CSF641 – P2P Computing

點對點計算

Routing on P2P Overlay Networks

吳俊興

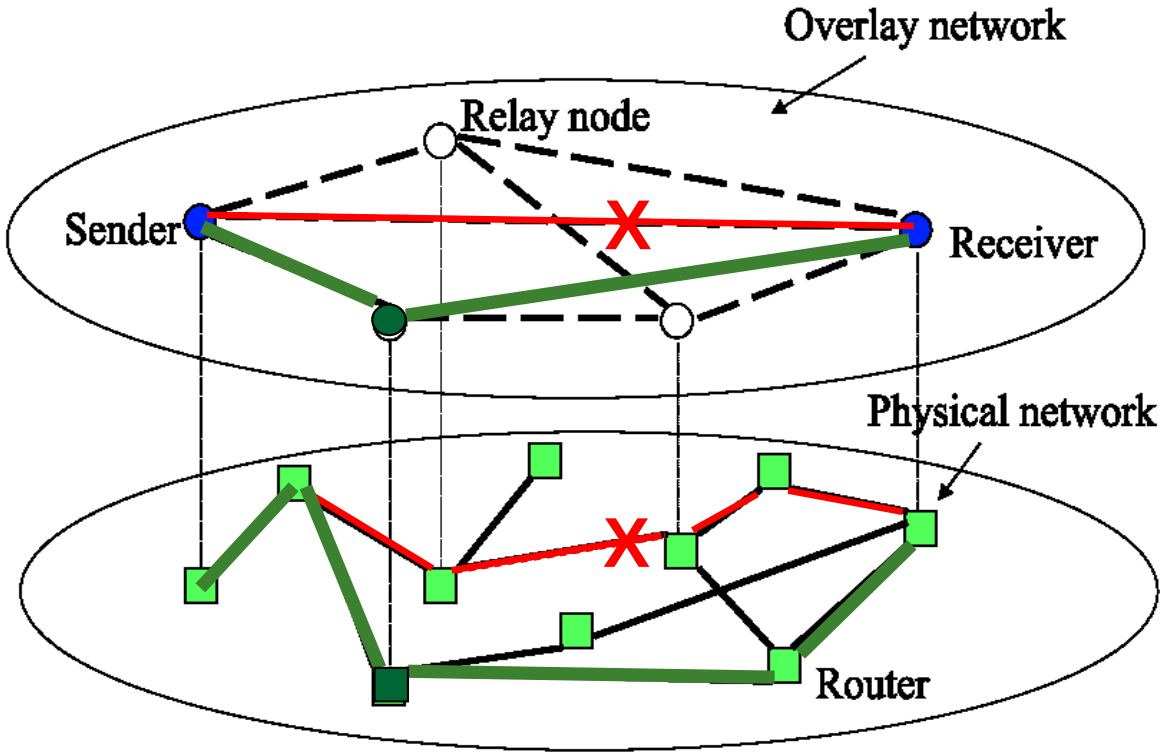
國立高雄大學 資訊工程學系

Outline

- Introduction to Overlay Networks
  - Churn in P2P Networks
- Proximity Routing: Exploiting Network Proximity
  - Geographic Layout
  - Proximity Routing
  - Proximity Neighbor Selection
- Resilient Routing: Exploiting Alternate Route
  - Detour
  - Resilient Overlay Networks (RON)
  - Path Diversity with Forward Error Correction (PDF)
  - Path Probe Relay Routing (PPRR)

Overlay Networks

- A network on top of another networks
  - Links on one layer are network segments of lower layers



Forward over a fast/reliable alternate route

Make the application control the routing

P2P and Overlay Networks

- The peering connections of P2P nodes form a kind of overlay networks on top of the IP network
  - Overlay node: peer of the P2P system
    - Overlay network address: peer's GUID
  - Overlay link: (TCP or UDP) connection between two nodes
- Users of a P2P network may not be directly connected to the overlay nodes
  - E.g. super-node or hierarchical designs
  - Open v.s. private overlay networks

Characteristics of P2P Overlay Networks

- Tens or thousands of peers may join or leave simultaneously
  - **Network churn**, dynamics of peer participation, makes the overlay network changed continuously
- A peer usually has tens of overlay links that are changed dynamically
  - **Link transience**: the overlay link between any two nodes may be established or disconnected arbitrarily

Characteristics of P2P Overlay Links

- One virtual hop may be many underlying hops away
  - Latency and cost vary significantly over the virtual overlay links
  - An overlay packet traveling a multiple-hop path to its destination peer may appear several times on an underlying link
  - Delay for one-hop path may be longer than two-hop path
- The connections for a peer node behind firewalls or NAT may be uni-directed
  - The underlying IP network usually induces a complete graph of connectivity

Challenges to P2P Overlay Routing

Node and ***link*** states of peers change more dynamically than those of IP routers. These make overlay routing become more difficult than traditional IP routing!

- ***Proximity Routing***: establish or choose a *better* overlay link or path to the destination peer
  - Performance metrics: latency, packet loss rate, bandwidth, load-balance
  - Related topics: topology-aware, network-aware, path-aware, link-aware routing
- ***Resilient Routing***: support efficient routing in the presence of network churn and dynamic link change
  - Related topics: fault-tolerant routing

Churn in P2P Networks

- There may be dozens of membership changes simultaneously
- Peers may crash without notice
- Depends on applications

Handling Churn in a DHT

Three common approaches

- Recovering from failures
- Routing around suspected failures
- Proximity neighbor selection

I. Recovering From Failures

- For correctness, maintain leaf set during churn
  - Also routing table, but not needed for correctness
- The Basics
  - Ping new nodes before adding them
  - Periodically ping neighbors
  - Remove nodes that don't respond
- Simple algorithm
  - After every change in leaf set, send to all neighbors
  - Called *reactive* recovery

