The Intra-domain BGP Scaling Problem

Danny McPherson danny@arbor.net
Shane Amante shane@level3.net
Lixia Zhang lixia@cs.ucla.edu

Objective

- Outline issues with BGP scalability caused by network path explosion
- Concerted focus on intra-domain routing
- Broad array of areas for concern/improvements
  - BGP protocol constraints
  - Implementation issues
  - Network architecture
  - External effects
- Conclusions

Agenda

- Objective
- Background, BGPisms
- What breaks first?
- A look Route Reflection
- Network Architecture Considerations
- Miscellaneous
- Conclusions

It’s All About Perspective!

- Most, if not all, of BGP scalability, stability analysis today is based on one or more views of **external** BGP sessions
- Internal BGP dynamics are very different, and very dependent on network design, vendor implementations, etc..
- More study of internal BGP views at various levels of internal BGP hierarchy (if exists) necessary (some underway)

BGPisms, Background Slides

Topology: The Bogey Man!

- BGP behavior dependent on topology
- Making connectivity (internal & external) richer SHOULD result in improved reliability
  - but instead may cause convergence delays of multiple minutes when routes flap
  - even in the absence of flap dampening
- This is a path hunting problem which won't go away until it is solved
  - Until then, it causes escalation of BGP update counts and convergence delay

eBGP - iBGP - eBGP

- Multiple offset update receipt or processing variance can trigger withdraw + new announcement where just new announcement would have otherwise been sufficient
- Can cause cascade of unnecessary path hunting
- Rich topological connectivity (internal or external) can result in badly behaved path selection and announcement, in race conditions prior to new correct state while withdrawals flood the global DFZ
- Behaves badly because of limited local knowledge - with exponential badness based on N^M, (where N is number of paths from a given AS to the end site, and M is the number of ASes in the path). M typically 4-6, N can be double digits

Minimum Route Advertisement Interval (MRAI)

- Needed to prevent runaway melt-down of router CPUs
- Has adverse effects when doing path hunting (legitimately)
- Need for negotiated and configurable timers for external and internal BGP, per peer and AFI/SAFI - environment-specific
- Interaction between successive run timers whose values differ can make things worse!
- Common default MRAI - **0 seconds**

BGP Impacting Factors

- Only best routes sent (currently)
- Even if multiple routes sent, only best installed in FIB
- Lack of information on alternative paths prevents look-ahead, also leads to update flooding whenever the best path changes
- Regardless of other improvements, delay of updates will be bounded above by speed of light in fiber (Nominal Velocity of Propagation ~200 km/s) && packet regeneration time
- Intermediate states in path hunting may be (and often are) completely bogus

IGP & BGP Interaction

- IGPs typically carry only NEXT_HOP and session reachability information for BGP
- iBGP NEXT_HOP is often ingress router loopback, ideally keep external interfaces out of IGP for stability reasons
- No use of IGP/BGP synchronization, using this would mean each router has to have full set of BGP routes in their IGP in order to preserve destination reachability
- IGP metrics often uses to populate BGP MEDs - or determine best ‘hot-potato’ location

iBGP Route Advertisement Rules

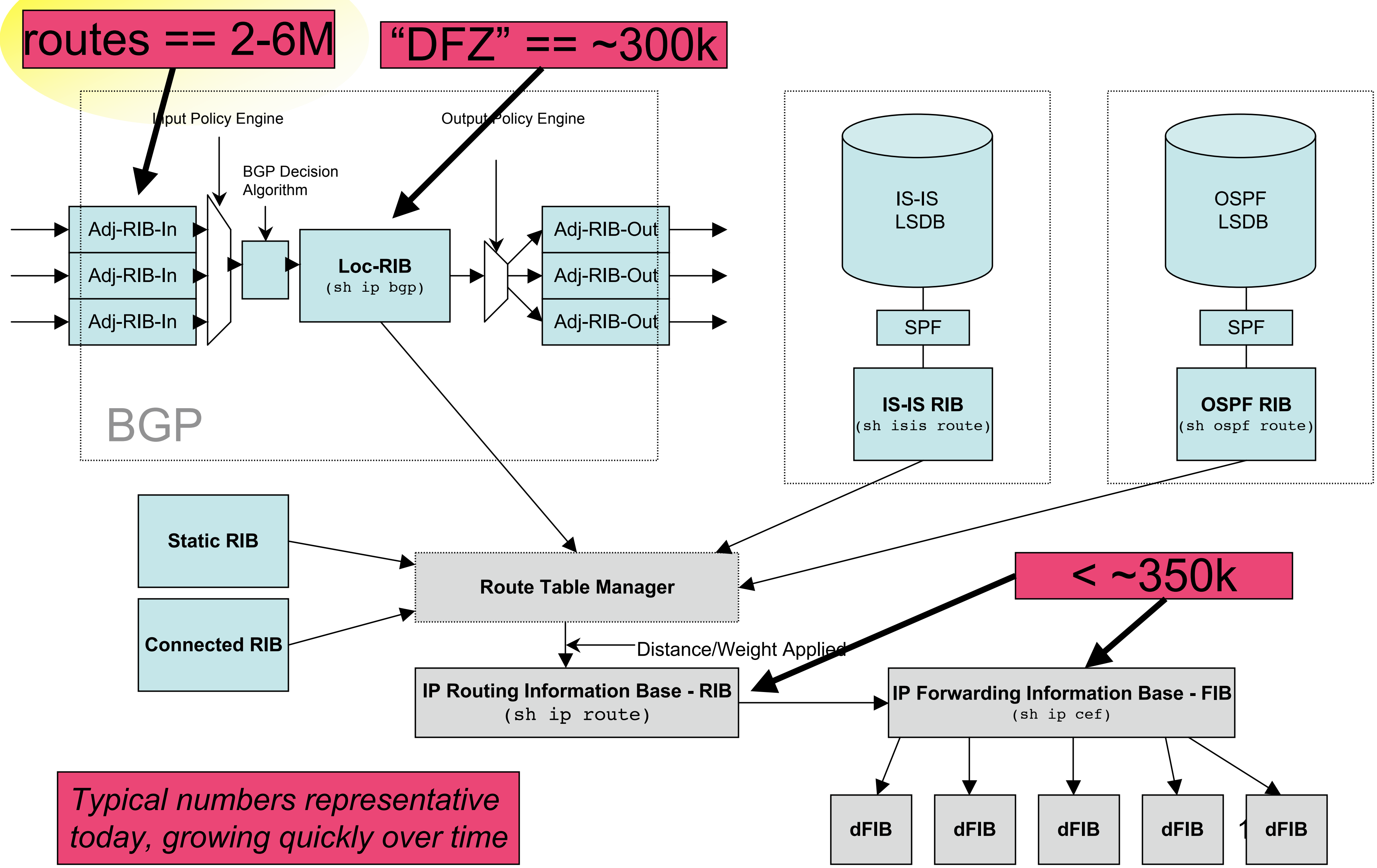
- iBGP rule: a route learned from one iBGP speaker can’t be propagated to another iBGP speaker - else routing information loops will occur
- As such, iBGP full mesh required:: N(N-1)/2 sessions
- This iBGP rule can be relaxed, however (with introduction of new path vectors)
  - **route reflection**; introduces cluster ID, cluster lists and originator ID attributes (serve as path vector)
  - **AS Confederations**; partition AS into sub-ASes, full mesh still required within sub-AS, introduces AS_CONFED_* attributes (serve as path vector)
  - Some ISPs use both RR and Confederations, some one, some neither
  - RRs can be used hierarchically within a routing domain

What Breaks First?

