Overlay Networks

Lecture 4

Further reading:
www.dsn.jhu.edu/publications/
A single, multi-purpose, IP-based network

- Each additional node increases its reach and usefulness (similar to any network)
- Each additional application domain increases its economic advantage
- Will therefore swallow most other networks
  - Happened: mail to e-mail, Phone to VoIP, Fax to PDFs
  - Started the process: TV, various control systems
  - Still to come: Cell phone networks

The art of design – a successful paradigm

- Keep it simple in the middle
  - Best-effort packet switching, routing (intranet, Internet)
- Smart at the edge
  - End-to-end reliability, naming

Could therefore adapt and scale

- Survived for 4 decades and counting
- Sustained at least 7 orders of magnitude growth

Standardized and a lot rides on it

- The basic services are not likely to change
New Applications Bring New Demands

• Communication patterns
  – From Point-to-point – to point-to-multipoint – to many-to-many
• High performance reliability
  – “Faster than real-time” file transfers
• Low latency interactivity
  – 150ms key stroke mirroring
  – 100ms for VoIP
  – 80-100ms for interactive games (remote surgery?)
• End-to-end dependability
  – From “Internet” dependability – to “phone service” dependability – to “TV service” dependability – to “remote surgery” dependability
• System resiliency
  – From E-mail fault tolerance – to financial transaction security – to critical infrastructure (SCADA) intrusion tolerance

So, What Can Be Done?

• Build specialized networks
  – Was done decades before the Internet
  – Think Cable/TV distribution (Satellite + last mile)
  – Extremely expensive
• Build private IP networks
  – Avoids the resource sharing aspects of the Internet, solves some of the scale issues
  – Expensive
  – Still confined to basic IP network capabilities
• Build a better Internet
  – Improvements and enhancements to IP (or TCP/IP stack)
  – “Clean slate design”
• Build overlay networks
The Overlay Paradigm

- Overlay paradigm:
  - In contrast to “keep it simple in the middle and smart at the edge”
  - Move intelligence and resources to the middle
    - Software-based overlay routers working on top of the internet
    - Overlay links translated to Internet paths
  - Smaller overlay scale (# nodes) ➔ smarter algorithms, better performance, and new services.

Initial Overlay Research

- Flexible Routing
  - RON – resilient routing using alternate paths [Andersen et al, 01]
  - XBone – flexible routing using IP in IP tunneling [Touch, Hotz, 98]
- Content Distribution
  - Yoid – host-based content distribution [Francis 00]
  - Overcast – reliable multicast for high bandwidth content distribution [Janotti et al, 00]
  - Bullet – multi-path data dissemination [Kostic et al 03]
- Multicast
  - ESM – provides application-level multicast [Chu et al, 00]
  - HTMP – interconnects islands of IP Multicast [Zhang et al, 02]
- Peer to Peer
  - Chord – logarithmic lookup service [Stoica et al, 01]
  - Kelips – O(1) lookup with more information stored [Gupta et al, 03]
- Group Communication
  - The Spread toolkit – scalable wide area group communication using an overlay approach [Amir, Danilov, Stanton, 06]
Outline

• The Overlay Network Paradigm
• First Steps
  – Low-latency reliable protocol
  – Spines – from Concepts to Systems
• The Quest for QoS
  – Almost-reliable, real-time protocol for VoIP
  – Almost-reliable, real-time protocol for Live TV
• Overlays on a Global Scale
  – The LiveTimeNet Cloud
• Going even faster
  – Reliability and timeliness
  – How fast can it get

End-to-End Reliability

• 50 millisecond network
  – E.g. Los Angeles to Baltimore
  – 50 milliseconds to tell the sender about the loss
  – 50 milliseconds to resend the packet
• At least 100 milliseconds to recover a lost packet
End-to-End Reliability

- 50 millisecond network
  - E.g. Los Angeles to Baltimore
  - 50 milliseconds to tell the sender about the loss
  - 50 milliseconds to resend the packet
- At least 100 milliseconds to recover a lost packet
  - Can we do better?

