Data Transfer and Data Access Protocols

Dr. Silvio Pardi (INFN-Napoli)

Belle II Italia 31/05/2016

SE Protocols in HEP

Assets:

- Gridftp, xrootd are very popular protocols for data transfer and data access in the international HEP community.
- SRM is one of the most used Storage Resource Manager.
- Data Federation as demonstrated to be an useful feature for analysis.

New opportunities:

- Http is world-wide most used protocol - native protocol in many storage systems.
- Cloud storage.
- Global Storage able to integrate permanent storage, cloud storage, ephemeral storage, cache system etc

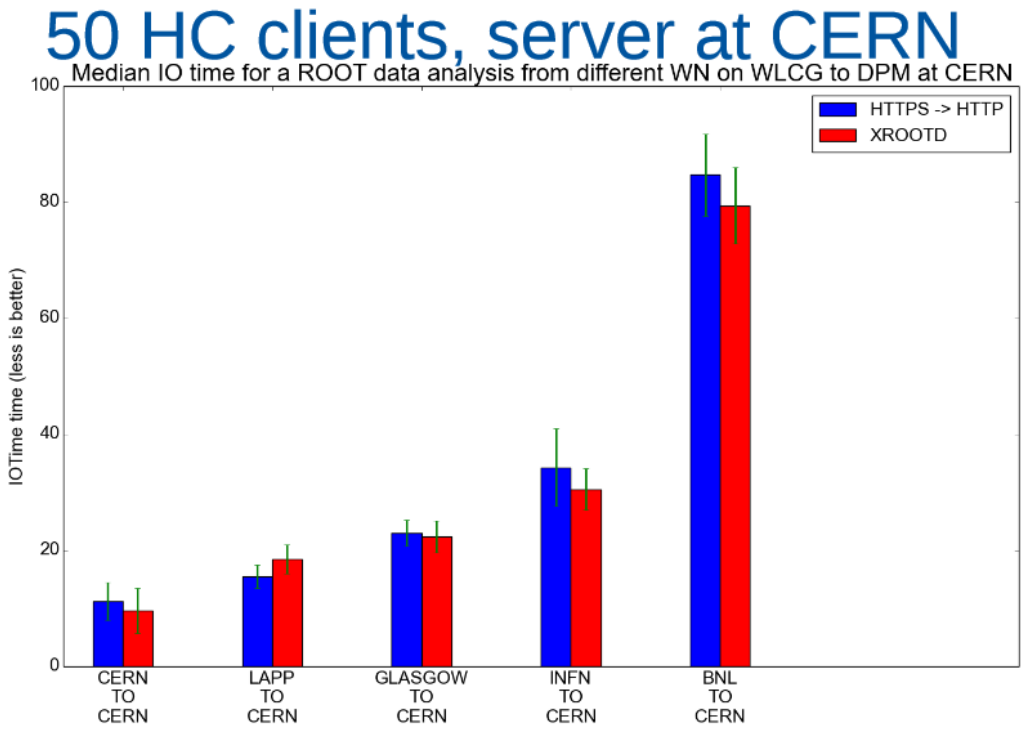
Nice overview on these topics

https://indico.cern.ch/event/433164/contributions/1930256/attachments/1220413/1783888/DM-WS-Lissabon.pdf

Http support for root

Presentation at CHEP 2015: “Protocol benchmarking on DPM - Use cases for HTTP and Xrootd”

http://indico.cern.ch/event/304944/session/3/contribution/188/attachments/578648/796810/KeeblProtocolComparison.pdf



Conclusion of the authors

According to the test cases, HTTP and xroot access to DPM have equivalent performance.

Root and Davix Tutorial

https://dmc.web.cern.ch/projects/davix/root-and-davix-tutorial

HTTP Deployment Task Force

The following have agreed to be involved

- Experiments - Atlas, CMS & LHCb
- Sites/Infrastructures - ASGC, BNL, CERN, GridPP, KIT, PIC & TRIUMF
- Storage systems - dCache, DPM, EOS, StorRM, xrootd
- WLCG monitoring is represented

https://twiki.cern.ch/twiki/bin/view/LCG/HTTPDeployment

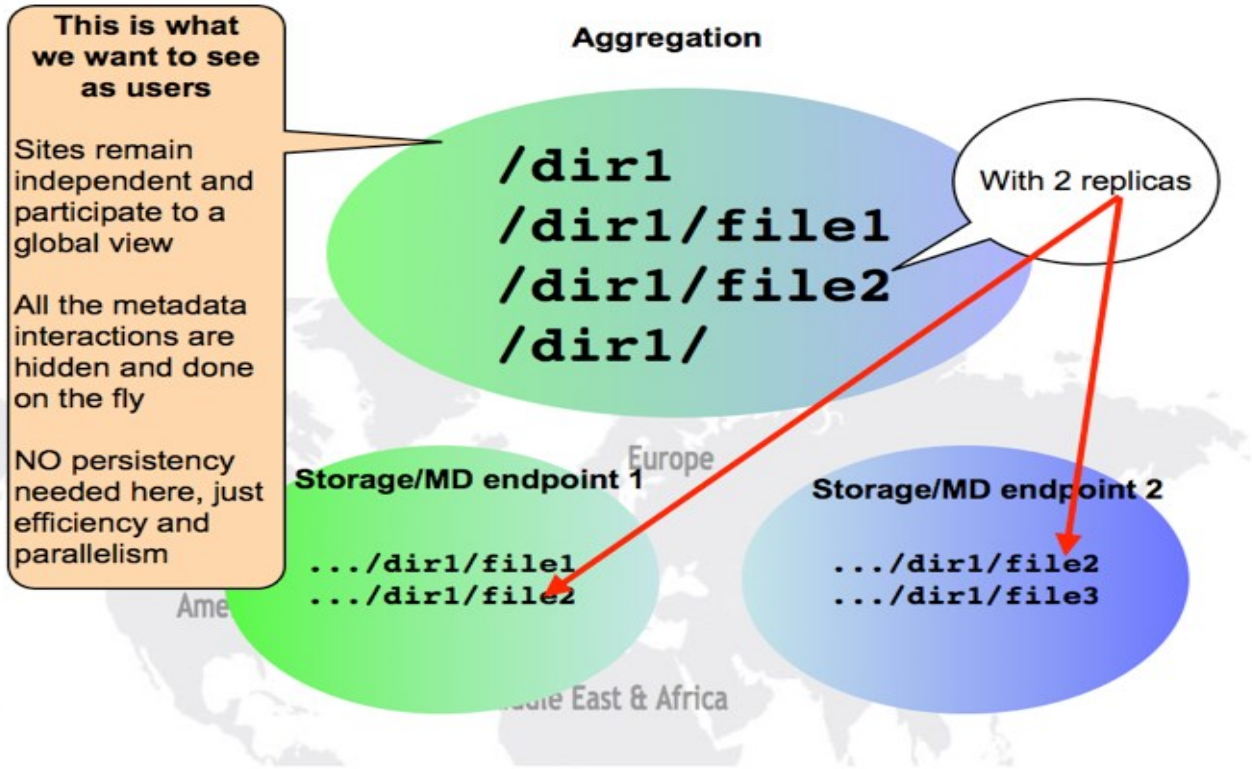
Http task force is now closed (start April 2015, last meeting 23rd Mar 2016)

A set of nagios plugin for the http/webdav has been released and integrated in the SAM monitoring.