II. Routing Around Failures

- Being conservative increases latency
  - Original next hop may have left network forever
  - Don't want to stall lookups
- DHT has many possible routes
  - But retrying too soon leads to packet explosion
- Goal:
  1. Know for sure that packet is lost
  2. Then resend along different path

III. Proximity Neighbor Selection (PNS)

- For each neighbor, may be many candidates
  - Choosing closest with right prefix called PNS
  - One of the most researched areas in DHTs
  - Can we achieve good PNS under churn?
- Remember:
  - leaf set for correctness
  - routing table for efficiency?
- Insight: extend this philosophy
  - Any routing table gives O(log N) lookup hops
  - Treat PNS as an optimization only
  - Find close neighbors by simple random sampling

Exploiting Network Proximity

Three basic approaches for exploiting proximity in structured DHT protocols

1. Geographic Layout
2. (Overlay) Proximity Routing
3. Proximity Neighbor Selection
   - Incurs only a modest additional overhead for organizing and maintaining the overlay network

1. Geographic Layout

- Topology-based nodeId assignment
- The nodeIDs are assigned in a manner that ensures that nodes that are close in the network topology are close in the nodeID space
- CAN

2. Proximity Routing

- The routing tables are built without taking network proximity into account
- The routing algorithm chooses a nearby node at each hop from among the ones in the routing table
  - a balance between making progress towards the destination in the nodeId space and choosing the closest routing table entry
  - Chord, CAN

3. Proximity Neighbor Selection

- Routing table construction takes network proximity into account
- Routing table entries are chosen to refer to nodes that are nearby in the network topology, among all live nodes with appropriate nodeIds
  - The distance traveled by messages can be minimized without an increase in the number of routing hops
  - Tapestry, Pastry

Resilient Routing: Examples

Exploit the alternate routes

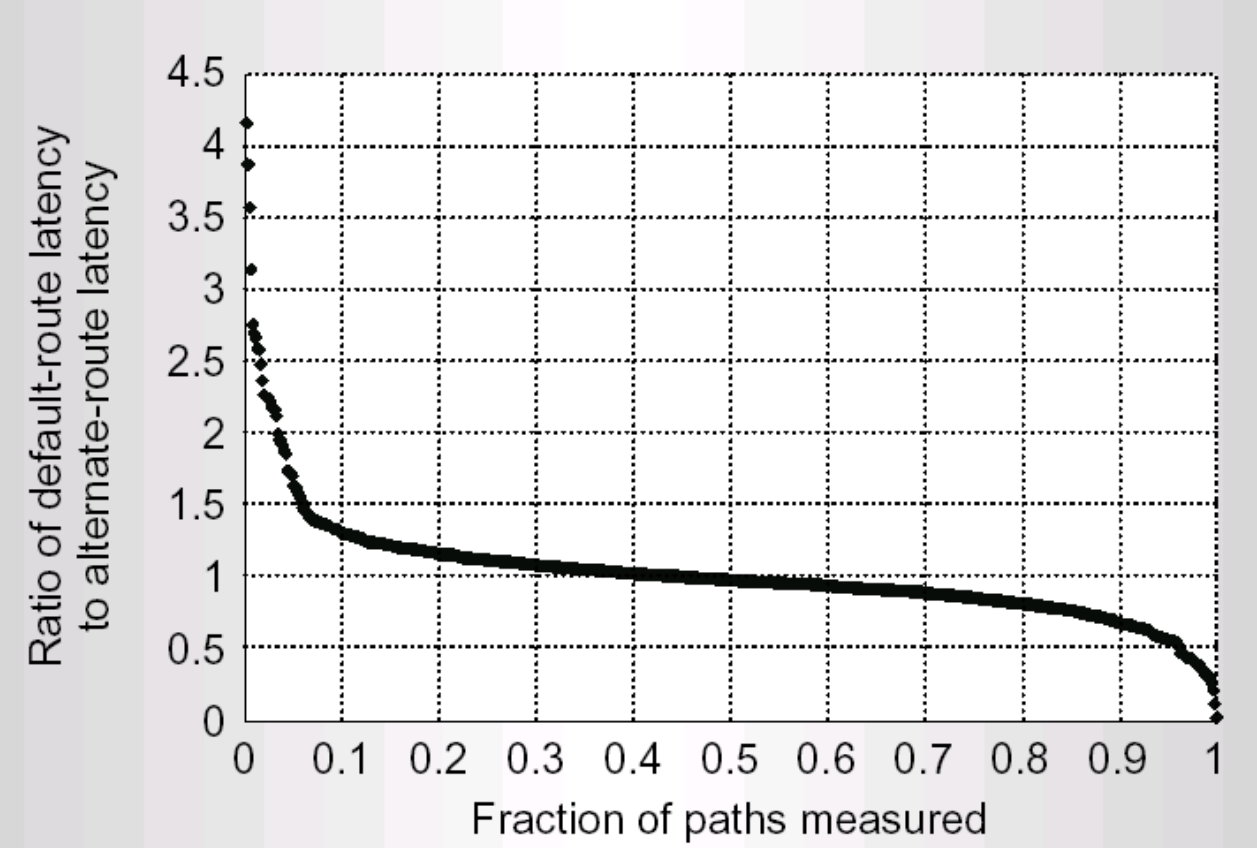
- Detour
- Resilient Overlay Networks (RON)
- Path Diversity with Forward Error Correction (PDF)
- Path Probe Relay Routing (PPRR)

Detour: Introduction

- Inefficiencies in routing and transport protocols in modern Internet
  - Call a route between two hosts inefficient when there is some alternate route with superior latency or packet drop rate
- Detour is a virtual Internet, in which routers “tunnel” packets over the commodity Internet
  - Intelligent routing and congestion control
- Collected a full set of pair-wiser latency and drop-rate measurements
  - Used 43 publicly available servers running the traceroute program
  - Over the course of 35 days
  - Randomly distributed the time intervals between these requests, with a mean of 15 minutes per host

