Internet Indirection Infrastructure (i3)

UCL Computer Science

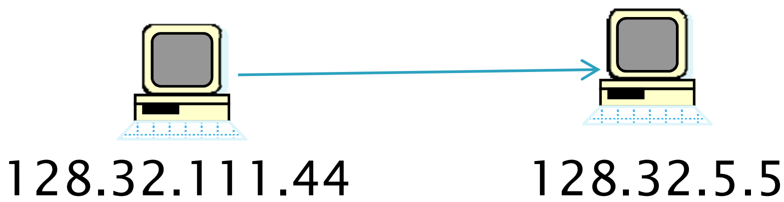
Giotsas Vasileios
Demertzis Foivos
Ntoulas Antonios
Radoi Alexandru

Introduction

- Today's Internet is built around a point-to-point communication abstraction
  - Scalability
  - Efficiency
  - Simplicity
- But…many applications would benefit from a more general communication abstraction:
  - Multicast
  - Anycast
  - Mobility

Introduction (2)

- Point-to-point communication :



  - Known address
  - Fixed location
  - Unicast operation

Introduction (3)

- Multicast, anycast, mobility:
  - Sending host no longer knows identity of receiving host
  - The location of the receiving host need not be fixed

➔ Fundamental mismatch between original point-to-point abstraction and multicast, anycast & mobility.

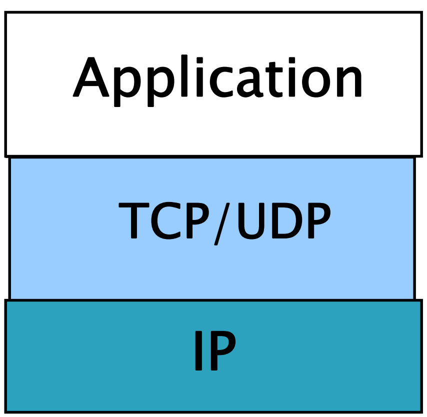
Motivation

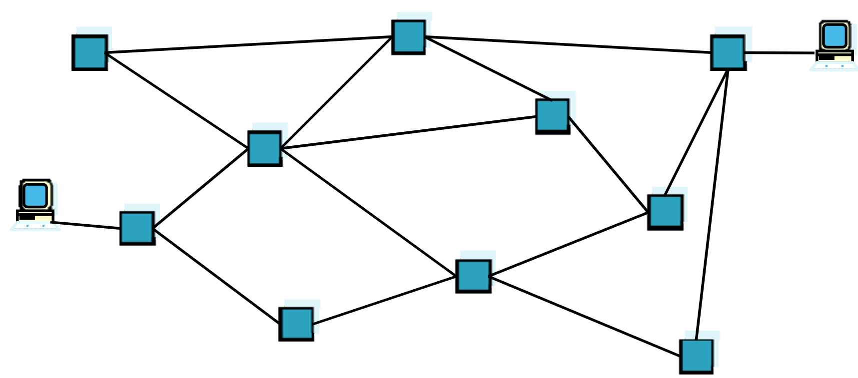
- Need an alternative communication abstraction
  - layer of indirection that decouples the sending hosts from the receiving hosts

- Existent solutions:
  - Network layer: IP multicast, mobile IP
    - Difficult to implement scalability
  - Application layer:
    - Disjoint functionality

i3 solution

- An additional overlay network:
  - On top of IP
    - Best effort service
  - general purpose and flexible rendezvous-based communication abstraction.

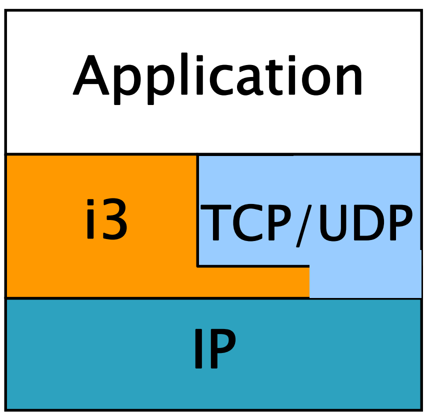




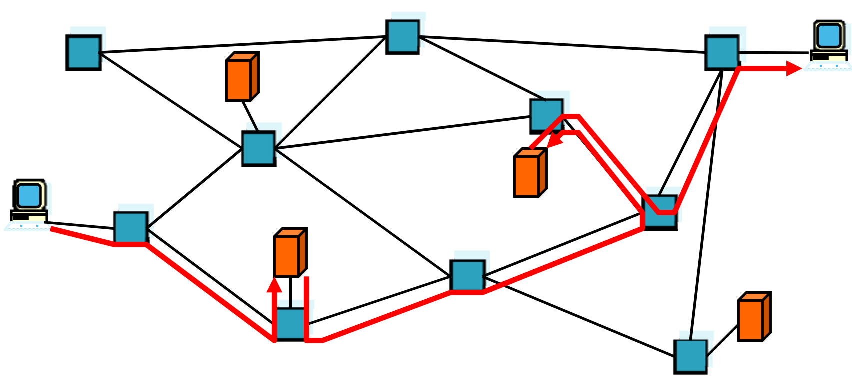
© Stoica, I

i3 solution

- An additional overlay network:
  - On top of IP
    - Best effort service
  - general purpose and flexible rendezvous-based communication abstraction.



