Lesson 5:
PREFIX-BASED DHT:
KADEMLIA

14/3/2018
PREFIX MATCHING DHT

- Plaxton, Rajamaran and Richa: **Routing: prefix or plaxton routing:**
  - a mechanism for the efficient diffusion of object over a network
  - published in 1997, before P2P systems came about!
- Basic ideas: a generalization of the routing on hypercubes.
  - map the nodes and the keys to numbers of m digits of a certain base
  - assign each key to the node with which it shares the longest prefix, if possible

```
base = 4, m= 6

321302  →  321333
```

- prefix matching DHT: a family of DHT
  - Pastry
  - Tapestry
  - Kademlia
IDENTIFIER SPACE

- $2^l$-bit identifiers (typically: $l = 128$) wrap-around at $2^l - 1 \leftrightarrow 0$
- Interpret identifiers to the base of $2^b$ (typically: $b = 4$, base 16)
- Identifiers described by a tree structure
  - Kademlia: $b=1$, binary tree
- pair keys on longer prefix matching nodes, if possible, or the node with numerically close Id.

![Kademlia Binary Tree Diagram]

$l=6$: 6-bit identifiers
$b=2$: base 4

key
node
managed by
The identifier tree is modelled by a tree

- depth of the tree \( \leq l \), \( l \) the length of the identifiers.
  - internal nodes correspond to identifier prefixes
- each node has \( b \) sons (one son for each digit in that base)
- leaves: keys & node ids;
- the routing table of each node includes some references to nodes of this tree
• \( b = \) symbol size = number of bits corrected at each step

• In Pastry \( b=4 \)

• In Kademlia \( b=1 \), at each routing step at least a bit is corrected

• The size of the node routing table and the number of look-up hops depends on \( b \)
  
  • Routing: \( O(\log_b(n)) \)

• **Remark:** some IP-level protocols may be considered prefix matching, where the matching is computed on the IP address of the nodes.
PLAXTON ROUTING TABLES

- **K-buckets**: list of references to nodes, stored in the routing table

- At each lookup step each node has the possibility to choose among K different contacts
  - K = 1 in Pastry
  - K ≈ 20 in Kademia

- A value K > 1 guarantees
  - an higher robustness and tolerance to faults
  - possibility to choose among alternative routing paths
  - possibility to search the key in parallel on different paths
    - **parallel routing**: a key K received by a node is sent in parallel to a set of nodes, taken from the k-bucket

- Chord has a single contact (finger) in each row of the routing table: very strict constraint on the finger to insert
• **iterative routing:**
  - node sending the lookup request manages the search process
  - at each routing step, that node waits for a reply
  - the received reply includes a notification of the next routing step

• **recursive routing:** look up passes from node to node without the intervention of the starting node

• Kademlia
  - iterative routing
THE KADEMLIA DHT

http://xlattice.sourceforge.net/components/protocol/kademlia/specs.html

• initially proposed by P. Maymounkov e D. Mazières (University of New York)

• its protocol is used by the largest public DHTs.
  • KAD network (emule)
  • Vuze network,
  • BitTorrent Mainline DHT (MDHT), BEP 5 specification

• a set of characteristics which are not offered by any existing DHT (Chord, Pastry,..)
  • routing information spreads automatically as a side-effect of lookups
  • flexibility to send multiple requests in parallel to speed up lookups by avoiding timeout delays (parallel routing)
  • iterative routing
• exploited by Emule and Bittorrent
THE XOR METRIC

"distance" between two objects: bitwise \( \oplus \) (XOR) operation on their identifiers (160 bit space), interpreted as an unsigned integer

\[
\text{ID} = \text{sha-1 (Communication Breakdown)} = a1174eb9d7b9150ac6077b3baa7d378486447a0d
\]

Node A
\[
\text{ID}_A = \text{sha-1 (194.29.169.2): 67a83db681412740c808c949761b9b4aec0a492}
\]

Node B
\[
\text{ID}_B = \text{sha-1 (194.29.160.5): e6954b6744885214bde257d3e41d5bc457bb3474}
\]

