Management of Massive Data in a Distributed Environment: issues and advances

Kokou Yétongnon

Laboratoire LE2I

Outline

- Data grid
- Stream data
- Location based data
- Web based massive data
- P2P and data management
  - What is P2P
  - Some work in our research group
    - HON P2P for multimedia data
    - Multimedia data streaming
    - Semantic matching and interoperability
- Conclusion

Data Grid

- Grid system: collection of heterogeneous computers and resources spanning multiple administrative domains to provide access to resources

- Computing grid widely used for collecting and managing scientific experiments Scientific experiments
  - Large collections of data

  (Cern particle physics, human genome data, etc…)
  - Heterogeneous data

Scientific data grid Issues

- Computing and processing power
- Storage space
- Storage format
- Access and security
- Querying

Scientific data grid Issues

- New distributed data structure
  - CERN's EDMS :
    - PDM (Product Data Management)
    - Engineering Data Management System
  - MDAS : Massive Data Analysis System

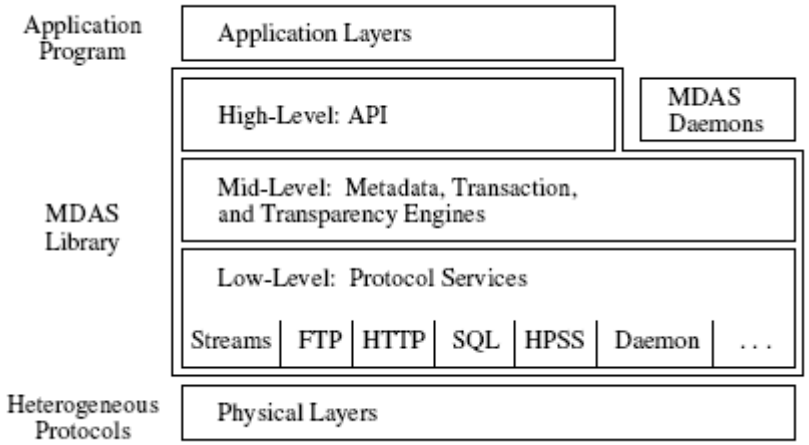
    San Diego Supercomputer Center 95-97, DARPA financed
- Manage resources in a heterogeneous distributed system
  - Metadata and data description
  - Detect available resources, storage spaces

Scientific data grid Issues

- Data format Heterogeneity

MDAS approach :

Middleware,
Layers architecture,
Transparency on multiple levels

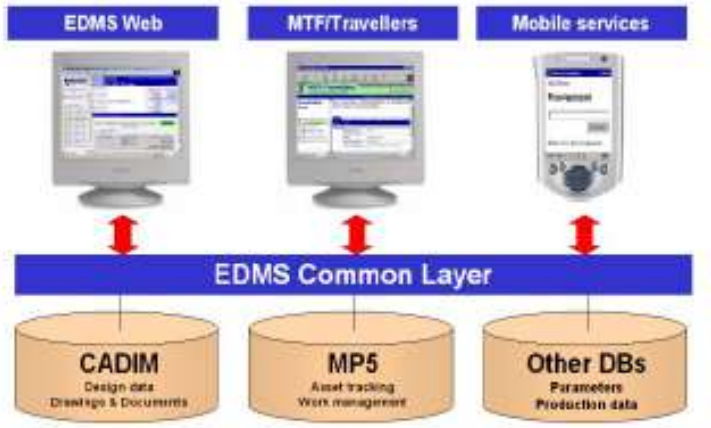


Scientific data grid Issues

- Data format Heterogeneity

  CERN's EDMS approach:

  PDM *System Architecture*



Design data & drawing documents

Asset tracking and work management → workflow ? …

DBMS→ Parameters & production data

Scientific data grid Issues

- Data querying and presentation
  - queries addressed to a common engine
  - intranet and internet Web based interface
- Security issues
  - authentication
  - authorization
- Data Replication and distribution

