Architecture for Optimistic Replication over P2P networks

Stéphane WEISS

March 5, 2012

Optimistic replication

Optimistic replication:

- Each site is uniquely identified and hosts data replicas,
- Modifications can be processed on any replicas,
- Modifications are sent to all other replicas,
- Received modifications are integrated.

Dissemination properties

Consistency relies on the following properties:

- Messages are delivered to all sites
- No message is delivered more than once
- Deliveries in causal order

P2P System

- Very large and unknown number of nodes
- Users are supposed to work at one node of the network
- Partial replication:
  - a document is only replicated on a subset of the nodes
- Any user can be access and modify any document

Basic problems

- Distribute the data
- Search a document
- Ensure that modifications will reach all nodes interested in one document exactly one time
- Deliver modifications in causal order
- Receive the minimum of modifications they are not interested in

Basic problems

- Distribute the data
- Search a document
- Ensure that modifications will reach all nodes interested in one document exactly one time
- Deliver modifications in causal order
- Receive the minimum of modifications they are not interested in

Diffusion “Many-to-Many” tackled by Pub/Sub approaches

Publish Subscribe model

- 2 roles:
  - Publisher
  - Subscriber
- 2 types:
  - Topic-based
  - Content-based

Publish Subscribe model

- 2 roles:
  - Publisher
  - Subscriber
- 2 types:
  - Topic-based
  - Content-based

P2P Pubsub

Network:

- Unstructured:
  - Partial view of the network,
  - Topic connectivity, small topic diameter, low node degree (Min-TCO)
- DHT
  - StoreSub:
    - Subscribers' interests are stored on the DHT
    - Publishers look for interested subscribers
  - StorePub
    - Publishers announce them-selves
    - Subscribers choose publishers

Spidercast [Chockler07]

In a Gossip protocol, each node:

- maintains a partial view of the network,
- periodically selects a few nodes from his local view to exchange some information,

Spidercast [Chockler07]

In a Gossip protocol, each node:

- maintains a partial view of the network,
- periodically selects a few nodes from his local view to exchange some information,

In Spidercast, each node

- periodically exchanges their knowledge about existing nodes and the topics,
- maintains a list of K nodes per topic he is interested in using 2 heuristics:
  - random: selects randomly a node that increases the number of K-covered topics,
  - greedy: selects a node that minimizes the number of topics that are not K-covered.

Messages propagation

About messages propagation for a given topic:

- an epidemic protocol can be used,
- properties ensured:
  - probabilistic guarantees that a message will be delivered to all nodes,
  - a message can be received several time,
  - causality?

Summary on Spidercast

- Creates a low diameter subgraph per topic
- Scalable (10,000 nodes, 1,000 topics, 70 subscriptions)
- What if all nodes from the same topic leave?

Summary on Spidercast

- Creates a low diameter subgraph per topic
- Scalable (10,000 nodes, 1,000 topics, 70 subscriptions)
- What if all nodes from the same topic leave?
- Can be used for systems where all nodes are active

Magnet [Girdzijauskas10]

- Based on two DHTs:
    - Uniform hash function (interest-aware membership, document availability)
    - Non-uniform hash function (OSCAR DHT)
- Creates a multicast tree per topic

Clustering users

OSCAR DHT:

- Cluster of users with similar subscriptions:

$$sim(s_1, s_2) = \frac{\mid s_1 \cap s_2 \mid}{\mid s_1 \cup s_2 \mid}$$

- Join next to the closest node
- Dynamic clustering

Propagation of changes

- Multicast tree with several roots
    - Reach all nodes
    - Deliver one time
- A priori: no message ordering

Summary on Magnet

- Pub/Sub based on two DHTs
- Scalable (10,000 nodes, 3,000 topics, 1 to 384 subscribers)
- Allows document persistence
- Mainly accessed in read
- Maybe too costly for small and/or dynamic group

Conclusion

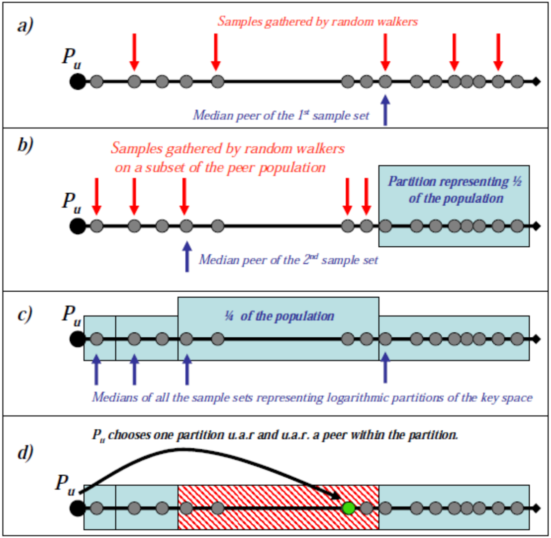
- Existing P2P Pub/Sub approaches can be used for STREAMS:
  - Spidercast for active collaboration
  - Magnet for large dissemination
- Open problems
  - Ensuring causality
  - Join procedure
  - Recovery mechanism

References

[Chockler07] G. Chockler, R. Melamed, Y. Tock, R. Vitenberg
SpiderCast: A Scalable Interest-Aware Overlay for Topic-Based Pub/Sub Communication
In *DEBS'07*, pages 14–25, Toronto, Ontario, Canada, June 2007. ACM Press.

[Girdzijauskas10] S. Girdzijauskas, G. Chockler, Y. Vigfusson, Y. Tock, R. Melamed
Magnet: Pratical Subscription Clustering for Internet-Scale Publish/Subscribe
In *DEBS'10*, pages 172–183, Cambridge, UK, July 2010. ACM Press.

Oscar DHT


a)
Samples gathered by random walkers
P_u
Median peer of the 1st sample set
b)
Samples gathered by random walkers
on a subset of the peer population
Partition representing ½ of the population
P_u
Median peer of the 2nd sample set
c)
P_u
¼ of the population
Medians of all the sample sets representing logarithmic partitions of the key space
d)
P_u chooses one partition u.a.r and u.a.r. a peer within the partition.
P_u