M & M: Freshman Experience

Yair Amir

Distributed Systems and Networks lab
Department of Computer Science
Johns Hopkins University

www.dsn.jhu.edu

Plan for This Week and Next

- Tangible: provide a pick into a decade of work from idea to research to realization in the commercial world:
  - Overlay networks
- Intangible:
  - Share a point of view / experience.
    - Research / engineering -> having a better tomorrow
    - Computer science as the purest form of engineering –> speed of change
    - Career paths

A bit about me

Making Communication Reliable A few Algorithms

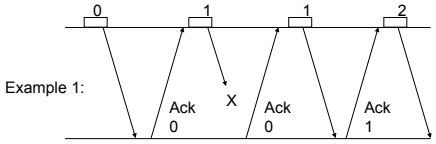
Causes for Message Loss in Networks:

- Buffer spill.
- Error detection in a packet.

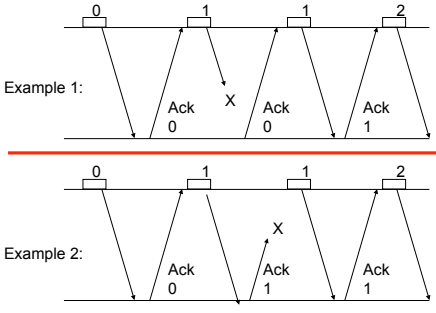
Protocols (algorithms):

- Send & Wait.
- Arpanet.
- Go back *n*.
- Selective Repeat.

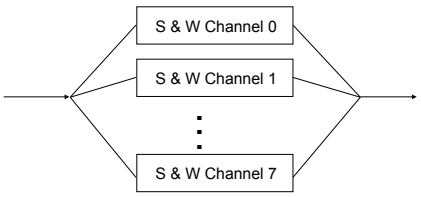
Send & Wait ARQ



Send & Wait ARQ (cont.)



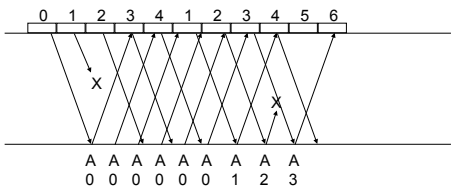
Arpanet ARQ



- Better line utilization than S & W.
- Unlimited memory required in theory.

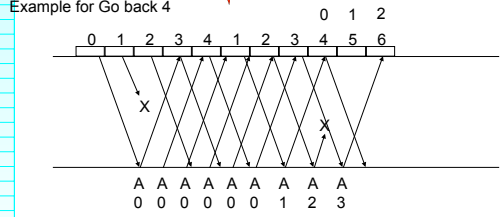
Go back *n* ARQ

Example for Go back 4



Go back *n* ARQ (cont.)

Example for Go back 4



- Good utilization
- limited memory required (one packet only).
- Full window is retransmitted in case of (one) error.

Selective Repeat ARQ

- Sliding window technique (as Go back *n*).
- Specifically indicating which packet is missing.
- Combines nacks and cumulative acks.
  - Acks acknowledge all messages with index of up to and including the ack value.
  - Nacks (negative acknowledgements) specifically request the messages with the indices in the nacks' values.
- Limited memory required (a full window).

Question: What if there is no feedback?

- A word about forward error correction (FEC), Internet loss patterns, etc.

Routing Approaches in Networks

- Distance vector routing
- Link state routing
- Inter-network routing

Distance Vector Routing

- Each router knows the id of every other router in the network.
- Each router maintains a vector with an entry for every destination that contains:
  - The cost to reach the destination from this router.
  - The first link that is on that least-cost path.
- Each router periodically sends its vector to its direct neighbors.
- Upon receiving a vector, a router updates the local vector based on the direct link's cost and the received vector.

Link State Routing

- Each router knows the id of every other router in the network.
- Each router maintains a topology map of the whole network.
- Each router periodically floods its direct links state (with its direct connectivity information).
- Upon receiving a vector, a router updates the local topology map and re-calculates shortest paths.