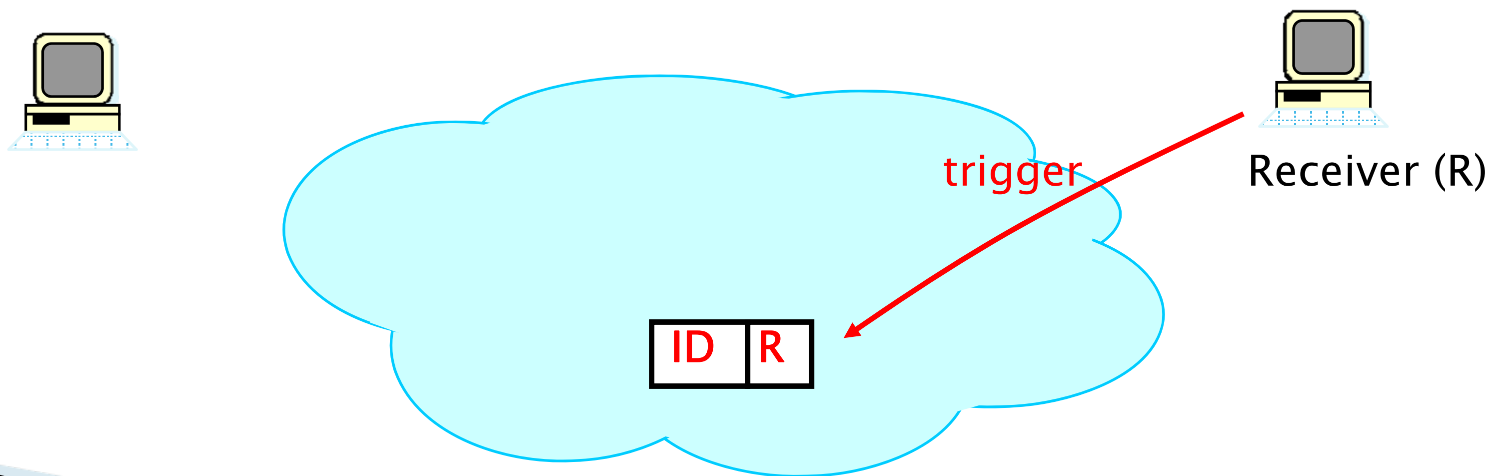
© Stoica, I



© Stoica, I

Rendezvous-Based Communication

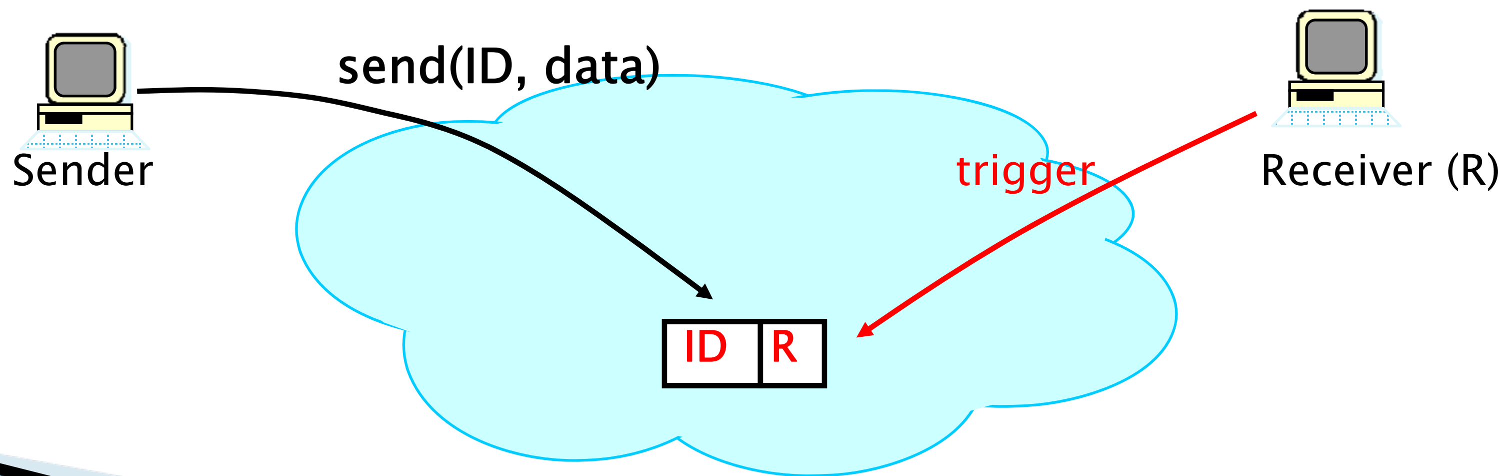
- Receivers use triggers to express their interest in packets
- Trigger (ID,R)
  - ID – Identifies the flow of packets
  - R – Address of the Receiver (usually IP address)



© Stoica, I

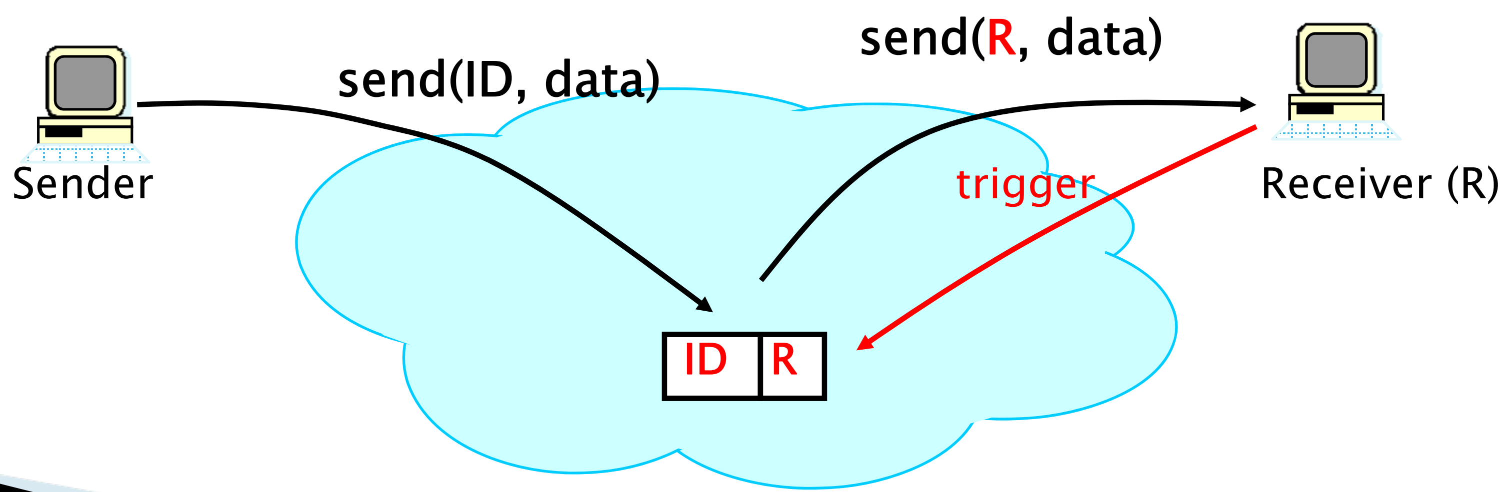
Rendezvous-Based Communication

- Sent packets are pairs of (ID,data)
  - ID – m bit identifier associating with trigger ID
  - Data – payload (usually IP packet payload)



