CERN's data stores soar to 530M gigabytes

CERN's data requirements are growing at massive rates: Every experiment with its Large Hadron Collider creates 3GB of data per second.

Lucas Mearian (Computerworld (US)) on 14 August, 2015 23:36

Since restarting in June after a two-year upgrade, CERN's Large Hadron Collider (LHC) has been recording about 3GB of data per second, or about 25 petabytes -- that's 25 million gigabytes -- of data per year.

Every time the LHC smashes particles together at near the speed of light in its 16 mile-long chamber, the shattered particles fly off in myriad of directions. Those particlaes leave behind traces in space, like footsteps in snow, which are recorded and later analyzed in a search for the most basic element of matter.

But unlike a camera, which absorbs light in order to produce a photo, the traces that result from particle collisions pass through the LHC's "detectors," leaving many points of interaction in their path. Every point represents an action at a point in time that can help pinpoint the particle's characteristics.

The detectors that record particle collisions have 100 million read-out channels and take 14 million pictures per second. It's akin to saving 14 million selfies with every tick of a watch's second hand.

Needles and haystacks

Guenther Dissertori, a professor of particle physics at CERN and the Swiss Federal Institute of Technology in Zurich, said the task of finding matter's most basic particle is vastly more difficult than finding that proverbial needle in a haystack.

"The search for the particle is more than a search for a needle in a haystack. We get 14 million haystacks per second - and unfortunately the needle also looks like hay," Dissertori said. "The amount of data produced at CERN was impressive 10 years ago, but is not as impressive as what's produced today."

Dissertori said CERN's public-private partnerships could solve the expected technological hurdles, including the need for new storage technologies that can save exabytes of data in the future.

Unlike Google or Amazon, two Internet companies that spend billions of dollars every year to develop new technology, CERN has limited money; it's funded by 21 member states and has an annual budget of around $1.2 billion.

"We have to be very creative to find solutions, Dissertori said. "We're forced to find the best possible ways to collaborate with [the IT] industry and get most out of it."

Almost since its founding, CERN has been developing ways to improve data storage, cloud-technologies, data analytics and data security in support of its research. Its technological advancements have resulted in a number of successful research spin-offs from its primary particle work, including the World Wide Web, hypertext language for linking online documents and grid computing.

Its invention of grid computing technology, known as the Worldwide LHC Computing Grid, has allowed it to distribute data to 170 data centers in 42 countries in order to serve more than 10,000 researchers connected to CERN.

Storing data, sharing data

During the LHC's development phase 15 years ago, CERN knew that the storage technology required to handle the petabytes of data it would create didn't exist. And researchers couldn't keep storing data within the walls of their Geneva laboratories, which already house an impressive 160PB of data.

CERN also needed to share its massive data in a distributed fashion, both for speed of access as well as the lack of onsite storage.

As it has the past, CERN developed the storage and networking technology itself, launching the OpenLab in 2001 to do just that. OpenLab is an open source, public-private partnership between CERN and leading educational institutions and information and communication technology companies, such as Hewlett-Packard and Nexenta, a maker of software-defined storage.

OpenLab itself is a software-defined data center that started phase five of its development cycle this year. That phase will continue through 2017 and tackle the most critical needs of IT infrastructures, including data acquisition, computing platforms, data storage architectures, compute provisioning and management, networks and communication, and data analytics.

A growing grid

In all, the LHC Computing Grid has 132,992 physical CPUs, 553,611 logical CPUs, 300PB of online disk storage and 230PB of nearline (magnetic tape) storage. It's a staggering amount of processing capacity and data storage that relies on having no single point of failure.

In the next 10 to 20 years, data will grow immensely because the intensity of accelerator will be ramped up, according to Dissertori.

"The electronics will be improved so we can write out more data packages per second than we do now," Dissertori said.

Every LHC experiment at the moment writes data on a magnetic tape at the order of 500 data packets per second; each packet is a few megabytes in size. But CERN is striving to keep as much data as possible on disc, or online storage, so that researchers have instant access to it for their own experiments.

"One interesting development is to see how can we implement it with data analysis within our cloud computing paradigm. For now, tests are ongoing on our cloud," Dissertori said. "I could very well imagine in near term future more things done in that direction."

Email address

Join

Tags storage CERN cloud storage

More about CERN Geneva Google Nexenta Technology

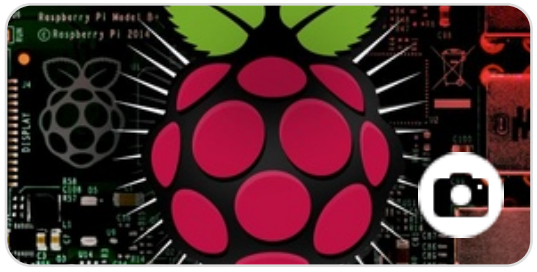
Read next

Google DeepMind advances AI with deep Q-network algorithm

Apple, IBM release 10 new apps, incorporating analytics and iBeacons

Security Watch: Nuix bolsters leadership team, focus on cybersec policy

CSO Online



In Pictures: Ultimate guide to Raspberry Pi operating systems, part 1

In Pictures: User guide to Windows 10

Windows 10: What we like, what we don't like