OSG STORAGE AND DATA MOVEMENT

Tanya Levshina

Talk Outline

- Data movement methods and limitation
- Open Science Grid (OSG) Storage
- Storage Resource Manager (SRM) and SRM Clients
- Storage Discovery Tools
- How to debug problems?
- Summary

Acknowledgments

This presentation is compiled from multiple sources:

- Brian Bockelman's lecture at the OSG Summer Grid School
- OSG Storage twiki (pages contributed by Ted Hesselroth, Doug Strain, Neha Sharma and others)
- Conversations with the experts (Alex Sim, Andrew Hanushevsky, Gabriele Garzoglio, Parag Mhashilkar, Derek Weitzel and others)

Grid Jobs and Data (I)

- Computation your are planning to do is often data driven and could be data intensive.
- Your job may require input files as well as output files.
- There are various ways to make your data available to your job:
  - bring your data with your job
  - bring your job to your data
- You will need to decide which approach is the most efficient in your case.

Grid Jobs and Data (II)

- Bring your data with your job:
  - rely on Condor-G for data transfer
  - use GlideinWMS
  - use SQUID cache on a site
- Bring your job to your data:
  - pre-stage input data at the shared POSIX-mounted storage available at the site
  - pre-stage input data at a Storage Element

Data Movement Methods

- Condor-G and GlideinWMS are covered in other presentations:

  http://indico.fnal.gov/contributionDisplay.py?contribId=18&sessionId=26&confId=3586

  http://indico.fnal.gov/contributionDisplay.py?contribId=27&sessionId=33&confId=3586

- In this presentation:
  - Squid
  - Shared storage area attached to Compute Element (Classic Storage Element)
  - OSG Storage Element

Squid

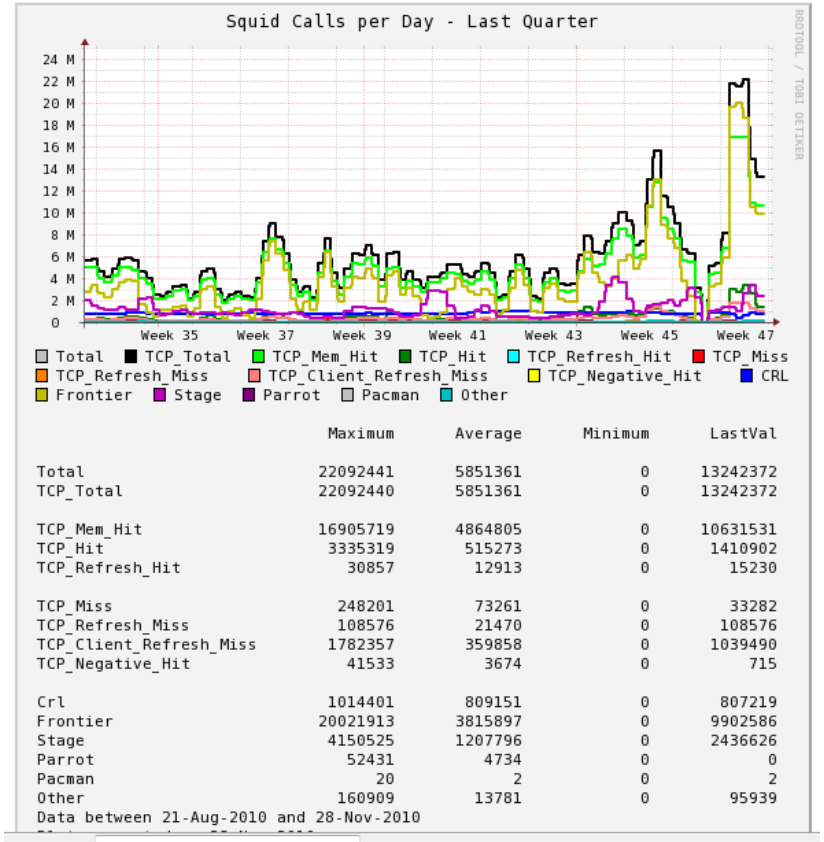
- Squid (http://wiki.squid-cache.org/FrontPage) is a web caching service:
  - downloads requests from http servers
  - improves response times by caching and reusing frequently-requested web pages
- Installed on several OSG Sites
- Mostly used on the OSG:
  - for CRL downloads
  - download common configuration files used by VO (CMS)
  - software download (CDF)

Squid Availability and Configuration on the OSG Sites

Environment variables are set on the site and could be access by a Grid job on the worker node:

- OSG_SQUID_LOCATION - The hostname (and optionally port) of the machine that is providing squid proxy services for the site. Set to "UNAVAILABLE "if squid is not provided.
- OSG_SQUID_CACHE_SIZE - controls the size the size in MB of the squid disk cache
- OSG_SQUID_POLICY - the cache replacement policy
- Allowed file size controls the size the largest HTTP message body that will be sent to a cache client for one request. Is not defined as OSG env. variable but set in Squid configuration (256 MB)

