Prefixes, Paths & Internet Routing System Scalability

ARIN 25

April 19, 2010

Danny McPherson danny@arbor.net
Chief Security Officer

Overview

ARBOR NETWORKS

- Number of discrete prefixes (i.e., “DFZ size”) in the routing system is only one measure of Internet routing system scale
- Number of unique “routes” or “paths” (prefix + attributes) associated with any given prefix dependent on many variables – numerous interactions with both interior and inter-domain routing scalability
- Systemic effects of new prefix introduction need to be considered during all phases of Internet engineering
  - protocol design
  - implementation
  - network architecture
  - policy development

BGP Overview

- BGP is the *de facto* protocol for inter-domain routing on the Internet – used to convey destination reachability to peers
  - *prefix* – set of destinations (e.g., 10.0.0.0/8)
  - *attributes* (e.g., AS_PATH, MED, Origin , etc.)
- Large number of loosely interconnected routing domains, represented as *autonomous systems (AS)*, make up global routing system
- *Path vector* elements employed for routing information loop detection
  - "AS path" inter-domain; route reflection or AS confederation attributes for intra-domain
- A BGP speaker only advertises best available path for a given prefix (currently)

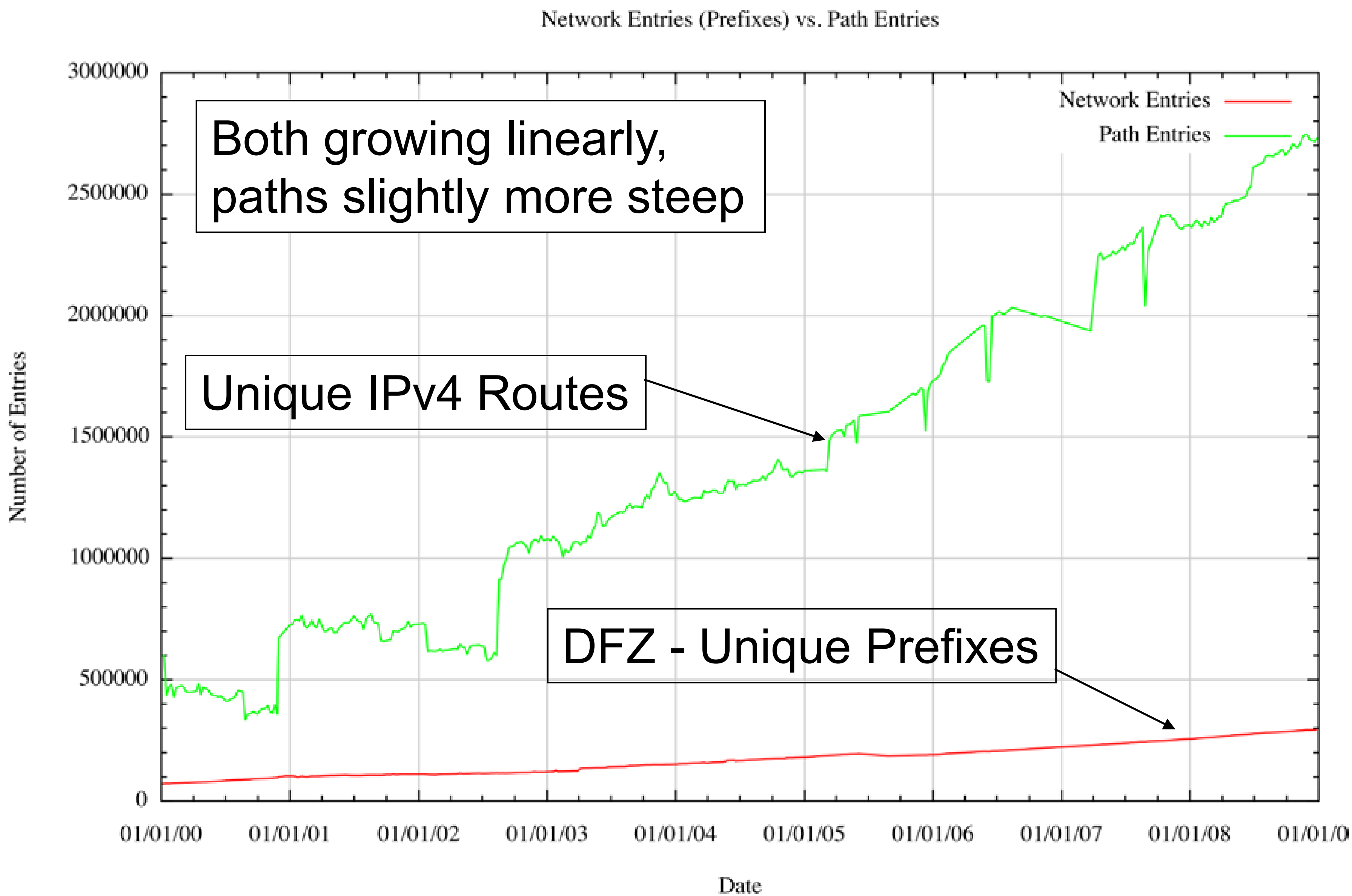
Topology: The Bogey Man!

