OGSA-DAI Today

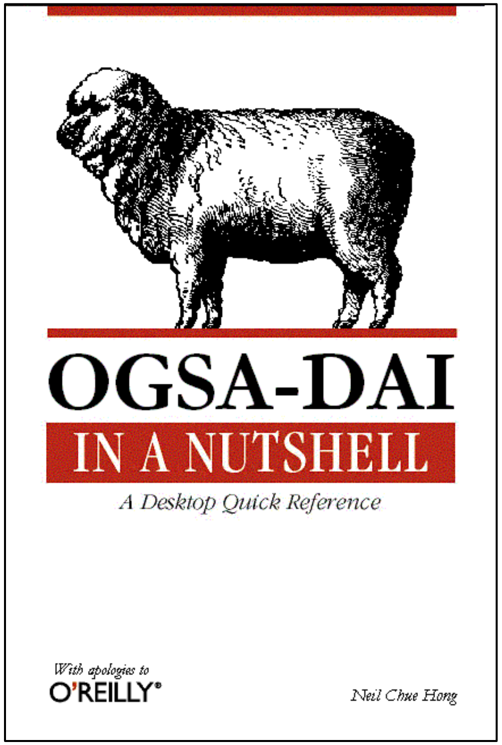
GridWorld 2006, Washington DC
11 September 2006

Outline

- What is it?
- “Let us out”
  - Exposing data to clients – the server’s perspective
- “Let us in”
  - Getting to the data – the client’s perspective
- “More, more more…”
  - Extending OGSA-DAI

OGSA-DAI in a nutshell

- An *extensible framework* for data access and integration
- Expose heterogeneous data resources to a grid through web services
- Interact with data resources
    - Queries and updates
    - Data transformation / compression
    - Data delivery
    - Application-specific functionality
- A base for higher-level services
    - Federation, mining, visualisation,…



OGSA-DAI

IN A NUTSHELL

A Desktop Quick Reference

With apologies to
O'REILLY®

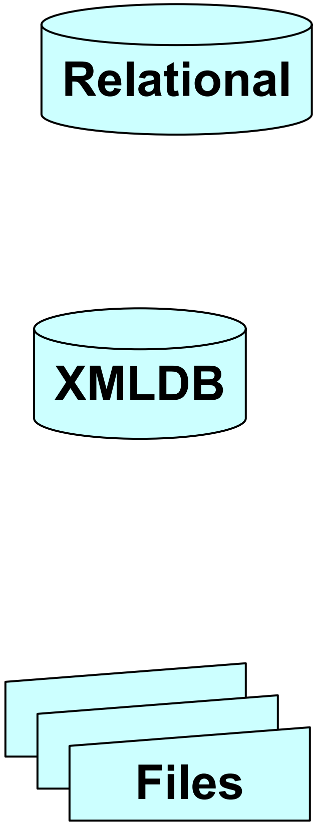
Neil Chue Hong

OGSA-DAI motivation

- Entering an age of data
  - Data Explosion
    - CERN LHC will generate 1GB/s = 10PB/y
    - VLBA (NRAO) generates 1GB/s today
    - Pixar generate 100 TB/movie
  - Storage getting cheaper
- Data stored in many different ways
  - Relational databases
  - XML databases
  - Text and binary files
- Need ways to facilitate
  - Data discovery
  - Data access
  - Data integration
- Empower e-Business and e-Science
  - The grid is a vehicle for achieving this

