Usage of Volatile Pools in Belle-II

Dr. Silvio Pardi

INFN-Napoli

DOMA / ACCESS Meeting

CERN -27/12/2018

INFN

Istituto Nazionale di Fisica Nucleare

SCoRES Project

Goal of the activity is to setup and test an HTTP Caching system and investigate how to integrate it in the HEP computing model. Pilot experiment is Belle II.

Activities are carrying on in the context of a project funded by GARR within a National call consisting in a 2Year fellowship.

Davide Michelino - project fellowship
Silvio Pardi – Project Tutor for INFN-Napoli

Consortium GARR

INFN
Istituto Nazionale di Fisica Nucleare

Caching laboratory with DPM

- DPM 1.9 with Dome will allow investigation of operating WLCG storage as a cache
- Scenarios
  - Data origin a regional federation of associated sites
  - Data origin the global federation
- A volatile pool can be defined which calls out to a stager on a miss
  - Caching logic implemented in a pluggable way
  - Hybrid cache/conventional setup
- Questions to investigate
  - Cache management logic
  - Different client strategies on miss
    - blocking read, async read, redirection to origin
  - Authentication solutions
  - Workflow adaptation for locality

CHEP 2016

We are trying to answer at these questions

Concept of DPM Volatile Pool

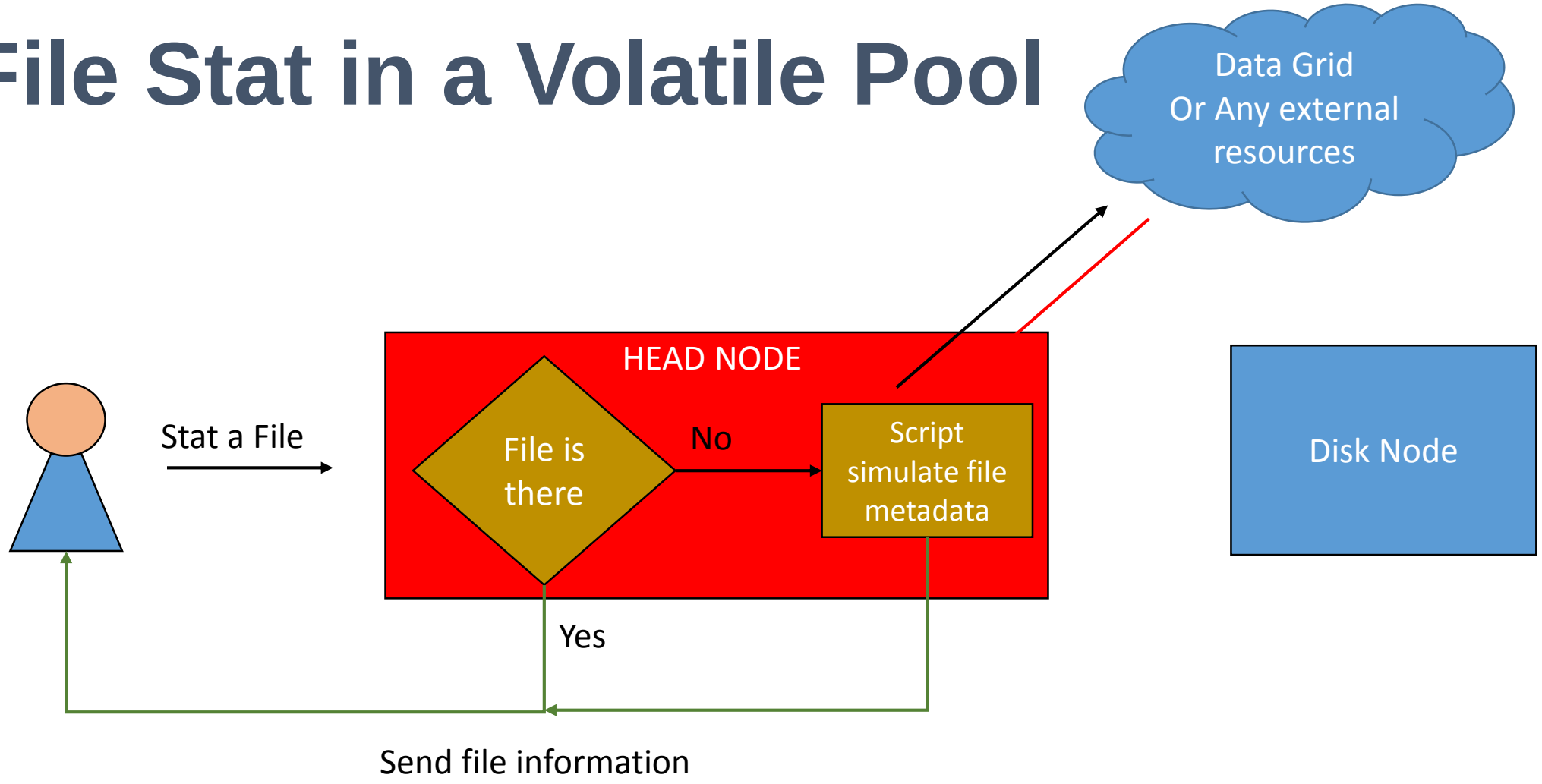
A **Volatile Pool** is a special storage area in a DPM system that can download files from external sources when clients ask for them.

Two main scripts configurable by the system admin:

- **A script running on DPM head node that manages the stat operations**
- **A script running in Disk Nodes responsible to get file from external sources**

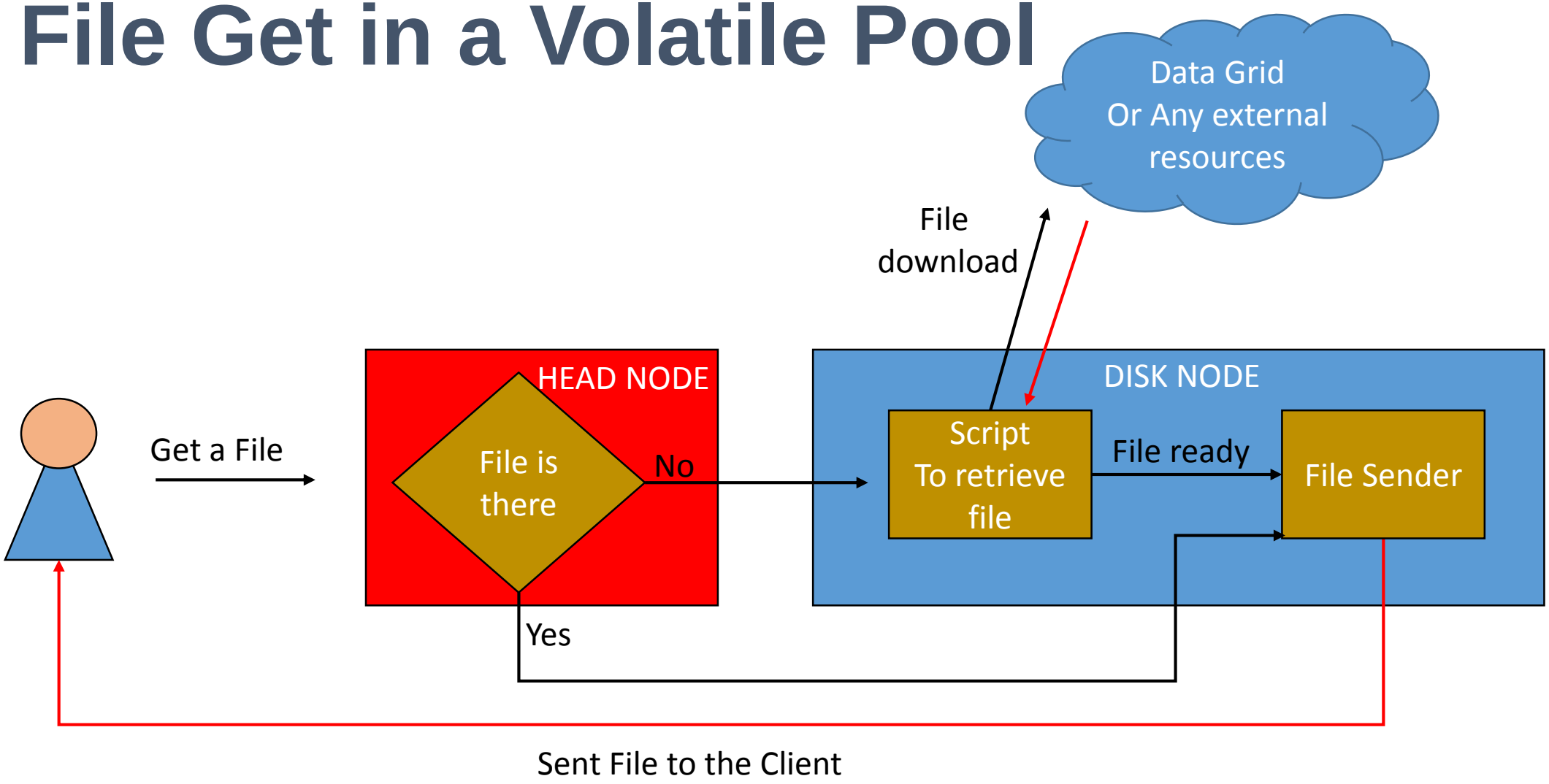
INFN
Istituto Nazionale di Fisica Nucleare

File Stat in a Volatile Pool



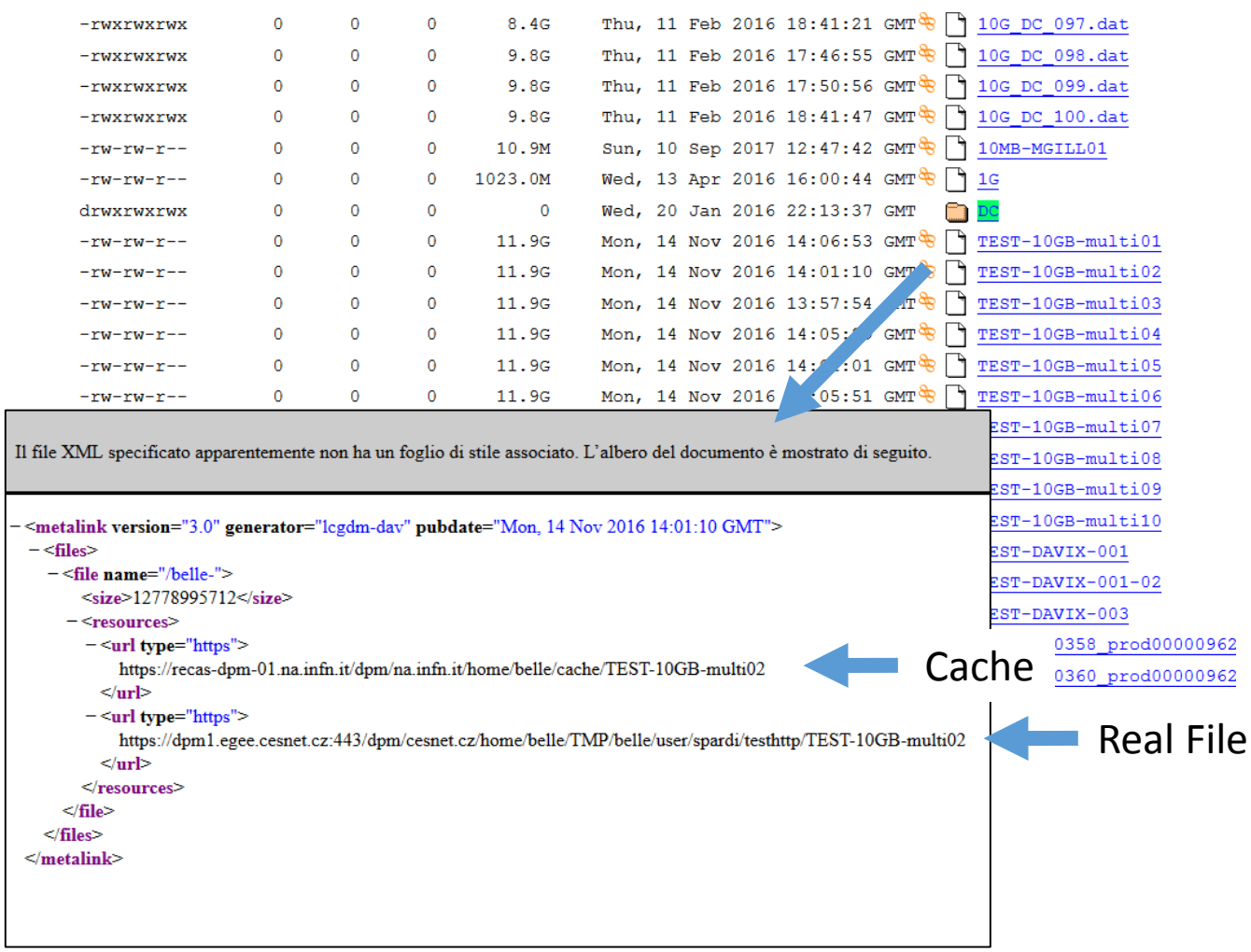
INFN
Istituto Nazionale di Fisica Nucleare

File Get in a Volatile Pool



INFN
Istituto Nazionale di Fisica Nucleare

Dynafed + Volatile Pool



What happen if we aggregate a set of standard http endpoints with a DPM Volatile Pool?

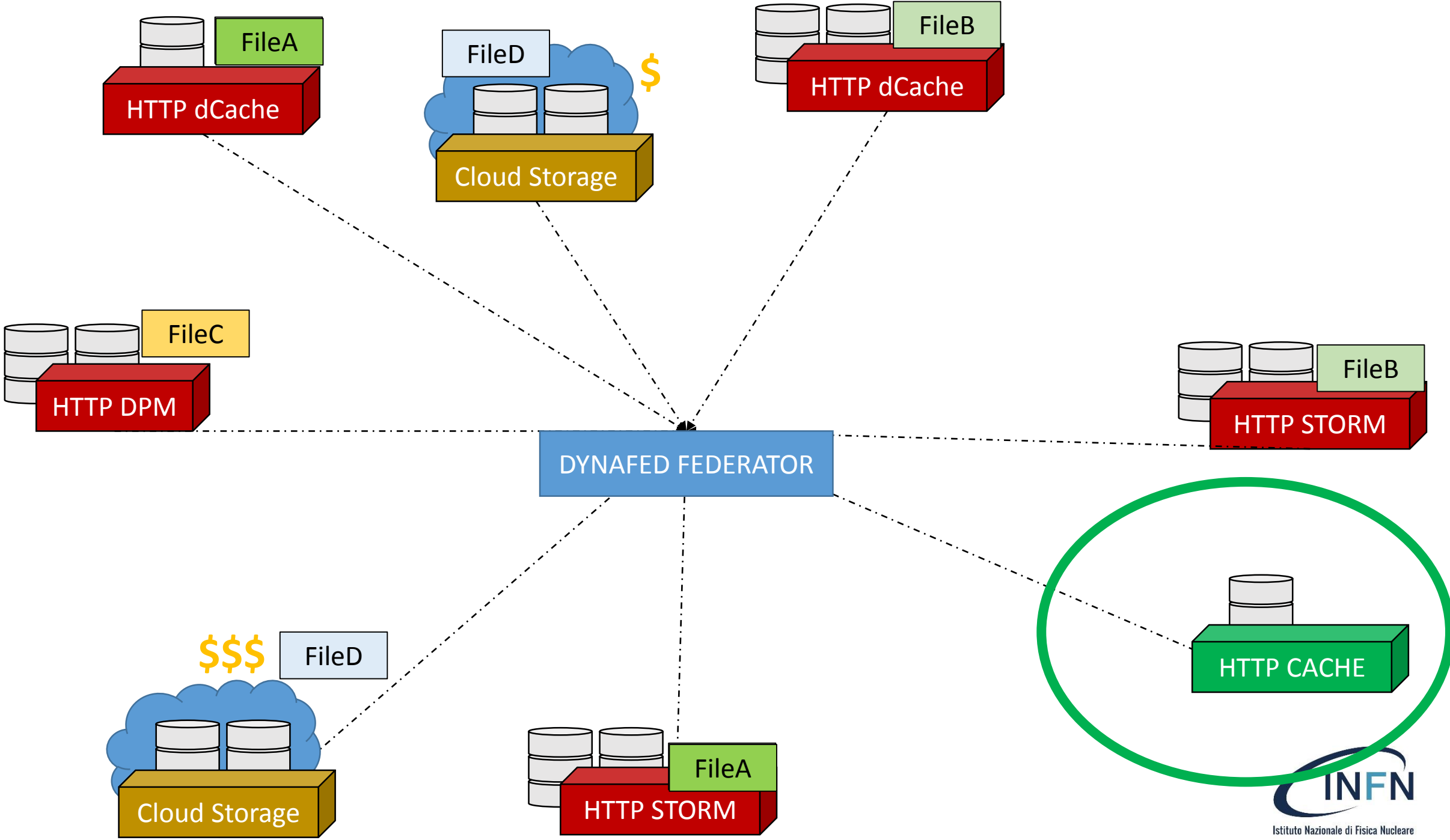
When Dynafed stats a file, it receive always a positive answer from the Volatile Pool.

