Questa storia appartiene alla categoria *Innovazione* e al dossier *Servizi infrastrutturali* , *International cooperations*

The academic cloud goes international

In search of solutions for scientists to handle big data, SWITCH cooperates with other research and education networks.

Testo: **Saverio Proto**, pubblicato il 30.08.2016

Scientists often have to work with large amounts of data, for example when analysing extensive datasets in the public domain. That is what SWITCHengines was developed for. A researcher developing a new search algorithm for the web can test his work against the Common Crawl dataset , which contains data collected from the Word Wide Web over many years. A biologist working on human genomes, meanwhile, can compare his local dataset against the public 1000 Genomes Project data.

It is easier to process big data nowadays than it used to be because there are many open-source tools to choose from, and compute capacity is available on demand from big commercial cloud providers like Amazon. Some public cloud providers have scientific datasets available for free, as long as you pay for the CPU time to process the data.

Scientists need two things for their work: a computer cluster with a high compute capacity to process the data plus storage space to hold the datasets and the results of the computation.

> *The SCALE-UP project contains a work package called Scientific Data Pools, which aims to offer storage for big datasets to make them available for other researchers – tailored to researchers' specific needs and budgets. The aim is to integrate this feature into the SWITCHengines service.*

> *Why not work together in Europe? Each institution could store the data with reduced redundancy.*

The good thing is that all National Research and Education Networks (NRENs) and research institutions have quite the same problems when hosting scientific datasets. Why not work together in Europe? Each institution could store the data with reduced redundancy, exploiting the ability to download a lost copy in an emergency. If all the institutions make the data accessible through a standard object storage API, it is easy to cooperate and reduce costs for data redundancy. SWITCH started a cooperation for such a pilot with GARR (the Italian NREN), the University of Zurich (UZH) and the Federal Institute of Technology in Lausanne (EPFL).

The project

Sull'autore

Saverio Proto

Saverio Proto is an OpenStack Cloud Engineer. He has been working for NRENs since 2011, first in Italy and then in Switzerland. He has significant experience in running critical infrastructures using open-source software. At SWITCH, he works in the Peta Solutions team, delivering an OpenStack-based cloud to the Swiss universities.

E-mail

SWITCHengines

SWITCHengines offers computing power suited to projects that do not have their own infrastructure and do not plan to acquire it. You can place your order online, and your made-to-measure processing and storage capacity is available immediately. You only pay for what you need. SWITCHengines is a service developed specially for the research community.

Betatest
Do you want to beta test the SWITCH public dataset service? Read the *tutorial* and get in contact with *SWITCH engineers*.

SWITCH chose Google Books Ngrams as the dataset for its tests. With its 5 TB size, it is big enough for a proof of concept and small enough to allow us to test our copy procedures quickly. EPFL provided the first copy of Google Books Ngrams to download. SWITCH downloaded the dataset over SWITCHlan and served it over the GÉANT network. To emulate a real production environment, we described our use case in these terms:

Each NREN

- must have read-only access to the datasets hosted on a remote site.
- should be able to sync the remote datasets to the local datasets at any time.
- is independent in terms of how it presents the dataset to its users.

> *SWITCH plans to enhance SWITCHengines with hosting for scientific datasets at the petabyte scale by the end of 2017.*

Synchronisation of the datasets between the sites is easy because the datasets only change by incrementing the data, never changing the existing data. Using standard object storage APIs like *swift* and *S3*, SWITC served the data successfully, and GARR and UZ were able to sync a full copy of the dataset.The experiment with a dataset of smaller size was useful to identify the software bugs in the existing open-source tools. SWITC , GARR, EPFL and UZ cooperated in fixing them and contributed the fixes to open-source projects during the pilot phase. Without the joint work of the engineers, it would have been much harder to fix all these bugs in such a short time.

Next steps

SWITC plans to enhance SWITC engines with hosting for scientific datasets at the petabyte scale by the end of 2017. The datasets will be accessible both for computation within SWITC engines and for access via SWITC lan in case of computation with existing computing clusters in Switzerland.

After this positive experience, we believe that cooperation with other institutions has a great value. SWITC is open to cooperations with other NRENs and institutions worldwide. International collaboration among NRENs, implementing services together, helps to reduce costs and to provide users with a better service. Know-how on software tools is built up much faster when working with an international team. First of all, sharing progress every step of the way forces the engineers to produce high-quality documentation, making it easy for anyone to step in and help. Cooperation also means you learn from each other. Discussing technical difficulties openly with other engineers who have a different point of view often resulted in a quick solution to problems that were holding us back.

More about *SWITCHengines*