Locality Awareness in Overlay Networks

J. L. Chiang

July 1, 2005

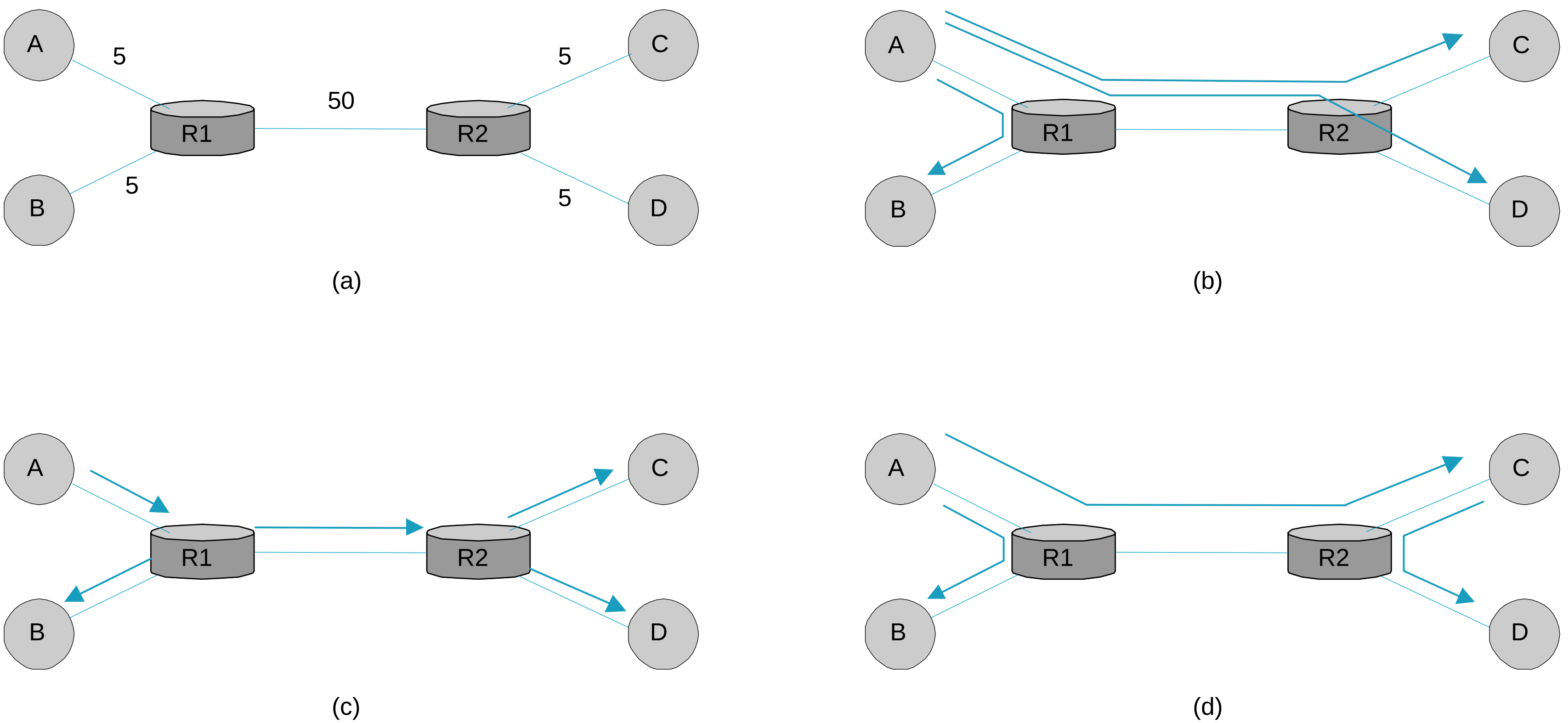
References

1. Y. H. Chu, S. G. Rao, S. Seshan, and H. Zhang, "A Case for End System Multicast," *IEEE JSAC, Vol. 20, No. 8*, 2002.
2. X. Y. Zhang, Q. Zhang, Z. Zhang, G. Song, and W. Zhu, "A Construction of Locality-Aware Overlay Network: mOverlay and Its Performance," *IEEE JSAC, Vol. 22, No. 1*, 2004.
3. S. Ratnasamy, M. Handley, R. Karp, and S. Shenker, "Topologically-Aware Overlay Construction and Server Selection," *IEEE INFOCOM* 2002.
4. Y. Liu, L. Xiao, X. Liu, L. M. Ni, and X. Zhang, "Location Awareness in Unstructured Peer-to-Peer Systems," *IEEE Trans. Parallel and Distributed Systems, Vol. 16, No. 2*, 2005.
5. M. Ripeanu, A. Iamnitchi, and I. Foster, "Mapping the Gnutella Network," *IEEE Internet Computing*, 2002.
6. T. Klingberg and R. Manfredi, Gnutella 0.6, http://rfc-gnutella.sourceforge.net/src/rfc-0_6-draft.html

Introduction

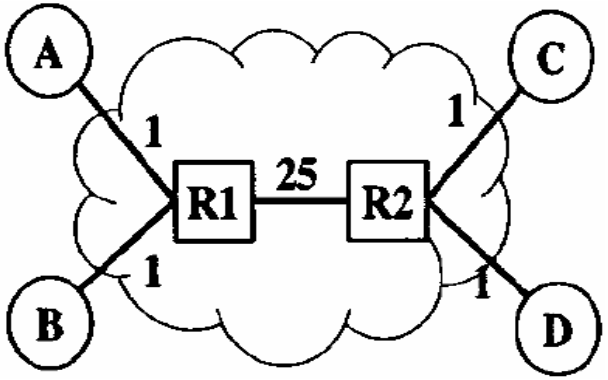
- Overlay
  - Each of the edges in an overlay corresponds to a unicast path between two end systems in the underlying Internet [1].
- Topology Mismatch
  - Nearby hosts in the overlay networks may actually be far away in the underlying network [2].
  - Application level connectivity is not congruent with the underlying IP-level topology. [3]
- End System Multicast

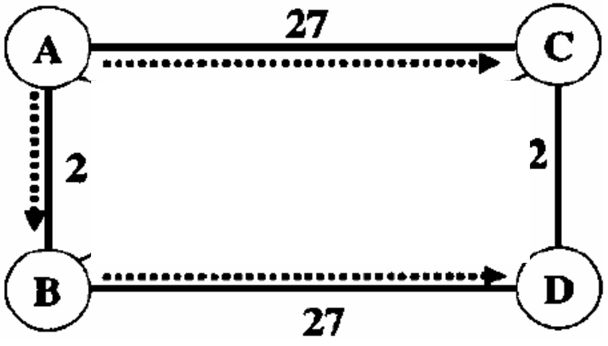
ESM



NARADA [1]

- Physical Network -> Mesh
  - The quality of the path between any pair of members is comparable to the quality of the unicast path between that pair of members
  - Each member has a limited number of neighbors in the mesh
- Mesh -> Spanning trees
  - DVMRP-like