Node C
\[
\text{ID}_C = \text{sha-1 (175.165.110.85): 3afbeba4271a07e962c87784a385538eba8a4882}
\]

\[
\begin{align*}
\text{ID}_A \text{ XOR ID} &= C6BF730F56FD077ECA87F7AF3D1C8E302884DE9F \\
\text{ID}_B \text{ XOR ID} &= 478205DE9331471E7BD52CE84E606C40DF1FF4E79 \\
\text{ID}_C \text{ XOR ID} &= 9BECA51DF0A312E3A4CF0CBF09F8640A3CCE328F
\end{align*}
\]

Node B is the closest (4<C and 4<9) and should store \((\text{ID}, 94.29.160.5, 3465)\)
XOR: IS REALLY A METRIC?

- $d(x,y) > 0$
- $d(x,y) = 0$ if and only if $x=y$
- $\forall x,y: d(x,y) = d(y,x)$ symmetry
- $d(x,y) \oplus d(y,z) = d(x,z)$ transitivity
- $d(x,y) + d(y,z) \geq d(x,z)$ triangular inequality
  - Follows from the previous one
- given $x$ and a distance $\Delta$, it exists a single $y$ such that $d(x,y) = \Delta$ unidirectionality
  - $x=1001$, $\Delta=0001$, the only point at distance D form $x$ is $y=1000$
WHY THE XOR METRIC?

- any peer in the same half of the ID space as the starting peer is closer to it than any peer from the other half.
- distance graph looks the same in both halves but shifted along the y-axis.
- the same is true when further separating each half into smaller fractions
CHORD METRIC IS NOT SYMMETRIC

• Chord calculates the distance between two peers by subtracting the identifiers numerically from each other

• Take into consideration that the identifier space is “wrapped”

\[
distance(A, B) = B - A \quad \text{(for } B \geq A) \\
= B - A + 2^N \quad \text{(for } B < A)
\]

that is

\[
distance(A,B) = (B - A + 2^N) \mod 2^N
\]

• For instance:

\[
distance(0, 11) = (11 - 0 + 16) \mod 16 = (27) \mod 16 = 11
\]
\[
distance(11, 0) = (0 - 11 + 16) \mod 16 = (5) \mod 16 = 5
\]

• Chord metric is not symmetric
WHY THE XOR METRIC?

why this metric?

- related to the identifier prefix:
  the larger is the prefix common to two nodes, the smaller is their distance computed by $\oplus$

- close” nodes are characterized by a long common prefix
  
  $000100 \oplus 000110 = 000010$ common prefix $0001$, distance $2$
  
  $010000 \oplus 000001 = 010001$ common prefix $0$, distance $17$

- Routing guided by prefix matching
Strictly related to the identifier tree and to prefix matching routing:

- assign a key $K$ to the node whose identifier is closer to $K$, according to the $\oplus$ metrics.
- key $\downarrow$ is assigned to a leaf $\circ$ in the left subtree, the closest leaf, while the numerically closest leaf is in the right subtree.
- according to $\oplus$ metric, the key is closer to any node in its subtree than to nodes in other subtrees.
- look-up guided by the tree structure.
WHY THE XOR METRIC?

- **symmetric:**
  - enables Kademlia to learn contacts from ordinary queries it receives.
  - helps in building the routing tables
  - non symmetric distances (Chord fingers are not symmetric) do not allow this

- **unidirectional:**
  - lookups for same key converge to the same path, and thus caching items along this path is good to avoid hotspots.
two leaves may be close in the tree and numerically close, but they are
distant according to the metrics $\oplus$

\[ 1000 \oplus 0111 = 1111 = 15, \]

numerical difference between 1000 and 0111 = 1
consider two identifiers $x$ and $y$ of length $L$ that
- share a common prefix of length $p$
- differ in the last $i=L-p$ bits

their distance (exclusive OR) will be such that

$$2^{i-1} \leq d(x,y) < 2^i$$

\[
\begin{align*}
X &= 0 1 0 1 1 0 \\
Y &= 0 1 1 1 1 0 \\
X \oplus Y &= 0 0 1 0 0 0, \quad d(x,y) = 2^3=8 \text{ (minimal distance)}.
\end{align*}
\]

\[
\begin{align*}
X &= 0 1 0 1 1 0 \\
Y &= 0 1 1 0 0 1 \\
X \oplus Y &= 0 0 1 1 1 1, \quad d(x,y) = 2^4-1 = 15 \text{ (maximal distance)}
\end{align*}
\]

this enables to pair the nodes of the subtree with an identifier range
DISTANCES AND IDENTIFIER TREE

