CHIMERA – A NEW NAME SERVICE for dCache

Tigran Mkrtchyan
for dCache Team

Motivations

dCache.ORG

The main reason to replace Pnfs with something new is to solve in advance its limitations:

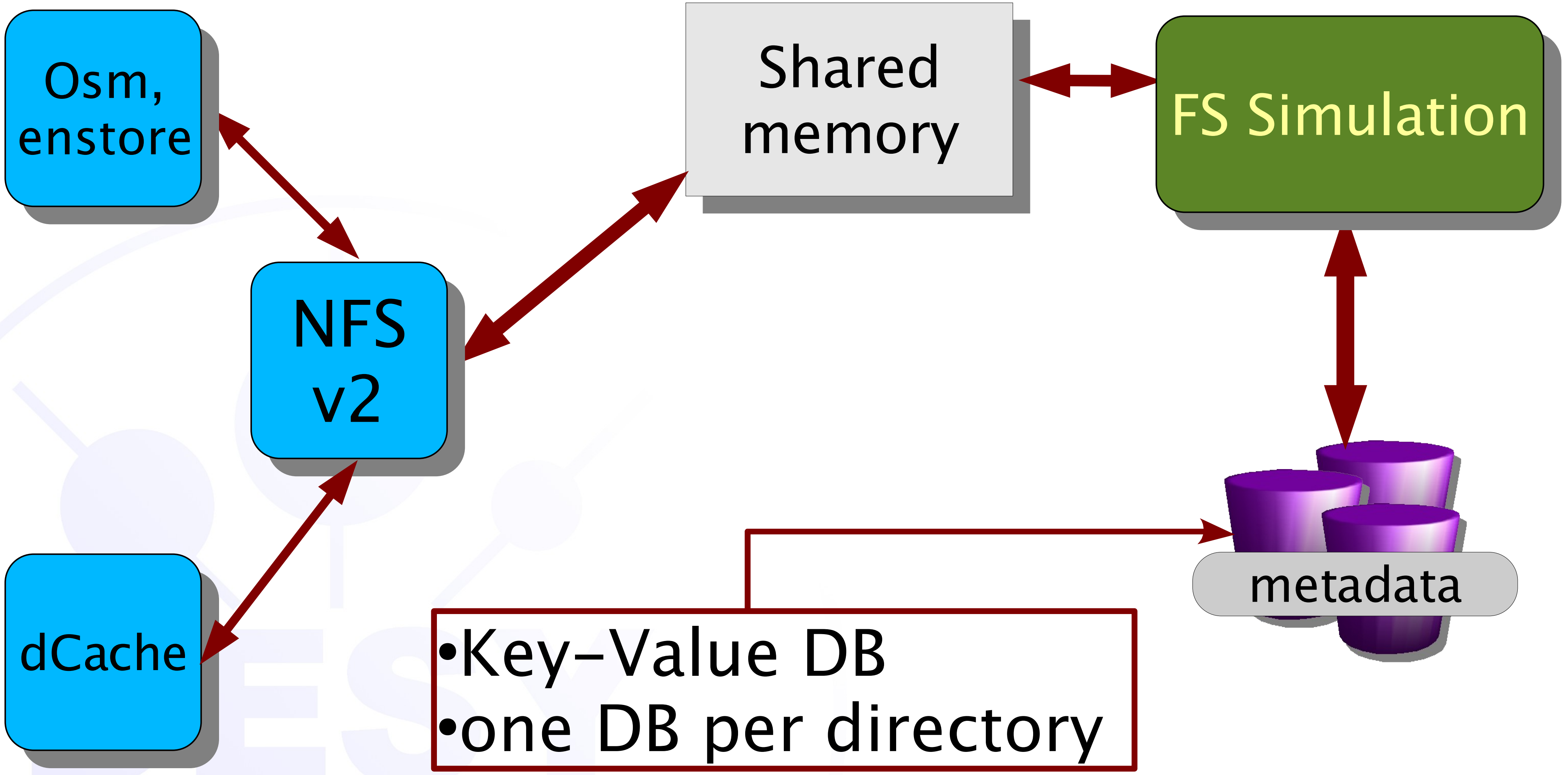
- Max. file size 2 Gb (NFSv2)
- Metadata access only through NFS
- Metadata stored as BLOB
- NFSv2 security (no security), No ACLs
- Identified as bottleneck

