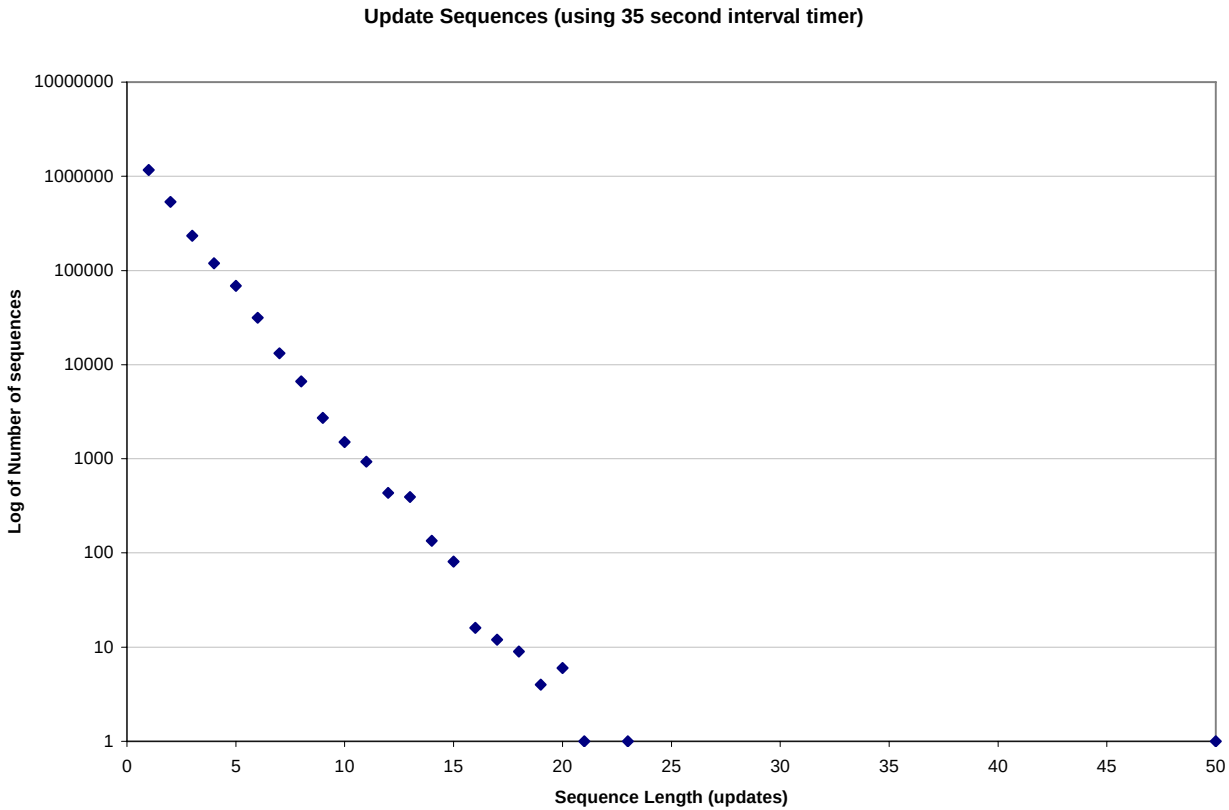
Elapsed time between received updates for the same prefix - minutes

Time Distribution of Updates



Elapsed time between received updates for the same prefix - seconds

Update Sequence Length Distribution



A “sequence” is a set of updates for the same prefix that are separated by an interval <= the sequence timer (35 seconds)

Path Exploration Damping (PED)

- A prevalent form of path hunting is the update sequence of increasing AS path length, followed by a withdrawal, all closely coupled in time

  {AA+, AA0, AA} *, AW

The AA+, AA0 and AA updates are intermediate noise updates in this case representing transient routing states

