

# File Catalog Development in Japan e-Science Project

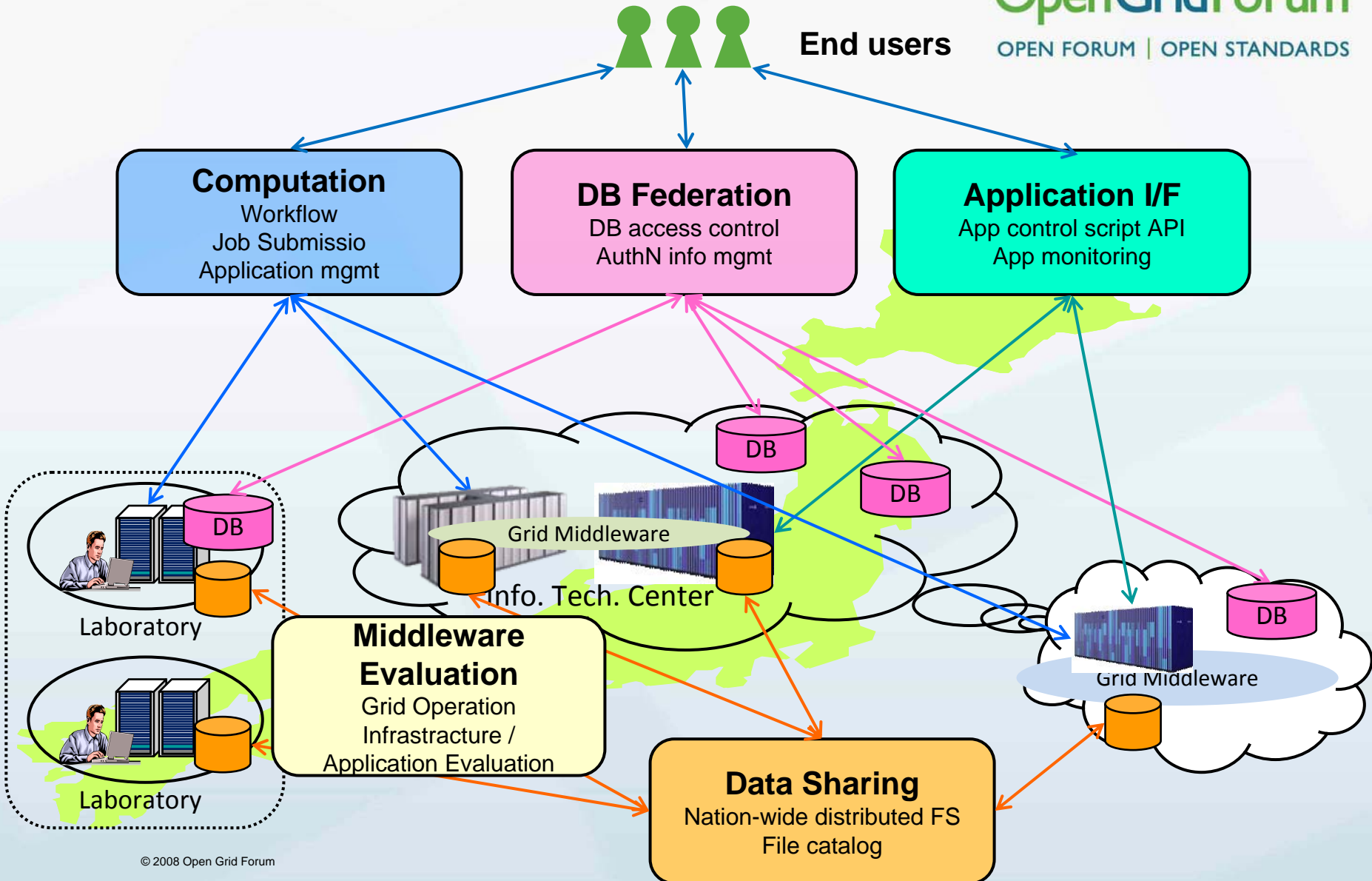
GFS-WG, OGF24 Singapore

Hideo Matsuda  
Osaka University

# Japan e-Science Project

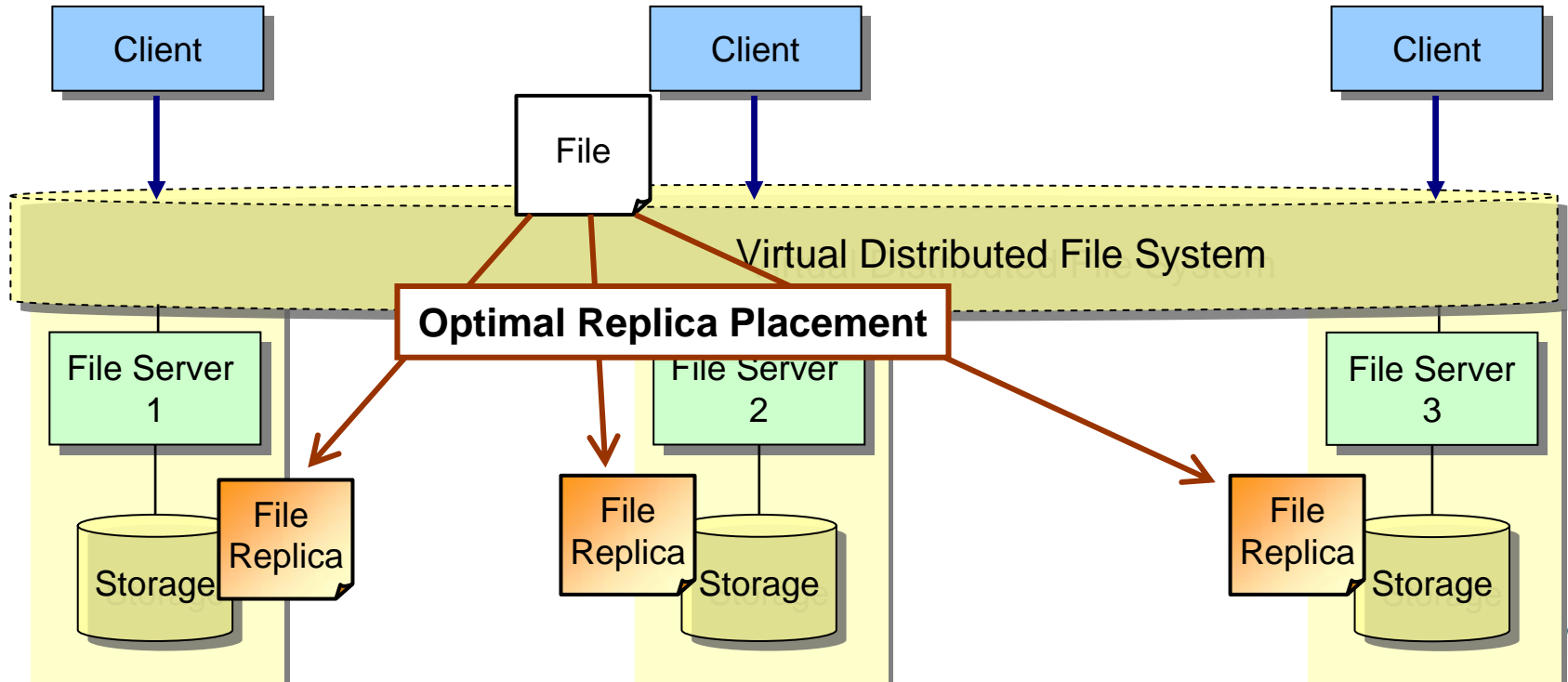
- 3.5 years project, starting from September 2008
- Sponsored by MEXT (the Ministry of Education, Culture, Sports, Science and Technology), Japan
- Two major sub-projects
  - System Software (Leader: Yutaka Ishikawa, Univ. Tokyo)
  - **Grid Software** (Leader: Ken-ichi Miura, NII)

