Dynamic Networks for Peer-to-Peer Systems

Pierre Fraigniaud
CNRS
Lab. de Recherche en Informatique (LRI)
Univ. Paris-Sud, Orsay

Joint work with Philippe Gauron (LRI)

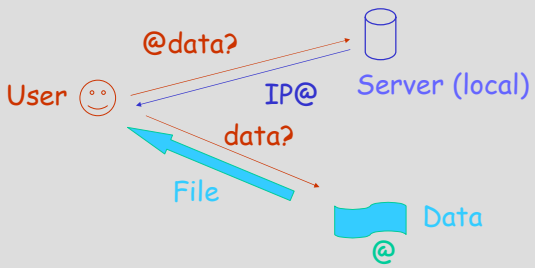
Peer-to-Peer Systems (P2P)

- Opposed to the master-slave model
- A group of users (computers) share a common space in a decentralized manner.
- Objectives :
  - Share data (music, movies, etc.)
  - Share resources (computing facilities)

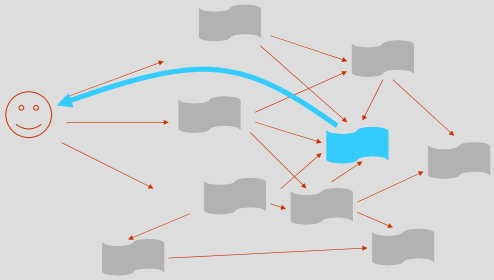
Main (Ideal) Characteristics

- No central server
- Cooperation between users
- Users can join and leave the system at any time
- Fault-tolerance
- Anonymity
- Security
- Self organization

Half-Decentralized Sytems



Decentralized Systems



Different Types of Lookups

- Flooding (e.g., Gnutella)
  - Pro : simple
  - Con : network load → non exhaustive
- Routing from A to B=h(d).
  - Pro : exhaustive
  - Difficulty : routing → Distributed Hash Tables (a.k.a. Content-Addressable Network)

Problem

Design a dynamic network (i.e., nodes join and leave at their convenience) in which look-up routing and updating are "efficient".

Constraints

- Fast updates
  Limited amount of control messages → small degree
- Fast lookups
  Short lookup routes → small diameter
- Balanced traffic
  No hot spot during lookup routing

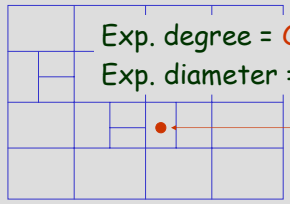
Distributed Hash Tables

- Data d → h(d) = key ∈ K
- Nodes = computers = users
- Arc (A,B) ⇔ A store the IP@ of B in its routing table
- Each computer stores a lookup table: key vs. IP@.
- Lookup routing performs on a key-basis

CAN

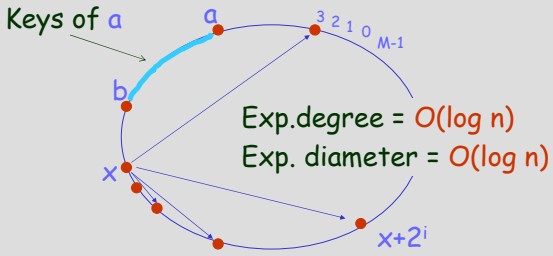
"Content-Addressable Network"

d-dimensionnal torus



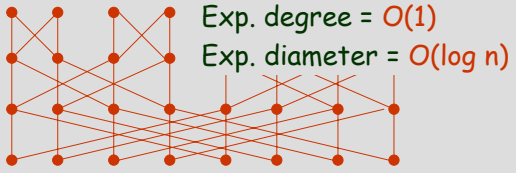
Chord

d-dimensional hypercube



Viceroy

Butterfly Network



Why yet another DHT?

- Most of the existing DHTs have expected degree at least $\Omega(\log n)$
- CAN has expected degree $O(d)$ but diameter $O(dn^{1/d})$
- Viceroy has degree $O(1)$ and diameter $O(\log n)$, but is based on relatively complex machineries.

