An Evaluation Study of Router FIB Aggregatability

B. Zhang, L. Wang, X. Zhao, Y. Liu, L. Zhang

draft-zhang-fibaggregation-02.txt

November 8, 2009

FIB Aggregation (FA)

- The idea of FA has floated around for some time now
- What is FA: if multiple adjacent RIB entries share the same nexthop, only install one entry in the FIB, e.g.
  - 1.0.0.0/9 and 1.128.0.0/9 → 1.0.0.0/8
  - If they share the same next hop, install 1.0.0.0/8 in FIB in place of 1.0/9 & 1.128/9
- Why FA: To reduce the FIB size

FIB Aggregation: Pros and Cons

- ✔ No impact to packet forwarding
  - Multi-homing, load balancing, TE all work the same.
- ✔ No change to routing protocols
  - Only a software upgrade, can be done per router
- ✔ Compatible with other proposed routing scalability solutions
  - LISP, APT, Virtual Aggregation, etc.
- ✘ Extra CPU processing time
- ✘ Potentially extra routable space
  - Packets to previously non-reachable destinations may be forwarded for a few more hops.
  - Whether, or how badly, it happens depends on the level of aggregation.

Why FA Can Be Effective

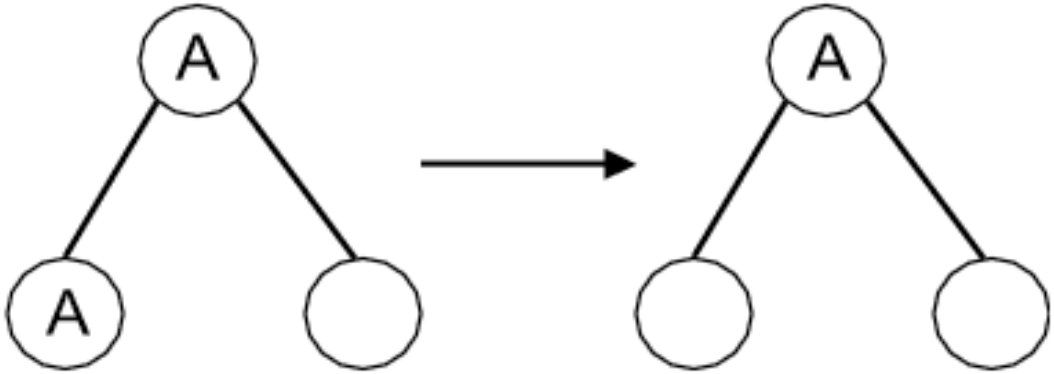
- FIB aggregation is opportunistic
- Our analysis show plenty of aggregatable opportunities
  - Prefixes allocated to the same RIR/country/ISP
  - Prefixes split from one original assignment
- Why these prefixes share the same next-hop
  - Prefixes announced far away are more likely to share the same next-hop than nearby prefixes.
  - Multi-homing and traffic engineering make a difference when traffic gets close to the destination, but may not to routers far away.

What we have done

- Refinement of the FA scheme
  - Four levels of prefix aggregation
    - each additional level can aggregate more but also adds more overhead
  - Efficient handling of routing changes
- Evaluation of FA's gains and costs.
  - Table size reduction.
  - Computation time.

Level-1 Aggregation

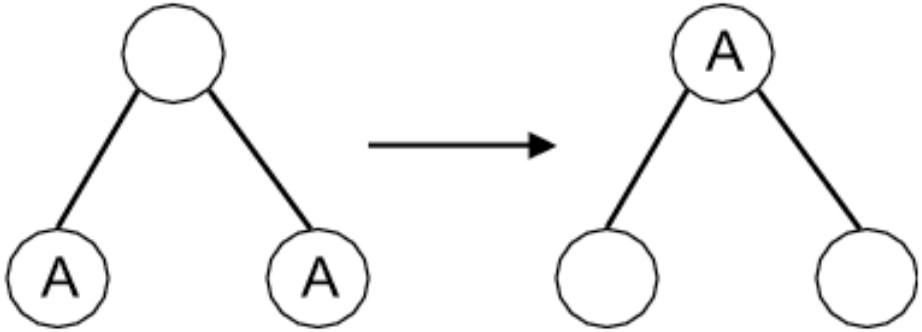
- Remove covered prefixes
  - Add no new prefix nor new routable space.



