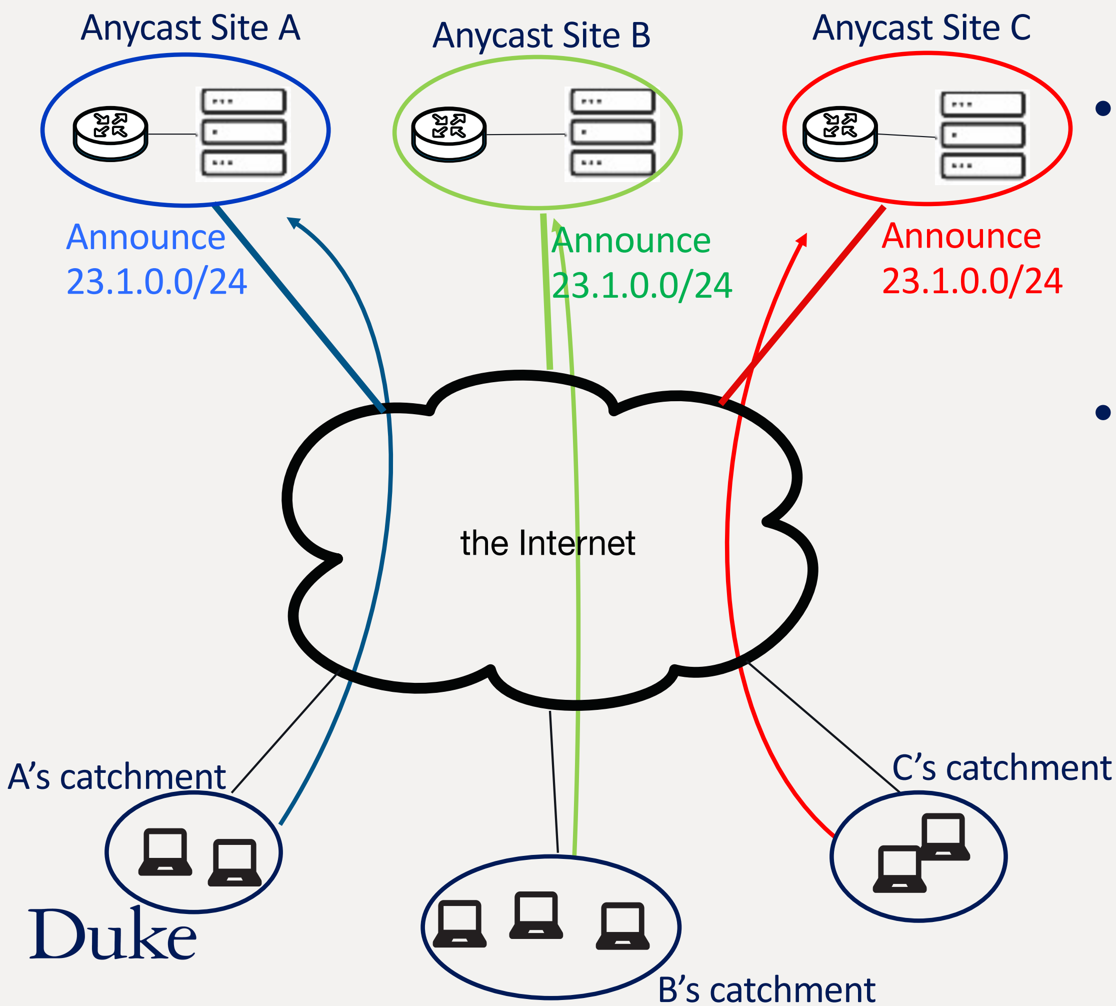
AnyOpt: Predicting and Optimizing IP Anycast Performance

Xiao Zhang, Tanmoy Sen, Zheyuan Zhang, Tim April, Balakrishnan Chandrasekaran, David Choffnes, Bruce M. Maggs, Haiying Shen, Ramesh K. Sitaraman, and Xiaowei Yang

Duke

IP Anycast



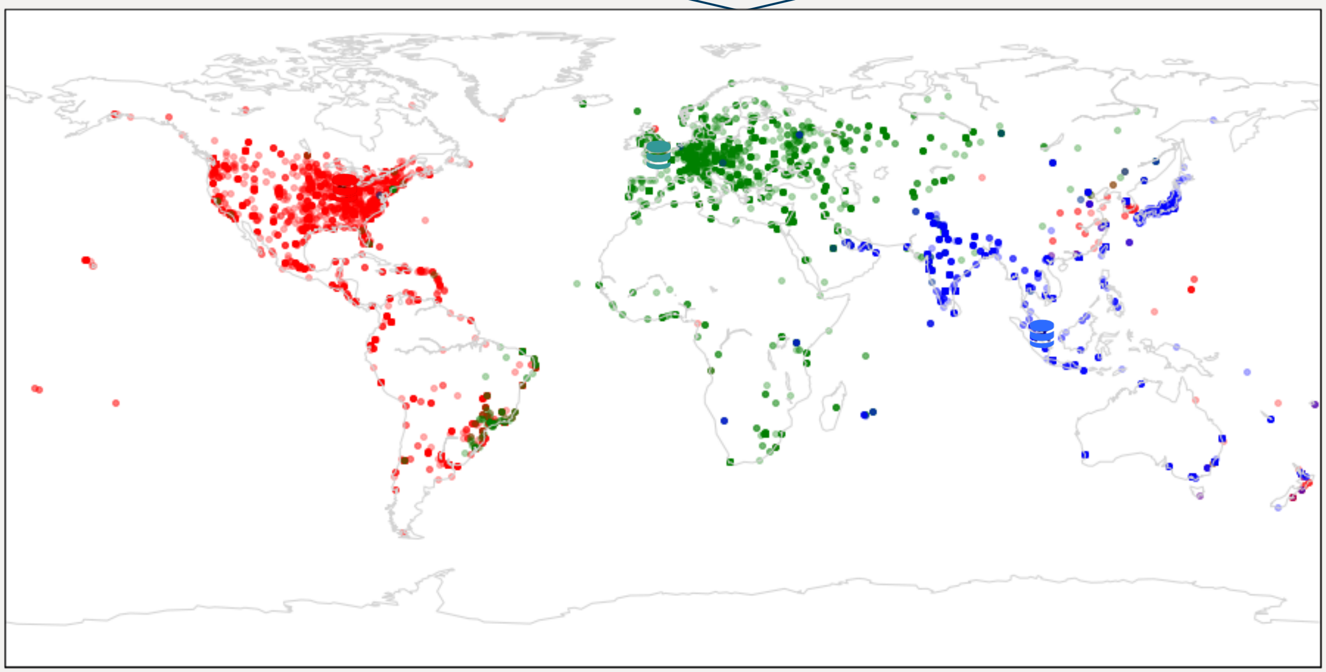
- IP anycast: the same IP prefix is announced from multiple locations
- Many services use IP anycast for performance and resilience
  - DNS
  - CDN
  - DDoS mitigation systems