Detour: Latency Measurement

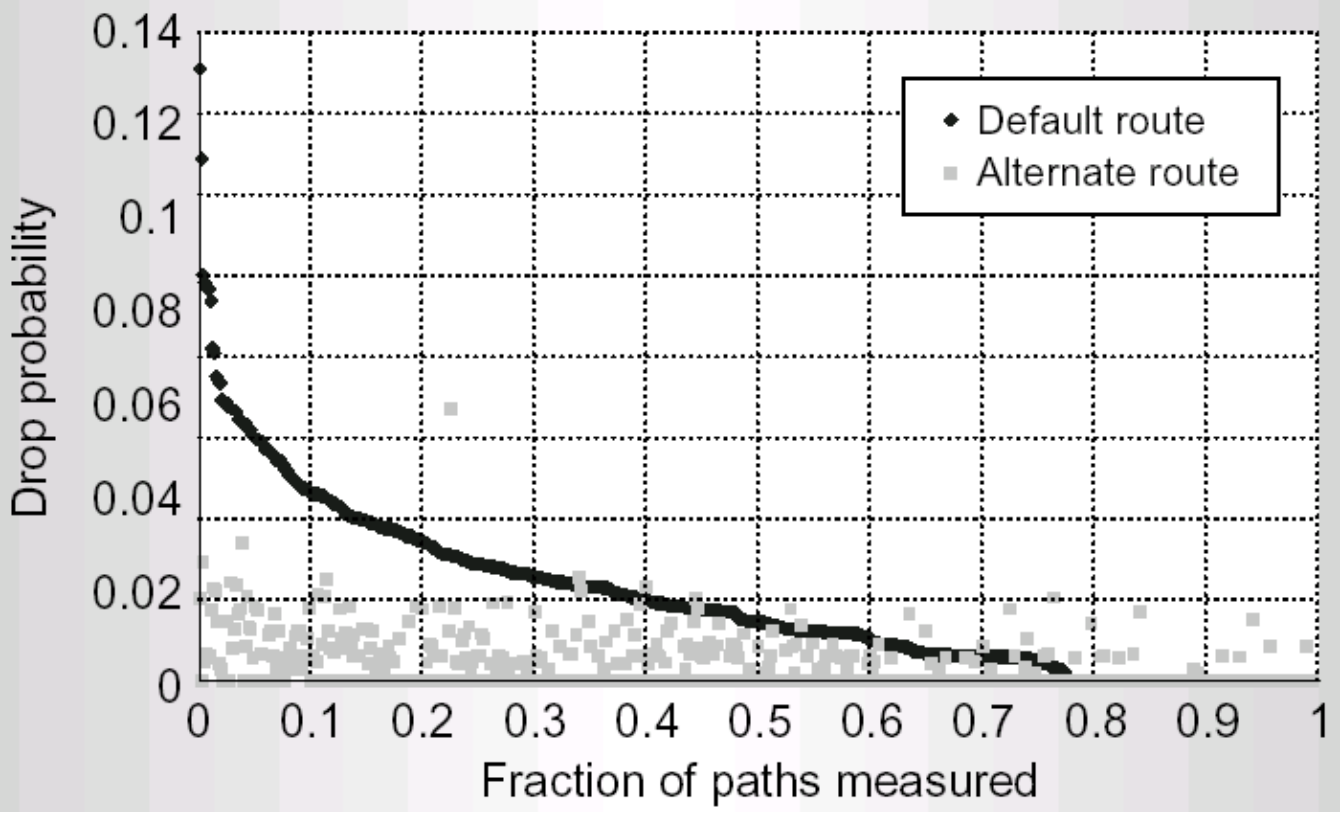
Ratio of best-alternate-route to default-route latency



- Half of the paths measured, there is a faster route
- For 15% of paths, there is an alternative that offers an improvement in latency better than 25%
- For more than 15% of paths, our alternate route choice will shave at least 25 ms from the RTT

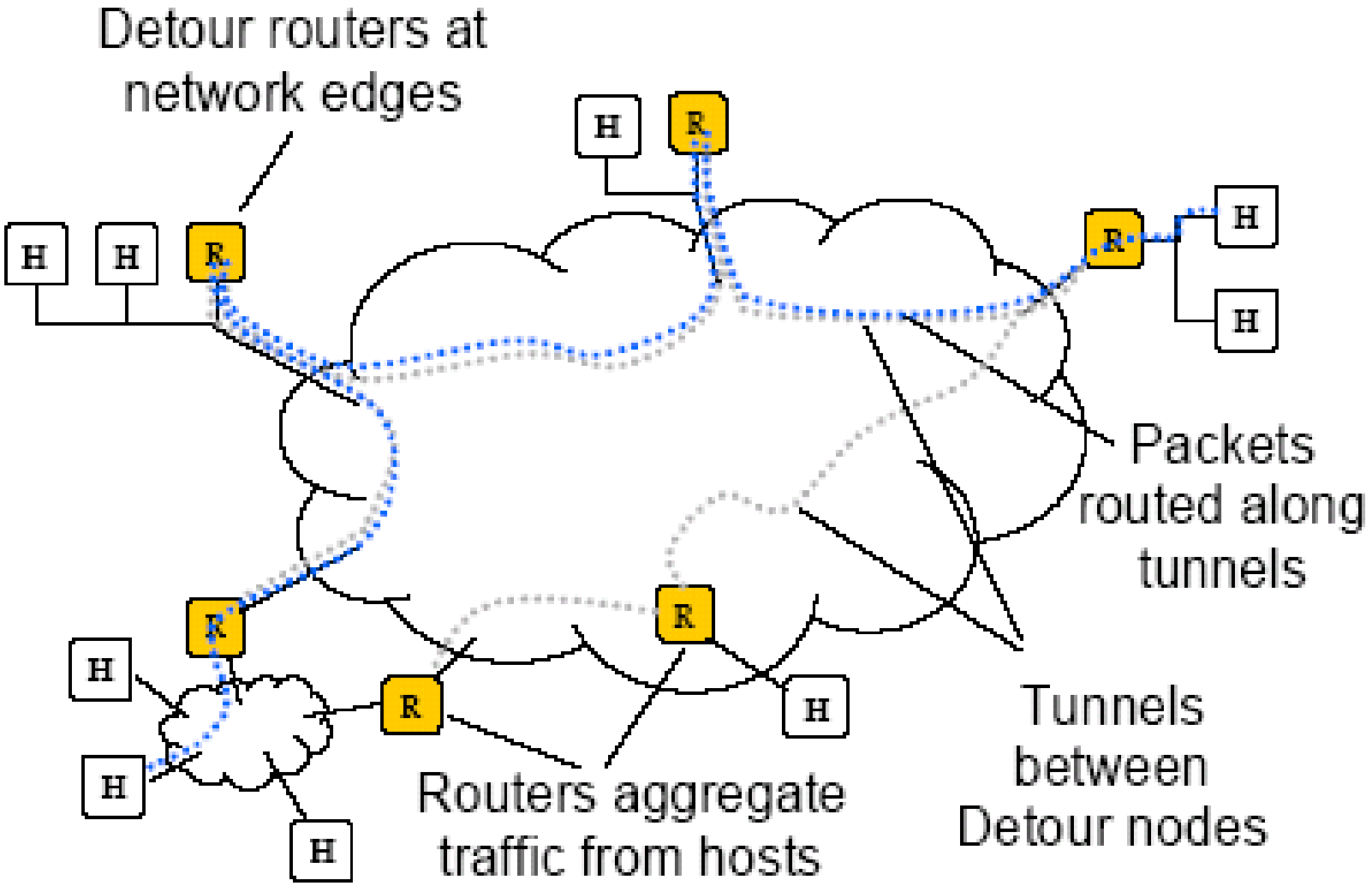
Detour: Drop-rate Measurement:

Average drop rates for default routes vs. best-alternative routes



- For almost 80% of the paths, an alternate route offers a lower probability of dropping packets
- In almost 50% of the paths, the improvement is a factor of six or better

Detour: Architecture



Detour: Summary

- Opportunities in Routing
  - Detour routers can exchange information about the measured latency, drop rate, and bandwidth available along their tunnels
  - Fluctuations
  - Dynamic multi-path routing
    - Hop to automatically balance loads in their system and avoid congestion before it occurs
  - Specialize routing decisions to the needs of different service classes
- Opportunities in Informed Transport
  - A detour router at the network edge can observe many different flows
    - For connection establishment
      - A Detour router can provide an informed round-trip time estimate for subsequent hosts using that path
      - The router may choose to retransmit the connection establishment request on the host's behalf
    - For slow-start
      - It could reduce or avoid these packet drops and consequent retransmits if the host know its fair share of the bottleneck

Resilient Overlay Networks: Introduction

- Motivation
  - Remedy today's IP routing problems
  - Distributed application-layer overlay
  - Cooperate to forward data
  - Exploiting redundancy in underlying Internet
- General Steps
  - Measure all links between nodes
  - Compute path properties
  - Determine best route
  - Forward traffic over that path

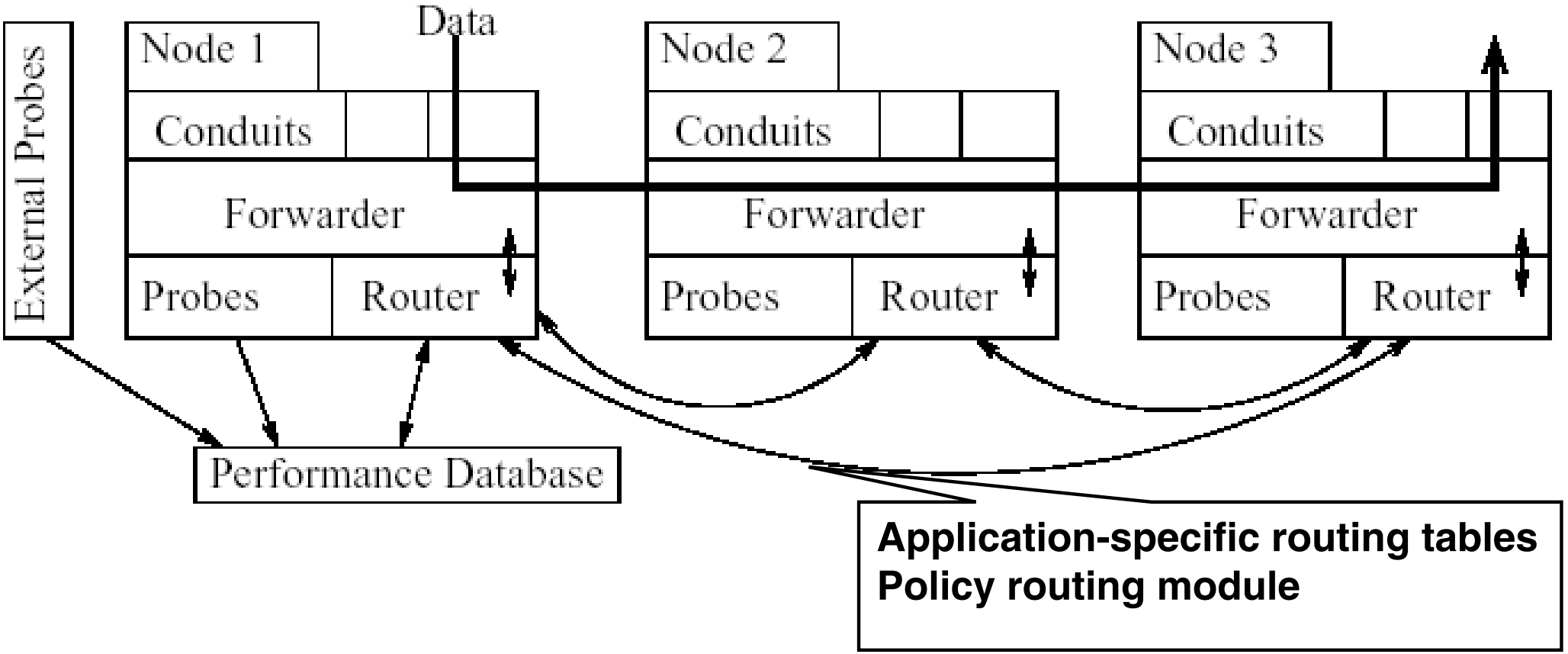
RON: Overview

- RON nodes can communicate with each other
  - Exchange information about quality of the paths via a routing protocol