Letter in the circle: next hop
Blank circle: prefix not in RIB

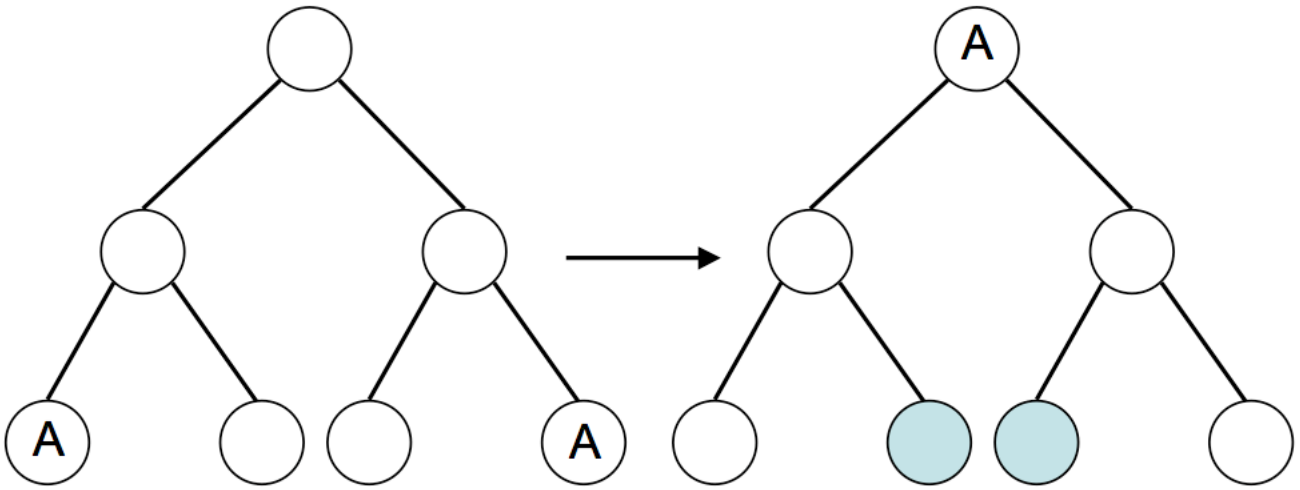
Level-2 Aggregation

- Combine sibling prefixes
  - Insert a new prefix, but the routable space remains the same.



Level-3 Aggregation

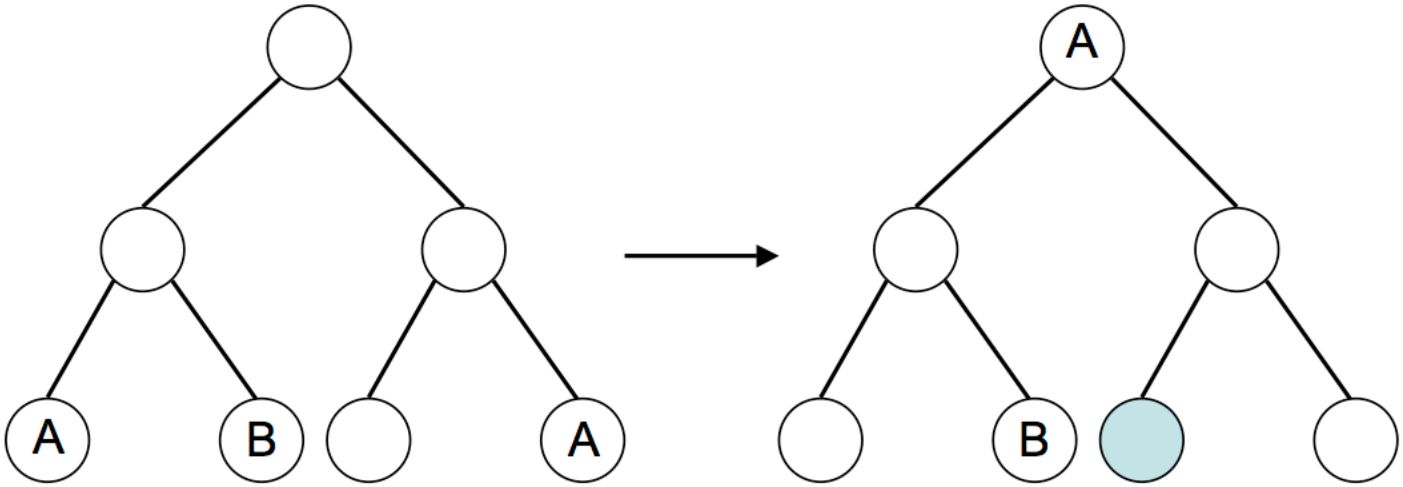
- Aggregate non-sibling prefixes
  - Packets heading to non-reachable destinations will be dropped when they get close to the destination or TTL expires.



