Scientific Data Management at the Environmental Molecular Sciences Laboratory

Daniel R. Adams, David M. Hansen, Kevin G. Walker

Pacific Northwest National Laboratory

John D. Gash

Lawrence Livermore National Laboratory

23 March 1998

Pacific Northwest National Laboratory

Operated by Battelle for the U.S. Department of Energy

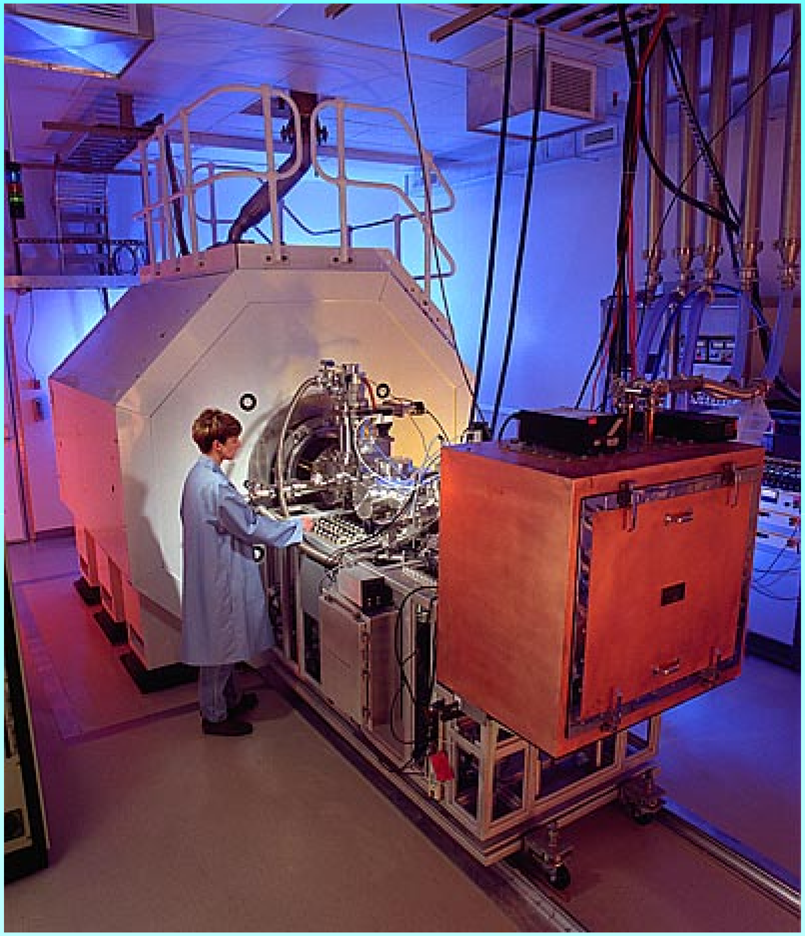
Environmental Molecular Sciences Laboratory

As a research organization, the EMSL seeks to

- **attain an understanding of the physical, chemical, and biological processes needed to solve critical environmental problems**
- advance molecular science in support of the DOE's long-term environmental mission.