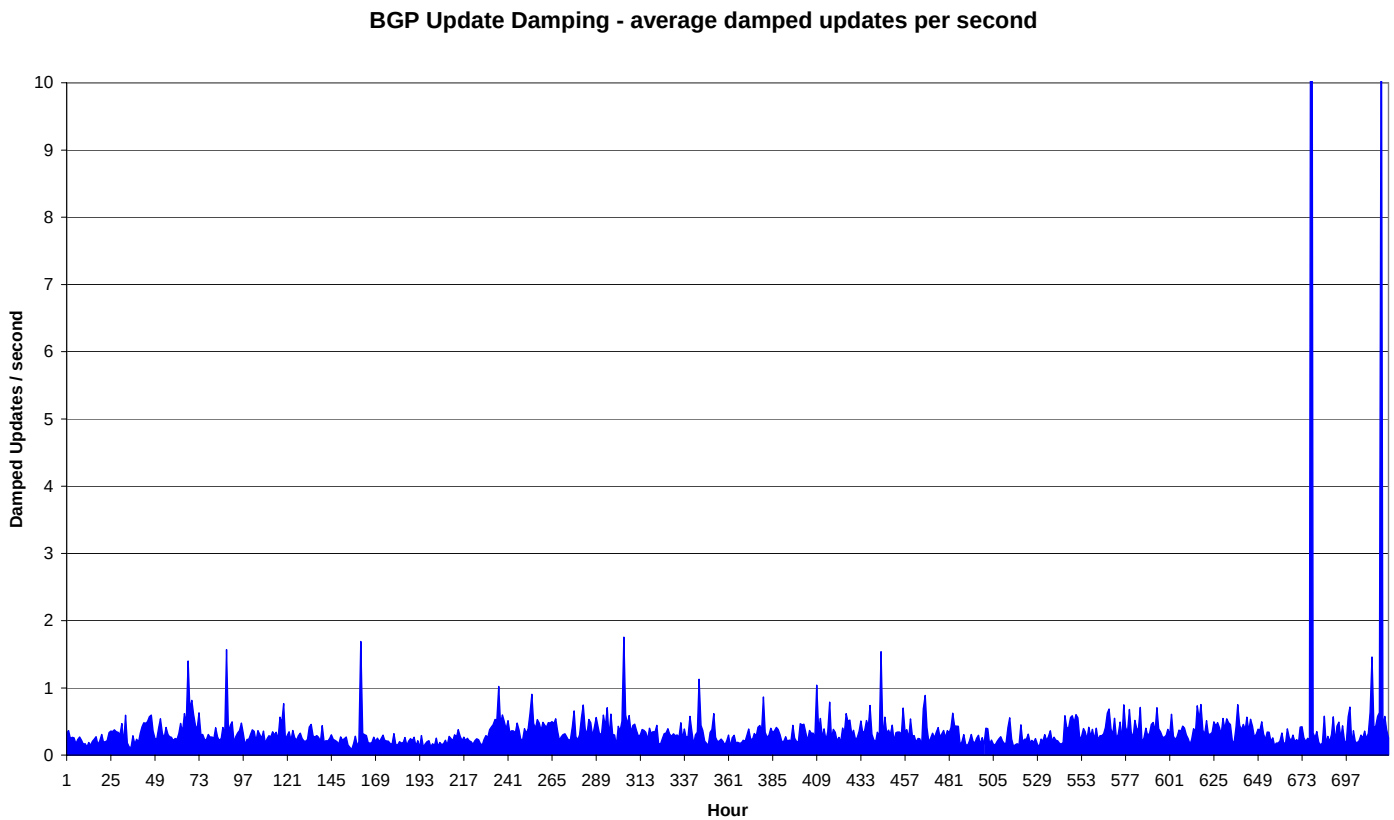
Can these updates be locally suppressed for a short interval to see if they are path of a BGP Path Exploration activity?

The suppression would hold the update in the local output queue for a fixed time interval (in which case the update is released) or the update is further updated by queuing a subsequent update (or withdrawal) for the same prefix