Locality Awareness

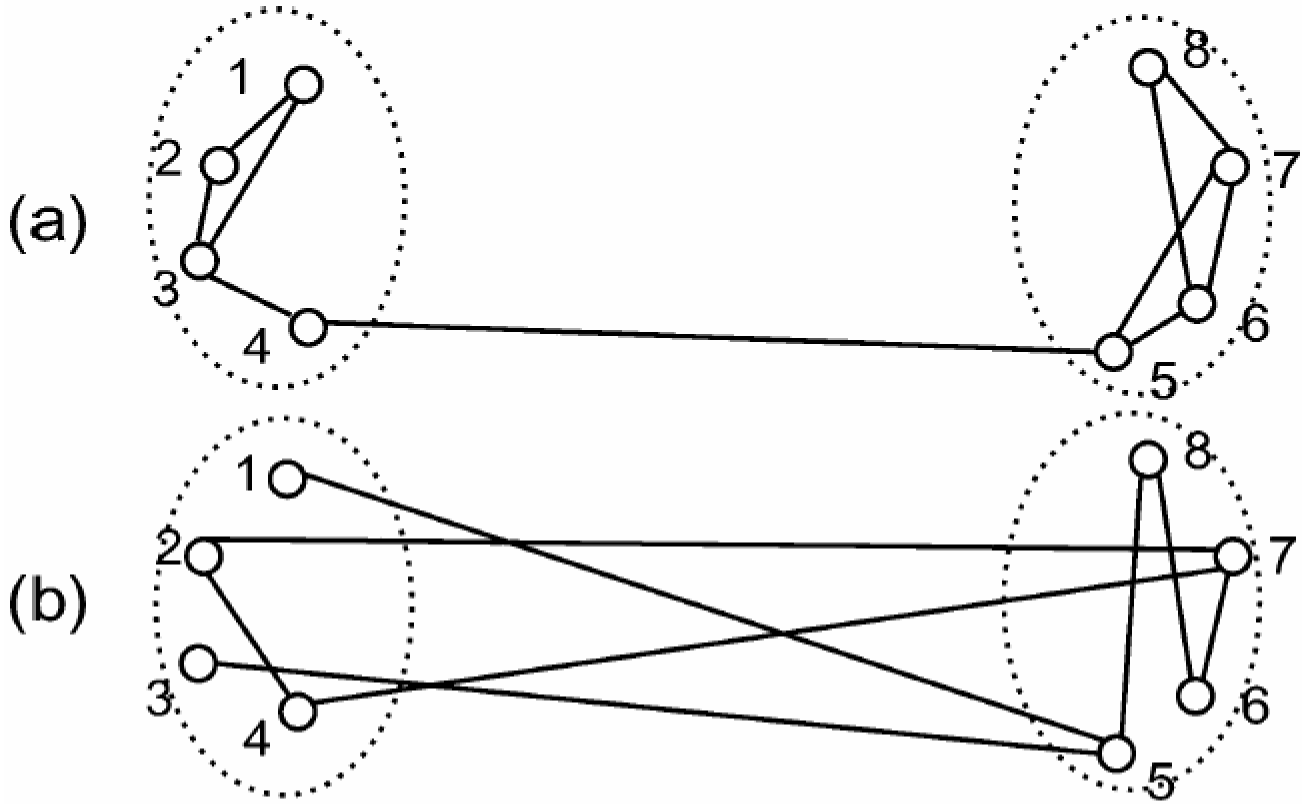
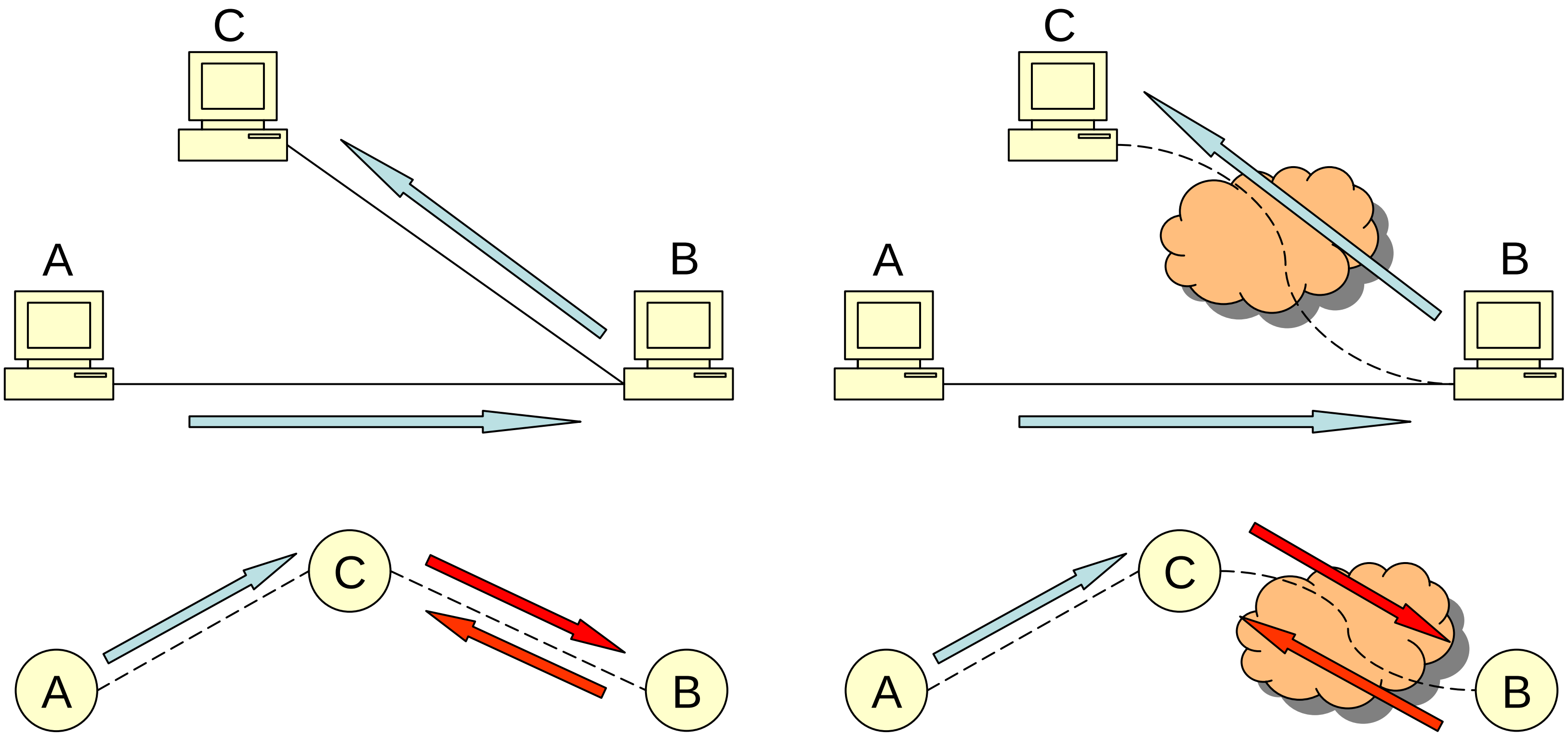


Fig. in [2]

Topology Mismatch

- Topology mismatch and blind flooding makes the unstructured P2P systems far from scalable [2].
- The mismatch problem leads to the same message traversing the same physical link multiple times [4].

Topology Mismatch



Topology Mismatch

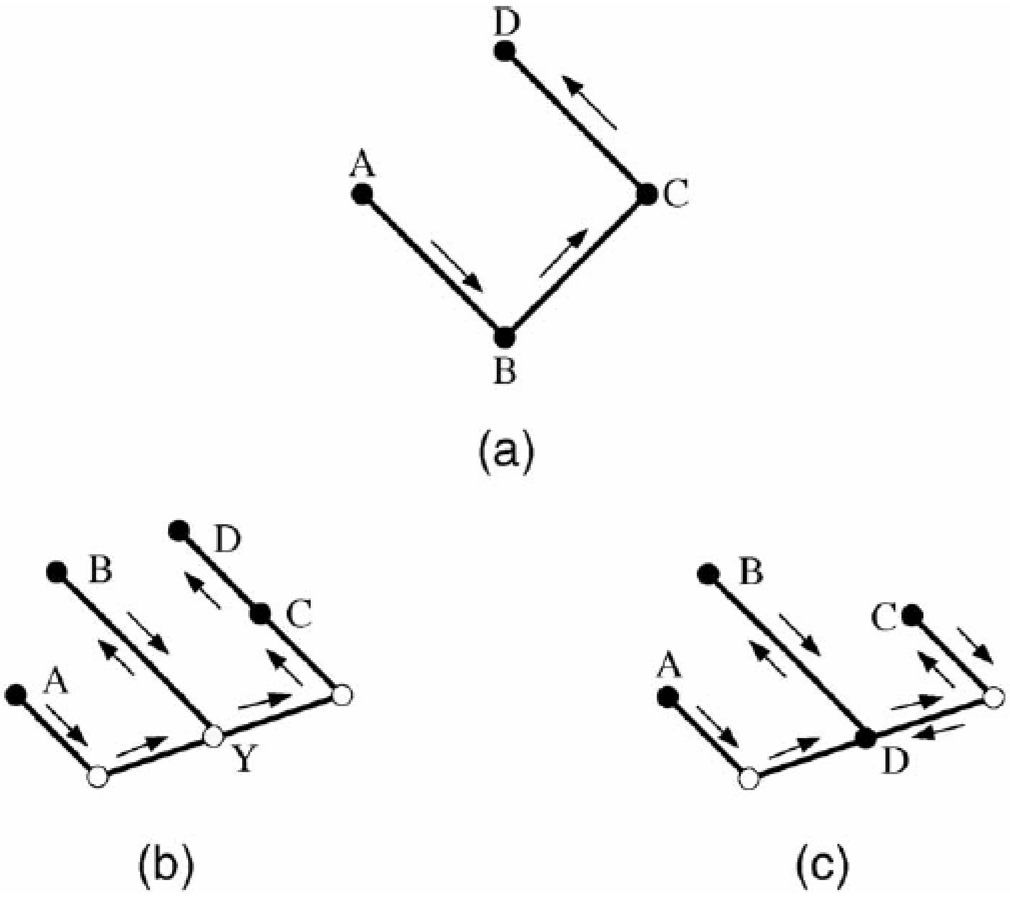


Fig. from [4]

Mapping the Gnutella Network [5]

- Only 2 to 5 percent of Gnutella connections link peers within a single AS.
- But, more than 40 percent of all Gnutella peers are located with the top 10 ASes.

- Most Gnutella-generated traffic crosses AS borders so as to increase topology mismatch cost.

Solutions

- Overlay construction should be aware of locality in the underlying topology.
- IP-address-based
  - Mapping accuracy
  - Searching scope
- Landmark-based [2,3]
  - RTT measurement
- Overlay optimization [4]