Data Stream

- Data changes very frequently
- Dynamic Data Streams instead of static data sets
- Storage issues :
  - space
  - data redundancy
  - update-delete problem,
- Query processing : continuous querying
  - overload, latency, QoS,
  - up-to-date answers
  - load balancing (data provided by multiple data sources),

Data Stream

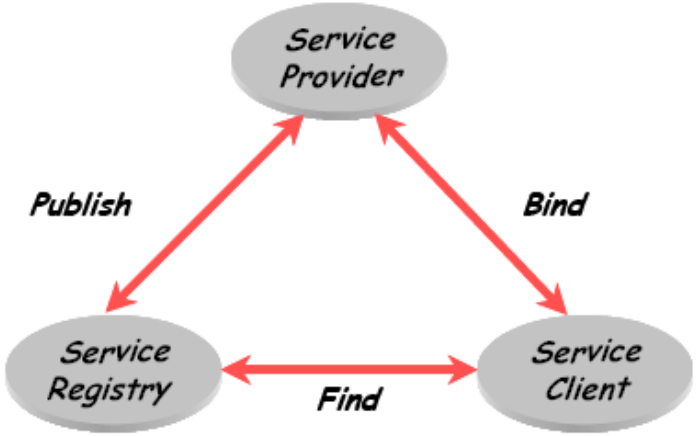
- SQL not expressive enough ?
  - PDT-SQL: Extension to handle sensor readings
  - Example : Pollution detection query
- Continuous querying
  - Ex: STREAM at Stanford [Continuous Queries over Data Streams]
    - Triggers, materialized views
    - Updates and deletion
    - Temp storage for query result
- Query Load Shedding based on QoS
  - MIT AURORA
    - tuples are dropped based on their content
    - tuples are dropped in a randomized fashion (randomly)

Massive data in the web

- Collections of human readable data and services
  - automate applications (B2B etc …)
  - semantic web and semantic web services
- Pervasive systems
- Ubiquitous computing and associated huge collection of data

Massive data in the web

- Web services
  - A “Library” providing data and services for other applications
  - Building blocks for creating distributed applications



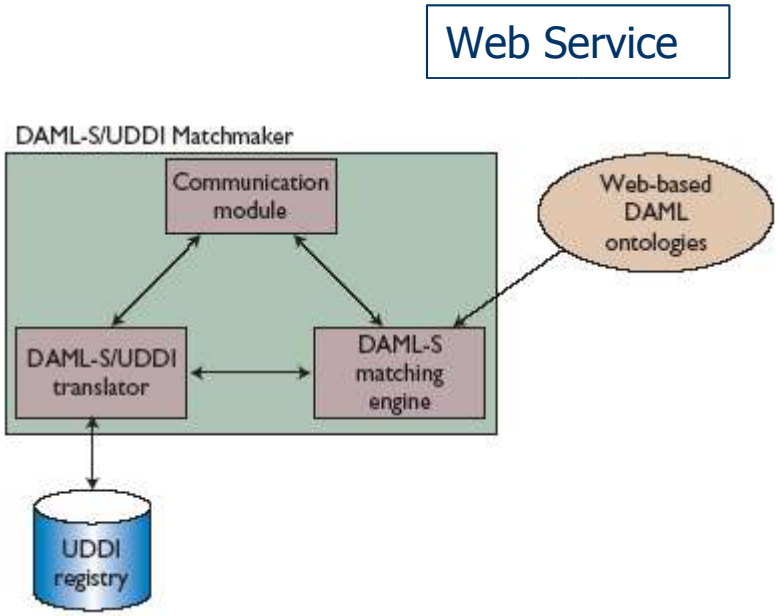
Communication : **SOAP** over **HTTP**
Description : **WSDL**
Publication and research: **UDDI**

Massive data in the web

- Web service issues
  - Automatic selection of adequate web service
  - Composition of web services
  - Web service similarity and semantic compatibility
  - Semantic aspects
- Evolution of semantic web services description