Internet Routing

- Routing Information Protocol:
  - **Distance vector protocol.**
  - **Hop count metric**
  - **Exchange is done every 30 seconds, fault detection every 180 seconds.**
  - **Cheap and easy to implement, unstable in the presence of faults.**
- Open Shortest Path First:
  - **Link state protocol.**
  - **Internal hierarchy for better scaling.**
  - **Optimization for broadcast LANs with routers on them. (A designated router represents the whole LAN) - Saves control messages and size.**

Internet Routing (cont).

- A hierarchical routing protocol that connects networks, each of which runs an internal routing protocol.
- OSPF or RIP are common internal protocols.
- BGP - Border Gateway Protocol -
  - **A path vector protocol with additional policy information for each path. Path vector protocols have the complete path in each entry and not only the next direct member.**
  - **Generally used as the hierarchical routing protocol.**

Overlay Networks: An Old-New Communication Paradigm for the Coming Decade

Yair Amir

Distributed Systems and Networks lab
Department of Computer Science
Johns Hopkins University

www.dsn.jhu.edu

Team Work!

- The one and only – Michal Miskin-Amir
- The advisor – Danny Dolev
- The professors –
  - M. Melliar-Smith, L. Moser, K. Birman, A. Brodsky, Y. Yemini
- The student-colleagues -
  - R. Borgstrom, J. Stanton, D. Shaw, J. Green, J. Schultz, T. Schlossnagle, A. Peterson
  - C. Nita-Rotaru, C. Danilov, C. Tutu, R. Caudy, A. Munjal, M. Hilsdale
  - N. Rivera, J. Lane, R. Musaloiu-Elefteri, J. Kirsch, M. Kaplan
- The go-to experts -
  - B. Awerbuch, A. Barak, G. Tsudik, S. Goose, A. Terzis, B. Coan, R. Ostrovsky
- The Hopkins professors -
  - B. Awerbuch, G. Mason, R. Kosaraju, S. Smith, M. Goodrich, R. Westgate
- The program managers - D. Maughan, T. Gibson, C. Landwehr

The Internet-based Communication Revolution

A single, multi-purpose, IP-based network

- It's everywhere
  - School, work, home,
  - airport, coffee shop, beach
  - train, plane, …
- It becomes like electricity
  - A basic, globally available service
  - Simply a plug in the wall
  - Or, better yet, without a plug …

The Internet-based Communication Revolution (cont.)

A single, multi-purpose, IP-based network

- Positive feedback
  - 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 in the past – telegraph? telex?
  - Currently happening: Phone to VoIP, fax to pdfs, various control systems
  - Still to come: TV, Cell phone networks
  - Future applications

The Internet-based Communication Revolution (cont.)

A single, multi-purpose, IP-based network

- The art of design – crucial!
  - Packet switching
  - Routing (intranet, internet)
  - End-to-end approach to reliability, naming.
- Could therefore adapt, scale and be economically viable
  - Survived for 4 decades and counting
  - Sustained at least 7 orders of magnitude growth
- It is standardized and a lot rides on it
  - The basic services offered by the Internet are not likely to change

New Applications Bring New Demands

- Communication patterns
  - From Point-to-point to point-to-multipoint.
  - Many point-to-multipoint, many many-to-many
- High performance reliability
  - "Faster than real-time" file transfers
- Low latency interactivity
  - 100ms for VoIP
  - 80-100ms for interactive games (remote surgery?)
  - 150ms key stroke mirroring
  - 185ms mouse-down feedback
- End-to-end dependability
  - "Internet" dependability to "phone service" dependability to "TV service" dependability to "remote surgery" dependability

New Applications Bring New Demands

- Communication patterns
  - From Point-to-point to point-to-multipoint.
  - Many point-to-multipoint, many many-to-many
- High performance reliability
  - "Faster than real-time" file transfer
- Low latency interactivity
  - 100ms for VoIP
  - 80-100ms for interactive games (remote surgery?)
  - 150ms key stroke mirroring
  - 185ms mouse-down feedback
- End-to-end dependability
  - "Internet" dependability to "phone service" dependability to "TV service" dependability to "remote surgery" dependability

So, What Can Be Done?