Requirements

- Unique file ID independent from name
- Path ID mapping
- Mechanism for clients to store metadata
- Callbacks on FS events
- platform independence (runtime + persistent)
- custom dCache integration
- multiple front-ends running in parallel
- Extendable/puggable front ends
- Extendable/pluggable ACL modules

Current approach

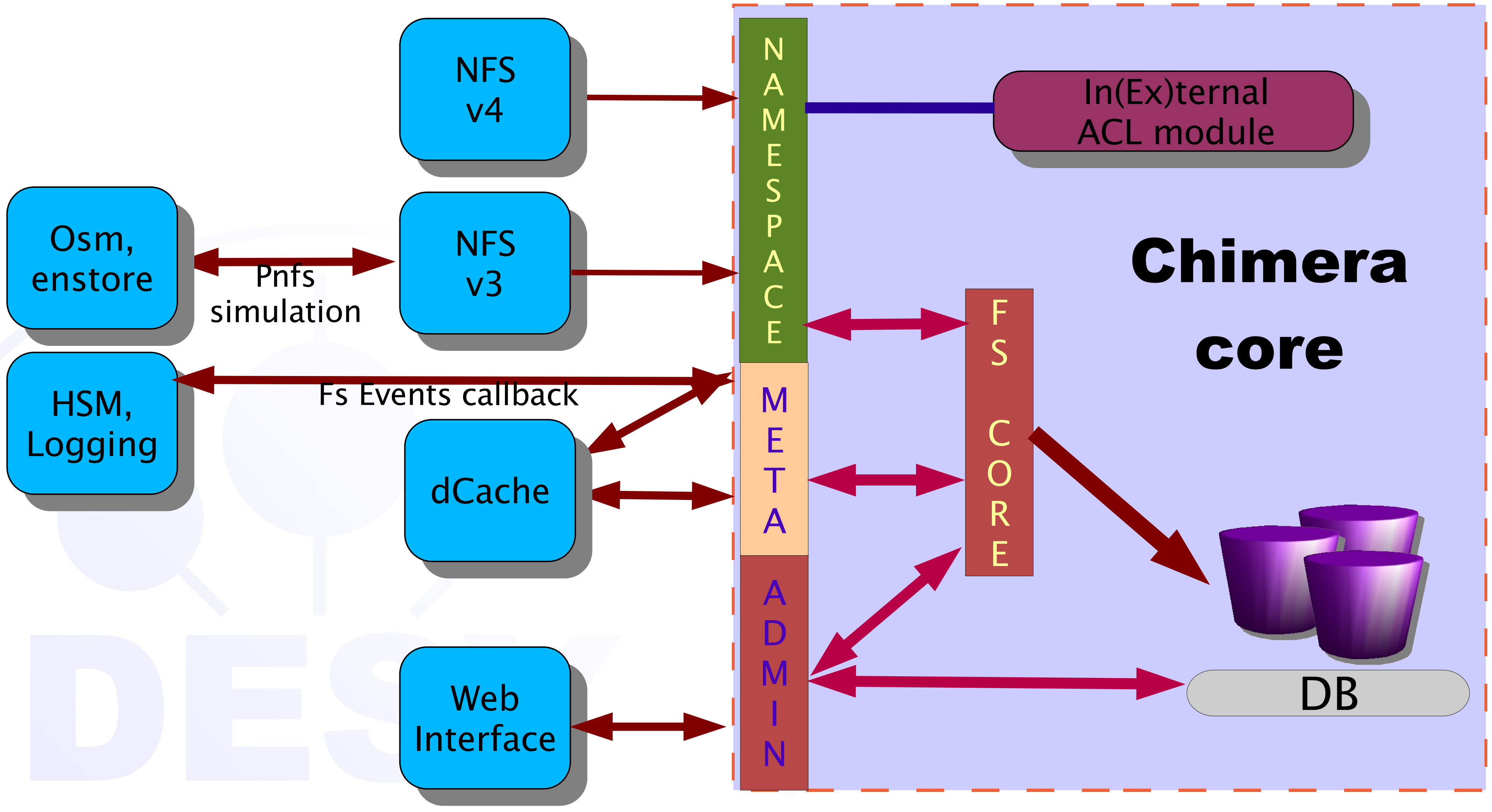
dCache.ORG



•Key–Value DB
•one DB per directory

Chimera approach