Blue nodes: extra routable space

Level-4 Aggregation

- Aggregate non-sibling prefixes
- allow "holes" of different nexthops under the aggregated prefix
  - We tried two algorithms, 4A and 4B.

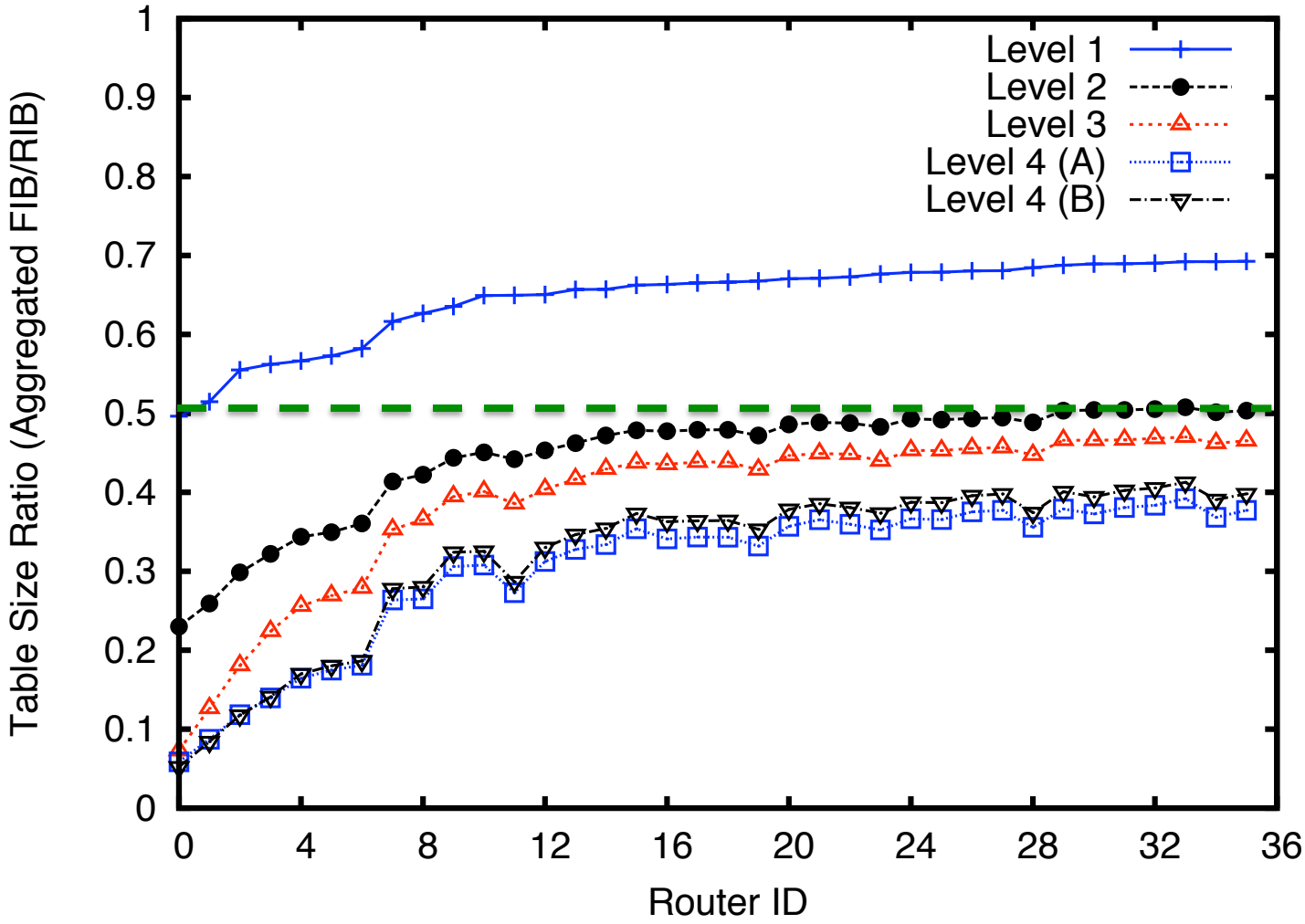


For details, see http://www.cs.arizona.edu/people/bzhang/paper/aggregate.pdf

Evaluation Methodology

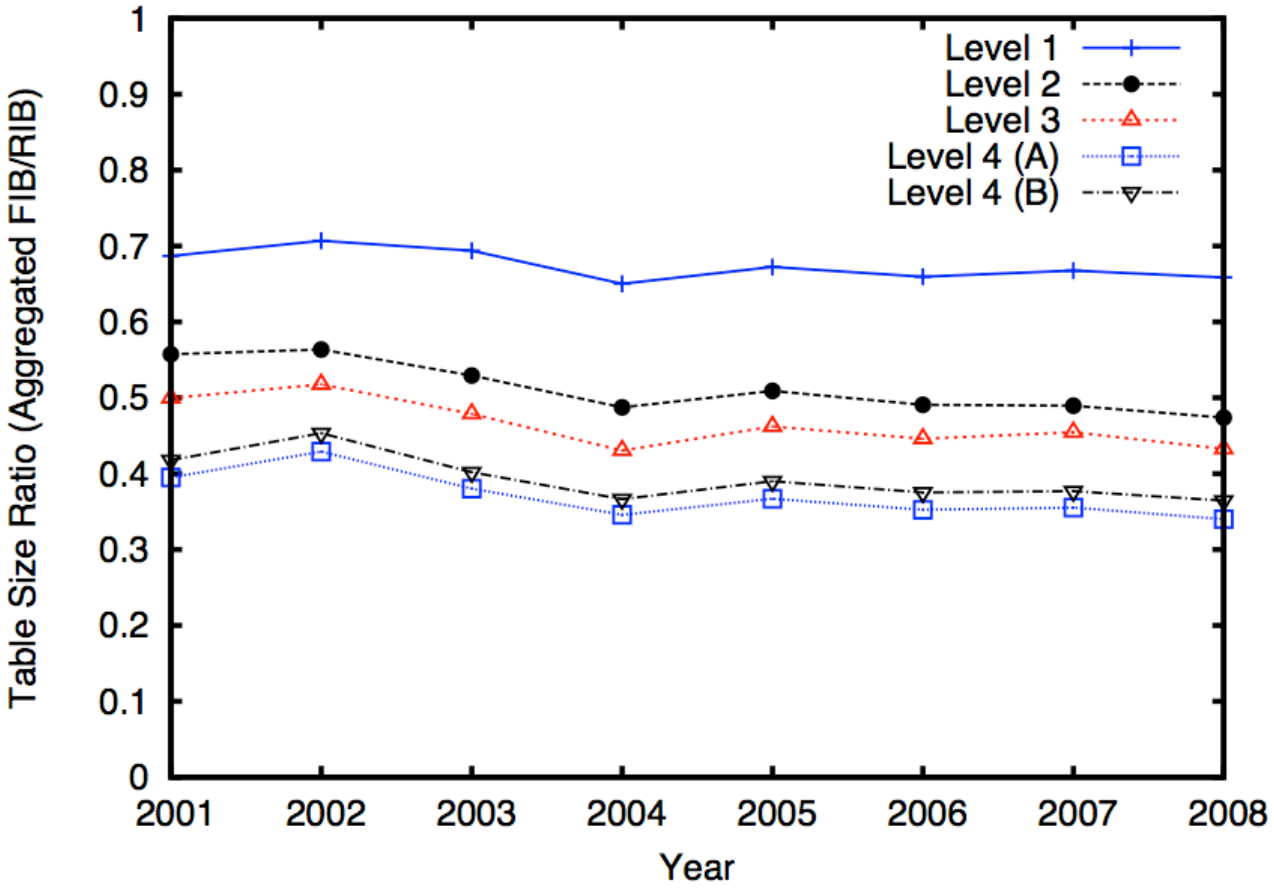
- Data Source: BGP routing tables and updates from RouteViews's Oregon collector.
- Assumption: prefixes with the same next-AS-hop use the same next-IP-hop.
  - Verified with 9 routing tables downloaded from route servers: one has 85%; the other 8 have 93% - 100% of prefixes that satisfy this assumption (Fig. 4 in the paper)
- Computation time is measured on a Linux machine.
  - an Intel Core 2 Quad 2.83GHz CPU (single thread process)
  - Comparing relative processing time of diff. aggreg. levels
- RIB/FIB: implemented as a Patricia Trie.

