From Overlays to Clouds: Inventing a New Network Paradigm

Yair Amir

Don P. Giddens Lecture

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
  - D. Obenshain, T. Tantillo
- The go-to experts -
  - B. Awerbuch, A. Barak, G. Tsudik, S. Goose, A. Terzis, B. Coan, R. Ostrovsky
- The entrepreneurs -
  - M. Khan, Y. Javadi, S. Goose
- The Hopkins professors -
  - B. Awerbuch, G. Mason, R. Kosaraju, S. Smith, M. Goodrich, R. Westgate
- The program managers –
  - D. Maughan, T. Gibson, C. Landwehr, H. Shrobe

The Internet Revolution
A Technical Perspective

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 Internet Revolution
A Technical Perspective

A single, multi-purpose, IP-based network

- 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

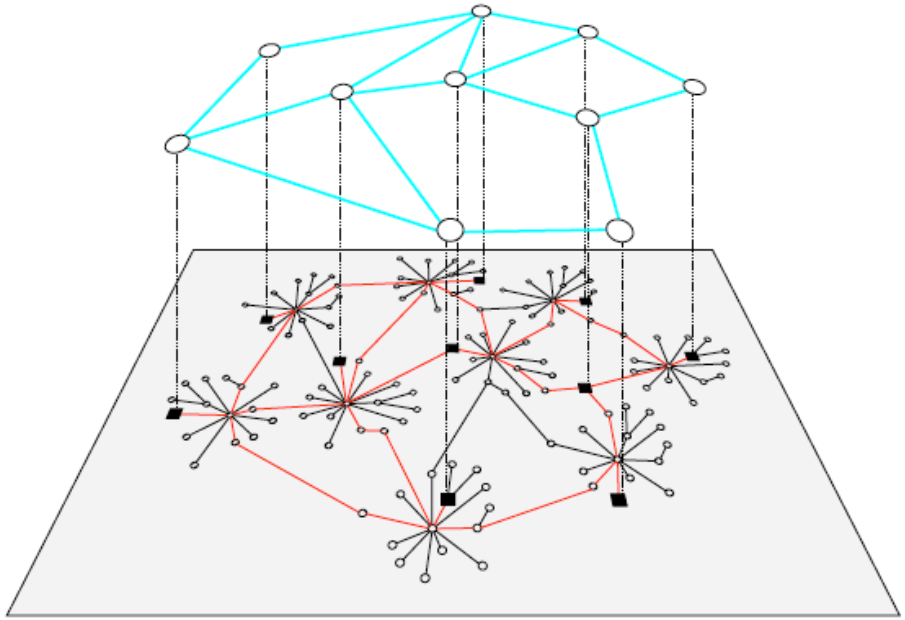
New Applications Bring New Demands

- 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.

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

Outline

- The Overlay Network Paradigm
- The DARPA Networking Challenge (99-03)
  - Overlay Architecture
  - Low-latency reliable transport
- The Siemens VoIP Challenge (03-06)
  - Almost-reliable, real-time transport
- The LiveTimeNet TV Challenge (08-…)
  - From Overlays to Clouds
  - Ultimate resiliency, automated monitoring and control
- The challenges ahead

The DARPA Challenge (99-03)

- The traditional paradigm (keep it simple in the middle and smart at the edge) works well for traditional applications in typical connectivity conditions
- But not so well for traditional applications in bad connectivity conditions
- And not so well for emerging applications in typical connectivity conditions

End-to-End Reliability

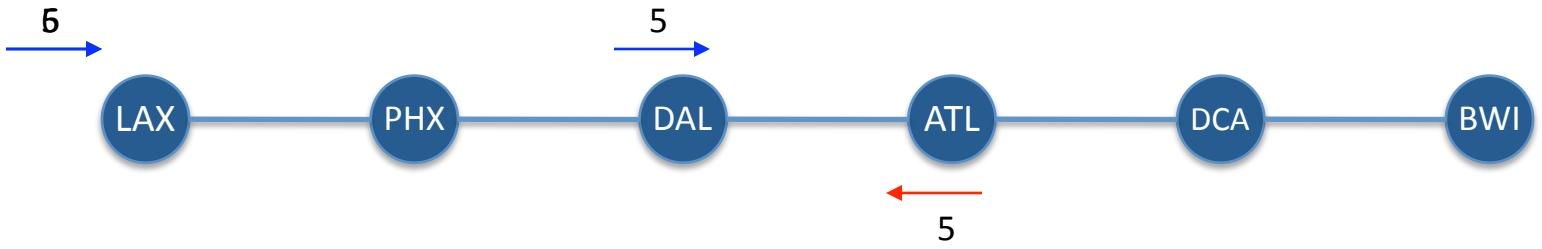
- 50 millisecond network
  - E.g. Los Angeles to Baltimore
  - 50 milliseconds to tell the sender about the loss
  - 50 milliseconds to resend the packet
- At least 100 milliseconds to recover a lost packet

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

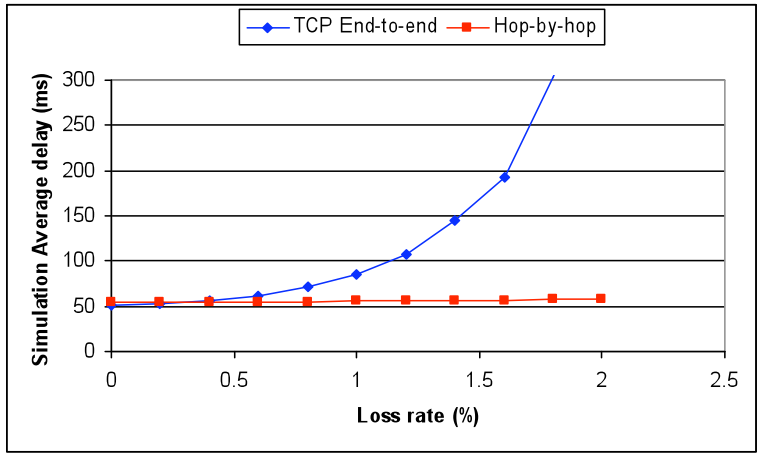
- 50 millisecond network, five hops
  - 10 milliseconds to tell node DAL about the loss
  - 10 milliseconds to get the packet back from DAL
- Only 20 milliseconds to recover a lost packet
  - Lost packet sent twice only on link DAL – ATL



