dCache Storage Resource Manager

Timur Perelmutov

For the dCache team

Edinburgh, November 2007

Outline

- dCache Project Topology
- SRM at a glance
- dCache Specific Concepts
- Deployment
- dCache SRM Configuration
- gPlazma
- Monitoring dCache SRM
- SRM and dCache Deployment & Upgrade Q&A
- SRM Network Usage and Firewalls
- SRM inter-dCache interactions

Project Topology : The Team

Head of dCache.ORG

Patrick Fuhrmann

Core Team (Desy and Fermi)

Bjoern Boettscher

Alex Kulyavtsev

Iryna Koslova

Dmitri Litvintsev

David Melkumyan

Dirk Pleiter

Martin Radicke

Owen Synge

Vladimir Podstavkov

Head of Development FNAL :

Timur Perelmutov

Head of Development DESY :

Tigran Mkrtchyan

External Development Contribution

Gerd Behrmann, NDGF

Jonathan Schaeffer, IN2P3

Andrew Baranovski, CEDPS

Ted Hesselroth, OSG

Support and Help

Greig Cowan, gridPP

Stijn De Weirdt (Quattor)

Maarten Lithmaath, CERN

Flavia Donno, CERN

Project Topology: Partners

Code contribution

besides DESY, FERMI

NDGF : ftp (protocol V2)

IN2P3 : HoppingManager

Integration. Verification

- CERN
- Open Science Grid
- d-Grid

SRM Functionality at a glance

- SURL
  - srm://fapl110.fnal.gov:8443/srm/managerv2?SFN=/ /pnfs/fnal.gov/data/test/file1
- Data Transfer functions
  - srmPrepareToPut: SURLs, Protocols->TURL
  - srmPrepareToGet : SURLs, Protocols->TURL
  - srmCopy: SourceSURLs ->DestinationSURLs
- Space Management functions
  - srmReserveSpace -> SpaceToken
    - srmPrepareToPut, srmCopy
  - SrmReleaseSpace<-SpaceToken
- Directory Functions
  - srmLs, srmMkdir,srmRm, srmMkDir, srmRmDir
- Permission Functions

TapeXDiskY vs. AccessLatency and RetentionPolicy

- From SRM v2.2 WLCG MOU
  - the agreed terminology is:
    - TAccessLatency {ONLINE, NEARLINE}
    - TRetentionPolicy {REPLICA, CUSTODIAL}
  - The mapping to labels ‘TapeXDiskY’ is given by:
    - Tape1Disk0: NEARLINE + CUSTODIAL
    - Tape1Disk1: ONLINE + CUSTODIAL
    - Tape0Disk1: ONLINE + REPLICA

AccessLatency support

- AccessLatency = Online
  - File is guaranteed to stay on a dCache disk even if it is written to tape
  - Faster access but greater disk utilization
- AccessLatency = Nearline
  - In Taped backed system file can be removed from disk after it is written to tape
  - No difference for tapeless system
- Property can be specified as a parameter of space reservation, or as an argument of srmPrepareToPut or srmCopy operation

SRM Client Server Interactions at a glance

SRM Reserve Space

1. srm-reserve-space requests a new reservation
2. While request status is “in progress”, update the status
3. Eventually status is changded “success”, Space Token is available,

Srmcp reads/writes a file(s)

1. srmcp issues get/put, gets request token back
2. while request status is “in progress”, update request status
3. once status is ready and TURL(s) is available perform transfer from/to TURL(s)
4. once transfer completes, set file status to “Done”

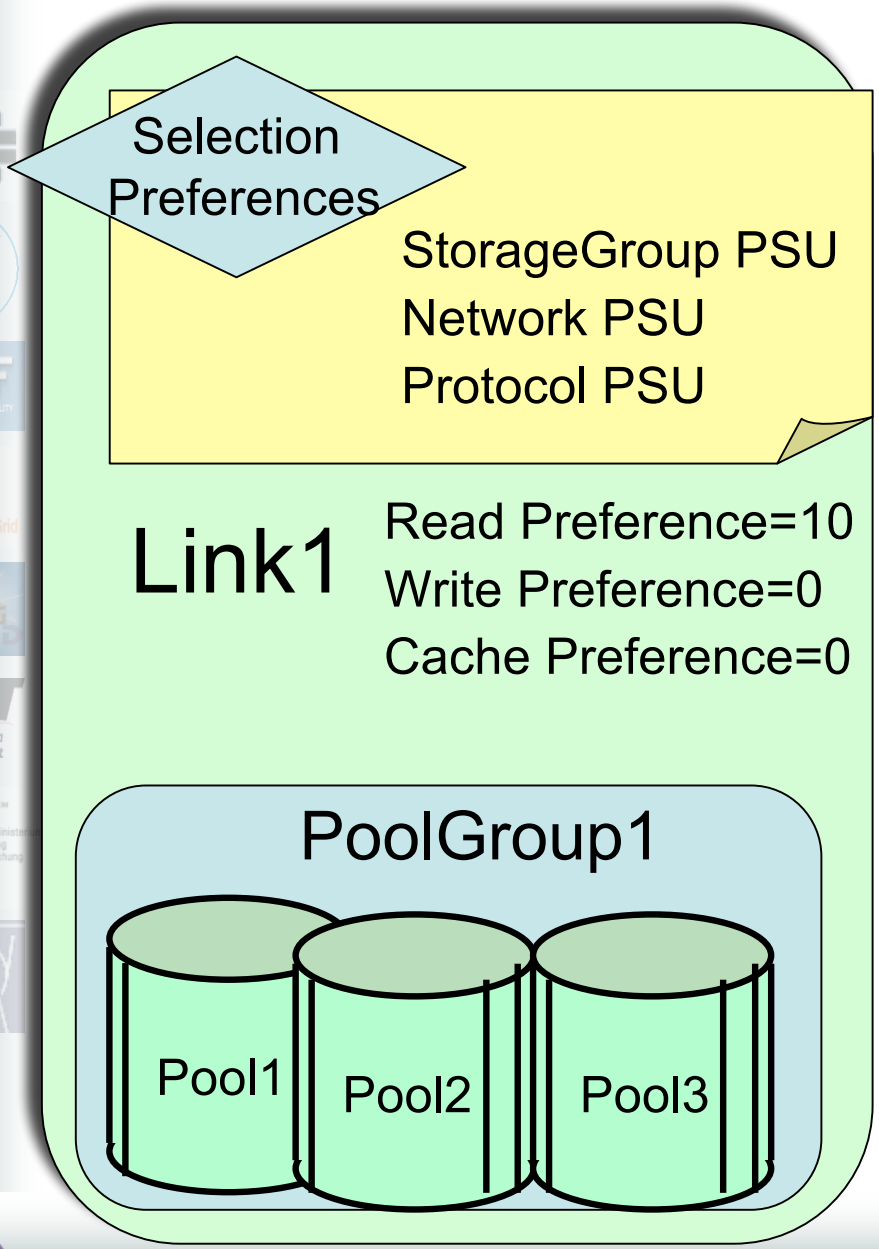
Srmcp copies a file from one SRM server to another

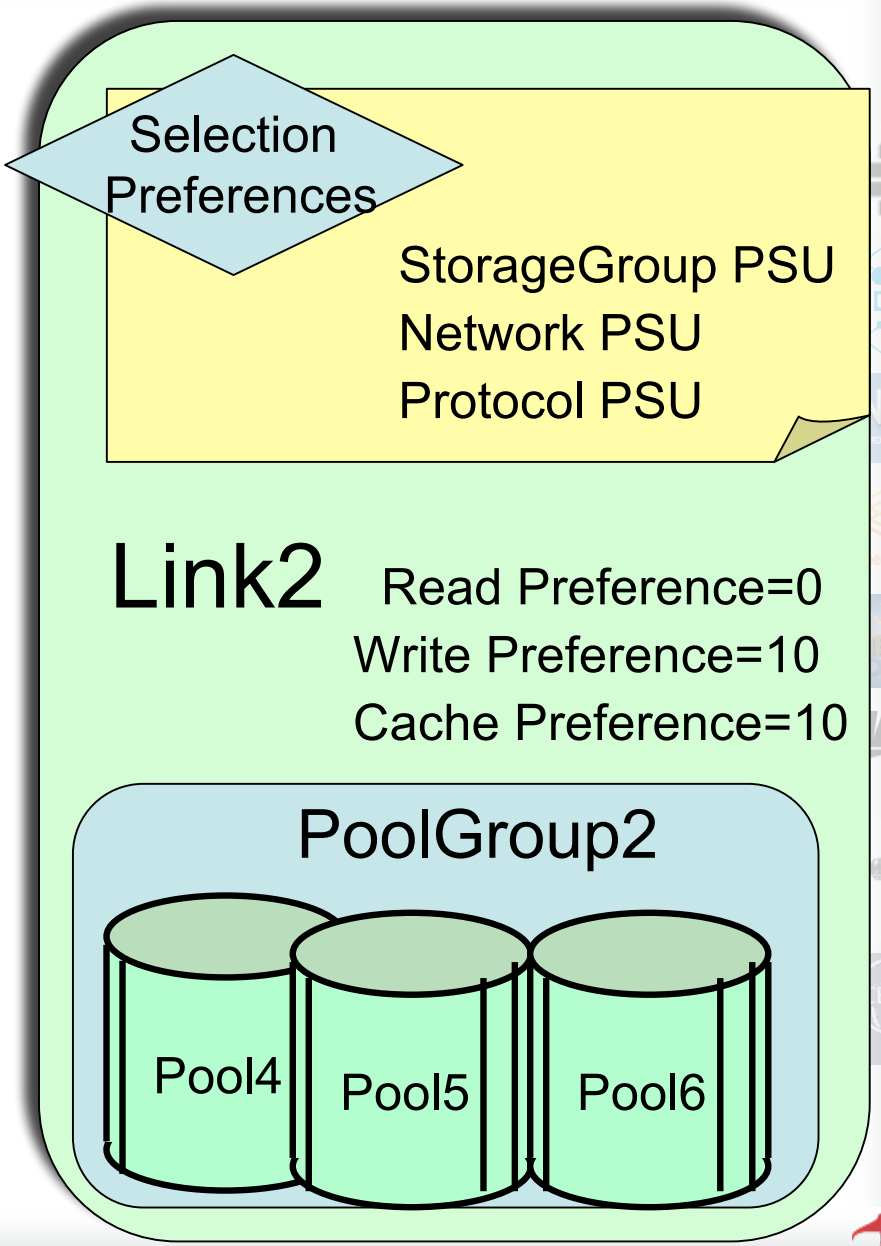
1. srmcp issues copy, gets request token back
2. while request status is “in progress”, update request status
3. once status is “Done”, transfer has completed, report result and exit.

dCache Specific Concepts Outline

- Disk Space Management
- Link Groups
- Space Reservations
- Putting Files in Space Reservations
- Movement of Files in Spaces
- Return of Space to Reservations

dCache Disk Space Management





Link Groups

Link Group 1 (T1D0)

replicaAllowed=false
outputAllowed=false
custodialAllowed=true
onlineAllowed=false
nearlineAllowed=true

