LBNL/SDM Contribution to Open Science Grid (OSG) and Earth System Grid (ESG)

Scientific Data Management Research Group
Computational Research Division
Lawrence Berkeley National Laboratory

SDM group contributions to OSG

- **BeStMan Support**
  - VO support
    - US ATLAS, US CMS, STAR, and
    - Other smaller ones such as LIGO, SBGRID, CERN EOS, etc.
    - Supported areas
      - Deployment and configuration
      - Scalability/performance
      - Compatibility/interoperation with dCache
      - General SE (storage and file system) needs
  - User support
    - SRM client usage with BeStMan
      - e.g. lcg-utils, FNAL SRM clients (srmcp), bestman srm clients (srm-copy)
    - Data replication middleware
      - e.g. FTS, PhEDEx
- **VO-requested feature addition and maintenance**
  - Collaboration with OSG RPM software team

OSG BeStMan deployments

- Number of sites: ~57
  - Unofficial OSG statistics based on BDII information
- Last version release on May 15, 2012
  - BeStMan2 v.2.2.1
- 8 RPM packages
  - bestman2-common-libs
  - bestman2-server
  - bestman2-server-libs
  - bestman2-server-dep-libs
  - bestman2-client
  - bestman2-client-libs
  - bestman2-tester
  - bestman2-tester-libs
- RPMs are all built from the source release.

BeStMan Source availability

- Source codes under BSD with Grant-back provision
  - Available via SVN: https://codeforge.lbl.gov/projects/bestman/
    - Official LBNL source code public repository, maintained by Berkeley Lab
  - Binary package (tar.gz file with configure) also available
- Plug-in source codes
  - For load balancing on transfer server lists
  - Maintained as source and a simple package
  - Available via SVN https://codeforge.lbl.gov/projects/bestmancontrib/
- OSG VDT package with pacman for bestman release
- OSG RPM package for bestman2 release

Ongoing support items

- Collaboration with OSG software team
- OpenJDK support, Java 1.7 support
- Transition to https layer from the current httpg
  - Collaboration with EMI
  - Implementation
  - Interoperation
  - Compatibility
  - Transition

Berkeley Storage Manager (BeStMan)

- **SRM v2.2 implementation – OGF standard (Aug. 2009)**
  - interoperable and compatible to other implementations
- **Works on existing storages with posix-compatible file systems**
  - NFS, GPFS, Lustre, HPFS, XroodFS, PVFS, PNFS, HFS+, …
  - Adaptable to special file systems and storages with customized plug-in
    - Site-specific customization on the file system i/o mechanism
    - Plug-in extension for external archival storage systems
    - E.g. MSS such as HPSS, REDDnet
- **Supports multiple transfer protocols**
  - GridFTP, FTP, HTTP/S
- **Load balancing for multiple transfer servers**
  - Also, supports customized plug-in for transfer server selection with custom policy

BeStMan

- Supports multiple storage partitions
  - Supports pre-defined static space tokens
  - Supports space reservation
- Supports Gateway Mode for faster performance
  - Jetty based web server container
    - Better performance in http connection handling
    - Scalable and configurable for heavier load
  - Scales well with some file systems and storages, such as Xrootd and Hadoop
- Authentication and authorization
  - Supports grid-mapfile
  - Supports GUMS server – SAML and XACML based
  - Supports limited access to the underlying file system
    - User access restriction to certain directory paths
  - Supports limited permissions on file access
    - User access control to files by owners/creators only

BeStMan

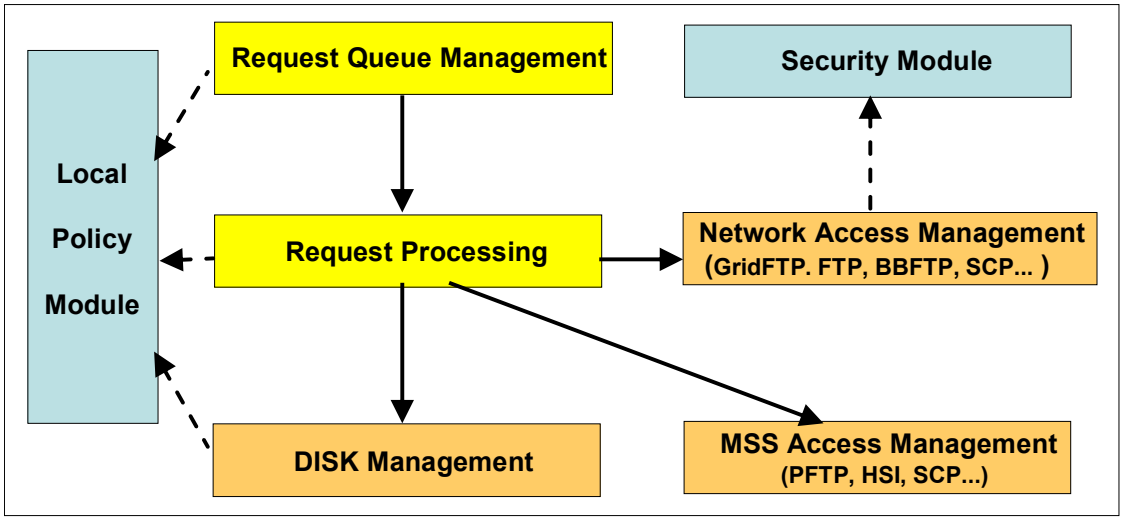
- **As data movement broker**
  - BeStMan manages multiple file transfers without user intervention when a request for large scale data movements of thousands of files is submitted.
    - Recovers from transient failures
    - Supports recursive directory transfer requests
    - Supports asynchronous status check
  - BeStMan verifies that enough storage space exists for file transfer requests
  - File movements from/to remote SRMs or GridFTP servers
  - E.g. STAR use case for data movements from NERSC/PDSF to BNL

BeStMan SRM Clients

- **Supports all interfaces in SRM specification**
  - **Interoperable and compatible to other SRM server and client implementations**
  - **Supports multiple transfer protocols**
- **Added functionality**
  - **User friendly command options**
    - **E.g. srm-copy –mkdir**
      creates recursive directories before transferring files into the target
    - **E.g. srm-copy –nooverwrite**
      avoids duplicate transfers when target file exists.
    - **E.g. srm-copy –gatewayfriendly**
      skips some redundant SRM calls for BeStMan Gateway.
      Works only for BeStMan Gateway mode.
  - **Supports 3rd party gridftp file transfers**
  - **Supports a bulk request**
    - **To reduce the load on the server, many single calls to SRM PUT requests are bundled together as one single request with many files in the request using –f option.**
  - **Available SRM Java API**
  - **Available SRM-Tester**

BeStMan Design

- Java implementation of SRM specification v2.2
- Designed to work with unix-based disk systems
- Adaptable to other file systems and storages via plugin mechanism
- MSS support to stage/ archive from/to its own disk
- Uses in-memory database (BerkeleyDB) for full mode



- Multiple transfer protocols
- Space reservation
- Directory management (no ACLs)
- Local Policy
  - Fair request processing
  - File replacement in disk
  - Garbage collection
- Can copy files from/to remote SRMs or GridFTP Servers
- Can copy entire directory recursively
  - Large scale data movement of thousands of files
  - Recovers from transient failures (e.g. MSS maintenance, network down)