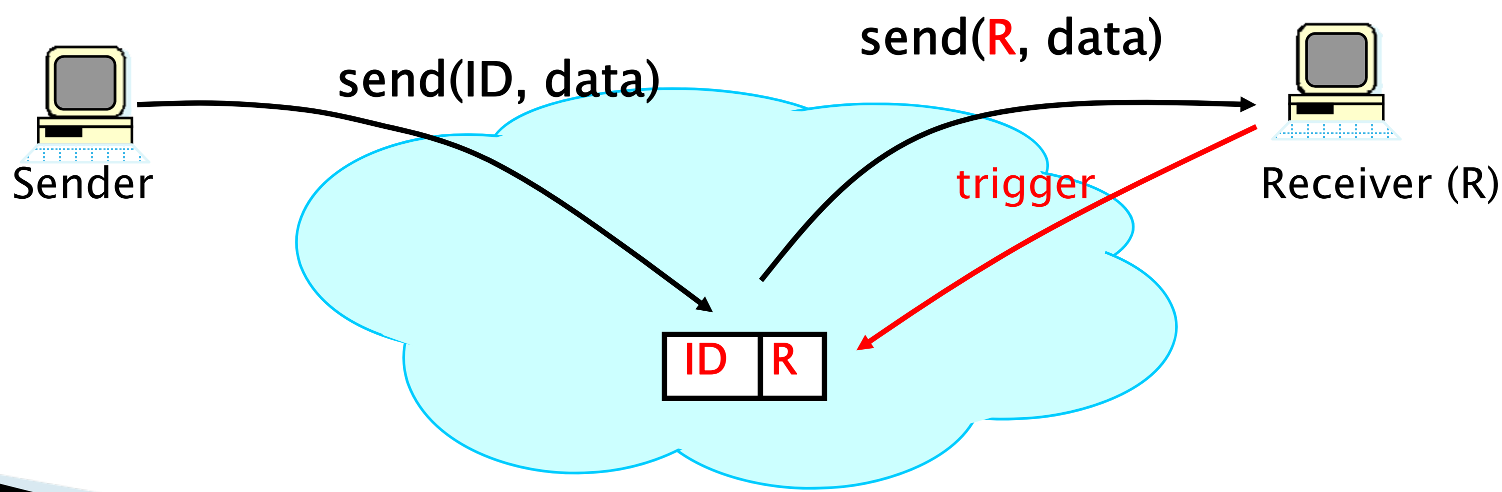
Rendezvous-Based Communication

- A packet (ID, data) will inserted into the overlay network and then forwarded by the i3 infrastructure to the corresponding node identified by trigger (ID,R)
- From there the packet will be forwarded via IP to the receiver



Rendezvous-based Communication

- ID represents the logical rendezvous between the sender's packets and the receiver's trigger
  ➔ Decouples the sender from the receiver



Overview

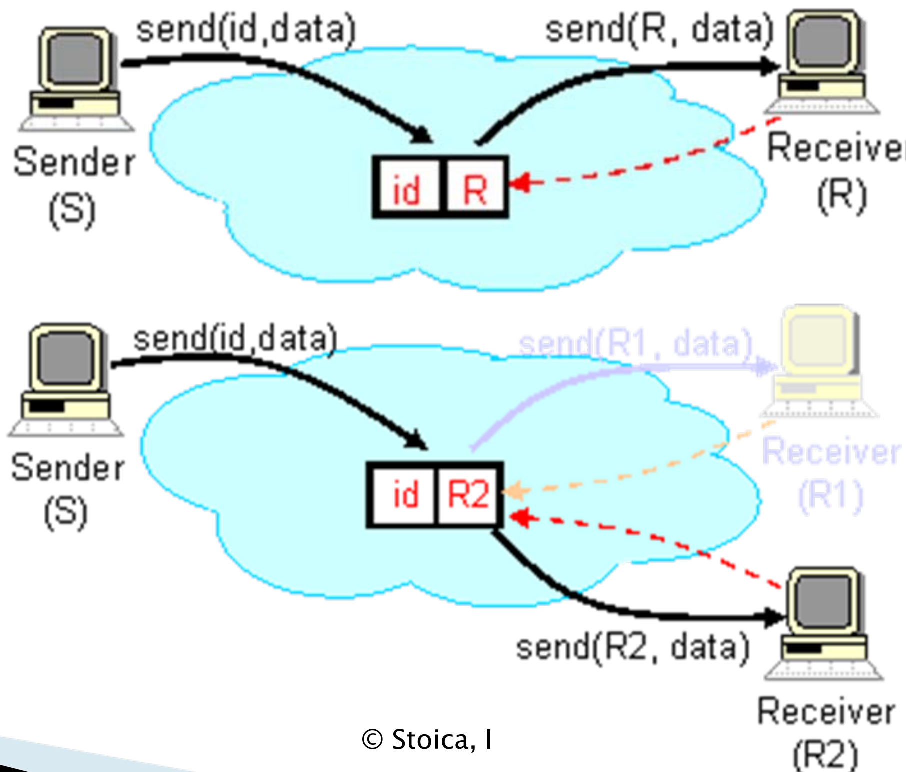
- i3 is an overlay network
    - consists of a set of servers that store triggers and forward packets using IP between i3 nodes and end hosts
    - each identifier is mapped to a unique i3 node

Overview (2)

- When a trigger (ID, R) is inserted it is stored on the i3 node responsible for this ID
- When a packet is inserted into the overlay network, it is routed by i3 to the node responsible for ID
- There it is matched against any triggers for that ID and forwarded (using IP) to all hosts interested in packets sent to that identifier

Mobility

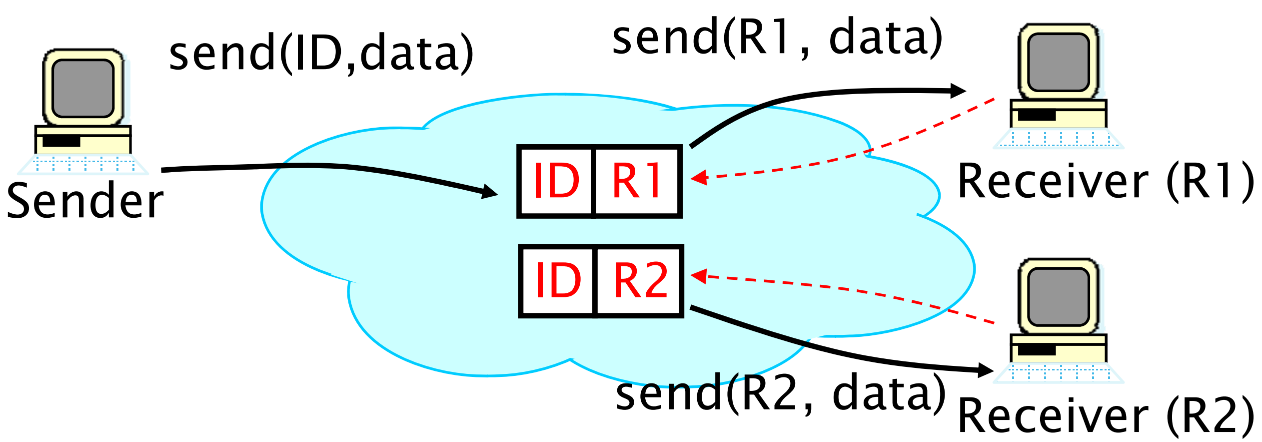
- When a host changes its address, the host needs only to update its trigger



