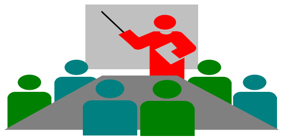
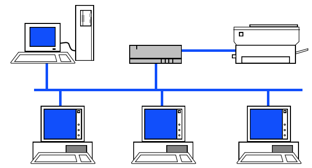




Distributed Systems
601.417
Overlay Networks

Department of Computer Science
The Johns Hopkins University

Overlay Networks

Lecture 4

Further reading:

www.dsn.jhu.edu/publications/

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
  - Already happened: mail to e-mail, Phone to VoIP, Fax to PDFs
  - Ongoing: 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 – the end-to-end principle
  - Keep it simple in the middle …
    - Best-effort packet switching, routing (intranet, Internet)
  - … and smart at the edge
    - End-to-end reliability, naming
- Could therefore adapt and scale
  - Survived for 5 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

A New Generation of Internet Applications

- 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
  - 100ms for VoIP
  - 80-100ms for interactive games
  - 65ms (one way) for remote robotic surgery, remote manipulation
- End-to-end dependability (availability, reliability)
  - From "e-mail" dependability – to "phone service" dependability – to "TV service" dependability – to "remote surgery" dependability
- System resiliency, security, and access control
  - From e-mail fault tolerance – to financial transaction security – to critical infrastructure (SCADA) intrusion tolerance

Addressing New Application Demands: Potential Approaches

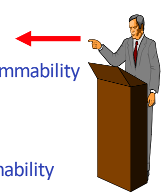
- Build specialized (non-IP) networks
  - Was done decades before the Internet (e.g. TV Infrastructure)
  - Extremely expensive
- Build private IP networks
  - Avoids the resource sharing aspects of the Internet, solves some of the scale issues
  - Expensive
  - Still limited by the basic end-to-end principle underlying the IP service
- Build a better Internet
  - Improvements and enhancements to IP (or TCP/IP stack)
  - "Clean slate design"
  - Long process of standardization and gradual adoption
- Build structured overlay networks

The Structured Overlay Network Vision

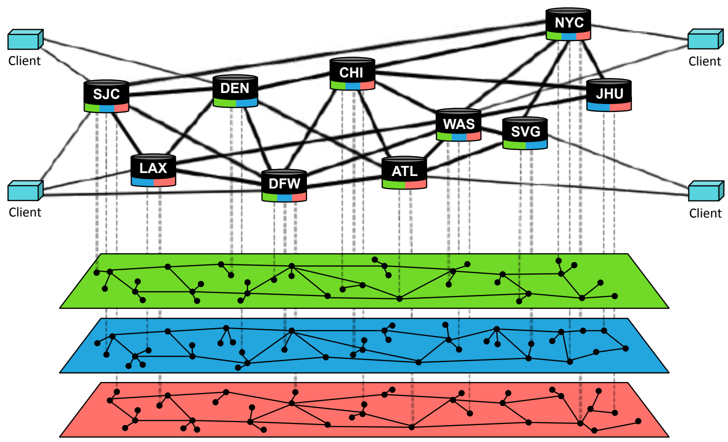
- Key idea: puts processing and context into the middle of the network, providing more flexibility and control
  - At overlay level
  - Underlying network maintains the end-to-end principle
- Three structured overlay network principles:
  - Resilient network architecture
  - Overlay node software architecture with global state and unlimited programmability
  - Flow-based processing

Outline

- A New Generation of Internet Services
- The Structured Overlay Network Vision
  - Resilient Network Architecture
  - Overlay Node Software Architecture with Global State and Unlimited programmability
  - Flow-based Processing
- First Steps and Benefits
  - Responsive Overlay Routing with a Resilient Network Architecture
  - Hop-by-Hop Reliability with Flow-based Processing and Unlimited Programmability
  - Spines – from Concepts to Systems
- The Quest for QoS
  - Almost-reliable real-time protocol for VoIP
  - Almost-reliable real-time protocol for Live TV
- Going even Faster
  - Remote Manipulation, Remote Robotic Surgery, Collaborative Virtual Reality
  - Dissemination Graphs with Targeted Redundancy
- Deploying Structured Overlays on a Global Scale
  - The Service Provider Approach



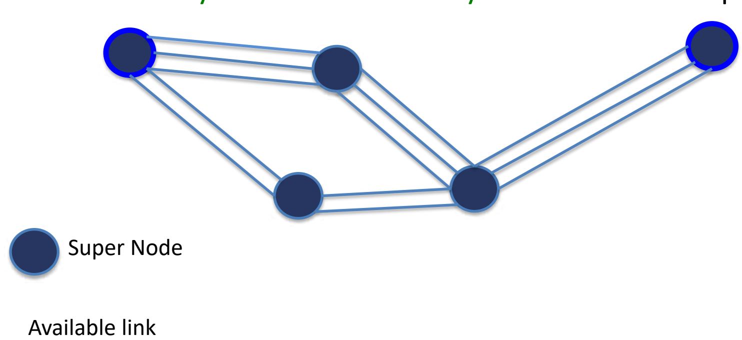
Resilient Network Architecture



U.S. portion of a resilient structured overlay network with overlay nodes located in strategic datacenters