What Breaks First?

- Considerable amount of focus on “DFZ size” - the number of unique ***prefixes*** in the global routing system - ultimate FIB size **is** considerable issue
- However, second issue is number of ***routes (prefix, path attributes)*** and frequency of change
- More routes == more state, churn; effects on CPU, RIBs && FIB
- *Routes growing more steeply than unique prefixes/DFZ*

Conceptual Router Architecture



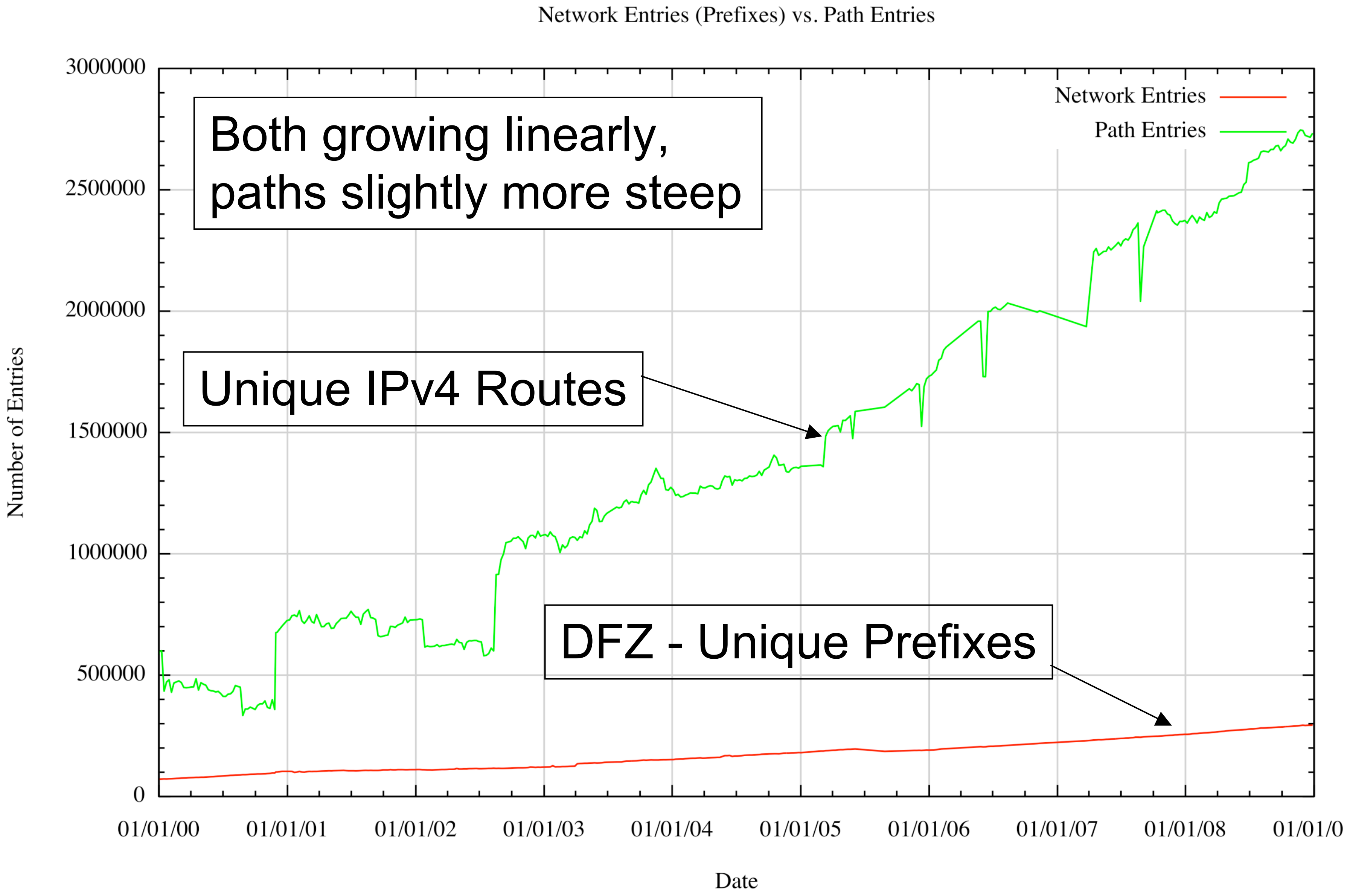
routes == 2-6M

“DFZ” == ~300k

< ~350k

Typical numbers representative today, growing quickly over time