Average Latency and Jitter

Simulation

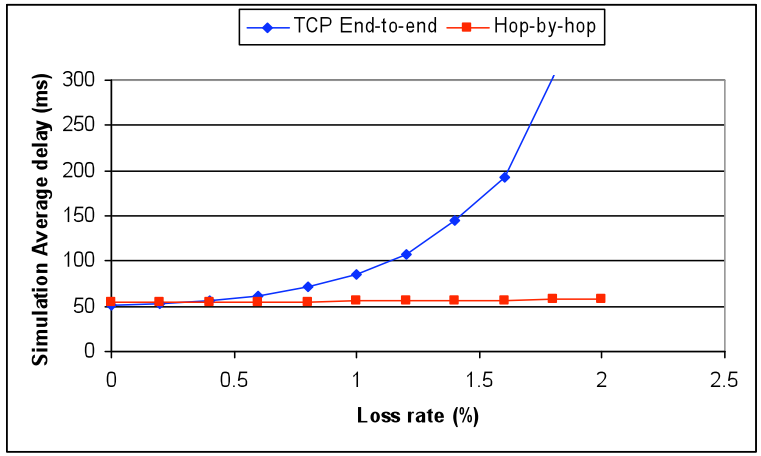
Latency



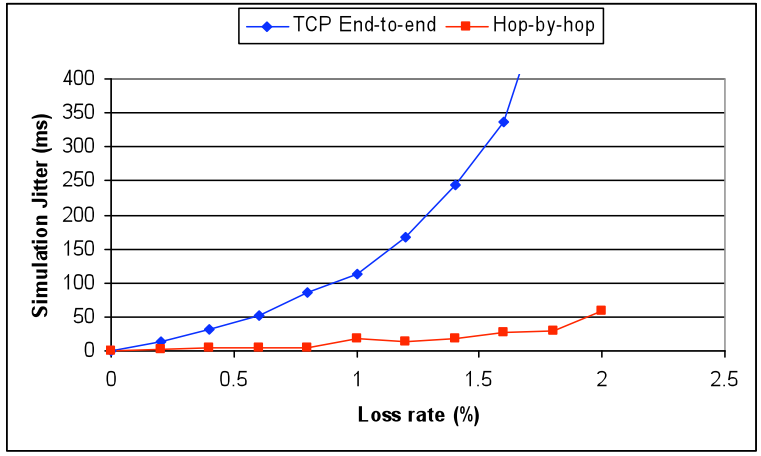
Average Latency and Jitter

Simulation

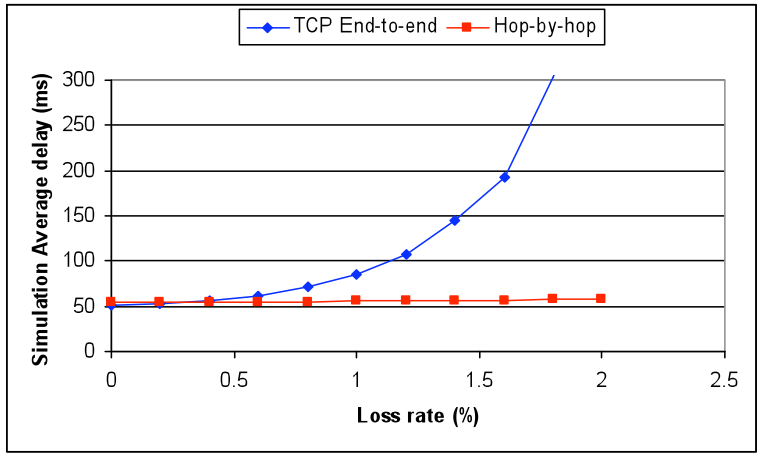
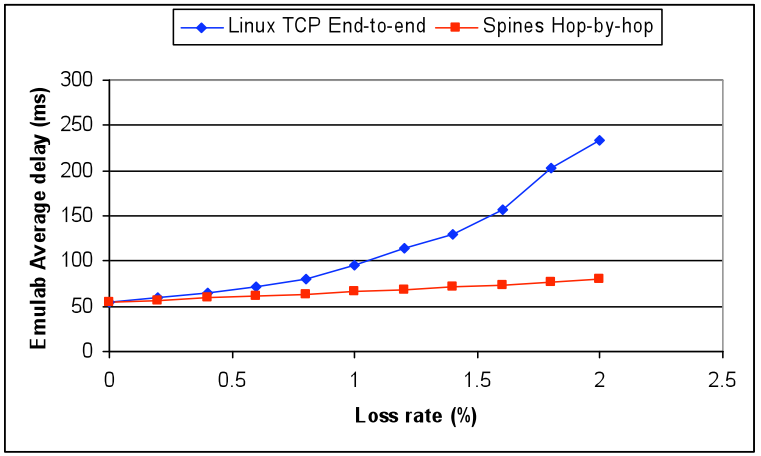
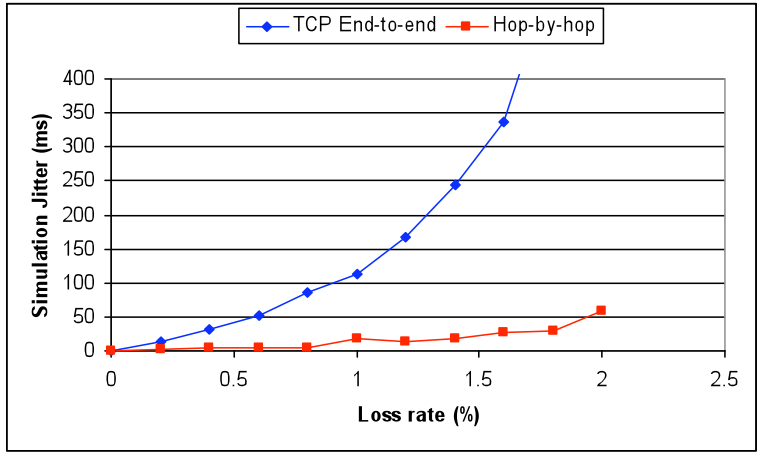
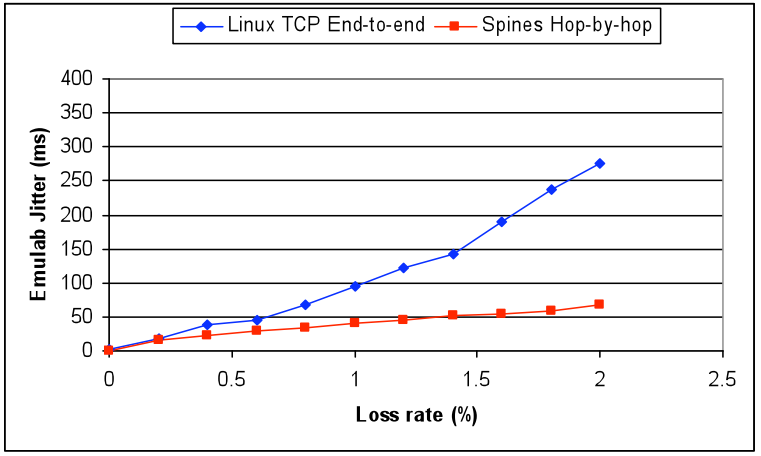
Latency



