Big-Data-Oriented Computing on a Mid-Size Data Center: the ADHOC Infrastructure

Augusto Tortora
Department of Physics "E. Pancini"
University of Naples "Federico II"
Naples, Italy
email: augusto.tortora@unina.it

Massimo Brescia
Department of Physics "E. Pancini"
University of Naples "Federico II"
Naples, Italy
email: massimo.brescia@unina.it

Stefania Conte
Department of Physics "E. Pancini"
University of Naples "Federico II"
Naples, Italy
email: stefania.conte@unina.it

Paolino Guida
Naples section
National Institute for Nuclear Physics
Naples, Italy
email: pguida@na.infn.it

Guido Russo
Department of Physics "E. Pancini"
University of Naples "Federico II"
Naples, Italy
email: guido.russo@unina.it

Bernardino Spisso
Naples section
National Institute for Nuclear Physics
Naples, Italy
email: spisso@na.infn.it

***Abstract*—This work presents the implementation of a mid-size Data Center devoted to Physics data analysis, realized in 2023-24 at the University of Naples, as the second upgrade to a previous Data Center built on 2007 during SCoPE project. Data Center had a first upgrade, during RECAS project, in 2014. After PNRR projects (STILES, CN1), the Data Center handles about 100 servers, 10.000 cores, 100 GPUs, 10 Pbytes of storage, and is used for big data analysis in astrophysics, high energy physics, material science and theoretical physics. The purpose of this work is to illustrate and demonstrate how it is possible to effectively and efficiently perform physical and astrophysical analyses in the mid-size Data Center, using Big Data, on unstructured datasets.**

Keywords*—*Data Center; Scientific Computing; Big Data; GPUs.

I. INTRODUCTION

Big-data analysis is becoming a common practice for many areas of research. So far, the handling of such data has been restricted to large Data Centers, however there is now a need for a larger distribution of data. This requirement comes from the request to have big data sets local to the Data Center, provided it is at least a mini- or mid-size Data Center.

The Physics University Department is organized in several research areas, and most of them have their own scientific data set to preserve and to analyze. In this paper we present the solutions adopted for the implementation of a mid-size Data Center, ready for big data analysis. In 2007, a previous Data Center was created from scratch (SCoPE project), and upgraded in 2014 (RECAS project [2]). Due the its nowadays obsolescence, a big transformation is underway, and it is here described.

The Data Center (Figure 1) is now operational, with 10 Pbytes of data, 80 servers with a total of 200 GPUs, about 15.000 cores, and is accommodating all different research needs of the Physics Department.

The main application area of the Data Center in the next 3 years will be in astrophysics, we named the center as Astrophysical Data HPC Operating Center (AD HOC).

Fig. 1. The Data Center at Dept. of Physics

The aim of this work is to demonstrate that for large unstructured datasets, it is possible to leverage a mid-size Data Center for astrophysical analyses. Section II outlines the requirements to achieve the set objectives; Section III presents the implementation of IT systems and Power and Cooling of the Data Center; Section IV showcases examples of applications in the field of astrophysics, and in Section V, conclusions and future developments are outlined.

II. REQUIREMENTS

The requirements for the Data Center come from the data types the users have to work with. A first class of users is the astrophysical scientists, which collects and analyzes images from ground based telescopes (ESO in Chile, GranTeCan at Canary Islands), space telescopes (e.g., Hubble, Web), and radio telescopes. These images are typically quite large, say 1 to 4 millions pixels each, and have to be processed with AI tools [6]. AI will recognize the various objects in the files, beside an image analysis packages can be used to compute physical properties, after proper calibration.

Another range of users is in the material properties physicists, who want -as an example- develop and test alternative jet fuels by introducing a state-of-the-art mathematical model that enables accurate estimations of the fuel consumption in jet engines [9]. These kind of computation requires the adoption of models with a lot of variables, generating thousands of tests which have to be stored and compared each other.

A third type of users is the high energy physicists. In a typical experiment (ATLAS at CERN, BelleII at KEK), the data analysis can be split in two steps: first, the sample of ”interesting” events has to be selected; then, the relevant quantities can be compared to the theory by studying the overall features of the sample have to be extracted. The data sets are of the order of several dozen of Terabytes [8].