Growth: Prefixes v. Routes

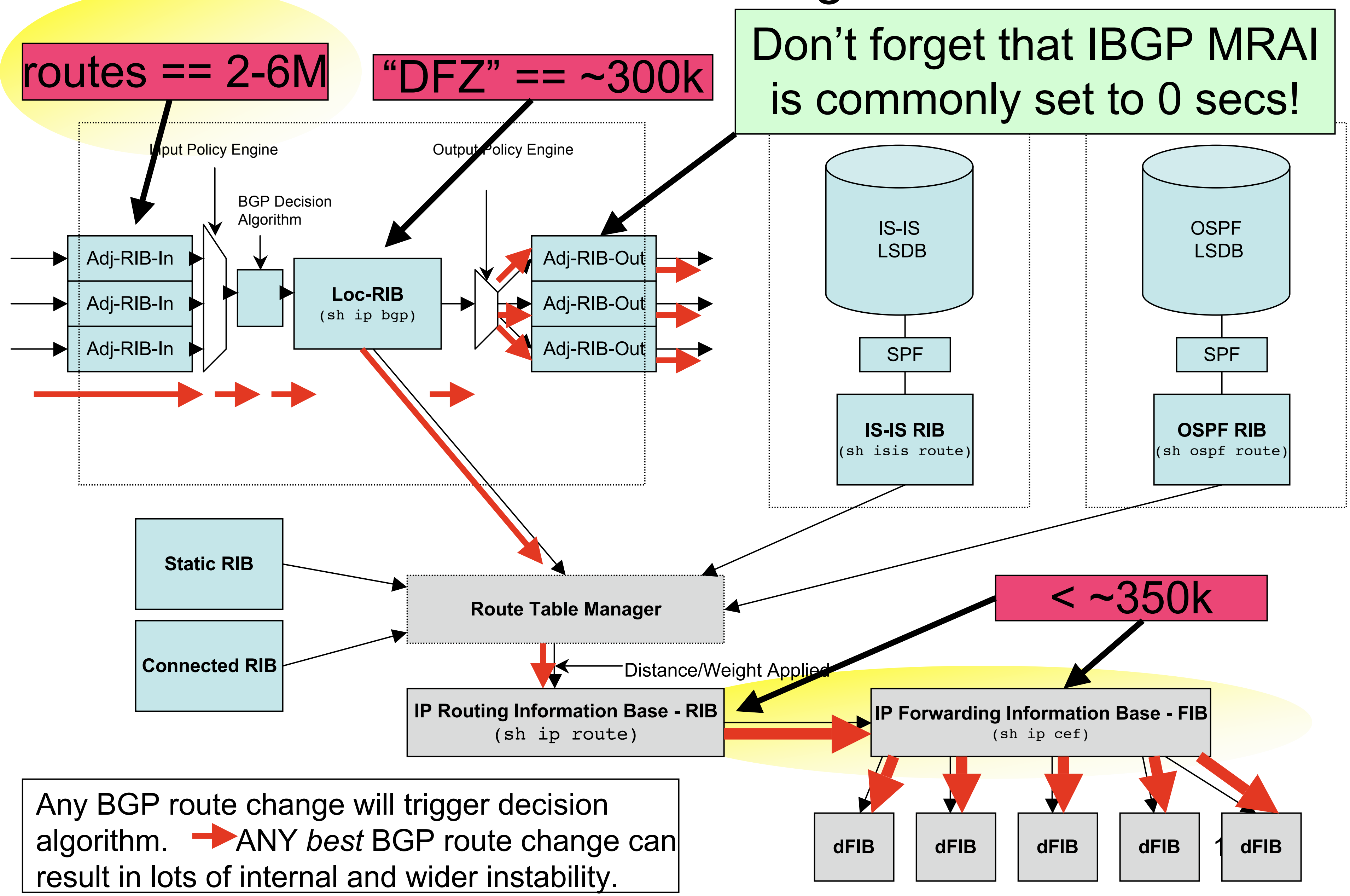


Both growing linearly, paths slightly more steep

Unique IPv4 Routes

DFZ - Unique Prefixes

ANY *Best* Route Change Means….



routes == 2-6M

“DFZ” == ~300k

Don’t forget that IBGP MRAI is commonly set to 0 secs!

< ~350k

Any BGP route change will trigger decision algorithm. → ANY *best* BGP route change can result in lots of internal and wider instability.

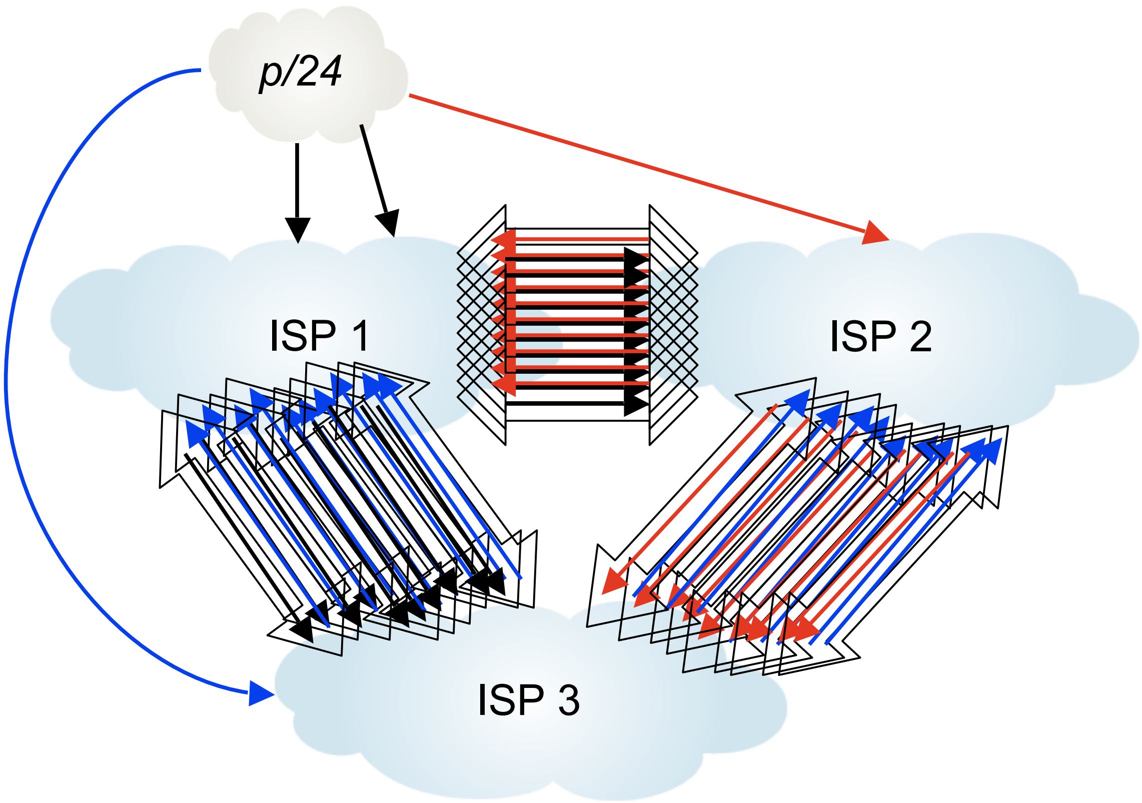
Why is # of unique routes increasing faster than # of prefixes?