Use case: BeStMan Gateway + Disk storage

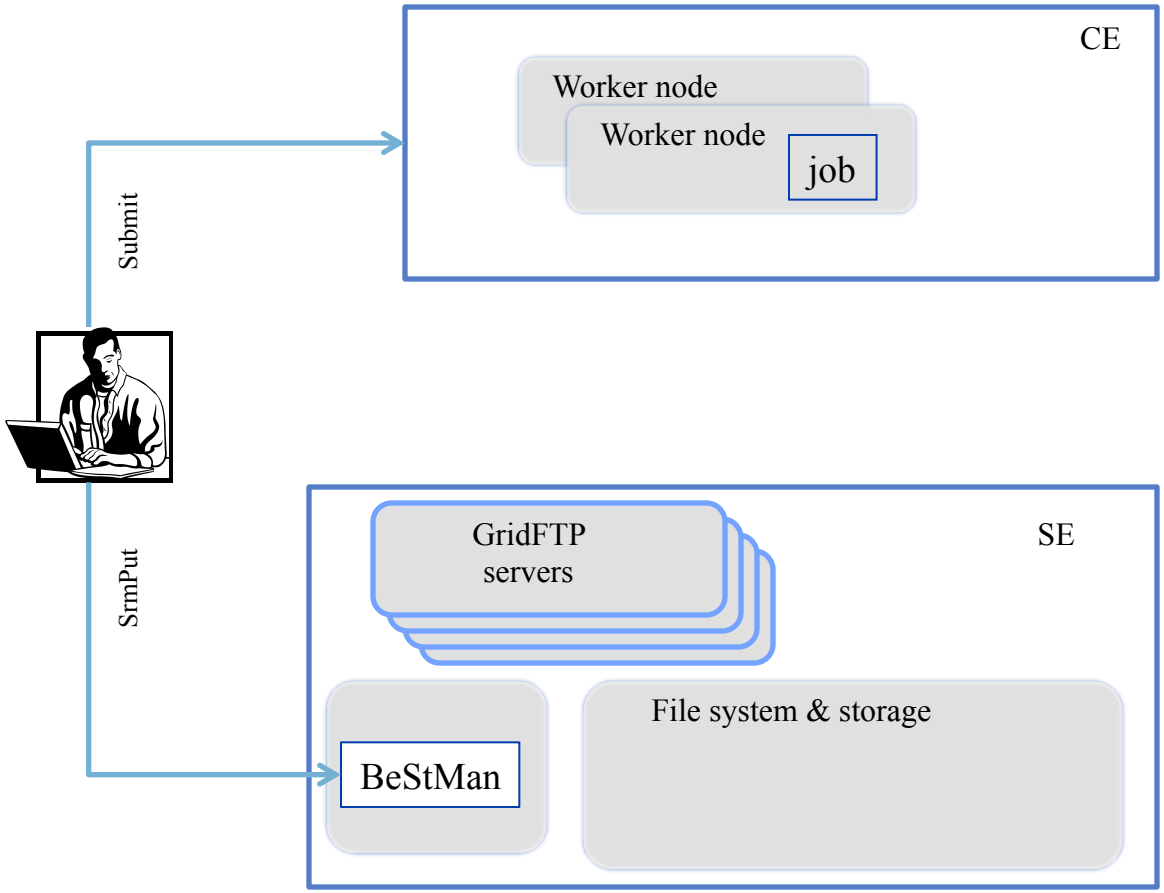


Image courtesy: Tanya Levshina

Use case: BeStMan Gateway + HDFS

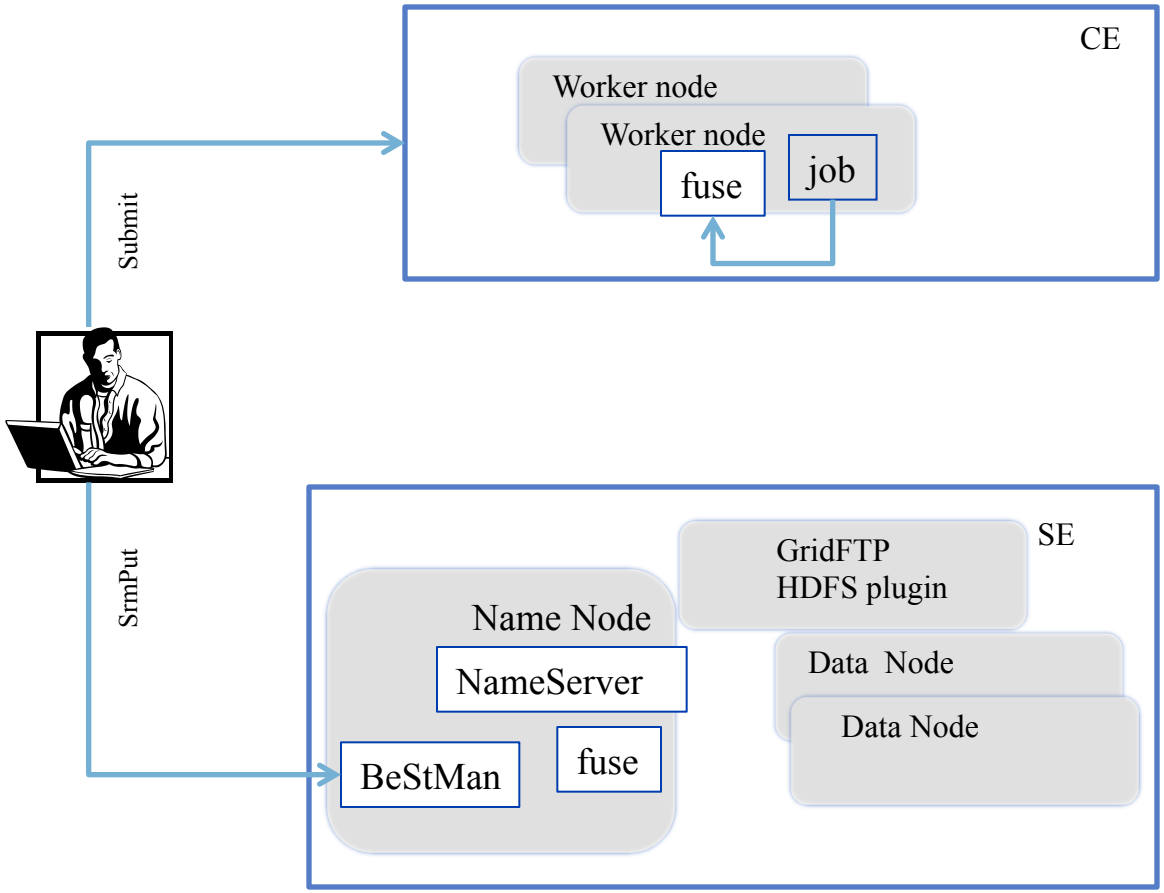


Image courtesy: Tanya Levshina

Use case: BeStMan Gateway + XrootD

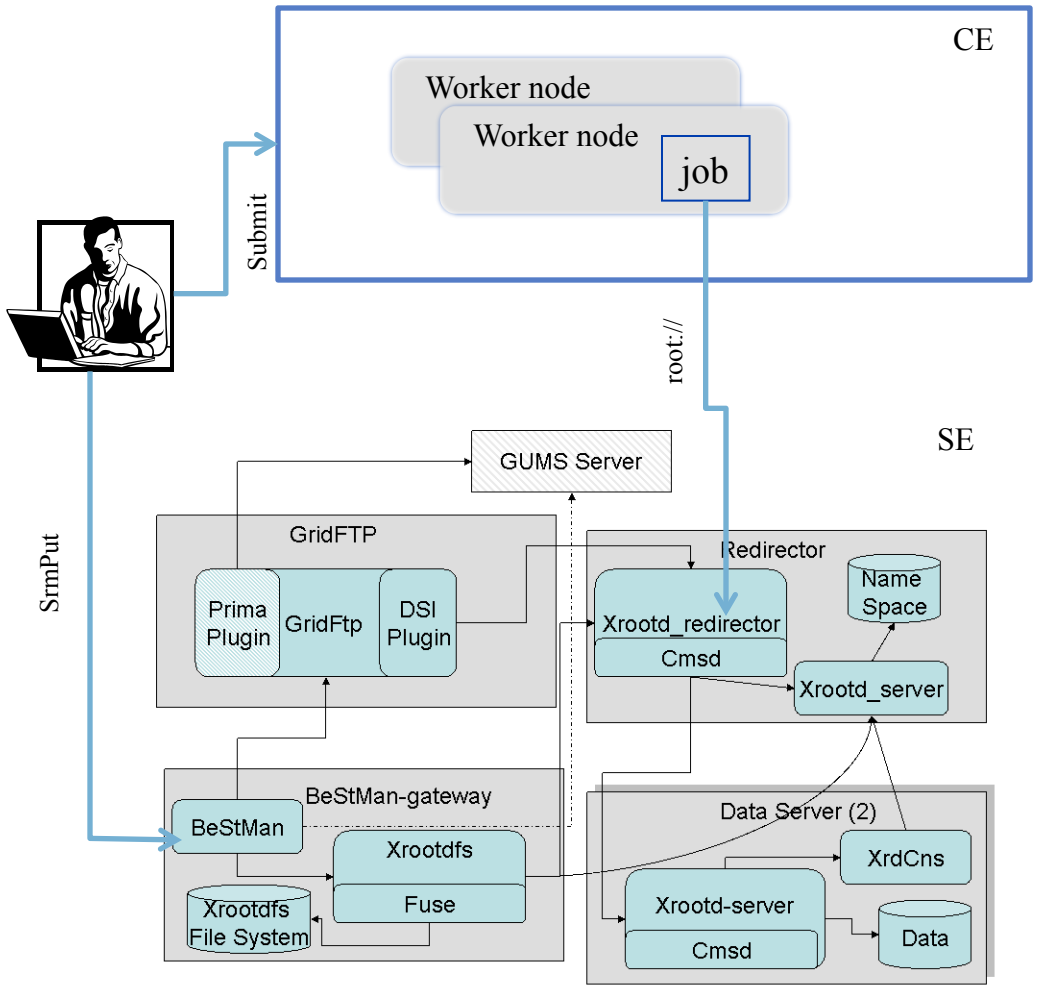
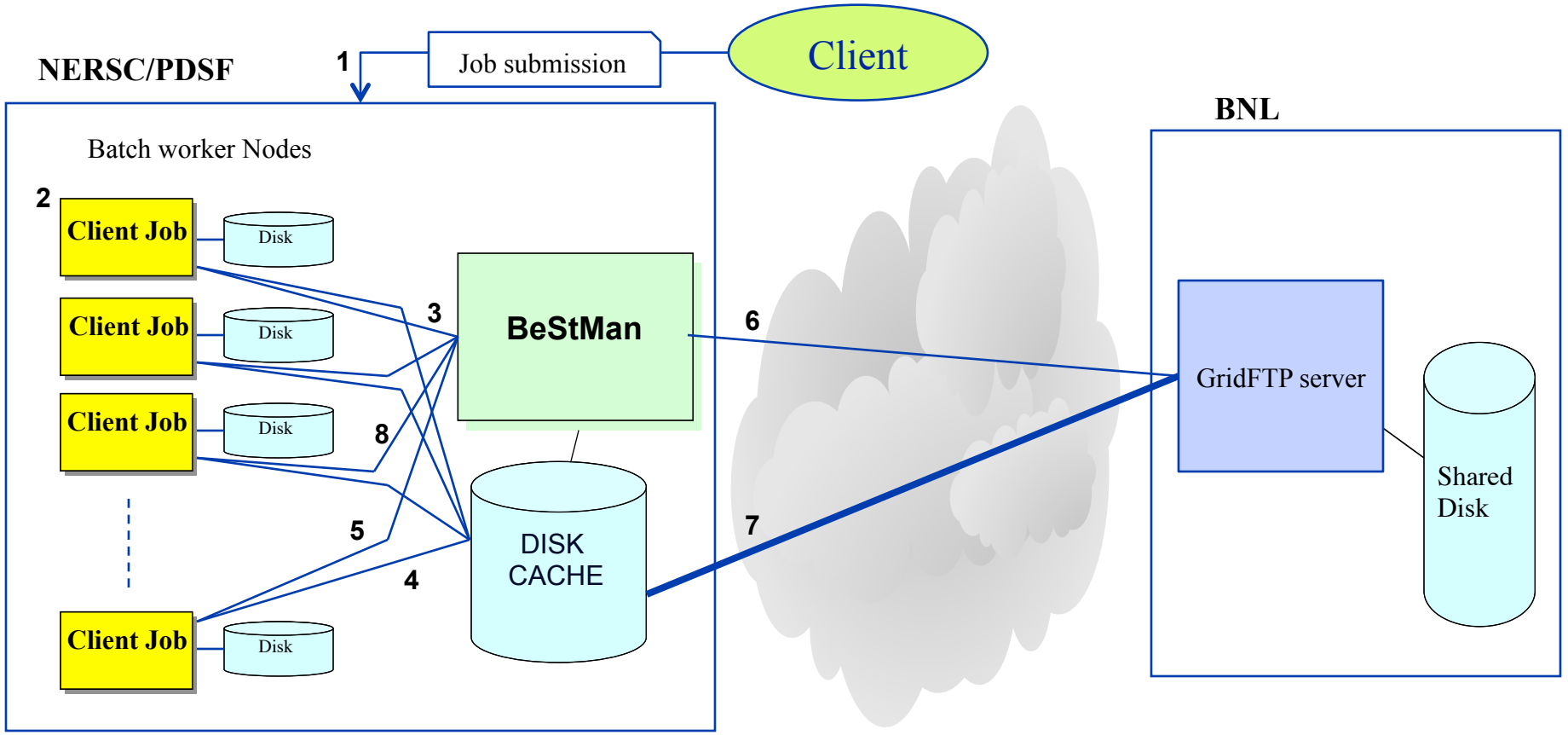


Image courtesy: Tanya Levshina

Use case: Job-driven data movement in STAR



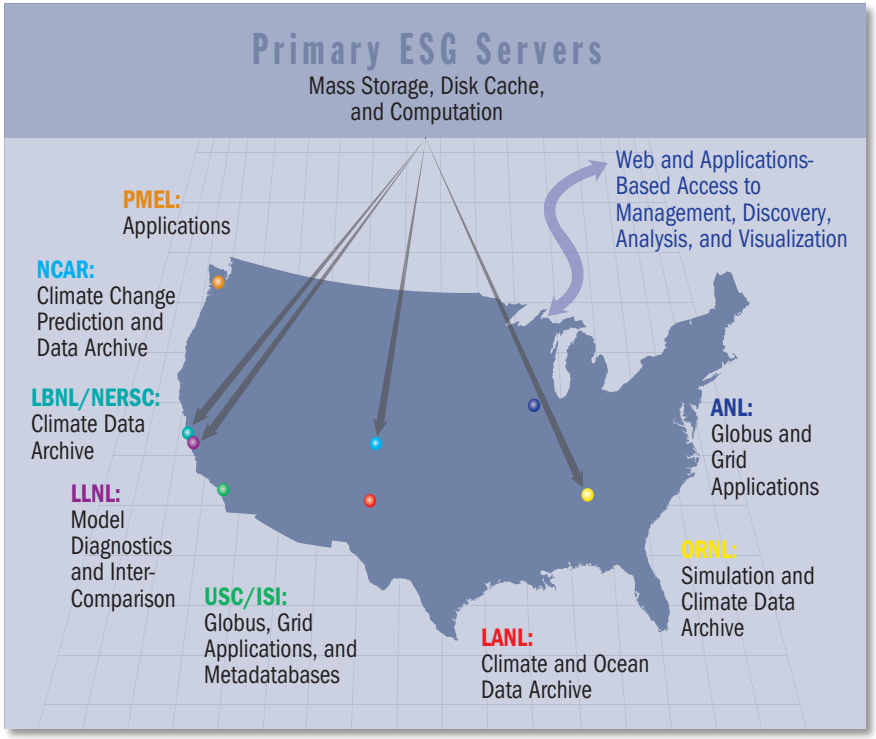
1. Client submits analysis job
2. Client jobs get created on the worker nodes, and create files
3. Jobs contact local bestman
   to move the result files to the remote storage repository
4. Client jobs (using bestman client) stage files into
   bestman managed locak disk cache via TURL
5. Client jobs (using bestman client) notify bestman
   for file staging completion
6. Local bestman contacts remote storage sites
7. Bestman transfers files to the remote sites via GridFTP
8. Client jobs check the status of the file transfers results
9. Client jobs finish upon successful status status

SDM group contributions to ESG

- Earth System Grid
- Berkeley Storage Manager (BeStMan)
  - BeStMan server deployments
    - At NCAR, LBNL/NERSC, ORNL and LANL
  - Support for customized MSS access for ESG Gateway
    - Support for customized site security and authentication
    - NCAR HPSS (previously MSS)
    - NERSC HPSS
    - ORNL HPSS
  - BeStMan access from ESG Gateway
    - SRM Java API and SRM clients from LBNL
- DataMover-Lite
  - ~4000 webstart downloads in the last year
  - ~60 downloads for stand-alone
- Bulk Data Movement and Climate data replications
- LBNL/NERSC ESGF P2P node

