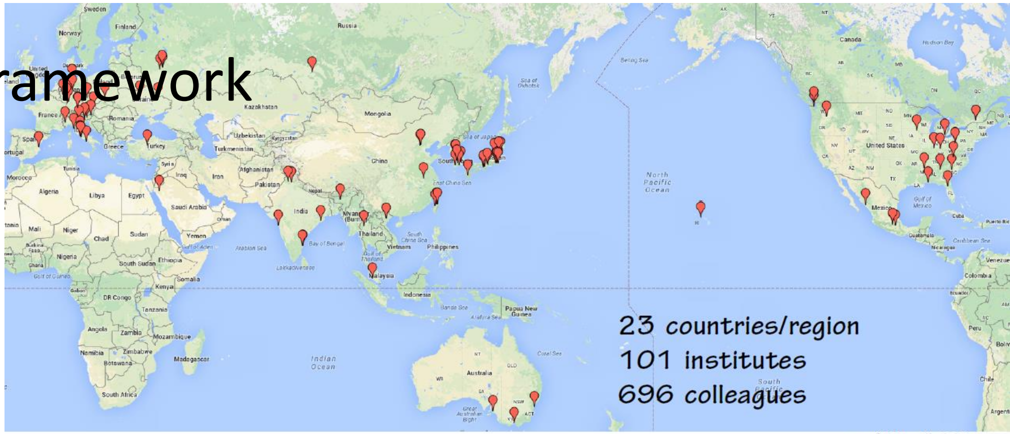
DPM in Belle II sites

Dr. Silvio Pardi
on behalf of the Belle II Computing Group

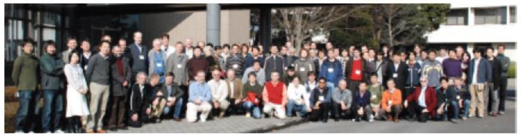
DPM Workshop 2016
LPNHE Paris
23/11/2016

The Belle II Experiment

- Main Site at KEK, Tsukuba – Japan
- Distributed Computing System Based on existing, well-proven solutions plus extensions
- VO name: belle
- DIRAC framework
- LFC for file catalogue
- AMGA for metadata
- Basf2, Simulation and Analysis framework
- Gbasf2, Grid Interface to Basf2
- FTS3 for data movement
- CVMFS for software distribution
- Grid and non Gird resources (ssh and CLOUD)



23 countries/region
101 institutes
696 colleagues

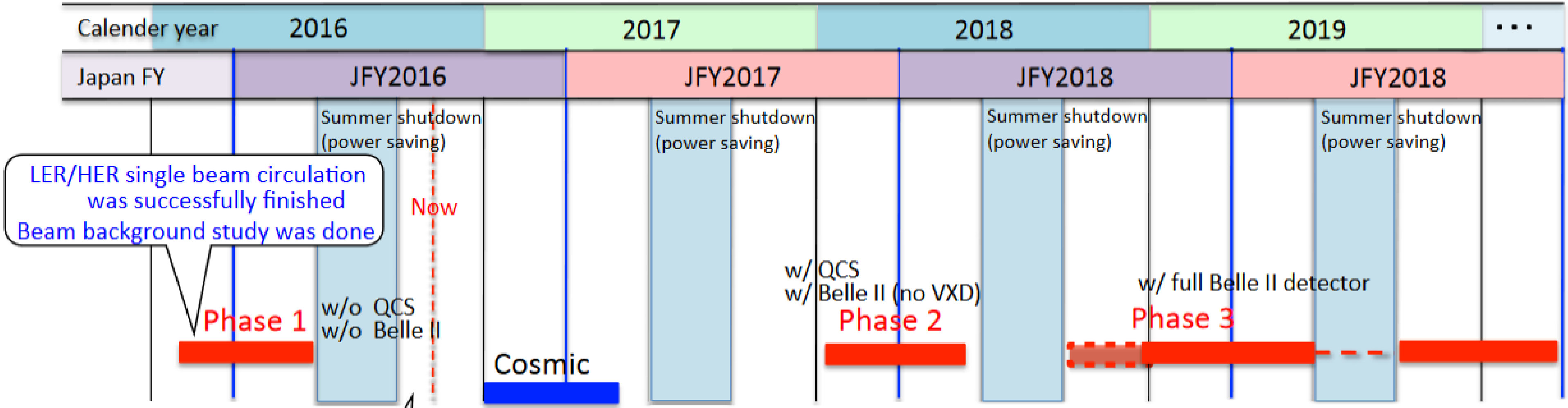


as of Nov. 7, 2016

Asia : ~45%	N. America : ~15%	Europe : ~40%
Japan : 159	US : 80	Germany : 99
Korea : 37	Canada : 20	Italy : 71
Taiwan : 20	Mexico : 10	Russia : 44
India : 34		Slovenia : 17
China : 26		Austria : 13
Australia : 31		Poland : 10
		Czech rep. : 6

others : < 6 colleagues / country

Belle II Experiment: Time line


Calender year
2016
2017
2018
2019
• • •
Japan FY
JFY2016
JFY2017
JFY2018
JFY2018
Summer shutdown (power saving)
Summer shutdown (power saving)
Summer shutdown (power saving)
Summer shutdown (power saving)
Now
Phase 1
w/o QCS
w/o Belle II
Cosmic
w/ QCS
w/ Belle II (no VXD)
Phase 2
w/ full Belle II detector
Phase 3