# Overview of e-Science Grid Software Project



# Nation-wide Distributed File System

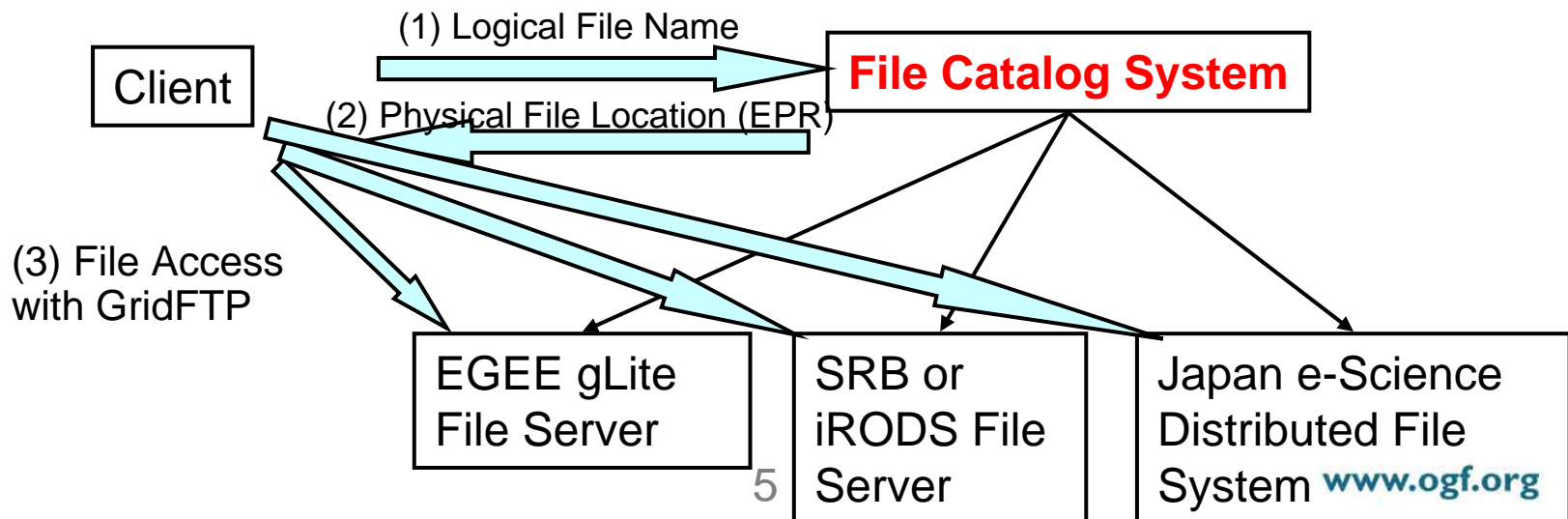
- Goal: Development of distributed file system technology spread over nation-wide with comparative performance of local fileserver
- Research Topics:
  - Optimal automatic placement of file replicas based on Gfarm 2.0.
  - Fault tolerance with file replicas



# File Catalog Service

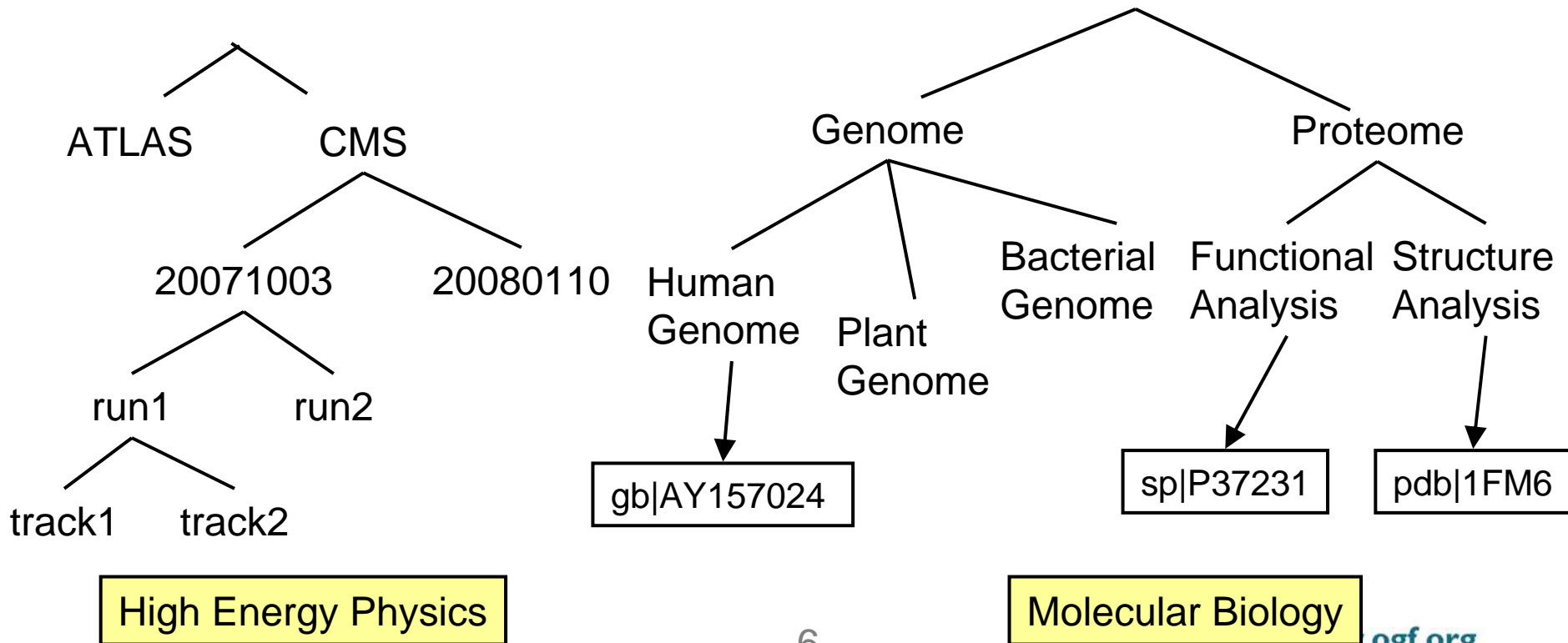
Goal: Development of interoperable file catalog service between heterogeneous Grid environments.

