

## 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



- 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.



- 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



- 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



- 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



- 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 3<sup>rd</sup> 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 unixbased 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)
- 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)



- Local Policy
  - Fair request processing
  - File replacement in disk
  - Garbage collection

## se 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, CRD, LBNL



## **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





- 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 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

Get Credential						
📅 Get Credential 🔤						
OpenID Login/Password						
OpenId login						
OpenID:						
Password:						
LoginName: UseDifferentName						
(If your login name is different than the openid, please click different name and provide loginname.)						
GetCredential Cancel Reset Credential Info. Delete Credential Renew proxy auto						



## **DML screen samples (2)**

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





## **DML screen samples (3)**

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

DateMover	-Life 👷	New York Street Street					900900 F
DML							
lle.	19174	Earth Syste	m Grid	Con the second	1.00	01	
Get Credential	Open	Choose Target Dis.	/1mp		Transfer	Cancel	Clear
bade							
Salact Highlig.	Spurce Uri http://cmip-dn. http://cmip-dn. http://cmip-dn. http://cmip-dn. http://cmip-dn. http://cmip-dn. http://cmip-dn. http://cmip-dn. http://cmip-dn. http://cmip-dn.	Target Uri rsuscs 3hr, Had rsuscs 3hr, Had	Status Jone Pending Pending Pending Pending Pending Perset Raphed Pending Pending Pending Pending	X. FisName           M. Haddard X. S. Rep (S. Difficil. 2009) 10101 10101 2010           M. Haddard X. S. Rep (S. Difficil. 2009) 10000 10101 20102           M. Haddard X. S. Rep (S. Difficil. 2009) 10000 10000 20102           M. Haddard X. S. Rep (S. Difficil. 2009) 10000 10000 20102           M. Haddard X. S. Rep (S. Difficil. 2009) 10000 10000 20102           M. Haddard X. S. Rep (S. Difficil. 2009) 10000 10000 20102           M. Haddard X. S. Rep (S. Difficil. 2009) 10000 10000 20102           M. Haddard X. S. Rep (S. Difficil. 2009) 10000 10000 10000000           M. Haddard X. S. Rep (S. Difficil. 2009) 10000000000000000000000000000000000	Experted Size 1006 7 7 7 7 7 7 7 7 7 7 7 7 7	Current	Size 1006 0 0 0 0 0 0 0 0 0 0 0 0 0
Summary Infor	mation	Detailed File Tr	ansfer Status In	formation			le!
Total Requested       9       8         Total Requested       9       8         Total Requested       9       8         Total Transferred       2       6         Total Failed       0       7         Total Pending       6       7         Total Cancelled       0       8         Total Cancelled       0       9         To							



#### HTTPS downloads from wget script file

• Sample for downloading second subset of files.

