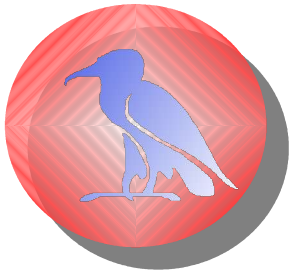
Insight dCache

Patrick

for the dCache people

Preliminaries

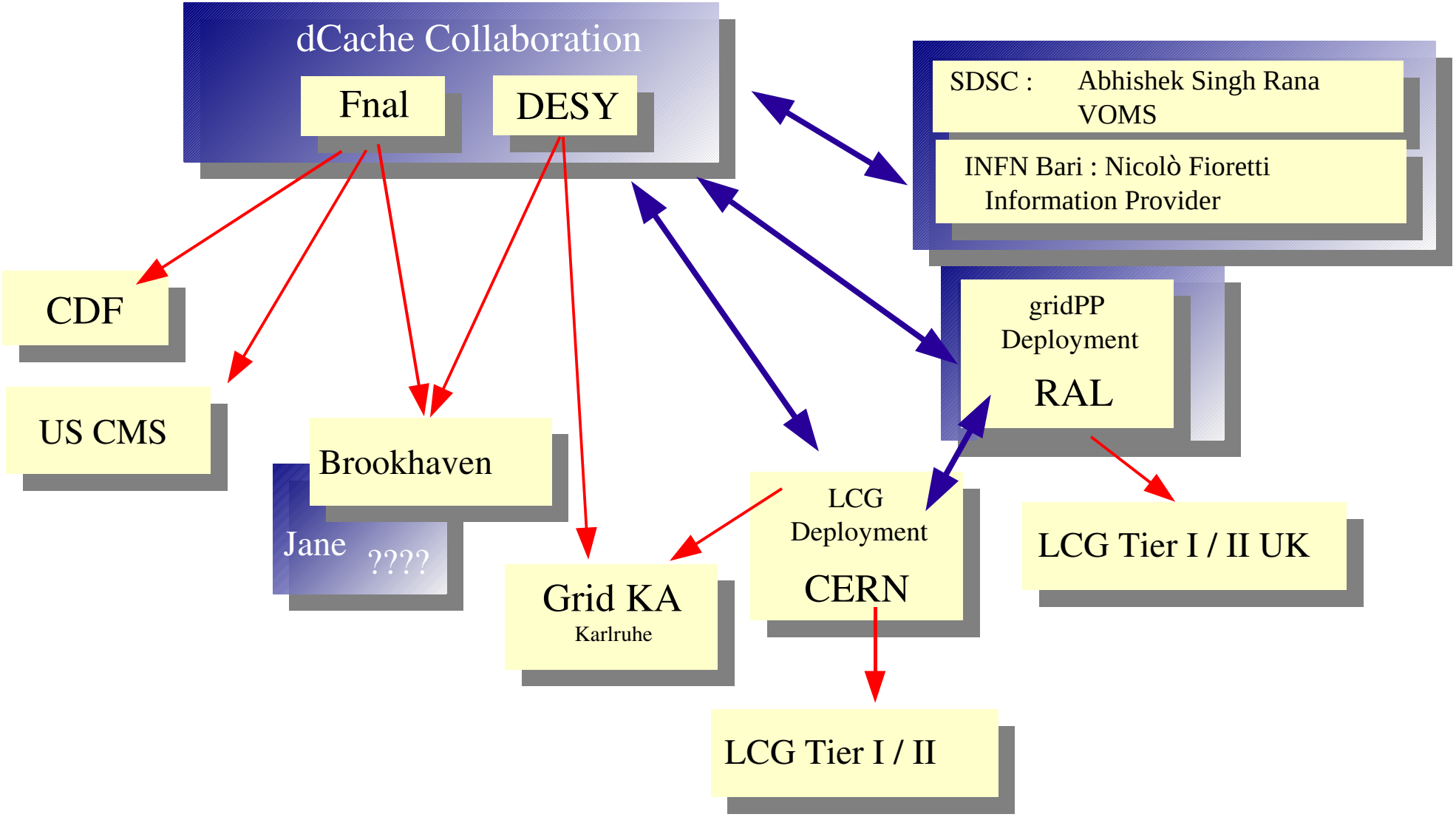


LCG Tier I / II

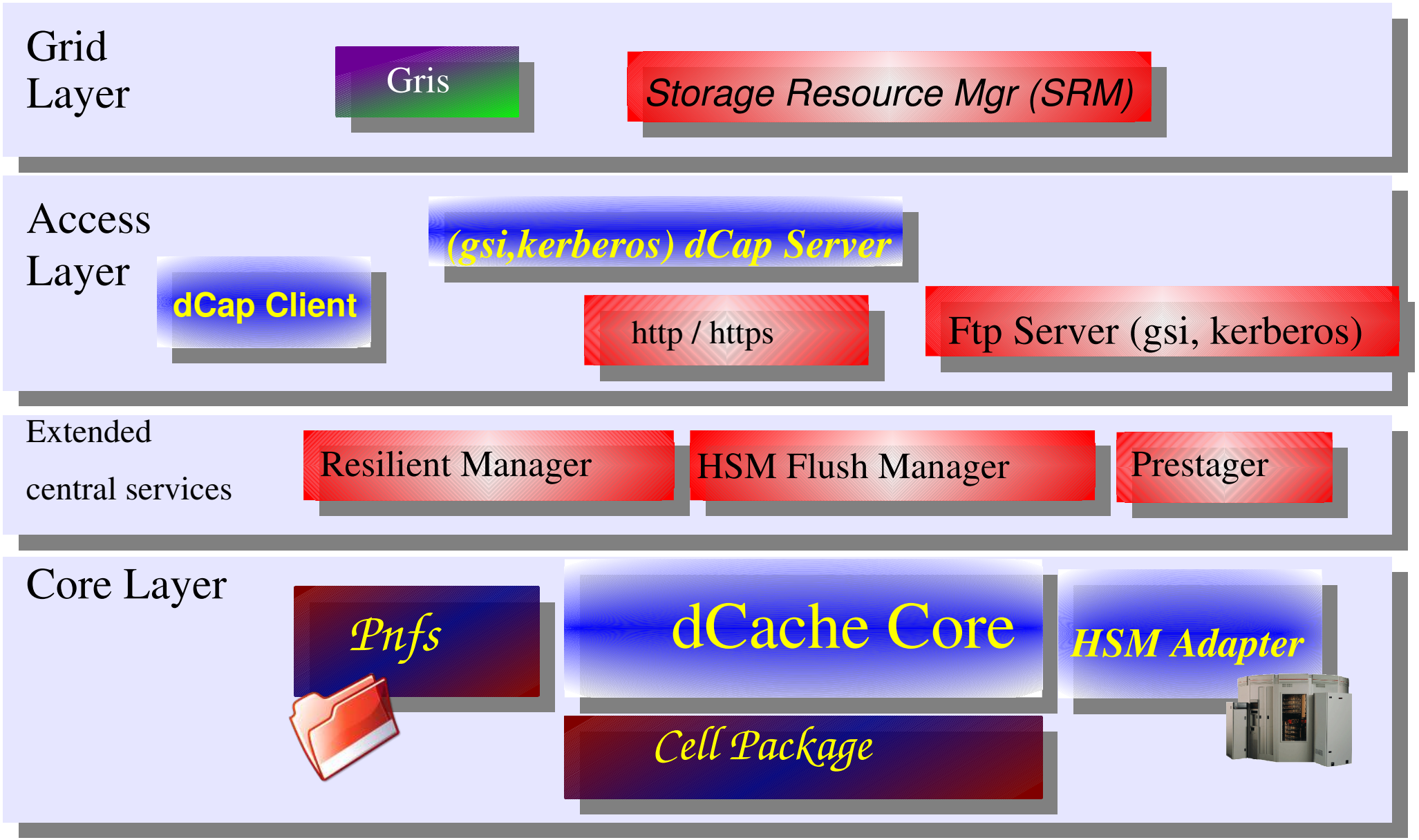
gridPP
Deployment
RAL

dCache Core

SRM


dCache Collaboration
Fnal
DESY
SDSC : Abhishek Singh Rana
VOMS
INFN Bari : Nicolò Fioretti
Information Provider
CDF
US CMS
Brookhaven
Jane ????
gridPP
Deployment
RAL
LCG
Deployment
CERN
Grid KA
Karlruhe
LCG Tier I / II UK
LCG Tier I / II

dCache module topology



http://www.dcache.org/manuals/Book

Lector, Editor and Author : Mathias de Riese

He needs input from all of you -> support@dcache.org

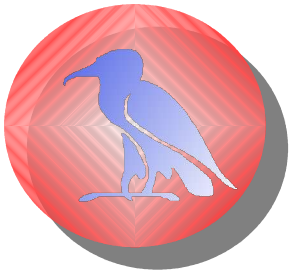
comments and suggestions
you possibly already described some topics

