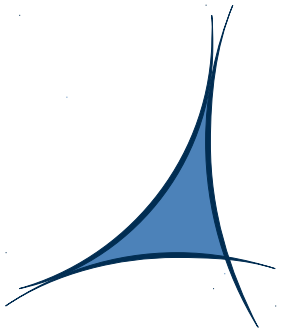


PIC
port d'informació
científica

PIC SRM deployment for a multi-VO based dCache storage

M. Caubet, E. Planas, E. Acción
PIC (Port d'Informació Científica)
Barcelona (Spain)

Institut de Física d'Altes Energies

Centro de Investigaciones Energéticas, Medioambientales y Tecnológicas

PIC
port
d'informació
científica

Outline

- PIC & dCache
- Old Single-SRM Setup
- Motivations
- New Multi-SRM Setup
- Future plans

PIC
port
d'informació
científica

PIC & dCache

- PIC (Port d'Informació Científica) is a collaboration between CIEMAT & IFAE

PIC
port d'informació
científica

Institut de Física
d'Altes Energies

GOBIERNO DE ESPAÑA
MINISTERIO DE ECONOMÍA Y COMPETITIVIDAD
Ciemat
Centro de Investigaciones Energéticas, Medioambientales y Tecnológicas

PIC
port
d'informació
científica

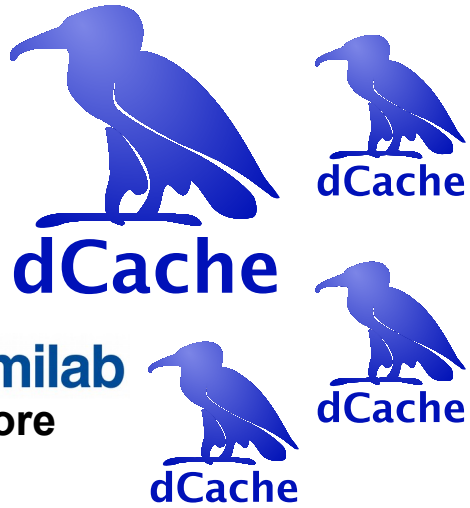
PIC & dCache

- Supported experiments:
  - WLCG Tier1: serving ATLAS, CMS (5.1%) and LHCb (6.5%)
  - ATLAS Tier2: 25% of the Federated Spanish Tier2 (IFIC/IFAE/UAM)
  - ATLAS Tier3 for IFAE
  - MAGIC Datacenter: Reference Data Center for all MAGIC data (international experiment)
  - IEEC/ICE MICE Cosmological Simulations: Tier0
  - PAU survey: Tier0 of PAUcam Data
  - DES analysis and scientific exploitation
  - EUCLID: One of Ten Science Data Centers (expected a storage growth on the PB scale) (european + U.S. american)
  - Ramping up CTA data acquisition from telescope prototype (Inter experiment)
- All experiments are managed with the same dCache system