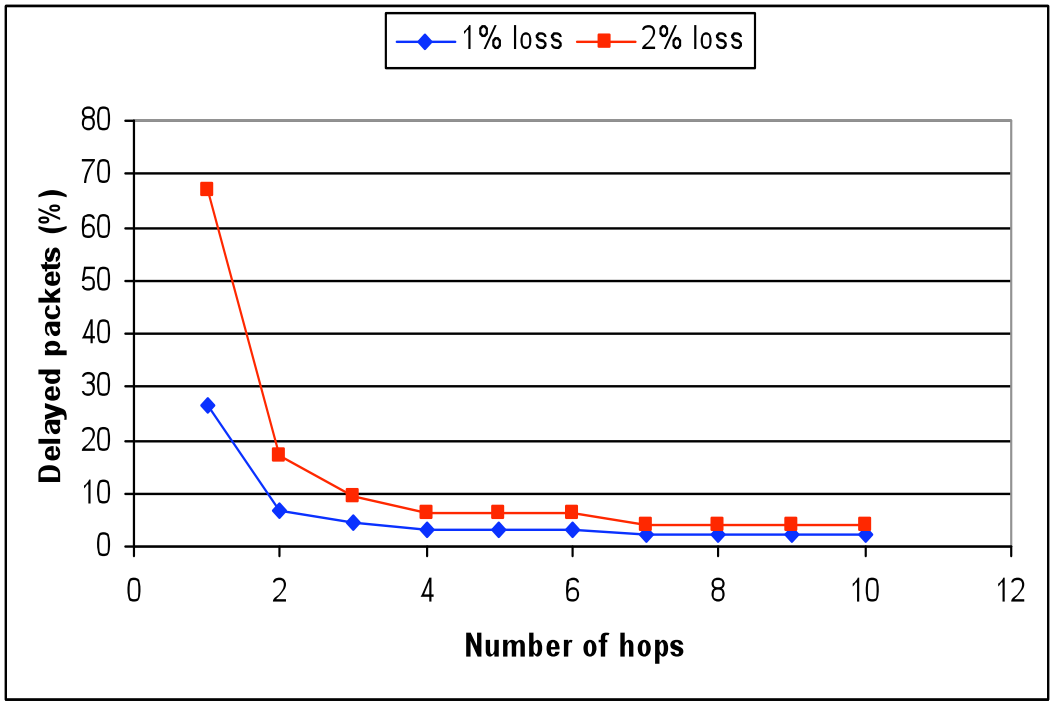
Jitter



Average Latency and Jitter

	Simulation	Spines on Emulab
Latency		
Jitter		

How Dense Should an Overlay Be?

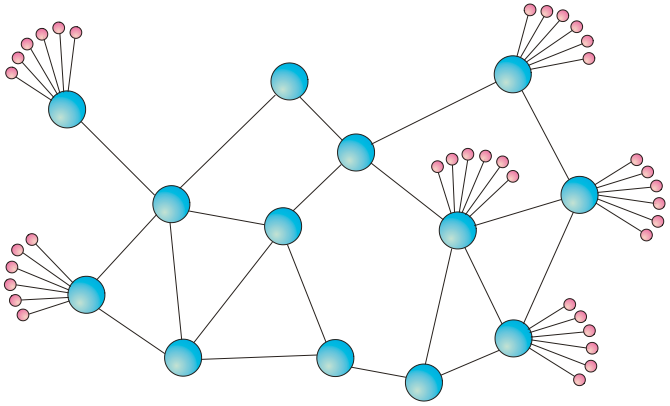


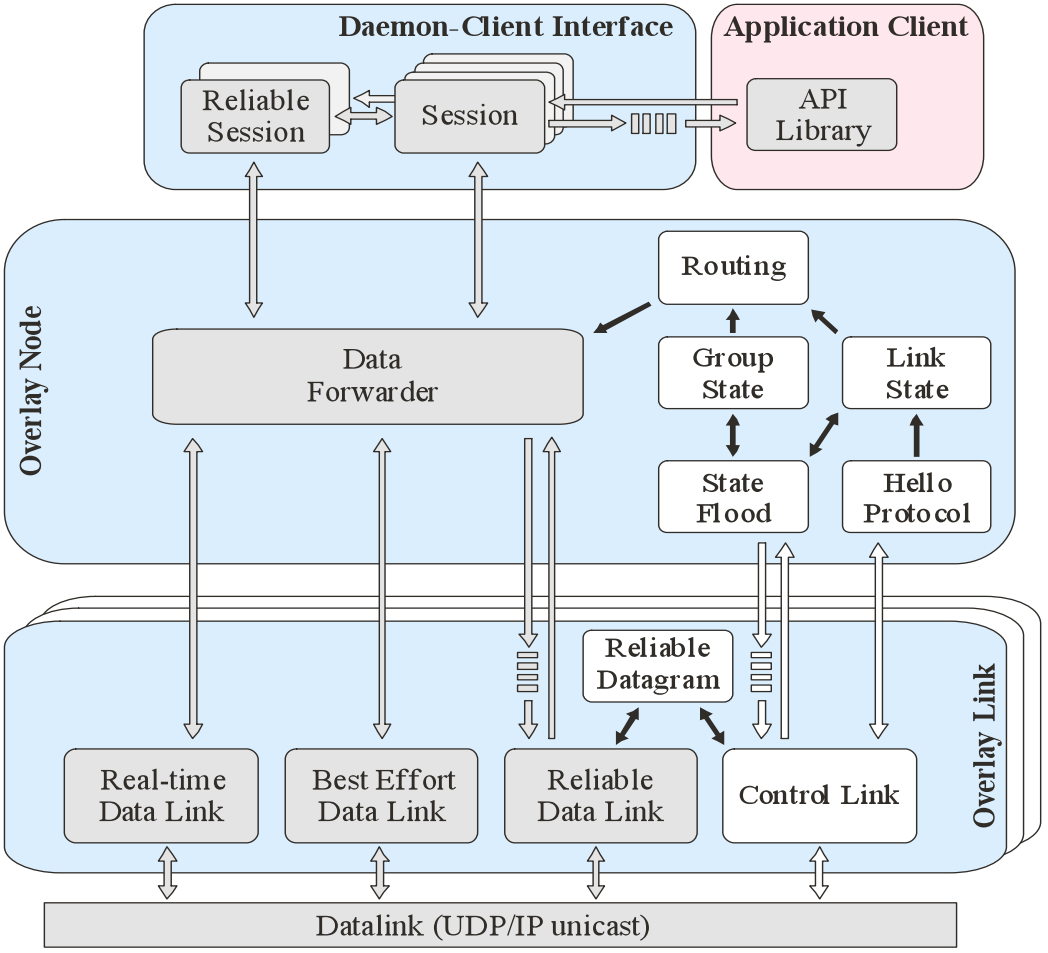
- 50 ms network divided evenly into x hops
- Delayed packets: arrive after more than 50+10ms

Spines: From Ideas to Reality

- 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)

The Spines Overlay Architecture



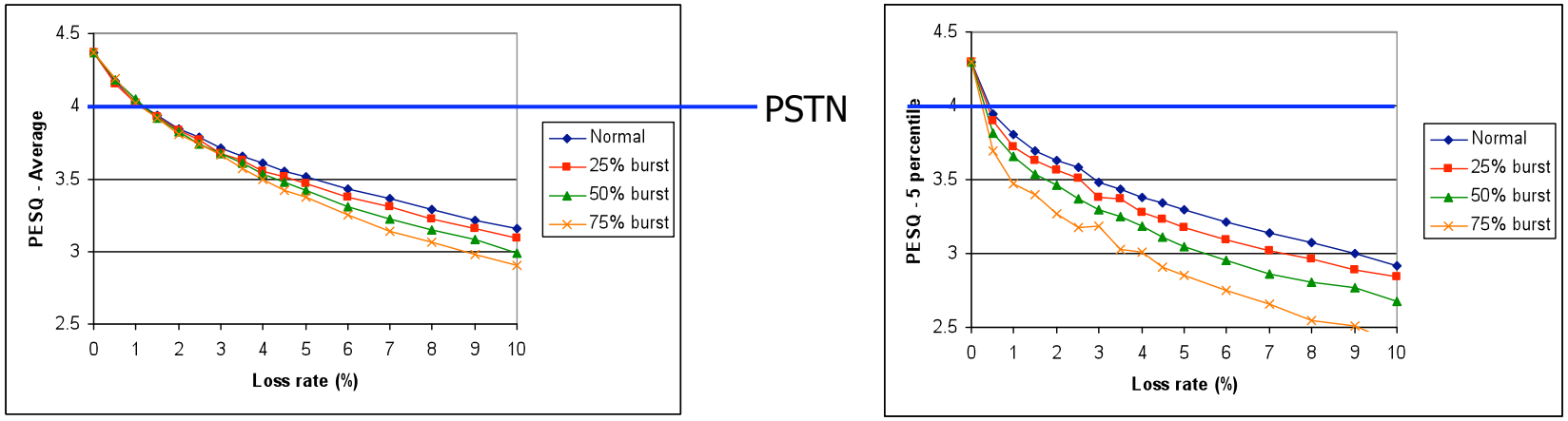


Outline

- The Overlay Network Paradigm
- The DARPA Networking Challenge (99-03)
  - Overlay Architecture
  - Low-latency reliable transport
- The Siemens VoIP Challenge (03-06)
  - Almost-reliable, real-time transport
- The LiveTimeNet TV Challenge (08-…)
  - From Overlays to Clouds
  - Ultimate resiliency, automated monitoring and control
- The challenges ahead