- BGP behavior dependent on topology
- Making connectivity (internal & external) richer SHOULD result in improved reliability
  - Instead, may cause [considerable] convergence delays when routes flap - even in the absence of flap dampening
  - Rich topological connectivity (internal or external) can result in bad path selection announcement/withdraw behavior, race conditions prior to new correct state while withdrawals flood the global DFZ
  - This is a path hunting problem which won't go away until it is solved (causes escalation of BGP update counts and convergence delay, among other things)

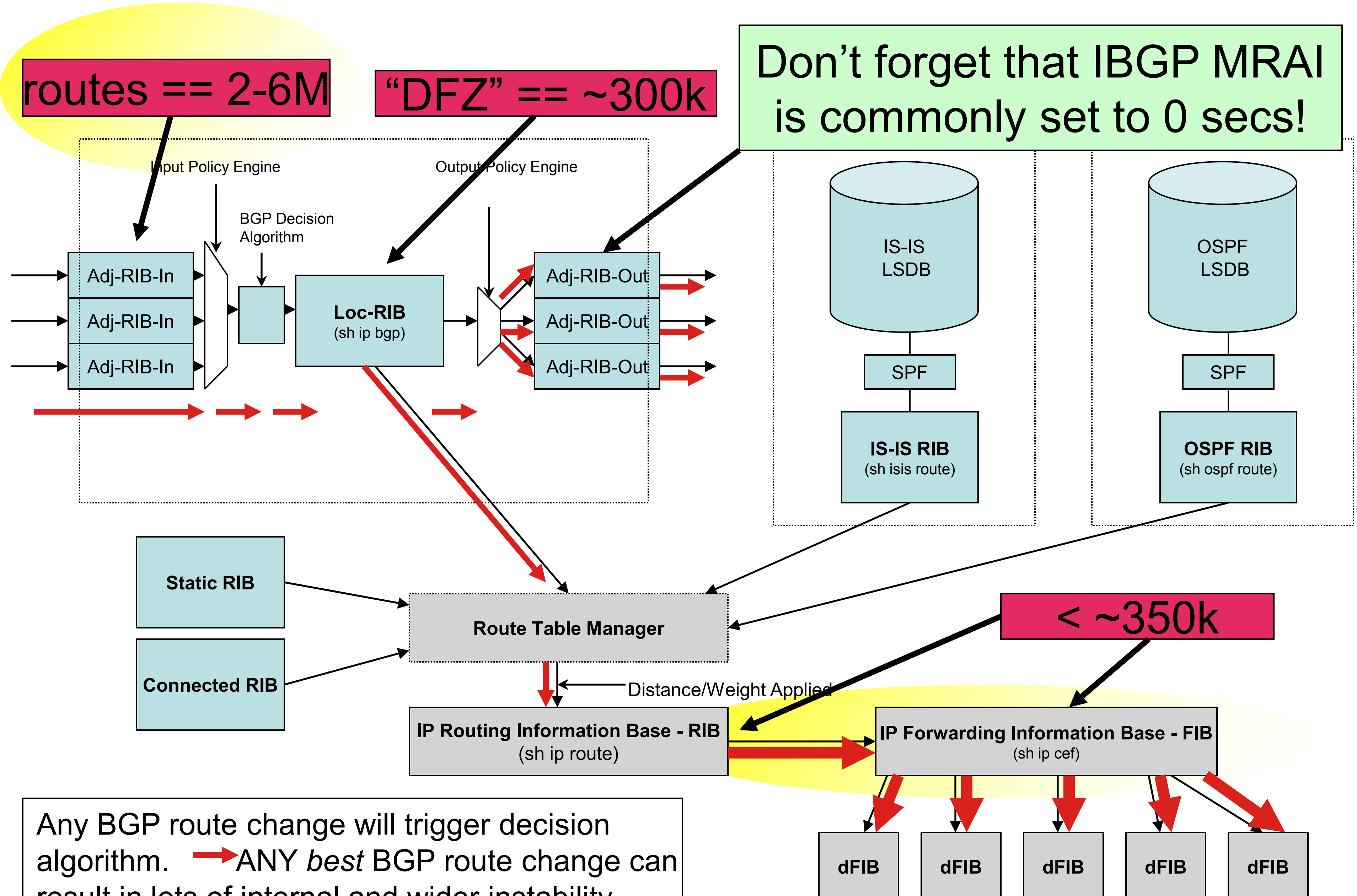
What Breaks First?