PIC
port
d'informació
científica

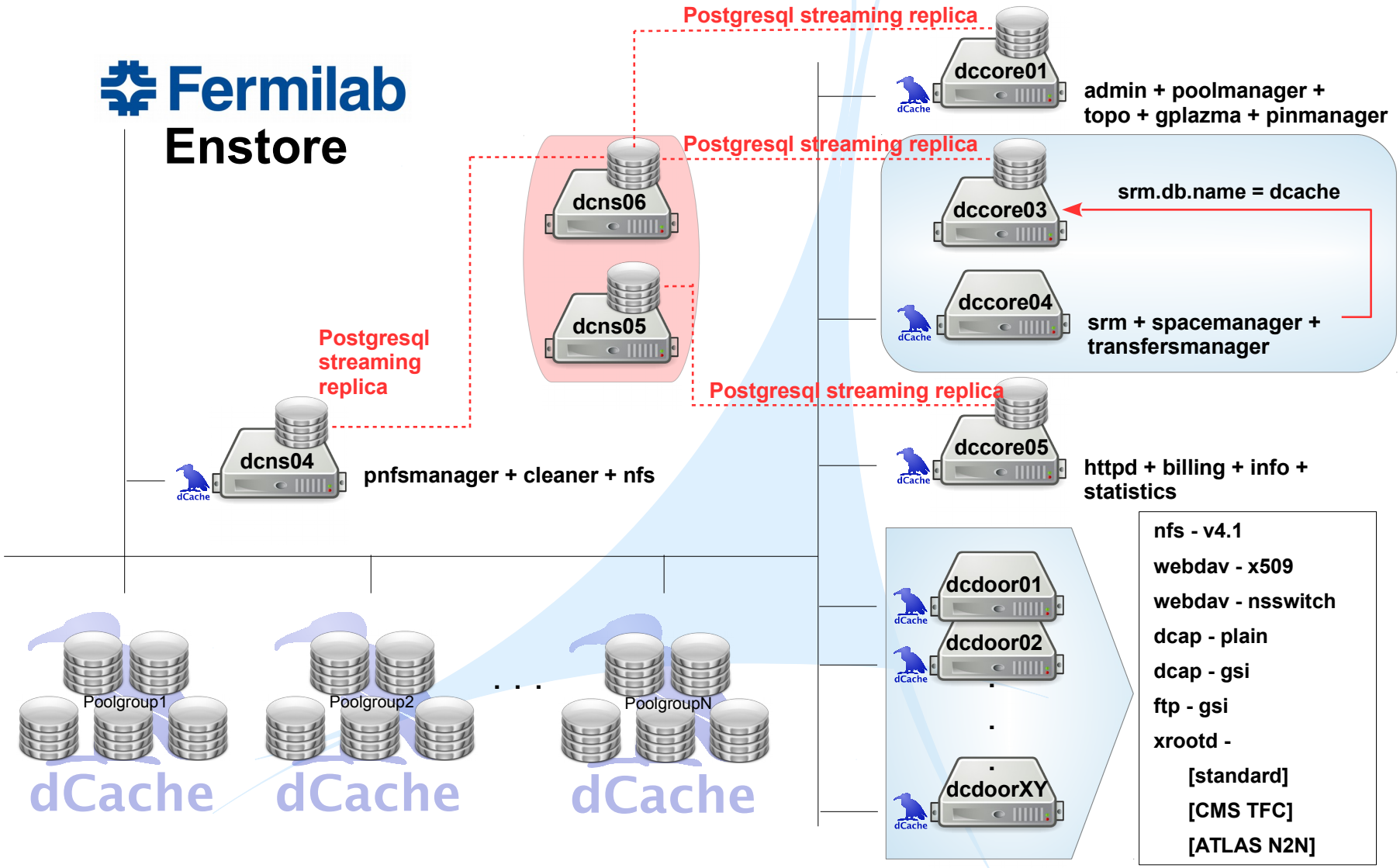
PIC & dCache

- Production + Development + Test
  - dCache 2.13.28-1
    - dCache system with ~6PB of disk
  - Fermilab Enstore 5.1.1-1
    - Enstore system ~12PB
    - Tapes: LTO4, LTO5, T10KC, T10KD
  - PostgreSQL:
    - 9.2.10-1+ in Enstore
    - 9.3.11-1+ Chimera PinManager & Billing
    - 9.5.0-2+ new installations (srm, spacemanager, transfermanagers)
- CMS Middleware Readiness
  - dCache 2.14.13-1
  - We will upgrade to dCache 2.15 soon