So that the metalink representing a file in Dynafed, will included always at least two links: the real URL, and the corresponding virtual copy in the cache (even if the latter does not exist yet)

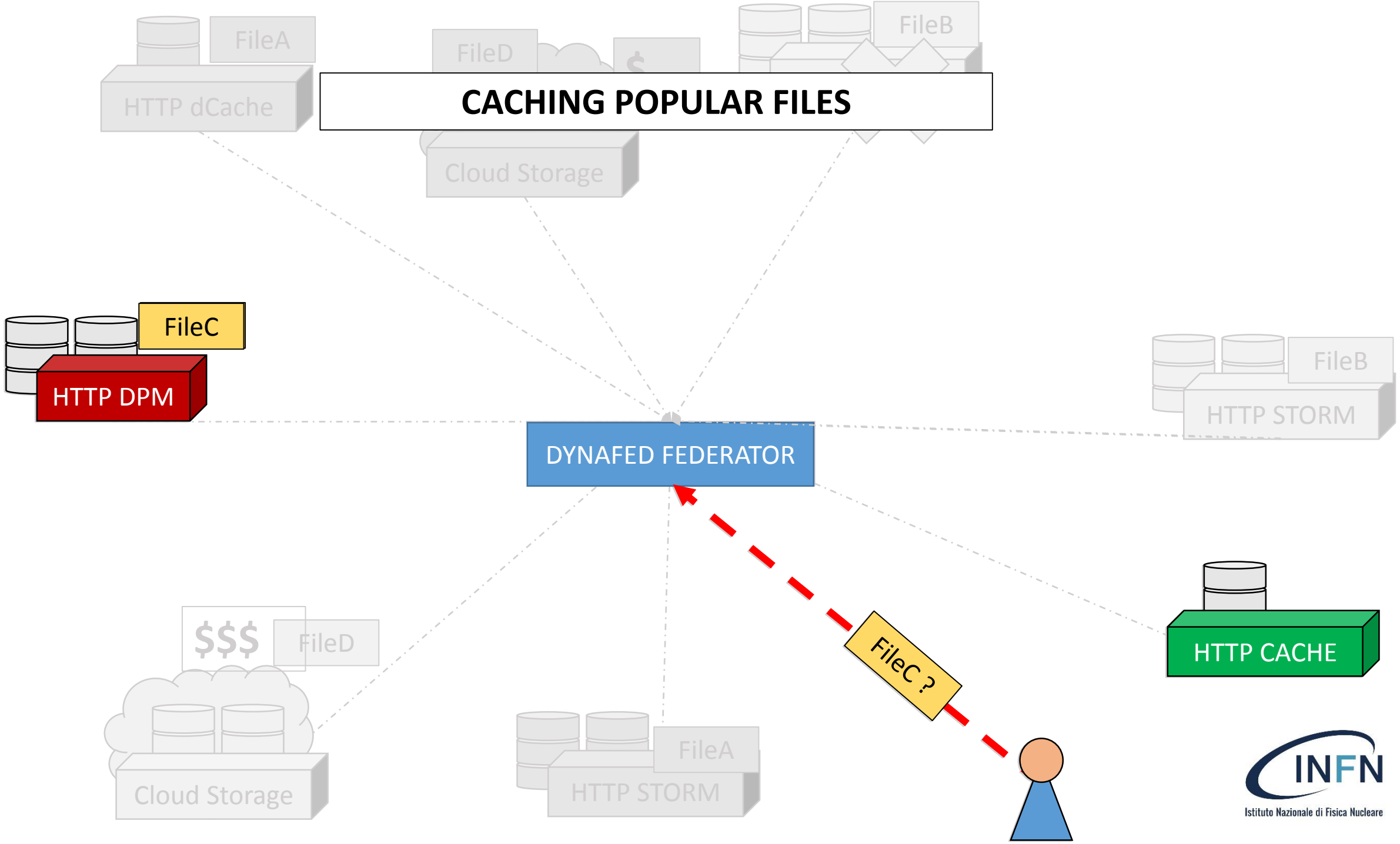
Moreover thanks to the GeoPlugin, Dynafed prioritize the cache copy if the Volatile Pool is local to the Client or close to it.

This combination allow to create a cache system

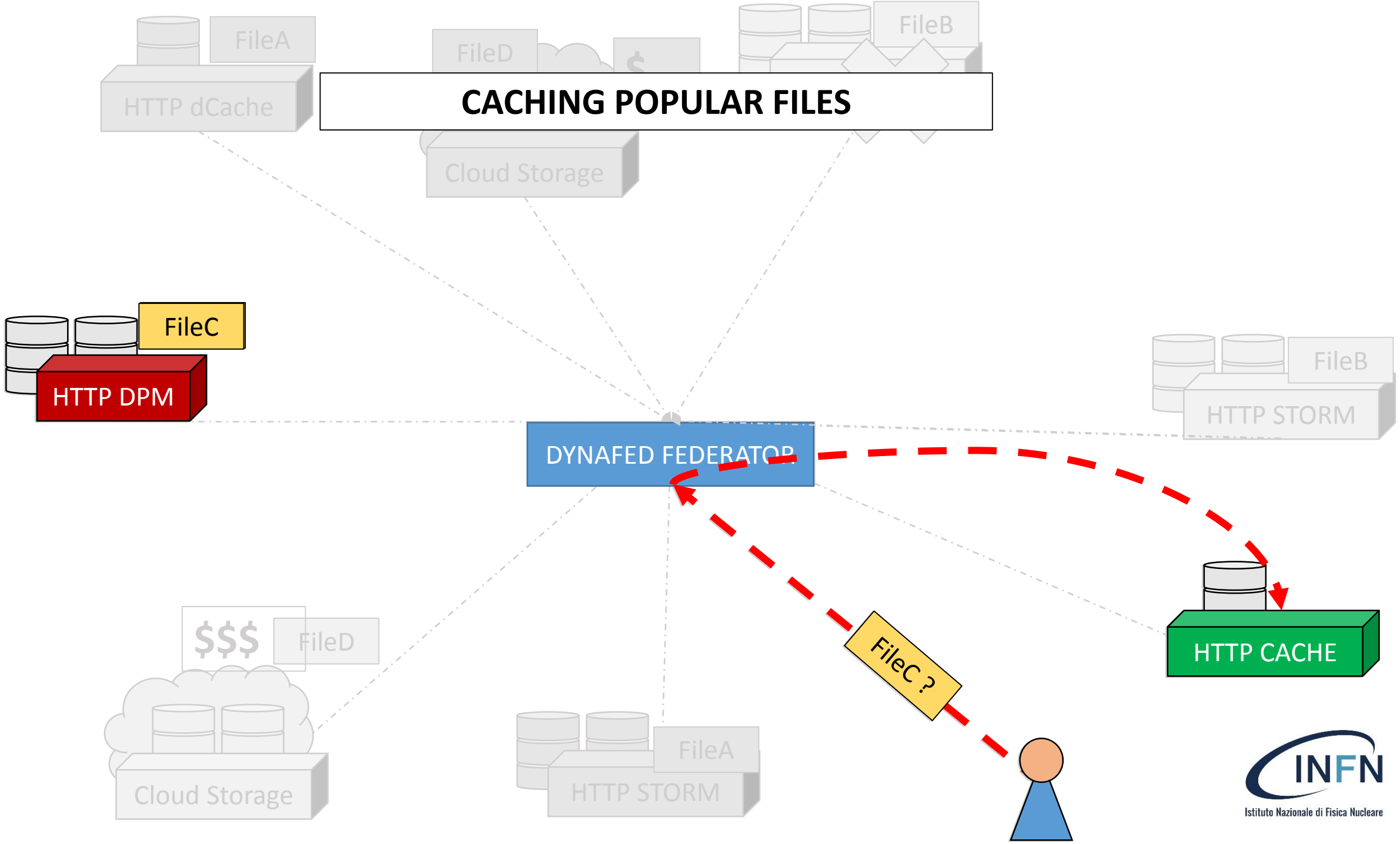
INFN
Istituto Nazionale di Fisica Nucleare


FileA
HTTP dCache
FileD
$
Cloud Storage
FileB
HTTP dCache
FileC
HTTP DPM
FileB
HTTP STORM
DYNAFED FEDERATOR
HTTP CACHE
$$$
FileD
Cloud Storage
FileA
HTTP STORM
INFN
Istituto Nazionale di Fisica Nucleare