Massive data in the web

- Comparison and matching



Communication module: receives the WS request or avertissement (inquiet or public)

Translator: stores avertissements in UDDI and sends match request to the matching engine

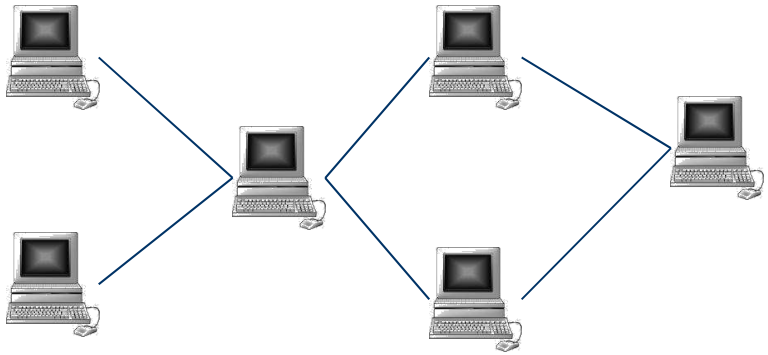
Matching engine : checks the ontologies for potencial match

Management of massive data

P2P and Data Management

Introduction

- A peer is a computer that behaves as a client and a server



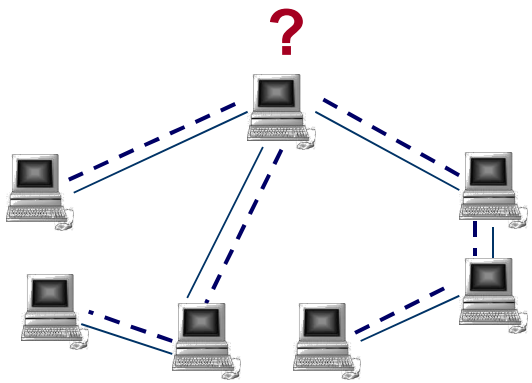
P2P-Network

- P2P Network allows peers to interact directly.