The Cell Environment

dCache Components

Pnfs Filesystem Setup

The PoolManager



Pool Selection Unit

Cost Module

The Resilient Manager

This and That

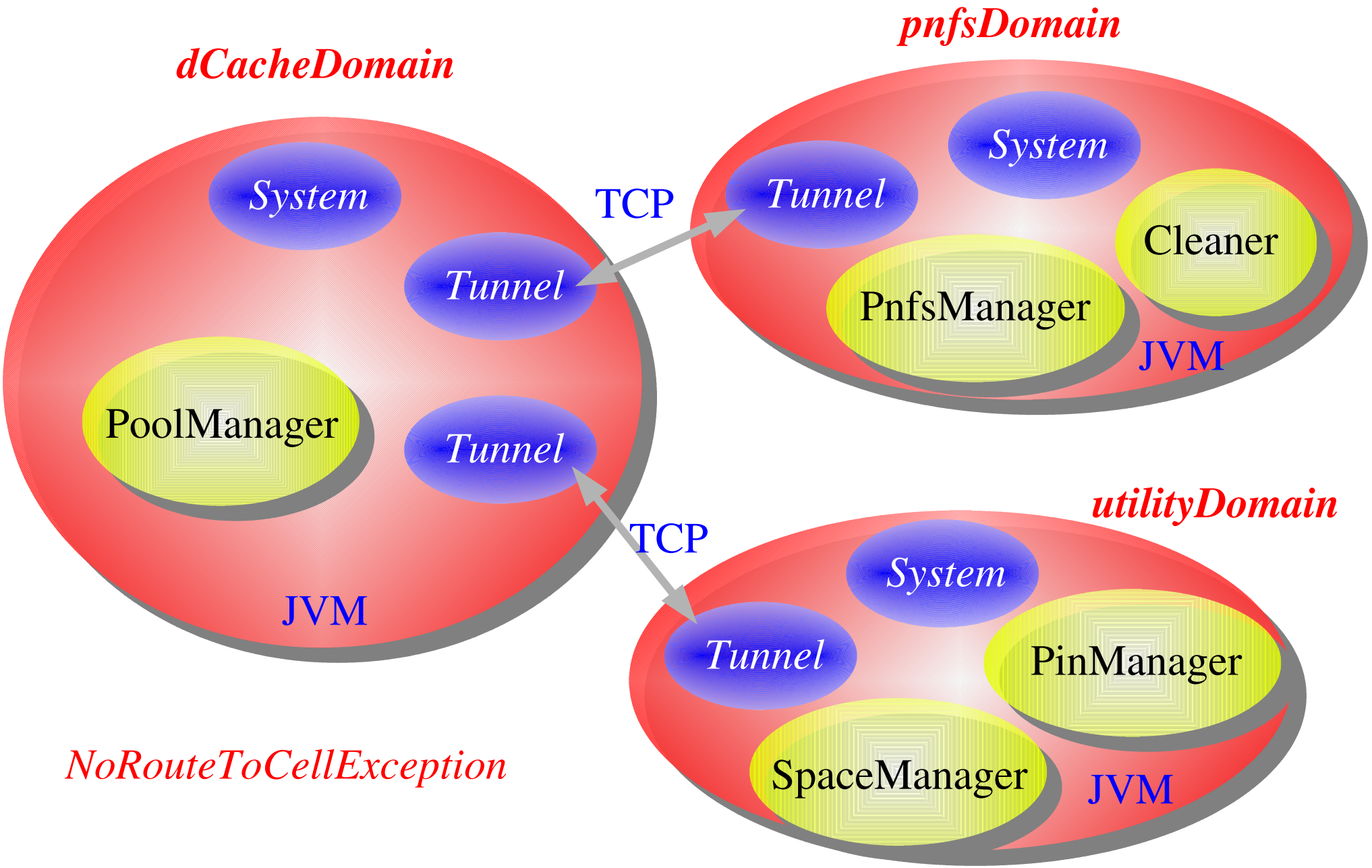
pnfs postgres

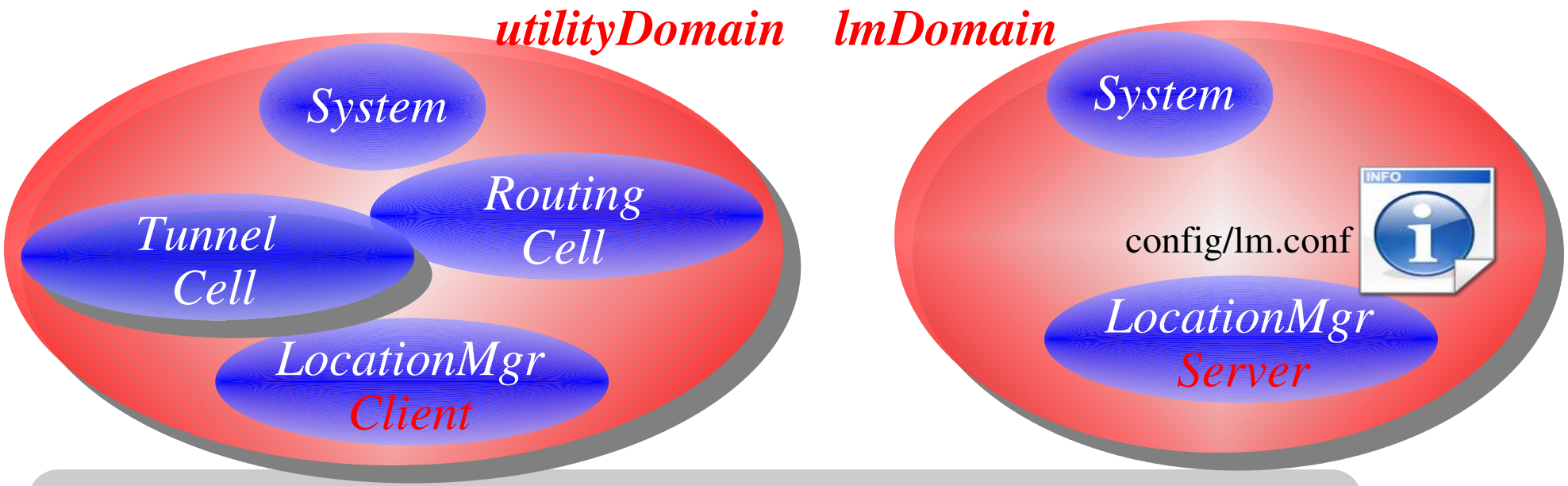
Internal Copy Manager

Coming up ...

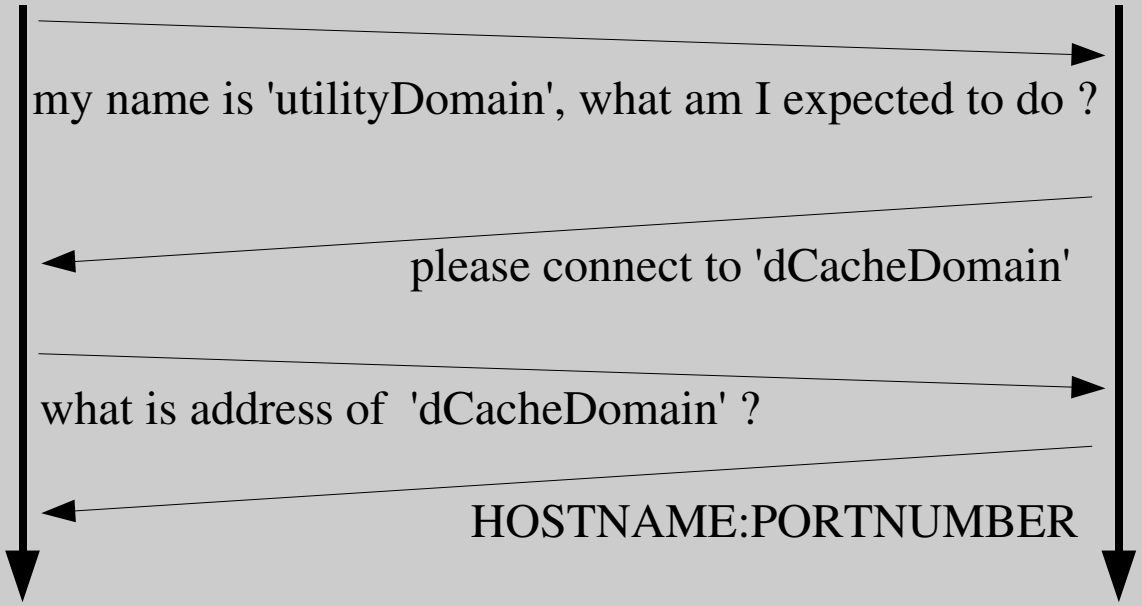
The Cell Environment

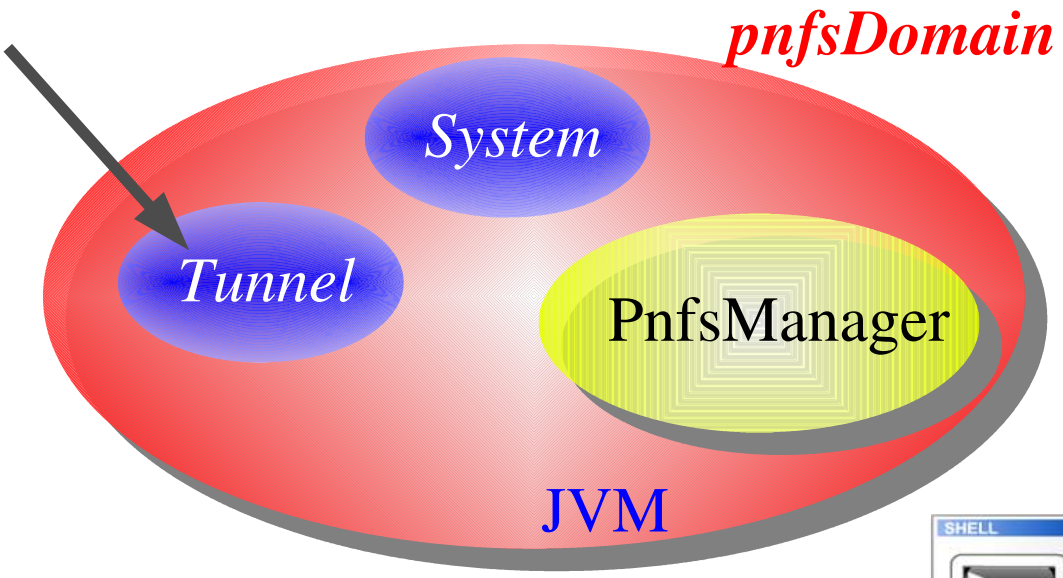
Cell / Domain Game





Location Manager Protocol





../d-cache/jobs/pnfs start

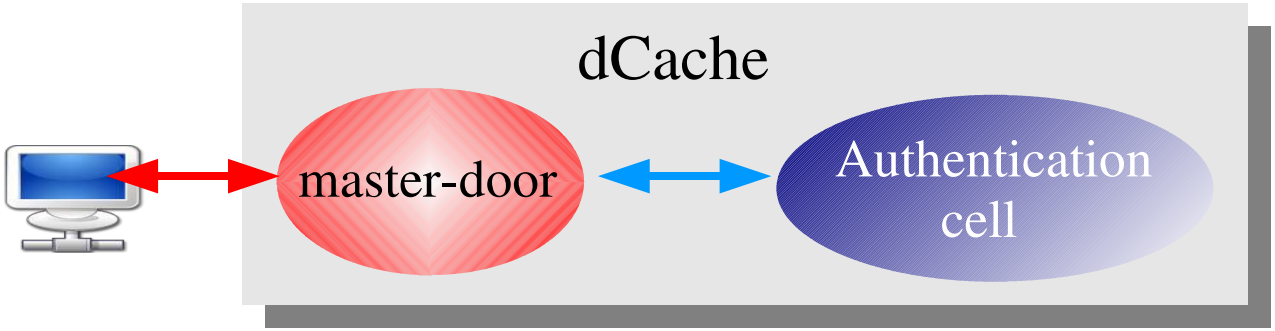
../d-cache/config/pnfs.batch

```
#
create diskCacheV111.namespace.PnfsManagerV3 PnfsManager \
        "diskCacheV111.util.OsmInfoExtractor \
         -threads=4   \
         -enableLargeFileSimulation \
         -default=${defaultPnfsServer} -pnfs=${pnfs} \
         -dbURL=jdbc:postgresql://localhost/dcache   \
         -jdbcDrv=org.postgresql.Driver  \
         -dbUser=srmdcache -dbPass=srmdcache”
#
```

config/dCacheSetup

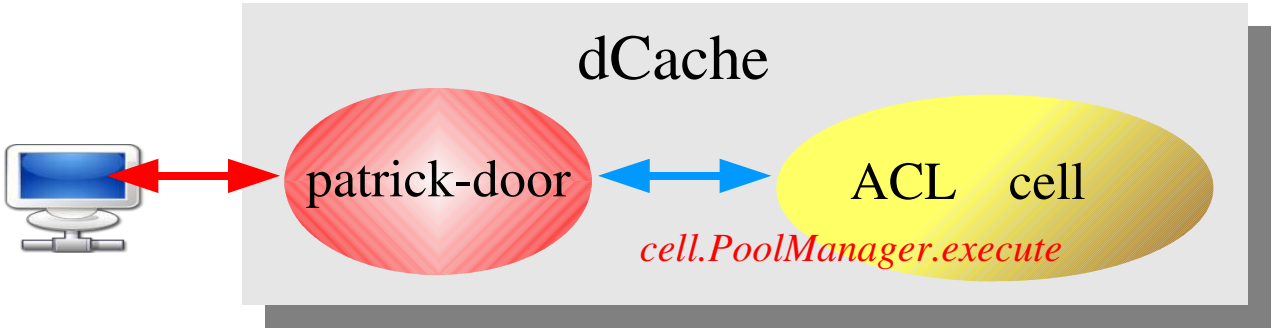
Cells : Login example

```
# ssh -l patrick -c blowfish -p 222223 localhost
Password:
```