LER/HER single beam circulation was successfully finished
Beam background study was done

CDC installation was completed
Now cabling is on-going

TOP installation was completed

In 2017

Dress rehearsal is planned
(concurrent running of different type process)

→ Raw data processing
→ MC production
→ User analysis

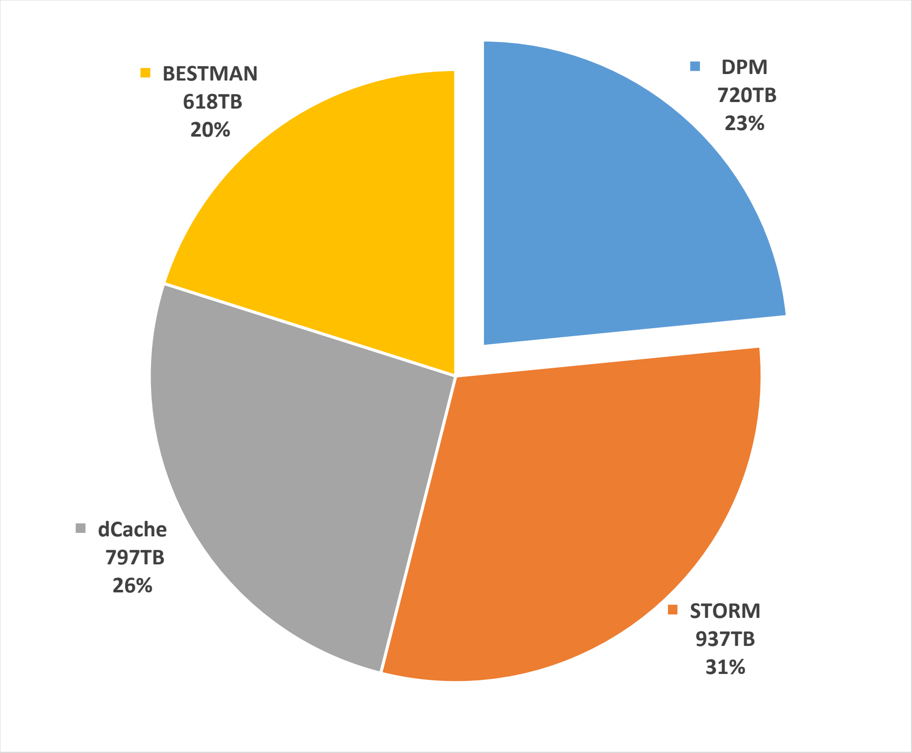
Fast Calibration is necessary

Current Storage Elements

24 Storage currently working

Backend Type:

- **10 DPM**
- 6 dCache
- 7 StoRM
- 1 bestman2



Reserved disk space for BELLE: **3.0 PB** of which **720 TB** managed with DPM 23% - (638TB in 2015)

Requirement for Storage Element

For each SE we require:

- The presence of BELLE Space Token (used to check the disk capacity assigned to the VO, and the current usage)

- The presence of the following directory structure and ACL settings (used to protect data from a misusage of native tools)

    - [root dir]/ (Role=Null R-X, Role=production/lcgadmin RWX)
    - [root dir]/DATA (Role=Null R-X, Role=production/lcgadmin RWX)
    - [root dir]/TMP (Role=Null RWX, Role=production/lcgadmin RWX)

Not all the SRM technologies offer the same features :DPM, SToRM and bestman2 allow to implement all the required ACL.

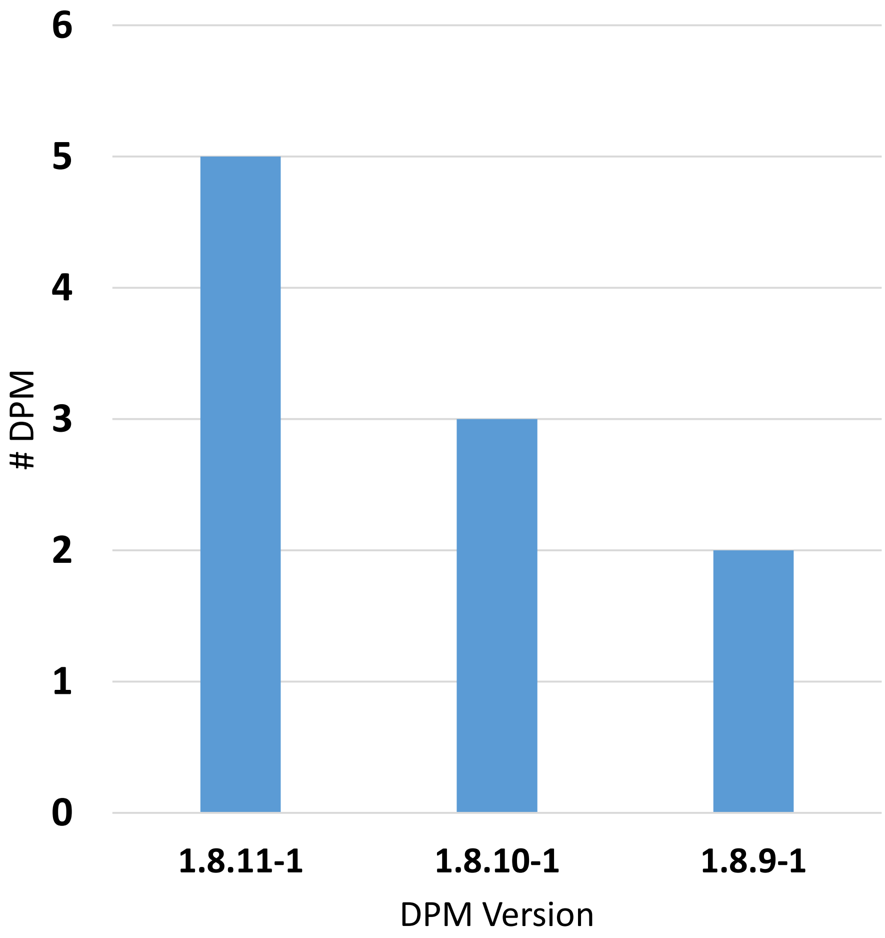
dCache based SEs seem not support the implementation of the full ACL rules required.