  - Build forwarding tables based on path metrics
    - latency, packet loss rate, available throughput
  - Each RON node obtains the path metrics
    - active probing experiments, passive observations
  - Designed to be limited in size
- Integrate routing and path selection with distributed applications more tightly
  - The ability to consult application-specific metrics in selecting paths
  - The ability to incorporate application-specific notions of what network conditions constitute a “fault”
- Provide a framework for the implementation of expressive routing policies, which govern the choice of paths in network
  - Packet classifying
  - Forwarding rate controls

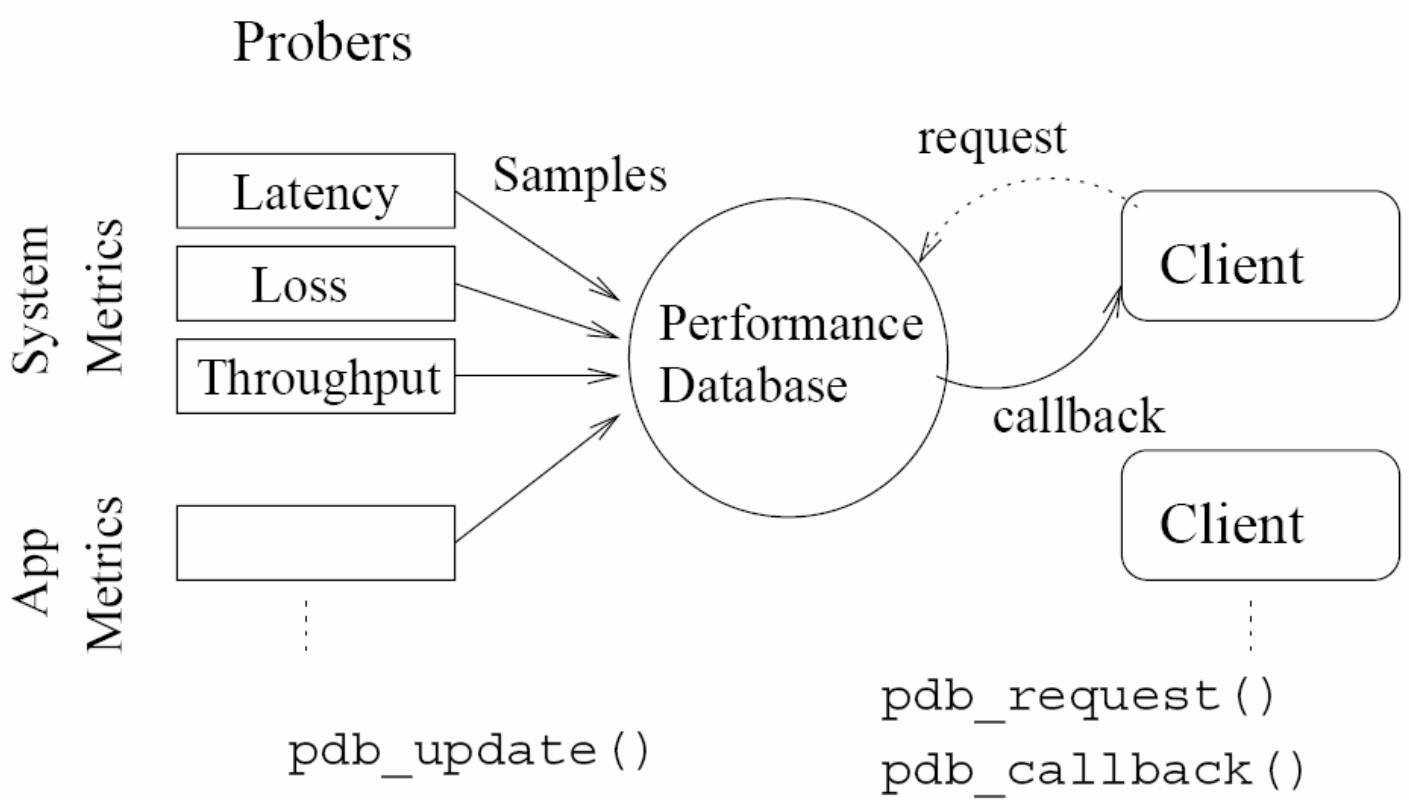
RON: Routing and Path Selection

- The entry node *tags* the packet's RON header
  - support multi-hop routing
  - tie a packet flow to a chosen path
- The small size of a RON
  - Maintain information for each virtual link
    - (i) latency, (ii) packet loss rate,(iii) throughput
  - Select the path that best suits the RON client
- Link-state dissemination
  - The default RON router uses a link-state routing protocol to disseminate topology information between routers
  - Information is sent via the RON forwarding mesh itself
  - Thus, the RON routing protocol is itself a RON client
- Path evaluation and selection
  - Every RON router implements *outage detection*
    - uses an active probing mechanism for this.
  - Every RON router implements three different routing metrics:
    - latency-minimizer, loss-minimizer, TCP throughput-optimizer

RON: Design and Implementation



RON: Performance Database



- PROBE_INTERVAL: 12 seconds
- PROBE_TIMEOUT: 3 seconds
- ROUTING_INTERVAL: 14 seconds