Path Exploration Damping

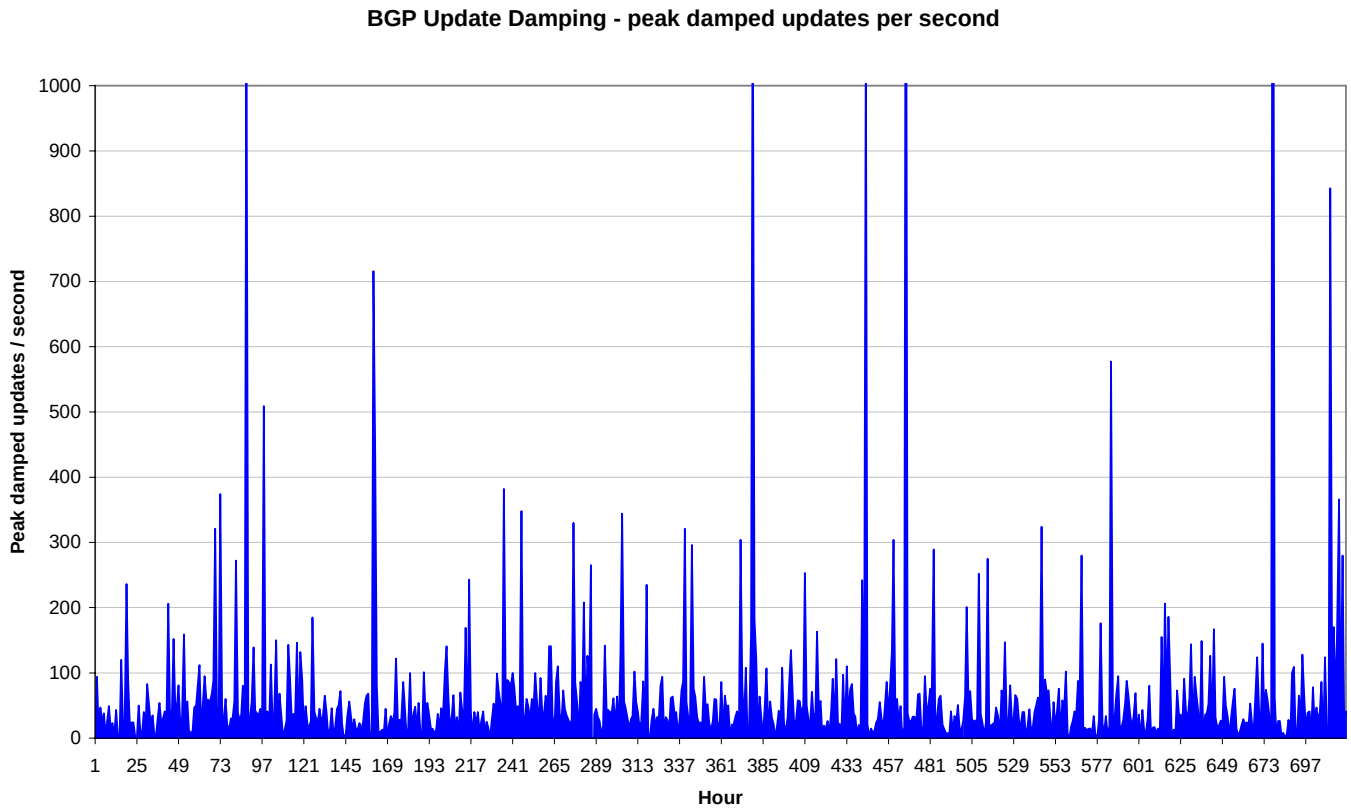
- Apply a 35 second MRAI timer to AA+, AA0 and AA updates queued to eBGP peers
- No MRAI timer applied to all other updates and all withdrawals
  - 35 seconds is used to compensate for MRAI-filtered update sequences that use 30 second interval

PED Results on BGP data



Path Exploration Damping applied to
BGP updates recorded at AS2.0, June 28 – July 12

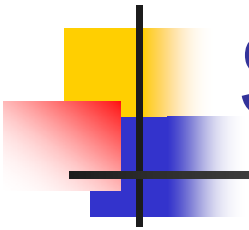
PED Results on BGP data



Path Exploration Damping applied to
BGP updates recorded at AS2.0, June 28 – July 12

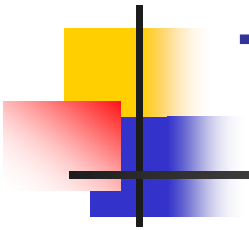
PED Results on BGP data

- 21% of all updates in the collection period would've been eliminated by Path Exploration Damping
- Average update rate for the month would fall from 1.60 prefix updates per second to 1.22 prefix updates per second
- Average peak update rates fall from 355 to 290 updates per second



Summary

- Much of the update processing load in BGP is in processing non-informative intermediate states caused by BGP Path Exploration
- Existing approaches to suppress this processing load appear to be too coarse to be very effective
- Some significant leverage in further reducing BGP peak load rates can be obtained by applying a more selective algorithm to the MRAI approach in BGP, attempting to isolate Path Exploration updates by the use of local heuristics



Thank You

Questions?