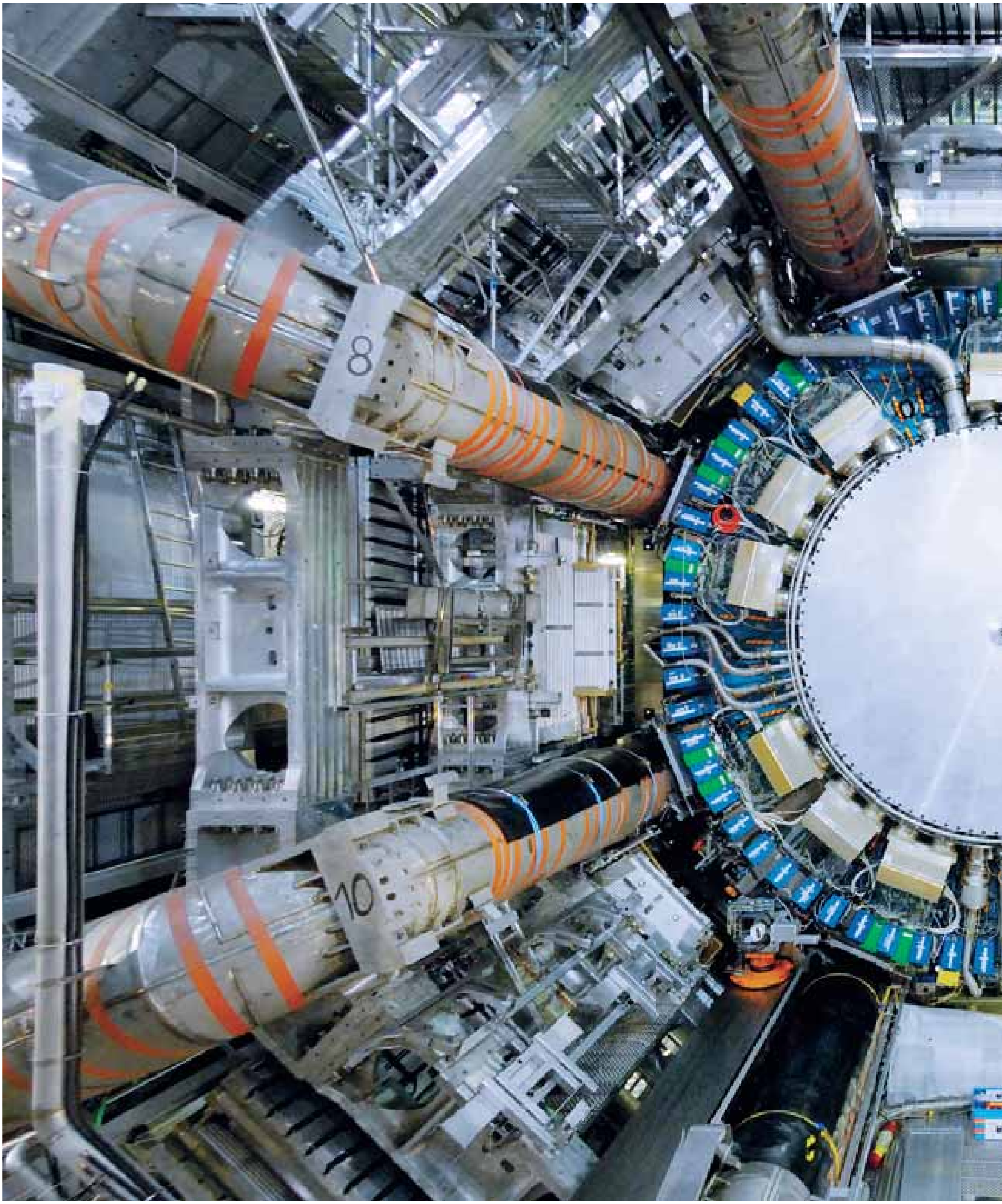
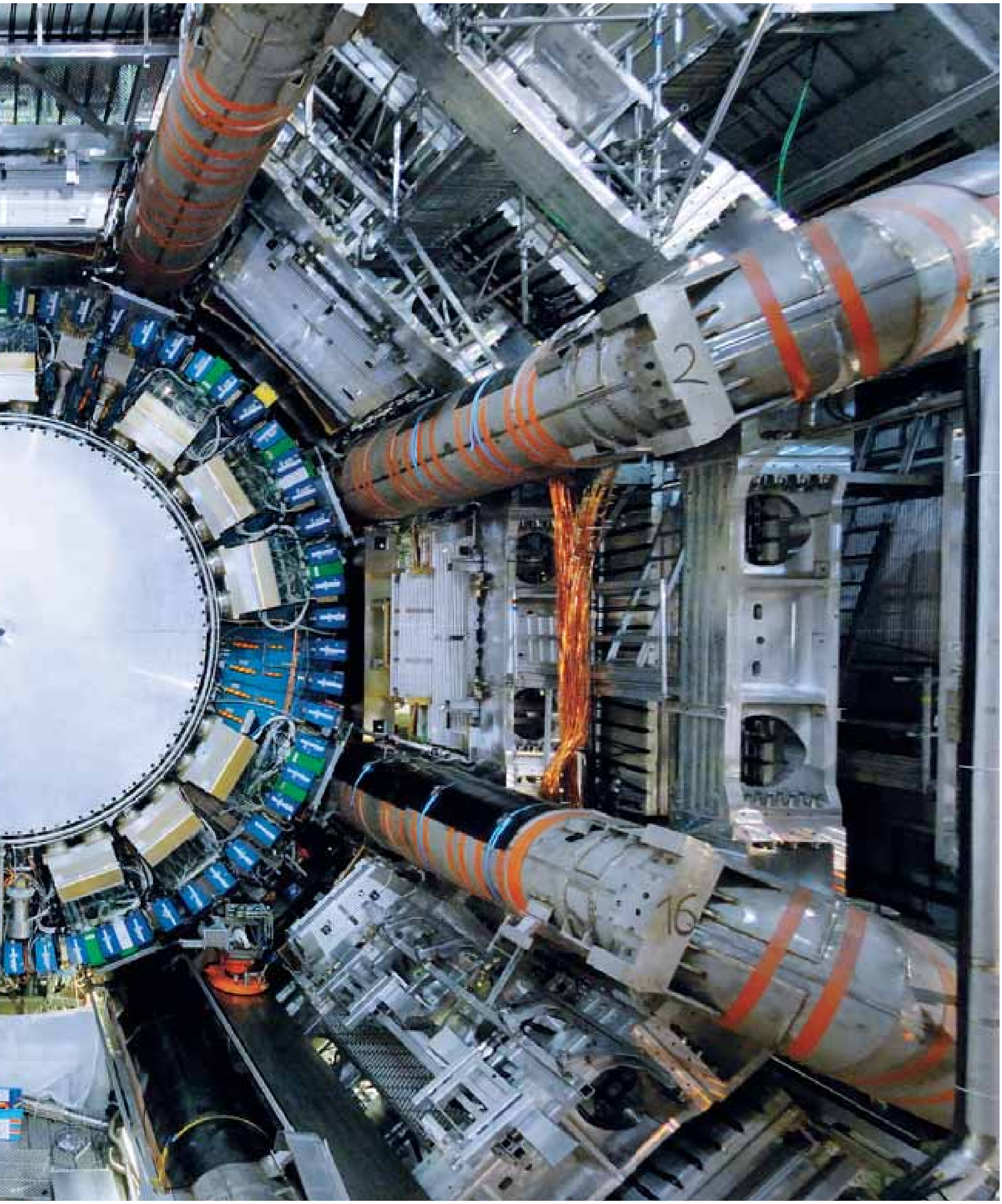
**➔ESSENTIALS:**
How to process hundreds of petabytes of data every year? With a network based on technology and trust