Dynamic Federations

The Dynamic Federations system allows to aggregate remote storage. The aggregation is performed on the fly, by distributing quick WebDAV queries towards the endpoints and scheduling, aggregating and caching the responses.

HTTP and WebDAV clients can browse the Dynamic Federation as if it were a unique partially cached name space, which is able to redirect them to the right host when they ask for a file replica.



Dynamic Federations is based on a open source project that allows to aggregate storages, that expose different protocols: HTTP/WebDAV, Cloud S3, moreover the system is generic enough to support others, if a suitable frontend exists.
https://svnweb.cern.ch/trac/lcgdm/wiki/Dynafeds
Currently we have multiple example of aggregated storage on a server at desy
LHCB testbed (14 storages)
http://federation.desy.de/fed/lhcb/
ATLAS testbed (~50 storages)
http://federation.desy.de/fed/browseatlas/

http://federation.desy.de/DynaFeds/The_Dynamic_Federations.html

Goal of the R&D work

- Study and overcome issues relate the usage of HTTP/WEBDAV with the Belle II Software for simulation and analysus Basf2.
- Configure and tune the HTTP/WEBDAV interface on Grid Storages.
- Implement a first example of http federation for Belle II with dynafed software.
- Test federation with dynafed component
- Learn more about this topic.

The work is ongoing with the precious support of several sites of Belle II collaboration, and with the teams of DPM, dCache and SToRM.

Basf2 with Davix support

The Belle II software for simulation and analysis is based on the standard ROOT I/O library

Since the release build-2016-03-04 distributed to each sites via CVMFS, basf2 supports natively http/webdav thanks to the introduction of TDavixFile.h library.

```
……
filelistSIG=['https://belle-dpm-01.na.infn.it/dpm/na.infn.it/home/belle/TMP/test.root']
inputMdstList(filelistSIG)
……
```

ROOT Parameters relate TDavixFile

Variables in system.rootrc

Davix.UseOldClient: no

Verbosity level of the external Davix library

Davix.Debug: 0

Davix.GSI.UserProxy: /my/path/my_proxy

Davix.GSI.UserCert: /my/path/my_cert

Davix.GSI.UserKey: /my/path/my_key

Davix.GSI.CAdir: /etc/grid-security/certificates

Davix.GSI.CACheck: y

Davix.GSI.GridMode: y

Davix.S3.SecretKey: secret

Davix.S3.AccessKey: token

#TTreeCache.Size 1.0 TO ENABLE CACHE

They are already configured properly in externals/vXXXX/Linux_x86_64/opt/root/etc/system.rootrc

The testbed infrastructure

STORGE DIRAC NAME	HOSTNAME	TYPE
DESY-DE	dcache-belle-webdav.desy.de	DCACHE
GRIDKA-SE	f01-075-140-e.gridka.de	DCACHE
NTU-SE	bgrid3.phys.ntu.edu.tw	DCACHE
SIGNET-SE	dcache.ijs.si	DCACHE
UVic-SE	charon01.westgrid.ca	DCACHE
Adelaide-SE	coepp-dpm-01.ersa.edu.au	DPM
CESNET-SE	dpm1.egee.cesnet.cz	DPM
CYFRONNET-SE	dpm.cyf-kr.edu.pl	DPM
Frascati-SE	atlasse.lnf.infn.it	DPM
HEPHY-SE	hephyse.oeaw.ac.at	DPM
Melbourne-SE	b2se.mel.coepp.org.au	DPM
Napoli-SE	belle-dpm-01.na.infn.it	DPM
CNAF-SE	ds-202-11-01.cr.cnaf.infn.it	STORM
McGill-SE	gridftp02.clumeq.mcgill.ca	STORM

In January 2016 we started the investigation to verify http/webdav performances

At now 14 Storages of the 23 SRM endpoints registered in DIRAC have activated the webdav interface.

3 different storages technologies represented **dCache, DPM, STORM**

DYNAFED Server in Napoli
https://dynafed01.na.infn.it/myfed/

/myfed/belle/ — dynafed01.na.infn.it/myfed/belle/

/myfed/belle/

Aggregate view of all the storages

Mode	Links	UID	GID	Size	Modified	Name
-rw-rw-r--	0	0	0	1000.0M	Thu, 05 Mar 2015 15:31:21 GMT	1G
drwxrwxr-x	0	0	0	0	Mon, 16 May 2016 15:32:19 GMT	DATA
drwxrwxrwx	0	0	0	0	Tue, 15 Sep 2015 23:55:07 GMT	DC
drwxrwxr-x	0	0	0	2	Thu, 05 Mar 2015 02:00:06 GMT	DC2014
drwxrwxrwx	0	0	0	0	Sun, 26 Apr 2015 22:04:28 GMT	DC2015
drwxrwxr-x	0	0	0	0	Fri, 07 Aug 2015 15:20:46 GMT	MC
drwxrwxr-x	0	0	0	0	Mon, 16 May 2016 15:32:59 GMT	TMP
-rwxrwxrwx	0	0	0	4	Tue, 03 Mar 2015 08:32:15 GMT	aaa
drwxrwxrwx	0	0	0	0	Wed, 01 Apr 2015 05:31:26 GMT	belle
drwxrwxrwx	0	0	0	0	Fri, 04 Mar 2016 15:18:36 GMT	bellehttpd
drwxrwxr-x	0	0	0	0	Wed, 01 Apr 2015 02:50:43 GMT	generated
drwxrwxr-x	0	0	0	0	Fri, 11 Mar 2016 06:56:55 GMT	group
drwxrwxrwx	0	0	0	0	Mon, 09 Feb 2015 05:44:27 GMT	hoge
drwxrwxrwx	0	0	0	0	Tue, 21 Jul 2015 12:10:19 GMT	monitor
drwxrwxrwx	0	0	0	0	Tue, 17 Feb 2015 07:34:10 GMT	raw
-rwxrwxrwx	0	0	0	5	Fri, 03 Jul 2015 14:31:03 GMT	silviotest30
-rwxrwxrwx	0	0	0	5	Fri, 03 Jul 2015 14:31:50 GMT	silviotest31
-rw-rw-r--	0	0	0	5	Tue, 07 Jul 2015 21:47:38 GMT	test-null-acl01
-rwxrwxrwx	0	0	0	12.3M	Wed, 22 Oct 2014 10:47:46 GMT	test.pippo
-rwxrwxrwx	0	0	0	185	Mon, 12 Apr 2010 07:47:50 GMT	testfile
-rw-rw-r--	0	0	0	19	Mon, 16 May 2016 08:37:03 GMT	testfile1
-rwxrwxrwx	0	0	0	5	Fri, 03 Jul 2015 09:19:48 GMT	testsilvio10
-rwxrwxrwx	0	0	0	5	Fri, 03 Jul 2015 09:25:40 GMT	testsilvio11
-rwxrwxrwx	0	0	0	5	Fri, 03 Jul 2015 14:09:05 GMT	testsilvio14
-rwxrwxrwx	0	0	0	5	Fri, 03 Jul 2015 14:10:29 GMT	testsilvio17
-rwxrwxrwx	0	0	0	5	Fri, 03 Jul 2015 09:11:35 GMT	testsilvio3
-rwxrwxrwx	0	0	0	5	Fri, 03 Jul 2015 09:15:40 GMT	testsilvio6

