CNAF experience in support of the JUNO distributed computing model

Andrea Rendina

Outline

- Introduction
- Early functional tests
- First network challenge
- Second network challenge
- Conclusions and future challenges

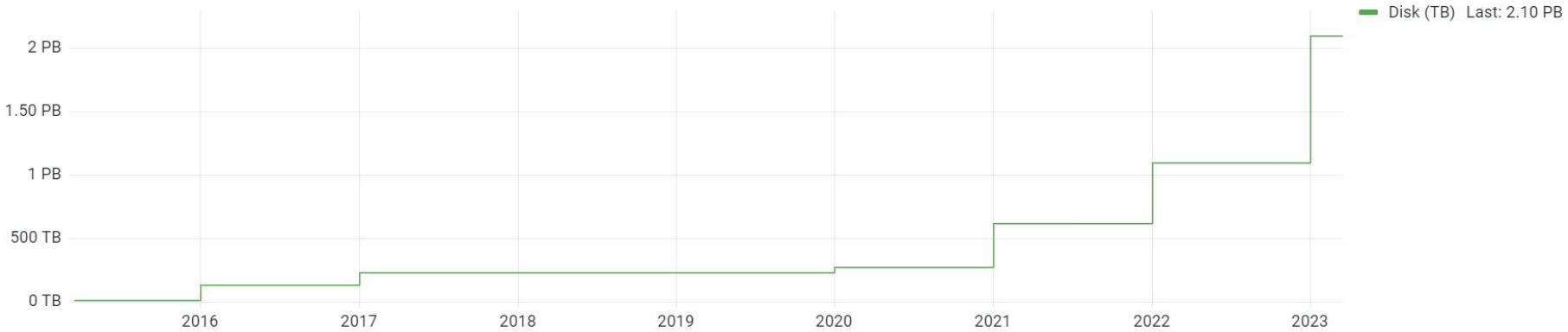
Introduction

CNAF Data Center

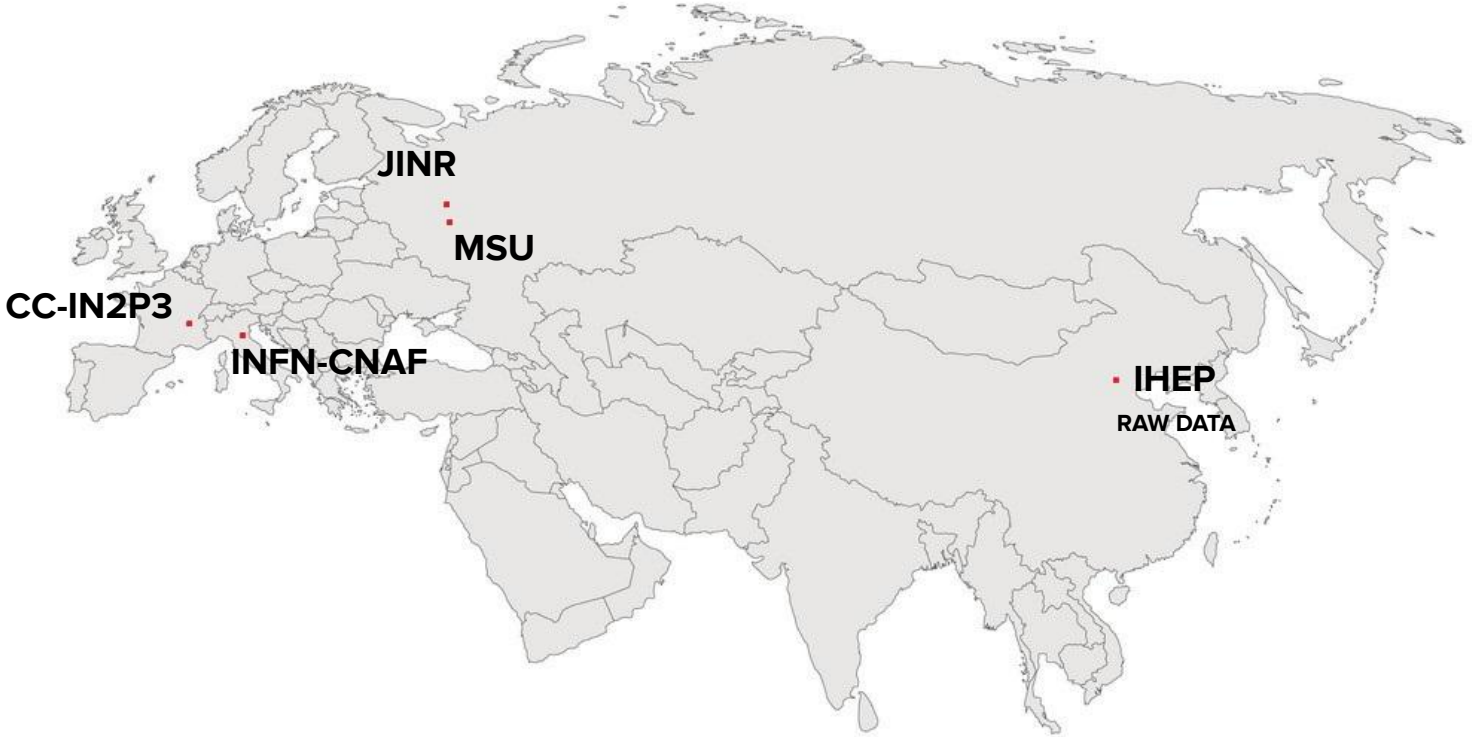
- CNAF hosts the main INFN data center, and the INFN Tier-1 in the WLCG e-infrastructure
- Provides services and resources to more than 60 scientific collaborations
  - LHC experiments so far the more demanding
  - ~50k cores, ~70 PB of disk, ~130 PB of tape
- Huge increase of resources foreseen in the coming years. By 2025:
  - ~130k cores, ~110 PB of disk, ~250 PB of tape
  - and even more (x10) from 2027 (HL-LHC)