DPM Storage in Belle II

SITE	HEAD NODE	COUNTRY
Melbourne-SE	b2se.mel.coepp.org.au	AUSTRALIA
Adelaide-SE	coepp-dpm-01.ersa.edu.au	AUSTRALIA
HEPHY-SE	hephyse.oeaw.ac.at	AUSTRIA
CESNET-SE	dpm1.egee.cesnet.cz	CZECH REPUBLIC
KISTI-SE	belle-se-head.sdfarm.kr	SOUTH KOREA
Frascati-SE	atlasse.lnf.infn.it	ITALY
Napoli-SE	belle-dpm-01.na.infn.it	ITALY
CYFRONET-SE	dpm.cyf-kr.edu.pl	POLAND
ULAKBIM-SE	torik1.ulakbim.gov.tr	TURKEY
IPHC-SE	sbgse1.in2p3.fr	FRANCE
MEX-SE	Under Implementation	MEXICO

DPM Survey 2016

- Xrootd supported by 9 SEs
- HTTP supported by all SEs
- rfio everywhere
- DPM version:
  - 5 sites v.1.8.11-1
  - 3 sites v.1.8.10-1
  - 2 sites v.1.8.9-1



7th MC Campaign

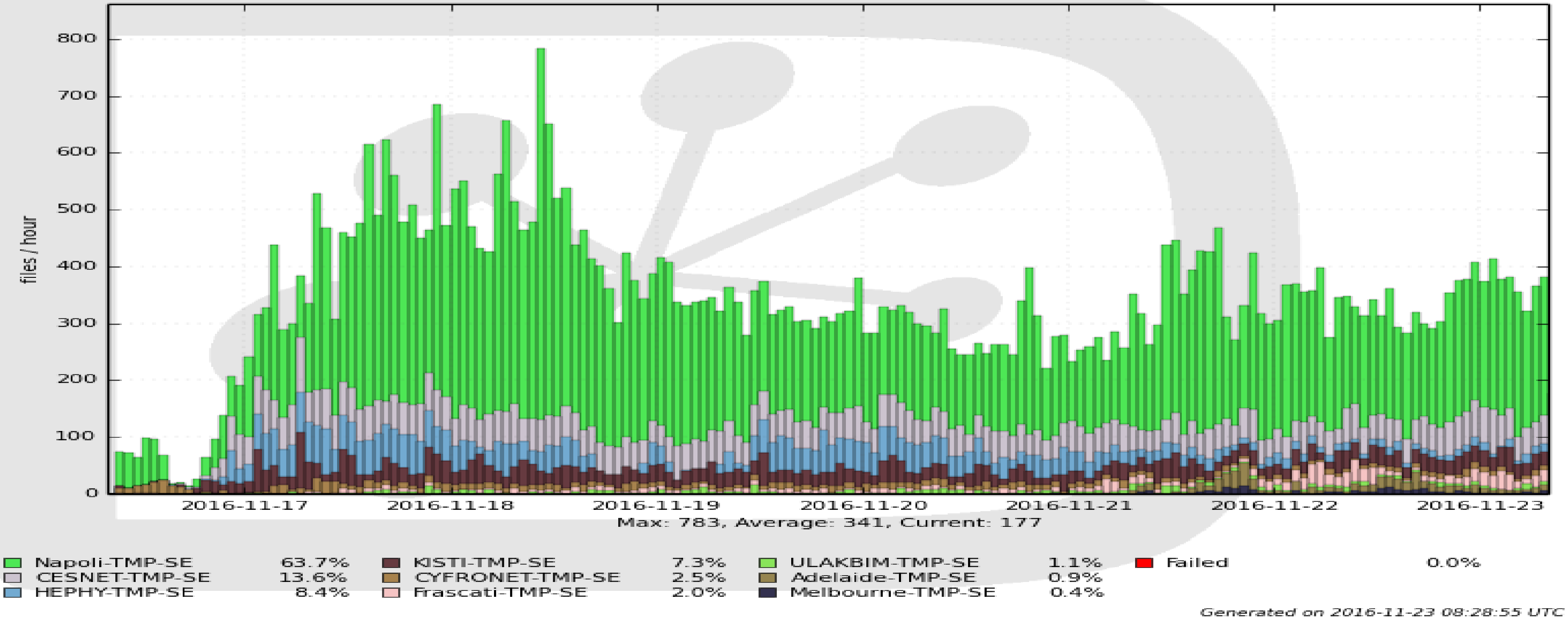
Started 1 November 2016

