The DiFX Correlator for the RAEGE Network

Javier González-García[1], Pablo de Vicente-Abad[1], Abel García-Castellano[1,2], José Manuel Blanco[3], Eduardo Alonso-Peleato[3]

Abstract For the RAEGE network to operate autonomously, the ERDF project YNART (Infrastructures for updating the radio telescopes at Yebes Observatory) has provided a scalable and flexible software VLBI correlator based on a High Performance Cluster. The initial design is intended for the correlation of VGOS data from four stations with a weekly observation cadence, but the design allows for the easy addition of more processing nodes to handle larger sessions in terms of the number of baselines. This correlator is now complete and operational. Although its main focus is on the RAEGE network, some available time can be offered to the geodetic VLBI community. Here we present the correlator, its architecture, and the hardware used, together with the tests carried out to check its correct functioning (real VLBI sessions) and benchmark analysis. We have tested different working scenarios as a function of the number of threads and explored the execution time when increasing the number of stations from three to ten.

Keywords correlation, DiFX, RAEGE, HPC

1 Introduction

RAEGYEB and RAEGSMAR, the first two stations of the RAEGE project, are already contributing to VGOS observations. The third station in the network, located in the Canary Islands, is about to become a reality. We present here the software correlator, installed at Yebes Observatory (Figure 1), which will allow the processing of observations done by the network. Since 2021, we have been operating a small cluster to train staff in the correlation process using DiFX and post-processing tools for geodetic applications such as HOPS, nuSOLVE, CALC/SOLVE, etc. During this period we processed about 100 experiments from various projects, including EU-VGOS and Intensive sessions, most of them involving three stations. This allowed us to dimension the new HPC so that we could estimate an observation time per correlation time factor of less than 1.

1. Yebes Observatory, Instituto Geográfico Nacional, Spain
2. Santa María RAEGE station, Azores, Portugal
3. Quasar Science Resources S.L., Spain

2 Hardware Description

The original design aimed to sustain the correlation load from a four-station network observing in full-rate VGOS mode (32 Gbps at each station) within a processing rate of 1 or less. Thus each observing second would produce 128 Gb of data.

The original design envisages three independent storage volumes (cdata1, cdata2, and cdata3). Each storage volume is composed of eight storage nodes, each one contributing with a couple of RAID-5 groups of six hard disk drives each, individually adding 10 GB of storage to the volume. Considering the space required to store the parity, the net available storage space in each volume is 728 TB, for a total of 2.1 PB. Both the technology and the size were chosen to reduce the cost per byte at the cost of lower I/O speed compared to other massive storage technologies. However, a good distributed file system allows parallel access to data, which reduces the significance of this penalty.

Fig. 1 Side view of the new HPC for DiFX correlation installed at Yebes Observatory.

Like many other correlators, we have selected BeeGFS, a parallel distributed file system that has proven good performance in HPC environments. This file system requires, in addition to the storage equipment itself, the existence of so-called metadata servers, functioning to provide the physical location of the chunks into which the data has been divided.

For the main computational tasks, a single chassis system with four independent nodes, each equipped with double Intel Xeon IceLake 4314 CPU, has been incorporated into the cluster. This CPU contains 16 cores for a total of 128. The processors will take the majority of the computational tasks (core threads in the DiFX terminology). However, the storage units can also be used for this role to increase the computational power to a maximum of 896, shared with the tasks required to read and provide the raw data from the stations.