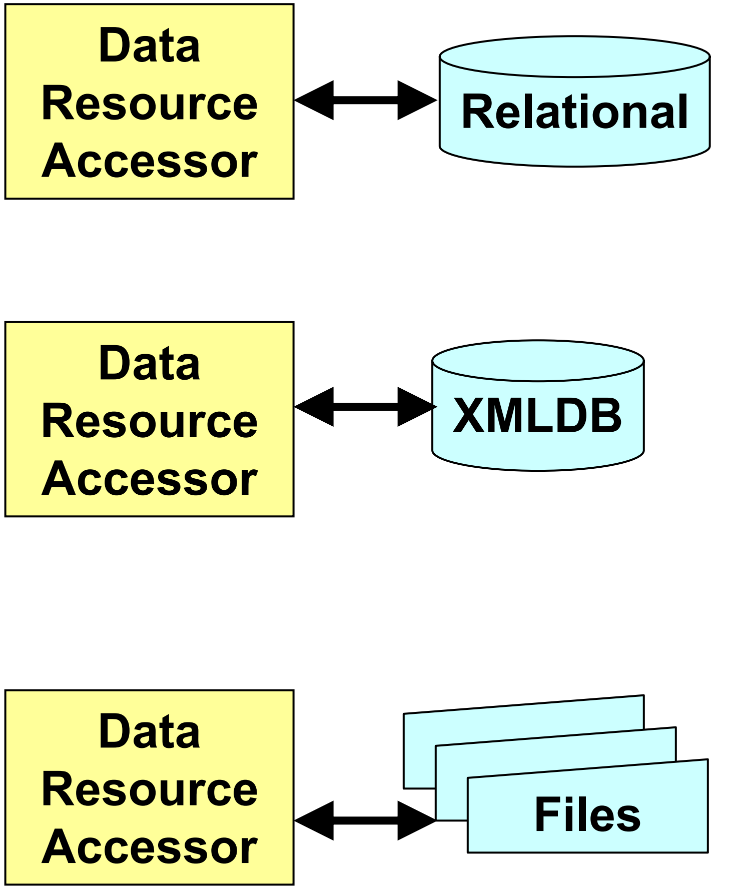
Data resources

- Relational
  - MySQL
  - Microsoft SQL Server
  - Oracle
  - IBM DB2
  - PostGres
  - HSQL
- XML
  - eXist
  - Xindice
- File system
  - SwissPROT
  - OMIM
  - Text
  - Binary