Remote data access for JUNO at CNAF

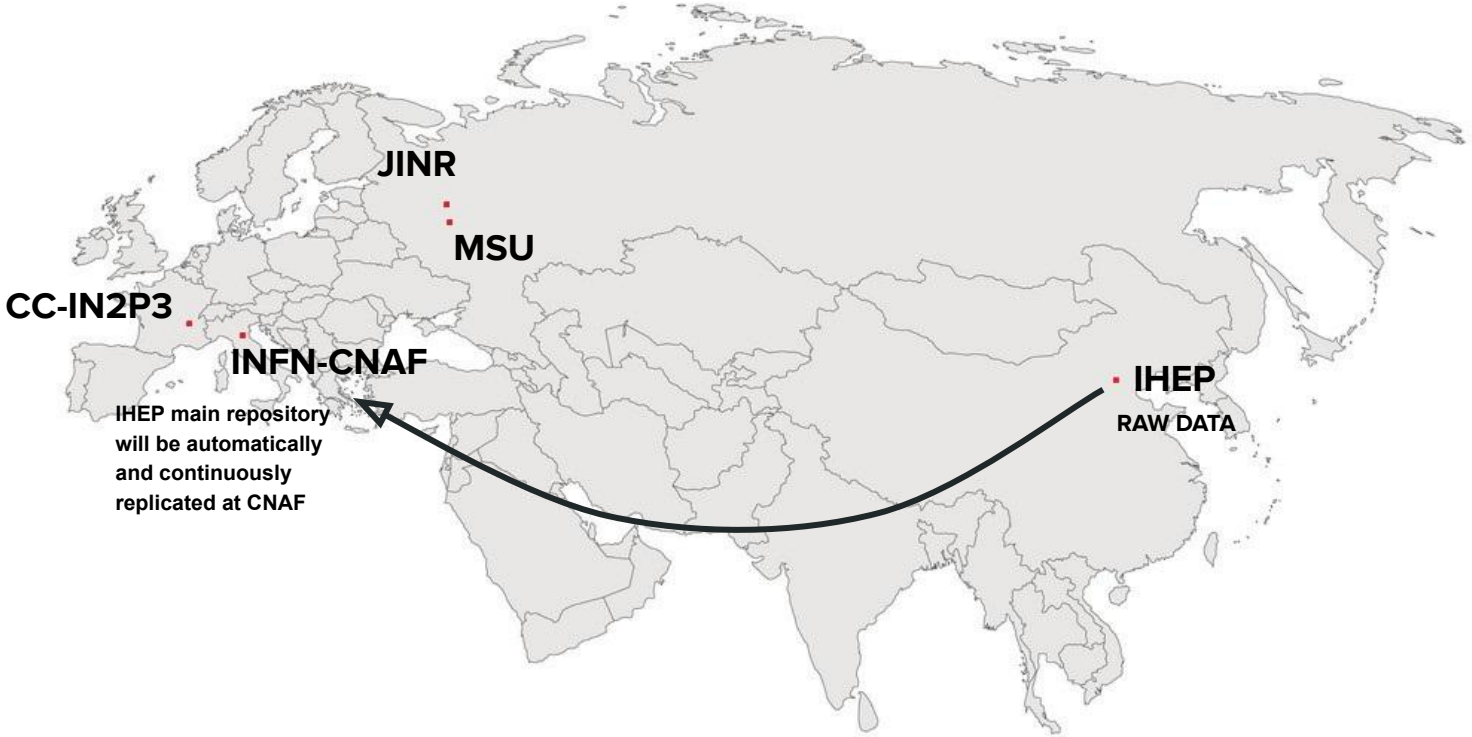
- Dedicated Fileset on GPFS
  - 1.1PB of disk quota dedicated to JUNO
  - Remote access thanks to StoRM WebDAV (https/davs)
  - 4 different Storage Areas to access data
  - AuthN/Z both with voms-proxy and IAM tokens
  - 4 StoRM WebDAV servers shared among several different experiments



