DPM Administration for Tier2s

Sophie Lemaitre (Sophie.Lemaitre@cern.ch)
Jean-Philippe Baud (Jean-Philippe.Baud@cern.ch)

Tier2s tutorial – 15 Jun 2006

DPM in a nutshell

- Description
- Installation
- DPM as a service
- Troubleshooting
- Documentation / support

DPM in a nutshell

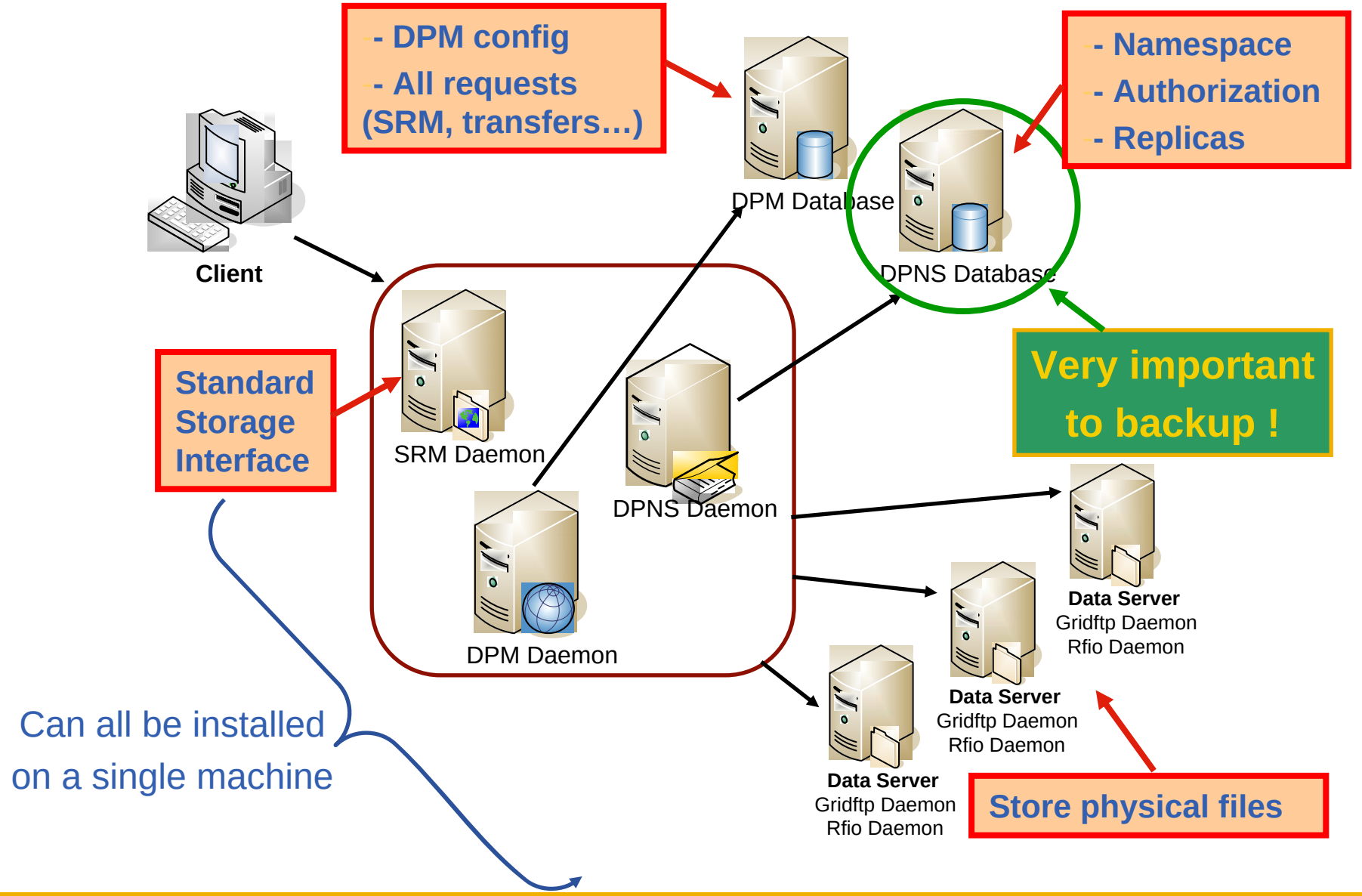
- **Description**
- **Installation**
- **DPM as a service**
- **Troubleshooting**
- **Documentation / support**

What is a DPM ?

