StreamGlobe

Adaptive Query Processing and Optimization in Streaming P2P Environments

A. Kemper, R. Kuntschke, and ***B. Stegmaier***

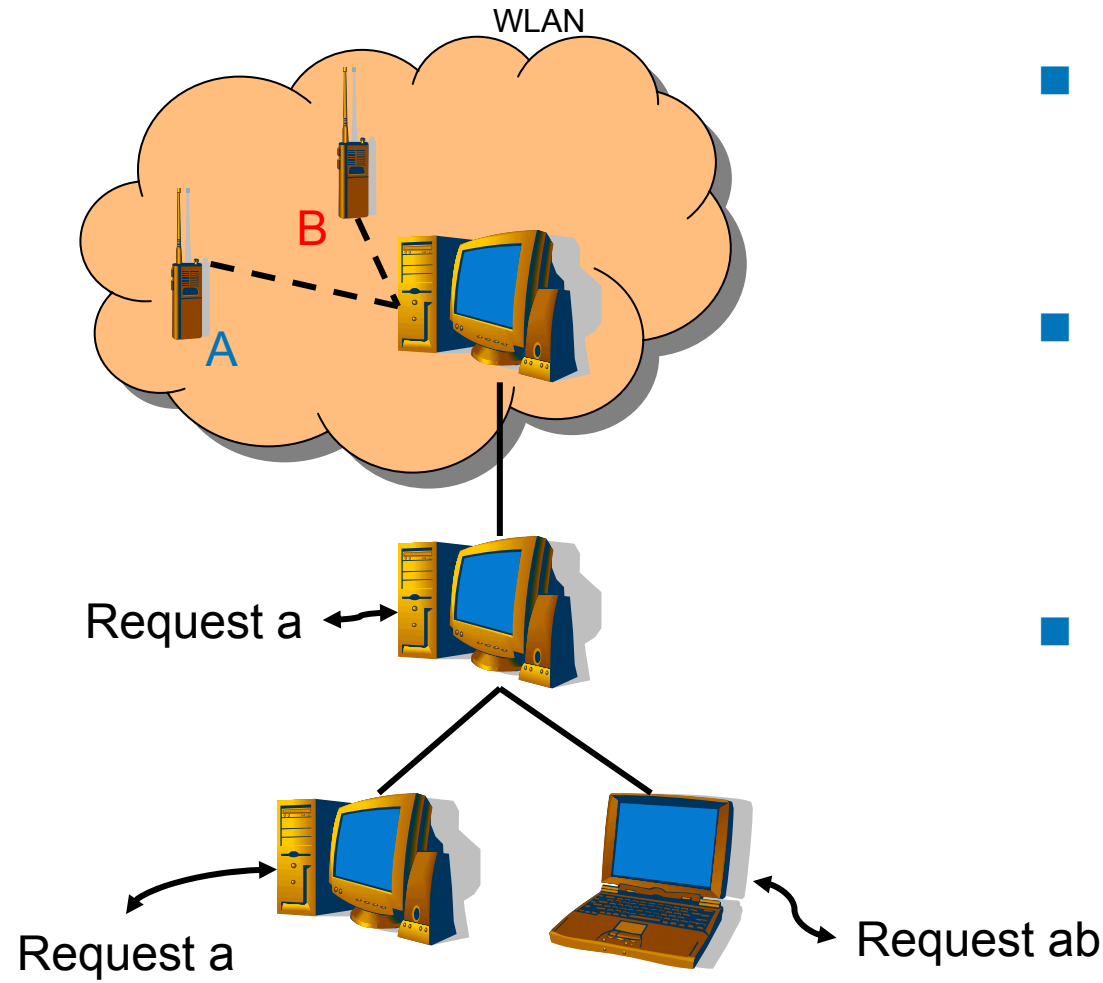
TU München – Fakultät für Informatik
Lehrstuhl III: Datenbanksysteme