Hop-by-Hop Reliability

- 50 millisecond network, five hops
  - 10 milliseconds to tell node DAL about the loss
  - 10 milliseconds to get the packet back from DAL
- Only 20 milliseconds to recover a lost packet
  - Lost packet sent twice only on link DAL – ATL
Average Latency and Jitter

Simulation

Latency

Jitter

Yair Amir

Fall 2014 / Week 4

13
How Dense Should an Overlay Be?

- 50 ms network divided evenly into x hops
- Delayed packets: arrive after more than 50+10ms
Spines – from Concepts to Systems

• The Spines Overlay Messaging system
  – An Overlay software router (daemon) on top of UDP
  – Running as a normal Internet application

• Easy to use programming platform
  – Transparent interface identical to the socket interface, giving TCP, UDP and IP Multicast functionality

• “Commercial grade” deployable system
  – Improving application performance over the Internet
  – Enabling new services
  – Open source (www.spines.org)

www.spines.org

[DSN03, NOSSDAV05, TOM06, Mobisys06, TOCS10]

• Daemons create an overlay network on the fly
• Clients are identified by the IP address of their daemon and a port ID
• Clients feel they are working with UDP and TCP using their IP and port identifiers
• Protocols designed to support up to 1000 daemons (locations), each daemon can handle up to about 1000 clients
The Spines Architecture

www.spines.org

Outline

- The Overlay Network Paradigm
- First Steps
  - Low-latency reliable protocol
  - Spines – from Concepts to Systems
- The Quest for QoS
  - Almost-reliable, real-time protocol for VoIP
  - Almost-reliable, real-time protocol for Live TV
- Overlays on a Global Scale
  - The LiveTimeNet Cloud
- Going even faster
  - Reliability and timeliness
  - How fast can it get
The Siemens VoIP Challenge

- Can we maintain a “good enough” phone call quality over the Internet?
- High quality calls demand **predictable** performance
  - VoIP is interactive. Humans perceive delays at 100ms
  - The best-effort service offered by the Internet was not designed to offer any quality guarantees
  - Communication subject to **dynamic loss, delay, jitter, path failures**

![Graph](image1)

- **Almost Reliable, Real-time Protocol for VoIP**

- Localized real-time recovery on overlay hops
  - Retransmission is attempted only once
- Each Overlay node keeps a history of the packets forwarded in the last 100ms
  - When the other end of a hop detects a loss, it requests a retransmission and moves on
  - If the upstream node still has the packet in its history, it resends it
- Not a reliable protocol
  - No ACKs. No duplicates. No blocking.
  \[
  \text{loss} = 2 \cdot p^2 \quad \text{retr}_\text{delay} = 3 \cdot T + \Delta
  \]
- Recovery works for hops shorter than about 30ms
  - This is ok: overlay links are short!
VoIP Quality Improvement

- Spines overlay – 5 links of 10ms each
- 10 VoIP streams sending in parallel
- Loss on middle link C-D

Real-Time Routing for VoIP

- Routing algorithm that takes into account retransmissions
- Which path maximizes the number of packets arriving at node E in under 100 ms?
- Finding the best path by computing loss and delay distribution on all the possible routes is very expensive
- Weight metric for links that approximates the best path

\[ \text{Exp}_\text{latency} = (1 - p) \cdot T + (p - 2 \cdot p^2) \cdot (3 \cdot T + \Delta) + 2 \cdot p^2 \cdot T_{\text{max}} \]
Real-Time Routing for VoIP

- Different routing metrics evaluated on random topologies generated by BRITE
  - On each topology, the nodes defining the diameter of the network (furthest apart) are chosen as sender and receiver
  - Random loss rate from 0% to 5% on half of the links
- Optimizing Exp_latency metric compared with:
  - hops: Number of hops in the path
  - latency: Delay of the path
  - loss: Cumulative loss on the path
  - greedy: Dijkstra algorithm that computes delay distributions at each iteration and selects the partial path with maximum delivery ratio
  - best path: Computed out of all the possible paths

Each point in the graphs is an average over 1000 different topologies generated with BRITE
- Our simulator could not compute best path for topologies with more than 16 nodes in a timely manner
Overlay Approach to VoIP

• Localized real-time protocol on overlay hops
  – Retransmission is attempted only once

• Flexible routing metric avoids currently congested paths
  – Cost metric based on measured latency and loss rate of the links
  – Link cost equivalent to the expected packet latency when retransmissions are considered

What About Live TV?