- **Disk Pool Manager**
  - Manages storage on disk servers
  - SRM support (1.1 and 2.1)

- **Deployment status**
  - ~50 DPMs in production
  - 70 VOs supported

Architecture

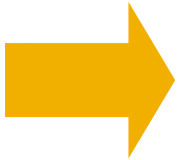


DPM strengths

- Easy to install/configure
  - Few configuration files
- Manageable storage
  - Logical Namespace
  - Easy to add/remove file systems
- Low maintenance effort
- Supports as many disk servers as needed
- Easy classic SE to DPM migration

Daemons

- DPM server
- DPM Name Server
- SRM servers
  - srm v1
  - srm v2
- RFIO server
- DPM-enabled GridFTP server
  - Normal GridFTP server slightly modified for the DPM



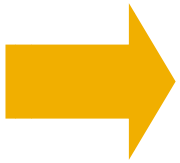
  - Can be used in place of the normal GridFTP server
  - **But** normal GridFTP server cannot be used for the DPM

DPM in a nutshell

- Description
- **Installation**
- DPM as a service
- Troubleshooting
- Documentation / support

Before installing

- **5 questions before installation:**
  - **For each VO, what is the expected load?**
    - Does the DPM need to be installed on a separate machine ?



Yes, this is recommended !

  - **How many disk servers do I need ?**
    - Disk servers can easily be added or removed later
  - **Which operating system ?**
  - **Which file system type ?**
  - **At my site, can I open ports:**
    - **5010 (Name Server)**
    - **5015 (DPM server)**
    - **8443 (srmv1)**
    - **8444 (srmv2)**
    - **5001 (rfio)**
    - **20000-25000 (rfio data port)**
    - **2811 (DPM GridFTP control port)**
    - **20000-25000 (DPM GridFTP data port)**

Operating System

- RPMs provided for:
  - SLC3
  - Itanium
  - Mac OS X
- DPM should be able to build on:
  - SLC4
  - Other 64 bit platforms

File Systems

- **Dedicated file systems**
  - **DPM cannot work properly if each file systems is not dedicated:**
    - Reported space will be incorrect
    - Thus, space reservation, garbage collection, etc. will not work as expected

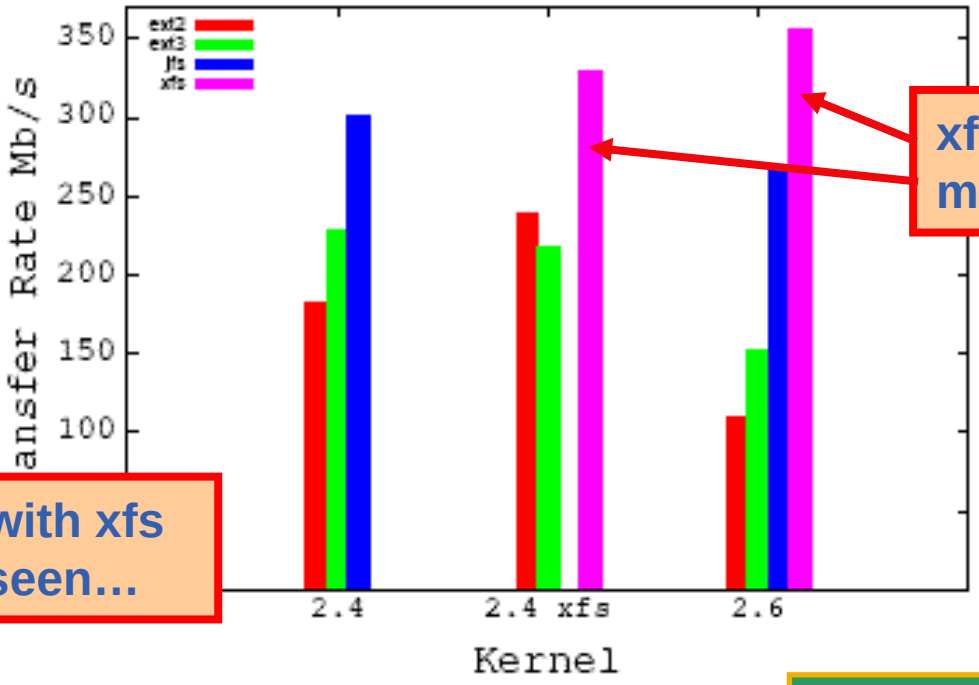
- **Supported file systems ?**
  - **ext2: not recommend**
  - **ext3, xfs: OK**
  - **reiserfs: ?**

- **NFS mounted file system ?**
  - **Problems seen:**
    - DPM sometimes hangs
    - File corruption

Which File System type ?

GridPP
UK Computing for Particle Physics

DPM: $N_f = N_s = 5$



xfs seems to be more performant

But, problems with xfs and kernel 2.6 seen…

From Greig A. Cowan (HEPIX 2006)

Small # failed file transfers for:

- ext2,3 with 2.6 kernel and jfs with vanilla 2.4 kernel.

Greig A Cowan HEPIX April 2006

ScotGrid

YAIM

- **YAIM configuration :**

Site-info.def

$MY_DOMAIN	$DPM_DB_PASSWORD
"$DPM_HOST:/data01 disk1.$MY_DOMAIN:/data02"	B_HOS
$DPM_HOST	**$DPMFSIZE**
$DPMPOOL	$SE_LIST
$DPM_FILESYSTEMS	**$SE_ARCH**
$DPM_DB_USER	$VOS

Default amount of space reserved for a file. Check VO typical file usage.