D2B

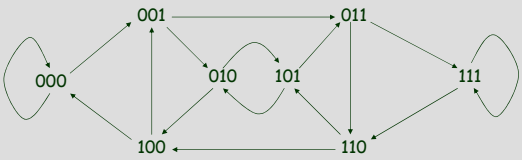
- Expected #key per node $O(|K|/n)$ $O(|K|\log n/n)$ with high probability.
- Expected degree $O(1)$; $O(\log n)$ w.h.p.
- Length of lookup route $O(\log n)$ w.h.p.
- Congestion minimal for a constant degree network: $O(\log n/n)$

Underlying topology

Based on the *de Bruijn* Network

$V = \{\text{binary sequences of length } k\}$

$E = \{(x_1x_2\ldots x_k)\rightarrow(x_2\ldots x_ky), y=0 \text{ or } 1\}$



Node and key labels

- Node = binary sequence of length $\leq m$.
- Key = binary sequence of length = m.

→ up to 2^m nodes and keys

In practice, set $m=128$ or even 256

- The key κ is stored by node x if and only if x is a prefix of κ.

Universal Prefix Set

Let W_i, i=1,…,q, be q binary sequences.

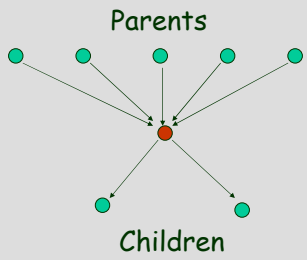
The set $S=\{W_1,W_2,\ldots,W_q\}$ is a *universal prefix set* if and only if, for any infinite binary sequence B, there is one and only one W_i which is a prefix of B.

Example: {0,11,100,1010,10110,10111}

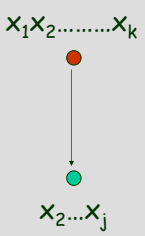
Remark: {e} where e is the empty sequence is a universal prefix set.

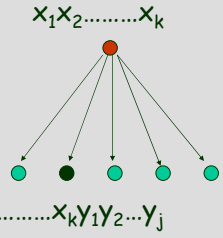
By construction, the set of nodes in D2B is a universal prefix set.

Routing Connections



Children Connections and Routing





The set $\{y_1y_2\ldots y_j\}$ is a UPS

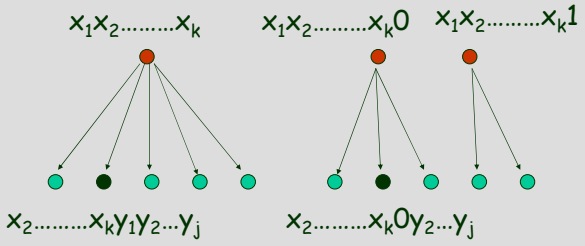
Join Procedure (1/3)

- A joining node u contacts an entry point v in the network;
- Node u selects a m-bit binary sequence L at random: its preliminary label;
- A request for join is routed from v to the node w that is in charge of key L;

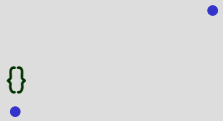
Join Procedure (2/3)

- Node w labeled $x_1x_2\ldots\ldots x_k$ extends its label to $x_1x_2\ldots\ldots x_k0$
- Node u takes label $x_1x_2\ldots\ldots x_k1$
- Node w transfers to u all keys K such that $x_1x_2\ldots\ldots x_k1$ is prefix of K.