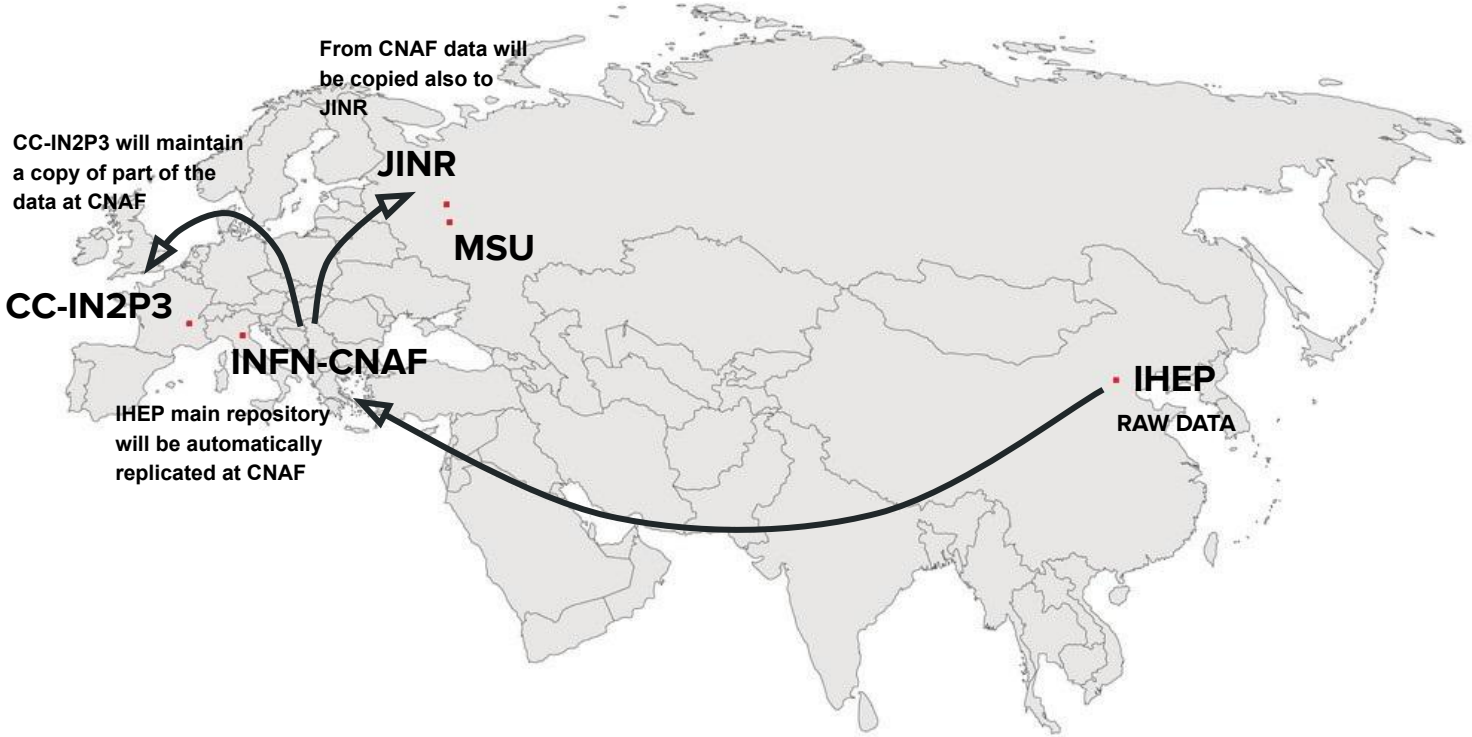
The experiment data flow



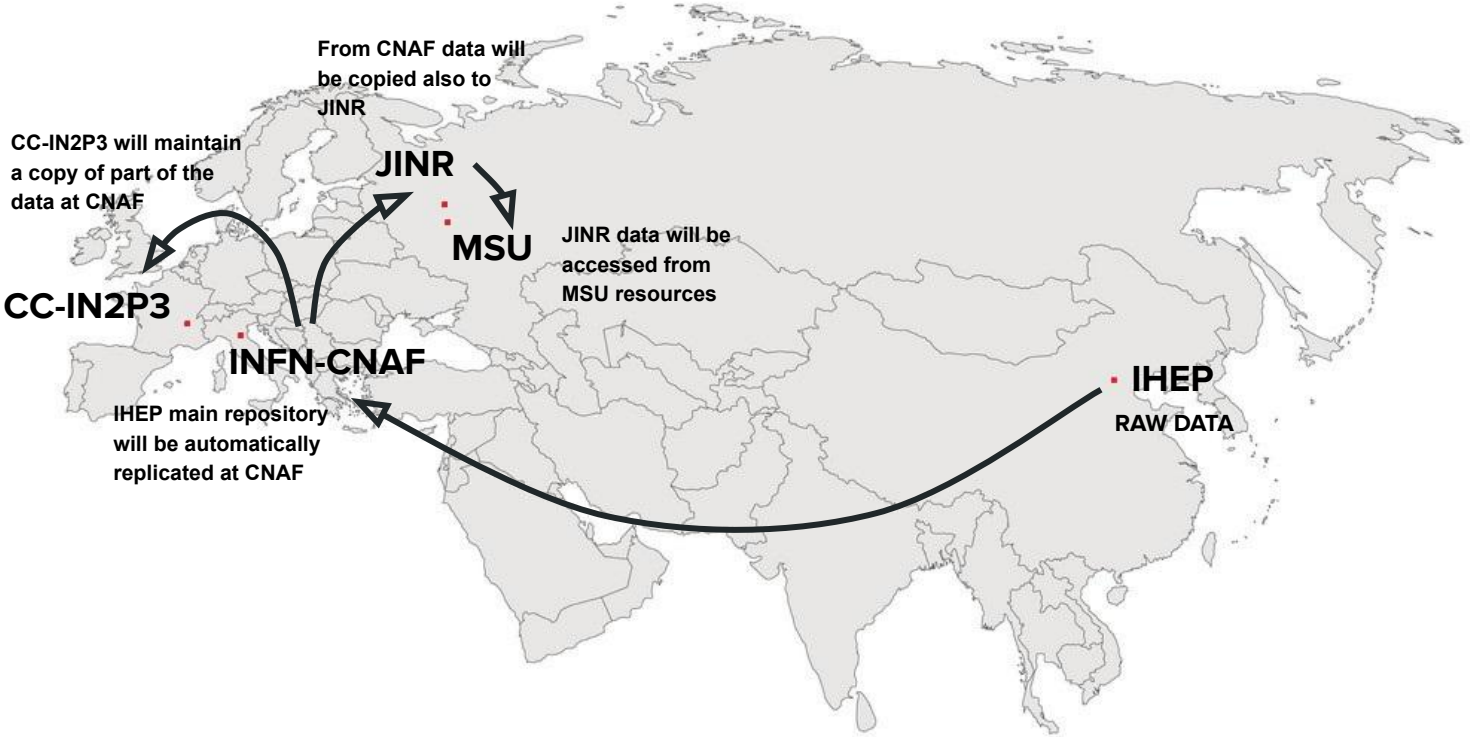
The experiment data flow



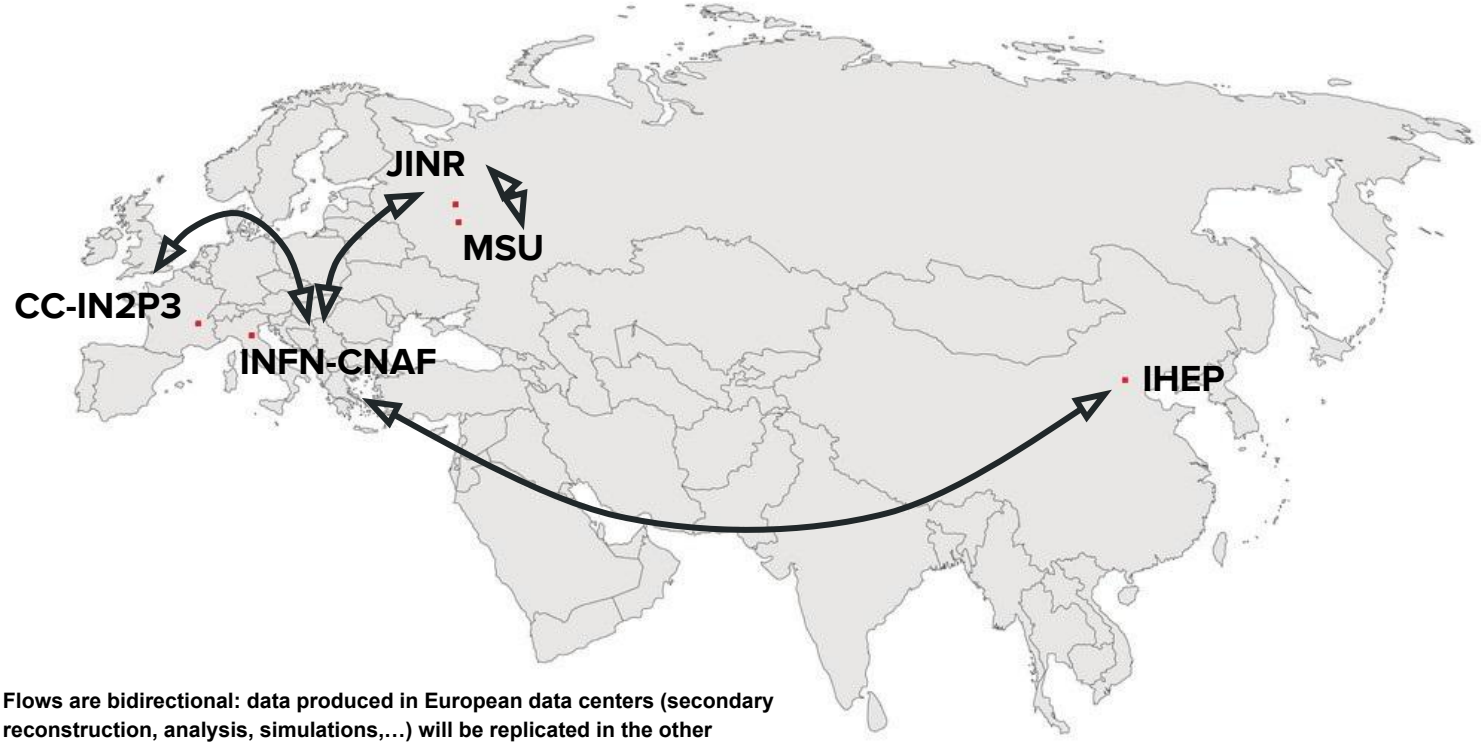
