CANONICAL

The Wellcome Sanger Institute enables sharing of genomic research worldwide with Canonical-supported Ceph storage

About the Wellcome Sanger Institute

- One of the premier centres of genomic research in the world
- Uses leading-edge DNA sequencing technology to carry out large-scale studies that cannot be conducted in most institutes
- Leads global scientific collaborations in areas including cancer, infectious disease, and cellular genetics

HIGHLIGHTS

- The Wellcome Sanger Institute deployed an OpenStack private cloud, and engaged Canonical to support its evolving Ceph object storage system
- With Canonical handling high-level issues, the Institute is able to operate a reliable Ceph platform with more than 20 PB of raw capacity
- Through the S3 protocol, Ceph enables the Institute's scientists to securely share genomic data worldwide

wellcome sanger institute

When it comes to genomic research, data storage holds the crown jewels. So the Wellcome Sanger Institute wanted expert support for its Ceph clusters.

Aiming to give its scientists greater flexibility, the Sanger Institute recently implemented an OpenStack private cloud to complement its existing batch computing infrastructure. The Institute initially opted for self-maintained Ceph storage, but as the service evolved, the Institute turned to Canonical for high-level support. With Canonical's help, the Sanger Institute has successfully scaled to over 20 PB of Ceph storage and unlocked the potential of the S3 protocol – enabling scientists to seamlessly share huge volumes of genomic data all around the world.

Challenge

As a global leader in genomic research, the Sanger Institute plays a key role in an array of large-scale collaborative studies. For example, the Institute is currently leading the Darwin Tree of Life Project – a UK-wide initiative to map the genetic code of all 60,000 complex species found in the British Isles.

Given the breadth and complexity of the Institute's research, its 500 scientists frequently have differing software requirements for their workloads. However, the Sanger Institute's traditional, batch computing approach significantly limited researchers' flexibility. Users only had access to a single computing environment, and could not be given the freedom to install the different applications or operating systems that they needed.

To overcome this obstacle, the Institute decided to implement an OpenStack private cloud. Through OpenStack, scientists can freely customise virtual machines to meet their specific needs.

Naturally, the new private cloud required object storage, and Ceph on Ubuntu emerged as the best option due to its reliability, scalability and compatibility with OpenStack deployments. Rather than deploy OpenStack and Ceph together as a single unit, the Institute was keen to keep the two components independent.

Peter Clapham, Informatics Support Group Team Leader at the Wellcome Sanger Institute, explains: "We wanted to keep the installation and configuration changes of OpenStack and Ceph discrete so that we could manage and update them separately. That way – by dissociating their requirements – we can be closer to the bleeding edge."

In 2018, the Wellcome Sanger Institute output almost 7,558 billion DNA bases a day, reading the equivalent of one human genome every 17 minutes. This was made possible by a data centre containing 32,000 compute cores, with a network backbone speed of 400 Gb/sec.

The Institute initially chose to deploy and maintain its Ceph storage in-house, starting with 3 PB of capacity. Yet demand for the new service proved to be far greater than expected.

Matthew Vernon, Principal System Administrator at the Wellcome Sanger Institute, comments: "We thought that our scientists might be interested in the facilities afforded by the RADOS Gateway. It turned out to be wildly popular."

The S3 protocol is an internet-based object storage service that enables users to store and retrieve data of any volume from any location. The protocol has become a de facto standard for accessing object storage. Collaborating on genomics projects with other institutes often involves securely sharing hundreds of terabytes of data. Traditionally, this meant posting physical discs – but with the object store service enabled by Ceph, scientists can make even the largest data volumes available digitally.

"We were eager to evolve the service," continues Peter. "We were scaling capacity to keep up with demand, but we also wanted to improve performance and unlock new opportunities for our scientists. To make that happen, we needed the support of a trusted partner."

> "Canonical support often connects us directly with the engineering team. Being able to consult with people who are so familiar with Ceph's internals has been highly valuable for resolving the thornier issues."
>
> Peter Clapham, Informatics Support Group Team Leader, The Wellcome Sanger Institute

Solution

After evaluating a range of providers, the Sanger Institute chose Canonical to support its Ceph implementation.

Matthew explains: "We started out from day one of our Ceph journey with Ubuntu as the underlying operating system, since we have a lot of experience with Ubuntu in-house. That made Canonical a logical choice."

Matthew adds: "We also wanted a certain amount of flexibility in what we could do with our Ceph storage. We found that Canonical's offering gave us the most freedom in terms of which Ceph versions we could run. If we identify any bug fixes that we need to apply, Canonical is very helpful in getting those fixes backported to the Ceph release we're using."

With Canonical's support, the Sanger Institute has already deployed several key upgrades to its Ceph service. The Institute has begun taking advantage of Ceph's BlueStore backend, it has made well-received improvements to its S3 interface, and it is currently working to achieve better utilisation of its storage capacity.

"Because of our own expertise with Ceph, we generally only need to talk to Canonical about the really complex issues," says Peter. "Support often connects us directly with Canonical's engineering team. Being able to consult with people who are so familiar with Ceph's internals has been highly valuable for resolving the thornier problems."

Peter continues: "Canonical understands that we're already experienced with Ceph, and while we don't raise many support tickets, those we do raise are far more in depth than average. Canonical provides the extra degree of knowledge and competence that our challenges require – and that has led to a trusted relationship."

When the Sanger Institute first deployed Ceph, it took a month to rollout. But thanks to improved in-house experience alongside Canonical support, the most recent 3 PB addition took just four days.

> "The work we do at Sanger isn't possible at most other institutes, but now that our data is more accessible than ever, it's reducing the barrier to entry for other scientists."
>
> Peter Clapham, Informatics Support Group Team Leader, The Wellcome Sanger Institute

Benefits

Today the Wellcome Sanger Institute is running 20 PB of Ceph storage in production across three clusters – providing its researchers with a reliable platform for their scientific workflows, and for seamlessly sharing immense volumes of genomic data worldwide.

"Being able to make data externally available through the standardised S3 interface is hugely powerful," comments Matthew. "The work we do at Sanger isn't possible at most other institutes, but now that our data is more accessible than ever, it's reducing the barrier to entry for other scientists. This is a great enabler not just for us, but for everyone that wants to be involved in these collaborations."

Compared to physical discs, managing dataflows through Ceph and S3 is significantly easier. The Institute no longer has to spend so many resources worrying about who is responsible for the discs, cataloguing, encryption, or border constraints on data movement – so staff can spend more time focusing on the projects themselves.

Looking to the future, the Wellcome Sanger Institute intends to continue building on its Ceph deployment. The next step will be to transition production machines from Ubuntu 16.04 to 18.04. And with Canonical's help, the Institute hopes to be able to more quickly adopt new technologies – such as erasure coding – to further develop the service.

"Storage effectively holds the crown jewels of the Institute," concludes Peter. "There's a huge amount of value held on the system. With Canonical as our trusted partner, we're confident that we can overcome even the most complex Ceph challenges and continue to create new opportunities for our scientists and the community."

Click the link below to learn more about Ceph: ubuntu.com/ceph

ubuntu
Delivered by Canonical