CS227 Topics in Database Management
Publish and Subscribe

Don Carney, Qianyong Chen, Allison Lee

March 20, 2001

I. Introduction

- Describe Publish-Subscribe at very high-level (abstract)
- figure - abstract picture of pub/sub
- Describe figure and introduce notion of a "profile"
- Why is there publish/subscribe - general discussion of problem (scalability)

A. The problem
   - What is the scalability problem
     - Polling
   - What is the asymmetry problem/circumstance
   - Why are scalability and asymmetry a cause for pub/sub
     - describe the difficulty of the traditional dissemination technology to be applied to wide-area network.

B. publish/subscribe models
   - topic-based (subject-based) publish/subscribe
   - content-based publish/subscribe
   - compare the publish/subscribe models, explain the efficiency and difficulties of these models

C. Context (how does pub/sub fit into framework of data delivery) (context can refer to models in previous section)
   - set up context and data delivery options with respect to pub/sub
   - describe push and pull
   - compare push and pull
     - push - data delivered to clients automatically
     - pull - clients have to pro-actively search for data
   - periodic and aperiodic (when is data sent - with or without schedule)

- unicast and 1-to-N - most pub/sub today is unicast (except on LANs)
- figure of delivery options and description
- where does publish subscribe fit in framework
- our definition of publish-subscribe and why we have defined it as such

D. Compare and Contrast traditional query-based systems with Publish/Subscribe

E. History of Publish/Subscribe (very high-level, low detail discussion)

- Publish/Subscribe has been put to some good uses
  - Information Bus
  - SIFT
  - Pointcast (media made a big deal of this but it failed)

II. Technology Issues

A. Broadcast, multicast, unicast

- describe broadcast, multicast, and unicast
- how do these three mechanisms impact the scalability of a publish-subscribe system
- Wide area vs. Local Area

B. Internet vs. Intranets

- Give definition of Internet and Intranets. How is Publish/Subscribe applied to Internet vs. Intranets?
- Give examples.
- Problems and solutions

C. Channelization (matchmaking data with clients)

- What are channels - can also be thought of as categories
- What is channelization in the context of publish-subscribe?
- Why is this a hard problem?
- discussion of Katz, McCanne solution

D. Profiles & Continuous Queries

- Profiles
  - Ways to express profiles
  - keywords
  - "better" queries
  - IR method (boolean-based & similarity-based)
- Filtering information
  - effectiveness of profiles vs. efficiency of filtering

- given large numbers of users and large amounts of data (as is typical of publish-subscribe systems) efficiency and scalability are important concerns.
    - IR community's solution does not scale
- Continuous Queries
    - studied in the database community
    - recent case studies
        * OpenCQ
        * NiagaraCQ
    - How do solutions scale?
- Reversing the Role of queries and data (Indexing Profiles)
    - SIFT was the first to reverse these roles
    - In traditional DB systems the data items are indexed and stored and queries are applied individually.
    - In dissemination systems, large numbers of queries are stored and indexed and the data is individually applied. The data becomes the query.
    - Indexing Profiles
        * Sift
        * XFilter

E. Mobile Computing

- Reliable vs. unreliable networks
    - mobile computing
    - delivery guarantee
- Scheduling (fundamental performance issue)
    - fundamental performance issue
    - Broadcast Disks
    - battery life of mobile devices

III. Software & System Architecture for Publish/Subscribe

- classical centralized architecture (Elvin)
- an extension of centralized architecture: distributed components are connected in a hierarchical structure.(TIBCO)
- distributed event service with a generic topology .(SIENA)

compare those architectures, explain the efficiency and the difficulty of scalability.

IV. Applications

- Introduction to this section
  What applications are developed using publish/subscribe technology and architecture?

