Exploiting Routing Redundancy via Structured Peer-to-Peer Overlays

Sep. 17, 2003
Byung-Gon Chun

Contents

• Motivation
• Resilient Overlay Routing
• Interface with legacy applications
• Evaluation
• Comparison

Motivation

• Frequent disconnection and high packet loss in the Internet
• Network layer protocol’s response to failures is slow

Quick recovery from route failures using structured P2P overlay
Resilient Overlay Routing

- Basics
- Route failure detection
- Route failure recovery
- Routing redundancy maintenance

Basics

- Use the KBR of structured P2P overlays [API]
- Backup links maintained for fast failover
- Proximity-based neighbor selection
- Proximity routing with constraints
- Note that all packets go through multiple overlay hops.

Failure Detection

- Failure recovery time ~ failure detection time when backup paths are precomputed
- Periodic beaconing
  - Backup link probe interval = Primary link probe interval * 2
- Number of beacons per period per node - log(N) vs. O(<D>) for unstructured overlay
- Routing state updates – log²N vs. O(E) for link state protocol

Failure Detection

- Link quality estimation using loss rate
  - \( L_n = (1-\alpha) L_{n-1} + \alpha L_p \)
- TBC - metric to capture the impact on the physical network
  - \( TBC = \text{beacons/sec} \times \text{bytes/beacon} \times \text{IP hops} \)
- PNS incurs a lower TBC

Structured overlays can do frequent beaconing for fast failure detection
How many paths?

- Recall the geometry paper
  - Ring - \((\log N)\)!
  - Tree - 1
- Tree with backup links
  - \(L^{\log_B N} = N^{\log_B L}\)

Failure Recovery

- Exploit backup links
- Two polices presented in [Bayeux]
  - First reachable link selection (FRLS)
    - First route whose link quality is above a defined threshold

Failure Recovery

- Constrained multicast (CM)
  - Duplicate messages to multiple outgoing links
  - Complementary to FRLS. Triggered when no link meets the threshold
  - Duplicate message drop at the path-converged nodes
    - Path convergence!
Routing Redundancy Maintenance

- Replace the failed route and restore the pre-failure level of path redundancy
- Find additional nodes with a prefix constraint
- When to repair?
  - After certain number of probes failed
  - Compare with the lazy repair in Pastry

- Thermodynamics analogy – active entropy reduction [K03]

Interface with legacy applications

- Transparent tunneling via structured overlays

Tunneling

- Legacy node A, B, Proxy P
- Registration
  - Register an ID - P(A) (e.g. P-1)
  - Establish a mapping from A’s IP to P(A)
- Name resolution and Routing
  - DNS query
  - Source daemon diverts traffic with destination IP reachable by overlay
  - Source proxy locates the destination overlay ID
  - Route through overlay
  - Destination proxy forwards to the destination daemon

Redundant Proxy Management

- Register with multiple proxies
- Iterative routing between the source proxy and a set of destination proxies
- Path diversity
Deployment

- What’s the incentive of ISPs?
  - Resilient routing as a value-added service
- Cross-domain deployment
  - Merge overlays
  - Peering points between ISP’s overlays
    - Hierarchy - Brocade

Simulation Result Summary

- 2 backup links
- PNS reduces TBC (up to 50%)
- Latency cost of backup paths is small (mostly less than 20%)
- Bandwidth overhead of constrained multicast is low (mostly less than 20%)
- Failures close to destination are costly.
- Tapestry finds different routes when the physical link fails.

Microbenchmark Summary

- 200 nodes on PlanetLab
- Alpha ~ between 0.2 and 0.4
- Route switch time
  - Around 600ms when the beaconing period is 300ms
- Latency cost ~ 0
  - Sometimes reduced latency in the backup paths – artifacts of small network
- CM
  - Bandwidth*Delay increases less than 30%
- Beaconsing overhead
  - Less than 7KB/s for beacon period of 300ms
Self Repair

- **RON**
  - Use one overlay hop (IP) for normal op. and one indirect hop for failover
  - Endpoints choose routes
  - $O(<D>)$ probes $D=O(N)$
  - $O(E)$ messages $E=O(N^2)$
  - Average of $k$ samples
  - Probe interval 12s
  - Failure detection 19s
  - 33Kbps probe overhead for 50 nodes (extrapolation: 56kbps around 70 nodes)

Tapestry ($L=3$)

- Use (multiple) overlay hops for all packet routing
- Prefixed routes
- $O(\log N)$ probes
- $O(\log^2 N)$ messages
- EWMA
- Probe interval 300ms
- Failure detection 600ms
- < 56Kbps probe overhead for 200 nodes

Comparison