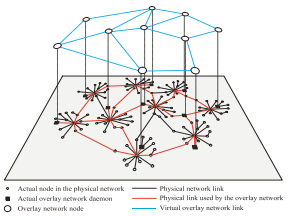
- Build specialized networks
  - Was done decades before the Internet
  - Think Cable/TV distribution (Satellite + last mile)
  - Extremely expensive
- Build private IP networks
  - Avoids the resource sharing aspects of the Internet, solves some of the scale issues
  - Expensive
  - Still confined to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

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

Overlay Networks

- Application-level routers working on top of an underlying physical network (the Internet).
- Overlay links consist of multiple physical links.
- Processing overhead
- Placement of overlay routers not optimal
- Smaller scale allows smarter algorithms with better performance
- New services not practically available in the Internet



Overlay Networks (cont.)

- Improved performance and new functionality:
  - Relatively small size (1000 nodes) enables maintaining state information
    - New insight: ~60 well-placed sites world-wide probably good enough
  - Optimized algorithms based on state information
    - From overlay network nodes
    - From the application
- Independent deployment:
  - No need for global coordination and standards
- "Old-New" communication paradigm:
  - The Internet started as an overlay over the phone network.

Early Overlay Network 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
  - Spread – scalable wide area group communication using an overlay approach [*Amir, Danilov, Stanton, 00*]

Outline

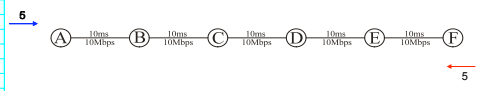
- The Internet Communication Revolution
- The Overlay Network Approach
- 1st hop (99-03):
  - The DARPA Fault Tolerant Network Challenge
    - Low latency reliability
    - Inventing an overlay architecture
- 2nd hop (03-06):
  - The Siemens VoIP Challenge
    - Almost-reliable, real-time transport
- 3rd hop (08- ..):
  - The LiveTimeNet TV Challenge
- Looking toward the future

The DARPA FTN Challenge

- How to improve reliable communication latency?
  - The Internet provides reasonable performance over reasonable connectivity conditions
  - In not-so-reliable, high-latency networks (think military networks), the performance of reliable communication over the Internet collapses
- Phenomenon inherent to end-to-end behavior of the Internet
  - Experienced in standard wide-area networks

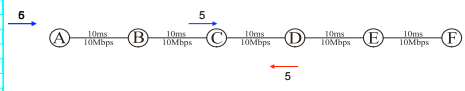
End-to-End Reliability

- 50 millisecond network, five hops
  - Think 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 ?

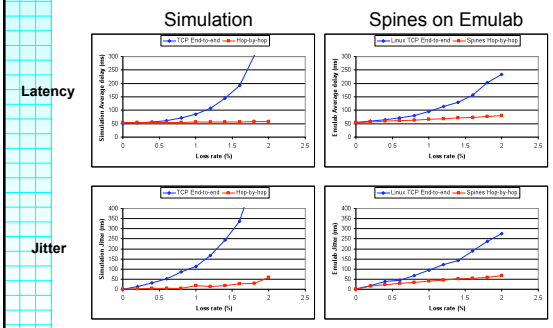


Hop-by-Hop Reliability

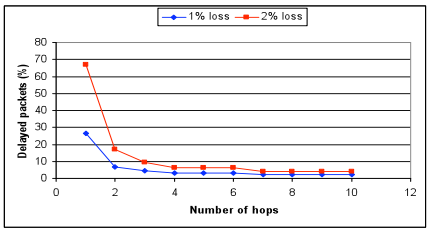
- 50 millisecond network, five hops
  - 10 milliseconds to tell node C about the loss
  - 10 milliseconds to get the packet back from node C
- Only 20 milliseconds to recover a lost packet
  - Lost packet sent twice only on link C – D
  - Where was packet 6 during the recovery of 5 ?



Average Latency and Jitter



How Dense Should an Overlay Be?