A. Newsgroups/Mailing lists

- How is publish/subscribe technology applied to newsgroups/mailing lists?
  - clients submits profiles
  - data is delivered from 1-to-N (multicast), through internet, intranets, or channels
  - applies pull, push (data delivered automatically) technologies
  - data can be delivered periodically or aperiodically
  - needs continuous operation - "24 by 7"
- Examples -
- Problems and solutions

B. Channels

- How is publish/subscribe technology applied to channels?
  - clients submits profiles
  - data is delivered from 1-to-N (multicast), through internet, or intranets
  - applies pull, push technologies
  - data can be delivered periodically or aperiodically
  - needs continuous operation - "24 by 7"
  - real-time
- Examples - Marimba's Castanet, Netscape's netcaster, Microsoft's CDF
- Problems and solutions

C. Software distribution

- How is publish/subscribe applied to software distribution?
  - 1-to-N (multicast), internet, intranets, channels
  - pull, push
  - periodic, aperiodic
  - continuous operation - "24 by 7"
  - real-time (for security trading software)
  - dynamic system update
- Examples - microsoft
- Problems and solutions
  - compared to Newsgroups/Mailing lists

D. Stock and News

- How is publish/subscribe applied to stock and news?
  - clients submit profiles
  - 1-to-N (broadcast, multicast), internet, intranets, channel
  - pull, push

- periodic, aperiodic
- continuous operation - “24 by 7”
- real-time

- Examples - Election results, USENET News, Pointcast, TIBCO, Vitria
- Problems and solutions
  - compared to Software distribution

E. Traffic information systems

- How is publish/subscribe applied to traffic information systems?
  - 1-to-N (broadcast, multicast), internet, channels
  - pull, push
  - periodic and aperiodic
  - mobile computing
  - real-time
- Examples
- Problems and solutions

F. Battlefield data dissemination

- How is publish/subscribe applied to battlefield data dissemination?
  - 1-to-N (mulitcast), unicast
  - continuous operation - “24 by 7”
  - real-time
  - mobile computing
- Problems and solutions

V. Products

Give examples of systems developed according to the above application to disseminate information and data:

A. Pointcast

- Describe what Pointcast is.
  - stock and news data distribution
  - Users send profiles to server then gather info.
- Include figure and describe.
- Describe its concepts and the technology it based on.
  - Based on pull, not really push
  - internet, intranets

B. SIFT (Stanford Information Filtering Tool)

- Describe what SIFT is.

  - stock and news data distribution
  - Disseminate info to users according to their profiles.
  - USENET News.
- Include figure and describe.
- Describe the technologies it based on?
  - Efficient matching of users' profiles and the info interested.
  - Index technology (on profiles) - can process large amount of info vs. large number of profiles.
  - The users get what they want only.

C. TIBCO (http://www.tibco.com)

- Describe what TIBCO is.
  - software distribution, stock and news data distribution
  - providing real-time e-business infrastructure software.
  - based on an architecture -The Information Bus- developed by Teknekron Software Systems, Inc.
  - guarantee delivery
- Include figure and describe.
- Describe its concepts and the technology is based on.
  - push
  - employs mulitcast protocol with each message containing a subject identifier
  - client side filter out messages based on the subject identifier

D. Vitria (http://www.vitria.com)

- Describe what Vitria is.
  - software distribution, stock and news data distribution
  - E-business platform for automating cross-enterprise business process.
  - Enable users to conduct business electronically across corporate networks and over the internet.
- Describe its concepts and the technology is based on.
  - Model
  - Automate
  - Monitor and Analyze

E. Channels

- What is Channels (web-casting)? (describe previously)
- Examples
  - Marimba's Castanet - send "Channels" from sever to clients, clients subscribe to the channels then periodic downloads the updates

- Netscape's netcaster - integrate push technology, web channels, and castanet channels mechanisms
- Microsoft's CDF (Channel Definition Format) - CDF files on sever specify client request, client retrieve files periodically

VI. Related Work

- Triggers & Active Databases (see comment in Franklin - Efficient Filtering...)
  - more general mechanism which can involve predicates over many data items and can initiate updates to other data items.
  - not optimized for fast matching of individual items to a large number of simple queries
  - see recent work: "Scalable Trigger Processing"

VII. Conclusions