Managed Storage @ GRID

or

why NFSv4.1 is not enough

Tigran Mkrtchyan
for dCache Team

What the hell do physicists do?

- Physicist are ***hackers*** – they just want to know how things works.
- In moder physics given cause does not produce same effect.
- Statistic is used to describe behavior.
- Physics data is IMMUTABLE : you keep it forever or you removed it, but you never FIX it!

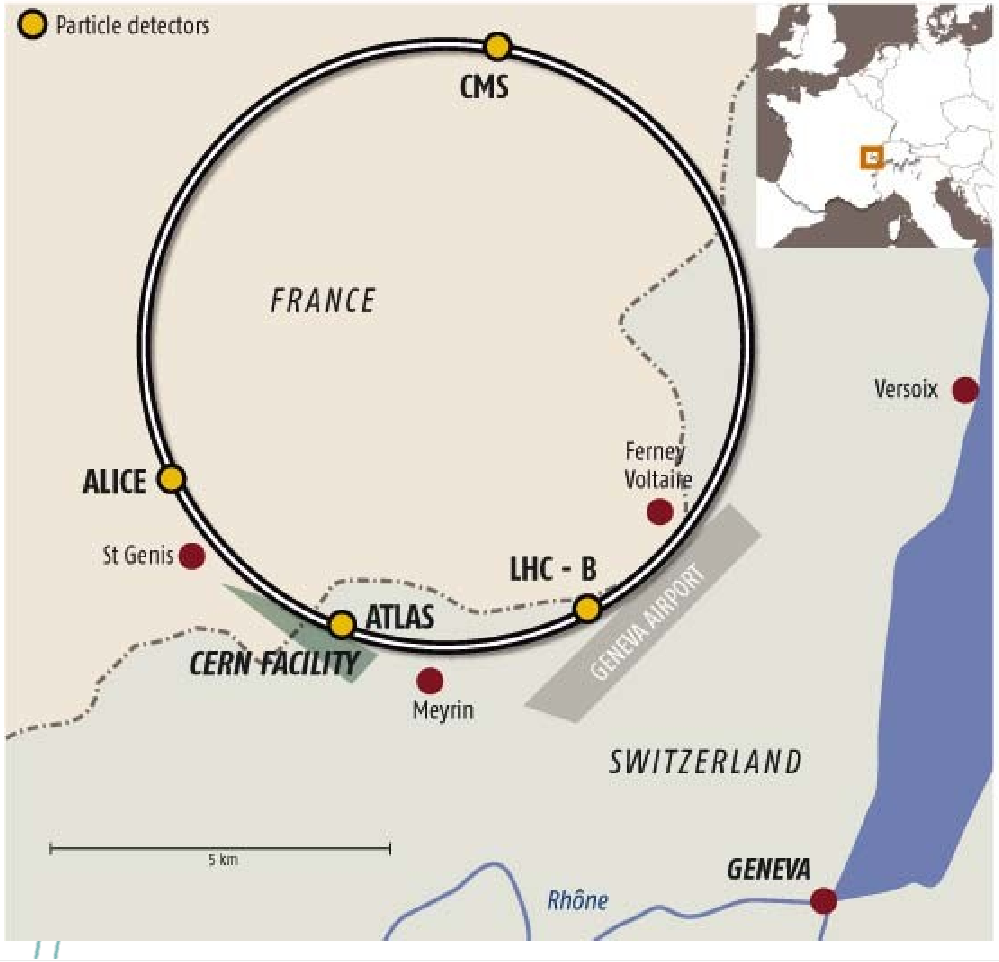
Right tool for right job

Large Hadron Collider:

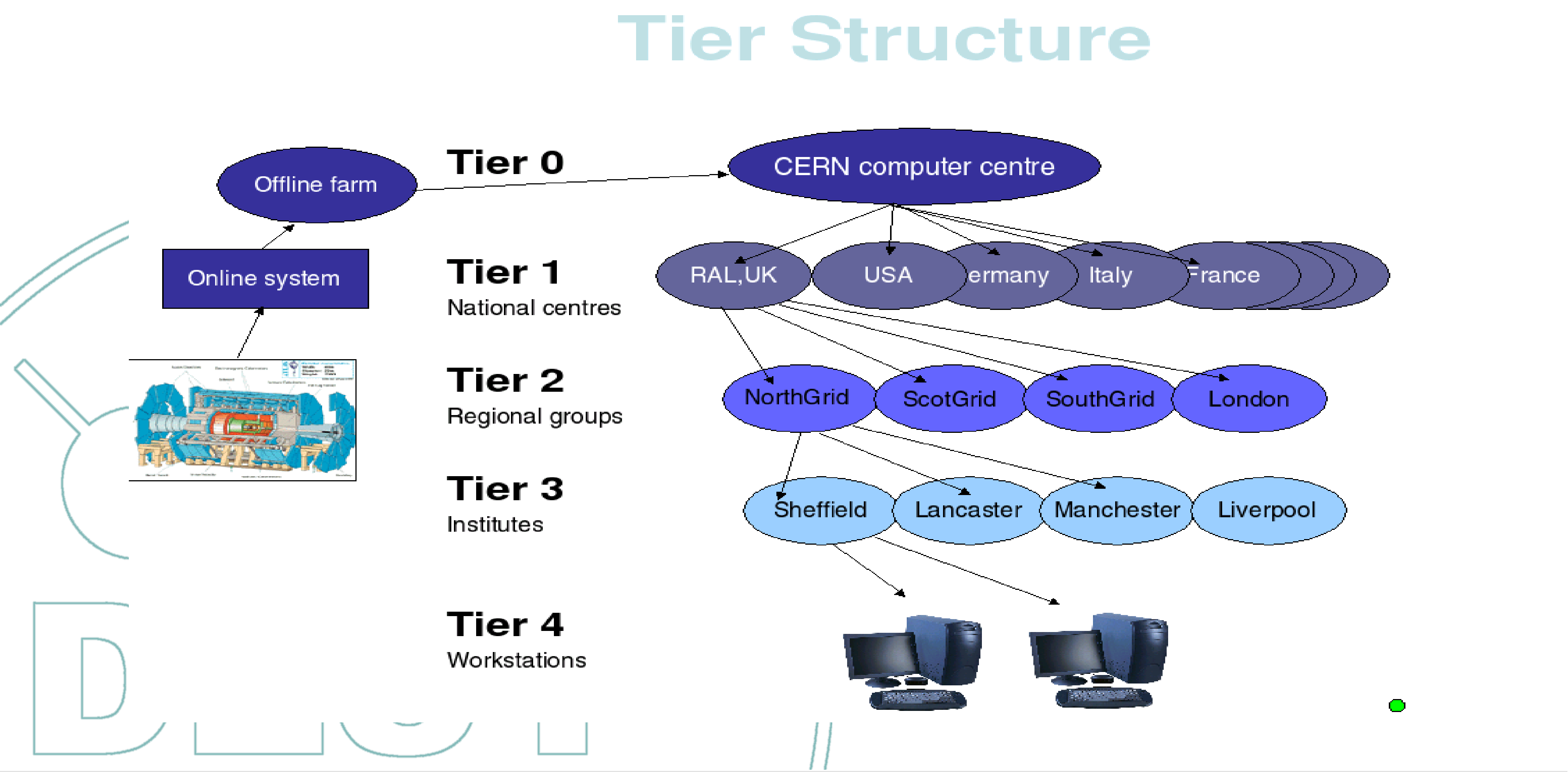
Expected start July 2008
800 million collisions per second
(25 km long)
Data rate ~ 1.5 GB per second
~15PB per year

LARGE HADRON COLLIDER

Four detectors around the 27-km-long accelerator will hunt for new particles, including the Higgs boson or "God particle"



Multiple tier model



Tier Structure

Offline farm

Online system

Tier 0

CERN computer centre

Tier 1

National centres

RAL,UK

USA

ermany

Italy

France

Tier 2

Regional groups

NorthGrid

ScotGrid

SouthGrid

London

Tier 3

Institutes

Sheffield

Lancaster

Manchester

Liverpool

Tier 4

Workstations

GRID as core infrastructure

GRID middleware applied to solve two major goals:

- Physical
  - space, power, cooling, connectivity
- Political
  - let regional investors to spend many for regional centers

What is a GRID ?

“The term Grid computing originated in the early 1990s as a metaphor for ***making computer power as easy to access as an electric power grid***.”

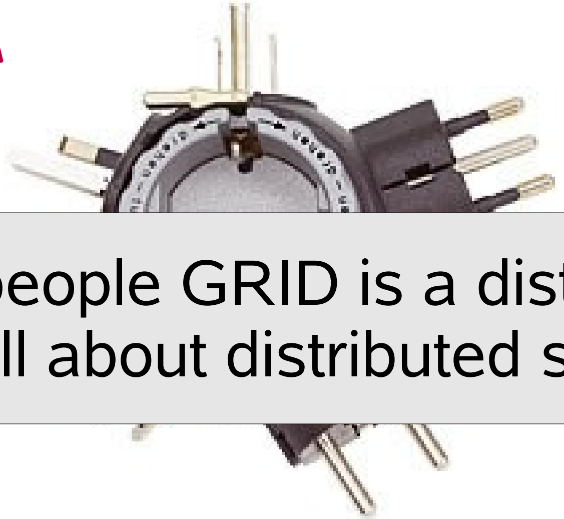
What is a GRID ?

“The term Gr ly 1990s
as a metaphor *easy to*
access

What is a GRID ?

“The term Gr

ly 1990s



While or most of the people GRID is a distributed CPU recourses, it's all about distributed storage!

Storage Resource Manager

To hide storage system implementation a top level management interface was defined - SRM.

SRM together with ***'Information Provider'***, which allows to query storage system called '***Storage Element (SE)***'

LCG

Open Science Grid

egee
Enabling Grids
for E-sciencE

Storage Resource Manager

Storage Resource Managers (SRMs) are middleware components whose function is to provide dynamic space allocation and file management on shared storage components on the Grid.

SRM interface defines following functions:

- Data Transfer
- File Pining/UnPining
- Space Management
- Request Status queries
- Directory operations
- Permission management

SRM Data Transfer

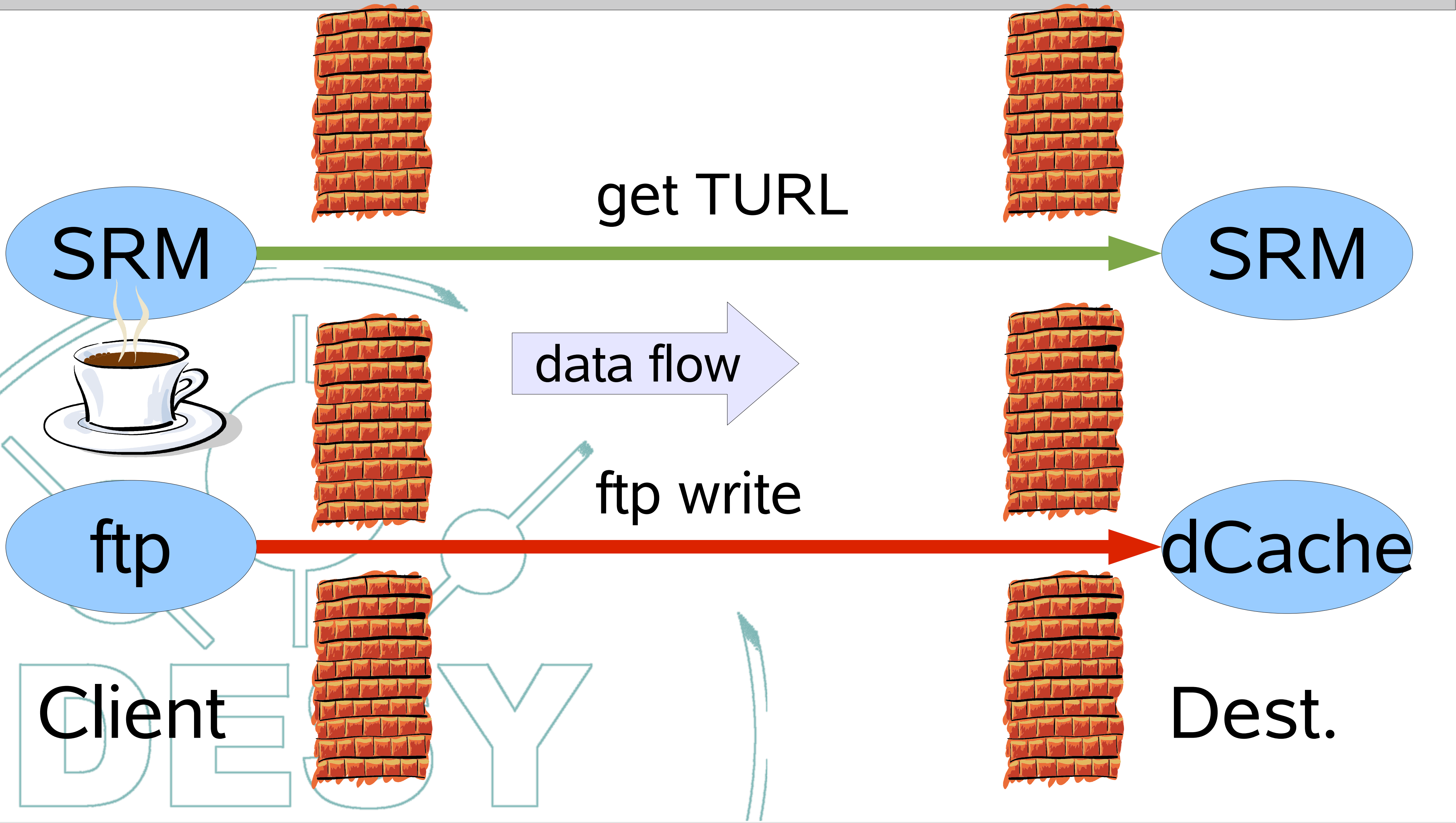
SRM data transfer based on two concepts: SURL and TURL.