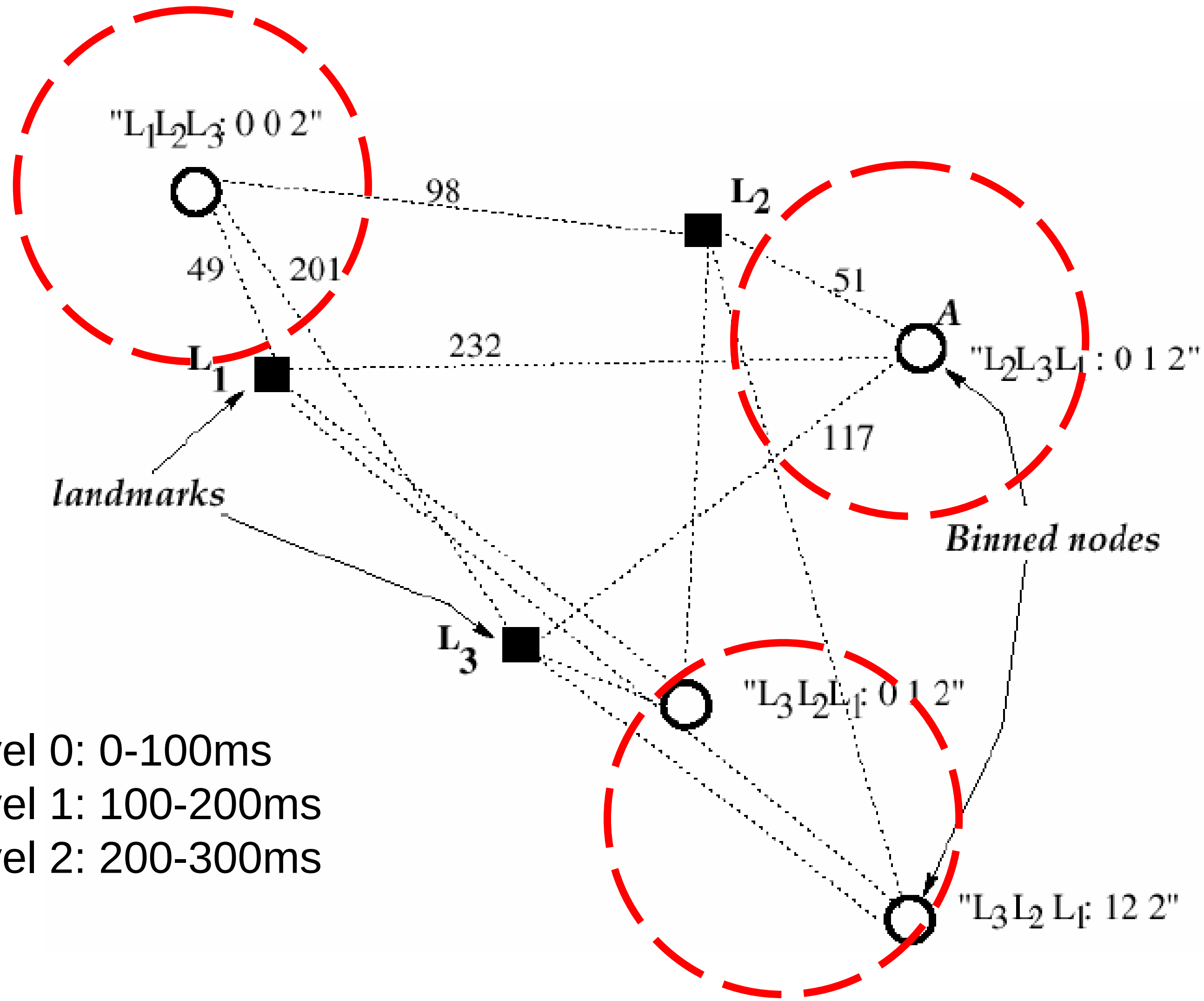
Landmark [3]

- A distributed binning scheme whereby nodes partition themselves into bins such that nodes that fall within a given bin are relatively close to one another in term of network latency.
- Stable landmark machines (unsuspecting participants such as DNS servers)

Distributed Binning

- A node measures its RTT to each of these landmarks and orders the landmarks in order of increasing RTT.
- The ordering of landmarks represents the "bin" the node belongs to.
- The absolute values of RTT can also be used to define a level vector to argument the landmark ordering of a node.

Distributed Binning



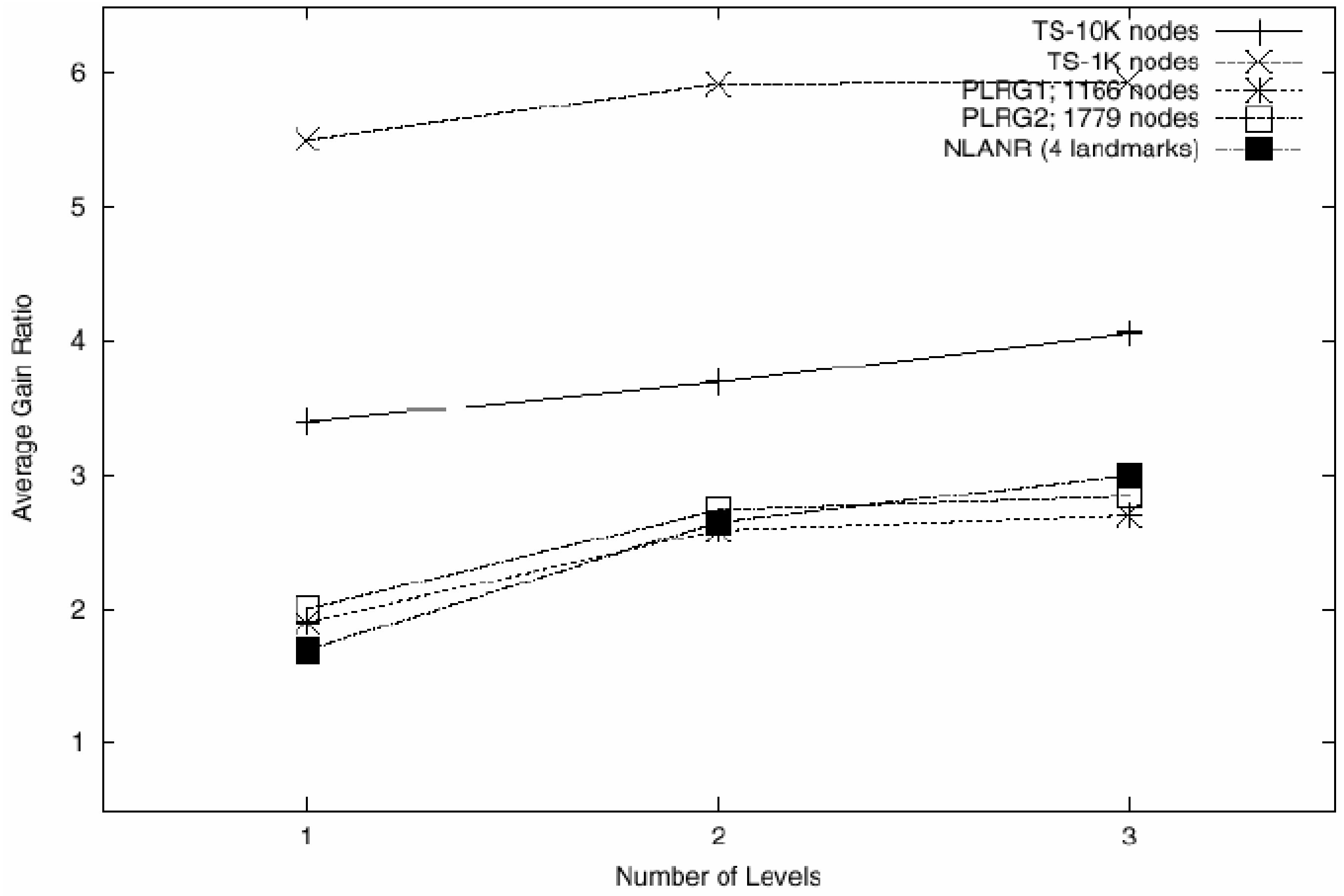
Level 0: 0-100ms
Level 1: 100-200ms
Level 2: 200-300ms

Performance

- The scheme is scalable since nodes need only have knowledge of a small set of landmarks.
  - 1 million nodes/10 pings/refresh per hour -> 2778 pings per second on each landmark [3].
  - 1600 DNS requests per second at f.root-servers.net

- Gain ratio = inter-bin latency / intra-bin latency

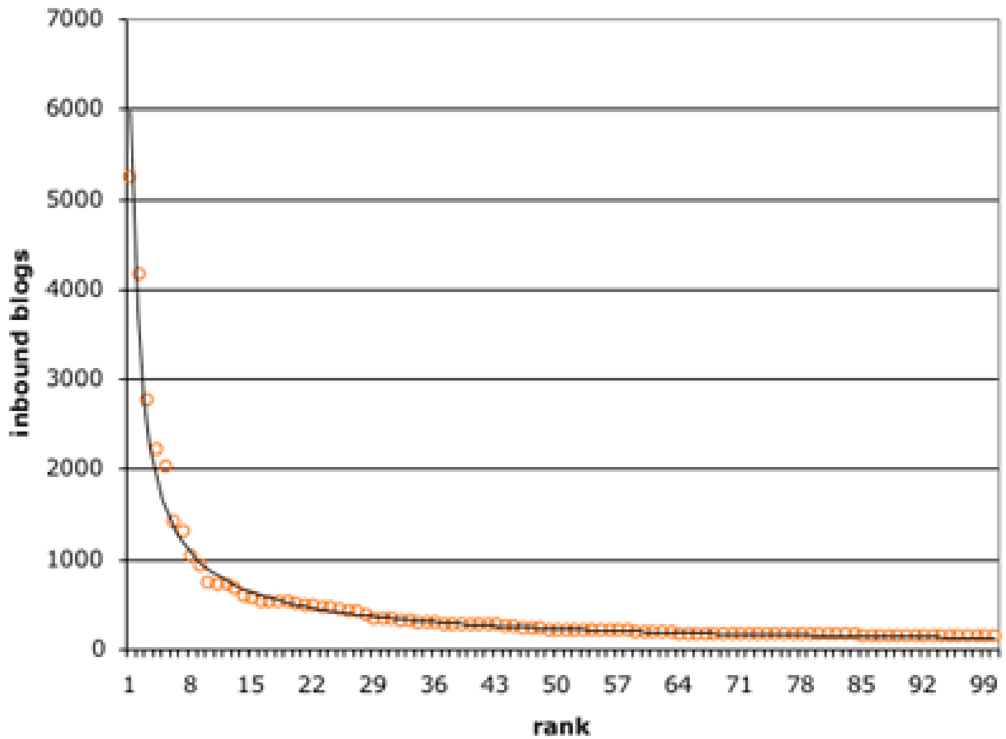
Gain Ratio



TS-10K nodes
TS-1K nodes
PLRG1; 1166 nodes
PLRG2; 1779 nodes
NLANR (4 landmarks)
Average Gain Ratio
Number of Levels
1
2
3
4
5
6