Request by nobody (nobody)
Powered by LCGDM-DAV 0.17.0 (New UI)

/myfed/PerSite/ — https://dynafed01.na.infn.it/myfed/PerSite/

/myfed/PerSite/

Mode	Links	UID	GID	Size	Modified	Name
drwxrwxrwx	0	0	0	0	Thu, 01 Jan 1970 00:00:00 GMT	Adelaide-SE
drwxrwxrwx	0	0	0	0	Thu, 01 Jan 1970 00:00:00 GMT	CESNET-SE
drwxrwxrwx	0	0	0	0	Thu, 01 Jan 1970 00:00:00 GMT	CNAF-SE
drwxrwxrwx	0	0	0	0	Thu, 01 Jan 1970 00:00:00 GMT	CYFRONET-SE
drwxrwxrwx	0	0	0	0	Thu, 01 Jan 1970 00:00:00 GMT	DESY-SE
drwxrwxrwx	0	0	0	0	Thu, 01 Jan 1970 00:00:00 GMT	Frascati-SE
drwxrwxrwx	0	0	0	0	Thu, 01 Jan 1970 00:00:00 GMT	GRIDKA-SE
drwxrwxrwx	0	0	0	0	Thu, 01 Jan 1970 00:00:00 GMT	HEPHY-SE
drwxrwxrwx	0	0	0	0	Thu, 01 Jan 1970 00:00:00 GMT	McGill-SE
drwxrwxrwx	0	0	0	0	Thu, 01 Jan 1970 00:00:00 GMT	Melbourne-SE
drwxrwxrwx	0	0	0	0	Thu, 01 Jan 1970 00:00:00 GMT	Napoli-SE
drwxrwxrwx	0	0	0	0	Thu, 01 Jan 1970 00:00:00 GMT	SIGNET-SE
drwxrwxrwx	0	0	0	0	Thu, 01 Jan 1970 00:00:00 GMT	UVic-SE

PerSiteView that shows the file system of each storage

Request by nobody (nobody)
Powered by LCGDM-DAV 0.17.0 (New UI)

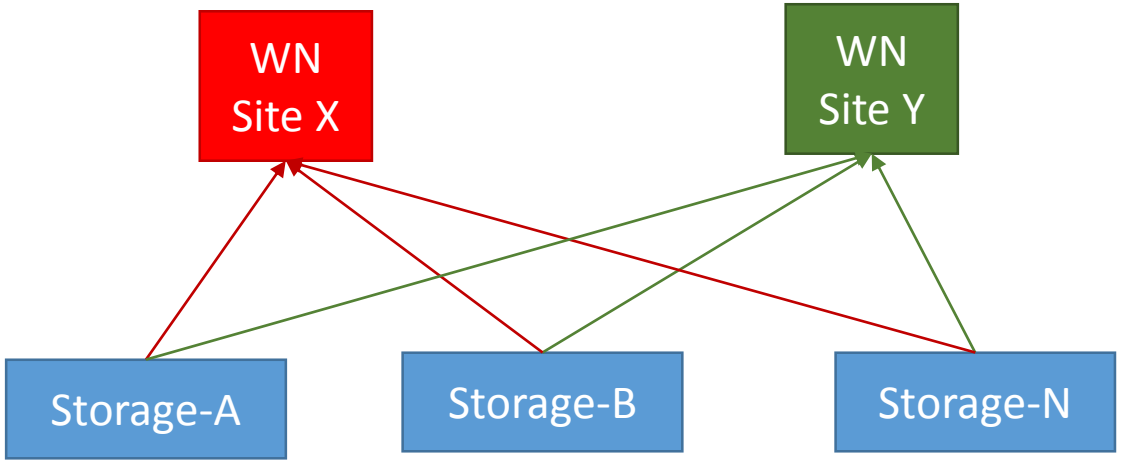
Performance Tests

Software used: bsaf2

Skim used: D+rho_skim.py

6 different tests done analysing 1 single file (1000 events, size 10MB)

- Local input (download via http)
- Local input (download via xrootd)
- Local input (download via lcg-cp)
- Input via http streaming
- Input via xrootd streaming
- Dynafed access



5 trials for each test

Test are performed from Grid WorkerNodes in different sites vs All the storage of type DPM and dCache, while storage running STORM are still in evaluation.

Performance Tests

For each test I collect information about:

Total time: Average, Max, Min, StdDev

RootInput time: Average, Max, Min, StdDev

Download Time: Average, Max, Min, StdDev

Latency from the client

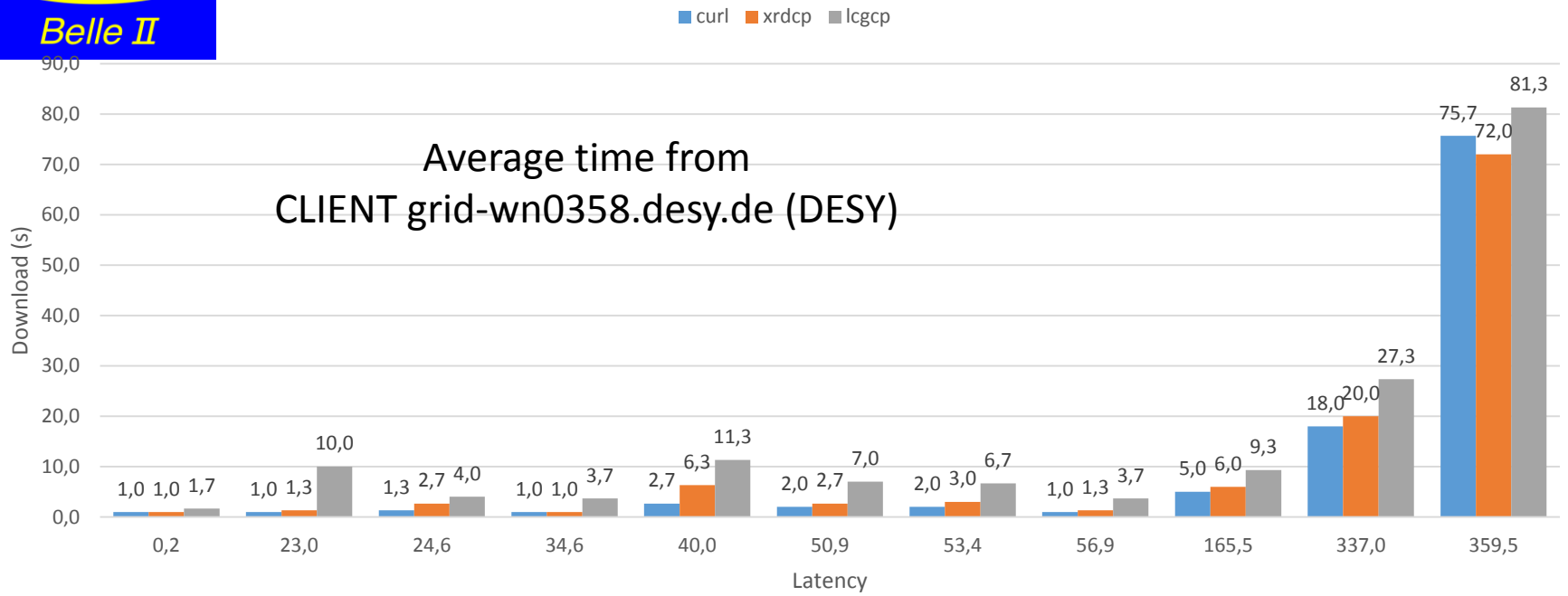
Network monitoring information through dstat software.