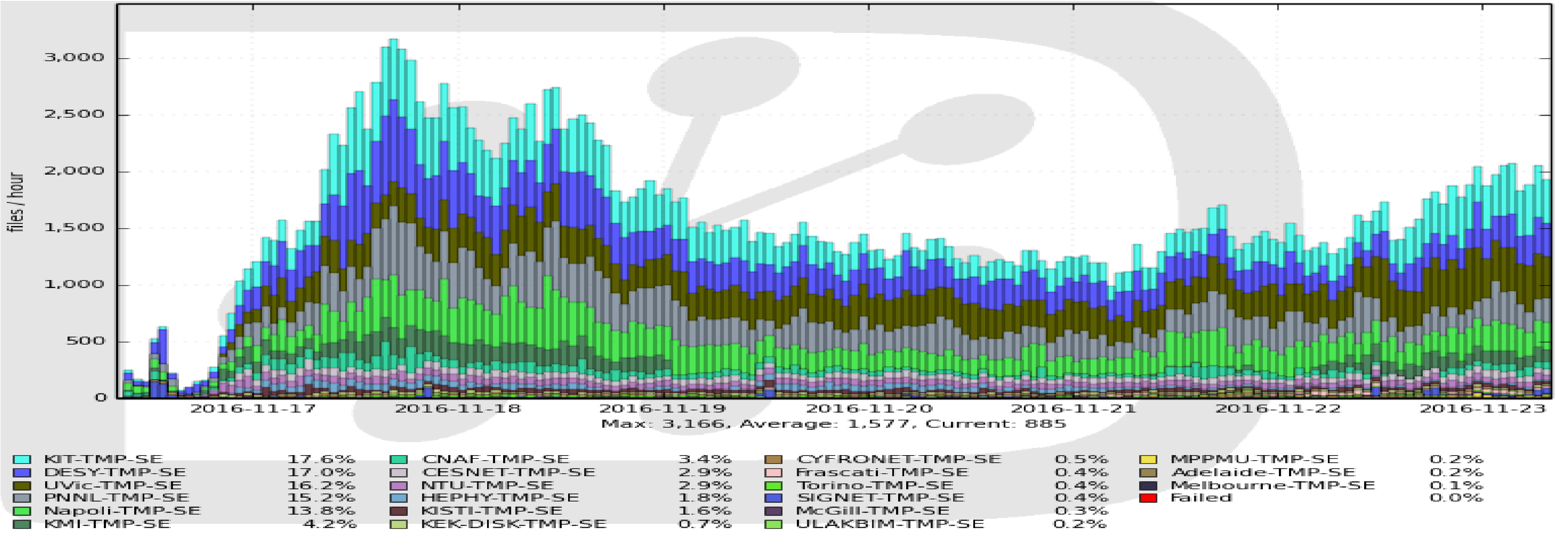
Production goals:

- Preparation for Phase II physics analysis
- Assessment of beam background impact on long term physics analysis
- Ongoing assessment of resource requirements

Suceeded Transfers by Destination

7 Days from 2016-11-16 to 2016-11-23





Http and Dynafed Server for Belle II

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
ULAKBIM-SE	torik1.ulakbim.gov.tr	DPM
CNAF-SE	ds-202-11-01.cr.cnaf.infn.it	STORM
McGill-SE	gridftp02.clumeq.mcgill.ca	STORM
ROMA3-SE	storm-01.roma3.infn.it	STORM

Dynafed server in Napoli in place since January 2016

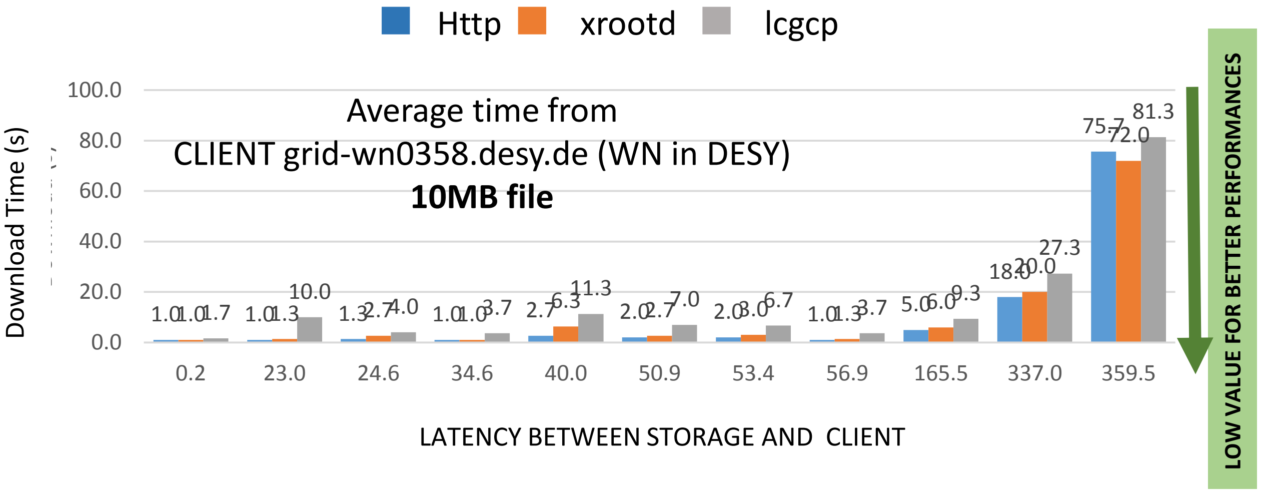
Testbed included 16 of the 24 SRM endpoints currently in production and registered in the DIRAC server.