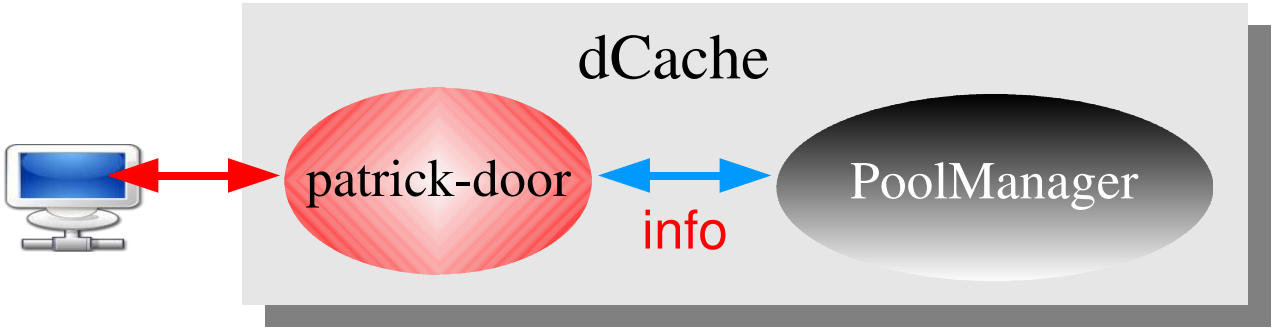

```
dCache Admin (VII) (user=patrick)

[DESY-HH] (local) patrick > cd PoolManager
```




```
dCache Admin (VII) (user=patrick)

[DESY-HH] (local) patrick > cd PoolManager
[DESY-HH] (PoolManager) patrick > info
PoolManager V [$Id: PoolManagerV5.java ...]
  SelectionUnit : $Id: PoolSelec...
     Write Threads : 0
     Read Threads : 0
```



(local) mode commands :

cd <cellName>[@<domainName>] # change destination cell

set timeout # sets timeout (sec) for cell communication

Useful System cell commands : (cd System@dCacheDomain)

ps -f # get all cells running in this Domain

ps -f <cellName> # get more details about the specified cell

route # show message routing table

create <cellClass> <cellName> [<Arguments>]

show message routing table

kill <cellName> # kill cell

Common cell commands (except for System cell) :

help # cell specific help text (all commands)

info # ask cell for brief status report

show pinboard 200 # get the most recent 200 debug messages

dCache Components

Ftp Server (gsi, kerberos)

Pnfs

dCache Core

Cell Package

dCache Core ->

PoolManager

Finds best pool for each request.

Cleaner

Forwards 'rm'requests from pnfs filesystem to pools.

PnfsManager

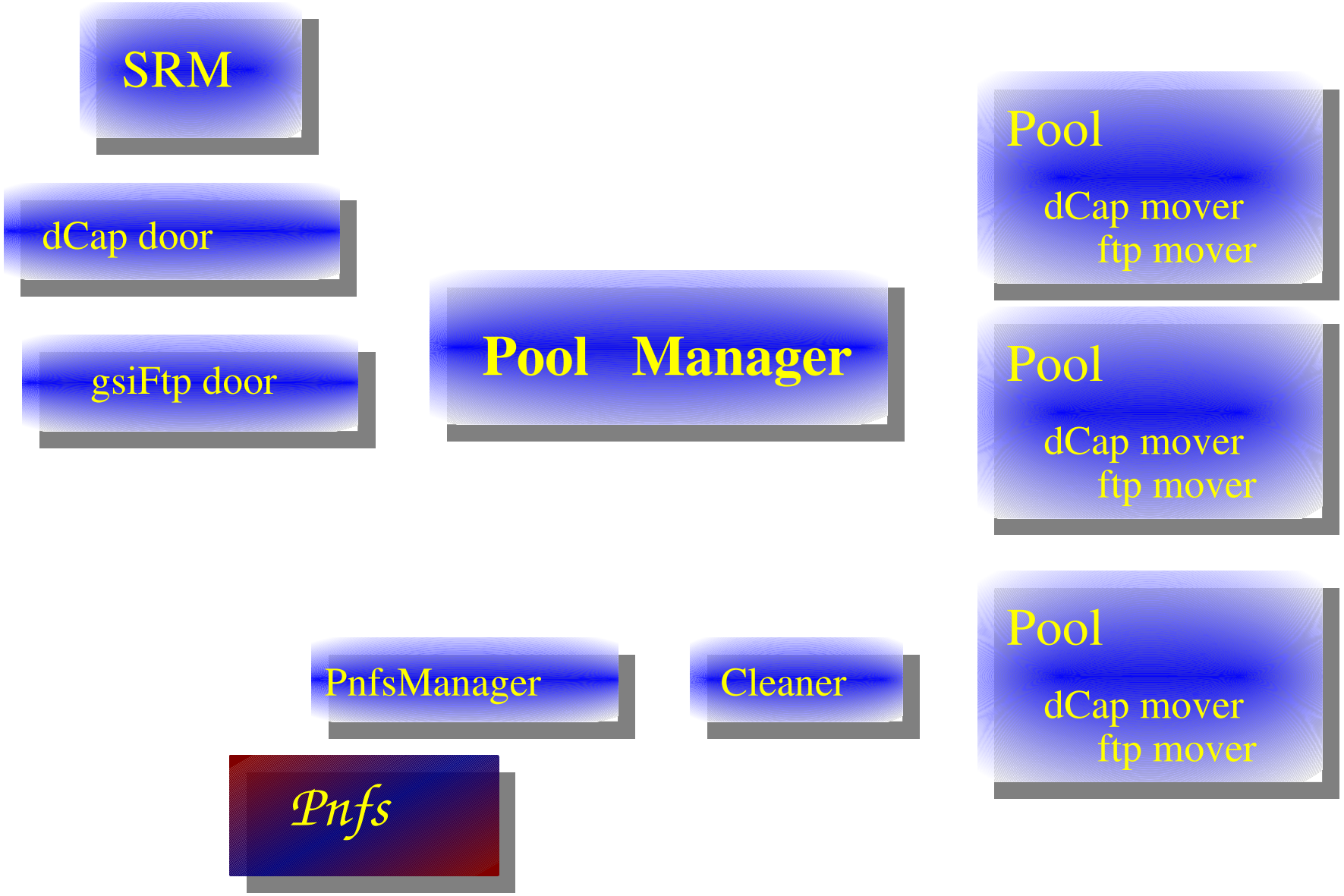
Interface between dCache and filesystem.

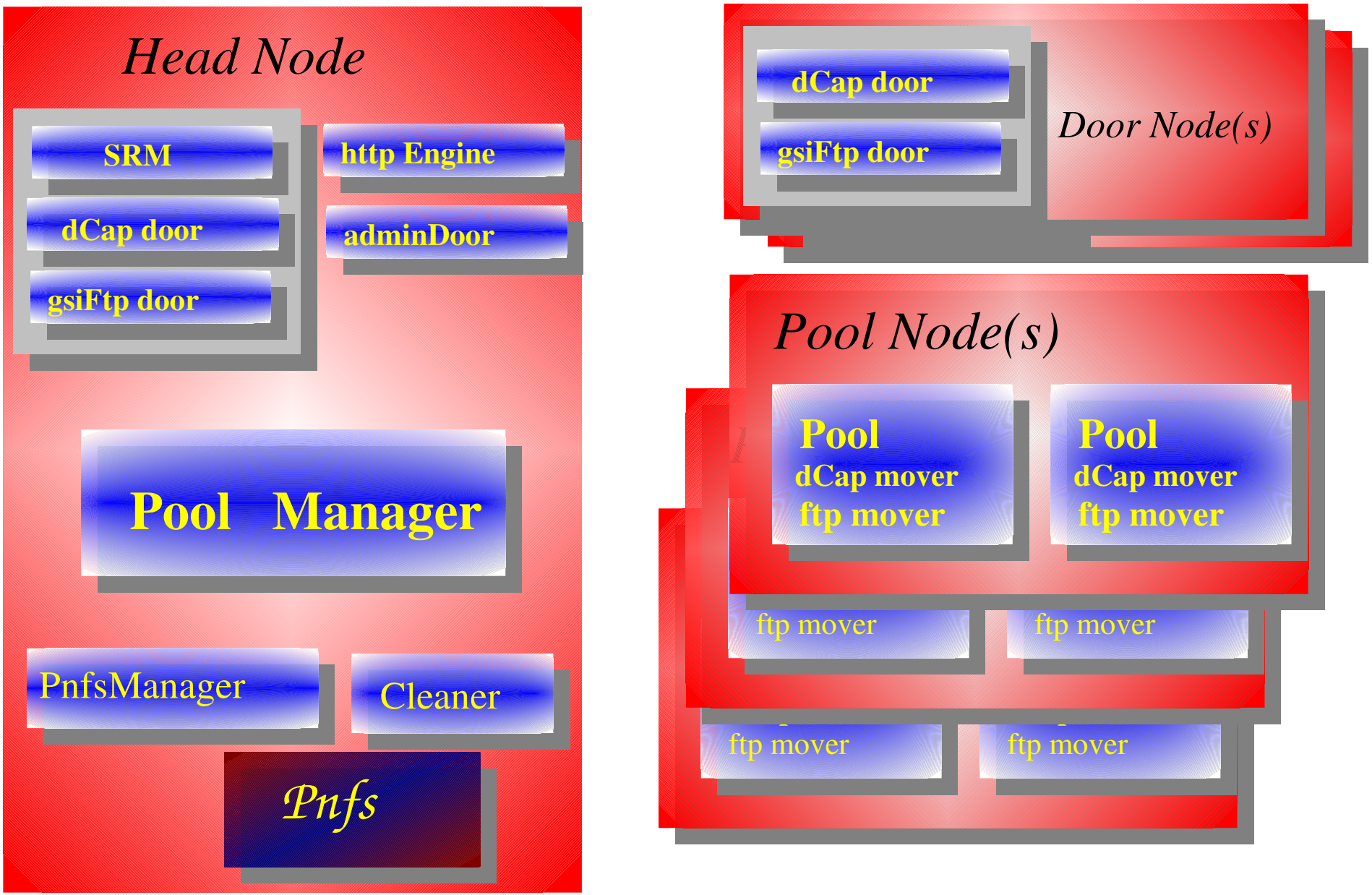
Door (Launcher)