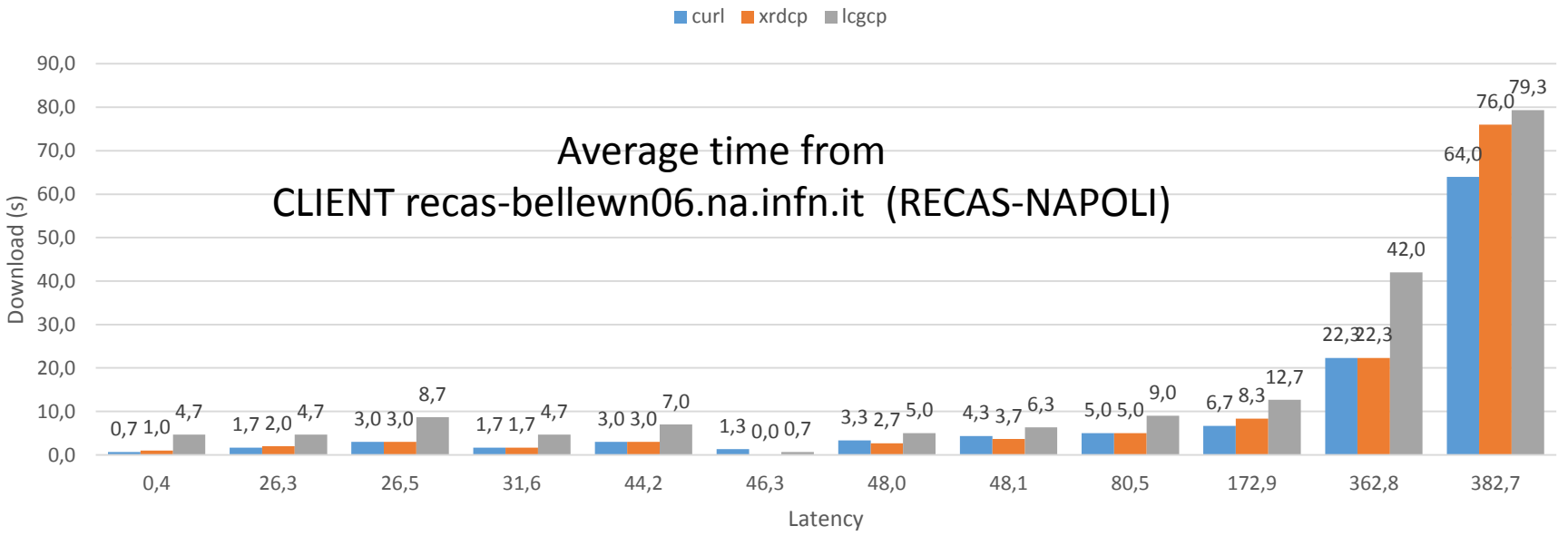
A LOT OF NUMBERS!

	AC	AD	AE	AF	AG	AH	AI	AJ	AK	AL	AM	AN	AO	AP	AQ	AR	AS	AT	AU	AV	AW	AX	AY	AZ	BA	BB	BC	BD	BE
1	min	ROOT-a	dev	max	min	4-xroot_	dev	max	min	xrdcp	dev	max	min	ROOT-a	dev	max	min	5-srm_c	dev	max	min	lcgcp-av	dev	max	min	ROOT-a	dev	max	min
2	0	0,39	0	0,39	0,39	13,667	0,7483	15	13	1	0	1	1	0,3833	0,0049	0,39	0,38	17,333	0,8	19	17	4,6667	0,4899	5	4	0,3867	0,0075	0,4	0,38
3	1	0,39	0	0,39	0,39	14,667	0,4899	15	14	2	0,4	2	1	0,39	0,004	0,4	0,39	18	0,8	18	16	4,6667	0,7483	6	4	0,39	0,004	0,39	0,38
4	2	0,39	0	0,39	0,39	15,667	0,4899	16	15	3	0,4	3	2	0,39	0,004	0,39	0,38	21,667	0,8	22	20	8,6667	1,0198	10	7	0,39	0	0,39	0,39
5	1	0,39	0	0,39	0,39	14,333	0,4899	15	14	1,6667	0,4899	2	1	0,39	0,004	0,4	0,39	17,667	1,3565	19	16	4,6667	1,6	7	3	0,39	0	0,39	0,39
6	2	0,39	0	0,39	0,39	16	0,4	17	16	3	0	3	3	0,3867	0,0049	0,39	0,38	19,667	0,9798	21	19	7	1,1662	8	5	0,3867	0,0075	0,4	0,38
7	1	0,39	0	0,39	0,39	6	0	6	6	0	0	0	0	-333,3	0	0	1000	6	0,4	7	6	0,6667	0,4899	1	0	-333,3	0	0	1000
8	3	0,39	0	0,39	0,39	15,667	0,4899	16	15	2,6667	0,4899	3	2	0,39	0,004	0,39	0,38	18	0	18	18	5	0	5	5	0,39	0,004	0,39	0,38
9	3	0,39	0	0,39	0,39	16,667	0,7483	18	16	3,6667	0,7483	5	3	0,39	0,004	0,39	0,38	19,667	0,4899	20	19	6,3333	0,4899	7	6	0,3833	0,0049	0,39	0,38
10	4	0,39	0,004	0,39	0,38	17,667	0,7483	19	17	5	0,4	5	4	0,39	0	0,39	0,39	22	0	22	22	9	0,4	9	8	0,39	0,004	0,39	0,38
11	6	0,39	0,004	0,4	0,39	21,667	2,4166	27	20	8,3333	2,3324	14	8	0,3867	0,0049	0,39	0,38	25,333	0,9798	26	24	12,667	0,4899	13	12	0,39	0,004	0,39	0,38
12	0	-333,3	0	0	1000	30	13,437	63	28	16,667	13,023	49	16	0,39	0,004	0,39	0,38	36,667	25,111	99	35	23	25,303	86	22	0,39	0	0,39	0,39
13	20	0,3867	0,0049	0,39	0,38	35,667	1,0955	38	35	22,333	1,4967	25	21	0,39	0	0,39	0,39	48,333	10,412	65	35	42	10,19	58	29	-333,3	0	0	1000
14	23	0,39	0,004	0,39	0,38	89,333	19,886	110	55	76	19,467	96	42	0,3867	0,0049	0,39	0,38	93	8,5884	105	82	79,333	8,1388	91	69	0,39	0,004	0,39	0,38
15																													