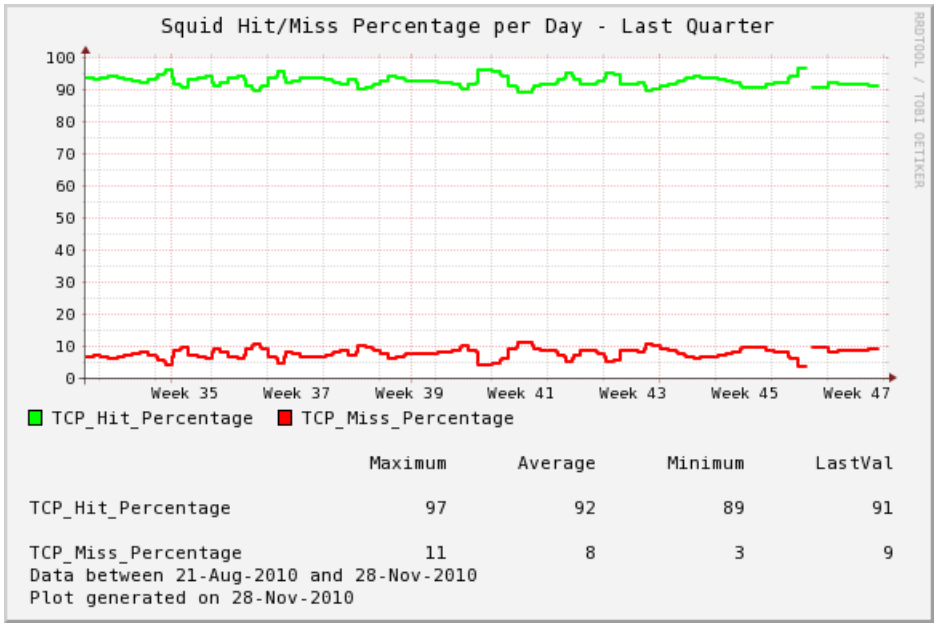
Squid Activities Monitoring



	Maximum	Average	Minimum	LastVal
Total	22092441	5851361	0	13242372
TCP_Total	22092440	5851361	0	13242372
TCP_Mem_Hit	16905719	4864805	0	10631531
TCP_Hit	3335319	515273	0	1410902
TCP_Refresh_Hit	30857	12913	0	15230
TCP_Miss	248201	73261	0	33282
TCP_Refresh_Miss	108576	21470	0	108576
TCP_Client_Refresh_Miss	1782357	359858	0	1039490
TCP_Negative_Hit	41533	3674	0	715
Crl	1014401	809151	0	807219
Frontier	20021913	3815897	0	9902586
Stage	4150525	1207796	0	2436626
Parrot	52431	4734	0	0
Pacman	20	2	0	2
Other	160909	13781	0	95939

Data between 21-Aug-2010 and 28-Nov-2010

From FermiGrid Monitoring page (http://fermigrid.fnal.gov/)



	Maximum	Average	Minimum	LastVal
TCP_Hit_Percentage	97	92	89	91
TCP_Miss_Percentage	11	8	3	9

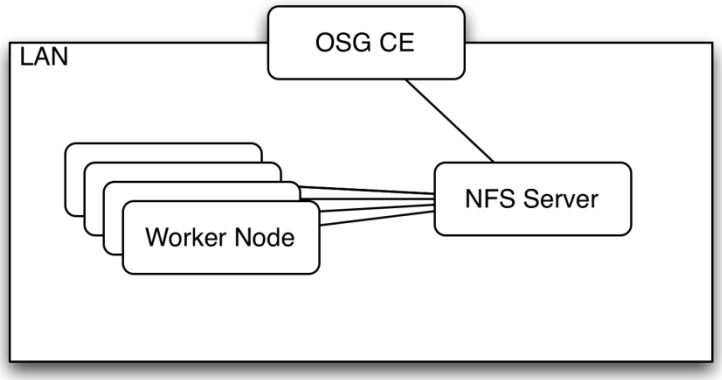
Data between 21-Aug-2010 and 28-Nov-2010
Plot generated on 28-Nov-2010

Potential Problems with Squid

- Currently Squid is running on a few sites
- Max files size is set to 256 MB
- The space management policy per VO is unknown
- In general case there is no security (additional work is needed to do checksum for software and data)

OSG Storage on CE

- OSG sites provide shared storage area:
  - POSIX-mounted storage (typically NFS)
  - is mounted and writable on the CE head node
  - readable and sometimes writable from WN
  - There are exceptions: UCSD doesn't provide access to OSG_DATA from the WN



Slide from B. Bockelman's talk at OSG Summer School

Potential Problems with classic SE

- Not scalable: heavy load on CE head node
- Could be too much for NFS server so ,probably, site will need high-performance filesystems (e.g Fermilab uses BlueArc)
- Most sites have quotas per VO (e.g Fermilab has limited OSG_DATA to 400 GB)
- Difficult to manage space.

Storage Element on OSG

- A SE is a cluster of nodes where data is stored and accessed: physical file systems, disk caches, hierarchical mass storage systems.
- Most sites have at least one SE.
- SE software manages storage and enforces authorization policies.
- Scalability and capacity of a SE significantly varies from site to site.
- A user interacts with a SE via a get or put of the file.
- A SE doesn't necessarily provide POSIX access.