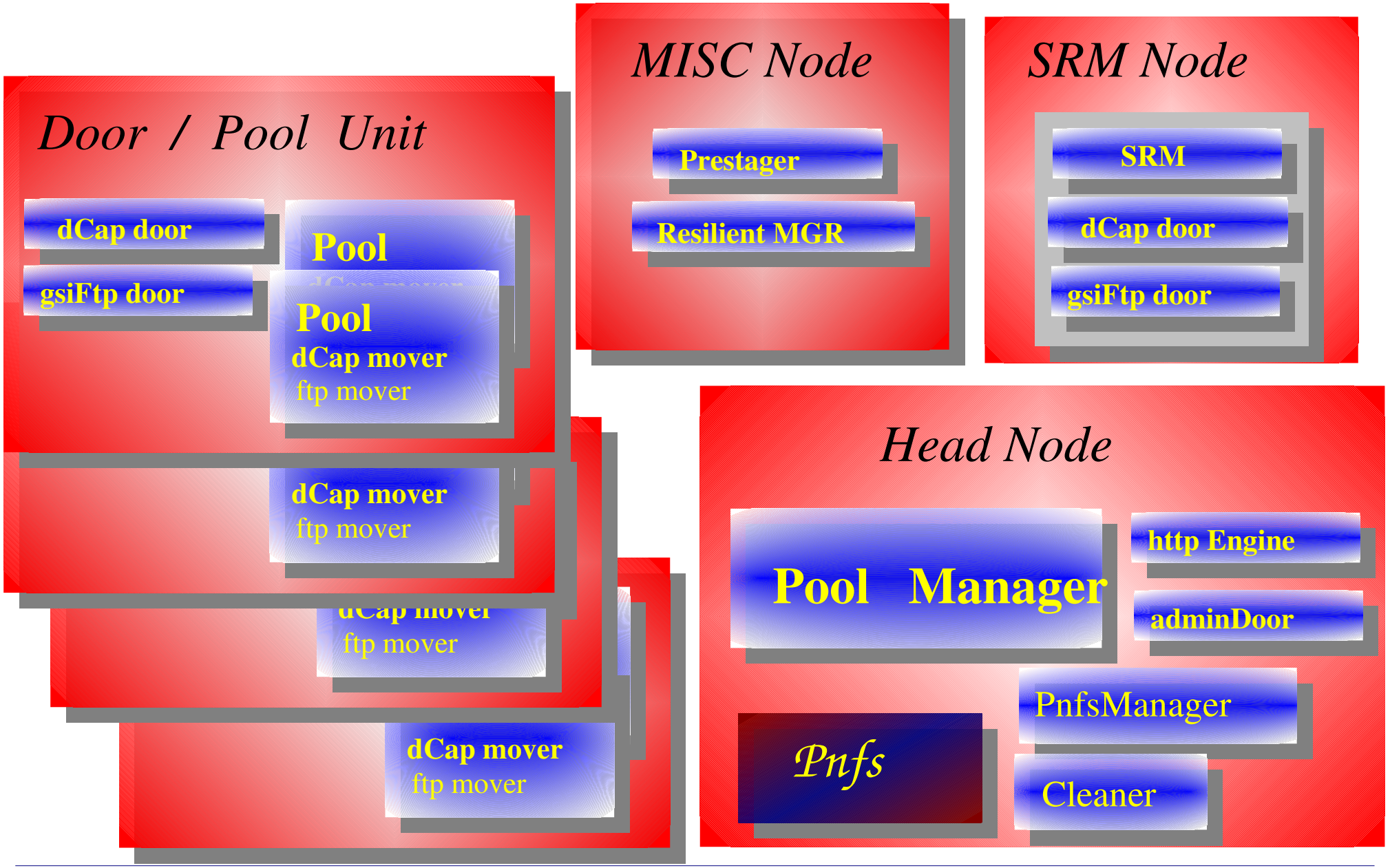
Starts appropriate door for incoming connections.

Pool

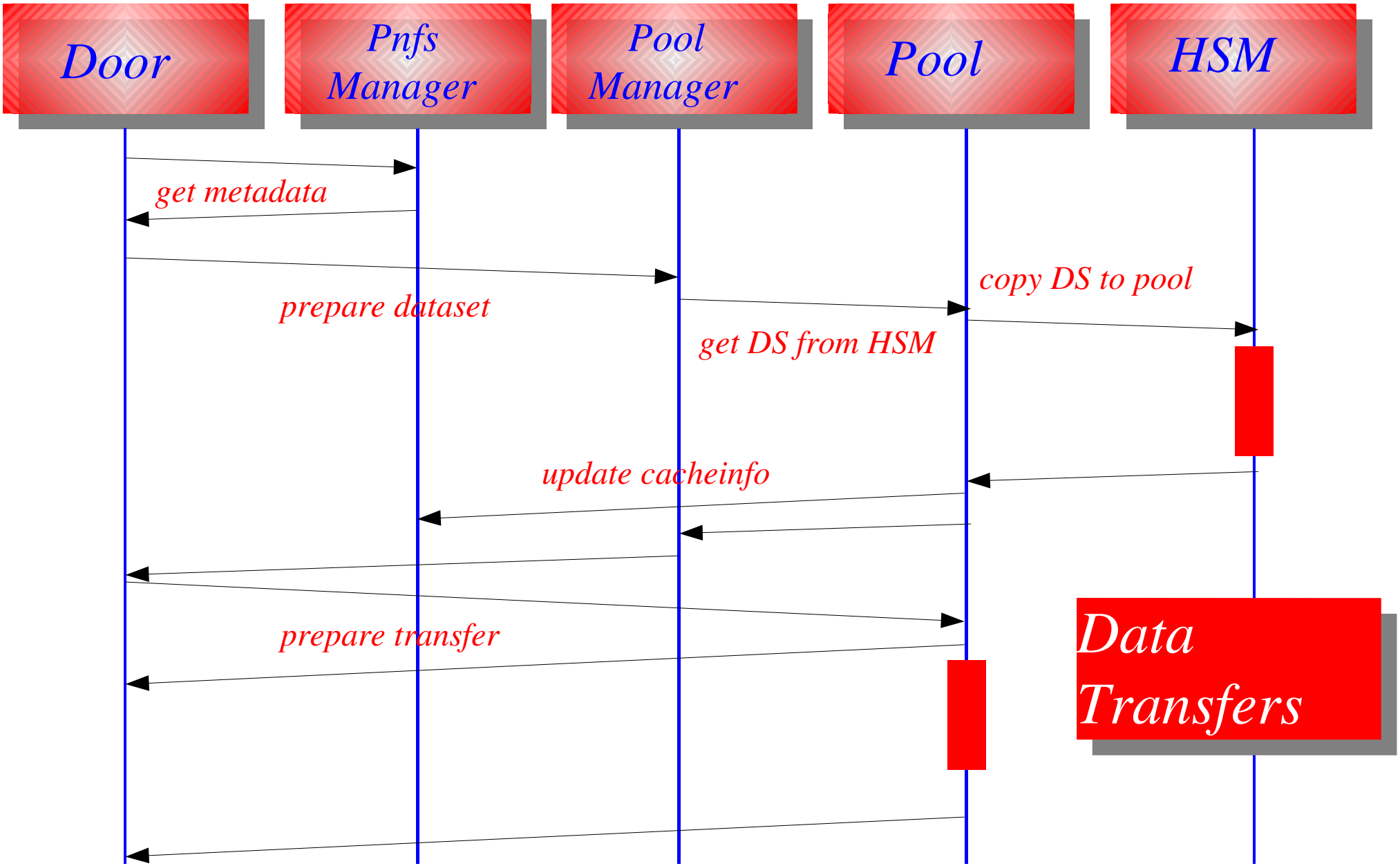
Does space management and launches 'movers'.


SRM
dCap door
gsiFtp door
Pool Manager
Pool
dCap mover
ftp mover
Pool
dCap mover
ftp mover
Pool
dCap mover
ftp mover
PnfsManager
Cleaner
Pnfs

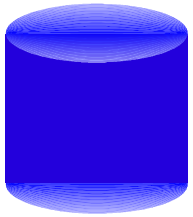

Head Node
SRM
dCap door
gsiFtp door
http Engine
adminDoor
Pool Manager
PnfsManager
Cleaner
Pnfs
dCap door
gsiFtp door
Door Node(s)
Pool Node(s)
Pool
dCap mover
ftp mover
Pool
dCap mover
ftp mover
ftp mover
ftp mover
ftp mover
ftp mover


Door / Pool Unit
dCap door
gsiFtp door
Pool
Pool
dCap mover
ftp mover
dCap mover
ftp mover
ftp mover
dCap mover
ftp mover
MISC Node
Prestager
Resilient MGR
SRM Node
SRM
dCap door
gsiFtp door
Head Node
Pool Manager
http Engine
adminDoor
PnfsManager
Pnfs
Cleaner