dCache.ORG



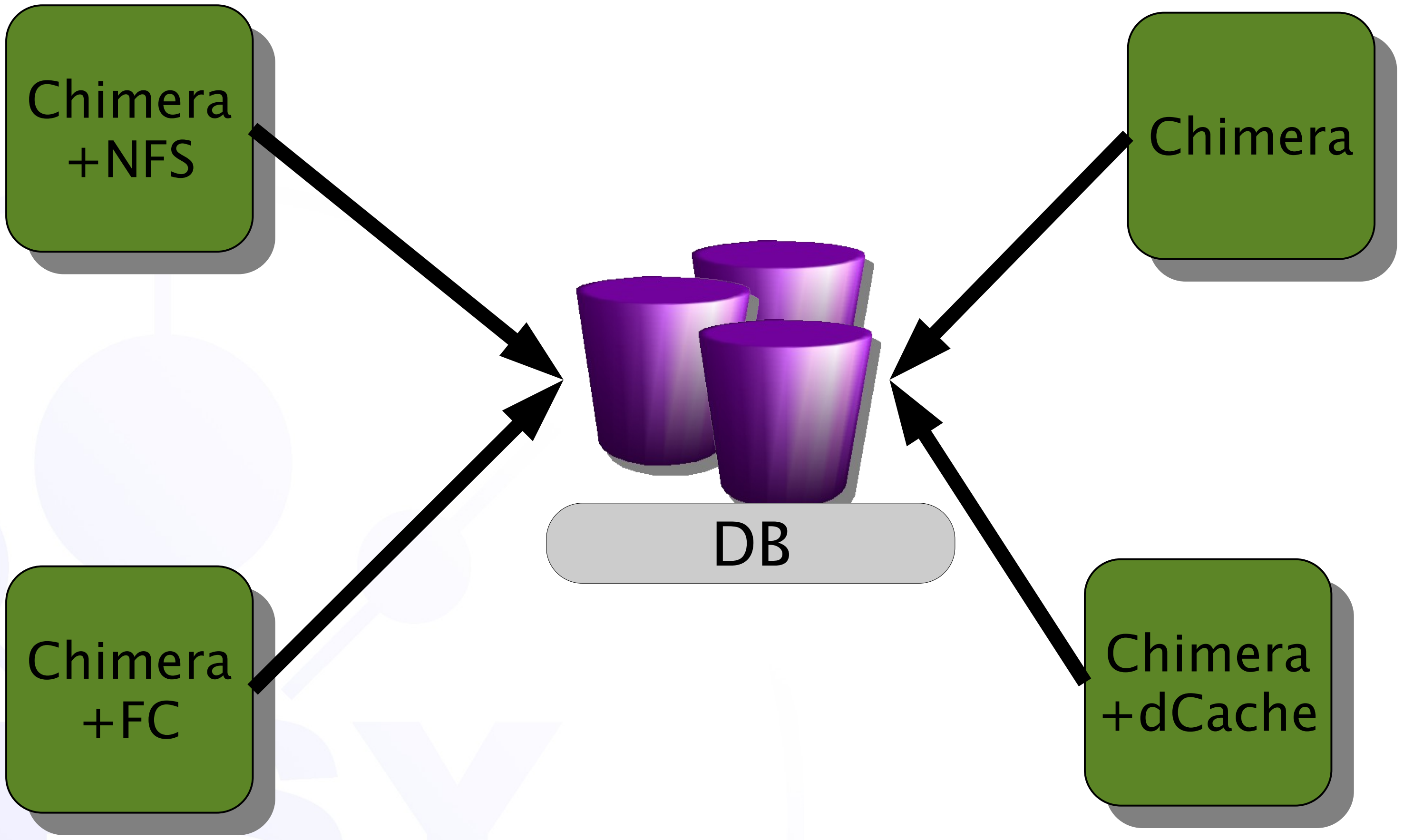
The Benefits

dCache.ORG

- Filesystem view and metadata separated
- UUID as pnfsid
- No NFS operations involved in API
  - stat over NFS end up with 3 ops (parent GATATTR LOOKUP, GETATTR)
- Pluggable authentication
  - unix, ACL, VOMS
- Extendable frontends
  - File browsers, replica catalogue, NFSv4

Chimera approach (II)

dCache.ORG



Benefits of RDBMS

dCache.ORG

- Well known
- Query Language