RON: Major Results

- RON was able to successfully detect and recover from 100% (in RON1) and 60% (in RON2) of all complete outages and all periods of sustained high loss rates of 30% or more
  - RON1: 12 nodes; RON2: 16 nodes
- RON takes 18 seconds, on average, to route around a failure & can do so in face of flooding attack
- RON successfully routed around bad throughput failures, doubling TCP throughput in 5% of all samples
- In 5% of the samples, RON reduced the loss probability by 0.05 or more
- Single-hop route indirection captured the majority of benefits in our RON deployment, for both outage recovery and latency optimization

RON: Summary

- Improved availability of Internet communication paths using small overlays
  - Layered above scalable IP substrate
  - RON provides a set of libraries and programs to facilitate this application-specific routing
- Experimental data suggest that approach works
  - Outage detection and recovery in about 15 seconds
  - Able to route around certain denial-of-service attacks
  - Performance
  - Routing behavior

PDF: Introduction

Path Diversity with Forward Error Correction (PDF) System for Packet Switched Networks

- Video streaming is a delay sensitive application
- Most schemes assume a single fixed path between the receiver and the sender throughout the session
- If congestion happens along that path, video suffers from high loss rate and jitter

PDF: Overview

Send packets simultaneously over ***multiple disjoint*** paths

- Executes traceroute to obtain underlying network information before setting up a communication channel
  - Find disjoint paths
- Sends packet with Forward Error Correction (FEC)
  - Reduce delay due to retransmission
  - Expense of bandwidth expansion

PDF: Setup a Communication Channel

1. Sender executes traceroute from itself to all relay nodes and receiver
2. Link latencies and router names are obtained
3. Sender instructs relay nodes to execute traceroute from themselves to receiver
4. Send the path information back to the sender
5. The sender, based on the information received , select the redundant path
6. Sender sends the setup packet to that selected relay node, containing flow ID, IP address and the port number of the receiver
7. The relay node builds up a table for forwarding packets
8. Each time, the sender attaches the flow ID in sending packets for relay node to where it should forward to

PDF: Redundant Path Selection

- PDF system may not need to find the best paths
  - Complexity increases due to active monitoring of probed packets and maintaining the link state information
  - Sending packets on two paths with the absolute lowest loss rates may not be necessary to achieve reasonable performance
- Select the redundant path
  - The weight can be thought of as delay, bandwidth, or loss rate. In this paper, weights denote the latencies
  - Two-step procedure
    - First, finds a set of redundant paths that are as disjoint as possible from the default path
    - Next, it selects the one that results in minimum latency
  - The sender either runs this algorithm at the beginning, or it can use the stored path provided from previous session

PDF: Summary

- A PDF system for delay sensitive applications over packet switched networks
- A scalable heuristic scheme for selecting a redundant path
  - Simulations done on Internet-like topologies
  - NS Simulations for comparing unipath and multipath scheme

PPRR: Introduction

- Challenges to BGP routing
  - Needs to scale to zillions of nodes—small overhead and fast convergence
  - Needs to do what AS (ISP) wants it to do—to implement a wide variety of policies
- BGP is not doing a good enough job
  - Misconfiguration
  - Hardware/software failure
  - AS has little incentive to do its best
  - Inconsistent policies among multiple AS
- PPRR: Path Probing Relay Routing
  - End-to-end path probing
  - Source-directed relay routing

