Kademlia: A Peer-to-peer Information System Based on the XOR Metric

Petar Maymounkov and David Mazières
New York University

(Presented In 1th International Workshop on P2P Systems 2002)

Amir H. Payberah (amir@sics.se)
Seif Haridi (haridi@kth.se)

BitTorrent™

Core Idea

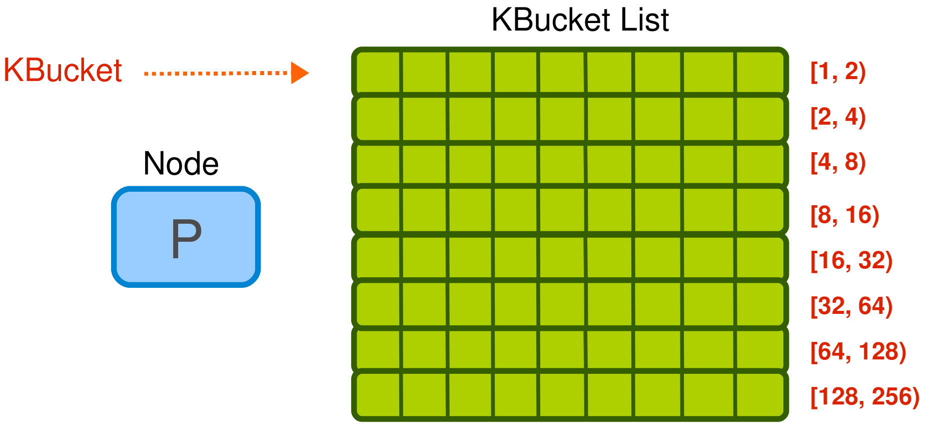
Definition

- Each object is stored at the k closest nodes to the object's ID.
- Distance between id1 and id2: $d(id1, id2) = id1 \text{ XOR } id2$
  - If ID space is 3 bits:

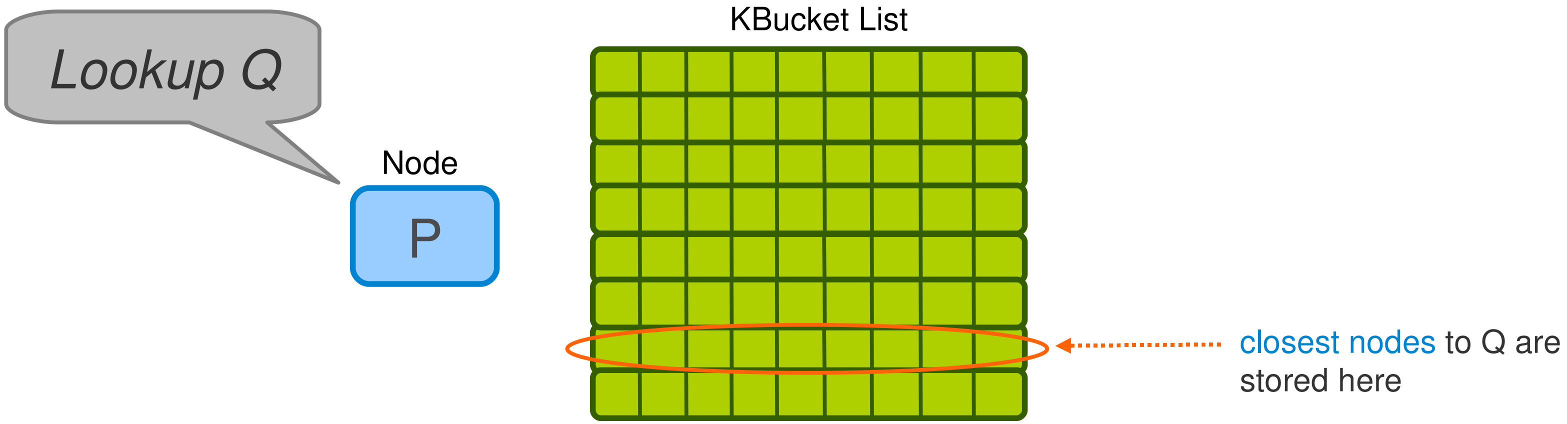
$$\begin{aligned} d(1, 4) &= d(001_2, 100_2) \\ &= 001_2 \text{ XOR } 100_2 \\ &= 101_2 \\ &= 5 \end{aligned}$$

Core Idea - 1

- Kbucket: each node keeps a list of information for nodes of distance between 2^i and 2^{i+1}.

