ALTO Survey

draft-hilt-alto-survey-00

Marco Tomsu,
Volker Hilt, Vijay Gurbani, Enrico Marocco, Ivica Rimac

72nd IETF Meeting

Outline

- How to select a good peer?
  - Application Level
    - Direct Measurement
    - ID Maps
    - Vivaldi
    - iPlane
    - Ono
  - Layer Cooperation
    - Provider Portal for Applications (P4P)
    - Oracle-based ISP-P2P Collaboration
    - ISP Driven Informed Path Selection (IDIPS)

Packet Dispersion Techniques

[Dovrolis et al., INFOCOM 2001]

Basic idea:
Estimate bottleneck bandwidth
e.g. from the **dispersion** experienced by back-to-back packets or packet trains
(fluid analogy)

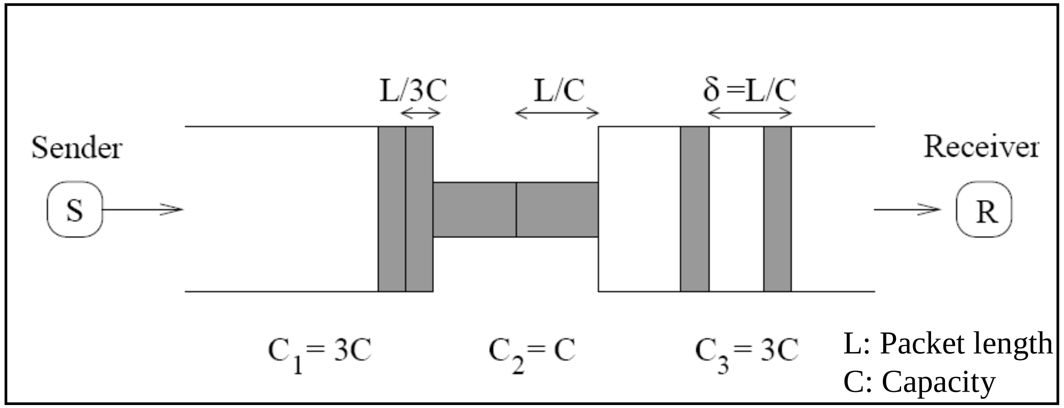
Practically:
Only the available bandwidth at a given time is measured (unused capacity)

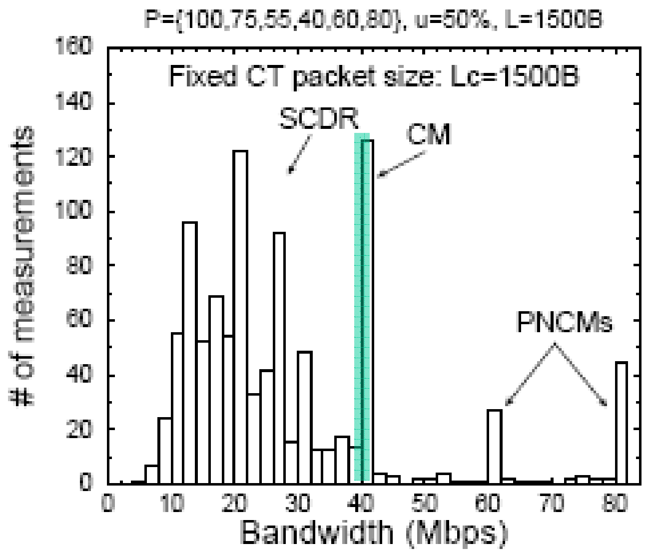
Interference:
Queuing delays (e.g. cross traffic) lead to measurements showing multi-modal behavior

Statistical + heuristic approaches to resolve
→ **Very good accuracy can be achieved**

Simple to implement on end points: Used for peer/path selection (BitTorrent), codec selection (Skype) …

Scalability issue: Suitable for a small candidate set of peers





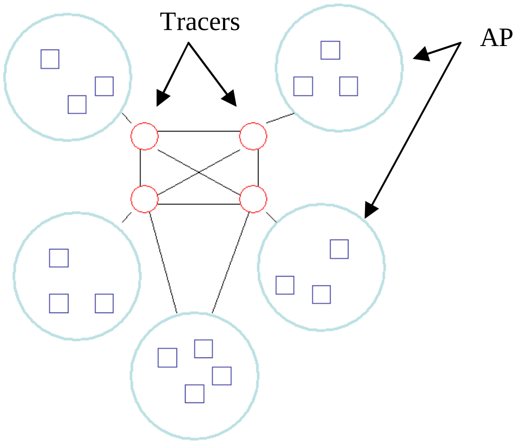
CM: Capacity Mode (desired measurement result)
SCDR: Sub Capacity Dispersion Range (queues increase dispersion)
PNCM: Post Narrow Capacity Modes (queues can decrease packet delay

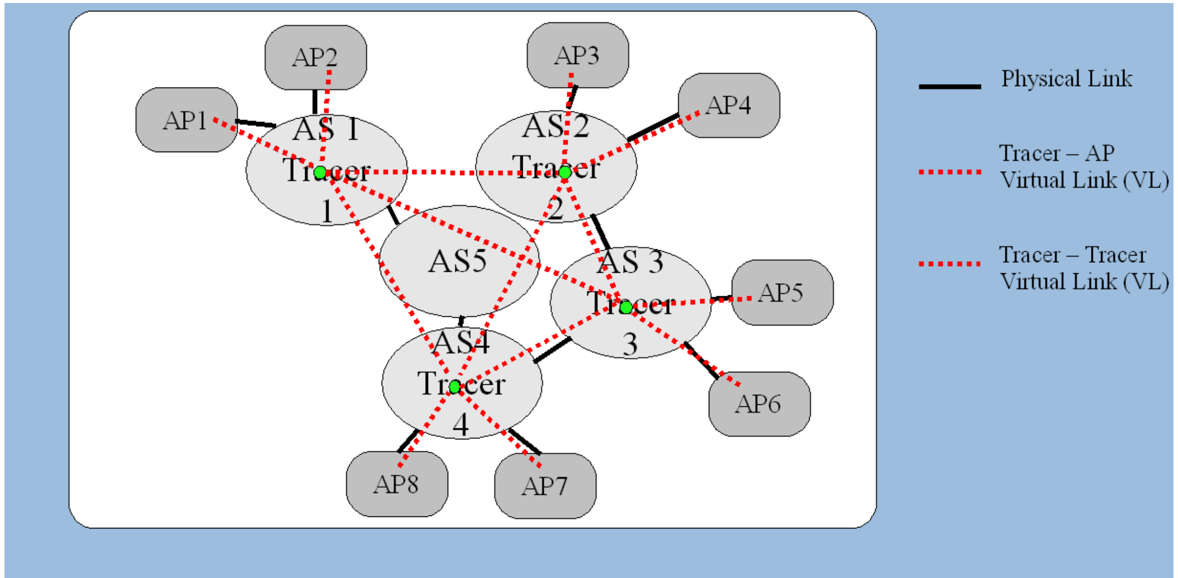
IDMaps

[Francis et al., IEEE/ACM ToN 2001]

Definitions:

1. Address Prefix (AP): Consecutive IP address range within which all hosts with assigned addresses are equidistant (with some tolerance) to the rest of the Internet.
2. Tracer: One or more special host(s) deployed near an AS. Inter-Tracer distance and AP->Tracer distances are measured.
3. Virtual Link (VL): Raw distance between two tracers, and between a tracer and AP.





distance(AP1,AP6)=distance(AP1,Tracer1)+distance(Tracer1,Tracer3)+distance(Tracer3,AP6)

Graphic source: Dragan Milic, University of Bern

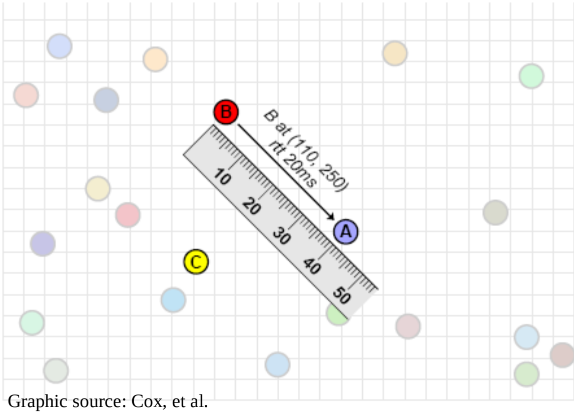
Drawbacks:

- Infrastructure support needed: at least one tracer per AS.
- Scalability: $O(n^2)$ as each tracer measures and stores RTT to all other tracers.
- Performance depends heavily on the placement and number of tracers.

Vivaldi

[Dabek, et al. SIGCOMM 2004]

A computes distance to B in coordinate space.

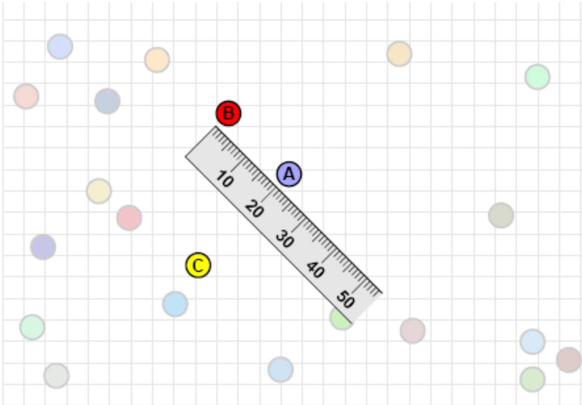


Graphic source: Cox, et al.
http://swtch.com/~rsc/talks/vivaldi-ccs.pdf

Vivaldi

[Dabek, et al. SIGCOMM 2004]

A adjusts coordinates so distance matches actual RTT.

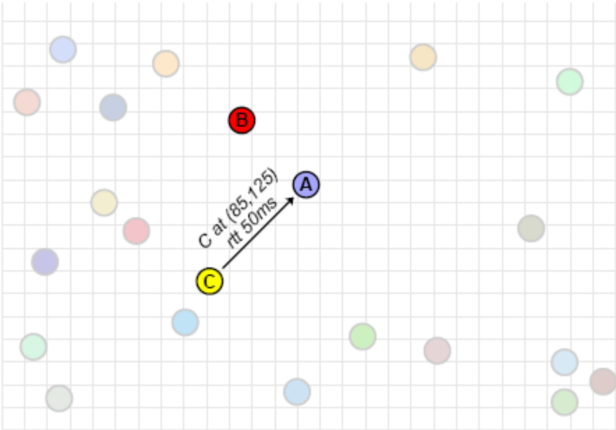


Graphic source: Cox, et al.
http://swtch.com/~rsc/talks/vivaldi-ccs.pdf

Vivaldi

[Dabek, et al. SIGCOMM 2004]

A obtains C's coordinates, RTT.

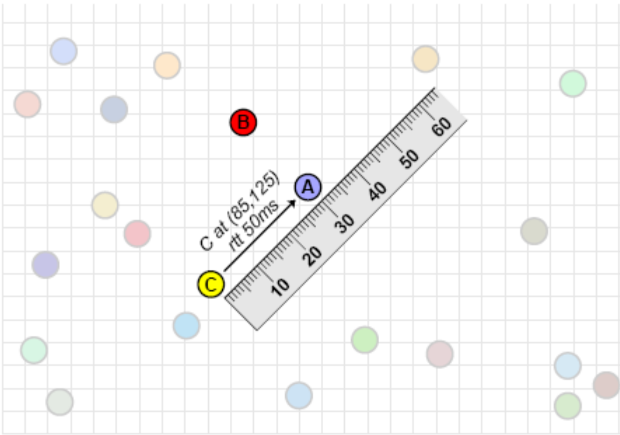


Graphic source: Cox, et al.
http://swtch.com/~rsc/talks/vivaldi-ccs.pdf

Vivaldi

[Dabek, et al. SIGCOMM 2004]

A computes distance to C in coordinate space.

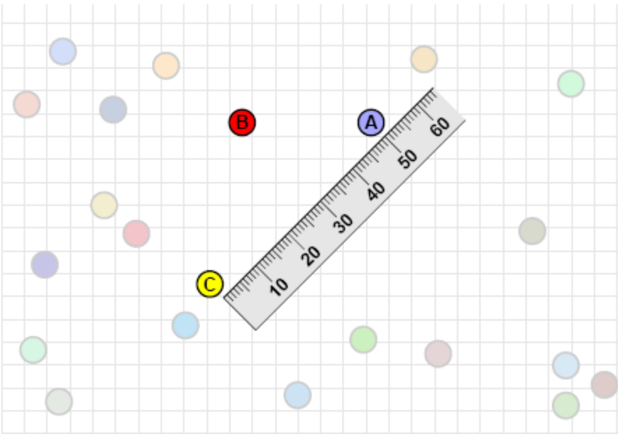


Graphic source: Cox, et al.
http://swtch.com/~rsc/talks/vivaldi-ccs.pdf

Vivaldi

[Dabek, et al. SIGCOMM 2004]

A adjusts coordinates so distance matches actual RTT.
(Now A is wrong distance from B.)

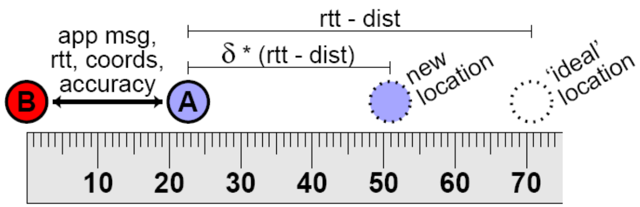


Graphic source: Cox, et al.
http://swtch.com/~rsc/talks/vivaldi-ccs.pdf

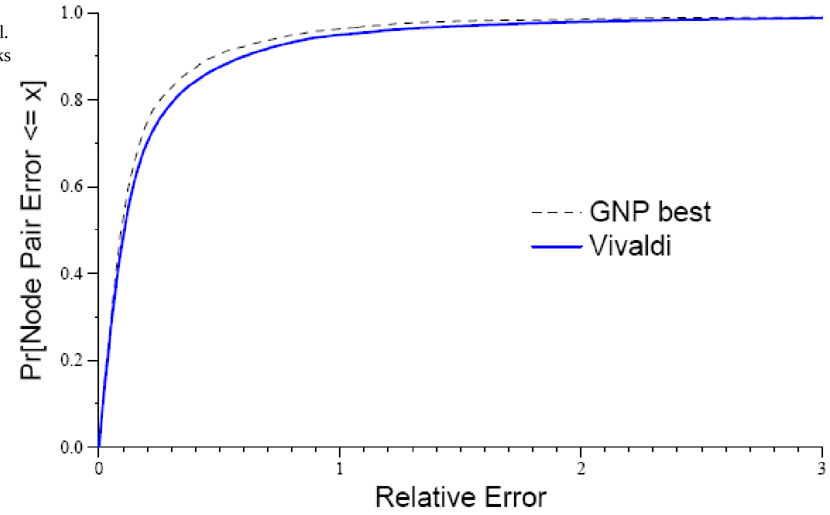
Vivaldi Algorithm

Given the coordinates, round trip time, and accuracy estimate of a node:

- Update local accuracy estimate.
- Compute 'ideal' location.
- Compute damping constant δ using local and remote accuracy estimates.
- Move δ of the way toward the "ideal" location.



Graphic source: Cox, et al.
http://swtch.com/~rsc/talks



Relative Error = | Actual RTT – Predicted RTT| / *min*(Actual RTT, Predicted RTT)

Data for plot: 1,000 node network initialized and allowed to converge. Then 1,000 new nodes added one at a time.

Used as plugin-in for Azureus (BitTorrent client)

Fundamental issue with Network Coordinates:
Triangular Inequality not always given

iPlane: An Information Plane for Distributed Services

[Madhyastha et al., USENIX OSDI 2006; http://iplane.cs.washington.edu/]

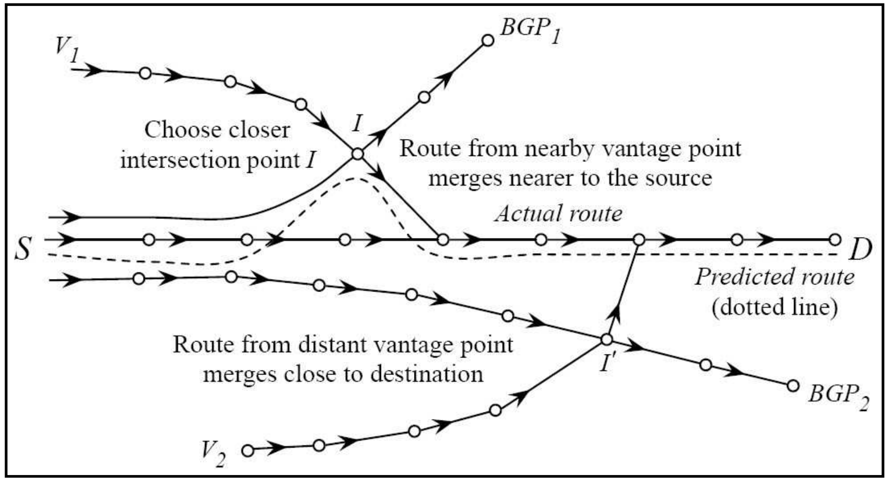
1. Builds a structured Internet atlas

- Uses PlanetLab + public traceroute servers
  ⇒ >700 distributed vantage points
- Clusters IP prefixes into BGP atoms
- Traceroutes from vantage points to BGP atoms
- Clusters network interfaces into PoPs

2. Annotates the atlas

- Latency, loss rate, capacity, avail. bandwidth
- Active measurements in the core
- Opportunistic edge measurements using
  a modified BitTorrent client

3. Predicting routes between arbitrary end-hosts

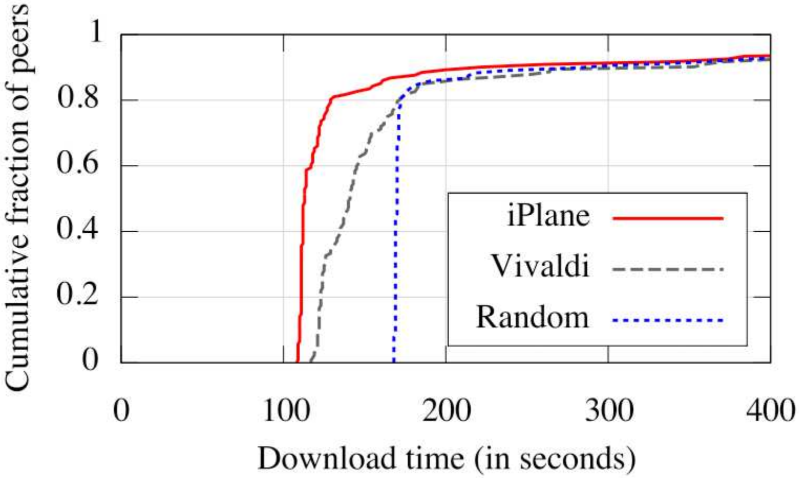


4. Predicting end-to-end path properties:

Latency	Sum of link latencies
Loss-rate	Product of link loss-rates
Bandwidth	Minimum of link bandwidths

A BitTorrent study case

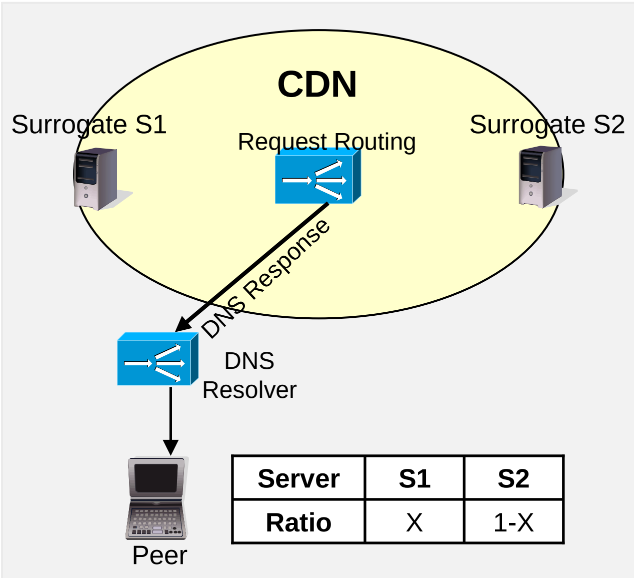
- 150 nodes swarm size
- 50 MB file size



Taming the Torrent (Ono Project)

[Choffnes and Bustamante, SIGCOMM 2008; http://www.aqualab.cs.northwestern.edu/projects/Ono.html]

- CDN-based oracle implementation for biased peer selection in BitTorent (Azureus plugin)
- Recycles network views gathered by CDNs (Akamai and Limelight)