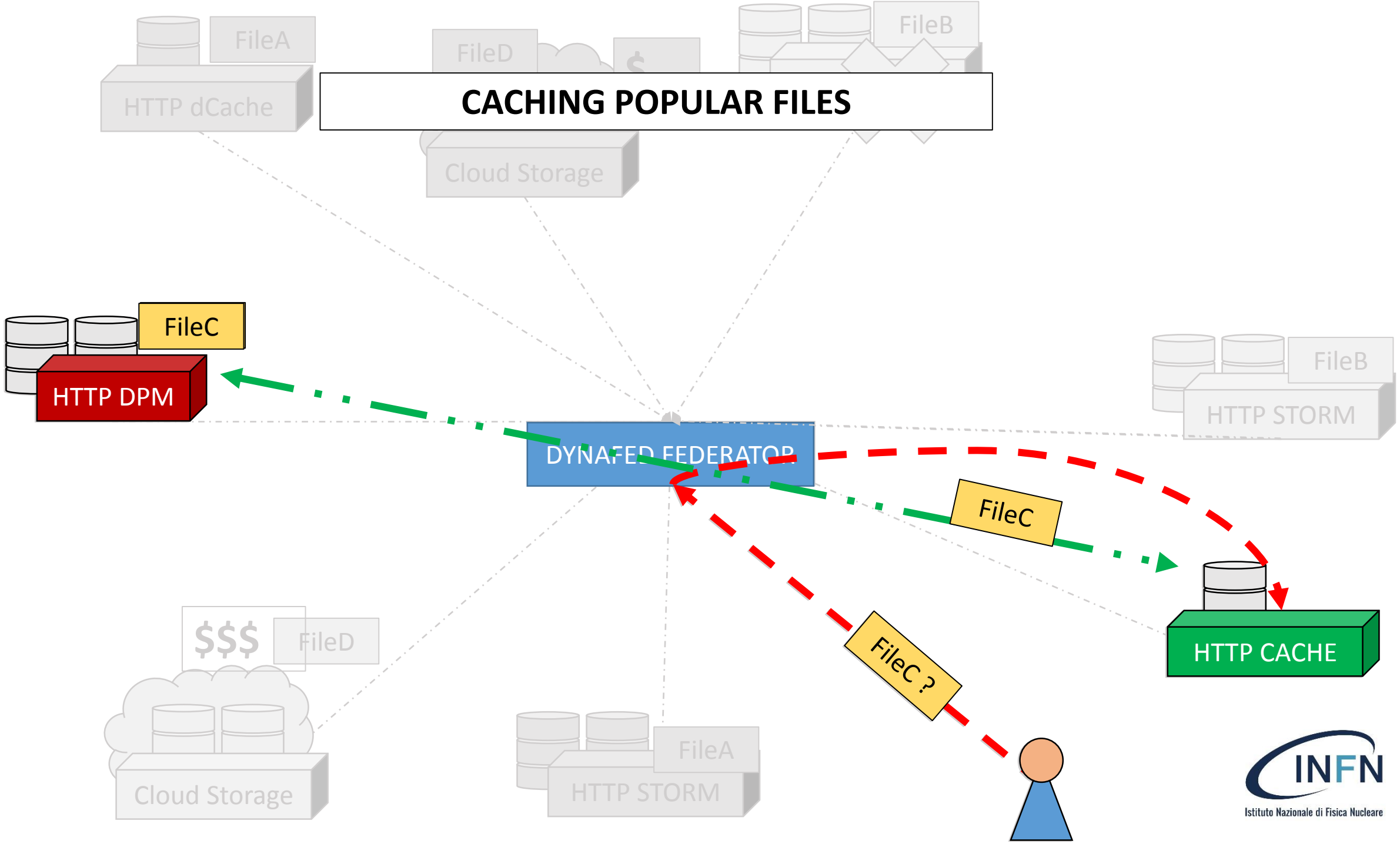
CACHING POPULAR FILES



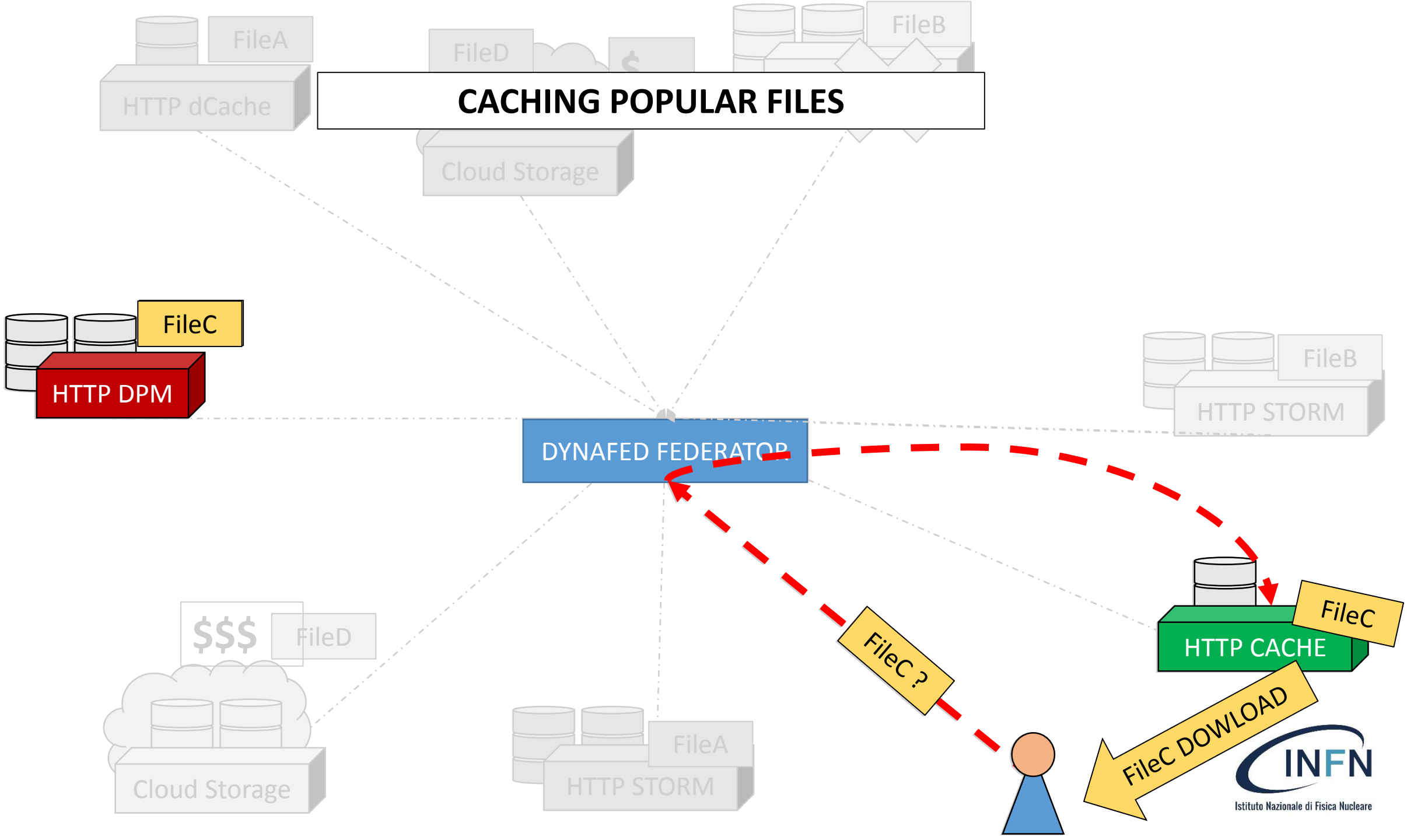
CACHING POPULAR FILES



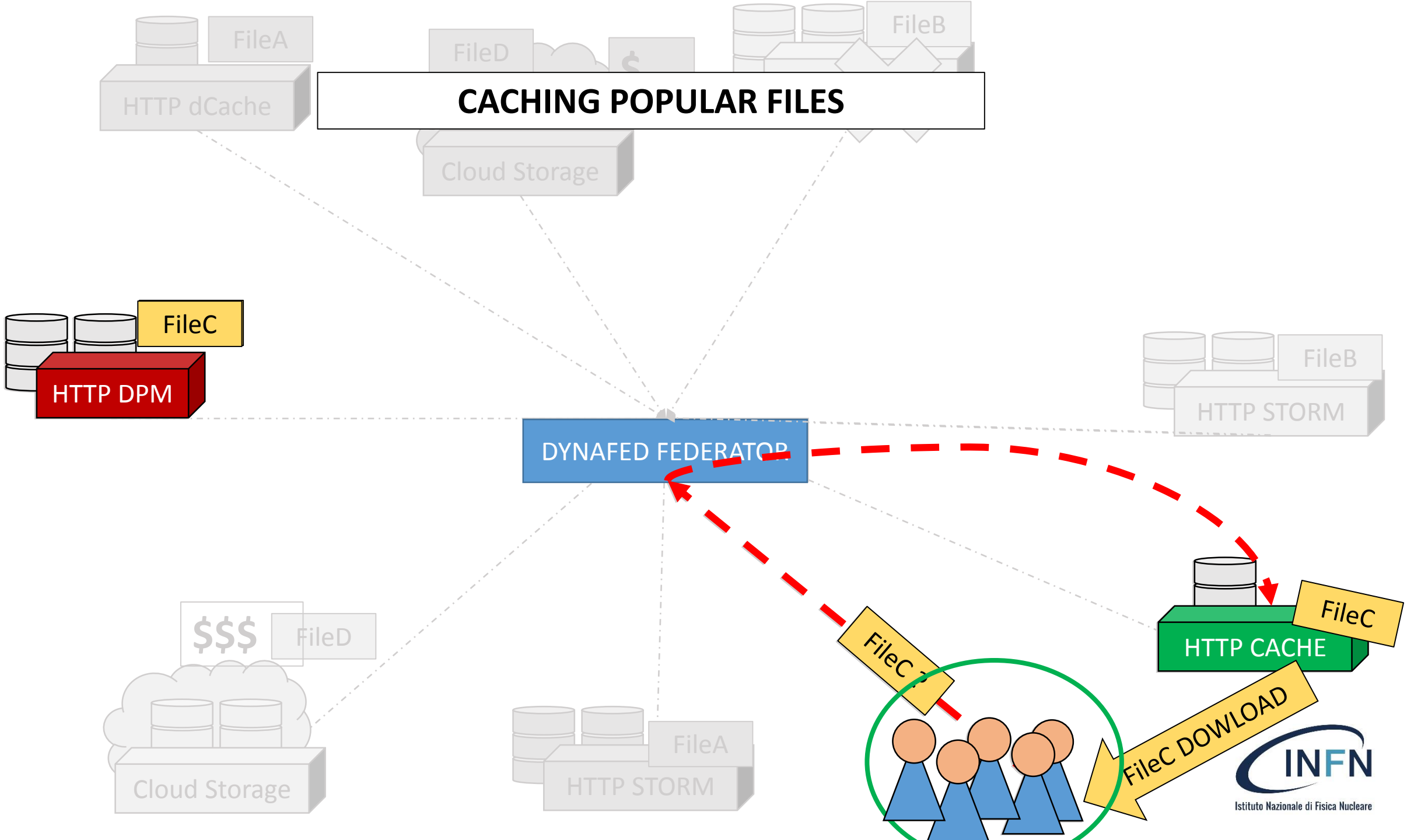
CACHING POPULAR FILES



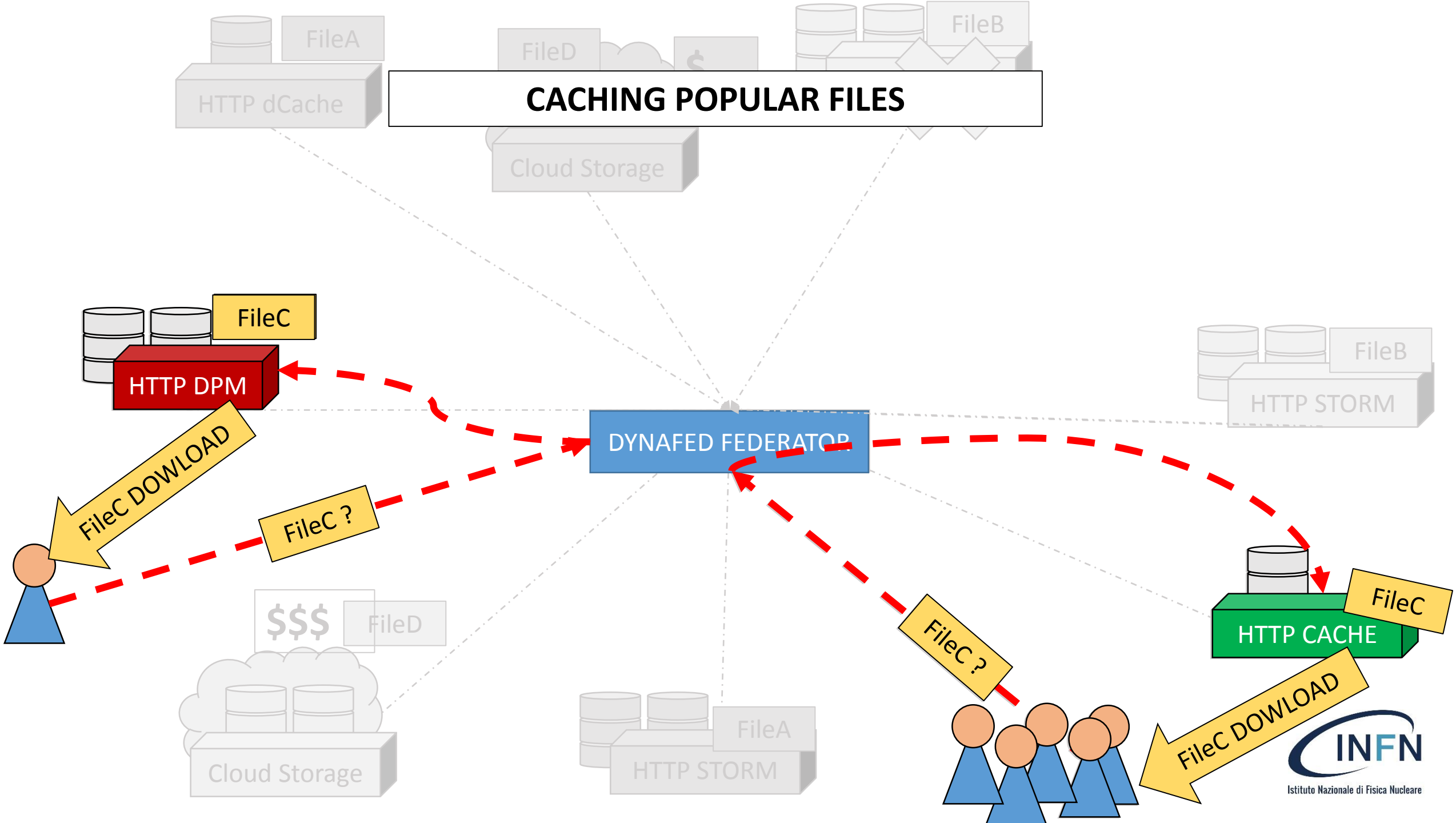
CACHING POPULAR FILES



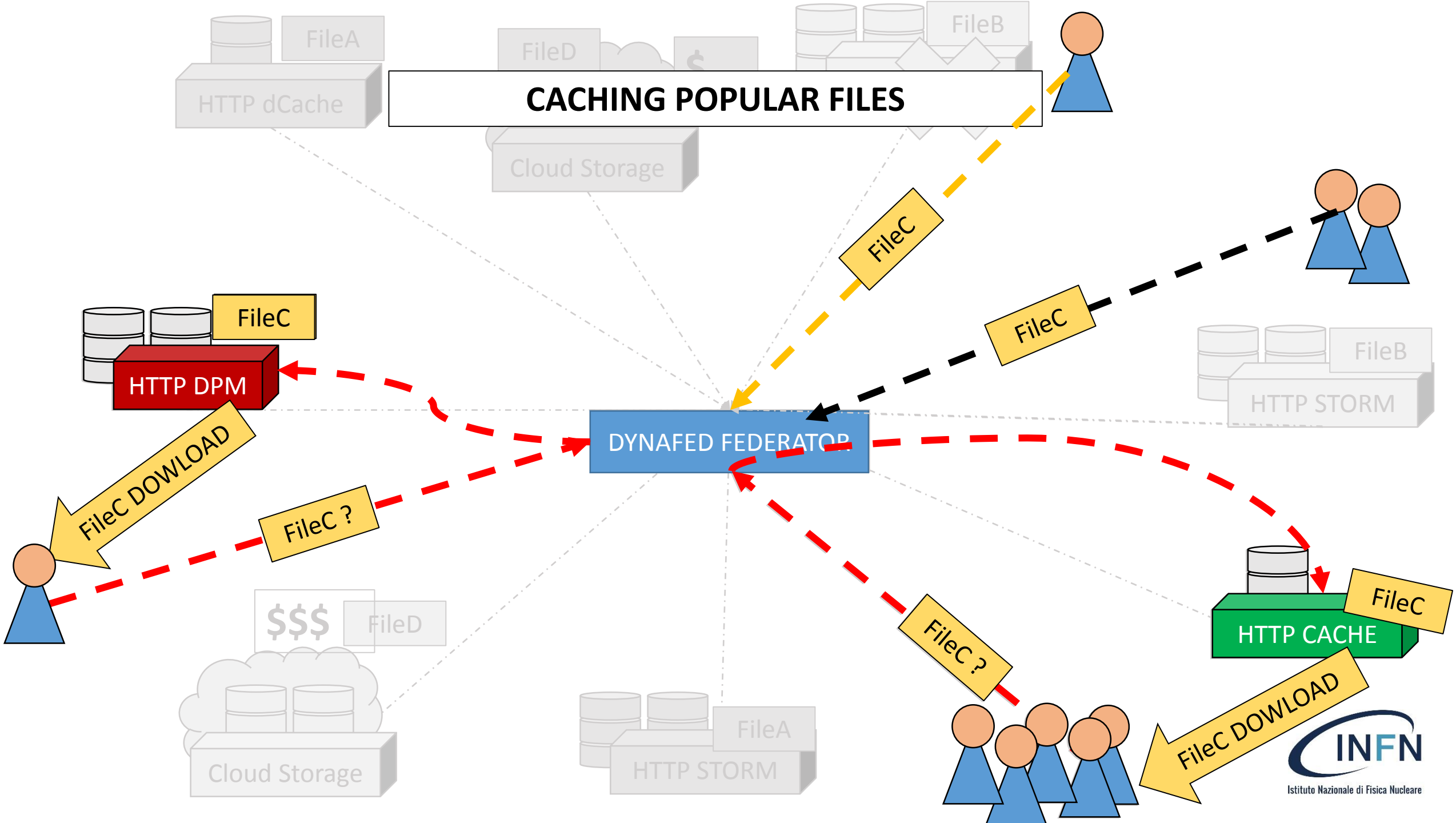
CACHING POPULAR FILES



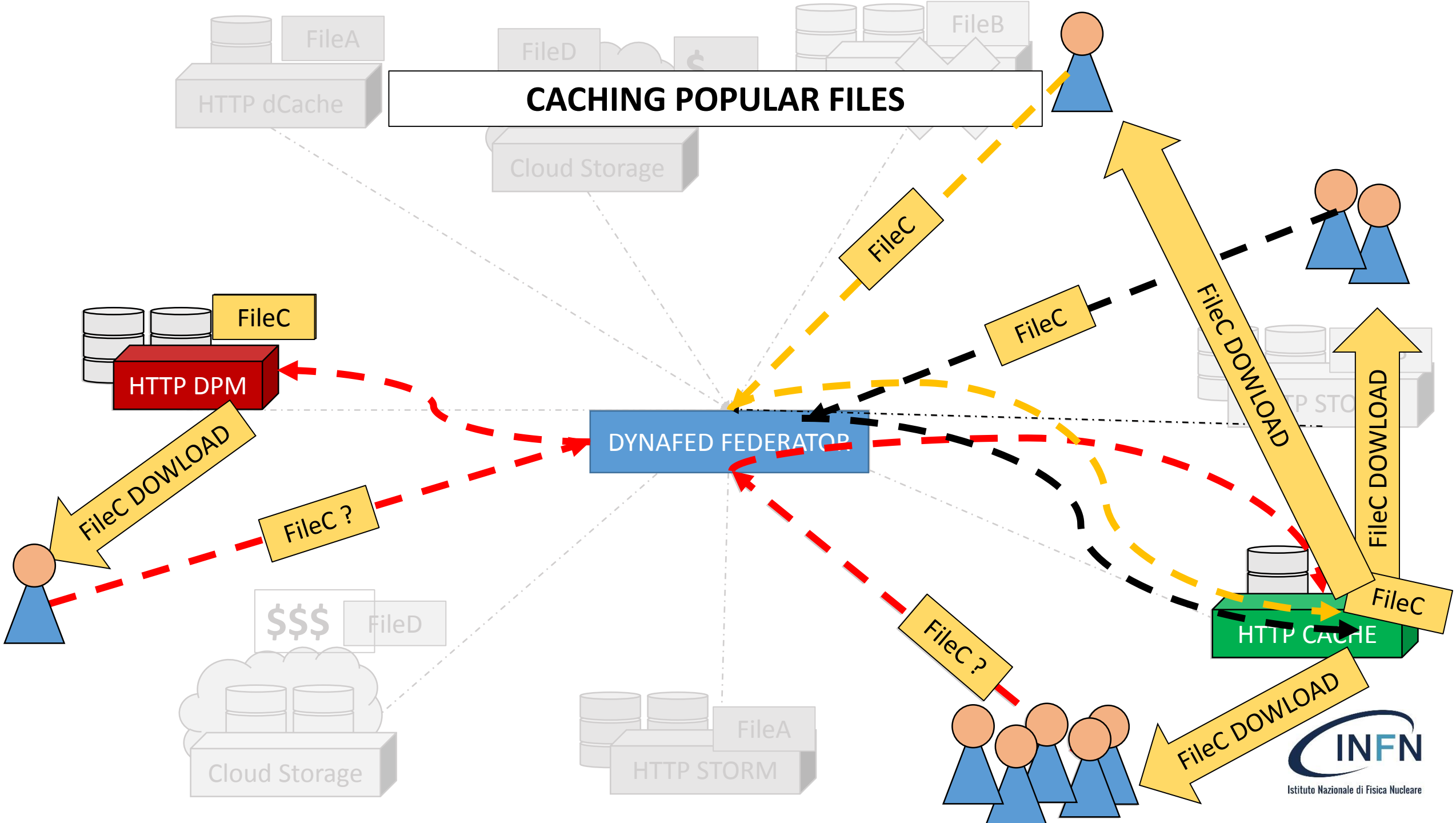
CACHING POPULAR FILES



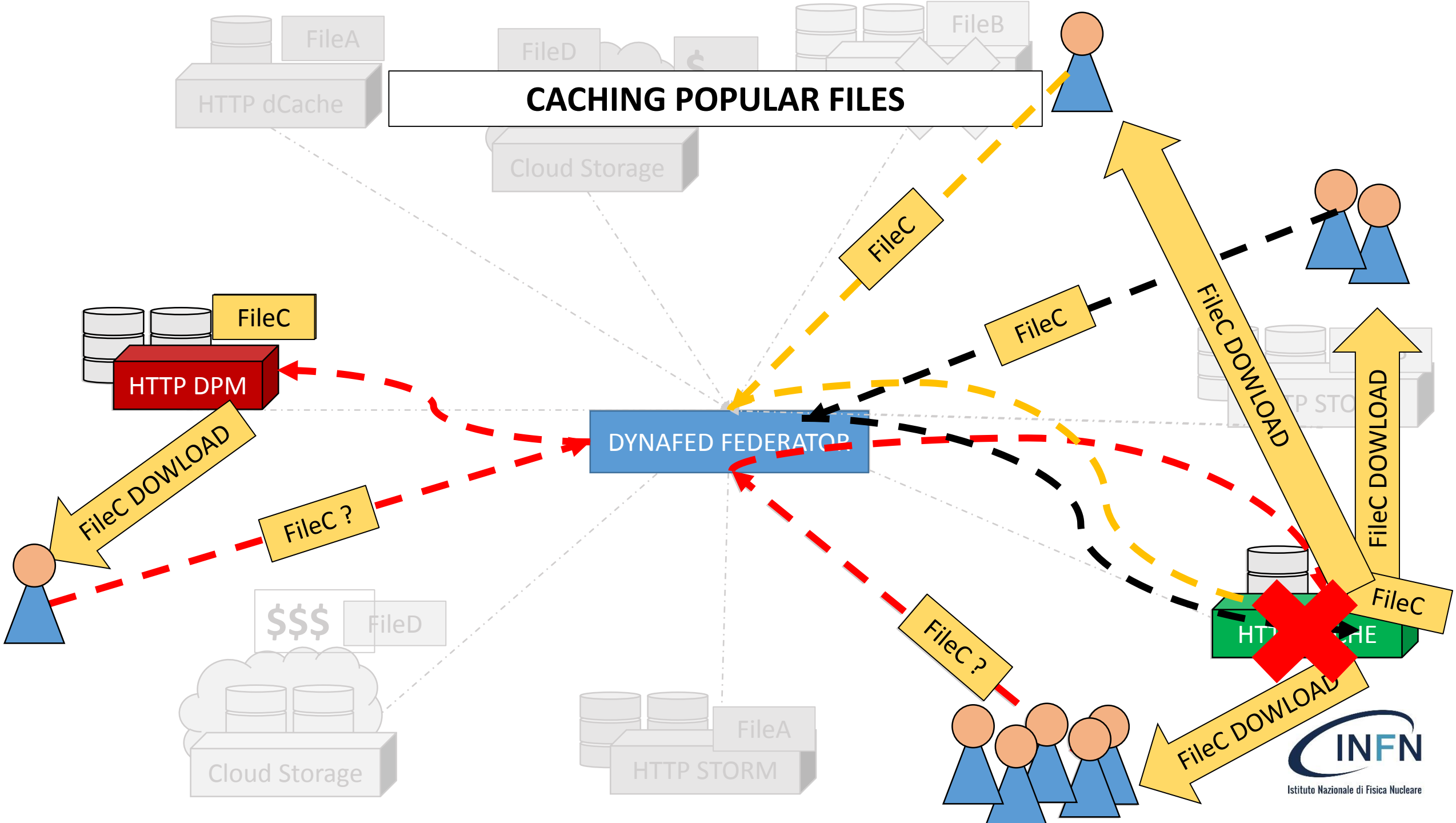
CACHING POPULAR FILES



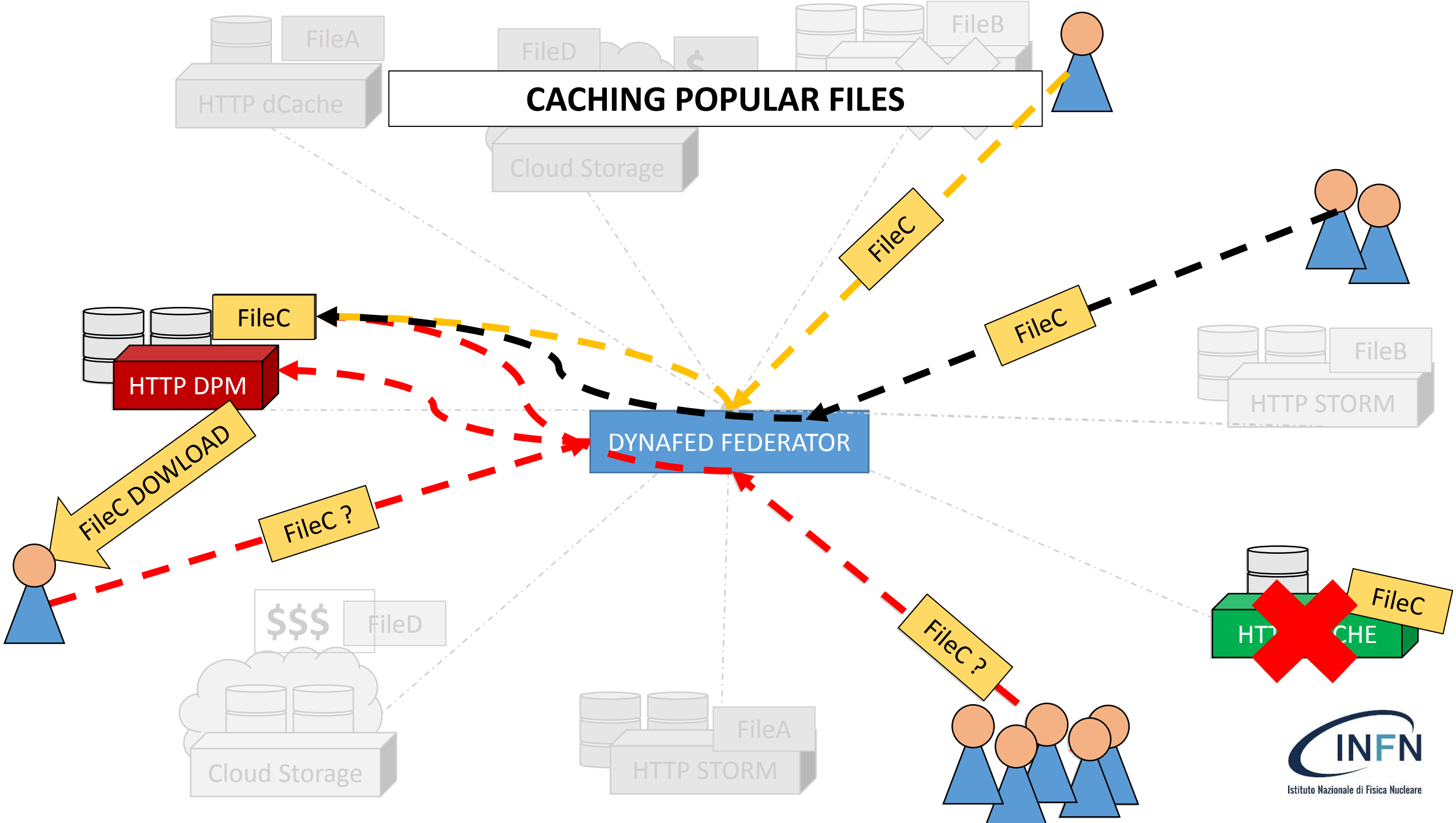
CACHING POPULAR FILES

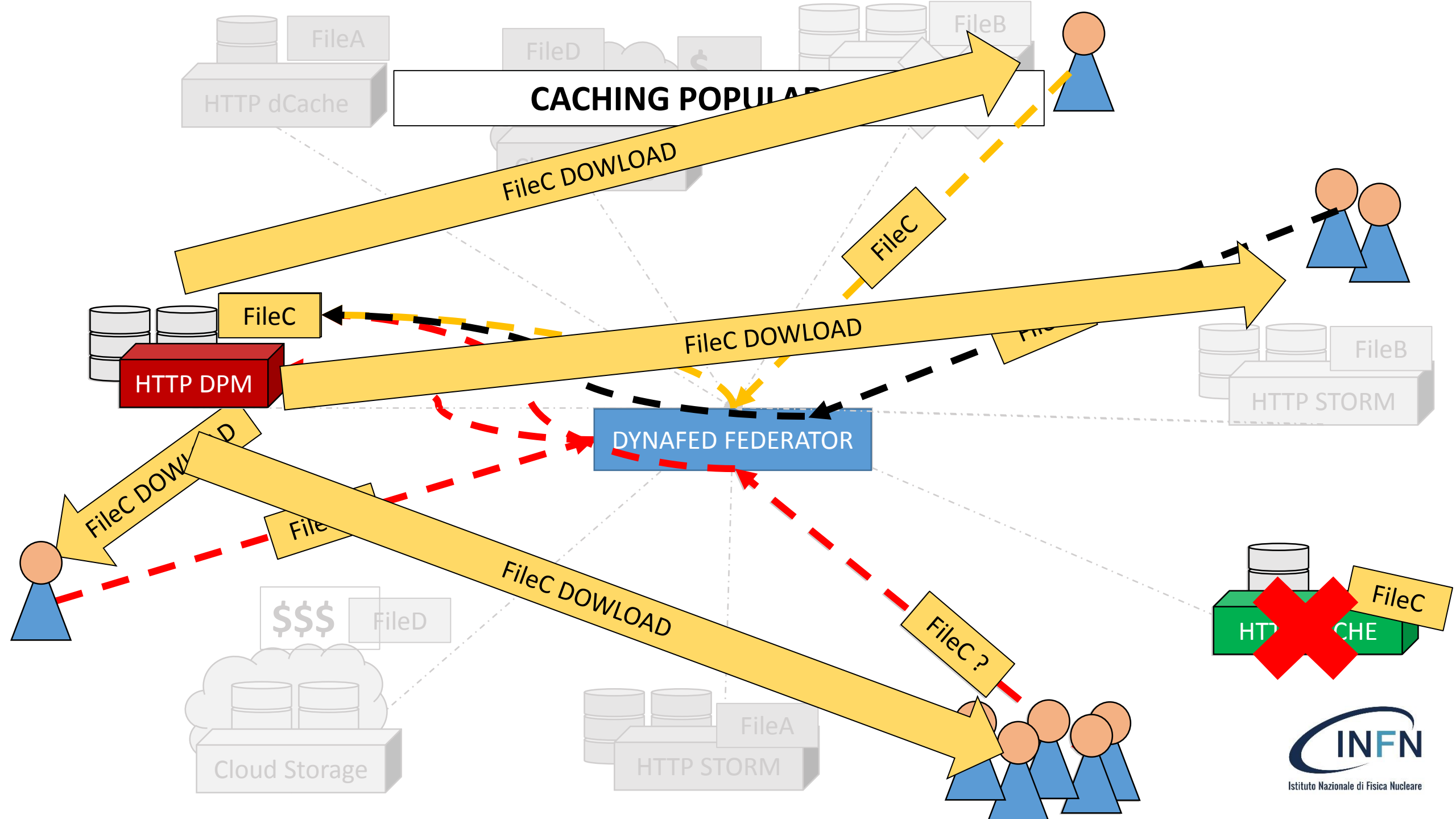


CACHING POPULAR FILES

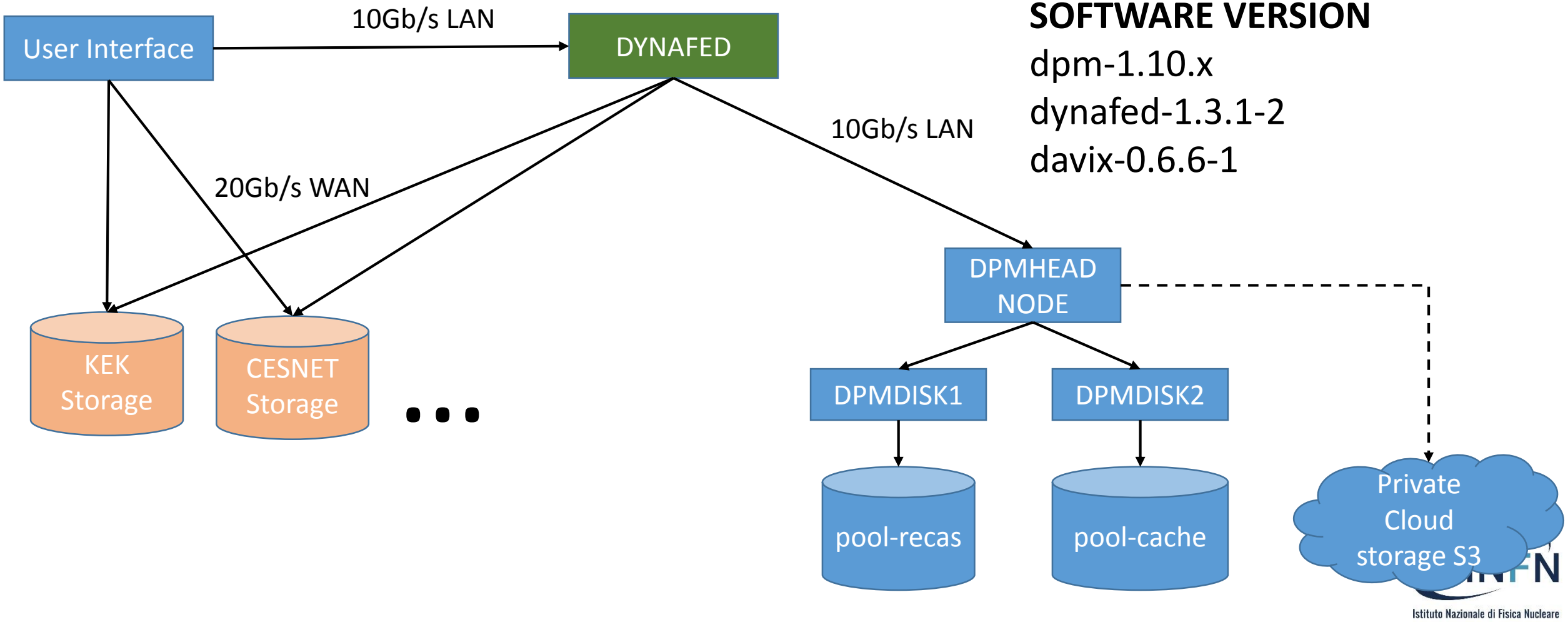


CACHING POPULAR FILES




CACHING POPULAR
FileA
HTTP dCache
FileD
FileB
FileC DOWLOAD
FileC
FileC
HTTP DPM
FileC DOWLOAD
FileB
HTTP STORM
DYNAFED FEDERATOR
FileC DOWLOAD
FileC DOWLOAD
FileC ?
$$$
FileD
Cloud Storage
FileA
HTTP STORM
FileC
INFN
Istituto Nazionale di Fisica Nucleare

The testbed



SOFTWARE VERSION

dpm-1.10.x

dynafed-1.3.1-2

davix-0.6.6-1

Dynafed Server for Belle II

#	STORGE NAME	HOSTNAME	TYPE
1	DESY-DE	dcache-belle-webdav.desy.de	DCACHE
2	GRIDKA-SE	f01-075-140-e.gridka.de	DCACHE
3	NTU-SE	bgrid3.phys.ntu.edu.tw	DCACHE
4	SIGNET-SE	dcache.ijs.si	DCACHE
5	UVic-SE	charon01.westgrid.ca	DCACHE
6	BNL-SE	dcbldoor01.sdcc.bnl.gov	DCACHE
7	Adelaide-SE	coepp-dpm-01.ersa.edu.au	DPM
8	CESNET-SE	dpm1.egee.cesnet.cz	DPM
9	CYFRONNET-SE	dpm.cyf-kr.edu.pl	DPM
10	Frascati-SE	atlasse.lnf.infn.it	DPM
11	HEPHY-SE	hephyse.oeaw.ac.at	DPM
12	Melbourne-SE	b2se.mel.coepp.org.au	DPM
13	Napoli-SE	belle-dpm-01.na.infn.it	DPM
14	ULAKBIM-SE	torik1.ulakbim.gov.tr	DPM
15	IPHC-SE	sbgse1.in2p3.fr	DPM
16	CNAF-SE	ds-202-11-01.cr.cnaf.infn.it	STORM
17	ROMA3-SE	storm-01.roma3.infn.it	STORM
18	KEK-SE	Kek-se03.cc.kek.jp	STORM
19	McGill-SE	gridftp02.clumeq.mcgill.ca	STORM

Testing Dynafed server in Napoli since Feb 2016

In January 2018 we installed the new new version of Dynafed on CENTOS-7

https://dynafed-belle.na.infn.it/myfed

19 Storages (about 75%)

Proxy generated by a robot certificate

Version on SL6 Still available
https://dynafed01.na.infn.it/myfed/

INFN
Istituto Nazionale di Fisica Nucleare

Cache Implementation via DOME

Script on the Head Node:

The implemented script recognizes if the requested path is a file or a directory then reply to the client consequently. The plugin retrieve as well the size of the real copy of the file.

Script on the Disk Node:

When a file is not in the cache, the disk node download the requested file from the datagrid by resolving the location via Dynafed. (Using Robot Certificate registerd in the VO)

INFN
Istituto Nazionale di Fisica Nucleare

Client Behaviour