Duke

Ideal anycast behavior

Ideal avg RTT 62 ms

- Anycast Site
- Client



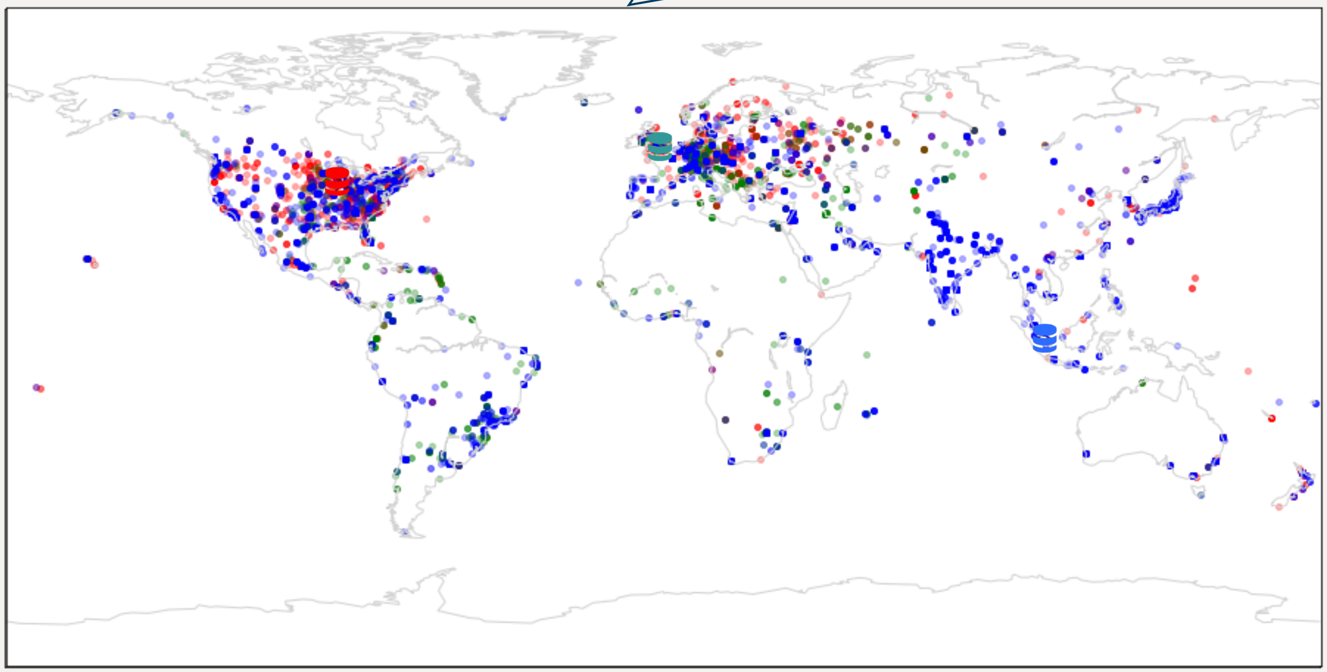
- A three-site deployment: Chicago, London, and Singapore

Duke

IP anycast does not minimize latency

Ideal avg RTT 62 ms

Actual avg RTT 133 ms



- Clients reach far-away sites

Duke

Measurement Related Work

- Around 1/3 queries suffer from serious anycast inflation over geographic distance latency; Li et al. [SIGCOMM’2018]
- Only 20%-35% of users experience serious anycast inflation. Calder et al. and Koch et al. [IMC’2015] [SIGCOMM’2021]
- Proactively measure the anycast catchment-Verfploeter Vries et al. [IMC’17]

Duke

Challenge

- A service provider needs to choose anycast sites
- BGP determines a site's catchment
- BGP is performance agnostic
- Increasing # of sites does not always reduce latency
  - E.g., Li et al. [SIGCOMM 2018], Kyle et al. [SIGCOMM 2020]

Q: How to choose a subset from potential anycast sites to minimize latency?

Estimating latency requires predicting catchment

Duke

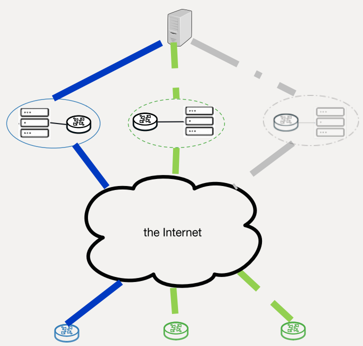
A Strawman Approach

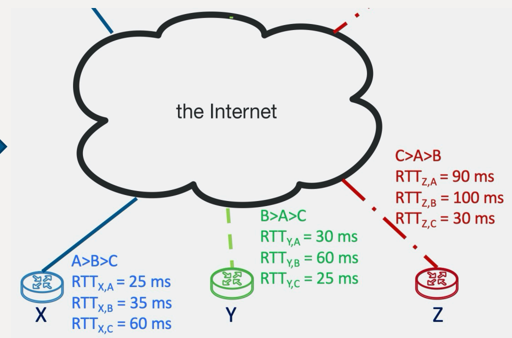
1. Experiment with all possible subsets of available sites
2. Measure each site's catchment and average client latency
3. Choose the subset with minimum average latency