"multidisk"

  - VO_<myVO>_DEFAULT **not used for DPM**
  - VO_<myVO>_STORAGE_DIR **not used for DPM**

- **On the DPM server:**
  - ./install_node site-info.def glite-SE_dpm_mysql
  - ./configure_node site-info.def glite-SE_dpm_mysql

- **On each disk server (except DPM server):**
  - ./install_node site-info.def glite-SE_dpm_disk
  - ./configure_node site-info.def glite-SE_dpm_disk

YAIM installation

- **YAIM installs**
  - **On $DPM_HOST:**
    - DPM server
    - DPM Name Server
    - SRM servers (srmv1 & srmv2)
  - **On each disk server specified in $DPM_FILESYSTEMS:**
    - RFIO server
    - DPM-enabled GridFTP server
- **YAIM creates**
  - **One pool including all file systems specified**
- **Only configuration files**
  - /opt/lcg/etc/DPMCONFIG : DPNS and DPM DB connection string
  - /etc/shift.conf : different servers/disk servers TRUST rules

How does it look ?

- **DPM Pool**

  export DPM_HOST=dpm01.cern.ch
  dpm-qryconf

Cleanup starts when 5% space left

Cleanup stops when 15% space left

```
POOL TutorialPool DEFSIZE 200.00M GC_START_THRESH 5
 GC_STOP_THRESH 15 DEFPINTIME 0 PUT_RETENP 86400
 FSS_POLICY maxfreespace GC_POLICY lru RS_POLICY fifo GID 0

                CAPACITY 52.77G FREE 43.79G ( 83.0%)
 data CAPACITY 17.27G FREE 14.24G ( 82.4%)
 disk01.cern.ch /storage CAPACITY 17.75G FREE 14.78G ( 83.3%)
 disk01.cern.ch /data CAPACITY 17.75G FREE 14.78G ( 83.3%)
```

- **Name Server**

  export DPNS_HOST=dpm01.cern.ch
  dpns-ls -l /dpm/cern.ch/home

Your domain

Supported VOs

Virtual gids

```
drwxrwxr-x   0 root     104            0 May 22 14:43 atlas
drwxrwxr-x   2 root     2688           0 May 22 14:53 dteam
drwxrwxr-x   0 root     105            0 May 22 14:43 geant4
drwxrwxr-x   0 root     103            0 May 22 14:43 lhcb
```

Test installation (examples)

- **lcg_utils (from a UI)**
  - **If DPM not in site BDII yet**
    - export LCG_GFAL_INFOSYS=dpm01.cern.ch:2135
    - lcg-cr –v --vo dteam –d dpm01.cern.ch file:/path/to/file
  - **Otherwise**
    - export LCG_GFAL_INFOSYS=my_bdii.cern.ch:2170
    - lcg-cr –v --vo dteam –d dpm01.cern.ch file:/path/to/file

- **rfio (from a UI)**
  - export LCG_RFIO_TYPE=dpm
  - export DPNS_HOST=dpm01.cern.ch
  - export DPM_HOST=dpm01.cern.ch
  - rfdir /dpm/cern.ch/home/dteam/myVO

Test installation (examples)

- **srmcp (from a UI)**
  - Note: doesn't work between DPM and other Storage Element
  - /opt/d-cache/srm/bin/srmcp file:////path/to/file srm://dpm01.cern.ch:8443/dpm/cern.ch/home/dteam/my_dir/my_file

- **globus-url-copy (from a UI)**
  - lcg-gt srm://dpm01.cern.ch/dpm/cern.ch/home/dteam/generated/2006-06-14/file5b517db2-30dc-4df0-9ac9-0e30433e10da gsiftp

    TURL returned:

    gsiftp://disk01.cern.ch/disk01.cern.ch:/data/dteam/2006-06-14/file5b517db2-30dc-4df0-9ac9-0e30433e10da.1.0
  - globus-url-copy file:/path/to/file gsiftp://disk01.cern.ch/disk01.cern.ch:/data/dteam/2006-06-14/file5b517db2-30dc-4df0-9ac9-0e30433e10da.1.0

DPM in a nutshell

- **Description**
- **Installation**
- **DPM as a service**
  - Administration
  - Monitoring
- **Troubleshooting**
- **Documentation / support**

Classic SE migration

- **Why migrating your Classic SE ?**
  - **Users: DPM supports SRM**
  - **Admins: DPM provides**
    - **Logical Namespace**
    - **Manageable storage (easy to add/remove disk space)**
    - **Automatic garbage collection**

- **Migrating to what ?**
  - **DPM**
    - Only metadata operation (Name Server population)
    - No physical file move needed

- **Script to run**

  ./migration classicSE_hostname classicSE_dir
  DPM_hostname DPM_dir DPM_pool

- **More details, see**
  https://twiki.cern.ch/twiki/bin/view/LCG/ClassicSeToDpm

Admin Tasks (1)

- **Modify configuration**

  dpm-modifypool --poolname myPool --gc_start_thresh 5 --gc_stop_thresh 10

  dpm-modifyfs --server disk01.cern.ch --fs /data --st RDONLY

- **Add a disk server**
  - ./install_node site-info.def glite-SE_dpm_disk
  - ./configure_node site-info.def glite-SE_dpm_disk
  - **Add it to TRUST rules in DPM server** /etc/shift.conf **file**