The experiment data flow



The experiment data flow

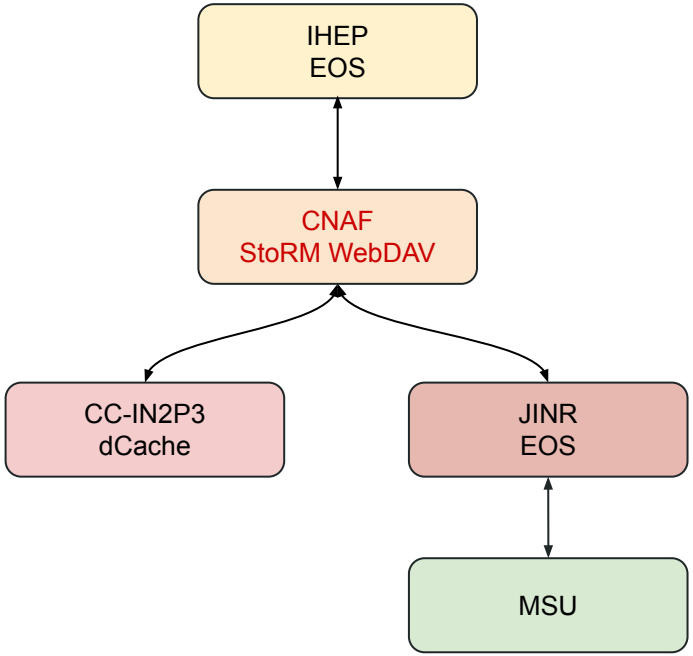


The experiment data flow



Flows are bidirectional: data produced in European data centers (secondary reconstruction, analysis, simulations,…) will be replicated in the other European data centers and at IHEP