- If the file is not in cache or not ready yet, the client receives a 202 Message that ask for waiting.
- Davix or gfal clients will retry after a n-seconds (retry_delay) up to max_retry.
- Then the file will be downloaded from the volatile pool

INFN
Istituto Nazionale di Fisica Nucleare

Federation Views

With Dynafed is possible to create multiple views by aggregating storage paths in different manner. Two new views as been added

- **myfed/PerSite/** Shows the file systems of each storage separately (without aggregation)
- **myfed/belle/** Aggregation of all the directory /DATA/belle and /TMP/belle/ + VOLATILE POOL
- **myfed/nocache/** Aggregation of all the directory /DATA/belle and /TMP/belle/ + WITHOUT VOLATILE POOL

/myfed/

Mode	Links	UID	GID	Size	Modified	Name
drwxrwxrwx	0	0	0	0	Thu, 01 Jan 1970 00:00:00 GMT	PerSite
drwxrwxrwx	0	0	0	0	Thu, 01 Jan 1970 00:00:00 GMT	belle
drwxrwxrwx	0	0	0	0	Thu, 01 Jan 1970 00:00:00 GMT	nocache

INFN
Istituto Nazionale di Fisica Nucleare

Webmail | DOMA / ACCESS Meeting (27 No | WLCG DOMA ACCESS Links - Go | https://dynafed-belle.na.infn.it/m

https://dynafed-belle.na.infn.it/myfed/belle/Raw/e0002/cosmic/r00013/sub00/cosmic.0002.00013.HLT3.f00000.root?metalink

App | Bookmarks | Data production shift | Computing ShiftMan | Computing DCOpera | Computing Operatio | Presentazione preven | Nuova cartella | Nuova c

This XML file does not appear to have any style information associated with it. The document tree is shown below.


```
▼<metalink xmlns="http://www.metalinker.org/" xmlns:lcgdm="LCGDM:" version="3.0" generator="lcgdm-dav" pubdate="Thu, 07 Jun 2018 10:30:36 GMT">
  ▼<files>
    ▼<file name="/belle/">
      <size>711396759</size>
     ▼<resources>
      ▼<url type="https">
         https://recas-dpm-01.na.infn.it/dpm/na.infn.it/home/belle/cache1/Raw/e0002/cosmic/r00013/sub00/cosmic.0002.00013.HLT3.f00000.root
       </url>
      ▼<url type="https">
         https://dcbldoor01.sdcc.bnl.gov:443/pnfs/sdcc.bnl.gov/data/bellediskdata/DATA/belle/Raw/e0002/cosmic/r00013/sub00/cosmic.0002.00013.HLT3.f00000.root
       </url>
      ▼<url type="https">
         https://kek2-se03.cc.kek.jp:8443/belle/DATA/belle/Raw/e0002/cosmic/r00013/sub00/cosmic.0002.00013.HLT3.f00000.root
       </url>
```


RAW DATA FILE - METALINK IN THE FULL VIEW

VOLATILE POLL FIRST IN THE LIST

Webmail | DOMA / ACCESS Meeting (27 No | WLCG DOMA ACCESS Links - Go | https://dynafed-belle.na.infn.it/m

https://dynafed-belle.na.infn.it/myfed/nocache/Raw/e0002/cosmic/r00013/sub00/cosmic.0002.00013.HLT3.f00000.root?metalink

App | Bookmarks | Data production shift | Computing ShiftMan | Computing DCOpera | Computing Operatio | Presentazione preven | Nuova cartella | Nuova ca

