Issues in P2P Systems and Content Distribution

Ernst Biersack
Institut Eurecom
erbi@eurecom.fr

With contributions from
P. Felber, G. Urvoy-Keller, K. Ross, L. Garces

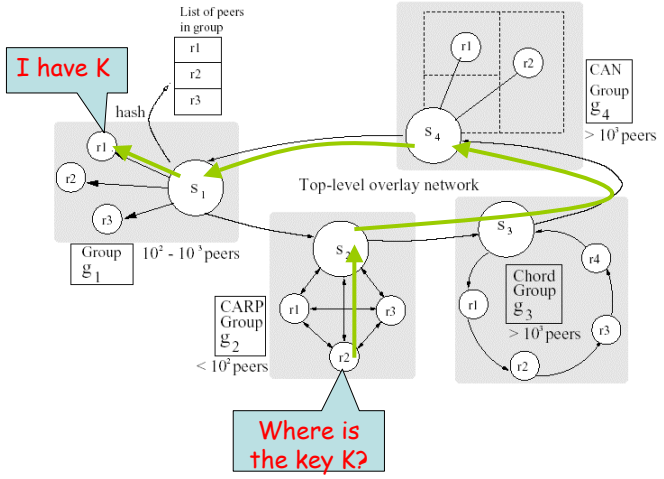
Overview: Issues

- Hierarchical DHTs
- Topology-Aware DHTs
- Scalable Content Distribution using P2P systems

Hierarchical DHTs

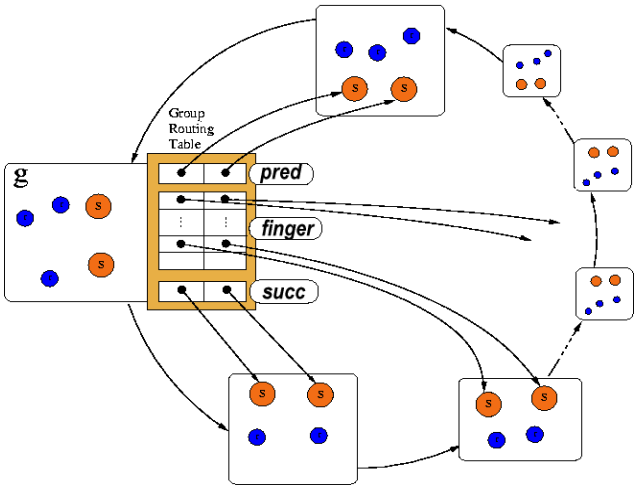
The Internet is organized as a hierarchy

- Can DHTs benefit from hierarchy?
  - Peers are organized in groups
  - Inter-group and Intra-group lookup scheme



Hierarchical DHTs

- Multiple rings among super-peers:



Hierarchical DHTs

- Advantages of hierarchical DHTs :
  - Exploit heterogeneity of peers: By designating the most reliable peers as super-nodes (part of multiple overlays), number of hops to locate a key can be significantly decreased
  - Topological awareness: Peers that are close in the Internet can be in the same group
  - Fewer lookup steps, since number of groups is orders of magnitudes smaller than total number of peers
  - Fewer maintenance messages in wide-area, since most of the overlay maintenance traffic will happen inside a group
  - Heterogeneity of DHTs: Use the DHT the is most appropriate for a given group size. Multiple overlays managed by possibly different DHTs (Chord, CAN, etc.)
  - Facilitates large scale deployment since groups are administratively autonomous (as in intra AS routing)

Hierarchical DHTs

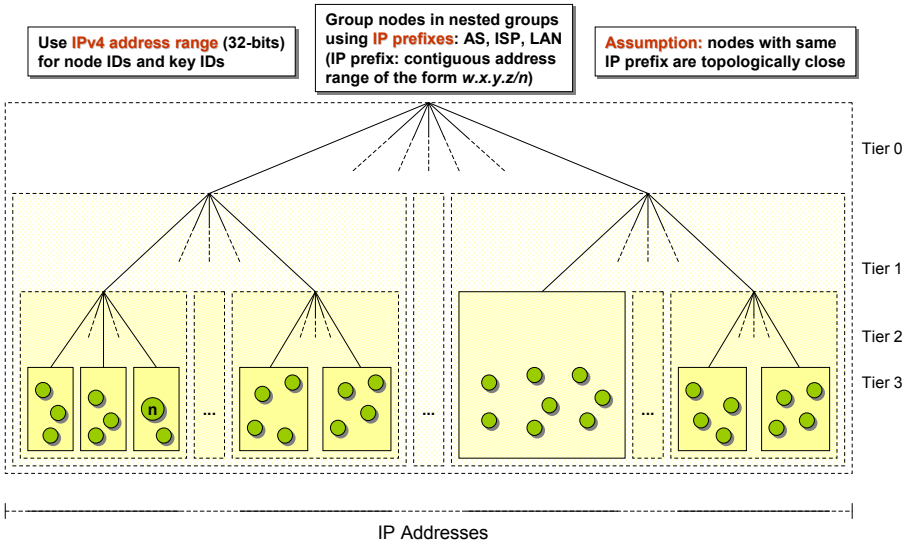
- Open Issues:
  - How can we deploy, maintain such architectures?
    - When to decide to split or merge groups
    - When to promote a node to become supernode