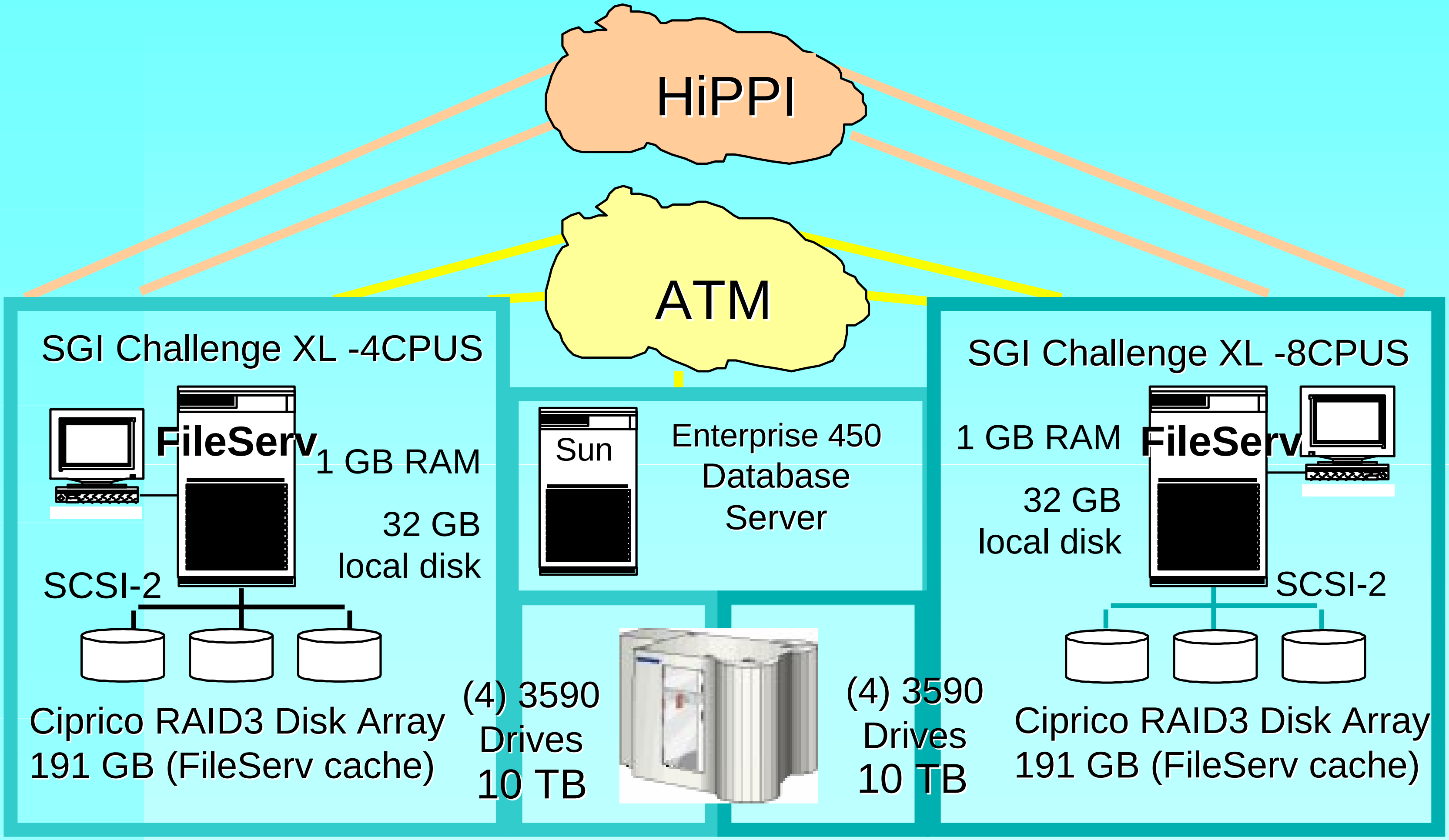
- As a national scientific user facility, the mission of the EMSL is to
  - **provide advanced and unique resources to scientists engaged in research on critical problems in the environmental molecular sciences**
  - educate young scientists in the molecular sciences to meet the demanding environmental challenges of the future.