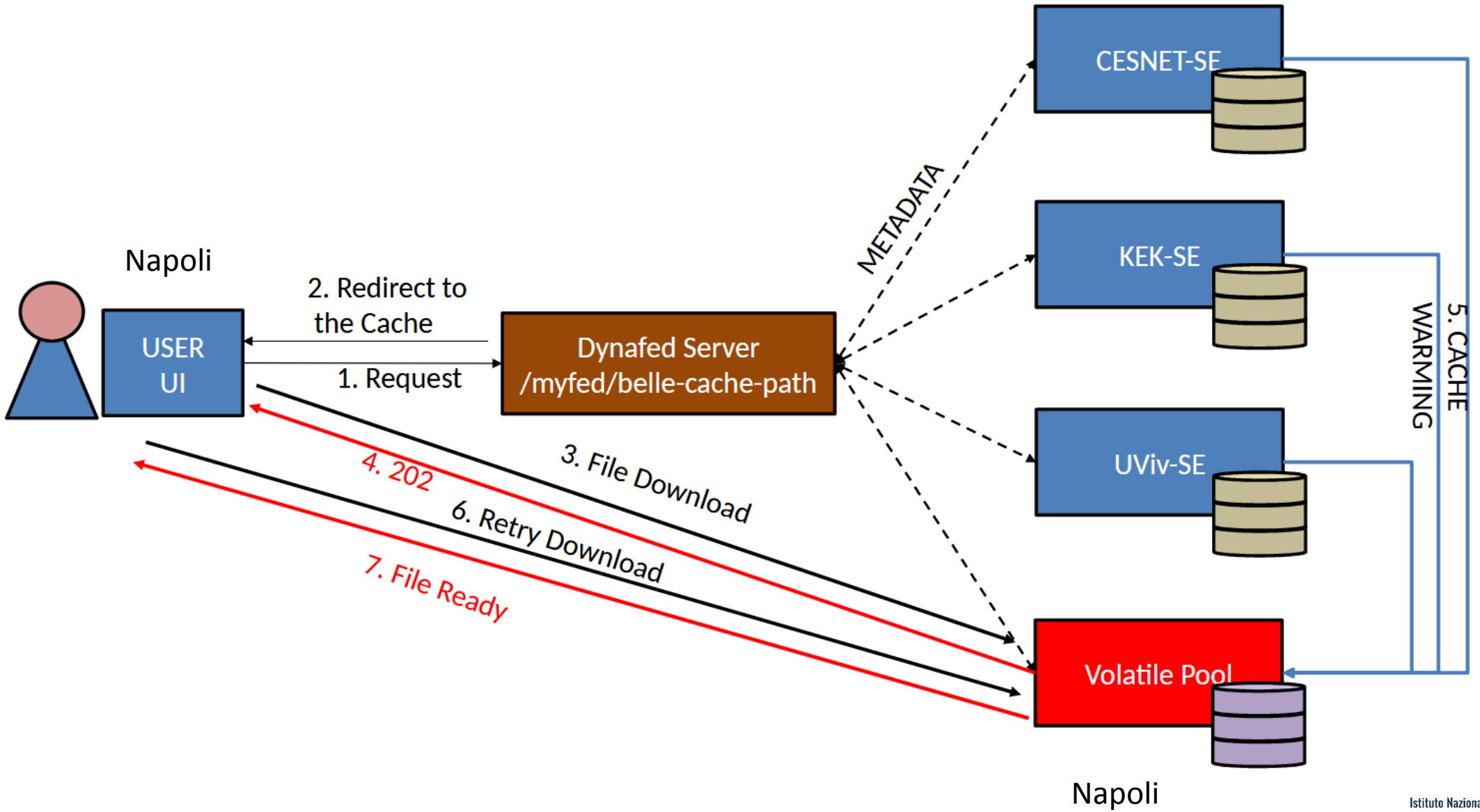
This XML file does not appear to have any style information associated with it. The document tree is shown below.


```
▼<metalink xmlns="http://www.metalinker.org/" xmlns:lcgdm="LCGDM:" version="3.0" generator="lcgdm-dav" pubdate="Thu, 07 Jun 2018 10:30:36 GMT">
  ▼<files>
    ▼<file name="/nocach">
      <size>711396759</size>
     ▼<resources>
      ▼<url type="https">
         https://dcbldoor01.sdcc.bnl.gov:443/pnfs/sdcc.bnl.gov/data/bellediskdata/DATA/belle/Raw/e0002/cosmic/r00013/sub00/cosmic.0002.00013.HLT3.f00000.root
       </url>
      ▼<url type="https">
         https://kek2-se03.cc.kek.jp:8443/belle/DATA/belle/Raw/e0002/cosmic/r00013/sub00/cosmic.0002.00013.HLT3.f00000.root
       </url>
     </resources>
    </file>
  </files>
```


RAW DATA FILE - METALINK IN NOCACHE VIEW

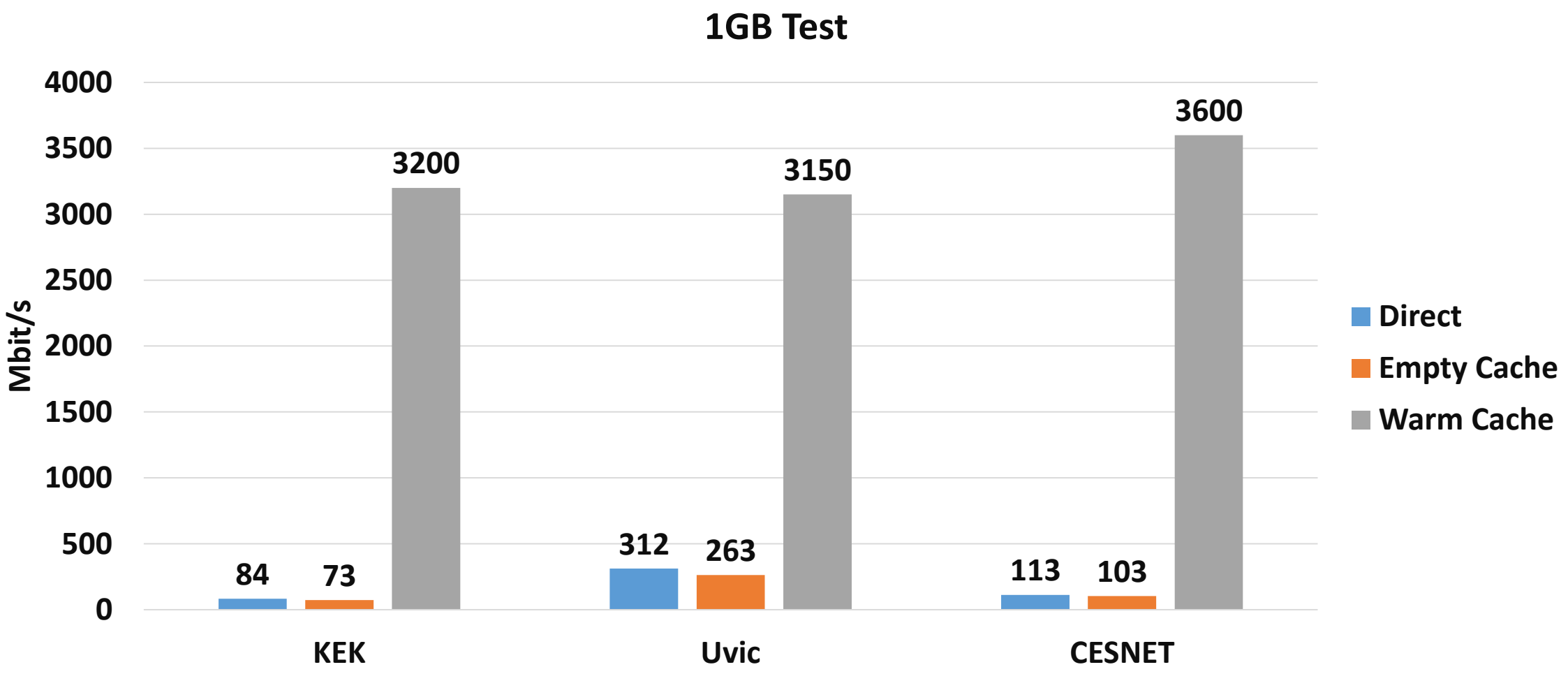
INFN Istituto Nazionale di Fisica Nucleare

Implementation Detail



File Download Test 1GB from a UI in Napoli

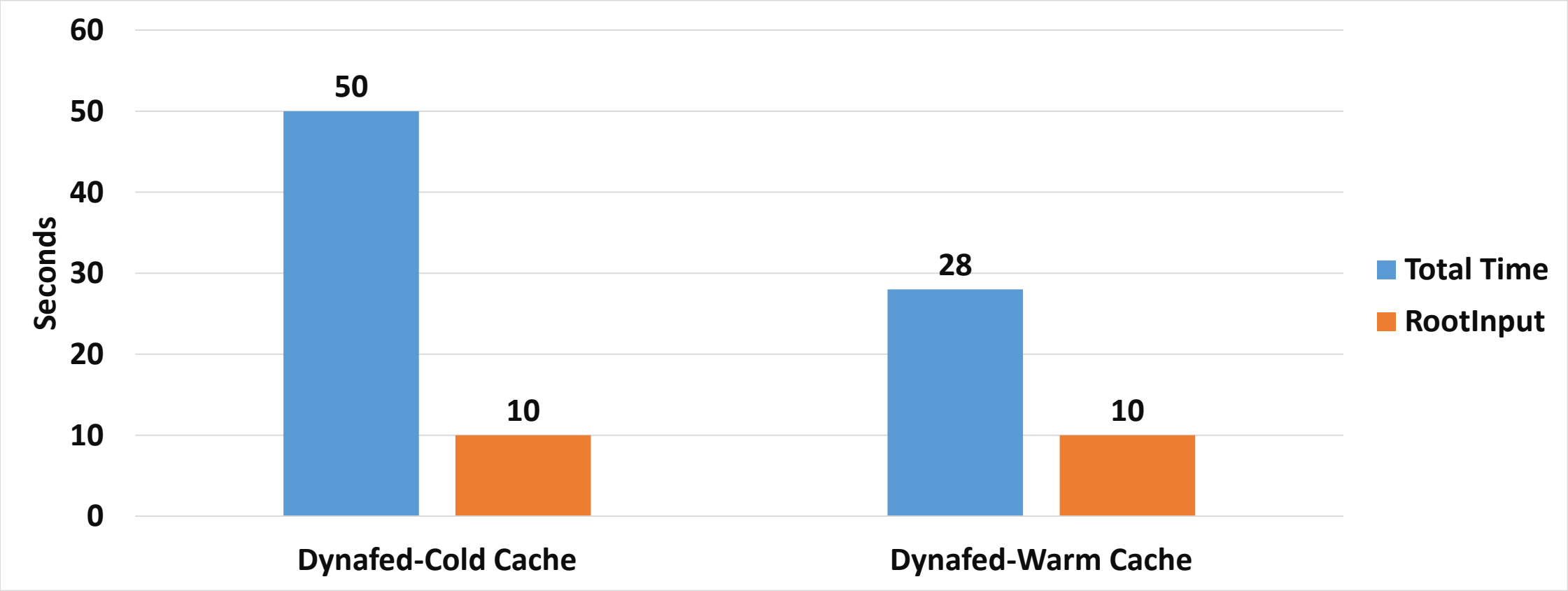
Mbit/s (Higher is better)



Local job reading file through dynafed

basf2 B2A602-BestCandidateSelection.py -i dav://dynafed-belle.na.infn.it/myfed/belle/MC/mdst_000028_prod00003102_task00000028.root

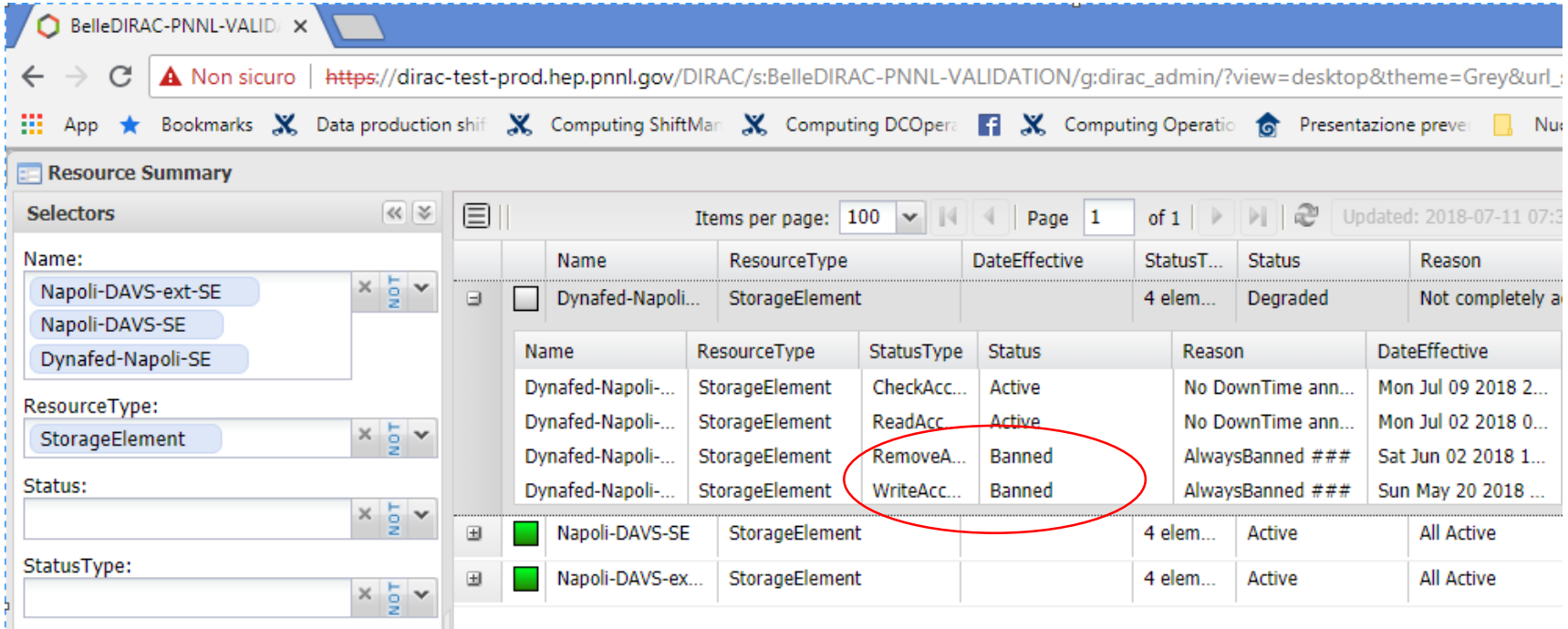
USER INTERFACE IN NAPOLI – PHYSICAL COPY AT KEK



How to use this object?

Using the DIRAC Validation server of Belle II we are investigating different approaches:

- Register the Volatile Pool among SEs (in that case we loss the benefit of dynafed)
- Register dynafed as a Storage (In that case DIRAC loss the control in writing)
- Make a special configuration for the HTTP endpoints registered in DIRAC in order to be used directly in writing and through Dynafed in reading.



INFN
Istituto Nazionale di Fisica Nucleare

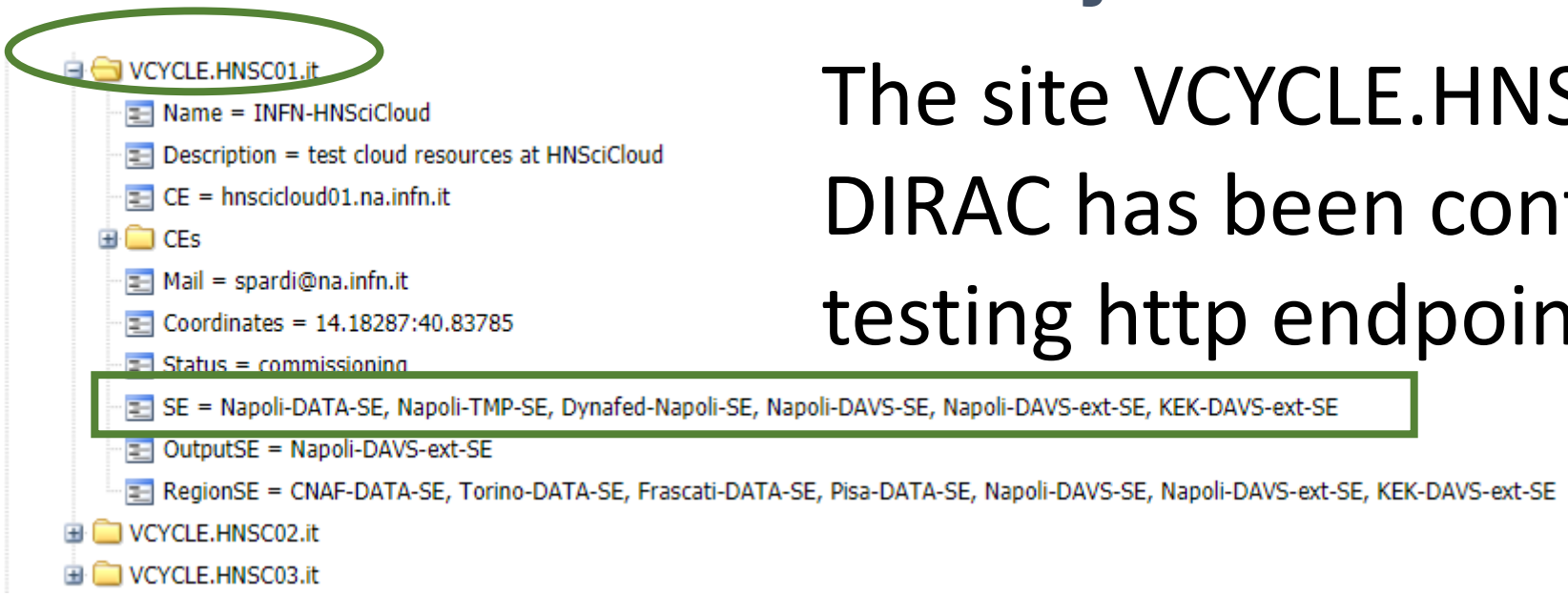
How to use this object?

Ongoing test are focussed on three main use-cases:

- DAVS protocol in DIRAC
- DAVS + Dynafed + DIRAC
- DAVS + Dynafed + DPM Volatile Pool (Cache) + DIRAC

INFN
Istituto Nazionale di Fisica Nucleare

How to use this object?



The site VCYCLE.HNSC01.it in PNNL DIRAC has been configured to use the testing http endpoint, included dynafed

We created a set of datasets locally with basf2 then we copied and registered it on KEK-DAVS-SE storage via **gb2_ds_put** command.

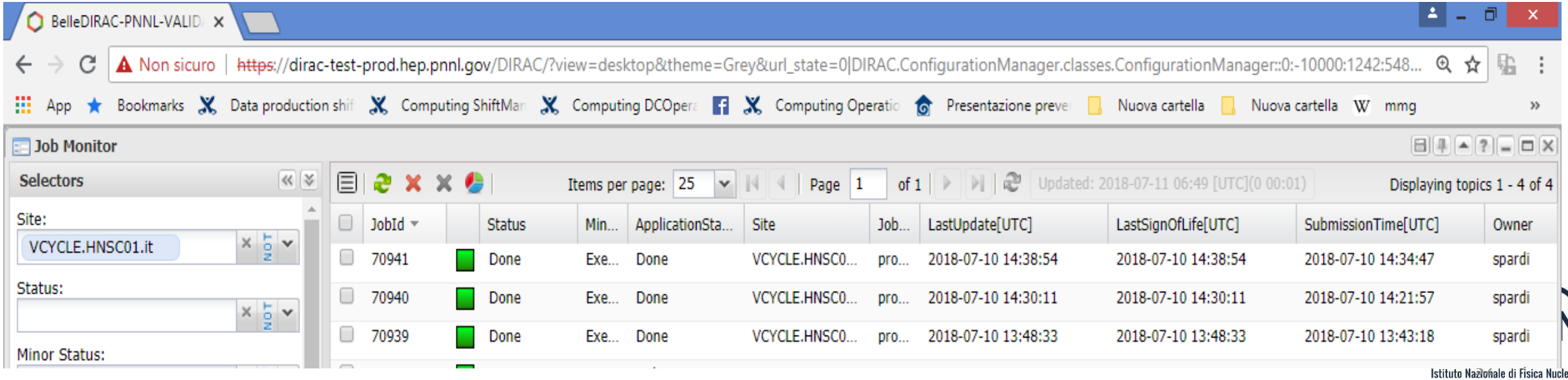
INFN
Istituto Nazionale di Fisica Nucleare

How to use this object?