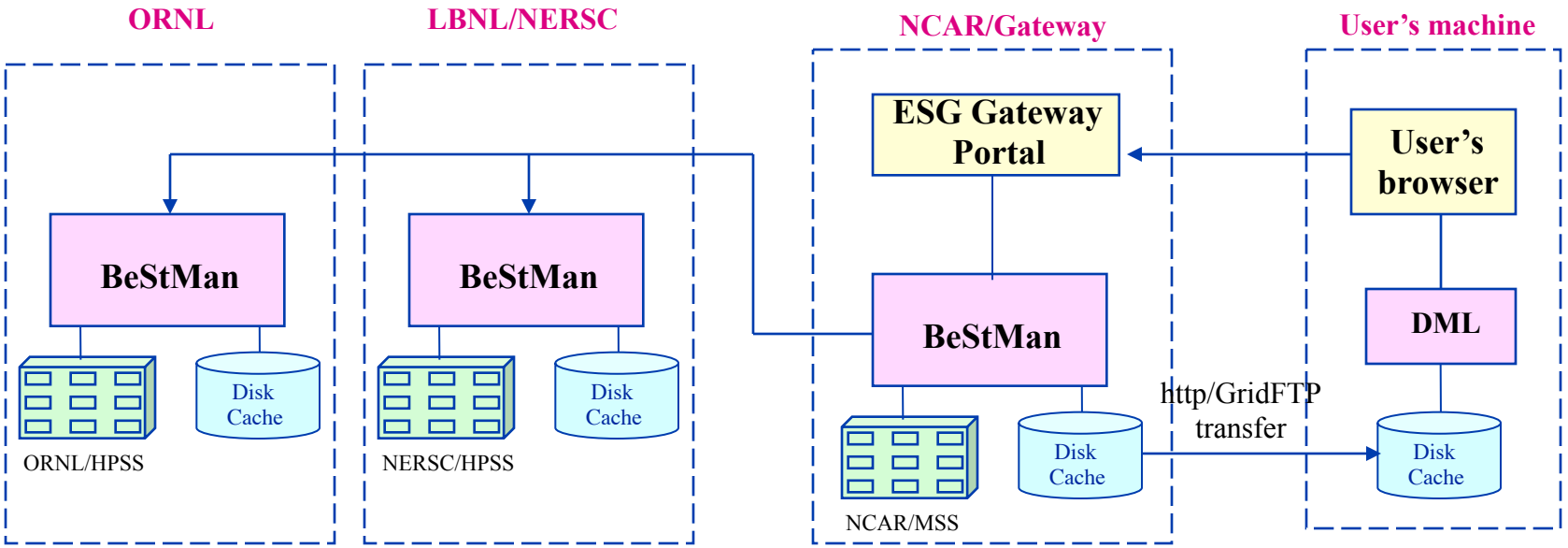
Earth System Grid

- **Earth System Grid (ESG)**
  - To support the infrastructural needs of the national and international climate community, ESG is providing crucial technology to securely access, monitor, catalog, transport, and distribute data in today's grid computing environment.
  - ANL, LANL, LBNL, LLNL, NCAR, ORNL, PMEL, USC/ISI
- **Project history**
  - ESG-I (1999-2001)
  - ESG-II (2001-2006)
  - ESG-CET (2006-2011)
  - ESGF (2012 -)
- **Production since 2004**
- **LBNL/NERSC contribution**
  - CCSM/CESM on HPSS
  - 36TB Replica of CMIP-3 (IPCC AR4)
  - 45TB Replica of CMIP-5 (IPCC AR5)
  - ~20TB of local CMIP-5 data



DataMover-Lite (DML)

- DML: ESG-specific versatile file download tool with simple graphical user interface
  - Works with ESG portals through Java web start as well as stand-alone program
  - Works with ESG authentication and authorization system
  - Works with ESG supported file transfers via http/https, gridftp, ftp and scp
- Current Scenario: simple HTTP/GridFTP download from ESG Gateway/P2P sites
  - User goes to ESG Gateway/P2P portal, selects files
  - (optional) Portal gets files into BeStMan disk from other MSSs or disks
  - Portal notifies user for files on disks
  - User uses DML to download files



DML Features

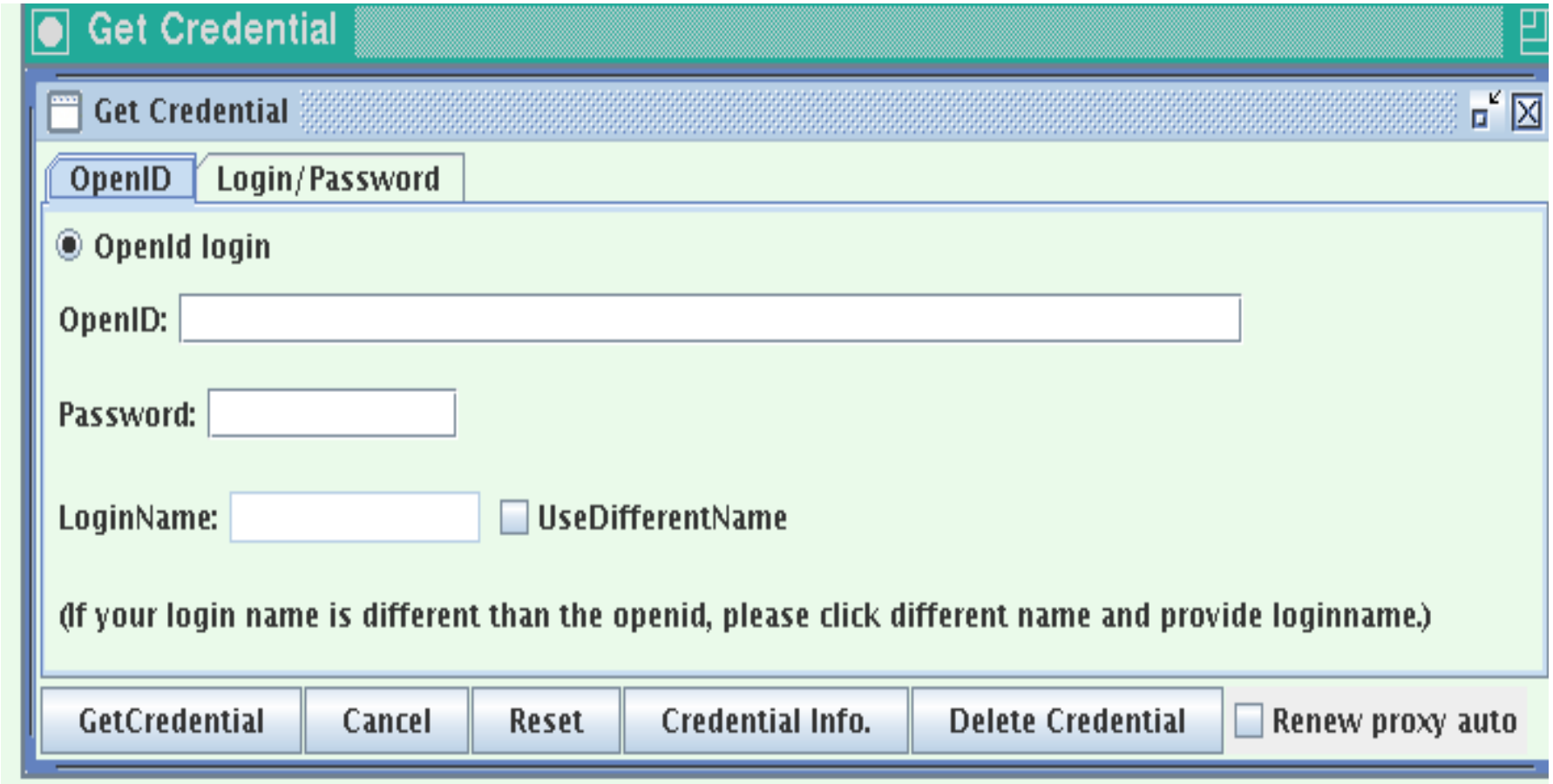
- **wget script integration with DML downloads**
  - **All http downloads from wget are integrated with DML webstart**
    - **Select the wget script as an input file of DML, and DML parses the wget script to download**
  - **File selection support from the wget download script for downloading subset of files within the request**
- **User friendly authentication**
  - **DML includes myproxy servers as a dropdown list that user can choose from and ESGF OpenID support**
  - **Automatic renewal of the user credentials for long-running transfer requests**
- **ESGF catalog browsing and search capability within DML**

DML Technology

- **DML HTTP parallel streaming capability**
  - Concurrently download multiple files, where downloading each file by splitting It into multiple blocks and streaming through multiple HTTP connections established with an ESGF data server
    - Block size can be as small as 1 MB
  - Partial file downloading from each https stream to compose a whole file
    - After all the blocks of the source file are streamed down, final target file is recreated
  - Partial file downloading from multiple replica to compose a whole file
    - supported when data replicas are available on multiple ESGF data nodes, and replica info is known in the catalog
  - Transfer error recovery mechanism

DML screen samples (1)

- **User login window**
  - **Retrieving Myproxy credential**

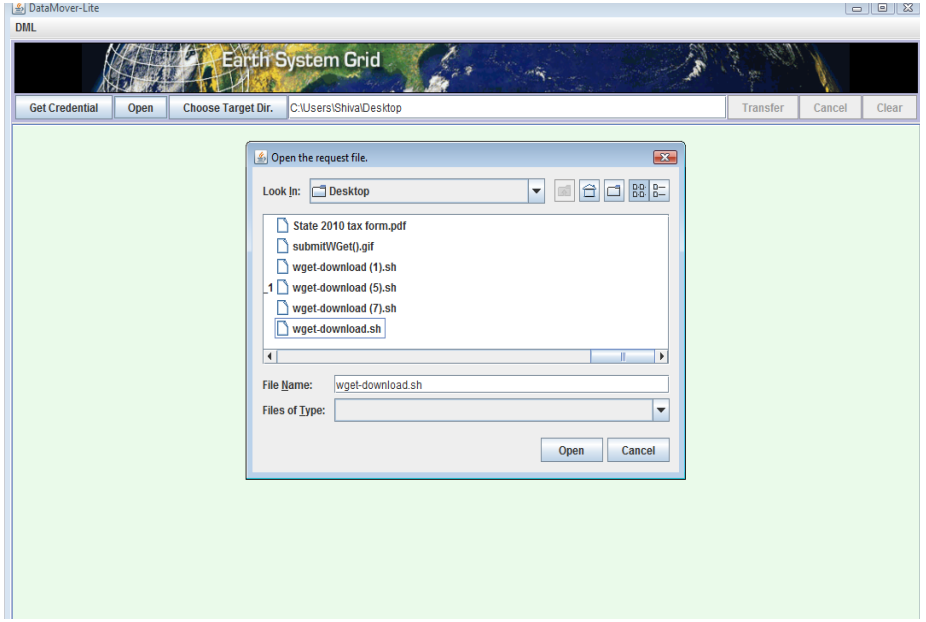


DML screen samples (2)

