Magnolia: a novel DHT architecture for keyword-based search

Why Magnolia?

- DHT based systems † big improvement over unstructured systems
  (1) O(log n) routing and lookup
  (2) Effective Load balancing
- Problem † Keyword search difficult because of hashing based lookup

“Seven Innovation Myths” h(title) = 1100100101 “Innovation”

Storage and Lookup

Insert
  Keyword hₜ, SiblingGroup ID
  Locate a sibling node via SIFT

Lookup
  Keyword hₜ
  Replies
  Group Broadcast or Multicast

Incremental Search
  Popularity based ranking
  Boolean Searches
  Multiple Attributes

Analytical Bounds
  Numbers of Nodes visited = O(mⁿ⁺2ᵐ⁺r)
    - 258 nodes for complete query
  Total Traffic Generated = O(mⁿ⁺2ᵐ⁺r)
    - Traffic ~ 268 + r units for complete query

Routing and Lookup = O(1)
  Numbers of Nodes visited = O(mⁿ⁺2ᵐ⁺r) + r
  Traffic ~ 268 + r units for complete query

Good Balancing Properties

- For a Zipf keyword distribution, final load variance can be computed from original distribution variance
- Predicted Theoretical Variance = 124

Question: What is the variance of load over all the nodes ?

V_{orig} = Keyword occurrence variance in input
V_{final} = Load variance over all nodes
k = number of keywords

Simulation Results

DBLP Paper Database: 50,000 paper title insertions
Total keywords = 412,040, Distinct Keywords = 40789

Max load ~ 120 keys
Max load ~ 15500 keys
Max load ~ 1030 keys
Max load ~ 120 keys

Load Variance over all the nodes drops from 27292 to 131

Similar Benefits for traffic distribution, max traffic per node, reply traffic.