- SURL - is a “site URL” which consists of “srm://host.at.site/<path>”.
- TURL - is the “transfer URL” that an SRM returns to a client for the client to “get” or “put” a file in that location. It consists of “protocol://TFN”, where the protocol must be a specific transfer protocol selected by SRM from the list of protocols provided by the client .

SRM behaves as a load balancer and redirector

de facto, GSI enabled FTP protocol is used for transfers

SRM PUT (ftp)



SRM

get TURL

SRM

data flow

ftp write

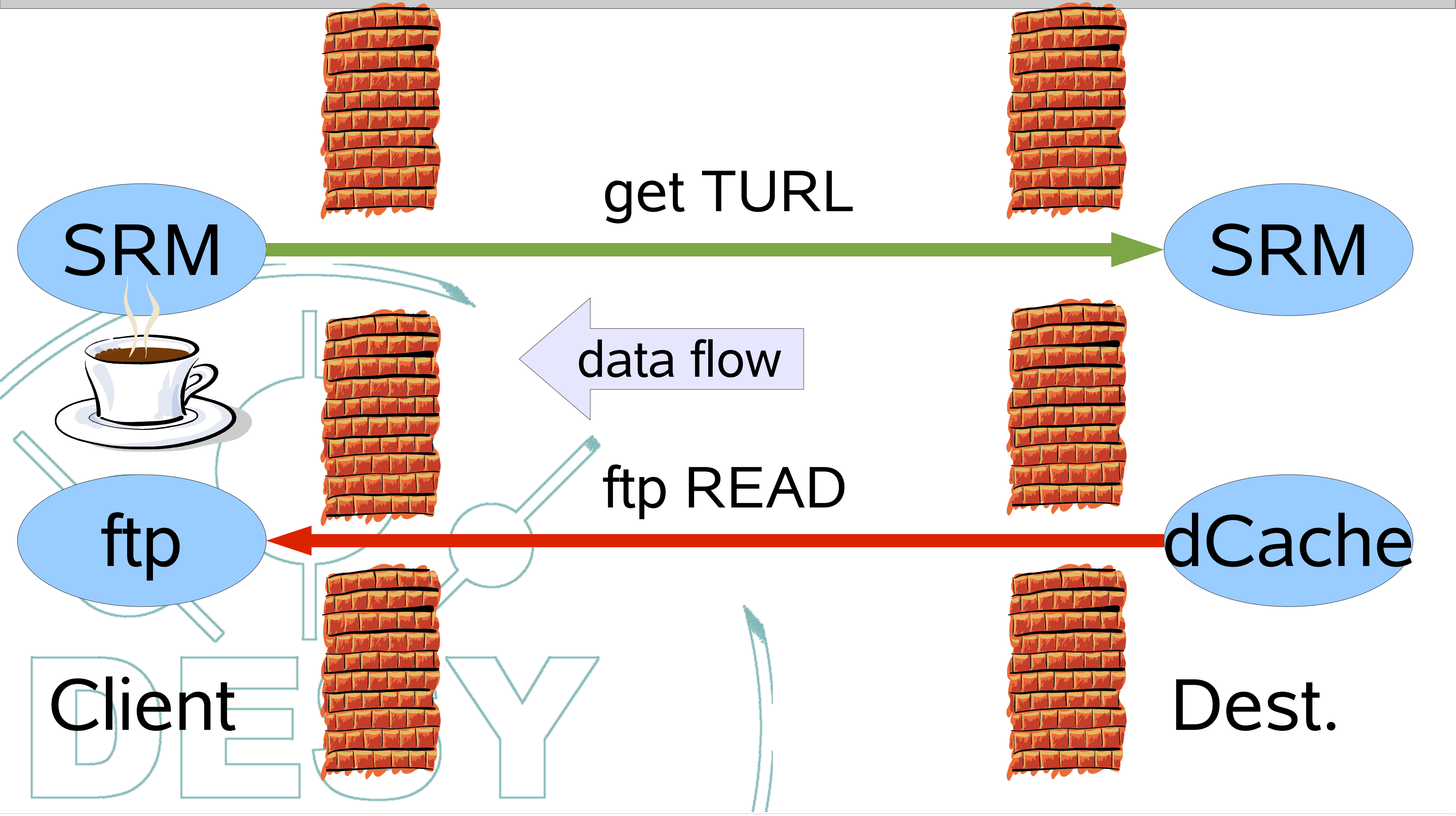
ftp

dCache

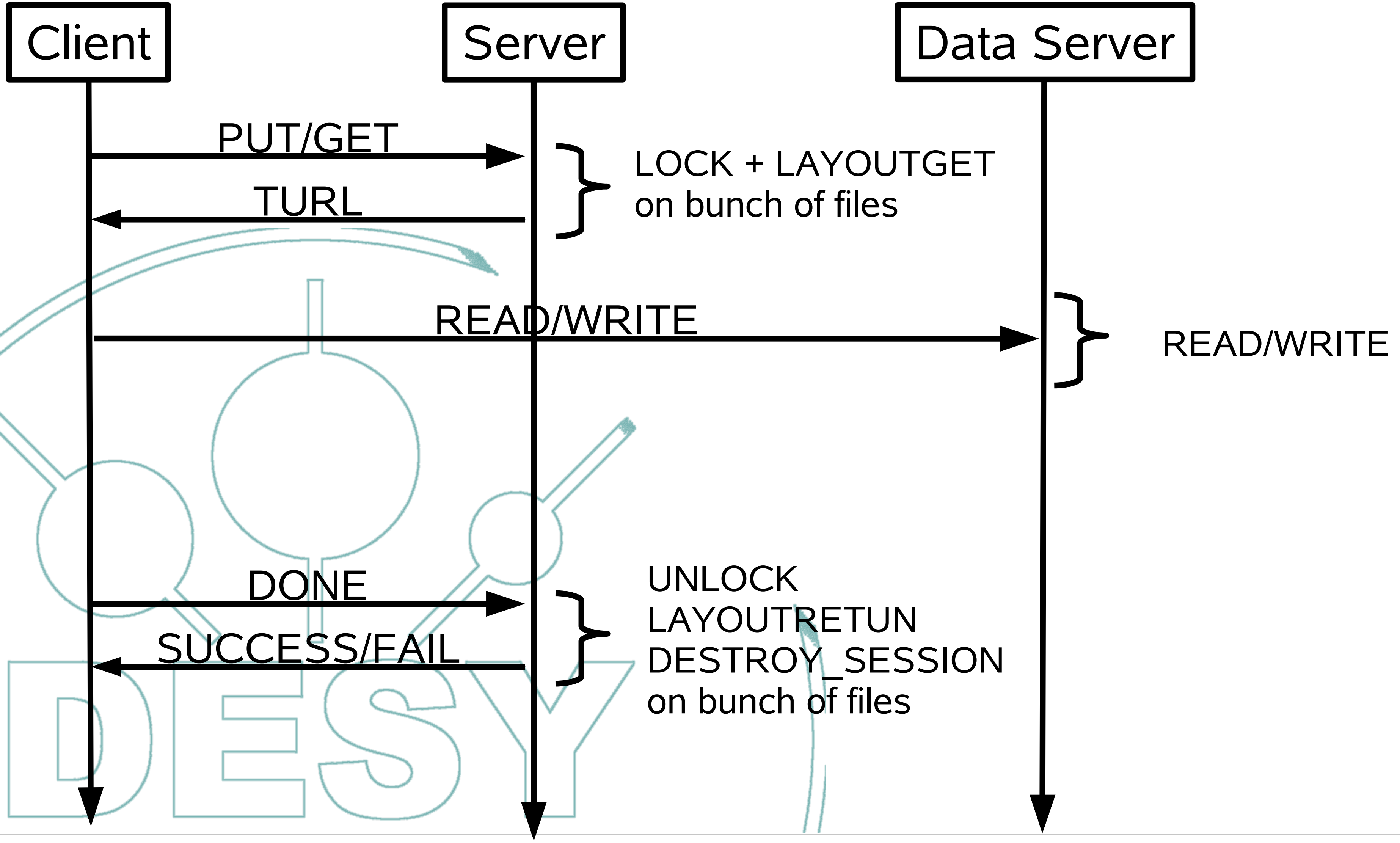
Client

Dest.

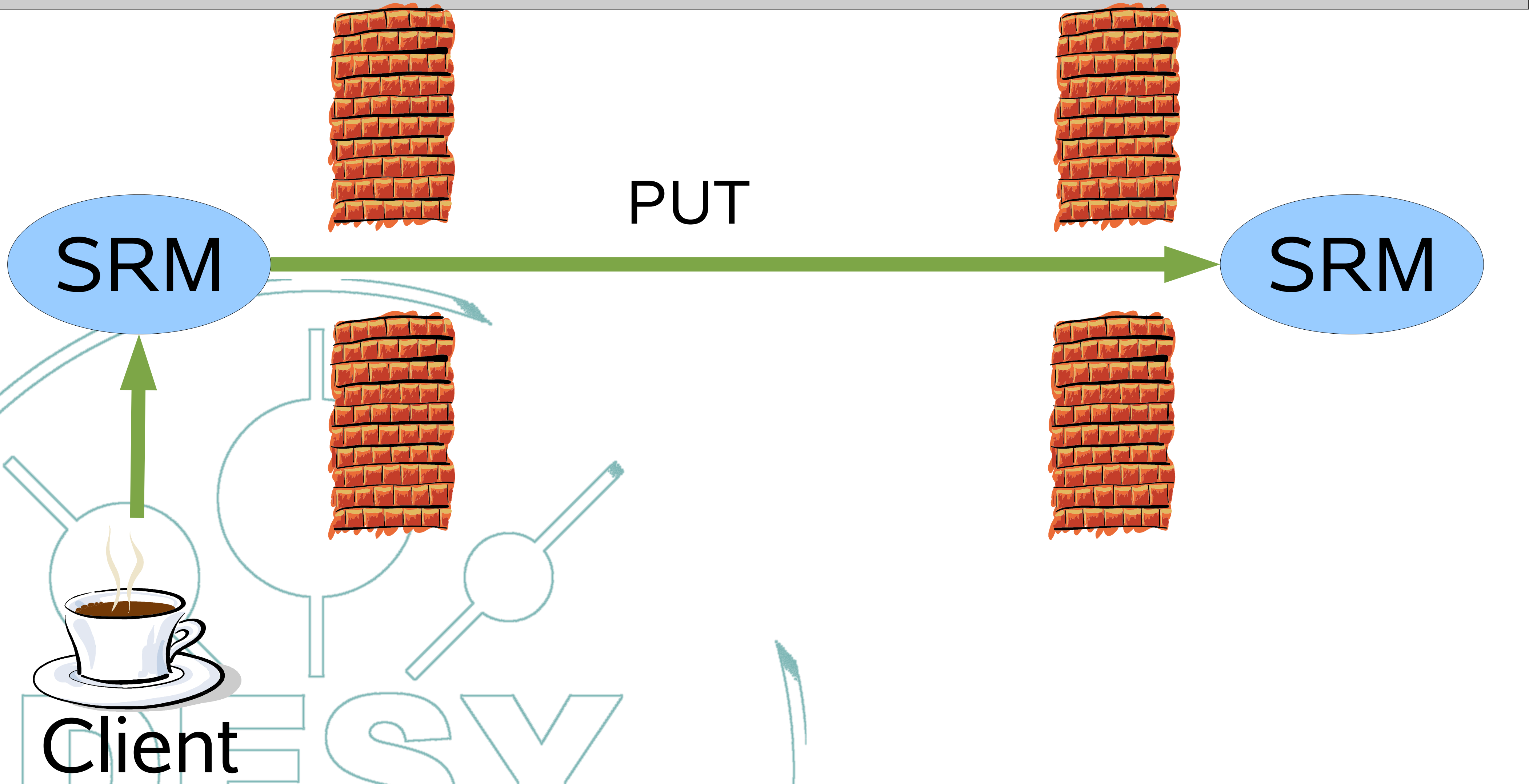
SRM GET (ftp)