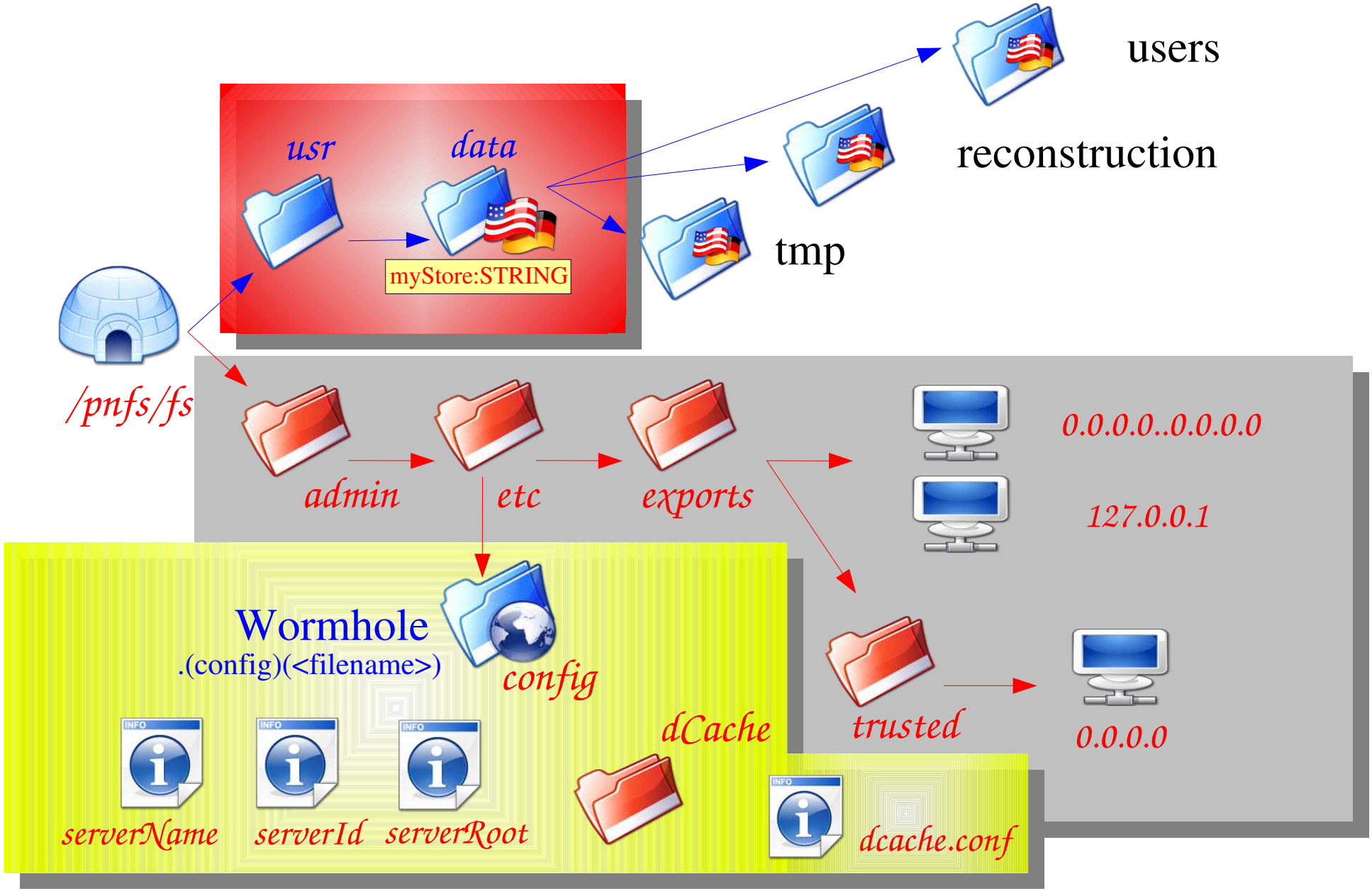
dCache transfer



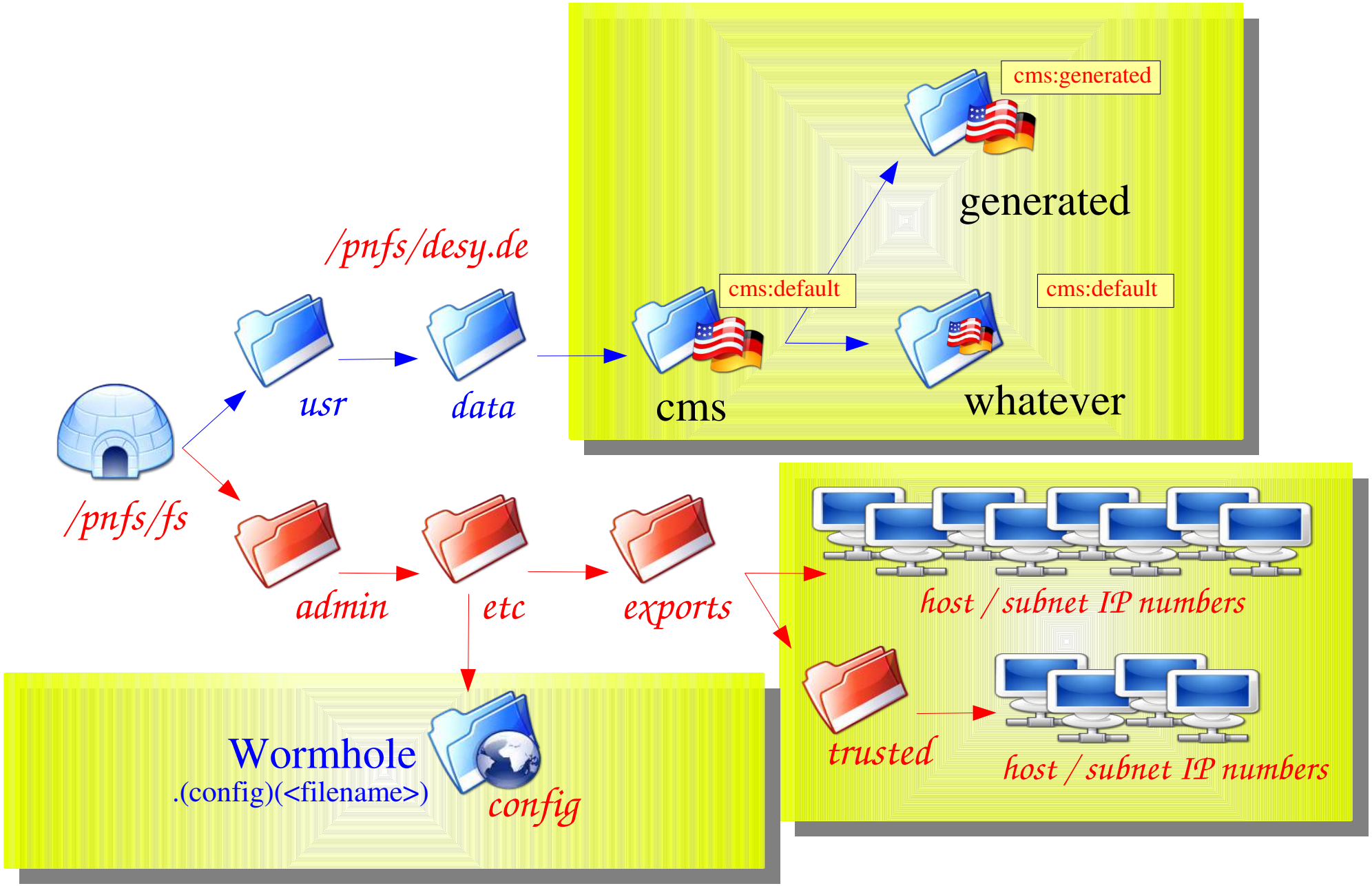
Pnfs Filesystem Setup



Pnfs : default setup



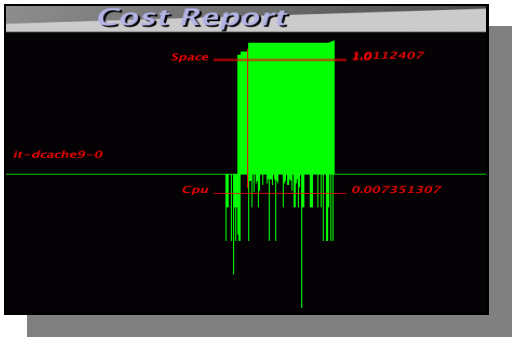
Pnfs : customized setup



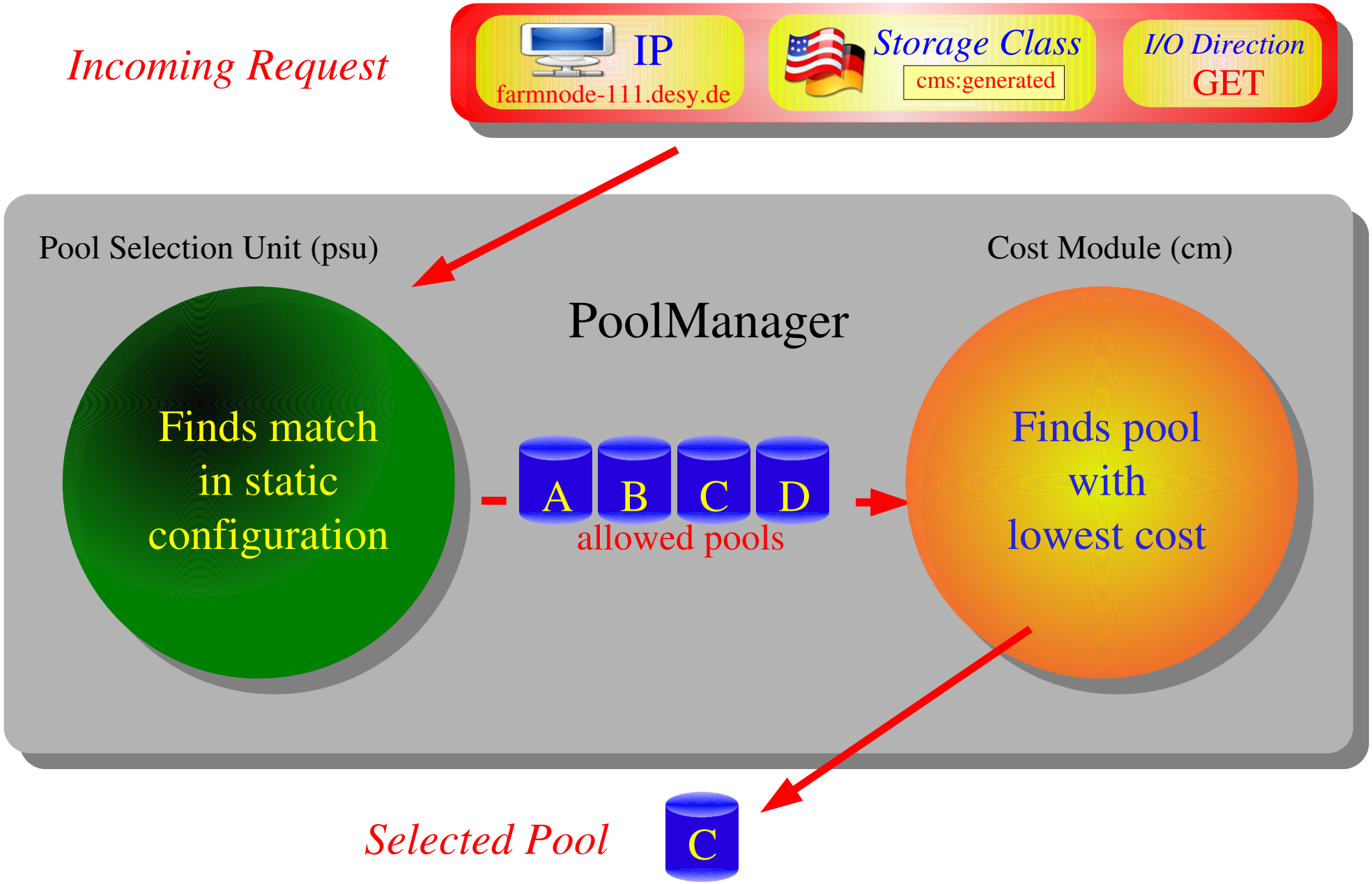
The PoolManager

Pool Selection Unit

Cost Module



PoolManager : request flow



PoolManager : static configuration

dCache.ORG

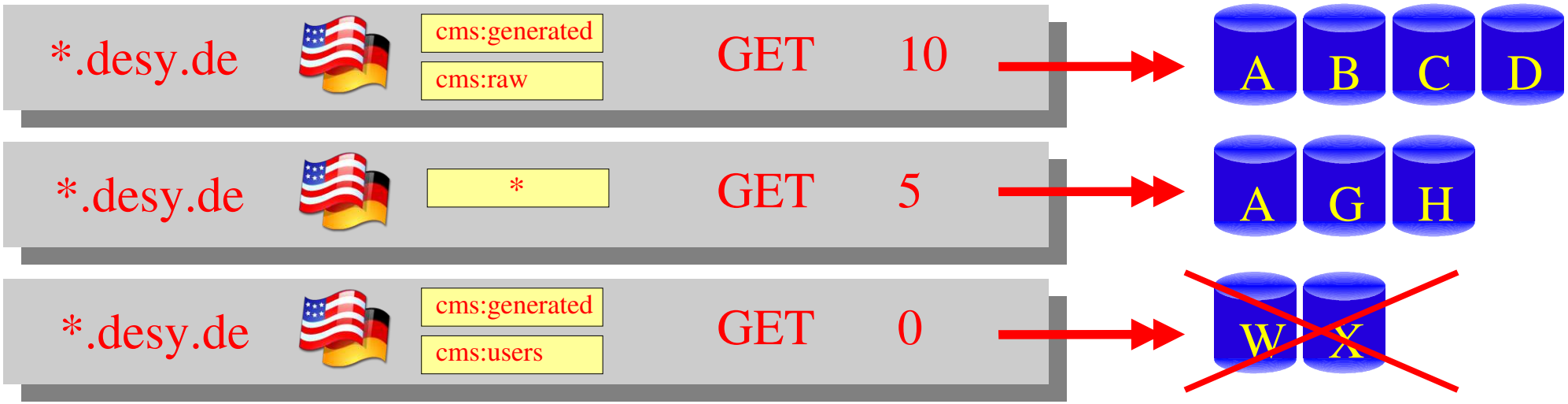
Incoming Request

!!! tries to find match in attraction table !!!