➔# of experiments is exponential in # of sites

Duke

AnyOpt’s Approach





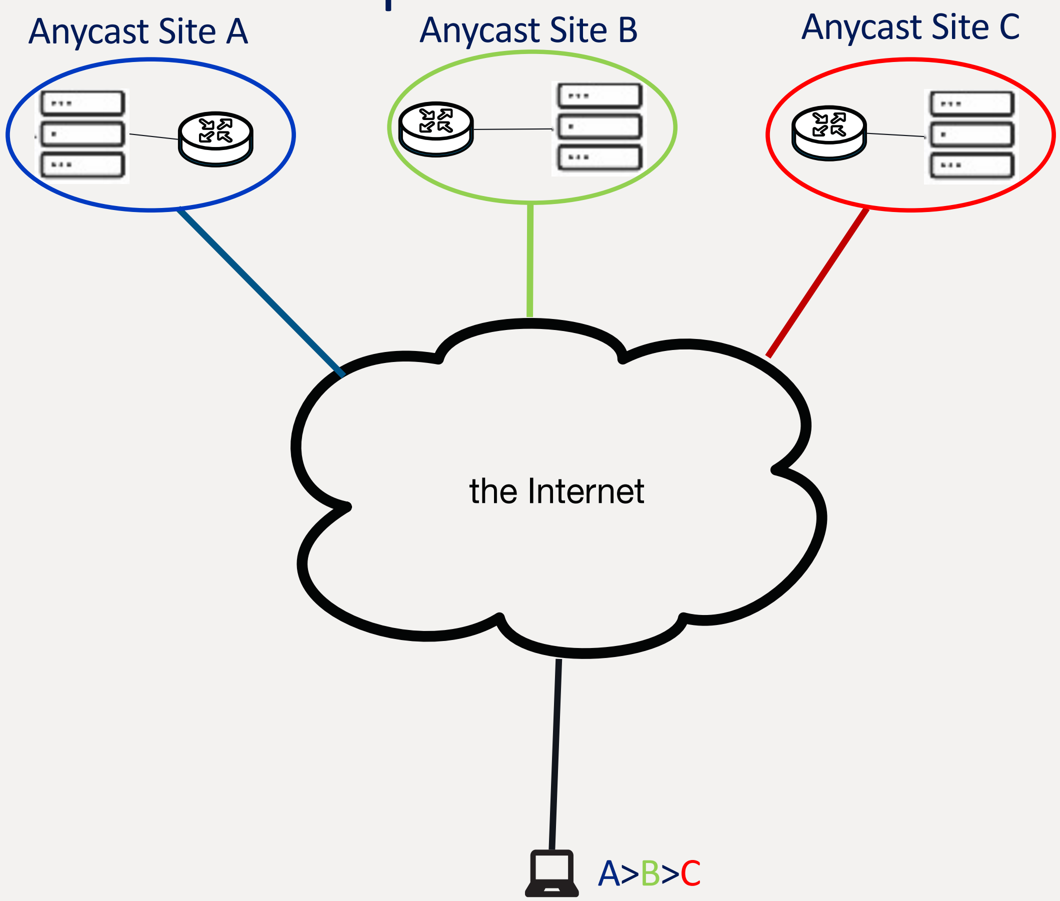
$$\min \sum_j \sum_i RTT_{j,i} \times (j\ chooses\ i\ ?\ 1:0)$$

- Measure → Model → Optimize
  - **Measure** a client’s preferences between each pair of anycast sites
  - **Model** a client’s route selection behavior as a linear preference order
  - **Solve** an optimization problem offline to minimize latency

Duke

Linear order observation and assumption

- A client's preferences form a linear order
  - E.g., A > B > C
- For any subset of the potential sites, a client will select its most preferred site
  - A, C → A
  - B, C → B
  - A,B,C → A