Most data sets, which were described above, must be processed primarily sequentially and are read from disk sequentially: images are read row by row, data acquired by sensors have to be processed in order of acquisition. This means that slow access disks can be used for storage with a good caching mechanism implemented. The data have to be moved rapidly, which means that high speed networking is mandatory. Eventually, data have to be processed in parallel, which means that we need a parallel file system. That part of the processing has to be done on GPUs, and/or exploiting the parallelism implemented with OpenMP.

These are the main requirements for implementation in the Physics Department Data Center. Among the requirements, the limits related to the budget available in Italian universities must also be considered.

III. THE IMPLEMENTATION

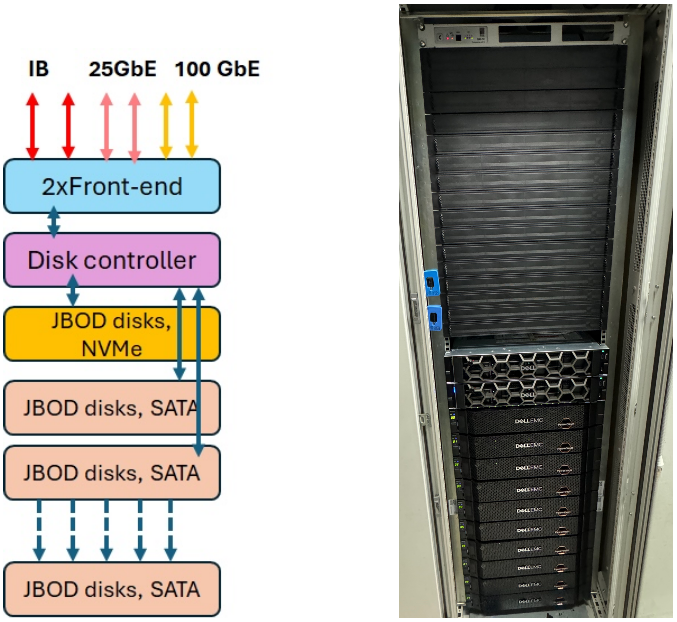
The requirements, as shown in the previous paragraph, had a significant impact on the implementation of the ADHOC Data Center. In this section, we will provide a description of the entire IT infrastructure (Computing, Network, and Storage), the Power and Cooling systems, the security criteria adopted for users and professionals, and the control room where the entire Data Center is monitored.

A. Storage

The storage capacity needed is achieved with 500 SATA disks, 22 TB each, and 100 NVMe disks, 7.5 TB each (raw capacity). This bulk disk capacity is organized in groups of 12 enclosures, with several of these groups connected to a single Fibre Channel Arbitrated Loop (FC-AL) controller. The NVME disks act as a caching area to speed up access. Two server every 2.5 PBytes act as a front-end for the users, each server is connected to the controller via FC-AL. The server itself is accessible via multiple networking options: 2x25 GbE, 2x100 GbE, 2x200 Infiniband. All the interfaces are duplicated towards different switches guaranteeing robustness and fail safe to the storage infrastructure. The file system is Lustre via Infiniband. In Figure 2 are shown the functional scheme of the described system (Figure 2a) and the real hardware involved in storage (Figure 2b).

B. Networking

The network in the Data Center is based on both Ethernet and Infiniband switches, interconnected as in Figure 3. This guarantees that no single point of failure exists in the network, indeed there is always a path from source to destination, even if a switch (or a board in the switch) fails. All apparatus are enterprise level with multiple power supplies. While the 25 GbE is used for user access, file transfer uses the 100 GbE network with an automatic fail-over from one to another. The InfiniBand is HDR, 200 GB/s is used mainly for OpenMP parallel jobs, and for the Lustre parallel file system.