© Stoica, I

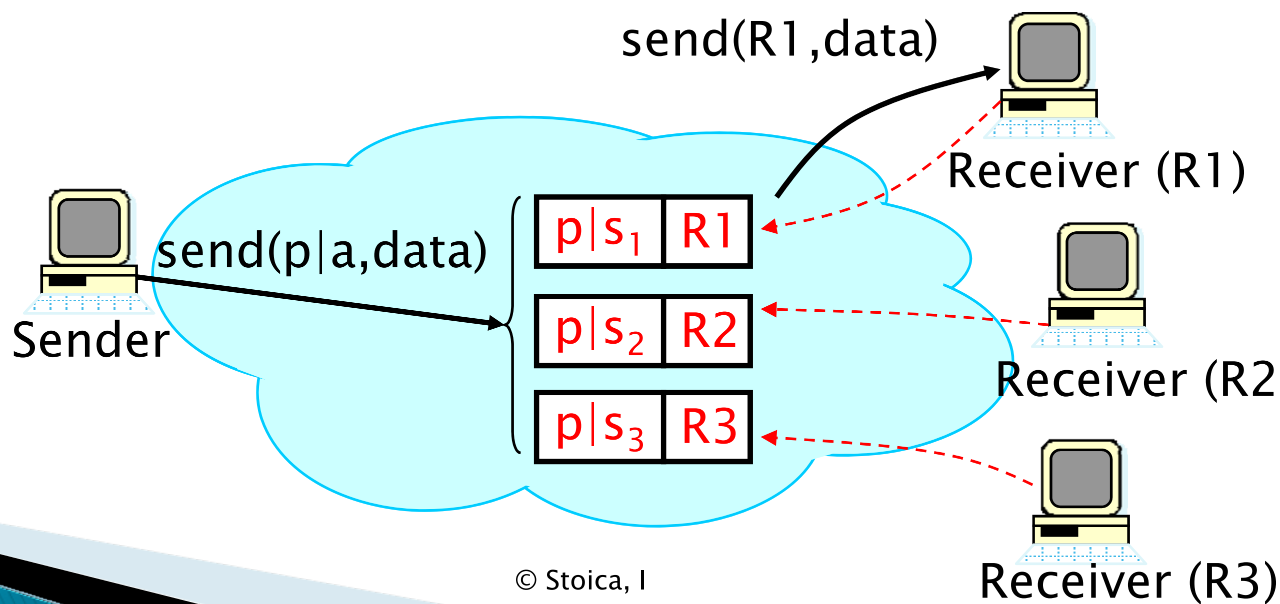
Multicast

- Any packet that matches ID will be forwarded to all the members of the group



Anycast

- Longest prefix matching
- Packet is delivered to a member of a group whose trigger identifier best matches the packet identifier
  - Triggers identifiers share a common prefix *p*

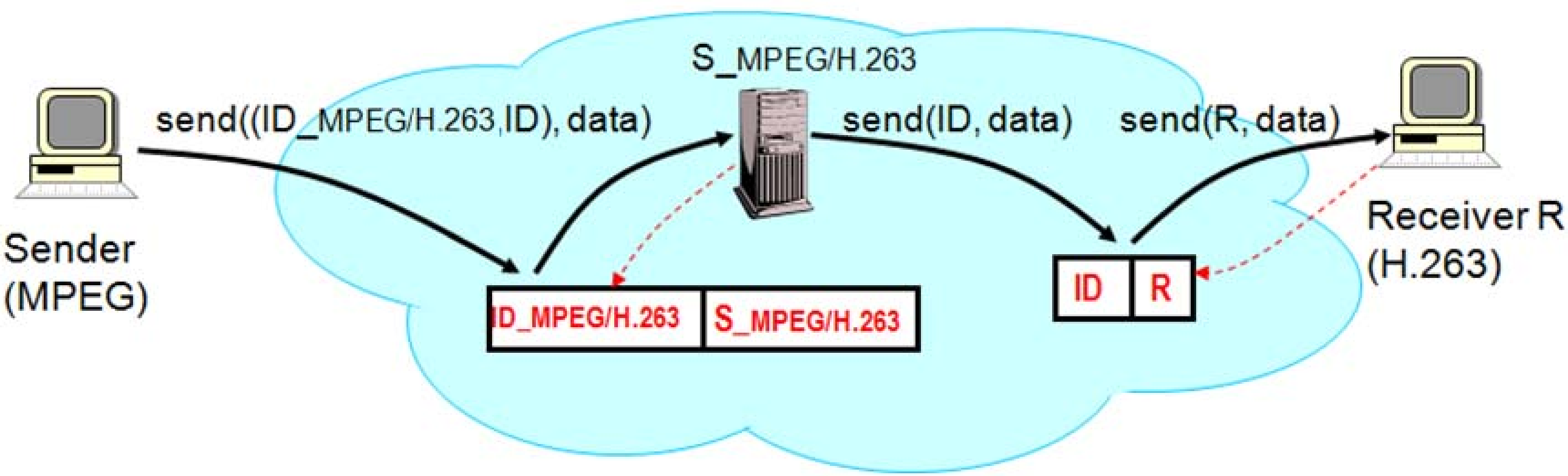


Service Composition

- Stack of identifiers
  - Identifier ID is replaced with a stack of identifiers
    - Packet p = (id_{stack},data)
    - Trigger t = (id, id_{stack})
  - Greater flexibility
  - Packet p is always forwarded based on the first identifier in the stack until it reaches the server storing the matching triggers for p
    - Matching server pops the head of the stack & forwards on the packet