PIC
port
d'informació
científica

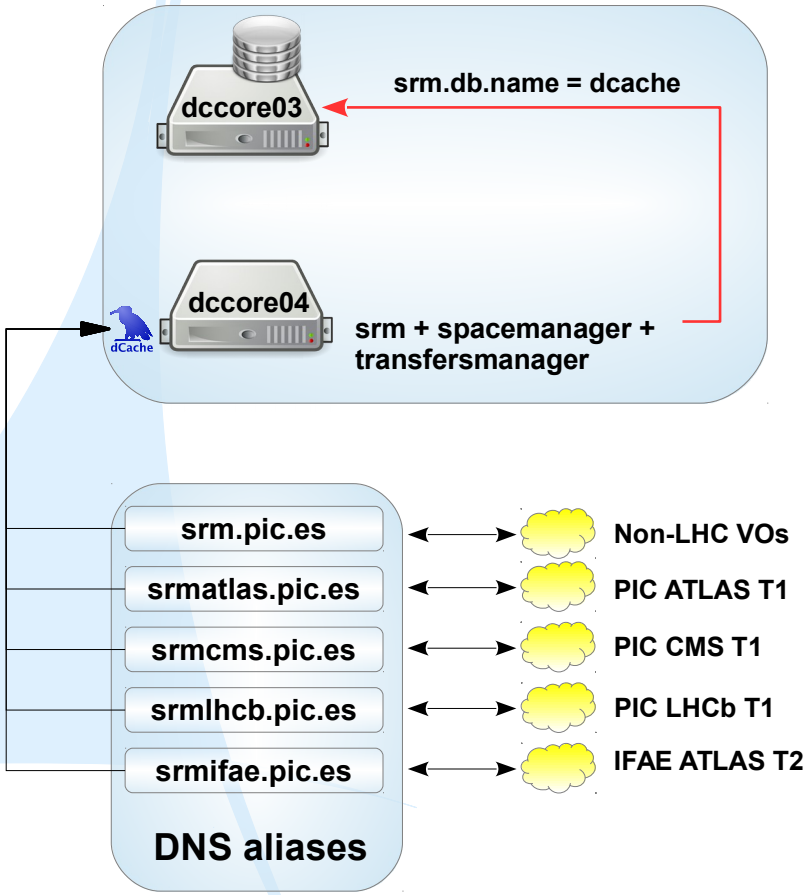
Old Single-SRM Setup



PIC
port
d'informació
científica

Old Single-SRM Setup

- Unique DB with srm, spacemanager & transfersmanagers.
- dccore04.pic.es with multiple DNS aliases
  - One for each T1 VO
  - One for IFAE T2
  - One generic for other experiments



PIC
port
d'informació
científica

Old Single-SRM Setup

- Problems:
  - VO overloading SRM will affect other VOs

    Huge values for some parameters in the configuration file → needed to handle load from all VOs with a single SRM!
  - SRM crash → SPF for all experiments
  - Not able to tune up SRM according to the experiment real use
  - No redundancy of the service
  - Non realistic published information. 3 options:
    - `dcache-info-provider` output hack in order to publish all SRM endpoints → method used at PIC
    - To run 5 info providers, 1 for each DNS alias
    - To change the Info Provider → GlueSEUniqueID not SRM end-point anymore (GlueSEUniqueID != GlueServiceUniqueID)

PIC
port
d'informació
científica

Old Single-SRM Setup


```
#!/bin/bash

file=/tmp/info.tmp.$$

dcache-info-provider > $file

rm -f ${file}.out > /dev/null 2>&1 ; cp $file ${file}.out > /dev/null 2>&1

# PIC T1

for i in srm.pic.es srmatlas.pic.es srmcms.pic.es srmlhcb.pic.es; do

       perl -p00e 's/\r?\n //g' $file | sed s/"dccore04.pic.es"/"$i"/g | sed
s/"srm.pic.es"/"$i"/g | sed -r s/"(GlueSE.*Size.*):.*"/"\1: 1"/g >> ${file}.out

Done

# IFAE T2

perl -p00e 's/\r?\n //g' $file | sed s/"dccore04.pic.es"/"srmifae.pic.es"/g | sed
s/"srm.pic.es"/"srmifae.pic.es"/g | sed s/"GlueSiteUniqueID=pic"/"GlueSiteUniqueID=ifae"/g |
sed -r s/"(GlueSE(Total|Used)NearlineSize).*"/"\1: 0"/g | sed s/"GlueSEArchitecture:
tape"/"GlueSEArchitecture: multidisk"/g | sed s/"mds-vo-name=resource"/"mds-vo-
name=local"/g>> ${file}.out

cat ${file}.out

Sleep 120

rm -f ${file}.out ${file} > /dev/null 2>&1
```

dcache-info-provider

PIC
port
d'informació
científica

Old Single-SRM Setup

```
[srm-${host.name}Domain]
dcache.java.memory.heap=8192m
[srm-${host.name}Domain/srm]
srm.db.host=srmdb.pic.es
srm.limits.db.queue=2000
srm.limits.db.threads=8
srm.limits.jetty-connector.acceptors=8
srm.limits.jetty.threads.max=3000
srm.limits.jetty.threads.queued.max=1500
srm.net.host=dccore04.pic.es
srm.net.listen=dccore04.pic.es
srm.net.local-hosts=dccore04.pic.es,srm.pic.es,srmatlas.pic.es,srmcms.pic.es,srmlhcb.pic.es,srmifae.pic.es
srm.persistence.keep-history-period=30
srm.request.get.lifetime=7200000
srm.request.get.max-requests=20000
srm.request.get.max-transfers=15000
srm.request.lifetime=86400000
srm.request.max-by-same-user=1000
srm.request.max-transfers=14000
srm.request.threads=20
```