- 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 function of
  - more prefixes in DFZ
  - richer internal and external interconnection topologies
- More routes == more state, churn; effects on CPU, RIBs && FIB
- Routes growing more steeply than unique prefixes/DFZ – *highly topologically dependent*

Growth: Prefixes v. Routes



ANY *Best* Route Change Means….



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

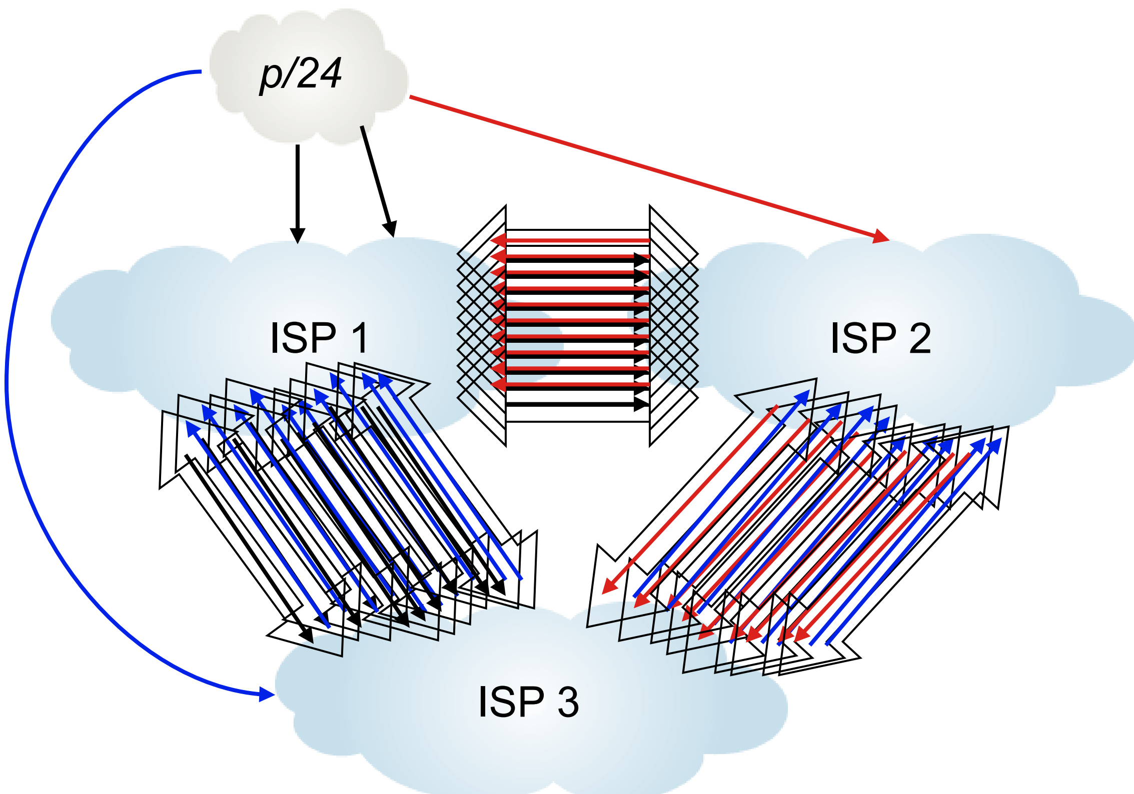
Why is # of unique routes increasing faster than # of prefixes?

