PROCEEDINGS OF SCIENCE

CNAF experience in support of the JUNO distributed computing model

Andrea Rendina,[*] **Lucia Morganti, Federico Fornari, Lorenzo Chiarelli, Donato De Girolamo and Stefano Zani**

INFN-CNAF,
viale Berti Pichat 6/2, Bologna, Italy

E-mail: andrea.rendina@cnaf.infn.it

The Italian WLCG Tier-1 located in Bologna and managed by INFN CNAF provides computing and storage resources to several research communities in the fields of High-Energy Physics, Astroparticle Physics, Gravitational Waves, Nuclear Physics and others. Among them, the Jiangmen Underground Neutrino Observatory (JUNO), devoted to the construction and operation of a neutrino detector located underground in Kaiping, Jiangmen in Southern China, will employ a computing infrastructure geographically distributed in Chinese, Russian, French and Italian data centers. The detector data rate is expected to be of the order of 2 PB per year, continuously transferred from the detector site to the INFN Tier-1 in Italy. To guarantee the optimal operations among all the aforementioned sites, a series of periodic network and data management challenges has been performed.

In this talk, the technologies involved to set up the cross-continent data transfer (e.g. StoRM WebDAV, EOS, dCache, FTS, Rucio) are presented, together with their performance.

International Symposium on Grids & Clouds (ISGC) 2023,
19-31 March, 2023
Academia Sinica Taipei, Taiwan

[*] Speaker

© Copyright owned by the author(s) under the terms of the Creative Commons
Attribution-NonCommercial-NoDerivatives 4.0 International License (CC BY-NC-ND 4.0).

1. Introduction

CNAF, located in Bologna, is the INFN National Center dedicated to Research and Development on Information and Communication Technologies [1]. It hosts the main INFN data center, providing services and resources to more than 60 scientific collaborations and representing the INFN Tier-1 in the WLCG e-infrastructure. Among them, the Jiangmen Underground Neutrino Observatory (JUNO) experiment [2] [3], devoted to the construction and operation of a neutrino detector located underground in Kaiping, Jiangmen in Southern China, is and will be supported by CNAF.

At CNAF, data of the JUNO collaboration are hosted in a dedicated GPFS fileset of 1.1 PB, remotely accessible thanks to the StoRM WebDAV transfer service and the protocol https/davs. In particular, two StoRM WebDAV servers are deployed at CNAF in support of the activities of JUNO and several other experiments. The JUNO collaboration can manage data stored at CNAF through 4 different storage areas and the authentication and authorization are based both on X.509 voms-proxy and IAM JSON Web Token, for reference see [4] [5].

The other involved sites of the collaboration are IHEP in Beijing, CC-IN2P3 in Lyon, JINR in Dubna and the Moscow State University (MSU). The start of the period of data taking, hereafter Run, is scheduled for the second part of 2023 and the entire Run will take more than 10 years. The expected amount of raw data produced each year is about 2 PB [6]. IHEP is designed to be the main repository of a full copy of the raw data, which will be totally, continuously and automatically replicated to CNAF. CC-IN2P3 and JINR will maintain a partial copy of the raw data. JINR data will be accessed also from MSU for end-user analysis. Figure 1 shows the data flow, which have to be bidirectional because the data produced in European data centers (secondary reconstruction, analysis, simulations) will be replicated in the other European data centers and at IHEP. The goal of the collaboration is to maximize the achievable throughput between the involved sites in order to guarantee the data availability in each site of the collaboration.