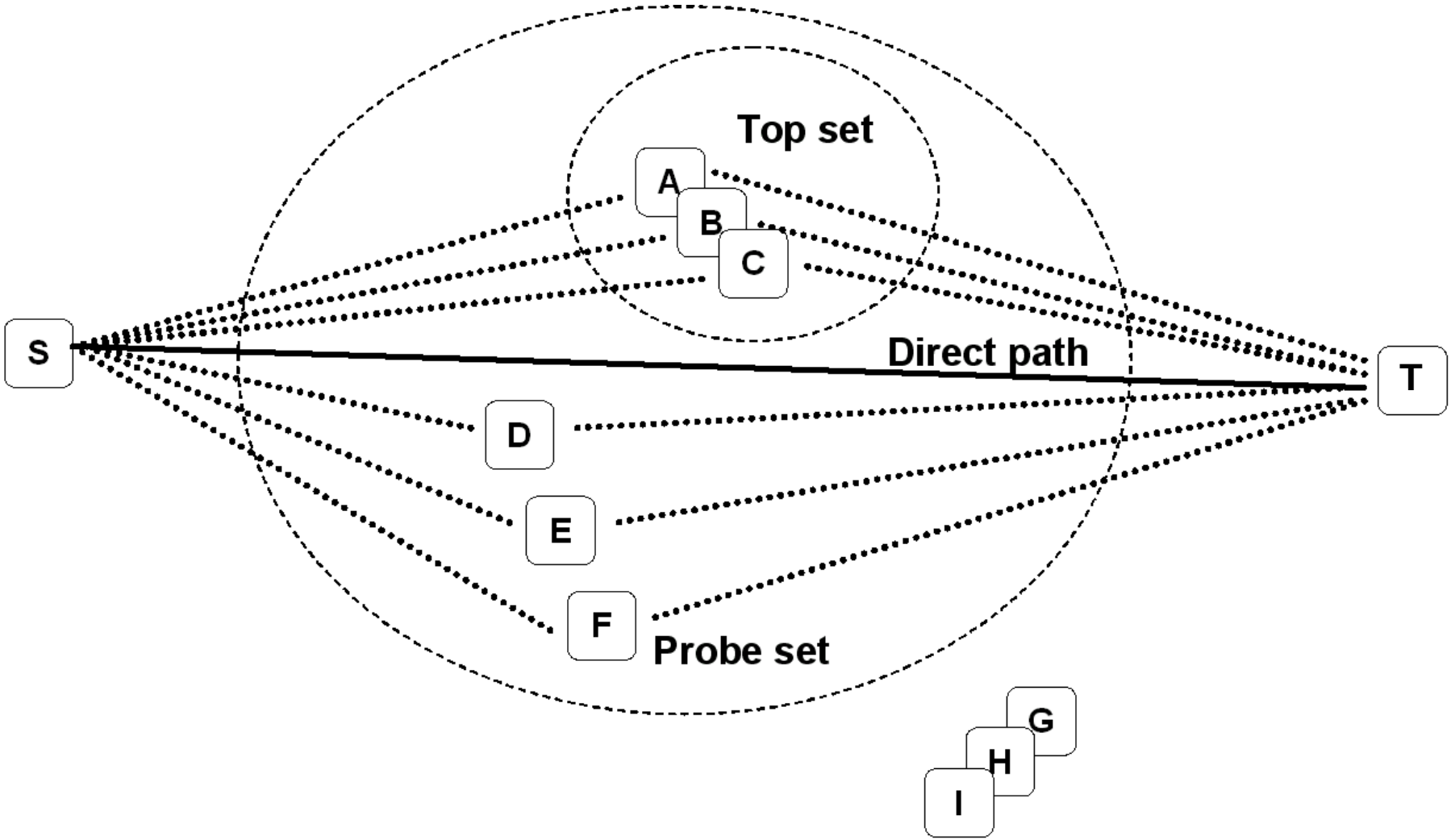
PPRR: Overview

- Participants only provide one simple service
  - Relay packets for other participants using, e.g., the Source Demand Routing Protocol (SDRP; see RFC1940) or tunneling
- Each participating node probes for its own application traffic on demand
  - More accurate probing because end-to-end
- Probing overhead depends on number of active sessions rather than network size

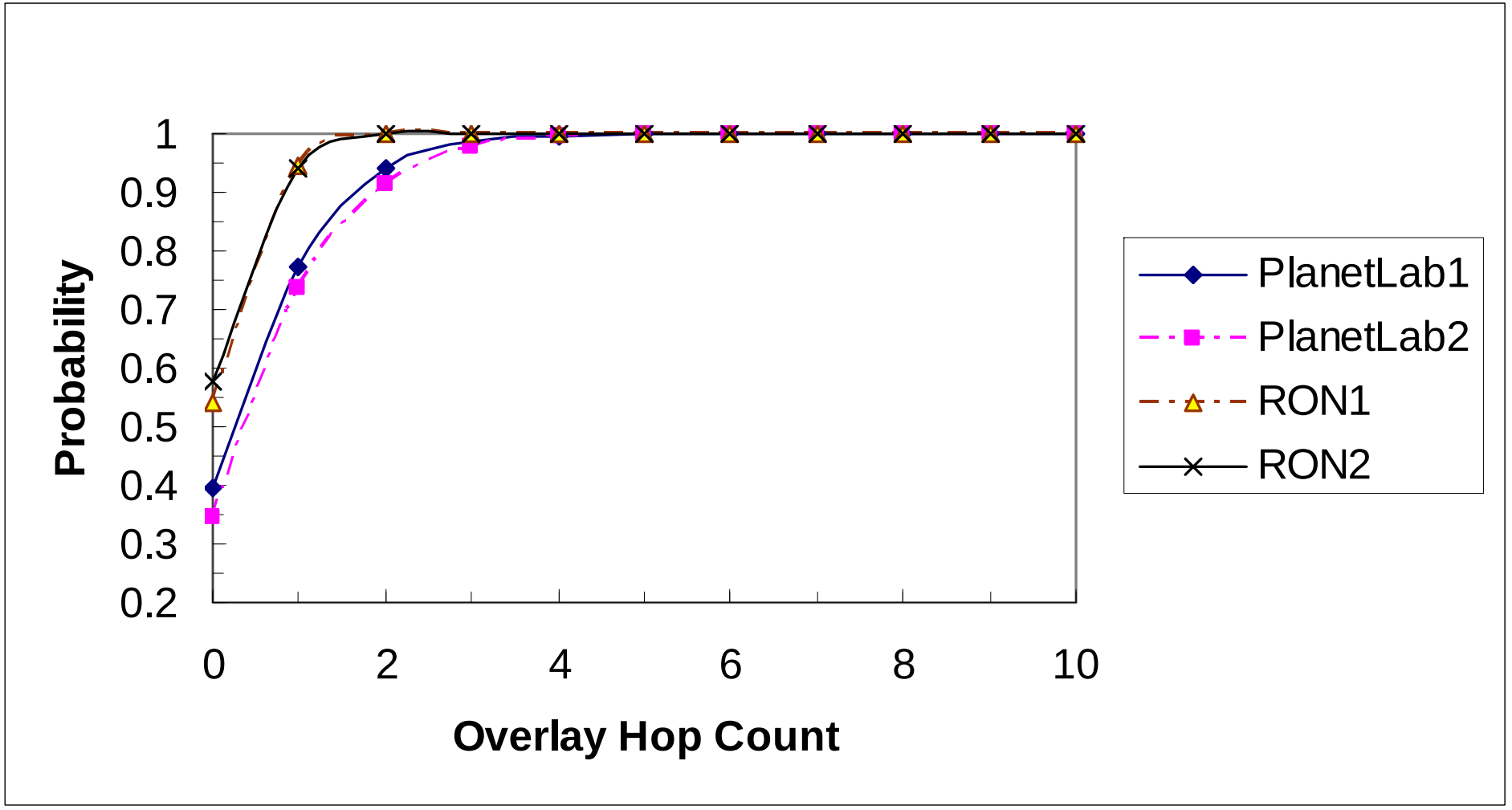
PPRR: Top-Set Probing Strategy

- For every active destination T, a node S maintains a Top Set, which contains the best viable routes to T known by S
- Also maintains a Probe Set containing
  - The direct path from S to T
  - The paths in Top Set
  - Periodically, randomly picked new paths
- S probes all paths in Probe Set and update Top Set periodically