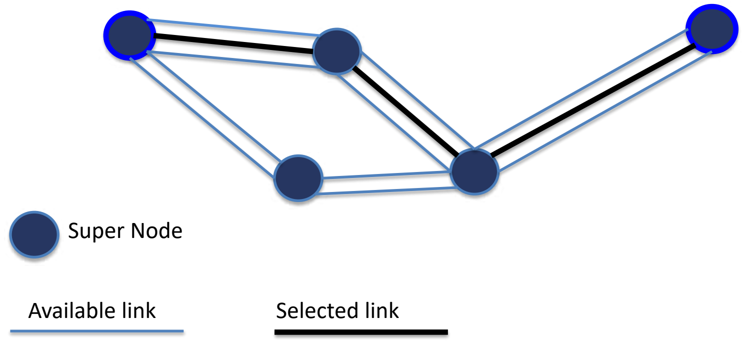
Responsive Overlay Routing with a Resilient Network Architecture

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



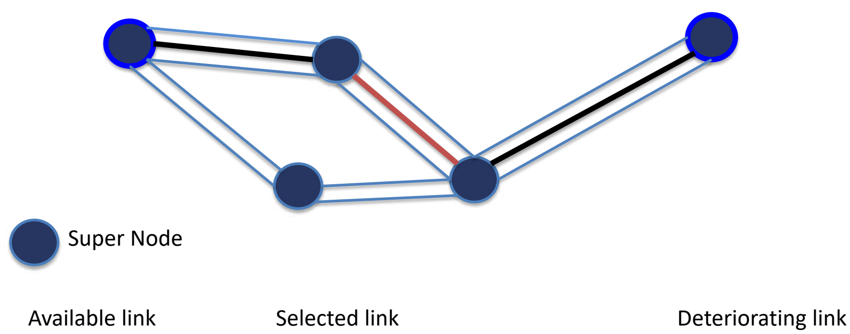
Responsive Overlay Routing with a Resilient Network Architecture

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



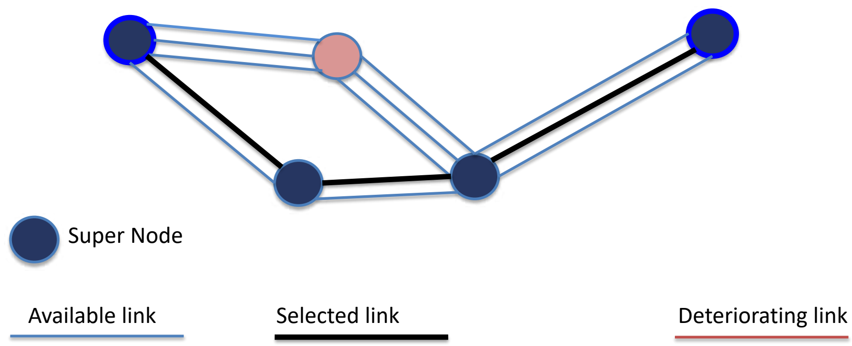
Responsive Overlay Routing with a Resilient Network Architecture

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



Responsive Overlay Routing with a Resilient Network Architecture

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



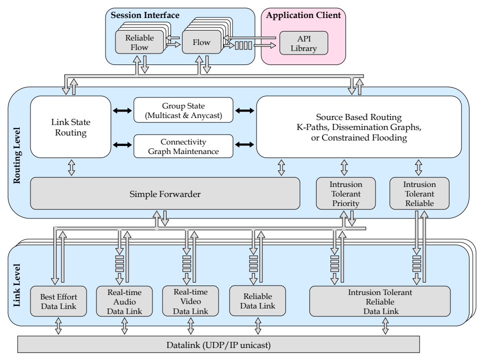
Yair Amir Fall 2021 / Week 4 13

Overlay Node Software Architecture

- Structured overlay messaging system
  - Running overlay software routers (daemons) on top of UDP as user-level internet applications
  - Using commodity servers in strategic datacenters
- Easy-to-use programming platform
  - API similar to the socket API
  - Additional, seamless API through packet interception
- Deployable
  - Vision partially realized by the Spines messaging system (www.spines.org) and its derivatives

Yair Amir Fall 2021 / Week 4 14

Overlay Node Software Architecture



- Global State
  – Possible due to the relatively small number of nodes (e.g. a few tens)
- Unlimited programmability
  – General purpose computers (or clusters) in datacenters
  – Flexible and extensible architecture

Flow-based Processing

- Leverages flow-specific context
  – Hop-by-hop recovery
  – De-duplication of retransmitted or redundantly transmitted packets in the middle of the network
  – Enhanced resiliency through flow-based fairness
- Allows different services to be selected for different application flows