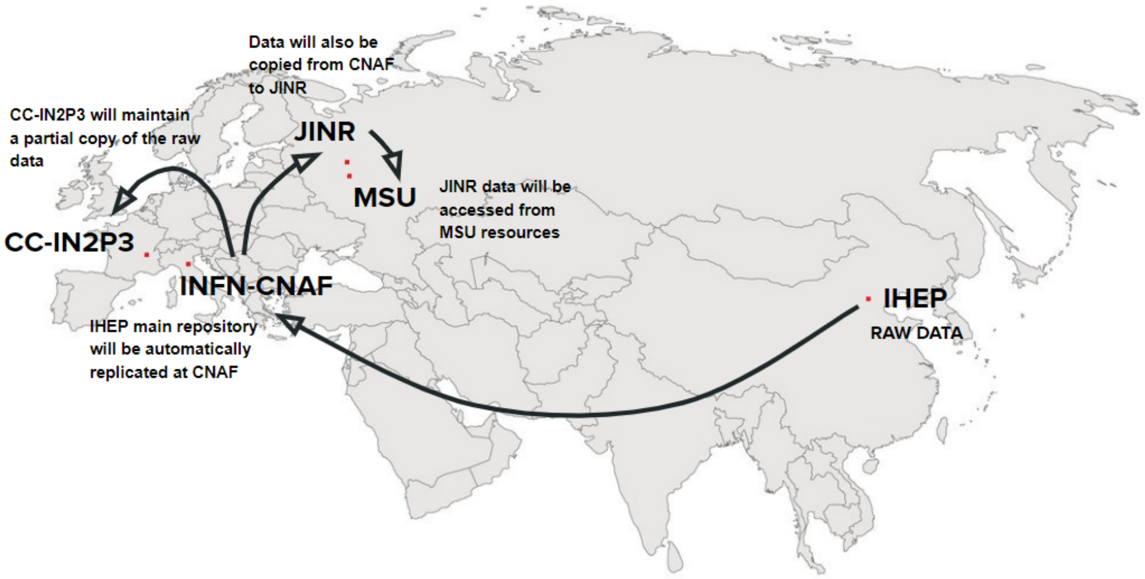
Data will also be copied from CNAF to JINR
CC-IN2P3 will maintain a partial copy of the raw data
JINR
MSU
JINR data will be accessed from MSU resources
CC-IN2P3
INFN-CNAF
IHEP main repository will be automatically replicated at CNAF
IHEP
RAW DATA

Figure 1: Experiment data flow

PoS(ISGC&HEPiX2023)006

2. Early functional tests

In order to properly perform the preliminary functional tests between the collaboration sites, a federated management model was deployed at CNAF based on FTS [7] and Rucio [8] in January 2022. A Rucio testbed (server, clients, ui, database, daemons) and an FTS one have been deployed with Docker-compose. The authentication to the Rucio server has been enabled via username and password, X.509 certificates and IAM tokens. An FTS web interface has been implemented, allowing the monitoring of the transfers submitted by Rucio. Depending on historical and technical reasons, each data center has deployed endpoints based on their preferred data transfer service. At IHEP and JINR some EOS endpoints [9] have been set up for this purpose, whereas at CC-IN2P3 a dCache one [10]. In this context, CNAF - with its StoRM WebDAV transfer service - will play a major role for the experiment data flow as detailed above in Section 1.

In view of the Globus toolkit retirement [11], all the preliminary transfers have been implemented with the https/davs protocol, using X.509 voms-proxy for the authentication and authorization to the endpoints and transferring small-sized files. Preliminary Third-Party Copies (TPCs) using the Rucio/FTS infrastructure were promising: Table 1 reports the successful directions explored via FTS transfers.

Third-Party Copies		**To**			
		CNAF	IHEP	IN2P3	JINR[1]
From	CNAF	Pull	Pull	Pull	
	IHEP	Pull	Pull	Pull	
	IN2P3	Pull	Pull	Pull	
	JINR[1]				

Table 1: Early functional tests using FTS - only successful directions are reported in the table (see Section 3 for details)
FTS tries to perform a pull-mode copy; if this fails, a push-mode copy is issued.

As a consequence, the collaboration decided to increase the number of concurrent transfers and also the files' size up to 100 MB, checking both pull and push mode [11] [12] with gfal-copy.

Third-Party Copies		**To**			
		CNAF	IHEP	IN2P3	JINR[1]
From	CNAF	Pull/Push	Pull	Pull/Push	
	IHEP	Push	Pull/Push	Pull/Push	
	IN2P3	Pull/Push	Pull/Push	Pull/Push	
	JINR[1]				

Table 2: Tests using 2.20.1 gfal-copy - only successful directions are reported in the table (see Section 3 for details).

[1] JINR endpoint was not enabled for TPCs at the time of the tests.

PoS(ISGC&HEPiX2023)006

The results in Table 2 show that all the third-party copies worked well between CNAF (StoRM WebDAV) and CC-IN2P3 (dCache), although several different and random errors occurred between CNAF and IHEP (EOS) in both directions.