- consider a leaf in the left subtree and a leaf in the right subtree
- the length of the shared prefix is $= 0$
- the distance varies:

$$2^3 \leq d < 2^4$$

$$0111 \oplus 1000 = 1111 = 15 \text{ (maximal, numerical distance is minimal)}$$

$$0111 \oplus 1111 = 1000 = 8 \text{ (minimal, numerical difference is high)}$$
The distance between the two nodes in the figure is minimal: they differ in only the last bit:

\[ 2^0 \leq d < 2^1 \]

\[ 0110 \oplus 0111 = 0001 = 1 \]
peers present in the network are much lesser than the identifiers: the space of identifier is very large and not all the identifiers are paired with a peer.

node tree: an unbalanced binary tree showing only the identifiers of peers present in the network (a subset of all the identifiers).

a leaf for each peer: the identifier of the leaf is a prefix of that of the peer.
A leaf in the node tree may correspond to a prefix P of the identifier of a peer if the overlay includes a single peer with that prefix P.

- 0011 uniquely identifies the red peer
- no other peer in the overlay with the same prefix
- The rest of the path is not useful to identify the peer
Routing table, the basic idea:

- maintains some contacts for each subtree/common prefix
- the figure, shows the contacts in the routing table of the red peer
- It includes 2 contacts for each subtree
THE ROUTING TABLE

- routing table: the rows are **k-buckets**, each one contains k contacts

- for any $0 \leq i < 160$, the i-th row of the routing table of node_ID contains up to k-buckets such as:
  \[ d_{\text{XOR}}(\text{node}_\text{ID}, \text{contact}_\text{ID}) \in [2^i, 2^{i+1}) \]

- each k-bucket corresponds to a subtree

- each row stores:
  
  (ID, IP, UDP port)
ROUTING TABLE

- each k-bucket corresponds to a prefix and covers a subset of the identifier space: the set of all the k-buckets cover the whole identifier space

- the first entries of the routing table correspond to peers sharing a long prefix with the owner of the routing table
  - may include a few contacts

- the last entries of the routing table correspond to peers sharing a smaller prefix, and cover a larger set of identifiers
  - may include a larger number of contacts, never more than K contacts

- the value of K is defined such that the probability that a crash of more of K nodes is a rare event

- nodes in each bucket are maintained ordered such that:
  - least recently contacted nodes are in the first positions of the list
Routing Tables Examples

Subtrees for node N0

Subtrees for node N6

Subtrees for node N15

Laura Ricci
Prefix Based DHT: Kademia
K-BUCKETS MANAGEMENT

- K-buckets: lists, least-recently seen nodes in the first positions.
- when a node receives any message from another node, it updates the appropriate k-bucket for the sender’s node ID

\[\text{if the sending node already exists in the k-bucket: move it to the tail of the list.}\]
\[\text{otherwise}\]
  \[\text{if the bucket has fewer than k entries: insert the new sender at the tail of the list}\]
  \[\text{otherwise}\]
    \[\text{ping the k-bucket’s least-recently seen node}\]
    \[\text{if the least-recently seen node fails to respond evict it from the k-bucket and insert}\]
    \[\text{the new sender at the tail.}\]
    \[\text{otherwise move it to the tail of the list, and the new sender’s contact is discarded.}\]
Esempio:

Tabella nodo 0

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>[1,2)</td>
<td>nodo1</td>
</tr>
<tr>
<td>1</td>
<td>[2,4)</td>
<td>nodo2 nodo3</td>
</tr>
<tr>
<td>2</td>
<td>[4,0)</td>
<td>nodo4 nodo5 nodo4</td>
</tr>
</tbody>
</table>

m=3; k=2

1. 6 manda un messaggio a 0, ma il k-bucket è pieno (k=2)
1. 0 interroga 4
1. 4 risponde e viene spostato al fondo (6 è fuori)
Esempio:

1. 6 manda un messaggio a 0, ma il k-bucket è pieno (k=2)
2. 0 interroga 4 (nodo in testa)
3. 4 non risponde: 6 è messo in fondo al k-bucket
K BUCKETS MANAGEMENT

Why this politics for bucket management?