- Primarily due to *denseness of interconnection outside of local routing domain*
  - Increased multi-homing from edges
  - Increased interconnection within core networks
- Each new unique prefix brings multiple unique routes into the system
- Function of routing architecture - internal BGP rules, practical routing designs, etc..
- More routes result in extraneous updates and other instability not necessarily illustrated in RIB/FIB changes

External Interconnection Denseness

- More networks interconnecting directly to avoid transit costs, reduce transaction latency, forwarding path security (e.g., avoid hostile countries / “cyberlock”),
  - More networks building their own backbones (e.g., CDNs), have presence in multiple locations
  - More end-sites and lower-tier SPs provisioning additional interconnections
  - SPs adding more interconnections in general to local traffic exchange and accommodate high-bandwidth capacity requirements
  - The “peer with everybody” paradigm
- Increased interconnections made feasible by excess fiber capacity and decreasing cost, offset transit costs
- More interconnections means more unique routes for a given prefix

External Interconnection Denseness



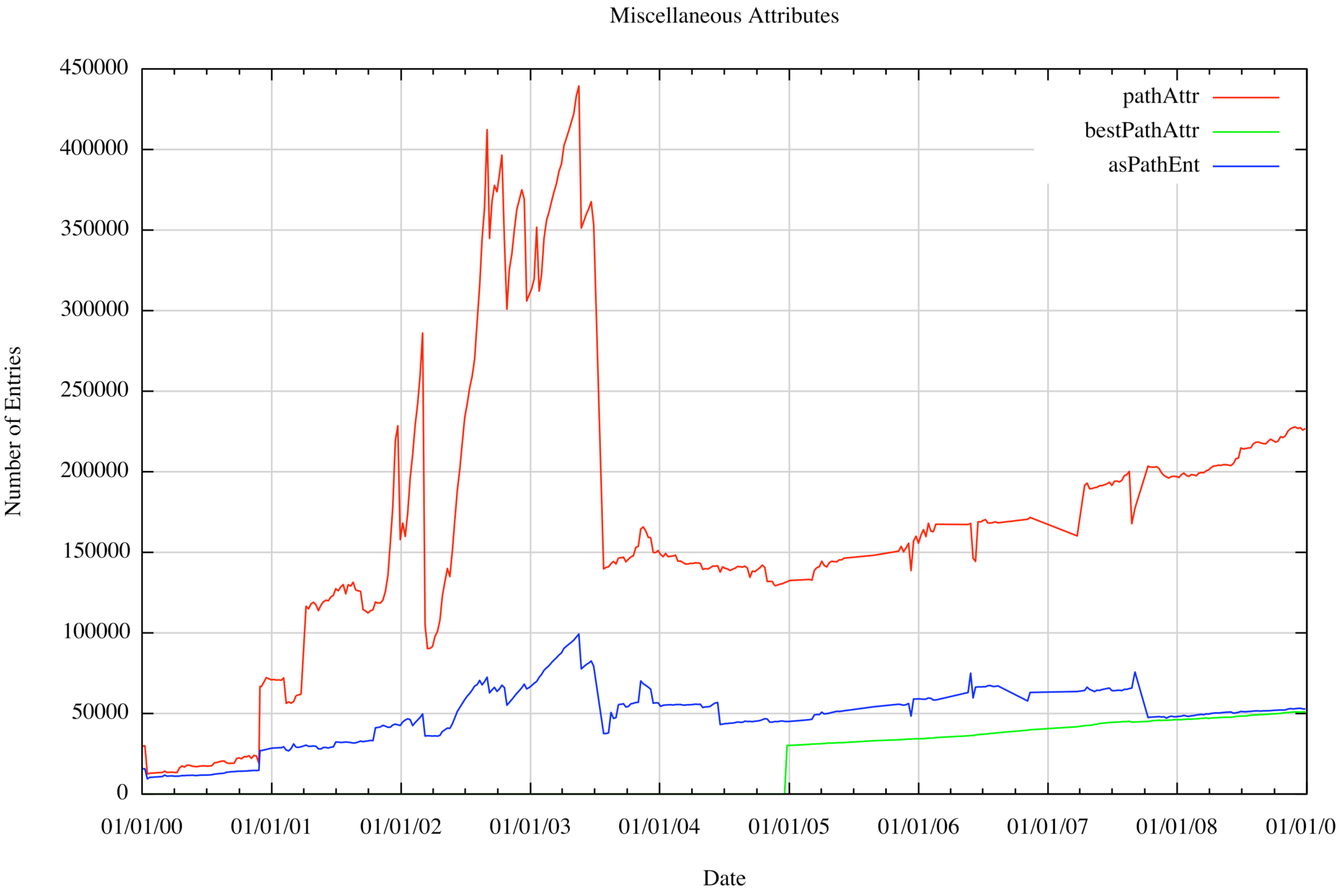
ISP1 - one unique prefix (*p*), 22 routes total on PE routers

- Consider N ASes: if an edge AS E connects to one of the N ASes, each AS has (N-1) paths to each prefix *p* announced by E
- When E connects to n of N ASes, each AS has at least n*N routes to *p*
  - In general the total number of routes to p can grow super-linearly with n
  - Edge AS multi-homing n times to the same ISP does NOT have this effect on adjacent ISPs
- It's common for ISPs to have 10 or more interconnects with other ISPs
  - when E connects to n ISPs, each ISP likely to see n*10 routes for p announced by E
- New ISPs in core, or nested transit relationships, often exacerbate the problem

Effects of Attribute Growth

- More unique attributes means more unique routes
- Results in less efficient update packing; more BGP updates, more BGP packets
- Common expanding attribute types
  - AS path
  - Communities
  - MEDs
  - Others (AFI/SAFIs, route reflection attributes)