Power Law Distribution

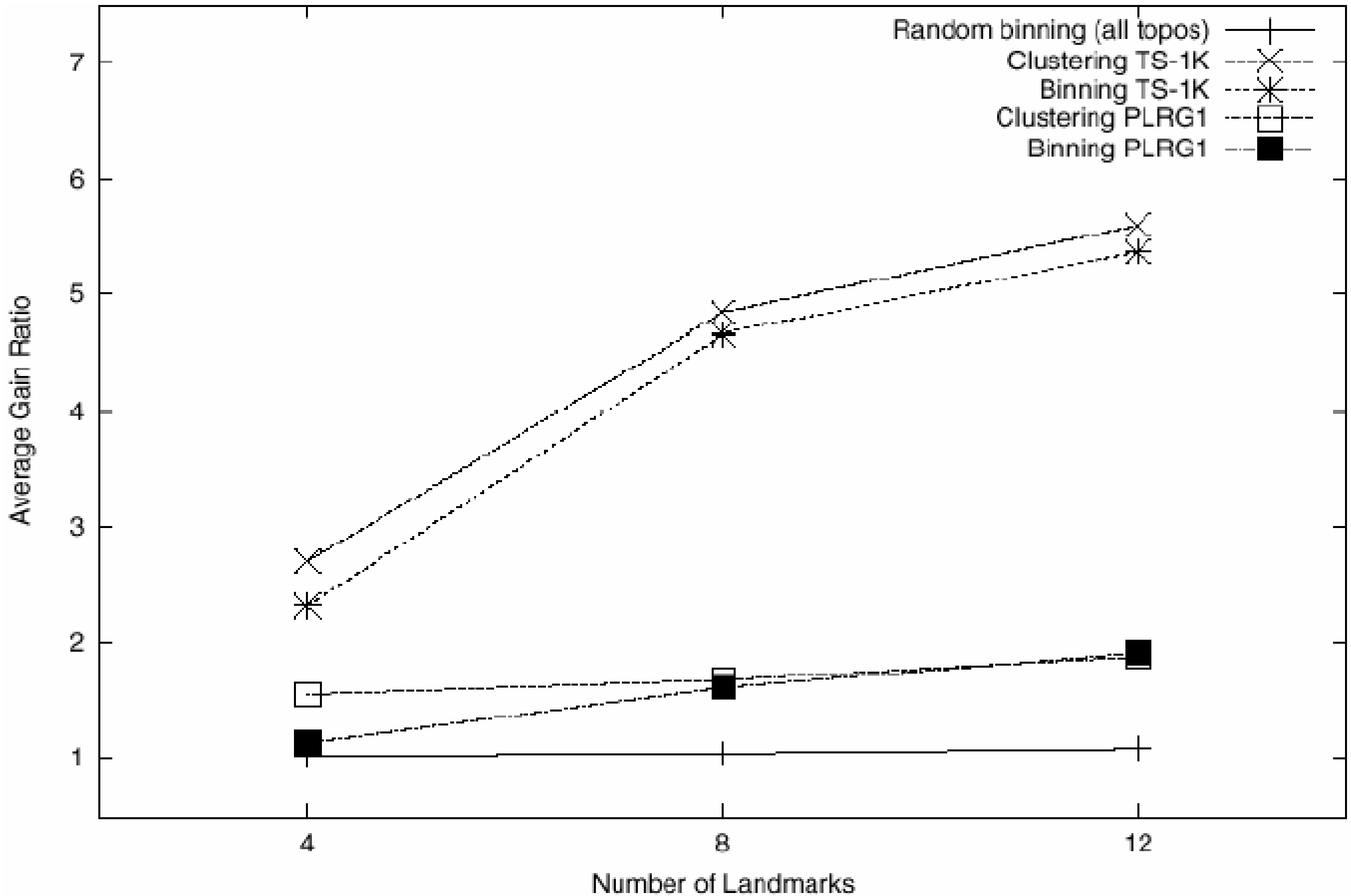
- 80/20 rule: 80% of the wealth is controlled by 20% of the population (large is rare and small is common).



Performance Evaluation

- Random binning (lower bound)
  - Using the same number of bins as generated by landmark-based binning scheme, each node selects a bin at random.
- Nearest-neighbor clustering (upper bound)
  - Each node is initially assigned to a cluster by itself.
  - At each iteration, the two closest clusters are merged into a single cluster until the required number of clusters are obtained.

Performance Evaluation



Overlay Construction

- Given a set of *n* nodes on the Internet, have each node picks any *k* neighbor nodes from this set, so that the average routing latency on the resultant overlay is low (assuming shortest path routing).

- NP-hard

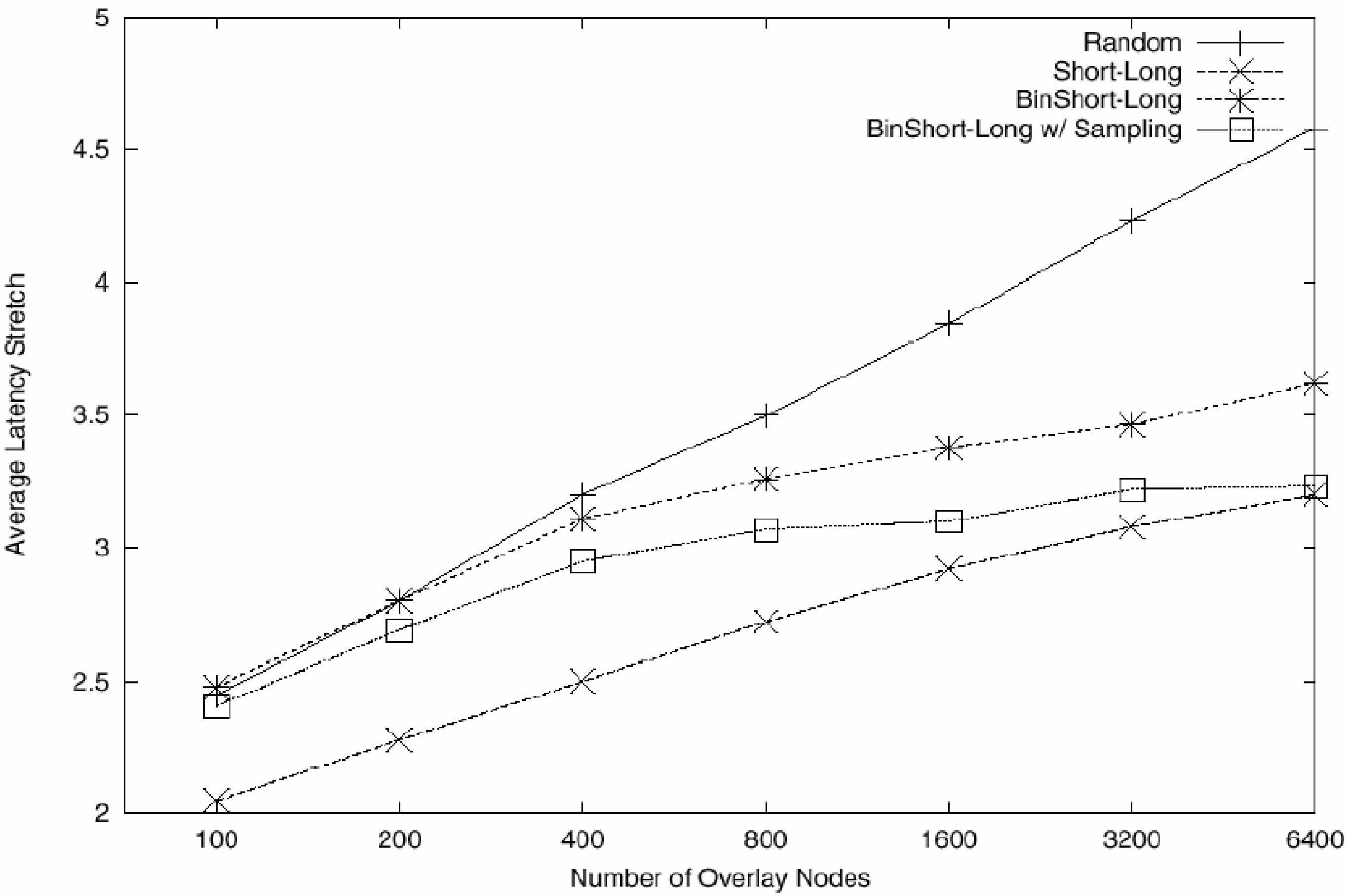
Heuristic Algorithm

- Short-Long
  - A node picks its k neighbors by picking the k/2 nodes in the system closest to itself and then picks another k/2 nodes at random.
  - k/2 closeby nodes for well-connected pockets of nearby nodes; k/2 random links for keeping graph connected and interconnecting different pockets of nodes
- Requirement of global knowledge of all other nodes

Performance Evaluation

- BinShort-Long
  - Use binning for picking nearby k/2 nodes
- BinShort-Long w/ sampling
  - Additionally sample RTT of bin nodes
- Average latency stretch
  - The ratio of the path latency using shortest path routing on the overlay to the path latency on the underlying network topology.