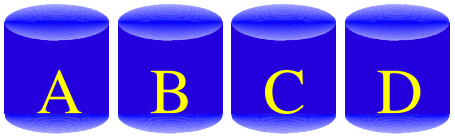
Netmask	Group of Storage Classes	I/O	Pref	LINK	Group Of Pools
*.desy.de	cms:generated, cms:raw	GET	10	→	A B C D
*.desy.de	cms:mc2004, cms:mc2005	GET	10	→	E F
*.desy.de	*	GET	5	→	A G H
*	cms:generated, cms:raw	GET	10	→	M N K
*.desy.de	cms:generated, cms:users	GET	0	→	W X

Incoming Request

!!! Result sorted by preference !!!



Selected Pool Candidates :



dCache -> Pools -> match

Birds Home

Pool Manager Database V3

Pools | Pool Groups | Selections | Selection Groups | Match

Simulated I/O Request

Store	dCache	Net
cms:generated@osm	*	farmnode-111

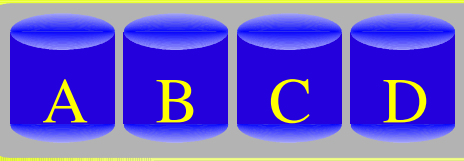
Select Request Type : **Read** Send Query

Selected Pools with attraction 0

A	B	C	D	-

Selected Pools with attraction 1

A	G	H	-	-




```
psu create pool A
psu create pool B
psu create pool C
psu create pool D
psu create pgroup ABCD-pools
psu addto pgroup ABCD-pools A
psu addto pgroup ABCD-pools B
psu addto pgroup ABCD-pools C
psu addto pgroup ABCD-pools D
```

*.desy.de

cms:generated

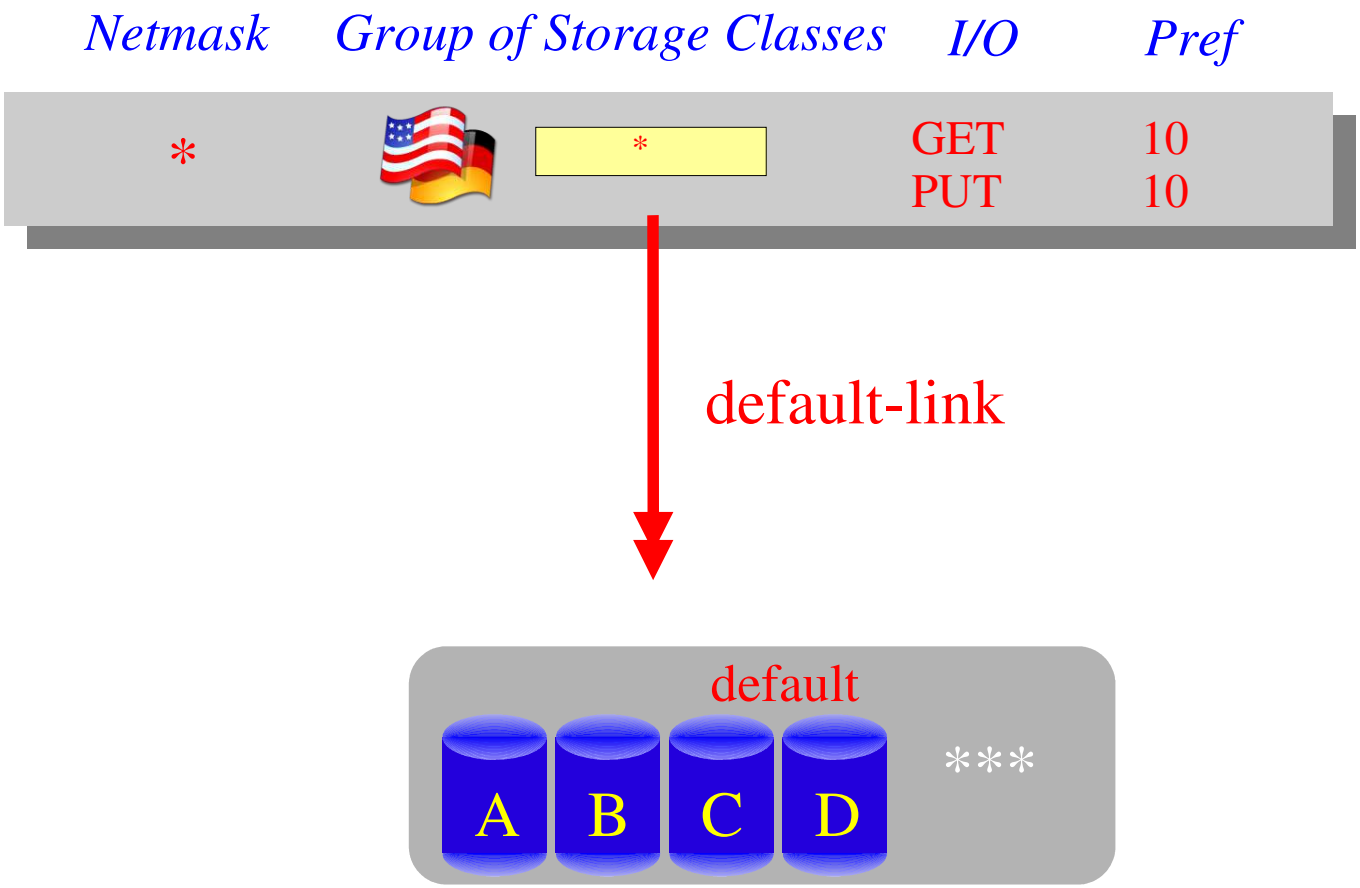
cms:raw

```
psu create unit -store cms:generated@osm
psu create unit -store cms:raw@osm
psu create ugroup gen_and_raw
psu addto ugroup \
        gen_and_raw cms:generated@osm
psu addto ugroup \
        gen_and_raw cms:raw@osm
psu create unit -net 255.255.0.0/131.169.0.0
psu create ugroup desy-net
psu addto ugroup \
        desy-net 255.255.0.0/131.169.0.0
```

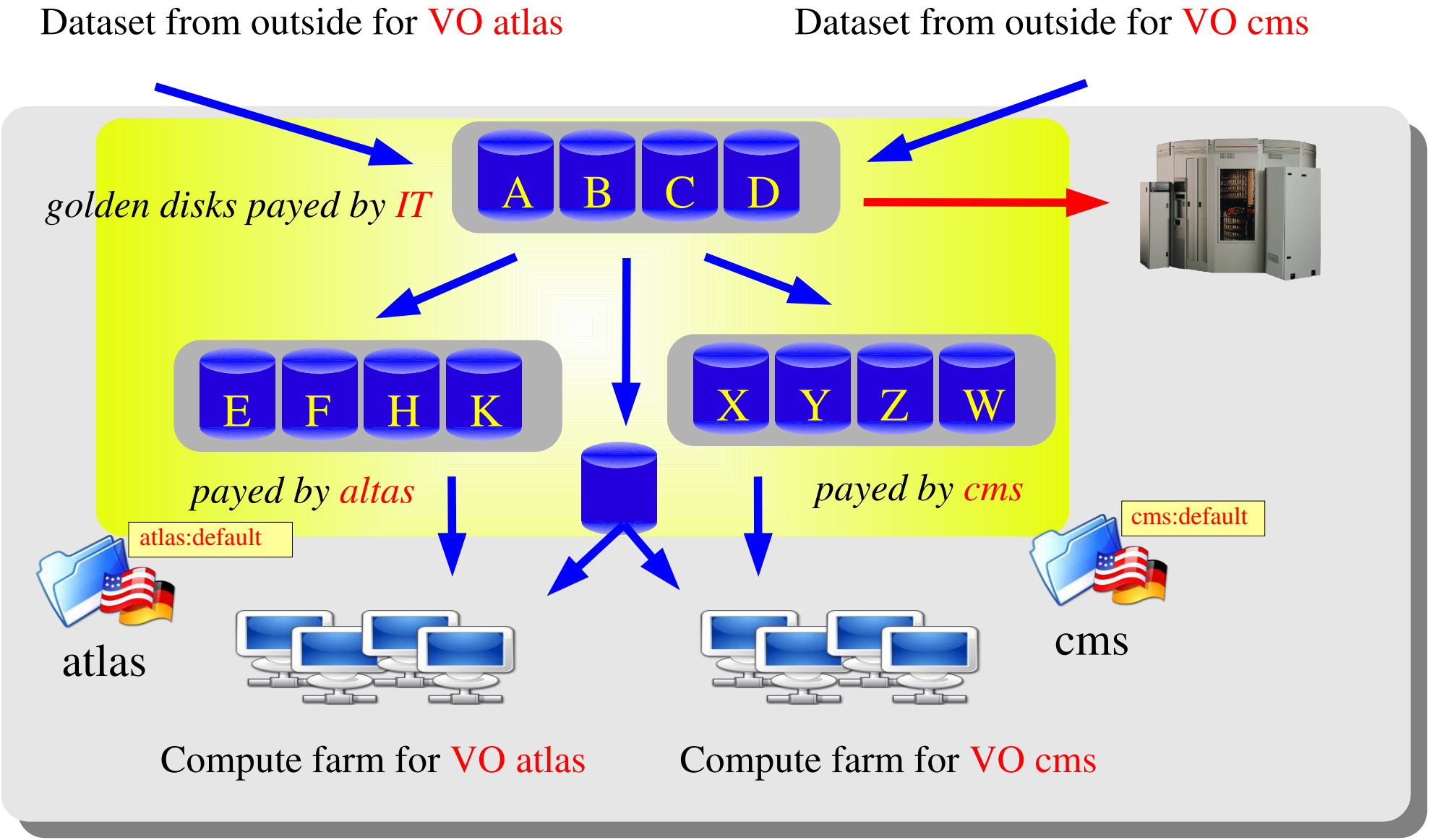
```
psu create link raw-gen-from-desy gen_and_raw desy-net
psu add link raw-gen-from-desy ABCD-pools
psu set link raw-gen-from-desy -readpref=10 -writepref=20 -cachepref=0
```

GET 10

PoolManager : default setup



Topology Example : the picture



Topology Example : formal description

dCache.ORG

Netmask	*Group of Storage Classes*	*I/O*	*Pref*	
*	*	GET PUT	10 10	A B C D write-pools
*.desy.de	atlas:default	GET PUT	10 10	E F H K atlas-pools
*.desy.de	cms:default	GET PUT	10 10	X Y Z W cms-pools
*.desy.de	atlas:default cms:default	GET	5	

Defining the pool groups

psu create pgroup **write-pools**
psu create pgroup **atlas-pools**
psu create pgroup **cms-pools**

psu create pool **XXX**
psu addto pgroup **xxx-pools XXX**

***

Defining the storage classes

psu create unit -store **atlas:default@osm**
psu create ugroup **atlas-store**
psu addto ugroup **atlas-store atlas:default@osm**