(a) Storage Scheme (b) Storage in real

Fig. 2. Architecture of the storage subsystem (replicated each 2.5 PBytes)

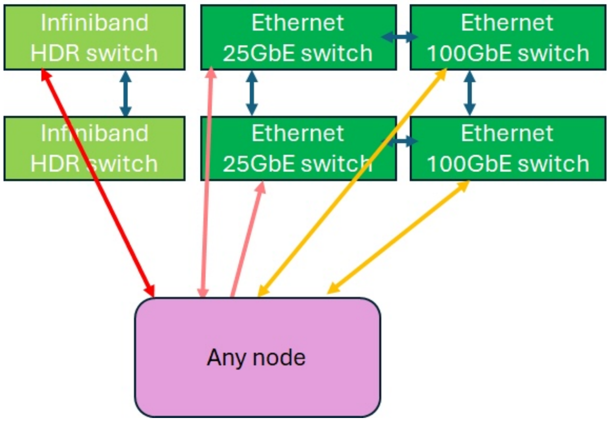


Fig. 3. Architecture of the network in the data center

The 25 GbE and 100 GbE networks act in parallel, and the 25 GbE network is also used as a backup network.

C. Computation

In the Data Center, there are two kind of nodes for computational tasks: HTC nodes with many cores (2x128 cores) and HPC nodes with a pair of nVidia GPUs. The HPC nodes are of two types: type A uses 2x84 cores, 3 Tbyte RAM, 2x H100 GPUs; type B uses 2x32 cores, 1 Tbyte of RAM each, 2x nVidia L40 GPUs. All type A nodes have 3 Tbyte of RAM each, as a specific request from most users, in particular theoretical physicists. The Figure 4 shows the three types of computational nodes in the Data Center.

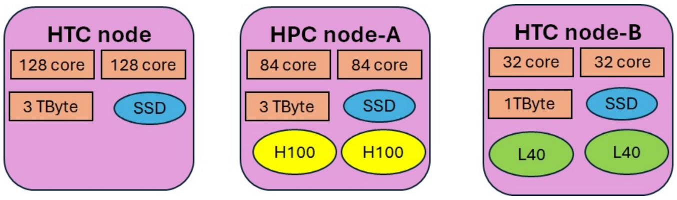


Fig. 4. The three types of computational nodes

A third type of computational node, HPC type C, was acquired and is used since 2022, and is based on a server with 0.77 Tbyte RAM and 4x nVidia V100 with nVlink interconnection. There are 32 nodes, for a total of 128 GPUs, with EDR 100 Gb/S Infiniband, which all together form a cluster for general purpose applications open also to students.

D. Security

All Data Centers are subject to potential hacker attack, and so is our. We have put in operation two firewalls, with strict access rules, using public IP addresses only for the User Interfaces servers. All accesses are logged, with disk space for 4 years of the operation before the circular buffer starts to be overwritten. Moreover, the physical access is strictly controlled via fingerprint and NFC smartcard, according to our ISO 9001 certification, again with a very long log history. Furthermore, 10 IP cameras are present with a NVR to check anomalous accesses.

E. Power and Cooling

Power and cooling issues are becoming very important in Data Centers, as they must ensure continuous up-time, with minimal interruptions, as well as reduced energy costs. The PUE (Power Usage Effectiveness) parameter is used to measure the energy efficiency of a Data Center. It is defined as the ratio between the total absorbed power (PT) and the power used by the IT equipment alone, and is therefore greater than one [7]. The lower it is, the better. Over the years, work has been done to reduce the PUE value, which in 2022 was 1.5, now reaches 1.3, which can still be reduced with further actions. Among the actions undertaken, one strategy is the adoption of free cooling systems, which is not very effective at the latitudes where the Data Center is located. Furthermore, it should be noted that new server hardware architectures allow the possibility of increasing the temperature of the water leaving the chillers. Another strategy to lower the PUE is the use of integrated monitoring systems, which collect the operating parameters of the Data Center services (IT, Power and Cooling, network) and allow optimal management of resources. As regards IT resources, the monitoring system takes into account the percentage of use of individual machines and the number of active processes. Energy data for IT systems is collected for each machine using Rittal's metered PSM (Power System Module) devices, which transfer the data via CMC (Computer Multi Control) controllers in each rack. It should be noted that all the Data Center equipment is powered by two 420KW UPS groups located in the dedicated 1MW electrical transform room. The two UPS units perform two functions: i) stabilize the electrical supply to the IT infrastructure; ii) provide energy to infrastructures in the event of a power failure from the electricity grid.

Proprietary software is used to monitor the UPS (Figure 5 shows the parameter collected), while the consumption data is compared with that reported by the general monitoring system.