McGill | Desy | Napoli

Analysis 1. File download with Http, Xrootd and Lcg-cp





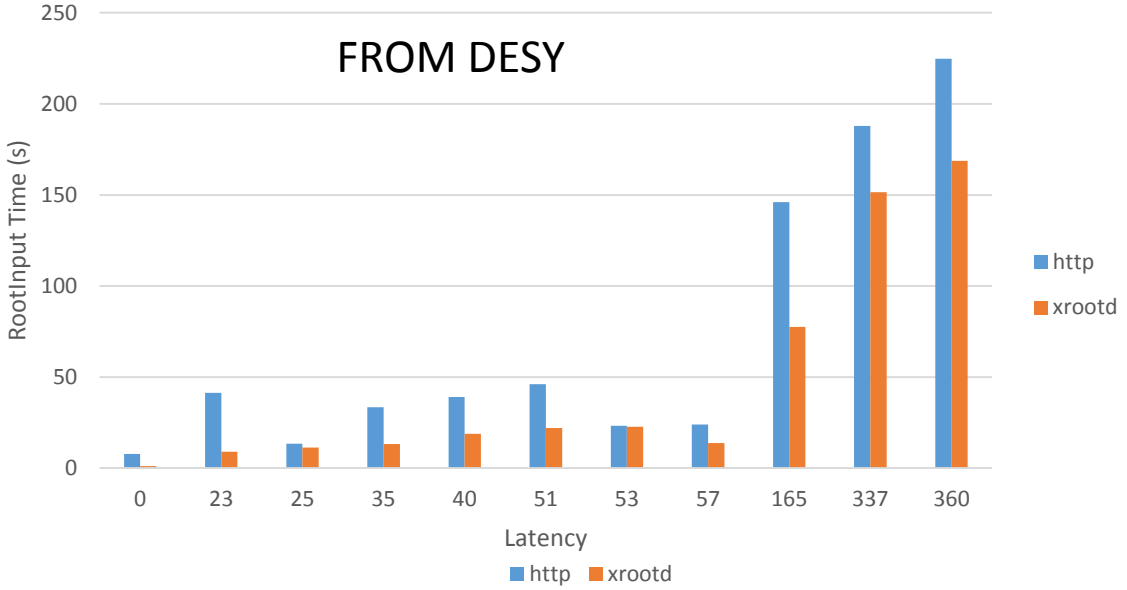
Description

File download performances in function of the latency from the two different Sites.

Comments

http, xrootd performs quite similar and in all cases performs better than lcg-cp through srm interface

