A federated file cache for Belle II

Dr. Silvio Pardi

INFN-Napoli

DPM Workshop 2018

Prague - 31/05/2018

INFN
Istituto Nazionale di Fisica Nucleare

SCOReS Project

Italian Acronym for: Study of a Caching system to optimize the usage of Opportunistic Resources and sites without pledged storage, for e-Science application(s) (SCOReS) end of the project 14 Feb 2019

Project funded by GARR within a National call consisting in a 2Year fellowship.

- Davide Michelino - project fellowship
- Silvio Pardi – Project Tutor for INFN-Napoli
- Prof. Guido Russo

Consortium GARR

INFN
Istituto Nazionale di Fisica Nucleare

Cache Use-Cases

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

Cache can affect performance in many scenarios:

- Increase performance of analysis jobs running on the same data-set.
- Cache improve performance for jobs running in sites geographically close to it
- Storage-Less Paradigms
- Cloud Storage
  - Limited bandwidth vs the clients
  - Limited number of free GET requests

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 Volatile Pool

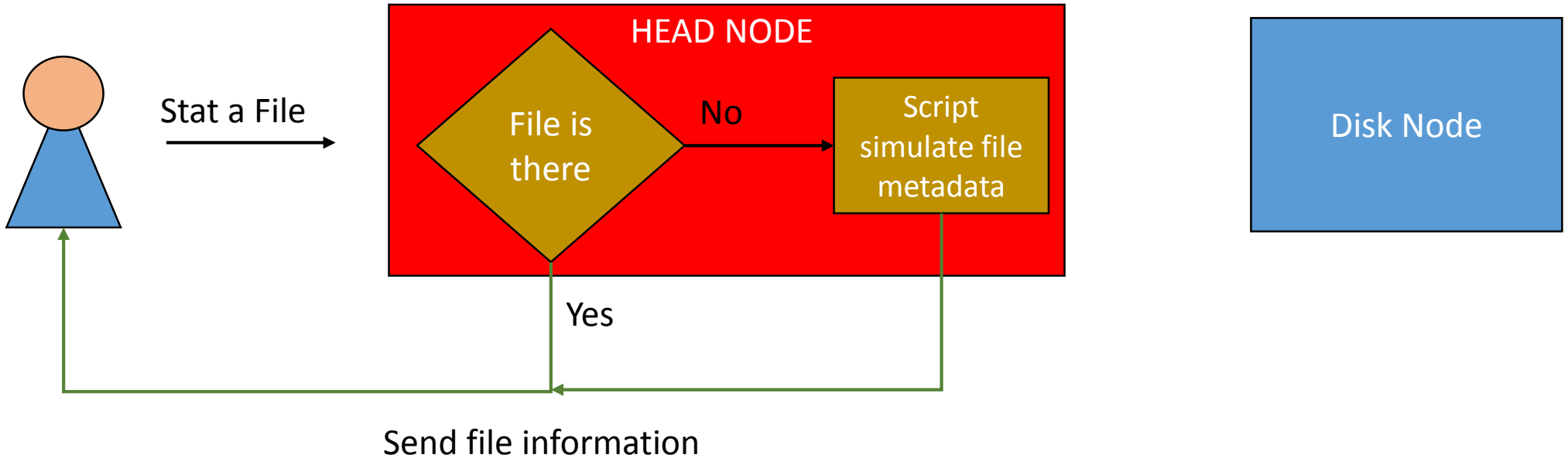
A **Volatile Pool** is a special pool that can download files from external sources.

Two main scripts:

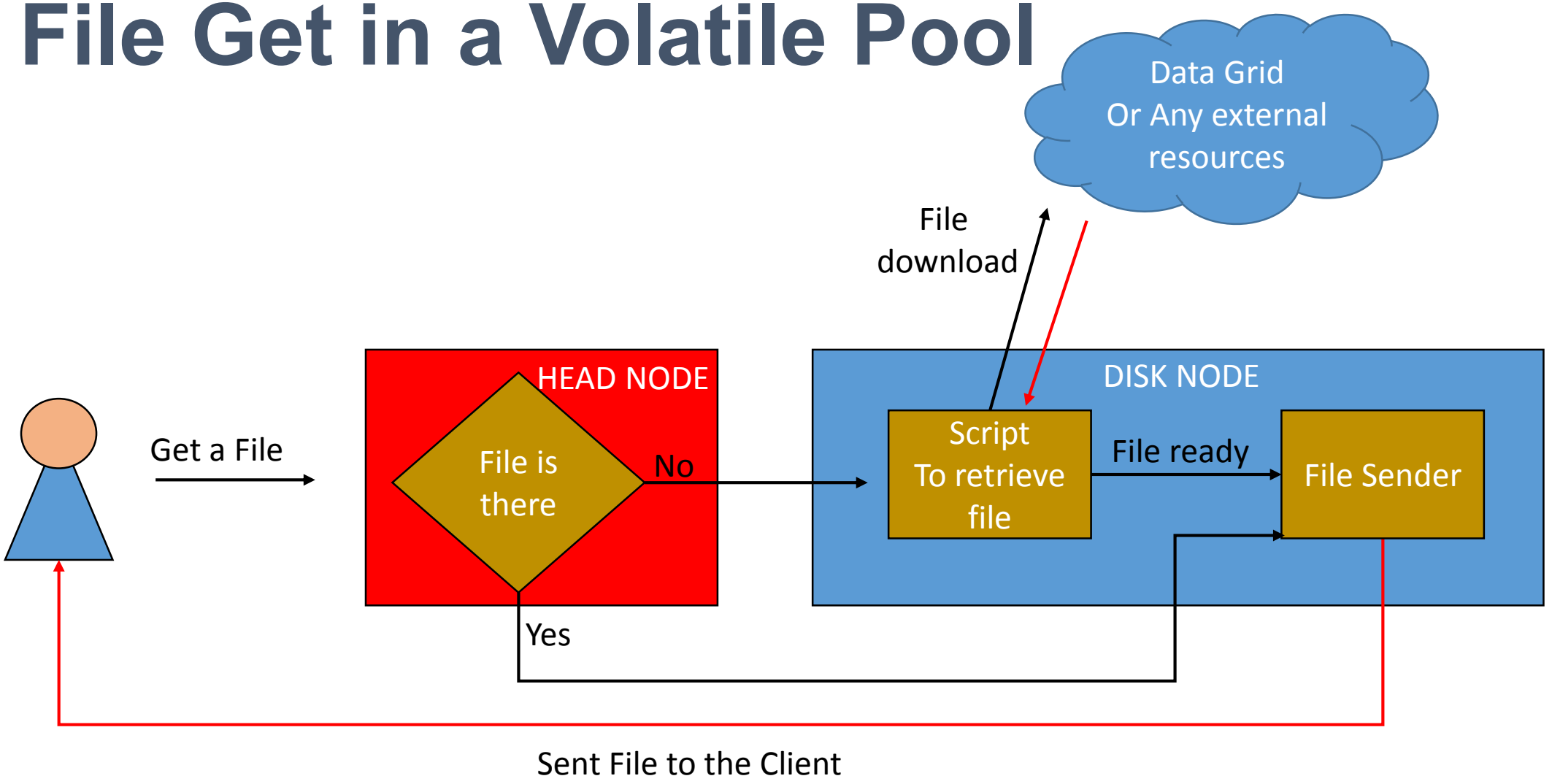
- **Script running on DPM head node that manage stat operation**
- **Script running in Disk Nodes responsible to get file from external sources**

