Optimizing Globus File Transfer with Metadata-defined Virtual Collection

Martin Margo

Presented for 2015 Globus World Conference

Argonne, IL

April 14-15 2015

GENERAL ATOMICS

Scientific Workflow Tools are Disjointed

- **Typical scientific workflow**
  - (1) Idea → (2) experiment → (3) analyze data → (4) organize file → (5) publish/share
- **Steps 4 – 5 are brittle in our scientific workflow**
- **Great tools exist to help steps 4-5. How to combine them?**
  - Nirvana to address organizing files (step 4)
  - Globus to address publishing or sharing files (step 5)
  - Glue Python script to synergize them

Nirvana Metadata Driven Data Management Tool

- **Flexible**
  - Add user defined metadata to files
  - Virtual Collection: dynamic, real time, customizable, and actionable
  - Feed targeted dataset to Globus file transfer
- **Transparent gateways**
  - Windows: Network drive share
  - Mac OS X: NFS mount point
  - Specific applications: direct access using Nirvana SDK
- **Easy to install and use**
  - Binary installer available
  - Pre-built virtual machine available for PoC and Demo

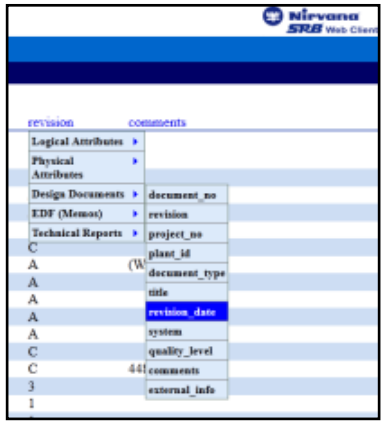
Nirvana Components

Client
User program interacts with MCAT and agents (Browser GUI, Python, Java, custom SDK)

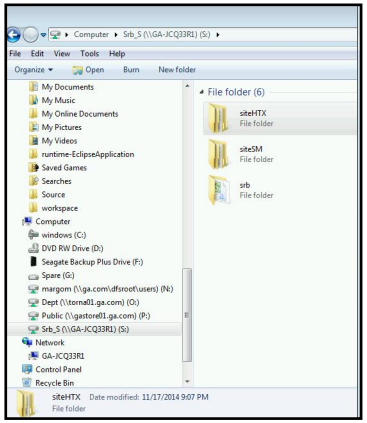
Metadata Catalog (MCAT)
Implements Global Namespace and the maps Data Objects to Storage Resources

Location Agent
Daemon directly accesses Storage Resource

Storage Resource
File or Object Storage System



Web Client
Command Line



CIFS & NFS Folders

Database
(Postgres, Oracle)

Storage Resource

• • •

Globus as High Performance File Transfer Tool

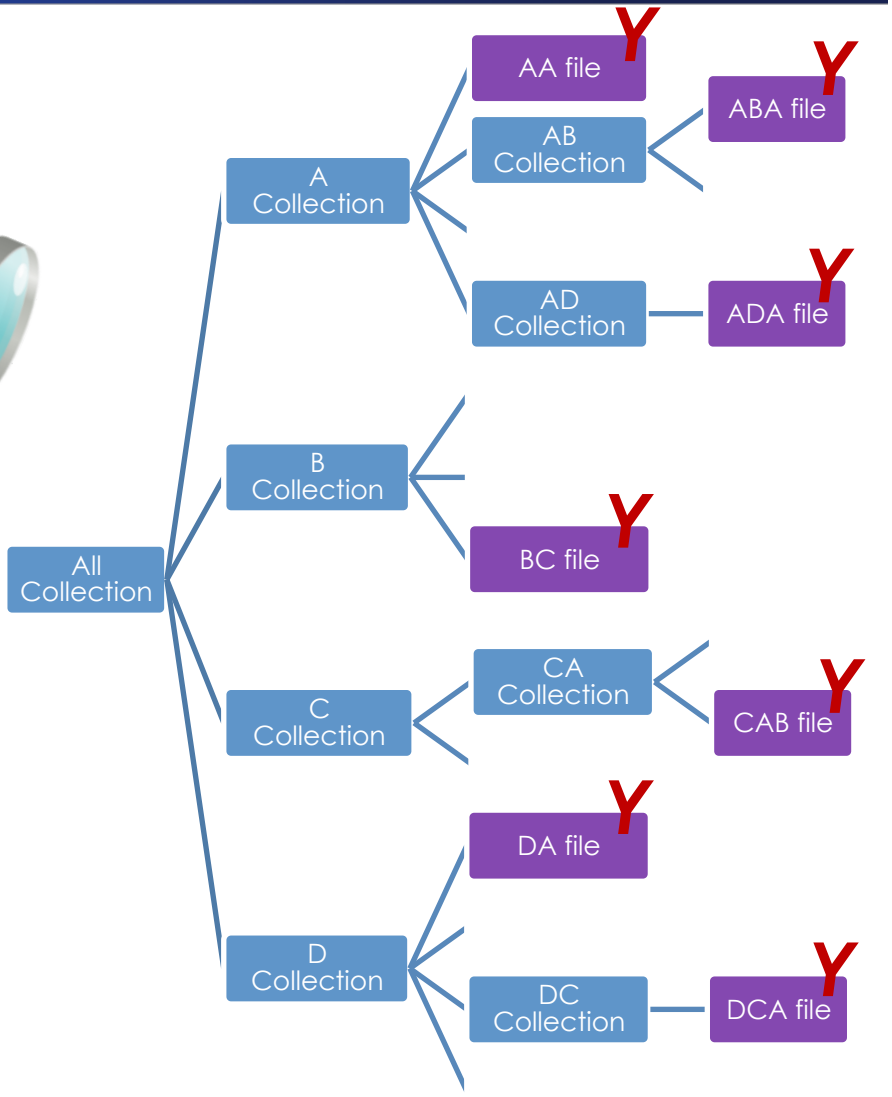
- **Globus**
  - Easy to use
    - ✓ Web GUI and drag & drop
  - High performance
  - Self service
  - Free for non-profit research organizations

