» INTERVIEW

WHEN RESEARCH IS BIG DATA AND COMPLEX COMPUTING

Interview with Antonio Zoccoli, vice president of INFN and responsible for the Computing and Networks division of the INFN's executive committee.

To enable epoch-making achievements like the discovery of the Higgs Boson and that of gravitational waves, but also to study the properties of cosmic rays and neutrinos, basic physics research handles enormous volumes of data and uses complex computing systems. For instance, in view of the huge amount of information produced by each collision between particle beams in the LHC accelerator at CERN, physicists have designed and developed a special infrastructure for the selection, storage and analysis of data. This continually evolving global infrastructure is a complex and organised system that incorporates different computing resources regardless of their geographical location or capacity. A worldwide computing network, known as the GRID, that harnesses the computing power and memory capacity of tens of thousands of different computers. The result is a computing power equal to that of 100,000 computers.

What are the essential requirements that guide the INFN's scientific calculations?
At the INFN we started performing scientific calculations when we had to analyse data from experiments in which we were taking part, so really it is something we have been doing ever since our first experiments. Right from the start, we recognised the importance of not just analysing data, but also of developing computing resources capable of performing Montecarlo simulations, a fundamental resource in scientific research. However, although scientific calculation was recognised as an important part of research activities, it was considered a secondary aspect in the planning and implementation of experiments until 20 or 30 years ago. Most experiments were designed irrespective of their computing needs: only later and depending on the circumstances were the computing instruments improved and the necessary infrastructure provided. There has definitely been a change of approach in recent years owing to the huge volumes of data produced by the LHC: the computing grid is now regarded as a

» INTERVIEW

fundamental part of the experiments, on a par with the detectors and the various scientific instruments. What we have witnessed in recent years is a real paradigm shift. Today it would be unthinkable to design an experiment without first knowing how much data will have to be handled or defining the appropriate procedure and infrastructure to analyse them. We will have to tackle this challenge over the coming years, since the LHC upgrade and subsequent HI-LUMI LHC project are two new experiments which are expected to generate 10 times more data than the LHC has done up until now.

The LHC is undoubtedly a driving force of development in computing resources for high energy physics. What are the specific research needs in this field and which solutions have been adopted?

The LHC has marked the turning point for computing infrastructure. Before it was designed, experiments could only rely on their own, extremely localised computing resources. With the start of the LHC project, instead, two goals were pursued right from the start. First: to provide enough computing capacity to analyse an unprecedented volume of data. Second: to allow all scientists participating in the experiments, based anywhere in the world, to access data so that calculations could be performed by the respective institutions. This meant the infrastructure had to be accessible from anywhere. The solution was the GRID, a worldwide computing infrastructure that literally encompasses the entire Planet. The name GRID comes from the analogy with the electricity grid. When you plug an electrical appliance in you certainly never have to think about having to build a electricity power station. Likewise, the GRID allows users to obtain a computer processing resource without having to know where it comes from. A network of computing sites connected via high-speed optical fibres and an interface that offers access to all users is no longer an infrastructure made up of individual resources, but a system. This is an entirely novel approach and the GRID is the first and only one of its kind in the world.

How has the INFN contributed?

The technological challenge has been addressed at a global level and the INFN has made a substantial contribution that has gone hand in hand with its participation in the LHC experiments. The challenge consisted in the need to allow the high energy scientific community to access the available resources and transmit massive data volumes in a very short time. With the problems associated with sharing data on such a large scale, such as authentication and data protection. In the end, we managed to develop the necessary hardware and software, with a significant contribution by the INFN in terms of manpower. The WLCG (Worldwide LHC Computing Grid) project is a collaboration of more than 170 computing centres in 42 countries. Its mission is to distribute and analyse the 50 Petabytes of data generated by the LHC in 2016 alone: a volume

» INTERVIEW

of data unparalleled in other disciplines and to which the term Big Data refers, not only for the huge volumes involved, but also to indicate their variability and the speed and flexibility with which they are transmitted and shared. The INFN has contributed with its researchers and specific skills to the implementation of the GRID and has been a key player in the process, concentrating most of its efforts between 2000 and 2010. In terms of scientific progress, this revolution produced its effects immediately. For the first time in the history of large experiments, scientific results can now be obtained just a few months after gathering data. The discovery of the Higgs Boson was the first tangible proof of this. As regards national resources, the process has led to the creation of a distributed computing grid in Italy in which the main centre, known as Tier1, is in Bologna and to which ten other Tier2 centres distributed nationwide are connected. This grid is part of the worldwide grid. It is connected to the Tier0 centre at CERN and to the other Tier1 and 2 centres around the world, in Europe, Asia, Japan, USA.

In addition to the LHC, the GRID supports experiments and collaborations in the field of data sharing and analysis. Which experiments benefit most?
At first the GRID was only used for analysing LHC data, the purpose for which it was originally developed. But then other experiments involving the analysis of large volumes of data, such as Belle II in Japan, and BES III in China, began to adopt the same approach. More and more experiments now rely on the GRID infrastructure, even in fields other than accelerator physics, such as large-scale international astroparticle physics research collaborations. In Italy, the Italian National Institute for Astrophysics (INAF) is the main body involved in such projects, in which the INFN also participates. I refer for instance to projects currently being developed such as the Cherenkov Telescope Array (CTA) or the Euclid satellite project. Then there is the Xenon experiment studying dark matter at the Gran Sasso National Laboratory of INFN. Given the significant increase in the volume of data to be analysed, researchers have asked to use the services of the LHC Tier1 and Tier2 sites. These experiments are rapidly expanding their scope and becoming increasingly international. Although they will continue to process less data than the LHC in the coming years, the GRID will still be an extremely valuable resource.

What are the prospects in Europe and worldwide?
We now face a double challenge. First, the infrastructure must move towards a new organisational model. The paradigm of GRID computing based on the connection of many CPUs via networks is no longer appropriate. Just to analyse the amounts of data generated by the next LHC upgrades we will need a larger infrastructure that uses Tier1 and Tier2 computing facilities, as well as machines with High Performance Computing capabilities. The infrastructure must be more general so that users within any branch of research can use it in the way most useful to them.

» INTERVIEW

If I need to perform data analysis that also involves the use of computing capacity, I must expand the opportunities offered by the infrastructure. Second, we will have to abandon the GRID approach and move towards CLOUD computing, a more flexible system in which resources can be used by users with different needs. Other experiments and research topics that are not necessarily relevant to the INFN must be able to access the infrastructure. The INAF, for instance, with which we are involved in ongoing experiments, but also the Italian Space Agency (ASI), and the Long Baseline Science sector that is currently particularly active in China.
For this process to be effective we must continue to develop the GRID for the future of the LHC. From the outset, we had to decide whether it is worth generalising the infrastructure to make it more flexible and integrated in a system, so that it is available to Italian research centres other than the INFN. We have the expertise needed to take this step and, with the support of the institutions, it would definitely be worthwhile in order to avoid the unnecessary dissipation of efforts and resources which would lead to fragmentation of the interests of the different scientific communities. This is exactly what happened before with the GRID, which was set up for us and is now a common asset.■