INFN
Istituto Nazionale di Fisica Nucleare

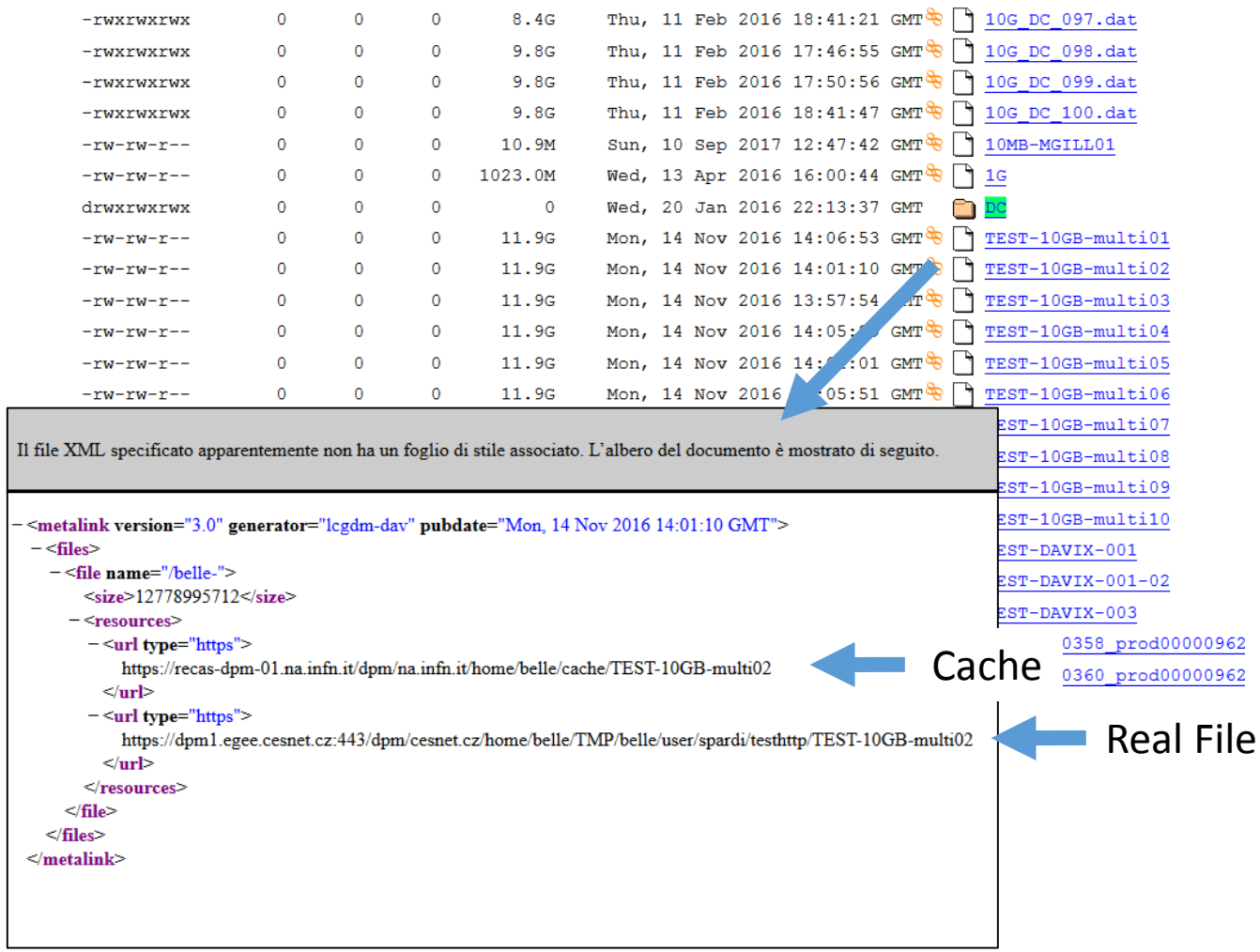
File Stat in a Volatile Pool



File Get in a Volatile Pool



Dynafed + Volatile Pool



What happen if we aggregate a standard http endpoint with a DPM Volatile Pool?

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

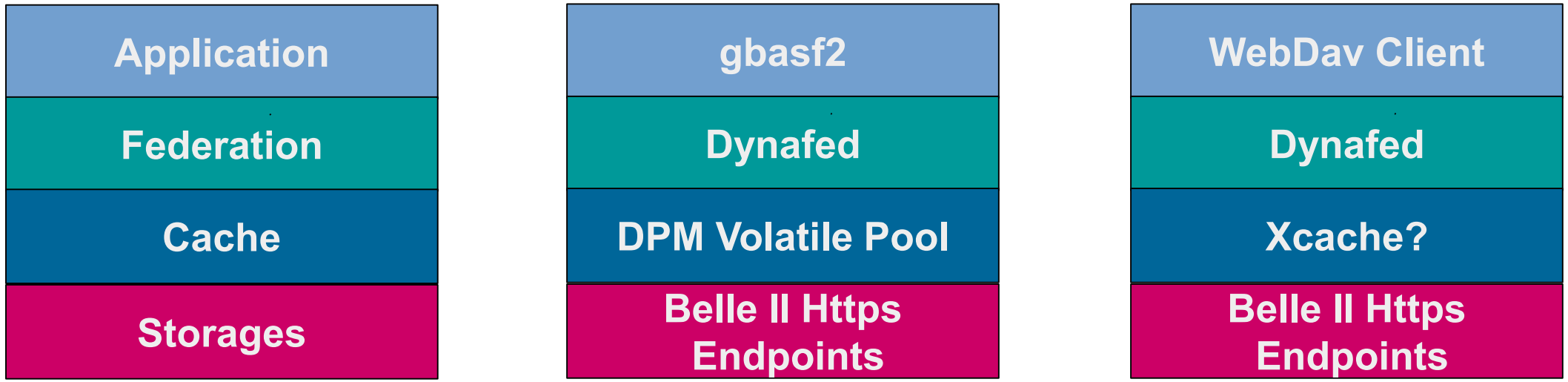
So that the metalink representing a file in Dynafed, will included always at least two link: the real URL and the corresponding virtual copy in the cache (even if the latter does not exist yet)

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

Dynafed and Cache: Model and implementation

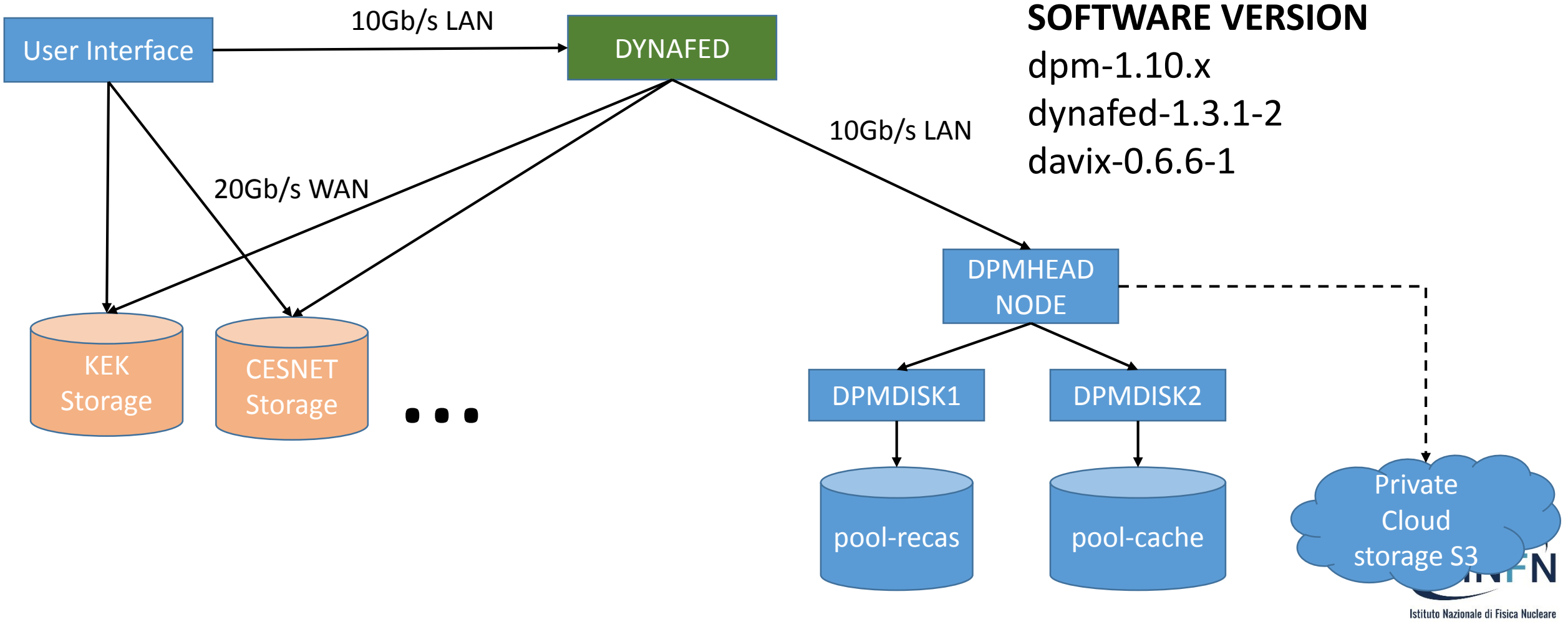


Test this model in Belle II require two steps:

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

INFN
Istituto Nazionale di Fisica Nucleare

The testbed



SOFTWARE VERSION

dpm-1.10.x

dynafed-1.3.1-2

davix-0.6.6-1

Istituto Nazionale di Fisica Nucleare

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 SRM production (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 recognize 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.