Submit jobs to DIRAC via gbasf2, taking advantage from the cache.

Early results:

In a protected environment, we replicated datasets to KEK-DAVS-SE and then we ran a set of simple analysis on HNSC resources, reading files from the http storage via Dynafed, using the volatile pool feature as well, experiencing the caching effect.



Current Status and ongoing activities

Up to now we mainly focussed on creating a working testbed, overcoming the issues and investigating how to introduce the cache element in the belle II computing model.

Last part 3 months of the SCoRES project will be dedicated in doing performance and resilience tests that should be ready by the end of February 2018 together with the characterization of the testbed.

Additional Initiatives

The ATLAS Team at INFN-Napoli is working with similar technologies in the context of ATLAS using Volatile Pool in combination with RUCIO. Preliminary results have been presented at CHEP18, more detailed and results will be presented soon.

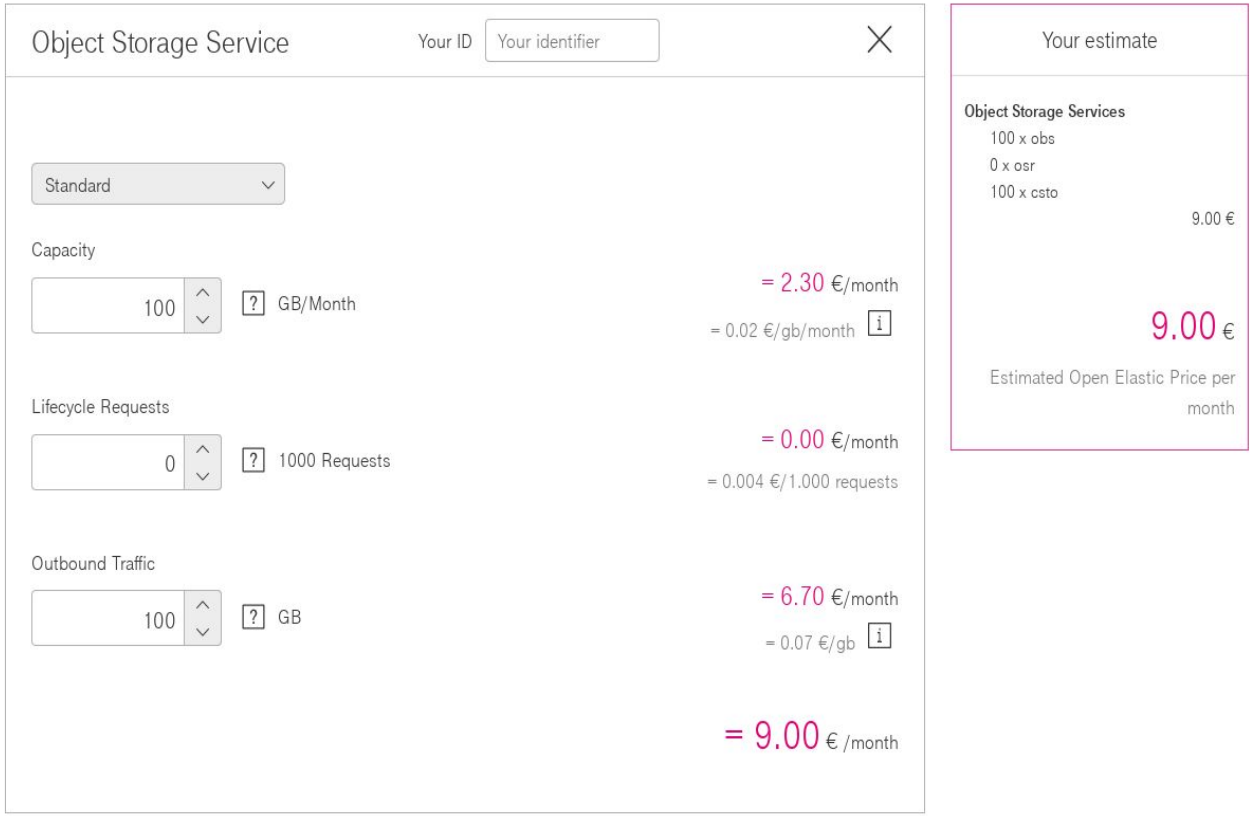
There are currently a set of new initiatives submitted in different context in Italy to support activities related this topic:

Included a research project named “HTTP in Physics (HTTPhy)” submitted within the national call PRIN 2017 (result expected by the end of the year).

I.Bi.S.Co. (Infrastructure for Big Data and Scientific Computing) is a new proposal submitted by several Italian institutions (including INFN and University Federico II) in the contest of the National CALL for datacenter extension.

Thank you

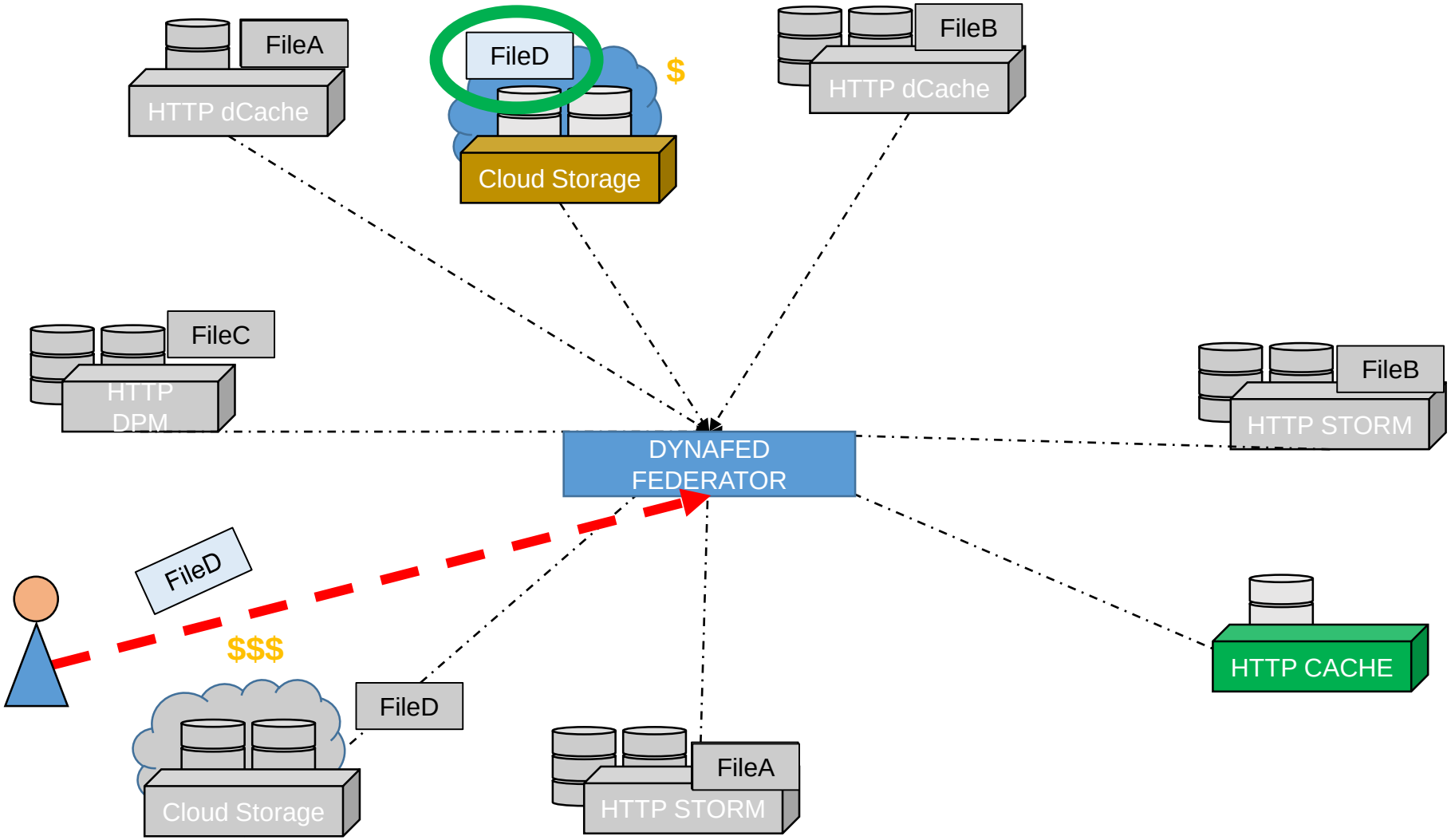
USE CASE Cloud Access



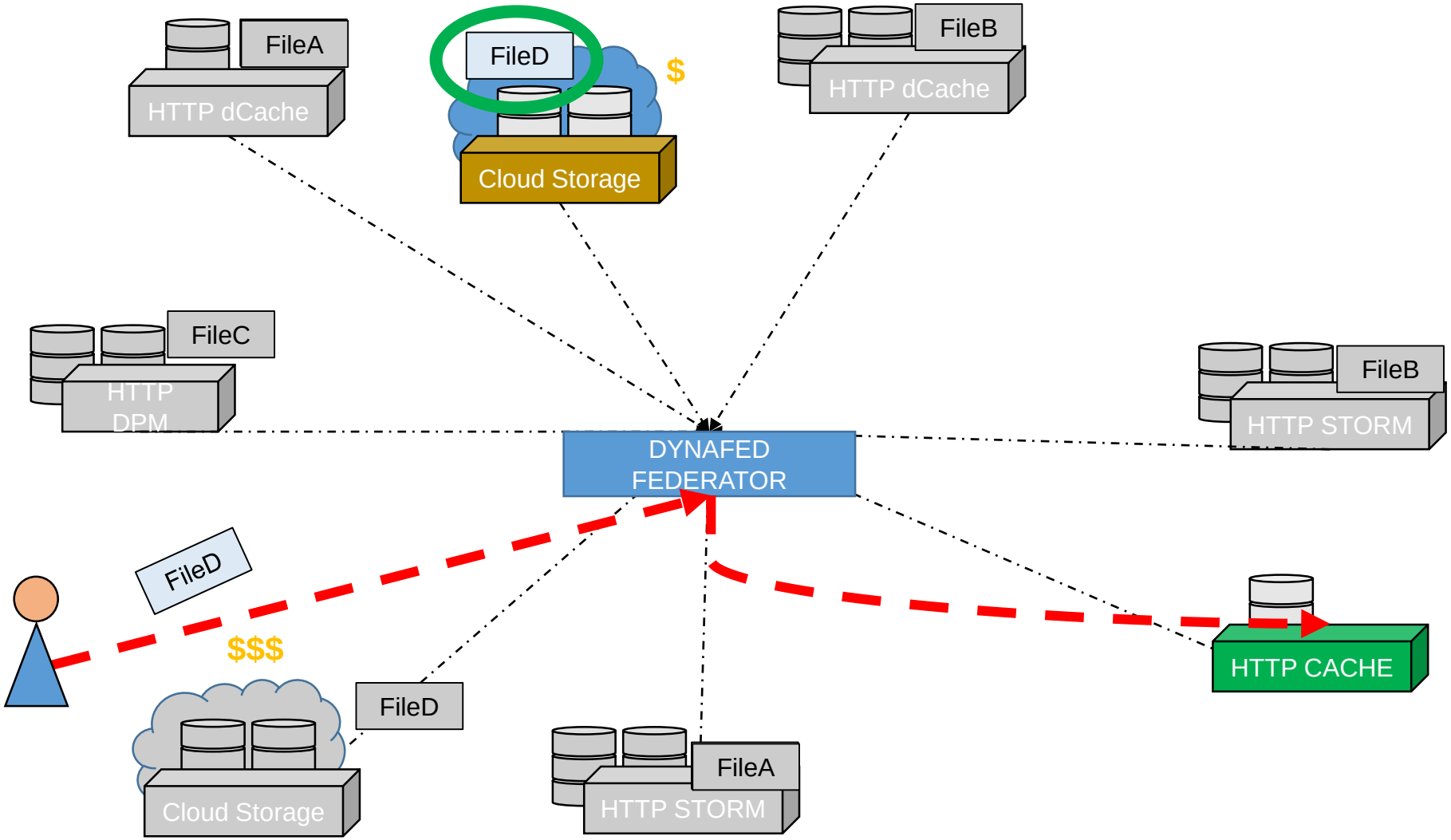
In the context of HNSC project we had the opportunity to simulate the cost of data access with a cloud storage.

Copy 100GB of data from an S3 bucket may cost up to **6.7 Euro**

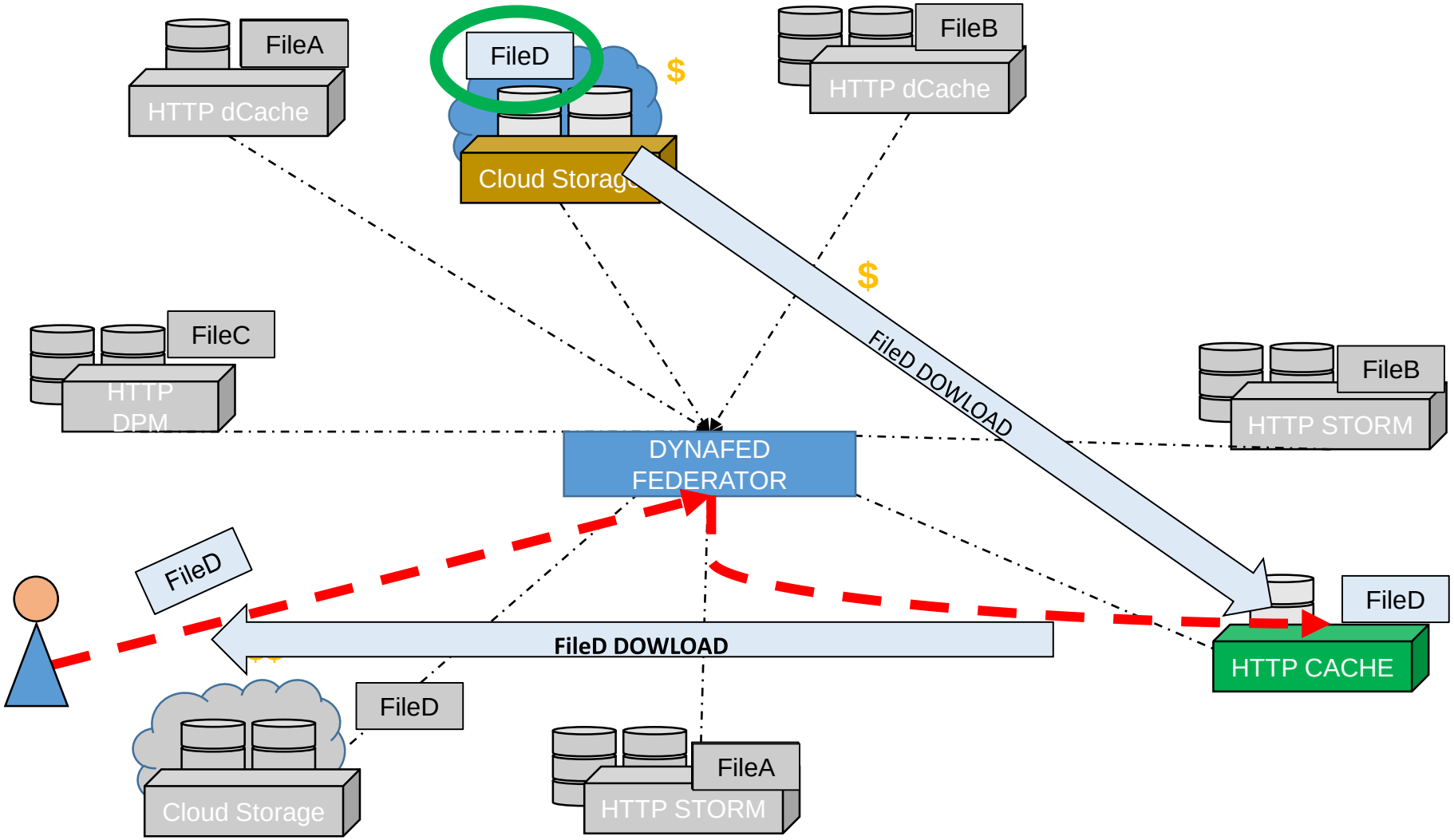
(Cloud T-system in Germania)