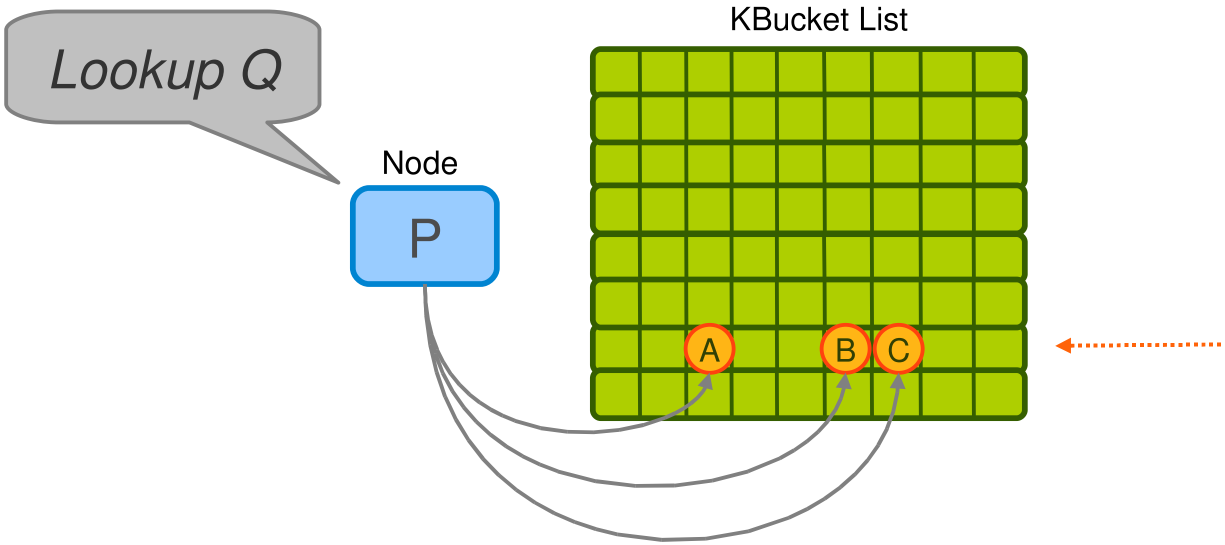


Core Idea - 2



- Closest nodes in ID space

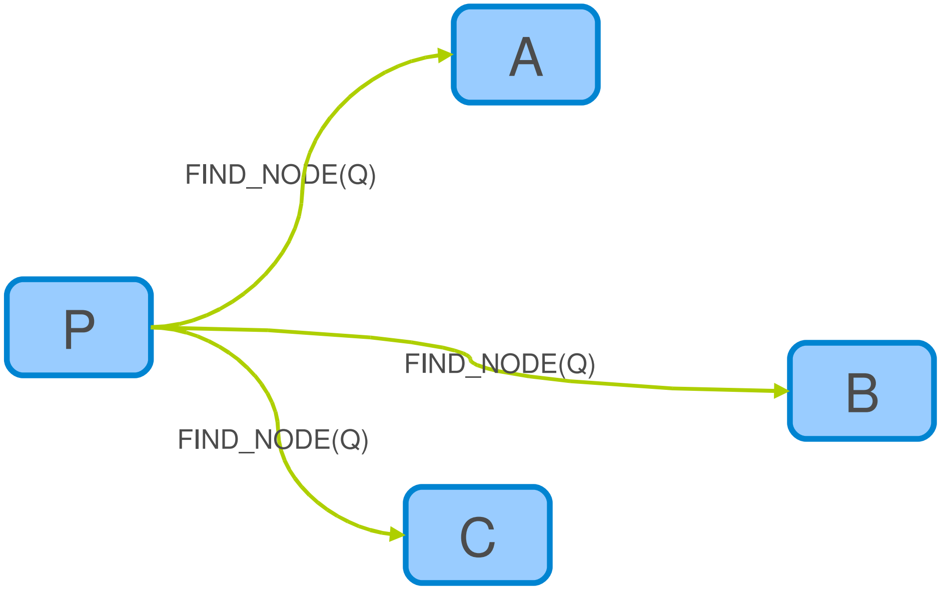
Core Idea - 3



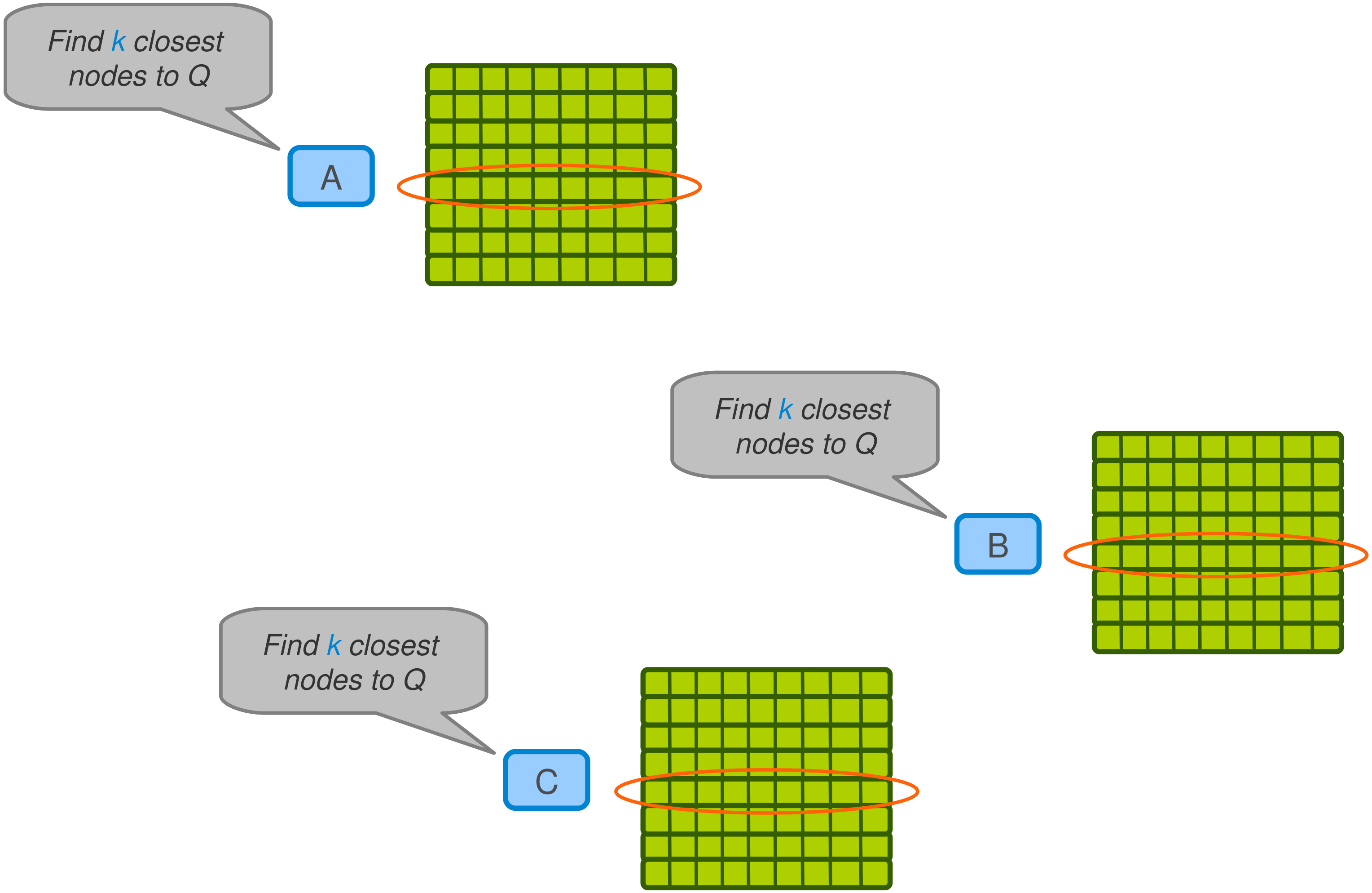