Server	S1	S2
Ratio	X	1-X

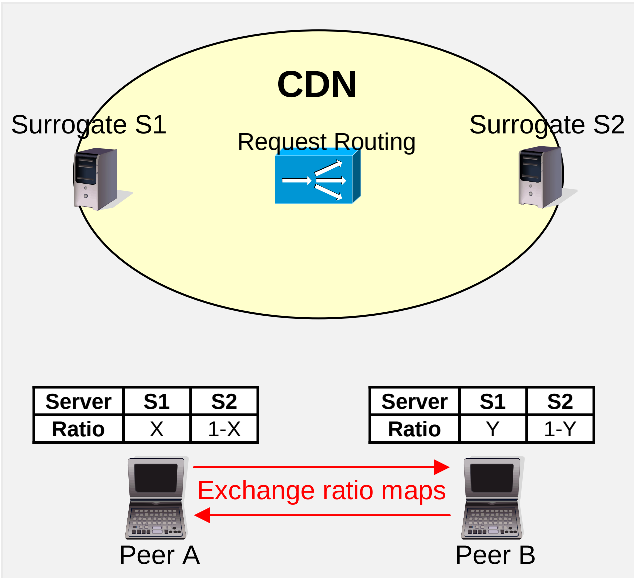
Peer–observed DNS redirection

- An Ono-enabled BT peer periodically looks up a list of CDN names
- The request routing system in the CDN triggers distance measurements (RTT) between the surrogates and the peer's local DNS server
- The peer is redirected to the "best" surrogate server
- The peer updates its redirection ratio map

Taming the Torrent (Ono Project)

[Choffnes and Bustamante, SIGCOMM 2008; http://www.aqualab.cs.northwestern.edu/projects/Ono.html]

- CDN-based oracle implementation for biased peer selection in BitTorent (Azureus plugin)
- Recycles network views gathered by CDNs (Akamai and Limelight)



Peer–observed DNS redirection

- An Ono-enabled BT peer periodically looks up a list of CDN names
- The request routing system in the CDN triggers distance measurements (RTT) between the surrogates and the peer's local DNS server
- The peer is redirected to the "best" surrogate server
- The peer updates its redirection ratio map

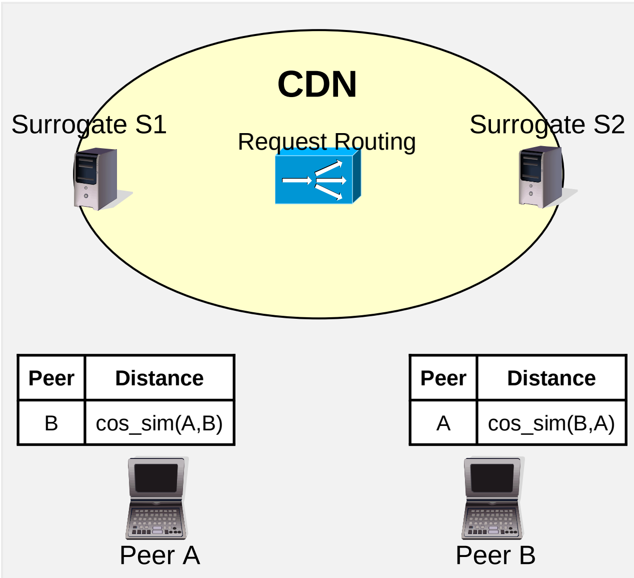
Biasing traffic

- Ono-enabled peers exchange ratio maps at connection handshake

Taming the Torrent (Ono Project)

[Choffnes and Bustamante, SIGCOMM 2008; http://www.aqualab.cs.northwestern.edu/projects/Ono.html]

- CDN-based oracle implementation for biased peer selection in BitTorent (Azureus plugin)
- Recycles network views gathered by CDNs (Akamai and Limelight)



Peer–observed DNS redirection

- An Ono-enabled BT peer periodically looks up a list of CDN names
- The request routing system in the CDN triggers distance measurements (RTT) between the surrogates and the peer's local DNS server
- The peer is redirected to the "best" surrogate server
- The peer updates its redirection ratio map

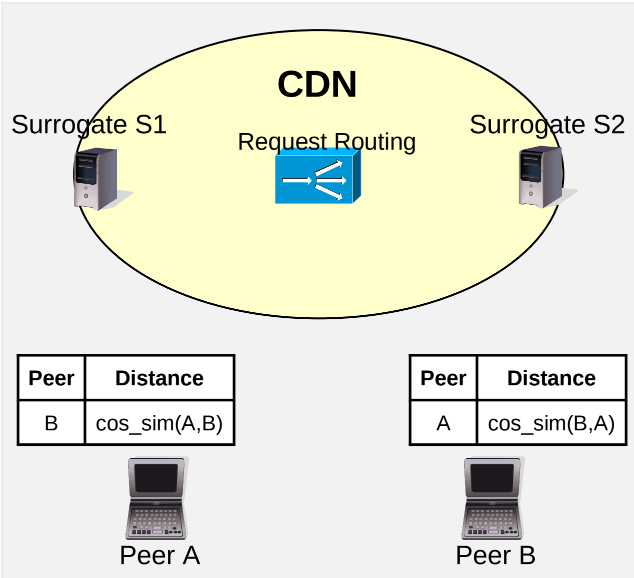
Biasing traffic

- Ono-enabled peers exchange ratio maps at connection handshake
- Peers are computing the cosine similarity of their redirection ratios (values on a scale of [0,1])
- A peer attempts to bias traffic toward a neighbor with similarity greater than a threshold (0.15)

Taming the Torrent (Ono Project)

[Choffnes and Bustamante, SIGCOMM 2008; http://www.aqualab.cs.northwestern.edu/projects/Ono.html]

- CDN-based oracle implementation for biased peer selection in BitTorent (Azureus plugin)
- Recycles network views gathered by CDNs (Akamai and Limelight)



Peer–observed DNS redirection

- An Ono-enabled BT peer periodically looks up a list of CDN names
- The request routing system in the CDN triggers distance measurements (RTT) between the surrogates and the peer's local DNS server
- The peer is redirected to the "best" surrogate server
- The peer updates its redirection ratio map

Biasing traffic

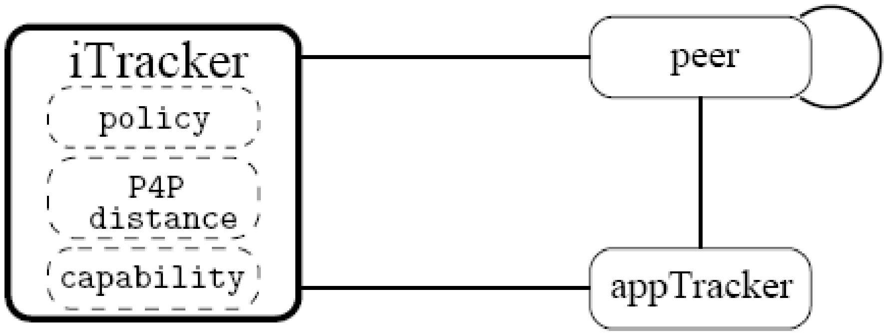
- Ono-enabled peers exchange ratio maps at connection handshake
- Peers are computing the cosine similarity of their redirection ratios (values on a scale of [0,1])
- A peer attempts to bias traffic toward a neighbor with similarity greater than a threshold (0.15)

Some measured BT results

- Download rate improvements of 31-207%
- 33% of the time selected peers are within a single AS

Provider Portal for Applications (P4P)

[Xie et al., SIGCOMM 2008]



P4P-distance interface:

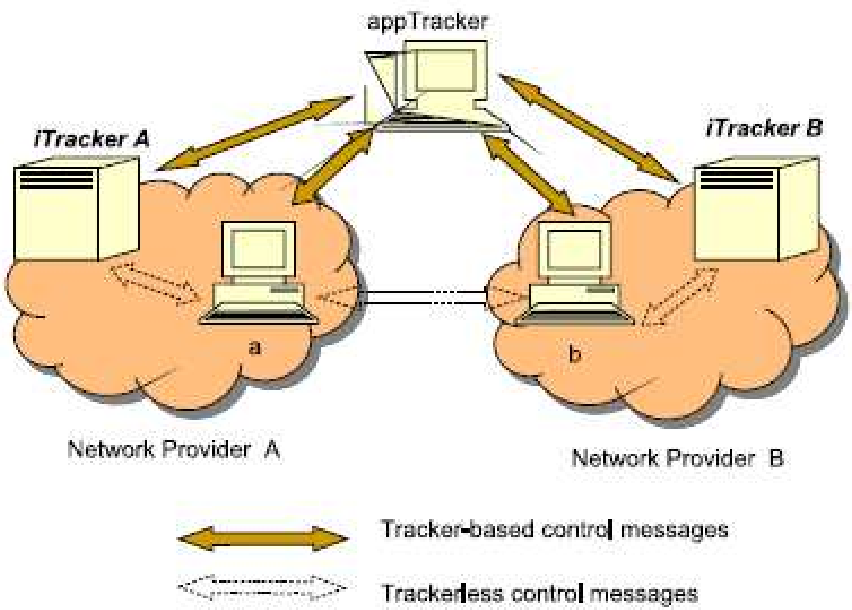
- IPs are mapped on PIDs (e.g. a PID represents a subnet)
- P4P-distance measured between PIDs

Policy interface:

- E.g. time-of-day link usage policy

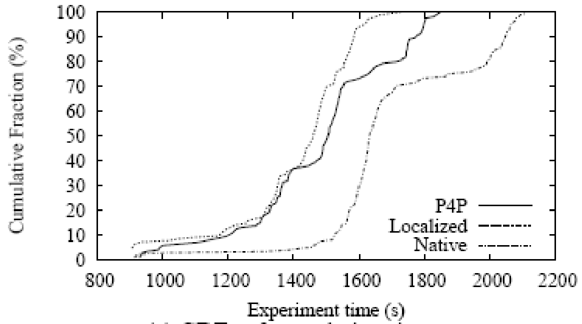
Capability interface:

- E.g. cache locations

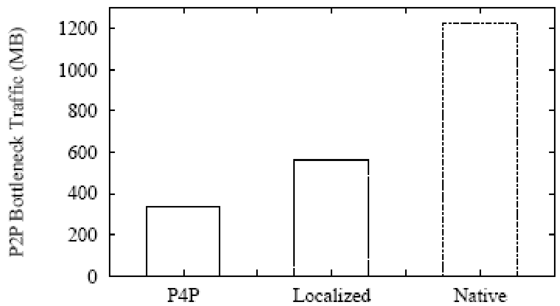


Simulations, PlanetLab experiments and field tests

http://openp4p.net/front/fieldtests



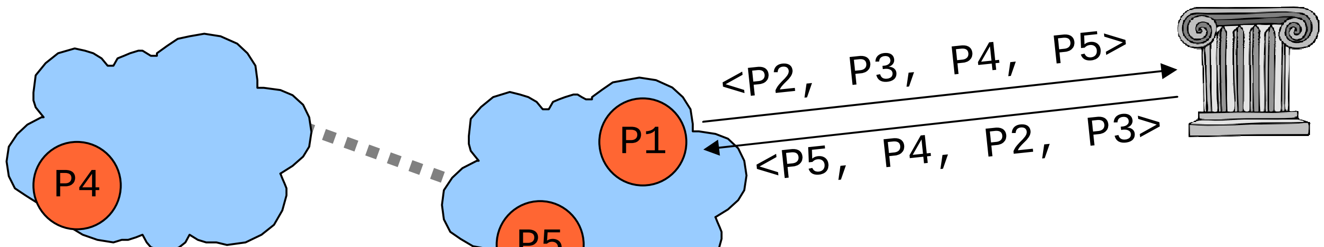
(a) CDFs of completion time.

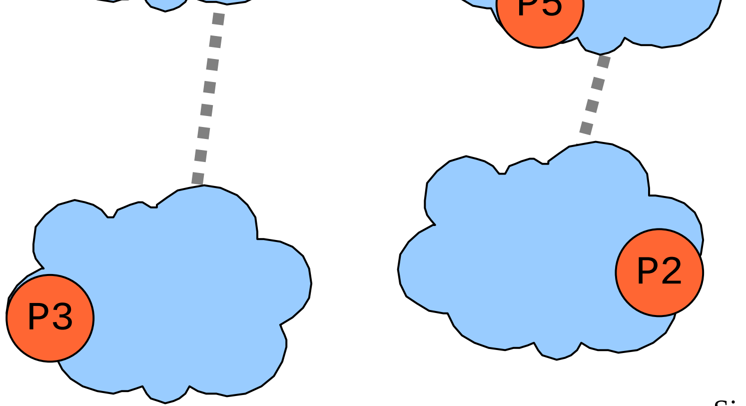


(b) P2P bottleneck traffic.

Oracle-based ISP-P2P Collaboration

[Aggarwal et al., SIGCOMM 2007, Aggarwal et al., IEEE GIS 2008]

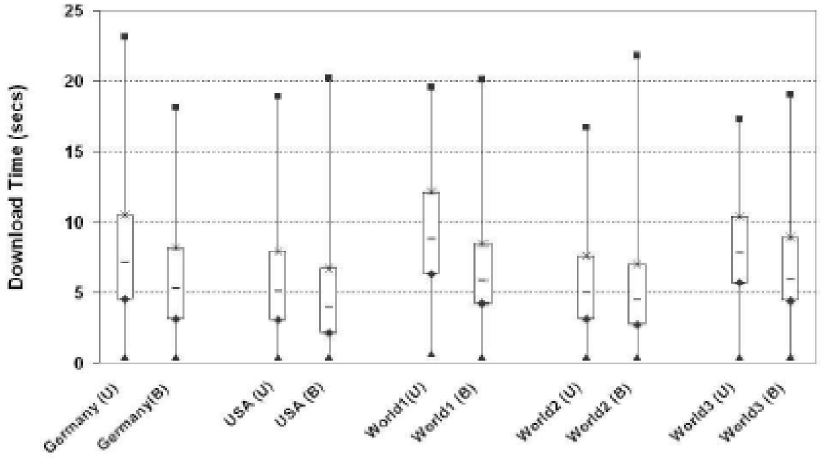




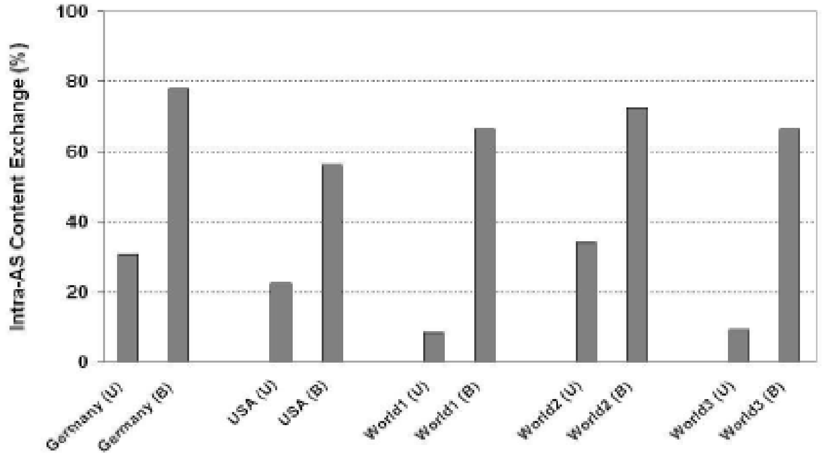
Ranking based on:

- Inside/outside of the AS
- Number of AS hops according to BGP path
- Distance to the edge of the AS according to IGP metric
- Geographic information (e.g. same PoP, same city)
- Performance information (e.g. expected delay, bandwidth)
- Link congestion

Simulations and PlanetLab experiments



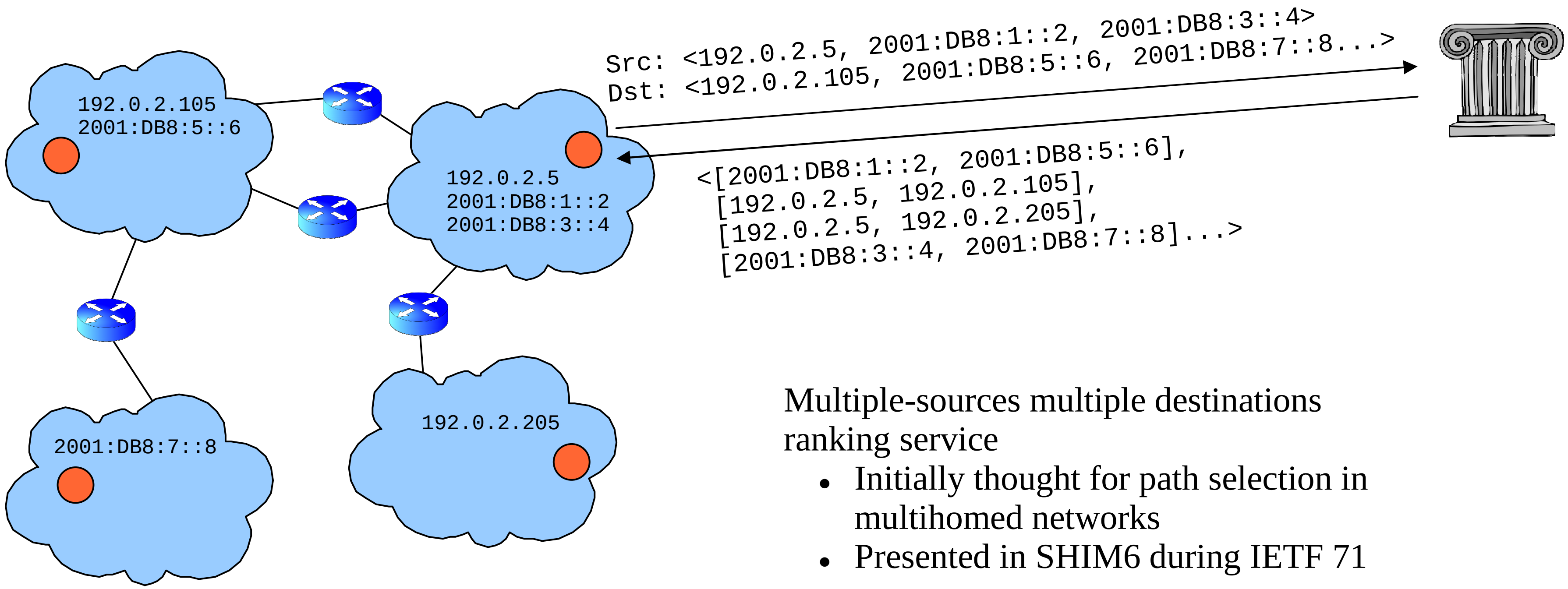
(a) File download time - box plot [36]



(b) Amount of intra-AS file exchange - bar plot

ISP Driven Informed Path Selection (IDIPS)

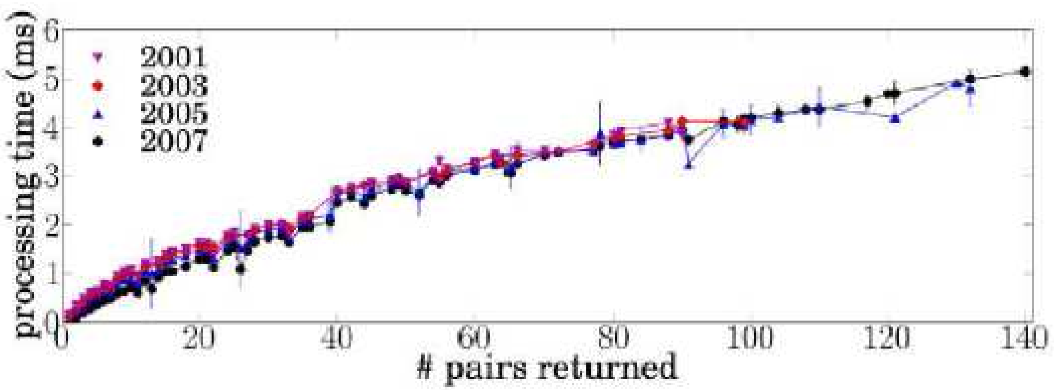
[draft-bonaventure-informed-path-selection, Saucez et al., ACM CoNEXT 2007]



Multiple-sources multiple destinations ranking service

- Initially thought for path selection in multihomed networks
- Presented in SHIM6 during IETF 71

Performance evaluation



Thanks!

Application Level

- Direct Measurement
- ID Maps
- AS Aware Peer-Relay Protocol (ASAP)
- Global Network Positioning (GNP)
- Vivaldi
- Meridian
- iPlane
- Ono

Layer Cooperation

- Provider Portal for Applications (P4P)
- Oracle-based ISP-P2P Collaboration
- ISP Driven Informed Path Selection (IDIPS)

More references can be found in the draft.

Annex

Global Network Positioning (GNP)

[Ng and Zhang, ACM IMW 2001, IEEE Infocom 2002]

Two part architecture:

1. Landmark operations.
2. Ordinary host operations.

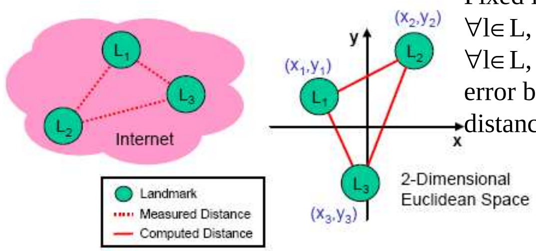


Fig. 2. Part 1: Landmark operations

Fixed landmarks, L, selected.
$\forall l \in L$, compute mutual distances.
$\forall l \in L$, compute coordinates by minimizing error between measured distance and computed distance: Minimize error($d_{i,j}$, $D_{i,j}$).