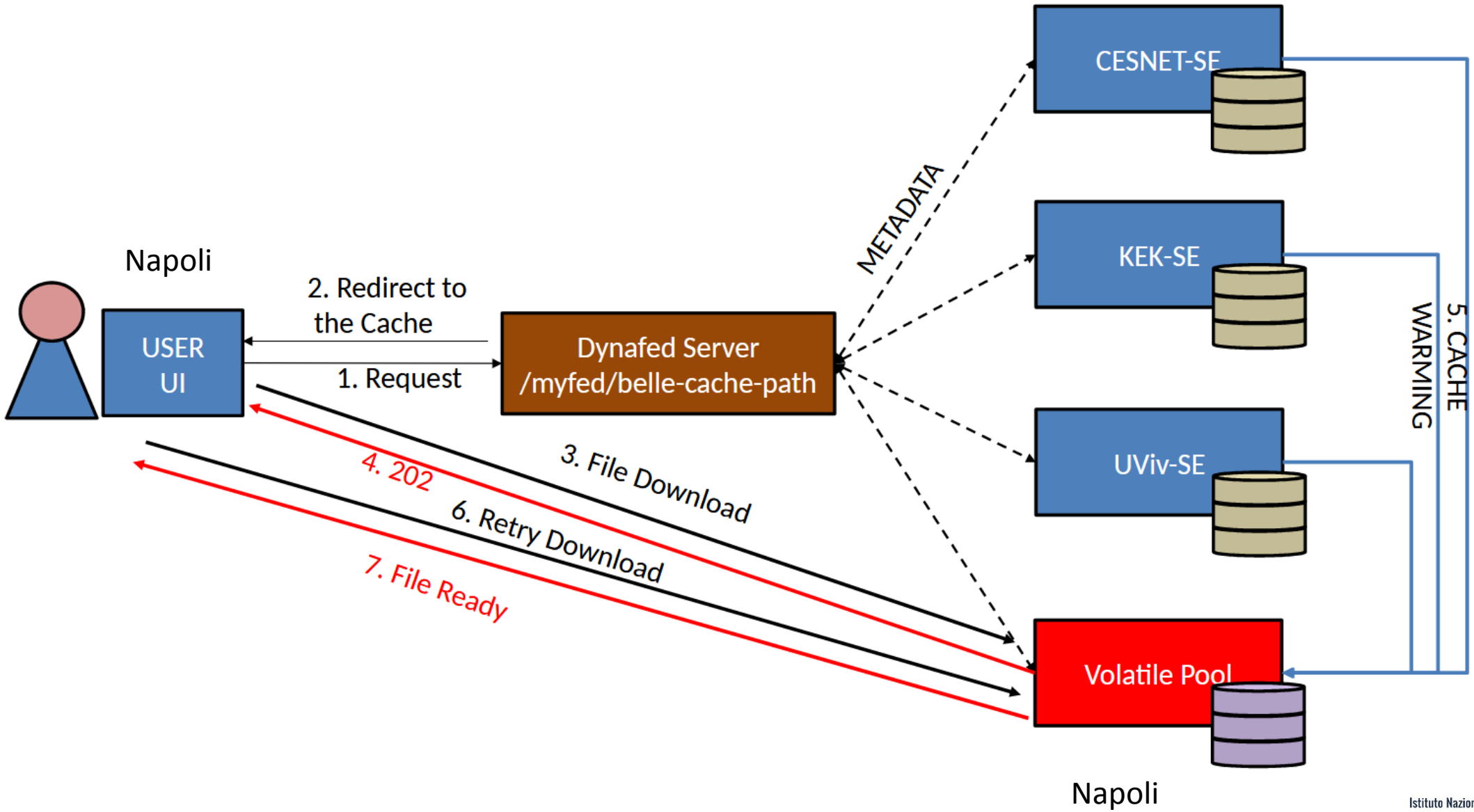
Script on the Disk Node:

When a file is not in the cache, the disk node download the requested file from the datagrid by resolving the location via Dynafed

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

Implementation Detail



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:

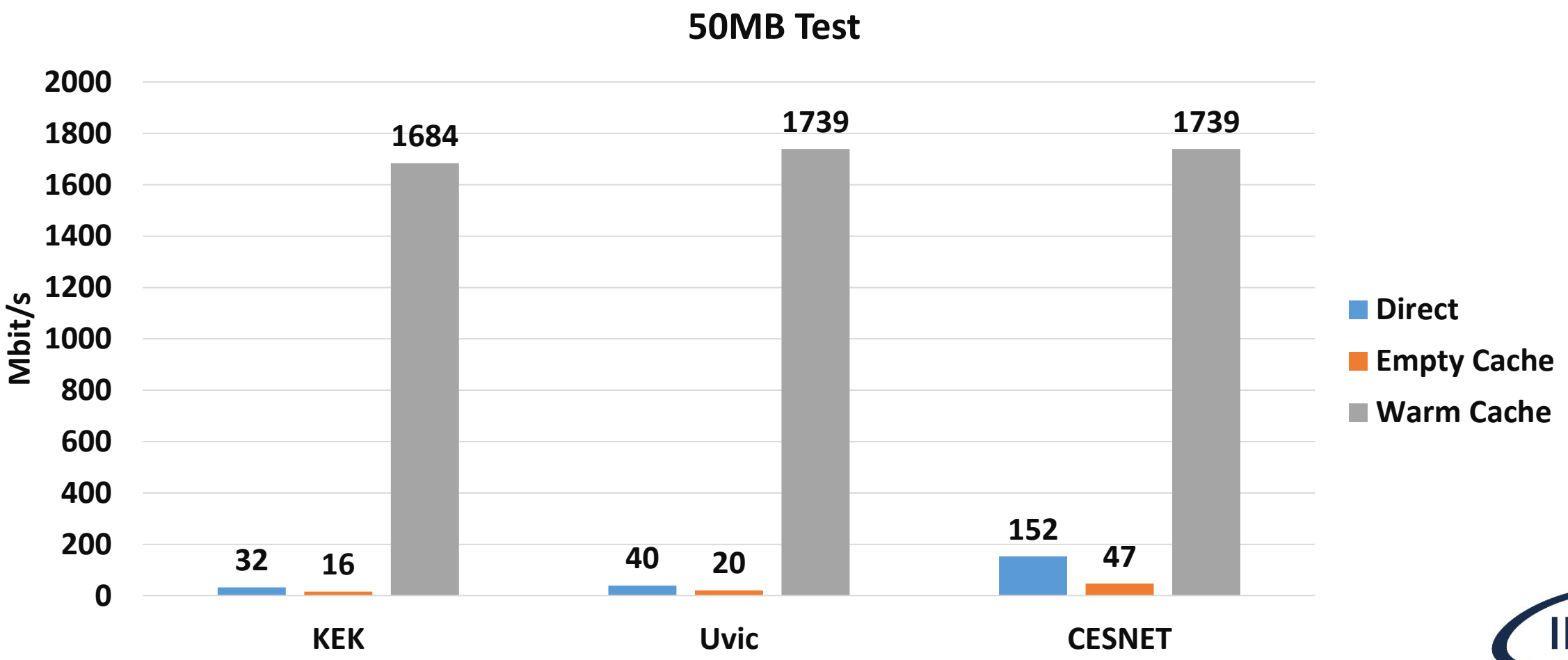
- File Download using the direct link to the remote storage
- File Download using Dynafed with Cold cache
- File Download using Dynafed with Warm cache