PPRR: Illustration of Top Set Systems



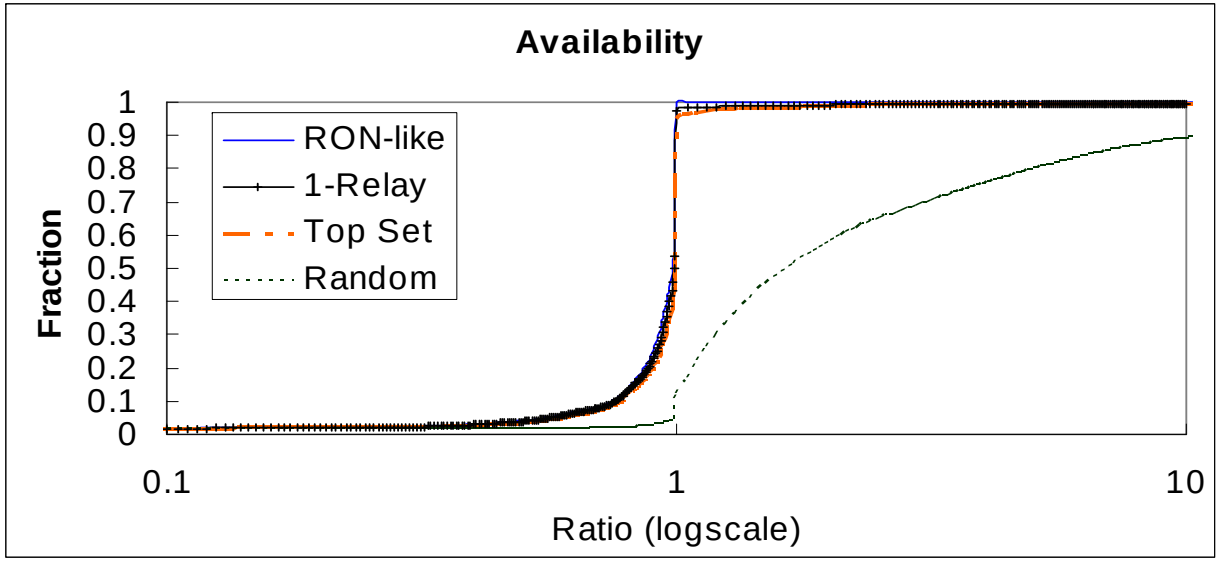
PPRR: One-Relay Paths Are Pretty Good

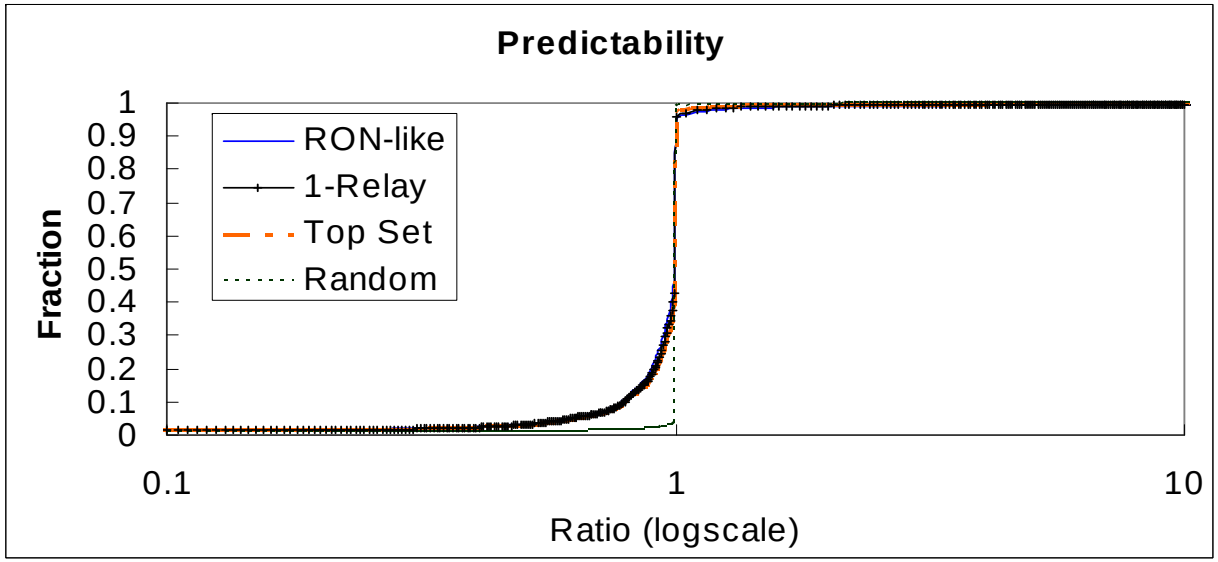


PPRR: Four Methods of Finding Alternative Paths to Be Compared

RON-like	Full-fledged link-state routing protocol
1-Relay	Best among all 1-relay paths
Top Set	Path found by Top Set System
Random	Typical 1-relay path chosen uniformly

PPRR: Performance





- Availability
  - Probability that we have found a path satisfying some performance requirement
  - Serves as an upper bound
- Predictability
  - Probability that the path we are using satisfies the performance requirement
  - More realistic metric but depends on quality of probing, how smart we are at predicting, etc

PPRR: Summary

- 1-relay paths are pretty good
  - Account for more than ¾ of the shortest paths
  - When direct path fails, almost always can find a 1-relay path that works
- Top Set System is a good approximation
  - Finds 1-relay path more than 95% of the time
  - Scales well to large networks
- Simple heuristics are quite satisfactory
  - Works about ¾ of the time
  - Still room for improvements, though

Summary

- There are several essential differences between IP networks and P2P overlay networks
- Overlay networks make the application control the routing
- Several new routing issues arise in P2P overlay networks
  - proximity routing
  - resilient routing