The experiment data flow



- IHEP main repository will be automatically replicated at CNAF
- CC-IN2P3 will maintain a copy of part of the data at CNAF
- From CNAF data will be copied also to JINR
- JINR data will be accessed from MSU resources
- Flows are bidirectional

Early functional tests

Rucio and FTS instances at CNAF

- Federated data management model
- Rucio testbed
  - 1.26.9 version
  - The service has initialised (server, clients, ui, database, daemons) with a Docker-compose
  - Authentication by Userpass, X.509 certificates and token IAM
- FTS testbed
  - 3.10.1 version
  - Web user-interface for monitoring
- JUNO deployment at CNAF in January 2022

Early functional tests with Rucio/FTS

TPC		TO			
		IN2P3	JINR	CNAF	IHEP
FROM	IN2P3	PULL		PULL	PULL
	JINR				
	CNAF	PULL		PULL	PULL
	IHEP	PULL		PULL	PULL

- https/davs protocol
- Small size files
- No errors between the sites for pull-mode copies, except JINR
  - JINR endpoint was not enabled for TPCs yet
- We decided to increase the size and check both pull and push mode for each transfer
- Increase the number or transfers

Early functional tests with gfal-copy 2.20.1

TPC		TO			
		IN2P3	JINR	CNAF	IHEP
FROM	IN2P3	PULL/PUSH		PULL/PUSH	PULL/PUSH
	JINR				
	CNAF	PULL/PUSH		PULL/PUSH	PULL
	IHEP	PULL/PUSH		PUSH	PULL/PUSH

- All the third-party copies work well between CNAF (StoRM WebDAV) and CC-IN2P3 (dCache)
- Several different errors between CNAF and IHEP (EOS) in both directions

First network data challenge

General situation between CNAF and IHEP

TPC		TO					
		CNAF			IHEP		
		PULL	PUSH	STREAM	PULL	PUSH	STREAM
FROM	CNAF	39 MB/s	81 MB/s	100 MB/s	263 KB/s **	*	263 KB/s **
	IHEP	4 MB/s **	1.89 MB/s **	2.5 MB/s **	50 MB/s	50 MB/s	299 KB/s

- *Constant error in push-mode copies from StoRM WebDAV to EOS
  - “SSLException while pushing [...]: Broken pipe (Write failed)”
  - (see this in slide 20)
- ** Frequent errors
  - “SocketTimeoutException while fetching [...]: Read timed out”