# BIG BANG, BIGGER DATA



At CERN, physicists are looking for nothing less than the fundamental principles of the universe. While doing so, CERN's particle accelerators generate more data than any other place on earth. A visit to the "world machine."

THE CERN BEARS MANY NAMES. For some, it is the “cradle of the Internet.” Here, Tim Berners-Lee proposed a project to his employer CERN, based on the principle of hypertext. His goal: to simplify the worldwide exchange of information between scientists – which could change nothing less than the whole world. Others call it the “birthplace of the touch screen.” It’s where Frank Beck and Bernt Stume invented the first touchscreen, designed to run CERN’s SPS control system. It had a simple display with buttons, which, when touched, helped the operator run it with little contact. Still others call the CERN a “world machine,” because that is exactly what they are doing here at the Conseil Européen pour la Recherche Nucléaire: dealing with fundamental issues, such as understanding the very first moments of our universe, looking for antimatter, or trying to comprehend dark matter.

Here, they trace the origin of the world by simulating the Big Bang with the Large Hadron Collider (LHC), the world’s largest and most powerful particle accelerator. Booted up in Sep-

tember 2008, the LHC consists of a 27-kilometer ring of superconducting magnets and a number of accelerating structures to boost the energy of the particles along the way.

## Collisions that generate petabytes of data per second

Inside the accelerator, two high-energy particle beams travel at 99.9 percent of the speed of light before they collide. The resulting debris comprises new particles, which fly out in all directions. This interaction creates an enormous flow of data. Up to nearly 1 billion particle collisions can generate up to 1 petabyte of data per second. Data is filtered in real time, selecting potentially interesting events, the so-called "trigger."

**Alberto Di Meglio,** Head of CERN's openlab: "We have to examine the emerging patterns"

**The CERN data center** in Meyrin, Switzerland, heart of CERN's entire scientific, administrative, and computing infrastructure

Physicists at CERN's data center in Meyrin, the heart of the lab's infrastructure, sift through these petabytes – mostly in real time – running complex algorithms to achieve structured data. Still, even after filtering out almost 99 percent of it, CERN expected to gather around 50 to 70 petabytes of data in 2018. A rather conservative estimate, since by the end of the year, they had 80 petabytes of data on tape.

And it's not going to be less: "The main driver for us until today was the so-called 'standard model of particle physics,' a very successful way to classify interacting particles," says Alberto Di Meglio, Head of CERN's openlab. The scientists' key mission was to complete this model, and it was, in the main, accomplished with the discovery of the Higgs boson. "That was the last missing piece," Di Meglio continues, "compared to the challenges ahead it was not difficult to 'listen' to the data and gain the relevant insights from it". Not difficult? "Yes, because we could describe exactly what we were looking for," says Di Meglio.

## Discarding the "noise" of data

"Listening to the data" – for Di Meglio, this means finding specific patterns in raw data-sets and organizing them, while discarding everything else that is too "noisy." Easy, as the head of CERN frankly explains, "We collect the data based on clearly defined requirements, while using extremely fast filtering techniques, implemented to discard 98 to 99 percent of that noise."

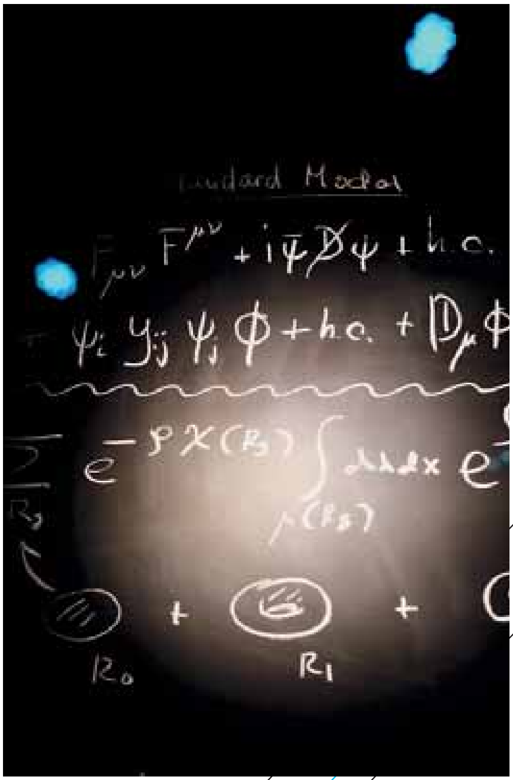
Because today, as Di Meglio adds, this standard model is fairly complete; everything it can describe has been validated in experiments. However, with only 5 percent of the universe explained, it merely represents a very small part of what exists. "We know there is more that we cannot describe properly. We don't have a model to illustrate the remaining 95 percent." That makes it difficult to determine which data to discard and which to use. Since there is always a risk of throwing away useful data/information, it is necessary to really listen to the data and what it contains, instead of hunting for specific patterns that correspond to a predefined search pattern. "We have to examine the emerging patterns, instead of diving into the data for something that we already know how to describe."

Already, the scale and complexity of data from the LHC is unprecedented and will continue to set new standards in the future. Data that needs to be stored, easily retrieved, and analyzed by physicists all over the world requires massive storage facilities, global networking, and immense computing power. But because CERN does not have the computing or financial resources to crunch all of the data on site, it turned to grid computing in 2002, in order to share the burden with computer centers around the world.

