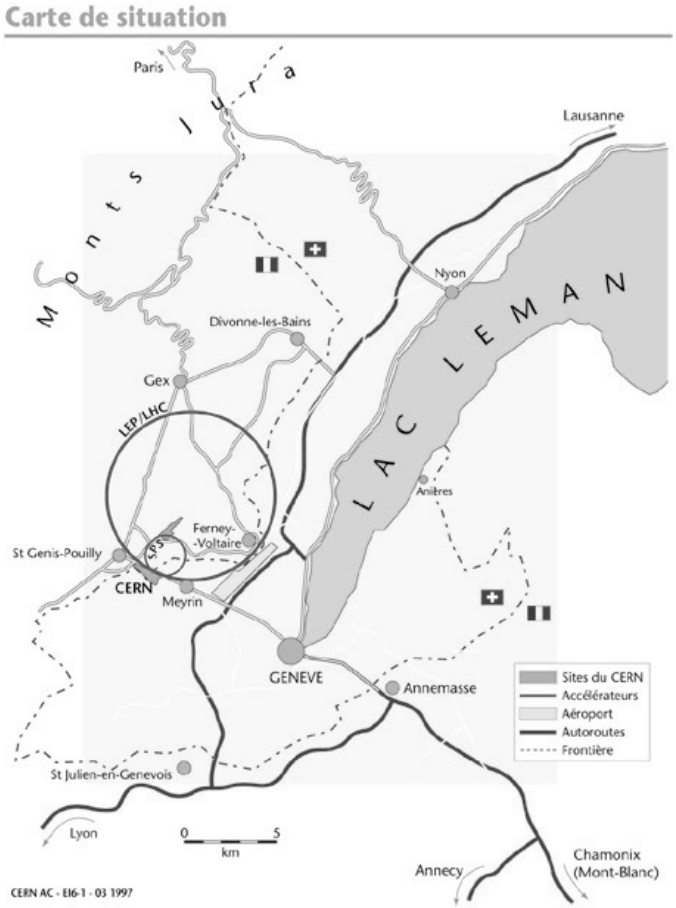
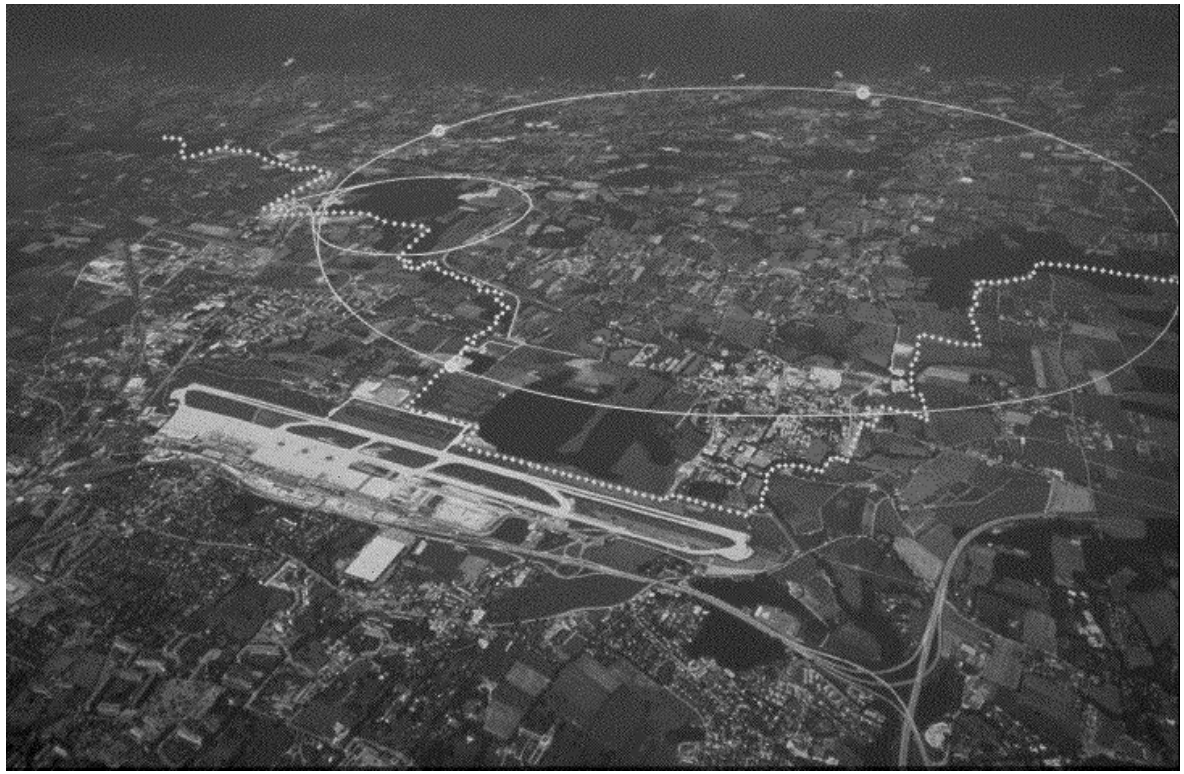
Data Management in the European DataGrid Project