  - Simple queries to get space usage, file numbers
- Backup
  - Some databases allows point in time recovery
- Consistency check
  - primary/foreign key
- Stored procedures
  - fs check

RDBMS as a back end

dCache.ORG

- Transactions for all views

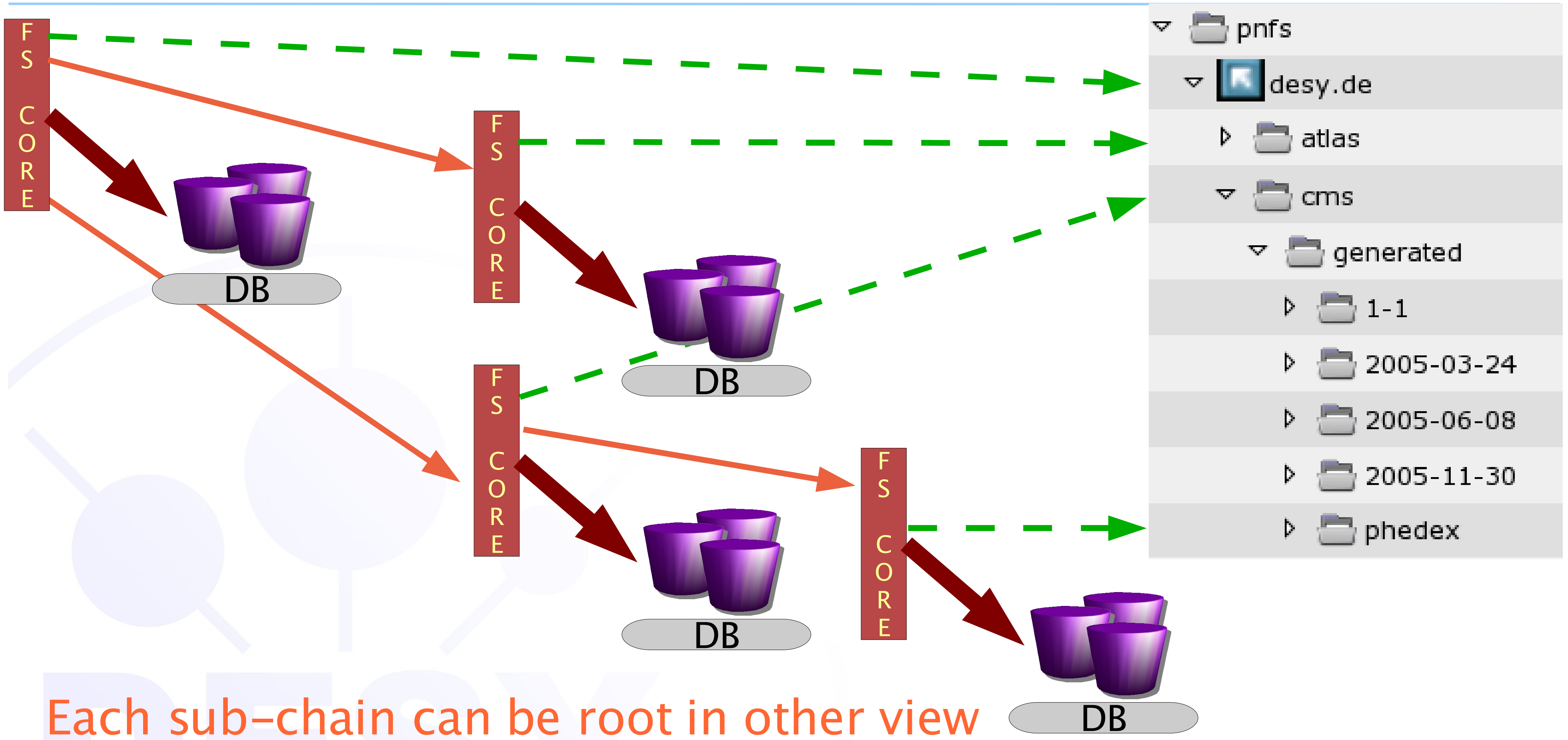
  only consistent state available to all frontends
- All attributes for a file available for SQL queries
- Different tables for data and directory structure