- Primarily due to denseness of interconnection outside of local routing domain
  - Increased multi-homing from edges
  - Increased interconnection within core networks
- Each new non-aggregated prefix (~PA) 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
- Highly topologically dependent

Disintermediation; nixing the middlemen

- More networks interconnecting directly to avoid transit costs, reduce transaction latency, forwarding path security or diversity (e.g., avoid hostile countries)
  - More networks building their own backbones (e.g., CDNs, 'hyper-giants'), have presence in multiple locations
  - More end-sites and lower-tier SPs provisioning additional interconnections, minimizing transit costs while state still there
  - Networks adding more interconnections in general to localize traffic exchange, accommodate high-bandwidth capacity requirements, and optimize performance
- Increased interconnections made feasible by excess fiber capacity and decreasing cost, offset transit costs
- More interconnections means more unique routes for a given prefix

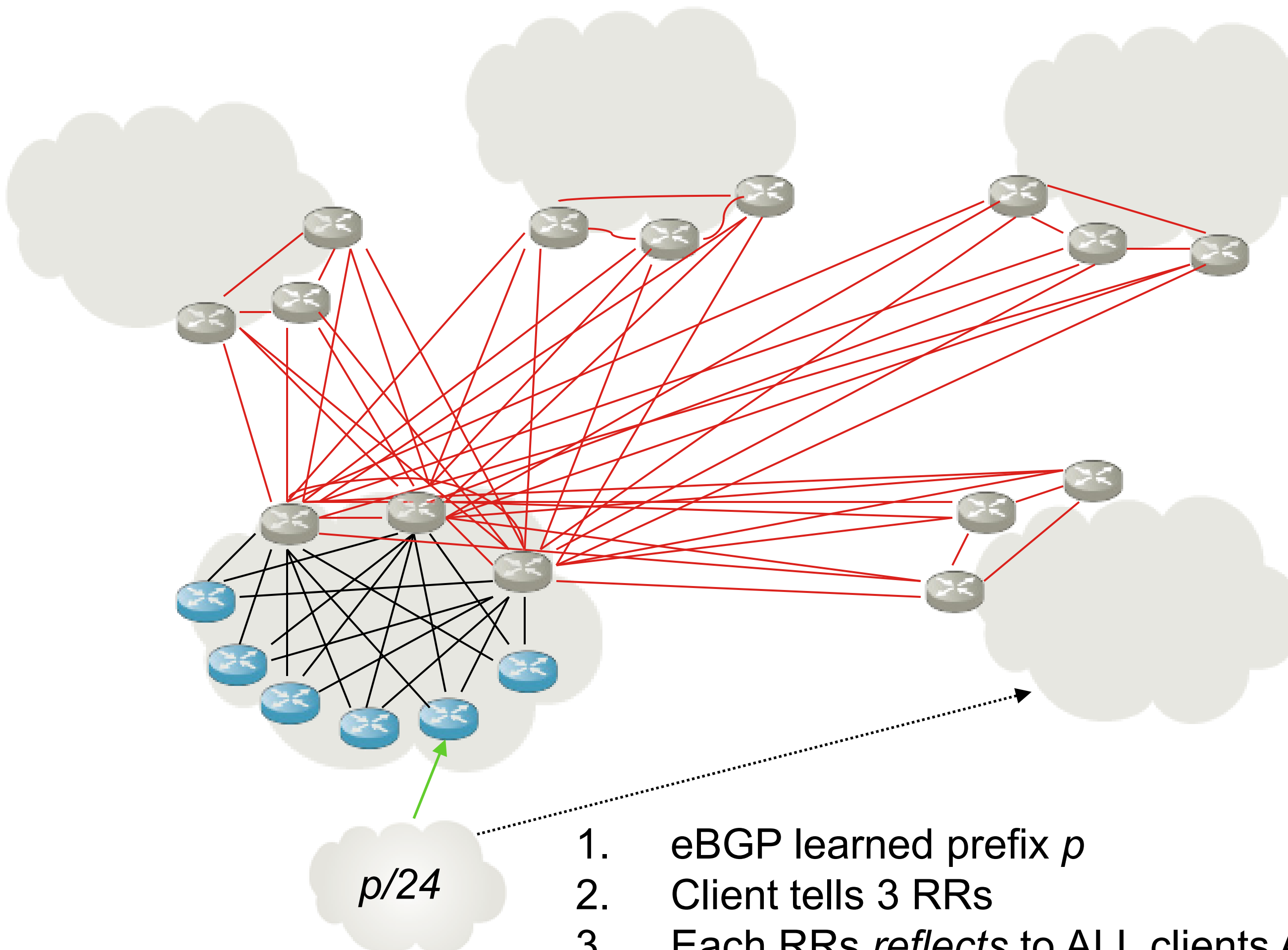
External Interconnection Denseness



ISP1 - one unique prefix (*p*), 22 routes total on PE routers, without intra-domain BGP effects

- 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

Route Reflection Illustrated



Those 22 routes total for p on the PEs result in 30 paths on EACH RR in simple network:

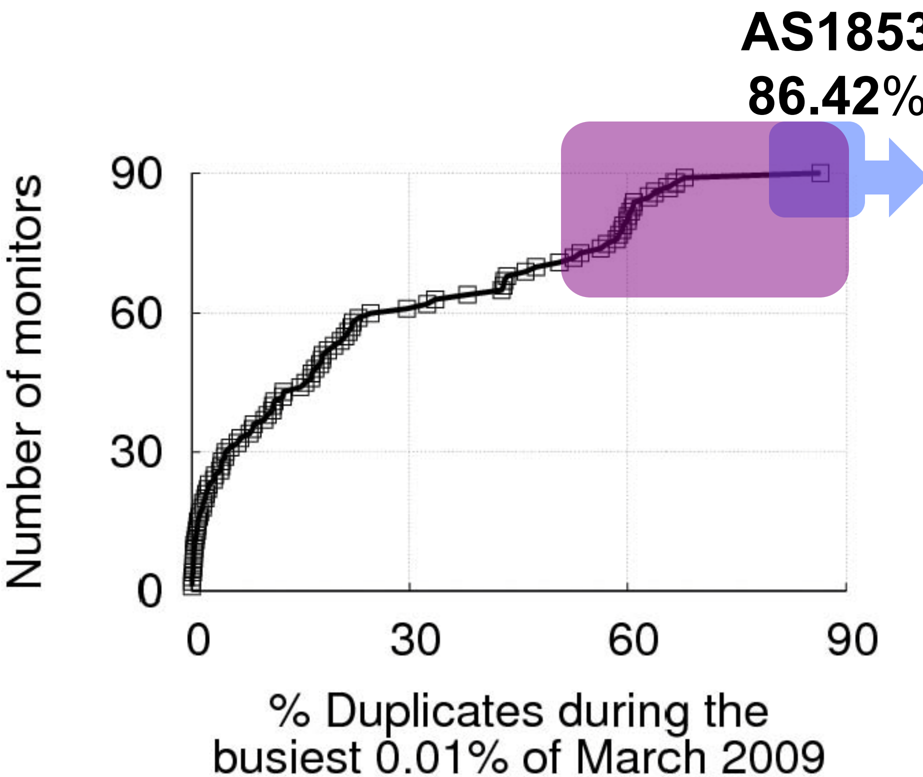
9 other clusters (pops)
* 3 RRs/cluster
+ 1 client path
+ 2 other RRs local

