Into the cloud... based on a true story

Specialising in online automatic translation with artificial intelligence

OVHcloud

SYSTRAN
beyond language

30,000+ hours
of computing on
NVIDIA Tesla V100 GPUs

50,000 intermediary models
stored in
Object Storage
for a volume of 30Tbits

A translation capacity
of **5 billion words**
per day

The background

SYSTRAN is a publisher of automatic translation solutions for business, and celebrated its 50th anniversary in 2018.

With more than 140 available language pairs, SYSTRAN services are customised to suit each customer's needs. Its solutions are used by a number of international companies, public organisations and translation agencies.

Since it was founded, SYSTRAN has always been a pioneer in automatic language processing. In late 2016, the company became a trailblazer with its launch of the first professional neural machine translation engine. This engine is based on the very latest advancements in deep learning technology, and is designed to improve the quality of instant translation.

As an OVHcloud customer from the very beginning, the publisher began working with OVHcloud in 2018 to design the solution now known as SYSTRAN Marketplace. This community platform offers the best translation models on the market, and they have been trained by multilingual experts from a range of different domains. The models are available in an on-premises cloud via professional translation tools, and are integrated into the customer's information system.

To approach this challenge, SYSTRAN adopted a community-based approach based on four pillars: technology, data, human expertise, and infrastructure. This has enabled them to adopt an open, responsible, highly available solution scaled for the web.

The challenge

Since 2016, automatic translation has considerably changed. Neural translation is an approach based on research into artificial intelligence, and more specifically, deep learning. It has become a standard, and is a successor to what is known as statistical translation. Statistical translation was essentially based on big data, as well as representation by experts on the rules that govern languages.

A lot of drastic changes have come with this transition. On a technological level, the algorithms required are constantly evolving, and come directly from huge public and private research laboratories. As a result of the neural approach, a general open-source standard has been developed and enforced. This means that replicable scientific progress can be made, and industrial development can be achieved almost instantaneously.

The volume of data required is lower than before, but the quality of the data is absolutely vital because the neural models will try to interpret any ‘noise’ as language rules. Big data makes it easy to forget that the information used to train translation models is produced by human translators. But even if this data is available online, it is still subject to author rights. And the quality of a model relies directly on investment in this data, so perfect traceability is a must. Without this attention to detail, it would be risky to trust translation models because they could be biased by their source data.

Human expertise was once put aside during the era of statistical translation, but is now returning in force. The algorithms are very powerful, so they need to be supervised by language specialists in different domains.

Finally, the neural approach has drastically changed computing infrastructure requirements. As is the case for any deep learning algorithm, during the training phase for models, specific graphic processing cards (GPUs) are required. On the other hand, during the inference phase, i.e. the use of models in production, the algorithms need servers that are optimised for processing, but with a relatively small volume of memory compared to previous generations. Regulations have also changed to further protect user rights, and this calls for particular attention to infrastructures hosting services that can translate confidential data.

Beyond the apparent simplicity linked to each of these changes — which are often reflected through performance demonstrations in very restrained use cases — fundamental changes are required to deliver a large-scale production chain that is responsible, transparent, and able to deliver the very best quality for each industry. The fundamental principle of this approach involves recognising the expertise of today's players, and linking them together to reach excellence.

SYSTRAN started investing in open-source technology in 2016 by co-founding OpenNMT, a a neural translation algorithm framework. This technology is now the most popular and active of its kind in the sector. It is used by thousands of researchers and industry professionals, who enrich the framework every day with their contributions. With this cutting-edge software building block, SYSTRAN's R&D teams have built comprehensive translation solutions that are specially designed for end users. Finally, the publisher developed a marketplace composed of several services. This fosters a community of experts who produce and share high-quality models, and they are also paid directly for their contributions.

They quickly realised that to build this platform, they would need to choose a flexible, robust and adaptable infrastructure. It would need to offer the computing power required to train neural motors. This environment would also need to be scalable in order to deploy models into production, respond to changing demands, and stay loyal to the responsible ethos of this community-based approach — all at a competitive price.

The solution

An open, secure and responsible platform, perfectly adapted to suit deep learning requirements

> *"OVHcloud quickly became our choice as a technological partner for hosting and operating our marketplace. OVHcloud's core values match the spirit of the marketplace. Our requirements in terms of flexibility and power drew us directly to the Public Cloud solution."*
>
> **Jean Senellart, CEO of SYSTRAN**