Size= xilion Bytes

Link1

Link2

Link Group 1 (T0D1)

replicaAllowed=true
outputAllowed=true
custodialAllowed=false
onlineAllowed=true
nearlineAllowed=false

Size= few Bytes

Link3

Link4

Space Reservation

Link Group 1

Space Reservation 1
Custodial, Nearline
Token=777
Description“Lucky”

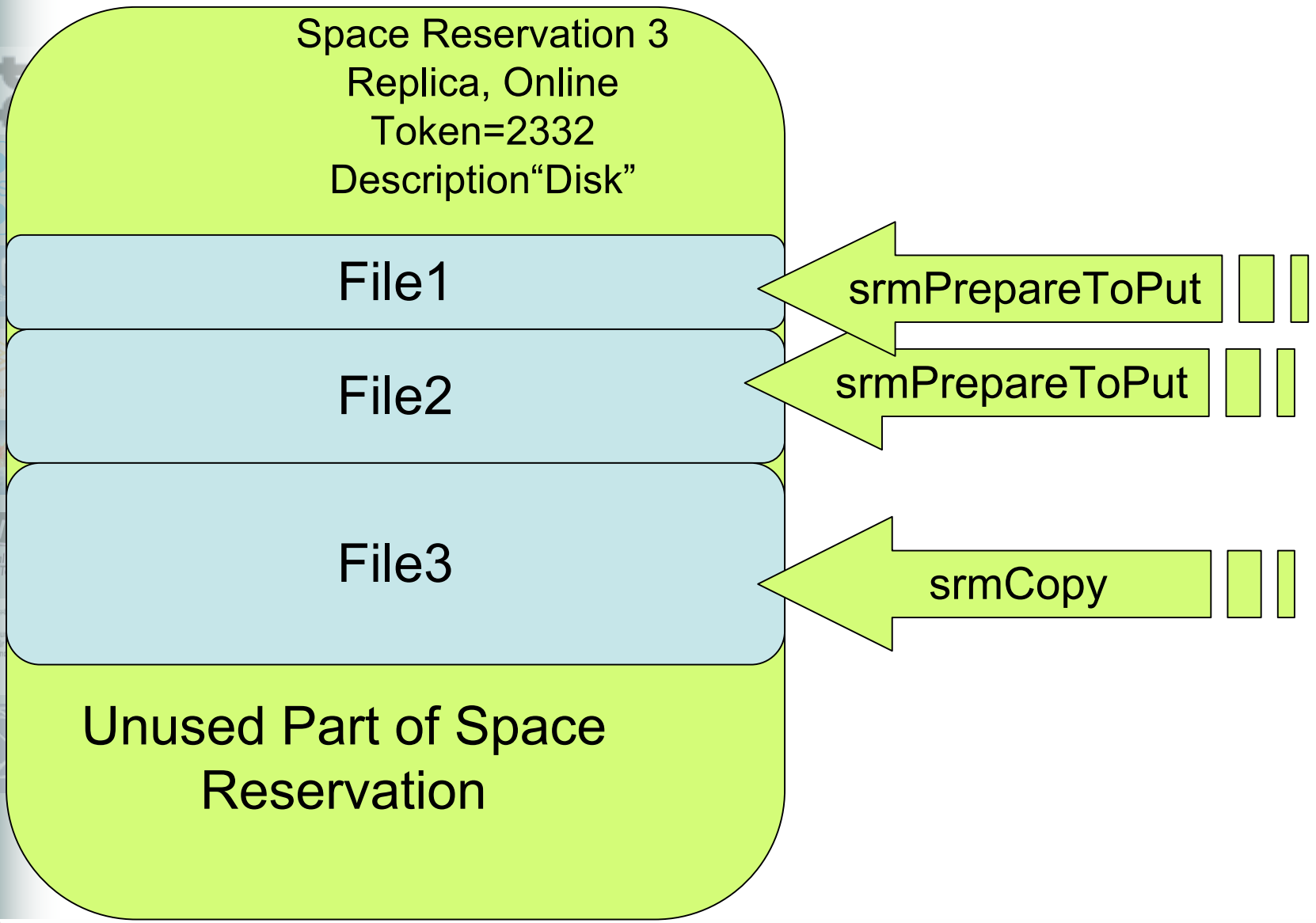
Space Reservation 2
Custodial, Nearline
Token=779
Description“Lucky”

Not Reserved

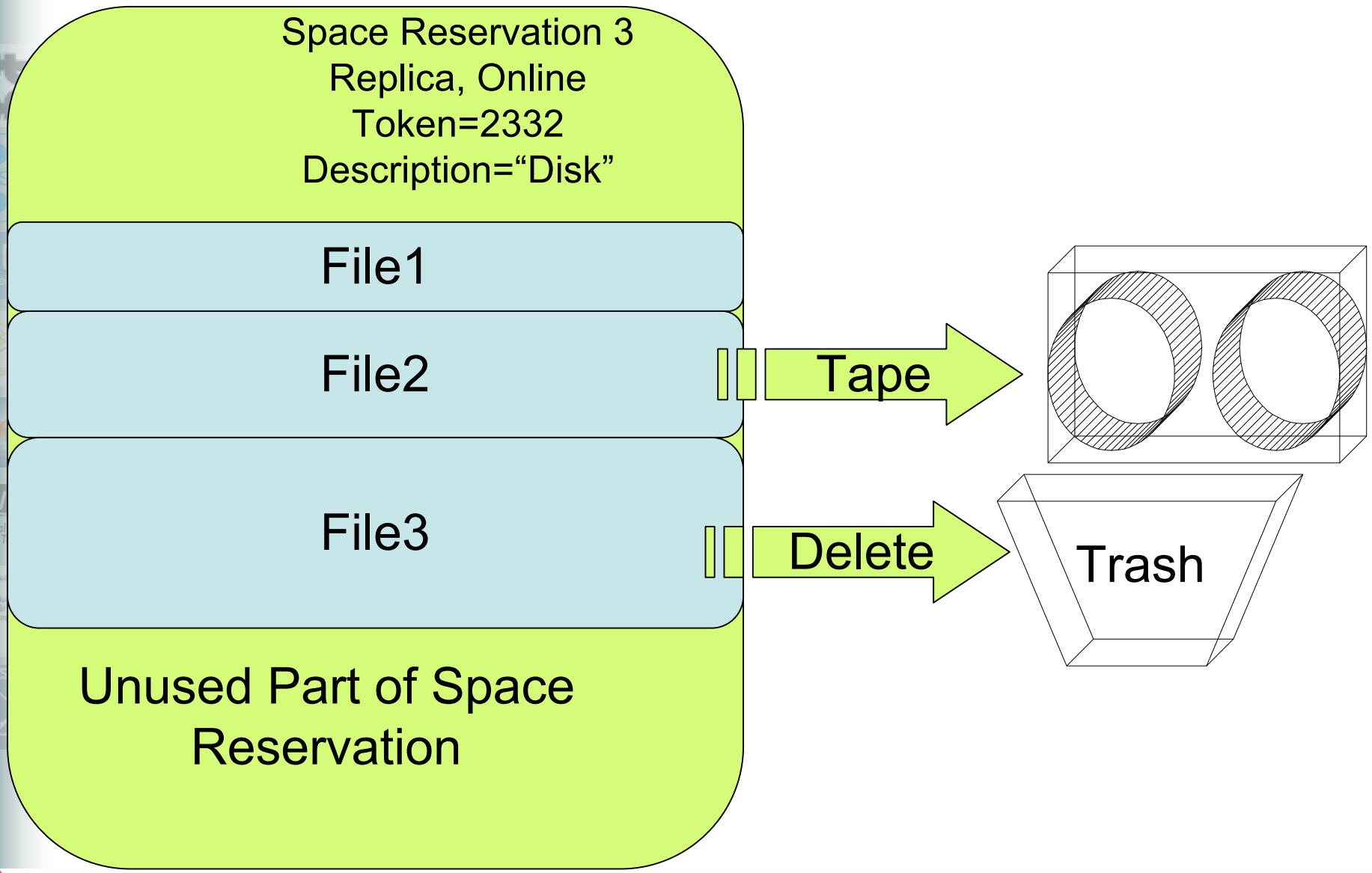
Link Group 2

Space Reservation 3
Replica, Online
Token=2332
Description“Disk”

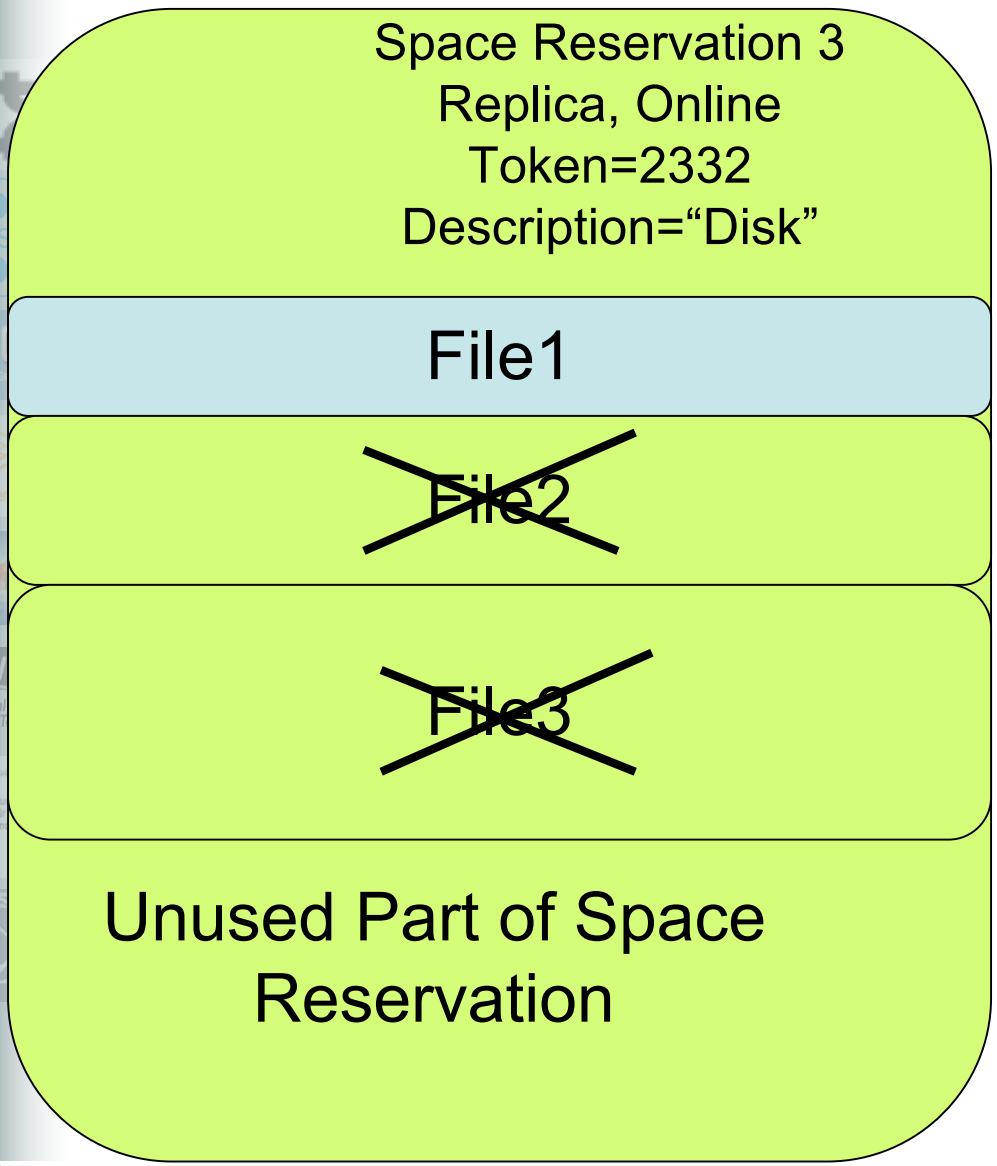
Putting Files in Space Reservations



Movement of Files in Spaces



Return of Space to Reservations



Deployment

- Separate SRM node
- Linux
- PostgreSQL
- Details:
  - http://www.dcache.org/manuals/Book/cf-srm-hrd-os.shtml
  - http://www.dcache.org/manuals/Book/cf-srm-psql.shtml