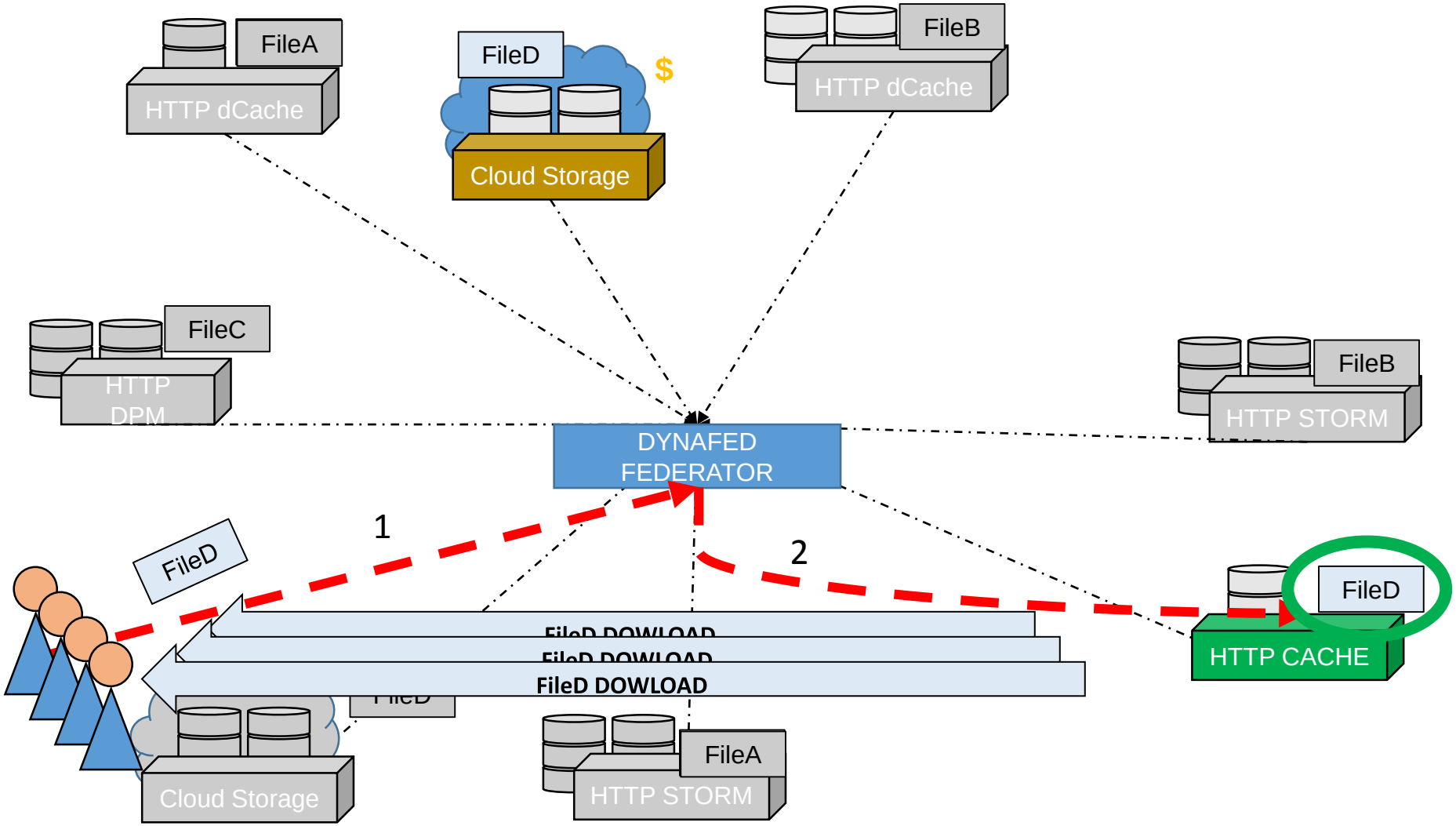
INFN
Istituto Nazionale di Fisica Nucleare



INFN
Istituto Nazionale di Fisica Nucleare



INFN
Istituto Nazionale di Fisica Nucleare



INFN
Istituto Nazionale di Fisica Nucleare

USE CASE Cloud Access

To manage this use-case we developed two filter plugins for Dynafed able to prioritize replicas in a different way rather than the geographical distance between client and storage:

- **Price Plugin**: Which allows to associate an arbitrary weight to storages
- **Default Plugin**: Which allows to set an endpoint as default storage for the host of a network

The combined usage of those two plugins allows to design new scenarios

INFN
Istituto Nazionale di Fisica Nucleare

USE CASE Cloud Access

Configurazione PRICE Plugin

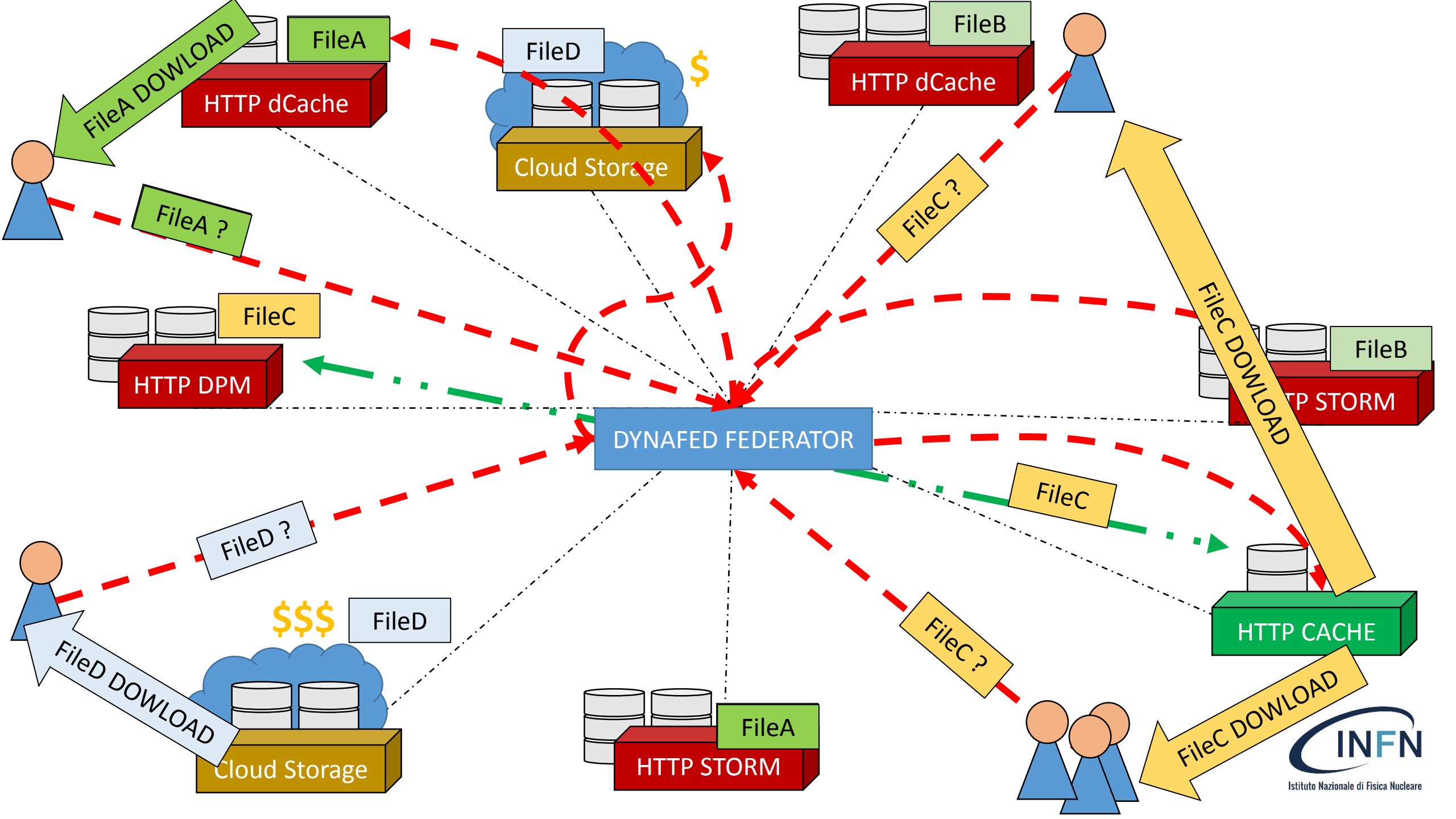
recas-dpm-01.na.infn.it	0.20 (CACHE)
dcache-belle-webdav.desy.de	0.40
kek2-se03.cc.kek.jp	0.50
dcache.ijs.si	0.50
charon01.westgrid.ca	0.50
dpm1.egee.cesnet.cz	0.50
davide.obs.otc.t-systems.com	0.80

Configurazione Default Plugin

131.169.168	recas-dpm-01.na.infn.it (DESY Network)
79.23.	kek2-se03.cc.kek.jp

	Total Size (GB)	Plugin	Costo I accesso	Costo II accesso	Costo III acceso
CLOUD	100	GeoIP	6,7 €	6,7 €	6,7 €
SCORES-CACHE	100	GeoIP+Price/Default	6,7 €	0	0

INFN
Istituto Nazionale di Fisica Nucleare


FileA DOWLOAD
FileA
HTTP dCache
FileD
$
Cloud Storage
FileB
HTTP dCache
FileA ?
FileC ?
FileC
HTTP DPM
FileC DOWLOAD
FileB
TP STORM
DYNAFED FEDERATOR
FileC
FileD ?
$$$
FileD
FileD DOWLOAD
Cloud Storage
FileA
HTTP STORM
FileC ?
HTTP CACHE
FileC DOWLOAD
INFN
Istituto Nazionale di Fisica Nucleare

Preliminary Tests Details (File Download)

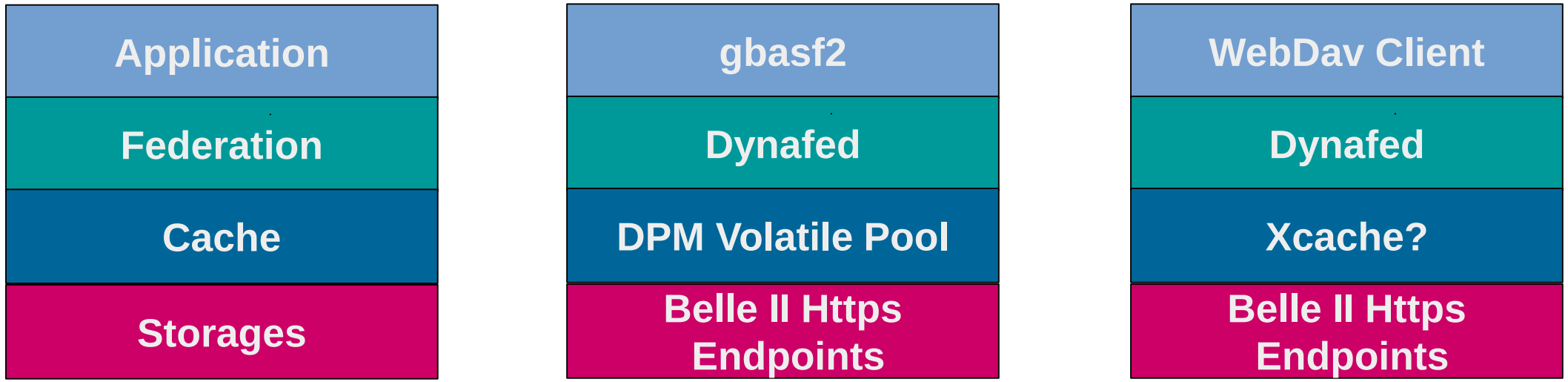
As preliminary test, we download from a **User Interface in Napoli** a set of Belle II files, stored in CESNET, KEK and UVic . Each file set is downloaded three times as follow:

- File Download using the direct link to the remote storage
- File Download using Dynafed with Cold cache
- File Download using Dynafed with Warm cache

Tests have been performed using files of different size: 50MB, 1GB

INFN
Istituto Nazionale di Fisica Nucleare

Dynafed and Cache: Model and implementation

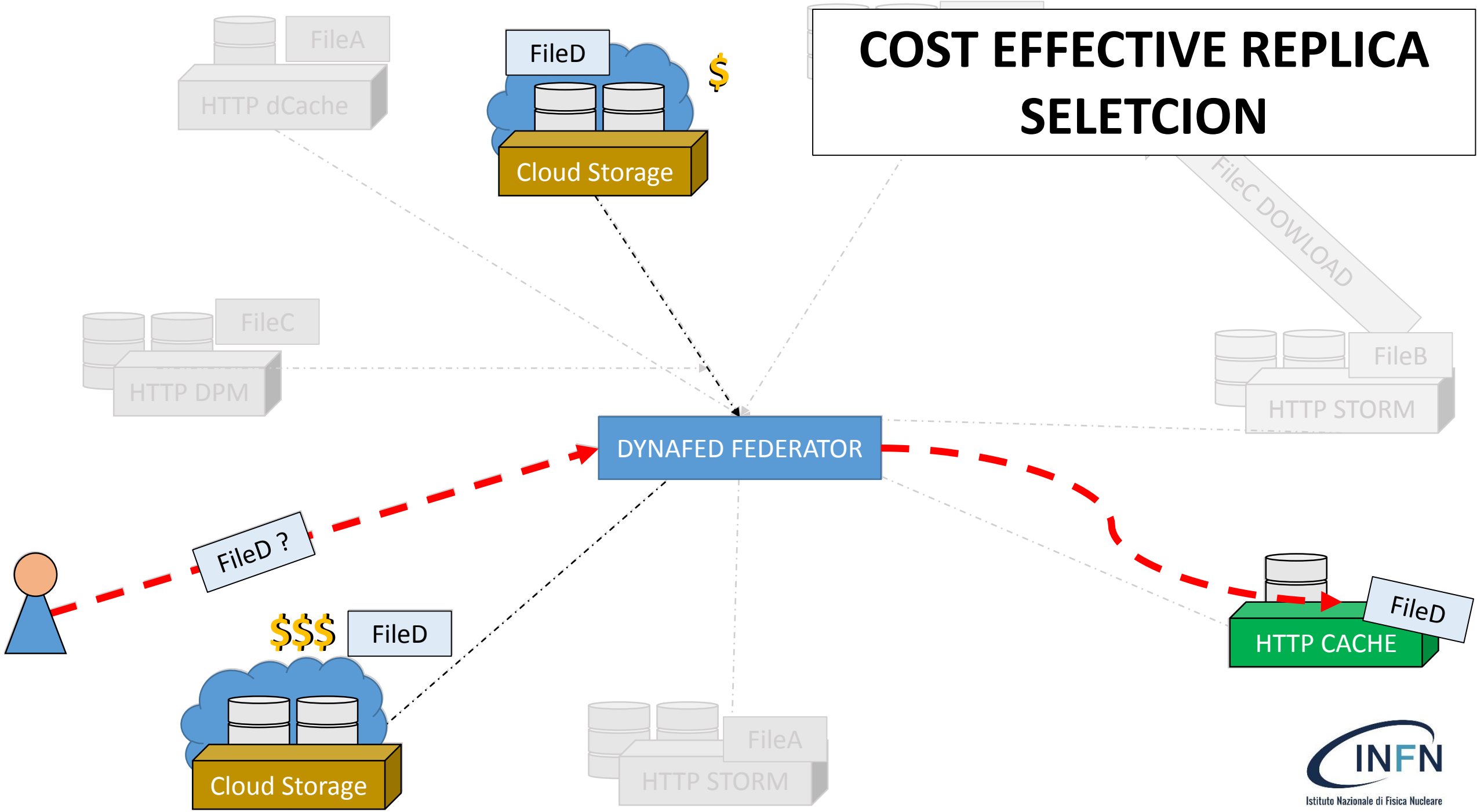


Test this model in Belle II require two steps:

- **Implement the caching system**
- **Study how to use HTTP/DAV in the application workflow**

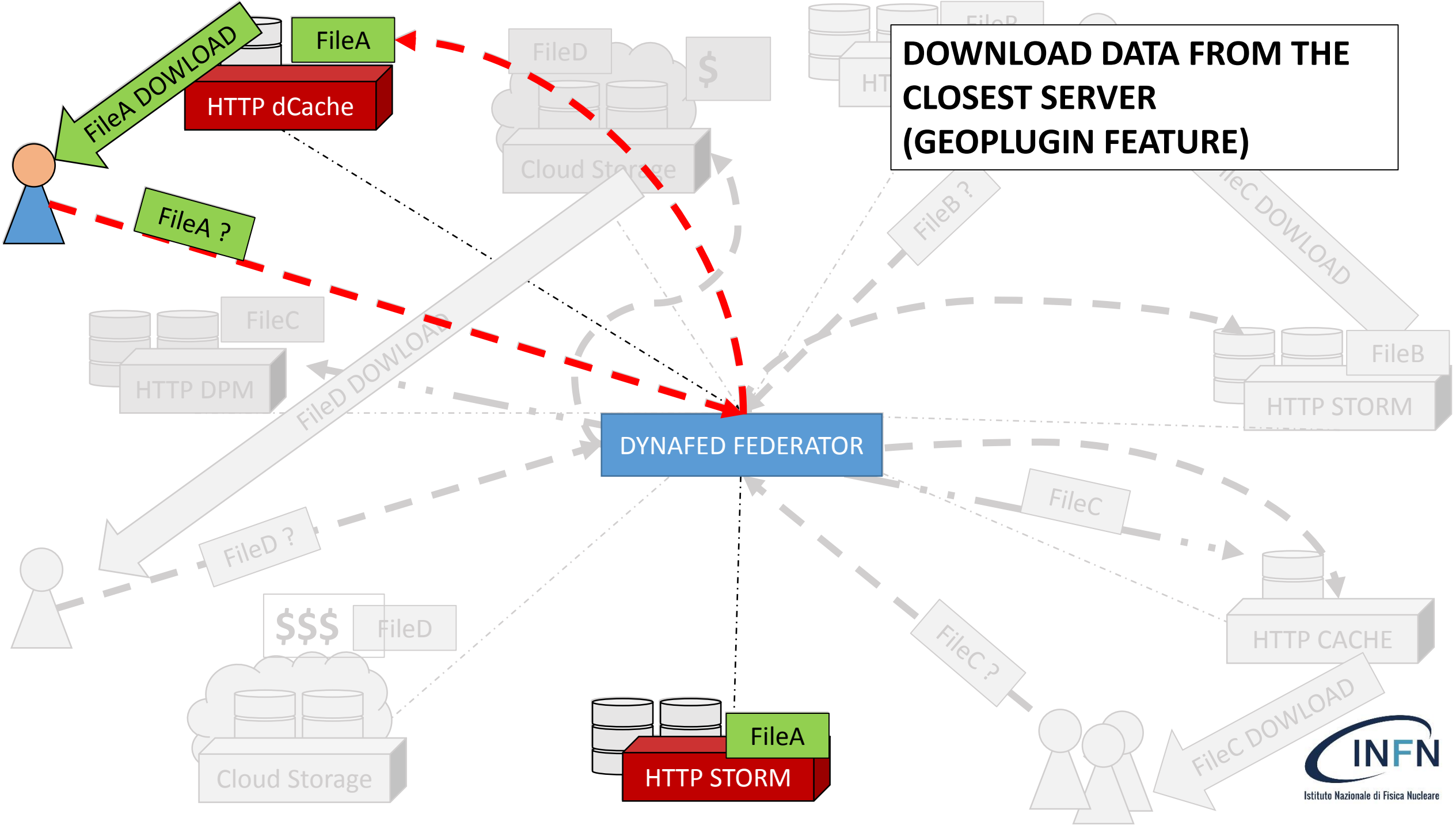
INFN
Istituto Nazionale di Fisica Nucleare