A technical solution combining power, flexibility and predictability

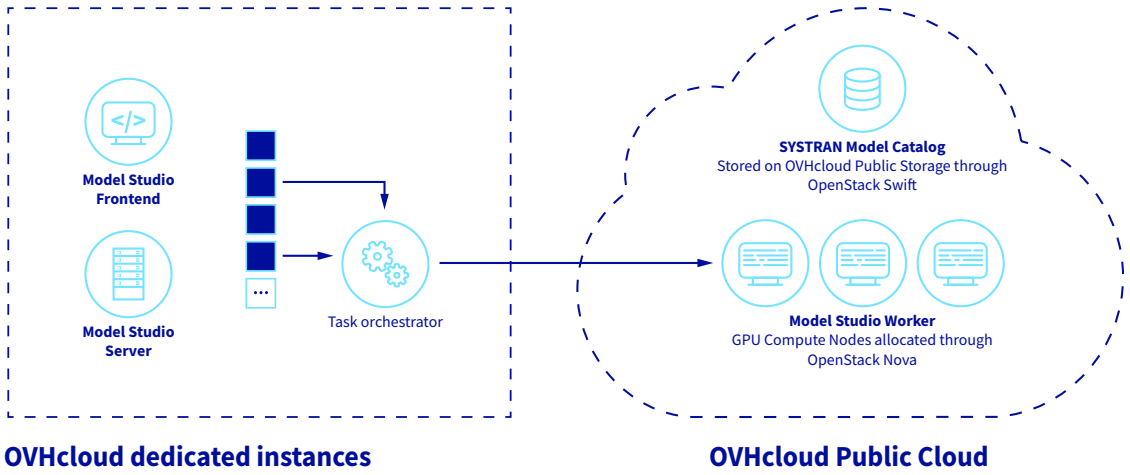
To ensure success for this project, SYSTRAN opted for the Public Cloud solution. Public Cloud solutions offer total cost management, and provide users with access to a wide range of servers and services. It also delivers the flexibility required to train neural models on demand, and generate varying volumes of translation over time.

SYSTRAN Model Studio — a unique solution developed by SYSTRAN to offer language and domain experts a way of training translation models themselves — requires on-demand access to the most powerful graphical processing units (GPUs) on the market. The constraint here was not instant availability for computing instances, as training for neural models is based on cycles that range from a few hours to a week.

Model Studio is a task orchestrator that can manage a sequence of iterations corresponding to a given training. It uses the OpenStack Nova API to launch compute instances dynamically.

In this diagram, the reliability of instances is essential. If an iteration failed, it would result in the failure of all the training associated with it, and days of computing would be lost.

Model Studio also needs a very high storage capacity, since each training iteration is a network of archived, tested neurons. A model represents billions of parameters, i.e. several gigabytes, which are stored in Object Storage via the OpenStack Swift service organised into containers.



This infrastructure was developed in one year. Over this period, SYSTRAN's teams were able to train hundreds of models using a pool based on NVIDIA DGX-1 servers, as well as complementary Public Cloud pools based on NVIDIA Tesla V100 GPU instances. The platform is now available to trainers in the marketplace, so that they can create their own models independently.

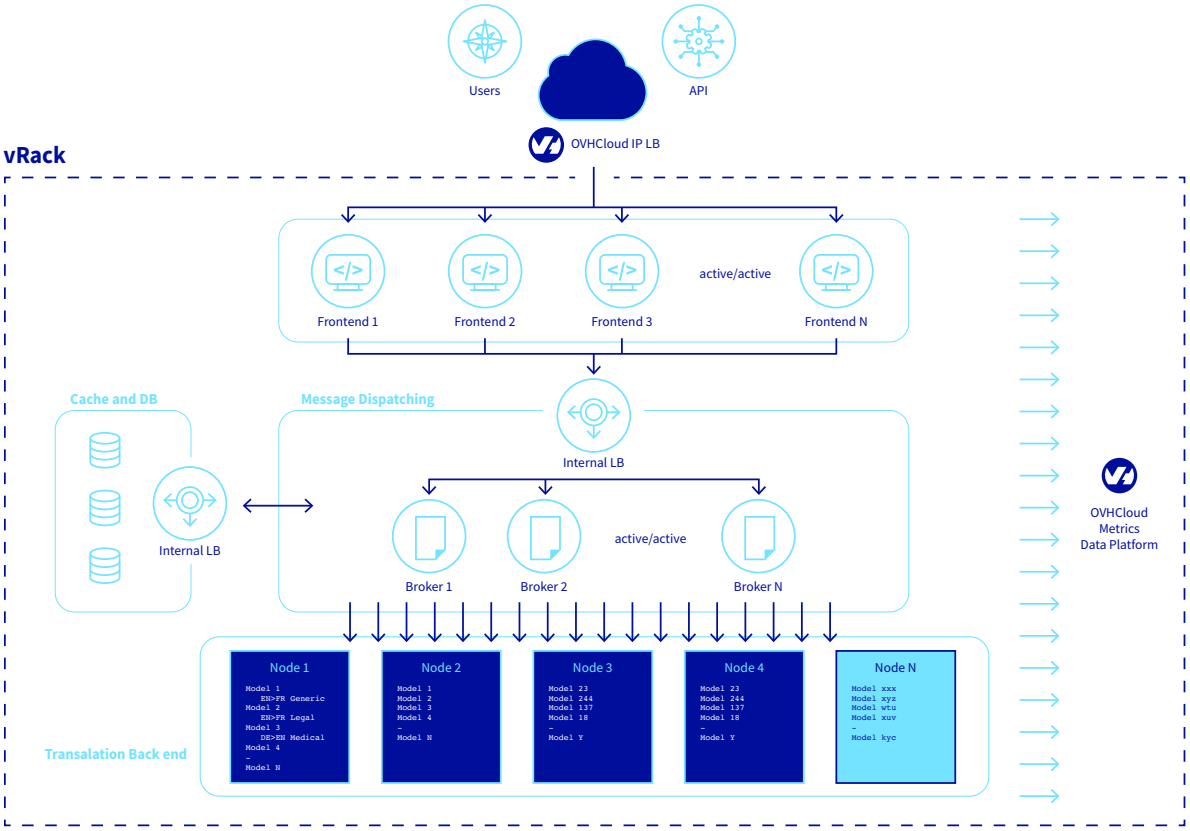
For inference, the opposite issue exists. The service needs to be available 24/7, and must adapt to changing volumes of requests in an instant, while using instances optimised for computing. In addition to this, each request must be processed in a few milliseconds, and requires a mixture of static and dynamic instances.