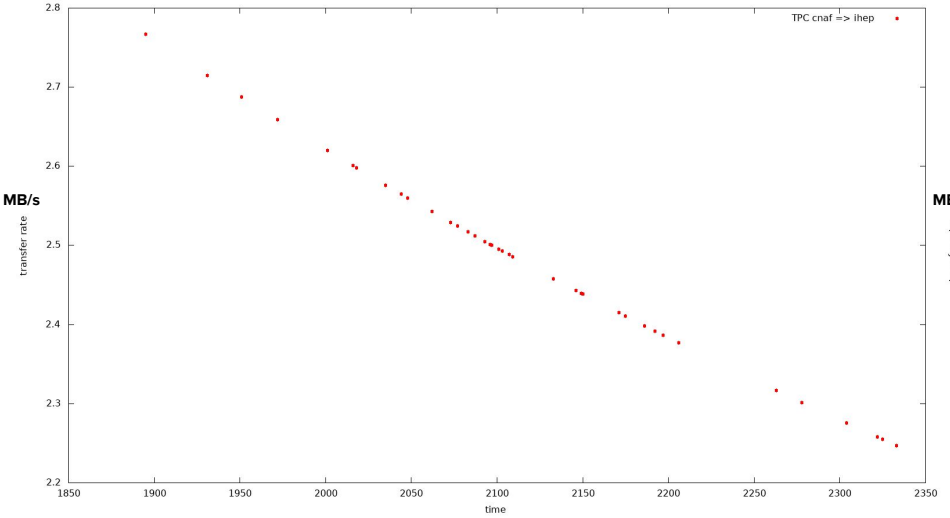
General situation between CNAF and IHEP

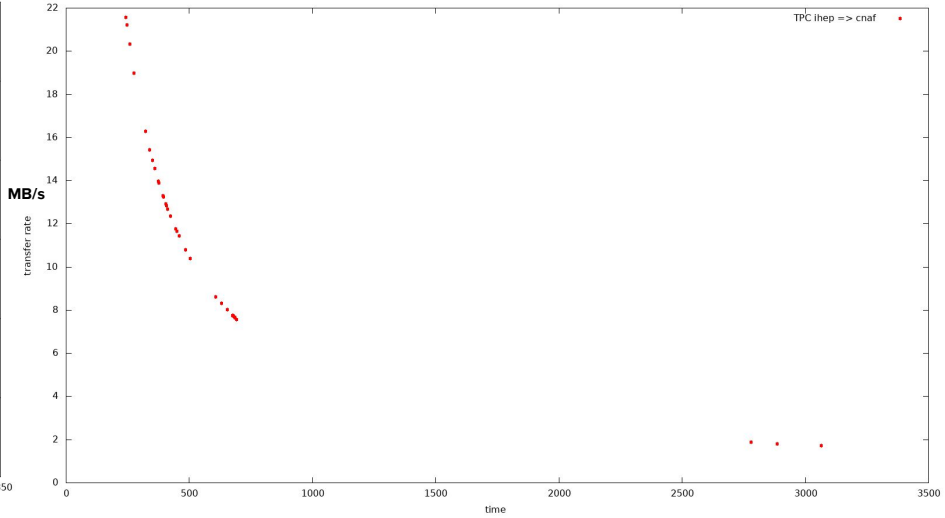
```
[root@junoeos01 ~]# tracepath -p 8443
xfer-archive.cr.cnaf.infn.it
1?:[LOCALHOST]                                  pmtu 1500
1:  gateway                                   4.773ms
1:  gateway                                   1.577ms
2:  202.122.37.209                            0.422ms
3:  no reply
4:  202.122.32.253                            1.203ms
5:  vpn1.ihep.ac.cn                           2.525ms
6:  cstnet-lhcone-gw.fra.de.geant.net         2.353ms
7:  cstnet-lhcone.fra.de.geant.net          151.010ms
8:  62.40.126.186                           150.457ms
9:  garr-lhcone-gw.gen.ch.geant.net         156.935ms
10:  ru-infn-cnaf-lhcone-l1-rx1.bo1.bo1.garr.net     159.870ms
11:  rx1.bo1-ru-infn-cnaf-lhcone-l2.bo1.garr.net     159.839ms
12:  ds-203-06-10.cr.cnaf.infn.it           159.455ms reached
Resume: pmtu 1500  hops 12 back 12
[root@ds-512 ~]# tracepath junoeos01.ihep.ac.cn
 1?: [LOCALHOST]                                        pmtu 9000
 1:  gw2-128.cr.cnaf.infn.it                             92.806ms
 1:  gw2-128.cr.cnaf.infn.it                              2.396ms
 2:  ru-infn-cnaf-lhcone-l1-rx1.bo1.bo1.garr.net          0.752ms
 3:  garr-lhcone-gw.gen.ch.geant.net                      3.966ms
 4:  geant-lhcone-gw.mx1.gen.ch.geant.net                10.033ms
 5:  62.40.126.178                                       18.654ms
 6:  cstnet-lhcone-gw.fra.de.geant.net                  160.967ms
 7:  192.168.200.1                                      161.517ms
 8:  192.168.200.2                                      160.402ms
 9:  no reply
```

- The network routes don’t change and the paths are symmetric
- MTU mismatch: 9000 vs 1500
  - INFN Tier-1 joined WLCG and so LHC OPN/ONE

General situation between CNAF and IHEP

- Quantify the maximum throughput with 40 parallel transfers of 5GB files
  - From 400 up to 700 Mb/s from CNAF to IHEP
  - About 4 Gb/s from IHEP to CNAF





Improving actions

- Increase the MTU on IHEP servers up to 9000
- Studying the PUSH copies from StoRM WebDAV to EOS (always failing)
  - Involving StoRM and EOS developers we found out that StoRM WebDAV sends the data and metadata together, whereas EOS is not able to manage this kind of transfers
  - EOS manages metadata and data independently
- Measure the maximum bandwidth achieved between the two sites using iperf tools
  - About 3Gb/s from one server to another in both directions
- Activate perfsonar instances for each site in order to constantly monitor the situation