http://www-db.in.tum.de/research/projects/StreamGlobe

Outline

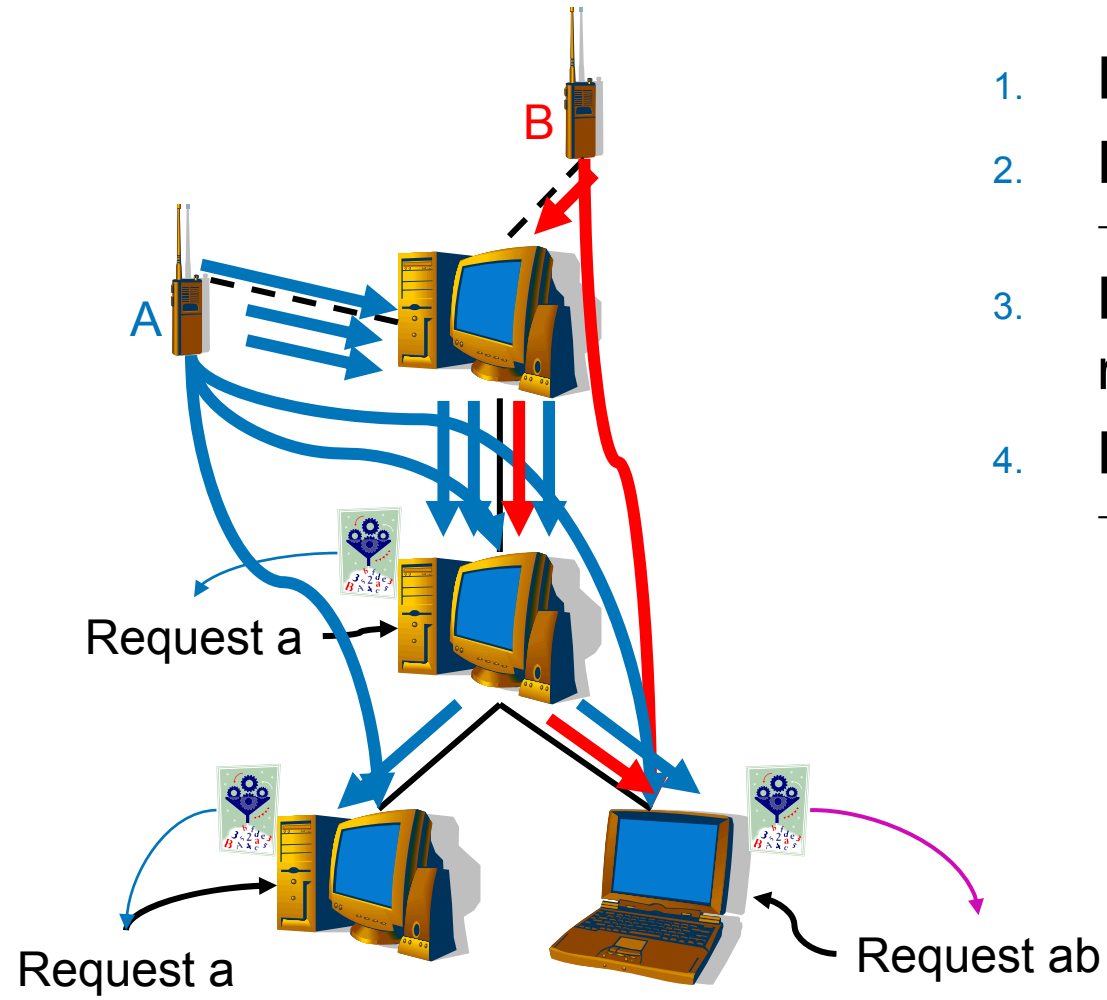
- Motivation
- StreamGlobe
  - The StreamGlobe Approach
  - Architecture Overview
- Current and Future Research
- Conclusion