The Siemens VoIP Challenge (03-06)

- 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



50ms network delay

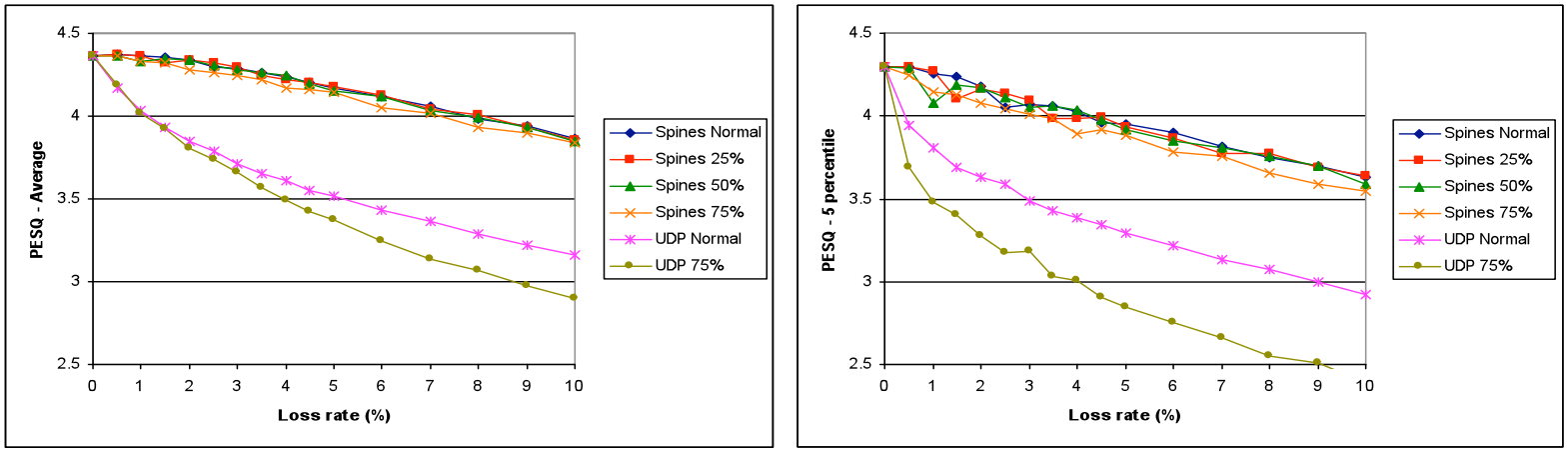
Real-time Recovery Protocol

- 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.

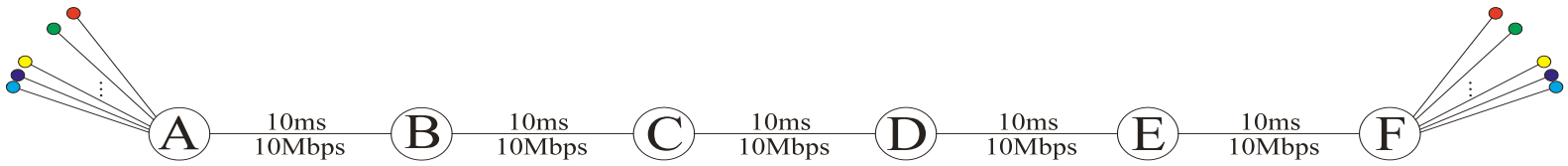
$$loss \approx 2 \cdot p^2 \qquad retr\_delay = 3 \cdot T + \Delta$$

- Recovery works for hops shorter than about 30ms
  - This is ok: overlay links are short !

VoIP Quality Improvement



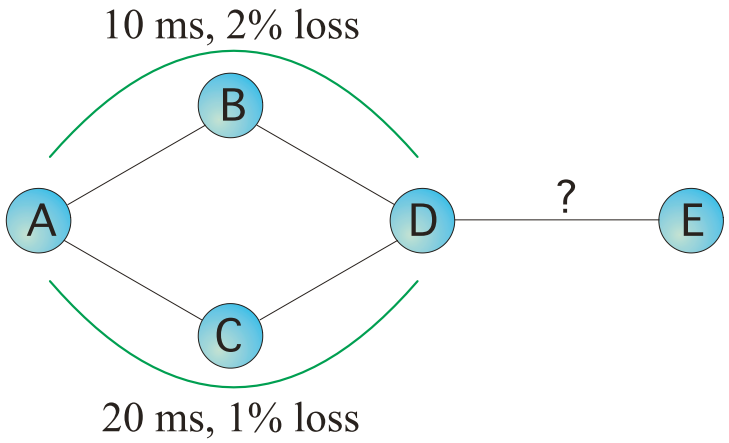
- Spines overlay – 5 links of 10ms each
- 10 VoIP streams sending in parallel
- Loss on middle link C-D



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

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

Overlay Approach to VoIP

- Localized real-time recovery 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

Outline

- The Overlay Network Paradigm
- The DARPA Networking Challenge (99-03)
  - Overlay Architecture
  - Low-latency reliable transport
- The Siemens VoIP Challenge (03-06)
  - Almost-reliable, real-time transport
- The LiveTimeNet TV Challenge (08-…)
  - From Overlays to Clouds
  - Ultimate resiliency, automated monitoring and control
- The challenges ahead

The LiveTimeNet TV Challenge (08-…)