dCache SRM Configuration Outline

- Pool Manager configuration
- Srm Cell configuration
- Srm Space Manager configuration
- Default Access Latency and Retention Policy
- Some admin commands

PoolManager.conf (1)

```
psu create unit -store  *@*
psu create unit -net    0.0.0.0/0.0.0.0
psu create unit -protocol  */*

psu create ugroup  any-protocol
psu addto ugroup any-protocol */*
psu create ugroup world-net
psu addto  ugroup world-net 0.0.0.0/0.0.0.0
psu create ugroup any-store
psu addto  ugroup any-store *@*

psu create pool w-fnisd1-1
psu create pgroup writePools
psu addto  pgroup writePools w-fnisd1-1

psu create link write-link world-net any-store any-protocol
psu set link write-link -readpref=1 -cachepref=0 -writepref=10
psu add link write-link writePools
```

Selection Units (match everything)

Ugroups

Pools and PoolGroups

Link

PoolManager.conf (2)

LinkGroup

```
psu create linkGroup write-LinkGroup
psu addto linkGroup  write-LinkGroup write-link
```

LinkGroup attributes
For Space Manager

```
psu set linkGroup custodialAllowed write-LinkGroup true
psu set linkGroup outputAllowed write-LinkGroup false
psu set linkGroup replicaAllowed write-LinkGroup false
psu set linkGroup onlineAllowed write-LinkGroup false
psu set linkGroup nearlineAllowed write-LinkGroup true
```

SRM Configuration

```
serviceLocatorHost=fapl110.fnal.gov
serviceLocatorPort=11111
```

Location of dCache Domain node

```
srmDbName=dcache
srmDbUser=srmdcache
srmDbPassword=<…>
```

Location of SRM Database

Do not modify srm.batch,
as it will be overwritten by the rpm upgrade

```
[root] #  /opt/d-cache/install/install.sh
[root] # /opt/d-cache/bin/dcache-core start
```

Install and start

```
[root] # ln -s
/opt/d-cache/libexec/apache-tomcat-5.5.20/logs/catalina.out
/var/log/srmDomain.log
```

Make a link To srm log

Details: http://www.dcache.org/manuals/Book/cf-srm-srm.shtm
http://www.dcache.org/manuals/Book/cf-srm-expert-config.shtml

SRM Space Manager Configuration

To reserve or not to reserve
Needed on SRM and DOORS!!!

```
srmSpaceManagerEnabled=yes

srmImplicitSpaceManagerEnabled=yes

SpaceManagerReserveSpaceForNonSRMTransfers=true
```

SRM V1 and V2 transfers Without prior space reservation

Gridftp without prior srmPut

Link Groups Authorization

```
SpaceManagerLinkGroupAuthorizationFileName=
"/opt/d-cache/etc/LinkGroupAuthorization.conf"
```

```
LinkGroup write-LinkGroup
/fermigrid/Role=tester
/fermigrid/Role=/production

LinkGroup freeForAll-LinkGroup
*/Role=*
```

Default Access Latency and Retention Policy

```
SpaceManagerDefaultRetentionPolicy=CUSTODIAL
SpaceManagerDefaultAccessLatency=NEARLINE
```

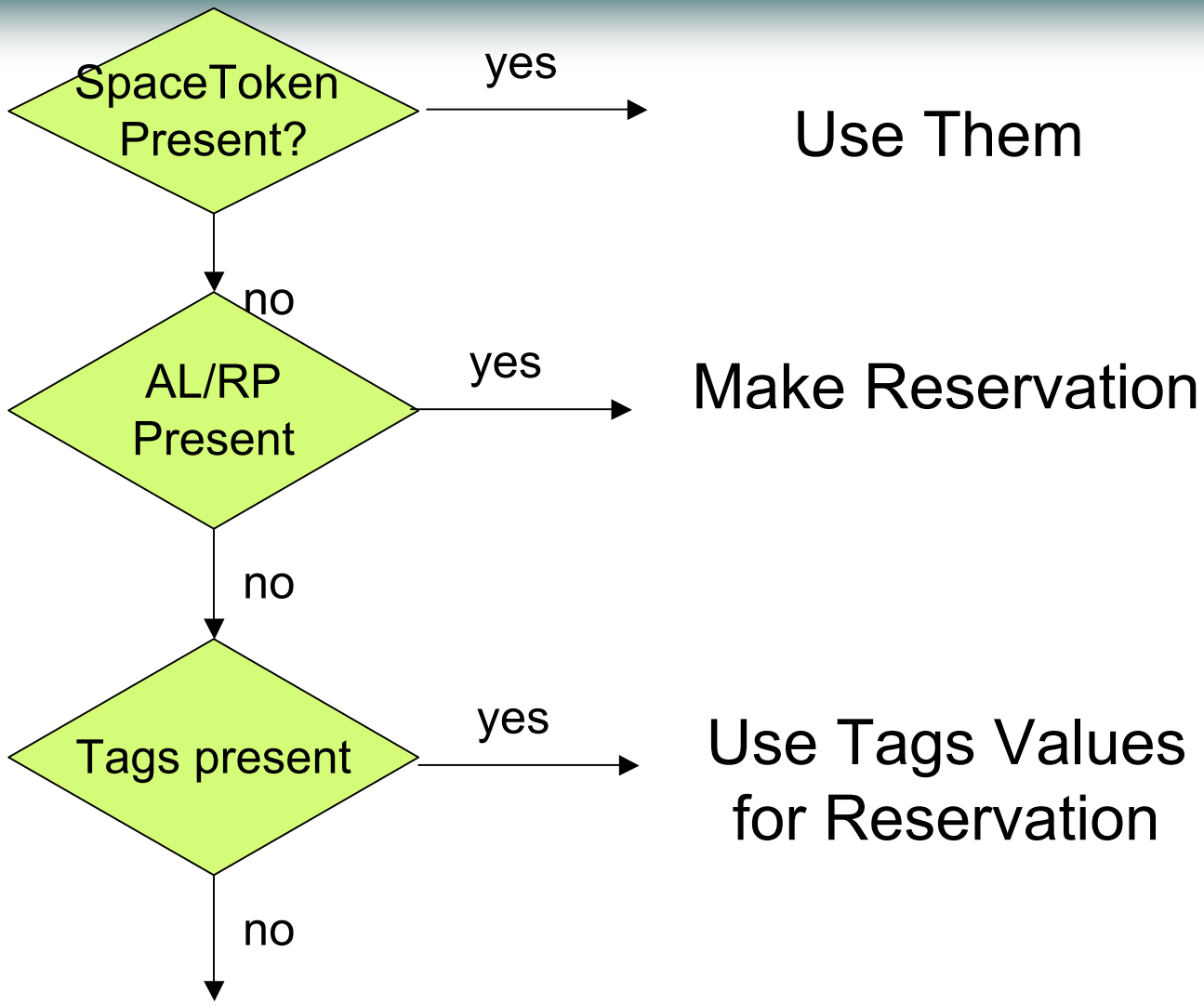
System Wide Defaults

Pnfs Path specific default

```
[root] #  cat  ".(tag)(AccessLatency)"
ONLINE
[root] #  cat  ".(tag)(RetentionPolicy)"
CUSTODIAL
[root] #  echo  NEARLINE > ".(tag)(AccessLatency)"
[root] #  echo  REPLICA > ".(tag)(RetentionPolicy)"
```

Details: http://www.dcache.org/manuals/Book/cf-srm-space.shtml

Space Type Selection



SRM Admin Commands

```
ls ["-l"] [<requestId>] # list current requests,
   including history of transitions in case of a single
   request with -l
cancel <requestId> # cancel a given request or file in
   a request
# Use double quotes for srm request ids:
# Example "-3983984034"
dir creators ls  [-l] # dir code used to be prone to
   blocking, this will allow to detect this
cancel dir creation  <path> # and to clean up without
   restart
```