Environmental Molecular Sciences Laboratory



Pacific Northwest
National Laboratory
Operated by Battelle for the U.S. Department of Energy

NWArchive Hardware Configuration



Pacific Northwest
National Laboratory
Operated by Battelle for the U.S. Department of Energy

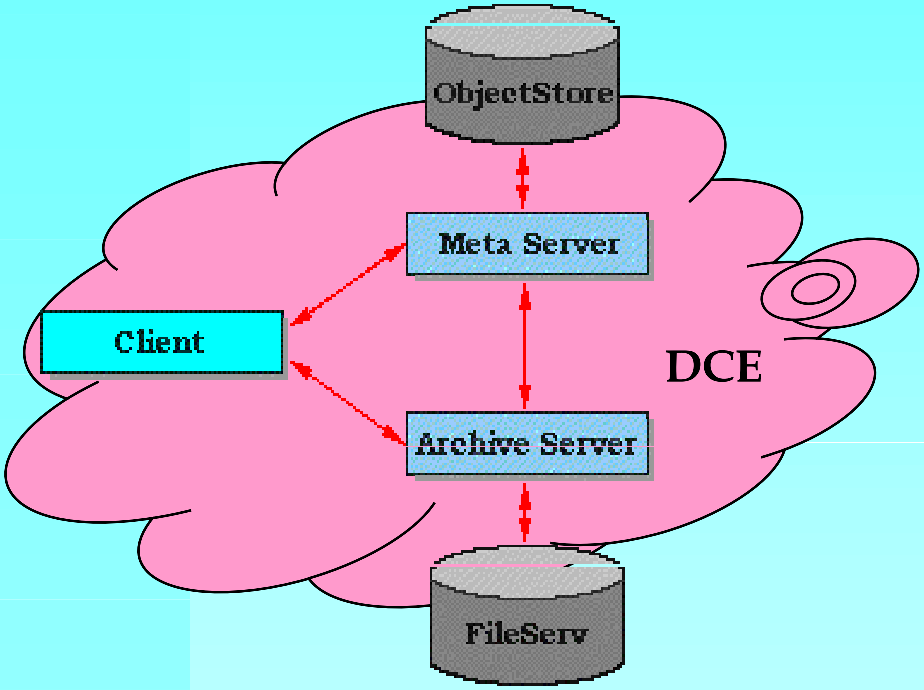
Scientific Data Management - Context

- Wide variety of data types produced
  - Computational chemistry
  - Experimental data
- Some data irreplaceable
- Must be accessible for 20+ years
- Must be able to find data
  - in spite of staff changes
  - without *a priori* knowledge of directory and naming conventions (and without grep)

Scientific Data Management - Design

- File oriented access
- Distributed, heterogeneous clients
- Distributed, homogeneous storage
  - but users see single file system
- Built using DCE
- Conservatively dependable
- Metadata Required
  - do not use predefined ontology
  - keep the metadata separate from the data

Scientific Data Management - Architecture



- Metaserver and database are key
  - maintain all metadata
  - maintain state of distributed transactions
  - balance storage distribution
- Archive Server
  - honors contracts
  - interacts with file system and transfers files

Pacific Northwest National Laboratory
Operated by Battelle for the U.S. Department of Energy

Scientific Data Management - Metaserver

- Distributed transactions implemented with “contracts”
- Metaserver and associated database
  - maintain state of contracts
  - implement unified file system view
  - maintain all metadata
  - implements access control
    - uses DCE groups
- Makes it possible to “hide” the file system

Scientific Data Management - Metadata

- Extensible, searchable metadata is key to making terabytes of data accessible for decades
- Three classes of metadata
  - File oriented
    - includes access notifications
  - Content oriented
    - Free text
    - Attribute=Value pairs (no predefined ontology)
  - Context oriented
- “Documentize” our metadata for searching

Scientific Data Management - Implementation

- Multiple threads x multiple vendors = multiple headaches
  - 2 stage delete
  - Solaris server added to mitigate DCE/ODI problems
- Clients
  - Command Line Interface
  - X interface (based on LLNL's xdir)
  - Web Interface (using Transarc's DFSWeb)

Scientific Data Management - Summary

- Client-Server architecture built on DCE
  - distributed transactions
  - authentication and authorization
- Metadata is key to long-term utility of system
  - extensible
  - file, content, and context
  - contemporaneous and *a posteriori* metadata
  - searchable
- “Just say ‘no’ to file systems”