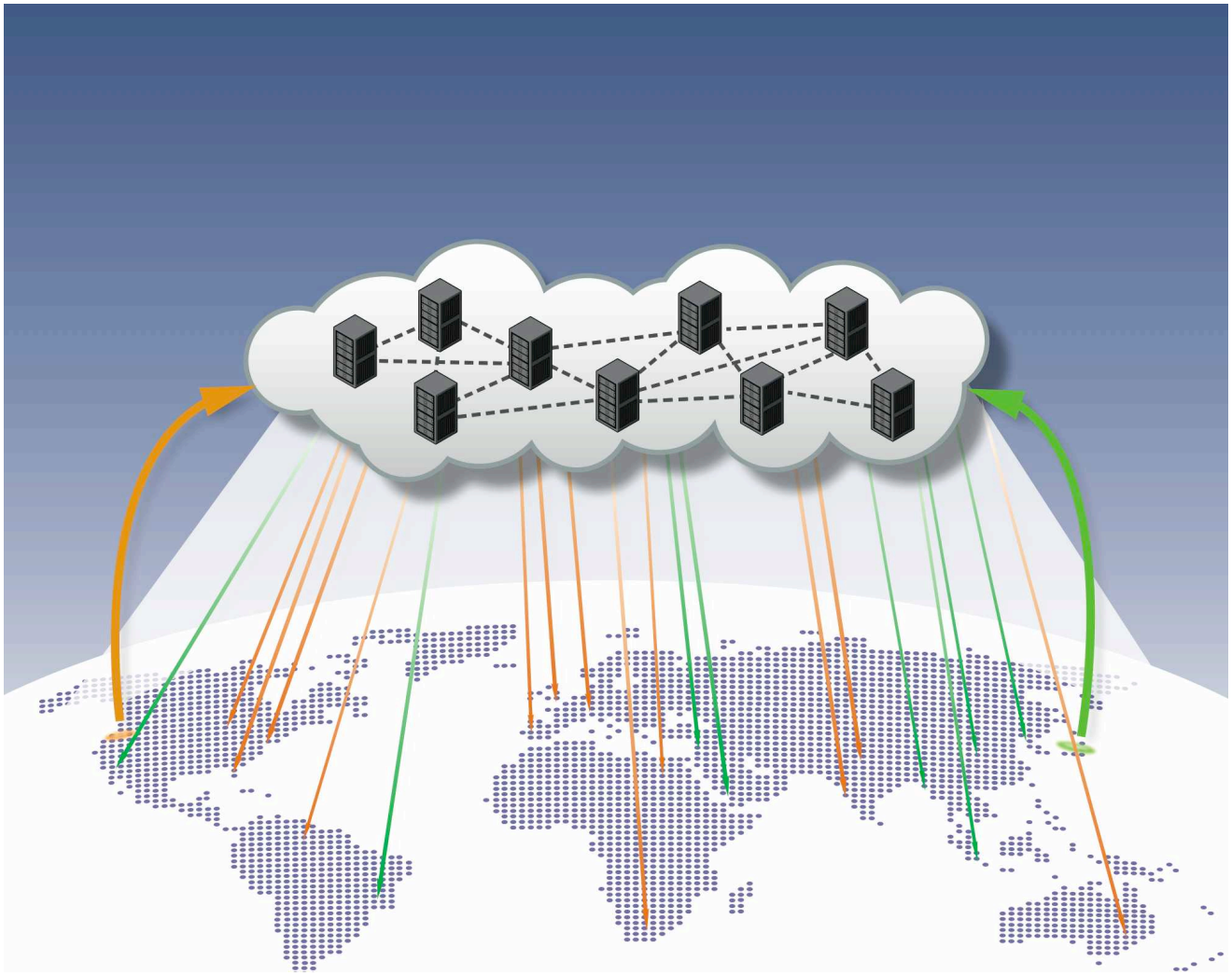
- Can the Internet be an underlying network for a live TV service?
  - Live channel transport (Business to Business)
  - The virtual cable company (Business to Consumer)
  - Next Generation TV (Interactivity)
- Requirements
  - Scalability: High capacity flows, many any-to-many flows
  - High availability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