Fig. 5. The UPS Monitoring Interface

As regards cooling, the CMC units present in each Rack control and manage the water/air heat exchangers inside the Liquid Cooling Package (LCP shown in Figure 6) columns located on the sides of each cabinet. The thermal power of each column is 20KW, guaranteeing a maximum temperature of 30 °C inside the Rack under full load conditions.

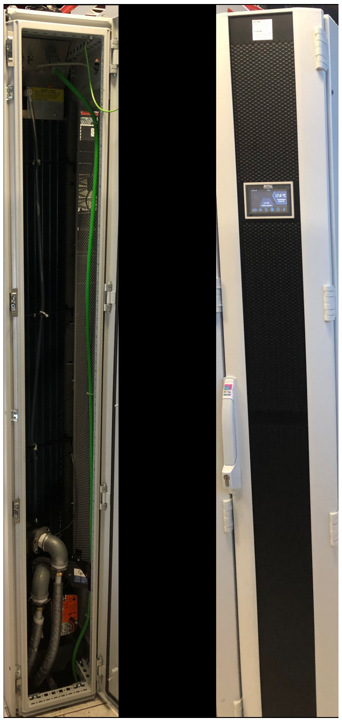


Fig. 6. Liquid Cooling Package - LCP

Water is supplied to the LCPs at a temperature of 10°C from the cooling system, consisting of two circuits. The first circuit cools the water returning from the Data Center through 2 parallel Chillers, each with a thermal capacity of 400KW. Each Chiller contains 2 independent scroll compressors, enabling modular operation based on the thermal load to be met. The second circuit starts from the inertial tank, regulates water pressure through 4 pumps equipped with inverters, and sends the cold fluid to the LCPs. In Figure 7 is possible to see the realization of external cooling plant.

Fig. 7. Data Center Hydraulic Plant

The Chillers, pumps equipped with inverters, and sensors on the system send energy consumption data and notifications of any faults to the monitoring system, which is based on a commercial product (by Siemens, Figure 8). Another commercial product (RiZone by RITTAL, Figure 9) collects data from LCPs and PDU's in each rack, both systems are integrated with a custom interface based on LibreNMS. Figure 10 represents an example screen that extracts data from IT resources and power and cooling systems.

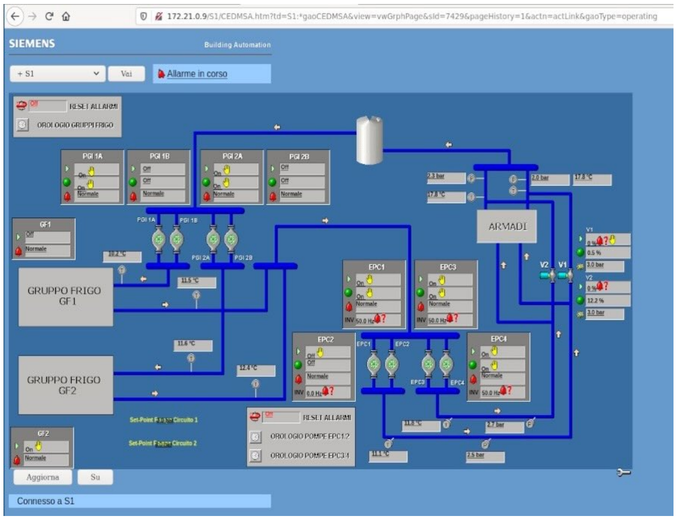
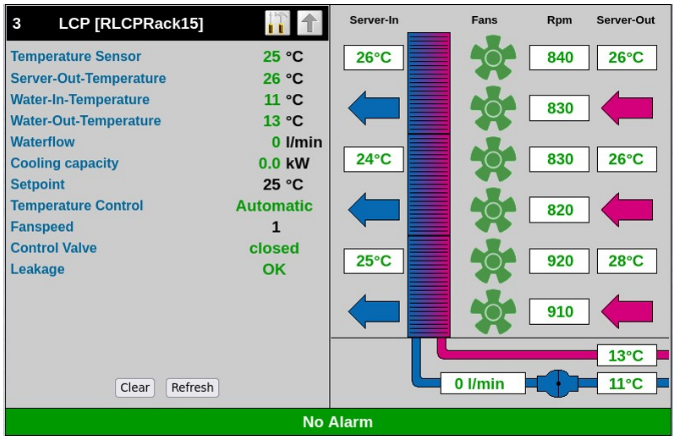


Fig. 8. The Siemens Monitoring for Pumps

F. The Control Room

The whole data Center is unattended, but operators work from a separate room, a Control Room (Figure 11) with a complete direct access to all resources. In this space expert operators perform training for young technicians and students of our University.