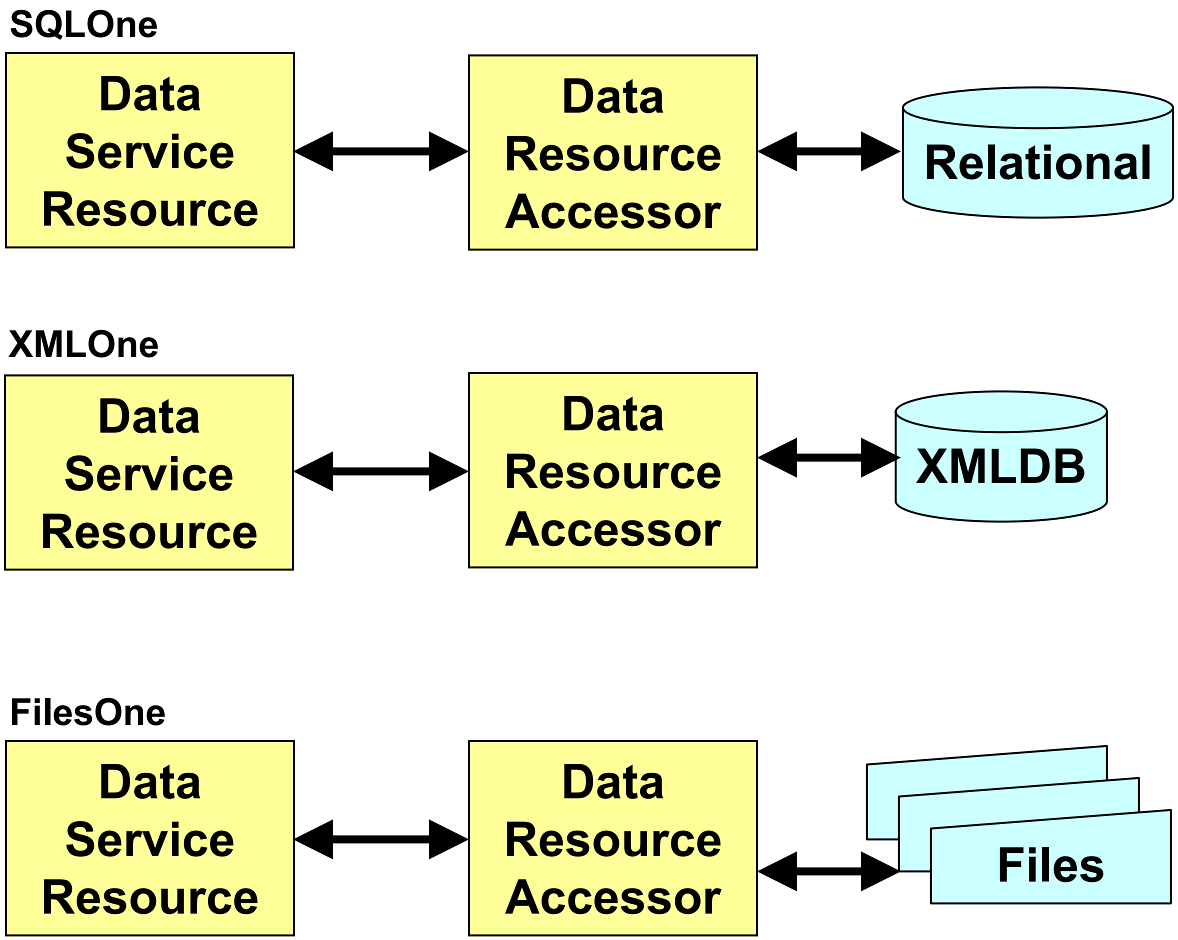


Data resource accessors

- Interfaces between data resources and OGSA-DAI
- Relational
  - JDBC drivers
  - `java.sql.*`
- XML
  - XMLDB API and compliant drivers
  - `org.xmldb.api.*`
- File system
  - Java file and directory utilities
  - `java.io.*`



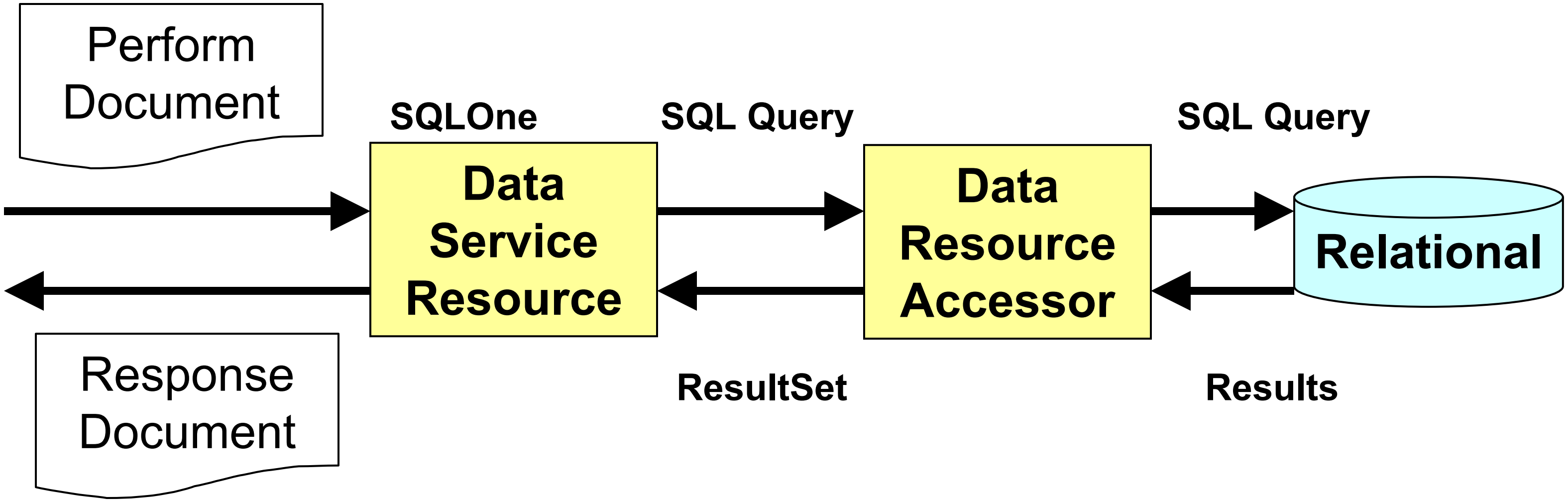
Data service resources



Data service resources

- OGSA-DAI’s core functionality
- Manages
    - Access to a data resource via a data resource accessor
    - Execution of data-related activities
    - Data caching and streaming of data to and from clients
    - Creation, access and termination of sessions
- Exposes data service resource properties
    - Information about a data resource
    - Information about supported activities
    - Information about current requests

Requests and responses



Requests and responses

- Request
  - A connected collection of activities that the data resource executes
  - Flow control – sequential or parallel execution of activities
  - XML perform document submitted by a client
- Activity
  - An individual data-related operation
  - 0 or more inputs and 0 or more outputs
- Response
  - Status of execution of a request possibly with result data
  - XML response document returned to a client
- OGSA-DAI engine
  - Parses requests, executes activities, builds responses