Performance Evaluation



TS-10K; # of levels = 1; # of landmarks = 12

Potential Issues

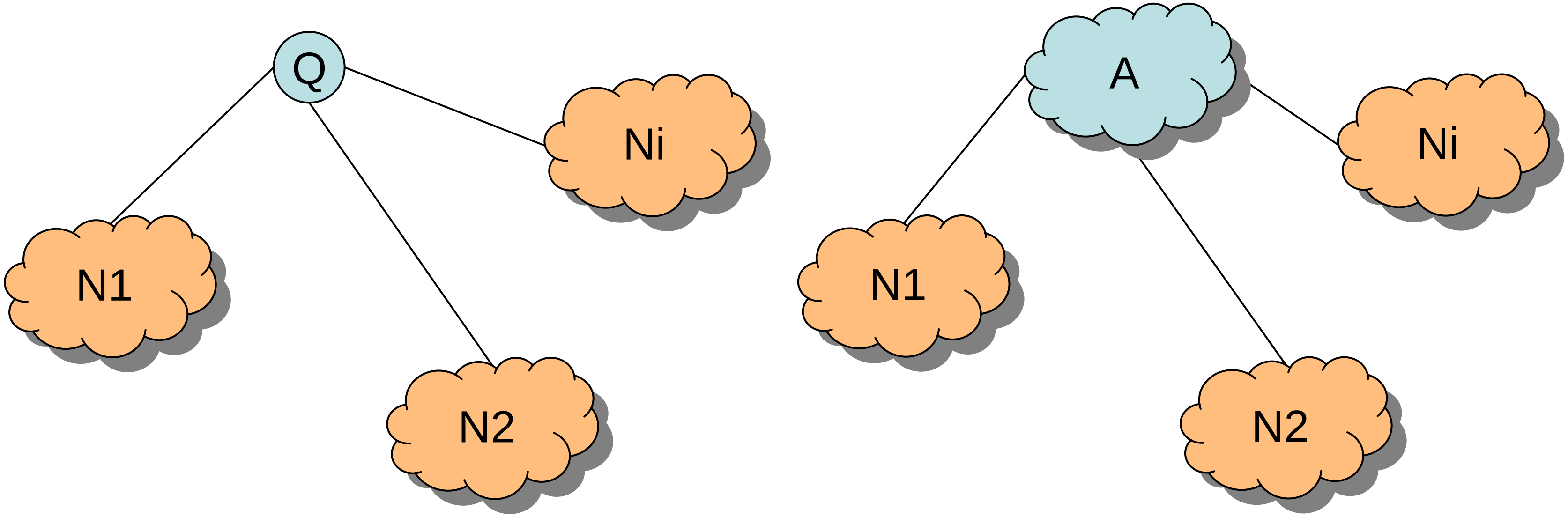
- On the construction of unstructured overlays,
  - It needs extra deployment of landmarks and produces some hotspots in the underlying network when the overlay is heterogeneous and large [2].
  - Nodes require the knowledge of other nodes in the same bin either through the landmark system or a bin leader for selecting nearby nodes.
    - Membership maintenance and message overhead
  - It require a match scheme on landmark orderings for the degree of similarity between two neighboring bins.
    - Maintenance of neighboring bins through the landmark system and keeping the whole network connected.

mOverlay [2]

- A group consists of a set of hosts that are close to each other.
- A desirable locality-aware overlay structure is that most links are between hosts within a group and only one or two links between two groups.
- A group is a self-organizing cluster of hosts with a group leader.
- The neighboring groups of a group act as the dynamic landmarks used in the grouping criterion.

Grouping Criterion

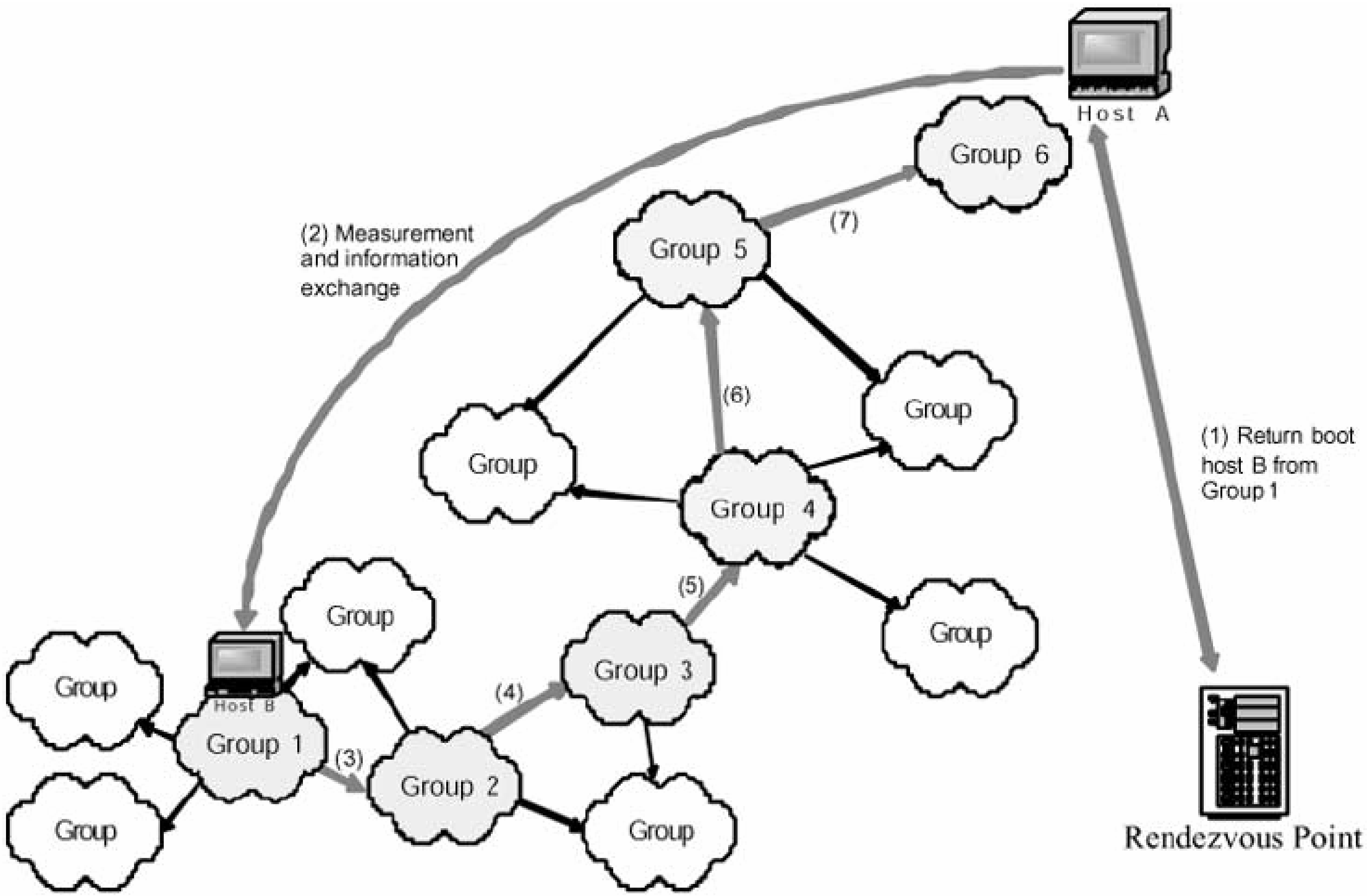
- When the distance between a new host Q and group A's neighboring groups is the same as the distance between group A and group A's neighboring groups, the host Q should belong to group A.



Components in mOverlay

- Locating process
  - Rendezvous Point (RP)
  - Boot hosts
  - Candidate group list (*M* neighboring groups)
  - Current closest group with *Dmin*
- Maintenance protocol
  - Local host cache (*H* group hosts)
  - Group leader

Locating Process



Maintenance

- Forming new groups
  - In the initialization stage
  - When the nearest group doesn’t meet the grouping criterion
- Information update by the group leader
  - Updating the host cache when a new host joins
  - Updating the group list when a nearby group is generated
- Information sharing by flooding in a group
  - e.g. distances to neighboring groups

Performance Analysis

- The complexity (distance) of locating the nearby group is of $O(\log N)$ [2].
- The local host cache provides robustness.
  - A special neighboring group is randomly selected to decrease the probability of disconnected graph.
- Scalability is achieved due to load balancing through random selection of boot hosts in RP.

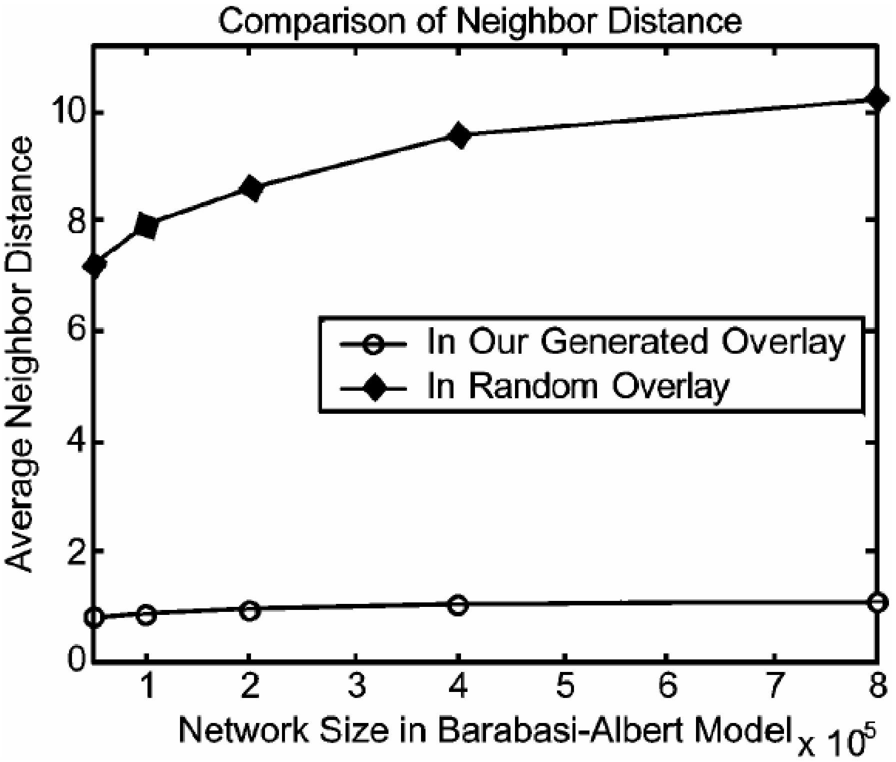
Performance Analysis

- Average neighbor distance
  - in locality-aware overlay
    - $$\bar{D} = \frac{D_i \frac{N \cdot n \cdot m}{2} + D_b \frac{N \cdot M}{2}}{\frac{N \cdot n \cdot m}{2} + \frac{N \cdot M}{2}} = \frac{D_i \cdot n \cdot m + D_b \cdot M}{n \cdot m + M}$$
  - In random connected overlay
    - $$\bar{D}' = \frac{D_i' n \frac{m'}{m''} + D_b' n}{n \frac{m'}{m''} + n} = \frac{D_i'(n-1) + D_b' n(N-1)}{nN-1} = \left[\frac{D_i'(n-1)}{D_b'(nN-1)} + \frac{n(N-1)}{nN-1}\right] \cdot D_b' \approx D_b'$$

N: # of groups
M: # of neighbor groups
m: # of neighbor hosts
Db: avg. distance between neighbor groups
Di: avg. distance between hosts in the same group
m': # of neighbor hosts in the same group
m": # of neighbor hosts in all other groups
D'b: avg. distance of intergroup links
D'i: avg. distance of intragroup links

Performance Evaluation

- The Barabasi-Albert model shows Power law distribution.
- *D'b* is fixed and determined by the underlyingnetwork.
- *Db* and *Di* depend on the overlay construction and are obtained through the simulation.
- The number of nodes and links are the same in both overlay.