Unique Attribute Growth



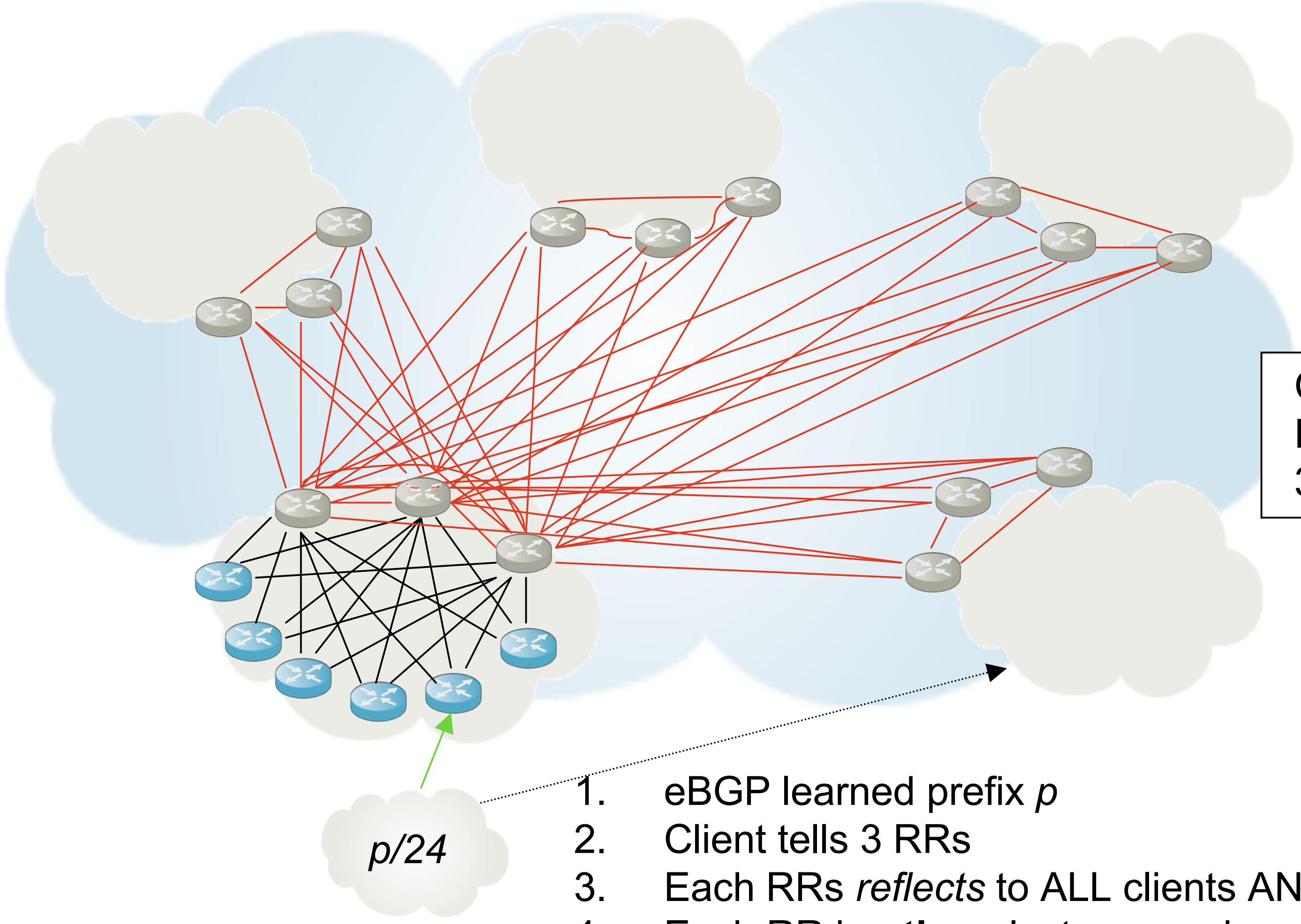
Copyright 2009 Level 3 Communications, Inc. All rights reserved.

A Peek Into Route Reflection

Route Reflection

- While route reflection (RR) does provide implicit aggregation by only propagating single “best route”
- RR may result in additional routing system state
- RR guidelines recommend that RR topology be congruent to IP network topology to avoid forwarding loops - difficult constraint in real networks (in general, RRs should not peer through client)
- Often 2-6 RRs per cluster, mirrors core or aggregation router physical or network layer interconnection topology
- Some ISPs have 3-4 tiers of RRs, **most just one**
- RRs within cluster typically fully meshed
- A RR client connects to multiple RRs
- Absent other attributes, closest eBGP learned route often preferred - result is that each RR advertises one route to all other BGP speakers at same “tier”
  - E.g., 5 interconnections with another AS, with 3 RRs per cluster, could result in 15 routes per RR for a single prefix!

Route Reflection Illustrated



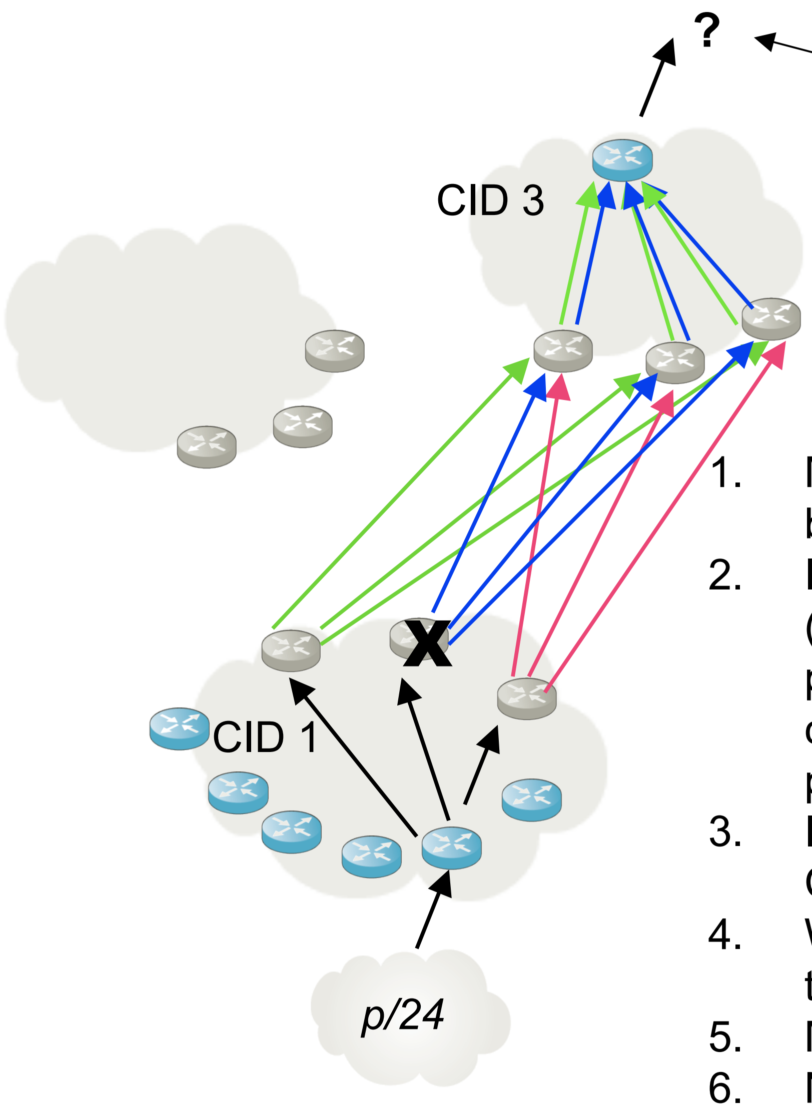
Client-Client Reflection
Full iBGP RR mesh
3 RRs per Cluster

1. eBGP learned prefix *p*
2. Client tells 3 RRs
3. Each RRs *reflects* to ALL clients AND normal e|iBGP peers
4. Each RR in **other** clusters now has 3 routes for prefix
5. IF edge AS multi-homes to another cluster, each RR will have 6 routes for prefix, etc..
6. ISPs commonly interconnect at 10 or more locations