closest nodes to Q are stored here

... and select α nodes from the appropriate kbucket

Core Idea - 4



Core Idea - 5



Find *k* closest nodes to Q

A

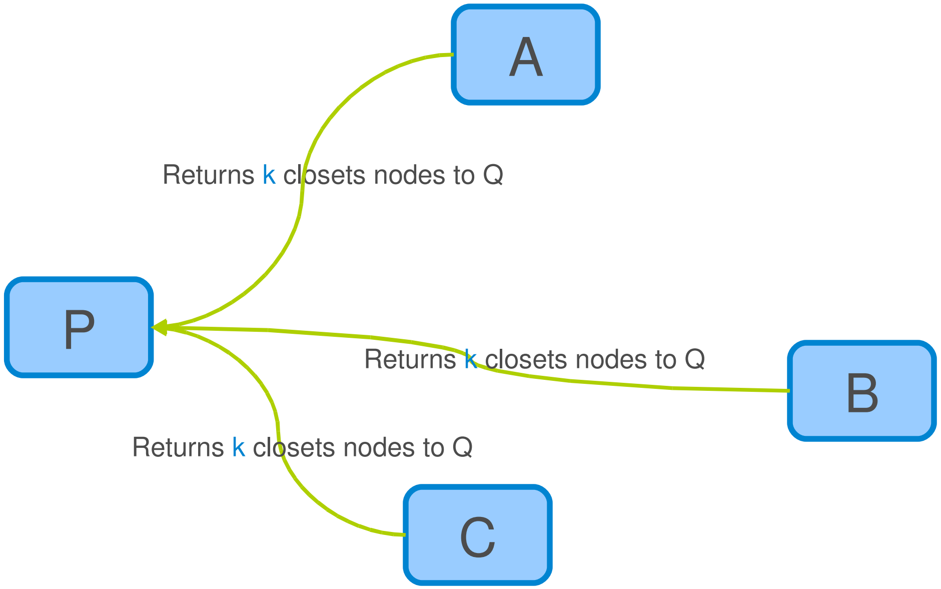
Find *k* closest nodes to Q