SAME FOR CMS

psu create unit -store ***@***
psu create ugroup **any-store**
psu addto ugroup **any-store *@***

Defining the network topology

psu create unit -net **0.0.0.0/0.0.0.0**
psu create ugroup **world-net**
psu addto ugroup **world-net 0.0.0.0/0.0.0.0**

psu create unit -net **131.169.0.0/255.255.0.0**
psu create ugroup **desy-net**
psu addto ugroup **desy-net 131.169.0.0/255.255.0.0**

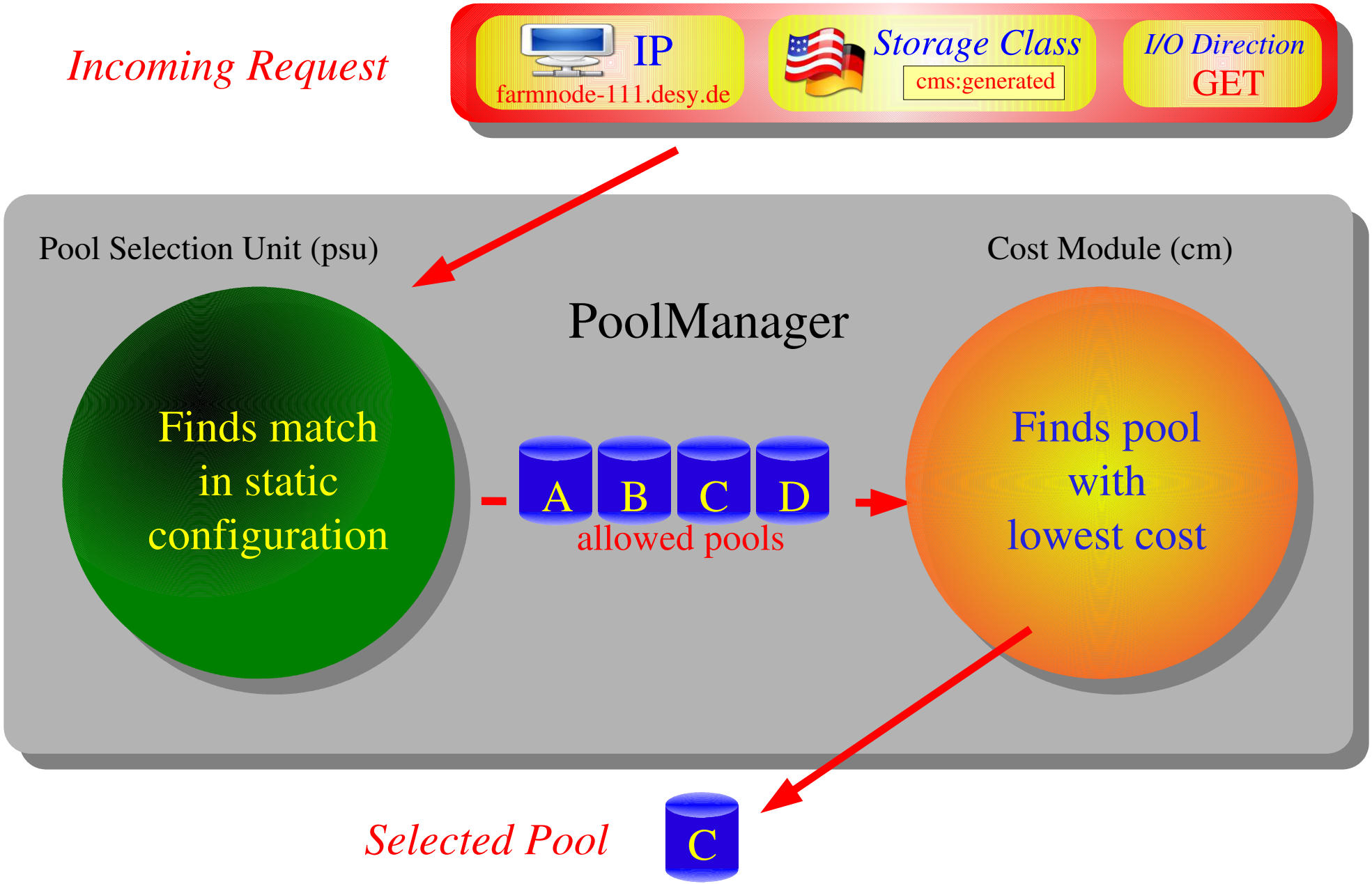
And the links

psu create link **write-link any-store world-net**
psu add link **write-link write-pools**
psu set link **write-link -writepref=10 -readpref=10 -cachepref=0**

psu create link **cms-link cms-store desy-net**
psu add link **cms-link cms-pools**
psu set link **cms-link -writepref=10 -readpref=10 -cachepref=10**

SAME FOR ATLAS

PoolManager : request flow



$$\text{Mover Cost (CC)} = \frac{\sum_{\text{Mover Queues}} \frac{\text{active} + \text{waiting}}{\text{maximum allowed movers}}}{\text{Number of Mover Queues}}$$

Mover Queues

from tape to disk	**rh set max active <num>**
from disk to tape	**st set max active <num>**
pool to pool (server)	**p2p set max active <num>**
pool to pool (client)	**pp set max active <num>**
client to/from pool	**mover -queue=<queueName> set max active <num>**

$$\text{Space Cost (SC)} = 1 + \frac{1}{\text{LRU file}} * 3600 * 24 * 7 * \text{<breakEven>}$$

Break Even : **set breakeven <breakEven>**

Total Cost per pool (tc) :

tc = cpu-factor * cc + space-factor * sc

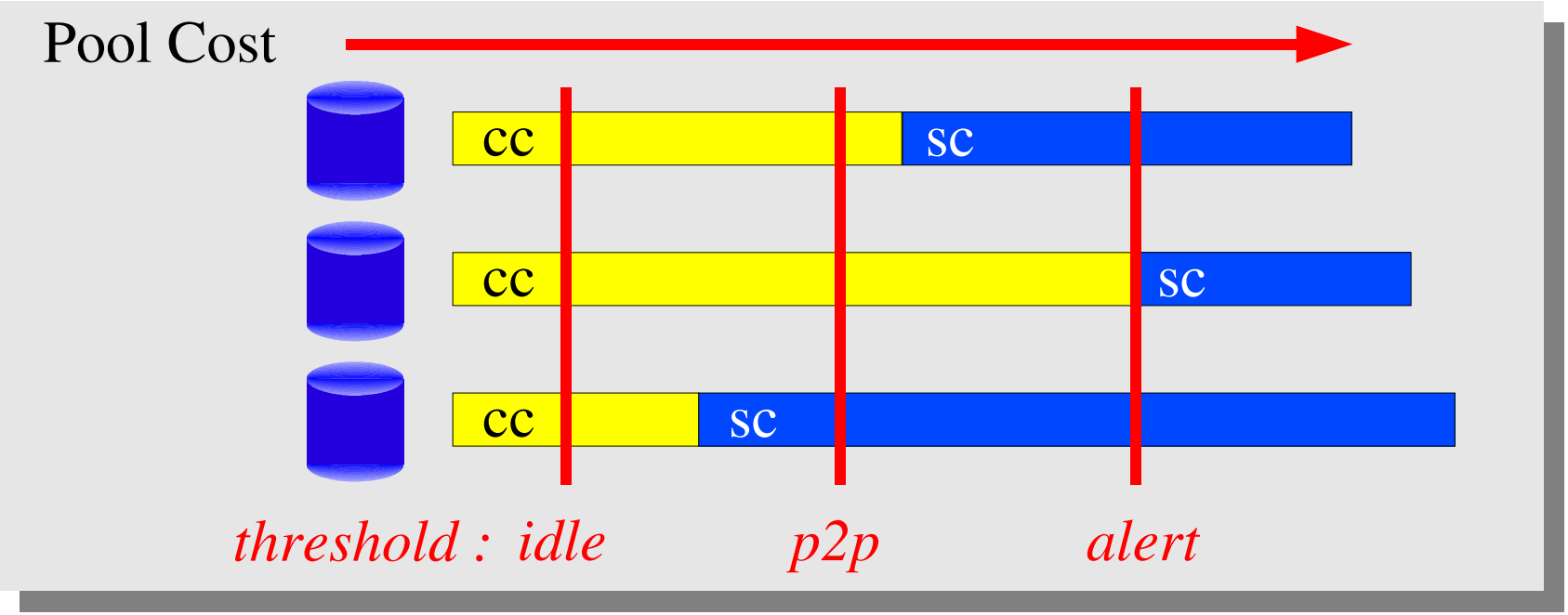
for write/p2p-server/restore

tc = cpu-factor * cc

for read/p2p-client

setting factors : **set pool decision -spacefactor=<factor> -cpufactor=<factor>**

Pool cost thresholds and actions



setting idle threshold : **set costcuts -idle=<idle-threshold>**
setting p2p threshold : **set costcuts -p2p=<p2p-threshold>**
setting alert threshold : **set costcuts -alert=<alert-threshold>**

idle : below : don't do load balancing on multiple file copies
p2p : above : start p2p transfers on high load of best pool
alert : above : stop p2p transfers on high load, do restore instead

1.6.5 :

(PoolManager) admin > *cm set magic on*

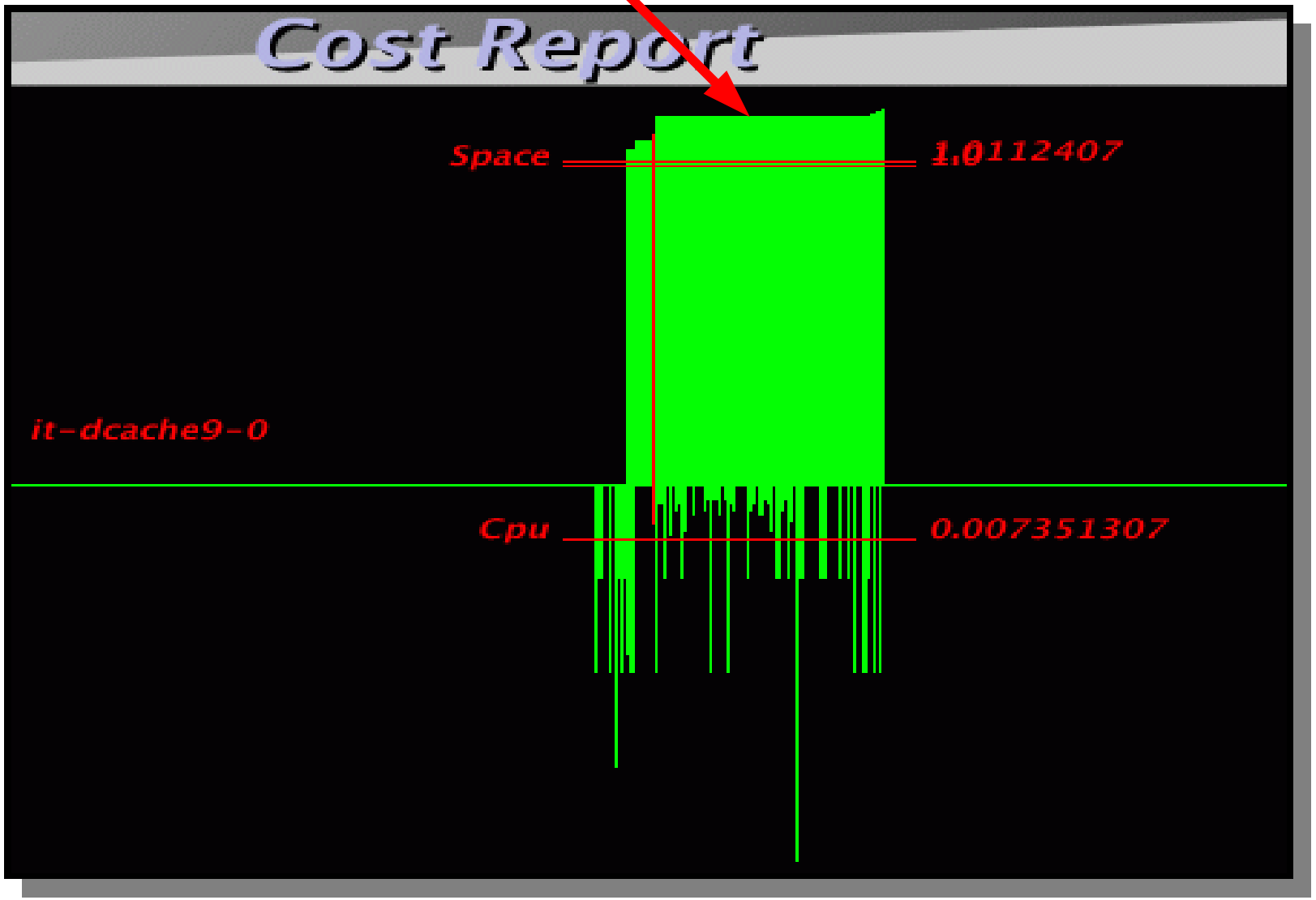
- command not saved with 'PoolManager save' command
- so either put it in PoolManager.conf and don't use save but reload
- or type it in after restart of the pool manager