RRs and Gratuitous Updates

- An RR crashes or a link failure changes network view of best path to BGP next hop
- New BGP route will be propagated to all BGP speakers because of change in RR cluster list, even if next hop and all other attributes and reachability are unchanged.
- Can occur with single or multiple RR tiers, can occur with common or unique cluster IDs
- When RR or link is available again, transitioning back to previous best path results in more BGP updates
- Other reasons for extraneous updates, research paper in the works w/Level(3), UCLA, Arbor
- An “avoid transition” mechanism is desirable for cluster lists of same length if all other attributes remain the same

Extraneous Updates



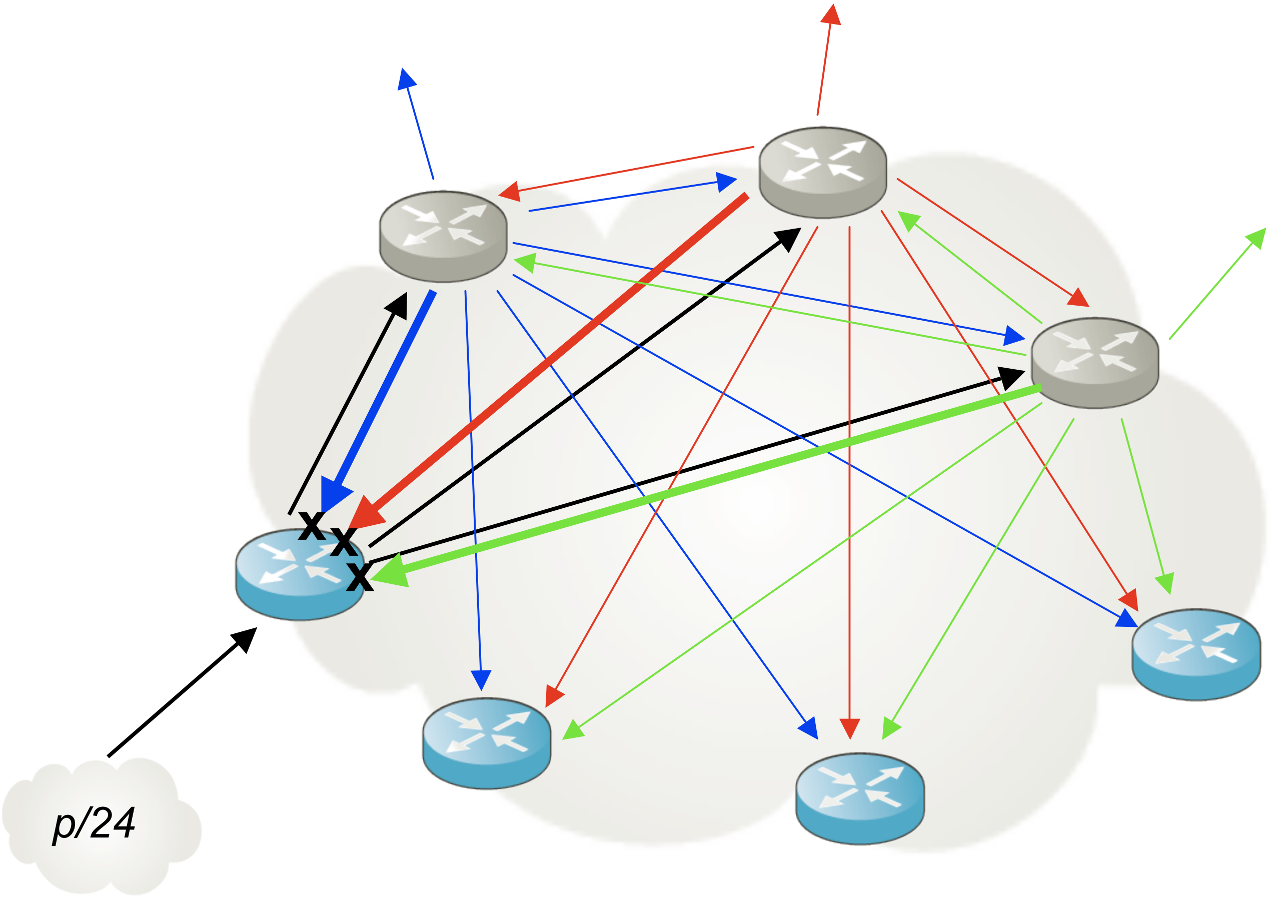
Duplicate external announcements, Flap dampening state per prefix, duplicates penalized accordingly

1. Middle RR in cluster 1 was preferred route for prefix *p* by RRs in cluster 3, it crashes
2. IF RRs in cluster 1 are using unique CIDs per RR (e.g., default router IDs), then RRs in cluster 3 must propagate new route (implicit withdraw for previous) to client, even though only cluster list contents changed, perhaps not even forwarding path
3. In mutli-tier RR, this can occur even with common CIDs for RRs within a cluster
4. When the failed router is restored, all routes will transition back
5. May trigger gratuitous eBGP updates as well
6. Need mechanism akin to eBGP Avoid best transition (RFC 5004) for iBGP cluster lists of same length when only cluster list values change

RR Advertisement Rules

- Change in specification from RFC 1966 to RFC 2796:
  - Change allowed an RR to reflect a route learned from a client back to that client
  - Change made to optimize local implementation (copying of updates task); no care given to system-wide effects
- Client now has to know it’s a client and “poison” received routes where Originator ID added by RR is equal to local BGP Router ID
- Consider example with 100k best routes from client with 3 RRs - client now has to discard 300k routes received from RRs that were reflected back to client, whether common or unique cluster IDs on RRs
- The updates are not benign - processing may delay legitimate update processing

RR Rule Change



1. *p/24* reflected from RRs back to originating client
2. Client expected to poison if Originator ID == Router ID
3. May not be issue with one prefix, but often 100k or more reflected back from each RR - all to be processed and discarded by client
4. A moderate RR implementation change led to high process cost at client