FIB Size Reduction



- RouteViews data from 2008.12.31
- Edge network routers get more FIB reduction than core networks
- The last few points are routers from tier-1s

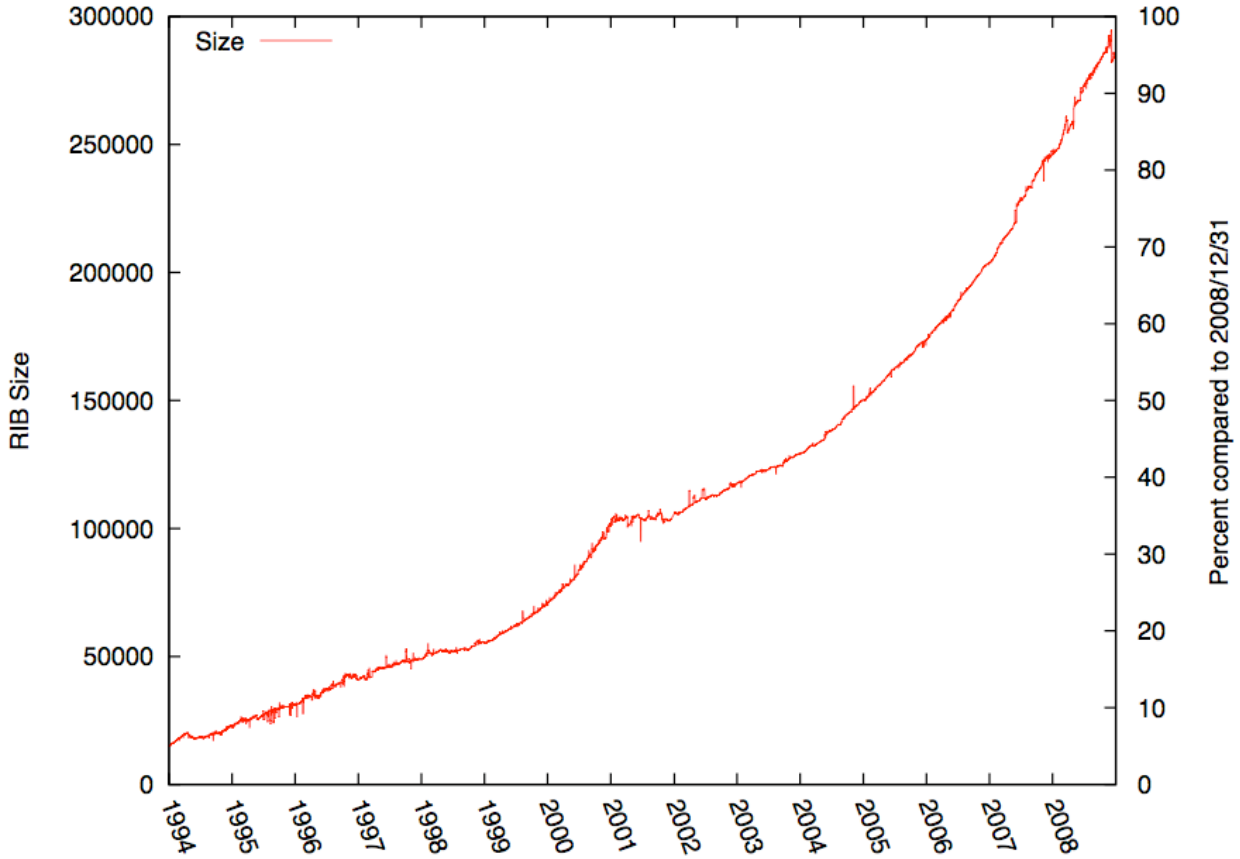
FIB Size Reduction Over Years



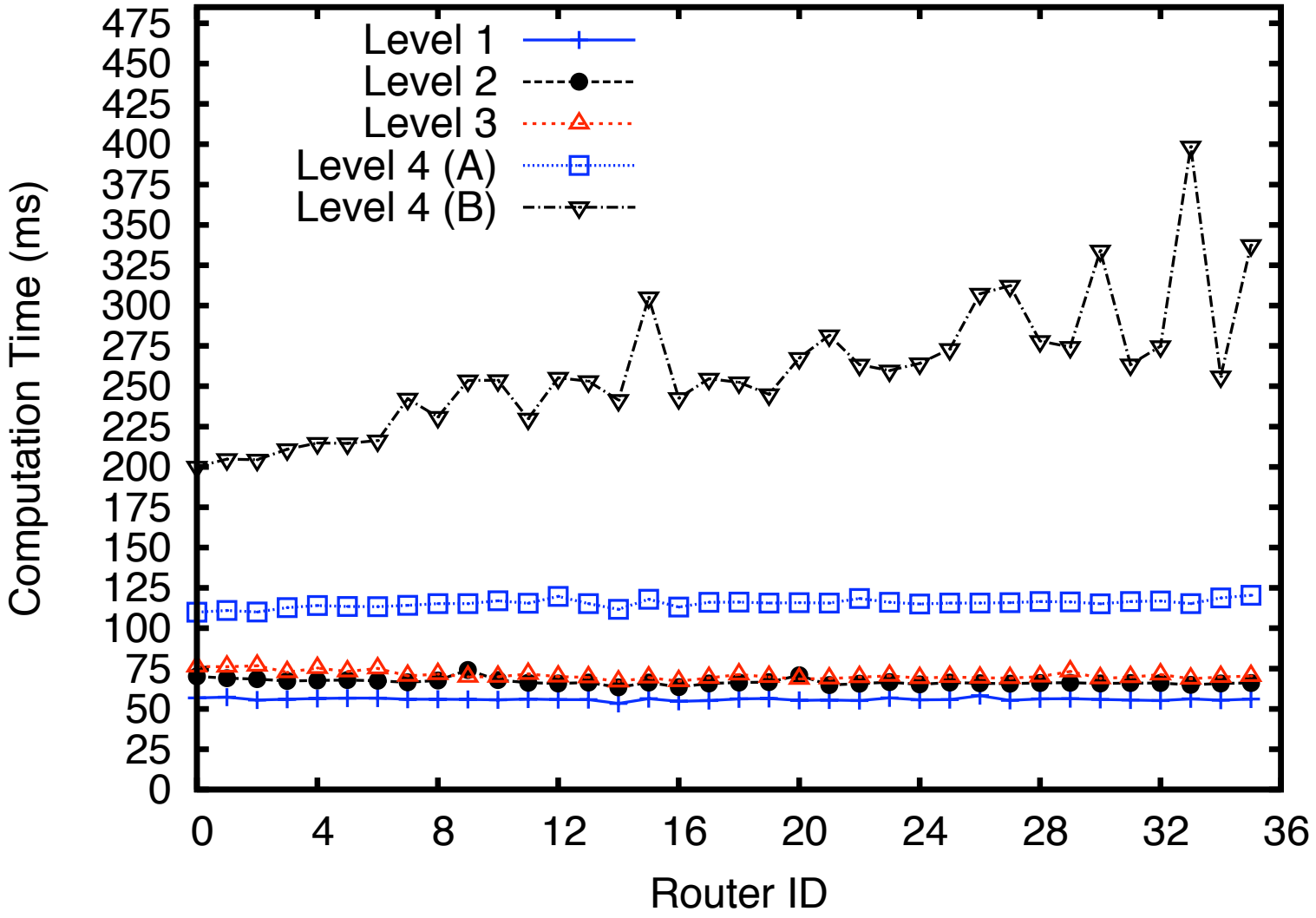
- Median of aggregated table size among all peers in each year.
- Slight decrease over years, may due to more prevalent TE and multi-homing.

What does the ratio mean?

- Take 2008.12.31 as 100%
  - 2006.10 (70%), 2004.08 (50%), 2000.06 (30%)
- If FIB size is an issue, FA can give routers quite a few more years of lifetime.