Nirvana Virtual Collection is Read Only Dynamic Collection that is Driven by Metadata

"List my files having tag Y"

Nirvana SRB

Client



Virtual Collection definition in SRB MCAT:

```
((DATA_OBJECT.collection_name = '/home') OR
(DATA_OBJECT.collection_name like '/home/*')) AND
(DATA_OBJECT.data_type = 'PDF') AND
(EXPRESSION.create_age <60) AND (EXPRESSION.create_age
> 30) AND
(Cluster.Name in
                ('XYZ', 'Default'))
```

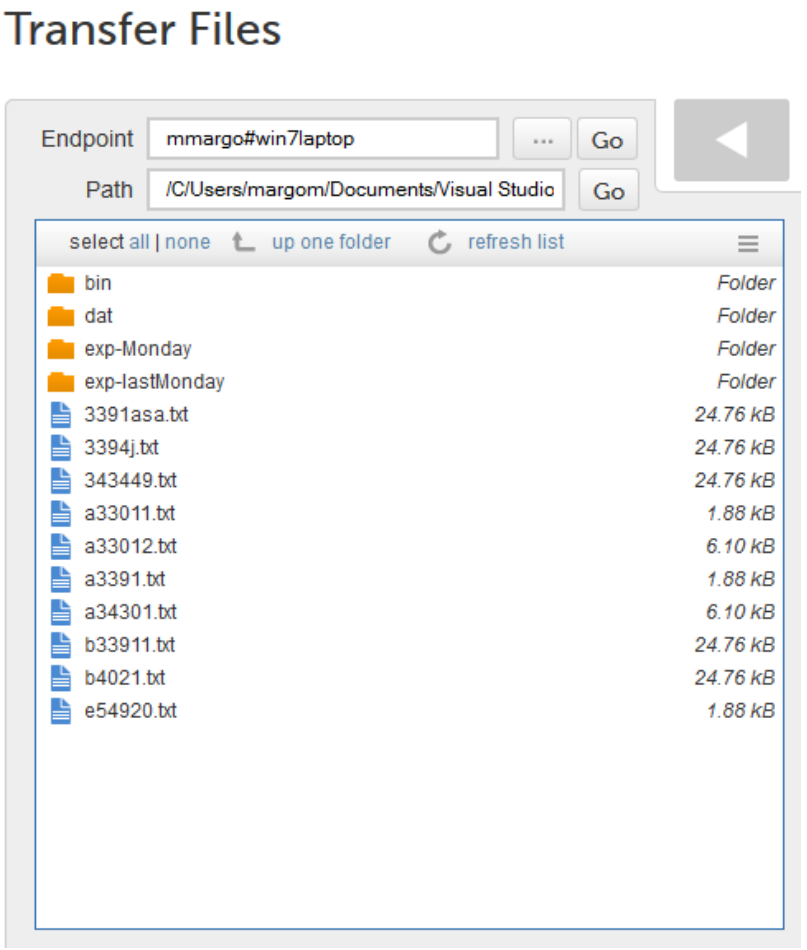
Steps

- **Create Nirvana Virtual Collection to organize research files based on user defined metadata**
  - Begin by registering research files in Nirvana
  - Create user defined metadata schema
  - Populate metadata on files during ingest or post ingest
  - Create virtual collection to share / publish

    Examples:
    - **Publish**: Type = PDF, experiment date not earlier than Mar2015, experiment run on cluster XYZ
    - **Verify**: Type = .dat | .out | .xml, experiment date not earlier than Mar 2015, experiment run on cluster XYZ

Steps (Cont.)

- **Easy route**
  1. Mount Nirvana virtual collections on local data mover server running Globus Endpoint via Nirvana FUSE driver / Nirvana NFS driver
  2. Export with Globus
  3. Inform 3rd party that data is available to be pulled



Steps (Cont.)

- **Optimized route**
  - Write glue Python script to
    1. Read Virtual Collection files from Nirvana using SDK API
    2. Use the resulting memory buffer to call Globus Transfer API
    3. Zero copy
    4. Well defined actors touching files

Conclusion

1. Use user defined metadata to drive policies
2. Use metadata to dynamically create logical groupings (virtual collection)
3. Use Globus to migrate data out of the Nirvana namespace

Nirvana

globus

Thank You!

Nirvana