- **Remove a disk server**
  - dpm-drain --server disk02.cern.ch --fs /data
  - dpm-rmfs --server disk02.cern.ch --fs /data
  - **Remove it from TRUST rules in DPM server** /etc/shift.conf **file**

Admin Tasks (2)

- **Load balancing**
  - **DPM automatically round robins between file systems**

- **Example**
  - **disk01: 1TB file system**
  - **disk02: very fast, 5TB file system**

  - **Solution 1:** **one file system per disk server**
    - A file will be stored on either disk, equally, if space left
  - **Solution 2:** **one file system on disk01**
    two file systems on disk02
    - A file will more often end up on disk02, which is what you want

Admin Tasks (3)

- **Replicate**
  - **Users and admins can do it**
  - **Useful if file often accessed**

    dpm-replicate --s P /dpm/cern.ch/home/dteam/my_file

  - **Note: as root, only in DPM > 1.5.7**

- **Drain**
  - **Set pool/file system as READ-ONLY**
  - **Let operations finish**
  - **Replicates to another pool/file system**
    - Pinned files
    - Permanent files

    Empty pool / file system can then be removed physically

    dpm-drain --poolname myPool
    dpm-drain --server disk01.cern.ch
    dpm-drain --server disk01.cern.ch --fs /data

  - **Note: only in DPM > 1.5.7**

Admin Tasks (4)

- **Remove Name Server entries for lost files**
  - **Use rfrm (as user or root)**
    - export DPNS_HOST=dpm01.cern.ch
    - export DPM_HOST=dpm01.cern.ch
    - rfrm /dpm/cern.ch/home/dteam/my_dir/my_file
  - **Removes**
    - Files on disk
    - Replicas and Logical Files Names in DPM Name Server
  - **Note: as root, only fully works in DPM > 1.5.7**

Quotas ?

- **Dedicate a Pool to a VO**
  - **By default, pool can be used by all supported groups/VOs**
  - **But, pool can be restricted:**

    dpm-modifypool --poolname VOpool --def_filesize 200M --gid VO_gid

    dpm-modifypool --poolname VOpool --def_filesize 200M --group VO_group_name
  - **Hard limit:**
    - What to do if pool is full ?

- **Quotas**
  - **Maximum space allowed per group/VO**
  - **Softer limit:**
    - Not bind to pools / physical file systems

- **Dynamic space reservation**
  - **Defined by user on request**
  - **Maximum value configurable**
  - **Admin maximum value could be higher than for standard user**

Not implemented yet

Monitoring

- **What to monitor ?**
  - **All processes running ?**
  - **All threads busy ? DPNS (see LFC)**
  - **All threads busy ? DPM**
    - 20 threads for fast operations
    - 20 threads for slow operations (get, put, copy)
    - Note: there are also
      - *1 garbage collector thread per disk pool defined*
      - *1 thread that removes the expired put request*

    Monitor separately
  - **RFIO/GridFTP**
    - Too many transfers to the same file system ?
  - **Number of requests (in dpm_db) per week/month ?**

- **GridView**
  - **Monitoring tool for GridFTP transfers**
  - **On each disk server, enable GridView**
    - Edit /opt/lcg/etc/lcg-mon-gridftp.conf
    - Change: LOG_FILE = /var/log/dpm-gsiftp/dpm-gsiftp.log
    - Restart service: service lcg-mon-gridftp restart

Not supported by YAIM yet

In the plan

- **Request DB cleanup script**
  - **Request DB only grows:**
    - Ex: two more rows in the dpm_db database, when opening a file
  - **Need script to remove the older requests**
  - **Question: how long should the requests be stored ?**
    - 1 week ?
    - 1 month ? (holidays…)
- **Implement quotas**
- **Implement dynamic space reservation**
- **Limitation of transfers to same file system**
- **Automated database backups for MySQL**
  - **MySQL replication**
  - **Save/backup data from the slave in a safe place (Tier1 ?)**

DPM in a nutshell

- Description
- Installation
- DPM as a service
- **Troubleshooting**
- Documentation / support

Log Files

- **DPM server**
  - /var/log/dpm/log
- **DPM Name Server**
  - /var/log/dpns/log
- **SRM servers**
  - /var/log/srmv1/log
  - /var/log/srmv2/log
- **RFIO server**
  - /var/log/rfiod/log
- **DPM-enabled GridFTP**
  - /var/log/dpm-gsiftp/gsiftp.log
  - /var/log/dpm-gsiftp/dpm-gsiftp.log

Permissions problem

- **DPM Name Server**
  - Based on same code as LFC
    - virtual uids,gids
    - /opt/lcg/etc/lcgdm-mapfile
      (if grid-proxy-init or simple voms-proxy-init)
    - VOMS support
  - No support for secondary groups yet
  - More details: see LFC tutorial

DPM in a nutshell

- Description
- Installation
- DPM as a service
- Troubleshooting
- Documentation / support

Another problem ?