Heinz.Stockinger@cern.ch

on behalf of the
EU DataGrid Data Management Working Group

CERN


Carte de situation
Paris
Jura
Monts
Lausanne
Nyon
LAC LEMAN
Divonne-les-Bains
Gex
LEP/LHC
Ferney-
-Voltaire
Anières
St Genis-Pouilly
SPS
CERN
Meyrin
GENEVE
Annemasse
Sites du CERN
Accélérateurs
Aéroport
Autoroutes
Frontière
St Julien-en-Genevois
Lyon
0
5
km
Annecy
Chamonix
(Mont-Blanc)
CERN AC - EI6-1 - 03 1997



The LEP accelerator

➔World's largest particle collider, ran for 11 years.

➔27 km circumference, 100 m underground

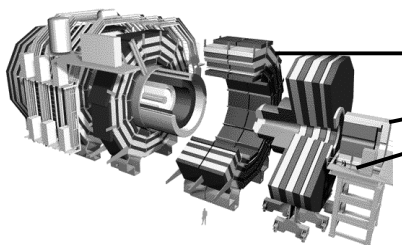
➔ Counter circulating beams of electron positron bunches

➔ Four experiments have confirmed Standard Model predictions to high precision

➔Maximum collision energy of 209 GeV

Many questions still remain ⇒ LHC

Multi-Tier Model



~100 MBytes/sec

Online System

Offline Farm

~PBytes/sec

~100 MBytes/sec

Tier 0

CERN Computer Center

~2.4 Gbits/sec

Tier 1

France Regional Center

UK Regional Center

Italy Regional Center

Fermilab

• • •

.6 - 2.4 Gbits/sec

Tier 2

Tier2 Center

Tier2 Center

Center

Center

Center

Tier 3

Institute

Institute

Institute

Institute

Physics data cache

100 - 1000 Mbits/sec

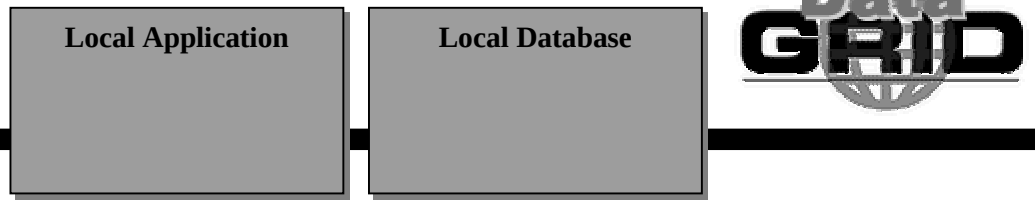
Workstations

Use Cases

- Testing of algorithms for data production
- Data Production & Simulation
- Physics Analysis