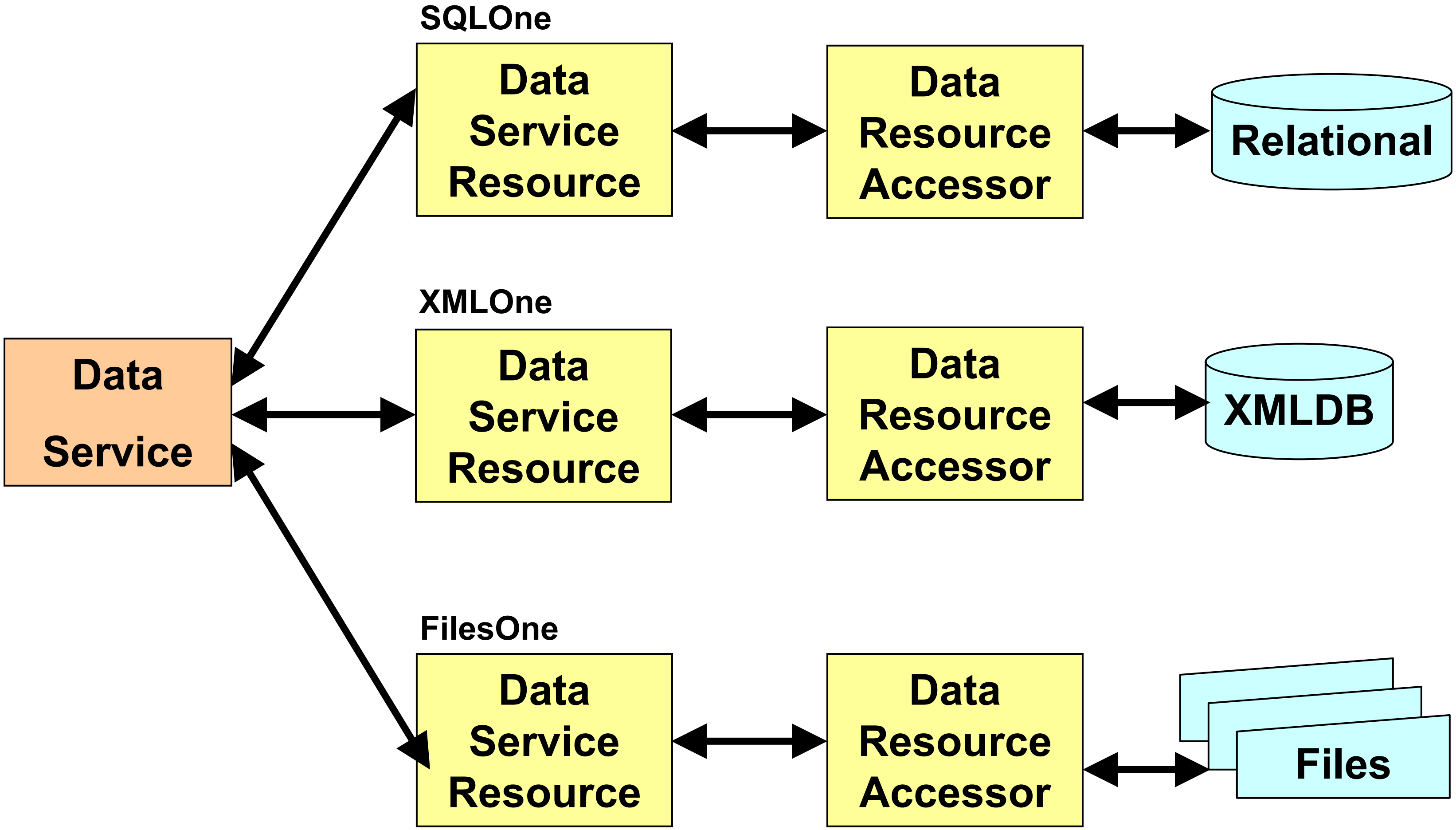
Activities

- Relational
  - SQL query, update, stored procedure, bulk load, extract logical and physical schema
  - Convert ResultSet to WebRowSet, ResultSet to CSV, ResultSet to bytes, relational database schema to XML
  - Project ResultSet or WebRowSet onto a column
  - Extract bytes from ResultSet
- XMLDB
  - Resource and collection management, XPath, XQuery, XUpdate, bulk load
- Files
  - List directory, create, read, write and update files
  - Index files, search indexed files

Activities

- Transformation and Compression
  - GZIP compression, ZIP archive
  - XSLT
  - Project CSV data onto a column
  - Distribute numerical data onto spaces
  - Create random sample of data
- Delivery
  - From and to URLs, files, GridFTP, remote data service resources, SOAP attachments
  - To servlets, SMTP, resource properties
- Factory
  - Create/destroy persistent/transient data service resources
- Relational multi-resources
  - Bag and resilient queries