Data Movement Methods Comparison

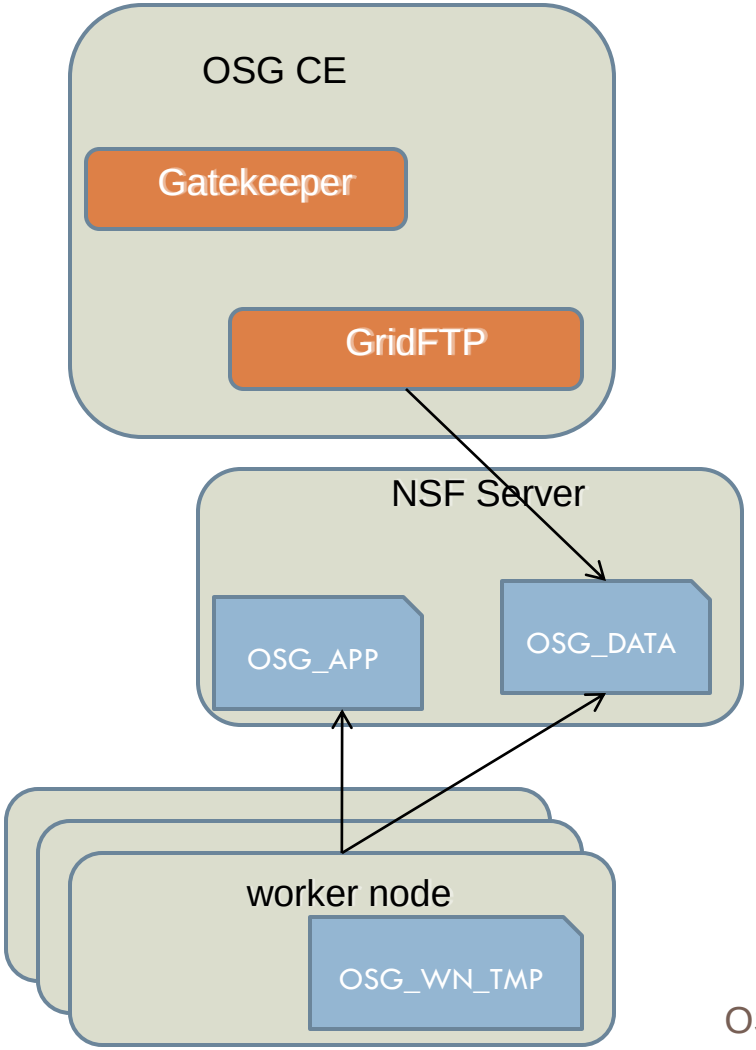
Type	Data flow	Limits	Potential Problems
Condor-G	Input and output files are transferred from the submission node to a gatekeeper node and then to the worker nodes.	File size <10 MB Small number of files	May produce heavy load on CE
classic SE (Shared storage attached to CE)	First, data is pre-staged into a shared area on a head node, then accessed from the worker nodes. In the majority of cases the access is POSIX compliant.	Limited by size of the shared area allocated for your VO	May produce heavy load on shared area server (NFS?), unknown policy of space management. Not all sites have OSG_DATA available from the worker nodes.
SQUID (a web caching service)	Execute wget command on the worker nodes. File may be pulled from the web or be in the squid cache already.	Works for input file only, limited file size, number of files, make sense only for files that are used by more than one job. No POSIX access. File size < 256 MB (at least at Fermi)	Not all the sites are using SQUID, no security.
GlideinWMS	Input and output files are transferred directly from the submission node to a worker node.	Network bandwidth between your host and a worker node. Sends files to WN all at once, possibly incurring in local disk space limitations.	Need to have related VO infrastructure
SE	Pre-stage data into SE, upload output data into SE.	No guarantee to be fully POSIX compliant.	For better performance, SE should be located on the same sites as CE where jobs are running

Limitations

- You may encounter a lot of problems and limit your choice of available sites if you are planning to:
  - transfer a large file > 10GB
  - have a large number of input/output files
  - expect POSIX access to data on a site

Classic SE Details

Classic Storage Element



- $OSG_APP - shared area, used for application installation, read access from a worker node.
- $OSG_DATA - shared area, for data that has a lifetime > job lifetime. Can be read-only on a worker node. Could be set to UNAVAILABLE.
- $OSG_WN_TMP - temporary storage area, local to the worker node and specific to a single job. Allocated size is less then 10 GB.

Before We Can Proceed…

- Let's assume that you already:
  - are a member of a particular VO
  - have OSG client installed
- Obtain voms proxy

  voms-proxy-init –voms osgedu
- Search Information services (myOSG, BDII, ReSS) or use discovery tools to find :
  - the site you want to use
  - GridFTP server, port and end point

How to Access a Classic SE?