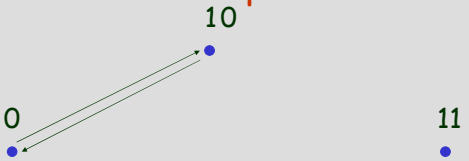
Join Procedure (3/3)

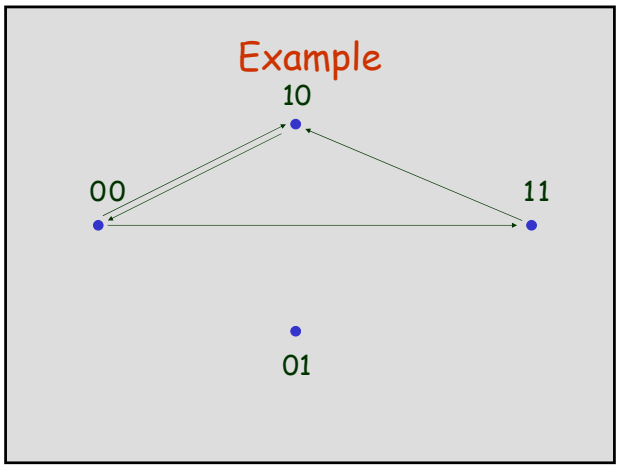


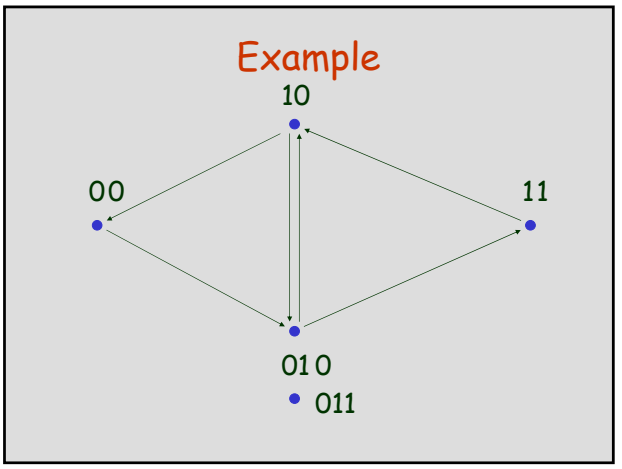
Example

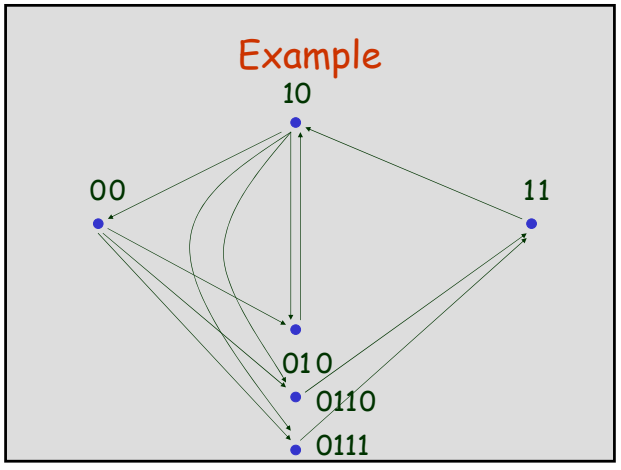


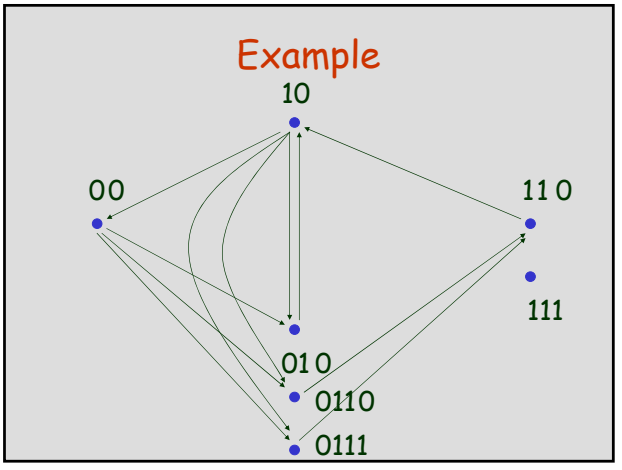
Example

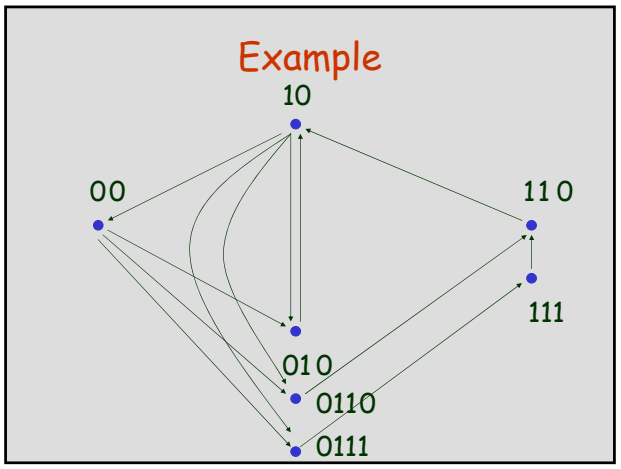


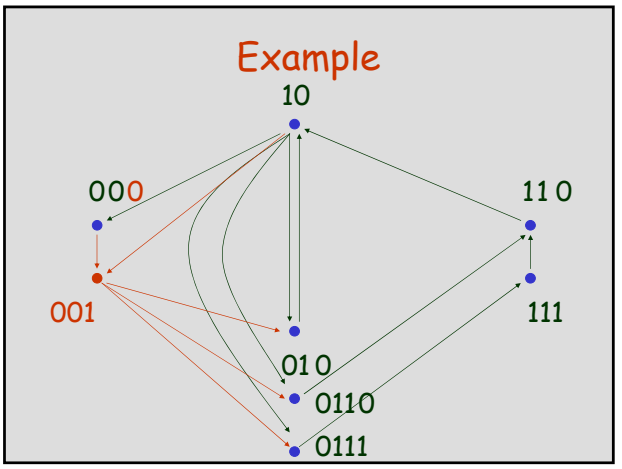

Example
10
00
11
01