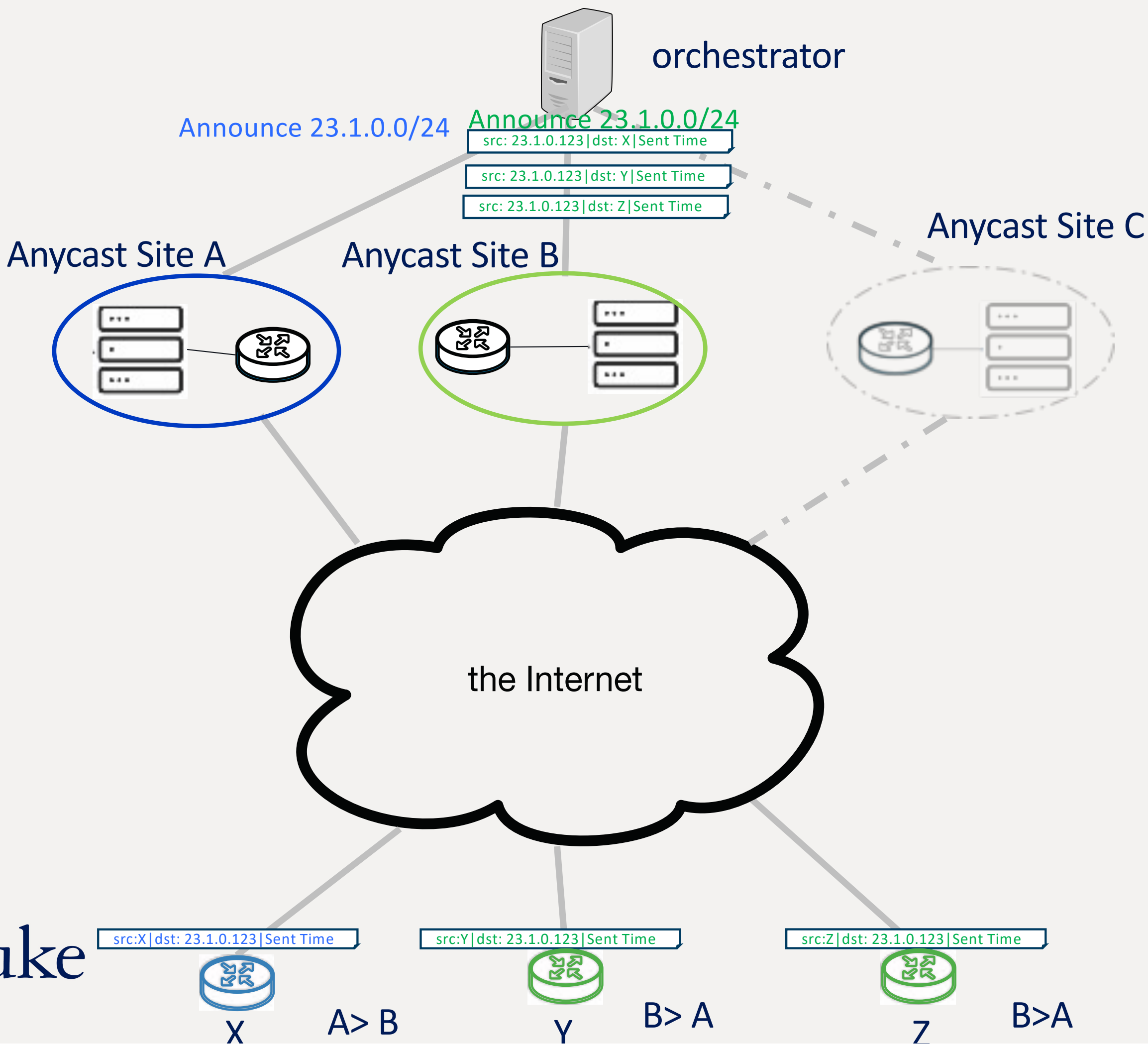
Duke

Catchment Related Work

- MPLS-based catchment control; Alzoubi et al. [TransWeb'2011]
  - Prefix-Anycast site mapping by MPLS
- Inference-based catchment prediction. Sermpezis et al. [SIGMETRICS'2019]
  - Based on BGP Table
  - And AS relationship

Duke

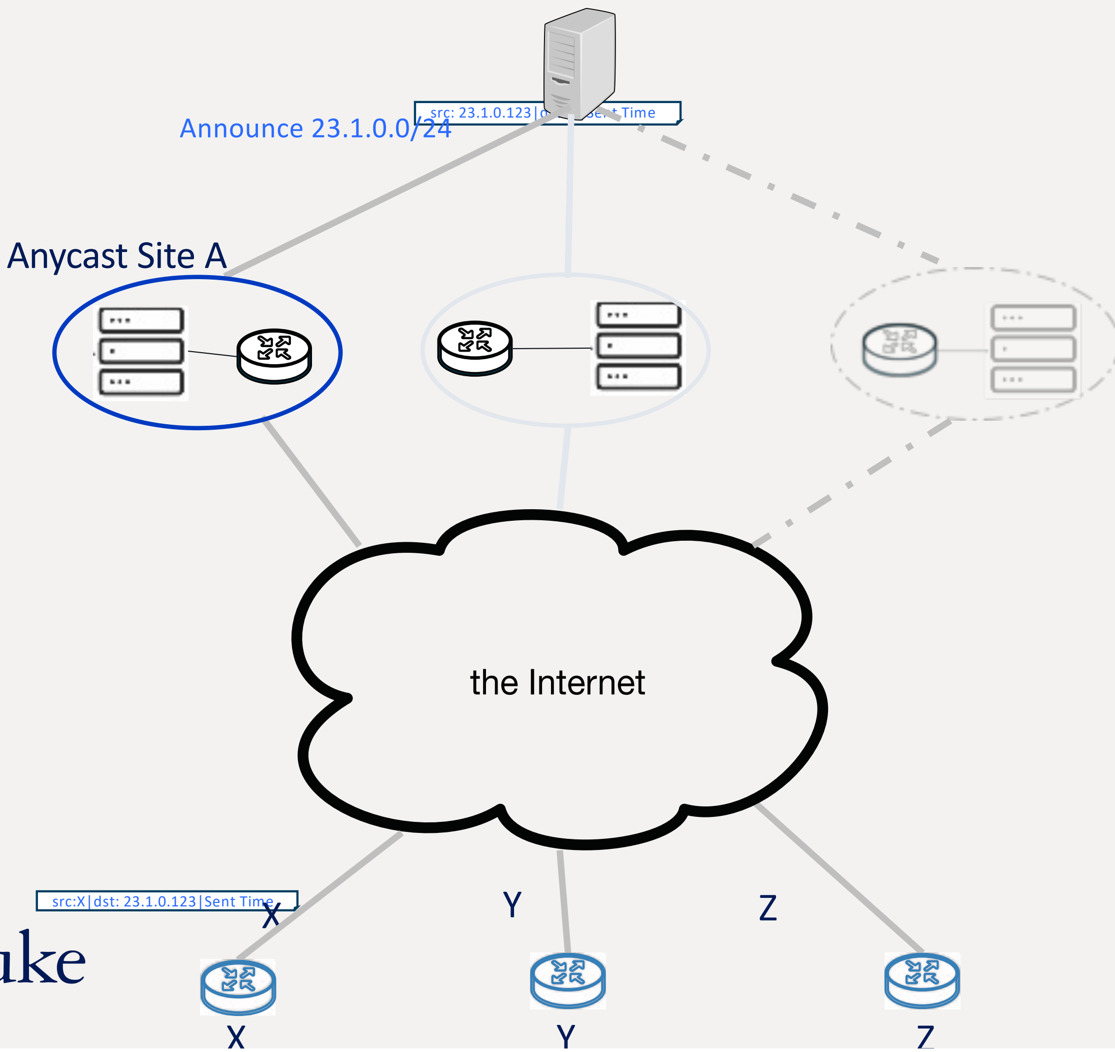
Pairwise site preference discovery



- Pair-wise comparison experiments
- Discover the preference order for all clients simultaneously
- → Reducing # of experiments to quadratic

Duke

RTT measurements



- Announce from one site
- Append sent time in ping

- Get RTT by T_{total}-T_{tunnel}

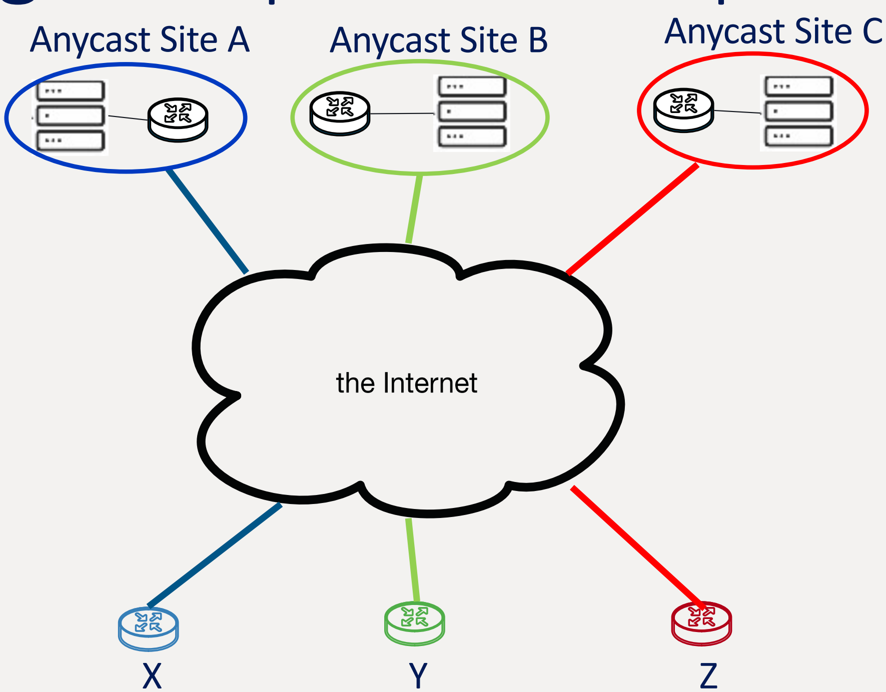
Duke

Testbed

- 15 sites around the globe
- Orchestrator connects to 15 sites with GRE tunnel
- 15,300+ router IP, 12,000+ /24 network prefixes, 5,300+ ASes