3 LCP [RLCPRack15]

Temperature Sensor 25 °C
Server-Out-Temperature 26 °C
Water-In-Temperature 11 °C
Water-Out-Temperature 13 °C
Waterflow 0 l/min
Cooling capacity 0.0 kW
Setpoint 25 °C
Temperature Control Automatic
Fanspeed 1
Control Valve closed
Leakage OK

Fig. 9. The Rittal Monitoring for LCP's

Fig. 10. The Data Center Monitoring Screen

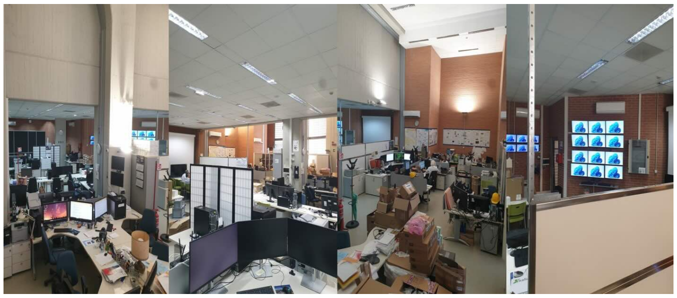


Fig. 11. The Data Center Control Room

IV. EXAMPLES OF APPLICATIONS

In this section, some of the applications of Big Data analytics taking place in the Physics Department's Data Center are showcased.

One of the astrophysical use cases foreseen to be supported by the Astrophysical Data HPC Operating Center (ADHOC) infrastructure is connected to the Italian participation to the V. Rubin telescope and its related LSST (Legacy Survey of Space and Time) data survey [1]. Rubin is an 8m class telescope characterized by a wide field of view and a high resolution, designed to carry out a survey of the so-called dynamic universe, mapping the entire sky visible in the southern hemisphere every few nights. This implies that approximately 20 TB of data will be collected each night. This poses extremely challenging aspects also from the technological point of view for the management and exploitation of the collected data. The camera will observe in six filters covering wavelengths from 320 to 1050 nm. The survey project foresees that about 90% of the observing time is dedicated to uniformly observe about 800 times 18000 deg2 of the sky (adding up on all 6 bands), during 10 years and will produce a co-added map up to magnitude 27.5 in r band. In Figure 12 is possible to see a 10-years map tiled over the entire southern sky by Rubin telescope and its related LSST.