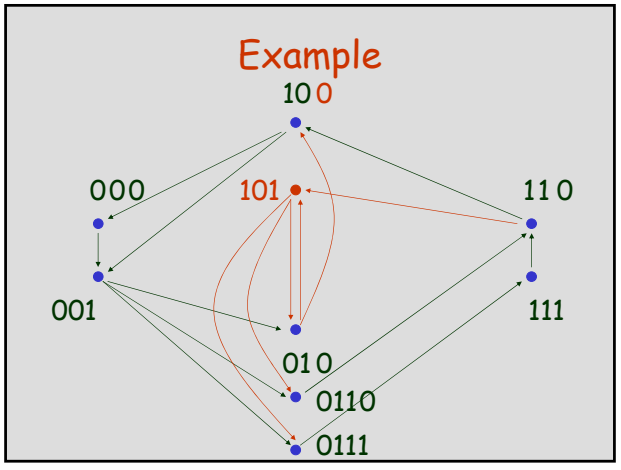

Example
10
00
11
010
011

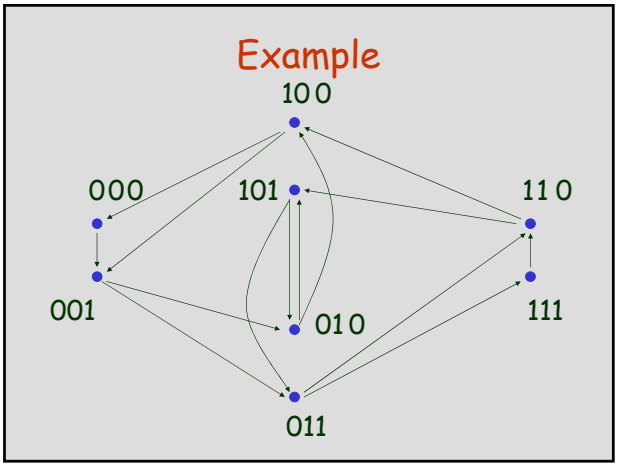

Example
10
00
11
010
0110
0111


Example
10
00
110
111
010
0110
0111


Example
10
00
110
111
010
0110
0111


Example
10
000
110
001
111
010
0110
0111


Example
100
000
101
110
001
111
010
0110
0111


Example
100
000
101
110
001
111
010
011

#keys per node (1/2)