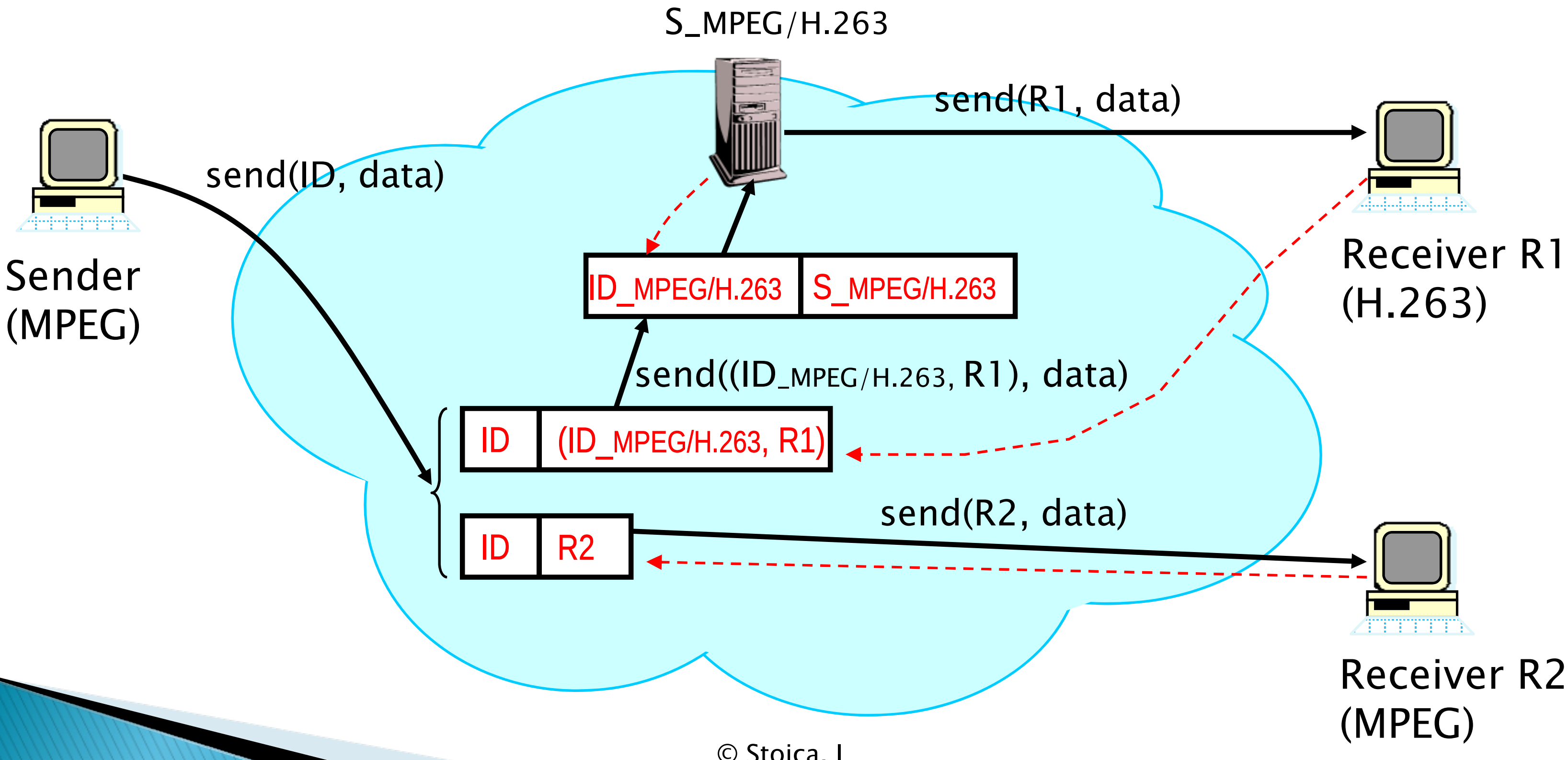
Service Composition (2)

- Some applications may require third parties to process data before it reaches the destination
- Receiver is not aware of data transformations



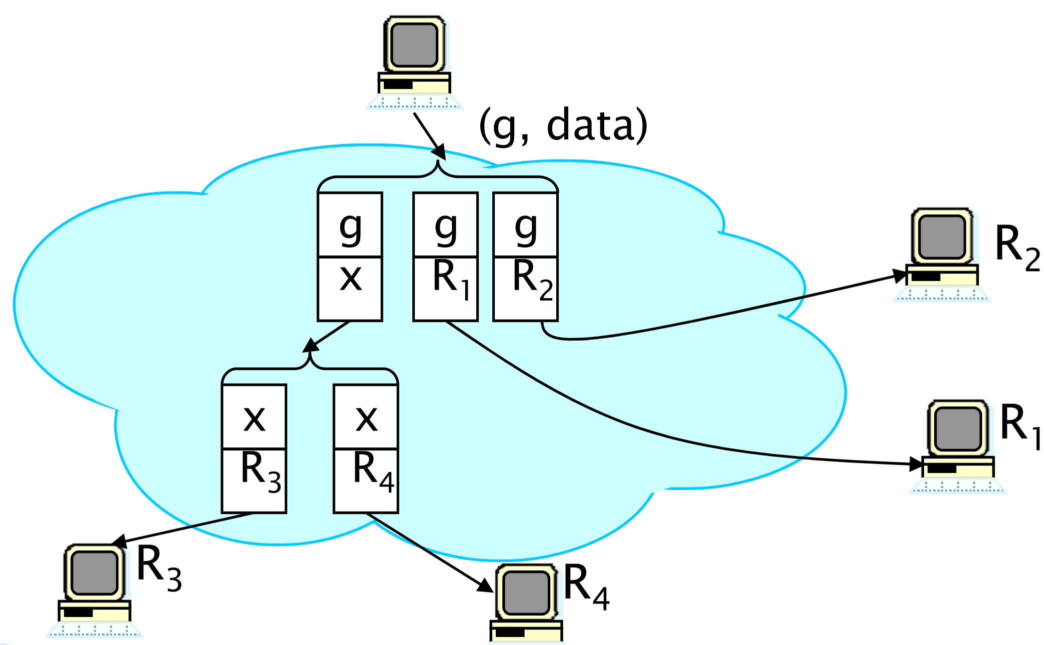
Heterogeneous Multicast

- Sender is not aware of the data transformations



Large Scale Multicast

- The multicast abstraction presented earlier does not scale to large groups
  - Identical identifiers are stored on the same i3 servers
- Use stack identifiers to create a hierarchy