B

Find *k* closest nodes to Q

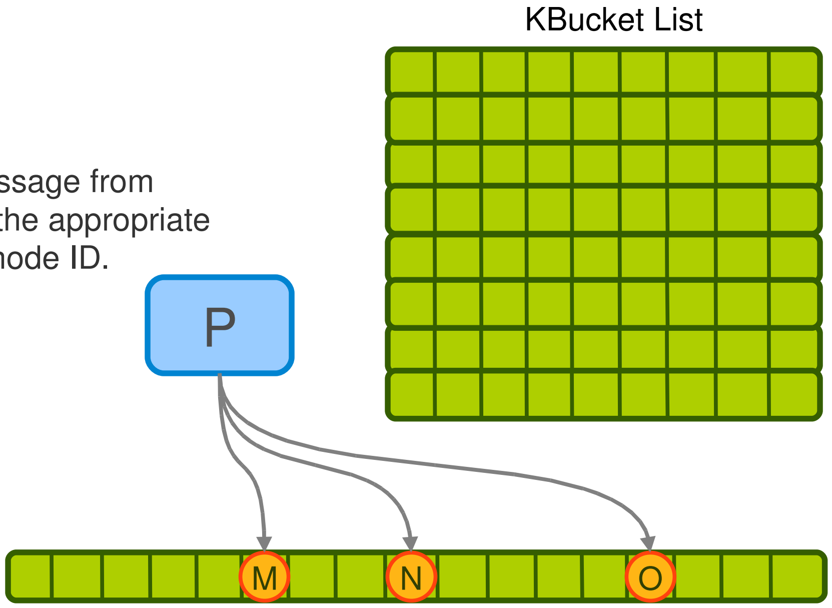
C

Core Idea - 6


A
Returns k closets nodes to Q
P
Returns k closets nodes to Q
B
Returns k closets nodes to Q
C