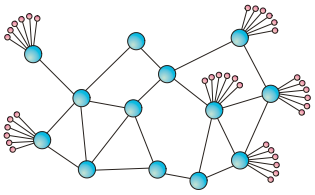


- 50 ms network divided evenly into x hops
- Delayed packets: arrive after more than 50+10ms

Spines: An Overlay Platform

- A messaging software system
  - Spines builds an overlay router (daemon) on top of UDP, running as a regular user application
  - System builders and researchers can build protocols within its framework, experimenting with routing, link protocols, etc.
- Transparent API
  - API similar to the socket interface, giving TCP, UDP and IP Multicast functionality
- A deployable platform
  - Improving application performance over the Internet
  - Enabling new services
  - Open source (www.spines.org)

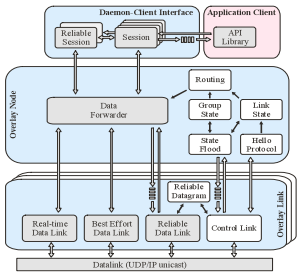
The Spines Architecture



[DSN03, NOSSDAV05, TOM06]

- 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 (cont.)



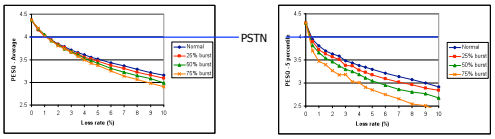
Outline

- The Internet Communication Revolution
- The Overlay Network Approach
- 1st hop (99-03):
  - The DARPA Fault Tolerant Network Challenge
    - Low latency reliability
    - Inventing an overlay architecture
- 2nd hop (03-06):
  - The Siemens VoIP Challenge
    - Almost-reliable, real-time transport
- 3rd hop (08- ..):
  - The LiveTimeNet TV Challenge
- Looking toward the future

The Siemens VoIP Challenge

- Can we maintain a good enough phone call quality over the Internet?
  - Metric
- People expect high quality and reliable telephony
  - While not impressed by constant good quality, users are easily disappointed by a few bad or dropped calls
- High quality calls demand predictable performance
  - 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
  - VoIP is interactive. Humans perceive delays at 100ms

Quality Degradation with Loss



50ms network delay

- G.711 – High quality voice codec
- PESQ – Standardized measure of voice quality
  - Desired quality is 4 PESQ, Chargeable quality is 3.5 PESQ
- Internet characteristics in 2003 [*Andersen et al, 03*]:
  - Average loss rate: 0.42% can be as high as 13%
  - Conditional loss probability (burstiness): 66%

Overlay Approach to VoIP

- Localized real-time recovery on overlay hops
  - Retransmission is attempted only once
- Flexible routing metric avoids chronically 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

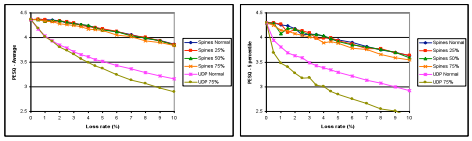
Real-time Recovery Protocol

- 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 blocking. No ACKs. No duplicates.

$$loss \approx 2 \cdot p^2 \qquad retr\_delay = 3 \cdot T + \Delta$$

- Able to recover packets only for hops shorter than about 30ms
  - That is ok: Overlay links are short !

VoIP Quality Improvements

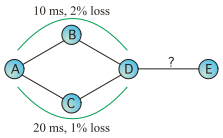


- Spines overlay – 5 links of 10ms each
- 10 VoIP streams sending in parallel
- Loss on middle link C-D

50ms network delay

Real-time Routing

- Routing algorithm that takes into account retransmissions
- Which path maximizes the number of packets arriving at node **E** in under 100 ms ?



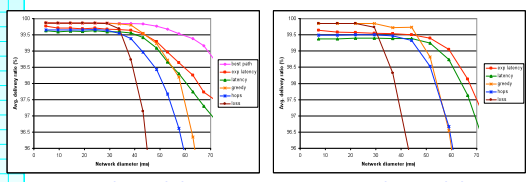
- Finding the best path by computing loss and delay distribution on all the possible routes is very expensive
- Weight metric for links that approximates the best path

$$Exp\_latency = (1-p)\cdot T + (p-2\cdot p^2)\cdot(3\cdot T + \Delta) + 2\cdot p^2 \cdot T_{\max}$$

Routing Evaluation

- Different routing metrics evaluated on random topologies generated by BRITE [*Medina et al, 01*]
  - On each topology, the nodes defining the diameter of the network (furtheest 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