Implementation

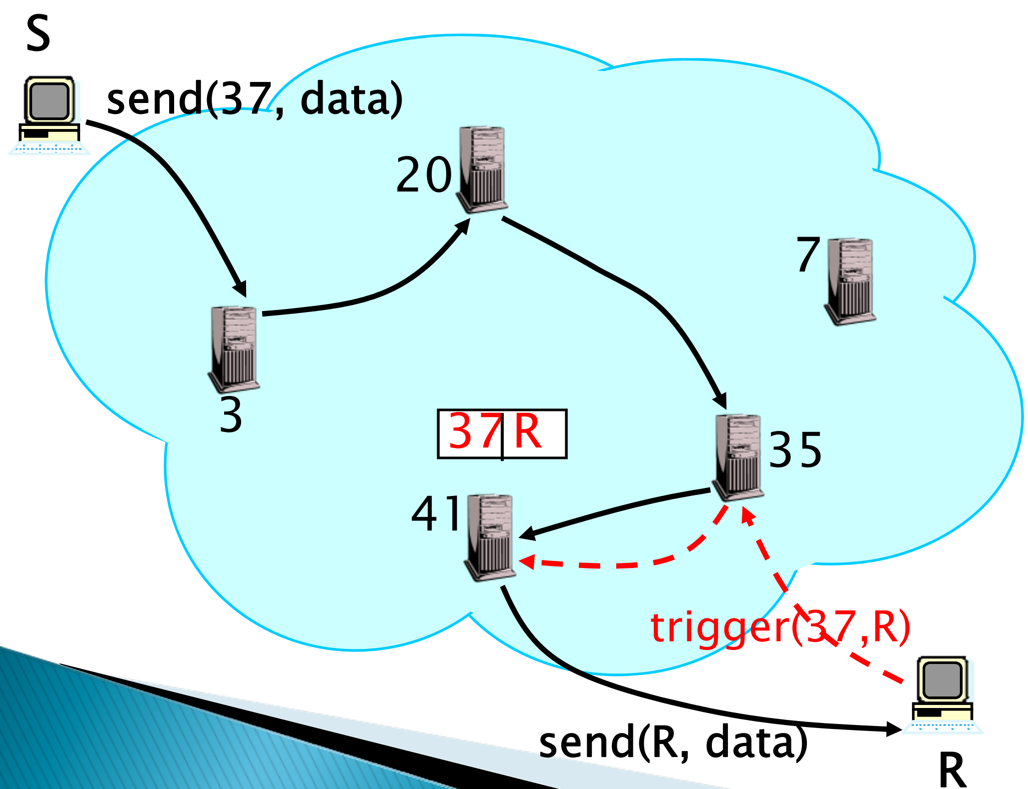
- i3 is implemented on top of Chord
  - circular identifier space
  - Each server has a unique identifier

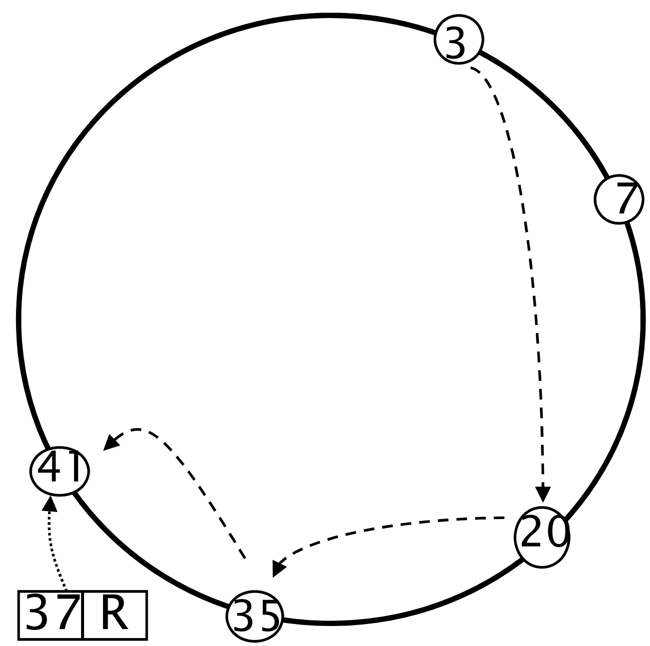
- Each trigger (ID, R) is stored on the node (server) responsible for ID

- Chord routing is responsible for finding the best matching trigger for packet (ID, data)
  - O(log n) hops to locate the responsible server for an arbitrary identifier (n = number of servers)

Design & Implementation (2)

- Receiver knows only node 35, sender knows only node 3
  - End hosts need to know only one i3 node





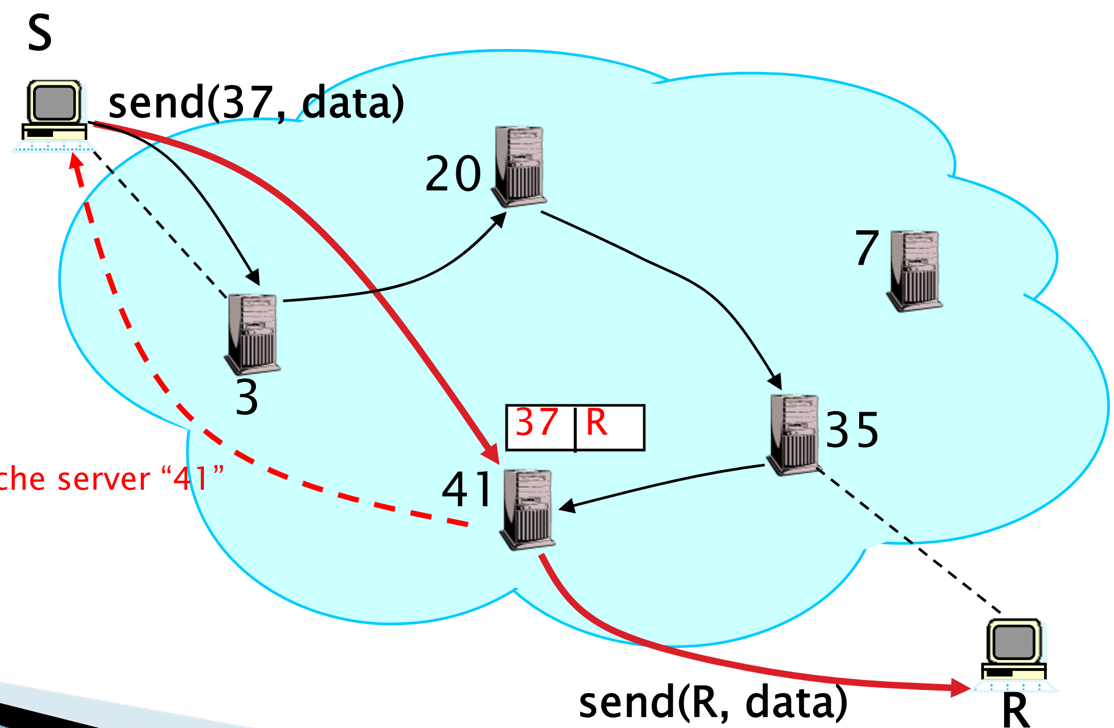
Properties

Inherits properties of the Chord backbone

- Robustness:
  - To prevent server failure (lost triggers)
    - It uses periodic refreshing of triggers
    - Backup triggers
    - Replication of triggers to immediate successors
- Self-Organizing
- Scalable