Example: End-to-End Reliability

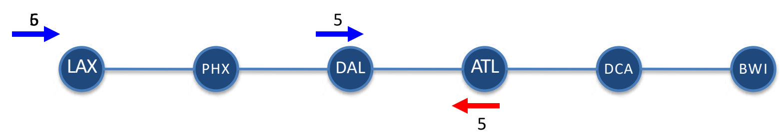
- 50 millisecond network
  - E.g. Los Angeles to Baltimore
  - 50 milliseconds to tell the sender about the loss
  - 50 milliseconds to resend the packet
- At least 100 milliseconds to recover a lost packet

Example: 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 with Flow-based Processing and Unlimited Programmability

- 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

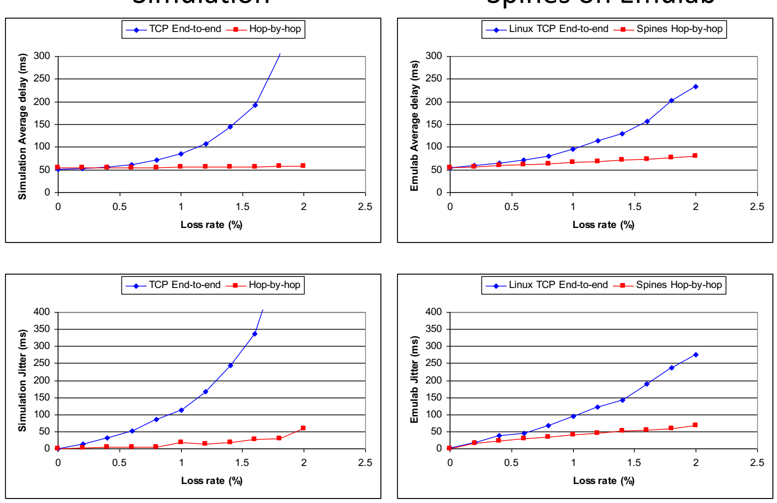


Average Latency and Jitter

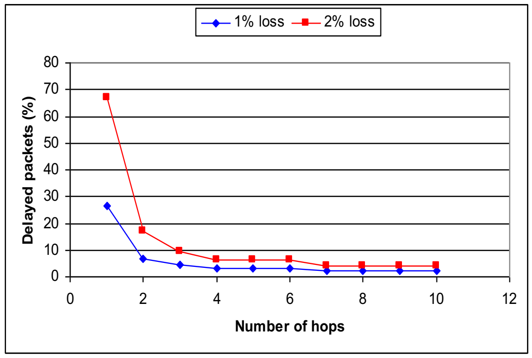
Simulation | Spines on Emulab

Latency

Jitter



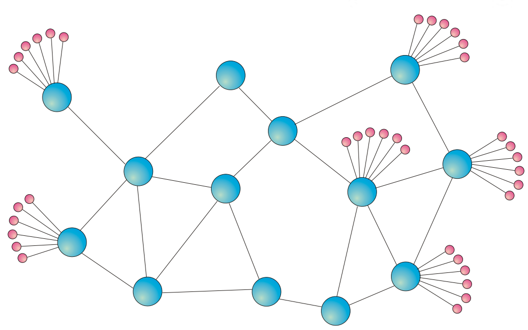
How Dense Should an Overlay Be?



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

Spines – from Concepts to Systems

www.spines.org

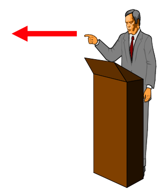


[DSN03, NOSSDAV05, TOM06, Mobisys06, TOCS10, LADIS12, ICDCS16, ICDCS17]

- 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

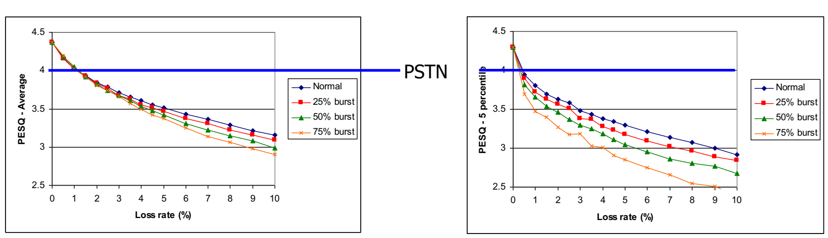
Outline

- Introduction
- The Structured Overlay Network Vision
  - Resilient Network Architecture
  - Overlay Node Software Architecture with Global State and Unlimited programmability
  - Flow-based Processing
- First Steps and Benefits
  - Responsive Overlay Routing with a Resilient Network Architecture
  - Hop-by-Hop Reliability with Flow-based Processing and Unlimited Programmability
  - Spines – from Concepts to Systems
- The Quest for QoS
  - Almost-reliable real-time protocol for VoIP
  - Almost-reliable real-time protocol for Live TV
- Going even Faster
  - Remote Manipulation, Remote Robotic Surgery, Collaborative Virtual Reality
  - Dissemination Graphs with Targeted Redundancy
- Deploying Structured Overlays on a Global Scale
  - The Service Provider Approach



