» INTERVIEW

CNAF AND THE NEW CHALLENGES OF THE BIG DATA AND SUPERCOMPUTING ERA

Interview with Gaetano Maron, director of CNAF, the INFN National Center for computing

CNAF, the INFN National Centre for information and communications technologies recently restarted its full operations, after the recovery operations made necessary by a serious flooding that hit its headquarters in Bologna last November. CNAF hosts the TIER1 centre for the LHC experiments and is a point of reference for many other experiments in which the INFN is involved. It deals with the research and development of innovative digital technologies for applications in different scientific disciplines and it is one of the most important centres of distributed computing in Italy. We spoke about CNAF's present and future projects, in the era of big data and supercomputing, with its director Gaetano Maron.

Starting a few weeks ago, a dedicated ultra-fast fibre link now connects the CNAF LHC-Tier1 centre to CINECA.

Last October, INFN signed a collaboration agreement with CINECA establishing the exclusive use of a part of CINECA's computing resources by both our theoretical and experimental researchers. This agreement has directly affected CNAF which maintains an interest to have sufficient computational processing power, albeit retaining in its LHC-Tier1 computing centre a relevant part of the data, of both LHC's experiments and the more significant experiments on astroparticle and nuclear physics in which INFN is involved. The agreement provides CNAF with approximately 15,000 CPU cores (corresponding to approximately 50% of the computing power now present at CNAF's LHC-Tier1) which help ensure an adequate processing capacity for at least the next two years. To provide access to the data of LHC-Tier1 to the computers available at CINECA, located about 10 km away from CNAF, we configured, in collaboration with GARR, a dedicated ultra fast fibre link with a capacity up to 1.2

» INTERVIEW

Terabits per second. The distance is such that the latency, i.e. the time it takes a bit of information to travel from one centre to another, is minimal, enabling the two centres to be considered each other's extension.
This expansion opportunity was exploited during the recovery operations of CNAF's data centre, after the flooding occurred in November last year when, due to the breakage of a pipe in Bologna's water mains, more than 500 cubic meters of water poured into the LHC-Tier1 rooms. The Centre suffered considerable damage both to the electrical infrastructure, and to the computing resources that were affected by the flooding for about 10%. LHC-Tier1 was in complete activity lock down for about three months, resuming gradually starting from February, and then returning to full speed in March. Once the minimum operations of the centre were restored, thanks to the link with CINECA, it was possible to provide CNAF users with a significant computing power that mitigated the lack of resources and the inconvenience caused by recovery works of the centre's complete functionality.

CNAF derives its name from the acronym of the name it had when it was established: national centre for the analysis of frames. CNAF has a long history, it was INFN's first National Centre
If there is a common denominator in the long history of CNAF linking the different activities carried out over the years, it is the information technology revolution.
CNAF was in fact born in 1962 as national centre for the analysis of frames, with the purpose of providing a central service for the digitisation and analysis of frames coming from the experiments that used bubble chamber detectors: in a nutshell, a refined and innovative scanner of yore, associated with a sophisticated digitized trace analysis system. With the evolution of the detection techniques of particles produced in high energy experiments, CNAF's role as a frame analysis centre waned, and in the '80s the centre's purpose was shifted, still on the wave of the information revolution, to another fundamental paradigm for INFN and for Italian research: communication networks. Up until 2000, CNAF first developed and then managed INFN's network that, in addition to connecting all INFN Divisions and National Laboratories, soon became a reference point for all research bodies, gradually developing into Italy's research network, now managed by GARR.
During the Noughties, CNAF completed its most recent transformation, made possible by its great experience in developing and integrating systems to support experimental physics, but also and above all by its will to embrace the best in research computing. CNAF was in fact given the task to design and manage the LHC-Tier1 computing centre to be dedicated first and foremost to LHC experiments, but it soon became a reference for most experiments in which INFN took part. At the same time, aware that a computing centre without research and development had little in

» INTERVIEW

terms of outlook, CNAF started multiple innovative activities in the field of geographically distributed systems, of primary importance for LHC computing , and soon becoming one of the main players for the development of the World wide LHC Computing Grid (WLCG) and the Italian grid infrastructure, still primary components of LHC computing today.

We know CNAF today also because it is an LHC-Tier1 centre, but CNAF also manages data of many other international experiments, as mentioned before

Yes, CNAF is not just a LHC-Tier1 centre, but it is INFN's computing centre dedicated to all the major experiments in which INFN is involved. LHC computing uses approximately 75% of the centre's computational storage resources, while the remaining 25% is dedicated to other experiments in high energy and in astroparticle physics. At the moment, the computing resources of our Tier1 are being used by more than 30 non-LHC related experiments, ranging from the LIGO and VIRGO interferometers for the search for gravitational waves, to detectors in space, such as the AMS-02 on the International Space Station to study the universe and its origins, to experiments in the underground INFN Gran Sasso Laboratories, such as DarkSide for the search for dark matter. Currently the quantity of data produced by all these experiments and stored in CNAF discs and magnetic tapes is over 60 Petabytes: this is certainly big data, in the current meaning of the word. The analysis of these data is done by the Tier1 internal processing systems, by those from other Tier1 within the WLCG collaboration that are linked to our centre with dedicated communication networks, but, as we have seen, also by computing centres that are external but integrated into our local network: like, indeed, CINECA, the RECAS centre in Bari and commercial cloud providers. Interesting from this last point of view is the pilot that we are conducting for the European H2020 project called Helix Nebula Science Cloud (HNSciCloud), where two independent providers supply WLCG experiments with considerable computing power.
But it is not only the amount of data that should attract attention: these are primarily invaluable and sometimes unique assets, not only to the experiments that produced them but also to INFN and, overall, to research. These are hence valuable assets that must be preserved over time and ensured to withstand technological change, and to be accessible even in the years ahead. For this reason, at CNAF we have a program, in collaboration with the other centres, of Long Time Data Preservation that develops techniques and tools to ensure the accessibility of the data stored in the centre over time.