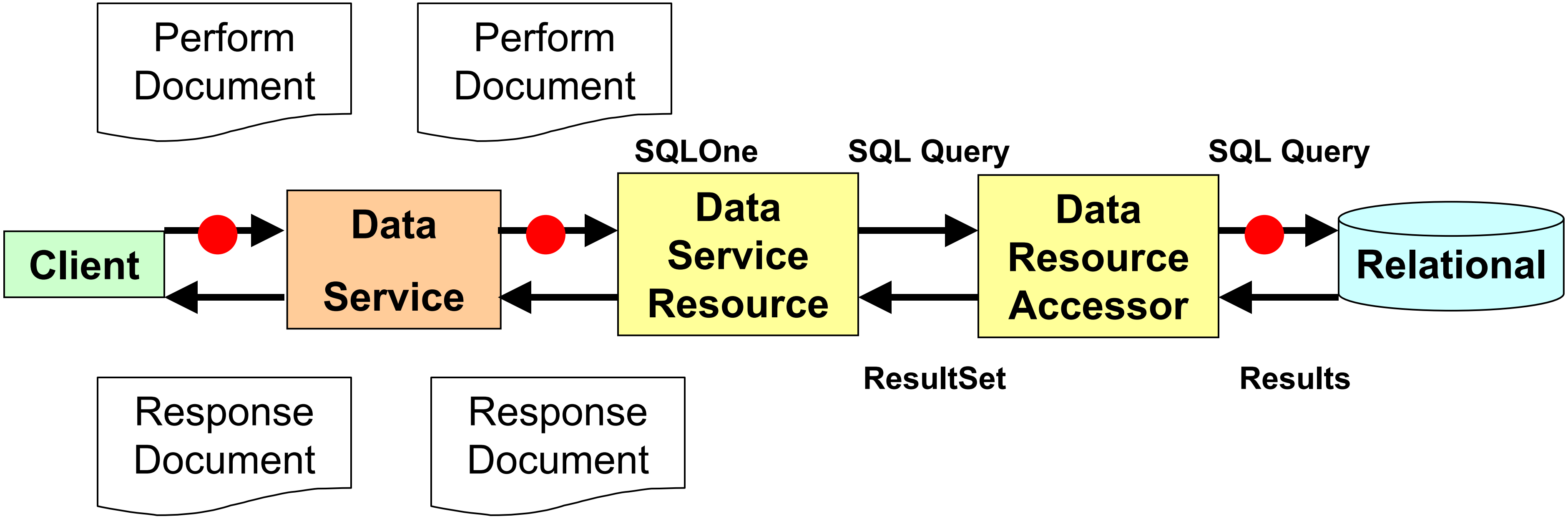
Data services



Data services

- Web services
- Expose 0..N data service resources to the outside world
- Two flavours
  - OGSA-DAI WSRF services
    - Compliant with the Web Services Resource Framework
    - Implemented using Globus Toolkit (4.0+)
  - OGSA-DAI WSI services
    - Compliant with vanilla WSDL
    - Implemented using Apache Axis (1.2.1 or 1.2RC3)