Introduction

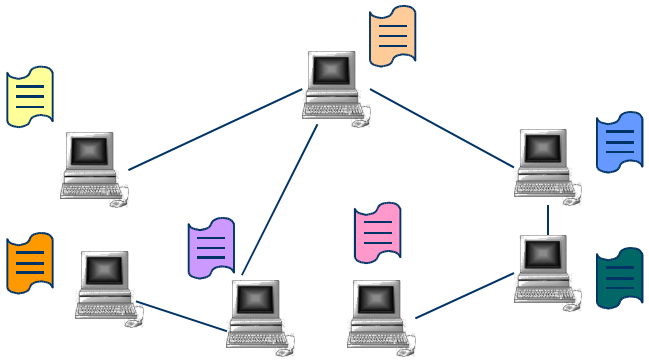
P2P Systems

Unstructured systems



- Each peer is unaware of the resources of other peers
- Flooding to search

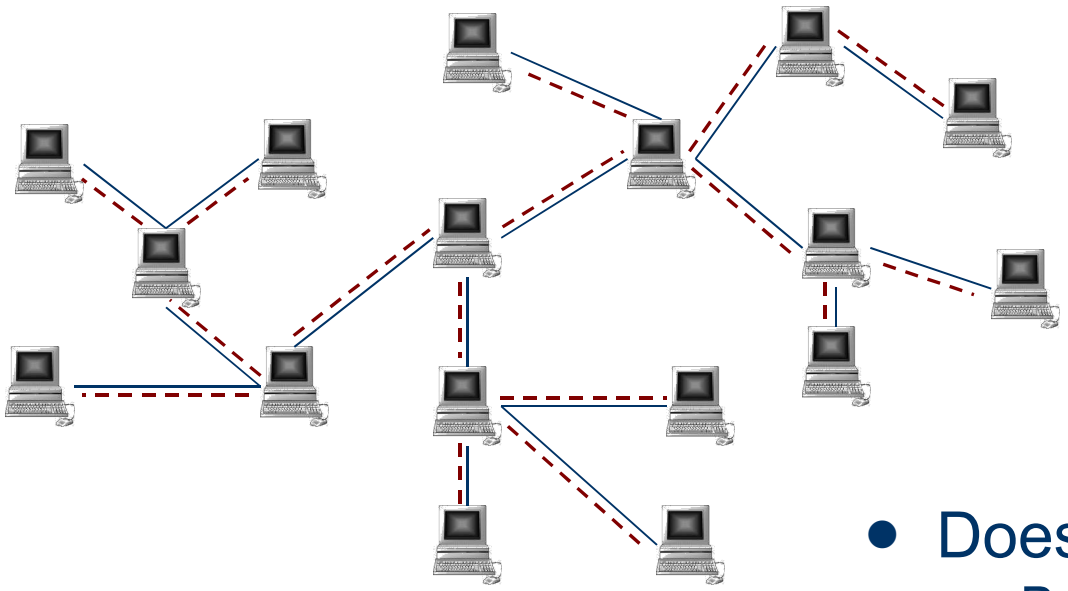
Structured systems



- Peers maintain information about resources the other peers offer.
  - DHT techniques
  - Clustering