- Luis Garces-Erice, Ernst W. Biersack, Keith W. Ross, Pascal A. Felber, and Guillaume Urvoy-Keller. **Hierarchical P2P Systems**. In *Proceedings of Euro-Par 2003*, Klagenfurt, Austria, 2003

Topology-Aware DHT

- Observation
  - P2P lookup services generally do not take topology into account
  - In Chord/CAN/Pastry, neighbors are often not locally nearby
- Goals
  - Provide small stretch: route packets to their destination along a path that mimics the router-level shortest-path distance
  - Stretch: delay DHT-routing / delay IP-routing
- Our solution
  - TOPLUS (TOPology-centric Look-Up Service), an "extremist design" to topology-aware DHTs
    - Node Ids are IP addresses
    - Nested groups
      - Based on IP prefixes that are obtained from BGP routing tables + some massaging

TOPLUS Architecture

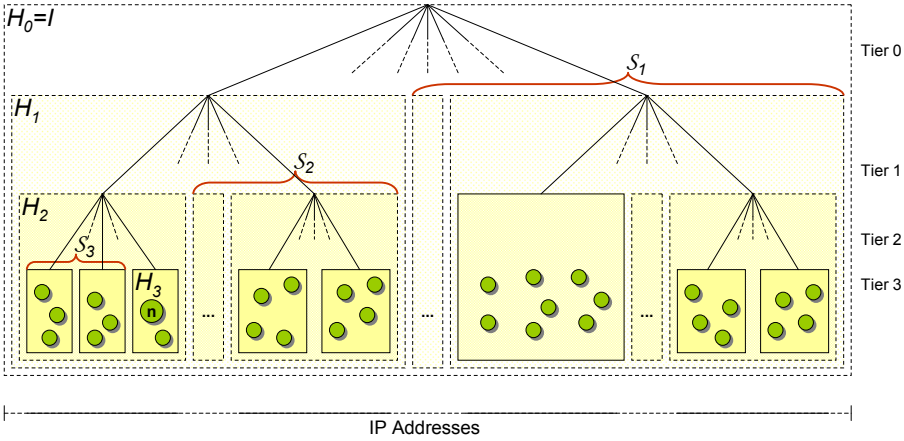


Node State

Each node *n* is part of a series of telescoping sets H_i with siblings S_i

Node *n* must know all up nodes in inner group

Node *n* must know one delegate node in each tier *i* set $S \in S_i$

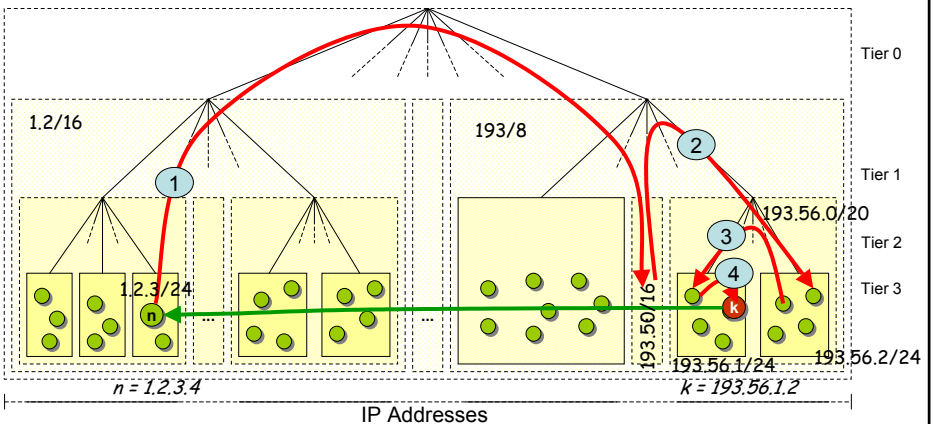


Prefix Routing Lookup

Compute 32-bits key *k* (using hash function)

Perform longest IP-prefix match against entries in routing table using XOR metric

Route message to node in inner group with closest ID (according to XOR metric)