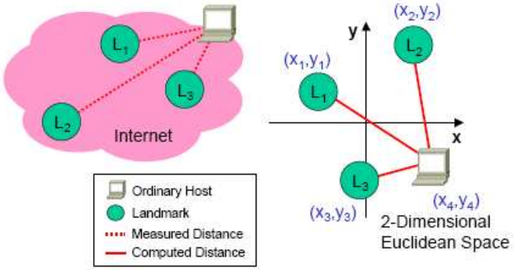


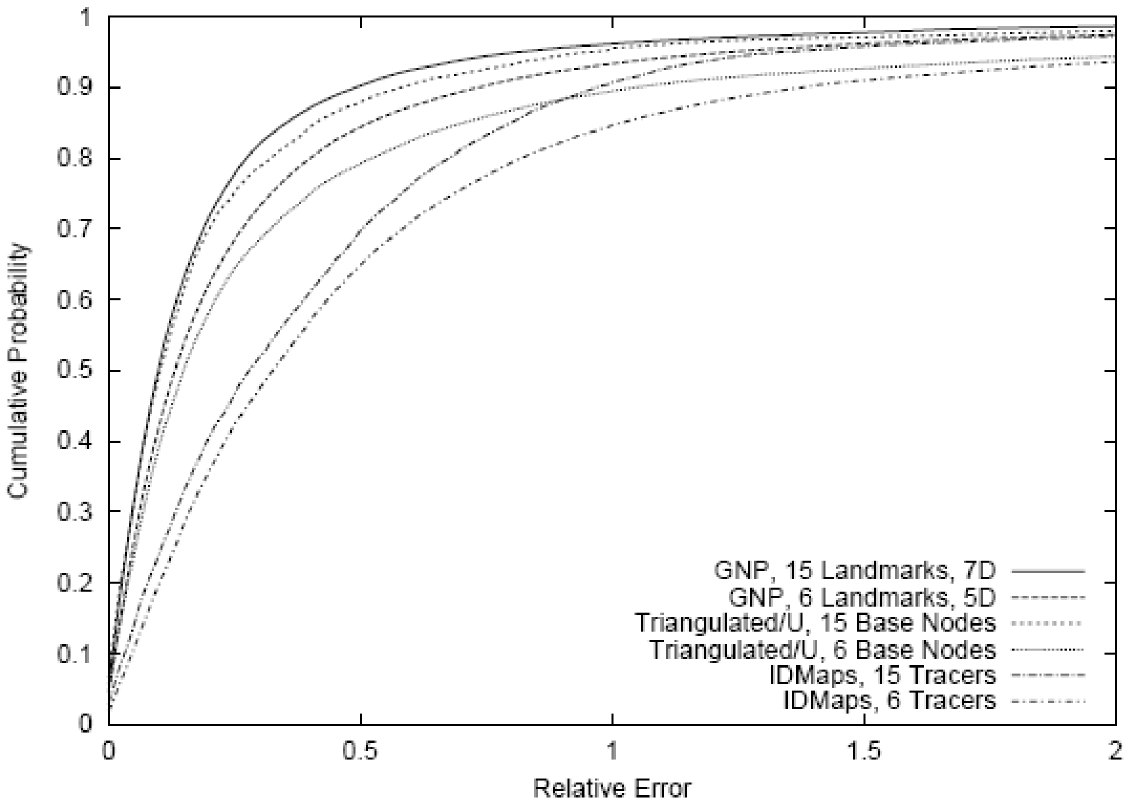
Fig. 3. Part 2: Ordinary host operations

Host, h, receives coordinates to all L landmarks.
Host, h, computes distance to all L landmarks.
Host computes own coordinates relative to L.
Compute own coordinates by minimizing error between measured distance from h to L_i and computed distance between h to L_i:
Minimize error($d_{h,Li}$,$D_{h,Li}$)

Issues in GNP:

- Coordinates not unique.
- Landmark failure and overload.
- Where to place landmarks?
- How many dimensions (diminishing returns after a certain number of dimensions.)

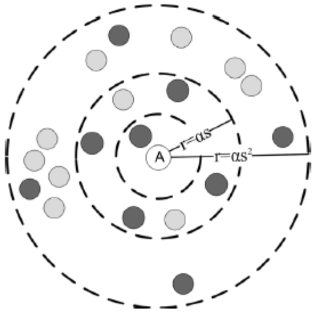
Results: With 15 landmarks, GNP predicts 90% of all paths with relative error of <= 0.5.



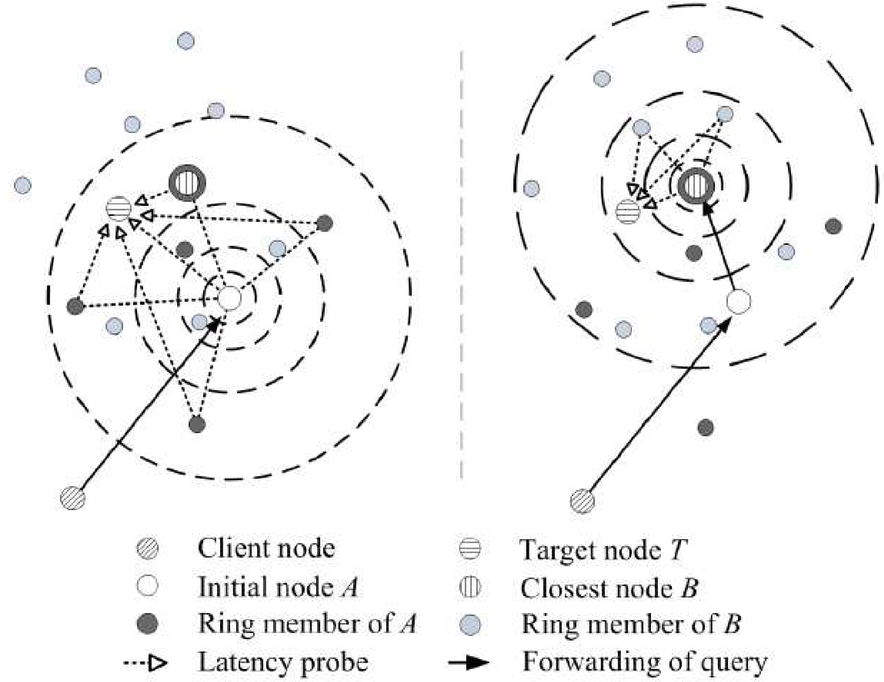
Meridian

[Wong, et al. SIGCOMM 2005]

No infrastructure support needed.