Exemplary Initial Situation



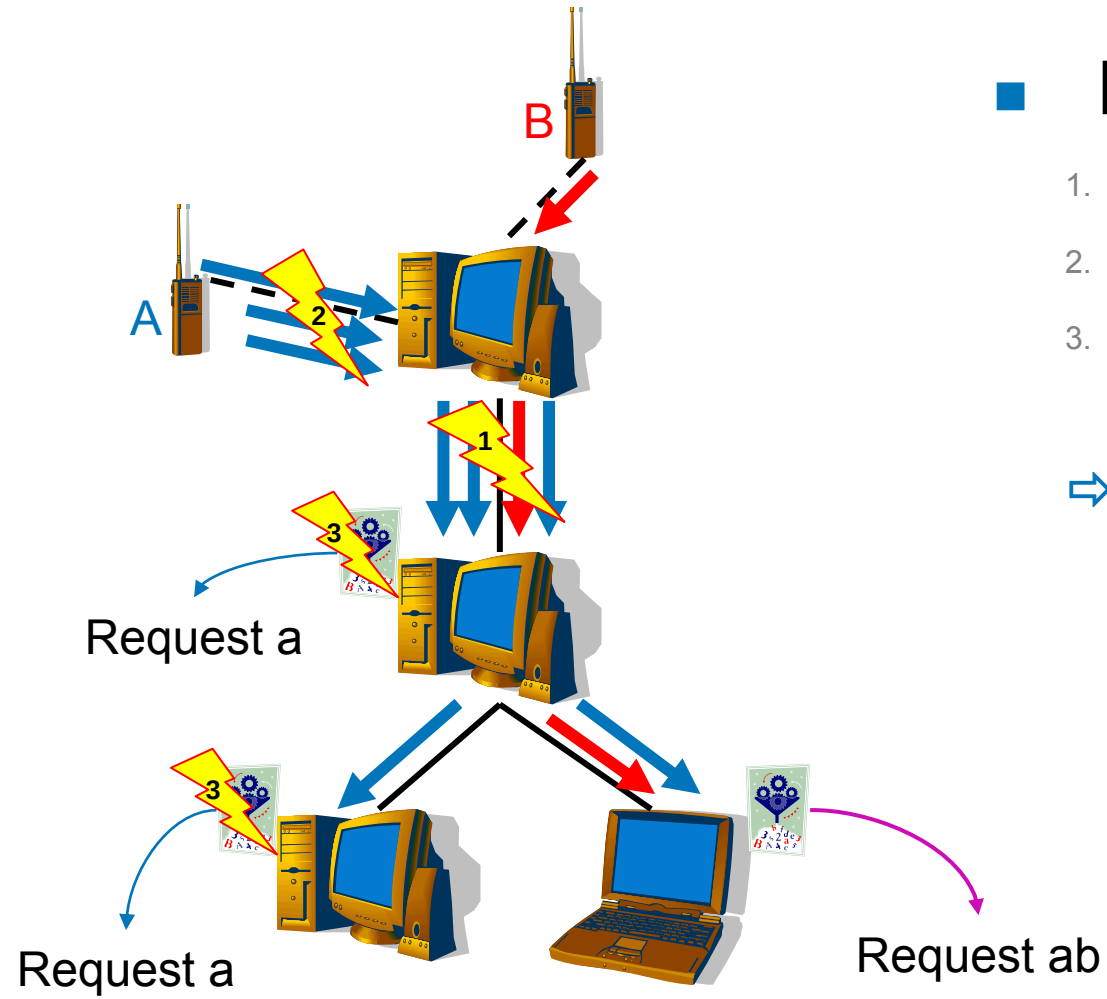
- Network
  - Consists of peers
  - Given or grown topology
- Data Sources
  - Provide XML data stream
  - Possibly infinite streams (e.g., sensor measurements)
- User requests
  - Continuous queries
  - Query language XQuery
  - Registered at a peer

General Traditional Approach



1. Register requests
2. Establish data transfer
   → Peers may connect arbitrarily
3. Process / Execute requests
4. Routing of streams
   → Map streams to network

General Traditional Approach (ctd.)



- Drawbacks
  1. Transmission of useless data
  2. Redundant transmissions
  3. Multiple request evaluation

⇨ Network congestion and processing overhead

Why StreamGlobe?

