Announcements

• HW3 due Friday
• P2 messages
Historical Context

• Research in distributed systems
  – No single point of failure
  – Distribute computation, storage, etc.

• Napster
  – Music file sharing
  – Centralized directory

• How to build distributed version of services like Napster?
P2P Architectures

• No centralized coordination

• Unstructured
  – Gnutella

• Semi/Structured
  – Kazaa
  – Skype
  – Chord/DHTs
Consistent Hashing

• Map keys to nodes
• nodeid = hash(node ip)

• K mapped to successor(k)
• Successor(k) = node equal to or follows K
Consistent Hashing Properties

• Designed for node join/leave with minimal churn in key mapping

• K/N keys per node

• K/N keys change hands during join/leave

• More precise statement in the paper
Locating the Keys

• Infeasible to keep a table of all the nodes
• With smaller table, we won’t cover all the nodes
• With just the successor node pointer, it could take a long time to traverse the circle
Finger Table

• m nodes at exponentially increasing distance starting with the successor

\[ s = \text{successor}(n+2^{i-1}) \text{ where } 1 < l \leq m \]

• More information about nearby key space
• Incomplete information about far away key space
Finger Table

finger[3].interval = [finger[3].start, 1)
Finding successor and predecessor

- How does node 3 find the successor and predecessor to key 1?
  - $\text{Succ}(1) = 1$
  - $\text{Pred}(1) = 0$
- $O(\log N)$ lookups
Node Join and Leave

• When node n joins Chord
  – Initialize predecessor and fingers of n
  – Update predecessors and fingers of existing nodes
  – Transfer keys

• Need external mechanism to find IP address of a Chord node

• Proactive or soft state triggers Leave
Node Failure

• When a node fails, need to update all entries that list that node across the network
• Maintain a small list of successors
• Replicate data on those successors
Load Balance

![Graph showing load balance comparison between two scenarios.](image)

VS

![Graph showing load balance comparison between two scenarios.](image)
Path Length
Lookup Latency (Testbed)

Typical Latency between the sites: 60ms
Improving Chord

• Not all the nodes are equally resourced or reliable
• Make it more aware of network topology and position
Services on top of key/value systems

• Chord is one example
  – Pastry, Kademlia, Tapestry, etc.
• Distributed Hash Table
  – Get/put key, value
• Distributed Storage
  – Traditional filesystem abstraction or key, value storage like Amazon S3
• Mirroring, indexing, etc.