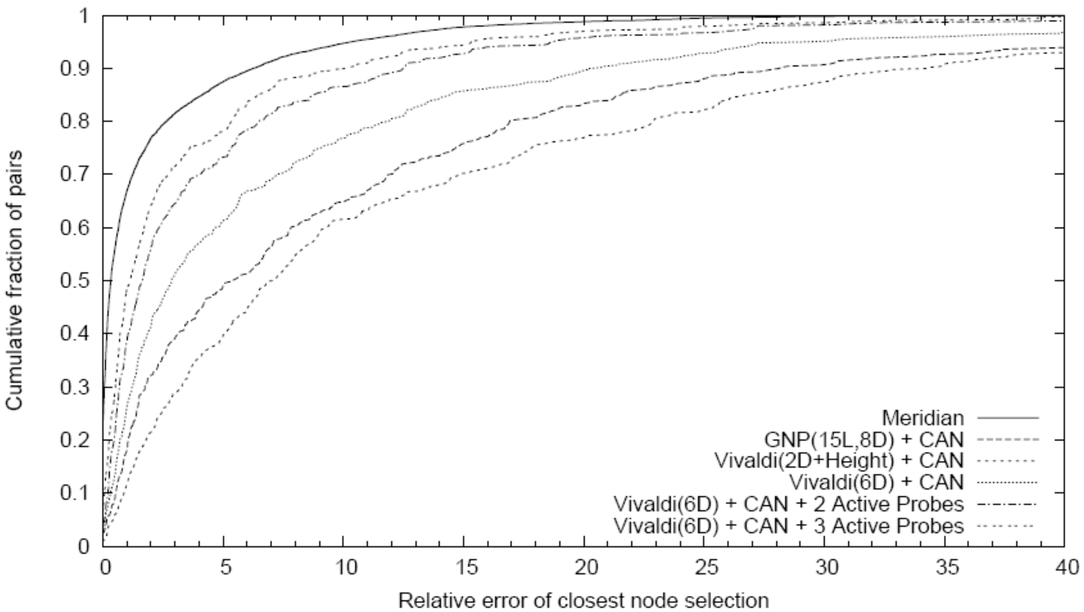


Each node keeps track of small fixed number of neighbors and organizes them in concentric rings, ordered by distance from the node.
k: number of nodes per ring (complexity O(*k*), so k should be manageable.
Nodes use a gossip protocol to maintain pointers to a sufficiently diverse set of nodes in the network.

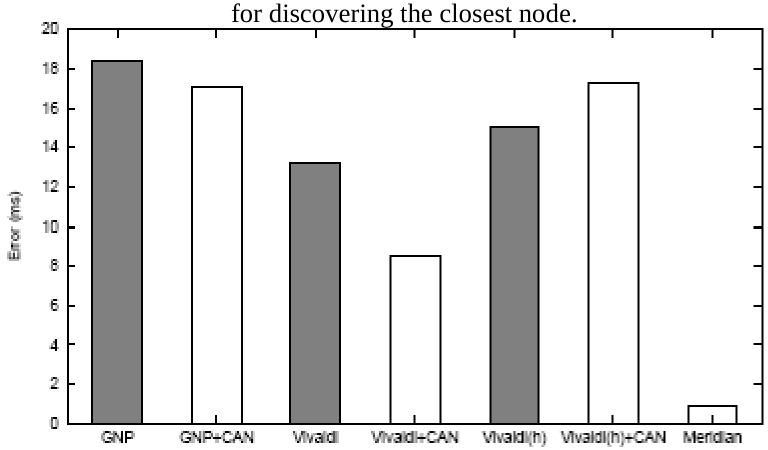


1. Client sends "closest node discovery to target T" request to A.
2. A determines latency, d, to T.
3. A probes ring members to determine latency to T.
4. Request forwarded to closest node and recurses from there.

Data for results: 2000 Meridian nodes, 500 target nodes, *k* = 16 nodes per ring, 9 rings per node.



Results show Meridian has lowest median error for discovering the closest node.

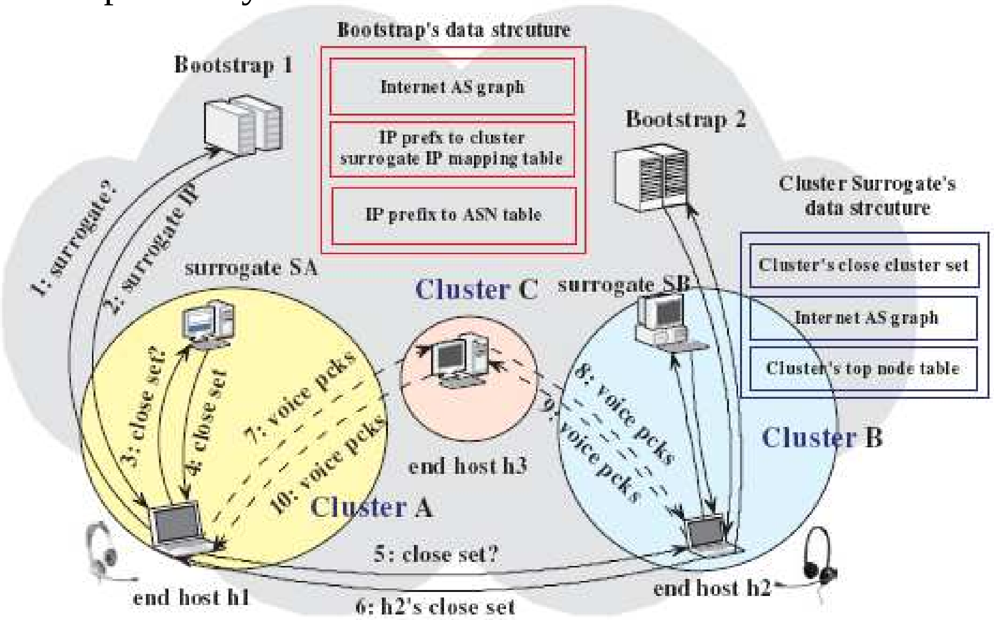


AS-Aware Peer-Relay Protocol (ASAP)

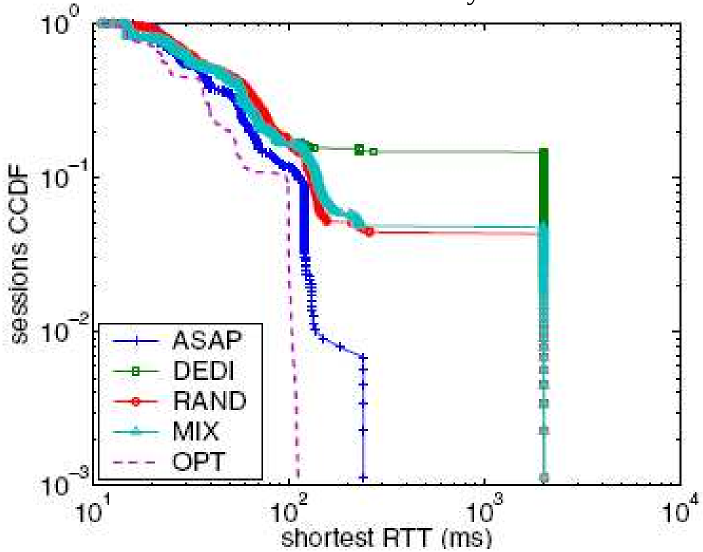
[Ren et al., IEEE ICDCS 2006]

Key principles:

- Bootstrap nodes have an up-to-date AS graph
- End hosts grouped in clusters based on their IPs
- Cluster surrogate nodes perform RTT measurements with clusters in same/close ASes and keep track of close clusters
- Relay negotiation based on cluster proximity and AS distance



Simulation analysis



DEDI: dedicated relays
RAND: random selection
MIX: 25% dedicated, 75% random
OPT: optimal selection