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



50ms network delay

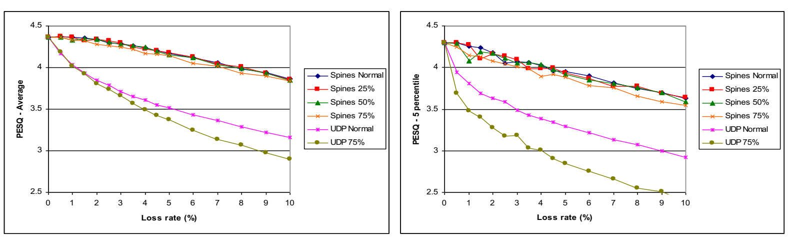
Almost-Reliable Real-time Protocol for VoIP

- 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.

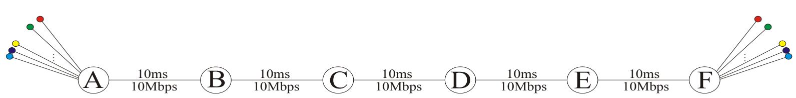
$$loss \approx 2 \cdot p^2 \qquad retr\_delay = 3 \cdot T + \Delta$$

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

VoIP Quality Improvement



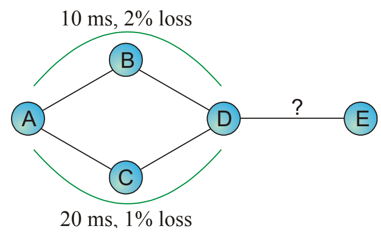
- Spines overlay – 5 links of 10ms each
- 10 VoIP streams sending in parallel
- Loss on middle link C-D



50ms network delay

Real-Time Routing 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

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

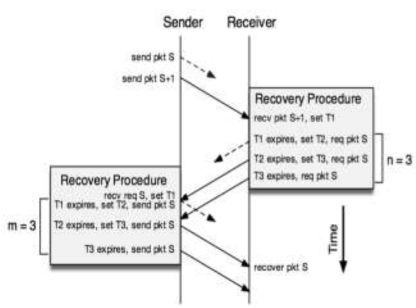
A Structured Overlay Approach to VoIP

- 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

The LiveTimeNet Live TV Challenge

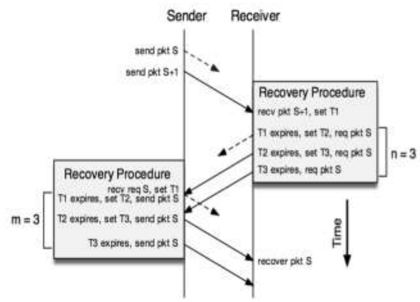
200ms one-way latency requirement, 99.999% reliability guarantee
40ms one-way propagation delay across North America

Almost-Reliable Real-Time Protocol for Live TV



NM-strikes overlay link protocol: guaranteed timeliness, "almost reliable" delivery

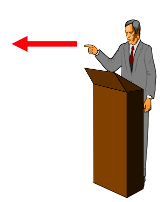
Almost-Reliable Real-Time Protocol for Live TV



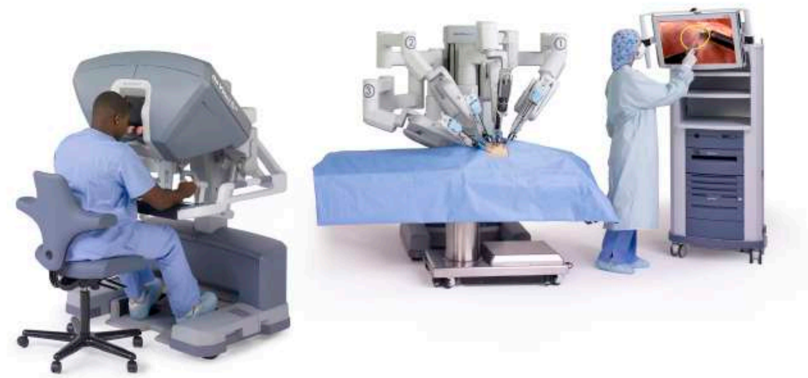
Network packet loss on one link (assuming 66% burstiness)	Loss experienced by flows on the LTN Network
2%	< 0.0003%
5%	< 0.003%
10%	< 0.03%

Outline

- Introduction
- The Structured Overlay Network Vision
  - Resilient Network Architecture
  - Overlay Node Software Architecture with Global State and Unlimited programmability
  - Flow-based Processing
- First Steps and Benefits
  - Responsive Overlay Routing with a Resilient Network Architecture
  - Hop-by-Hop Reliability with Flow-based Processing and Unlimited Programmability
  - Spines – from Concepts to Systems
- The Quest for QoS
  - Almost-reliable real-time protocol for VoIP
  - Almost-reliable real-time protocol for Live TV
- Going even Faster
  - Remote Manipulation, Remote Robotic Surgery, Collaborative Virtual Reality
  - Dissemination Graphs with Targeted Redundancy
- Deploying Structured Overlays on a Global Scale
  - The Service Provider Approach