In such an HPC based on independent nodes that do not share a common bus, the network performance is critical. For this reason, low-latency networks are a typical choice in VLBI correlators, Infiniband being a common choice. We have selected a combination of 200 and 100 Gbps technologies. Three networks with different technologies and purposes are used in our correlator. The remote stations transfer their data via the so-called frontend nodes or receiving nodes. They run the e-transfer tools, currently jive5ab and etd. The connection to the outside is via the Spanish NREN RedIRIS through a dedicated 100 GbE connection that connects the Observatory to a node of the RedIRIS backbone. Only the front nodes are exposed to the outside; from there inwards, we find a spine-leaf architecture based on three Mellanox IB switches, where one of them acts as a spine and two others as a leaf, providing a connection with fewer and more predictable hops between all the devices compared to a multilayer architecture. Within this network, there is a segment of IB HDR 200 Gbps which is used for connections between switches, as well as between a switch and those machines where the expected traffic will be one-to-many. This is the case for metadata nodes and frontend nodes, which need to communicate with each of the storage nodes to manage raw data traffic both in write and read operations, as well as with the compute nodes when reading level 0 data is required. The other segment uses the slightly cheaper IB EDR 100 Gbps technology to connect the multiple storage nodes and compute nodes. It would be too expensive to use IB 200 Gbps on all nodes, and for cost savings, it was decided to use this technology only on the metadata nodes that have to serve a multitude of requests for access to the Level 0 data stored on the storage nodes. The third network is the management network. This is a fully private 1GbE network.

In addition to the systems involved in the correlator itself, it was necessary to install a separate power supply and air conditioning system.

3 Software

The entire HPC was designed and built to run the DiFX correlation software, and for this reason no HPC manager was installed, although the possibility of including one in the future is being considered. The DiFX 2.5.4 version was agreed upon de facto for the correlation of VGOS-type experiments, and a trunk version is also maintained. As this is a correlator for geodetic VLBI, the post-processing tools are the usual ones for the geodetic VLBI community: HOPS (v3.24 rev 3757) and nuSolve 0.8.2. A PolConvert (v1.7.8) installation is also available. Two tools are used to manage e-transfers: jive5ab and etransfer (etc/etd). These tools run as a service on the frontend nodes, accepting requests on ports 2620 to 2629 (jive5ab) and 4000 to 4009 (etd) from networks supported by the inbound firewall. ETD allows the server to specify on which port it wants to receive the data, which is very handy because just by specifying the command port, a given station is already assigned a UDP port. In the case of jive5ab it is not possible for the service to control the data port, so the allocation of free UDP ports can be a bit more chaotic.

4 Benchmarking and Experience

The installation of the correlator was completed in November 2023 with the installation of the final software. A series of test runs was carried out in November and December to get an idea of the performance achieved. The tests were performed in isolation on each subsystem. That is, for storage, we used a BeeGFS utility that allows us to measure the write speed locally and independently of the network, as each process is executed with data generated on each storage node. We configured this test to write 200 GB of data to a file. The average write speed found was 9 Gbps on each of the RAID5 groups that form a storage pair on a node. For HDD disks with a SATA3 interface, the expected write speed can be around 3 Gbps, so with groups of six disks, the penalty for mounting RAID5 versus RAID0 is noticeable. However, given a volume of eight nodes per station, this is more than sufficient to maintain a correlation ratio below 1 for VGOS observations at 32 Gbps with three participating stations. Network tests are performed with Infiniband's own utility, obtaining speeds of 197 Gbps and 96 Gbps for the IB HDR200 and IB EDR 100 networks respectively, thus close to the theoretical limit. In the real application, the writing/reading of data is done through the network. We have therefore carried out a test bench to see the system performance under normal operation, isolating the data input from the station. The test bench consists of transferring files of ascending size starting from 1,024 MB and increasing in the power of 2. For this, we use the *dd* POSIX command with the dummy input in the following way:

```
dd if=/dev/zero of="$file_name"
bs=256M count=$cuentas
conv=fdatasync
```

We have compared the results with an alternative method using *fallocate* and *cp* with similar results, thus giving confidence in the measurements. The results can be seen in Figure 2. The writing performance of up to 20 GB is difficult to measure, but we could determine that writing speeds for typical VGOS files have an optimal throughput better than 20 Gbps. This is far beyond the current e-transfer capabilities of the stations.