- Wget script generation from the registered ESG portal
- Wget script import to DML for downloading files

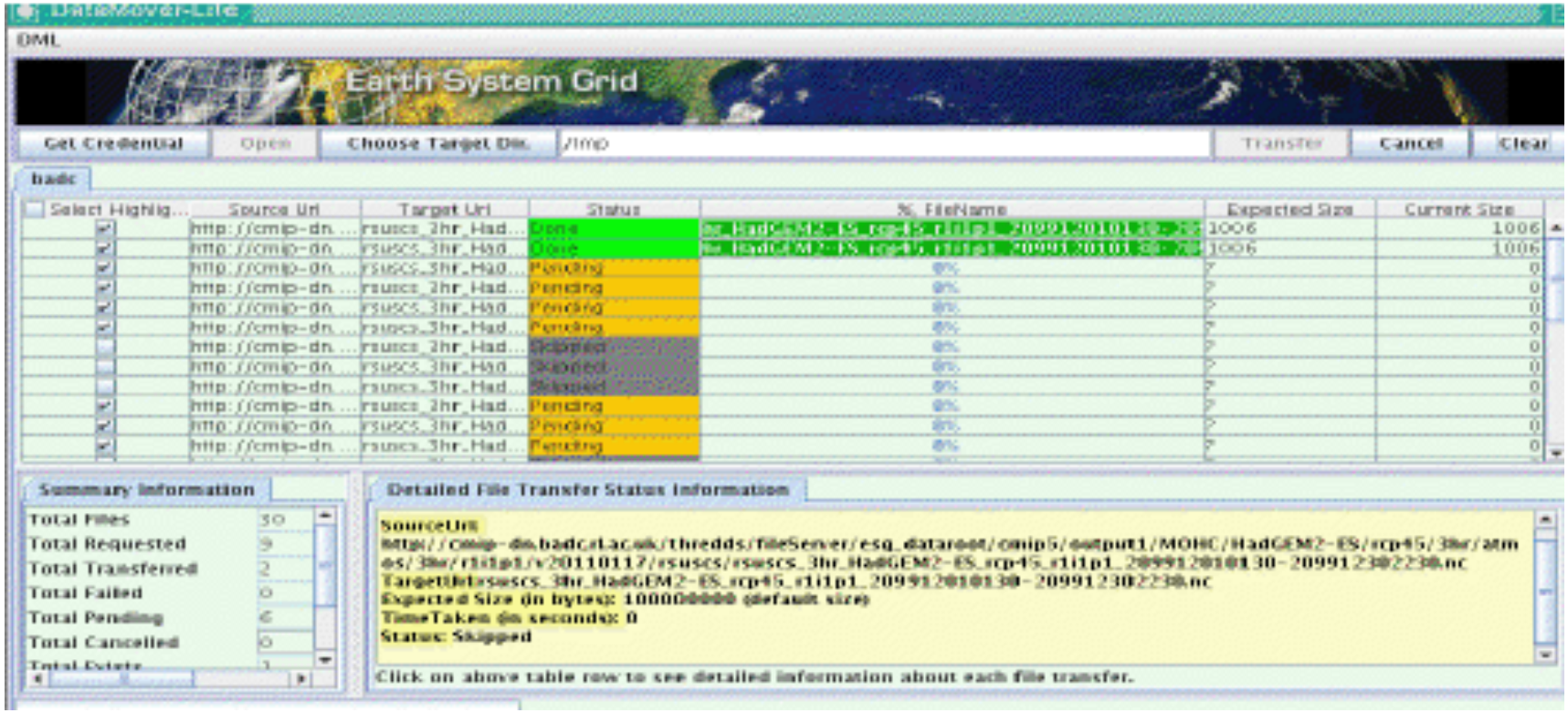
ESG-PCMDI Portal

DML



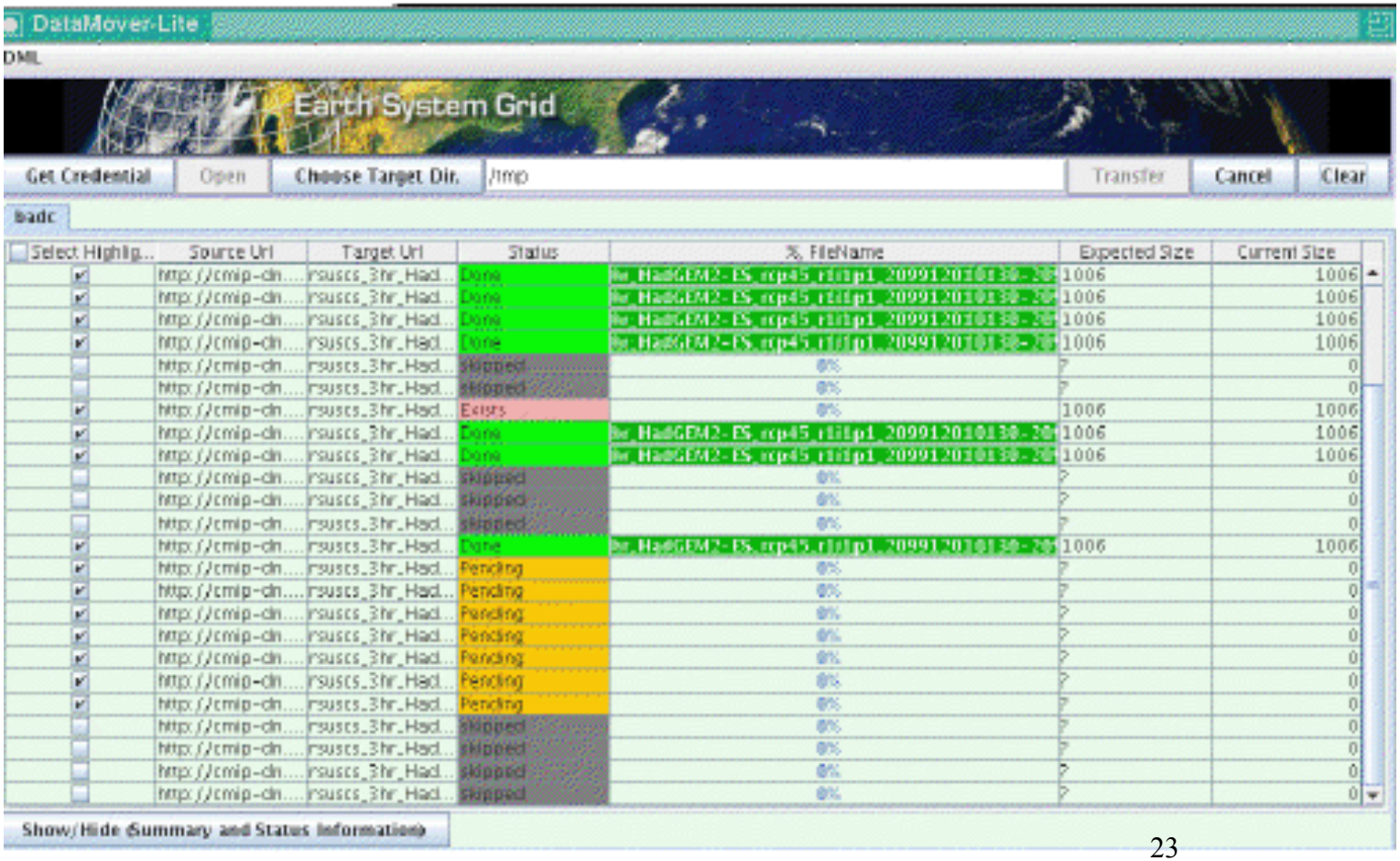
DML screen samples (3)

- HTTPS downloads from wget script file
- Active downloads with file information displayed