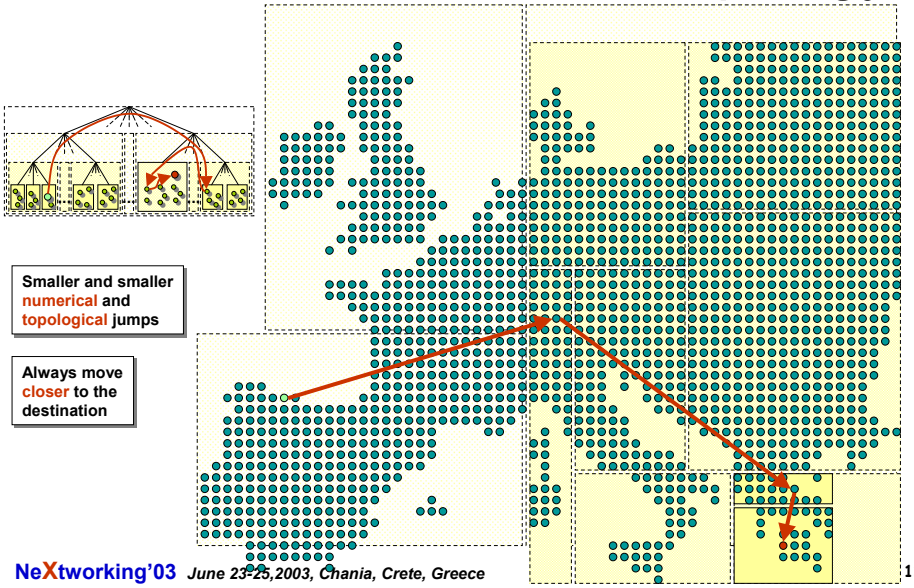


Number of hops < H+1, H = height of tree

Routing with XOR Metric

- Refinement of longest IP prefix match, based on XOR metric
- To lookup key *k*, node *n* forwards the request to the node in its routing table whose ID *j* is closest to *k* according to XOR metric
  - Let $j = j_{31}j_{30}...j_0$ — $k = k_{31}k_{30}...k_0$ $\quad d(j,k) = \sum_{i=0}^{31} |j_i - k_i| \cdot 2^i$
  - Note that closest ID is unique: $d(j,k) = d(j',k) \Leftrightarrow j = j'$
  - Example (8 bits)
    - k = 10010110
    - j = 10110110 d(j,k) = 2^5 = 32
    - j' = 10001001 d(j',k) = $2^4 + 2^3 + 2^2 + 2^1 + 2^0$ = 31

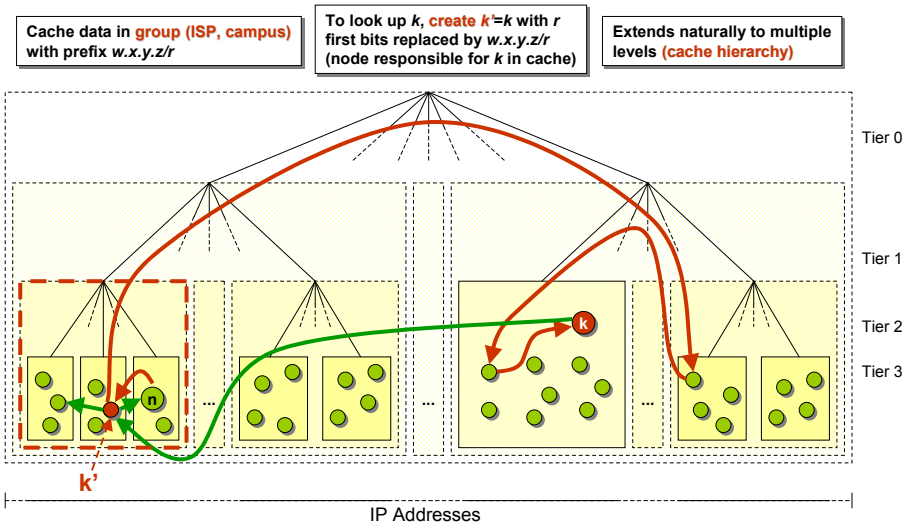
TOPLUS and Network Topology



TOPLUS: Performance

- 250,252 distinct IP prefixes from from Oregon, Michigan University and Routing registries from Castify, RIPE
  - 47,000 tier-1 groups, 10,000 of which have subgroups
  - 11 tiers
- Use King to estimate delay between arbitrary nodes
  ➔ Stretch: 1.17
- Can modify prefix trees (do aggregation) to reduce number of tier-1 groups
  - 16-bit regrouping: tier-1 prefix a.b.c.d/r, with r>16 is moved to tier-2 and a new 16-bit prefix is inserted at tier-1: Stretch: 1.19
  - 8-bit regrouping: tier-1 prefix a.b.c.d/r, with r>16 is moved to tier-2 and a new 8-bit prefix is inserted at tier-1: Stretch: 1.28
    ➔Tradeoff between routing table size and stretch