Remote Robotic Surgery



The Remote Robotic Surgery Challenge

130ms **round-trip** latency requirement (65ms one way latency)
80ms round-trip propagation delay across North America

Addressing the Challenge:
Dissemination Graphs with Targeted Redundancy

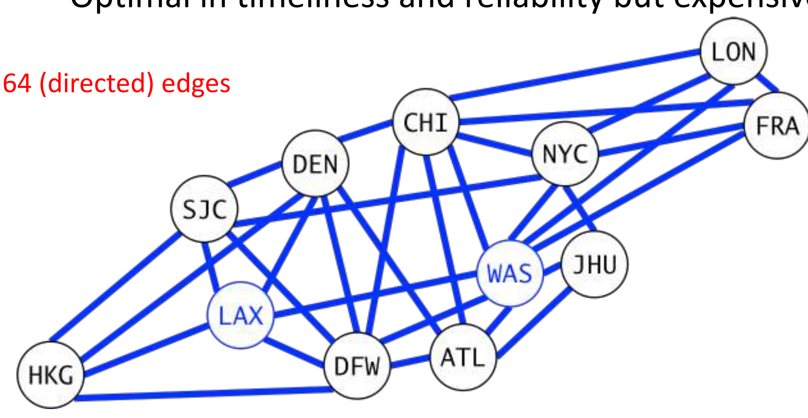
- Stringent latency requirements give much less flexibility for buffering and recovery
  - No more than a single recovery on a single hop
- Core idea: Send packets redundantly over a subgraph of the network (a dissemination graph) to maximize the probability that at least one copy arrives on time

How do we select the subgraph (subset of overlay links) on which to send each packet?

Initial Approaches to Selecting a Dissemination Graph

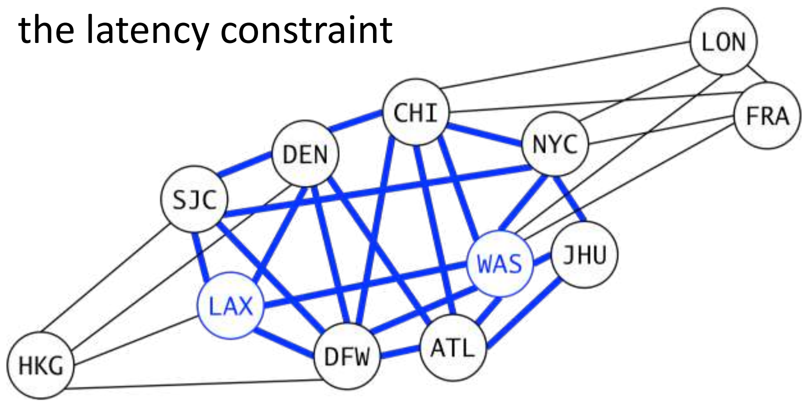
- Overlay Flooding: send on all overlay links
  - Optimal in timeliness and reliability but expensive

64 (directed) edges



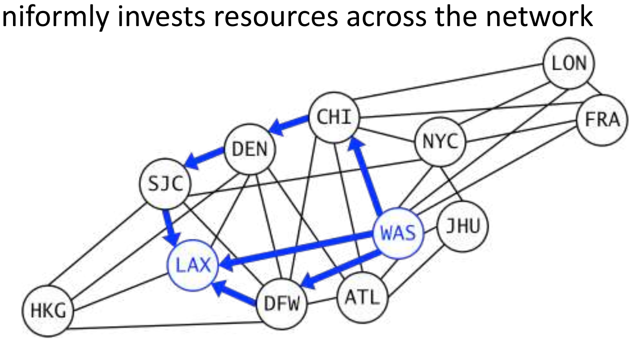
Initial Approaches to Selecting a Dissemination Graph

- Time-Constrained Flooding: flood only on edges that can reach the destination within the latency constraint



Initial Approaches to Selecting a Dissemination Graph

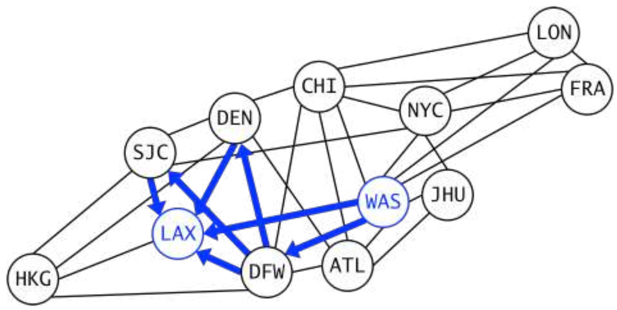
- Disjoint Paths: send on several paths that do not share any nodes (or edges)
  - Good trade-off between cost and timeliness/reliability
  - Uniformly invests resources across the network



Selecting an Optimal Dissemination Graph

Can we use knowledge of current network conditions to do better?