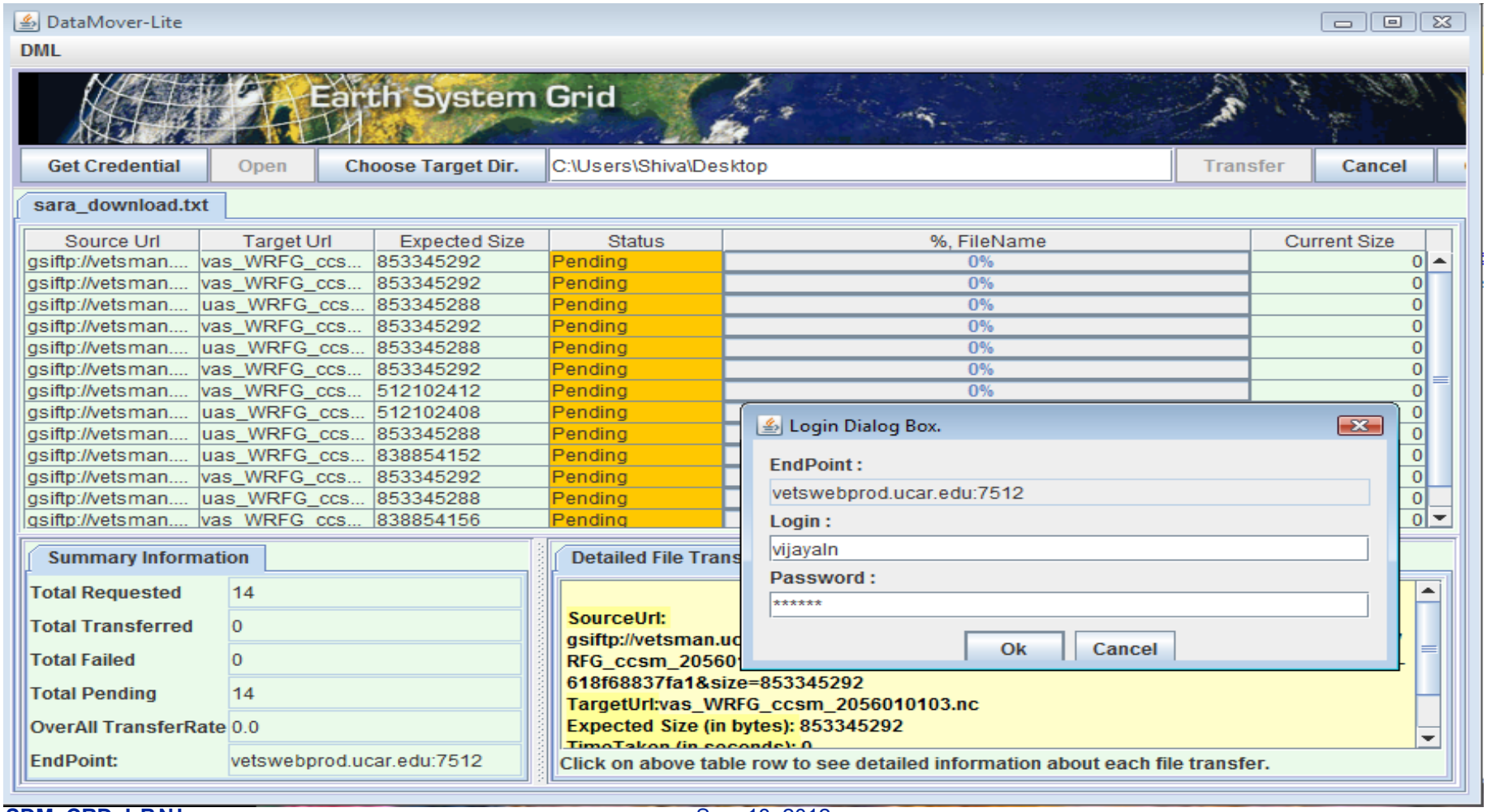
DML screen samples (4)

- **HTTPS downloads from wget script file**
  - Sample for downloading second subset of files.



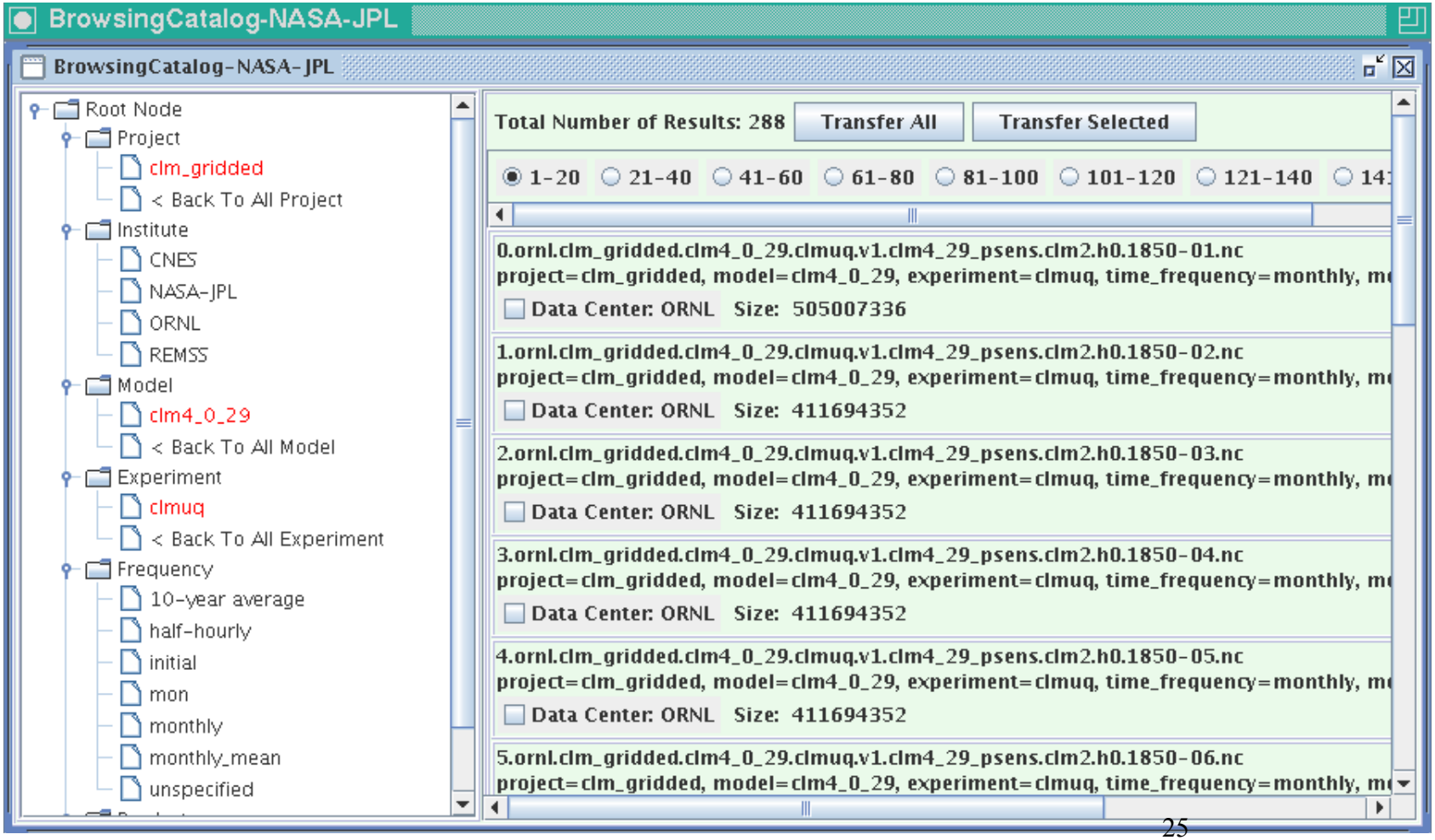
DML screen samples (4)