What are you doing at CNAF nowadays?

CNAF is based on four main tasks that also represent the strategic pillars of its mission: scientific computing dedicated to the support of INFN research activities; innovation and development; IT services

» INTERVIEW

of general interest for INFN, including all administrative services; technology transfer to the public and private sectors.

A centre such as CNAF would quickly lose its drive, if next to the management of scientific computing of experiments there was not an important set of innovation and development activities, which would open and experiment new paths and instruments to be then proposed in the daily management of the centre but, also, as a technology transfer to the industrial world, to corporations, etc.. Approximately 30% of human resources at CNAF are dedicated to these innovation activities, which are focused on two main lines: the development of new geographically distributed systems that are set in the European strategy for an Open Science Cloud (for example, the INDIGO Data Cloud H2020 project coordinated by INFN, but also the latest Extreme Data Cloud, European Open Science Cloud Hub, Deep Hybrid Data Cloud, etc.) thus continuing the decade long tradition of development and integration that began with the GRID projects; and the software development to support external projects at the centre, and the internal ones, but also the INFN experiments requiring on-line or off-line expert advice for their software (for example the LHC experiments, but also KM3Net, Euclid etc.).

Along these two strategic pillars – scientific computing, and innovation and development – there has been a increasingly important transfer activity of our scientific and technological developments toward open society, public administration, but also toward our manufacturing and services productive system, at both regional and national levels, and in Europe too. To optimize, but above all, to propose more heterogeneous skills deriving from this activity, we established a technological transfer laboratory called TTLab, with the INFN Divisions of Bologna and Ferrara. The objective is to promote the transfer of technology in physics, information technology, mechanics and electronics to companies located in the region. TTLab is accredited in the Emilia Romagna Region as a laboratory for industrial research and, as such, it can take part, together with companies form the region, in regional funding initiatives, which aim to encourage this process of exchange and integration between industry and research organizations (for example, the POR-ERDF projects).

Finally, it should be noted that CNAF also has a part in the supply of information technology services aimed at the administration management of INFN (accounting, management, documents, etc.), and a vast collection of public utility services that allow our institution to operate better (by supporting to research activities, websites, calendars, notes etc.).

What are the next challenges for CNAF, next to supercomputing and big data?

Data management plays a central role in our activities because the data of the experiments should certainly be effectively accessible to researchers from all over the world, but at the same time they also

» INTERVIEW

represent a precious asset to be preserved and maintained over time. The property of the data must therefore remain within our scientific community, which must provide for their management.
With the future enhancement of LHC (the High Luminosity LHC project, Ed.), this role will be even more prominent because the amount of data to manage will be much greater than the current one: a possible strategy could be having not many data centres distributed around the world, interconnected by very fast networks to reduce the need for data movement, with the objective of reducing the overall costs of the system. The main challenge for CNAF, but also for INFN, will therefore be to keep abreast of these demands and to be able to offer to the WLCG community an Italian data centre. The complementary aspect of this challenge lies in the computational and analytical capacity demanded by these data, which by definition are more flexible and can be provided within the WLCG world, including CNAF of course, by commercial clouds, but also by supercomputing centres. In this last case, the main problem is to be able to use these High Performance Computing (HPC) centres with analysis codes that we generally use in our Tier1 and Tier2, but this depends highly on the type of architecture and the type of processors that the HPC centres use, and it is therefore necessary to consider these case by case. A case in point is the CNAF-CINECA connection, mentioned before, which anticipates the strategic objective of putting the two communities into close contact. In this case the architectures and processors used by CINECA's Marconi supercomputer are substantially the same as those used in the WLCG world.
According to the European strategies, in Italy this supercomputing and big data pairing will be expressed at its best by 2021, when the INFN and CINECA's computing centres will be moved to Bologna's Tecnopolo, in an area of approximately 6,000 square meters, made available by the Emilia Romagna Region, and that will become the new site of the Data Centre of the European Centre for Medium-range Weather Forecast (ECMWF), which will leave the current location in Reading, England: hence becoming Italian research's largest scientific computing hub.
A great opportunity for the INFN and for the CNAF: a big challenge that – if we manage to compete in and win – will not only allow us to be a data centre in the High Luminosity LHC "data lake", but also to attract projects, and therefore funds, in the world of innovative information technology services (big data custody, analytics, deep and machine learning, etc.), to be developed for or with the world of industry, together with other scientific disciplines, favouring the ones with a more direct impact on the health and lives of citizens. ■