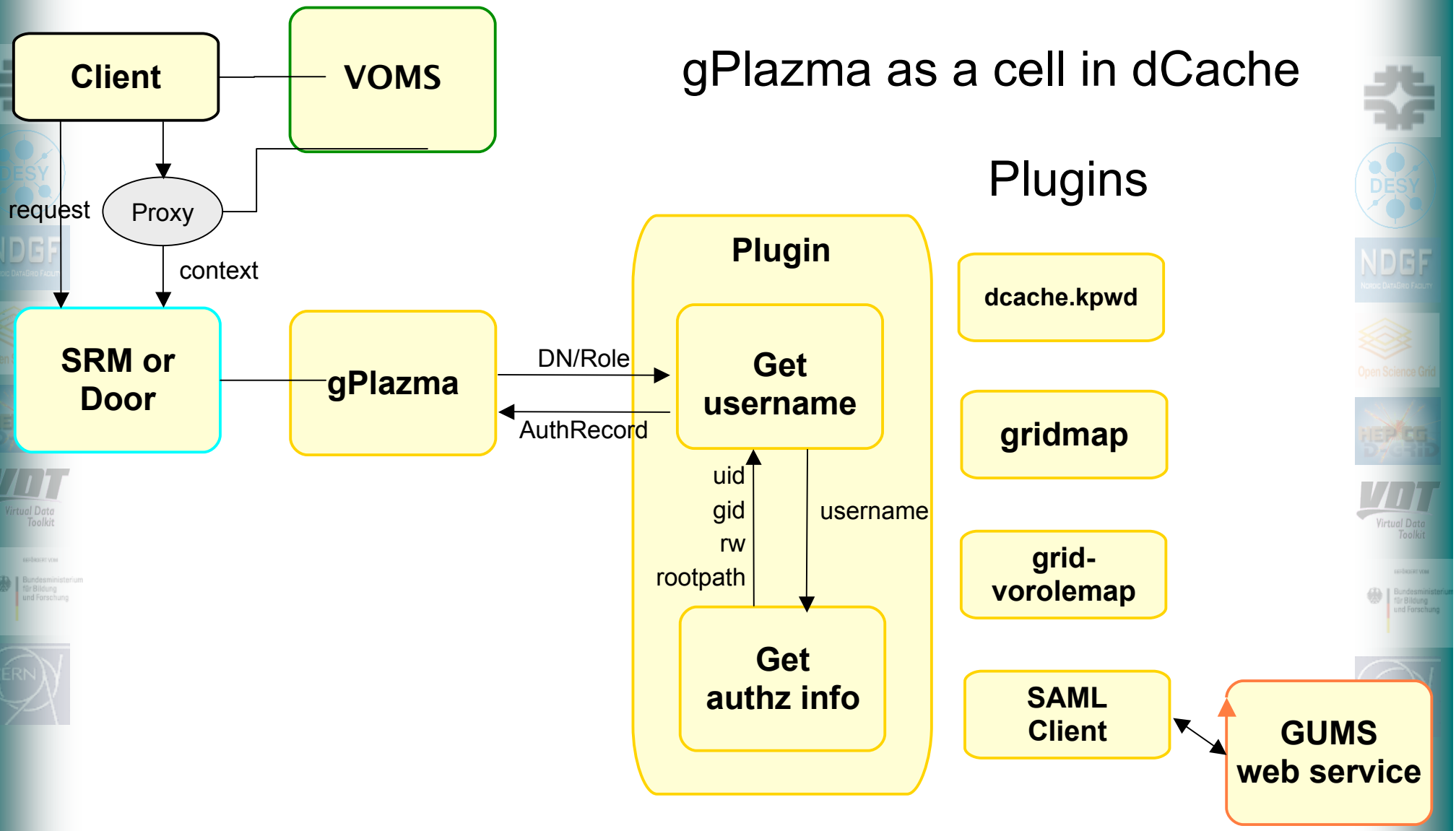
SrmSpaceManager admin command

```
reserve   [-vog=voGroup] [-vor=voRole] [-
   acclat=AccessLatency] [-retpol=RetentionPolicy] [-
   desc=Description]  [-lgid=LinkGroupId] [-
   lg=LinkGroupName] <sizeInBytes> <lifetimeInSecs (use
   quotes around negative one)> # create a new
   reservation

release  <spaceToken> [ <bytes> ] # release the space
   reservation identified by <spaceToken> # release
   existing  reservation

ls  [-l] # list reservations and link groups

update link groups # trigger update now, which is
   otherwise performed every 3 min

listFilesInSpace  <space-id> # what are the files
   already written into this space
```

gPlazma Outline

- gPlazma Architecture
- gPlazma Policy File
- dCache.kpwd
- grid-vorolemap
- gPlazma Use Cases

gPlazma Architecture



gPlazma Policy File

${ourHomeDir}/etc/dcachesrm-gplazma.policy

```
# Switches
saml-vo-mapping="ON"
kpwd="ON"
grid-mapfile="OFF"
gplazmalite-vorole-mapping="OFF"

# Priorities
saml-vo-mapping-priority="1"
kpwd-priority="3"
grid-mapfile-priority="4"
gplazmalite-vorole-mapping-priority="2”

# dcache.kpwd
kpwdPath="/opt/d-cache/etc/dcache.kpwd”
# grid-mapfile
gridMapFilePath="/etc/grid-security/grid-mapfile"
storageAuthzPath="/etc/grid-security/storage-authzdb”
# Built-in gPLAZMAlite grid VO role mapping
gridVoRolemapPath="/etc/grid-security/grid-vorolemap"
gridVoRoleStorageAuthzPath="/etc/grid-security/storage-authzdb”
# SAML-based grid VO role mapping
mappingServiceUrl=
          "https://gums.oursite.edu:8443/gums/services/GUMSAuthorizationServicePort"
```

dCache.kpwd

The dcache.kpwd file is used to map a user's DN to a local username, then a second mapping is performed to obtain the uid, gid, read-write privilege, and rootpath from the username. If the mappings succeed, file system access is controlled by use the uid and gid in the context of unix file permissions, and checks of the read-write privilege and that the path of the transfer is within the designated rootpath.

In this method both the username and the resulting set of permissions are contained in the same file.

dcache.kpwd:

```
# Mappings for 'cmsprod' users
mapping "/DC=org/DC=doegrids/OU=People/CN=Ted Hesselroth 899520" cmsprod
mapping "/DC=org/DC=doegrids/OU=People/CN=Shaowen Wang 564753" cmsprod


# Login for 'cmsprod' users
login cmsprod read-write 9801 5033 / /pnfs/fnal.gov/data/cmsprod /pnfs/fnal.gov/data/cmsprod
        /DC=org/DC=doegrids/OU=People/CN=Ted Hesselroth 899520
        /DC=org/DC=doegrids/OU=People/CN=Shaowen Wang 564753
```

grid-vorolemap

In this method the mapping to the username is done from the concatenation of the user's DN with the user's Role (or, more precisely, with the user's Fully Qualified Attribute Name).

The mapping of the user's DN and role to a username is in the grid-vorolemap file.

/etc/grid-security/grid-vorolemap:

```
"/DC=org/DC=doegrids/OU=People/CN=Ted Hesselroth 899520" "/cms/uscms/Role=cmsprod" uscms01

"/DC=org/DC=doegrids/OU=People/CN=Keri Pembrook 651725" dzero

# Wildcards for the DN are permitted.

"*" "/cms/uscms/Role=cmsprod" cmsprod

"*" "/cms/uscms/Role=analysis" analysis
```

The mapping of username to the user's set of permissions is through the storage-authzdb file.

etc/grid-security/storage-authzdb:

```
authorize cmsprod read-write 9811 5063 / /pnfs/fnal.gov/data/cms /
authorize dzero   read-write 1841 5063 / /pnfs/fnal.gov/data/dzero /
```

gPlazma Cases Use

Roles for Reading and Writing

In this use case there is write privilege for cmsprod role and read privilege for analysis and cmsuser roles.

/etc/grid-security/storage-authzdb:

```
authorize cmsprod  read-write  9811 5063 / /pnfs/fnal.gov/data /
authorize analysis read-only  10822 5063 / /pnfs/fnal.gov/data /
authorize cmsuser  read-only  10001 6800 / /pnfs/fnal.gov/data /
```

User Accounts

Each DN is mapped to a unique username and each username is mapped to a unique uid, gid, and rootpath.

/etc/grid-security/grid-vorolemap:

```
"/DC=org/DC=doegrids/OU=People/CN=Selby Booth" cms821
"/DC=org/DC=doegrids/OU=People/CN=Kenja Kassi" cms822
"/DC=org/DC=doegrids/OU=People/CN=Ameil Fauss" cms823
```

/etc/grid-security/storage-authzdb for version 1.7.0:

```
authorize cms821 read-write 10821 7000 / /pnfs/fnal.gov/data/cms821 /
authorize cms822 read-write 10822 7000 / /pnfs/fnal.gov/data/cms822 /
authorize cms823 read-write 10823 7000 / /pnfs/fnal.gov/data/cms823 /
```

gPlazma Use Cases 2

User Accounts with wildcards

Starting in dCache 1.8 the above permission mapping may be done with a single line using wildcards.

/etc/grid-security/storage-authzdb for version 1.8:

```
authorize cms(\d\d\d) read-write 10$1 7000 / /pnfs/fnal.gov/data/cms$1 /
```

Blacklisting

A user or VO may be blacklisted by entering a “-” instead of a username in grid-vorolemap.

/etc/grid-security/grid-vorolemap:

```
"/DC=org/DC=doegrids/OU=People/CN=Ted Hesselroth 899520" "/cms/uscms/Role=cmsprod" -
```

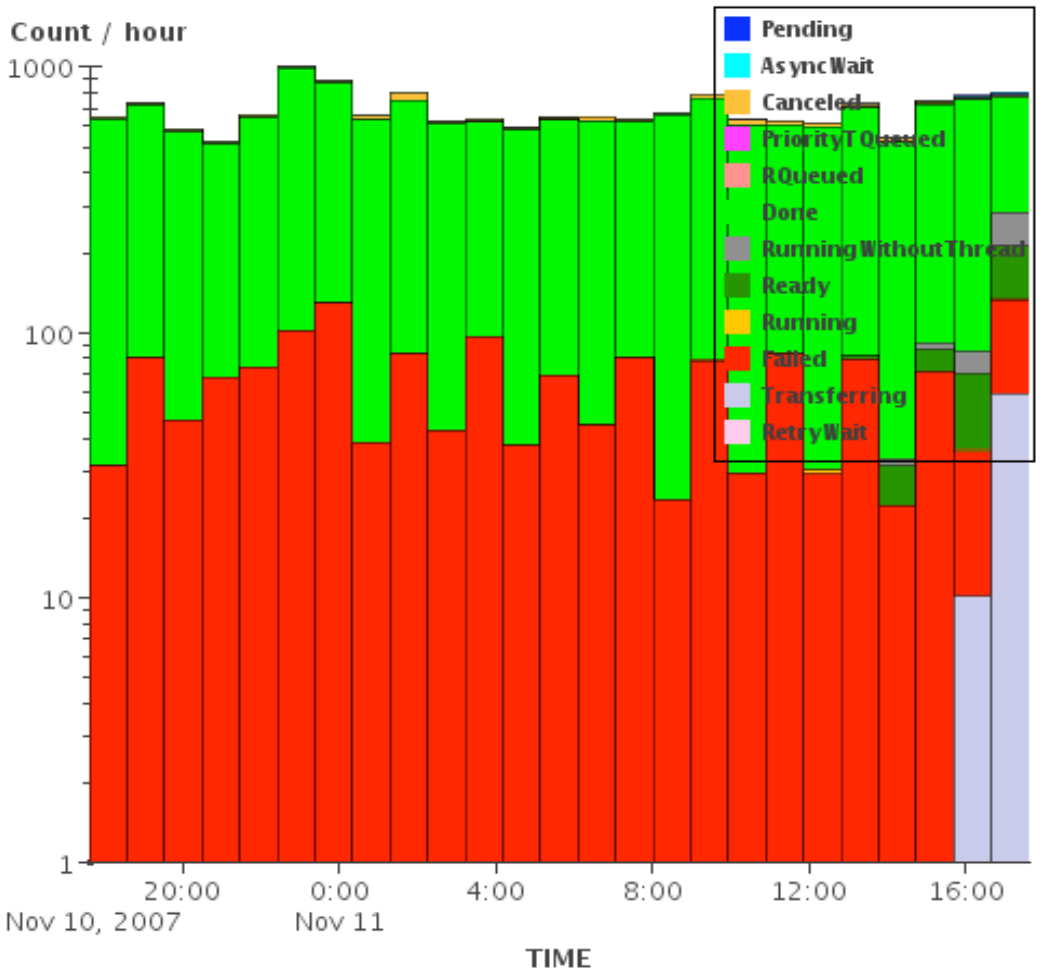
Monitoring Outline

- SRMWatch
- Log Files
- SRM Problem Diagnostics

SRMWatch

- SRM stores requests in SQL DB
- Web Based Monitoring application
- Runs in Tomcat
- Simple to Configure
- Allows to Search Requests
- Details at http://www.dcache.org/manuals/Book/cf-srm-monitor.shtml