Analysis 2. Http vs Xrootd Streaming

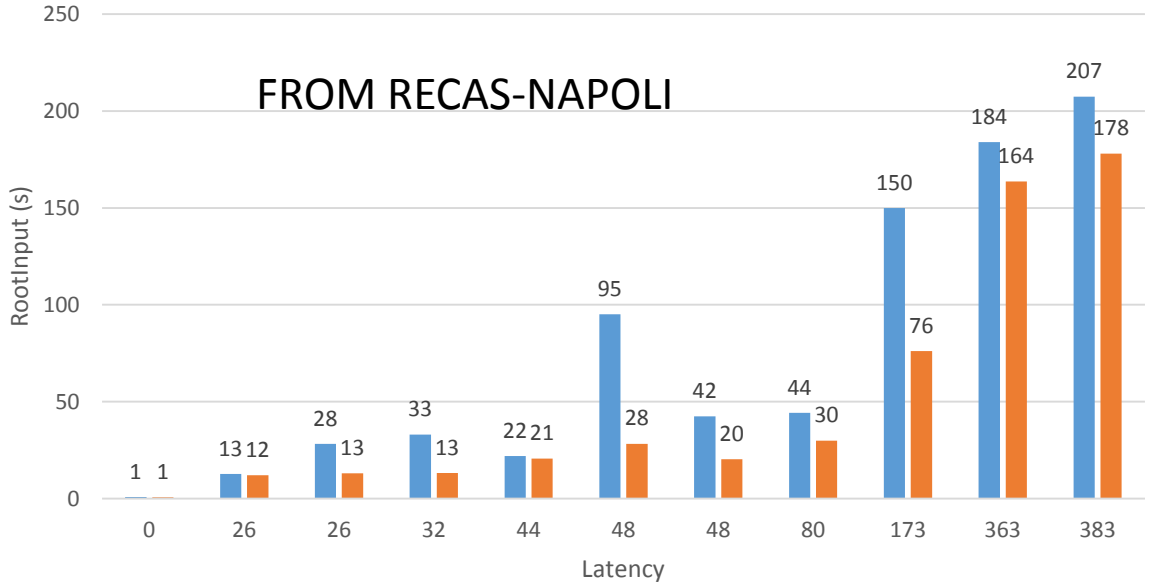


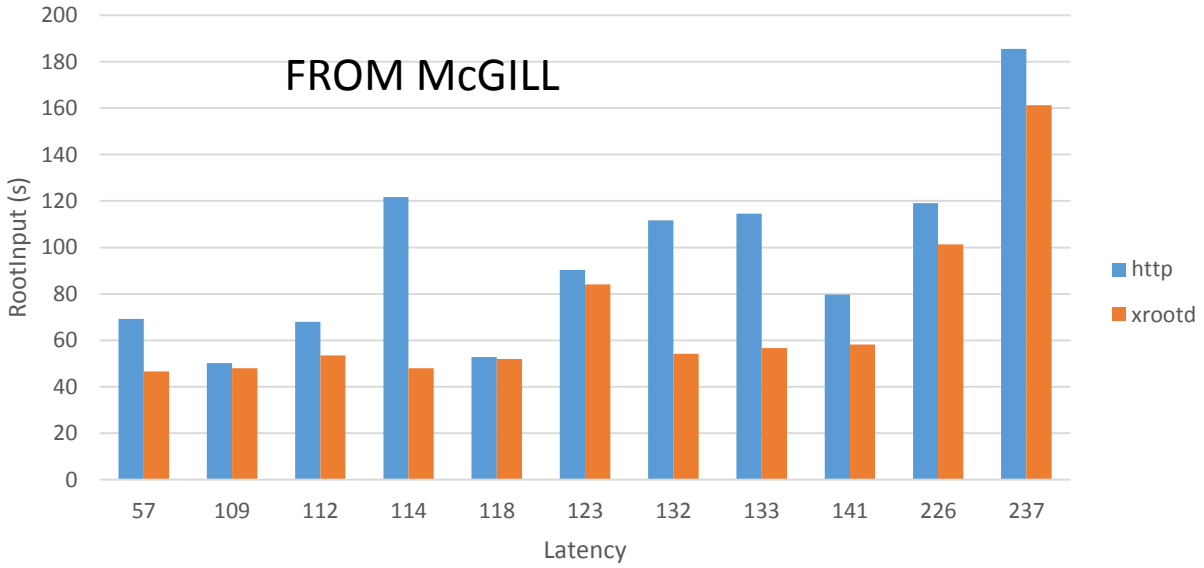
Description

File streaming performances in function of the latency from 3 different Sites. (RootInput function)

Comments

http, xrootd difference goes from 0% to 50%. The behaviour is affected not just from the latency but also by the storage configuration.





Analysis 2. Http vs Xrootd Streaming

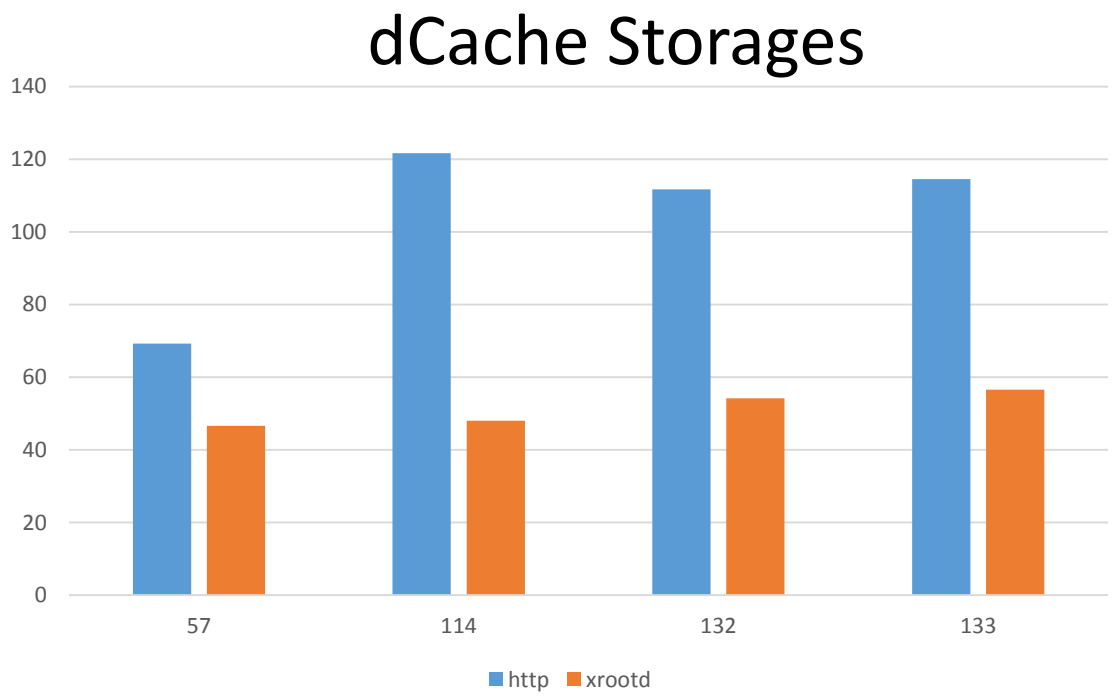
Description

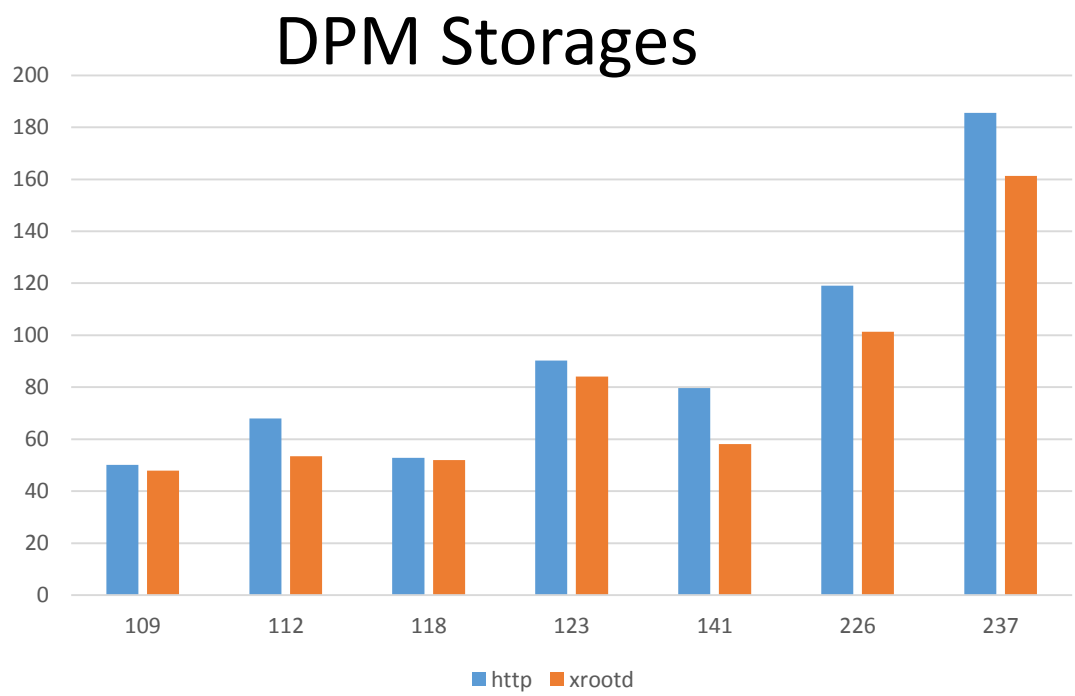
File streaming performances in function of technologies and latency (McGILL case study)

Comments

In case of dCache Storages, http, xrootd differ of about 50% in most cases.

In case of DPM Storages the two protocols performs quite similar in most cases.