- **GridFTP downloads**

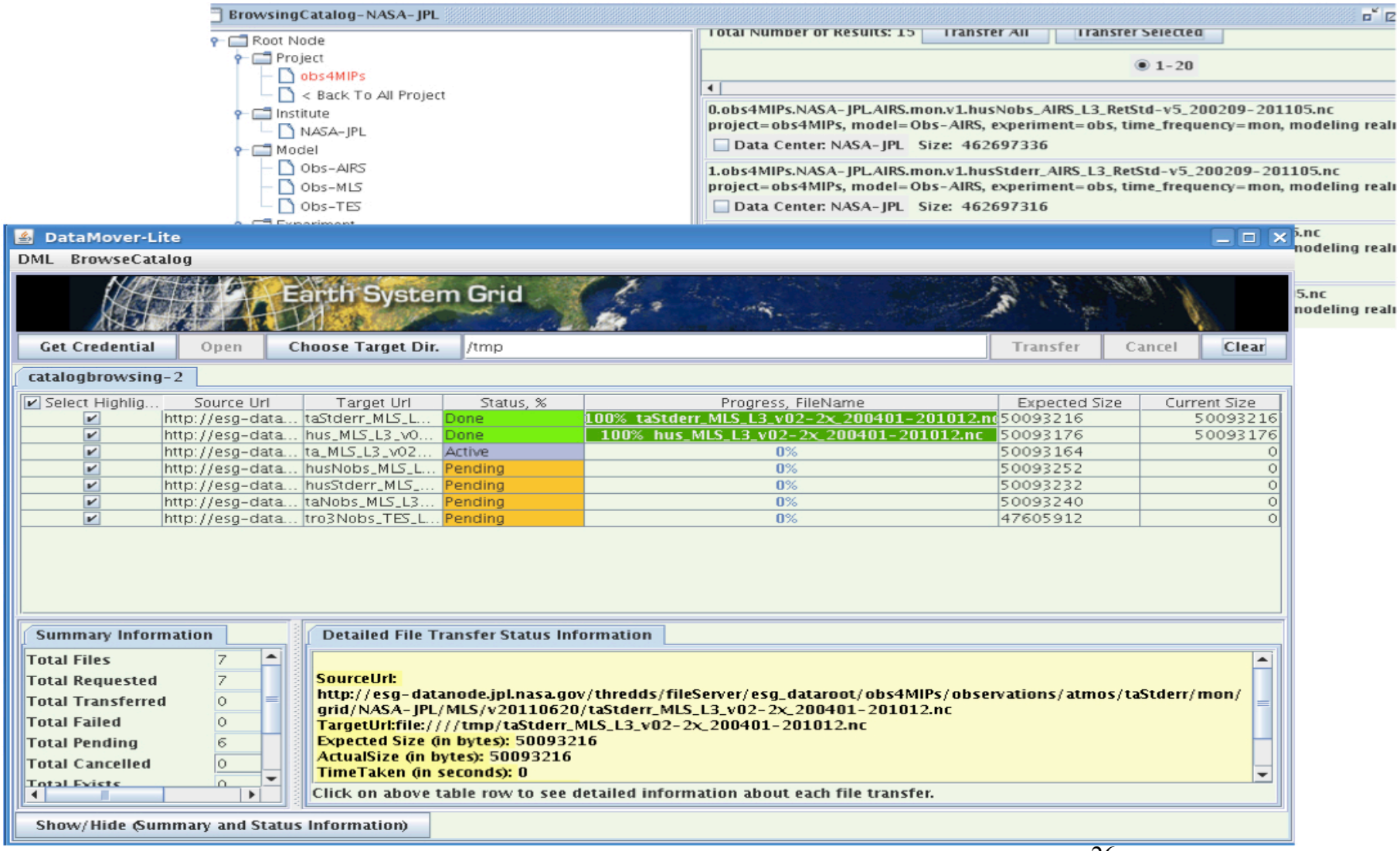


ESGF Catalog Browsing Integration

Enable user friendly search criteria for selecting files.



ESGF Data catalog browsing capability and Data downloading interface



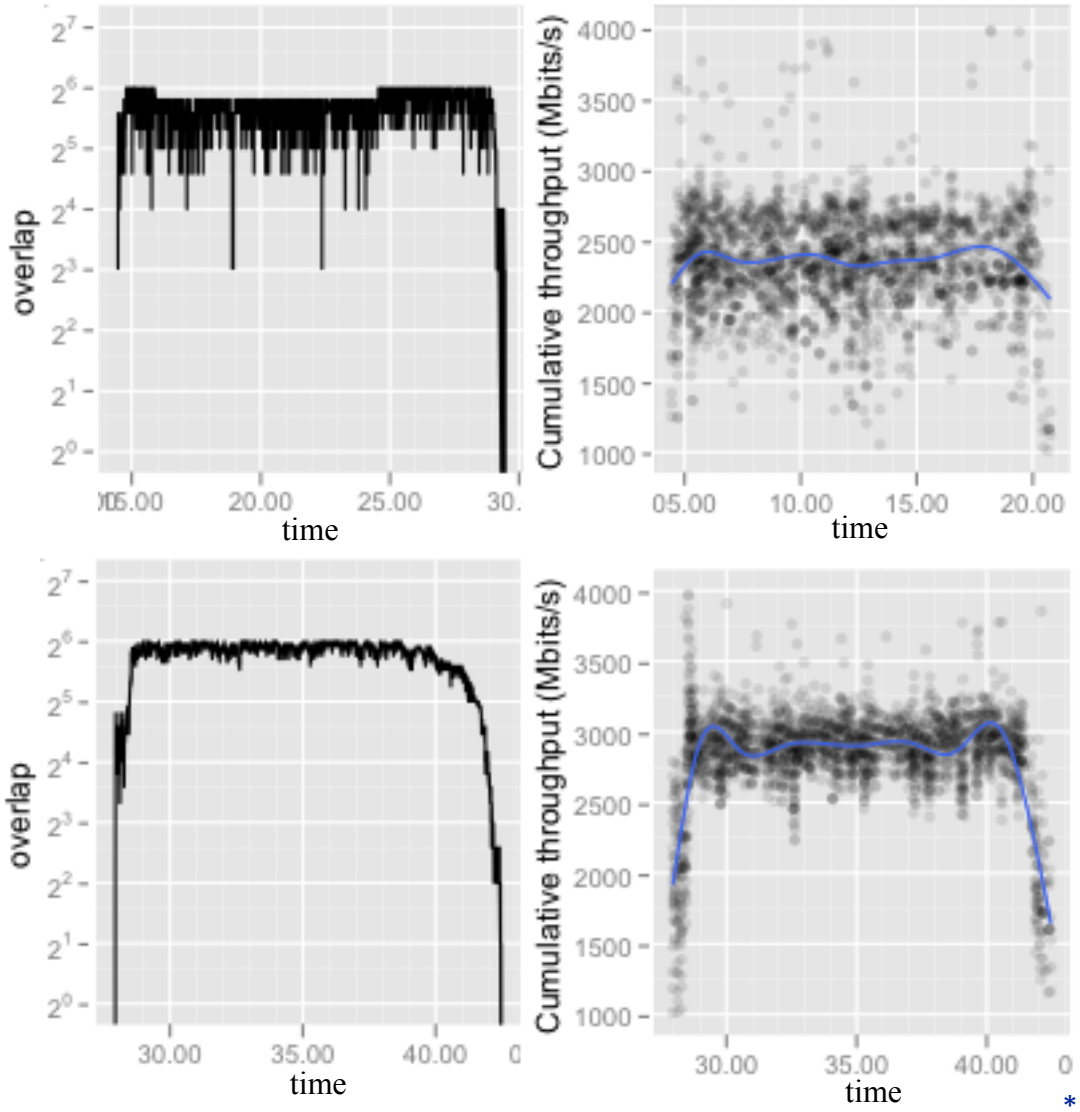
Bulk Data Mover (BDM)

- **Scalable bulk data transfer management tool**
  - **Designed for climate community (Earth System Grid) needs**
    - **Efficient and reliable transfer management from user's point of view**
    - **Simple to install and maintain as a novice user**
    - **Scalable to large in volume, and large in number of files**
    - **Efficient handling on extreme variance in file sizes**
    - **Scalable to future performance expectations**
      - **Network performance improvements – 100Gbps and beyond**
      - **Storage performance improvements – distributed, parallel, SSD, etc.**
      - **Multiple transfer protocol support**

BDM Technology

- **High performance using a variety of techniques**
  - **Multi-threaded concurrent transfer management**
    - **Contribute to more transfer throughput, including both network and storage (overlapping storage I/O with the network I/O)**
  - **Transfer queue management**
  - **Single control channel management for multiple data transfers**
  - **Load balancing on multiple transfer servers**
  - **GridFTP library supports data channel caching and pipelining**
- **Performance Adaptability (experimental)**
  - **Adaptable transfer management to the dynamic end-to-end bandwidth and system performance changes**
  - **Dynamic tuning: setting control parameters dynamically for throughput optimization**
    - **Does not require a complex model for parameter optimization**
    - **Does not depend on external profilers for active performance measurements**
    - **Adapts to changing environments**