From Overlays to Clouds

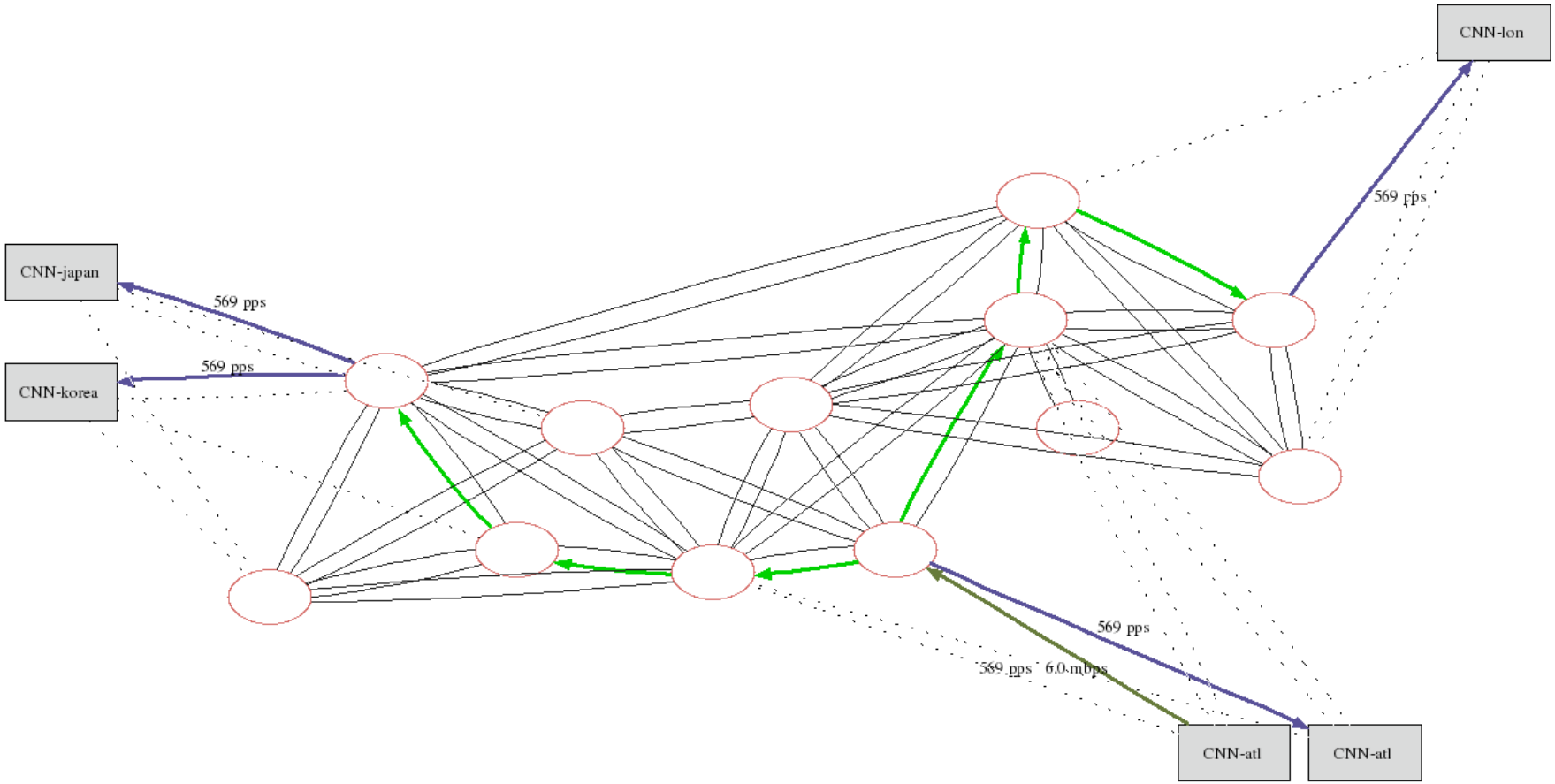
The service provider point of view

- 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

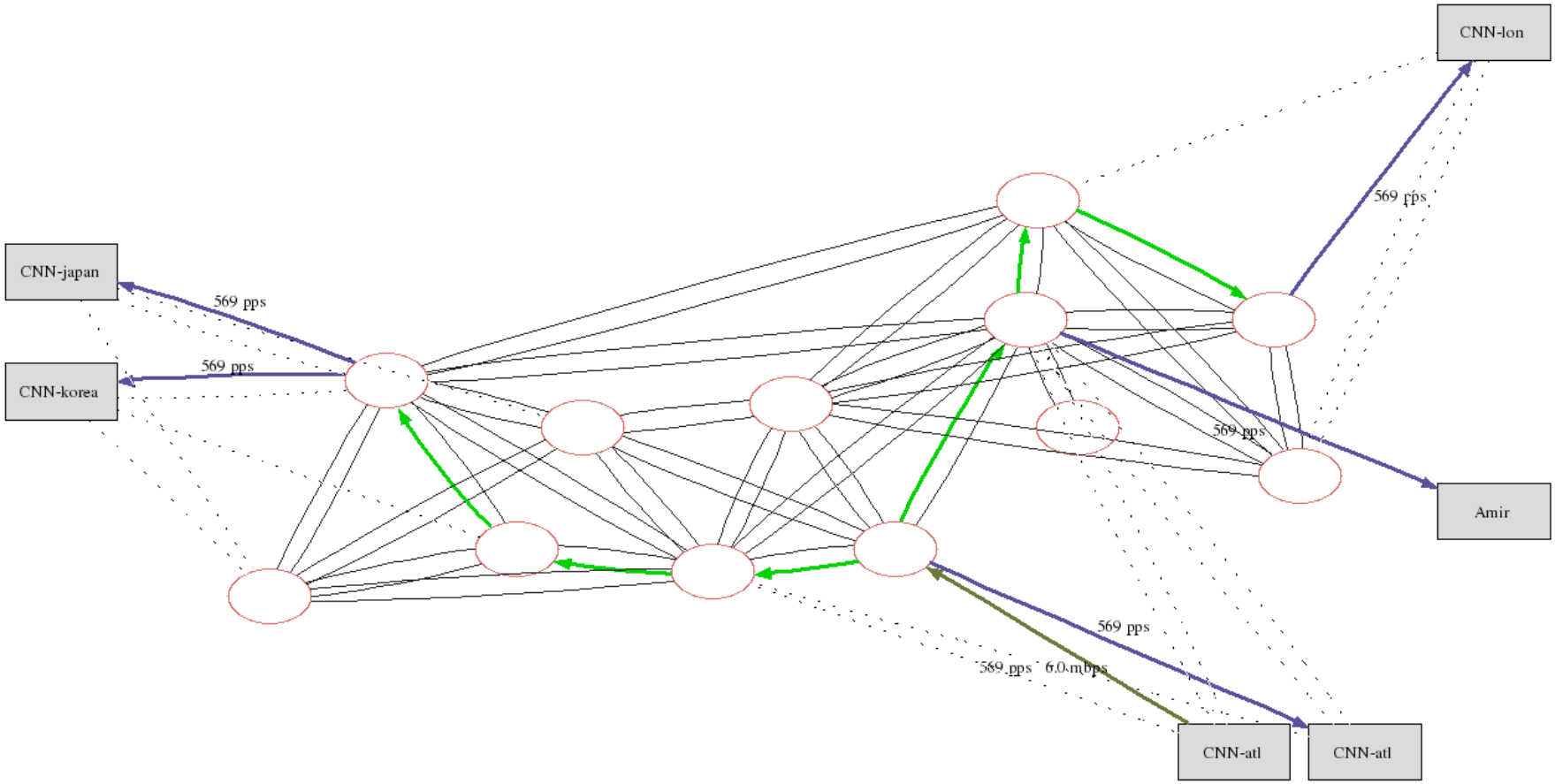
The LiveTimeNet Cloud



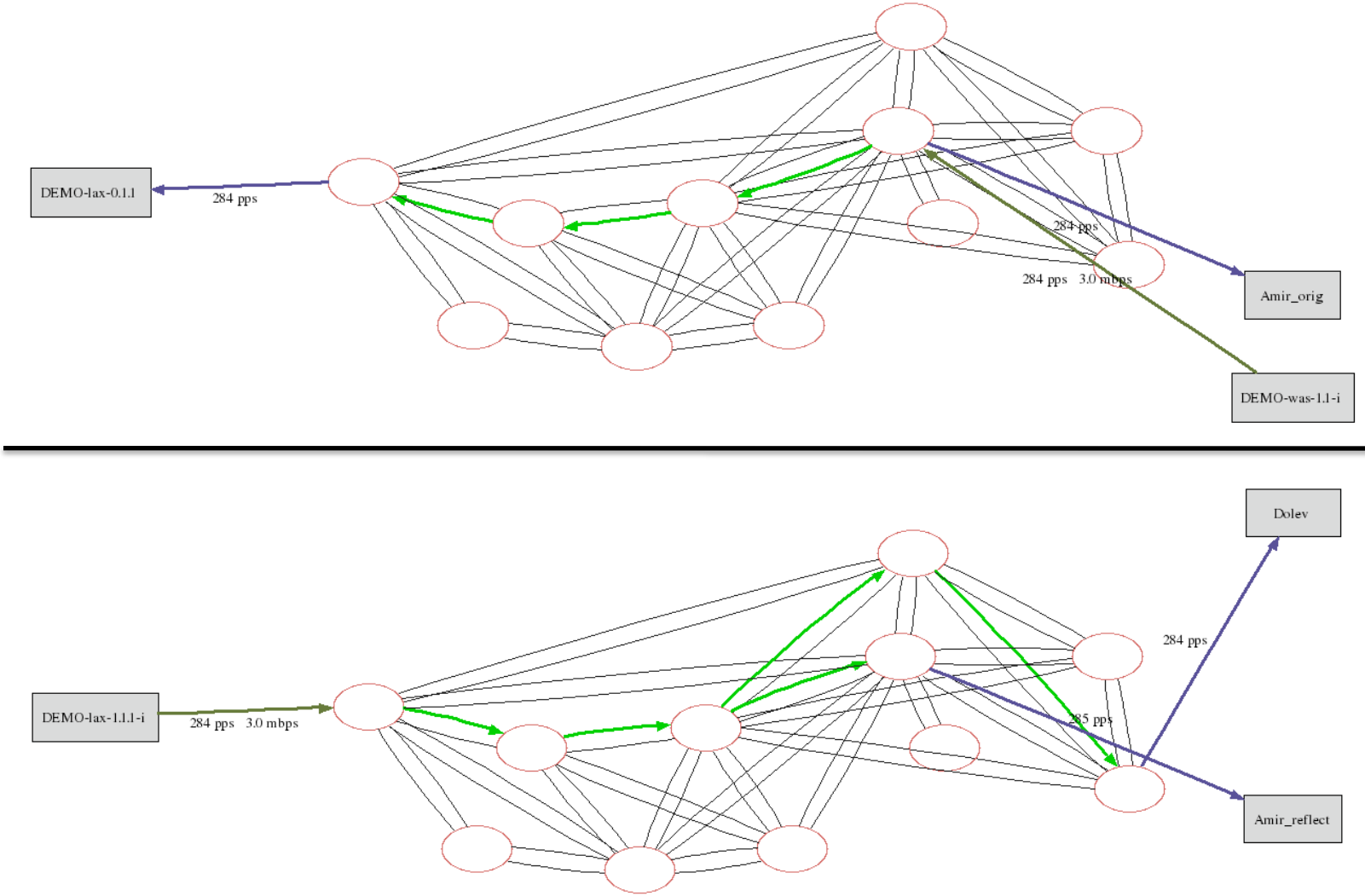
Time for a Demonstration



Time for a Demonstration



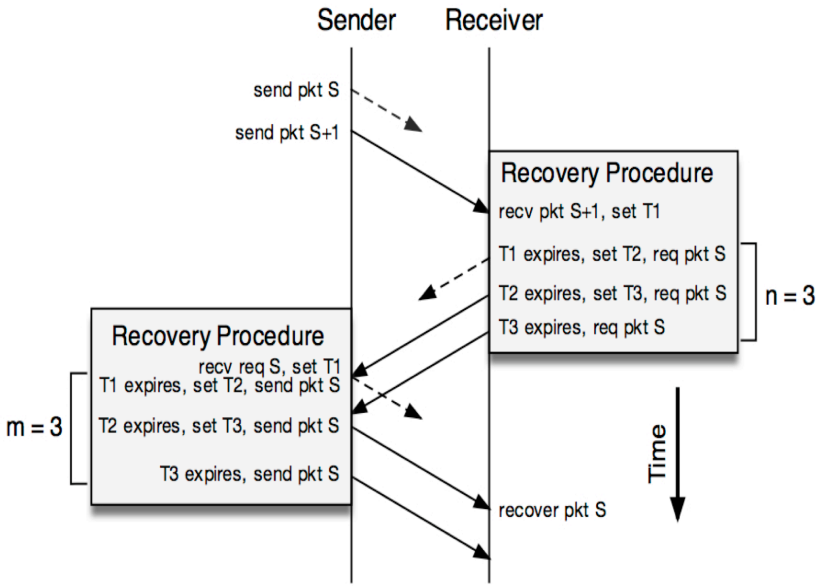
Another Demonstration



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 HD-TV
  - Overlay level: responsive overlay routing
  - Cloud level: NxWay failover for overlay routers