Job Type	*User*	*Scheduled*	*Input Data Set*			*Output*
			Known	*Size*	*Spans Datatypes*	
Testing	Physicist	No	Yes	Small Subset	No	Data Subset
Production	Manager	Yes	Yes	All Events	No	Data Set
Analysis	Physicist	No	Partially	Small to All	Yes	Histogram/Tag

Architecture



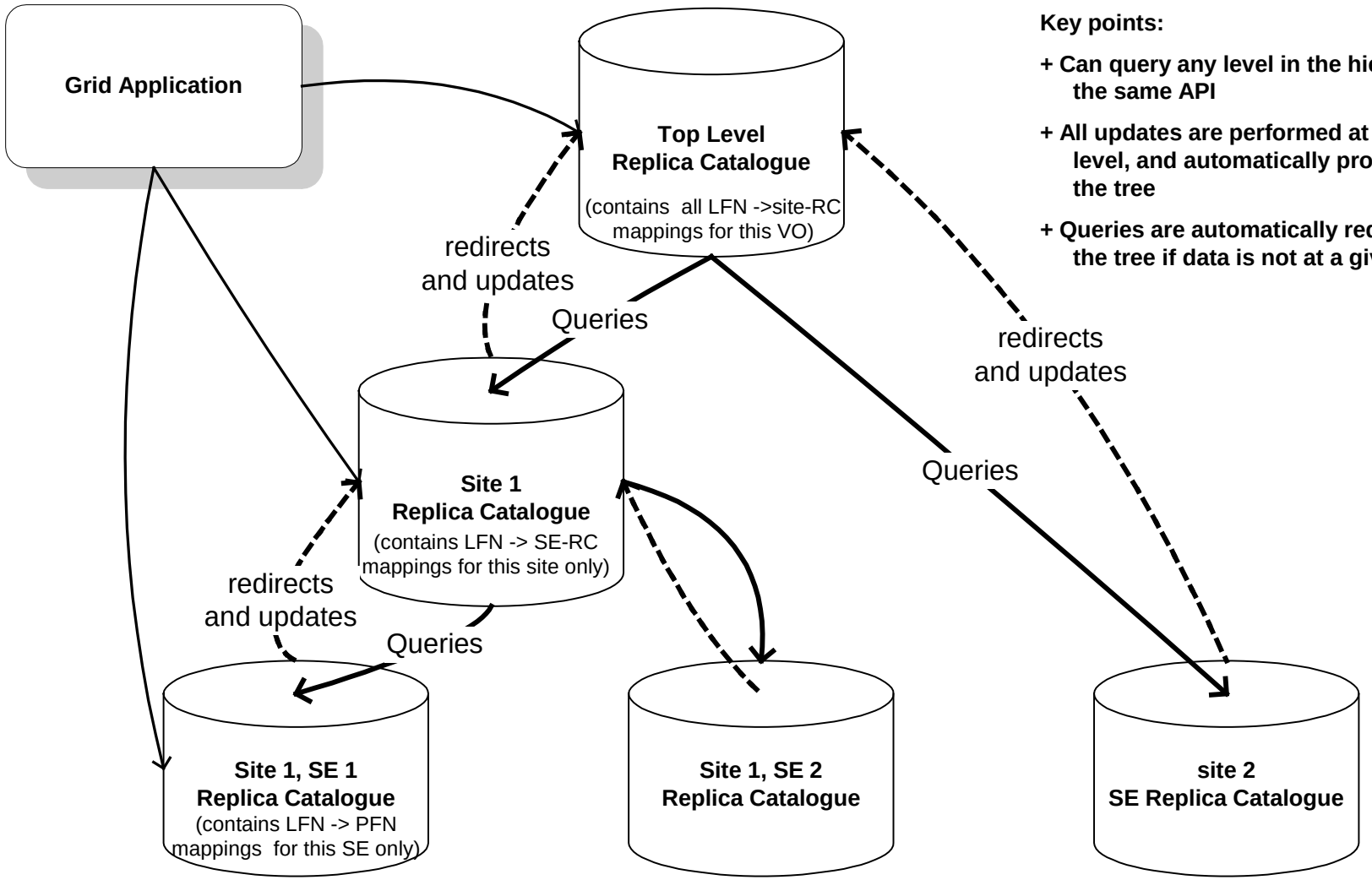
GDMP Replication Functionality

- Subscribing to a remote site
  - to get notified when new files are created and made public
- Publishing new files
  - to make files available and accessible to the Grid
- Obtaining a remote site's file catalog
  - for failure recovery
- Transferring files from remote site to local site