## A network of trust

The result, the Worldwide LHC Computing Grid (WLCG), is a distributed computing infrastructure arranged in tiers, giving a community of over 10,000 physicists almost real-time access to LHC data. The 170 computing centers in 42 countries, including Antarctica, are classified as Tier-0 centers located at CERN and in Hungary, where the data recording, reconstruction, and distribution happen; Tier-1 sites offer permanent storage, re-processing, and analysis, and Tier-2 centers provide simulation and end-user analysis.

The WLCG allows seamless access to computing resources, which include data storage capacity, processing power, sensors, visualization tools, and more. Users make job requests from one of the many entry points into the system. A job entails the processing of a requested set of data, using software provided by the experiments.

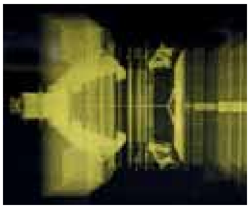


**Part of the equation** of the "Standard Model of particle physics" – a theory describing three of the four known fundamental forces in the universe, as well as classifying all known elementary particless

**The Globe of Science** and Innovation at CERN: its outer shell, resembling a finely spun cocoon, is designed to protect the building from the sun and the elements, just like Earth's atmospheric layer

## THE LARGE HADRON COLLIDER (LHC)

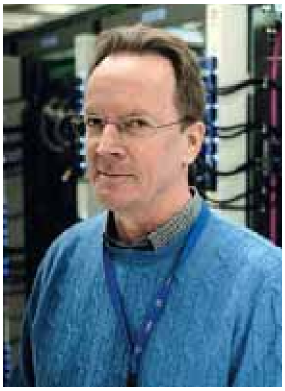
is the most complex machine ever built. It produces enormous amounts of very complex data. That's why CERN established CERN openlab in 2001: to help ensure that members of its scientific community have access to the very latest ICT solutions, thus helping them to gain scientific insights from this data and push back the frontiers of physics. CERN openlab's mission is to accelerate the development of cutting-edge ICT solutions for the worldwide LHC community and wider scientific research. The projects carried out are currently addressing topics related to data acquisition, computing platforms, data storage architectures, computer provisioning and management, networks and communication, and data analytics.



**A detail display** of ATLAS, one of the four major experiments at the LHC at CERN. ATLAS is designed to exploit the full discovery potential of physics opportunities that the LHC provides.

The computing grid establishes the identity of the user, checks credentials, and searches for available sites that can provide the resources requested. Users do not have to worry about where the computing resources are coming from – they can tap into the grid's computing power and access storage on demand, says the head of WLCG, Ian Bird.

"One of the underlying principles of WLCG is that it's a federated infrastructure, in the sense that each user has a single identity that is recognized everywhere," Bird says. Basically, all scientists register and are given credentials, allowing them to submit their work to the cloud. That forms a network of trust between all of these computer centers, built on the certificates issued by trusted authorities. Bird: "There is a whole set of rules and conditions under which these certificates are issued. For instance, the identity of the applicant has to be verified in person."



**Ian Bird,** Head of the Worldwide LHC Computing Grid: "The only reason we can run that federated network is because we trust each other"

### Security, done by people

That validation process is the fundamental part of CERN's network of trust, Bird continues, adding, "That puts us in a much better situation than we were, when we are a bunch of individual computer centers." Security is done by people rather than by technology, he says. "Of course, there is a layer of technology, but ultimately the only reason we can run that federated network is because we trust each other to issue the credentials in a way everyone agrees upon – a network of contacts that every business out there should have."

CERN is pushing boundaries, technology-wise, data-wise, and business-wise. "We are relying on new ways of working, as well as on new ways of analyzing information, from machine learning to deep learning," says Alberto Di Meglio. Increasing the speed at which an analysis is made so that it can be done in real time is what all of the experiments are about. The goal is to shrink the separation between the online and the offline world. "Today it is all about collecting the data, reducing it as much as possible, and analyzing what's left offline. Future experiments will have to process much more data much faster in real time." That's why the scientists at CERN are trying to understand how advanced applications can fill the gap, since the trend toward collecting information from sensors, wearable devices, and machines, as well as the need to analyze that massive amount of information in real time, are both growing.

This task concerns more than CERN. It will be critical for every science and business in the future. Industries are confronted with the need to understand this information very fast in order to make decisions, says Di Meglio. "We work to investigate these kinds of techniques, and to adapt and deploy them." To understand the universe, as well as the data pouring out of it. ✕

**The accelerating cavity** of CERN's first accelerator, the Synchrocyclotron