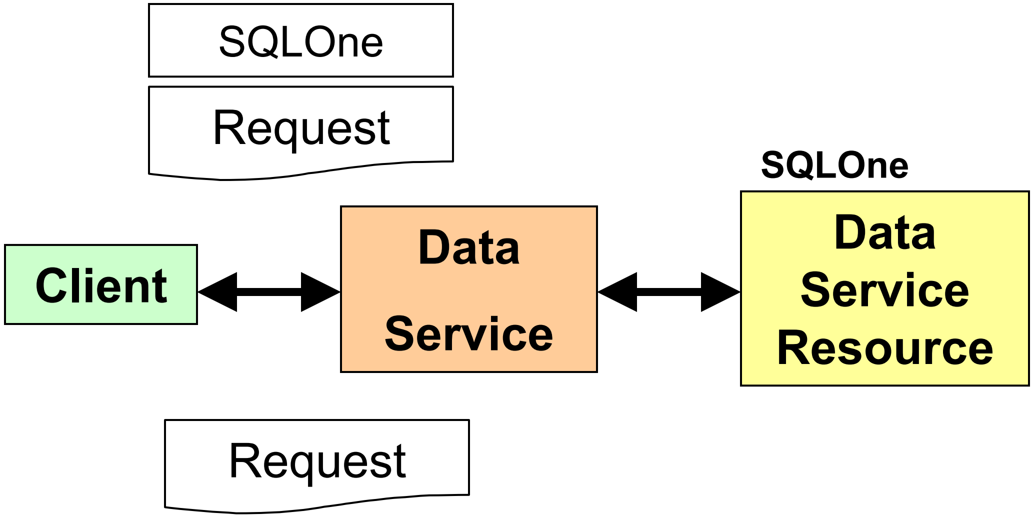
Clients



● Authorization points

Identifying a data service resource

http://host:port/services/wsrf/DataService



SQLOne

Request

SQLOne

Client

Data Service

Data Service Resource

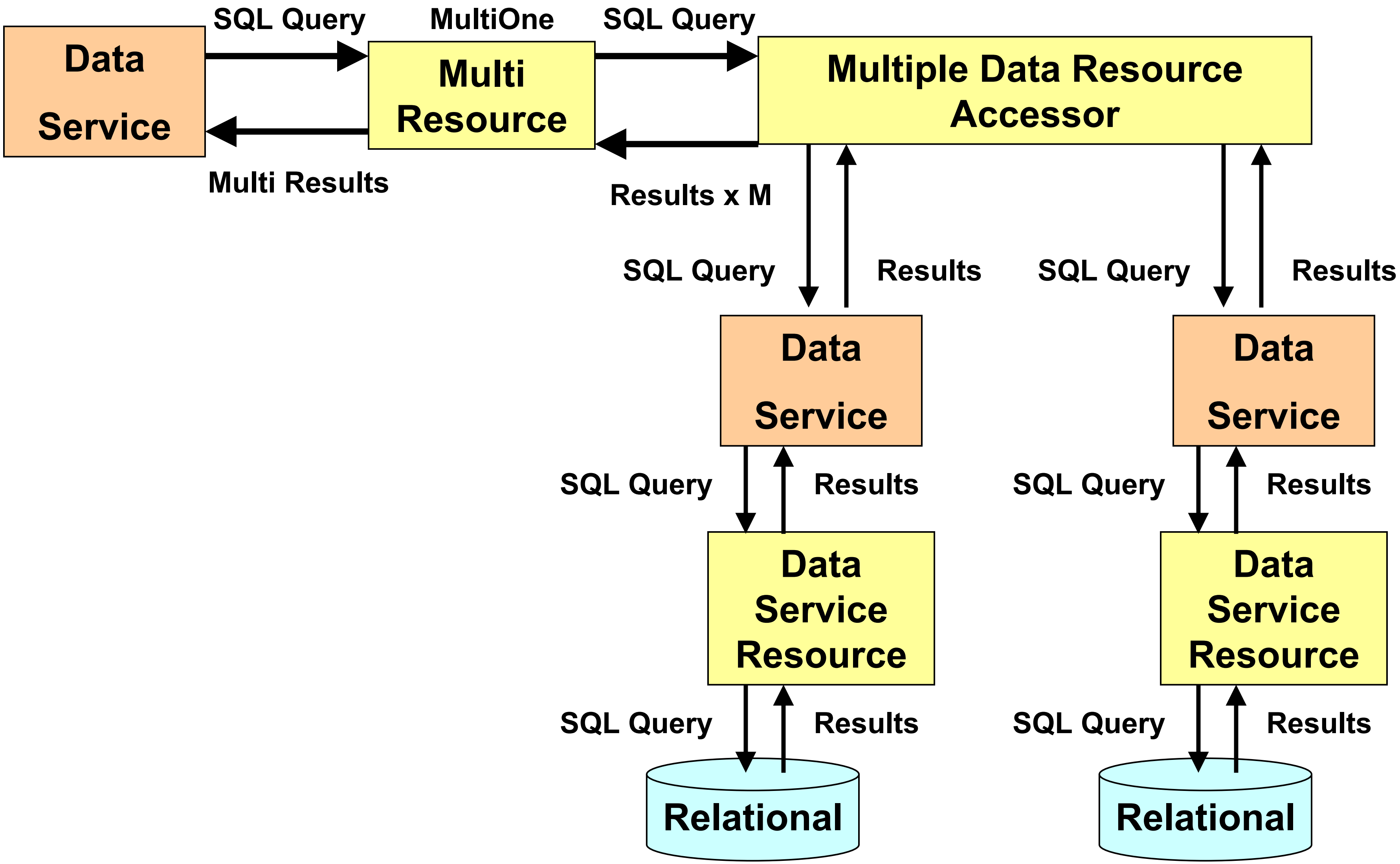
Request

http://host:port/services/axis/DataService/DAISQLOne

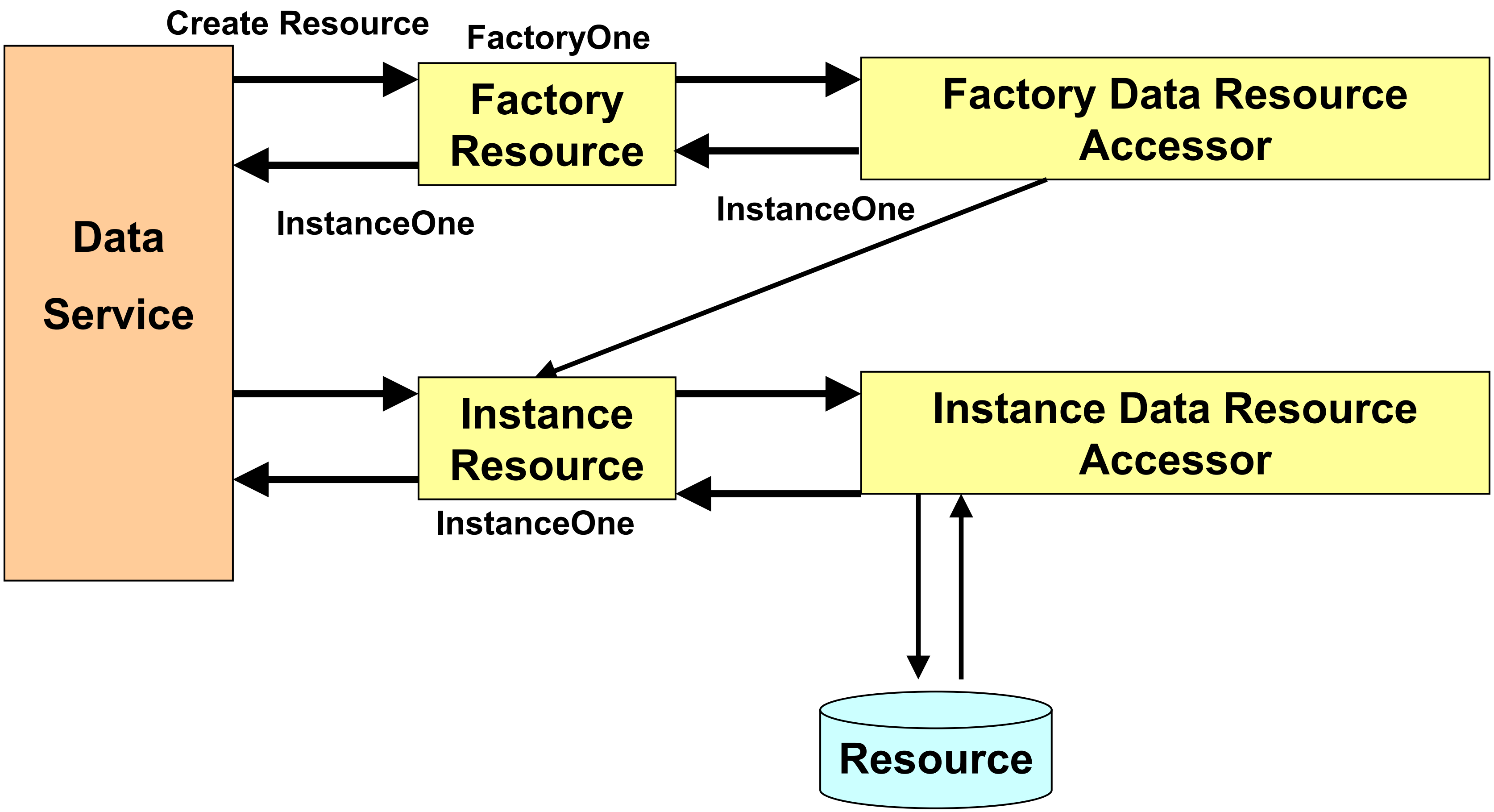
Clients and the client toolkit

- Clients interact with data services via SOAP over HTTP
  - Deduce service interface from service WSDL description
  - Construct SOAP request to invoke operation
  - Parse SOAP response from service
  - Resource identification scheme must be assumed from WSDL namespace
- OGSA-DAI client toolkit:
  - Construct and submit requests in Java not XML
    - Toolkit handles SOAP request construction and response parsing
  - Renders OGSA-DAI service types transparent
  - Java abstractions of
    - Data services
    - Data service resource IDs and session IDs
    - Requests and responses
    - Activities

Relational multi-resources



Factory resources



Extending OGSA-DAI

- Application-specific data resource accessors
  - Expose local or remote data resources
  - Expose virtual resources created by aggregation or integration
  - Create/destroy of persistent/transient data service resources
- Application-specific activities
  - Can be resource specific e.g query or update
  - Or generic e.g. transformation, compression, delivery, resource management, monitoring
- Application-specific authorization
  - Resource access
  - Activity execution

OGSA-DAI today - what's in a name?