Partitioned Replica Catalog (RC)

- Scalability requirements (e.g. CMS 2006)
  - RC with 50 million files spread across dozens of institutions
- Catalogs are independent and autonomous
  - Each site has RC containing a list of local files only
  - But linked together through a hierarchy or graph
- *Pros*
  - Scalable: because load is spread among multiple RCs
  - Minimal latency: Access is localized
  - No single point of failure. Local and independent admin.
- *Cons*
  - Must traverse hierarchy to locate data
  - Potentially complex update propagation
  - Potentially stale data due too lazy synchronization

Sample RC Hierarchy



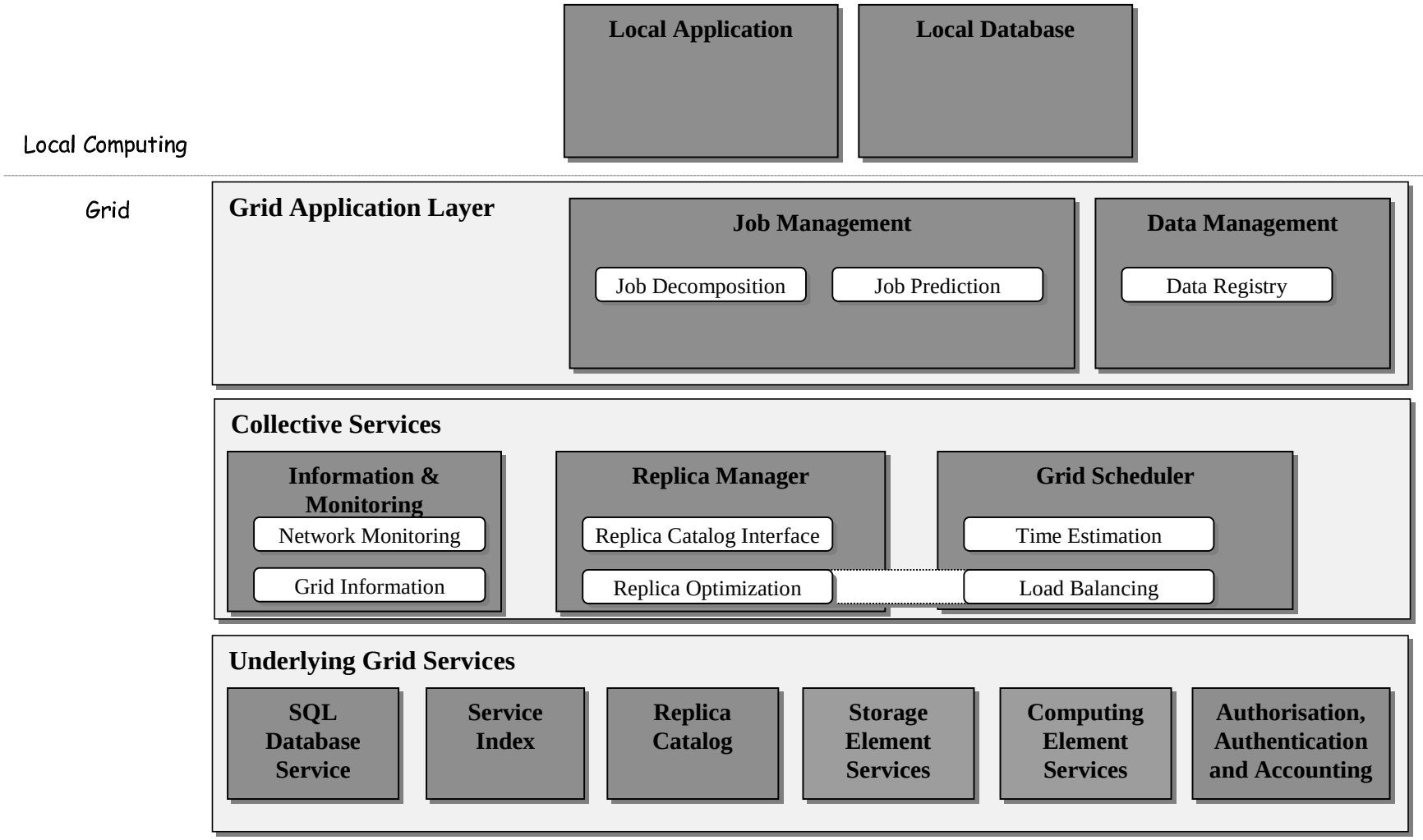
Key points:

+ Can query any level in the hierarchy with the same API

+ All updates are performed at the SE-RC level, and automatically propagated up the tree

+ Queries are automatically redirected up the tree if data is not at a given level

Optimisation Architecture



Example Optimisations

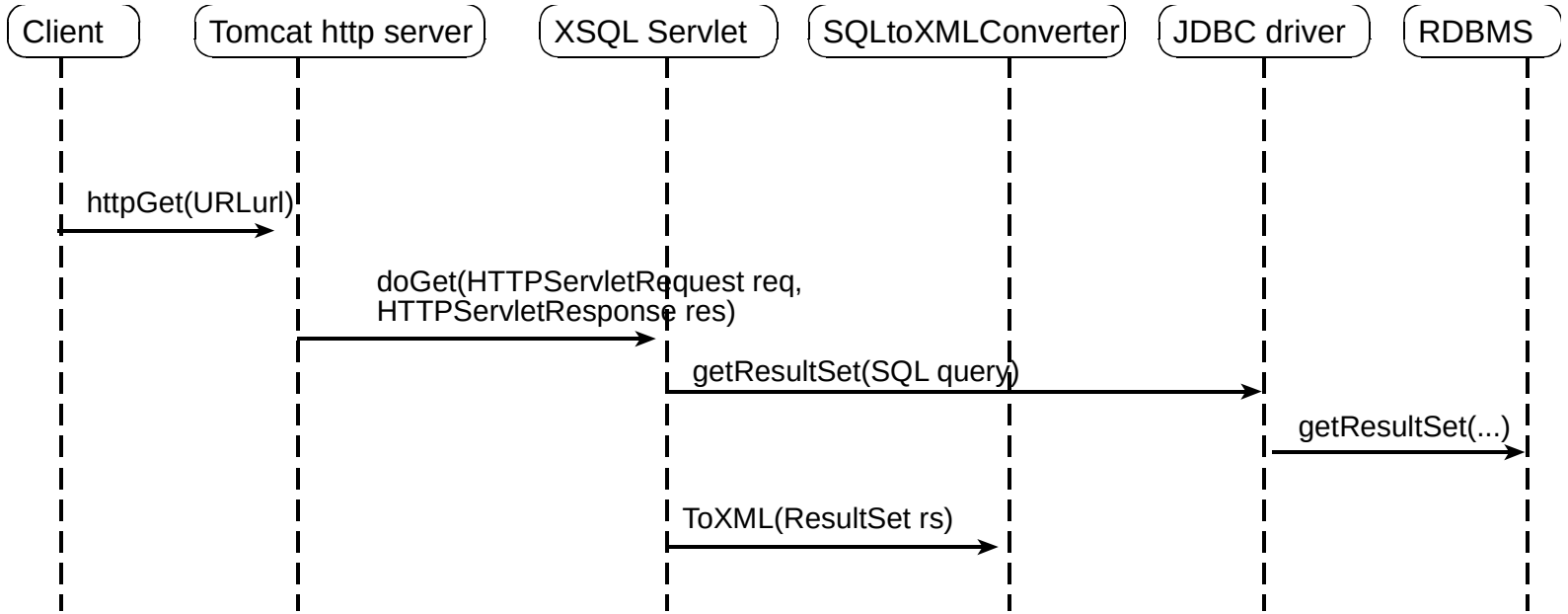
- Data Access Time Estimation
  - Help Scheduler estimate cost of data movement
  - Uses Network Monitoring Service
    - to discover the current network bandwidth situation
  - Uses Storage Device Monitoring Service
    - to discover the current load on the devices
  - Uses Replica Catalog
    - to discover the amount of replication of the required files
- Offline Optimisation
  - Monitor usage of logical files on the Grid as a whole
  - Try to match the supply of replicas to the demand for them