Example of Request Count vs Time Plot



Log Files

- SRM Domain logs into

`/opt/d-cache/libexec/apache-tomcat-5.5.20/logs/catalina.out`

- What to look for:
  - “Exception”, “Error”
- Messages are preceded with a timestamp and a source:

  11/11 17:53:05 Cell(RemoteGsiftpTransferManager@srm-fapl110Domain) :

  Using time stamps it is often possible to correlate the events in different dCache components
- Sources of the errors:
  - SrmSpaceManager: Space Reservation problem
  - RemoteGsiftpTransferManage: SRM Negotiated TURL, but could not transfer, possible cert problems on the pools
  - PinManager: could not pin file for SRM Get
- SRM Cells Communicate with
  - PnfsManager, PoolManager, LoginBroker, doors, pools
  - “No Route To Cell” is often an indication, that other services can not be reached

SRM Problem Diagnostics

- Blame the user, if does not help, then
- Check monitoring pages, logs
- Check database
- If specific types of the requests are stuck
  - Srm put, copy (pull) => Check SrmSpaceManager, dir creation
  - SRM Get, copy (push) => Check PinManager
- Check that other components (PnfsManager, PoolManager) perform
- Check the system health, load, memory usage, etc.
- Make sure that all certs are up to date
  - Update of Server or CA Certs requires SRM restart
- If this does not help, create a ticket

SRM and dCache Deployment & Upgrade Q&A

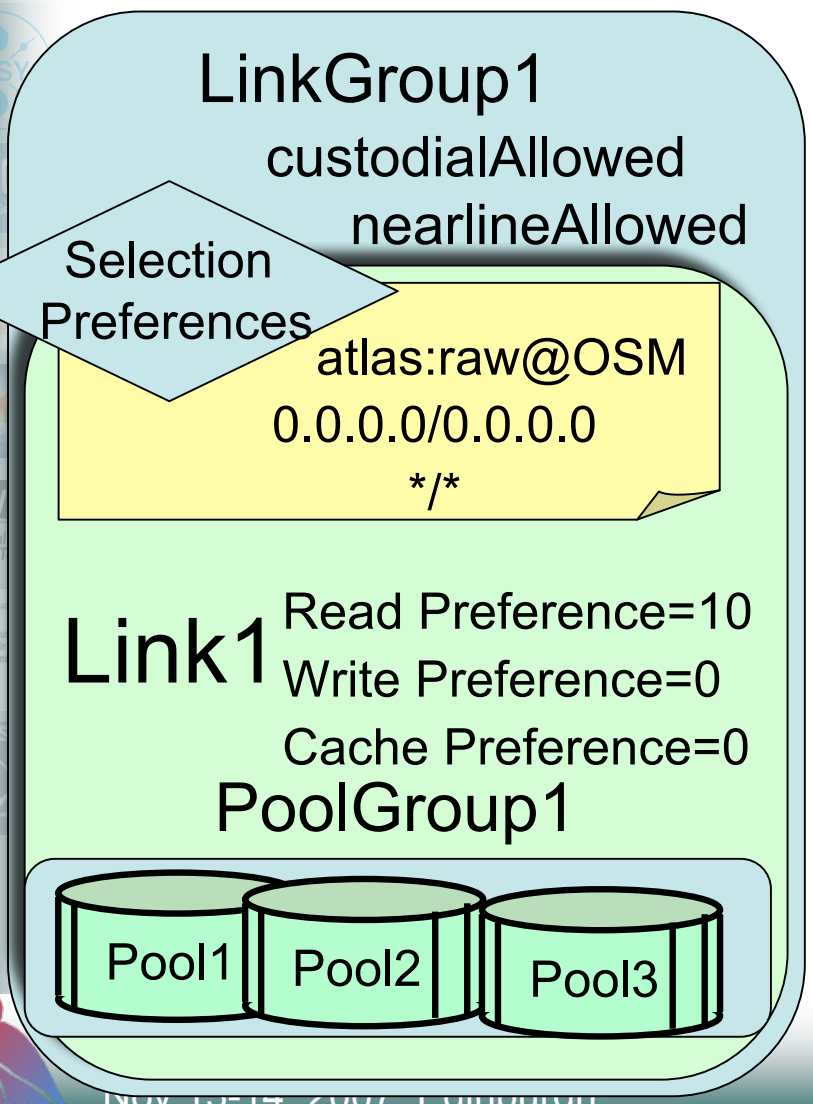
- What is the 1.7.0 -> 1.8.0 upgrade procedure
- What are the Recommendations for disk pool setup
- Load Balancing. What services can you run and on how many nodes?

What is the 1.7.0 -> 1.8.0 upgrade procedure

- Regular RPM Upgrade
- Must be performed on all Nodes
- PoolManager and SRM might need extra configuration
- Pool upgrades are transparent, no migration scripts
- Might require adding new tags to Pnfs Trees, which can take long time to complete
- ! Files going to tape should be flushed to Tape before the upgrade !
- Link Groups and Storage Tags might conflict, it is best to test your configuration before production deployment
- SRM V2.2 functionality is present only in the new 1.8 client rpm

Example of the conflicting Configuration

/pnfs/nesc.ed.ac.uk/data/atlas:
".(tag)(OSMTemplate)"=atlas
".(tag)(sGroup)"=generated



1. Reserve (CUSTODIAL, NEARLINE) =>
   TOKEN1

2. srmPrepareToPut (TOKEN1,
/pnfs/nesc.ed.ac.uk/data/atlas/File1) =>
TURL=giftp://server1/File1

3. gridftpStore(file:///data,
TURL=giftp://server1/File1) =>
Can Not Find Pool for atlas:generate@OSM

What are the Recommendations for disk pool setup

- Links not in LinkGroups will be excluded from Space Reservations
- SrmSpaceManager manages only write spaces
- dCache SrmPrepareToGet ignores space tokens
- Read Only Pools should be outside LinkGroups

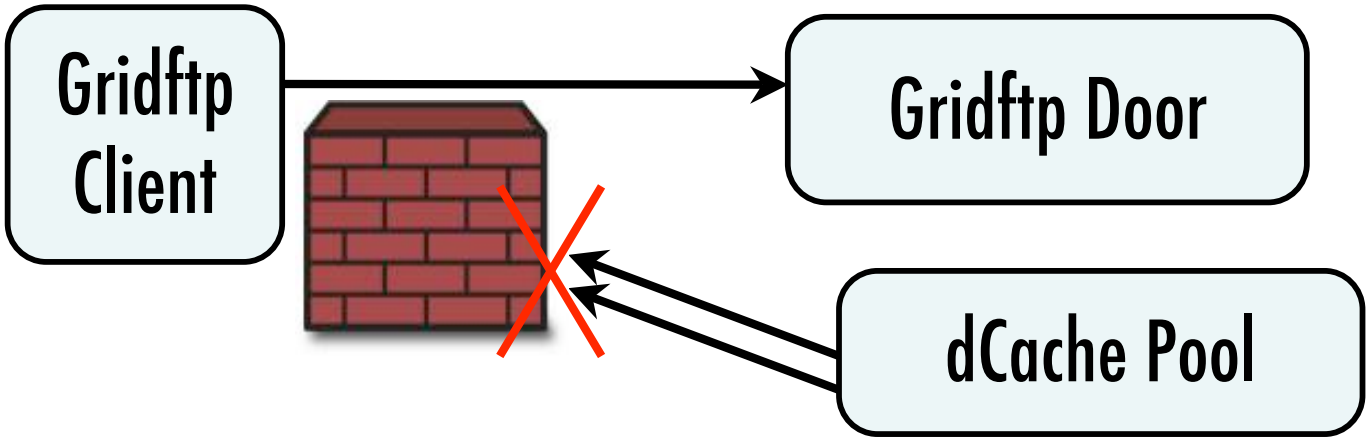
- Pools - Unlimited Number
- Doors - Unlimited Number
- LoginBroker collects information about the doors, used by SRM
  - If you want doors excluded from usage by SRM, remove LoginBroker option (door batch)
- Services that can’t be distributed (yet)
  - PnfsManager
  - PoolManager
  - SRM Services

SRM Network Usage and Firewalls Outline

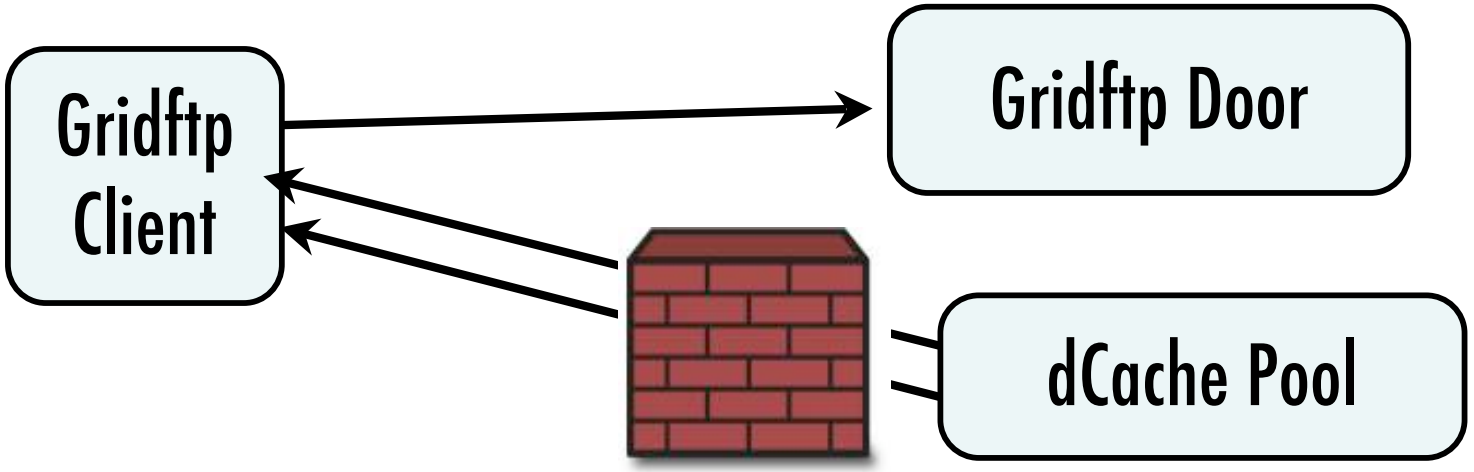
- Gridftp Get and Firewalls
- SRM Get Network Flows
- SRM Put Network Flows
- SRM Copy in Pull Mode Network Flows
- SRM Copy in Pull Mode Network Flows