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

INFN
Istituto Nazionale di Fisica Nucleare

File Download 50MB

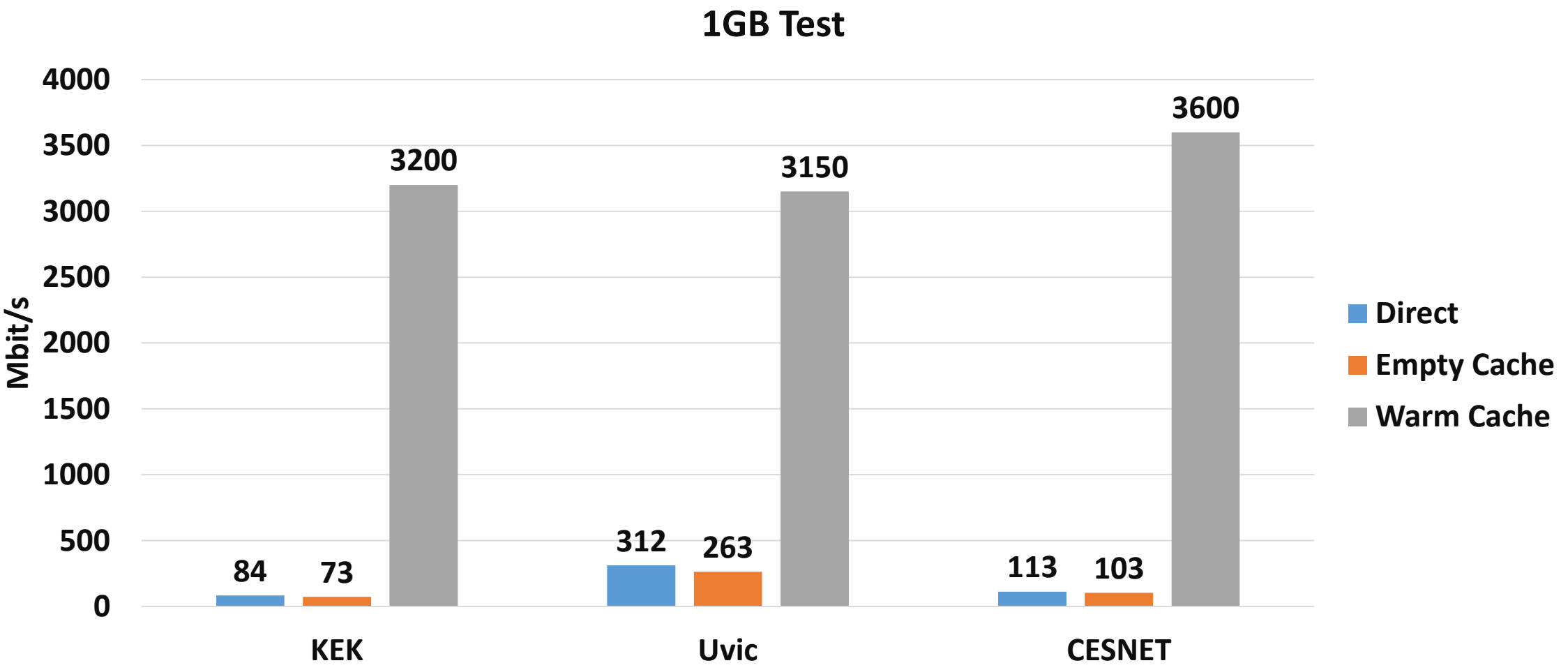
Mbit/s (Higher is better)



INFN
Istituto Nazionale di Fisica Nucleare

File Download Test 1GB

Mbit/s (Higher is better)



Istituto Nazionale di Fisica Nucleare

Local job reading file in streaming

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

Total time 0m50.516s With Dynafed+Cache (cold cache)

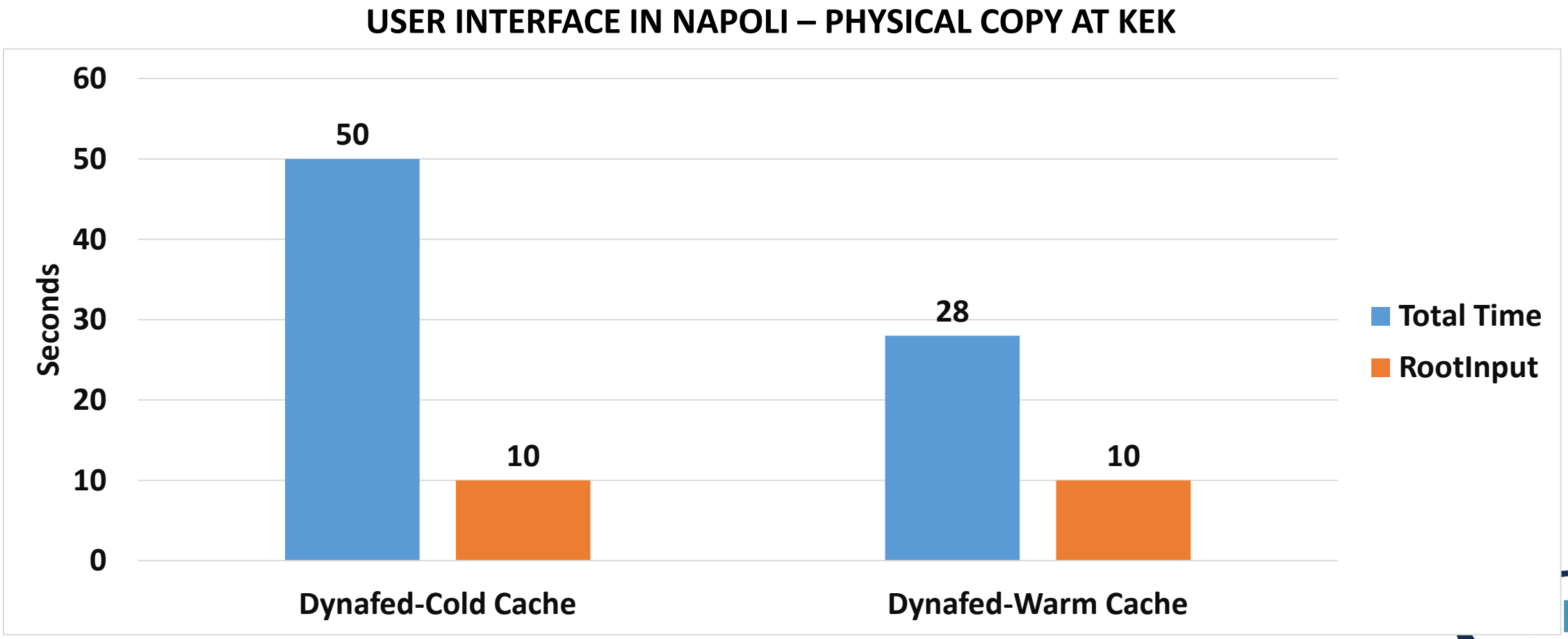
RootInput 10.30(s) -1.03 per event

Total time 0m28.993s With Dynafed+Cache (warm cache)

RootInput 9.24(s) – 0.92 per event

Test - Local job reading file in streaming

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



Status of the R&D activity

A minimal set of components has been setup to create an http caching system in a federated environment.

There are still some aspects to investigate.

Massive stress tests have to be performed to validate the setup check the stability the whole solution.

Davide Michelino has been selected in the "Future Talent Programme" promoted by GEANT and will present the work at the next TCN18 in June.

INFN
Istituto Nazionale di Fisica Nucleare

Status of the R&D activity: Issue

Webmail | DPM Workshop 2018 (31 | /myfed/belle/ | https://dynafed-belle.na | dCache File System

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

App Bookmarks Data production shif Computing ShiftMan Computing DCOpera Computing Operatic Presentazione preve Nuova cartella Nuova cartella mmg

/myfed/belle/

Mode	Links	UID	GID	Size	Modified	Name
drwxrwxr--	0	0	0	0	Fri, 02 Mar 2018 14:48:15 GMT	DATA
drwxrwxr-x	0	0	0	0	Thu, 24 Sep 2015 18:25:16 GMT	DC
drwxrwxr-x	0	0	0	0	Fri, 23 Mar 2018 13:20:50 GMT	DR3
drwxrwxrwx	0	0	0	0	Wed, 16 Nov 2016 05:56:50 GMT	Data
drwxrwxrwx	0	0	0	0	Tue, 08 May 2018 19:37:34 GMT	MC
drwxrwxr--	0	0	0	17.0M	Sun, 11 Feb 2018 21:40:02 GMT	TMP
-rw-rw-r--	0	0	0	0	Thu, 24 May 2018 13:11:48 GMT	aa1
drwxrwxrwx	0	0	0	0	Wed, 07 Jan 2015 05:00:56 GMT	data
drwxrwxr-x	0	0	0	0	Tue, 15 Mar 2016 22:01:49 GMT	ddm_test
drwxrwxrwx	0	0	0	0	Wed, 16 Nov 2016 05:57:21 GMT	group
-rw-rw-r--	0	0	0	10	Fri, 20 Apr 2018 14:03:16 GMT	kfox-hello.txt
drwxrwxr-x	0	0	0	0	Tue, 15 Sep 2015 04:29:24 GMT	monitor
-rw-rw-r--	0	0	0	0	Fri, 25 May 2018 15:11:05 GMT	spardi