  (allows to have a view based on different tree, e.g. *Spaces*)

Pure JDBC allows to be database implementation independent, nevertheless, vendor specific driver can be used.

Filesystem chaining

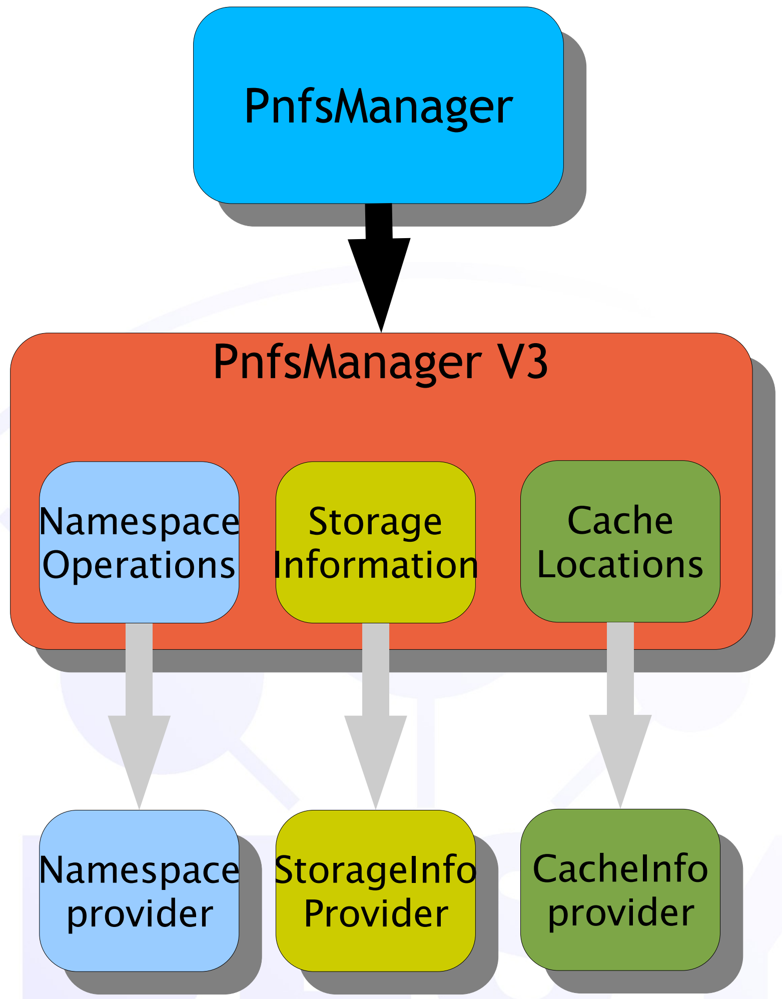
dCache.ORG



Each sub-chain can be root in other view

dCache adaptation

dCache.ORG



- Chimera provides:
  Namespace
  Storage Information
  file locations

- dCache provides:
  file locations (aka “Companion”)