Gridftp Get and Firewalls

Client Behind Firewall

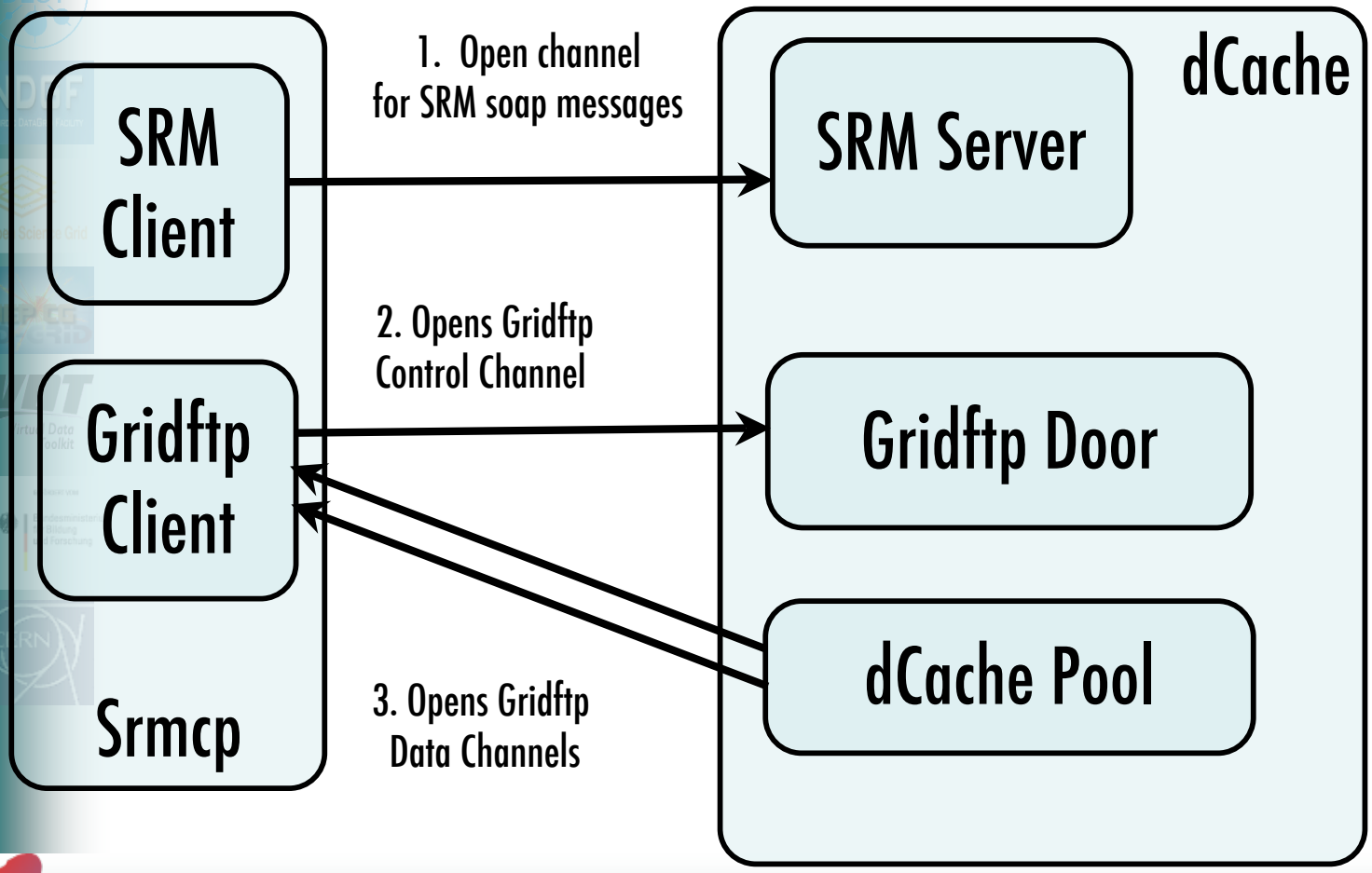


Pool Behind Firewall



SRM Get Network Flows

srmcp srm://dCache:8443/dir1/file1 file:////tmp/file1

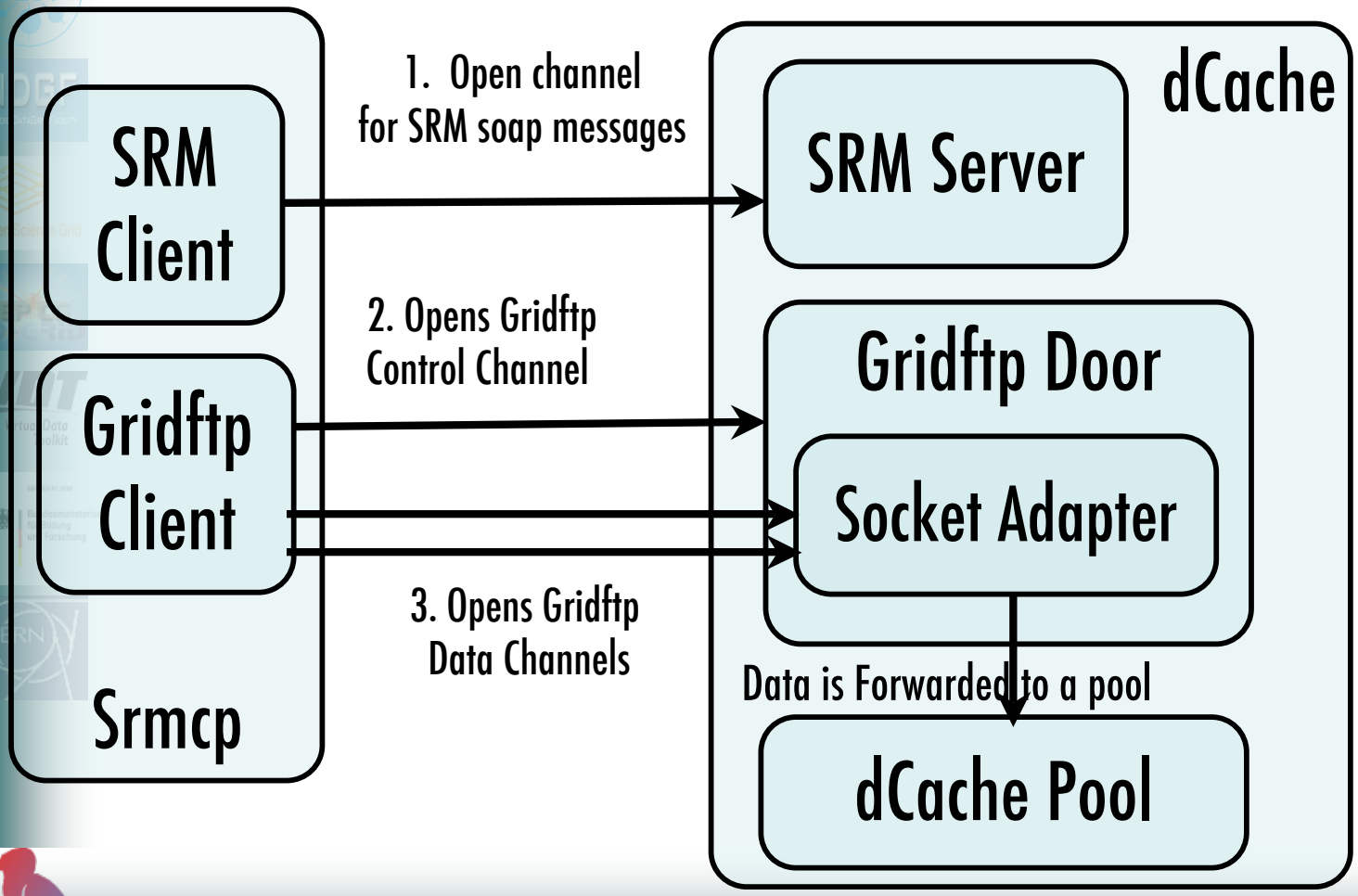


Firewalls:

- Srmcp Client port range needs to be open and configured when behind a Firewall
- Pools can be behind firewall
- Srm server port must be open on srm node
- Gridftp server port must open on Gridftp door

SRM Put Network Flows

srmcp file:////tmp/file1 srm://dCache:8443/dir1/file1

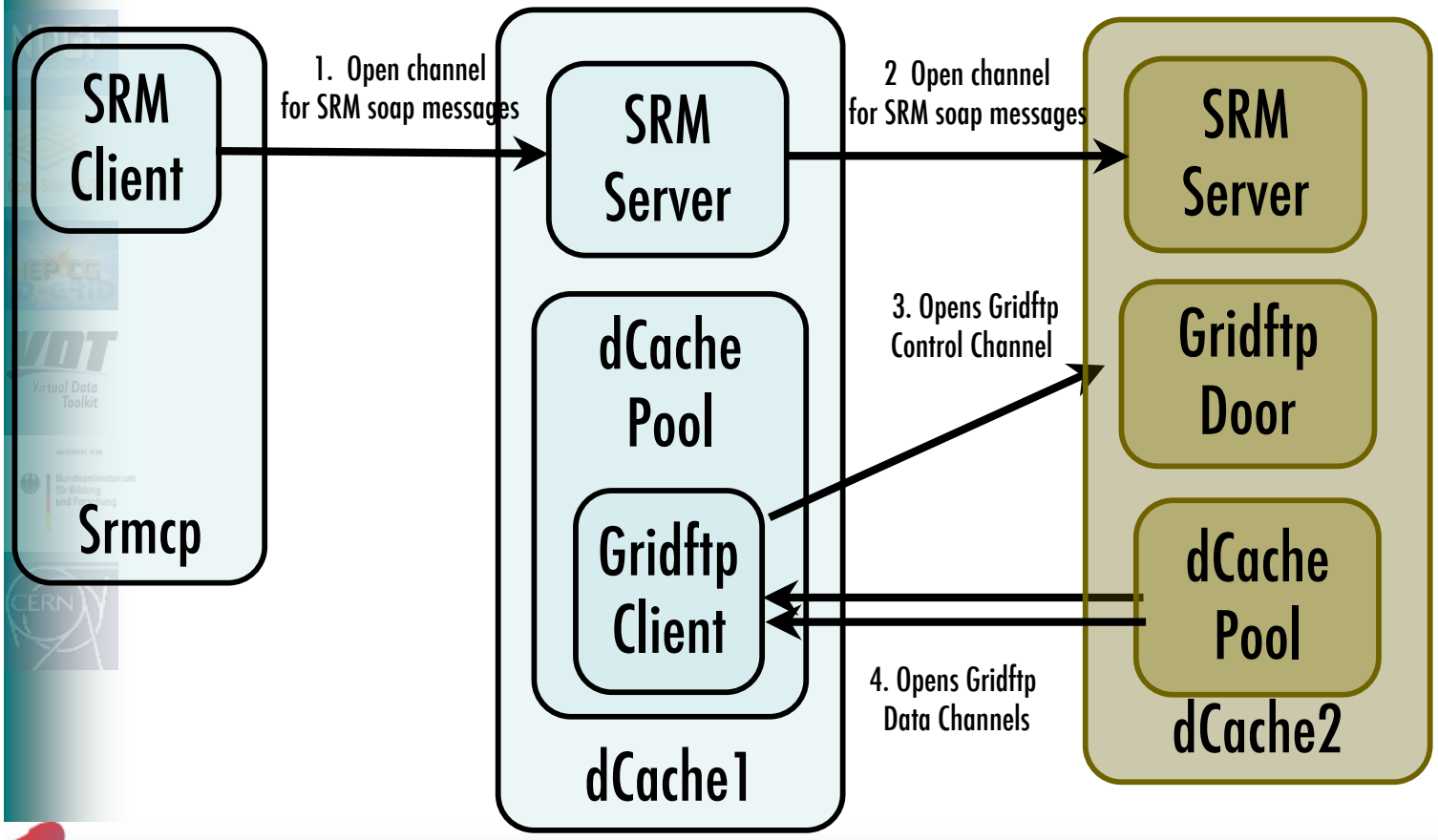


Firewalls:

- Srmcp Client can be behind a Firewall
- Pools can be behind firewall
- Srm server port must be open on srm node
- Gridftp port and port range for data must be configured on Gridftp door

SRM Copy in Pull Mode Network Flows

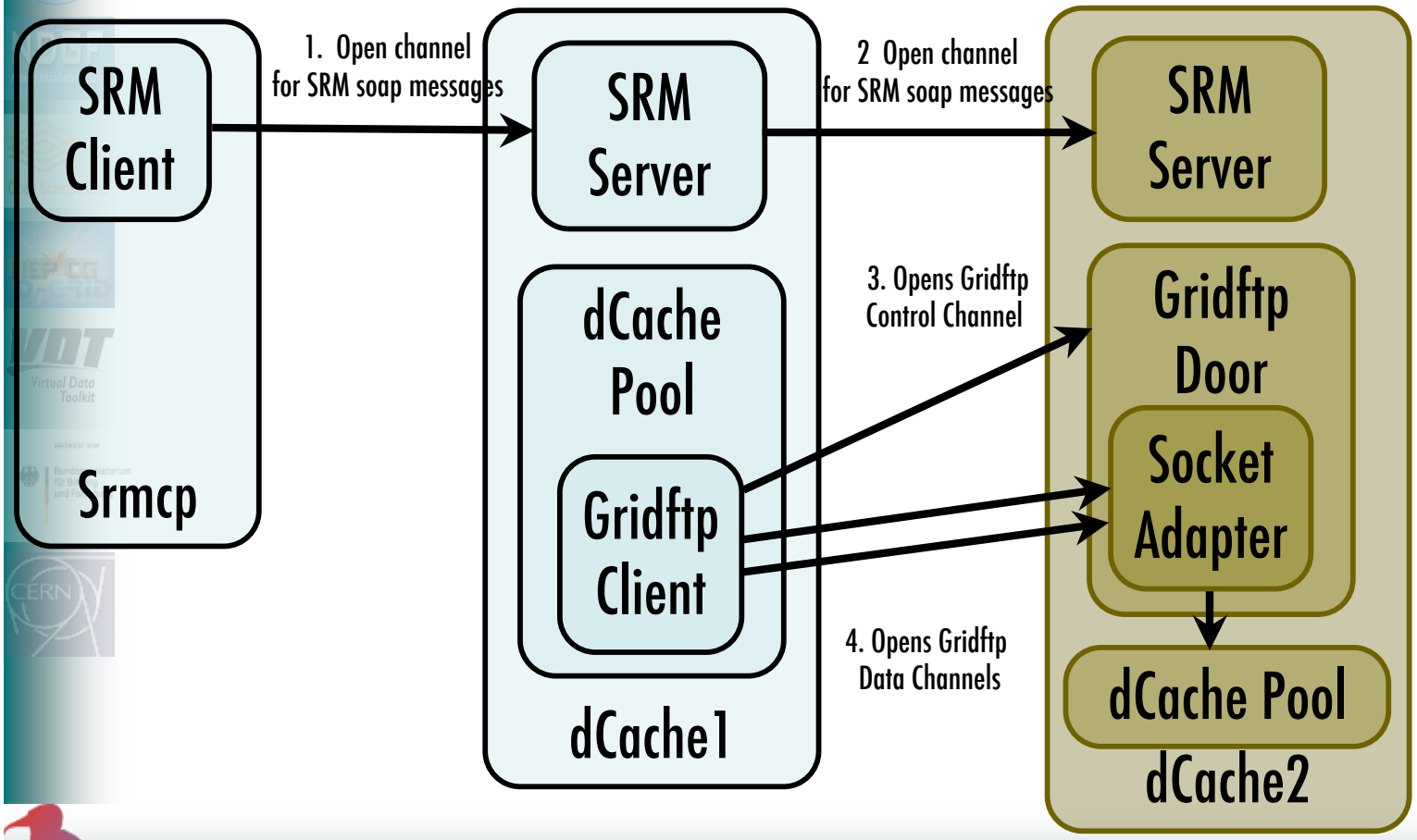
srmcp srm://dCache2:8443/dir1/file1
srm://dCache1:8443/dir1/file1.copy



Firewalls:

- Srmcp Client can be behind a Firewall
- dCache1 Pools need a port range set up and configured
- Srm servers ports must be open on srm nodes
- dCache2 Gridftp server port must be open on Gridftp door node

srmcp -pushmode srm://dCache1:8443/dir1/file1
srm://dCache2:8443/dir1/file1.copy



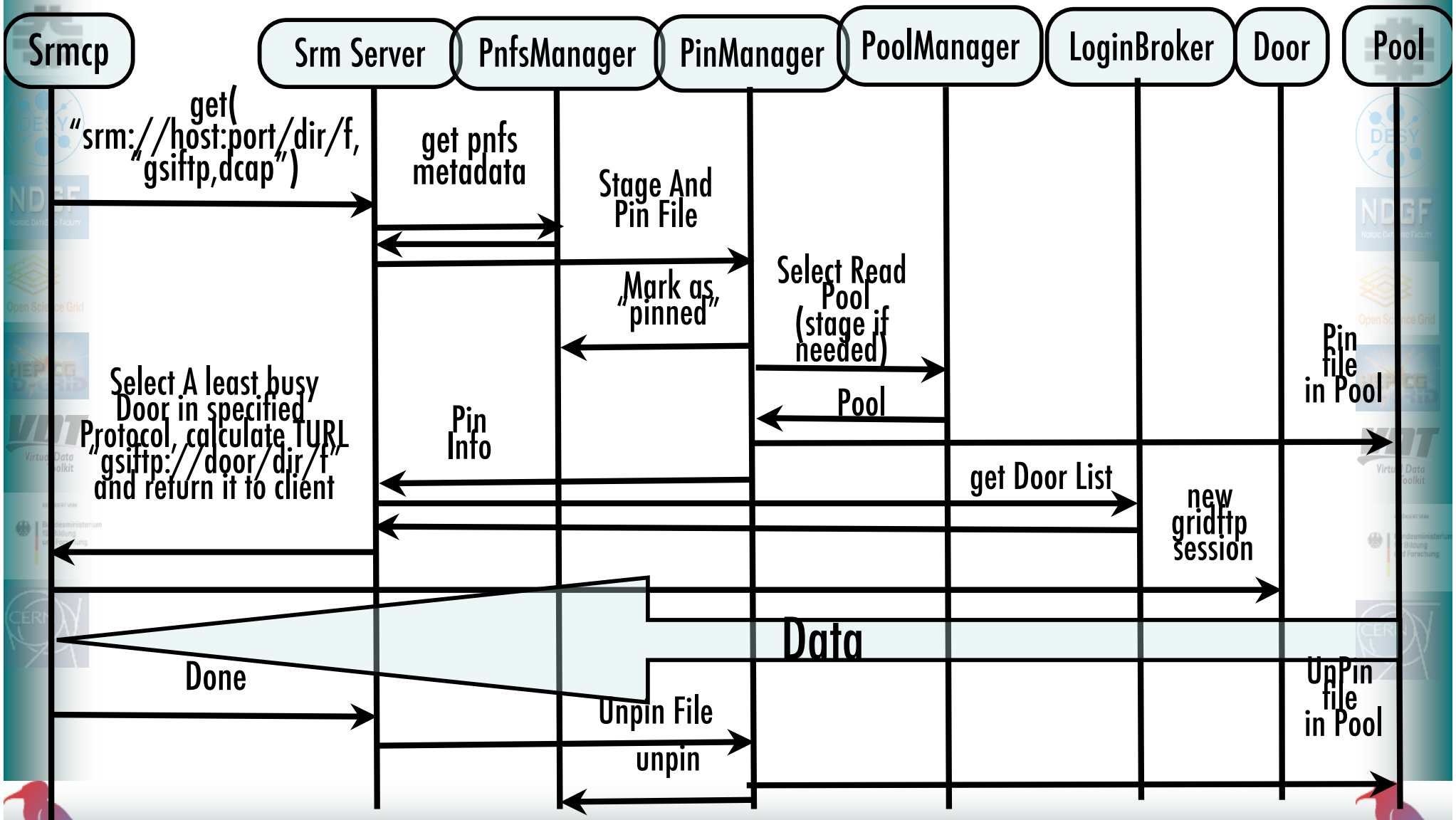
Firewalls:

- Srmcp Client can be behind a Firewall
- dCache1 Pools and dCache2 Pools can be behind firewals
- Srm servers ports must be open on srm nodes
- dCache2 Gridftp server must have a port range opne and configured port must be open on Gridftp door node

