Distributed Hash Tables

CS653, Fall 2010
Implementing insert/retrieve: distributed hash table (DHT)

- Hash table
  - data structure that maps “keys” to “values”
  - essential building block in software systems

- Distributed Hash Table (DHT) similar, but spread across the Internet

- DHT Interface:
  - insert(key, value)
  - lookup(key)

- DHT: overlay (on top of internet) of nodes
  - each DHT node supports single operation:
    - given input key, route messages toward node holding key
DHT in action
DHT in action
DHT in action

Operation: take key as input; route messages to node holding key
DHT in action: insert()

Operation: take key as input; route messages to node holding key
DHT in action: insert()

**Operation**: take key as input; route messages to node holding key
DHT in action: insert()

Operation: take key as input; route messages to node holding key
DHT in action: lookup()

Operation: take key as input; route messages to node holding key
DHT Design Goals

- An “overlay” network with:
  - flexible mapping of keys to physical nodes
  - small network diameter
  - small degree
  - local routing decisions

- A “storage” or “memory” mechanism with
  - best-effort persistence (soft state)
Example DHT: Chord

- m-bit identifier space: $2^m$ ids
- Identifiers ordered on identifier ring (Chord ring) modulo $2^m$
- Each key, $k$, maps to node on ring
  - # key values may be > # nodes
  - key $k$ maps to node on ring with next largest id: successor ($k$)

Example: 10 node Chord with $m=6$ storing 5 keys.
Basic lookup

- node \( i \) forwards to node \( i+1 \), until node holding \( k \) is reached
- memory: \( O(1) \), each node need only know one neighbor
- routing time: \( O(N) \)
Accelerating Lookups

- Lookups accelerated by maintaining additional routing information.
- Each node maintains a routing table with (at most) $m$ entries (where $N=2^m$) called finger table.
- The $i^{th}$ entry in the table at node $n$ contains the identity of the first node, $s$, that succeeds $n$ by at least $2^{i-1}$ on the chord.
- $s = \text{successor}(n + 2^{i-1})$ (all arithmetic mod 2).
Routing by halving distance

- Node uses finger tables to halve the ID space distance to destination in each step
Churn

- **Churn**: nodes join, leave system
  - stored keys must be move to (on leave) or moved from (join) neighboring nodes
  - finger tables must be updated
  - too much churn: too much overhead

- **Joins**: Node n joins by asking any node m for n’s successor

- **Voluntary leaves**
  - Transfer keys to successor, notify predecessor

- **Failure handling**: maintain r backup successors
  - r tuned to the mean-time-to-failure of nodes
DHTs discussion

- What systems can you build using DHTs?
- How to reduce stretch? [Plaxton97]
- How to support range requests or partial matches between request and key?
- What real applications use DHTs today?
  - Why or why not?
- Pros and cons of a centralized index?
Indirection
Indirection:

Indirection: rather than reference an entity directly, reference it ("indirectly) via another entity, which in turn can or will access the original entity.

"Every problem in computer science can be solved by adding another level of indirection"

-- Butler Lampson
Multicast: one sender to many receivers

- **Multicast**: act of sending datagram to multiple receivers with single “transmit” operation

- **Question**: how to achieve multicast

Network multicast

- router actively participate in multicast, making copies of packets as needed and forwarding towards multicast receivers
Internet Multicast Service Model

multicast group concept: use of *indirection*
- host addresses IP datagram to multicast group
- routers forward multicast datagrams to hosts that have “joined” that multicast group
Multicast via Indirection: why?

- Naming and forwarding in IP tailored for point-to-point communication
- Indirection
  - Provides flexible naming
  - Decouples sender from receivers (and their joins and leaves)
Mobility and Indirection

How do you contact a mobile friend?

Consider friend frequently changing addresses, how do you find her?

- search all phone books?
- call her parents?
- expect her to let you know where he/she is?

I wonder where Alice moved to?
Mobility and indirection:

- mobile node moves from network to network
- correspondents want to send packets to mobile node
- two approaches:
  - *indirect routing*: communication from correspondent to mobile goes through home agent, then forwarded to remote
  - *direct routing*: correspondent gets foreign address of mobile, sends directly to mobile
**Mobility: Vocabulary**

**home network**: permanent "home" of mobile (e.g., 128.119.40/24)

**permanent address**: address in home network, *can always* be used to reach mobile (e.g., 128.119.40.186)

**home agent**: entity that will perform mobility functions on behalf of mobile, when mobile is remote

**wide area network**
Mobility: more vocabulary

**permanent address:** remains constant (e.g., 128.119.40.186)

**visited network:** network in which mobile currently resides (e.g., 79.129.13/24)

**care-of-address:** address in visited network. (e.g., 79.129.13.2)

**foreign agent:** entity in visited network that performs mobility functions on behalf of mobile.

**correspondent:** wants to communicate with mobile

**wide area network**
Mobility: registration

End result:
- foreign agent knows about mobile
- home agent knows location of mobile
Mobility via Indirect Routing

home network

home agent intercepts packets, forwards to foreign agent

wide area network

foreign agent receives packets, forwards to mobile

visited network

3

4

correspondent addresses packets using home address of mobile

1

2

mobile replies directly to correspondent

3

4
Indirect Routing: comments

- mobile uses two addresses:
  - **permanent address**: used by correspondent (hence mobile location is *transparent* to correspondent)
  - **care-of-address**: used by home agent to forward datagrams to mobile
- foreign agent functions may be done by mobile itself
- **triangle routing**: correspondent-home-network-mobile
  - inefficient when correspondent, mobile are in same network
Indirect Routing: moving between networks