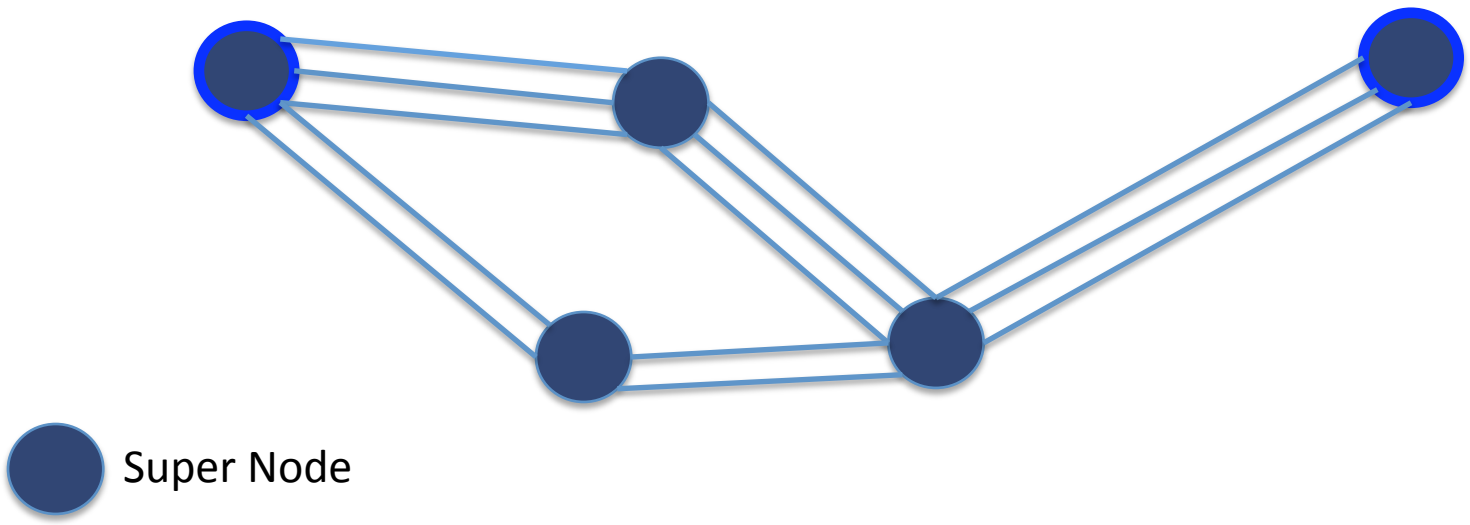
Real-time Protocol for HD-TV



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%

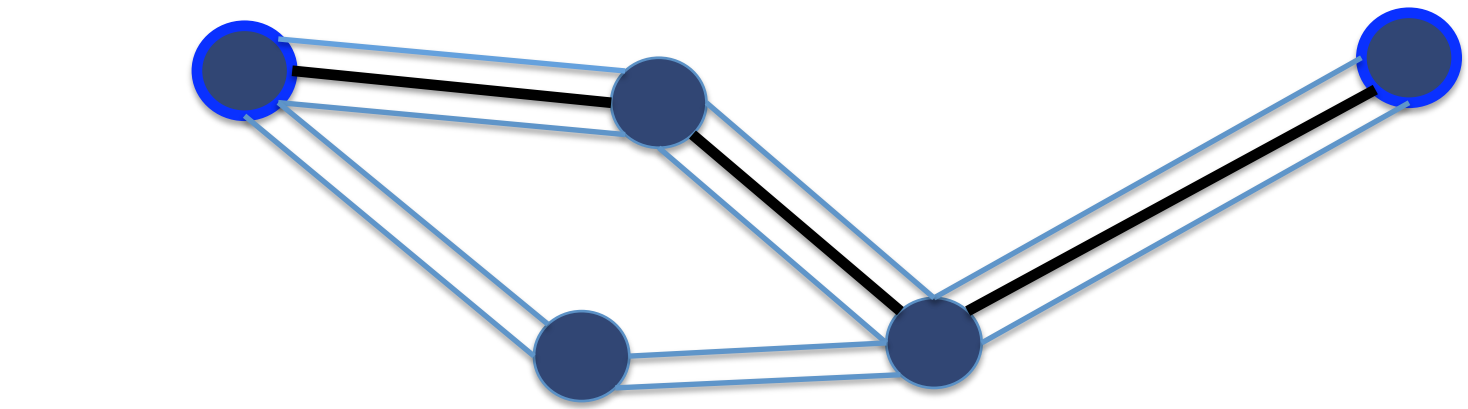
Responsive Overlay Routing

- 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



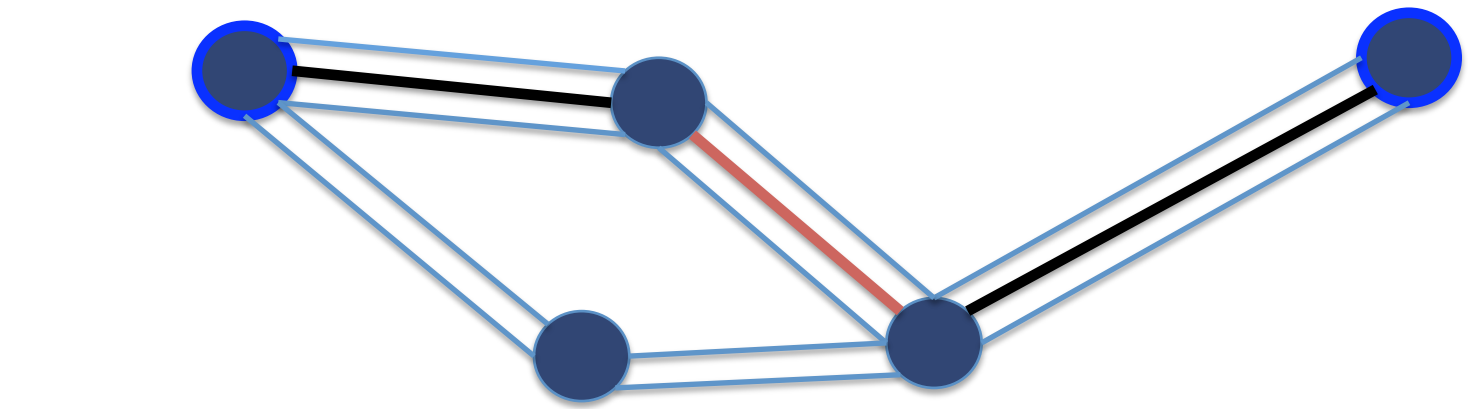
Super Node

Available link

Selected link

Responsive Overlay Routing

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



Super Node

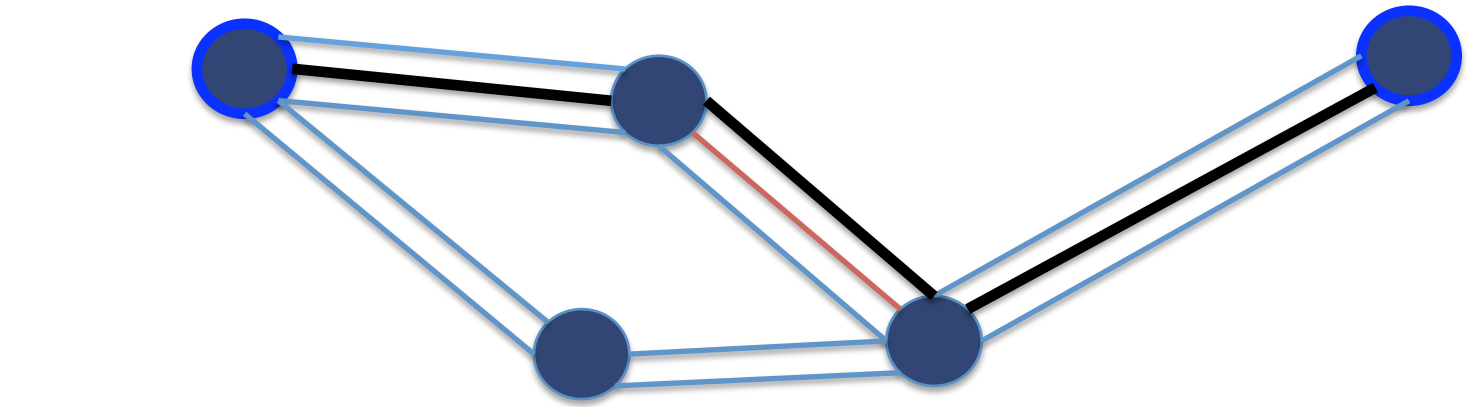
Available link

Selected link

Deteriorating link

Responsive Overlay Routing

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



Super Node

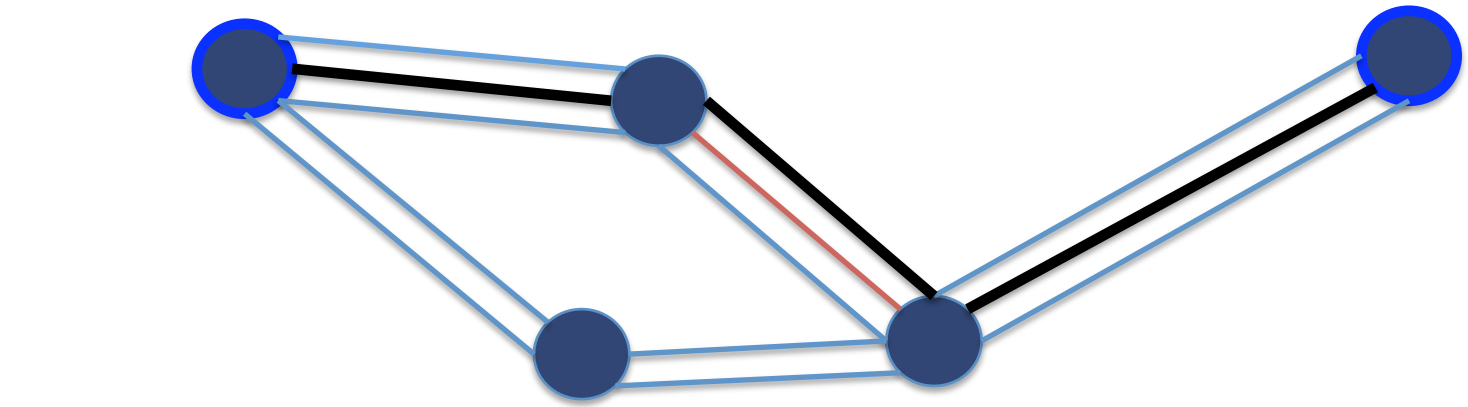
Available link

Selected link

Deteriorating link

Responsive Overlay Routing

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



Super Node

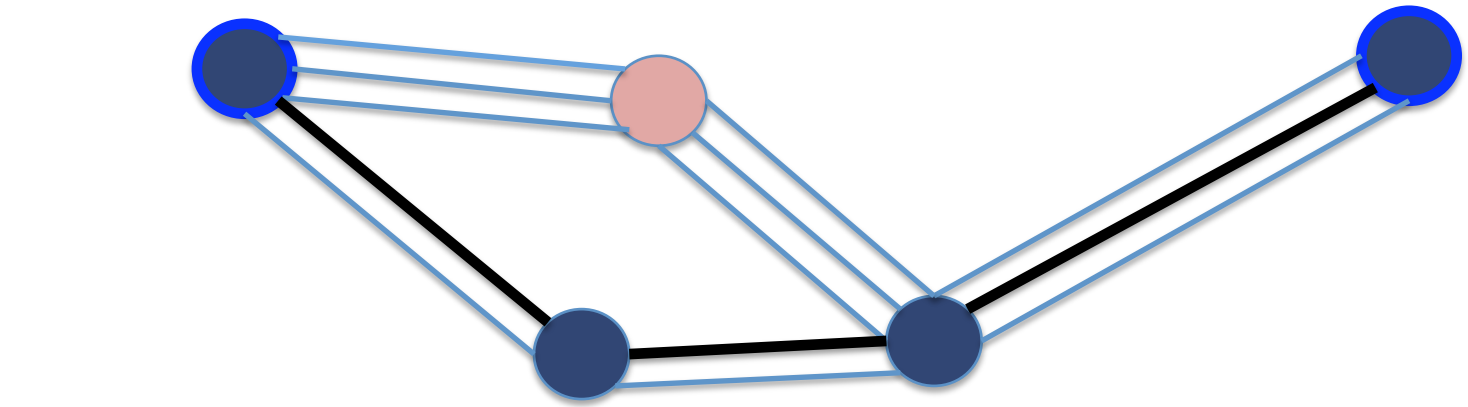
Available link

Selected link

Deteriorating link

Responsive Overlay Routing

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



Super Node

Available link

Selected link

Deteriorating link

Outline

- The Overlay Network Paradigm
- The DARPA Networking Challenge (99-03)
  - Overlay Architecture
  - Low-latency reliable transport
- The Siemens VoIP Challenge (03-06)
  - Almost-reliable, real-time transport
- The LiveTimeNet TV Challenge (08-…)
  - From Overlays to Clouds
  - Ultimate resiliency, automated monitoring and control
- The challenges ahead

The Challenges Ahead

- Resiliency - all the way to intrusion tolerance
  - Resilient clouds
  - Critical infrastructure
  - SKYDA (SCADA in the Sky)
- Timeliness and quality – no end to that
  - Remote manipulation
  - Remote surgery
  - Remote music training ?