- x axis: minutes
- y axis: percentage of peers that, being on-line for x minutes, are on line also for next x+60 minutes

percentage of nodes in a Gnutella network which remains online in the next hour as a function of their uptime.

- the longer a node remains on-line, the higher is the probability that it remains online for a even longer interval of time
K BUCKETS MANAGEMENT

- Why this politics for bucket management?
  - always prefer the contacts which are have been present in the network for a longer time, because, with high probability, they will remain in the network also in the next period.
  - least recently seen eviction
    - deleted nodes are those in the first position of the k-buckets list, that are those least recently contacted.
K-BUCKETS PERIODIC REFRESHMENT

- the k-buckets are refreshed for each query passing through the node
  - if a node has left the network, new information received from the queries “refreshes” the k-bucket list.

- However, it may be the case that a k-bucket is not refreshed for a given period of time, due to the lack of messages from nodes in the range covered by the k-bucket

- For this reason, a refresh is periodically executed (once each hour)
  - Kademlia chooses an identifier belonging to the range covered by the bucket at random and search that identifier
  - If the node with that identifier sends a reply it is inserted in the k-bucket
• \( \alpha \): number of nodes to which the query is propagated, at each routing step
• iterative Routing
• an example of routing for \( \alpha = 1 \)
Routing at a glance

Black node: query source (0011)
Orange Node: query target (1110)
Green Node: nodes known from a bucket of the black nodes

Prefix Match Routing: at each step the common prefix with the key increases
ROUTING AT A GLANCE

Black node : query source (0011)
Orange Node : query target(1110)
Green Node : nodes known form the back nodes (in its routing table)
Blue Node : returns further nodes to the black node

Prefix Match Routing: at each step the common prefix with the key increases
Black node: query source (0011)
Orange Node: query target (1110)
Green Node: nodes known from the back nodes (in its routing table)
Blue Node: returns further nodes to the black node

Prefix Match Routing: at each step the common prefix with the key increases
Routing at a Glance

Black node: query source (0011)
Orange Node: query target (1110)
Green Node: nodes known form the back nodes (in its routing table)
Blue Node: returns further nodes to the black node

Prefix Match Routing: at each step the common prefix with the key increases
The blue node 0011 looks for the red node 1101.
It has a reference to the green nodes 1001 e 1110, and
\[\text{dist (1101, 1001)} = 4, \text{dist}(1101, 1110)= 3\]
it is possible that the node 1001 which is more distant from the target,
has a reference to the target, while the closest node 1110 has no reference
to the target

parallel routing: the blue node sends the request to both nodes
• Consider a node $x$ and two further nodes $y$ e $z$, with

$$\text{dist}(y, x) < \text{dist}(z, x).$$

$z$ knows $x$, while $y$ does not know $x$.

• The dispatch of the query to the node which is closest to the target not necessarily implies the smaller path toward the target.

• Routing: the query is sent to the $\alpha > 1$ nodes closest to the target.

• The unidirectionality of the $x$-or metric guarantees that all the paths converge toward the target.
ROUTING: AN OBSERVATION

- for larger row numbers, the number of possible choices decreases exponentially
- in row $i+1$ we have $1/2$ the choices we had in row $i$, because of more contraints on the prefix
- for larger row numbers the distance to the nearest neighbor increases exponentially
- the distance of the source to the target is approximately equal to the distance in the last step.
Kademlia protocol consists of 4 primitive (not iterative) operations, defined as Remote Procedure Calls (RPCs) exploiting UDP:

- **PING v→w**
  - probe node w to see if its online

- **STORE v→w (Key, Value)**
  - instructs node w to store a <key, value> pair

- **FIND_NODE v→w (T)**
  - In: T, 160-bit ID
  - Out: the recipient of the RPC returns k contacts (<IP:Port, NodeID>) that it knows to be “closest” to T

- **FIND_VALUE v→w(T)**
  - In: T, 160-bit ID
  - Out: if a value corresponding to T is present, the associated data is returned, otherwise it is equivalent to FIND_NODE and a set of k triples is returned.
NODE LOOK UP

select k-closest = α contacts from the non-empty k-bucket closest to the key

if there are fewer than α contacts in that bucket then
    select k-closest = k-closest U closest contacts from other buckets.
    closestNode = the closest node in k-closest

repeat
    send in parallel, asynchronously FIND_NODE (or FIND_VALUE) to the contacts in k-closest
    each contact, if live, should normally return k nodes, add them to k-closest
    add to k-closest the new received nodes
    update closestNode
    select from k-closest α closest contacts which have not been queried yet and send them FIND_NODE
    until no node closer to the target than closestNode is returned
    send in parallel asynchronously FIND_NODE (or FIND_VALUE) to the k closest nodes it has not already queried

return the k closest nodes
The algorithm that Kademlia uses for locating the k nodes nearest to a key:

- is iterative
- exploits the basic function defined in the previous slides
- intuitively, at each iteration, the XOR metric is reduced by $\frac{1}{2}$ and results in smaller size k-buckets
P looks for the key Q (which is the identifier of a node or of a content)

- looks in the bucket-list the $a$ nodes closest to Q
- looks in the k-bucket closest to the key and not empty. If it includes less than $a$ nodes ($\alpha=3$), looks in close buckets
- selected contacts may belong to different k-buckets
LOOK_UP \((Q)\)

Here are stored the nodes closest to \(Q\)

\[ P \text{ selects } \alpha \text{ nodes from the selected bucket} \]
P sends the query in parallel to all the selected nodes, through the RPC FIND_NODE(Q)
• Each contacted node finds out, in turn, k nodes closer to the key
• Each node may exploit a different bucket of its routing table.
Iterative Routing:

- each node returns the results to P
- the results are inserted in a list which is ordered on the basis of the distance between the node and Q
- P continues the routing process through the results obtained from P
LOOK_UP (Q)

P updates its k-buckets list with the received information.

Information received A, B, C

P again selects $\alpha$ nodes from the received information.

If it obtains nodes closer to the target with respect to the preceding nodes, it performs look-up on these nodes.

Otherwise, it chooses further nodes from those which have not been contacted before.
FIND_NODE (Q)

Terminate when a round of FIND_NODE(T) fails to return any closer nodes
To store a *(key, value)* pair, a participant locates the *k closest nodes* to the key and sends them STORE RPCs.
- data is replicated on *k* nodes

In any case, consider following scenarios
- some of the *k* nodes (or all) that initially get the *(key, value)* pair leave the network
- new nodes enter the network with an identifier closer to some published key than the nodes on which the key-value pair was originally published

**Re-publishing mechanism:**
- each node re-publishes *(key, value)* pairs as necessary to keep them alive.

For Kademlia’s file sharing application, the original publisher of a *(key, value)* pair is required to **republish it every 24 hours**. Otherwise, *(key, value)* pairs expire 24 hours after publication.
NODE JOIN

- **new** (joining node) borrow an alive node’s ID off-line (bootstrap node **boot**)
- initial routing table of **new** has a single k-bucket containing **new** and **boot**.
- **new** performs FIND_NODE(\textit{new}) through **boot** to learn about other nodes
  - finds some nodes close to itself. Some low index k-buckets are filled
- other nodes start to know **new** and insert it in their routing tables
- **new** performs FIND_NODE(\text{ID}) for identifier ID in k-buckets further away than its own k-bucket generate a node identifier at random
- k-buckets are subsequently enriched with the information received in the queries passing through **new**

Note the flexibility of this procedure with respect to the joining procedure of Chord!
Maintenance

- refresh k-buckets for which there was no contact within a certain time, e.g. an hour
- refresh means lookup of random ID in bucket.