Core Idea - 7

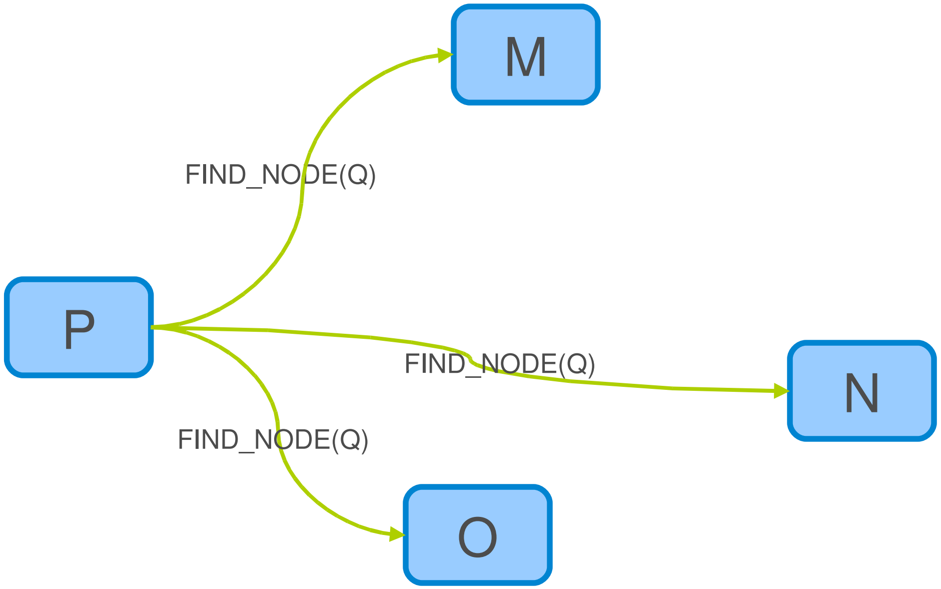
When P receives any message from another node, it updates the appropriate kbucket for the sender's node ID.