TOPLUS: On Demand Caching



TOPLUS Summary

- Issues
  - Non-uniform population of ID space (requires bias in hash to balance load)
  - Correlated node failures
- Advantages
  - Small stretch
  - IP longest-prefix matching allows fast forwarding
  - On-demand P2P caching straightforward to implement
  - Can be easily deployed in a "static" environment (e.g., multi-site corporate network)
  - Can be used as benchmark to measure speed of other P2P services

- Luis Garces-Erice, Keith W. Ross, Ernst W. Biersack, Pascal A. Felber, and Guillaume Urvoy-Keller. **Topology-Centric Look-Up Service**. To appear in Proc. Networked Group Communications, Sept. 2003

Scalable Video Distribution

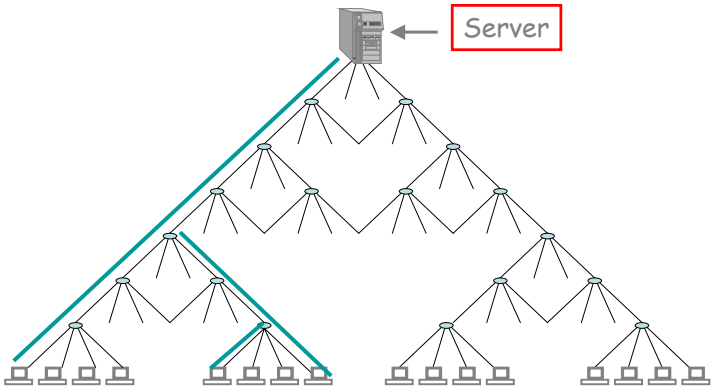
- Assume large number of clients that ask for same video almost simultaneously

Scalable Video Distribution

- Different models:
  - Server-Push or Open loop paradigm
    - **Broadcast schemes** with start-up latency
    - Broadcast schemes with Prefetching for **Zero start-up latency**
      - **Catching**: Retrieve missing initial part via dedicated Unicast or Multicast channel
  - Client-Pull or Closed loop paradigm
    - **Batching schemes** with start-up latency
    - Batching schemes with Prefetching for **Zero start-up latency**
      - **Patching**: Retrieve missing initial part via a dedicated Unicast channel or Multicast channel

Scalable Video Distribution

- Multicast distribution tree



Scalable Video Distribution

- Model:
  - Single source *pushes* data via multicast
  - Routers are multicast-capable:
    - Copy and forward
- Challenges
  - Native Multicast Routing not widely deployed
  - Multicast congestion control due to heterogeneity of receivers

Scalable Video Distribution Using P2P

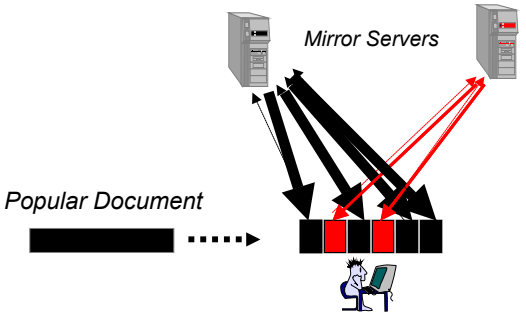
- Splitstream, P2Cast, and others propose to build overlay multicast distribution tree among participating peers
- Is building MC overlay trees a good idea?
  - Peers not as "stable" as routers
    - ➔Multicast tree may frequently get disrupted and must be rebuilt
  - **Peers have lots of storage**
    - ➔Can do "file-and-forward" (D. Cheriton, NGC 2000 keynote)

Scalable Video Distribution Using P2P

- Separate control and data actions:
- New clients needs to do 2 things
  - **Control**: Ask for names of peers close to him that are willing to serve him (can use DHT such as TOPLUS)
  - **Data**: **Pull** data from
    - one client, or
    - multiple clients simultaneously (parallel access)

Scalable Video Distribution

- Parallel-access to stored data [P. Rodriguez, A. Kirpal, and E. W. Biersack. Parallel-Access for Mirror Sites in the Internet. In Proc. Infocom 2000]
  - Speeds-up download times
  - Avoids complex server selection
  - Performs load balancing and increases fault-tolerance



Scalable Video Distribution Using P2P

- Parallel Download of files is implemented today in various tools such as
  - Morpheus, OpenCola, or BitTorrent
- Usefulness of Parallel Download for live video distribution should be further investigated

Summary

- Divide and conquer applied to DHTs
  - Hierarchy and proximity
- Harness the full power of P2P systems ("file-and-forward") for live streaming

Papers at: http://www.eurecom.fr/~btroup/BPublished/bib.html