- **Main LFC/DPM documentation page**
  - https://twiki.cern.ch/twiki/bin/view/LCG/DataManagementTop
- **DPM Admin Guide**
  - https://twiki.cern.ch/twiki/bin/view/LCG/DpmAdminGuide
- **Troubleshooting page**
  - https://twiki.cern.ch/twiki/bin/view/LCG/LfcTroubleshooting

Still no clue ?

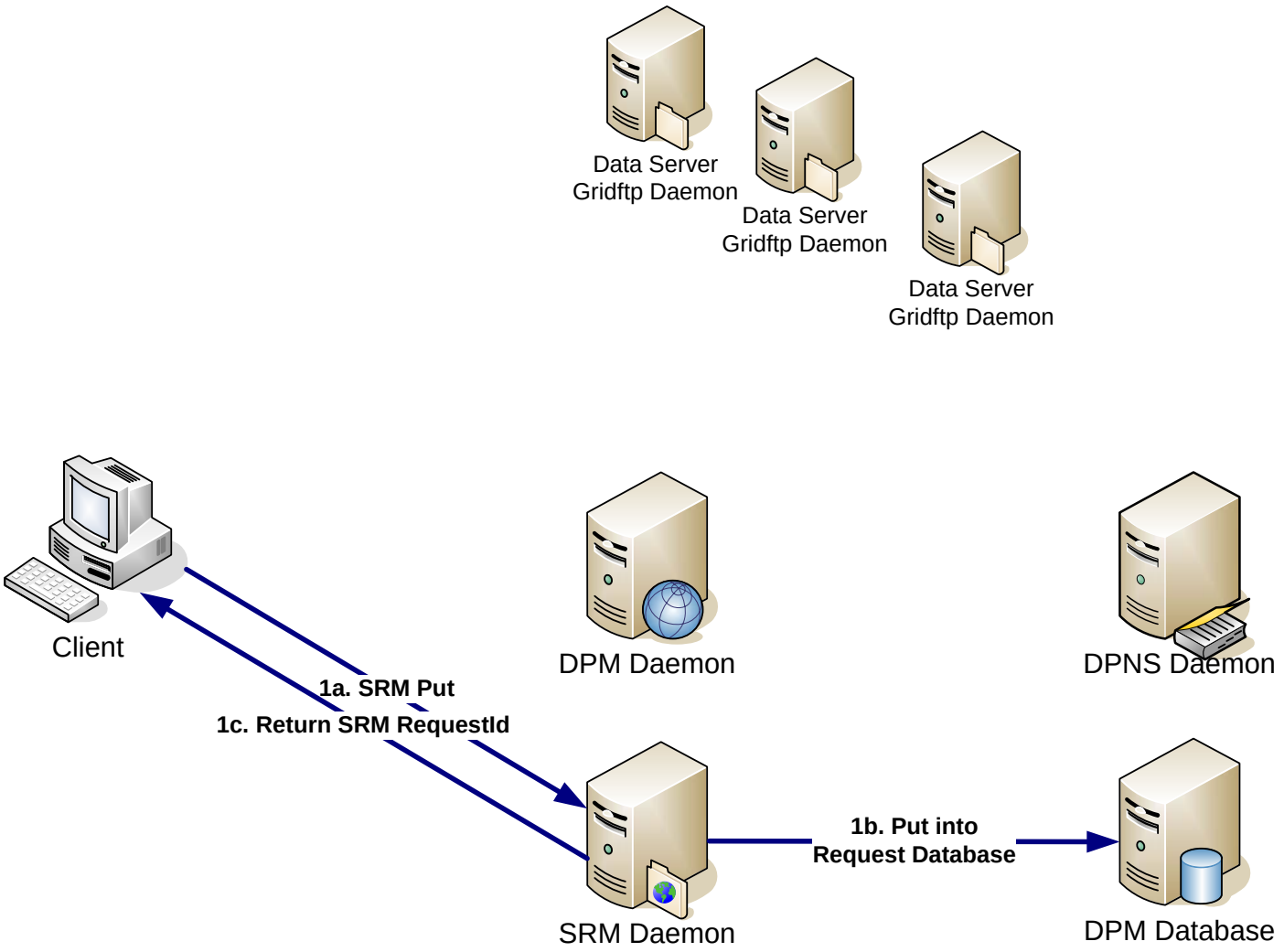
- **Contact GGUS helpdesk@ggus.org**
  - your ROC will help
  - If needed, DPM experts will help

Questions ?