dCache Layout for SRM

PIC
port
d'informació
científica

Motivations

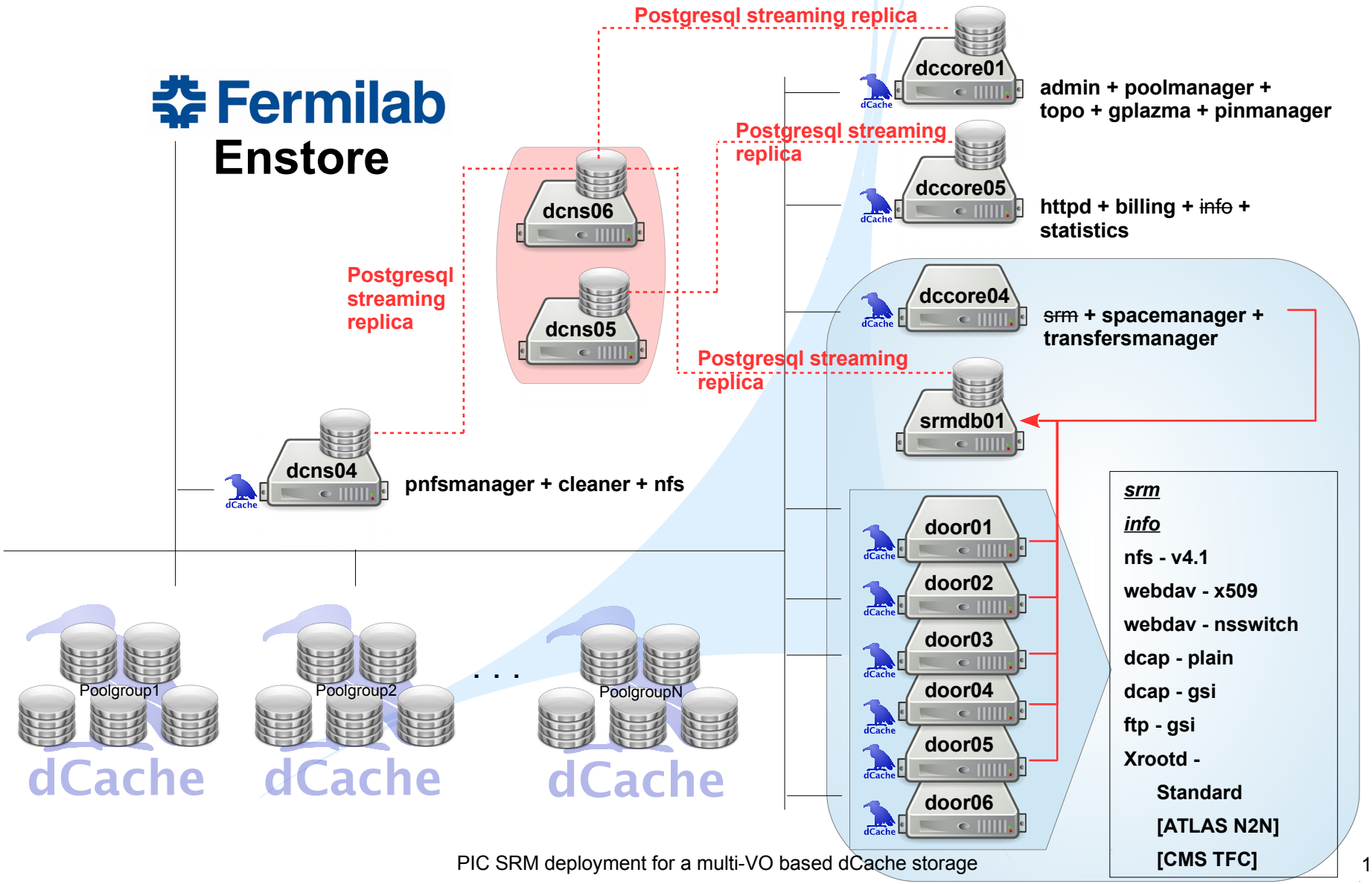
- SRM isolation:
  - SRM overload / SRM crash caused by a VO should not affect other LHC experiments anymore.
  - SRM / storage interventions → need to declare downtimes per VO.
  - Independent monitoring for LHC experiments and reflect reality.
  - Tune up SRM service according to the VO activity.
- Balancing load accross different SRM servers.
- Easy SRM publishment of BDII information and use the default `dcache-info-provider`.
- (Future) Ensure availability of the service → service redundancy