Simulation Results



15 nodes; 30 links 50 nodes; 100 links

- 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 highlights

- Localized recovery on overlay links
  - Takes care of sudden increases in loss
- Routing metric equivalent to the expected packet latency when retransmissions are considered
  - Avoids long term congested paths

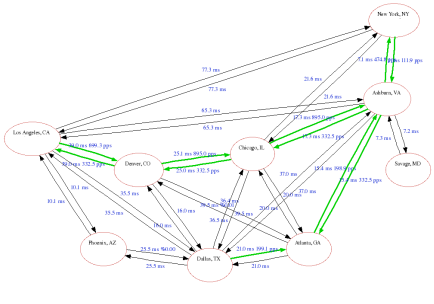
Outline

- The Internet Communication Revolution
- The Overlay Network Approach
- 1st hop (99-03):
  - The DARPA Fault Tolerant Network Challenge
    - Low latency reliability
    - Inventing an overlay architecture
- 2nd hop (03-06):
  - The Siemens VoIP Challenge
    - Almost-reliable, real-time transport
- 3rd hop (08- ..):
  - The LiveTimeNet TV Challenge
- Looking toward the future

The LiveTimeNet TV Challenge

- Can the Internet be an underlying network for a live TV service?
  - Live channel transport (Business to Business)
  - The virtual cable company (Business to Consumer)
  - Next Generation TV (Interactivity)
- Requirements
  - Scalability: High capacity flows, many any-to-many flows
  - High availability and uniform delivery
  - Huge amount of bandwidth
- Technology trends
  - Cheap long-haul access bandwidth
  - Broadband Internet connectivity to the home
  - Multi-core computer architecture

Lets See a Demonstration

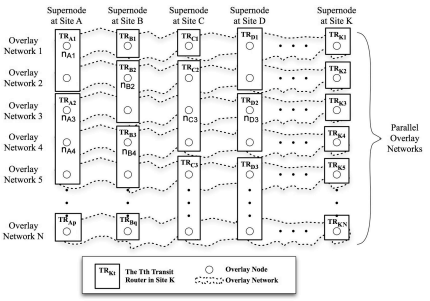


LiveTimeNet Link Usage

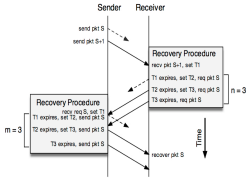
Addressing the Technical Challenge

- Scalable overlay network architecture
  - Scalable with the number of overlays
- Three levels of protection
  - Link level: Near-reliable real-time protocol
  - Overlay level: Policy-based routing per overlay
  - NxWay failover for overlay routers
- Service Overlay Network approach
  - Guaranteed capacity with admission control
  - Carefully selected "neutral" sites
  - Served by several top-tier Internet providers

Scalable Overlay Networks



Near-Reliable Real-Time Protocol



Network packet loss on one link (assuming 66% burstiness)	Loss experienced by flows on the LTN Network
2%	< 0.0003%
5%	< 0.003%
10%	< 0.03%

Outline

- The Internet Communication Revolution
- The Overlay Network Approach
- 1st hop (99-03):
  - The DARPA Fault Tolerant Network Challenge
    - Low latency reliability
    - Inventing an overlay architecture
- 2nd hop (03-06):
  - The Siemens VoIP Challenge
    - Almost-reliable, real-time transport
- 3rd hop (08- ..):
  - The LiveTimeNet TV Challenge
- Looking toward the future

Looking Toward the Future

Subjectively interesting "details":

- Overlay security, content protection, resiliency, survivability, management.
- Specialized overlay technologies
  - Remote Surgery as an example
    - 80ms response time, impeccable reliability

Looking Toward the Future (2)

Big picture trends:

- A single Internet swallowing most other networks
  - Substantial percentage of phone calls today are VoIP
  - ATT asks to be released from PSTN commitment (2/2010).
  - TV / VoIP is a factor of 100 = 1.5 years * 6 = about a decade. Hence, TV in 10 years where VoIP is today.
  - Cable/telco and eventually cell phone providers will have hard time avoiding becoming just a (great) pipe.
- The Internet is undergoing a rapid change
  - 150 autonomous networks carry 50% of Inter-AN traffic in 2009 down from several thousands in 2007!! (Arbor Networks report)