Data Distribution in Large-Scale Distributed Systems

Roberto Baldoni
MIDLAB Laboratory
Università degli Studi di Roma “La Sapienza”

ReSIST: Resilience for Survivability in IST
First Open Workshop
Budapest 21-3-2007

What is a Large-Scale Distributed System?
What is a large-scale distributed systems?

Internet-scale Applications
- unmanaged environment
- Shortlife peers
- High churn

Enterprise Data centers
- managed environment
- longlife peers
- low churn

Scalable QoS-Constrained Applications
- partially managed environment
- shortlife peers at network edges, longlife peers in the core
- high churn only at network edges, low churn in the core
What is the ideal software substrate for Large-Scale Distributed Systems?

Each application has requirements that impact the design of the overlay.

Overlay Networks Substrate as superimposition of graphs

- **Structured overlay**
  - small-world
  - ring
  - Chord [SMKKB03]
  - hypercube
  - ring
  - Pastry [RD01]

- **Unstructured overlay**
  - Skip-list
  - random
  - GosSkip [GHHKFR06]

- **Hybrid overlay**
  - aggregation
  - tree
  - random ring
  - ring
  - BISE [TCS06]
Using publish/subscribe systems for Data Dissemination

**Publishers:** produce data in the form of **events**.

**Subscribers:** declare interests on published data with subscriptions.

Each **subscription** is a filter on the set of published events.

An **Event Notification Service** (ENS) notifies to each subscriber every published event that matches at least one of its subscriptions.

Interaction between publishers and a subscribers is **decoupled in space, time** and **flow**.

---

Two main models are considered in the literature

**Topic-based selection**

- Each event published in the system is tagged with a **topic** that completely characterizes its content.
- Each subscription contains a topic which the subscriber is interested in.

**Content-based selection**

- Each event published in the system is a collection of pairs `<attribute, value>`
- Each subscription is a conjunction of constraints over attributes.
Scalable Data Distribution based on Overlay networks

Internet-scale Applications
• Scribe [CDKR02], Pastry...
• Sub2Sub [VRKS06]
• TERA [BBQQVT07]

Enterprise Data centers
• BISE [TCS06]
• QuickSilver [OB07]

Scalable QoS-constrained applications
• Data Distribution Service (OMG)
• Control Plane (P2P SIP)

Internet-Scale Data Distribution

- In a peer-to-peer environment peers play both the roles of publishers/subscribers and event brokers.

- Trivial solution to the problem of event dissemination:
  - Each event is broadcasted in the network.
  - Subscription-based filtering is performed locally.

- This usually implies a great waste of resources (on the network and on the nodes)

- The semantics of the publish/subscribe paradigm can be leveraged to confine the diffusion of each event only in the set of matched subscribers without affecting the whole network (traffic confinement)
Internet-Scale Data Distribution: Traffic confinement

Traffic confinement can be realized solving three problems:

- **Interest clustering**
  Subscribers sharing similar interests should be arranged in a same cluster; ideally, given an event, all and only the subscribers interested in that event should be grouped in a single cluster.

- **Outer-cluster routing**
  Events can be published anywhere in the system. We need a mechanism able to bring each event from node where it is published, to at least one interested subscriber.

- **Inner-cluster dissemination**
  Once a subscriber receive an event it can simply broadcast it in the cluster it is part of.

---

Current solutions: Scribe

- **Scribe** [Castro et al., IEEE Journal on Selected Areas in Communications n.8 v.20, 2002]
  - Topic-based publish/subscribe implemented on top of DHTs.
  - For each topic a single node is responsible to act as a rendez-vous point between published events and issued subscriptions.
  - Problems:
    - single points of failure
    - hot spots
    - partial traffic confinement
Current solutions: Sub-2-Sub

- **Sub-2-Sub** [Voulgaris et al., International Workshop on Peer-to-Peer Systems, 2006]
  - Content-based publish/subscribe
  - Complex three level infrastructure.
  - Employs clustering: brokers with similar interests are clustered in a same overlay.
  - Similarity is calculated checking intersections among subscriptions.
  - Problems:
    - depending on subscription distribution a huge number of distinct overlays must be maintained
    - the number of overlay networks a single node participates to is not proportional to the number of subscriptions it stores
TERA: Topic-based Event Routing for p2p Architecture

- A two-layer infrastructure:
  - All clients are connected by a single overlay network at the lower layer (general overlay).
  - Various overlay network instances at the upper layer connect clients subscribed to same topics (topic overlays).

- Event diffusion:
  - The event is routed in the general overlay toward one of the nodes subscribed to the target topic.
  - This node acts as an access point for the event that is then diffused in the correct topic overlay.

TERA: outer-cluster routing

- Event routing in the general overlay is realized through a random walk.
- The walk stops at the first broker that knows an access point for the target topic.

<table>
<thead>
<tr>
<th>topic</th>
<th>AP</th>
</tr>
</thead>
<tbody>
<tr>
<td>t</td>
<td>B1</td>
</tr>
<tr>
<td>y</td>
<td>B6</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>topic</th>
<th>AP</th>
</tr>
</thead>
<tbody>
<tr>
<td>x</td>
<td>B1</td>
</tr>
<tr>
<td>a</td>
<td>B5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>topic</th>
<th>AP</th>
</tr>
</thead>
<tbody>
<tr>
<td>e</td>
<td>B4</td>
</tr>
<tr>
<td>h</td>
<td>B4</td>
</tr>
</tbody>
</table>
TERA: Architecture

OMPs: Newscast, Cyclon, etc.

TERA Results: Outer-cluster routing

- We want every topic to appear with the same probability in every APT, regardless of its popularity.
TERA Results: traffic confinement

Which is the probability for an event to be correctly routed in the general overlay toward an access point?

- Depends on:
  - Uniform randomness of topics contained in access point tables.
  - Access point table size.
  - Random walk lifetime.

Random Walk success rate.

Conclusions

- Scalable Data Distribution based on Overlay networks for Internet-Scale applications
  - What is a large scale distributed systems
  - P2P Overlay networks as the ideal substrate for
    - Internet-scale applications
    - Enterprise datacenter applications
    - Scalable QoS-constrained applications

- TERA: Topic-based Event Routing for p2p Architecture
  - outer-cluster routing

- Joint activities within RESIST
  - Composing gossiping: a conceptual architecture for designing gossip-based applications. R. Baldoni, H. L. J. Pereira, E. Rivière (Submitted paper)
  - Looking for a Definition of Dynamic Distributed Systems, R. Baldoni, M. Bertier, M. Raynal, and S. Tucci-Piergiovanni (submitted paper)