Invest more resources in more problematic regions:



Selecting an Optimal Dissemination Graph Problem Definition

- We want to find the best trade-off between cost and reliability (subject to timeliness)
  - Cost: # of times a packet is sent (= # of edges used)
  - Reliability: probability that a packet reaches its destination within its application-specific latency constraint (e.g. 65ms)
- **Client perspective**: maximize reliability achieved for a fixed budget
- **Service provider perspective**: minimize cost of providing an agreed upon level of reliability (SLA)

Selecting an Optimal Dissemination Graph

- Solving the proposed problems is NP-hard
  - Without the latency constraint, computing reliability is the two-terminal reliability problem (which is #P-complete) [Val79]
  - Computing optimal dissemination graphs in terms of cost and reliability is also NP-hard
  - Exact calculations (via exhaustive search) can take on the order of tens of seconds for practical topologies – cannot support fast rerouting

Data-Informed Dissemination Graphs

- Goal: Learn about the types of problems that occur in the field and tailor dissemination graphs to address common problem types
- Collected data on a commercial overlay topology (www.ltnglobal.com) over 4 months
- Analyzed how different dissemination-graph-based routing approaches (time-constrained flooding, single path, two disjoint paths) would perform (Playback Overlay Network Simulator)

Data-Informed Dissemination Graphs

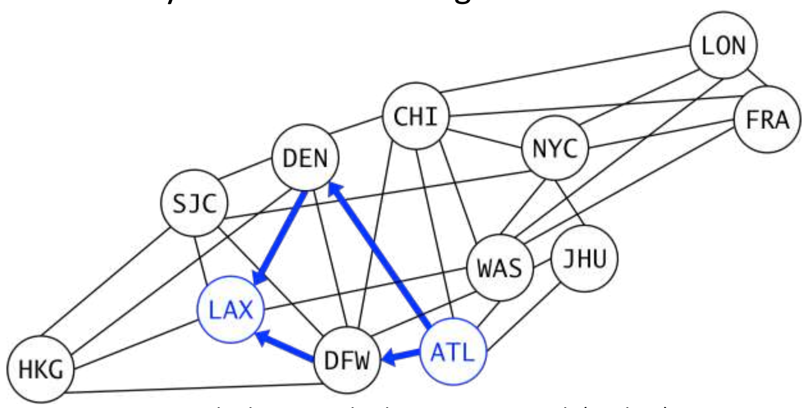
- Key findings:
  - Two disjoint paths provide relatively high reliability overall
    - Good building block for most cases
  - Almost all problems not addressed by two disjoint paths involve either:
    - A problem at the source
    - A problem at the destination
    - Problems at both the source and the destination

Dissemination Graphs with Targeted Redundancy

- Overall approach:
  - Pre-compute four graphs per flow:
    - Two disjoint paths (static)
    - Source-problem graph
    - Destination-problem graph
    - Robust source-destination problem graph
  - Use two disjoint paths graph in the normal case
  - If a problem is detected at the source and/or destination of a flow, switch to the appropriate pre-computed dissemination graph
  - Converts optimization problem to classification problem

Dissemination Graphs with Targeted Redundancy: Case Study

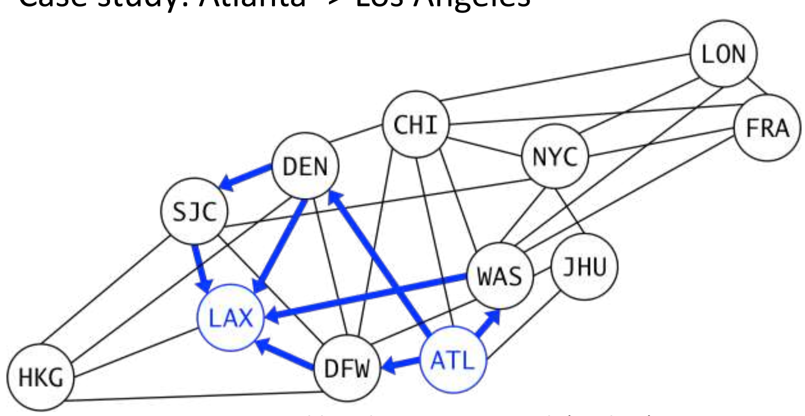
- Case study: Atlanta -> Los Angeles



Two node-disjoint paths dissemination graph (4 edges)