  Not possible yet

➢ Thanks to the dCache developers this is possible!

PIC
port
d'informació
científica

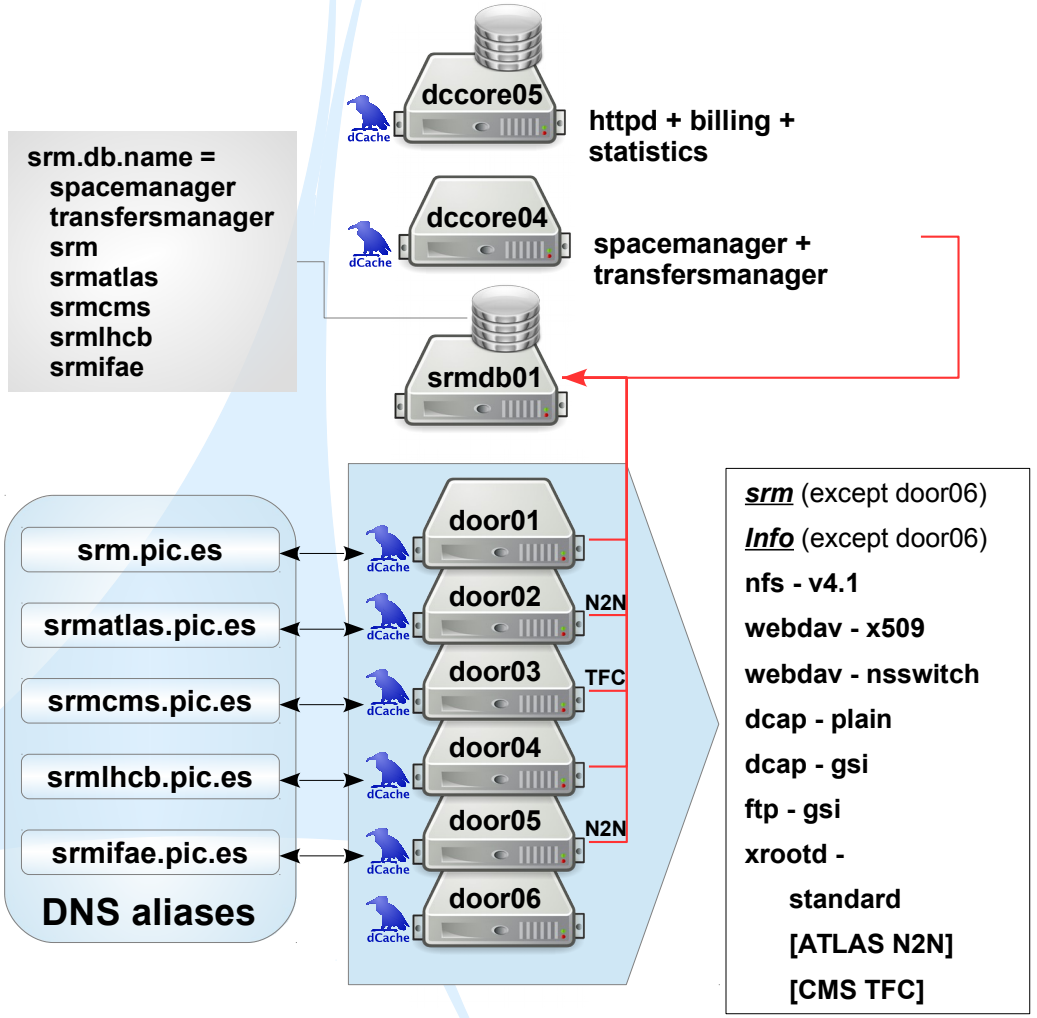
New Multi-SRM Setup



PIC
port
d'informació
científica

New Multi-SRM Setup

- 5 SRM services
  - 1 for each LHC experiment
  - 1 for all non-LHC experiments
- Unique PostgreSQL Instance
  - 1 DB for each SRM
  - 1 DB per spacemanager
  - 1 DB per transfersmanager
- 1 info service per SRM, running in the same host as his SRM
  - 5 info services
  - info service not running in dccore05 anymore
- We have gathered VO-specific services (i.e. XRootD, WebDAV) in order to run them all in the same host.