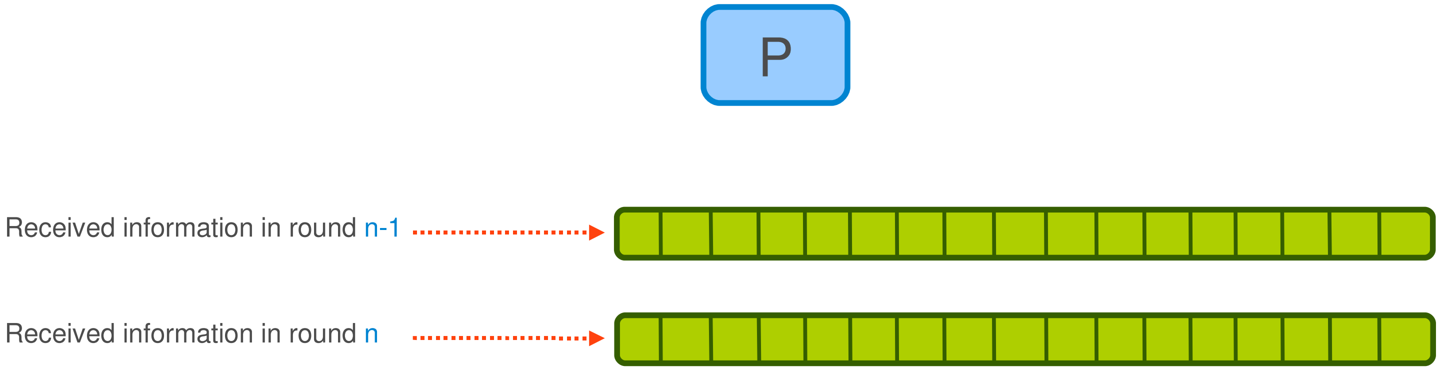
Received information from A, B and C

... again select α nodes from the received information

Core Idea - 8



Core Idea - 9



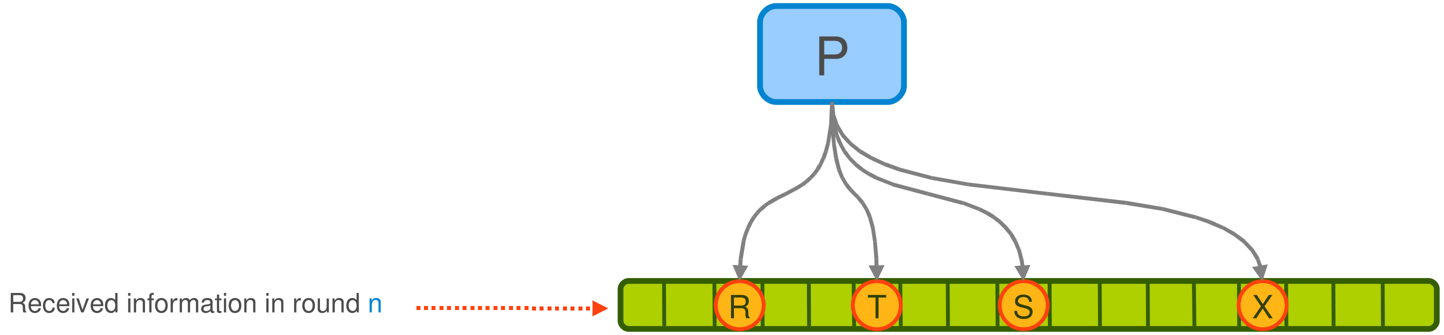
Received information in round n-1

Received information in round n

Repeats this procedure iteratively until received information in round n-1 and n are the same.

Core Idea - 10

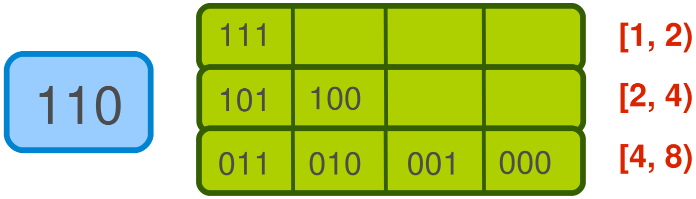
P resends the FIND_NODE to k closest nodes it has not already queried ...



Let's Look Inside of Kademlia

Node State

- Kbucket: each node keeps a list of information for nodes of distance between 2^i and 2^{i+1}.
  - 0 <= i < 160
  - Sorted by time last seen.

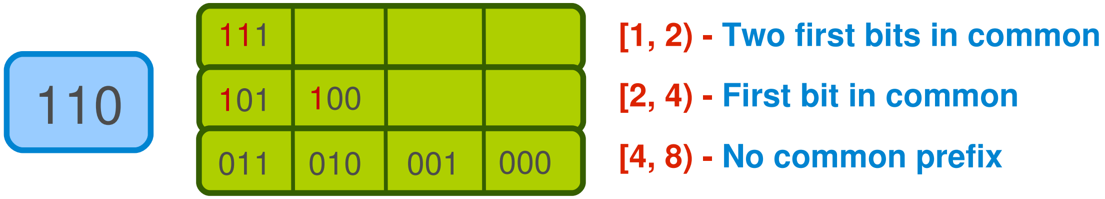


110

111				[1, 2)
101	100			[2, 4)
011	010	001	000	[4, 8)

Node State

- Kbucket: each node keeps a list of information for nodes of distance between 2^i and 2^{i+1}.
  - 0 <= i < 160
  - Sorted by time last seen.



110

111				[1, 2) - Two first bits in common
101	100			[2, 4) - First bit in common
011	010	001	000	[4, 8) - No common prefix

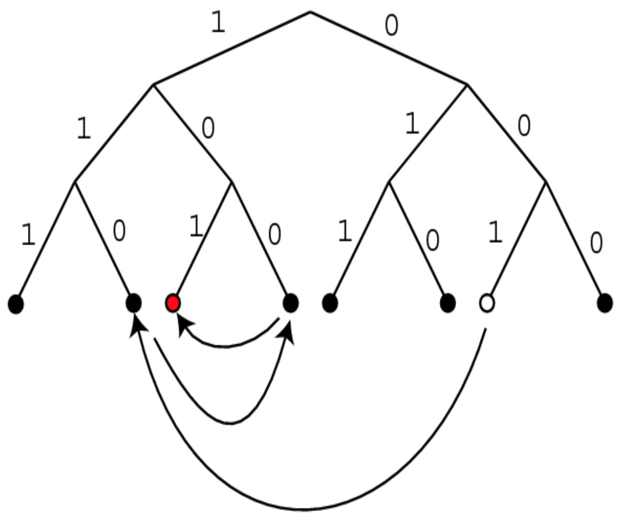
Kademlia RPCs

- PING
  - Probes a node to see if it is online.
- STORE
  - Instructs a node to store a <key, value> pair.
- FIND_NODE
  - Returns information for the k nodes it knows about closest to the target ID.
  - It can be from one kbucket or more.
- FIND_VALUE
  - Like FIND_NODE, ...
  - But if the recipient has stored they <key, value>, it just returns the stored value.