The entry point of the SYSTRAN Translate platform infrastructure is a load balancer. It has a crucial role, as it balances the load between the different services host in the datacentres, and protects the application against DDoS attacks. This hardware also ensures that the infrastructure is scaled in the event of traffic spikes. Finally, it guarantees high availability for the service and optimises response times.

In July 2019, the infrastructure was made up of 75 Public Cloud GPU instances. This is secured via the vRack, a private connection made in OVHcloud.

To go even further, the teams also added a dynamic component to the service. It is based on Kubernetes, and enables the teams to scale their infrastructure flexibly and instantly.

The infrastructure is monitored by the managed Metrics Data Platform (currently available in France only). This platform monitors the components in real time, and also ensures quick response times and volumes of translation for all language pairs and models.



A platform based on open standards

The development of the entire marketplace infrastructure was made much easier with OVHcloud services. They are all equipped with open-source APIs, and this guarantees that our development teams will quickly find them easy to use.

> *"Our choice to invest in open-source solutions guarantees that our end users benefit from the very best technology on the market, and it also benefits our developers and marketplace contributors because they are not locked in by proprietary technologies."*
>
> **Yannick Douzant, Director of Products & Technologies at SYSTRAN**

SYSTRAN develops and manages all of the neural translation code in the OpenNMT project. OVHcloud takes the same approach, and opted for a number of open standards for its Public Cloud service. In addition to the ease of use it offers, this open-source approach is an important component of the common philosophy shared by the two companies surrounding software development.

A responsible approach

> *"OVHcloud has an eco-friendly approach to server design, operating a system exclusively using water-cooling technology, and implementing a policy that involves developing green energy and recycling components when their hardware has passed its lifecycle. This was a key criteria in our choice for the marketplace infrastructure."*
>
> **Jean Senellart, CEO of SYSTRAN**

The data is protected, and there is an assurance that it will not leave European soil, to ensure compliance with the General Data Protection Regulation (GDPR).

The result

With the technology used and the support provided by OVHcloud experts, it only took two weeks for SYSTRAN's technical teams to deploy the SYSTRAN Translate service and put it online.

Just five months after its launch, the service has already helped more than 1 million users from 190 countries translate billions of words. It is very popular in Europe — particularly in France, the UK, Belgium and Germany.

The automatic translation service works in more than 40 languages, and offers 400 translation models. A year from now, the objective is to reach 5,000 models with the expanding community of experts.

And this is just the beginning, because SYSTRAN Translate only represents the first building block of a new solution aimed at professionals: SYSTRAN Marketplace. Its purpose is to offer the largest catalogue of specialised models, with the widest range of translation solutions deployed either on-premises or in the cloud, privately or publicly. It is designed to respond to all kinds of needs and volumes, with the same level of quality.

OVHcloud is a global, hyper-scale cloud provider that offers businesses industry-leading performance and value. Founded in 1999, the group manages and maintains 30 datacentres across four continents, deploys their own fibre-optic global network and controls the entire hosting chain. Relying on their own infrastructures, OVHcloud offers simple and powerful solutions and tools that put technology at the service of business, and revolutionise the way that our more than one million customers around the world work. Respect for individuals, freedom and equal opportunities for access to new technology have always been firmly rooted principles of the company. *"Innovation for Freedom".*