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

In addition we included an **S3** Amazon Free storage

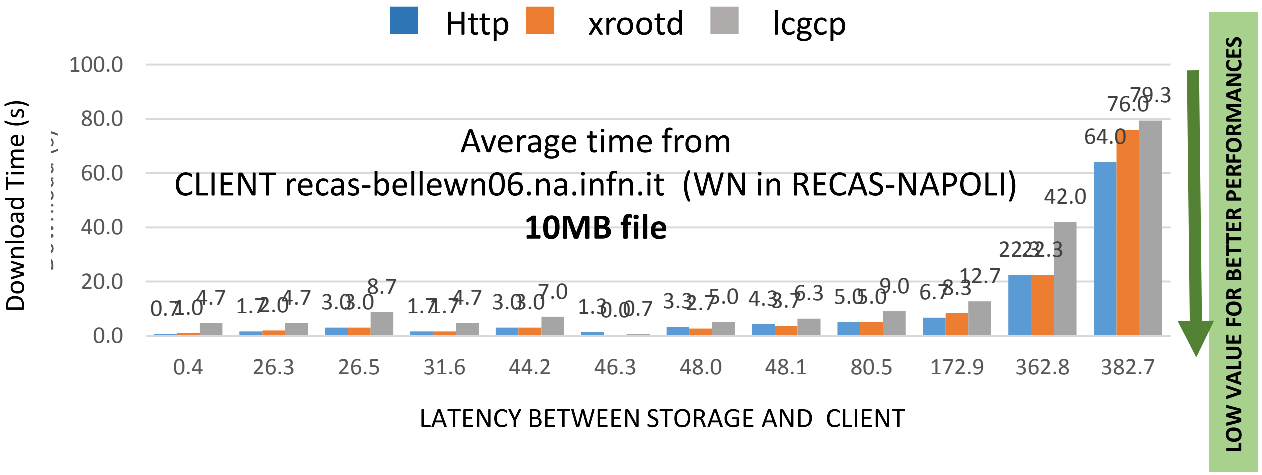
https://dynafed01.na.infn.it/myfed/

File Transfer protocols: Download



Average time from
CLIENT grid-wn0358.desy.de (WN in DESY)
10MB file

LOW VALUE FOR BETTER PERFORMANCES



Average time from
CLIENT recas-bellewn06.na.infn.it (WN in RECAS-NAPOLI)
10MB file

LOW VALUE FOR BETTER PERFORMANCES

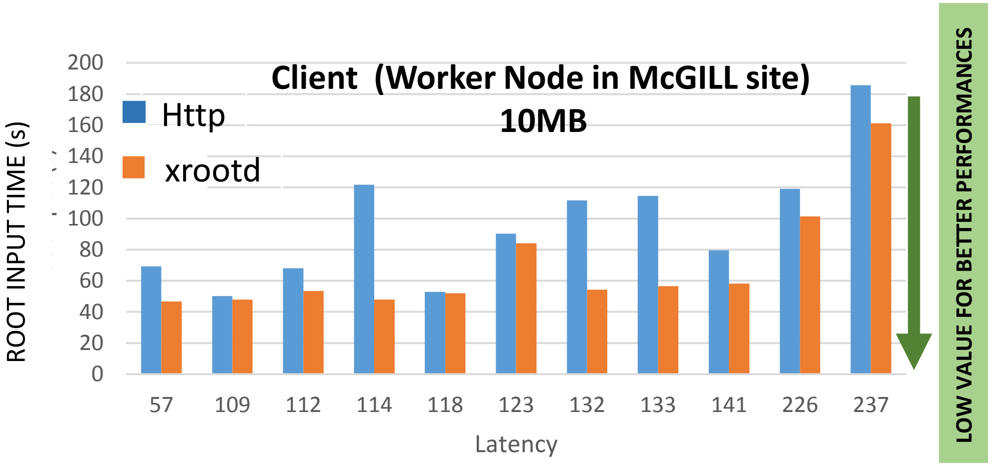
NB. We started with download to test transfers with different protocols under controlled circumstances

Description
File download performances in function of the latency from the two different Sites.
(Performance tuning with HTTP)

Test Analysis
http, xrootd performs quite similar in the case of file download.
Graphs show the overhead added by the SRM interface using lcg-cp command with gridftp.