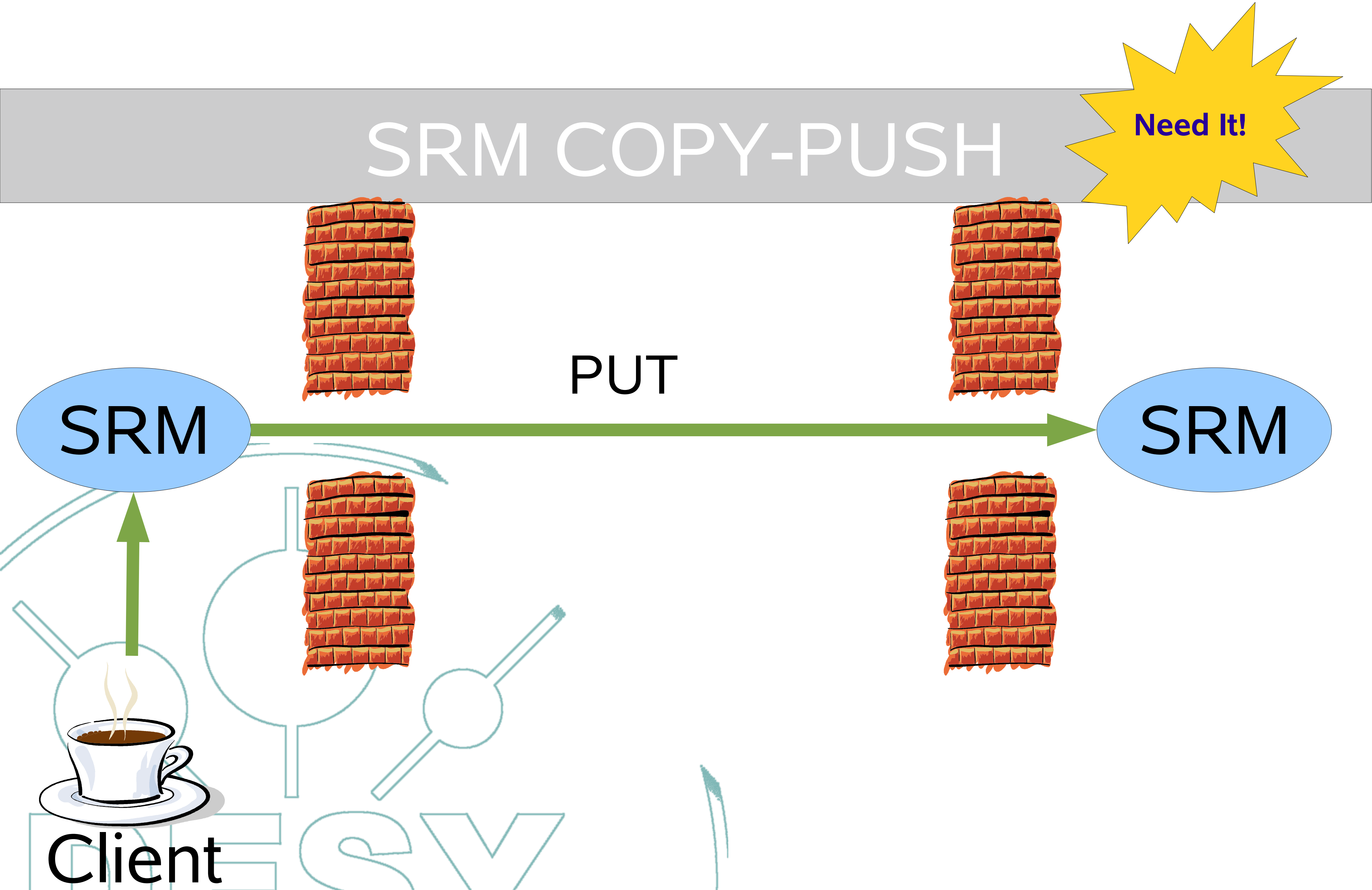
SRM for pNFS people



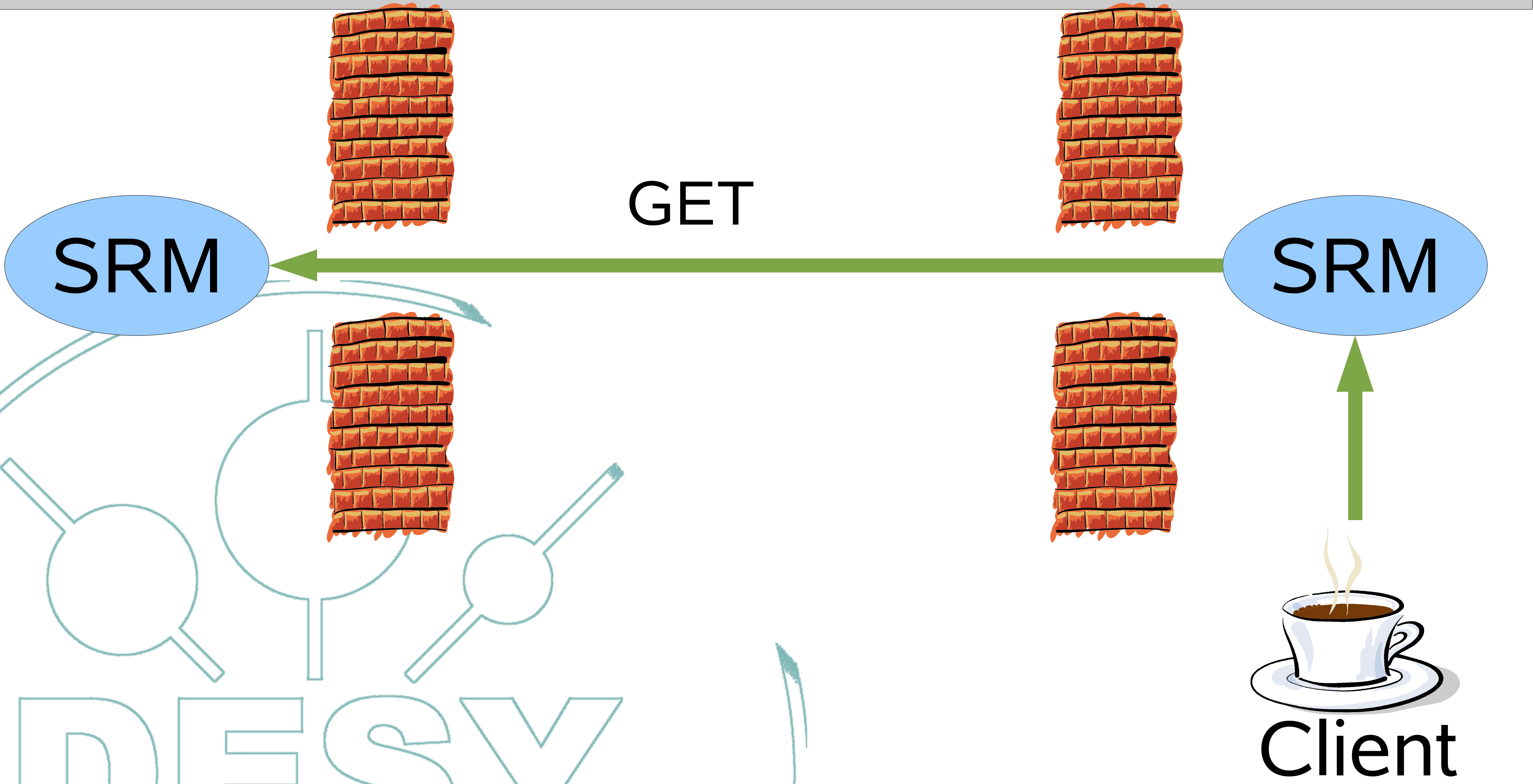
SRM COPY-PUSH



SRM COPY-PUSH

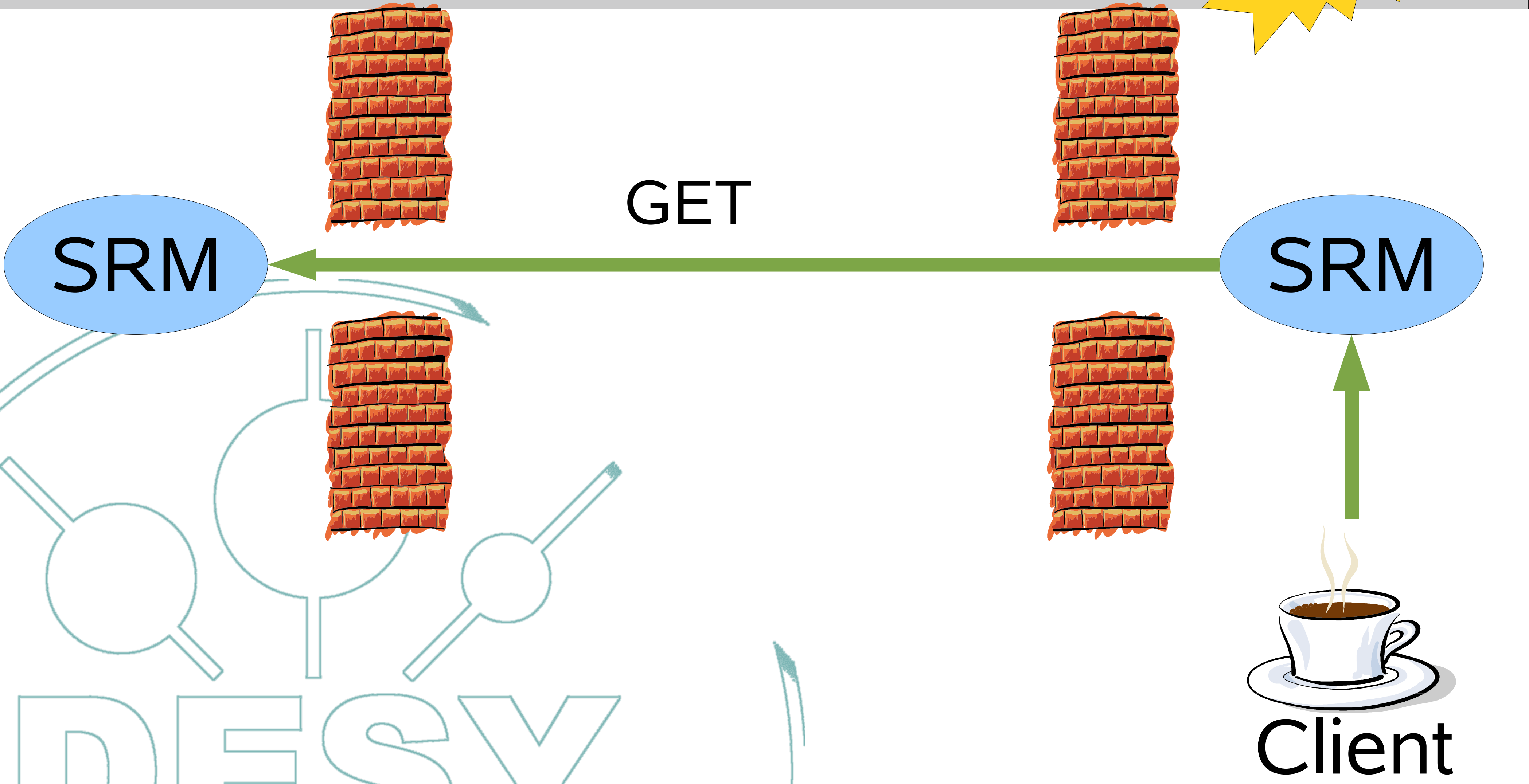


SRM COPY-PULL



SRM COPY-PULL

Need It!