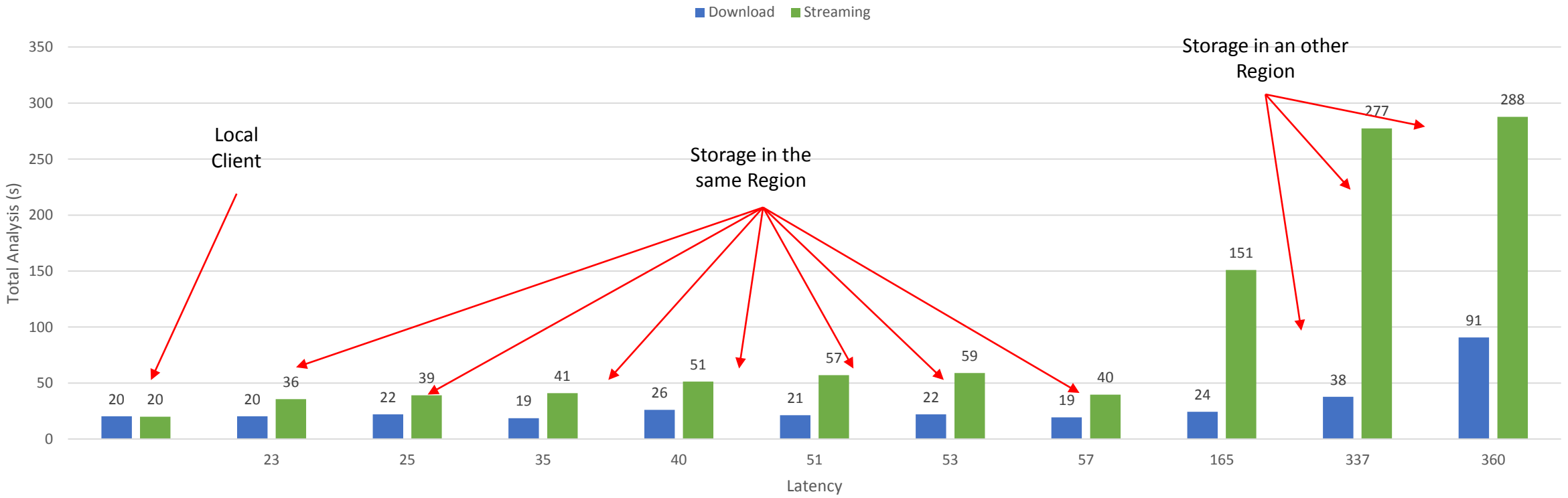
Analysis 3. Download vs Streaming

Description

Total time needed to complete the Job analysis in case of streaming and in case of local download of the file, in function of latency (Test from Desy)

Observation

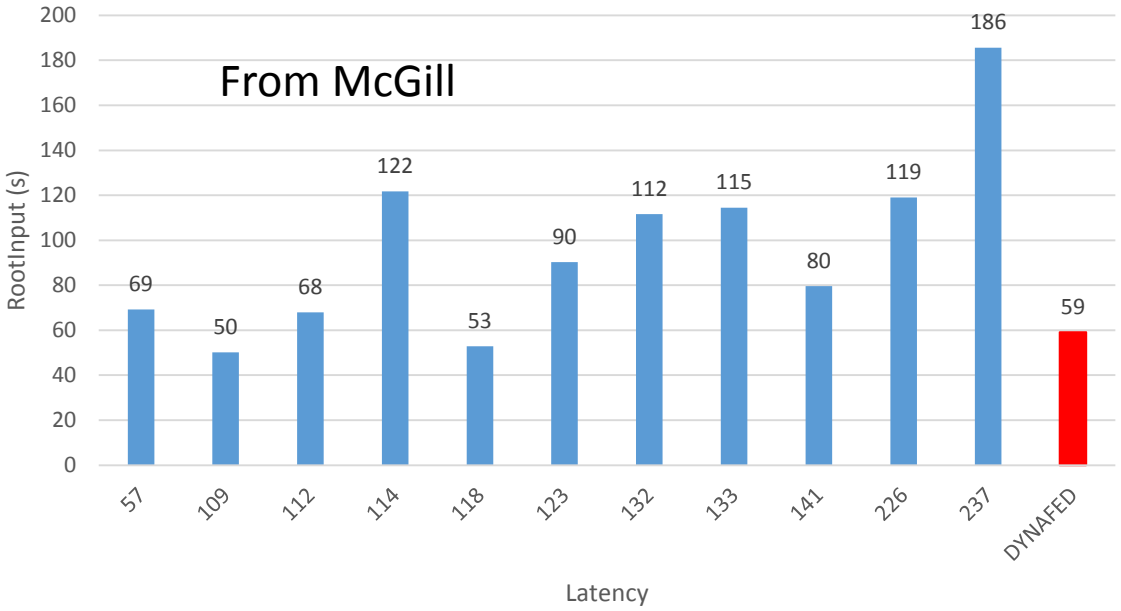
When the client is local to the storage the streaming of a single file perform like the download. In case of same Geographic Region the impact of streaming is around 50%

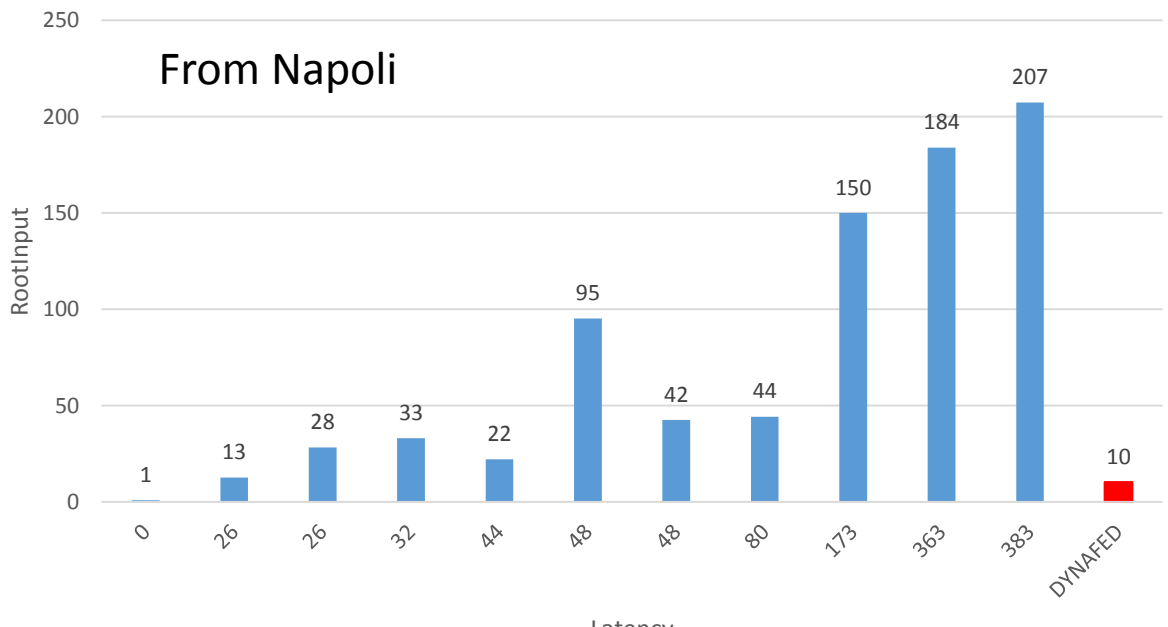


Analysis 4. Dynafed performances

Thanks to the aggregation feature provides by Dynafed, we can address a specific file and his replicas with a single url:
http://dynafed01.na.infn.it/myfed/belle/TMP/belle/user/spardi/testhttp/mixed_e0001r0009_s00_BGx1.mdst.root
For the testbed 50 files has been copied in the 14 storages part of Dynafed. It can be used also for future tests.