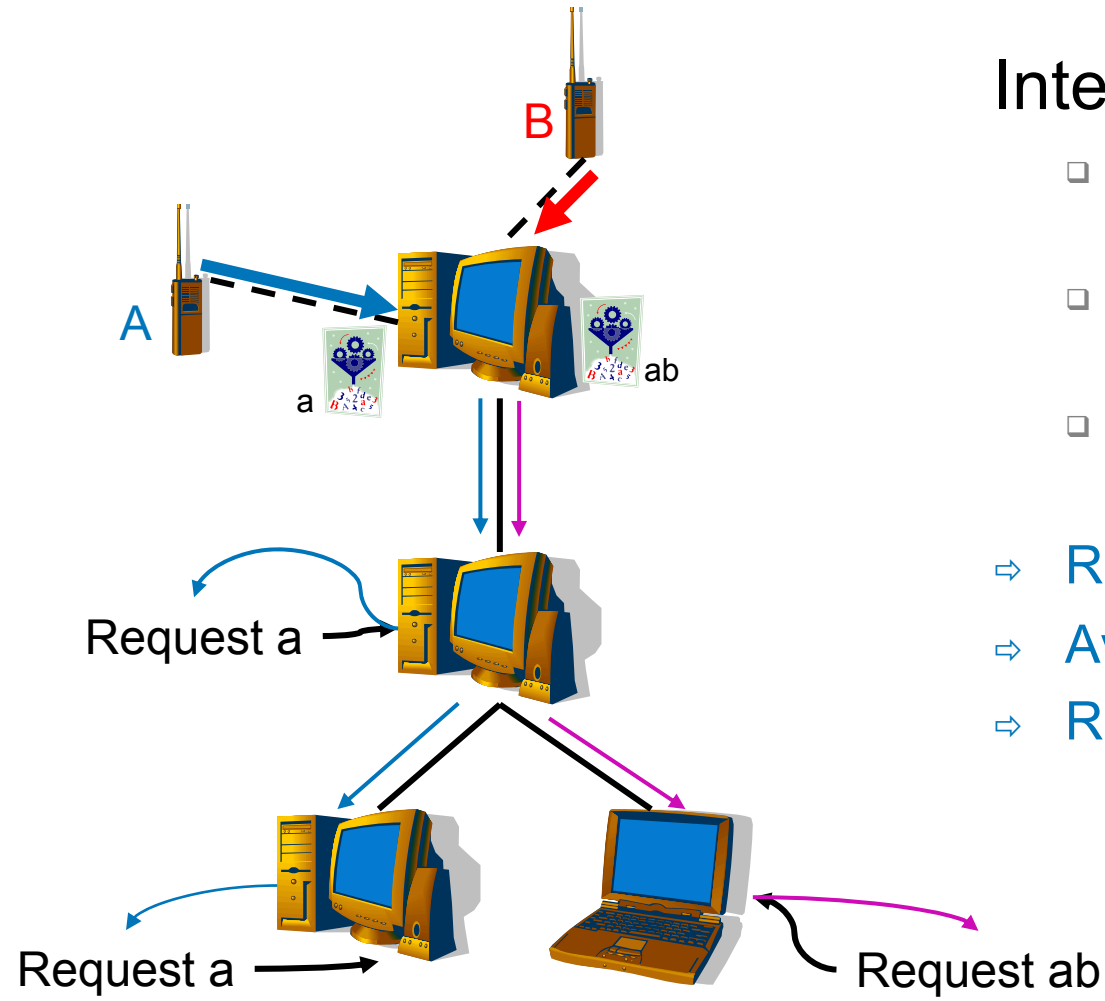
- Other Systems / previous work
  E.g. Cougar, TelegraphCQ, Multicast techniques:
  - Focus on specific aspects (e.g., query optimization)
  - Tailored to specific domains

- StreamGlobe
  - Contribution is combination of techniques:
    In-network query processing combined with routing
  - Constitutes a generic infrastructure
    - Independent of domain
    - Efficient data stream transformation and distribution

Outline

- Motivation
- **StreamGlobe**
  - **The StreamGlobe Approach**
  - **Architecture Overview**
- Current and Future Research
- Conclusion

The StreamGlobe Approach



Intelligent Routing

- Multicast routing techniques
  Data Stream Clustering
- Push query execution into network
- Multi-query optimization

