Case Study:
CNIO (Centro Nacional de Investigaciones Oncológicas)

the SERVER LABS
the IT architects

CNIO's cancer research gains speed in the Cloud

In the summer of 2010 CNIO (Centro Nacional de Investigaciones Oncológicas), Spain's leading organisation on cancer research, asked The Server Labs to undertake a comprehensive feasibility study identifying the possibilities to move its vast data processing to the Cloud.

Background – an escalating data processing bottleneck
CNIO is facing a problem shared by many laboratories around the world dedicated to genomic science research. Whilst in the past sequencing was the slowest and most costly step in projects, new DNA sequencer technologies have been speeding up the process to a point where data are produced much faster than in-house computational capacity can processes them. In short the bottleneck had shifted from creating DNA reads to their post-read processing. In order to shorten the processing time and memory requirements specialised algorithms had been developed.

Nevertheless, the sheer amount of data to handle forces laboratories like CNIO to allocate an ever-increasing amount of their budget and manpower to amplify their computational infrastructure.

CNIO is now considering using the Cloud to shift its data processing to the pay-as-you-go model of public clouds, which combined with open-source software, will allow them to scale their analysis clusters and run more experiments in parallel while limiting the cost for hardware and manpower to a minimum.

Using the public Cloud reduces the time to conduct an experiment from up to several days to a few hours only.

Tangible benefits of using the public Cloud – automation and reproducibility
Running an experiment requires a number of steps - configuration of compute resources, tailoring scripts and data processing – all with much manual intervention. In order to reproduce an experiment (or e.g. audit it by a third party – usually required for publishing) the process would have to be repeated step by step.

Cloud computing makes it possible to overcome the computational bottleneck, allowing compute resources to be allocated and released on-demand whenever sequencing data is available.

Using a cloud architecture makes it possible to process a theoretically unlimited number of reproducible genomic experiments in parallel.

An architecture using the public Cloud simplifies these tasks, automates the process, and can even be configured to create audit trails making publishing or sharing data much easier. Previously sharing data between members of a research lab or between research centres has been a manual process at CNIO.

CNIO's genomic science projects process…

100 to 150 sequencing lanes per year
generating each…
30 gigabyte of entry data (average)
making up a total of…
3 to 4.5 terabyte in processing requirements p.a.

2 hrs of processing per lane to execute a typical workflow in Amazon EC2

The public Cloud architecture enables the automation of publishing data. Results are made available as soon as the processing is complete and are stored in a backed up location in the Cloud which means that CNIO does not have to hold backups in-house.

The Server Labs has proven with their feasibility study that it is possible to use the public Cloud by building the required architecture for large scale data processing.

Case Study:
CNIO (Centro Nacional de Investigaciones Oncológicas)

the SERVER LABS
the IT architects

Where does this solution apply?

The Server Labs believes that the outcome of this project applies to any kind of project requiring a reproducible computational environment on demand.

It is particularly significant to corporations – like e.g. research institutes, pharmaceuticals, telecoms companies, banks, media and universities - generating large amounts of data and with a need to share, publish and backup this data outside of the organization.

Technical milestones summarised:

Open-Source sequencing environment on the Amazon Cloud:
The Server Labs have automated the provisioning, configuration and installation of an open-source based environment which runs on the Amazon cloud. That environment contains leading open source software used in sequencing research labs around the world, also the Burrows-Wheeler Alignment Tool, Novoalign, and SAM and BED tools.

Highly scalable on-demand job processing based on Rightscale: The Server Labs have leveraged Rightscale's Rightgrid technology to enable batch processing of sequencing experiments which can now be run in parallel by launching new backend-worker nodes to reduce the total processing time. It also allows "scaling-up" the number of resources at a rate and number that suits CNIO's needs and then "scale-down" once the number of jobs have been processed.

Durable result data on Amazon S3:
Data processed on the cloud can be automatically published and held on S3 enabling CNIO can share it with partners, collaborating institutions, etc. Since the data is persisted between data centres by Amazon, S3 liberates them from having to create backups on site.

The Server Labs
The Server Labs (TSL) is a specialist IT Consultancy and Development Company and a leading authority in Cloud Computing services. Founded in 2004, The Server Labs focuses on the design and implementation of IT architectures and advanced software engineering projects, working with the most advanced solutions and technologies and offering its clients cost-effective, scalable and high performance solutions.

TSL's customer groups are predominantly large and medium-sized corporations, which share a growing need for cost effective and scalable IT solutions. TSL has offices in Spain, Germany and the UK. Most recently TSL started partnering with Amazon and RightScale to facilitate the adoption of Cloud Computing in Europe.

More information about The Server Labs is available on www.theserverlabs.com