Store Data

- The <key, value> data is stored in k closest nodes to the key.

Lookup Service



Step1 001

000				[1, 2)
010	011			[2, 4)
110	100	111		[4, 8)

Step2 110

111				[1, 2)
100				[2, 4)
011	010	001	000	[4, 8)

Step3 100

101				[1, 2)
111	110			[2, 4)
001	000	010	011	[4, 8)

Maintaining Kbucket List (Routing Table)

- When a Kademlia node receives any message from another node, it updates the appropriate kbucket for the sender’s node ID.

- If the sending node already exists in the kbucket:
  - Moves it to the tail of the list.
- Otherwise:
  - If the bucket has fewer than k entries:
    - Inserts the new sender at the tail of the list.
  - Otherwise:
    - Pings the kbucket’s least-recently seen node:
    - If the least-recently seen node fails to respond:
      - it is evicted from the k-bucket and the new sender inserted at the tail.
    - Otherwise:
      - it is moved to the tail of the list, and the new sender’s contact is discarded.

Maintaining Kbucket List (Routing Table)

- Buckets will generally be kept constantly fresh, due to traffic of requests travelling through nodes.
- When there is no traffic: each peer picks a random ID in kbucket's range and performs a node search for that ID.

Join

- Node P contacts to an already participating node Q.
- P inserts Q into the appropriate kbucket.
- P then performs a node lookup for its own node ID.

Leave And Failure

- No action!
- If a node does not respond to the PING message, remove it from the table.

Kademlia and other DHTs

Kademlia vs. Chord

- like Chord
  - When $\alpha = 1$ the lookup algorithm resembles Chord's in term of message cost and the latency of detecting failed nodes.
- Unlike Chord
  - XOR metric is symmetric, while Chord's metric is asymmetric.

Kademlia vs. Pastry

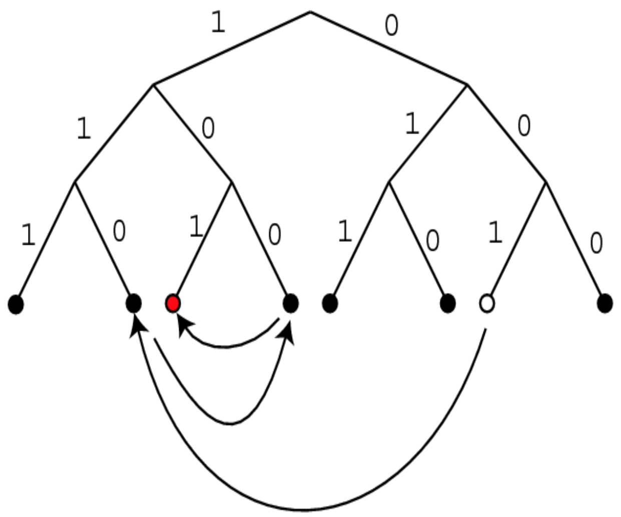
- like Pastry
  - The same routing table.

Pastry	Node 001 routing table				Kademlia
P = 2	000				[1, 2)
P = 1	010	011			[2, 4)
P = 0	110	100	111	101	[4, 8)

- Unlike Pastry
  - $\alpha = 3$ by default in Kademlia, while $\alpha = 1$ in Pastry.

DONE!

A Page To Remember



Step1 001

000				[1, 2)
010	011			[2, 4)
110	100	111		[4, 8)

Step2 110

111				[1, 2)
100				[2, 4)
011	010	001	000	[4, 8)

Step3 100

101				[1, 2)
111	110			[2, 4)
001	000	010	011	[4, 8)

References

- [1] Maymounkov, P. and Mazières, D. 2002. Kademlia: *"A Peer-to-Peer Information System Based on the XOR Metric"*. In Revised Papers From the First international Workshop on Peer-To-Peer Systems (March 07 - 08, 2002). P. Druschel, M. F. Kaashoek, and A. I. Rowstron, Eds. Lecture Notes In Computer Science, vol. 2429. Springer-Verlag, London, 53-65.

Question?