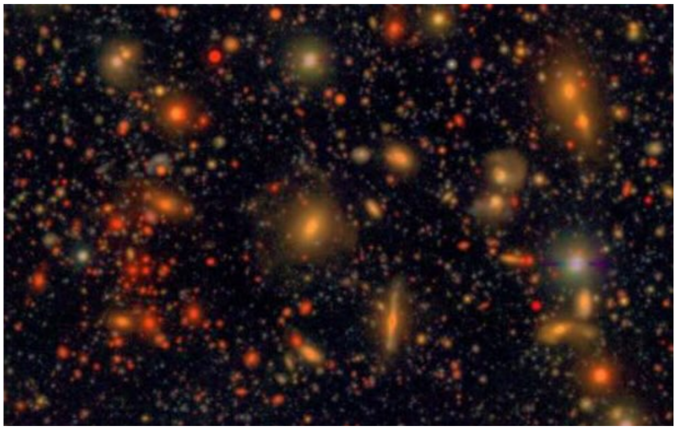


Fig. 12. LSST 10-Year Map Tiled over the Entire Southern Sky

These data will result in a database (of about 300 Petabytes) that will include about 37 billion observations of 20 billion galaxies, 17 billion stars, 6 million solar system objects and will serve most of the primary science programs. The Italian participation is mostly based on the in-kind contribution program, consisting in a series of scientific contributions provided by our community and formally agreed with the project Board, which will return data rights for our scientists, making them able to exploit LSST data products in real time, i.e., without waiting for the data public releases. In particular, one of the most interesting data analysis contributed pipeline to Rubin-LSST community is the analysis of crowded stellar fields (Galactic Bulge, Globulars, Magellanic Clouds) through specific software tools like the PSF Daophot/Allframe suite. This software is extremely computing demanding and processing time consuming, requiring parallel computing paradigms as well as machine and deep learning methods aimed at characterizing stellar clusters and homogeneously deriving key parameters e.g., age, distance, reddening, metallicity, etc.) for globular and open clusters in the entire survey footprint. Another crucial example of astrophysical use case is the support to the national research for two of the most important astrophysical projects of the next decade, ELT and SKA. ELT [11] [4] is an optical/nearIR telescope with a 39m primary dish, the largest of its kind ever built or planned. ELT is being built by ESO, and will be located atop Cerro Armazones, a 3000m peak in the Chilean desert. ELT is designed to exploit the full power of Adaptive Optics that removes atmospheric disturbances so as to reach the full resolution obtainable from the mirror and becoming able to produce data 5 times sharper and deeper than even JWST will do from space. The SKA Observatory (SKAO [12]) will comprise two radio interferometers. The low frequency antenna array (SKA-Low) will reside in Western Australia and the mid-frequency dish array (SKA-Mid) will be hosted in South Africa's Karoo region. SKA-Low will be made of 512 stations, each hosting 256 dual-polarized Log-Periodic antennas, distributed over an area of 65 km in diameter and operating over the 50-350 MHz range. SKA-Mid will comprise 197 dishes distributed over a region of 150 km in diameter, operating at frequencies from 350 MHz to 15 GHz, and will include the dishes of the precursor facility MeerKAT. Once completed, the SKA arrays will provide an order of magnitude improvement in sensitivity, one to two orders of magnitude increase in survey speed over the state of the art, as well as unprecedented image fidelity [5]. In Figure 13 there is an example of the main science drivers exploited by the SKA telescope.

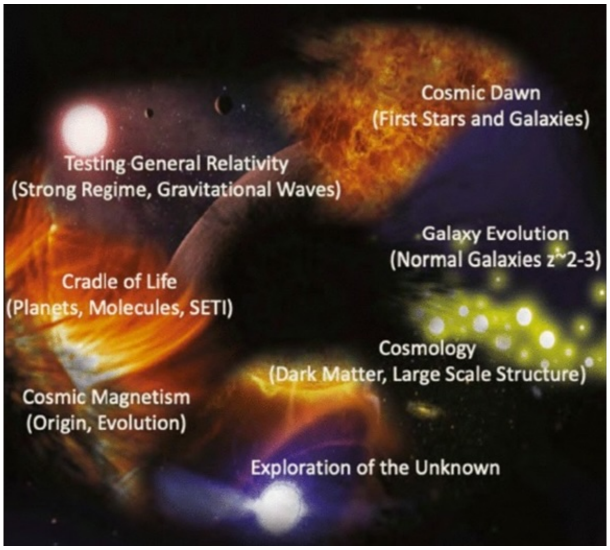


Fig. 13. Main Science Drivers Exploited by the SKA Telescope

The computing and storage support offered by our infrastructure will achieve a synergy between these two projects, since multi-frequency, multi-messenger approach is now recognized as a pillar of modern astronomy. Obvious examples of data products which will be supported by our computing infrastructure are the resource demanding multi-wavelength approach to the study of the reionization epoch, of the galaxy-AGN co-evolution, of the nature of Dark Matter, of the birth of stars and planets and ultimately the search of habitable planets. In Figure 14 is shown an example of a resolved stellar populations in a representative sample of the Universe. The ELT will offer the exciting prospect of reconstructing the formation and evolution histories of a representative sample of galaxies in the nearby Universe by studying their resolved stellar populations [ESO - ELT Science].

There is the need for combined observations at all wavelengths, crucially including data from both radio and optical telescopes. In fact, with astronomy rapidly becoming a science of Big Data, our goal here is to make available the computing

Fig. 14. Resolved Stellar Populations in a Representative Sample of the Universe.[ESO - ELT Science]

facility and the analysis tools that are required to analyse data from ELT, SKA and other observatories, in synergy with the ESO archive and the SKA Regional Centres. A main objective will be to exploit Italian expertise for innovative data-mining techniques (generically referred to as Machine Learning) that have not yet become “standard tools” for data analysis. One example is the mentioned LSST (Figure 15) survey project, for which we proposed an accurate Point Spread Function (PSF) reduction, based on a customised version of Allframe code [10], of the most crowded areas in the sky (the Galactic Bulge and Plane, the Milky Way globular clusters, the Magellanic Clouds, a few massive dwarf spheroidals).