COST EFFECTIVE REPLICA SELETCION

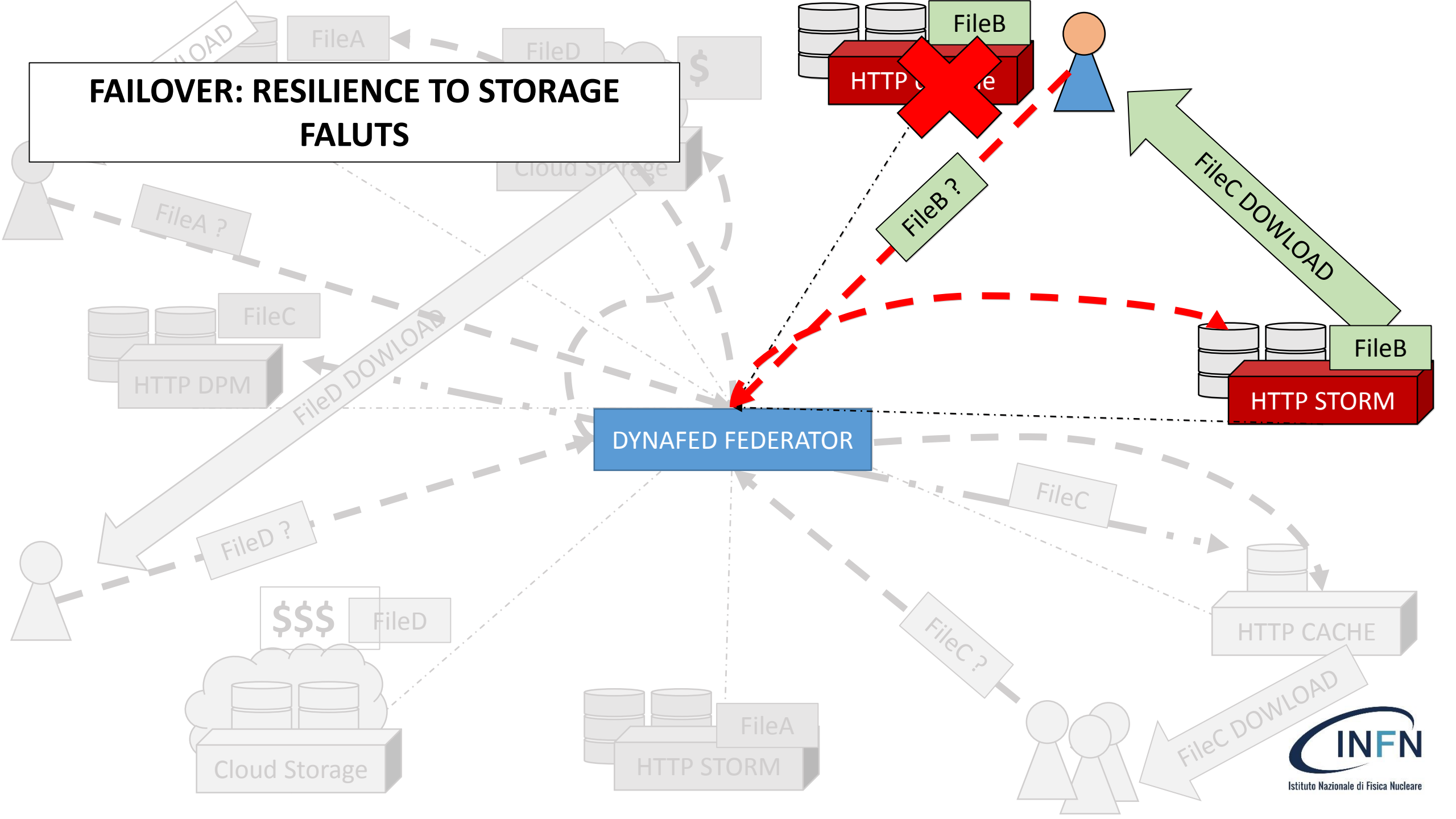


INFN
Istituto Nazionale di Fisica Nucleare

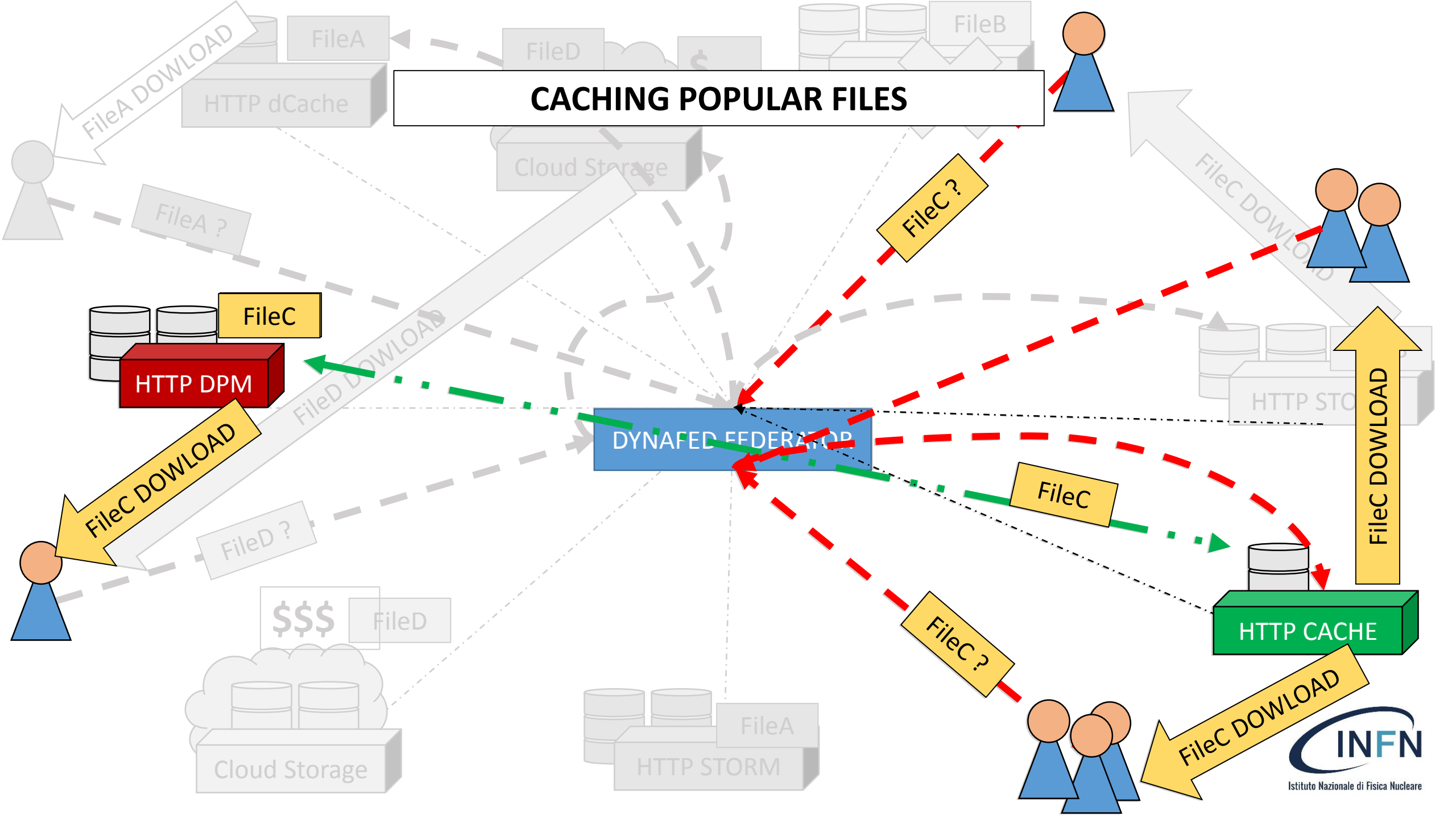
DOWNLOAD DATA FROM THE CLOSEST SERVER (GEOPLUGIN FEATURE)



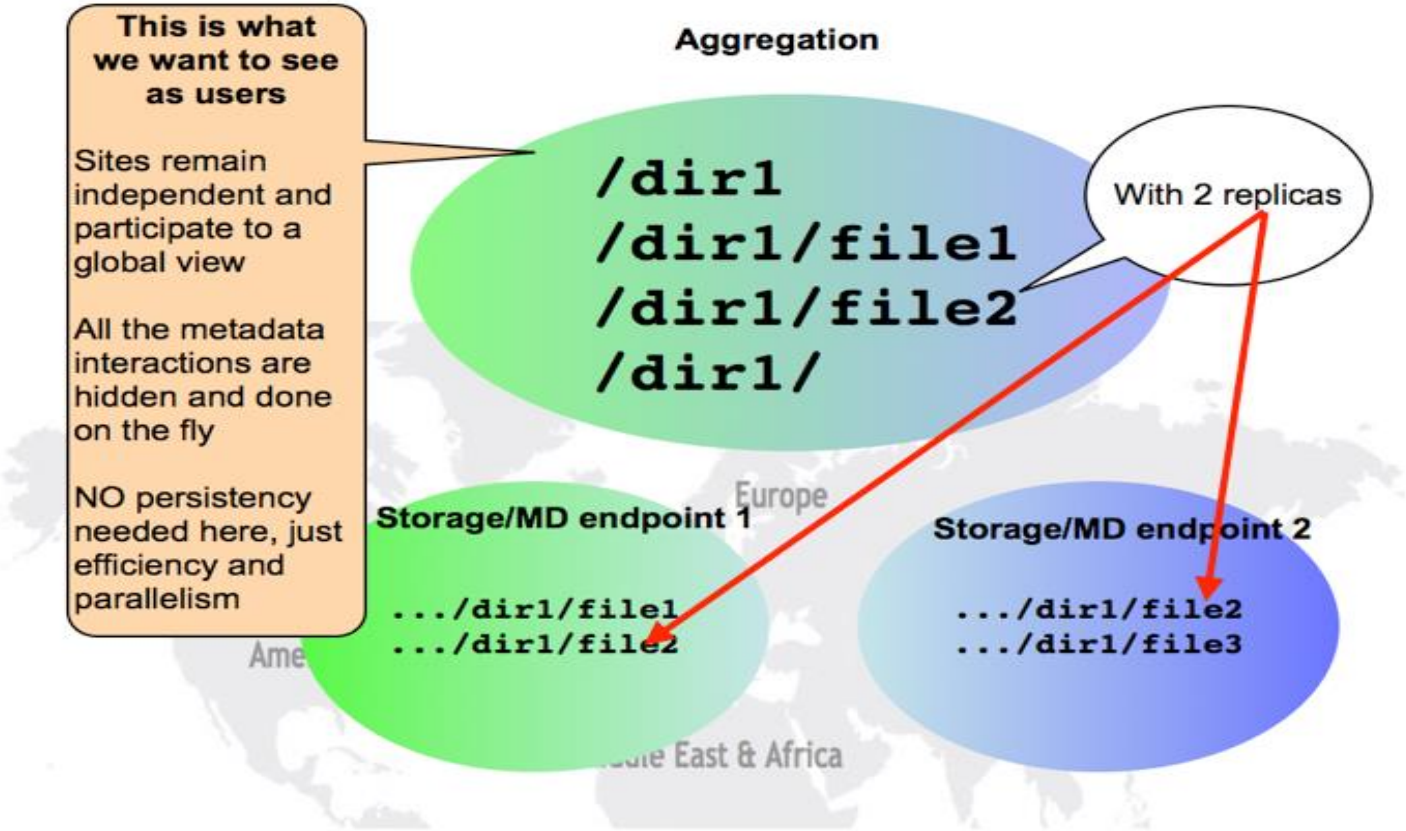
FAILOVER: RESILIENCE TO STORAGE FALUTS



CACHING POPULAR FILES



DYNAFED



Dynamic Federations system.
It can aggregate namespaces of different type of storages

- HTTP/Webdav Storage
- S3 storage
- NFS
- LFC
- Others

Storage aggregation is made on the fly
File metadata are cached on the Dynafed machine.
For the client point of view, Dynafed works as a redirector:
When a client ask for a file to it will be redirect the one of the available replicas.

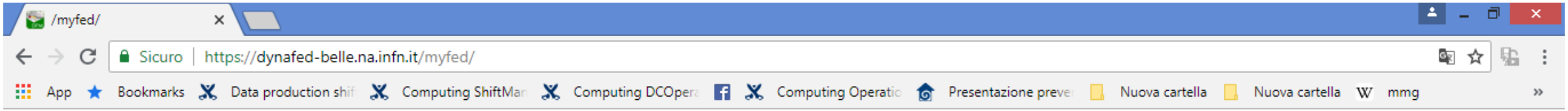
Dynafed file representation: Metalink

Sicuro | https://dynafed-belle.na.infn.it/myfed/belle/user/spardi/testhttp/mixed_e0001r0008_s00_BGx1.mdst.root?metalink

App | Bookmarks | Data production shif | Computing ShiftMar | Computing DCOpera | Computing Operatio | Presentazione preve | Nuova cartella | Nuova cartella | mmg

This XML file does not appear to have any style information associated with it. The document tree is shown below.

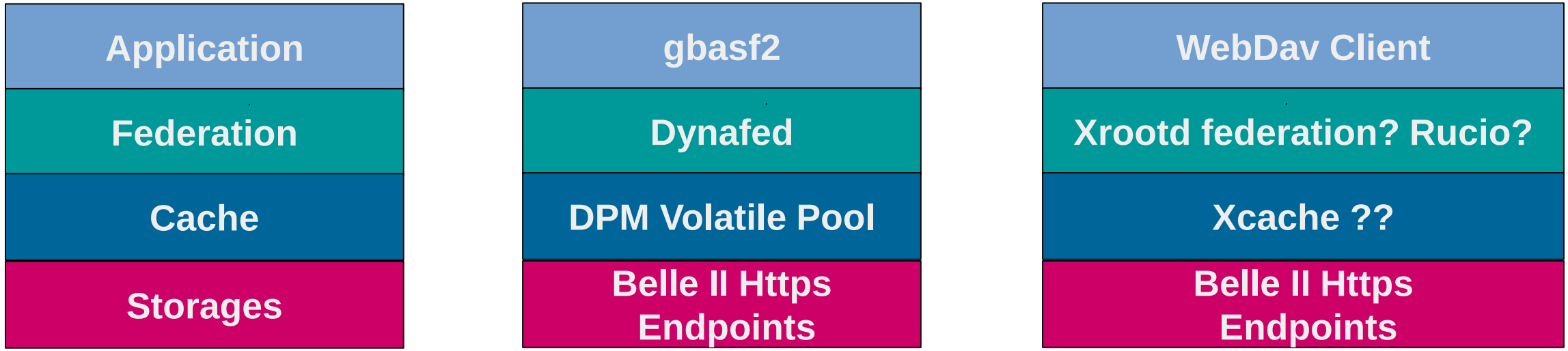

```
▼<metalink xmlns="http://www.metalinker.org/" xmlns:lcgdm="LCGDM:" version="3.0" generator="lcgdm-dav" pubdate="Wed, 13 Apr 2016 13:49:21 GMT">
  ▼<files>
    ▼<file name="/belle/">
      <size>11528882</size>
     ▼<resources>
      ▼<url type="https">
         https://kek2-se03.cc.kek.jp:8443/belle/TMP/belle/user/spardi/testhttp/mixed_e0001r0008_s00_BGx1.mdst.root
       </url>
      ▼<url type="https">
         http://bgrid3.phys.ntu.edu.tw:2880/pnfs/phys.ntu.edu.tw/home/belle/TMP/belle/user/spardi/testhttp/mixed_e0001r0008_s00_BGx1.mdst.root
       </url>
      ▼<url type="https">
         https://b2se.mel.coepp.org.au:443/dpm/mel.coepp.org.au/home/belle/bellescratchdisk/belle/TMP/belle/user/spardi/testhttp/mixed_e0001r0008_s00_BGx1.mdst.root
       </url>
      ▼<url type="https">
         https://dpm.cyf-kr.edu.pl:443/dpm/cyf-kr.edu.pl/home/belle/TMP/belle/user/spardi/testhttp/mixed_e0001r0008_s00_BGx1.mdst.root
       </url>
      ▼<url type="https">
         https://hephyse.oeaw.ac.at:443/dpm/oeaw.ac.at/home/belle/TMP/belle/user/spardi/testhttp/mixed_e0001r0008_s00_BGx1.mdst.root
       </url>
      ▼<url type="https">
         https://dpm1.egee.cesnet.cz:443/dpm/cesnet.cz/home/belle/TMP/belle/user/spardi/testhttp/mixed_e0001r0008_s00_BGx1.mdst.root
       </url>
```



/myfed/

Mode	Links	UID	GID	Size	Modified	Name
drwxrwxrwx	0	0	0	0	Thu, 01 Jan 1970 00:00:00 GMT	belle
drwxrwxrwx	0	0	0	0	Thu, 01 Jan 1970 00:00:00 GMT	belle-nocache

Dynafed and Cache: Model and implementation



Two challenges: User HTTP in the application workflow and implement a caching system

INFN
Istituto Nazionale di Fisica Nucleare