This is a directory but during aggregation Dynafed cached it as a file

App Bookmarks Data production shif Computing ShiftMan Computing DCOpera Computing Operatic Prese

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="Fri, 2
  <files>
    <file name="/belle/">
      <size>0</size>
      <resources>
        <url type="https">
          https://recas-dpm-01.na.infn.it/dpm/na.infn.it/home/belle/cache1/spardi
        </url>
        <url type="https">
          https://charon01.westgrid.ca:2880/pnfs/westgrid.uvic.ca/data/belle/belledisk/belle/TMP/belle/spardi
        </url>
      </resources>
    </file>
  </files>
</metalink>
```

11:40
31/05/2018

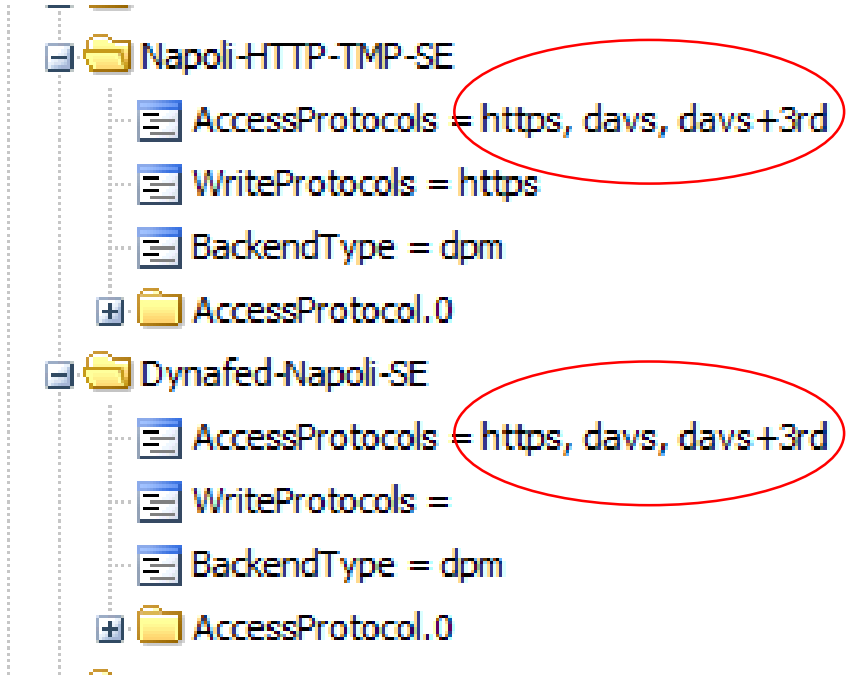
INFN Istituto Nazionale di Fisica Nucleare

Client Behaviour

- If the cache is not ready, 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

Question: would be possible to redirect the client to the next URL after the 202 message? (only if the client ask for it)

Cache in the Belle II Analysis Workflow via DIRAC



Andrea Spiezia BSc Thesis

Belle II uses the concept of dataset and datablock.

A datablock is a collection of file, a dataset is a collection of datablock.

User uses the logical name of the datablock as input for analysis jobs.

The framework gbasf2 is responsible to complete file lookup using AMGA as metadata catalog and LFC as file catalog.

The final URL is obtained by concatenating the storage information stored in DIRAC configuration which contain the access protocol as well.

To test the usage of DAVS in the whole chain, two storages has been crated in the DIRAC configuration of the validation server in BNL.

- Napoli-HTTP-TMP-SE
- Dynafed-Napoli-SE

DAVS protocol in a gbasf2 analysis

Ongoing test are focussed on three main use-cases:

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

At now we are still working on the first use-cases trying to resolve some Input File Resolution issue.

INFN
Istituto Nazionale di Fisica Nucleare

FilterPlugin for dynafed

Goal: ordering file replicas in the Dynafed metalink on the base of predefined matrix costs associate to a set of endpoints.

Use cases:

- If we aggregate a set of S3 storages we would like to prioritize the cheapest ones
- Any other ordering politics that can be expressed as a function cost.

Salvatore Lanzilli - BSc Thesis

INFN
Istituto Nazionale di Fisica Nucleare

FilterPlugin

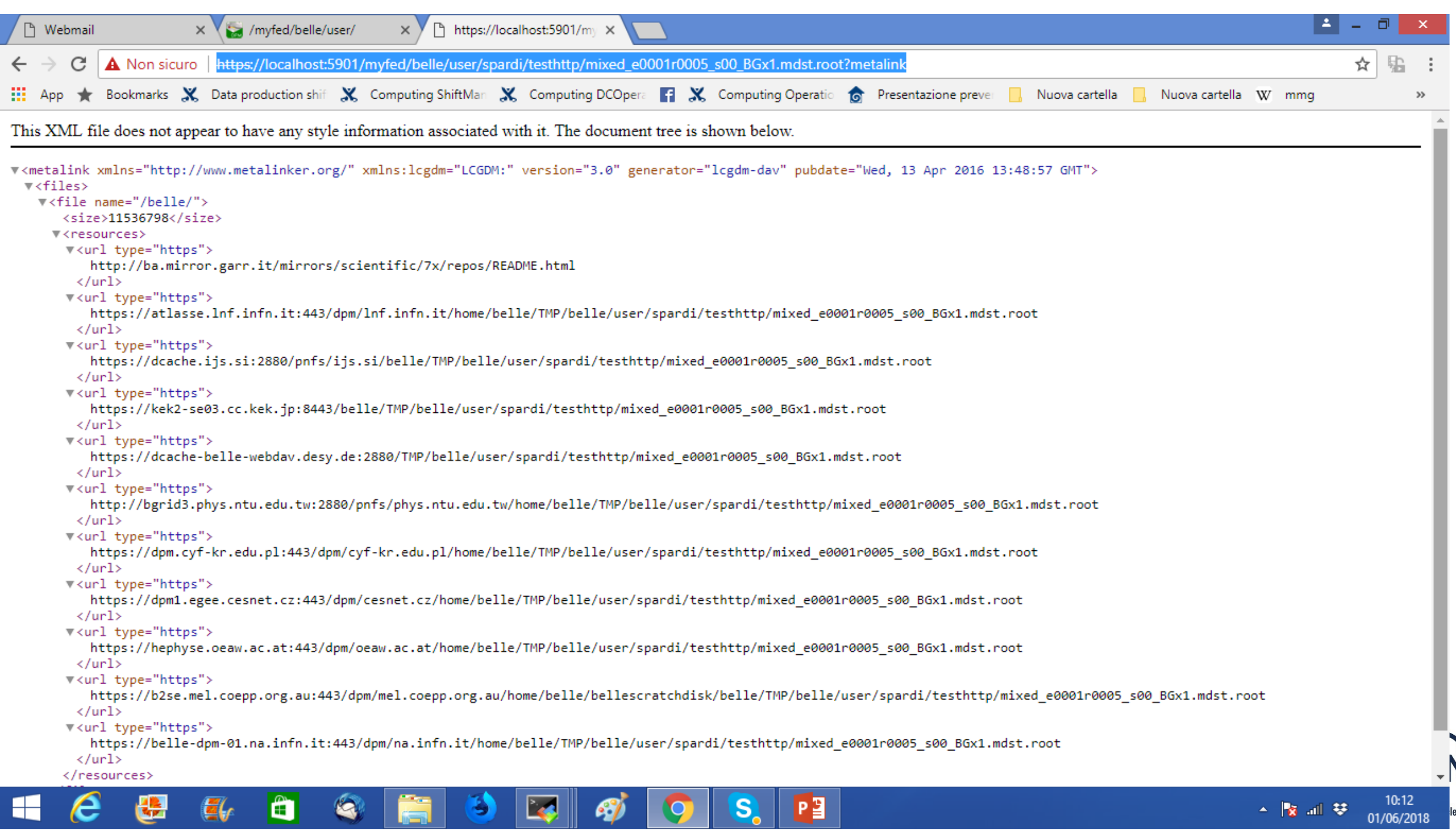
/etc/ugr/ugr.conf

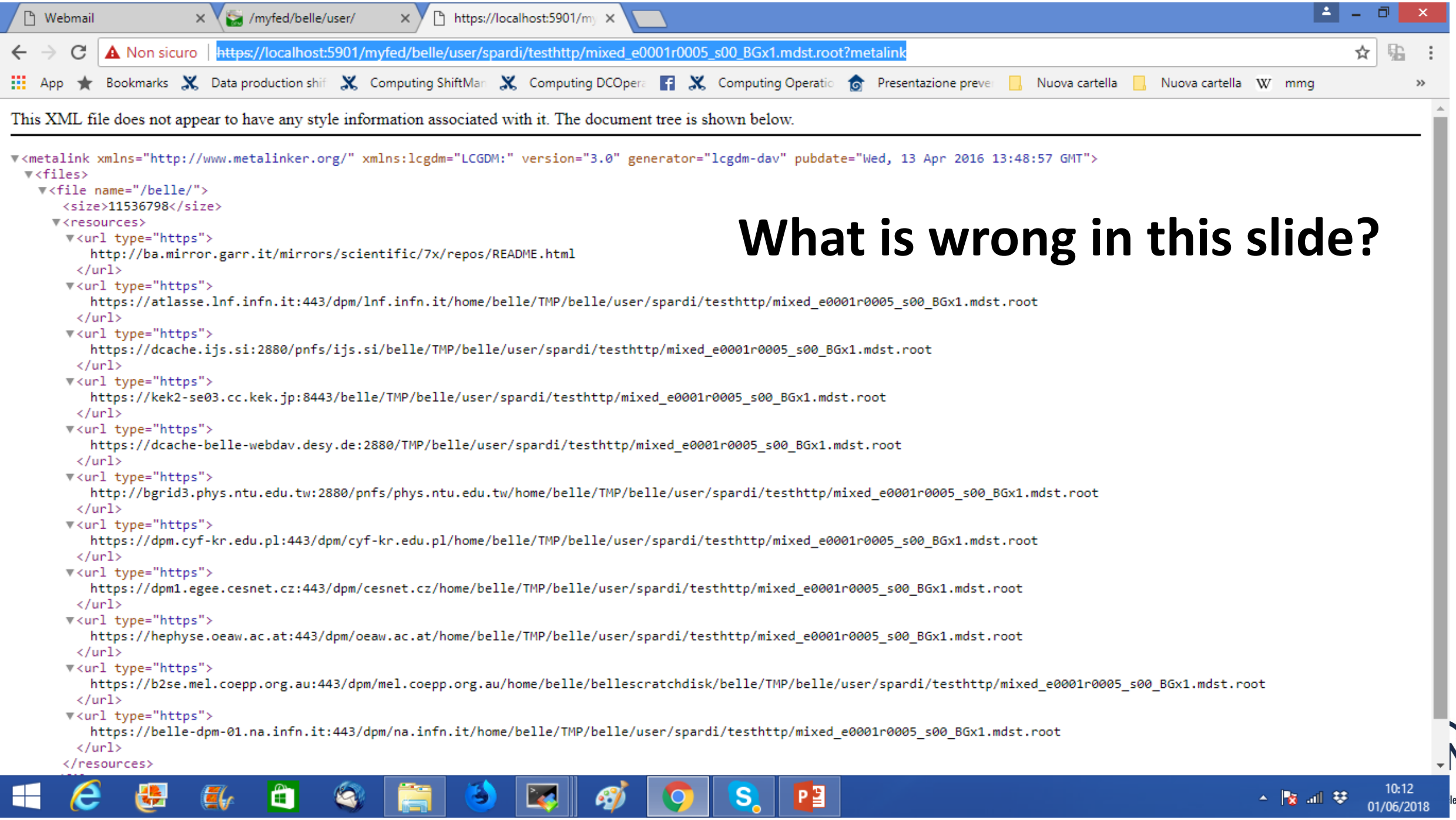
```
locplugin.*.cli_certificate: /tmp/x509up_u0
locplugin.*.cli_private_key: /tmp/x509up_u0
locplugin.*.xlatepfx: /belle-cache-path/ /
#locplugin.*.xlatepfx: /site-base-path/belle /
locplugin.*.cli_proxy_cert: /tmp/x509up_u0