Potential Issues

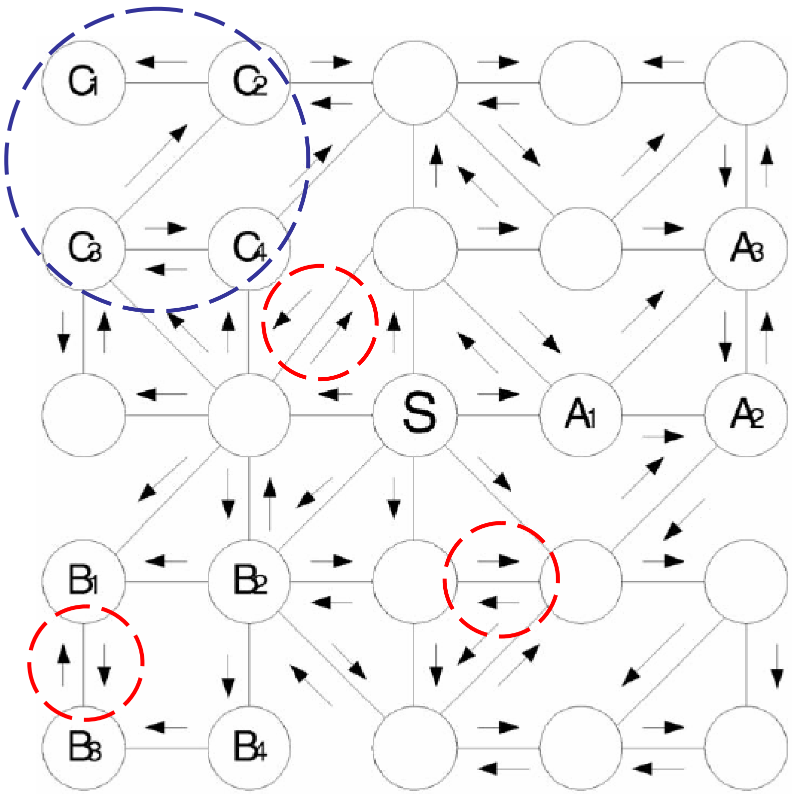
- Overhead on maintenances of local hosts and neighboring groups
- Overhead on messages between hosts and groups due to information updates

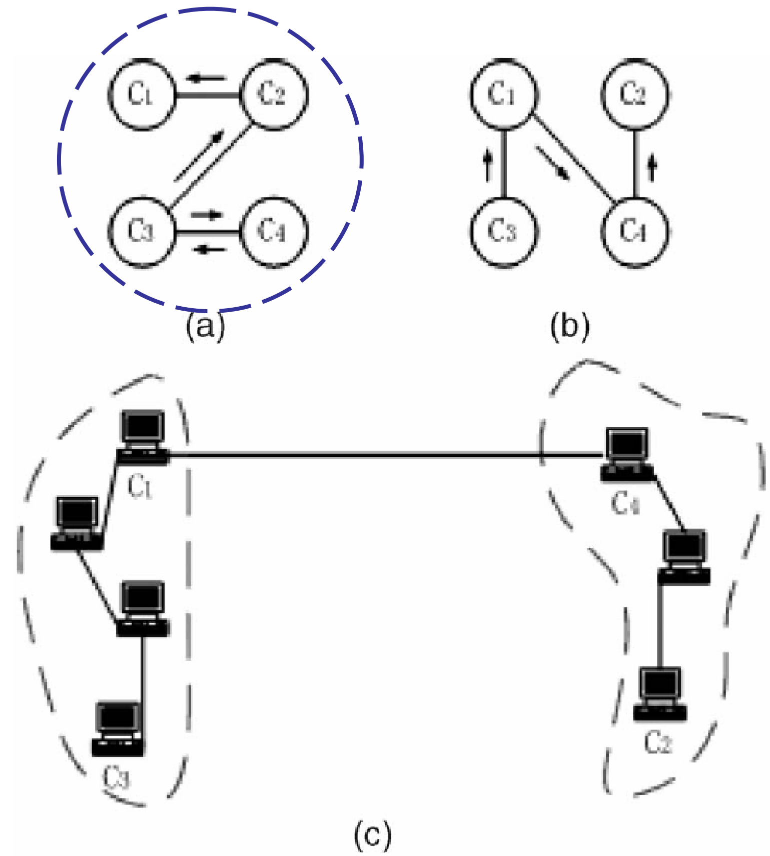
- Dynamic landmarks but a rendezvous point

Location-Aware Topology Matching

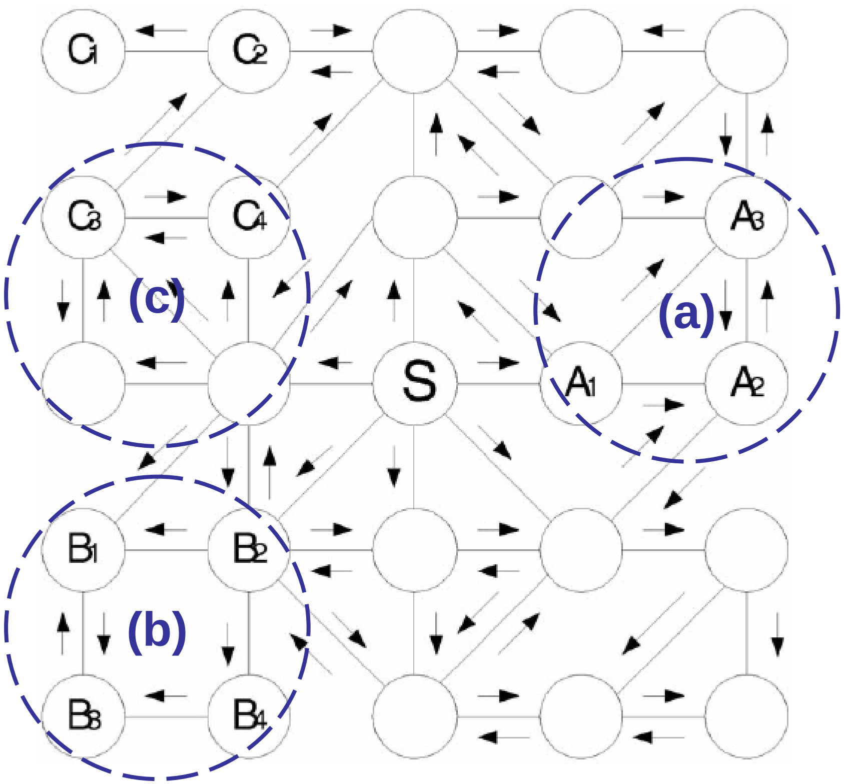
- LTM [4] builds an efficient overlay by disconnecting slow connections and choosing physically closer nodes as logical neighbors while still retaining the search scope and reducing response time for queries.