Duke

Solving the optimization problem



- Pairwise comparison → all clients' preference orders
- Measure a client j's RTT to a site i: RTT_{ji}
- → Simple facility location problem with clients' preference orderings [RSUE1987]

Input to the optimization problem

X	Y	Z
$RTT_{X,A}$ = 25 ms	$RTT_{Y,A}$ = 30 ms	$RTT_{Z,A}$ = 90 ms
$RTT_{X,B}$ = 35 ms	$RTT_{Y,B}$ = 60 ms	$RTT_{Z,B}$ = 100 ms
$RTT_{X,C}$ = 60 ms	$RTT_{Y,C}$ = 25 ms	$RTT_{Z,C}$ = 30 ms
A>B>C	B>A>C	C>A>B

$$\min \sum_{j=X,Y,Z} \sum_{i=A,B,C} RTT_{j,i} \times (j\ chooses\ i ? 1 : 0)$$

Duke

Theoretical Underpinnings

- Scenario 1:
  - Route selection based only on preference orders among neighbors
- Scenario2:
  - Announce from only tier-1 transit providers
  - Route selection based on <AS path, neighbor id>
- Consistent with “valley-free” BGP routing model [Gao&Rexford2001]
- However, a linear order may not exist for all valley-free BGP routing policies

Duke

BGP Implementation tie breaks with arrival time

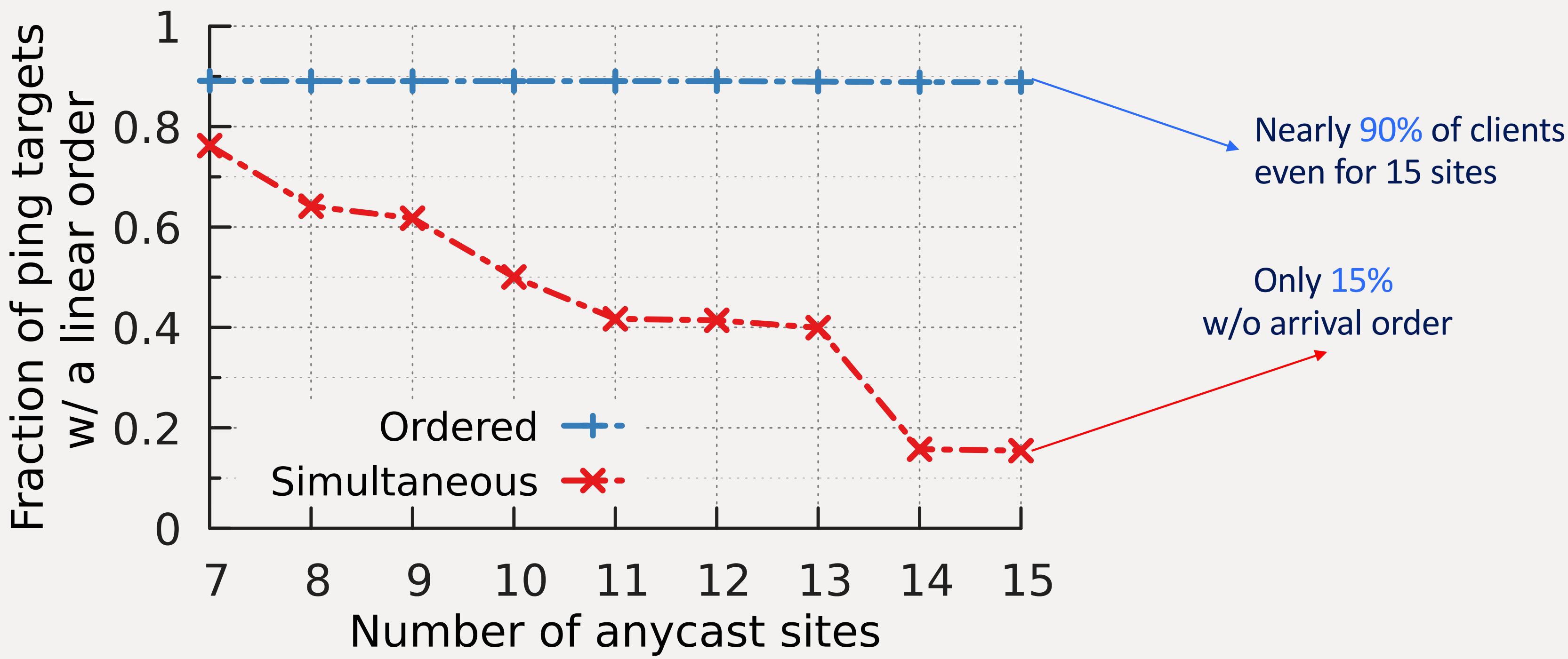
- BGP specification [RFC 4271]
  - Local preference
  - AS_PATH
  - Origin of prefix
  - MED
  - Type of BGP session
  - Interior cost
  - Router id
  - Neighbor address

- Cisco & Juniper Implementation
  - Local preference
  - AS_PATH
  - Origin of prefix
  - MED
  - Type of BGP session
  - Interior cost
  - Arrival time
  - Router id
  - Neighbor address