• Is it more or less demanding than VoIP?
  – The ear is more sensitive than the eye
  – But we really want to see a clear picture on our large-screen HD TVs (less tolerance for error)

• How demanding is it?
  – Personal experience: could not notice problems with up to 100 misses per million with MPEG-2 encoding, 20 misses per million with H.264 encoding.
  – Broadcast standard: 5-6 errors per million for Standard TV, about 1 error per million for HD TV.

• What is Live?
  – Common TV transport systems usually add a few seconds.
  – Live service for interviews requires less than half a second delay
  – End-to-end transport window is therefore about 150-200ms.
Almost Reliable, Real-time Protocol for Live TV

<table>
<thead>
<tr>
<th>Network packet loss on one link (assuming 66% burstiness)</th>
<th>Loss experienced by flows on the LTN Network</th>
</tr>
</thead>
<tbody>
<tr>
<td>2%</td>
<td>&lt; 0.0003%</td>
</tr>
<tr>
<td>5%</td>
<td>&lt; 0.003%</td>
</tr>
<tr>
<td>10%</td>
<td>&lt; 0.03%</td>
</tr>
</tbody>
</table>
Outline

- The Overlay Network Paradigm
- First Steps
  - Low-latency reliable protocol
  - Spines – from Concepts to Systems
- The Quest for QoS
  - Almost-reliable, real-time protocol for VoIP
  - Almost-reliable, real-time protocol for Live TV
- Overlays on a Global Scale
  - The LiveTimeNet Cloud
- Going even faster
  - Reliability and timeliness
  - How fast can it get

Overlays on a Global Scale

The service provider point of view
- A service rather than software or hardware
- Control over where overlay nodes are located
- Multiple network providers in each overlay node (Super Nodes)
- Guaranteed capacity with admission control
- Monitoring and Control – near automation
The LiveTimeNet Cloud

Time for a Demonstration
Addressing the Technical Challenge

- Scalable overlay network architecture
  - Parallel overlays
- Real-time monitoring and control
  - Automated – take the human out of the loop
- Three levels of protection
  - Link level: real-time protocol for Live TV
  - Overlay level: responsive overlay routing
  - Cloud level: NxWay failover for overlay routers

Responsive Overlay Routing

- Utilizes multiple Tier 1 IP backbones
- Optimized overlay paths determine selected links
- Automatically and instantaneously switch to a better path
Responsive Overlay Routing

- Utilizes multiple Tier 1 IP backbones
- Optimized overlay paths determine selected links
- Automatically and instantaneously switch to a better path

![Diagram of Responsive Overlay Routing]

- Super Node
- Available link
- Selected link
- Deteriorating link

Yair Amir
Fall 2014 / Week 4

19
Responsive Overlay Routing

- Utilizes multiple Tier 1 IP backbones
- Optimized overlay paths determine selected links
- Automatically and instantaneously switch to a better path

Available link  Selected link  Deteriorating link

Super Node

Yair Amir  Fall 2014 / Week 4
Outline

• The Overlay Network Paradigm
• First Steps
  – Low-latency reliable protocol
  – Spines – from Concepts to Systems
• The Quest for QoS
  – Almost-reliable, real-time protocol for VoIP
  – Almost-reliable, real-time protocol for Live TV
• Overlays on a Global Scale
  – The LiveTimeNet Cloud
• Going even faster
  – Reliability and timeliness
  – How fast can it get

Multiple Node-Disjoint Paths

Send messages along 2 node-disjoint paths
Approx. 2x as expensive as using a single path
Each message is delivered in time as long as it arrives on EITHER path in time
We can even send messages along 3 node-disjoint paths, for ~3x the cost.

OR, we can use erasure coding! Split the message into 3 parts of size 1/2.
If any 2 arrive, original message can be recreated - Costs ~1.5x, lower resilience.