- Current file catalog systems (LFC (EGEE gLite), MCAT (SRB), etc.) does not have interoperability to each other.
- Development of standardized file catalog based on **RNS (Resource Namespace Service)** specification.



# File Catalog in e-Science

- File Catalog can be used for not only file-location management but also **metadata** in e-Science since metadata is often described with *hierarchical representation* in many sciences.



# Metadata Management using File Catalog

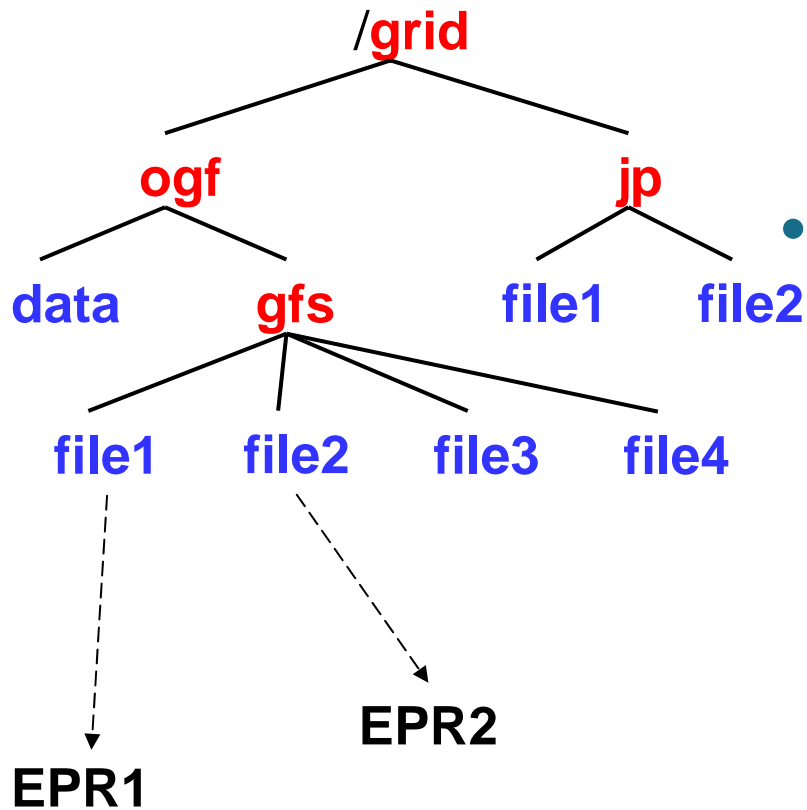
- Currently metadata are mainly stored in File Catalogs using their **hierarchical namespace** functionality.
  - gLite: LFC, Fireman
  - iRODS (SRB): ICAT
  - Globus: RLS
  - NAREGI: Gfarm
- It is not easy to exchange metadata over different Grid middlewares.