Properties (2)

- Routing Efficiency:
  - An overlay network is less efficient than direct IP routing
  - Sender caches the i3 server's IP address
  - Send all subsequent packets to that server directly



Properties (3)

- Triangular routing problem
  - Use of public and private triggers
    - Public triggers for initial rendezvous
    - Private triggers used as location aware triggers
- Legacy applications:
  - i3 is best effort → existent UDP applications can work without modifications
  - End hosts run an i3 proxy that translates between UDP and i3
- Anonymity:
  - Eavesdropping on packets will not reveal receiver's address

Security issues

- Eavesdropping by inserting a trigger with the same id as the target
- Solution: Use public & private triggers, also periodically change the private triggers

- Trigger hijacking: a malicious user can alter and remove triggers by knowing the (id,address)
- Solution: Server inserts two triggers, (id,x) and (x,S) instead of (id,S), where x is secret

Security (2)

DoS attacks:

- On end hosts: insert hierarchy of triggers, all of them point to the victim
  - Solution: Challenge the sender of the trigger to verify its originator
- On the infrastructure: create trigger loops, trigger dead-ends, trigger flooding…
  - Solution: Loop detection by sending a random nonce packet and check if it returns
    - Drop public triggers in case of flooding attack

Simulation Results

- Goal: Evaluate Routing efficiency

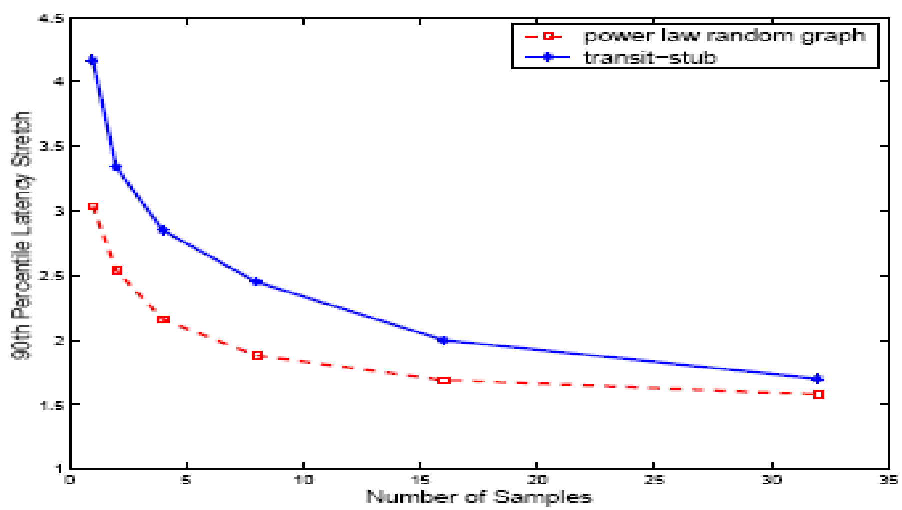
Testbed:

- Two different topologies based on:
  - Power-law random graph
  - Transit-stub
- Delays pre-assigned between links
- 16384 i3-servers

Simulation Results(2)

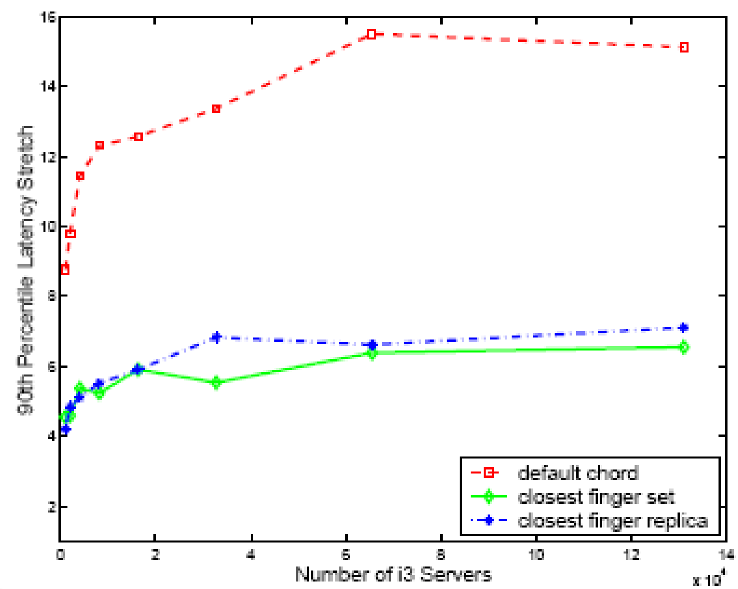
End – to – end latency stretch

- x axis: number of probes to find the closest server
- y axis: The inter-node latency of i3 over the IP counterpart



Simulation Results (3)

- Chord ensures overlay length is O(logN) hops
- Latency though can be quite large depending on the geographical network distance
- Two heuristics to alleviate this problem:
    - Closest finger replica
    - Closest finger set
  - To route a packet, select closest node in terms of network distance.



Simulation Results (4)

Performance:

- 32 nodes over a shared 1 Gbps Ethernet
- 256-bit identifiers
- Trigger insertion: 80,000 triggers/sec
- i3 header for one ID 48 bytes
- Throughput of data forwarding:
  - 35,500 pps (0 byte payload)
  - 23,300 pps (1,400 byte payload); 261 Mbps

Related Work

- Mobile IP
  - Transparently dealing with problems of mobile users
  - Enables hosts to stay connected to the internet, regardless of their location, without needing to change their IP address
  - Similarities
    - it requires no changes to applications/software of non mobile hosts/routers
    - It requires no modifications to IP addressing format
    - Triangle problem constitutes a real issue
    - Security issues – e.g. connection hijacking

Related Work

- Mobile IP (cont.)
  - Differences
    - It does not require additional large scale infrastructure
    - It relies on tunneling rather than creating a whole new protocol layer
    - (Robustness) Home agent failure will lead to collapse of communications
    - Complexity increases when mobile hosts are constantly moving