Need to test with FTS "transfers"

File Read protocols: streaming with HTTP vs xrootd



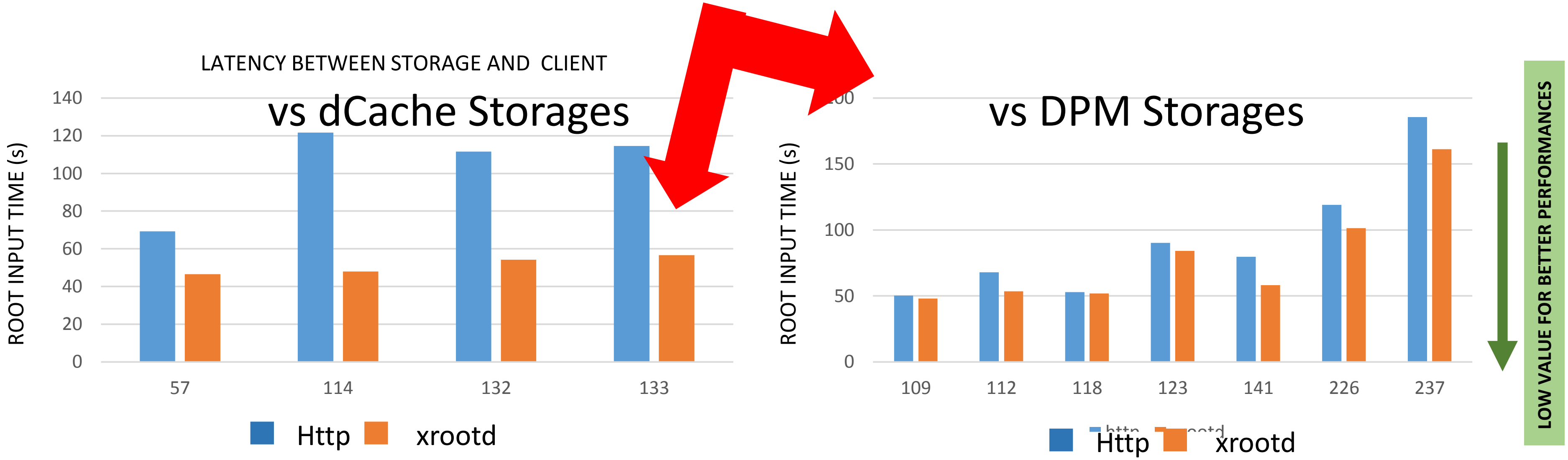
Description

File streaming performances, using a basf2 analysis job

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.



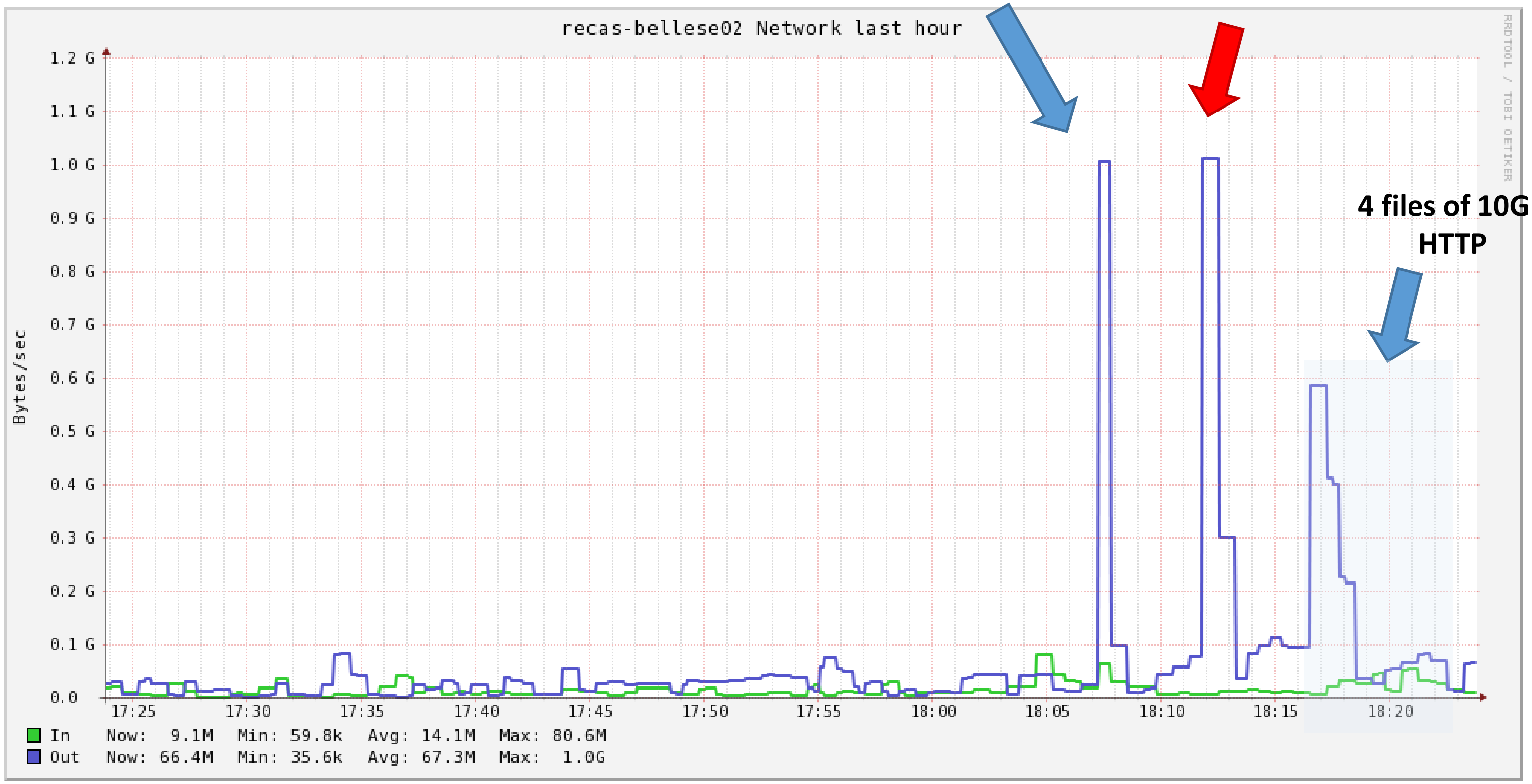
Third-Part Copy between two storage in Napoli