SQL Database Service Introduction

- Requirements
  - Meta data storage and retrieval
  - Query, insert, delete, update, create table, drop table
  - Easy to use, efficient and scalable
- Unified Grid enabled front end to RDBMS
  - Well defined language and platform neutral network protocol between client and server
  - Enables interoperability and flexibility
    - C/C++, Java and Python clients ...
    - to access data in local or remote ...
    - MySQL, Oracle, DB/2 or Postgres database ...
    - with clients @ Linux or Win32 and servers @ Solaris

Architecture & Interaction

- Classic loosely coupled 3-tier model on the web
  - ala Google, amazon.com, CERN web phonebook



Use Case: Retrieve meta data about LFNs

Given: Table RCatalog

```
LFN                                | Size
-----------------------------------------
host1.cern.ch/somepath/file1 | 10000000
host2.cern.ch/somepath/file2 | 50000000
```

1. Client query via HTTP GET to
   `http://sql.cern.ch/getLogicalFileMetaData?catalog="RCatalog"`

2. "getLogicalFileMetaData" query template on server:

```
<xsql:query xmlns:xsql="urn:oracle-xsql" connection = "myconnection"
    select * from {@catalog}
</xsql:query>
```

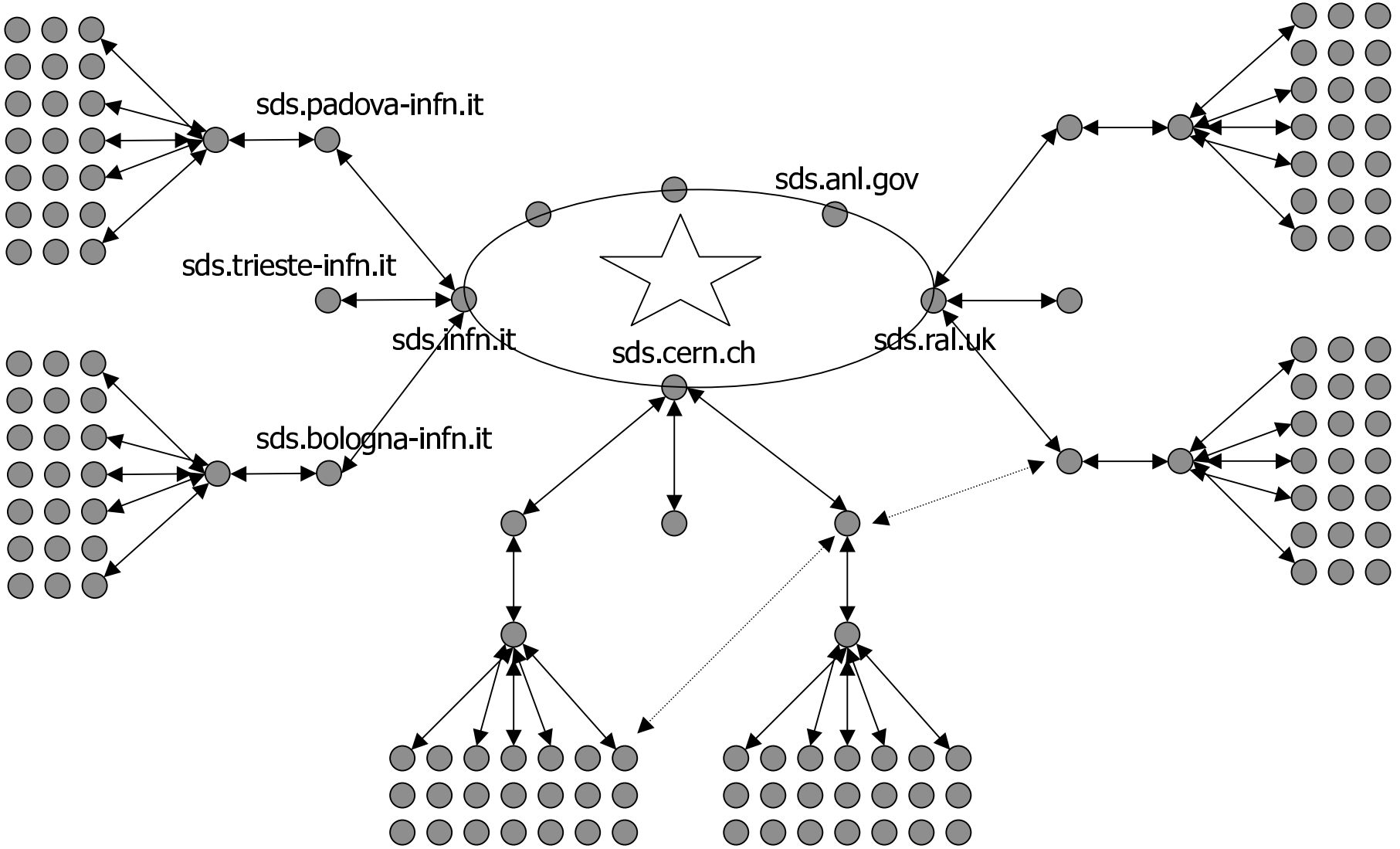
3. ResultSet returned as canonical XML:

```
<ROWSET>
          <ROW>
                    <lfn> host1.cern.ch/somepath/file1 </lfn>
                    <size> 10000000 </size>
          </ROW>
          <ROW>
                    <lfn> host2.cern.ch/somepath/file2 </lfn>
                    <size> 50000000 </size>
          </ROW>
</ROWSET>
```

Service Index Introduction

- Before talking to a service, client (app) needs to discover
  - where information sources and sinks are available
  - how to talk to them
- Scalability problem
  - Small scale environments
    - apps can be configured to “know” what is out there and how to talk to it