And furthermore…

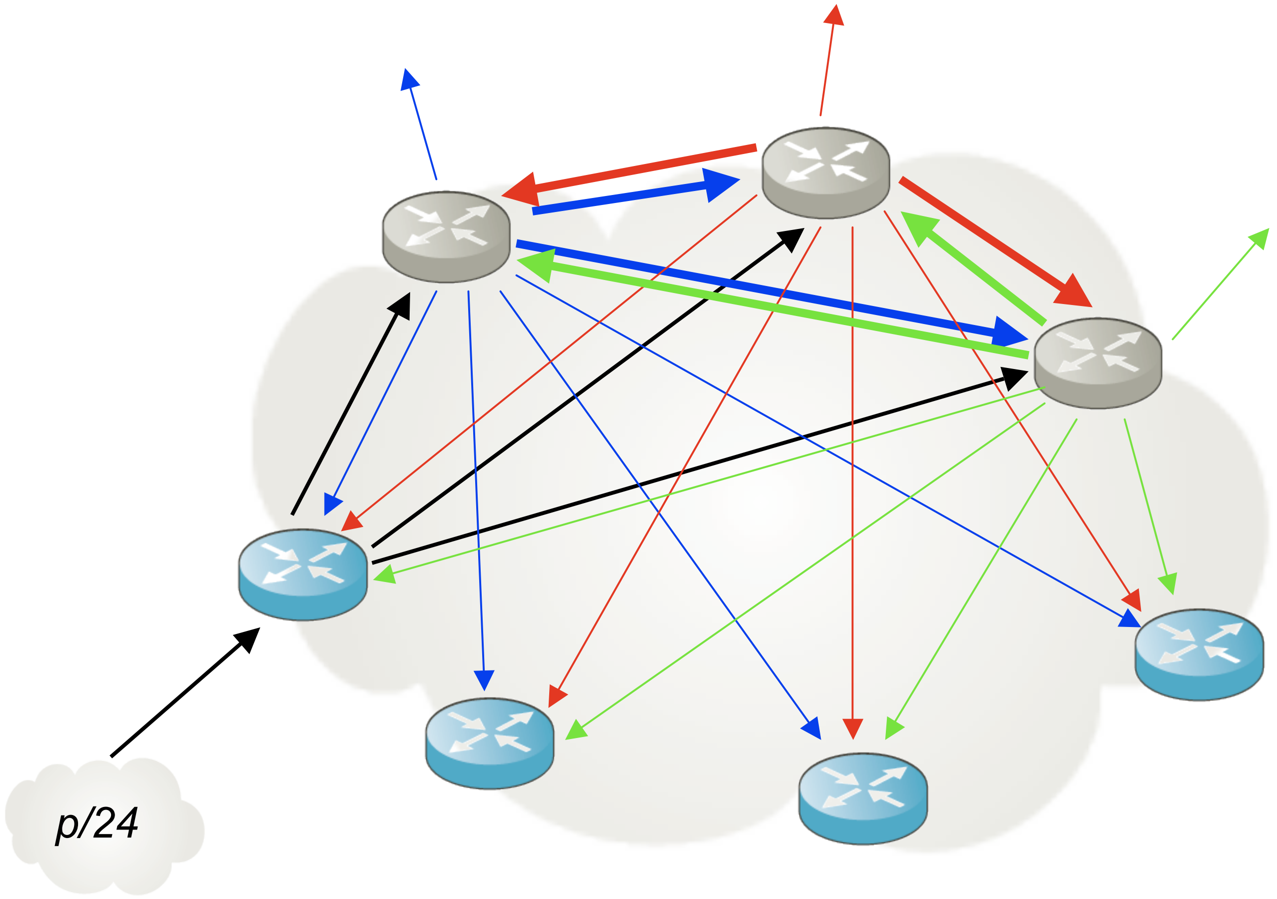
- Proposed IP VPN technique aims to exploit this behavior to minimize *local* configuration
  - Define community (ACCEPT_OWN) to allow acceptance of routes (not poison) by client, even if Originator ID equals local Router ID, if community present
  - Allows **upstream** RR to distribute routes between VRFs on **local** PE
  - Saves having to configure local inter-VRF redistribution policies on each PE
- In fairness, different *overlay* RRs are often used for IP-VPN address families…
- draft-ietf-l3vpn-acceptown-community

Network Architecture Considerations

RR Cluster IDs

- Unique Cluster IDs per RR within a given cluster can result in significant number of extraneous routes
  - Each RR will maintain routes from other RRs sourced from clients within cluster versus discarding - even if RR is NOT in forwarding path (i.e., useless)
  - E.g., A client with 3 RRs in cluster and 100k “best routes” means 300k Adj-RIB-In entries on *each* RR
  - Client-client reflection v. full-client iBGP mesh within cluster may or may not help this
  - Note: RRs within cluster usually fully-meshed because of external peers, configuration templates, etc..
- More unique attributes, less update packing ability, more state, more churn

Effects of Unique Cluster IDs



1. Common deployment model: each RR has a unique cluster IDs within cluster (default to RID).
2. Result is each RR storing redundant routes from other RRs within same cluster
3. May not be issue with one prefix, but if lots of prefixes, can be very significant needless overhead
4. With common cluster ID RRs would poison each others routers based on cluster list path vector
5. *Further optimization might be for RR configuration knob to identify iBGP RR peers within same cluster - or ORF iBGP-like model; to avoid update advertisement for client prefixes*

Network Architecture Effects

- Placement of peers v. customers, etc..
- Number of RRs per cluster
- Additional RR hierarchy
- Common v. unique cluster IDs
- Client-Client reflection v. full client mesh
- Overlay Topologies for other AFs
- IP Forwarding path congruency?
- Resetting attributes on ingress (e.g., community resets, MED resets) to optimize update packing, but may result in more routes (as local “best”)
- More low-end routers > more BGP speakers > more unique routes - effects of economic climate?

- **Operators: LOTS of room for improvement here**

Miscellaneous

New BGP Address Families

- New address families carried in BGP:
  - Higher BGP load
  - Change to **BGP** code base
  - Often on same routes and global “Internet” routers
- Example BGP AFs/SAFs include:
  - IP6
  - IP-VPN
  - BGP Flow Specification
  - Pseudo Wires
  - L2VPN
  - 2547 Multicast VPNs
- In fairness, many (most?) of these non-IPv4 unicast AFs employ overlay RR topologies rather than the native BGP topology
  - *Note: reasonable where PE-PE MPLS LSPs or tunnels exist, but for native hop-by-hop IP Network layer forwarding strong consideration should be given to topology, forwarding loops, etc..*
- Is this better than running another protocol? Perhaps. Perhaps not….