belle-dpm-01.na.infn.it vs t2-dpm-01.na.infn.it orchestrated from a User Interface.

```
[spardi@gridui TEST]$ davix-cp -P grid https://belle-dpm-01.na.infn.it/dpm/na.infn.it/home/belle/TMP/belle/user/spardi/testhttp/10G https://t2-dpm-01.na.infn.it/dpm/na.infn.it/home/belle/spardi/test-10G

0 (0 bytes/sec)

...

11899305984 (349979587 bytes/sec)


[spardi@gridui TEST]$ lcg-cp srm://belle-dpm-01.na.infn.it/dpm/na.infn.it/home/belle/TMP/belle/user/spardi/testhttp/10G srm://t2-dpm-01.na.infn.it/dpm/na.infn.it/home/belle/spardi/test-10G-01 -v

Source URL for copy: gsiftp://recas-bellese02.na.infn.it/recas-bellese02.na.infn.it:/SE02b/belle/2016-06-07/10G.14792856.0

Destination URL: gsiftp://atlasse13.na.infn.it/atlasse13.na.infn.it:/SE13c/belle/2016-11-14/test-10G-01.275414253.0

# streams: 1

            0 bytes      0.00 KB/sec avg      0.00 KB/sec inst

  ...

12778995712 bytes 388848.38 KB/sec avg 356277.09 KB/sec inst
```

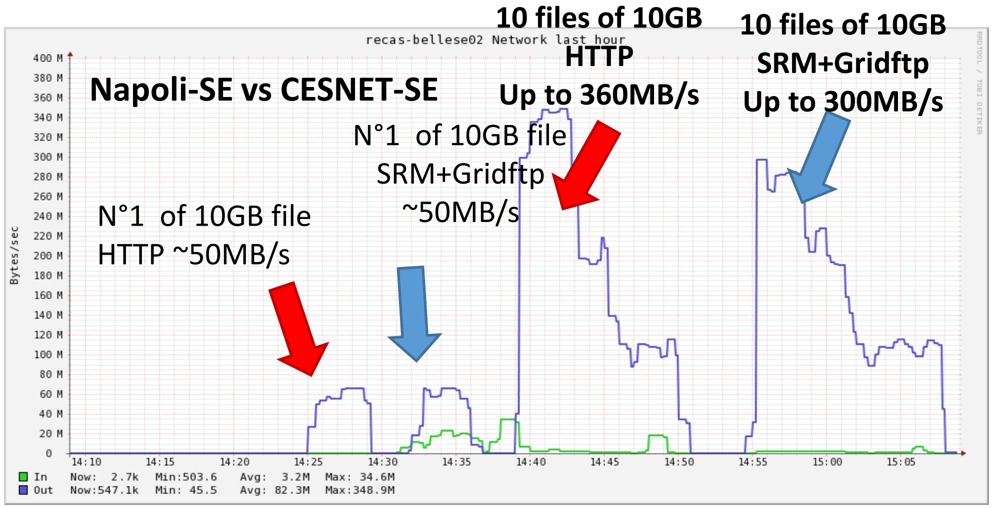
Third-Part Copy between two storage in Napoli

belle-dpm-01.na.infn.it vs t2-dpm-01.na.infn.it orchestrated from a User Interface.



Third-Part Copy with different technologies

Preliminary tests performed with success:
DPM vs DPM: Napoli-SE vs CESNET-SE
DPM vs STORM: Napoli-SE vs CNAF-SE
DPM vs dCache: Napoli-SE vs DESY-SE



Napoli-SE vs CESNET-SE

N°1 of 10GB file
HTTP ~50MB/s

N°1 of 10GB file
SRM+Gridftp
~50MB/s

10 files of 10GB
HTTP
Up to 360MB/s

10 files of 10GB
SRM+Gridftp
Up to 300MB/s

Performance must be checked.