  - Large scale dynamic multi org. environment requires
    - automatic registration of services and their properties
    - facilities to search for matching services
    - facilities to invoke services

Web of Services



Sample Service Description

```
<service>
        <mandatory>
                <URL> x-gridrep://eff555.cern.ch:55555 </URL>
                <type>  replica-manger </type>
                <owner> atlas </owner>
                <email> atlassupport@cern.ch </email>
                <protocol version = “1.2”
                        <authentication mech = “CRAM-MD5”/>
                        <authentication mech = “OTP”/>
                </protocol>
        </mandatory>
        <optional>
                <catalog>
                        <type> atlas-triggerstudies </type>
                        <runs
                                id-from  = “10000”
                                id-to    = “20000”
                                date-from= “2000-12-08”
                                date-to  = “2001-02-20”
                                status   = “disk resident”
                                size     = “2 TB”
                        />
                </catalog>
        </optional>
</service>
```

Summary

- Complex Multi-Tier Model
  - From very large computing center to home institute
- Use Cases & Architecture
  - Classification leads to modules & optimisation potentials
  - Layers, Services, Grid Protocols
- Replication System
  - Maintain copies of files across WAN for better data locality
  - Access Time Estimation for improved job scheduling
- SQL Database Service & Service Index
  - Unified Grid enabled front end to RDBMS
  - Discover info sources and sinks and how to talk to them
- For more information
  - http://www.eu-datagrid.org/
  - http://cern.ch/grid-data-management/