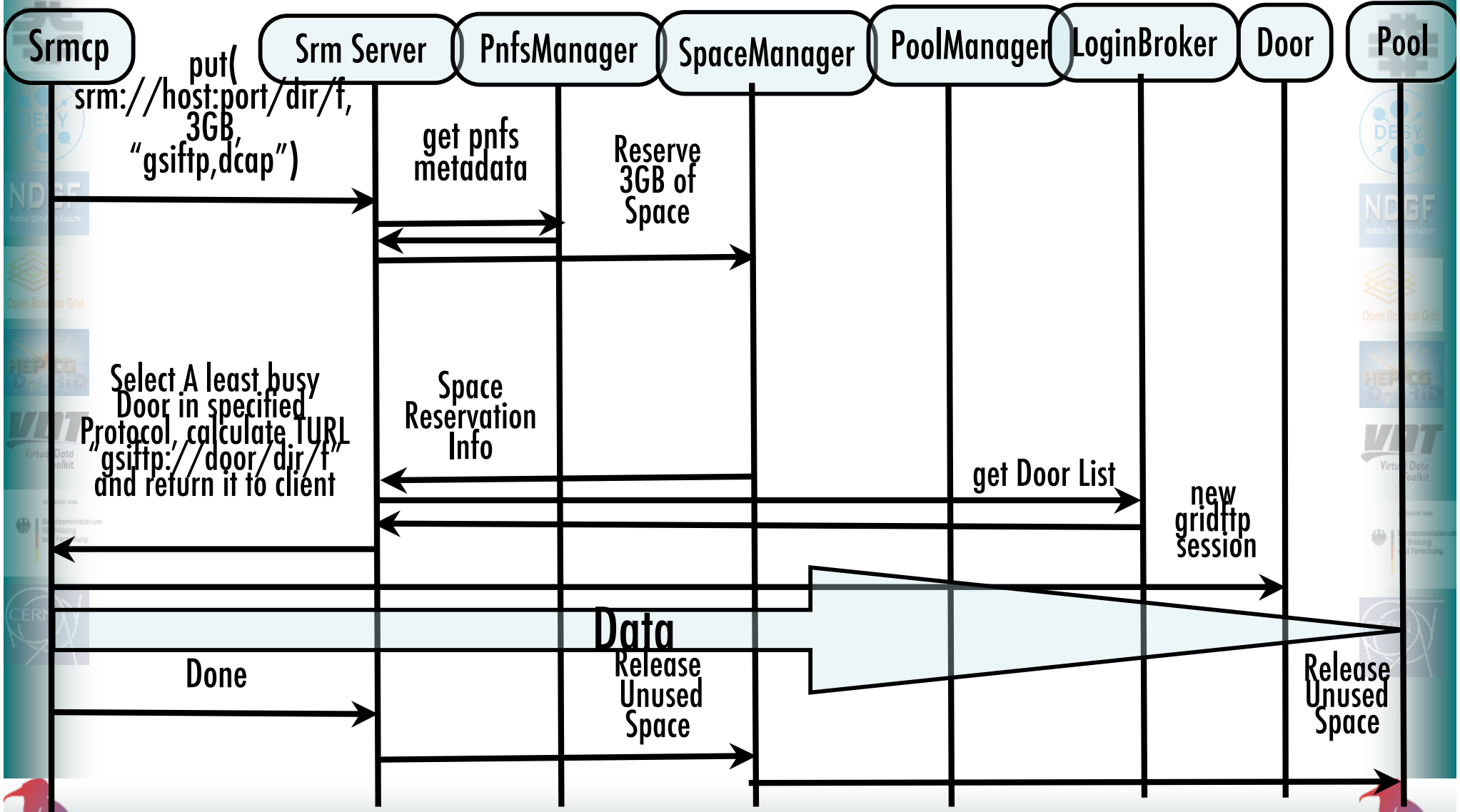
SRM inter-dCache interactions outline

- SRM Get details
- SRM Put details
- SRM Copy details

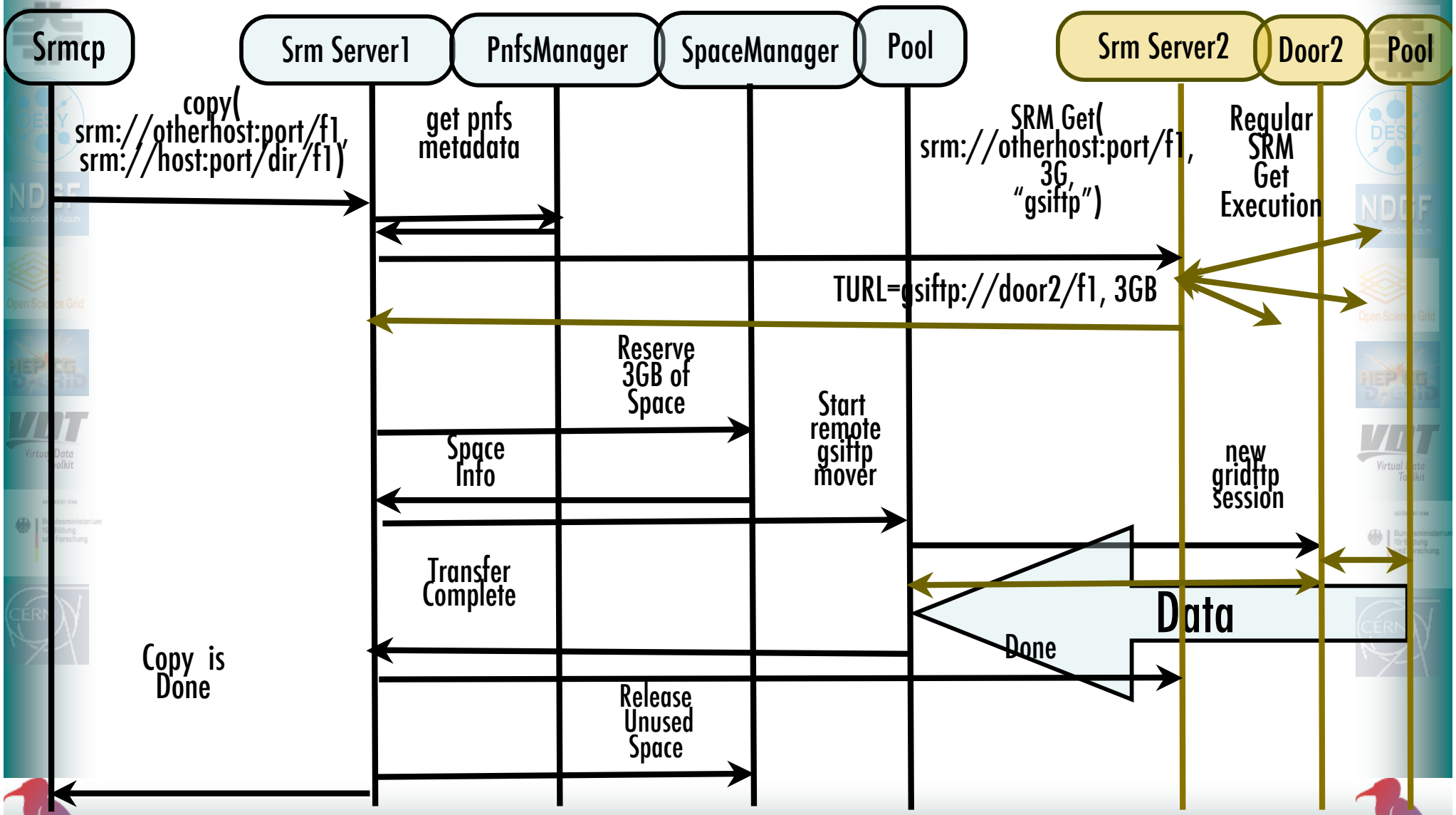
SRM Get Details



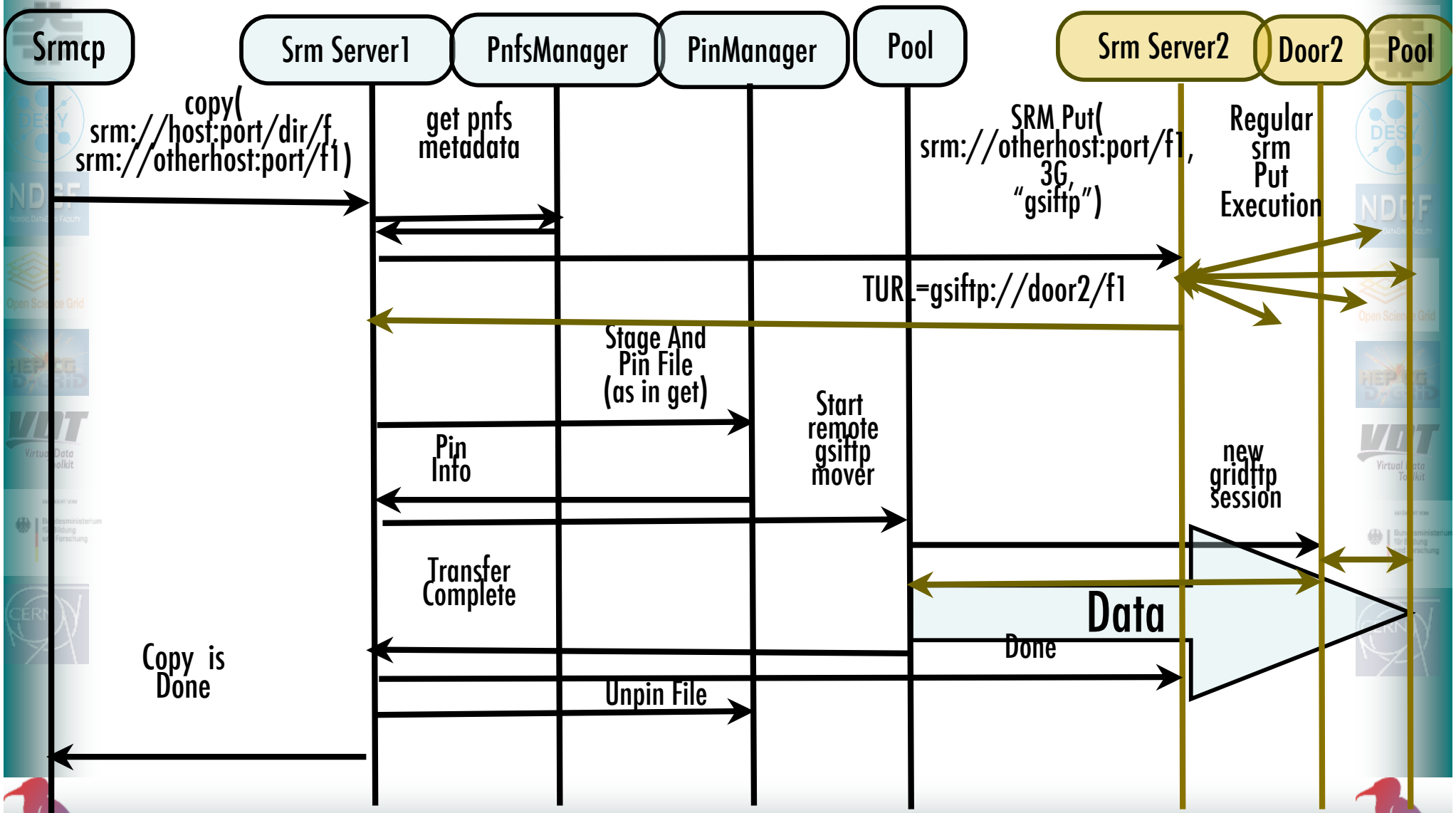
SRM Put Details



SRM Copy in pull mode details



SRM Copy in push mode detailes



References

- dCache www.dcache.org
- dCache SRM http://srm.fnal.gov
- SRM Working Group http://sdm.lbl.gov/srm-wg/
- SRM V2.2 spec http://sdm.lbl.gov/srm-wg/doc/SRM.v2.2.html