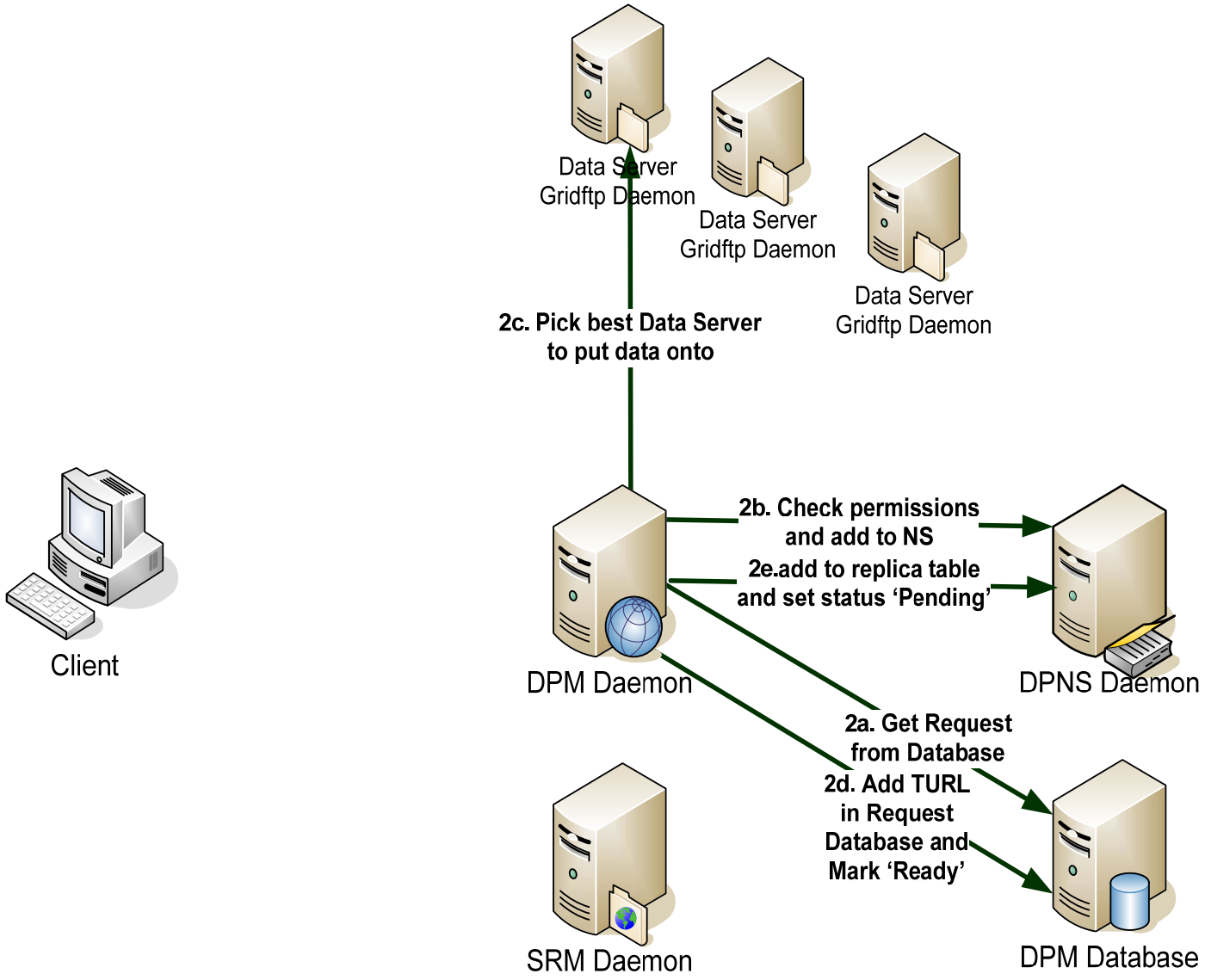
helpdesk@ggus.org

Sophie Lemaitre (Sophie.Lemaitre@cern.ch)
Jean-Philippe Baud (Jean-Philippe.Baud@cern.ch)