Extended file locations

dCache.ORG

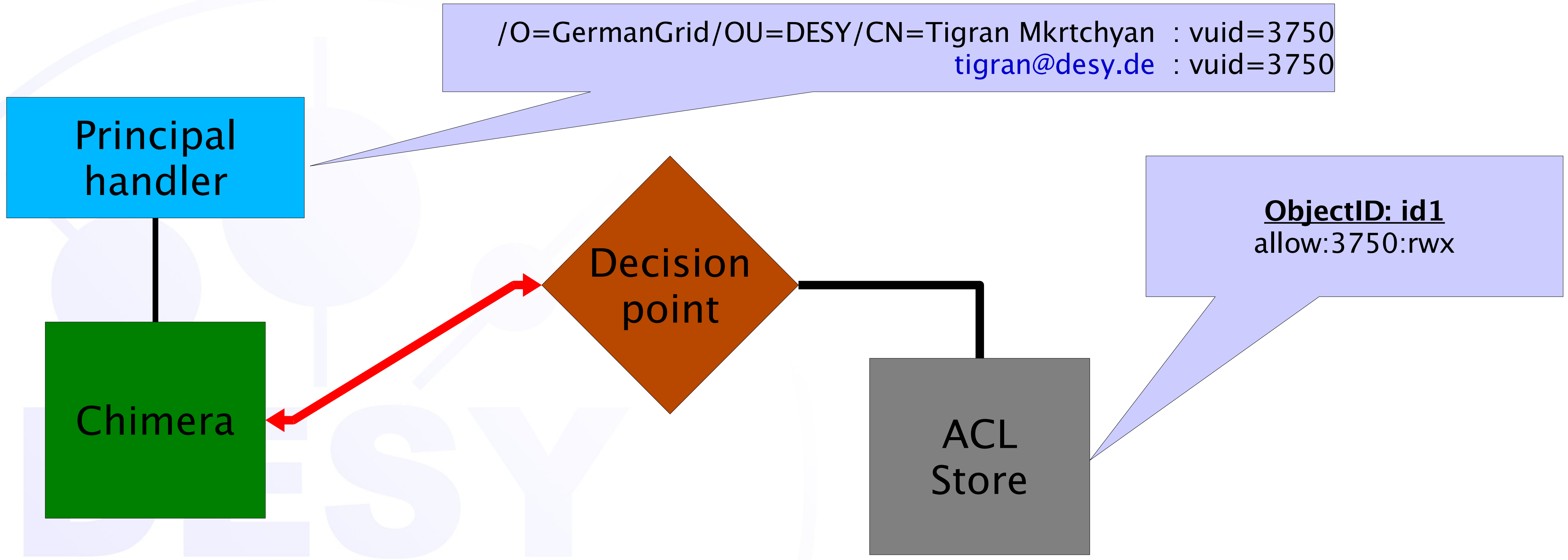
- Internal (disk) and External (tape) locations handled the same way
- location status (ON-LINE/OFF-LINE)

  Some pools can be disabled for some files
- location priority

After more experience will be ported to companion

ACL (See DP and DM talk)

- User always mapped to the same vuid
- ACLs based on vuid


/O=GermanGrid/OU=DESY/CN=Tigran Mkrtchyan : vuid=3750
tigran@desy.de : vuid=3750
Principal handler
Decision point
ObjectID: id1
allow:3750:rwx
Chimera
ACL Store

NFS v3 or v4?

dCache.ORG

Main difference between v3 and v4 is:

- Compound RPC calls

  Stat produces 3 RPC calls in v4 and only 1 in v4
- GSS authentication

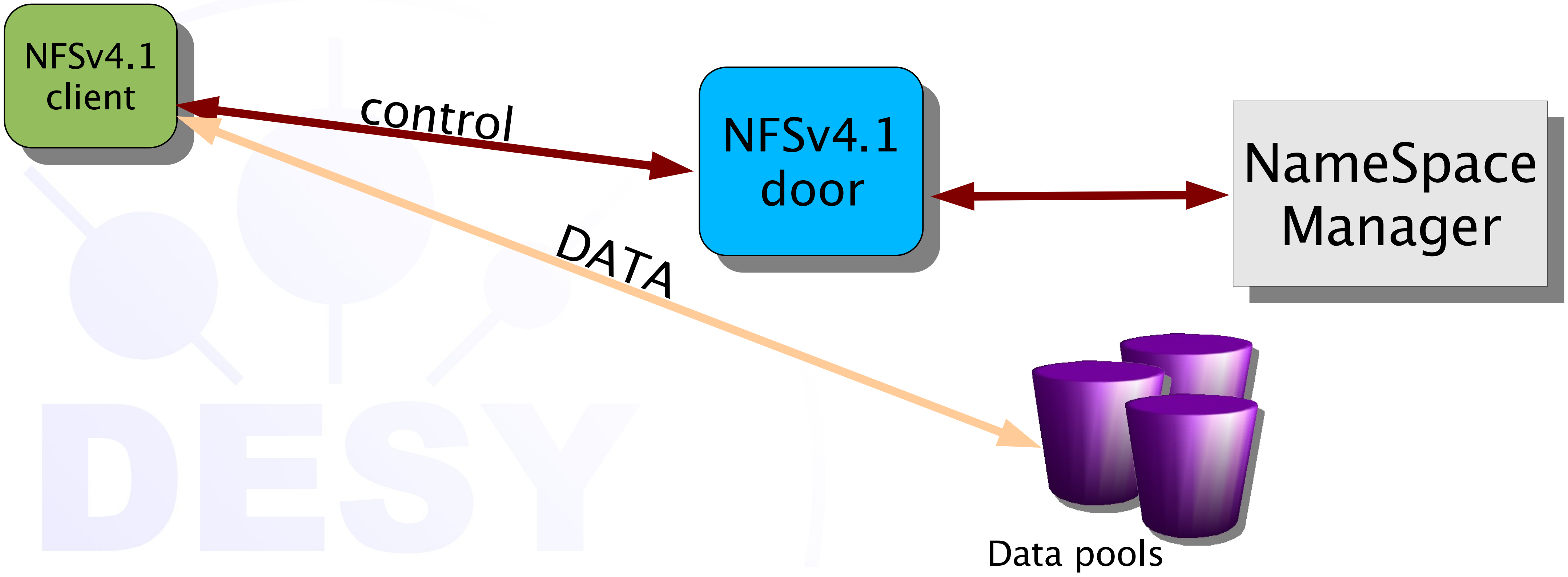
  Built in mandatory security on file system level
- ACLs
- OPEN/CLOSE notation

  Opposite to v3, v4 has clear ***BEGIN*** and ***END*** of IO operation
- 'Dead' client discovery

  The client have to periodically notify server, that hi sill accesses a file

or v4.1?

v4.1 makes possible to split control and IO:



And more ...

dCache.ORG

	NFSv4.1	SRM
Security Protocol negotiation	OK	NA
Data Protocol negotiation	OK	OK
Client crash recovery	OK	NA
Server crash recovery	OK	NA
ACLs	OK	UNIX-like

Chimera NFSv4.1

dCache.ORG

- compatible with existing clients (Linux, Solaris)
  (both with special kernels)
- Supported by other vendors
- Not official, greatly changing with v4.1 spec
- Needs some dCache internal changes
- Will be released as soon as spec finally defined.

NFSv4.1 called pNFS,
which makes lot of confusion

Current Status

dCache.ORG

- 200 file creates per second per thread
- Tested with ORACLE, PostgreSQL
- Full working beta (NFSv3/4, dCache)
- Used by myself in dCache development
- In test evaluation by FNAL
- Compatible with existing clients (dccp, osm)

ToDo

dCache.ORG

- Full scalable tests
  (looking for a volunteers)
- Distributed transaction log
- Migration mechanism for existing installations

Chimera?

dCache.ORG

In Greek mythology, a fire-breathing animal with a lion's head and foreparts, a goat's middle, a dragon's rear, and a tail in the form of a snake; hence any apparent hybrid of two or more creatures.