- Let's assume the site you want to use: GridUNESP
- You have find out that the CE gatekeeper is running on: ce.grid.unesp.br:2811
- The location of OSG_DATA is /osg/data/ and you should be able to use osgedu directory
- Copy file using globus-url-copy from your local file (e.g. /tmp/my_test) to GridUNESP $OSG_DATA area:

  globus-url-copy file:///tmp/my_test
  gsiftp://ce.grid.unesp.br:2811/osg/data/osgedu/my_test

- Submit Condor-G job and access $OSG_DATA from the worker node

OSG SE Details

Storage Element Components

- A Storage Element (SE) is installed separately from Compute Element
- A typical SE has the following components:
  - Distributed File System
    - NFS, GPFS, PVFS, Lustre (POSIX access)
    - HDFS, xrootd (POSIX-lite with fuse)
    - dCache
  - GridFTP server(s)
  - Namespace service
  - Storage Resource Manager endpoint

SRM Protocol

- Storage Resource Manager (SRM) is a protocol for Grid access to a SE
- The protocol itself is a collaboration between Berkeley Lab, Fermilab , Jeffersonlab, CERN, RAL and INFN.
- SRMs are middleware components that manage shared storage resources on the Grid
- SRM Functions include:
  - Space Management
  - Data Transfer
  - Directory and Permission
  - Status

SRM Glossary

- SURL - is the Site URL. It identifies the file inside a SE. The format is

  srm://<host>:<port>/[<web service path>?SFN=]<path>

- SFN - a site specific file name for a replica, eg:

srm://gw015k1.fnal.gov:8443/srm/v2/server\?SFN=/data/xrootdfs/public/fermilab/test_1

srm://gwdca04.fnal.gov:8443/srm/managerv2\?SFN=/pnfs/fnal.gov/data/fermilab/test_1

- TURL, or Transfer URL points to where the file is physically located. It returns by SRM in response to an SRM client request to copy the file.

gsiftp://gw015k1.fnal.gov//data/xrootdfs/public/fermilab/test_1

gsiftp://gw018k1.fnal.gov:5000//mnt/hadoop/fermilab/test_1

Storage Elements Zoo

OSG : Number of sites providing Storage Elements: 49

- dCache : 12
- BeStMan: 37
  - HDFS: 6
  - Xrootd: 3
  - Lustre : 3
  - GPFS: 2
  - RedDnet: 1
  - All other sites: Local disk, NFS

WLCG:

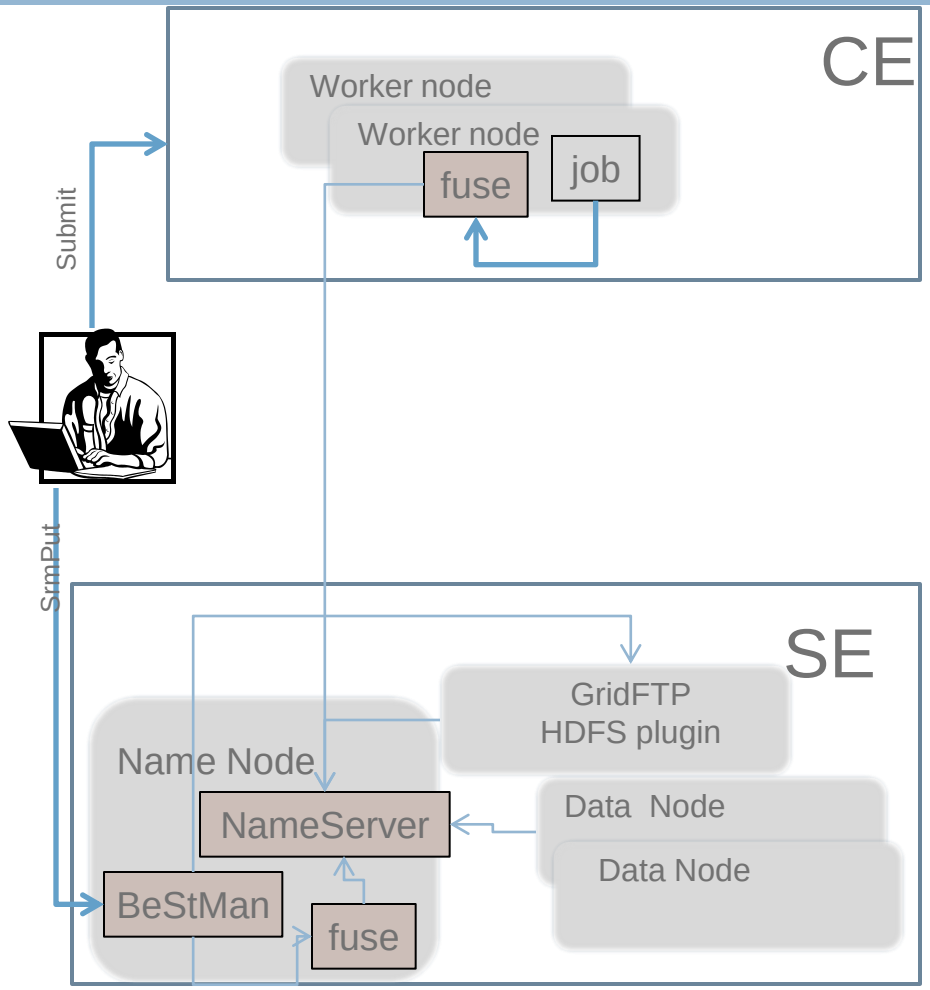
- Castor
- dCache
- DPM
- StoRM
- BeStMan