SRM

GET

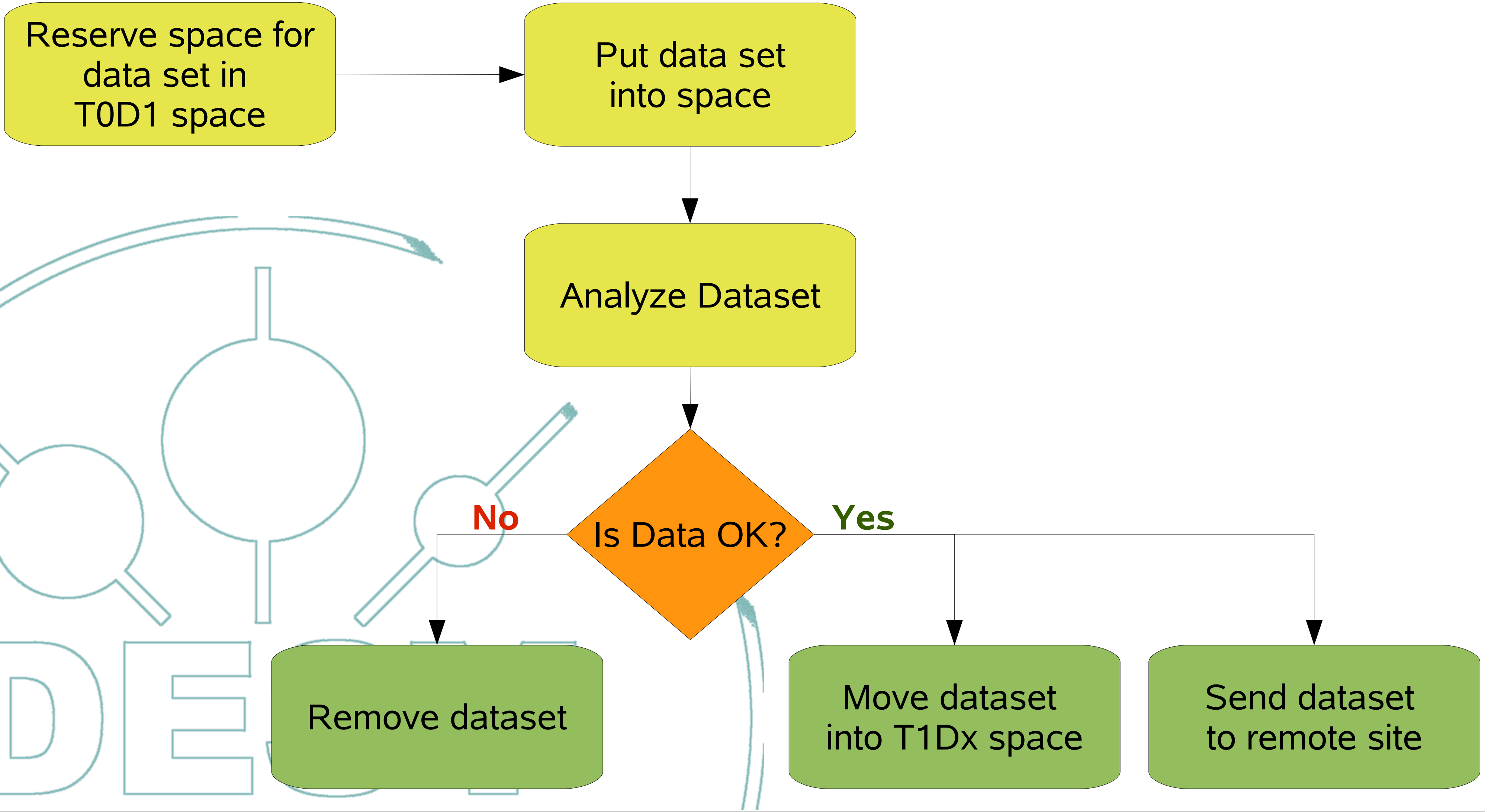
SRM

Client

SRM Space Management

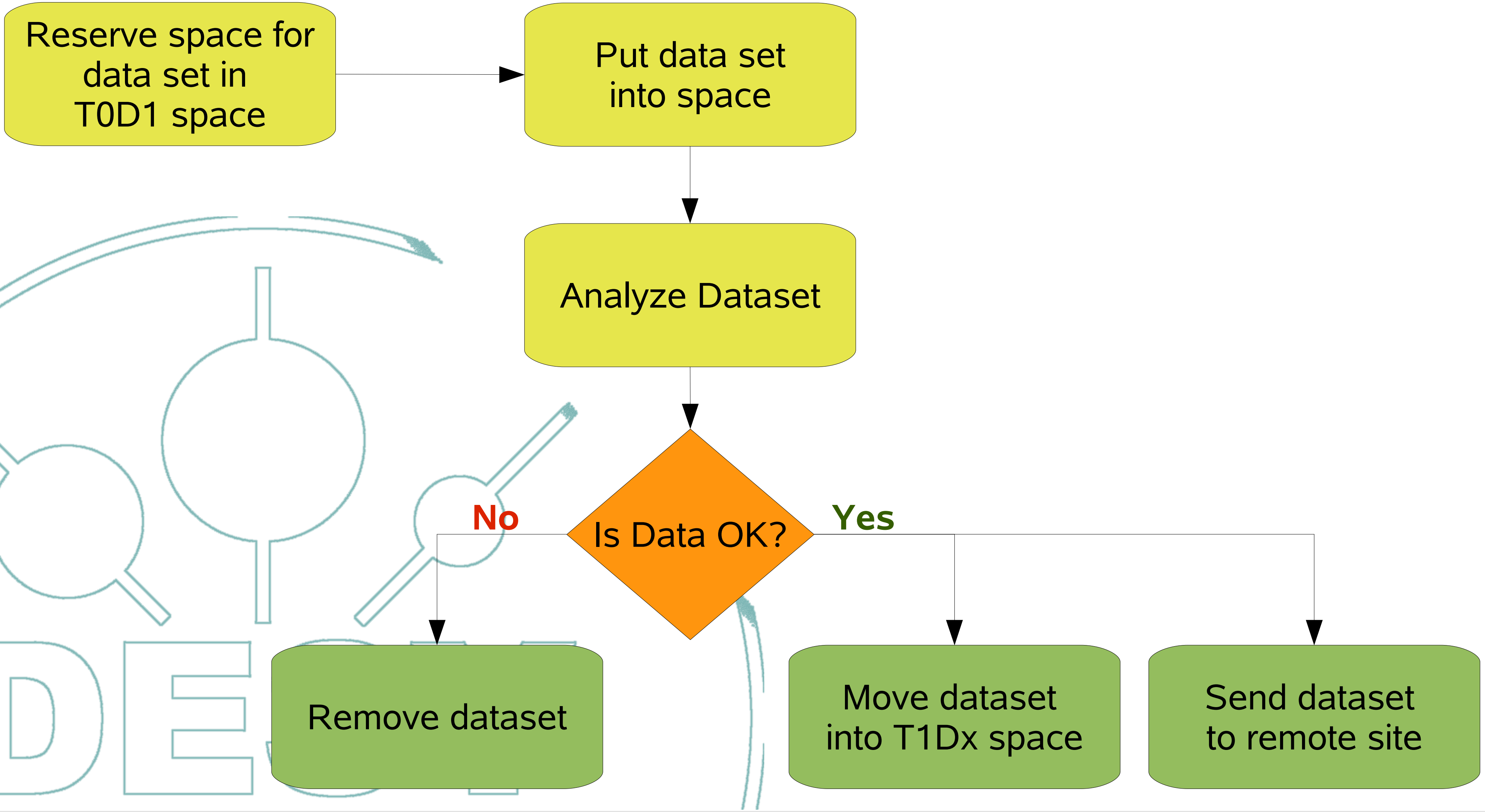
- allows to reserve space prior the transfer
  - Quota system, where you never get “file system full”
- has three space descriptions and allows transitions between them:
  - CUSTODIAL, ONLINE (Tape1Disk1)
  - CUSTODIAL, NEARLINE (Tape1Disk0)
  - REPLICA, ONLINE (Tape0Disk1)

SRM Space Management (use case)


Reserve space for data set in T0D1 space
Put data set into space
Analyze Dataset
Is Data OK?
No
Yes
Remove dataset
Move dataset into T1Dx space
Send dataset to remote site

SRM Space Management (use)

Need It!



Reserve space for data set in T0D1 space

Put data set into space

Analyze Dataset

Is Data OK?

No

Yes

Remove dataset

Move dataset into T1Dx space

Send dataset to remote site

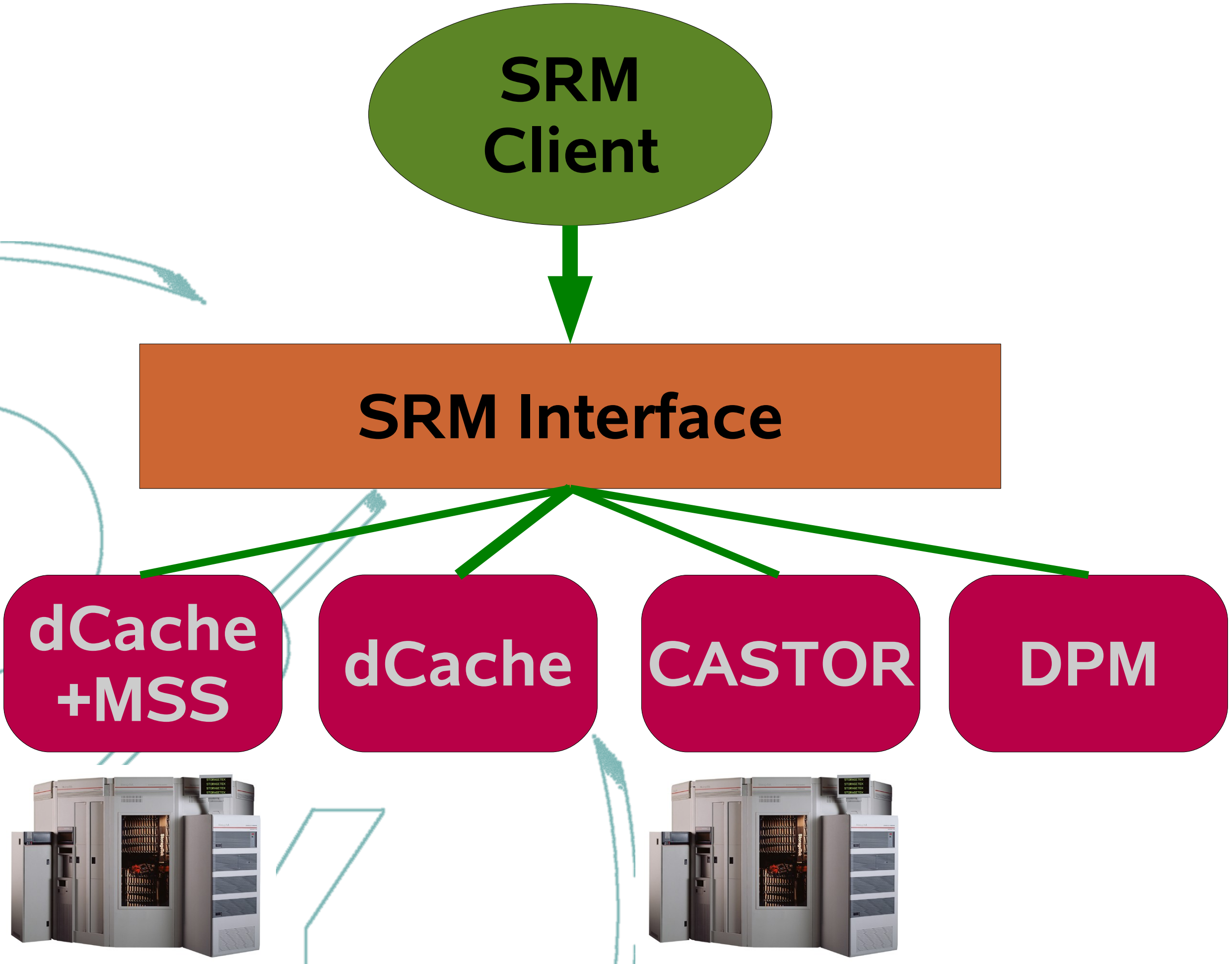
GRID Security

Need It!

- X.509 based certificates
- extensions for Virtual Organizations (VO) support
- no trusted hosts

```
subject   : /O=GermanGrid/OU=DESY/CN=Tigran Mkrtchyan/CN=proxy
issuer    : /O=GermanGrid/OU=DESY/CN=Tigran Mkrtchyan
identity  : /O=GermanGrid/OU=DESY/CN=Tigran Mkrtchyan
type      : proxy
strength  : 512 bits
timeleft  : 11:59:40
=== VO desy extension information ===
VO        : desy
subject   : /O=GermanGrid/OU=DESY/CN=Tigran Mkrtchyan
issuer    : /C=DE/O=GermanGrid/OU=DESY/CN=host/grid-voms.desy.de
attribute : /desy/Role=NULL/Capability=NULL
timeleft  : 11:59:40
```