⇨ Reduce network traffic

⇨ Avoid redundant transmissions

⇨ Reduce processing cost

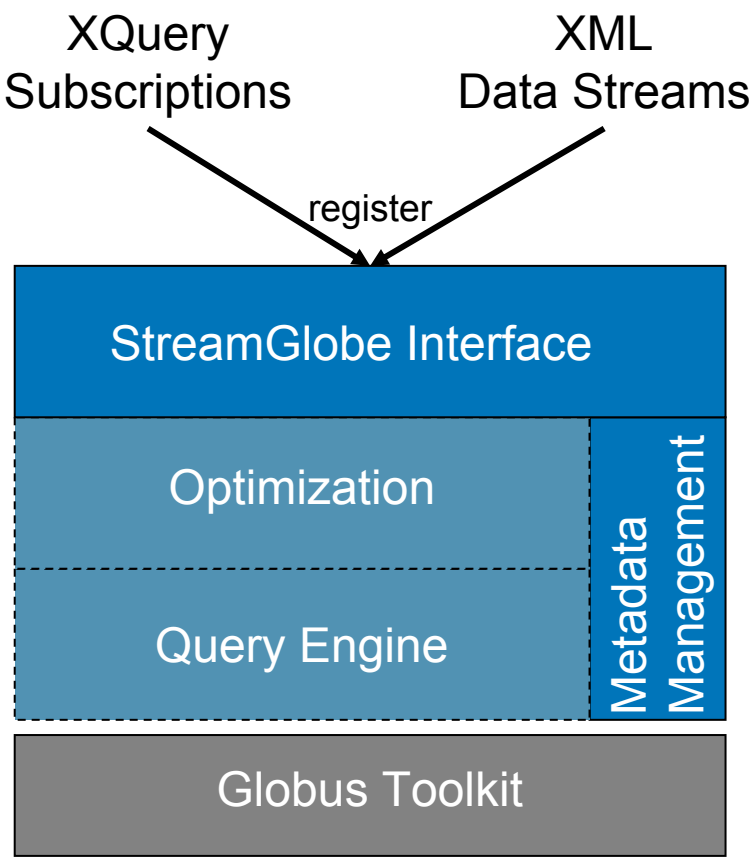
Basic Concepts

- P2P Network Topology
  - No arbitrary communication
    → Communication via *transfer paths*
  - No fixed P2P topology

- Classification of peers
  - Thin-Peers
  - Super-Peers
  - ⇨ Constitution of a super-peer backbone