Comments
Dynafed thanks to the geoip-plugin is able to chose a convenient replica for the client, however it seems that not discriminate storages in the same country.





Storage Tuning

- **DPM Tuning:** Easy to configure. Best practice: NSSecure should be switched off, and the plain http port should be opened on the firewall of disk-nodes.
- **dCache Tuning:** We opened two ticket related the tuning of dCache for http streaming. [www.dcache.org #8932], [www.dcache.org #8936] Currently they involved several Belle II sites using dCache.
- **SToRM Tuning:** Currently there is an issue relate the http streaming, More specifically the error “*** Break *** segmentation violation” occur after the read of 100events. It seems a systematic limit, at now SToRM team at CNAF is working on it.

Demo at WLCG Meeting (8-9 October 2016 S.Francisco)

WLCG storage, Cloud resources and volatile storage into HTTP/WebDAV-based regional federations

F.Furano, R.Sobie, S.Pardi, A.De Salvo, O.Keeble

The goal of this demonstrator is to evaluate regional federations of stable or volatile storage services that can be seen as a unique read/write multi-tier entity, using HTTP protocols for HEP applications.

The endorsement of the computing coordinators of involved experiments is needed to present the Demo.

Other Ideas for investigation

Investigate the implementation a Cache-system in addition to Dynafed. Issue: Squid does not work with SSL (Man in the middle) Varnish seems able to do that.

Extend dynafed to cloud storage.

New plugin for data choice in dynafed?

Volunteer researcher can use http for data access working with their files in order to found issues. Stefano Lacaprara from Padova (Italy) has show some interest.

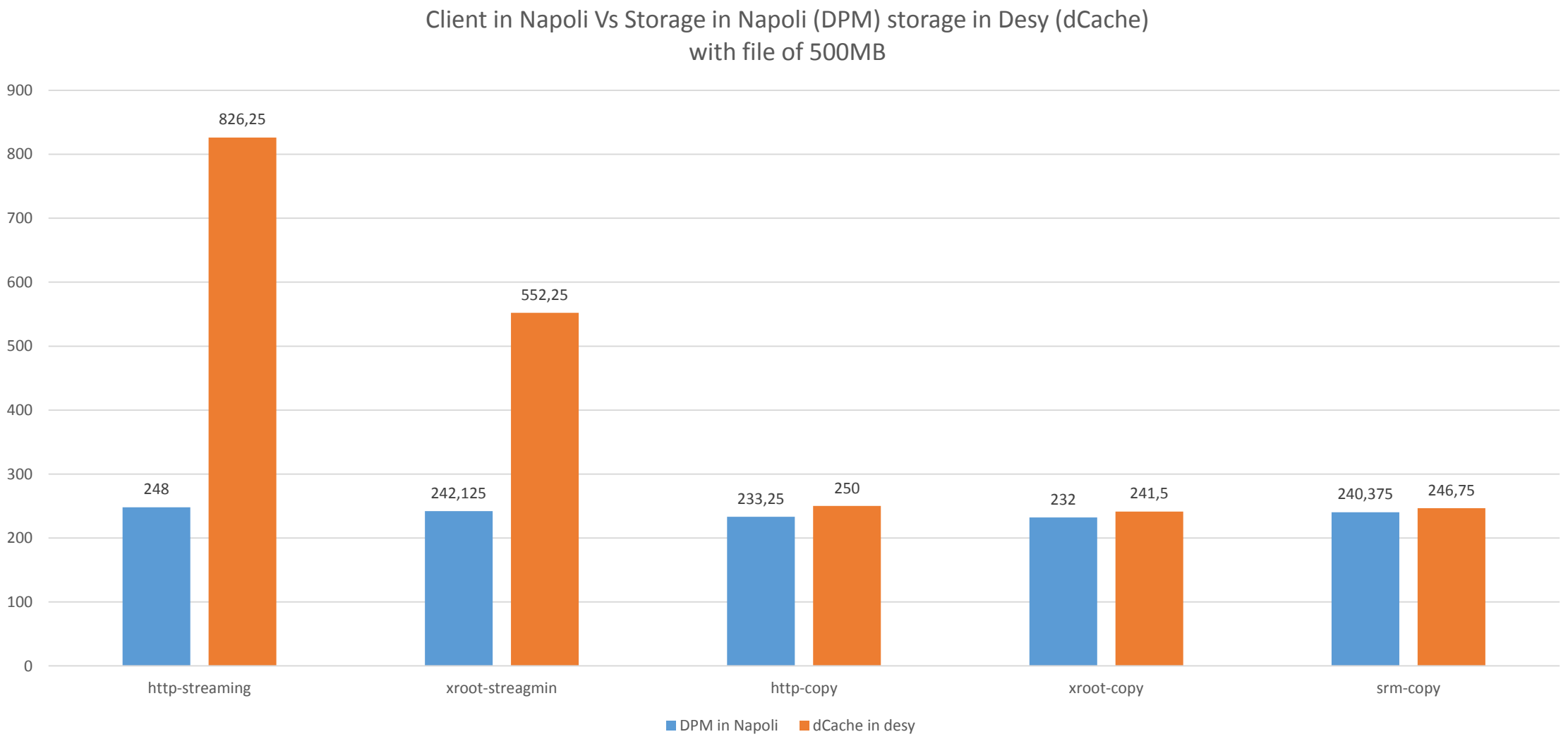
Conclusions (1/2)

- Since build 2016-03-04 Basf2 software is fully compliant with http url.
- 14 Belle II storages currently support properly http/webdav, others are coming.
- A first example of dynafed server has been setup in Napoli other services could be created in other Sites using the same configuration file.

Conclusion (2/2)

- In case of file download http/Webdav access has shown to have the same performances of xrootd and perform better that standard lcg-cp
- Http streaming can perform quite well, with similar performances to respect xrootd, but storage tuning in needed.
- Comparison among streaming and copy-and-analysis strategy has shown a drop of about 50% of performances when the client works in the same Region of the server. Caching system would mitigate this behaviors.
- Dyafed federation has demonstrate to be able to manage Global Storage, and to redirect the client to a convenient copy. Probably more test have to be done to better predict his behaviors.
- More test with multiple file will complete the snapshot.

BACKUP


Client in Napoli Vs Storage in Napoli (DPM) storage in Desy (dCache)
with file of 500MB
900
800
700
600
500
400
300
200
100
0
826,25
552,25
248
242,125
233,25
250
232
241,5
240,375
246,75
http-streaming
xroot-streagmin
http-copy
xroot-copy
srm-copy
DPM in Napoli
dCache in desy