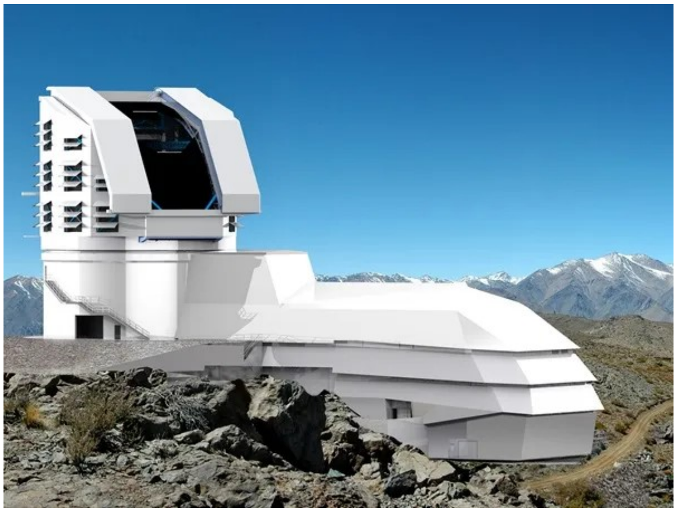


Fig. 15. A depiction of the completed LSST observatory. [LSST Project/NSF/AURA]

This kind of reduction is particularly demanding both on computing time and hardware resources. With the proposed new version of Allframe a single master list of stars can be simultaneously measured on all the available images. This translates in a more accurate and fast photometry, since the positions of the centroids and the fluxes are cross-correlated among all the images. This is basically a different approach from the standard “forced-photometry” strategy, since in the latter case the stars of the input list are measured individually on the target images. This different data mining strategy requires a HPC based computing framework, where, in particular, powerful GPUs could be used to obtain the needed high level of parallelism to perform the simultaneous photometric reduction. The Allframe reduction needs a non-linear minimization of (N x M x S) matrix, where N is the number of images, M the number of the input stars and S the size of the involved CCD images, and it creates 2 working copies for each input image. Our typical run consists of the parallel reduction of the available DECam mosaics, each of them being made of 61 CCDs. The mosaics (Figure 16) are reduced in about 15 hours, and this means that an individual 2k x 4k (S = 8 Mpixel) CCD is reduced in about 15 minutes.

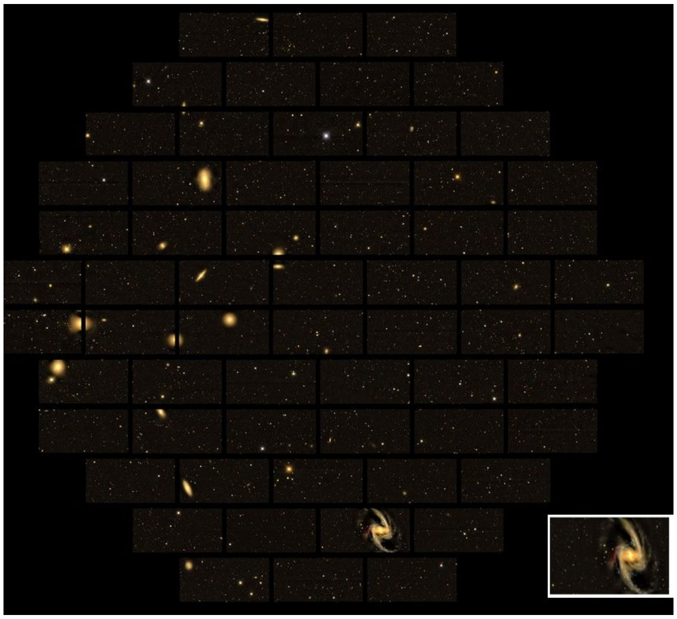


Fig. 16. An example of mosaic images with a zoom on NGC 1365. [DECam Interactive]