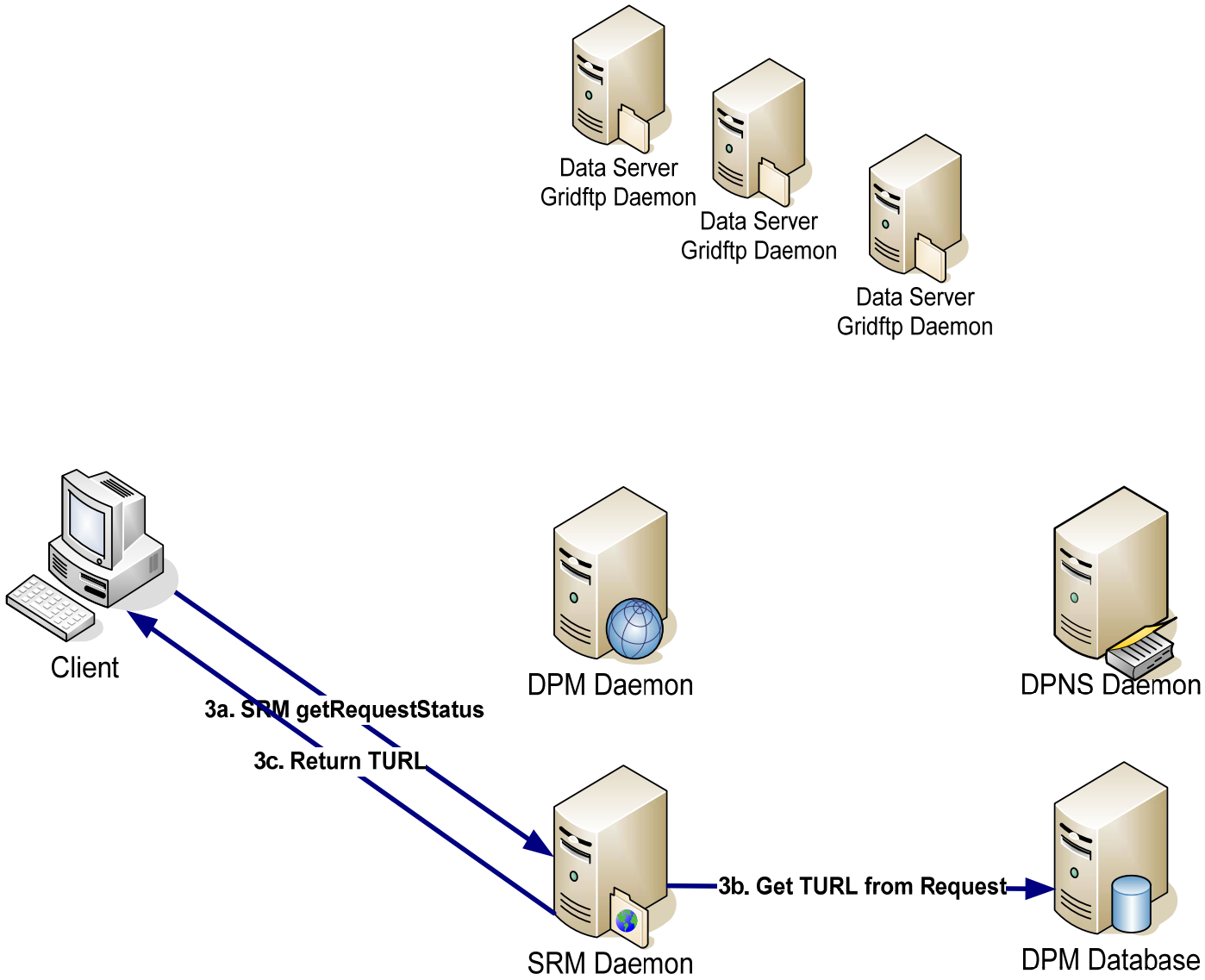
Example: SRM put processing (1)



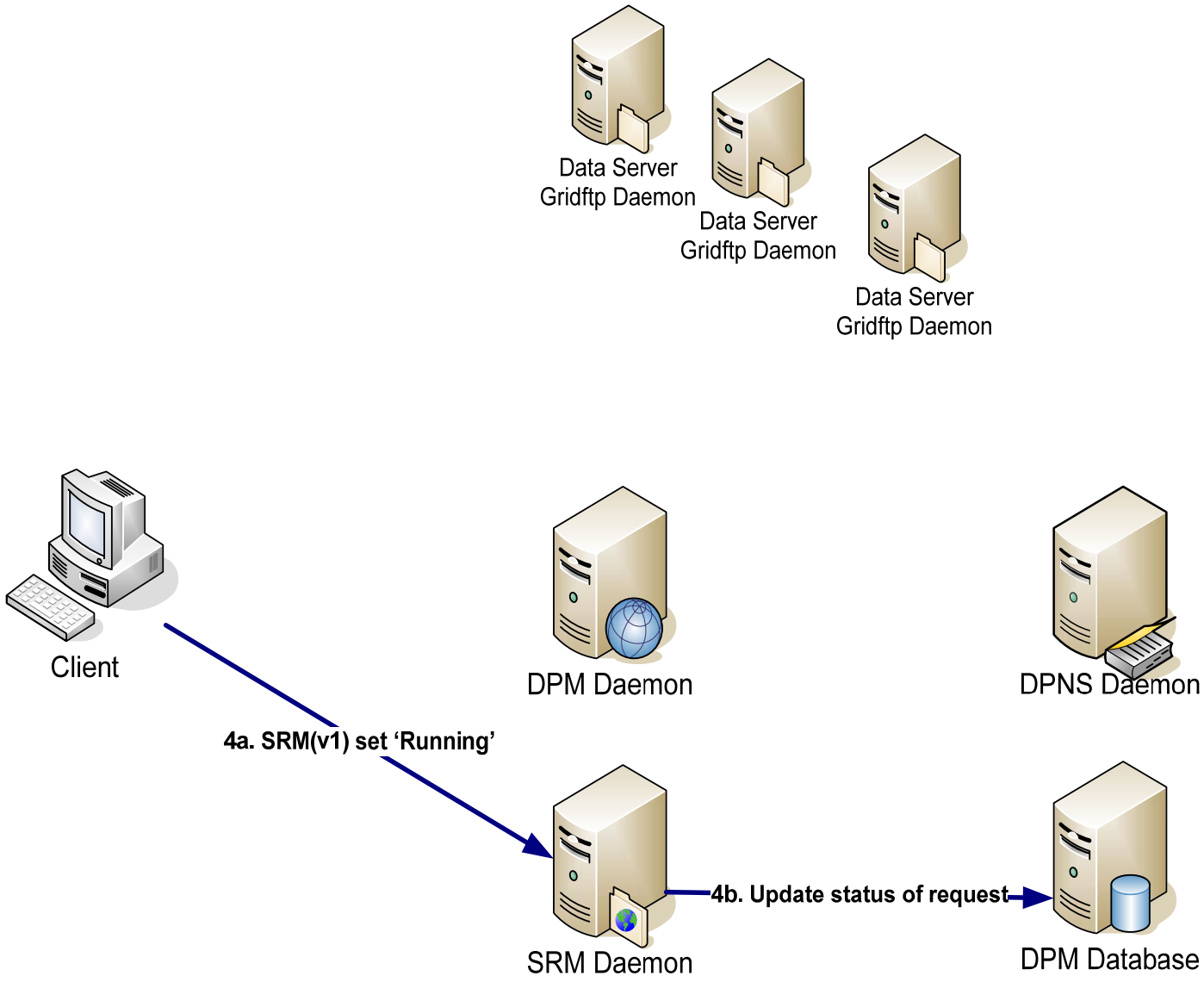
Example: SRM put processing (2)