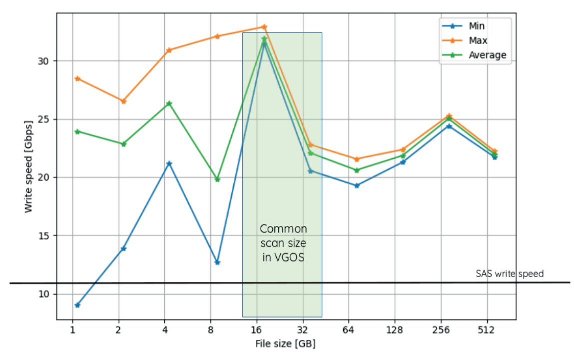

Fig. 2 Storage speed benchmark.

The Yebes Observatory began in 2021 with a small cluster built with surplus machines in which DiFX was installed to gain experience and, as a prototype, to size a cluster built ad-hoc. Intensive sessions have been correlated there since 2022, and since January 2024 we have migrated all operations to the new correlator. We therefore have some reference values to which to compare the performance of the new cluster. The table in Figure 3 shows the prototype's specifications.

Hostname	Product Name	CPU	No. Cores	No. Sockets	Hyper threading	Installed RAM
alnilam	HPE ProLiant Gen 10	Intel Xeon Silver 4110 @ 2.10 GHz	8	1	Yes	128 GB
alnitak	HPE ProLiant Gen 9	Intel Xeon(R) CPU E5-2620 v4 @ 2.10GHz	8	1	Yes	16 GB
mintaka	Supermicro XL	Intel(R) Xeon(R) CPU E5-2630 v2 @ 2.60GHz	6	2	Yes	32 GB

Other considerations:
- HPC nodes are connected via 10 GbE network.
- Datastreams reads from local storage in Supermicro XL server (mintaka).
- DiFX 2.5.4 is installed in NFS shared file system.

Fig. 3 HPC prototype specifications summary.

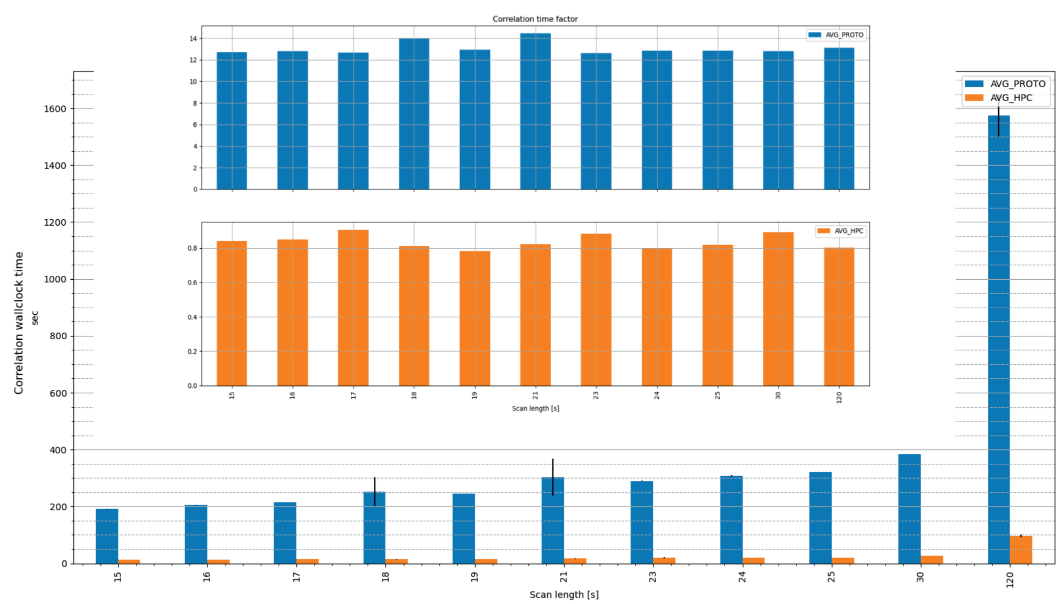

Fig. 4 Two bar plots showing performance comparison for ten Intensive-like sessions. Orange bars correspond to the new HPC, whereas blue bars correspond to the prototype.