Dissemination Graphs with Targeted Redundancy: Case Study

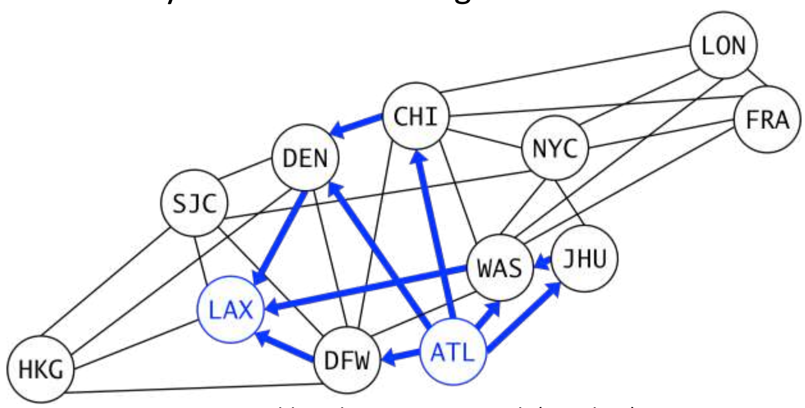
- Case study: Atlanta -> Los Angeles



Destination-problem dissemination graph (8 edges)

Dissemination Graphs with Targeted Redundancy: Case Study

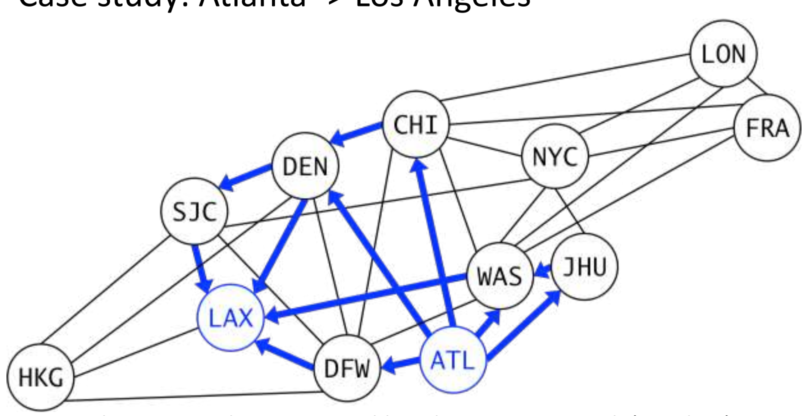
- Case study: Atlanta -> Los Angeles



Source-problem dissemination graph (10 edges)

Dissemination Graphs with Targeted Redundancy: Case Study

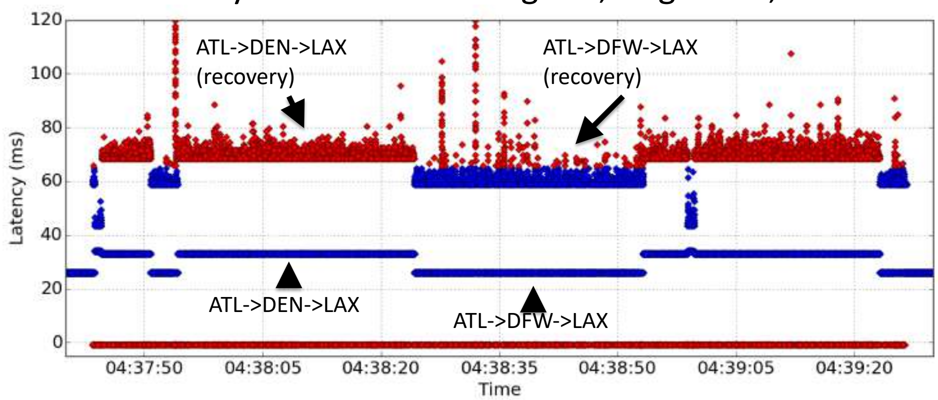
- Case study: Atlanta -> Los Angeles



Robust source-destination-problem dissemination graph (12 edges)

Dissemination Graphs with Targeted Redundancy: Case Study

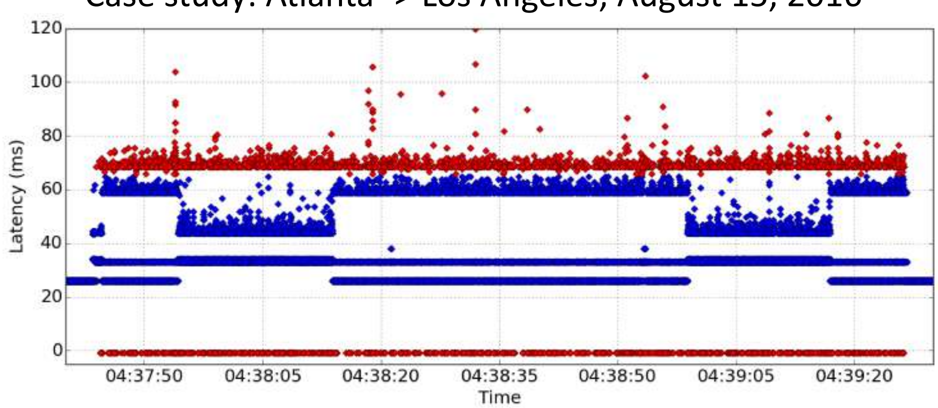
- Case study: Atlanta -> Los Angeles; August 15, 2016



Packets received and dropped over a 110-second interval using dynamic single path (27,353 lost/late packets, 5 packets with latency over 120ms not shown)