Storage & Caching

- to store a value, locate the k closest nodes to the ID of the node via lookup and then store (replicate) the value at these nodes.
- values are considered soft-state and need refreshing.
- values are cached at the first node on a path that did not know it.

Leave

- the node leave does not require further operations
  - if a node does not reply, it will be discarded from the k-buckets
PERIODIC TASKS

- Each node periodically publishes the pair \(<\text{key,values}>\) to guarantee the persistence of the data inserted in the overlay.

- The periodic publication mechanism has been introduced for:
  - avoid data loss as a consequence of the voluntary leave or of the crash of a node.

- Some optimizations are defined to decrease the number of exchanged messages:
  - if a node receives a STORE, it suppose that the STORE has been sent to the close neighbours and does not publish again the key in the next hour.
Chord routing table is rigid, has only one way information flow
- complicates recovery process
- incoming traffic cannot be used for reinforcing routing table.
- less fault-tolerance
Kademlia defines a flexible routing table:

- symmetric distances
- alternative paths toward a node, possibility of parallel lookups
- managing the routing table has a lower cost
- locality: store round-trip-time together each contact and choose the contact with lower round trip time

The symmetric metrics enables each node to enrich its routing table through the query.

On the contrary, in Chord:

- if a node $x$ receives a query from $y$, $y$ has in its finger table a reference to $x$, but $x$ may be not a finger of $y$.
- the information included in a received query cannot, in general, be exploited to enrich the finger table.
KADEMLIA: SUMMARY

Strengths
- low control message overhead
- tolerance to node failure and leave
- capable of selecting low-latency path for query routing
- provable performance bounds

Weaknesses
- non-uniform distribution of nodes in ID-space results into imbalanced routing table and inefficient routing
- balancing of storage load is not truly solved
- originally underspecified, plethora of different implementations
- hard to provide analytical results
- non-deterministic results of routing (time, neighbourhood)
# DHT: COMPARISON

<table>
<thead>
<tr>
<th>CAN</th>
<th>Chord</th>
<th>Kademlia</th>
<th>Koord</th>
<th>Pastry</th>
<th>Tapestry</th>
<th>Viceroy</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Foundation</strong></td>
<td>Multi-dimensional space (d-dimensional torus)</td>
<td>Circular space (hypercube)</td>
<td>XOR metric</td>
<td>de Bruijn graph</td>
<td>Plaxton-style mesh (hypercube)</td>
<td>Plaxton-style mesh (hypercube)</td>
</tr>
<tr>
<td><strong>Routing function</strong></td>
<td>Maps (key,value) pairs to coordinate space</td>
<td>Matching key and nodeID</td>
<td>Matching key and nodeID</td>
<td>Matching key and nodeID</td>
<td>Matching key and prefix in nodeID</td>
<td>Suffix matching</td>
</tr>
<tr>
<td><strong>System parameters</strong></td>
<td>Number of peers N, number of dimensions d</td>
<td>Number of peers N</td>
<td>Number of peers N, base of peer identifier B</td>
<td>Number of peers N</td>
<td>Number of peers N, base of peer identifier B</td>
<td>Number of peers N</td>
</tr>
<tr>
<td><strong>Routing performance</strong></td>
<td>$O(dN^d)$</td>
<td>$O(\log N)$</td>
<td>$O(\log_B N)$ + small constant</td>
<td>Between $O(\log \log N)$ and $O(\log N)$, depending on state</td>
<td>$O(\log \log N)$</td>
<td>$O(\log \log N)$</td>
</tr>
<tr>
<td><strong>Routing state</strong></td>
<td>$2d$</td>
<td>$\log N$</td>
<td>$B\log_B N + B$</td>
<td>From constant to $\log N$</td>
<td>$2B\log_B N$</td>
<td>$\log_B N$</td>
</tr>
<tr>
<td><strong>Joins/leaves</strong></td>
<td>$2d$</td>
<td>$(\log N)^2$</td>
<td>$\log_B N$ + small constant</td>
<td>$\log N$</td>
<td>$\log_B N$</td>
<td>$\log_B N$</td>
</tr>
</tbody>
</table>