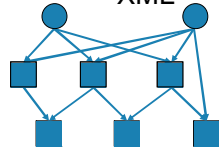


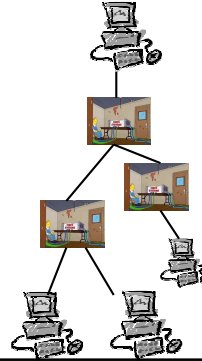
## Mesh-Based Content Routing using XML



Paper by  
Alex C. Snoeren, Kenneth Conley, and David K. Gifford  
SOSP '01

Presented by  
Micah Sherr  
(with several slides from Alex Snoeren)  
January 25th, 2007

## But first, a primer on IP Multicast



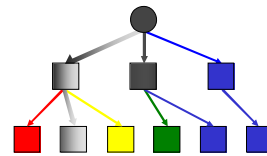
- Clients subscribe to multicast stream
- Routers only forward to interfaces which have subscribers (forms a routing tree)
- IPv4 addresses reserved for multicast
  - Global scope: 224.0.0.0-238.255.255.255
  - Limited scope: 239.0.0.0-239.255.255.255
- Protocols for group management (IGMP)
- Protocols for multicast delivery (PIM, DVMRP)

## Problems with IP Multicast

- Packet loss problematic (reliability fixes often adds significant latency)
- Link/node failure cascades down multicast tree
- All subscribers receive same data stream
- Security and firewall issues



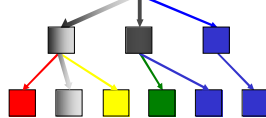
## Mesh-Based Content Routing using XML



- Reliably serve time-critical data streams
- Assumptions:
  - Time-critical data: low-latency more important than bandwidth
  - Clients interested in different content
  - Network (Internet) is failure-prone and lossy

## Motivating Example: Air Traffic Control (ATC) Streams

- Diverse client requests
  - Flights below 30,000 feet
  - UAL flights taking off from PHL
- Time-critical data
  - Available runways for landing at NWK
- Little tolerance for loss
  - JetBlue134 heading towards flight path of NWA513

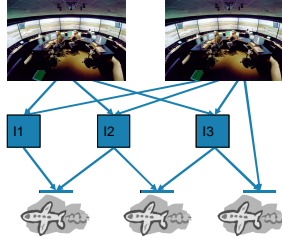


## Mesh-Based Content Routing using XML

- XML Routing: Packets tagged with XML descriptors
  - supports content-based routing
  - publish/subscribe logic in the network
- Mesh-Based Overlay Network
  - redundancy == fault-tolerance
  - redundancy provides low-latency

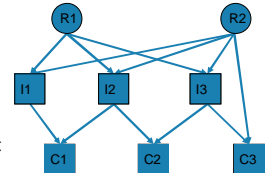
## 10,000 ft overview (no pun intended)

- Root routers (R1,R2) produce content (information providers)
- Internal routers (I1,I2,I3) are intermediary nodes in the network (redundancy providers)
- Clients (C1,C2,C3) are information consumers



## XML Routing

- Root routers *publish tagged XML data streams*
- *Clients subscribe to certain components*
- Internal routers prune content (logic in the network)
  - No need to forward data that isn't needed
  - Requires efficient XML parsing / querying (XQuery or XPath)



## ATC Stream

**Raw Encoding:**  
153014022245CCZVTZ UAL1021 512 290 4928N12003W

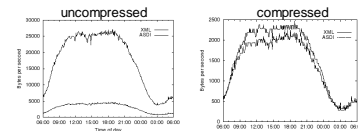
**XML Encoding:**

```

    <?xml version="1.0"?>
    <messageid>153014022245CCZVTZ</messageid>
    <flight>
      <d>UAL1021</d>
      <flightleg status="active">
        <speed type="ground">512</speed>
        <altitude type="reported" mode="plain">290</altitude>
        <coordinate><lat>4928N</lat><lon>12003W</lon></coordinate>
      </flightleg>
    </flight>
  
```

## XML Cost Overhead

- Bandwidth Cost
  - Use compression to mitigate XML bloat
  - XML ATC compresses 10X better than raw ATC data
- Processing Cost
  - Parsing XML and XPath query requires little overhead
    - ~ 65 usecs for parsing
    - ~ 5-15 usecs for query
  - Highly dependent on data stream



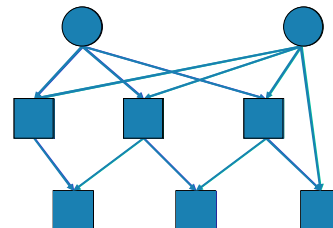
## What about reliability? Handling Failure in the Network

- Component repair or replacement
  - Takes time to detect failure and identify replacement
  - Synchronizing replacement takes additional time
  - Probably not ideal when *JetBlue134* heading towards flight path of *NWA513*