Dissemination Graphs with Targeted Redundancy: Case Study

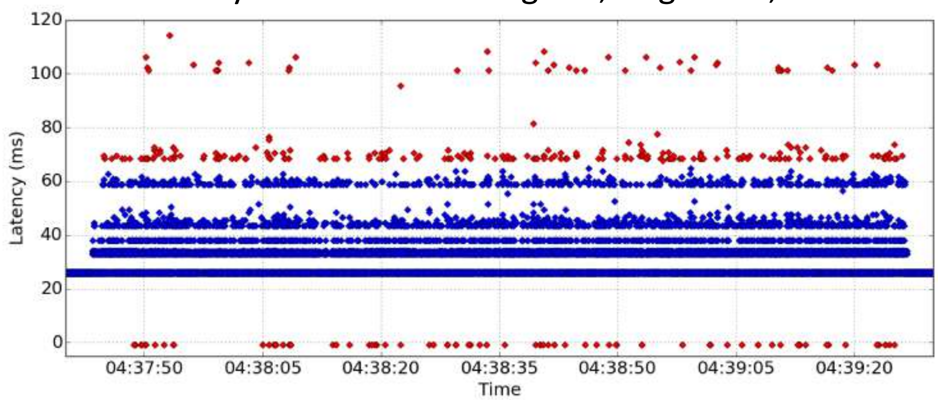
- Case study: Atlanta -> Los Angeles; August 15, 2016



Packets received and dropped over a 110-second interval using dynamic two disjoint paths (5,100 lost/late packets, 15 packets with latency over 120ms not shown)

Dissemination Graphs with Targeted Redundancy: Case Study

- Case study: Atlanta -> Los Angeles; August 15, 2016



Packets received and dropped over a 110-second interval using our dissemination-graph-based approach to add targeted redundancy at the destination (338 lost/late packets)

Dissemination Graphs with Targeted Redundancy: Results

- 4 weeks of data collected over 4 months
  - Packets sent on each link in the overlay topology every 10ms
- Analyzed 16 transcontinental flows
  - All combinations of 4 cities on the East Coast of the US (NYC, JHU, WAS, ATL) and 2 cities on the West Coast of the US (SJC, LAX)
  - 1 packet/ms simulated sending rate

Dissemination Graphs with Targeted Redundancy: Results

Routing Approach	Availability (%)	Unavailability (seconds per flow per week)	Reliability (%)	Reliability (packets lost/ late per million)
Time-Constrained Flooding	99.995883%	24.90	99.999863%	1.37
Dissemination Graphs with Targeted Redundancy	99.995864%	25.02	99.999849%	1.51
Dynamic Two Disjoint Paths	99.995676%	26.15	99.999103%	8.97
Static Two Disjoint Paths	99.995266%	28.63	99.998438%	15.62
Redundant Single Path	99.995223%	28.89	99.998715%	12.85
Single Path	99.994286%	34.56	99.997710%	22.90

Results: % of Performance Gap Covered (between TCP and Single Path)

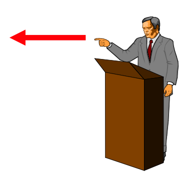
Routing Approach	Week 1 2016-07-19	Week 2 2016-08-08	Week 3 2016-09-01	Week 4 2016-10-13	Overall	Scaled Cost
Time-Constrained Flooding	100.00%	100.00%	100.00%	100.00%	**100.00%**	**14.350**
Dissem. Graphs with Targeted Redundancy	94.19%	99.19%	98.00%	99.50%	**98.97%**	**2.203**
Dynamic Two Disjoint Paths	80.91%	71.34%	47.73%	73.46%	**70.74%**	**2.197**
Static Two Disjoint Paths	-76.72%	50.89%	53.58%	40.79%	**39.50%**	**2.194**
Redundant Single Path	54.12%	37.25%	4.89%	59.10%	**45.75%**	**2.000**
Single Path	0.00%	0.00%	0.00%	0.00%	**0.00%**	**1.000**

Applications: Remote Manipulation

Video demonstration: www.dsn.jhu.edu/~babay/Robot_video.mp4

Outline

- Introduction
- The Structured Overlay Network Vision
  - Resilient Network Architecture
  - Overlay Node Software Architecture with Global State and Unlimited programmability
  - Flow-based Processing
- First Steps and Benefits
  - Responsive Overlay Routing with a Resilient Network Architecture
  - Hop-by-Hop Reliability with Flow-based Processing and Unlimited Programmability
  - Spines – from Concepts to Systems
- The Quest for QoS
  - Almost-reliable real-time protocol for VoIP
  - Almost-reliable real-time protocol for Live TV
- Going even Faster
  - Remote Manipulation, Remote Robotic Surgery, Collaborative Virtual Reality
  - Dissemination Graphs with Targeted Redundancy
- Deploying Structured Overlays on a Global Scale
  - The Service Provider Approach



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
- Guaranteed capacity with admission control
- Monitoring and Control – near automation

The LTN Global Communications Cloud