BeStMan

- Berkeley Storage Manager (BeStMan)
  - Developed by the Scientific Data Management Group at LBNL
  - Implements SRM v2.2
  - Provides load balancing front-end for transfer servers
  - Works on top of any disk-based POSIX-compliant file-systems
- BeStMan-Gateway supports subset of SRM v2.2 without internal queuing or space management

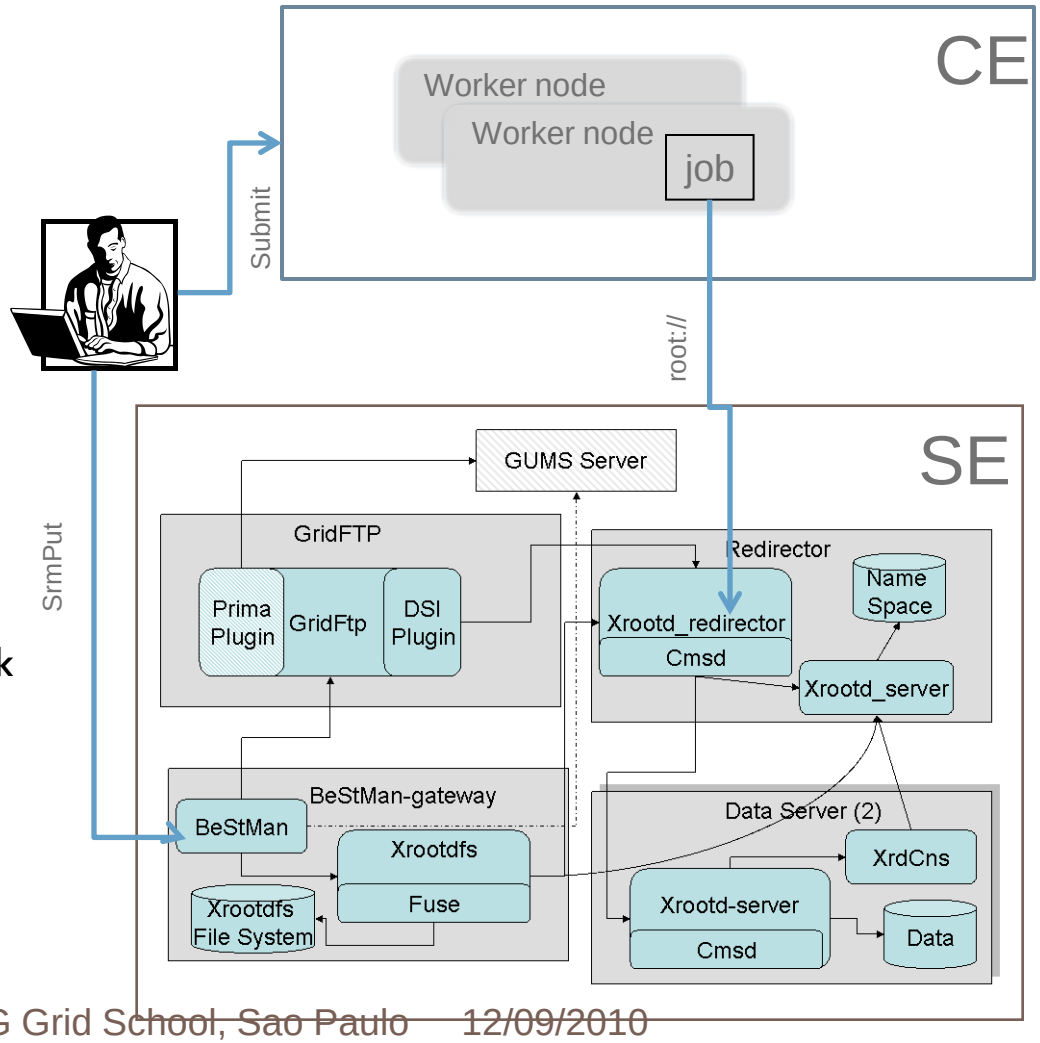
BeStMan-gateway/HDFS

- Hadoop Distributed File System is developed in the Apache project.
- Creates multiple replicas of data blocks
- Distributes them on data nodes throughout a cluster
- Consists of two major components:
  - Namenode: central metadata server.
  - Datanode: file servers for data
- Runs on commodity hardware
- Requires FUSE to hook with BeStMan, GridFTP –HDFS plugin
- Installed on multiple CMS Tier-2 sites