SRM – Uniform Data Access



SRM Client

SRM Interface

dCache +MSS

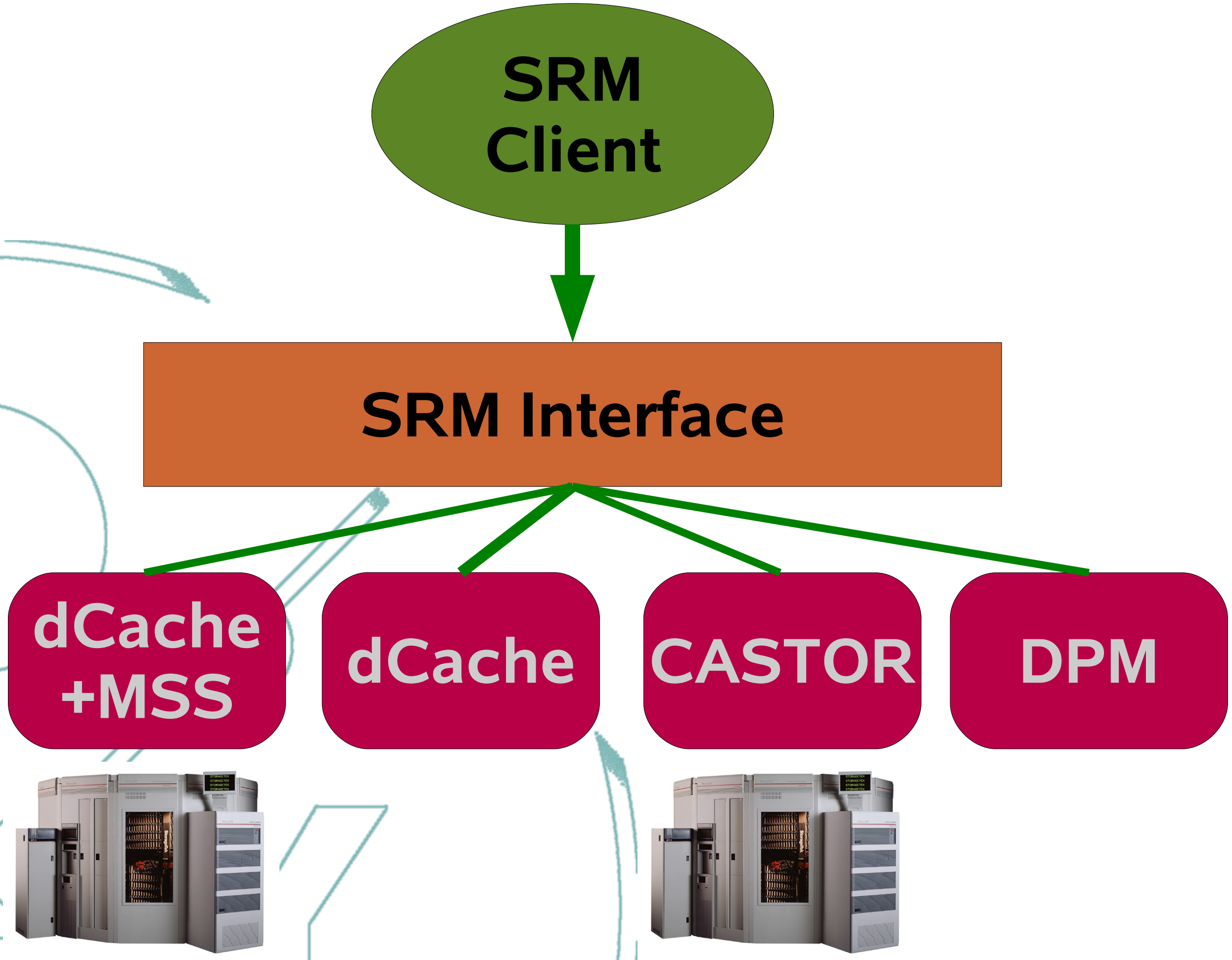
dCache

CASTOR

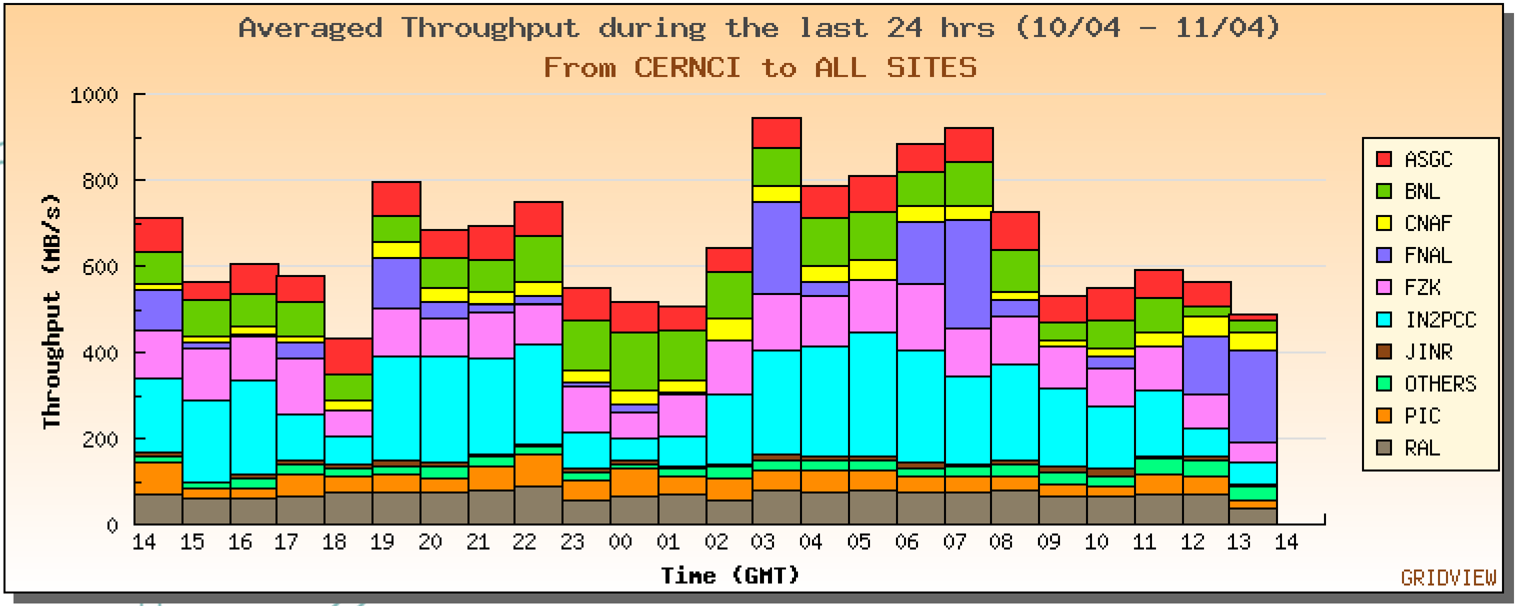
DPM

SRM – Uniform Data Acce

Need It!



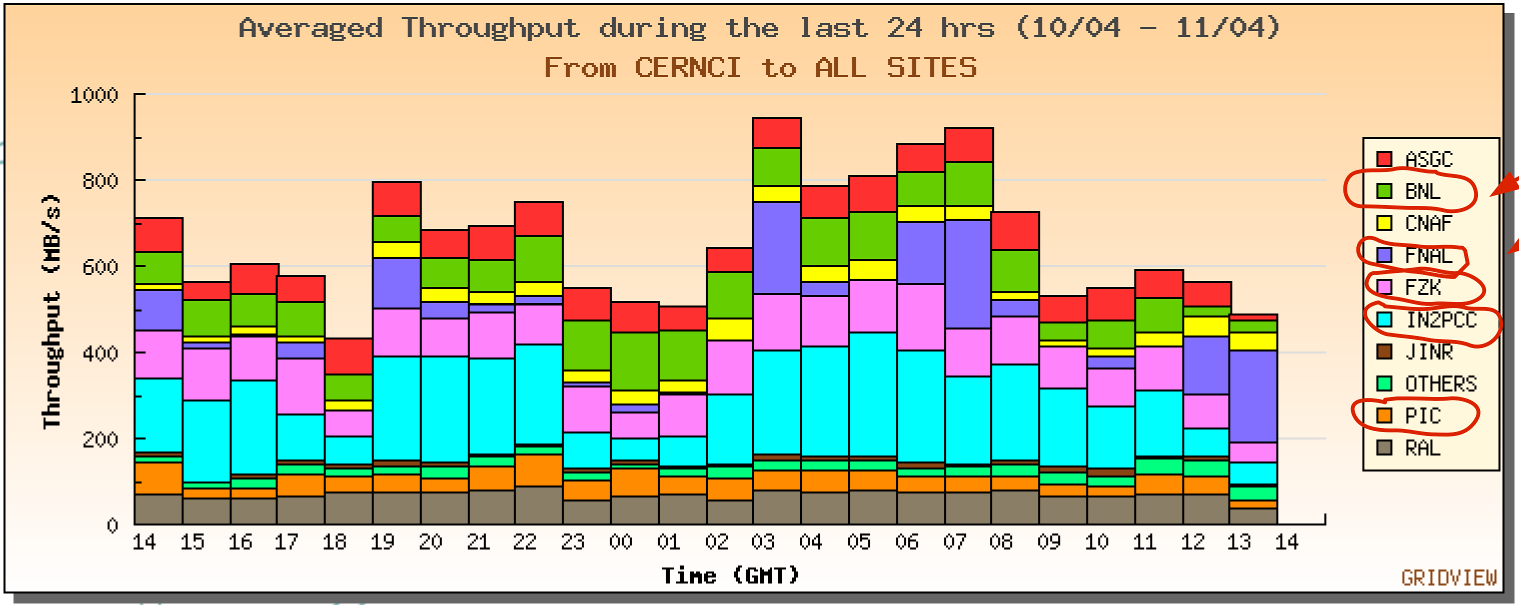
Mission i~~m~~Possible



We are doing well!

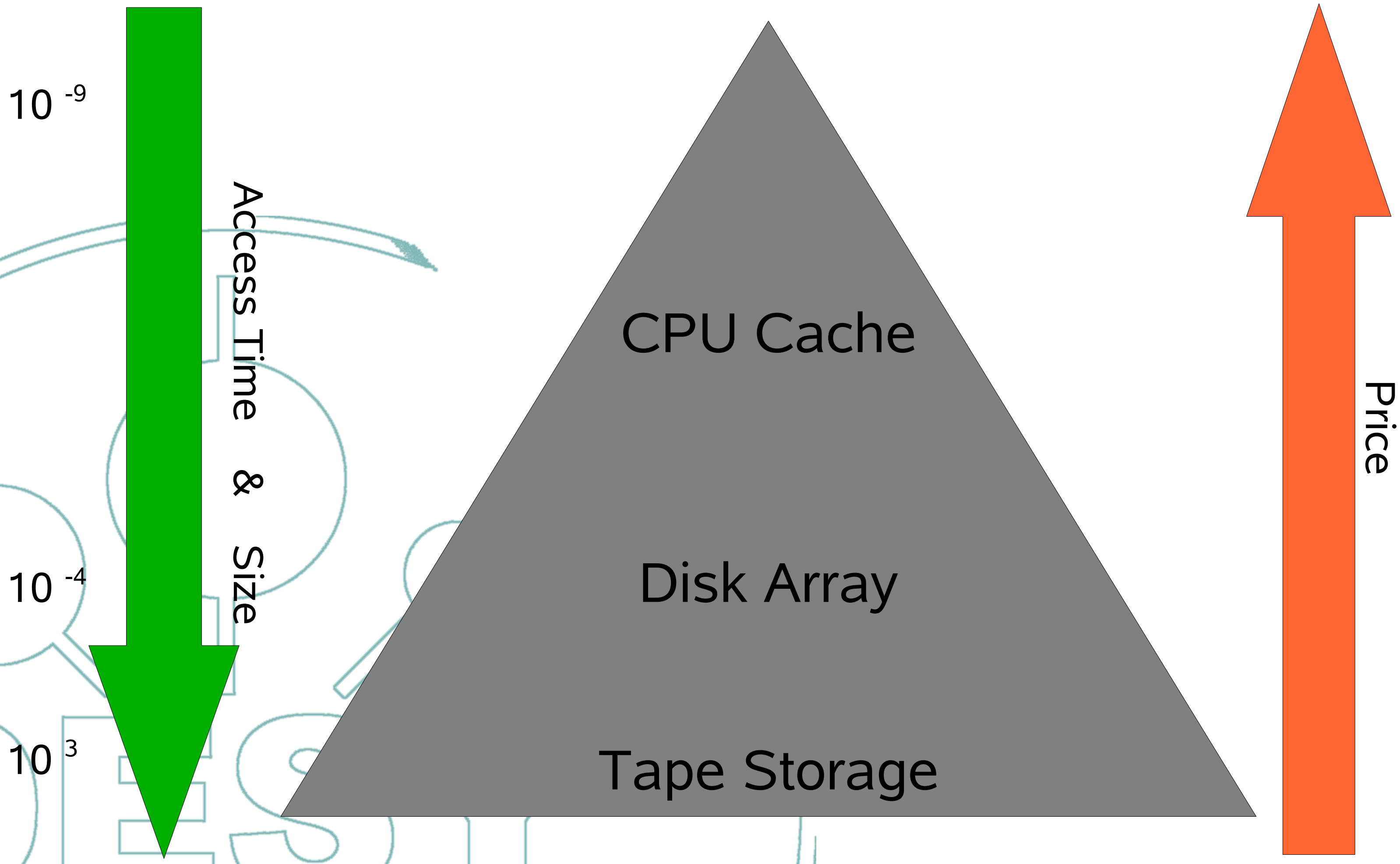
Mission i~~m~~Possible

dCache installations



We are doing well!