Announce a prefix from two sites in both orders

Duke

Total Order Preserving



Duke

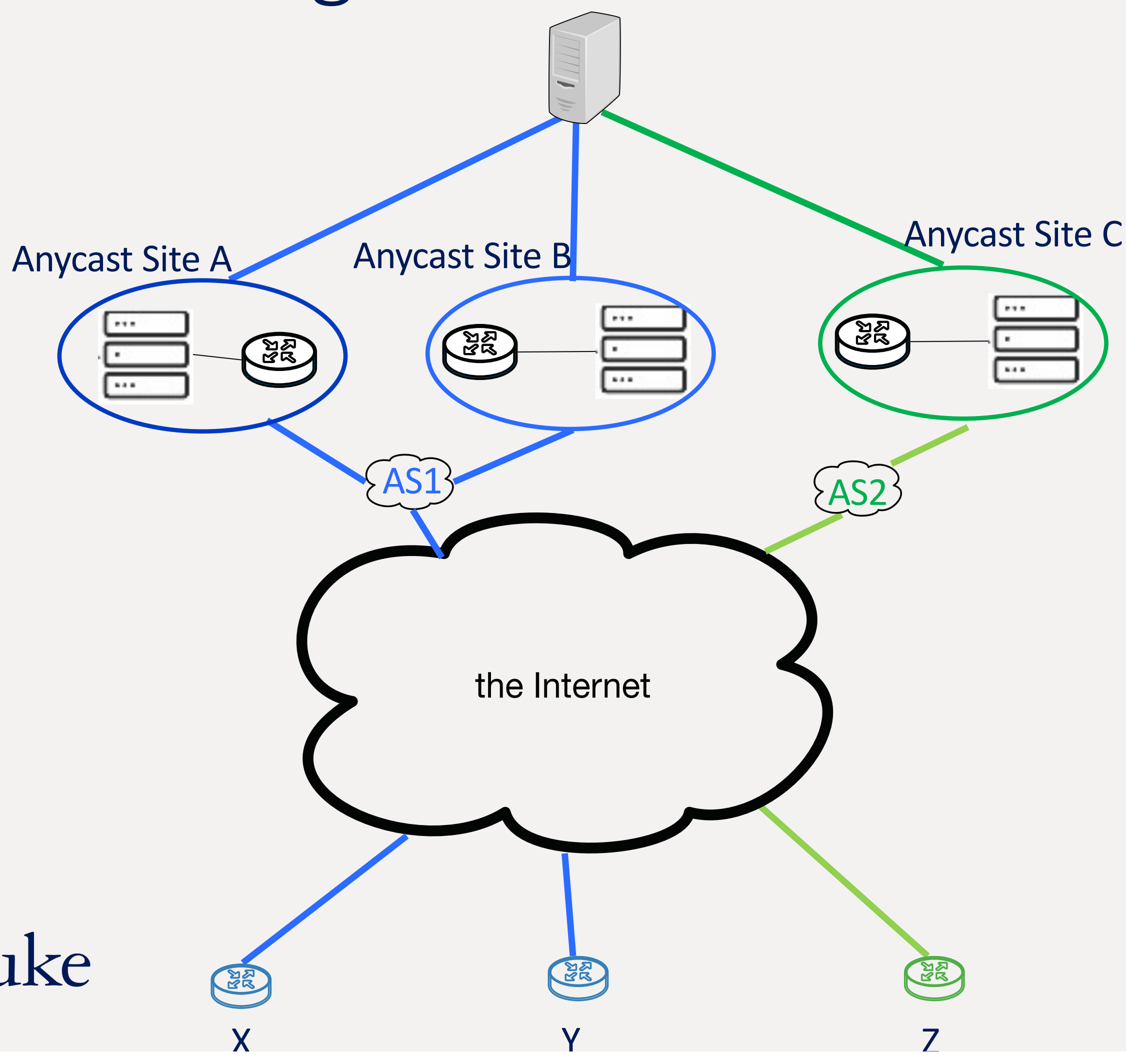
Scalability

- # of experiments is quadratic in terms of # of sites

- Example: 15 sites, 210 (i.e., 15*14) BGP experiments

Duke

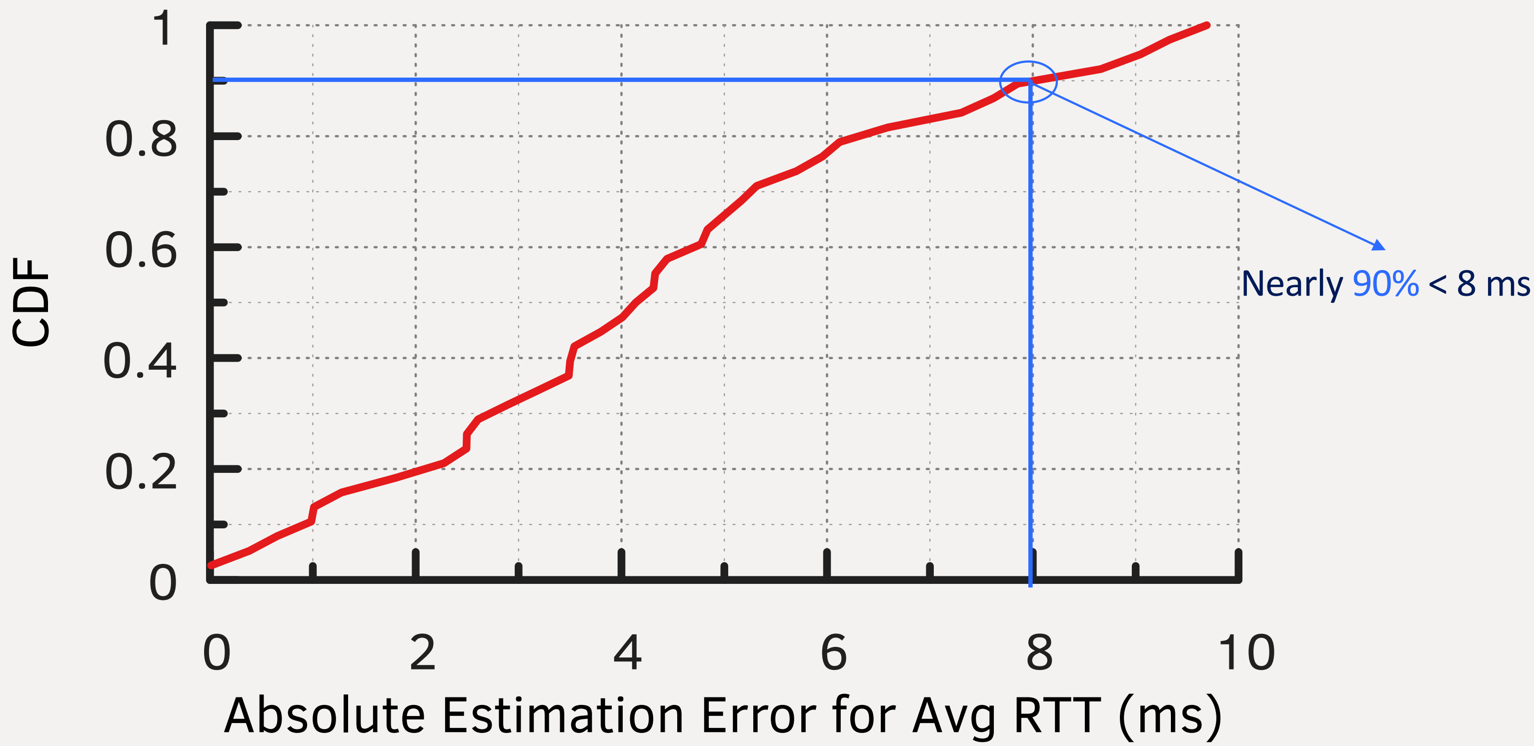
Scale to larger networks