- suppose mobile user moves to another network
  - registers with new foreign agent
  - new foreign agent registers with home agent
  - home agent updates care-of-address for mobile
  - packets continue to be forwarded to mobile (but with new care-of-address)

- mobility, changing foreign networks transparent: ongoing connections can be maintained!
Mobility via Direct Routing

1. Correspondent requests, receives foreign address of mobile
2. Correspondent forwards to foreign agent
3. Wide area network
4. Foreign agent receives packets, forwards to mobile
5. Mobile replies directly to correspondent
Mobility via Direct Routing: comments

- overcomes triangle routing problem
- non-transparent to correspondent:
  correspondent must get care-of-address from home agent
  - what happens if mobile changes networks?
Mobile IP

- RFC 3220

- Has many features we’ve seen:
  - home agents, foreign agents, foreign-agent registration, care-of-addresses, encapsulation (packet-within-a-packet)

- Three components to standard:
  - agent discovery
  - registration with home agent
  - indirect routing of datagrams
Mobility via indirection: why indirection?

- Transparency to correspondent
- “Mostly” transparent to mobile (except mobile must register with foreign agent)
  - transparent to routers, rest of infrastructure
  - practical concern: if egress filtering is in place in foreign networks (since source IP address of mobile is its home address): spoofing?
An Internet Indirection Infrastructure

Motivation:
- Today’s Internet is built around point-to-point communication abstraction:
  - send packet “p” from host “A” to host “B”
  - one sender, one receiver, at fixed and well-known locations
- … not appropriate for applications that require other communications primitives:
  - multicast (one to many)
  - mobility (one to anywhere)
  - anycast (one to any)
- We’ve seen indirection used to provide these services
  - idea: make indirection a “first-class object”
Internet Indirection Infrastructure (i3)

- Change communication abstraction: instead of point-to-point, exchange packets by name
  - each packet has an identifier ID
  - to receive packet with identifier ID, receiver R stores trigger (ID, R) in network
  - triggers stored in network overlay nodes
Service Model

- **API**
  - `sendPacket(p);`
  - `insertTrigger(t);`
  - `removeTrigger(t); // optional`

- **Best-effort service model (like IP)**
- **Triggers periodically refreshed by end-hosts**
  - Q: what is this approach called?
- **Reliability, congestion control, flow-control implemented at end hosts, and trigger-storing overlay nodes**

packet `p`: (ID, data)
trigger `t`: (ID, addr)
Discussion

- Trigger is similar to routing table entry
- Essentially: application layer publish-subscribe infrastructure
- Unlike IP, end hosts control triggers, i.e., end hosts responsible for setting and maintaining “routing tables”
- Provide support for
  - mobility
  - multicast
  - anycast
  - composable services
Mobility

- Receiver updates its trigger as it moves from one subnet to another
  - mobility transparent to sender
  - location privacy
Mobility

- Receiver updates its trigger as it moves from one subnet to another
  - mobility transparent to sender
  - location privacy
Multicast

- Unifies multicast and unicast abstractions
  - multicast: receivers insert triggers with same ID
- Application naturally moves between multicast and unicast, as needed
  - “impossible” in current IP model
Anycast (cont’d)

- Route to any one in set of receivers
- Receiver $i$ in anycast group inserts same ID, with anycast qualifications
- Route to receiver with best match between $a$ and $s_i$
Composable Services

- Use stack of IDs to encode successive operations to be performed on data (e.g., transcoding)
- Don’t need to configure path between services
Composable Services (cont’d)

- Both receivers and senders can specify operations to be performed on data

```plaintext
send(ID, data)
send((ID_MPEG/JPEG, R), data)
```

![Diagram of communication between sender and receiver](image)
Heterogeneous Multicast

- Both receivers and senders can specify operations to be performed on data
Discussion of I3

- How would receiver signal ACK to sender? what is needed?
- Does many-to-one fit well in this paradigm?
- Security, snooping, information gathering: what are the issues?
- In-network storage to handle disconnection?
Indirection: Summary

We’ve seen indirection used in many ways:
- multicast
- mobility
- Internet indirection

Uses of indirection:
- sender does not need to know receiver id - do not want sender to know intermediary identities
- elegant
- transparency of indirection is important
- performance: is it more efficient?
- security: important open issue for I3