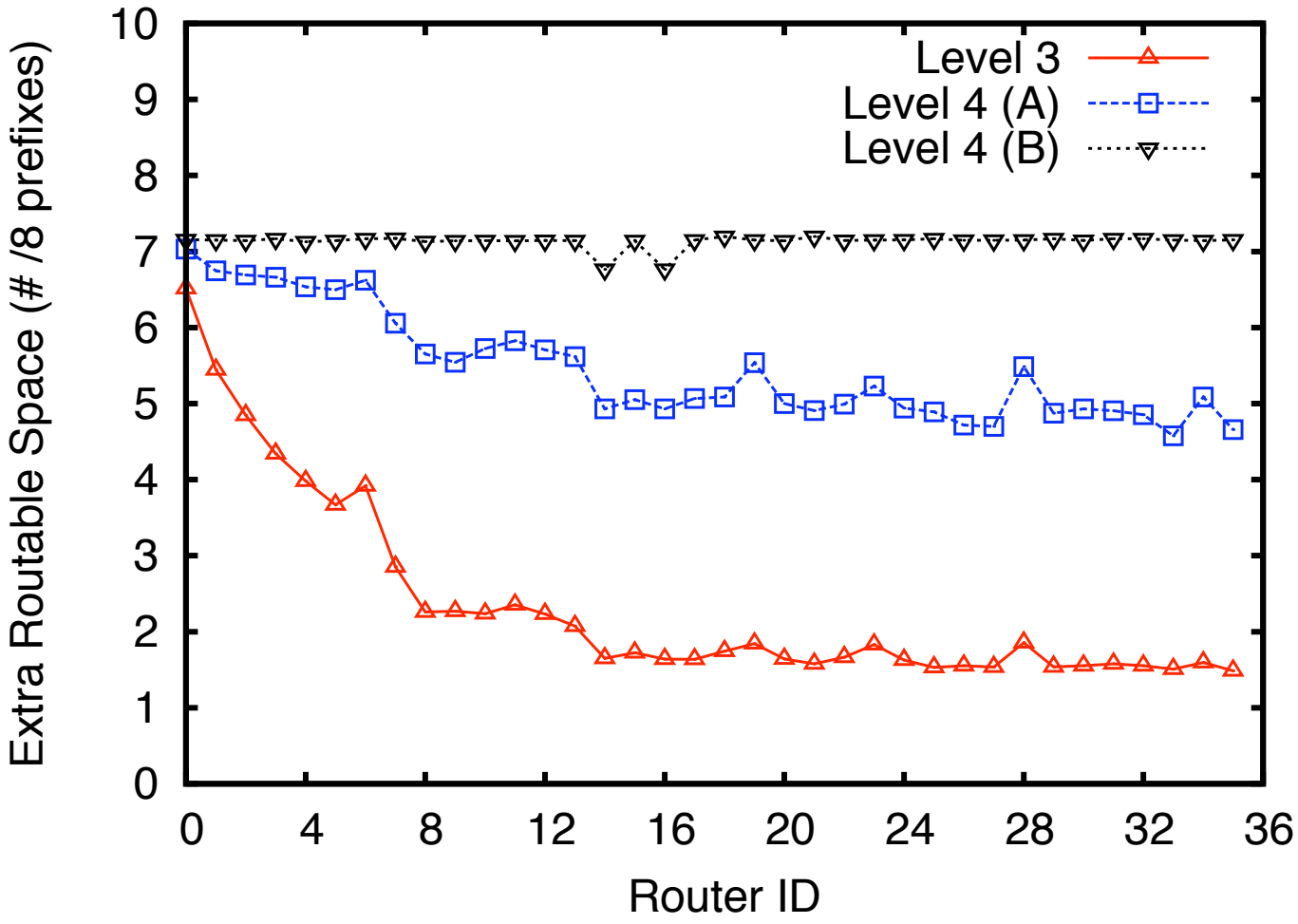


Computation Time



- Each algorithm labels every prefix as either IN-FIB or NON-FIB.
- No optimization attempted on the algorithm or implementation.

Extra Routable Space



- Extra routable space is measured by the number of /8 blocks (117 total in the routing table, < 6%).
- More table size reduction, more extra routable space.

Handling Routing Updates

3 approaches to handling routing changes to keep computation overhead low:

1. Operators choose an appropriate level of FA.
2. Incrementally update the aggregated FIB
   - Minimize computation, not care table size.
   - Need to de-aggregate or re-aggregate part of the tree.

   then Re-run full FIB aggregation periodically.
   - The trigger can be a timer, a threshold on FIB size, and/or current router CPU load.
3. A small number of prefixes are responsible for a large number of routing updates. Excluding them from FA can save CPU cycles.