- Two-level
  - Provider-level (6 ASes)
    - 30
  - Intra-AS level (No Arrival Order Issue)
    - 13
- 43 BGP experiments in total

Duke

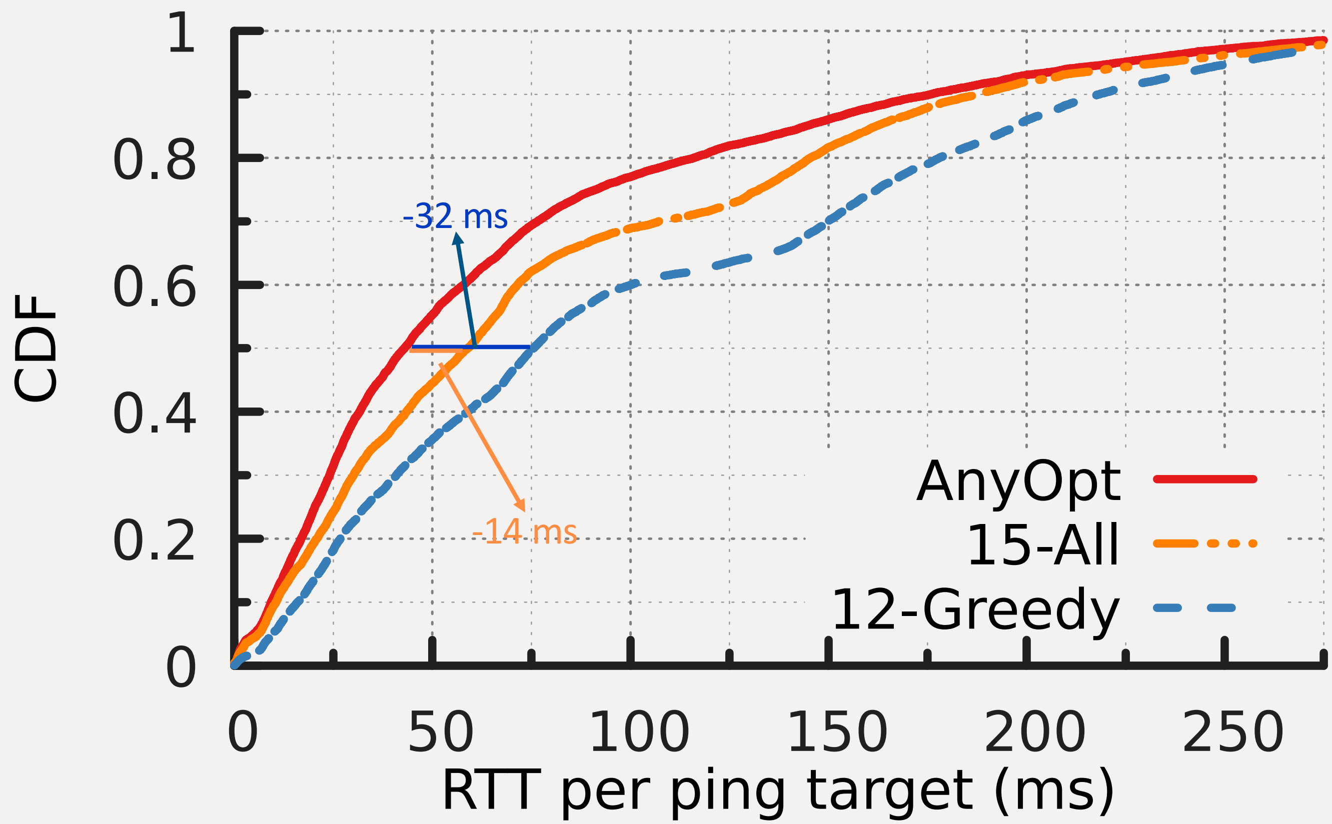
RTT estimation based on the catchment



- Deployed 38 random configurations
- Measure the actual RTTs
- Compare with the predicted RTTs

