Using SCTP Multihoming for Fault Tolerance & Load Balancing

Armando L. Caro Jr. and Janardhan R. Iyengar

Paul D. Amer, Gerard J. Heinz, Randall R. Stewart

http://pel.cis.udel.edu

P.E.L.
Protocol · Engineering · Laboratory
University of Delaware

Cisco Systems

Overview

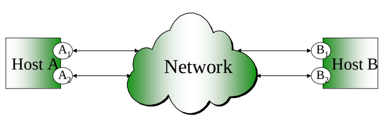
History

- Originally intended for telephony signaling
- Overcomes several TCP & UDP limitations
- Designed as a general purpose transport protocol
- Became an IETF Proposed Standard in October 2000 (RFC2960) under the SIGTRAN working group
- Handed to Transport Area working group for continued work

Features

- Reliable data transfer
- Ordered and unordered data delivery
- Multistreaming – no head-of-line blocking
- Multihoming

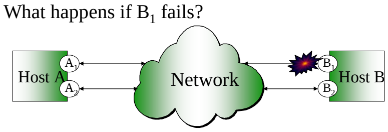
Multihoming



- 4 possible TCP connections:
  - (A_1,B_1) or (A_1,B_2) or (A_2,B_1) or (A_2,B_2)
- 1 SCTP association:
  - $(\{A_1,A_2\},\{B_1,B_2\})$
  - Primary destinations for A & B (e.g., A_1 & B_1)
  - Heartbeats determine reachability of idle destinations
  - Failover to an alternate destination if primary fails

Failover

- What happens if B_1 fails?



- TCP connection: (A_1,B_1)
  - Connection dies
- SCTP association: $(\{A_1,A_2\},\{B_1,B_2\})$
  - Failover to B_2 (ie, traffic temporarily migrates to B_2)
  - Upon B_1's restoration, traffic migrates back to B_1

Changeover

- Sender decides primary destination address for traffic
- Sender can change primary destination address during an active association
- Utilities (pending further research):
  - End-to-end mobility (with ADD/DELETE IP extension)
  - End-to-end load balancing

University of Delaware

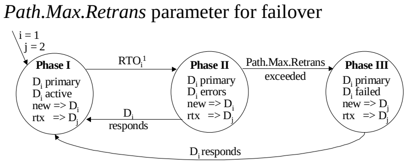
Adaptive Failover Mechanism

Motivation

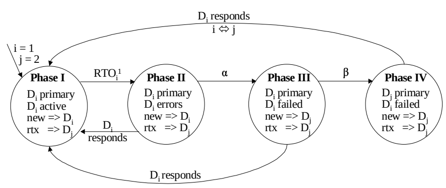
- End-to-end connectivity can suffer during net failures
  - Internet path outage detection and recovery is slow (shown to take as long as 3 to 30+ mins)
  - Path outages are common in mobile networks
- Network fault tolerance should cope with dynamic network conditions and varying applications needs
- Current SCTP failovers are not adaptive to application requirements and network conditions

Current Research

- SCTP's current failover mechanism uses the *Path.Max.Retrans* parameter for failover



- We propose a two-level (α-β) threshold failover mechanism
  - α - maintain primary, but failover temporarily
  - β - change the primary, making failover permanent



- Two-level threshold mechanism provides added control over failover actions
- Using ns-2, we have modeled SCTP data transfer latency to a dual homed destination with failovers incorporating the two-level threshold mechanism
- We are investigating the relationships between the thresholds and the network parameters to develop an adaptive failover mechanism

Future Research

- Develop an adaptive failover mechanism for SCTP
- Incorporate application requirements in the mechanism

Related Research

- Resilient Overlay Networks (RON)
  - Allows a small group of Internet applications to detect and recover from path outages within several seconds
- Rocks: Reliable Sockets
  - Protects apps from path failures common to mobile computing such as link failures and IP address changes
- Migrate
  - End-to-end framework for Internet mobility that supports rebinding of endpoints for established TCP connections
  - Fine-grained server failover mechanism of long-running connections
- Migratory TCP (M-TCP)
  - Mechanism to migrate live TCP sessions to a redundant server upon server overload, network congestion, etc.

End-to-End Load Balancing

Motivation

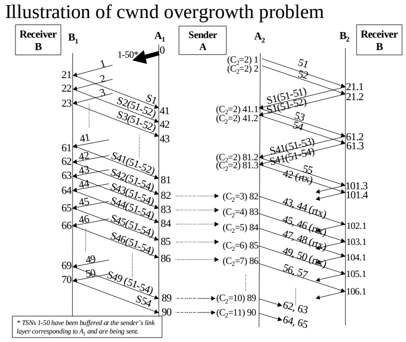
- Exploit all network resources visible at transport layer
- Perform *fine-grain* load balancing in transport layer
- Avoid replication of work and *coarse-grain* load balancing in applications

Issues

- Scheduling of traffic on multiple paths
- Reordering introduced by the sender
- Loss detection & recovery
- Congestion control: shared or separate?

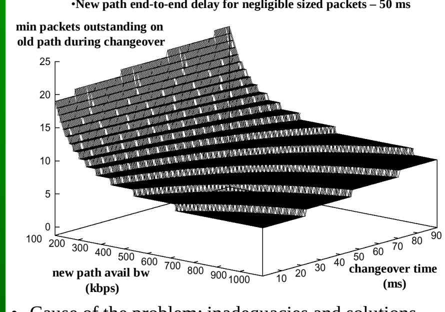
Current Work: Reordering Issues

- Reordering due to changeover causes spurious fast retransmissions and congestion window overgrowth
- Occurrence of reordering increases due to sender introduced route changes
- Illustration of cwnd overgrowth problem



- Is this problem a "corner case"? …NO!

Using:
•Old path avail bw – 500 kbps
•Old path end-to-end delay for negligible sized packets – 50 ms
•New path end-to-end delay for negligible sized packets – 50 ms



- Cause of the problem: inadequacies and solutions
  - Inadequacy: Retransmission ambiguity

    Solution: Rhein Algorithm (variation of Eifel Algorithm) Distinguish between acks for transmissions and retransmissions
  - Inadequacy: Congestion control is unaware of changeover

    Solution: Changeover Aware Congestion Control (CACC) Algorithms - prevents spurious fast retransmits

Future Research

- Investigate new loss detection and recovery techniques, which are robust to changeover
- Investigate dynamic shared bottleneck detection techniques for congestion control
- Investigate algorithms for scheduling traffic on multiple paths