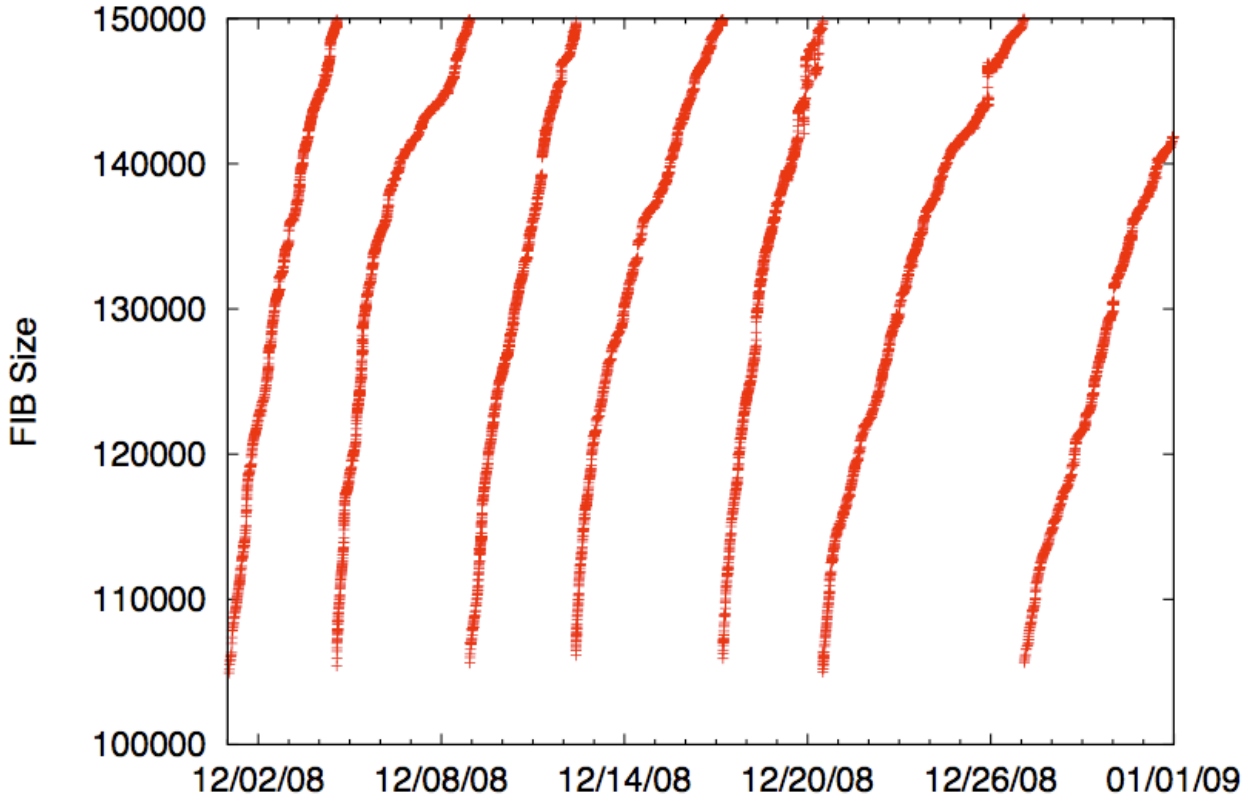
Update Processing Time

Algorithms	T_RIB(s)	t_RIB(us)	N_FIB	n_FIB	p_FIB	T_FIB(s)	t_FIB(us)
Original	4.30	0.593	2914020	2914020	1.000	2.60	0.892
Level-1	5.85	0.806	2904630	2921335	1.005	2.53	0.866
Level-2	5.96	0.822	2901530	2940178	1.013	2.45	0.833
Level-3	5.98	0.824	2900389	2941398	1.014	2.42	0.823
Level-4A	6.10	0.841	2897450	2942969	1.016	2.33	0.792
Level-4B	6.41	0.880	2913988	3388764	1.162	2.61	0.770

T_RIB: total RIB processing time;
t_RIB: average RIB processing time per routing update;
N_FIB: total number of FIB updates;
n_FIB: total number of prefixes affected in the FIB;
p_FIB: average number of affected prefixes per FIB update;
T_FIB: total FIB processing time;
t_FIB: average FIB processing time per affected prefix

- Using one month of BGP updates in 2008.12.
- Not all updates cause FIB changes (e.g., same nexthop).
- Some updates change the un-aggregated FIB, but not the aggregated FIB. (N_FIB)

Periodical Re-Aggregation



- Using one month of BGP updates of one router in 2008.12
- Full Level-4 aggregation after table size reaches 150K (50% of full table); otherwise incrementally update the aggregated FIB.
- Need run full aggregation only 7 times in a month.

Conclusion

- FA can effectively reduce FIB size
  - For large ISPs (whose FIBs probably least aggregatable), table size reduction by 30-70%, depending on the level of aggregation
- FA's computation overhead seems manageable
  - and can be controlled by incremental update plus periodic re-aggregation

- **Looking for Routing tables from operational routers for further evaluation!**

More Details

- A draft paper:
  - http://www.cs.arizona.edu/people/bzhang/paper/aggregate.pdf
- Internet Draft
  - http://www.ietf.org/id/draft-zhang-fibaggregation-02.txt

- Comments and suggestions are welcome!

- **Looking for Routing tables from operational routers for further evaluation!**