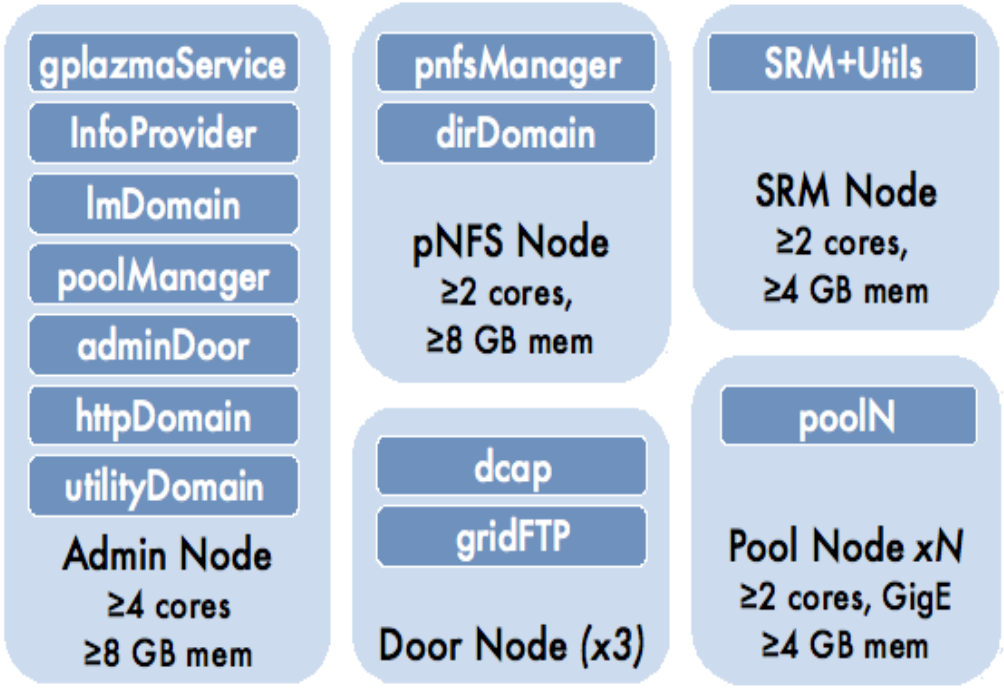
BeStMan-gateway/Xrootd

- Xrootd (developed at SLAC, contribution from CERN, others)
  - is designed to provide access
    - POSIX-like
    - via root framework (root://)
    - via native commands (xrdcp,…)
  - Allows cluster globalization
  - Allows unix-like user/group authorization as well as X509 authentication.
  - Requires FUSE, XrootdFS to hook with BeStMan, GridFTP DSI plugin
- Currently is used by many ATLAS and ALICE T2 sites , recommended for all Atlas T3



dCache

- dCache is a distributed storage solution developed at DESY, Fermilab and NGDF
- dCache supports requesting data from a tertiary storage system
- nfs-mountable namespace
- Multiple access protocols
- Replica Manager
- Role-based authorization
- Information Provider
- CMS Tier-1 and some ATLAS and CMS Tier-2 are using it.



Picture from Ted Hesselroth's
(from presentation: "Installing and Using SRM-dCache"

SRM Clients

SRM Clients

- Available from VDT (OSG-Client, wn-client)
- SRM-Fermi-Client commands
  - developed and maintained at Fermilab
  - access any Storage Element that complies with the SRM 1 or 2.2 specification
- SRM-LBNL-Client commands
  - developed at LBNL
  - access any SRM v2.2 based storage components
- LCG-utils is a suite of client tools for data movement written for the LHC Computing Grid.
  - based on the Grid File Access Library
  - access any SRM v2.2 based storage components
  - May use logical file names and require a connection to a BDII-based catalog for some commands

Argh…Which Client Should I Use?

- Sorry, I don't know
- Each has pros and cons, e.g:
  - Lcg-utils are most efficient and can deal with catalogs and bdii, but don't provide some useful commands like ping or rmdir.
  - Srm lbnl commands are very verbose but handle errors and exit codes better then fermi client or lcg-utils
- Your VO may already provide recommendations
- Try them all and select your favorite

Discovery

How Do I Find a SE?

- In order to use a SE you need to know the following:
  - SURL
  - Whether your VO is authorized to access the storage
- Information Services (BDII, ReSS, OSGMM)
  - Is this information reliable?
  - How do I query it?
- Discovery tools help to query BDII for storage related information

Discovery Tools

- These tools allow to search BDII and find relevant storage information for a particular VO that includes:
  - storage elements and corresponding site names
  - surl
  - available space
  - mount point to a SE on a WN
- Included in the OSG client and the wn-client VDT package

Example: how to find sites that support my VO.

Find all sites and SURLs that support your VO:

get_surl --vo Engage --show_site_name --show_storage_element_id

SITE NAME STORAGE ELEMENT ID SURL

UCSDT2 bsrm-1.t2.ucsd.edu srm://bsrm-1.t2.ucsd.edu:8443/srm/v2/server?SFN=/hadoop/engage/TESTFILE

UCR-HEP charm.ucr.edu srm://charm.ucr.edu:10443/srm/v2/server?SFN=/data/bottom/cms/TESTFILE

CIT_CMS_T2 cit-se.ultralight.org srm://cit-se.ultralight.org:8443/srm/v2/server?SFN=/mnt/hadoop/osg/engage/TESTFILE

GLOW cmssrm.hep.wisc.edu srm://cmssrm.hep.wisc.edu:8443/srm/managerv2?SFN=/pnfs/hep.wisc.edu/data5/engage/TESTFILE