1..6.6 (upcoming)

fixed

Pool cost example : desy zeus pools

1.154 = 30 Days (access date of LRU file)



Resilient Manager

Takes over control of dataset multiplicity

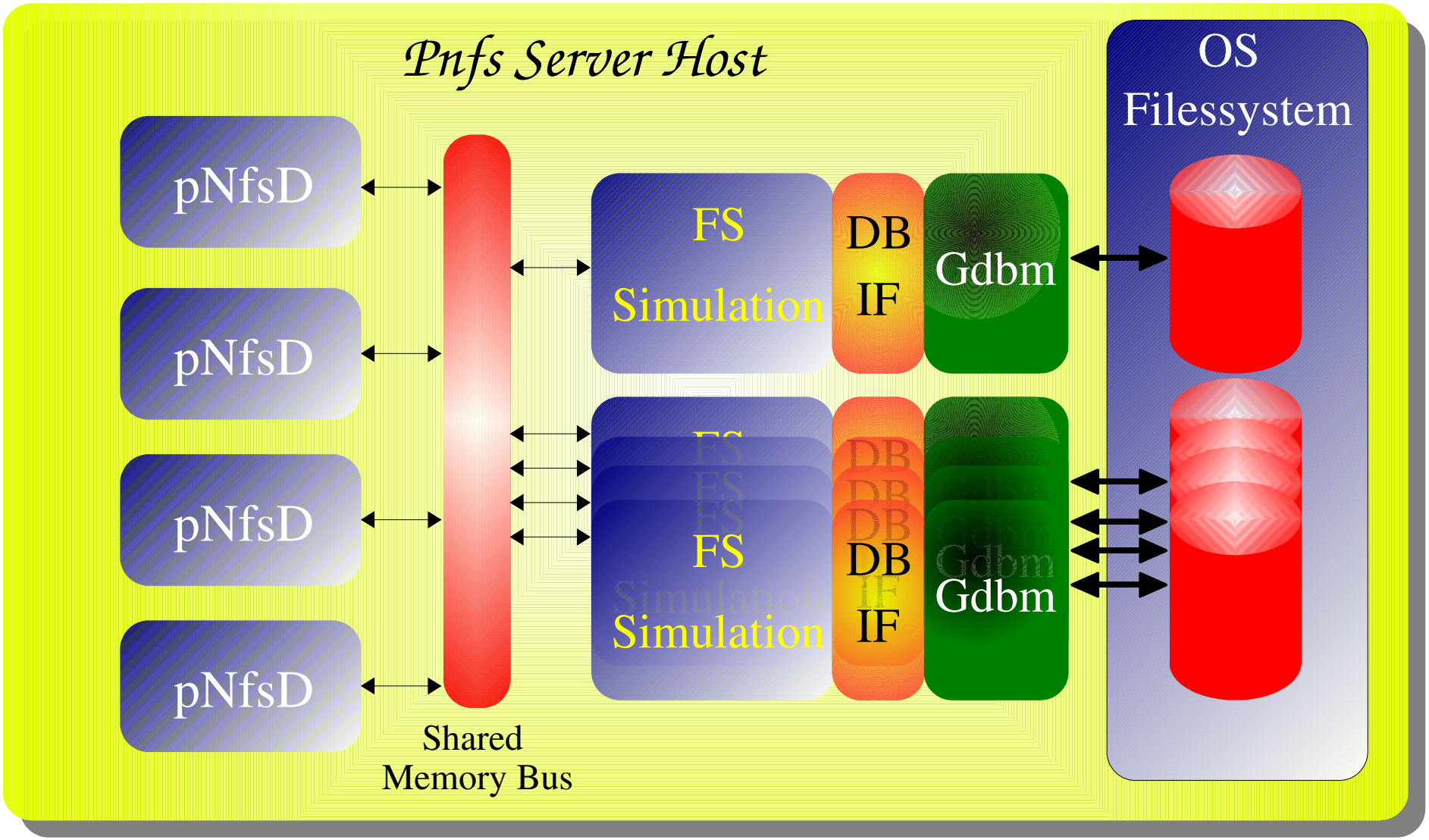
- makes sure, number of replicas between two given numbers.
- allows controlled draining of files from one pool to many other pools for maintenance reasons.
- up to now : HSM connected and replica pools are mutual exclusive.
- ResilientManager managed pools have to be 'lfs=precious'.
- Code already in distribution.
- Configuration still not easy enough.
- Configuration and Documention coming with 1.6.6.

This and That

Copies datasets from one pool to exactly one other pool

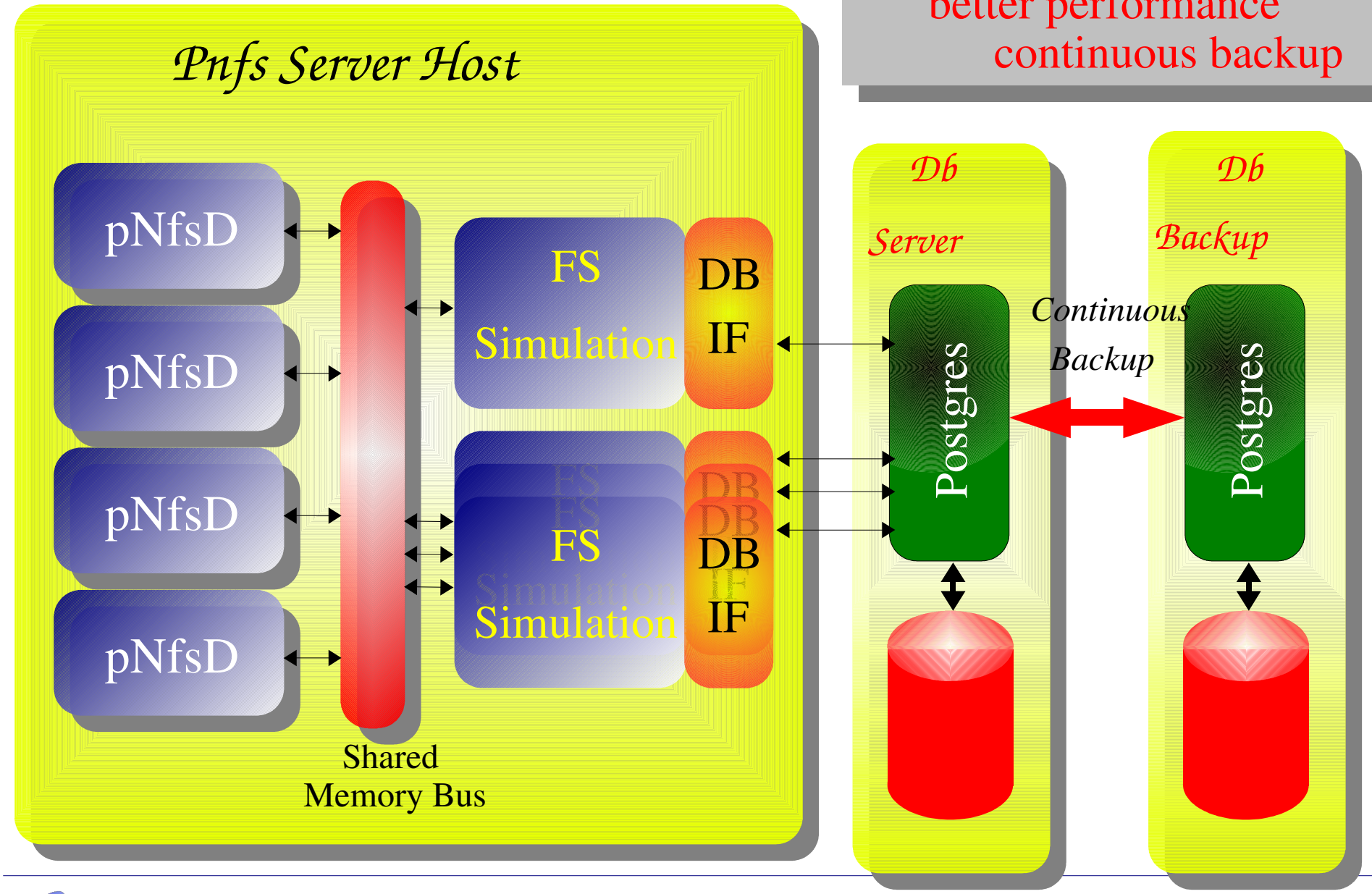
- still experimental, but turned out to work not only for DESY :-)
- needed only in 'non resilient' systems.
- code already in all 1.6.5 distributions
- documented in '*dCache, the Book*'
- may copy all files or just precious files.
- number of concurrent copy operations configurable
- destination pool needs enough (empty or removable space) to store all data from source pool.

Pnfs Issues : current design



pnfs : postgres solution

no db size limitation
better performance
continuous backup



pfns : details

Migration scripts (gdbm -> postgres) available

It's not possible to split existing gdbm/postgres databases

For pnfs files > 2G, and 'LargeFileSimulation' enabled :

- filesize in pnfs is '1 byte' because of nfs2 limitations
- size for nfs3 should be correct (code still experimental)

Coming soon .....

you are here

High speed mover queues

Scaling Web Monitors

Centralized HSM Flush
(disk spec aware)

Smart Restore (Staging)

nfs3

improved cacheinfo service

Chimera (db based pnfs)

SRM : Space reservation

SRM : Version 2 ...

Scalable StorageElement

dCache using VOMS for authorization (authentication)

Improved Documentation (dCache, the Book)