Flooding

- Popular query processing technique in P2P

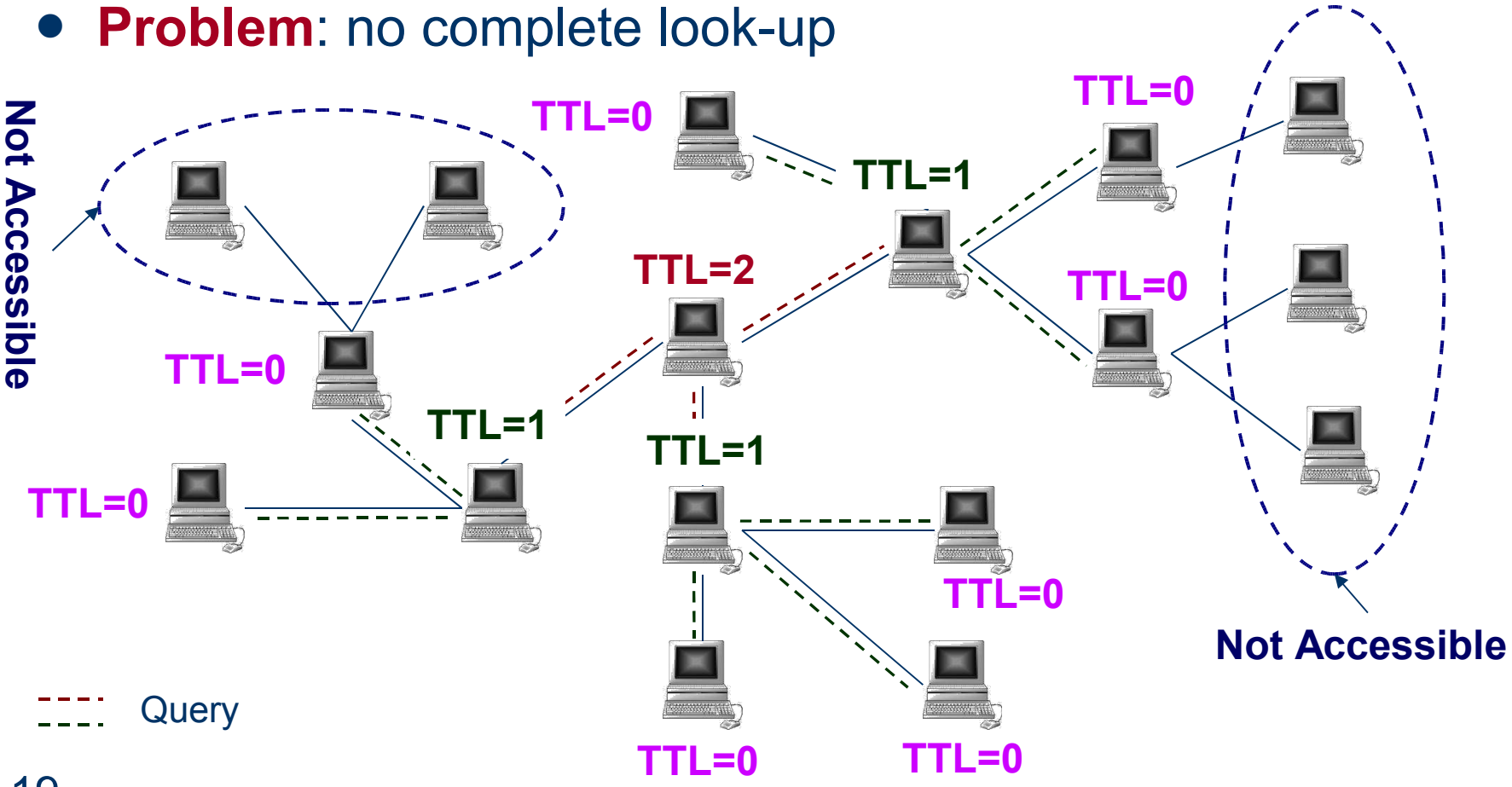


Query

- Does not scale
  - Bandwidth consumption
  - Peers load

Flooding

- **Solution for flooding** : TTL (Time-To-Live)
- **Problem**: no complete look-up

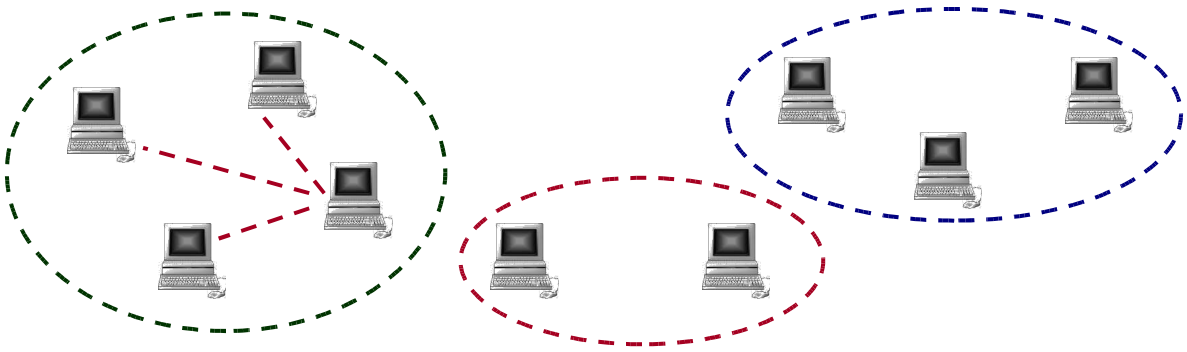


DHT (Distributed Hashing Table)

- **DHT Organizes data** is a key space
  - avoid flooding the network
  - complete look-up
- BUT perform poorly on
  - approximate query
  - range queries
  - complex queries

Clustering

- Clustering **Organizes peers** with common interests forming clusters



- Routing queries **only** to peers that are **more likely** to have answers may be more efficient
- Limitation of flooding to only a group of peers

Clustering

- Peers can be organized according to
  - Content (semantic or low level)
  - Type of resources (image, video…)
  - Network parameters
  - Application needs
  - …

P2P research in our group

- P2P hybrid network for multimedia data management
- Multimedia streaming in P2P.
- Semantic based matching and interoperability in P2P

Conclusion

- New emerging network based information systems
- How to describe information for semi-automatic and automatic processing?
- Can database tools be used to manage the data?
- Managing the dynamic aspect of these systems
  - continuous query in stream data
  - join and leave in P2P systems
  - dynamic aspect in semantic web services

Thank you