BNL-ATLAS dcsrm.usatlas.bnl.gov srm://dcsrm.usatlas.bnl.gov:8443/srm/managerv2?SFN=/pnfs/usatlas.bnl.gov/osg/engage/TESTFILE

Firefly ff-se.unl.edu srm://ff-se.unl.edu:8443/srm/v2/server?SFN=/panfs/panasas/CMS/data/engage/TESTFILE

FNAL_FERMIGRID fndca1.fnal.gov srm://fndca1.fnal.gov:8443/srm/managerv2?SFN=/

GridUNESP_CENTRAL se.grid.unesp.br srm://se.grid.unesp.br:8443/srm/v2/server?SFN=/store/engage/TESTFILE

Example: how to find SE mount point on the worker node

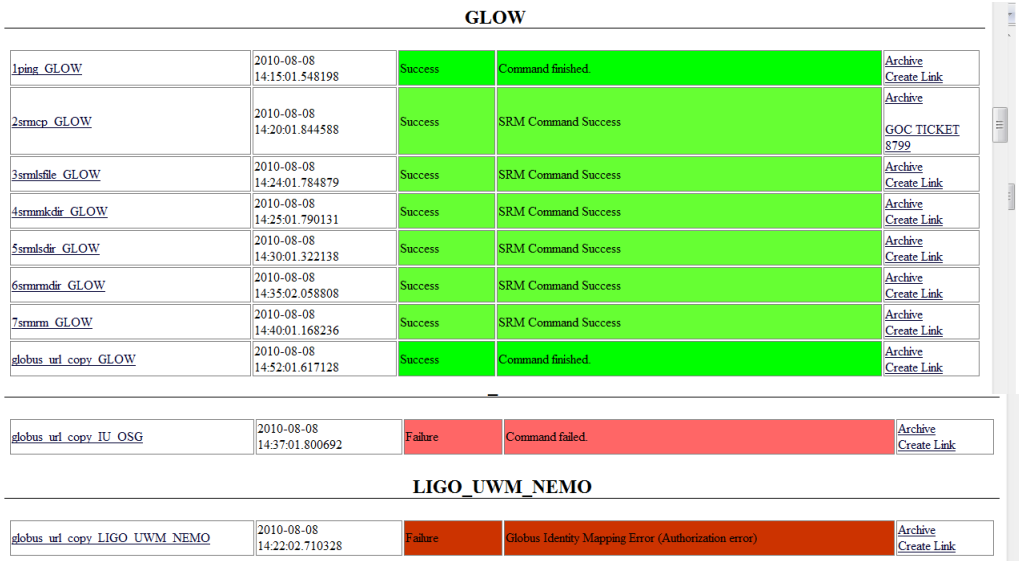
Find if SE has POSIX-like access to the data from the worker node:

get_mount_path --vo Engage --storage_element_id se.grid.unesp.br

COMPUTE ELEMENT ID	MOUNT POINT
ce.grid.unesp.br:2119/jobmanager-pbs-default	/store/engage
ce.grid.unesp.br:2119/jobmanager-pbs-long	/store/engage
ce.grid.unesp.br:2119/jobmanager-pbs-short	/store/engage

Pigeon Tools

- Discovery tools only help you to query the information from BDII.
- There is no guarantee that you will be able to access the SE or to transfer a file.
- Pigeon tools (created on top of Discovery tools) help a non-owner VO to debug site problems.
- Will be available as RSV probes for VOs



GLOW

1ping_GLOW	2010-08-08 14:15:01.548198	Success	Command finished.	Archive Create Link
2srmcp_GLOW	2010-08-08 14:20:01.844588	Success	SRM Command Success	Archive GOC TICKET 8799
3srmlsfile_GLOW	2010-08-08 14:24:01.784879	Success	SRM Command Success	Archive Create Link
4srmmkdir_GLOW	2010-08-08 14:25:01.790131	Success	SRM Command Success	Archive Create Link
5srmlsdir_GLOW	2010-08-08 14:30:01.322138	Success	SRM Command Success	Archive Create Link
6srmrmdir_GLOW	2010-08-08 14:35:02.058808	Success	SRM Command Success	Archive Create Link
7srmrm_GLOW	2010-08-08 14:40:01.168236	Success	SRM Command Success	Archive Create Link
globus_url_copy_GLOW	2010-08-08 14:52:01.617128	Success	Command finished.	Archive Create Link
globus_url_copy_IU_OSG	2010-08-08 14:37:01.800692	Failure	Command failed.	Archive Create Link

LIGO_UWM_NEMO

globus_url_copy_LIGO_UWM_NEMO	2010-08-08 14:22:02.710328	Failure	Globus Identity Mapping Error (Authorization error)	Archive Create Link

Problems …

- User's mistakes :
  - Don't have a right certificate proxy, proxy has expired
  - CA certificates, CRLs are not being updated on your local computer
  - Misspelled source or target names
- Information Service has wrong information:
  - A site doesn't really support my VO
  - SURL is wrong
  - Size of available storage area is wrong
- SE is misconfigured:
  - Proxy credentials are mapped to a user id that does not exist on the SE
  - Permissions are wrong on the end path directory
  - CAs ,CRLs, etc. misconfiguration at the SE