The typical number of stars is of the order M = 250k, while the number of images is N = 460. According to our experiments, a typical Allframe run with these numbers needs approximately 120 hours to be accomplished, and a single Allframe run needs approximately an allocated 1.5 GB RAM. The disk space used by the process is dominated by the disk space needed for the images, that is 460 x 3 images x 31MB = 42 GB. The final catalogues are of the order of 7MB for each CCD image, therefore 460 x 7MB = 3 GB for the whole set. The Rubin-LSST Camera is made of 189 4k x 4k CCDs and at the end of survey, Rubin-LSST will provide about 850 visits for each individual pointing. Therefore, we estimate that an

Allframe run over 450 individual Rubin-LSST CCD images, requiring about 480 hours of a single core computing time, 3 GB of RAM, 85 GB of disk space for the input images and their working copies and 6 GB of disk space for the permanent data products. These numbers have to be multiplied by the 189 CCDs that make the Rubin-LSST Camera. This has to be considered the maximum size of the "final" Allframe run that would be parallelized.

V. CONCLUSION AND FUTURE WORK

We have described the realization of a computing center, mainly devoted to astrophysical applications, built up on a previous system, by replacing all energy-consuming devices with new powerful yet not-so-much expensive hardware. A simplified storage architecture with 10 Pbyte capacity, and an efficient cooling system, have been realized. The Data Center is operational, and a few applications have been illustrated.

In the future, our goals are to increase the computational power, and to maintain the Data Center operations with minimal downtime.

REFERENCES

[1] Ž. Ivezić et al., "LSST: From Science Drivers to Reference Design and Anticipated Data Products", The Astrophysical Journal, Vol. 873, No. 2, pp. 111-144, doi:10.3847/1538-4357/ab042c, Marc 2019.
[2] G. Russo, G. B. Barone, G. Carlino, and G. Laccetti, The ReCaS Project Naples Infrastructure, in High Performance Scientific Computing Using Distributed Infrastructures", High Performance Scientific Computing Using Distributed Infrastructures: Results and Scientific Applications Derived from the Italian PON ReCaS Project, pp. 57-71, ISBN 978-981-4759-70-0, 2017.
[3] G. Laccetti et al., High performance scientific computing using distributed infrastructures. Results and scientific applications derived from the italian PON ReCaS project, in High Performance Scientific Computing Using Distributed Infrastructures, World Scientific Publishing Co., ISBN 978-981-4759-70-0, 2017.
[4] M. Kissler-Patig and M. Lyubenova, An Expanded View of the Universe, a high-level document outlining the E-ELT's science case, Edited by the E-ELT Project Scientist, Dec 2009, updated in 2011.
[5] R. Braun, T. L. Bourke, J. A. Green, E. F. Keane, and J. Wagg, "Advancing Astrophysics with the Square Kilometre Array", Advancing Astrophysics with the Square Kilometre Array, Giardini Naxos, Sicily, Italy, June 8-13, 2014, vol. 215, p. 174-182, published in 2015.
[6] D. Baron, "Machine learning in astronomy: A practical overview", arXiv preprint arXiv:1904.07248, 2019.
[7] M. Sharma, K. Arunachalam, D. Sharma, "Analyzing the data center efficiency by using PUE to make data centers more energy efficient by reducing the electrical consumption and exploring new strategies", Procedia Computer Science, vol. 48, pp. 142-148, 2015.
[8] R. Jones, D. Barberis, "The ATLAS computing model", Journal of Physics: Conference Series, Vol.119, No.7, p. 072020, IOP Publishing, 2008.
[9] V. Singh and S. K. Sharma, "Fuel consumption optimization in air transport: a review, classification, critique, simple meta-analysis, and future research implications" European Transport Research Review, vol. 7, no. 2, pp. 1-24, 2015.
[10] P. B. Stetson, "The Center of the Core-Cusp Globular Cluster M15: CFHT and HST Observations, ALLFRAME Reductions", Publications of the Astronomical Society of the Pacific, Volume 106, p. 250, doi:10.1086/133378, 1994.
[11] European Southern Observatory (ESO)- The Extremely Large Telescope (ELT). [Online]. Available at: https://elt.eso.org
[12] Explore SKA Observatory. [Online]. Available at: www.skatelescope.org