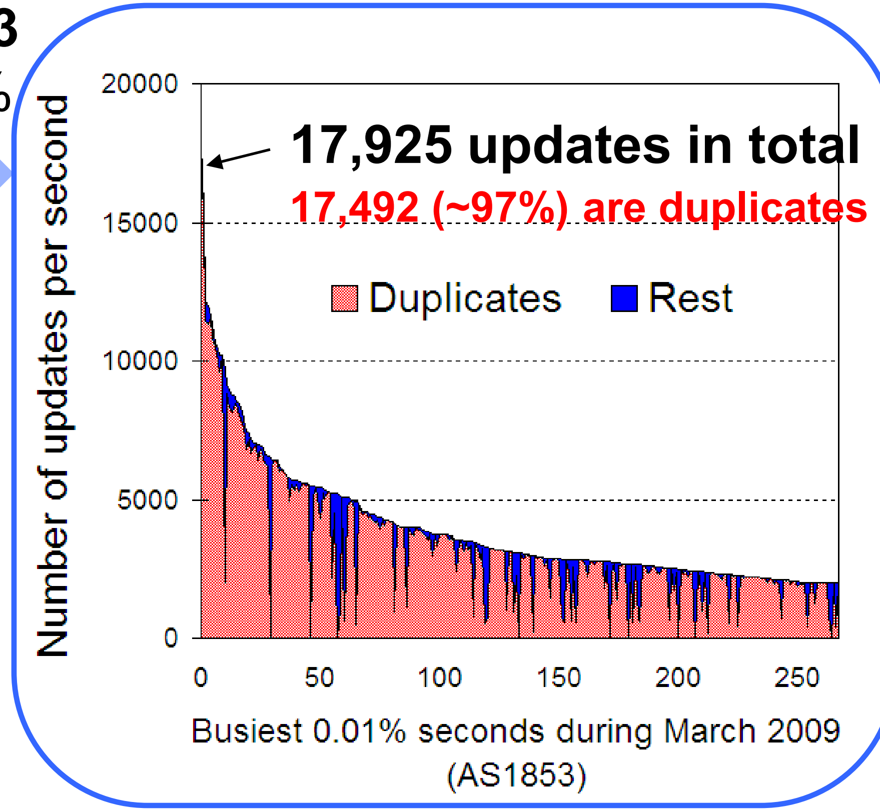
30 paths for p per RR!

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

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

Duplicates are responsible for most traffic during busiest times – PAM '10, Park et al.





Illustrates that duplicates are responsible for the majority of router processing loads during their busiest times

86.42% of the total updates during the busiest 267 seconds are duplicates

Conclusions

- # routes (v. unique prefixes) effects everything, increasing over time and more steeply than DFZ
- Mechanics of multi-homing no different for v4 v. v6, – a route table slot = FIB slot, but doesn't necessarily reflect systemic dynamics that impact FIB I/O, etc.
- Beyond mechanics of FIB hardware size, this is where things will break or strain the system
- Just because an update doesn't make it into the RIB doesn't mean it's benign (e.g., route reflection back to client, etc..)
- Possibilities for protocol, implementation, network architecture improvements
- Operators, implementers, scalable routing designs, policy development folk need to consider these factors