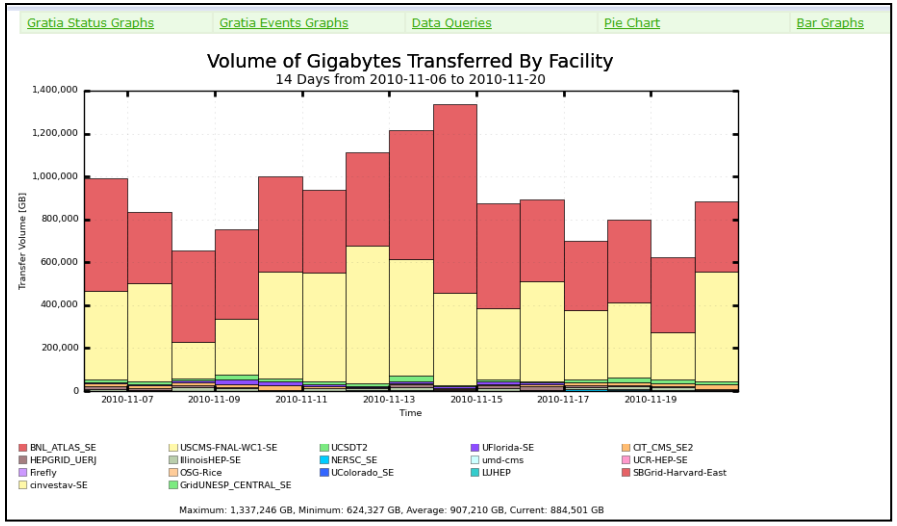
Things to Think About Before You Start

Open Science Grid

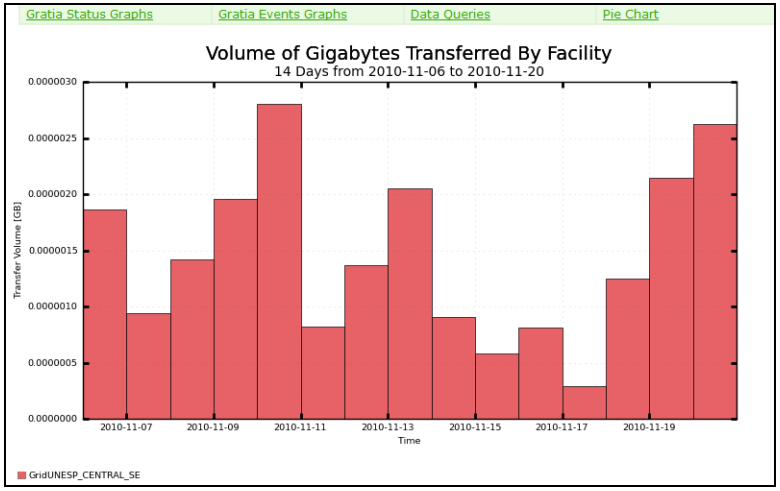
- How much total space do you need for your data?
- How long do you want to keep this data in storage?
- How much data is read by an individual job?
- How is the input data for an individual job subdivided into individual files?
- What kind of output data do you produce, and how much per job?
- How do you keep track of input and output data?
- Where do you want to ship your output data?

Gratia Transfer Probes

- Included in BeStMan, dCache VDT distribution
- Reports to OSG Gratia Accounting System
- Generates accounting information about file transfers, source, destination, size of the file and owner



http://t2.unl.edu/gratia/xml/facility_transfer_volume



http://t2.unl.edu/gratia/xml/facility_transfer_volume?facility=GridUNESP_CENTRAL_SE

Storage Documentation

Generic Storage documentation

https://twiki.grid.iu.edu/bin/view/ReleaseDocumentation/Storage

Storage for End User

https://twiki.grid.iu.edu/bin/view/ReleaseDocumentation/StorageEndUser

Discovery tools

https://twiki.grid.iu.edu/bin/view/ReleaseDocumentation/OSGStorageDiscoveryTool

Client Tools

- SRM clients
  - LCG Utils: https://twiki.grid.iu.edu/bin/view/ReleaseDocumentation/LCGUtils
  - LBNL: https://twiki.grid.iu.edu/bin/view/ReleaseDocumentation/LBNLSrmClient
  - Fermi: https://twiki.grid.iu.edu/bin/view/ReleaseDocumentation/FermiSrmClientCommands
- FTP clients
  - globus-url-copy http://www.globus.org/toolkit/docs/4.0/data/gridftp/rn01re01.html
  - Uber FTP https://twiki.grid.iu.edu/bin/view/Storage/StorageUberFTP

Summary

- It's very important to understand your workflow and choose the right data management solution
- Public Storage is not always easy to access, be patient while debugging the problems. Usually, after fixing initial problems the data could be successfully moved to/from SE (DZero is a good example).
- OSG Storage group is ready to help!
- Active mailing list:

  osg-storage@opensciencegrid.org
- GOC tickets:

  https://ticket.grid.iu.edu/goc/open/