dCache - Background



10^{-9}

10^{-4}

10^{3}

Access Time & Size

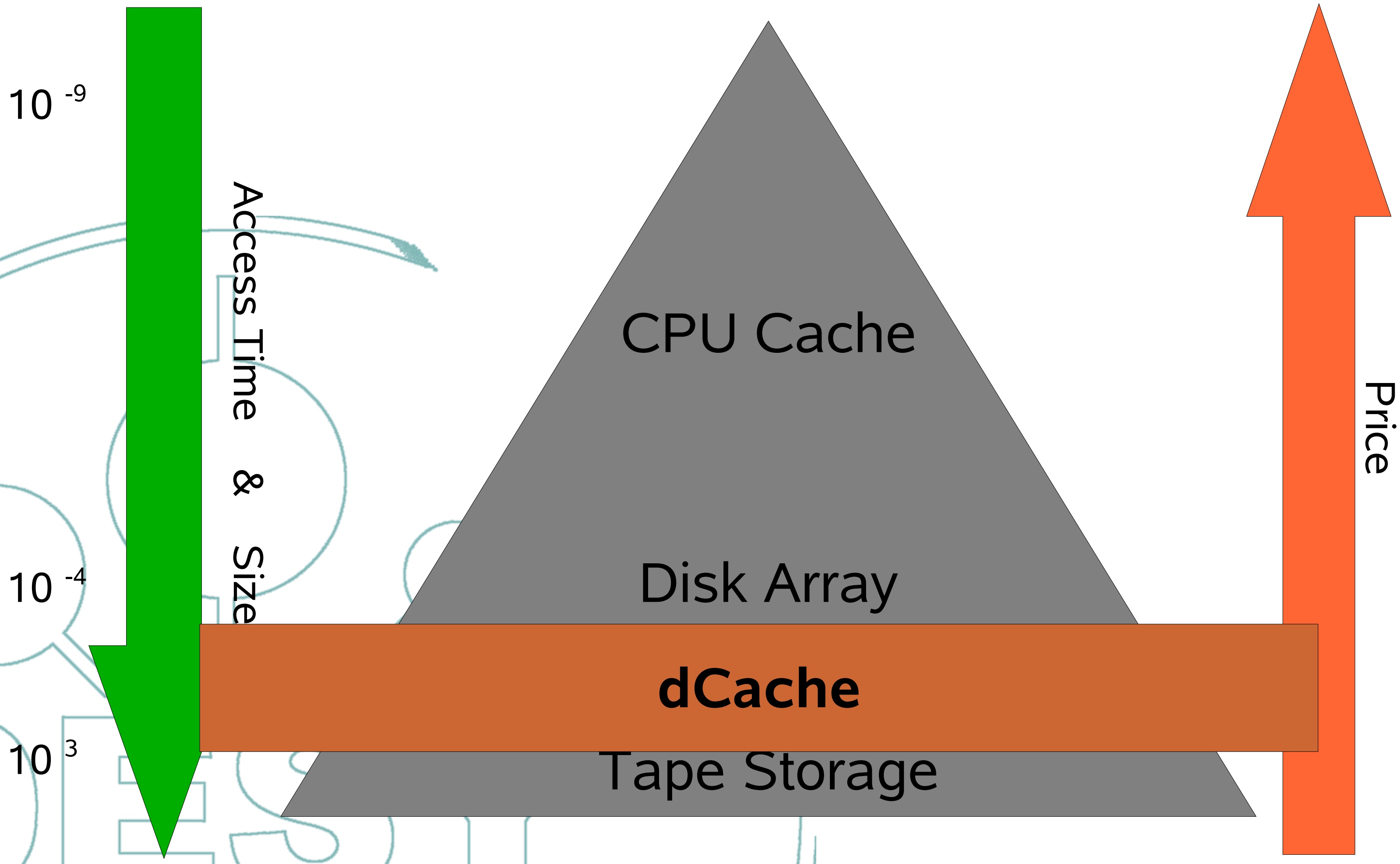
CPU Cache

Disk Array

Tape Storage

Price

dCache - Background

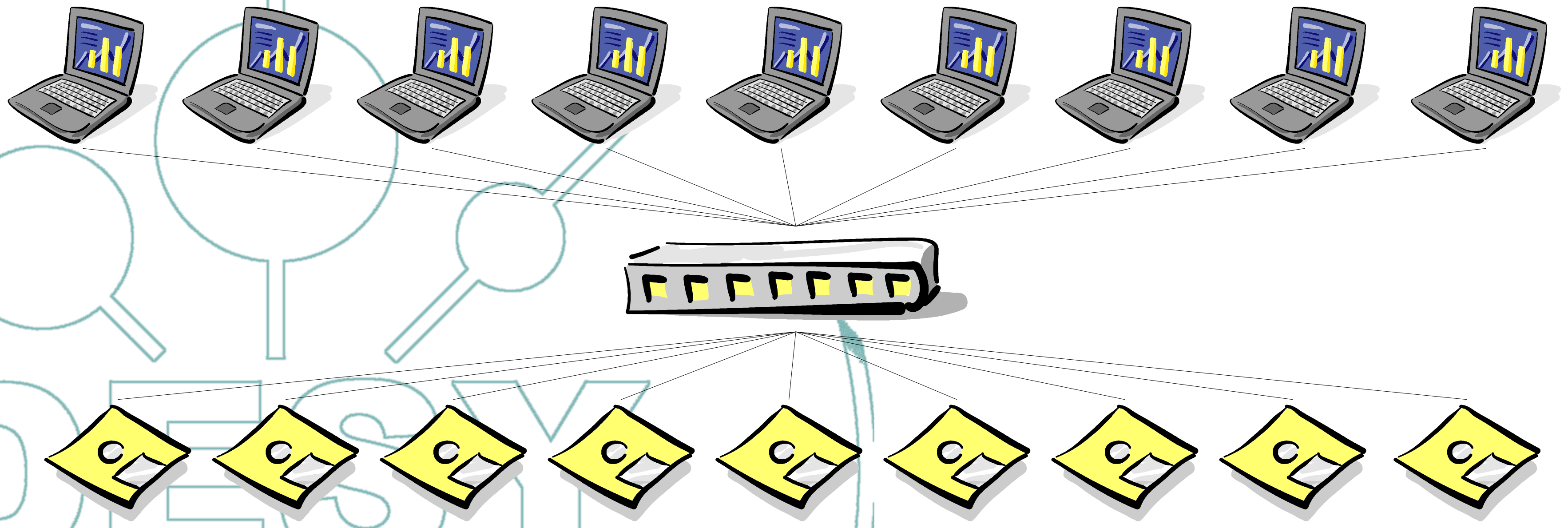


The goal of the project is:

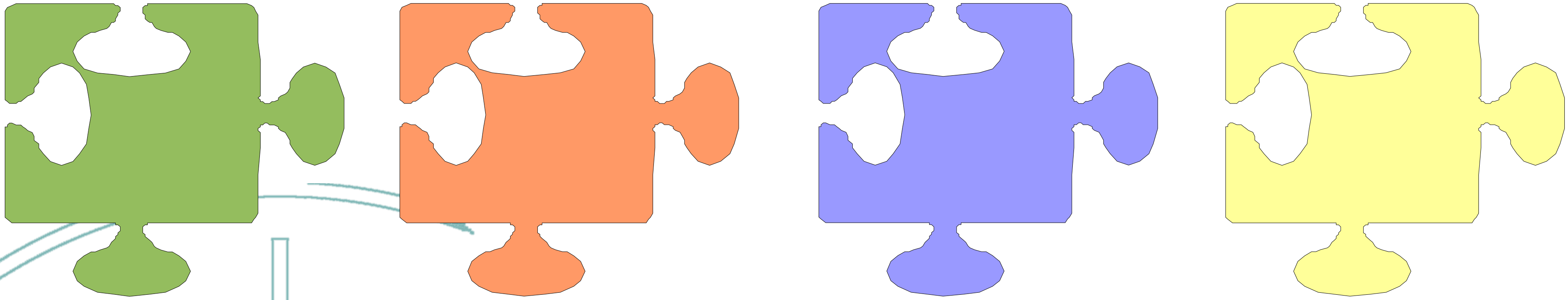
- to share and optimize access to non-sharable storage devices, like tape drives,
- make use of slower and cheaper drive technology without overall performance reduction,
- to provide a system for storing and retrieving huge amounts of data, distributed among a large number of heterogeneous server nodes, under a single virtual filesystem tree with a variety of standard access methods.

Requirement is:

to provide a system for storing and retrieving huge amounts of data, distributed among a large number of heterogeneous server nodes, under a single virtual filesystem tree with a variety of standard access methods.

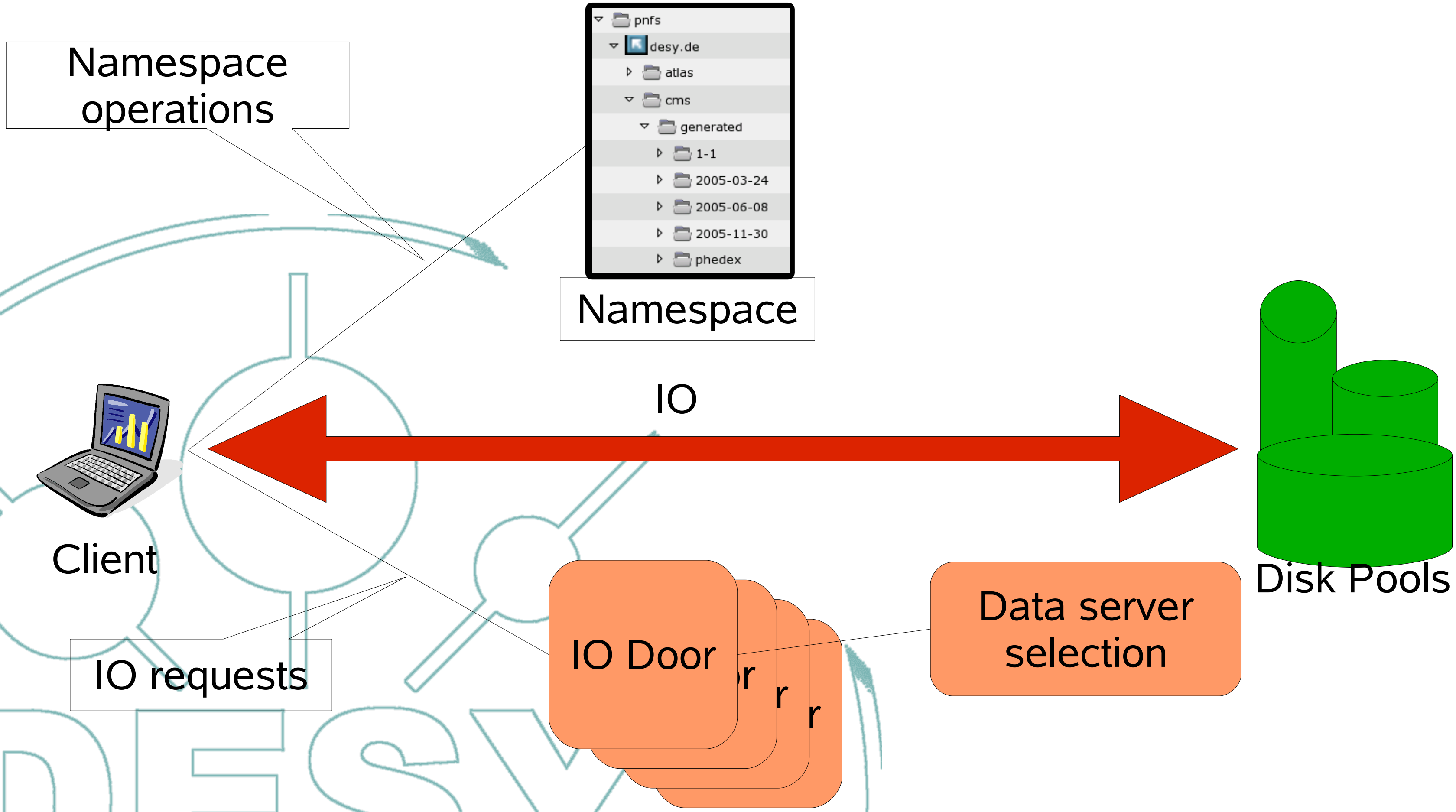


dCache Design



- Name Space Provider
  - size, owner, acls, checksum, ...
- Pool Selection Unit
- Protocol Specific Doors
- Multiprotocol Pools
  - can talk several protocols simultaneously

dCache Design



Namespace operations

pnfs
desy.de
atlas
cms
generated
1-1
2005-03-24
2005-06-08
2005-11-30
phedex

Namespace

IO

Client

Disk Pools

IO requests

IO Door

Data server selection

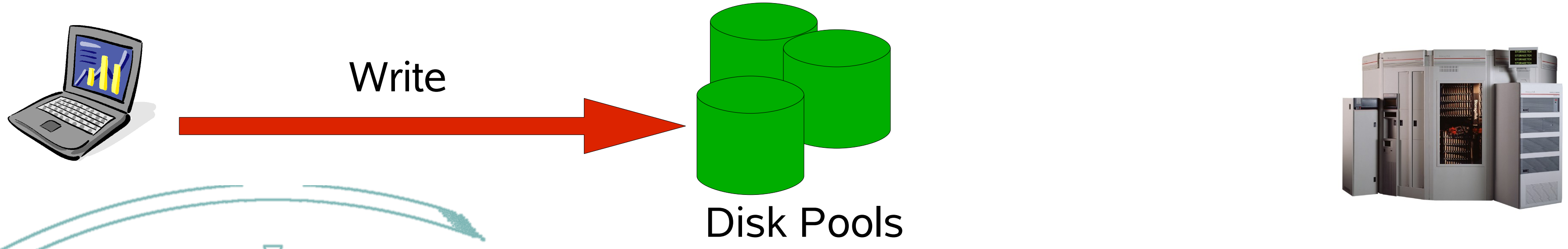
dCache Design

- Pools are grouped into PoolGroups
- PoolGroup selected by flow direction, 'path'(file set), protocol and client IP
- Pool selected by ***cost***, where cost is