- April 28th 2006
- It's not release 8
  - The lingering spectre of OGSI
- It's release 2.2
  - OGSA-DAI WSRF 2.2
    - Runs under Globus Toolkit 4.0.1 and 4.0.2
    - Patched version bundled with Globus Toolkit 4.1.0
  - OGSA-DAI WSI 2.2
    - Runs under Apache Axis 1.2.1 on Tomcat
    - Runs under Axis 1.2RC3 on Tomcat
    - Runs under OMII 3.0.1
    - Patched version bundled with OMII 3.
- http://www.ogsadai.org.uk

Resource management

- Transient data service resources
  - Exist only in memory
  - No associated configuration files on the server
  - Associated activity to create new transient resources
- Resource withdrawal
  - An activity to withdraw a data service resource
    - Optionally remove its configuration files from the server

Relational multi-resources

- A data service resource that aggregates multiple relational data service resources
- Aggregated resources can be local or remote
- Associated activities
  - Submit a query to every aggregated resource and return a bag of the results
  - Submit a query to every aggregated resource and return the results from the first one that completes
  - Remove duplicate rows from a bag of the results

Activities…

- Data conversion activities
  - Convert ResultSet to WebRowSet
  - Convert ResultSet to CSV
  - Convert BLOB from ResultSet column into bytes
- Relational meta-data activities
  - Retrieve logical database schema
  - Convert logical database schema to XML
  - Retrieve physical database schema
- XMLDB activities
  - XQuery

…and some more

- Projection and transformation activities
  - Remove duplicate rows from a WebRowSet
  - Project a ResultSet onto a column name or index
  - Project a WebRowSet onto a column name or index
  - Project CSV values onto a column index
  - Distribute numeric values onto spaces
  - Generate a random sample of input data
  - Write a stream of bytes to a temporary file and output a reference to this file
- Delivery activities
  - Throw data away
  - Write data to a resource property
  - Deliver data to a SOAP attachment

Then there's…

- eXist data resources and XQuery
- BLOBs
  - Improved support for BLOBs in SQL query and update activities
  - Activity to dump BLOBs into temporary files server-side
- Security
  - Resource and activity authorization
  - GSI Secure Conversation message-level security for inter-service communications using data transport
    - OGSA-DAI WSRF only
- Revamped logging, exceptions and internationalization
- Usage
  - Publication of an initial set of usage scenarios and best practice

…and…

- Bundled third-party JARs
  - OGSA-DAI WSI
    - JARs required to compile OGSA-DAI source distribution JARs required to run OGSA-DAI clients
  - OGSA-DAI WSRF
    - Non-Globus Toolkit JARs required to compile OGSA-DAI source distribution
    - JARs required to run OGSA-DAI clients
- Benchmarking and performance
  - WebRowSet – up to 35% better than in 2.0
  - resultSetToCSV activity – CSV yields up to 65% improvement compared to WebRowSet
  - Binary data transfer
    - Takes 25% of the time to transfer a binary 8MB file using SOAP attachments instead of in SOAP body
    - Improvements due to smaller data size and smaller SOAP messages
    - Limited by I/O performance rather than CPU
  - Numerous other improvements

Summary

- OGSA-DAI is an extensible framework for building data access and integration applications
- The OGSA-DAI layer cake
  - Data
  - Data resource accessors
  - Data service resources
  - Data services
  - Clients
- Extending OGSA-DAI
  - Data resource accessors
  - Activities
  - Authorization

Further information

- The OGSA-DAI project site
  - http://www.ogsadai.org.uk
- The DAIS-WG site
  - http://forge.gridforum.org/projects/dais-wg
- OGSA-DAI users mailing list
  - users@ogsadai.org.uk
  - General discussion on OGSA-DAI, data and the grid
- Formal support for OGSA-DAI releases
  - http://www.ogsadai.org.uk/support
  - support@ogsadai.org.uk
- OGSA-DAI training courses

OGSA-DAI