						10/24/03/09/09/03/04
		Earth Syste	m Grid	Con and the		Cir V
Set Credentia	Open	Choose Target Dir	/1mp		Transfer	Cancel Cle
idc						
Select Highlig.		Target Uri	Status	3, FileName	Expected Size	Current Size
×	http://cmip-dn	rsuscs_3hr_Had	Done	8r_HadGEM2-ES_rcp45_r1r1p1_209912010130-20	1006	100
1	http://cmip-dn	rsuscs_3hr_Had	Done	8r_HadGEM2-ES_rcp45_rtitp1_209912010130-20	1006	100
M	http://cmip-dn	rsuscs_3hr_Had	Done	8c_HadGEM2+ES_rcp45_r1/tp1_209912010138-26	1006	100
<b>K</b>	http://cmip-ch	rsuscs.3hr.Had	Done	8c;HadGEM2-FS; r(p45;r1)1p1;209912010138-28	1006	100
	http://cmip-dn	rsuscs.3hr.Had	stopped	6%	7	
	http://emip-ch	rsuscs_3hr_Had	stopped	8% BX	7	
E.	http://cmip-ch	rsuscs_3hr_Had	Exists	<b>8%</b>	1005	100
¥	http://cmip-dn	rsuscs_3hr_Had	Done	Br_HadGEM2+ES_rcp45_r1i1p1_209912010130-20	1006	100
×	http://cmip-dn	insuscs_3hr_Had	Done	Int_HadGEM2+ES_rcp45_r1i1p1_209912010130-20	1006	100
	http://cmip-dn	rsuscs_3hr_Had	skipped	0%	2	
	http://cmip-dn	rsuscs_3hr_Had	skipped	6%	2	
	http://cmip-ch	rsuscs.3hr.Had	stipped	6%	7	
E.	http://cmip-ch	rsusts.3hr.Had	Dane	br. HadGEM2+ES. rcp45.r1/1p1.209912010130-201	1006	100
<b>F</b>	http://cmip-ch	rsusts.3hr.Had	Pencing	0%	7	
E.	http://cmip-dn	rsuscs_3hr_Had	Pending	0%	7	
V	http://cmip-dn	rsuscs_3hr_Had	Pencing	8%	2	
×	http://cmip-dn	rsuscs_3hr_Had	Pencang	e%.	7	
×	http://cmip-dn	rsuscs_3hr_Had	Pencing	e%.	2	
E.	http://cmip-dn	rsuscs.3hr.Had	Pencing	0%	7	
<b>F</b>	http://cmip-dn	rsuscs.3hr.Had	Pending	<b>8</b> %	7	
	http://cmip-ch	rsusts.3hr.Had	shipped	80 Ø%	7	
	http://cmip-ch	rsusts.3hr.Had	skipped	0%	7	
	http://cmip-dn	rsuscs_3hr_Had	skipped	e%.	7	
	http://cmip-dn	rsuscs_3hr_Had	skipped	8%	8	



## **DML screen samples (4)**

#### GridFTP downloads

🖆 DataMover-Lite						
DML						
	Ear	th System	Grid		and the second	
Get Credential	Open C	hoose Target Dir.	C:\Users\Shiva\Des	top	Transfer Cancel	
sara_download.txt			P			
Source Url	Target Url	Expected Size	Status	%, FileName	Current Size	
gsiftp://vetsman vas	s_WRFG_ccs	853345292	Pending	0%	0 🔺	
gsiftp://vetsman vas	s_WRFG_ccs	853345292	Pending	0%	0	
gsiftp://vetsman ua	s_WRFG_ccs	. 853345288	Pending	0%	0	
gsiftp://vetsman vas	s_WRFG_ccs	853345292	Pending	0%	0	
gsiftp://vetsman ua	s_WRFG_ccs	. 853345288	Pending	0%	0	
gsiftp://vetsman vas	s_WRFG_ccs	853345292	Pending	0%	0	
gsiftp://vetsman vas	s_WRFG_ccs	512102412	Pending	0%	0	
gsiftp://vetsman ua	s_WRFG_ccs	. 512102408	Pending	🗳 Login Dialog Box.		
gsiftp://vetsman ua	s_WRFG_ccs	. 853345288	Pending			
gsiftp://vetsman ua	S_WRFG_CCS.	838854152	Pending	EndPoint :	0	
gsittp://vetsman vas	S_WRFG_CCS	853345292	Pending	vetswebprod ucar edu:7512	0	
gsittp://vetsman ua	S_WRFG_CCS.	020054456	Pending			
gsiftp://vetsman vas WRFG ccs 838854156 Pending			Pending	Login :		
Summary Information Detailed			Detailed File Trans	vijayaln		
Total Requested	14			Password :		
Total Transferred	0		SourceUrl:			
Total Failed	0		RFG_ccsm_20560	Ok Cancel		
Total Pending	14		618f68837fa1&size=853345292			
OverAll TransferRate	0.0		Expected Size (in bytes): 853345292			
EndPoint:	vetswebprod.u	car.edu:7512	Click on above tabl	e row to see detailed information about each file tr	ransfer.	