################################################
# Include all the config files contained in the
# specified directory
# Remember that the ugr config parameters are not positional!
#
INCLUDE /etc/ugr/conf.d/

################################################
#
# Price Plugin
glb.filterplugin[]: libugrpriceplugin.so price_plugin /etc/ugr/conf.d/endpoints_price.conf

################################################
```

1. spardi@aau01:~ 2. root@dynafed01:/etc/ugr

```
sasybucket.obs.otc.t-systems.com 0.50
davide.obs.otc.t-systems.com 0.80
atlasse.lnf.infn.it 0.03
dcache.ijs.si 0.14
kek2-se03.cc.kek.jp 0.26
ba.mirror.garr.it 0.0
~
```

INFN
Istituto Nazionale di Fisica Nucleare


Webmail
/myfed/belle/user/
https://localhost:5901/my
Non sicuro
https://localhost:5901/myfed/belle/user/spardi/testhttp/mixed_e0001r0005_s00_BGx1.mdst.root?metalink
App
Bookmarks
Data production shif
Computing ShiftMan
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 13:48:57 GMT">
<files>
<file name="/belle/">
<size>11536798</size>
<resources>
<url type="https">
http://ba.mirror.garr.it/mirrors/scientific/7x/repos/README.html
</url>
<url type="https">
https://atlasse.lnf.infn.it:443/dpm/lnf.infn.it/home/belle/TMP/belle/user/spardi/testhttp/mixed_e0001r0005_s00_BGx1.mdst.root
</url>
<url type="https">
https://dcache.ijs.si:2880/pnfs/ijs.si/belle/TMP/belle/user/spardi/testhttp/mixed_e0001r0005_s00_BGx1.mdst.root
</url>
<url type="https">
https://kek2-se03.cc.kek.jp:8443/belle/TMP/belle/user/spardi/testhttp/mixed_e0001r0005_s00_BGx1.mdst.root
</url>
<url type="https">
https://dcache-belle-webdav.desy.de:2880/TMP/belle/user/spardi/testhttp/mixed_e0001r0005_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_e0001r0005_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_e0001r0005_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_e0001r0005_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_e0001r0005_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_e0001r0005_s00_BGx1.mdst.root
</url>
<url type="https">
https://belle-dpm-01.na.infn.it:443/dpm/na.infn.it/home/belle/TMP/belle/user/spardi/testhttp/mixed_e0001r0005_s00_BGx1.mdst.root
</url>
</resources>
10:12
01/06/2018

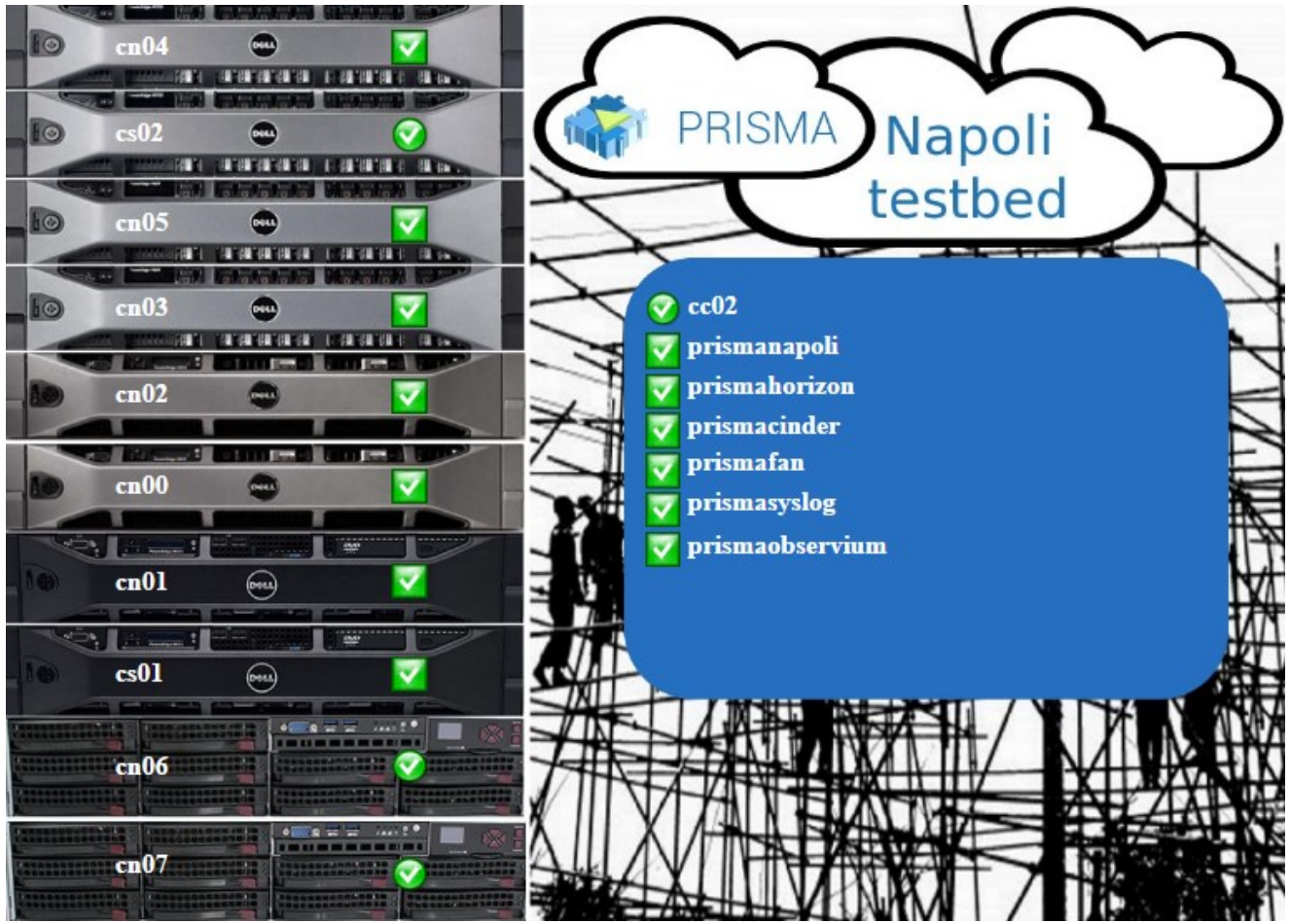


What is wrong in this slide?


Webmail
/myfed/belle/user/
https://localhost:5901/my
Non sicuro
https://localhost:5901/myfed/belle/user/spardi/testhttp/mixed_e0001r0005_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 13:48:57 GMT">
<files>
<file name="/belle/">
<size>11536798</size>
<resources>
<url type="https">
http://ba.mirror.garr.it/mirrors/scientific/7x/repos/README.html
</url>
<url type="https">
https://atlasse.lnf.infn.it:443/dpm/lnf.infn.it/home/belle/TMP/belle/user/spardi/testhttp/mixed_e0001r0005_s00_BGx1.mdst.root
</url>
<url type="https">
https://dcache.ijs.si:2880/pnfs/ijs.si/belle/TMP/belle/user/spardi/testhttp/mixed_e0001r0005_s00_BGx1.mdst.root
</url>
<url type="https">
https://kek2-se03.cc.kek.jp:8443/belle/TMP/belle/user/spardi/testhttp/mixed_e0001r0005_s00_BGx1.mdst.root
</url>
<url type="https">
https://dcache-belle-webdav.desy.de:2880/TMP/belle/user/spardi/testhttp/
</url>
<url type="https">
http://bgrid3.phys.ntu.edu.tw:2880/pnfs/phys.ntu.edu.tw/home/belle/TMP/b
</url>
<url type="https">
https://dpm.cyf-kr.edu.pl:443/dpm/cyf-kr.edu.pl/home/belle/TMP/belle/use