PIC
port
d'informació
científica

New Multi-SRM Setup

- Benefits:
  - Finally we have SRM isolation for LHC experiments!
    - SRM overload will not affect other LHC experiments anymore
    - SRM crash will affect only 1 LHC VO
    - We are now able to declare downtimes per VO
    - Now we are able to customize each SRM
  - SRM load is being balanced accross different SRM servers and we can monitor them independently.
  - We can easily publish SRM BDII information for each VO:
    - Nevertheless we still need a small hack because `srm.net.listen` is not working → a fix was provided last week (dCache 2.13.29)

```
#!/bin/bash
fqdn=$(hostname)
if [ "${fqdn}" == "door01.pic.es" ]; then aliasname="srm.pic.es"; fi
if [ "${fqdn}" == "door02.pic.es" ]; then aliasname="srmatlas.pic.es"; fi
if [ "${fqdn}" == "door03.pic.es" ]; then aliasname="srmcms.pic.es"; fi
if [ "${fqdn}" == "door04.pic.es" ]; then aliasname="srmlhcb.pic.es"; fi
if [ "${fqdn}" == "door05.pic.es" ]; then aliasname="srmifae.pic.es"; fi

dcache-info-provider | sed s/"${fqdn}"/"${aliasname}"/g
```

PIC
port
d'informació
científica

Multi-SRM Setup

- Cons:
  - SPF per VO. Can be solved by:
    - Easily moving the service to a different door host (i.e., door06)
    - Adding a second SRM per VO, R.R. Load Balance + H.A.

      Actually service redundancy is not possible
  - Info provider showing the dCache space * N published SRMs
    - We need to publish 1 SE with N SRMs (GlueSEUniqueID not SRM endpoint anymore)

      Paul told that will help us with this if we need it
  - For VOs using BDII, we need to hide the LHC SRMs
    - Use `*.loginbroker.tags` to unpublish protocols
      - But SAM OPS still using BDII
        - New GOCDB tags allow us to unpublish LHC SRMs

PIC
port
d'informació
científica

Future Plans

- Customize & tune up each SRM service depending on the experiment
- Fix dCache Information Provider with correct information about SRMs
- Service redundancy
  - Whenever possible
    - And only if LHC experiments do not discontinue the use of SRM


OBLIGATORIO SEGUIR TODAS
LAS NORMAS DE SEGURIDAD
The LHC Tier1 at PIC

Questions?

PIC
port
d'informació
científica

Backup Slides

PIC ATLAS T1 Info Provider Settings

info-provider.dcache-architecture=tape
info-provider.http.host=dcip.pic.es
info-provider.paths.tape-info=/var/lib/dcache/tape-info.xml
info-provider.se-name=PRODUCTION PIC ATLAS SRM
info-provider.se-unique-id=srmatlas.pic.es
info-provider.site-unique-id=pic

IFAE ATLAS T2 Info Provider Settings

info-provider.dcache-architecture=tape
info-provider.http.host=dcip.pic.es
info-provider.paths.tape-info=/var/lib/dcache/tape-info.xml
info-provider.se-name=PRODUCTION IFAE ATLAS SRM
info-provider.se-unique-id=srmifae.pic.es
info-provider.site-unique-id=ifae

```
<constants>
  <!--+
      |   GlueSiteUniqueID [1.3, 2.0] a unique reference for your site.
      |   This must match the GlueSiteUniqueID defined in other services.
      |
      +-->
  <constant id="SITE-UNIQUE-ID">ifae</constant>
  <!--+
      |   GlueSEUniqueID [1.3, 2.0] your dCache's Unique ID.  Currently,
      |   this MUST be the FQDN of your SRM end-point.
      +-->
  <constant id="SE-UNIQUE-ID">srmifae.pic.es</constant>
```