Results of Managed Transfers

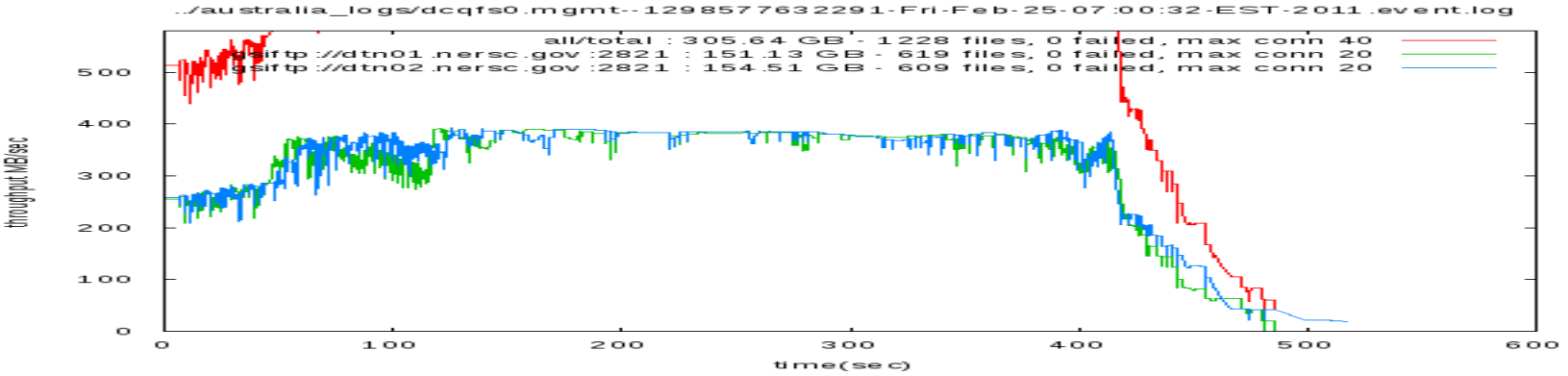


The number of concurrent transfers on the left column shows consistent over time in well-managed transfers shown at the bottom row, compared to the ill or non-managed data connections shown at the top row. It leads to the higher overall throughput performance on the lower-right column.

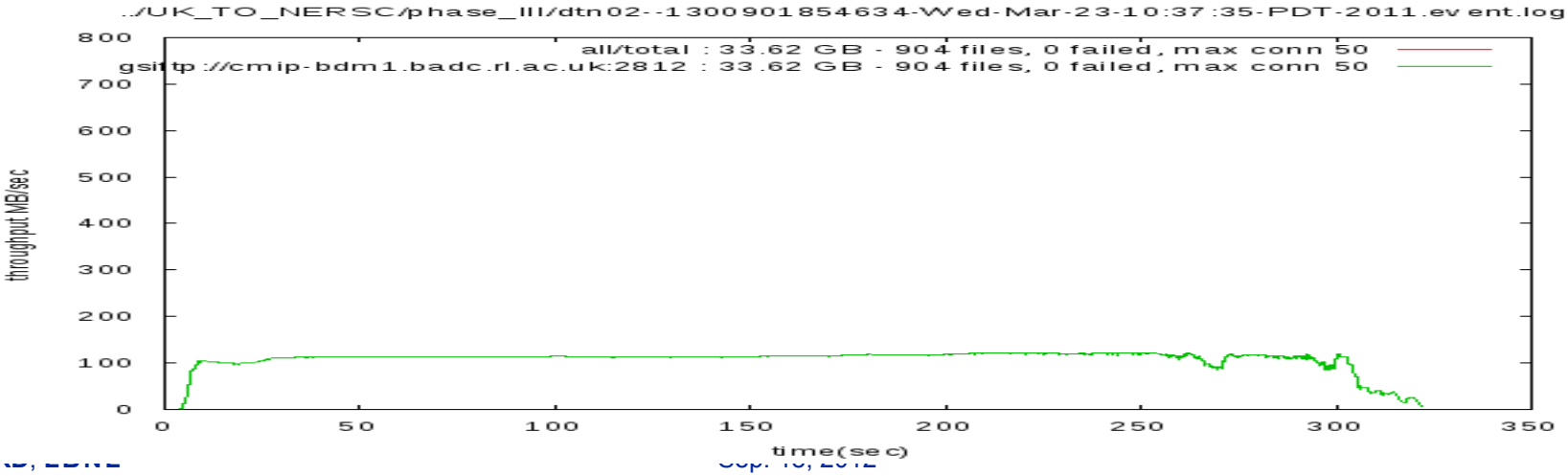
* Plots generated from NetLogger

Sample BDM runs (1)

- **BDM performance plot for data transfers from NERSC to ANU on 2/24/2011**
  - **~5.6 Gbps (700MB/sec) on average with ~6 Gbps at the peak**

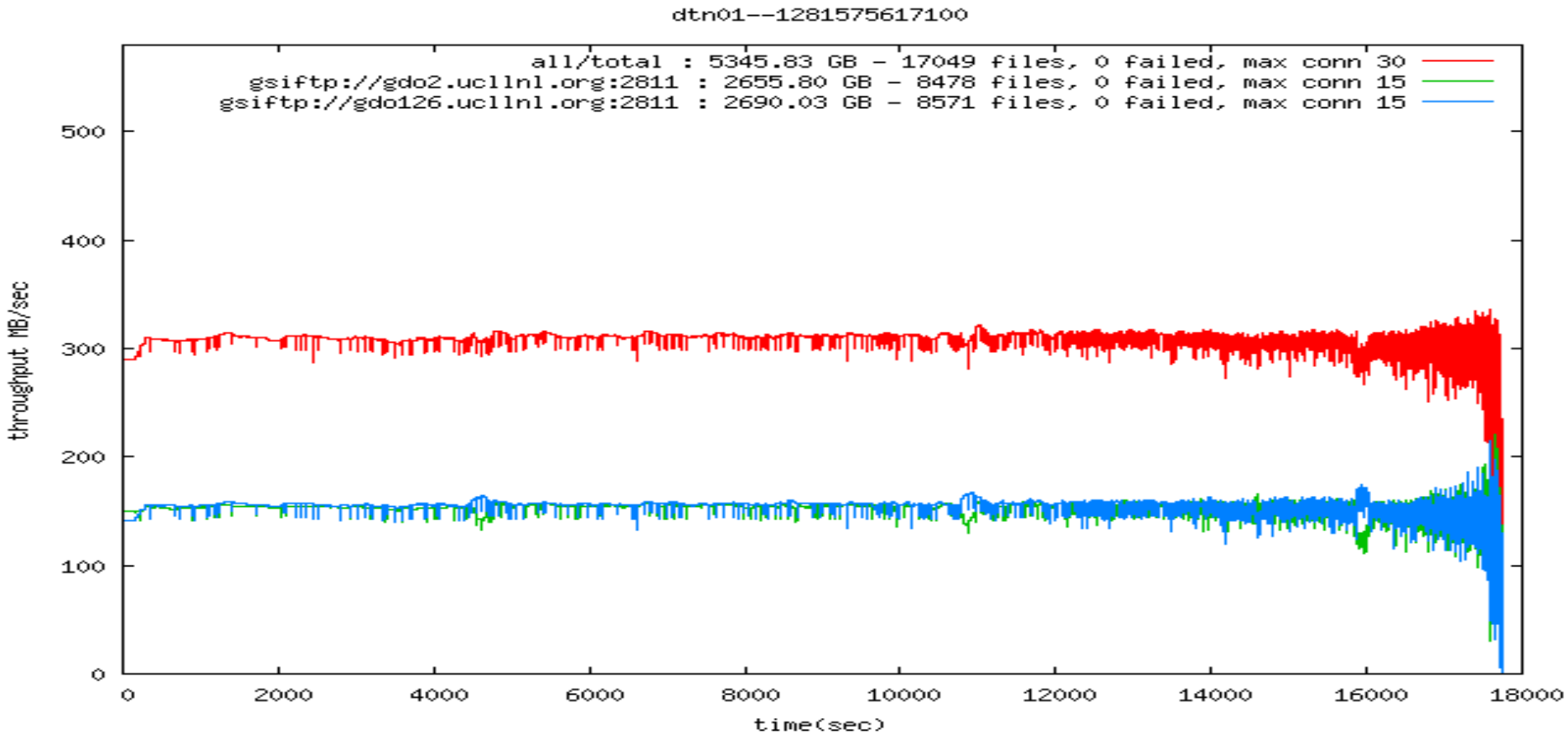


- **BDM performance plot for data transfers from BADC to NERSC on 2/24/2011**
  - **~0.9 Gbps (110 MB/sec) on average**



Sample BDM runs (2)

- **BDM performance plot for data transfers from LLNL to NERSC on Aug. 2010**
  - **~2.4 Gbps on average**



Summary

- **SDM contributions to Earth System Grid**
  - Berkeley Storage Manager (BeStMan)
  - DataMover-Lite
  - Bulk Data Movement and Climate data replications
  - LBNL/NERSC ESGF P2P node
- **SDM contribution to Open Science Grid: BeStMan Support**
  - VO support: US ATLAS, US CMS, STAR, LIGO, SBGRID, CERN EOS, etc
  - User support: SRM client usage with BeStMan
  - Middleware support: Data replication middleware, FTS, PhEDEx
  - VO-requested feature addition and maintenance
  - BeStMan server and client tools are integral part of OSG