</url>
<url type="https">
https://dpm1.egee.cesnet.cz:443/dpm/cesnet.cz/home/belle/TMP/belle/user/
</url>
<url type="https">
https://hephyse.oeaw.ac.at:443/dpm/oeaw.ac.at/home/belle/TMP/belle/user/
</url>
<url type="https">
https://b2se.mel.coepp.org.au:443/dpm/mel.coepp.org.au/home/belle/belle
</url>
<url type="https">
https://belle-dpm-01.na.infn.it:443/dpm/na.infn.it/home/belle/TMP/belle/user/spardi/testhttp/mixed_e0001r0005_s00_BGx1.mdst.root
</url>
</resources>
This entry has been added by the filter plugin.
The idea is try to investigate if we can take advantage to have a default storage (maybe a cache?) added by the filter (if activated) for a set of not well defined use-cases :D
10:12
01/06/2018

Thank you

Facilities



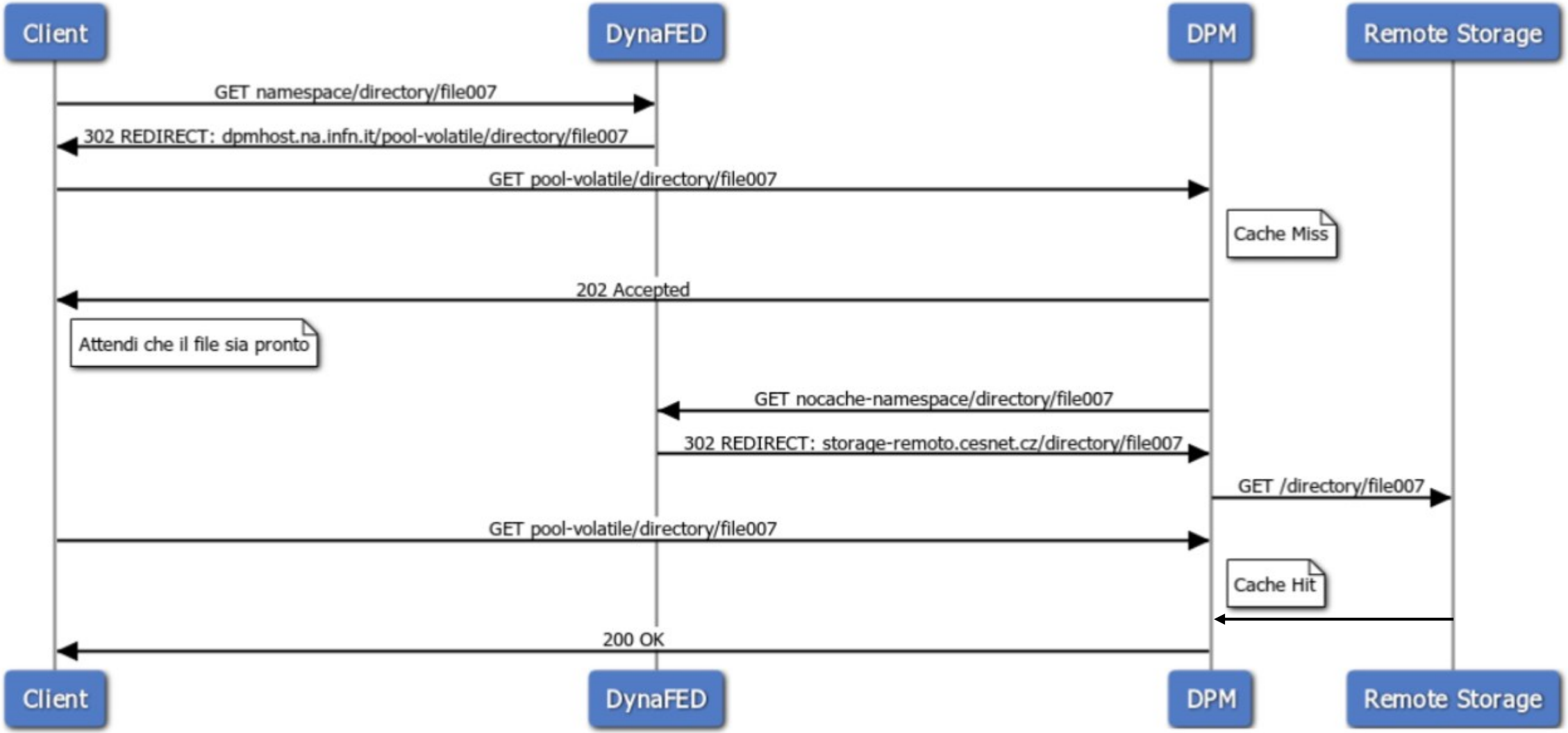
The project will integrate the caching system in the RECAS-Napoli infrastructure supporting belle II and Atlas experiment. The goal is to create a pilot system and if possible a pre-production services

For the testbed we can take advantage from a local cloud based on Openstack, with the following characteristics

- 2 Server (tot 80 Cores to store the collective service)
- 384 cores for computation
- 88TB Raw Data
- 10Gbps Network

INFN
Istituto Nazionale di Fisica Nucleare

OUR IMPLEMENTATION



Status of the R&D activity: Issue 2

Sicuro | https://dynafed-belle.na.infn.it/myfed/belle/MC/release-01-00-02/DB00000294/BG15/prod00004167/s00/e0000/phase2/r00000/beambgx2/s

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

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="Sat, 19 May 2018 09:36:51 GMT">
  <files>
    <file name="/belle/">
      <size>1363986205</size>
      <resources>
        <url type="https">
          https://dcache-belle-webdav.desy.de:2880/DATA/belle/MC/release-01-00-
          02/DB00000294/BG15/prod00004167/s00/e0000/phase2/r00000/beambgx2/sub00/bgoverlay_000001_prod00004167_task00000001.root
        </url>
        <url type="https">
          https://dcachewebdav-kit.gridka.de:2880/pnfs/gridka.de/belle/disk-only/DATA/belle/MC/release-01-00-
          02/DB00000294/BG15/prod00004167/s00/e0000/phase2/r00000/beambgx2/sub00/bgoverlay_000001_prod00004167_task00000001.root
        </url>
        <url type="https">
          https://belle-dpm-01.na.infn.it:443/dpm/na.infn.it/home/belle/DATA/belle/MC/release-01-00-
          02/DB00000294/BG15/prod00004167/s00/e0000/phase2/r00000/beambgx2/sub00/bgoverlay_000001_prod00004167_task00000001.root
        </url>
        <url type="https">
          https://recas-dpm-01.na.infn.it/dpm/na.infn.it/home/belle/cache1/MC/release-01-00-
          02/DB00000294/BG15/prod00004167/s00/e0000/phase2/r00000/beambgx2/sub00/bgoverlay_000001_prod00004167_task00000001.root
        </url>
        <url type="https">
          https://kek2-se03.cc.kek.jp:8443/belle/DATA/belle/MC/release-01-00-
          02/DB00000294/BG15/prod00004167/s00/e0000/phase2/r00000/beambgx2/sub00/bgoverlay_000001_prod00004167_task00000001.root
        </url>
        <url type="https">
          https://kek2-se03.cc.kek.jp:8443/belle/TMP/belle/MC/release-01-00-
          02/DB00000294/BG15/prod00004167/s00/e0000/phase2/r00000/beambgx2/sub00/bgoverlay_000001_prod00004167_task00000001.root
        </url>
      </resources>
    </file>
  </files>
</metalink>
```


