Lossless prefix aggregation for forwarding

Artificial Ornithology Lab

Maria Matějka • 13 February 2024

CZ.NIC | CZ DOMAIN REGISTRY

Find a difference

- 2001:db8:dada::/48 via 2001:db8:f00::ba1
- 2001:db8:dadb::/48 via 2001:db8:f00::ba1

Find a difference

- 2001:db8:dada::/48 via 2001:db8:f00::ba1
- 2001:db8:dadb::/48 via 2001:db8:f00::ba1
- Yes, it's the 48th bit

Find a difference

- 2001:db8:dada::/48 via 2001:db8:f00::ba1
- 2001:db8:dadb::/48 via 2001:db8:f00::ba1
- Yes, it's the 48th bit
- What about 2001:db8:dada::/47 via 2001:db8:f00::ba1 …
  …yes, it's the same

Find a difference

- 2001:db8:dad8::/48 via 2001:db8:f00::ba1
- 2001:db8:dad9::/48 via 2001:db8:f00::ba1
- 2001:db8:dada::/48 via 2001:db8:f00::bad
- 2001:db8:dadb::/48 via 2001:db8:f00::ba1
- 2001:db8:dadc::/48 via 2001:db8:f00::ba1
- 2001:db8:dadd::/48 via 2001:db8:f00::ba1
- 2001:db8:dade::/48 via 2001:db8:f00::ba1
- 2001:db8:dadf::/48 via 2001:db8:f00::ba1

Equivalent set of routes

- A covering route for the same nexthops
  2001:db8:dad8::/45 via 2001:db8:f00::ba1

Equivalent set of routes

- A covering route for the same nexthops
  2001:db8:dad8::/45 via 2001:db8:f00::ba1
- And a more-specific route for the other one
  2001:db8:dada::/48 via 2001:db8:f00::bad

Aggregating prefixes

- Static optimal-result algorithm exists[1]
- Dynamic almost-optimal-result algorithm exists[2]
- Finding minimal set of prefixes equivalent to input
- No misroutings!

[1] https://ieeexplore.ieee.org/document/749256
[2] https://doi.org/10.1145/2079296.2079325

Use cases

- Saving big ASICs/TCAMs from overflow
  - from steady growth
  - from accidental mispropagation of a million /48's
- More efficient usage of smaller ASICs/TCAMs

Preliminiary results

- IPv6 can be aggregated to approx. 50k to 100k prefixes
- IPv4 can be aggregated to approx. 100k to 250k prefixes
- Partially depends on actual number of nexthops
- Partially depends on location
- Data: voluntarily contributed full route dumps

IPv6 Aggregation in different parts of the world

- Example: CZ.NIC routing data from Anycast DNS
- Full BGP in London and Frankfurt
- Both agregate from 195k down to 65k
- $\sim$50k prefixes kept intact
- Resulting prefix set difference: $\sim$10k prefixes
- $\Rightarrow$ 15% of the whole result is location-dependent

IPv4 Aggregation in different parts of the world

- Example: CZ.NIC routing data from Anycast DNS
- Full BGP in London and Frankfurt
- Both agregate from 950k / 930k down to 220k / 200k
- $\sim$70k prefixes kept intact
- Resulting prefix set difference: $\sim$70k prefixes
- $\Rightarrow$ 30% of the whole result is location-dependent

Next steps

- Large-scale data analysis to verify preliminary estimations
- Check actual forwarding performance in ASICs
- Finish the implementation
- Test aggregation on route reflectors for iBGP

Some provocative questions

- Can we afford to route suboptimally?
- Which prefixes to divert?
- Consistent degradation on route reflector?

Some provocative questions

- Can we afford to route suboptimally?
- Which prefixes to divert?
- Consistent degradation on route reflector?

- Or shall we aim for less scattered assignments instead?

Data provided by

- Tom Bird, Portfast Ltd., AS 8916
- Cathal Mooney, Wikimedia Foundation, AS 14907
- CZ.NIC, AS 25192
- Thomas King, AS 31451
- Daniel Wagner, DE-CIX RnD, AS 205530
- Fredy Kuenzler, Init7, AS 13030, 196620
- Tobias Fiebig, AS 59645, 211286

Topic consulted with

- Igor Putovny, Ondrej Zajicek, Jiri Hudecek and others at CZ.NIC
- Daryll Swer, https://www.daryllswer.com/
- various other people under my tweet[3]

[3] https://twitter.com/marenamat/status/1739769026451026051

QED

Maria Matějka • 13 February 2024

CZ DOMAIN REGISTRY