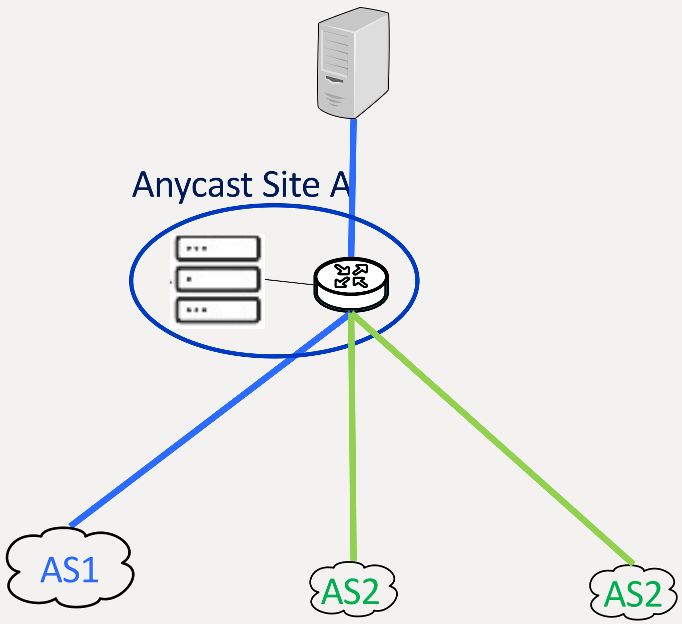
Duke

Performance Comparison

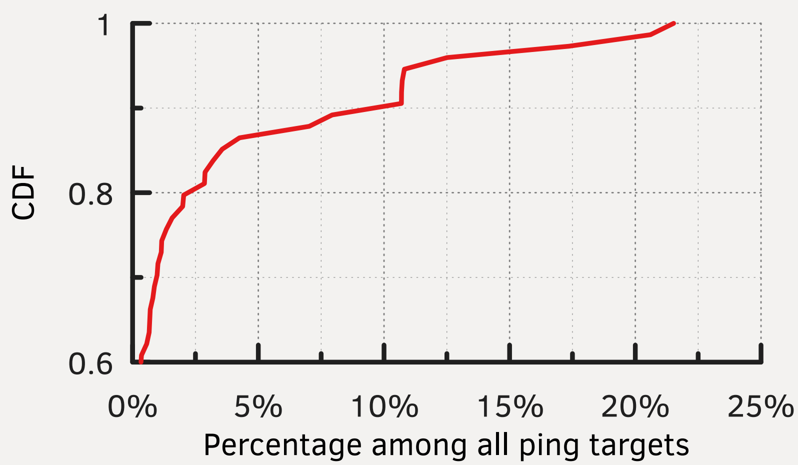


Duke

Peering Link Measurement



- Each site has
  - One transit link e.g., AS1
  - + other peering links e.g., AS2

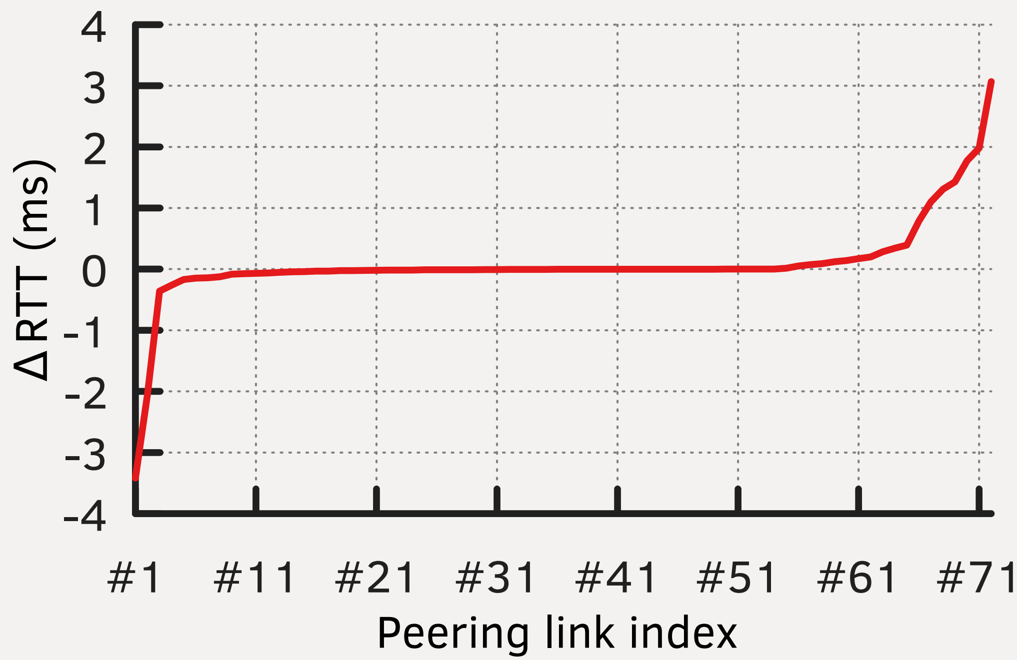


Duke

Incorporating Peering Links

- 72 peering links

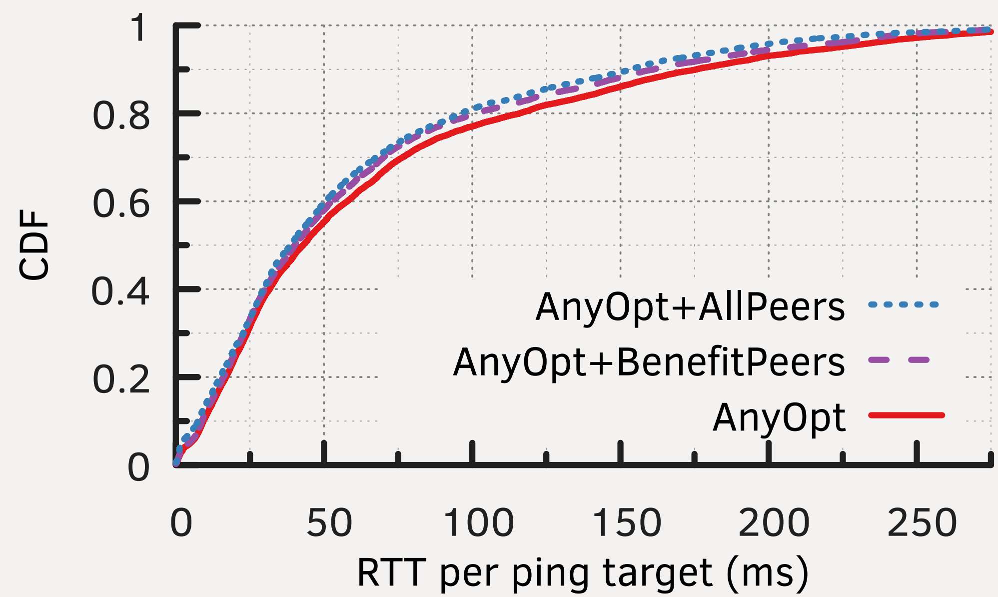
- Adding a peering link does not always improve the average RTT for clients in that cone



Duke

Incorporating Peering Links

- Adding peering links can reduce the median RTT by 7ms compared to the AnyOpt conf in our setting



Duke

Contributions

- The linear order assumption: empirical evidence and theoretical justification
- AnyOpt: a system to predict anycast catchment and optimize anycast configurations
- Evaluation using a real-world testbed

Duke

Future Work

- Scale to larger network;
  - Akamai DNS with hundreds of sites
- Optimize for other objectives
  - Robustness
  - Load balance
- Accurate prediction with peering links

Duke