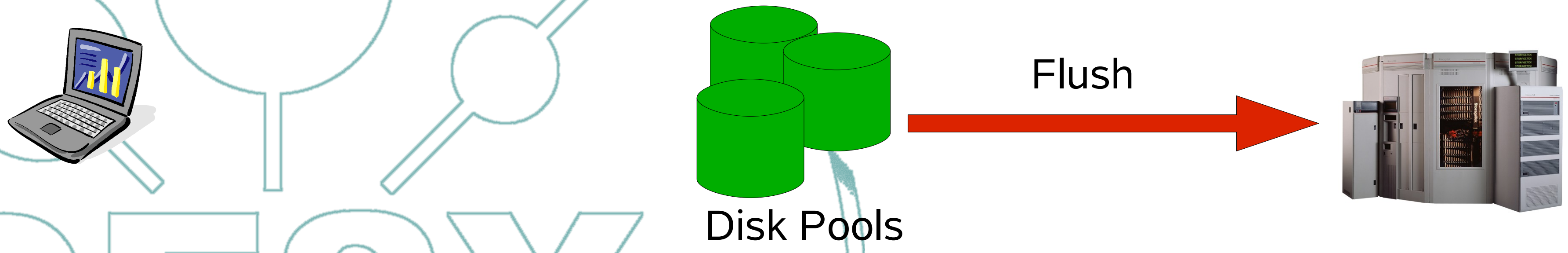
n*<CPU cost> + m*<space cost>

n=1, m=0 : fill network bandwidth first
n=0, m=1 : fill empty servers first

MSS connectivity

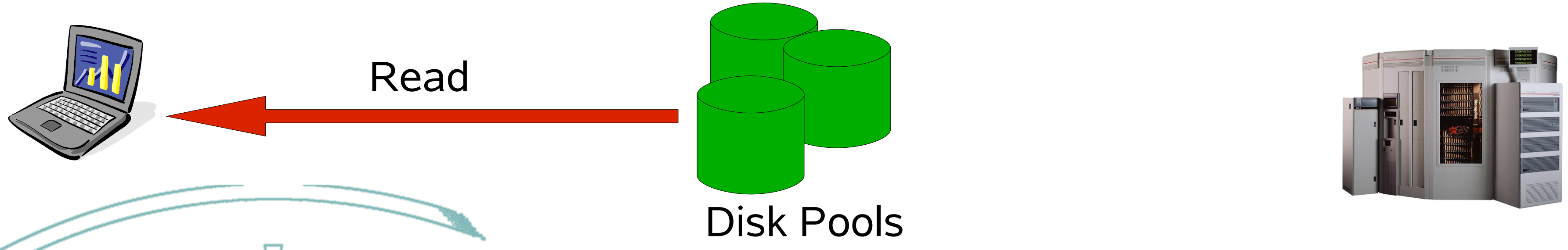


Files arrives to a pool and declared as ***Precious***



Precious files flushed according policy - time, size, number of files.

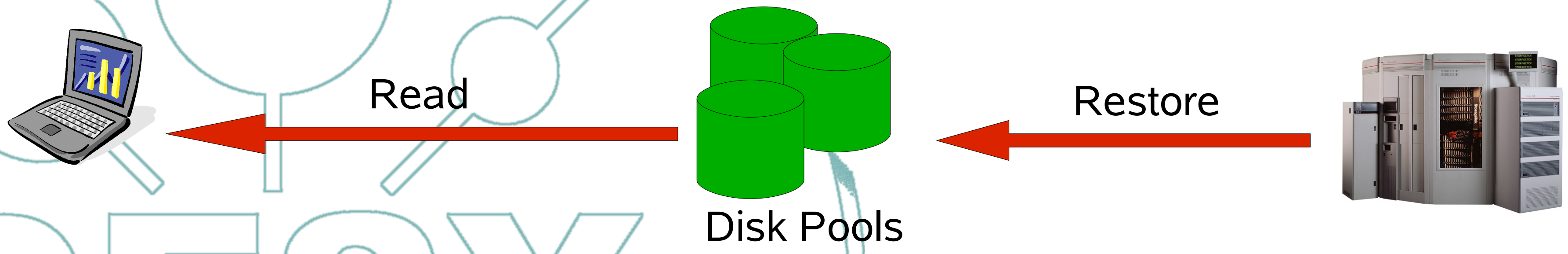
MSS connectivity



Read

Disk Pools

Cached files can be delivered immediately



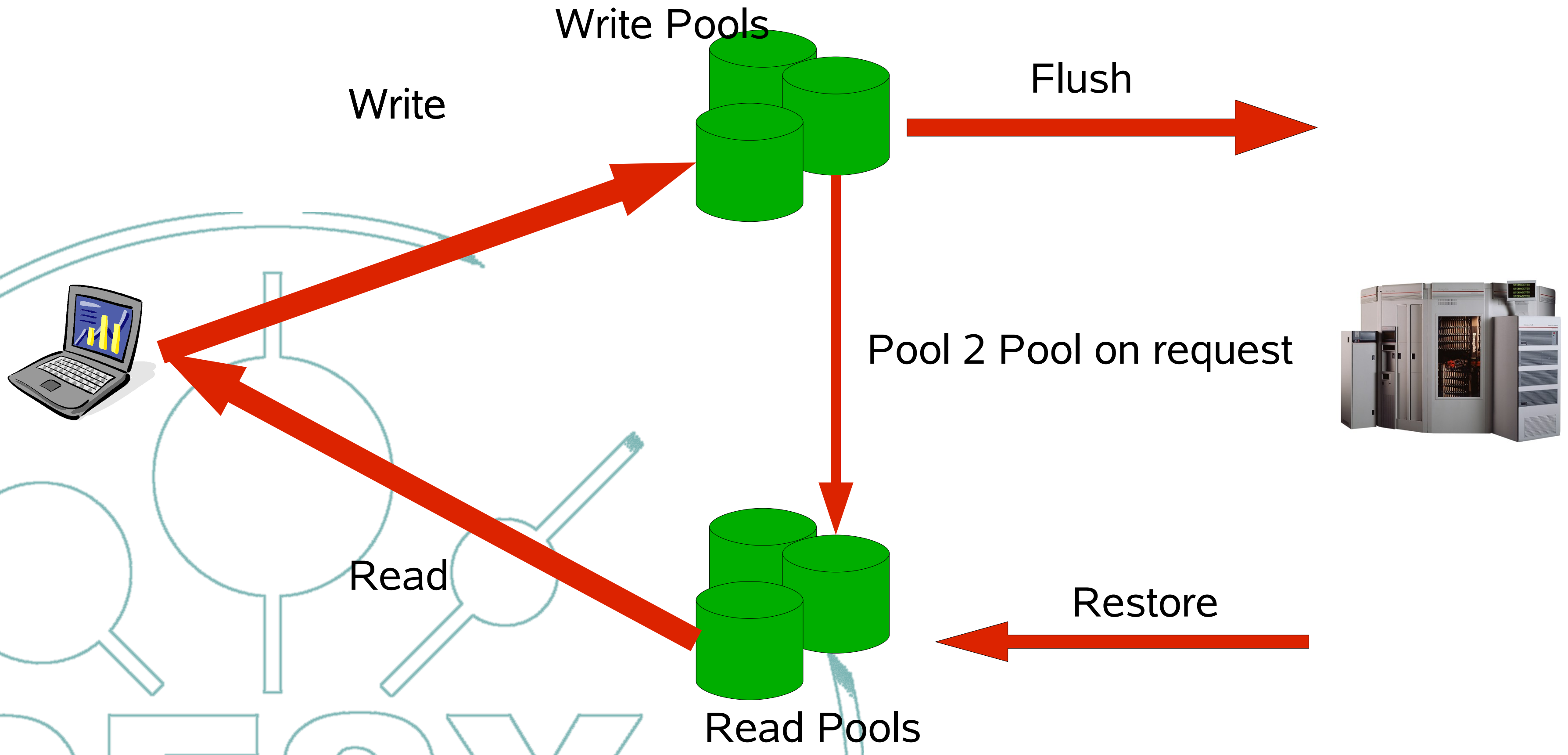
Read

Disk Pools

Restore

Missing files retrieved from the MSS first

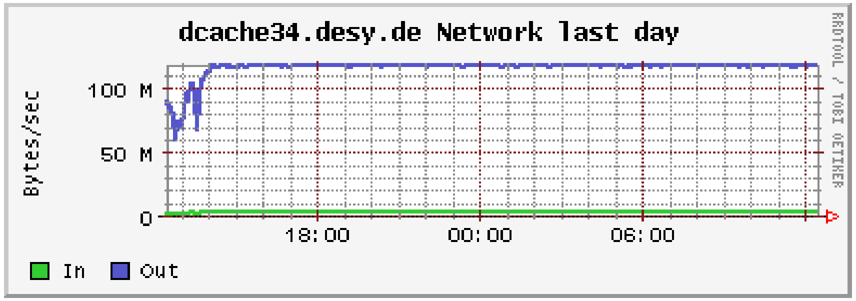
MSS connectivity



Current Status

- dCache let us build very large (capacity and bandwidth wise) storage system with small, independent building blocks
- building block need to provide:
  - JVM >= 1.5 (all components are Java based)
  - local filesystem
  - network Interface

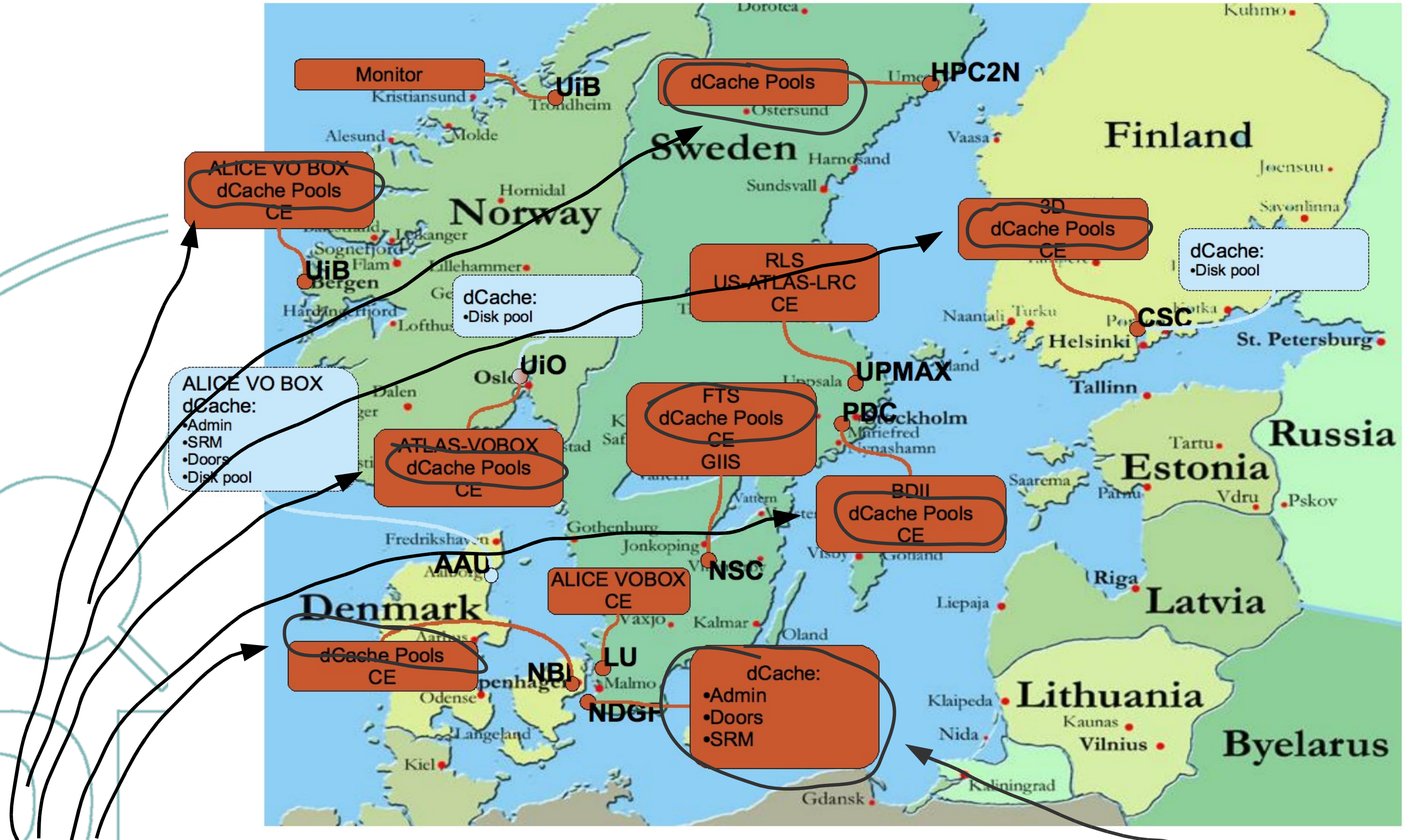
no IO penalty while using Java



Current Status

- Project started June of 2000 as a join effort of DESY and FNAL
- First prototype April 2001
- In Production since March 2002
- Supported local access Protocols: dcap, xrootd
- Supported WAN access Protocols: ftp, http
- Deployed on AIX, Linux (x86, Power, x64), Solaris (Sparc, AMD)
- Run over country border
- Has an interface to OSM, Enstore, HPSS , TSM, DMF
  - easy to add any other MSS
- Largest Installation 2PB (FNAL)
- ~1800 pools
- ~1.2 GB/s WAN (Peak rate – 2.5 GB/s!)
- 60 TB/day read (100000 files!)
- 2 TB/day write (8000 files)