Randomly this entry disappear. By reloading the page, we see it appearing and disappearing.

INFN Istituto Nazionale di Fisica Nucleare

Dynafed Setup

Two views configured:

1. Aggregation of a set of Belle II storage endpoints [path /belle]

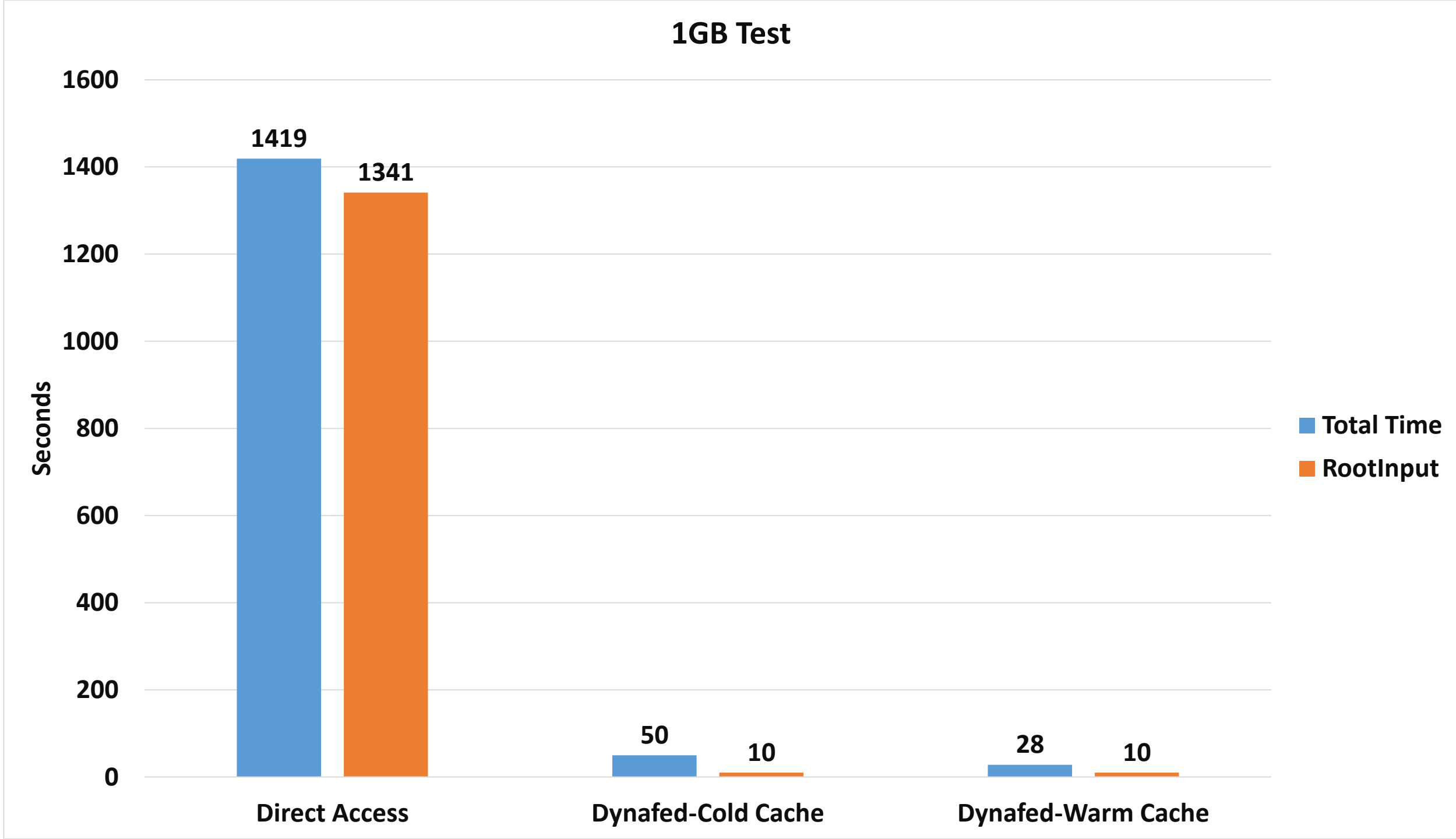
2. Aggregation of a set of Belle II storage endpoints + with the cache endpoint in Napoli. [path /belle-cache-path]

```
Example configuration for the view that include cache
….
locplugin.*.xlatepfx: /belle-cache-path/ /
….
glb.locplugin[]: /usr/lib64/ugr/libugrlocplugin_dav.so CESNET-SE 5 https://dpm1.egee.cesnet.cz:443/dpm/cesnet.cz/home/belle/TMP/belle/MC/merge1/

glb.locplugin[]: /usr/lib64/ugr/libugrlocplugin_dav.so SCOReS-CacheSE 5 https://recas-dpm-01.na.infn.it/dpm/na.infn.it/home/belle/cache/
```

Behaviour: in the example before, Dynafed creates a metalink with two endpoints, even in the file is not yet in the cache.

If the geoip plugin is activate the first endpoint for a client in Napoli will be always the local cache.

INFN
Istituto Nazionale di Fisica Nucleare


1GB Test
1600
1400
1200
1000
800
600
400
200
0
Seconds
1419
1341
50
10
28
10
Direct Access
Dynafed-Cold Cache
Dynafed-Warm Cache
Total Time
RootInput