Between December 2022 and December 2023, 68 VGOS-Intensive sessions were correlated with this prototype. Approximately half of these were standard Intensives, with the usual frequency configuration introduced by [2], a duration of one hour and scan times between 15 and 30 seconds. The other half consisted of ten-minute observations with three scans of durations between 120 and 300 seconds. The SWIN files of these observations can be found at the IVS Data Centers with prefix codes y22+DOY, y2+DOY+a, y23+DOY, and y3+DOY+a. We have repeated a reduced number of these sessions in the new correlator for comparison, with very satisfactory results shown in Figure 4.

We generated dummy VGOS data recorded at 8 Gbps to stress the system by increasing the number of antennas from three to ten (Figure 5). The correlation time scales linearly, instead of quadratically, as the number of baselines increases; so, we conclude that for data at 8 Gbps, HPC resource exhaustion is not reached at a number of ten antennas or less. This is more than sufficient to meet the objectives of the RAEGE project, where only four antennas are envisaged.

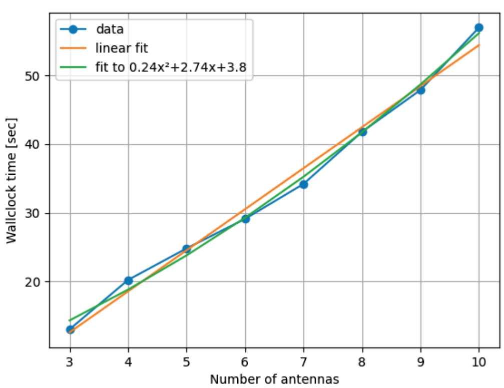

Fig. 5 Correlation time for a network of three to ten VGOS stations observing at 8 Gbps.

Using the same dataset we also explored the parametric space of thread and cores involved in correlation in terms of wall clock time, i.e., the time experienced by the operator since the start command is issued until the last job is finished. This value can differ from the CPU times. The full HPC can use up to 224 cores, although there is not an even distribution between nodes. A small subset is equipped with 32 cores in a single CPU, whereas most of the machines have only 16 (storage nodes). This explains the discontinuity in the horizontal axis (threads) visible in the colormap.

Since January 2024 all the operations have been done with the new HPC. So far we have processed 17 sessions corresponding to the IVS Intensive calendar in 2024. Sessions with three stations take about 30 minutes to be correlated (no post-processing) and four station sessions a bit longer, 40 minutes.

5 Conclusions

We have acquired the expertise to handle the correlation and post-processing software used for the VGOS observations. We have designed, built, and validated a cluster for the correlation of RAEGE network experiments, processing 17 sessions that correspond to the IVS Intensive calendar of 2024. Although the correlator load is currently low, the lack of correlator operators prevents us from taking on more workload. However, in the future, we plan to add more staff in order to be able to offer correlator time for other IVS master schedule experiments.

Acknowledgements

The RAEGE correlator was funded by ERDF (European Regional Development Funds) project YNART.

References

1. A. T. Deller, S. J. Tingay, M. Bailes, and C. West, "DiFX: A Software Correlator for Very Long Baseline Interferometry Using Multiprocessor Computing Environments", Astronomical Society of the Pacific, https://www.jstor.org/stable/10.1086/513572, pp. 318–336, 2007.
2. A. Niell, J. Barrett, A. Burns, R. Cappallo, B. Corey, M. Derome, et al., "Demonstration of a broadband very long baseline interferometer system: A new instrument for high-precision space geodesy." Radio Science, 53, 1269–1291, 2018. https://doi.org/10.1029/2018RS006617