Improving actions

- Increased the MTU on IHEP servers up to 9000 (April 2022)
- The single transfer rates improved a lot
- The amount of errors decreased, but it was still quite high
- Quantify again the maximum throughput with 40 parallel transfers of 5GB files
  - About 7 Gb/s from IHEP to CNAF and on the other way around
    - But these values are very fluctuating

Second network data challenge

Established procedure

- Use iperf3 in order to measure the maximum bandwidth between the sites
- Issue 10, 40 and 100 parallel transfers of 5GB files in order to quantify the maximum achievable throughput
  - Adding up the rates of the single file transfers
  - The authN/Z is always done with a JUNO valid voms-proxy
- The tests have been performed in January and February 2023

iperf tests

iperf3		TO			
		IN2P3	JINR	CNAF	IHEP
FROM	IN2P3			10Gb/s	
	JINR			6.5Gb/s*	
	CNAF	10Gb/s	10Gb/s		3Gb/s
	IHEP			3Gb/s	

- iperf3 tests show good results in general
- CNAF and CC-IN2P3 are in LHC-ONE and the results were as expected
- * Peak achieved with 40 parallel streams
- * 5.2 Gb/s average

gfal-copy tests (version 2.21.2)

TPC		TO					
		CNAF			CC-IN2P3		
		10	40	100	10	40	100
FROM	CNAF				10.9Gb/s	15.8Gb/s	7.83Gb/s
	CC-IN2P3	3.53Gb/s	9.04Gb/s	7.82Gb/s			

- Very good results as expected in both directions
- In particular for 40 parallel transfers
- StoRM WebDAV and dCache manage very well the TPCs between them

gfal-copy tests (version 2.21.2)

TPC		TO					
		CNAF			JINR		
		10	40	100	10	40	100
FROM	CNAF				6.1Gb/s	12.8Gb/s	3.1Gb/s
	JINR	1.74Gb/s	3.57Gb/s	3.90Gb/s			

- Very good results from CNAF to JINR
  - No errors occur
- Lower results from JINR to CNAF
  - Also from CNAF to JINR with 100 parallel transfers
  - It is very probable that at JINR there is only one data transfer EOS server

gfal-copy tests (version 2.21.2)

TPC		TO					
		CNAF			IHEP EOS		
		10	40	100	10	40	100
FROM	CNAF				282Mb/s	1.70Gb/s	1.92Gb/s
	IHEP EOS	1.97Gb/s	6.14Gb/s	7.79Gb/s			

- 75% failed from CNAF to IHEP EOS
  - Low maximum throughput achieved
- 20% failed from IHEP EOS to CNAF
  - But the maximum throughputs achieved are good
- A single transfer file in push or pull mode has a rate of about 50MB/s

gfal-copy tests (version 2.21.2)

TPC		TO					
		CNAF			IHEP StoRM WebDAV		
		10	40	100	10	40	100
FROM	CNAF				952Mb/s	1.82Gb/s	1.53Gb/s
	IHEP StoRM WebDAV	1.74Gb/s	2.28Gb/s	2.86Gb/s			

- No errors occur
- Low maximum throughput achieved
  - It is very probable that at IHEP there is only one StoRM WebDAV server
- A single transfer file in push or pull mode has a rate of about 10MB/s

Conclusions and future challenges

Conclusions and results

- In general, the connectivity between the involved sites of the collaboration shows good results both with iperf tool and issuing the proper amount of parallel file transfers
- The MTU mismatch caused “TIMEOUT exceed” errors and the package fragmentations decreased a lot the rate of a single transfer
  - From 50MB/s up to 2MB/s
- The MTU alignment didn’t fix all the problems related to the transfers from CNAF StoRM WebDAV and IHEP EOS
- Push-mode failures from StoRM WebDAV to EOS don’t allow to exactly measure the maximum achievable throughput
- StoRM WebDAV and dCache manage very well third-party copies

Future challenges

- New StoRM WebDAV release will fix the TPCs in push-mode from StoRM WebDAV to EOS
- Changing the authN/Z from voms-proxy to IAM tokens will improve the efficiency of the transfers
  - Avoid macaroons requests to each transfer server
- Increasing the number of StoRM WebDAV servers at IHEP could improve a lot the maximum throughput between the two sites
- Align the EOS release of IHEP to that one at JINR
  - From CNAF StoRM WebDAV to JINR EOS there are only errors for push-mode copies
  - No errors occur in the other way round

Thanks