$x_1x_2...x_k \rightarrow x_1x_2...x_k$**.....**

#keys per node (2/2)

- Devide K in n/(c log n) intervals, each containing c log n |K|/n keys.
- Let X = #nodes in interval I starting at x
- n Bernouilli trials with probability p = c log n/n
- Chernoff bound:

$$Prob(|\sum X_i - np| > k) < 2e^{-k^2/3np}$$

→ $Prob(|X - c \log n| > (3c)^{1/2} \log n) < 2/n$

→ W.h.p., there is at least one node in I

→ W.h.p., a given node manages $O(|K| \log n/n)$ keys

Lookup routing

Node $x_1x_2.....x_k$ looks for key $\kappa_1\kappa_2.....\kappa_m$

→ $x_2.....x_k\kappa_1...\kappa_h$

→ $x_3.....x_k\kappa_1...\kappa_h\kappa_{h+1}.....\kappa_{h+r}$

→ $x_4.....x_k\kappa_1...\kappa_h\kappa_{h+1}.....\kappa_{h+i}$

→ $x_5.....x_k\kappa_1...\kappa_h\kappa_{h+1}.....\kappa_{h+i}\kappa_{h+i+1}...\kappa_{h+i+s}$

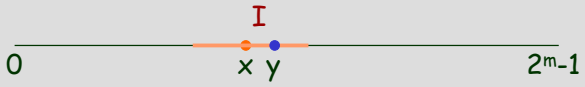
→ $x_6...x_t$

→ $x_7...x_t\ \kappa_1.....\kappa_d$

At most k hops to reach the node in charge of the key $\kappa_1\kappa_2.....\kappa_m$

Length of node label (1/2)

$x_1x_2...x_k \rightarrow x_1x_2...x_k$**.....**



$|I| = c\ |K| \log n/n$

Length of node-label (2/2)

$\text{Prob}(|X - c \log n| > (3c)^{1/2} \log n) < 2/n$

→ W.h.p., at most $O(\log n)$ nodes in I

→ x manages at least $|I|/2^{O(\log n)}$ keys

→ $k \leq m - \log|I| + O(\log n)$

→ $k \leq O(\log n)$

→ W.h.p., a lookup route is of length $O(\log n)$

Degree and congestion

- W.h.p., degree = $O(\log n)$ using similar techniques (expected degree $O(1)$)
- Congestion = proba that a node is traversed by a lookup from a random node to a random key = $O(\log n/n)$

  (Minimum possible for a constant-degree network)

Summary: Expected properties

	Update	Lookup	Congestion
CAN	$O(d)$	$O(dn^{1/d})$	$O(d/n^{1-1/d})$
Small world	$O(1)$	$O(\log^2 n)$	$O(\log^2 n/n)$
Chord	$O(\log n)$	$O(\log n)$	$O(\log n/n)$
Viceroy	$O(1)$	$O(\log n)$	$O(\log n/n)$
D2B	$O(1)$	$O(\log n)$	$O(\log n/n)$

Extensions

- d-dimensional D2B
  - Degree = d
  - Lookups = log n / log d
- Fault-tolerance
- Mapping the physical topology

References

[1] I. Abraham, B. Awerbuch, Y. Azar, Y. Bartal, D. Malkhi, and E. Pavlov. A Generic Scheme for Building Overlay Networks in Adversarial Scenarios. In Int. Parallel and Distributed Processing Symposium (IPDPS), April 2003.

[2] P. Fraigniaud and P. Gauron. The Content-Addressable Network D2B. Technical Report, LRI, Univ. Paris Sud, Jan. 2003. http://www.lri.fr/~pierre

[3] M. Kaashoek and D. Karger. Koorde: A simple degree-optimal distributed hash table. In Int. Peer-to-peer Processing Symposium (IPTPS), Feb. 2003.

[4] M. Naor and U. Wieder. Novel Architecture for P2P Applications: the Continuous-Discrete Approach. To appear in ACM Symp. on Parallelism in Algorithms and Architectures (SPAA), June 2003.