Current Status (NorduGrid)



Monitor

UiB

dCache Pools

HPC2N

ALICE VO BOX
dCache Pools
CE

UiB

dCache:
- Disk pool

RLS
US-ATLAS-LRC
CE

UPMAX

3D
dCache Pools
CE

CSC

dCache:
- Disk pool

UiO

ALICE VO BOX
dCache:
- Admin
- SRM
- Doors
- Disk pool

ATLAS-VOBOX
dCache Pools
CE

FTS
dCache Pools
CE
GIIS

PDC

BDII
dCache Pools
CE

NSC

AAU

ALICE VOBOX
CE

LU

dCache Pools
CE

NBI

NDGF

dCache:
- Admin
- Doors
- SRM

dCache Pools

dCache Core Service

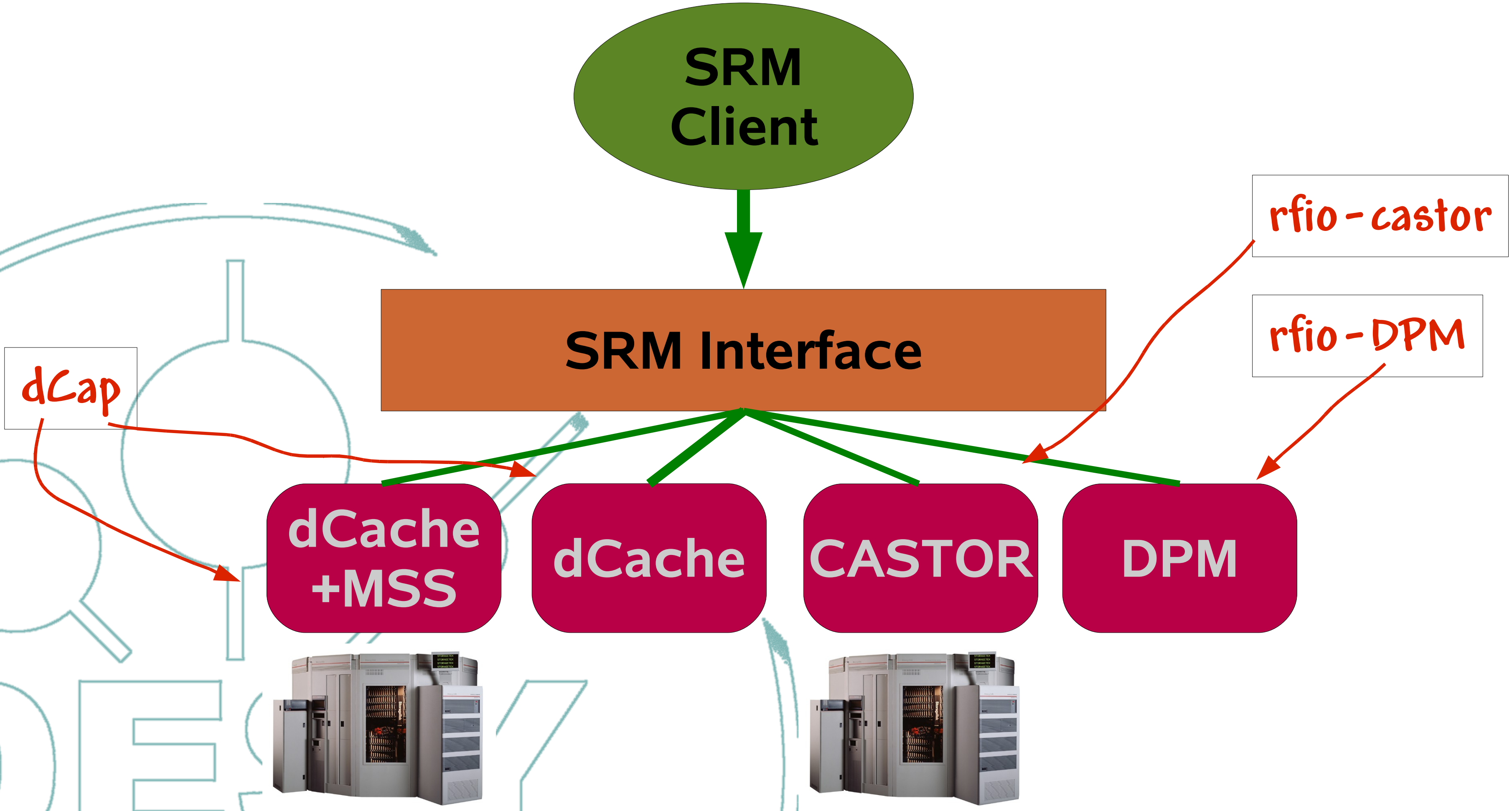
Pnfs != pNFS

The dCache's **Namespace** provider called Pnfs:

Perfectly Normal File System

developed in 1997 and currently replacement.

Uniform Data Access



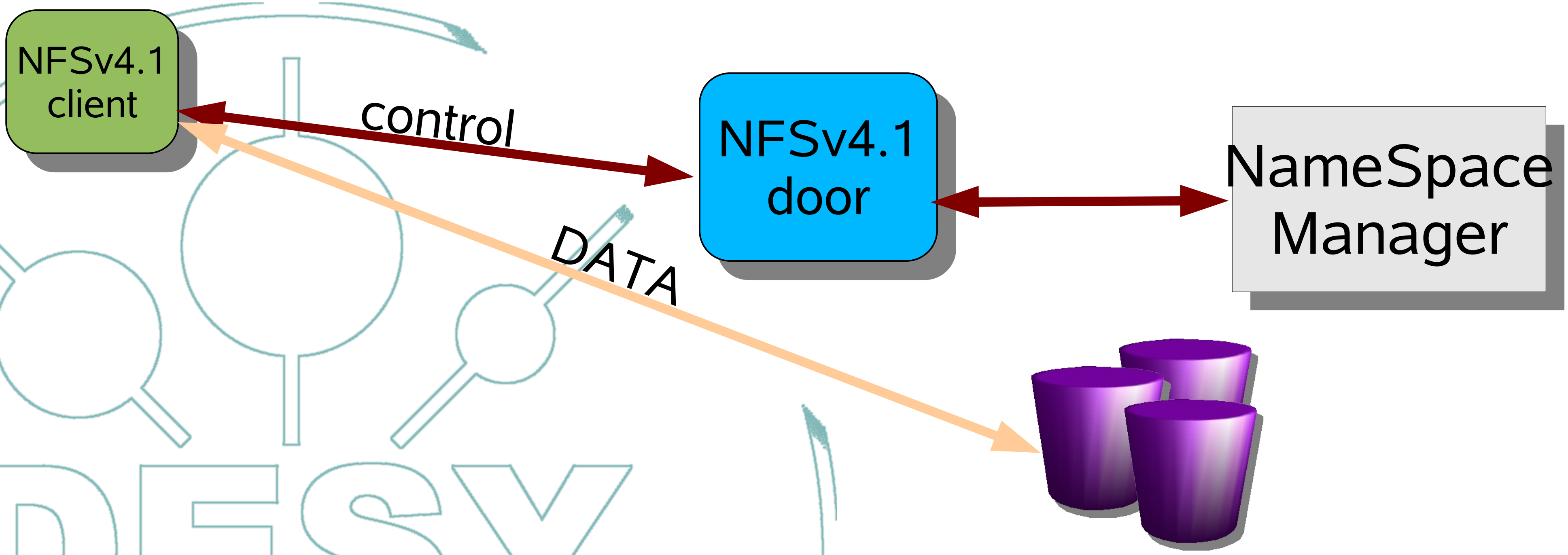
Why new protocols

- There is a three 'popular' protocols used in High Energy Physics:
  - dCap – dCache Access Protocol
  - rfio – Remote File IO
  - xroot – eXtended ROOT IO

- all protocols was designed, while NFSv2/3 was not distributed
- existing distributed solutions not fit well
  - and expensive (all of them)
  - and require special hardware
  - or require special OS/kernel versions

NFSv4.1

- fit well to dCache (and others) architecture
- Open Standard Protocol supported by industry NFSv4.1
- Client comes 'for free' with Operating System



NFSv4.1 client

control

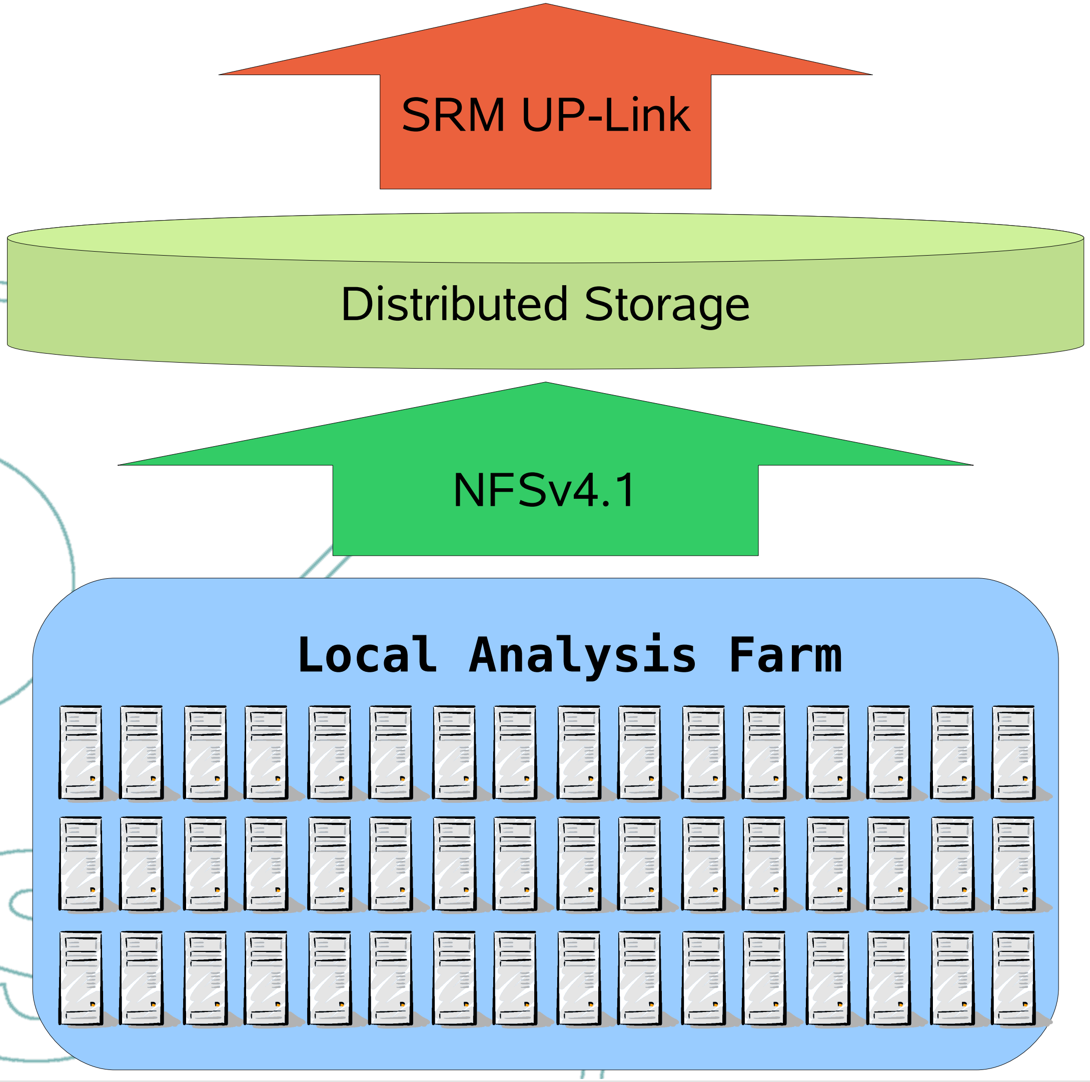
DATA

NFSv4.1 door

NameSpace Manager

The Vision:

Need It!



References:

- www.dCache.ORG
- SRM V2.2 spec. http://sdm.lbl.gov/srm-wg/doc/SRM.v2.2.html
- NFSv4.1 spec. http://www.nfsv4-editor.org/

Special Tanks to:

Andy Adamson (CITI)
Benny Halevy (Panasas)
Lisa Week (SUN)
Sam Falkner (SUN)
Robert Gordon (SUN)