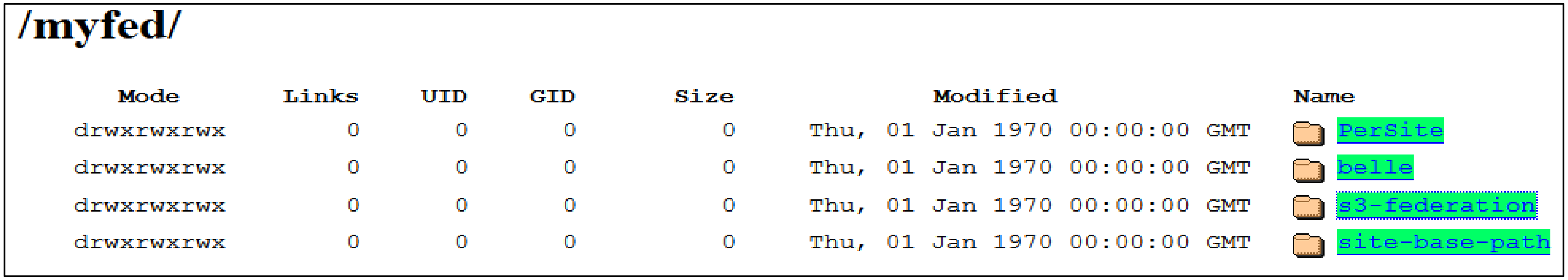
However from the first tests, http protocols behaviours seem consistent, and 3rd-part copy seems to work properly. (no traffic on client)

N.B Test could be affected by the local configuration at CESNET
(2 Disk nodes 10 Gbp + 2 Disk nodes 1Gbps
Randomly chosen at each data transfer)

Http 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/
- **myfed/site-based-path/** Aggregation of all the root directory of different storages
- **myfed/s3-federation/** Testing area for cloud storage



```
/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     s3-federation
  drwxrwxrwx           0      0      0          0   Thu, 01 Jan 1970 00:00:00 GMT     site-base-path
```

```
Example
#Unic Path Configuration
glb.locplugin[]: /usr/lib64/ugr/libugrlocplugin_dav.so Site01-Napoli-DATA-SE 5 https://belle-dpm-
01.na.infn.it:443/dpm/na.infn.it/home/belle/DATA/belle/
glb.locplugin[]: /usr/lib64/ugr/libugrlocplugin_dav.so Site01-Napoli-TMP-SE 5 https://belle-dpm-
01.na.infn.it:443/dpm/na.infn.it/home/belle/TMP/belle/

glb.locplugin[]: /usr/lib64/ugr/libugrlocplugin_dav.so Site01-Frascati-DATA-SE 5
https://atlasse.lnf.infn.it:443/dpm/lnf.infn.it/home/belle/DATA/belle
glb.locplugin[]: /usr/lib64/ugr/libugrlocplugin_dav.so Site01-Frascati-TMP-SE 5
https://atlasse.lnf.infn.it:443/dpm/lnf.infn.it/home/belle/TMP/belle/
```

Http experience from Users

Scientists from Padova : Alessandro Mordà and Stefano Lacaprara, are performing part of their analysis using the DPM storage in Napoli with HTTP interface.

Data movement done with different protocols.

- KEKCC -> KEK-SE : gfal-copy file:///basepath_KEK/test_file srm://kek-se03.cc.kek.jp/belle/test/test_file
- KEK-SE -> NA-SE : gfal-copy srm://kek-se03.cc.kek.jp/belle/test/test_file https://belle-dpm-01.na.infn.it/dpm/na.infn.it/home/belle/test1/test_file **(SRM STORM vs HTTP DPM)**
- NA SE -> PD: gfal-copy* *https://belle-dpm-01.na.infn.it/dpm/na.infn.it/home/belle/test1/test_file file:///basepath_PD/test_file

Performance User-based:
Data Transfer tests between NA-SE and PD-UI with HTTP protocol shown a transfer rate up **49MB/s** with single 1GB transfer.

Http experience from Users

For Analysis jobs two options considered:

- Copy files from DPM of NAPOLI to UI in PD (especially for signal MC files)
- Run the analysis scripts by directly accessing the input files stored in NA (what we would like to do for background samples, for which we don't need to run many times the scripts to optimize the analysis).

The time performances are the following: running analysis script on the same file:

- with local access in PD: **1m07s**
- with remote access to NA using https: **1m45s**
- with remote access to NA using root: **1m47s**

NA SE used ad transfer point to move files from KEK servers to Padova: time performances are quite good, once the size of the transferred files is optimized (about 1 Giga is the optimal one at least for background samples).

Once larger MC datasets will be available (after the current MC7 production campaign) background files will probably be accessed directly to NA, copying only the signal samples in PD.

Conclusion

- DPM is largely used from the Belle II community.

- The ongoing MC 7th is stressing the whole Belle II Computing Infrastructure including DPM storages that are properly working.

- More tests are ongoing in order to understand HTTP performance in different scenarios. The good HTTP support offered by DPM simplify the protocol exploration.

- Really appreciate is the proactive support of the DPM team.