Related Work

- IP Multicast
  - A method of forwarding IP datagrams to a group of interested receivers
  - Similarities
    - Connectionless service – evidence of deployment only on UDP
    - Security –real concern
    - Best-effort service, so reliability & congestion control are complicated

Related Work

- IP Multicast (cont.)
  - Differences
    - IP network is responsible for routing while in i3, end hosts have more control over routing – provides more flexibility (heterogeneous multicast)
    - Commercially implemented and used for streaming media; however still not widely available
    - Requires changes in the software of network equipment and end hosts (IGMP protocol)
    - Cannot switch on the fly from unicast to multicast
    - State maintained on routers – per flow

Related Work

- IP Anycast
  - Provides anycast operations at the IP layer
  - All the members of the anycast group share the same IP address
  - Similar to i3, nodes & routers do not require any special software/firmware
  - Not transparent to applications – hosts need to be preconfigured to receive packets
  - Unlike i3 that uses application level metrics, packets are sent to the closest host in terms of routing distance

Related Work

- Tuple space
  - Rendezvous based communication – use of tuples
  - Shared memory– distributed system
    - Similar to Publish-Subscribe-Notify as i3
    - Hard to implement on a large scale
    - Nodes explicitly ask for data packets – low speed communication
    - Matching operations – more powerful than longest prefix match
    - Cannot perform service composition

Related Work

- FARA
- Active Networks
- Intentional Naming System (INS)
- MPLS
- …

Critical appraisal

- Based on Chord – shares all the advantages and weaknesses of it
  + Robustness, Efficiency & Scalability
  – Network partition, SHA-1 proven to have collisions

- Implementation – overlay network
  - Real benefits
    - No state needs to be stored by network equipment
    - Incremental deployment
    - Application transparency (unicast, multicast, anycast & mobility) – Provides abstraction for communication
  - Disadvantages
    - difficult to deploy (complexity and cost)

Critical appraisal

- Not clear how TCP communication would work
  - How to initiate and maintain a TCP session in i3?
  - Packet IDs that a sender sent on a server identify a particular flow
    - What happens to the TCP session when more receivers join?
    - Flow control and congestion control?

- i3 overlay requires that all ids that share their first k-bits of the identifier be stored on the same i3 server
  - For load balancing ids are split between multiple servers
  - It is not clear how the routing works in this case

Critical appraisal

- Despite use of heuristics to improve routing efficiency, latency inflation still exists – difficult to provide low latency for end node
  - Limitation for time sensitive applications like streaming and other multimedia applications
- Probing the network and comparing the RTTs
  - Not realistic – increase network load

Critical appraisal

- Mobility
  - \+ Rendezvous based communication
    - Allows simultaneous mobility for both sender & receiver
  - – When a client moves, some outstanding packets could still be routed to the old IP
    - Client could lose these packets
    - What if another client will connect to the old IP address
      - Data integrity violation

Critical appraisal

- Security – major flows
  - Reliant on the i3 infrastructure
    - Introduces a lot of new vulnerabilities
    - What if an i3 node gets compromised?
    - Use of private triggers to prevent eavesdropping
      - Malicious user can just eavesdrop the initial packet exchange where private triggers are inserted into the network
      - Use of public key cryptography to exchange private triggers – increases the complexity even more

Critical appraisal

- Security – more major flows
  - Preventing DoS – solutions proposed are naïve
    - Challenging every sender when hierarchy of triggers is inserted in the overlay network
      - This could only aggravate the DoS attack by making the i3 node do even more work
    - Loop detection – send random packet and see if returns
      - If this is performed for every new chain of triggers inserted it could take forever – what if I just joined a multicast VoIP conference?

Critical appraisal

- Simulation
  - Provided results of the implementation of the overlay network – evaluated only point to point communication
    - The focus of the paper is on multicast, anycast and mobility – no evidence of evaluating in the simulation
  - Other useful communication models that were not even considered to be evaluated
    - Triangular routing problem
    - Node failures
    - Use of trigger chaining
  - No result comparison to other models (mobile IP, IP multicast)

Critical appraisal

- Brilliant idea!!!

- Might work as long as
  - Nobody is going to use it
  - Someone, somewhere is going to pay for deploying it

References:

Ion Stoica, Robert Morris, David Karger, M. Frans Kaashoek, and Hari Balakrishnan, Chord: A Scalable Peer-to-peer Lookup Service for Internet Applications, ACM SIGCOMM 2001, San Deigo, CA, August 2001, pp. 149-160

Ion Stoica, Robert Morris, David Liben-Nowell, David R. Karger, M. Frans Kaashoek, Frank Dabek, Hari Balakrishnan, Chord: A Scalable Peer-to-peer Lookup Protocol for Internet Applications.

Ion Stoica, Daniel Adkins, Shelley Zhuang, Scott Shenker, Sonesh Surana, "Internet Indirection Infrastructure," Proceedings of ACM SIGCOMM, August, 2002

CARRIERO, N. The Implementation of Tuple Space Machines. PhD thesis, Yale University, 1987.

DABEK, F., KAASHOEK, F., KARGER, D., MORRIS, R., AND STOICA, I. Wide-area cooperative storage with cfs. In Proc. ACM SOSP'01 (Banff, Canada, 2001), pp. 202-215.

WETHERALL, D. Active network vision and reality: lessons form a capsule-based system. In Proc. of the 17th ACM Symposium on Operating System Principles (SOSP'99) (Kiawah Island, SC, Nov. 1999), pp. 64-79.

"IP Multicast". Network World. © 2006. http://www.networkworld.com/details/502.html

"Internet Protocol (IP) Multicast". Cisco Systems © 2006. http://www.cisco.com/univercd/cc/td/doc/cisintwk/ito_doc/ipmulti.htm

"Mobile IP". Cisco System. © 2007. http://cisco.com/en/US/docs/ios/12_0t/12_0t1/feature/guide/MobileIP.html

"Introduction to Mobile IP". Golden G. Richard III, Ph.D. University of New Orleans. www.cs.uno.edu/~golden/mobile_ip/mobile_ip.PPT

"Case Study: Tuple Space". Ian Foster. © 1995. http://www.wotug.org/parallel/books/addison-wesley/dbpp/text/node44.html

"INS: Intentional Naming System". M. I. T. Computer Science and Artificial Intelligence Laboratory. http://nms.csail.mit.edu/projects/ins/

"Deploying IP Anycast". Kevin Miller. Carnegie Mellon Network Group. © 2003. www.net.cmu.edu/pres/anycast/Deploying%20IP%20Anycast.ppt