- Hierarchical organization
  → *Speaker-peer* responsible for certain subnet

StreamGlobe Peer Architecture

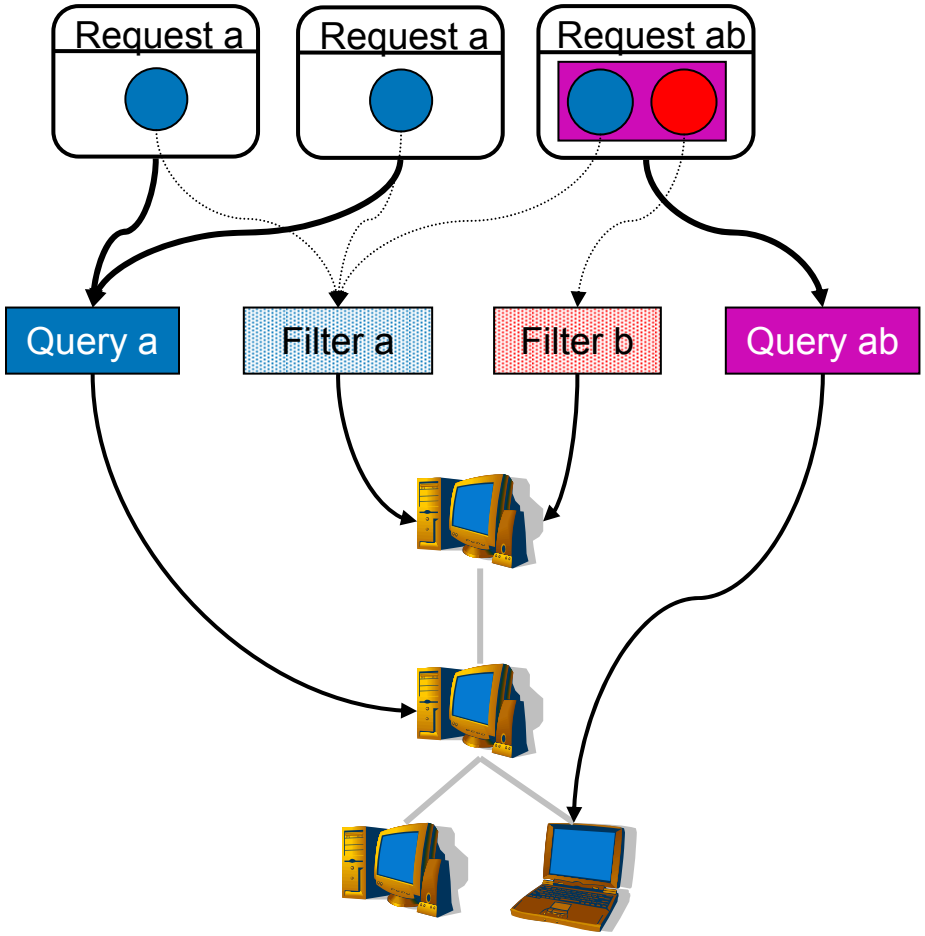


- Based upon *Open Grid Services Architecture (OGSA)*
- Integration similar to OGSA-DAI or OGSA-DQP
- Layers as grid-services
- Availability according to peer capabilities
- Message exchange via RPC and notifications
- Data stream transfer via direct TCP connections

Optimization

- Goals
  1. Registration of arbitrary subscriptions at any peer
  2. Achieve good distribution of data streams
  3. Optimize evaluation of many subscriptions
- Achievement
  - Pushing query execution into the network
    → (1) and (3)
  - Multiquery optimization
    → (3)
  - Early filtering of data streams resp. evaluation of subscriptions
    → (2)
  - Data stream clustering
    → (2)

Multi-Query Optimization



- Performed by speaker-peer
- Analyze subscriptions and streams
  - Common subqueries
  - Re-usability of streams
  - Based on properties of subscriptions / streams
- Computes
  - Filters and queries
  - Data stream clustering
  - Execution locations

Query Execution

- Basic concepts
  - Streaming evaluation and push-based techniques
  - Preclude unbounded buffering by requiring window constraints
  - Extensibility by means of mobile code

- Evaluation of subscriptions with *FluX*
  - Designed for streaming processing of XQuery
  - Event-based extension to XQuery
  - Usage of schema information for buffer minimization

→ *Visit my talk at the VLDB: Tomorrow, Research Session 6: XML(II)*

Outline

- Motivation
- StreamGlobe
  - The StreamGlobe Approach
  - Architecture Overview
- **Current and Future Research**
- **Conclusion**

Current and Future Research

- Current Research
  - Optimization techniques
  - Extension of FluX

- Future Research
  - Quality-of-Service management
  - Explicit load balancing
  - Load shedding techniques
  - Construction of overlay network

  …

Conclusion

StreamGlobe

- Exploiting in-network query processing capabilities
- In combination with data stream clustering

⇨ Minimization of network traffic

- Query execution with FluX

⇨ Efficient and scalable execution of subscriptions

- Multi-query optimization

⇨ Parallelization and load balancing in the network

Related Work

- Aberer, Cudré-Mauroux, Datta, Despotovic, Hauswirth, Punceva, Schmidt. *"P-Grid: a self-organizing structured P2P system"*. SIGMOD Record 32(3), 2003
- Arasu, Babcock, Babu, Datar, Ito, Motwani, Nishizawa, Srivastava, Thomas, Varma, Widom. *"STREAM: The Stanford Stream Data Manager"*. Data Engineering Bulletin 26(1), 2003
- Carney, Cetintemel, Cherniack, Convey, Lee, Seidman, Stonebraker, Tatbul, Zdonik. *"Monitoring Streams – A New Class of Data Management Applications"*. VLDB 2002
- Chandrasekaran, Cooper, Deshpande, Franklin, Hellerstein, Hong, Krishnamurthy, Madden, Raman, Reiss, Shah. *"TelegraphCQ: Continuous Dataflow Processing for an Uncertain World"*. CIDR 2003
- Cherniack, Balakrishnan, Balazinska, Carney, Cetintemel, Xing, Zdonik. *"Scalable Distributed Stream Processing"*. CIDR 2003
- Krämer, Seeger. *"PIPES – A Public Infrastructure for Processing and Exploring Streams"*. SIGMOD 2004
- Madden, Shah, Hellerstein, Raman. *"Continuously Adaptive Continuous Queries over Streams"*. SIGMOD 2002
- Sellis. *"Multiple-Query Optimization"*. TODS 1988
- Yang, Garcia-Molina. *"Designing a Super-Peer Network"*. ICDE 2003
- Yao, Gehrke. *"The Cougar Approach to In-Network Query Processing in Sensor Networks"*. SIGMOD Record 31(3), 2002