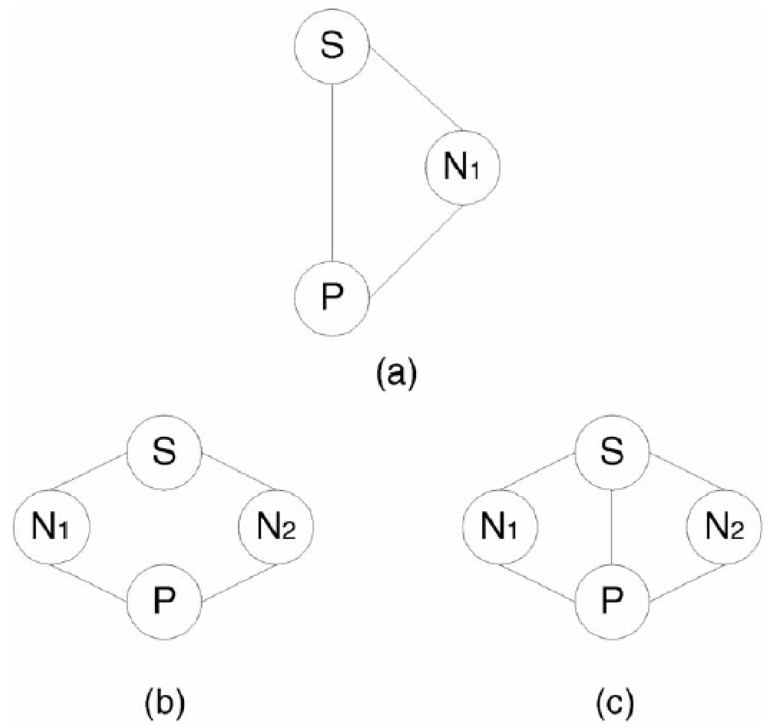
Observations





Observations





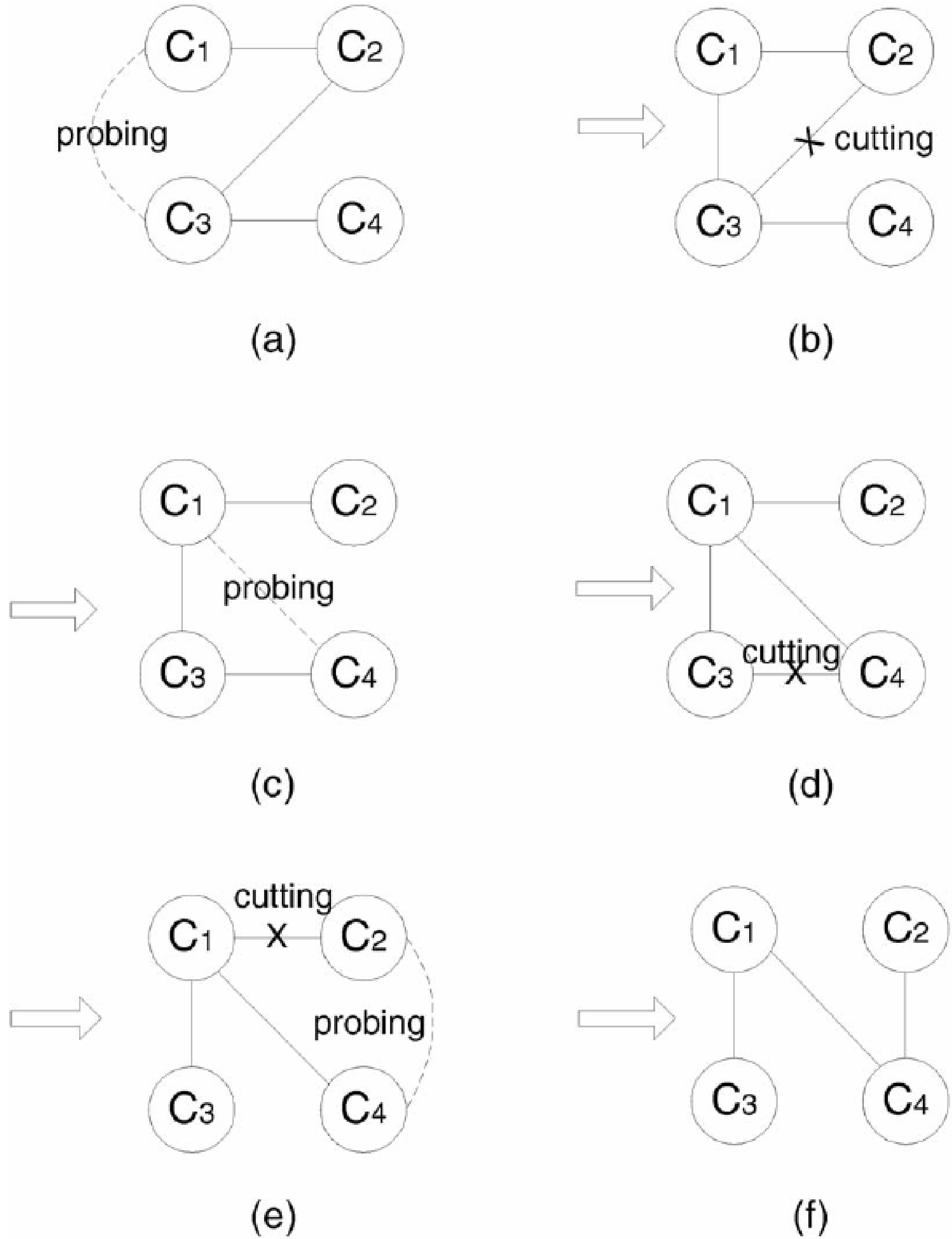
LTM Operations

- TTL2-Detector Flooding
  -

Source IP Address	Source Timestamp	TTL1 IP Address	TTL1 Timestamp
0 … 3	4 … 7	8 … 11	12 … 15

- Slow connection cutting
  - Will-cut list
  - Cut list
- Source peer probing

- An example of LTM

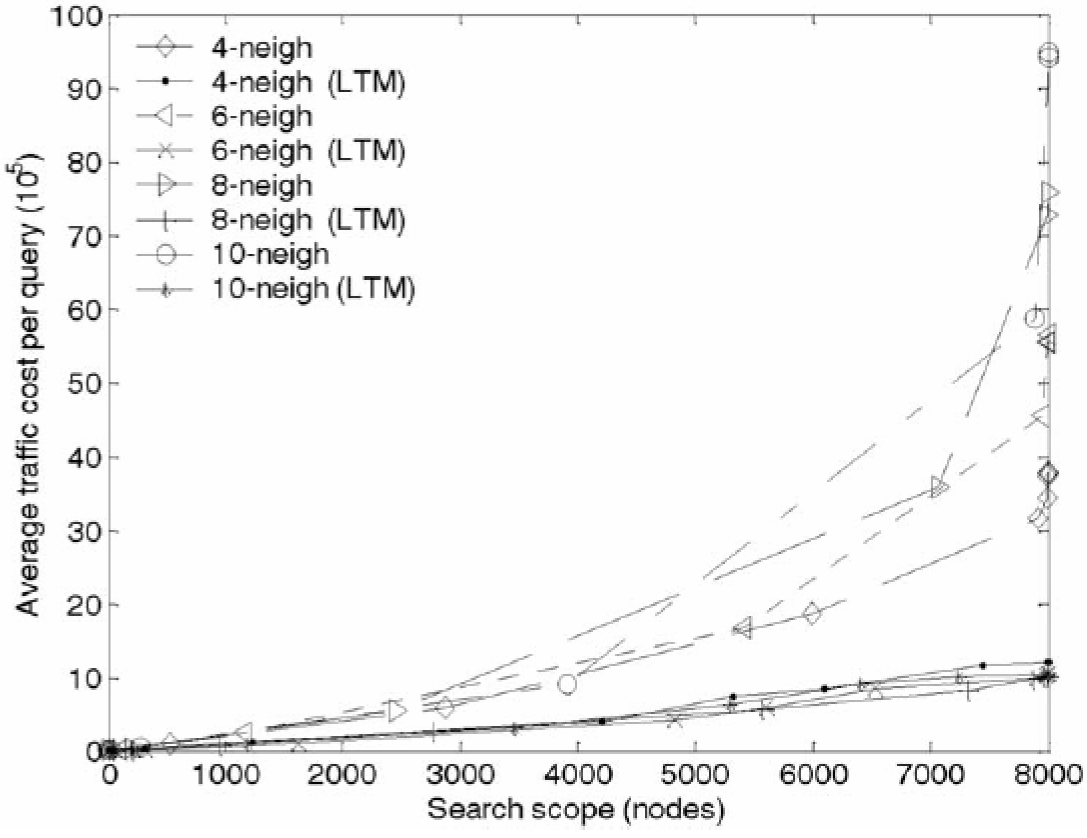


Performance Evaluation

- The per minute traffic overhead incurred by LTM (TTL2 Detector) is $O(n)$, n is the number of peers in the overlay [4].

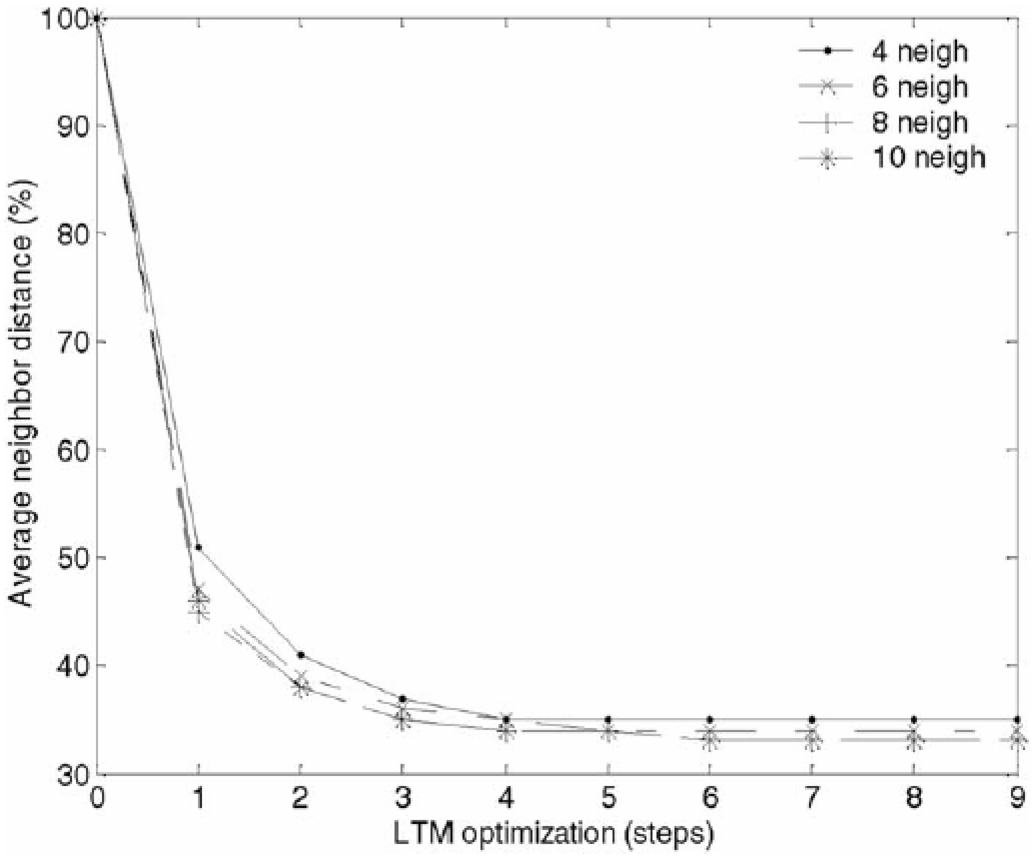
- The of 8,000 nodes on top of underlying topology of 22,000 nodes created by BRITE.

