Bayeux: An Architecture for Scalable and Fault Tolerant Wide-area Data Dissemination

By
Shelley Zhuang, Ben Zhao, Anthony Joseph, Randy Katz, John Kubiatowicz
Introduction

- Multimedia Streaming typically involves a single source and multiple receivers.
- Unicast and IP multicast not feasible.

Solution

- Application Level Protocols
- Build network of unicast connections and construct distribution trees over it
Introduction

• Bayeux protocol incurs minimum delay and bandwidth penalties and handles fault at both links and routing nodes.

• Utilizes prefix based routing of Tapestry, which is an application level routing protocol.

• Organizes receivers into a tree rooted at the source.

• Provides load balancing across replicated root nodes.
Tapestry

Routing Layer

- Incremental Routing of overlay messages
- Each node has map of multiple levels with each level having a number of entries
- Any destination will be found in logN hops
- Each entry has 3 matches for a given suffix

Data Location

- Each object is associated with a location root
- Server sends publish message to the root
- At each hop, object_id and server_id is stored
- For multiple copies, mapping sorted by distance from node.
Tapestry

Benefits

• Powerful Fault Handling
• Scalable
• Proportional Route Distance
Bayeux Base Architecture

Bayeux session identified by <session_name,UID>

**Session Advertisement**
- Hash the above tuple into a 160 bit unique identifier
- Root or source server creates a file using the identifier
- Advertise it
- Receive messages from interested client

**Tree Maintenance**
- JOIN and TREE messages
- When a router receives TREE message it adds new member to its list of receivers.
- LEAVE and PRUNE messages
Evaluation of Base Design

We compare Bayeux algorithm against IP multicast and naïve unicast.

Performance Metrics

• Relative Delay Penalty: The increase in delay that applications incur while using overlay routing.

• Physical Link Stress: Measure of how effective Bayeux is in distributing network load across multiple links.

For a majority of pair wise connections, RDP is low.

Stress Value is number of duplicate packets going through a link. In Bayeux, overall distribution of link stress is lower and naïve unicast has a much larger tail.
Scalability Enhancements

Source specific model has scalability drawbacks

Tree Partitioning

- Idea is to create multiple roots and partition receivers
- Add Bayeux root nodes to tapestry network
- Put object O in each of the root nodes
- Let each root node advertise O to the tapestry chosen location node
- On JOIN, client gets O from its nearest root node
- No need of periodic advertisements by roots
- See Graph for number of join request handled per root as number of roots increase
Scalability Enhancement

Receiver Identifier Clustering

- Aim is to reduce packet duplication
- Delivery of packets approaches destination digit by digit
- Local nodes should share longest possible suffix
- Packet duplication is thus delayed till LAN is reached thus bandwidth consumption at intermediate nodes is reduced
Fault Resilient Packet Delivery

- At each router, every outgoing hop has 2 backup pointers
- See figure for reachability comparison with IP

- Another aspect of tapestry is hierarchical routing
- Each hop decreases expected number of next hops by a factor equal to the base of tapestry identifier
- Paths converge to the destination in $\log N$ hops
- Intentionally fork of duplicates onto secondary and primary paths expecting them to merge quickly
Fault Resilient Packet Delivery

- Proactive Duplication
- Application Specific Duplication
- Prediction Based Selective Duplication
- Explicit Knowledge Path selection
- First Reachable Link selection

NOTE:

Each of the first three create duplicate packets. But the duplicates converge quickly.

Duplicate suppression is done using sequence numbers.
First Reachable Link Selection

• Delivers packets with high reliability in face of link failures
• No packet duplication
• Overhead in the form of Bandwidth used for transmitting membership information
• Size of membership state transmitted decreases for routers further away from the root node
• Delay for multicast data directly proportional to size of member state transmitted
Conclusion

• Bayeux is an architecture for Internet Content distribution that leverages Tapestry an existing fault tolerant routing infrastructure.

• Bayeux shows that efficient network protocol can be designed with simplicity while inheriting desirable properties from underlying application infrastructure.