SDM, CRD, LBNL



#### Enable user friendly search criteria for selecting files.

BrowsingCatalog-NASA-JPL	
BrowsingCatalog-NASA-JPL	
P- □ Root Node	Total Number of Results: 288 Transfer All Transfer Selected
Clm_gridded Sack To All Project	● 1-20 ○ 21-40 ○ 41-60 ○ 61-80 ○ 81-100 ○ 101-120 ○ 121-140 ○ 14:     ▲
P 🚍 Institute │ — 🖺 CNES	<pre>0.ornl.clm_gridded.clm4_0_29.clmuq.v1.clm4_29_psens.clm2.h0.1850-01.nc</pre>
	project=clm_gridded, model=clm4_0_29, experiment=clmuq, time_frequency=monthly, m Data Center: ORNL Size: 505007336
	1.ornl.clm_gridded.clm4_0_29.clmuq.v1.clm4_29_psens.clm2.h0.1850-02.nc project=clm_gridded, model=clm4_0_29, experiment=clmug, time_frequency=monthly, m
- Clm4_0_29	Data Center: ORNL Size: 411694352
└─ └`) < Back To All Model	2.ornl.clm_gridded.clm4_0_29.clmuq.v1.clm4_29_psens.clm2.h0.1850-03.nc project=clm_gridded, model=clm4_0_29, experiment=clmuq, time_frequency=monthly, m
Clmuq Sack To All Experiment	Data Center: ORNL Size: 411694352
Frequency     10-year average	3.ornl.clm_gridded.clm4_0_29.clmuq.v1.clm4_29_psens.clm2.h0.1850-04.nc project=clm_gridded, model=clm4_0_29, experiment=clmuq, time_frequency=monthly, m
- half-hourly	Data Center: ORNL Size: 411694352
– 🗋 initial – 🗋 mon	4.orni.cim_gridded.cim4_0_29.cimuq.v1.cim4_29_psens.cim2.n0.1850-05.nc project=clm_gridded, model=clm4_0_29, experiment=clmuq, time_frequency=monthly, mo
- D monthly -	Data Center: ORNL Size: 411694352
monthly_mean     msnecified	5.ornl.clm_gridded.clm4_0_29.clmuq.v1.clm4_29_psens.clm2.h0.1850-06.nc project=clm_gridded, model=clm4_0_29, experiment=clmuq, time_frequency=monthly. m(



#### ESGF Data catalog browsing capability and Data downloading interface

BrowsingCatalog-NASA-JPL			d <sup>e</sup>				
P Root Node Project Obs4MIPs	Octal Number of Results: 15 Transfer All Transfer Selected      Octal Number of Results: 15 Transfer All Transfer All Transfer Selected      Octal Number of Results: 15 Transfer All Transfer A						
C = C = C = C = C = C = C = C = C = C =							
Obs-AIRS     Obs-MLS     Obs-TES	1.obs4MIPs.NASA-JPL_AIRS.mon.v1.husStderr_AIRS_L3_RetStd-v5_200209-201105.nc project-obs4MIPs, model-Obs-AIRS, experiment-obs, time_frequency-mon, modeling re Data Center: NASA-JPL Size: 462697316						
🛃 DataMover-Lite			nodeling rea				
DML BrowseCatalog							
Earth System Grid			5.nc nodeling rea				
Get Credential Open Choose Target Dir. /tmp		Transfer Ca	incel Clear				
catalogbrowsing-2							
Select Highlig Source Url Target Url Status, %  Mttp://esg-data taStderr_MLS_L Done 100% taStde  Mttp://esg-data hus MLS_L3_v0 Done 100% hus	Progress, FileName rr_MLS_L3_v02-2x_200401-201012.nd MLS_L3_v02-2x_200401-201012.nc	Expected Size 50093216 50093176	Current Size 50093216 50093176				
http://esg-data ta_MLS_L3_v02 Active	0%	50093164	0				
http://esg-datahusNobs_MLS_L Pending	0%	50093252	0				
http://esg-data hussiderr_MCS Pending	0%	50093232	0				
http://esg-data tro3Nobs_TES_L Pending	0%	47605912	0				
Total Files       7       7         Total Requested       7       7         Total Requested       7       7         Total Transferred       0       9         Total Filed       0       9         Total Failed       0       1         Total Pending       6       6         Total Cancelled       0       1							
Total Exists Click on above table row to see detailed inform	nation about each file transfer.						
Show/Hide Summary and Status Information)							
		26					



- 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



### • 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 lowerright column.

<sup>\*</sup> 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

../UK\_TO\_NERSC/phase\_III/dtn02--1300901854634-Wed-Mar-23-10:37:35-PDT-2011.ev ent.log





## Sample BDM runs (2)

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

dtn01--1281575617100





- 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