Traffic Cost v.s. Search Scope



One-step LTM optimization

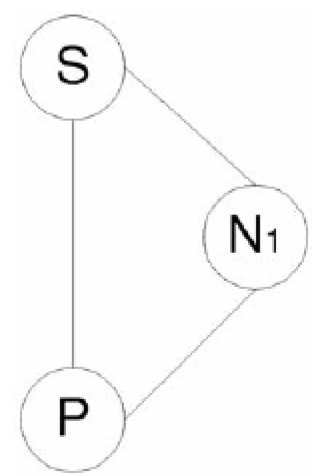
Avg. Neighbor Distance



Potential Issues

- Independent operations on each peer may lead to graph disconnection.
  - Forward and backward latency may vary on the overlay link, which may consists of two different path in the underlying network.

- P keeps N1P and discards SP
- N1 keeps PN1 and discards SN1
- S is disconnected from P and N1



Potential Issues

- LTM doesn't include construction of overlay networks but only performs optimization on established overlay topologies.
  - The performance of resulted overlay is limited by the TTL2 Detector and is mainly decided by the given topology.
  - There is a tradeoff on *k* of the TTL-*k* Detector in terms of the level of optimization and control overhead.

Conclusions

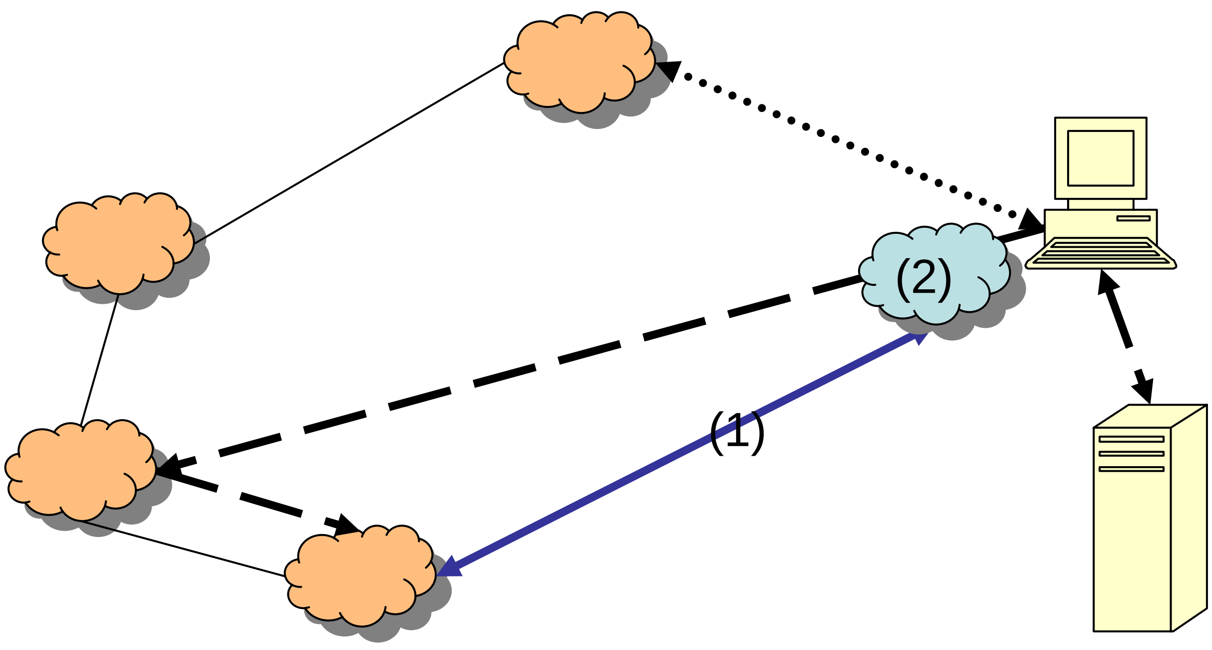
- Locality awareness greatly improves maintenance and searching performance in overlay networks.
- Clustering [2, 3] is practical for reducing messages and shortening searching latency in modern P2P systems.
- The localized distributed scheme [4] avoids well-known entry points but may result in convergence problems.

Discussions

- Dynamics of peers
  – Landmark-based schemes provide short-timescale locating process for use of long-term network services regardless of dynamic peer joining and leaving.
  – Topology optimization schemes require gradually operations on changes of membership.

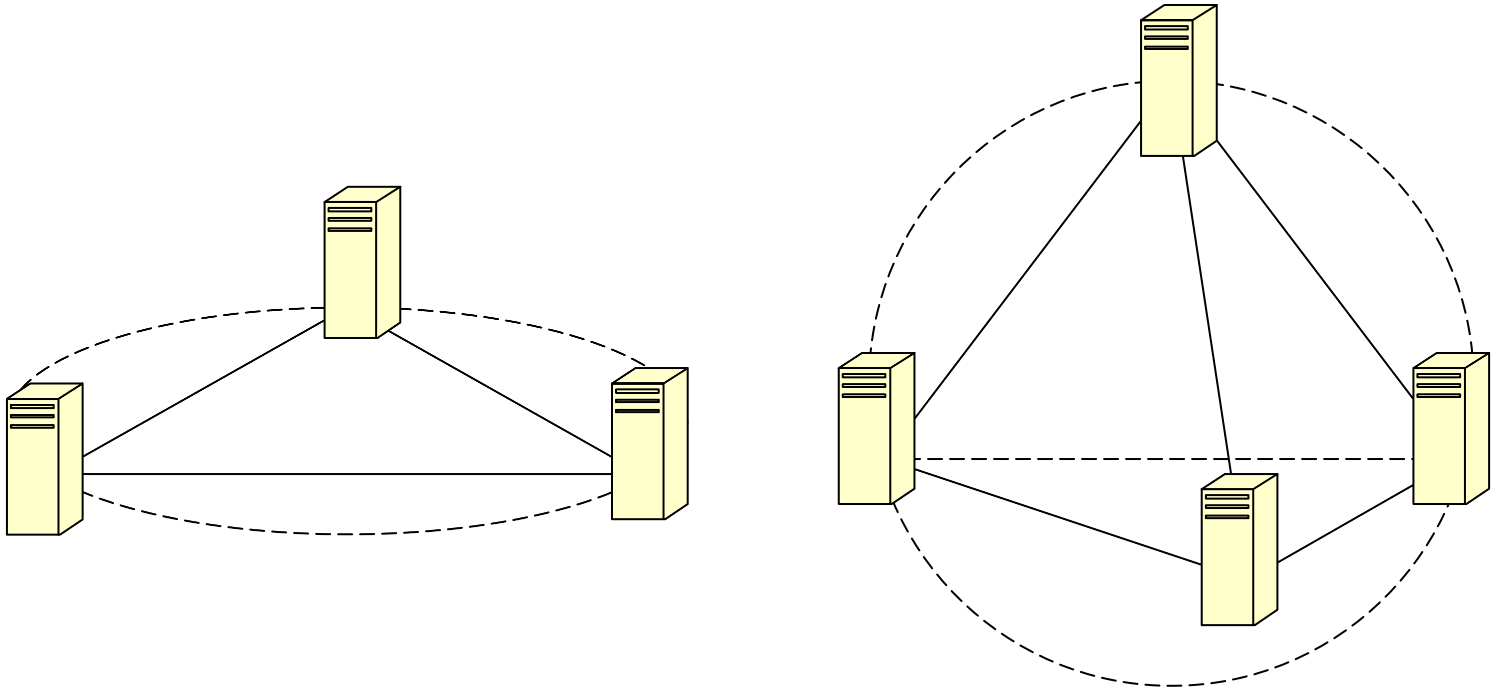
Discussions

- The avg. distance between landmarks and largest distance between a peer and the landmark should be of the same order.

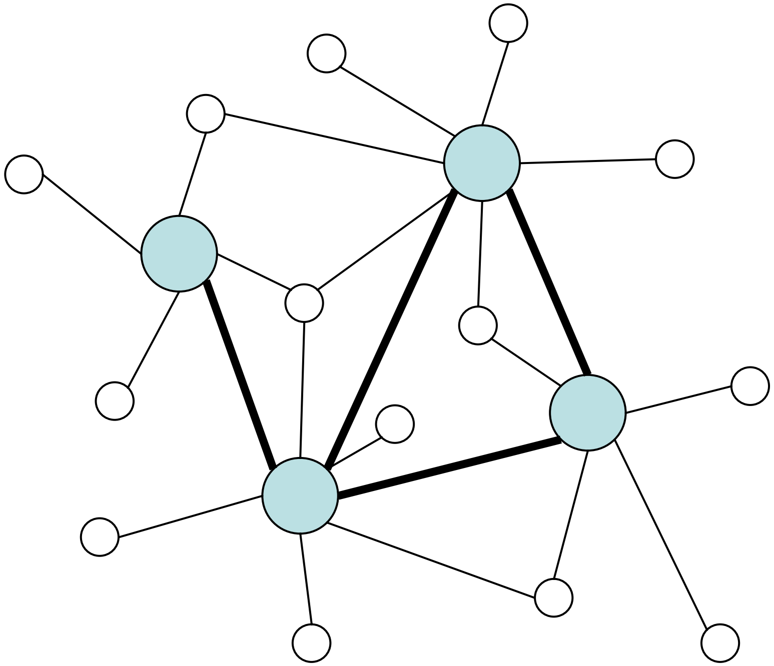


Discussions

- Landmarks should be evenly spread in the coverage of the overlay network.



Superpeers



A superpeer connects to 10-100 peers and 1-10 other superpeer(s).
A peer connects to 3-10 superpeers.

Hostcache (10-20 Superpeers)