# Resource Namespace Service (1)



- RNS lets you map any resource into single, **hierarchical namespace**
- Resources are referred to in a form of EndpointReference (WS-Addressing)
- RNS Specification is published as GFD-R-P.101  
<http://www.ogf.org/documents/GFD.101.pdf>
- RNS implementation is available from U.Virginia and U.Tsukuba.

# Resource Namespace Service (2)

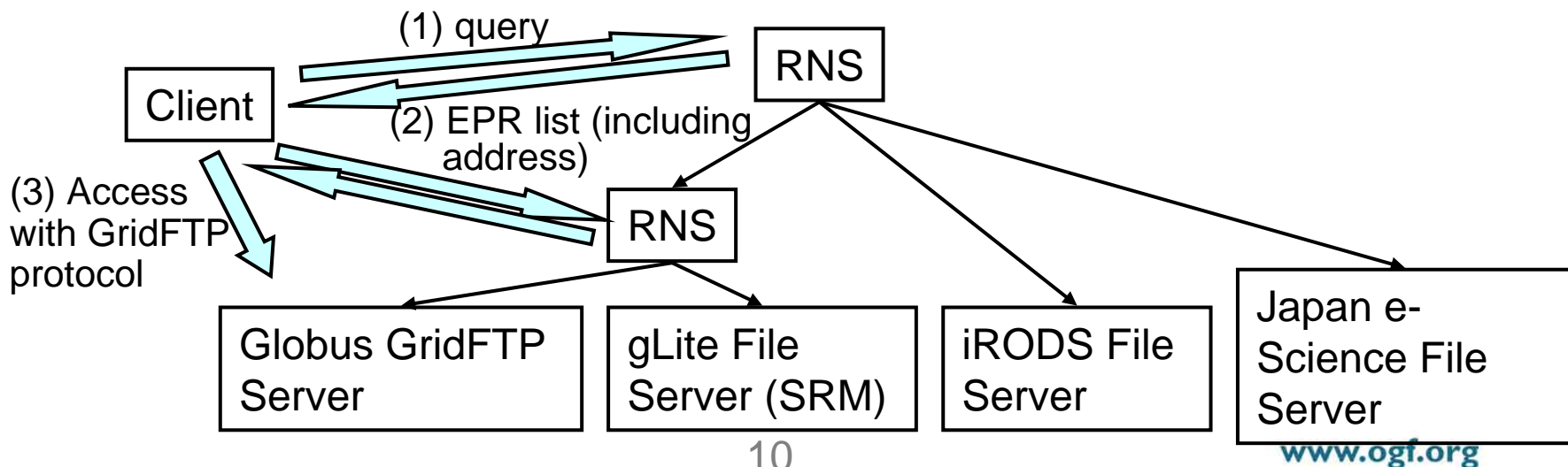


EPR: Endpoint Reference

- Hierarchical namespace management that provides name-to-resource mapping
- Basic Namespace Component
  - **Virtual Directory**
    - Non-leaf node in hierarchical namespace tree
  - **Junction**
    - Name-to-resource mapping that interconnects a reference to **any existing resource** into hierarchical namespace

# Development of File Catalog System (Plan)

- RNS can interconnect a reference to **any existing resource** into hierarchical namespace
  - Most of Grid middlewares have GridFTP for data transfer
- Use RNS as a **standardized** File Catalog
- Use GridFTP URL “gsiftp://.../” as the address of Endpoint Reference.



# Comparison with gLite LFC

## Comments from Erwin Laure (OGF22 GFS-WG)

- add EPR: RNS is missing the detailed attributes of the replicas.
- query EPR: The attributes of a namespace entry should be defined, allowing specialized queries and lookups.
- RNS lacks bulk operations, sessions, transactions. Adoption of those may improve performance.
- Access control and VO management are also not introduced yet.

# Comparison with iRODS

## Comments from Reagan Moore (OGF23 GFS-WG)

- Applications now manipulate structured information. iRODS can generate and manipulate structured information with micro-services.
- Multiple standards for describing structured information.

# Summary

- Standardized File Catalog is useful for federating heterogeneous Data Grids.
- Need to establish File Catalog Profile for interoperation of different File Catalogs (and for its standardization).