Measurements of bandwidth and throughput will be detailed in the following sections.

3. First network challenge

Given the critical situation between CNAF and IHEP, a deeper investigation was needed to figure out the problems and the possible improvements. For these reasons, we traced the performances for each copy mode from one site to the other and vice versa (Table 3).

Third-Party Copies (MB/s)		**To**					
		CNAF			IHEP		
		Pull	Push	Stream	Pull	Push	Stream
From	CNAF	39	81	100	0.263 **	*	0.263 **
	IHEP	4.0 **	1.89 **	2.5 **	50	50	0.299

Table 3: General situation between CNAF and IHEP
*Constant error in push-mode copies from StoRM WebDAV to EOS
**Frequent errors

A lot of errors occurred in both directions with a "timeout exceeded" error, namely `SocketTimeoutException while fetching [...]: Read timed out`. Furthermore, for files with size greater than 1 MB, there were constant errors in the push-mode copies from CNAF (StoRM WebDAV) to IHEP (EOS) with the `SSLException while pushing [...]: Broken pipe (Write failed)` error. In fact, StoRM WebDAV sends the data and metadata together, whereas EOS is not able to manage this kind of transfers because it manages metadata and data independently. However, as explained in Listings 1 and 2, the network routes from CNAF to IHEP and vice versa were symmetric and this excluded the possibility of package-route issues.

Listing 1: tracepath from CNAF to IHEP

```
[root@ds-512 ~]$ tracepath junoeos01.ihep.ac.cn
 1?: [LOCALHOST]                                      pmtu 9000
 1:  gw2-128.cr.cnaf.infn.it                            0.760ms
 1:  gw1-128.cr.cnaf.infn.it                            0.796ms
 2:  ru-infn-cnaf-lhcone-l2-rx1.bo1.bo1.garr.net        0.718ms
 3:  garr-lhcone-gw.gen.ch.geant.net                    3.954ms
 4:  geant-lhcone-gw.mx1.gen.ch.geant.net              10.388ms
 5:  62.40.126.178                                     18.818ms
 6:  cstnet-lhcone-gw.fra.de.geant.net                163.385ms
 7:  192.168.200.1                                    166.804ms
 8:  192.168.200.2                                    161.033ms
 9:  no reply
```

PoS(ISGC&HEPiX2023)006

Listing 2: tracepath from IHEP to CNAF

```
[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
```

Instead, an MTU mismatch was found out because CNAF servers, as members of the WLCG network infrastructure, have Maximum Transfer Unit (MTU) equal to 9000, whereas IHEP servers MTU was set to 1500.

We also decided to quantify the maximum achievable throughput issuing 40 parallel transfers of 5 GB files. Adding up each single transfer rate, we found out very low results. In particular, we achieved from 400 up to 700 Mbit/s from CNAF to IHEP and about 4 Gbit/s in the other direction. On this trail, we also computed the maximum bandwidth between the two data centers using the iperf3 tool [13], reaching about 3 Gbit/s from one site to the other in both directions.

Being aware of the good connectivity shown by the iperf tests, together with the JUNO community we undertook the following improving actions:

- increase the MTU on the EOS IHEP server up to 9000;
- analyze the PUSH copies from StoRM WebDAV to EOS, because they always failed;
- activate perfsonar [14] instances for each site in order to constantly monitor the situation.

In particular, in April 2022 the MTU of the EOS server at IHEP was increased up to 9000. This change led to the following facts:

- the single transfer rates improved from a few MB/s up to 50 MB/s;
- the total amount of errors decreased, although remaining quite high;
- the maximum throughput with 40 parallel transfers of 5 GB files reached about 7 GB/s from IHEP to CNAF and on the other way around, but the values were very fluctuating and therefore not trustworthy.

4. Second network challenge

On the basis of what we had observed during the preliminary and functional tests and the first network challenge, the JUNO experiment community decided to implement an established procedure in order to compute the bandwidth and the maximum achievable throughput between all the sites of the collaboration.