- **Redundancy**
  - Redundant network components
  - Redundant data
  - Bandwidth and synchronization overhead

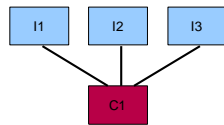
## Mesh Networks

- Leverage redundancy to achieve fault-tolerance and low-latency



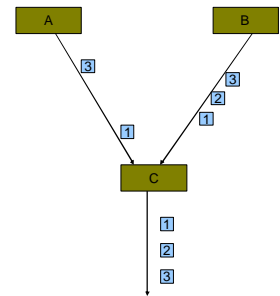
## k-Resiliency

- Every node connected to  $k$  parents, receives duplicate packet stream from each parent
- If graph is acyclic, minimum cut of mesh is  $k$ 
  - mesh resilient to  $(k-1)$  node or link failures
- Requirements for  $(k-1)$ -resilient mesh network:
  - acyclic
  - node needs  $k$  parents
  - paths to parents should be distinct



## Low Latency through Redundancy

- Three ways to improve latency:
  - Increase speed of light (hard to do)
  - Use forward error correction (losses come in bursts)
  - Use redundancy (bandwidth overhead)
- Redundancy reduces latency by
  - Using first arriving packet
  - Prevents need for retransmissions



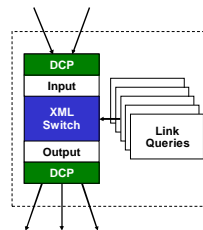
## Diversity Control Protocol (DCP)

- Reassembles packet streams from 1+ senders
- Application Serial Numbers (ANs)
  - associated with packet content, not sender
  - generated at root routers, remain identifiable throughout mesh
  - incremented with each packet
- DCP is a reliability protocol
  - retransmissions sent if missing AN
  - packets buffered and sent in-order at hop
- Supports datagram and stream modes



Figure 4: DCP Packet Header

## Putting it all together: The XML Router



- Input DCP component
  - Maintain parent set
  - Assemble data stream
- XML Switch
  - Parse incoming XML stream
  - Route XML packets based on link queries
- Output DCP component
  - Manage client subscriptions
  - Package and distribute XML streams to clients

## Coping with Loss

- If receiver timeouts waiting for next AN, it transmits *NACK* to all senders
  - Like TCP fast retransmit, timeout interval shorter if future AN received
- Senders resend upon receiving *NACK*
- Assuming independence of sender failures, probability of loss is  $f^k$
- Senders periodically request *ACKs* from clients
  - limits queuing
  - achieves rapid resynchronization

## Mesh Formation and Maintenance: Adding Routers and Clients

1. Initialize set  $S$  to be root routers
2. For each node in  $S$ , send *join request* and *remove node from S*
  - a. If node accepts join, add it to parent set  $P$ . If  $|P|=k$ , stop.
  - b. If node rejects join, ask it for a list of its children, and add them to  $S$ .
3. If  $|S|>0$ , goto 2.

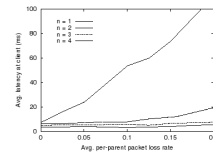
## Mesh Formation and Maintenance: Mesh Repair

- If parent fails, node attempts to join new parent
- Must preserve acyclic mesh:
  - routers keep *level* that is one greater than all of its parents
  - during recovery, node N will join P if  $N_{\text{level}} < P_{\text{level}}$
- Recovers ( $k-1$ ) resilience
- Results in mesh that flattens out over time



## Evaluation

- Implemented small mesh (6 nodes)
- Key findings:
  - Redundancy reduces loss exponentially
  - Redundancy reduces average latency



## Mesh approach outperforms TCP and erasure codes

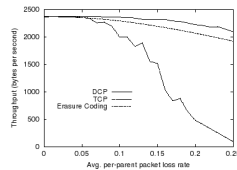
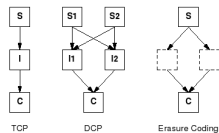


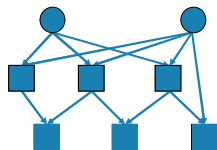
Figure 10: Observed throughput of a two-tier mesh with uniform link loss rates using both 1-resilient DCP and TCP. The stream is served in 262-byte chunks at a rate of 2381 bytes per second. DCP downloads utilize two parents at each tier while TCP can support only one at each tier. We also plot the expected performance of a simple channel-based erasure code using two disjoint, two-hop paths.

## Some limitations

- AN Generation: AN sequences from different root routers must be identical
  - Make ANs partially-ordered; each root router has its own sequence; client performs synchronization
  - block fingerprint matching
- Flow Control: Bandwidth/latency differ at various points in the mesh
- Large jitter could require the use of ACKs: results in ACK implosion

## Summary

- Certain applications require low-latency reliable multicast
- XML Routing enables flexible content-based routing
- Mesh-based Overlay Networks provide both fault-tolerance and low-latency



## Questions?