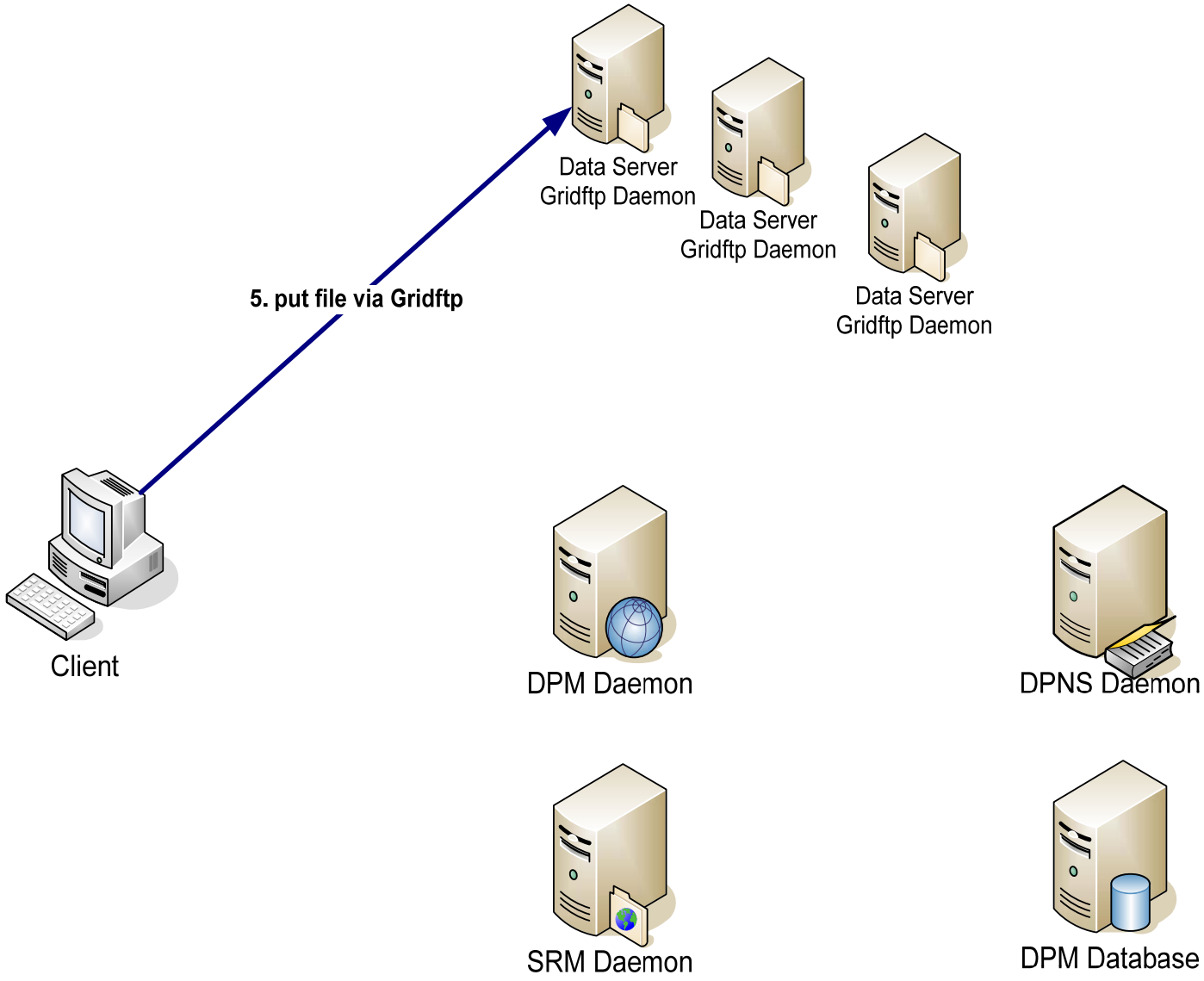
Example: SRM put processing (3)



Example: SRM put processing (4)



Example: SRM put processing (5)



Example: SRM put processing (6)