Effects on Routing Security

- Each route has to be authorized on per-peer basis, all viable routes need to be pre-enumerated
- Ideally, policy considers both AS_PATH and prefix per-peer; today most policy only prefix per-peer (prefix-based ACLs) IF at all
- Origin AS filtering alone provides very little benefit (can be spoofed, permits route leaks)
- Very little to no inter-provider filtering
- More routes means more policies that need to be configured, more routes that need to be authorized
- Explicit BCP 38 or anti-spoofing must factor feasible routes as well, else asymmetry will break forwarding

Additional IDR Work

- Work on ways to add new paths (versus remove extraneous ones)
  - In order to enable route analytics (e.g., draft-ietf-grow-bmp)
  - Mitigate BGP route oscillation (RFC 3345)
  - iBGP Multi-path
- Trade-off is expense of extra state versus oscillation reduction and iBGP multi-path support

Other BGP Issues

- BGP Wedgies
- Non-transitive attributes result in best path change, duplicate update propagation to eBGP peers
- Persistent Route Oscillation Condition
- RR topology congruency guidelines
  - Per-AF topologies changing mindset
  - Multiple RRs makes difficulty
  - IGP Metric constraints

Conclusions

- # routes (v. unique prefixes) effects everything, increasing over time and more steeply than DFZ
- This is where things will break first
- Just because an update doesn’t make it into the RIB doesn’t mean it’s benign
- Possibilities for protocol, implementation, network architecture improvements
- Operators, implementers, scalable routing designs need to consider these factors

Acknowledgements

- Level(3) Communications
- Ricardo Oliveira, Dan Jen, Jonathan Park & rest of UCLA team
- Keyur Patel @Cisco
- Craig Labovitz @Arbor (early work on stability)
- Halpern, Morrow, Rekhter, Scudder, BD for new and previously agreeing and dissenting views on the content in the slides, and recommended improvements

EOF

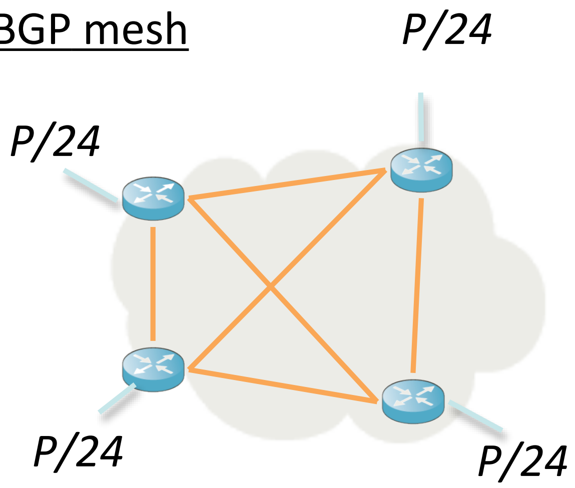
Internal Route Amplification

- Look at different architectures and evaluate them according to:

  + **RIB-in scaling**: number of entries per prefix in RIB-in

  + **Path redundancy**: number of possible BGP paths to a prefix; path redundancy is a rough upper bound of the churn involved in path exploration

<u>The iBGP mesh</u>



- Assume an iBGP mesh w/ **n** routers, in this case n=4
- A prefix P being received in eBGP at each border router
- Each border router will have **n** routes to reach P

- **RIB-in scaling = n = 4**
- **Path redundancy = n =4**

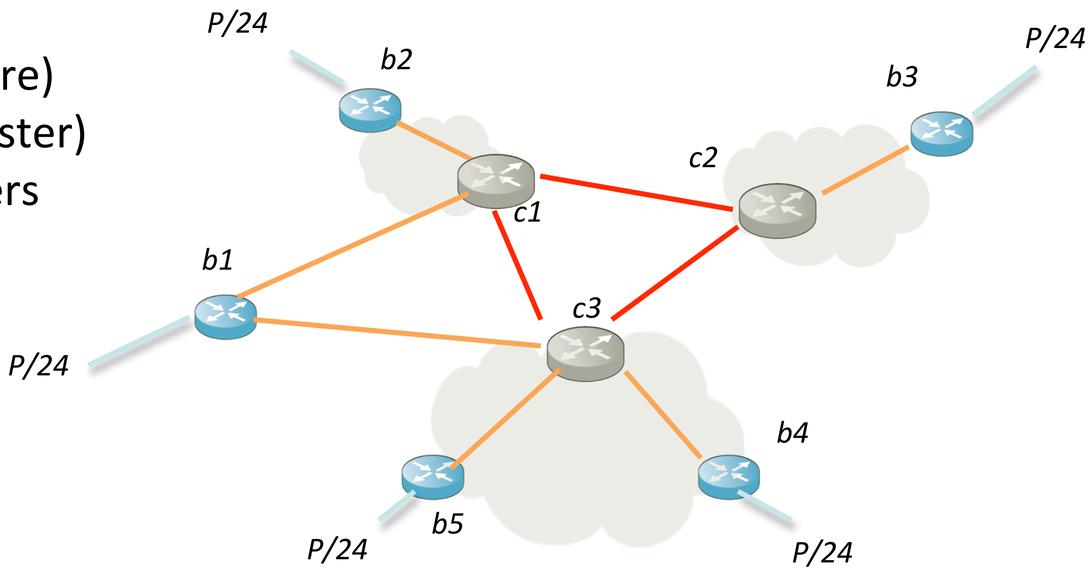
The single level RR

- N clusters connected in a mesh (N=3 here)
- Cluster size C (number of clients per cluster)
- Each border router connects to D clusters

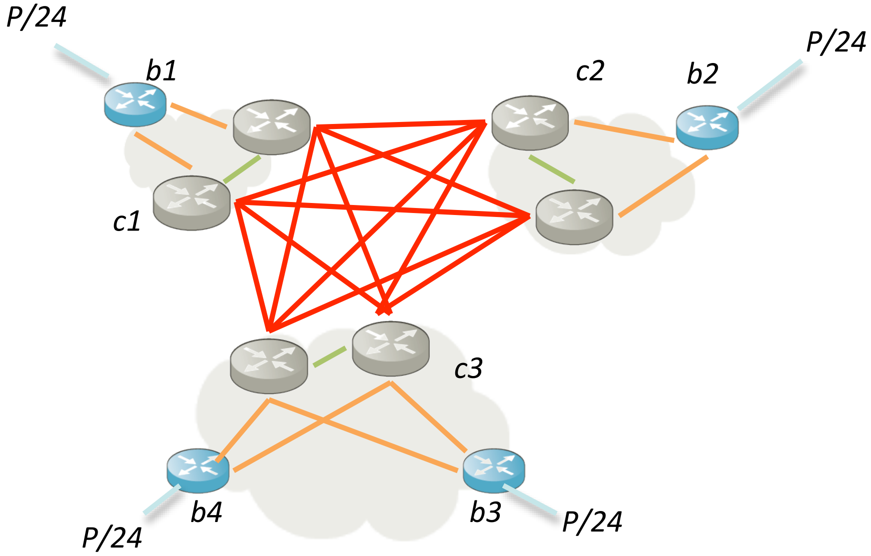
- **RIB-in scaling = D+1**

  3 (for *b1*), **4** (for *c1* RR)
- **Path redundancy ~ D*N*C**

  7 (for *b1*), **6** (for *c1* RR)



Adding redundancy in RRs per cluster…



- B RRs per cluster

- **RIB-in scaling = D*B+1**

  3 (for *b1*), **5** (for RRs)
- **Path redundancy ~ D*B*N*C**

  13 (for *b1*), **6** (for *c1* RR)