Indeed, in January and February 2023 the following procedure basically composed by two steps has been performed:

- use the iperf3 tool in order to measure the maximum bandwidth;
- issue 10, 40 and 100 parallel transfers of 5 GB files in order to quantify the maximum achievable throughput, adding up the rates of the single file transfers (the authentication and authorization mechanism still being based on voms-proxy for JUNO Virtual Organization).

The outcome of the iperf tests is summarized in Table 4.

iperf3		**To**			
(Gbit/s)		CNAF	IHEP	IN2P3	JINR*
From	CNAF		3.0*	10	10
	IHEP	3.0*			
	IN2P3	10			
	JINR	6.5**			

Table 4: iperf3 tests
*Results shown by perfsonar instances
**Peak achieved with 40 parallel streams, 5.2 Gbit/s average

As expected, the transfers concerning CNAF and CC-IN2P3 performed very well, since both are part of the WLCG network infrastructure. In general, despite the long distance, particularly between CNAF and IHEP, iperf shows very good results for the connectivity between CNAF and the other data centers.

For the second step of the challenge, we issued the transfers using the 2.21.2 version of gfal-copy, with authentication and authorization based on voms-proxy.

The tests between CNAF and CC-IN2P3, also in this case, showed very good results. As shown by Table 5, the achieved throughput was very high in both directions. In particular, the results issuing 40 parallel transfers were remarkably good. Also, no error occurred in all the performed tests and this means that StoRM WebDAV and dCache manage very well the third-party copies between each other.

The tests from CNAF to JINR yielded very good results, reaching 12.8 Gbit/s with 40 parallel transfers. On the other hand, we found out lower results from JINR to CNAF, also with 100 parallel transfers (3.90 Gbit/s). However, no errors occurred for all the performed tests. In light of our investigation, we believe that increasing the number of EOS servers at JINR should improve a lot the maximum throughput from JINR to CNAF.

In relation to the IHEP situation, a lot of transfers failed in both directions. In particular, the 75% of transfers from CNAF to IHEP EOS failed and this is one of the reasons why the achieved throughput values were quite low (1.70 Gbit/s, 1.92 Gbit/s). Additionally, the failure rate from IHEP to CNAF was 25% and the maximum achieved throughput values were good (6.14 Gbit/s,

PoS(ISGC&HEPiX2023)006

Throughput (Gbit/s)		**To** CNAF		
		10	40	100
From	IHEP EOS	1.97	6.14	7.79
	IHEP StoRM WebDAV	1.74	2.28	2.86
	CC-IN2P3	3.53	9.04	7.82
	JINR	1.74	3.57	3.90

Throughput (Gbit/s)		**From** CNAF		
		10	40	100
To	IHEP EOS	0.28	1.70	1.92
	IHEP StoRM WebDAV	0.95	1.82	1.53
	CC-IN2P3	10.9	15.8	7.83
	JINR	6.10	12.8	3.10

Table 5: Achieved throughput between CNAF and the other sites using 10, 40 and 100 parallel streams for the data transfer

7.79 Gbit/s). Furthermore, a single file transfer rate, both in push or pull mode, averages on about 50 MB/s.

At the end of 2022, a StoRM WebDAV server was installed at IHEP and we repeated the tests using such endpoint. This time, no errors occurred during all the performed tests, but lower maximum throughput values were observed. A single transfer file in push or pull mode showed an average rate of about 10 MB/s. As a matter of fact, increasing the number of StoRM WebDAV servers at IHEP should improve a lot the maximum throughput from and towards CNAF.

5. Conclusions

In this work, we described the activity carried out at INFN CNAF data center in support of the JUNO distributed computed model, with the goal of ensuring the optimal operations among all the sites involved in the data flow of the experiment (CNAF, IHEP, CC-IN2P3, JINR, MSU).

In general, the connectivity between the involved sites of the collaboration showed good results both using iperf tool and issuing the proper amount of parallel file transfers. At the very beginning, an MTU mismatch between the StoRM WebDAV servers located at CNAF and the EOS server at IHEP caused "TIMEOUT exceeded" errors. Indeed, packet fragmentation affected a lot the average rate of a single transfer, bringing it from 50 MB/s down to 2 MB/s. Unfortunately, the MTU alignment did not fix all the problems related to the transfers from CNAF to IHEP. Besides, the push-mode failures from StoRM WebDAV to EOS prevented the exact measurement of the maximum achievable throughput between the two sites. Despite the encountered problems between

PoS(ISGC&HEPiX2023)006

EOS and StoRM WebDAV, the tests between the latter and dCache at CC-IN2P3 have proven that these two services are able to properly manage the third-party copies among themselves.

Hopefully, the new StoRM WebDAV release (1.12.22) should fix the TPCs in push-mode from StoRM WebDAV to EOS. Also, the WLCG authentication and authorization protocol migration from X.509 certificates to JSON Web Tokens will improve the efficiency of the transfers, avoiding the need for macaroons requests to each transfer server [12] [15]. Increasing the number of StoRM WebDAV servers at IHEP should improve a lot the maximum throughput from and towards CNAF. In the end, aligning the IHEP and JINR EOS releases should decrease the total amount of failures, since the errors observed from CNAF StoRM WebDAV to JINR EOS only involve push-mode copies and no errors ever occurred the other way round.

PoS(ISGC&HEPiX2023)006

References

[1] Lucia Morganti et al. “Large Scale Data Handling experience at INFN-CNAF Data Center”. In: *Proceedings of 41st International Conference on High Energy physics — PoS(ICHEP2022)*. Vol. 414. 2022, p. 205. DOI: 10.22323/1.414.0205.

[2] Fengpeng An et al. “Neutrino physics with JUNO”. In: *Journal of Physics G: Nuclear and Particle Physics* 43.3 (2016), p. 030401.

[3] JUNO collaboration et al. “JUNO physics and detector”. In: *Progress in Particle and Nuclear Physics* 123 (2022), p. 103927.

[4] Brian Bockelman et al. “WLCG Token Usage and Discovery”. In: *EPJ Web of Conferences*. Vol. 251. EDP Sciences. 2021, p. 02028.

[5] Andrea Ceccanti, Enrico Vianello, and Francesco Giacomini. “Beyond X. 509: token-based authentication and authorization in practice”. In: *EPJ Web of Conferences*. Vol. 245. EDP Sciences. 2020, p. 03021.

[6] Giuseppe Andronico. “China-EU scientific cooperation on JUNO distributed computing”. In: *EPJ Web of Conferences*. Vol. 245. EDP Sciences. 2020, p. 03038.

[7] AA Ayllon et al. “FTS3: new data movement service for WLCG”. In: *Journal of Physics: Conference Series*. Vol. 513. 3. IOP Publishing. 2014, p. 032081.

[8] Martin Barisits et al. “Rucio: Scientific data management”. In: *Computing and Software for Big Science* 3 (2019), pp. 1–19.

[9] A Joachim Peters, Elvin Alin Sindrilaru, and Geoffrey Adde. “EOS as the present and future solution for data storage at CERN”. In: *Journal of Physics: Conference Series*. Vol. 664. 4. IOP Publishing. 2015, p. 042042.

[10] Patrick Fuhrmann and Volker Gülzow. “dCache, storage system for the future”. In: *Euro-Par 2006 Parallel Processing: 12th International Euro-Par Conference, Dresden, Germany, August 28–September 1, 2006. Proceedings 12*. Springer. 2006, pp. 1106–1113.

[11] Brian Bockelman et al. “Bootstrapping a new LHC data transfer ecosystem”. In: *EPJ Web of Conferences*. Vol. 214. EDP Sciences. 2019, p. 04045.

[12] Brian Bockelman et al. “Third-party transfers in WLCG using HTTP”. In: *EPJ Web of Conferences*. Vol. 245. EDP Sciences. 2020, p. 04031.

[13] Ajay Tirumala. “Iperf: The TCP/UDP bandwidth measurement tool”. In: *http://dast. nlanr. net/Projects/Iperf/* (1999).

[14] Brian Tierney et al. “perfsonar: Instantiating a global network measurement framework”. In: *SOSP Wksp. Real Overlays and Distrib. Sys* 28 (2009).

[15] Arnar Birgisson et al. “Macaroons: Cookies with Contextual Caveats for Decentralized Authorization in the Cloud”. In: Jan. 2014. ISBN: 1-891562-35-5. DOI: 10.14722/ndss.2014.23212.

PoS(ISGC&HEPiX2023)006