safespring

SOLUTION BRIEF

Sunet Drive is a GDPR compliant file sync and share solution

For research and education

Table of contents

About Sunet

The Swedish National Research and Education organisation (NREN) Sunet, was founded in the 1980s as a research project for Swedish computer scientists which paved the way for internet in Sweden.

Today Sunet connects 750 000 users and 110 organizations via 8400 km of fiber and also provides other services to support science and research.

About Safespring

Safespring is a provider of infrastructure- and platform-as-a-service, and is the provider of Sunets cloud offering to the Swedish academic sector.

By working closely with Sunet, Safespring has built large scale cloud solutions for Sunet's customers. Safespring's Services is also available on the OCRE framework agreement for all European NREN and it's members.

Background

Increased international cooperation has become one of the driving factors to influencing the success of our innovation ecosystem.

With this comes an ever-increasing amount of data being collected and analyzed, and the challenge of storing and archiving the data. Scientists always find solutions by either solving their storage needs themselves by establishing their own shadow IT in the project, or by using storage services from central IT departments. There are several problems with this local approach:

- Does the storage solution align with the data management requirements from the funding body?
- How can central IT cater for flexible collaboration on the data between researchers on campus and external actors?
- Who pays for the storage of the data once the research project is over?

Public cloud services with a simple billing model and collaboration features have solutions for these challenges but are hard to use for European researchers. This especially applies if the data produced by the project is somehow sensitive. There is always an option to run a private cloud solution but not all universities have an IT department staffed to implement and maintain such a solution, especially when it comes to establishing high standards of physical data center security and other administrative processes.

By growing accustomed to modern cloud storage services, the researchers have a good idea what they want but no way of achieving it and still be compliant to GDRP or other European legislations.

Needs and requirements

Universities all over Europe have common needs and requirements regarding a storage solution.

IT SHOULD BE EASY to use and administrate for all involved parties.

INITIAL COST should be covered by the universities, with an easy and transparent billing model for projects requiring storage of large datasets.

COLLABORATION between institutions, both on and off campus, and external actors should be flexible but still secure.

DATA should be able to be handled from the primary data sources up to archiving of published results.

THE SOLUTION should be based on open standards and API to prevent a vendor lock-in.

Classification

All these requirements are then fit into a general model of lifecycle-managing research data, based on a combination of classification parameters: on-premise or stored in the cloud, smaller or larger files, as well as sensitivity of the data to be stored.

Once these parameters have been determined, the data can be accessed either through a classical file system or modern apps that work on mobile and desktop devices, while always being able to determine where the data is stored physically.

Sunet Drive then enables a seamless lifecycle management of the data during the project execution, through the data retention period, to archiving of the data.

Solution

Sunet Drive is a managed storage solution installed in the university's local IT-environment.

Sunet Drive is a managed storage solution which is physically installed in the university's local data centers and is built on the trusted open-source projects Nextcloud, OpenStack and Ceph. It uses SAML2 federated login to tie together collaborating researchers all over the academic sector. It is built to handle data on petabyte scale and uses Sunet's high performance NREN network for file transfers between universities.

It is designed to solve the needs and requirements listed above to become a smart and long-term solution for Swedish universities to handle their increasing storage demands without compromising on legislative issues such as Schrems II.

Design

Sunet Drive enables universities to provide a storage solution with the same flexibility as many researchers have grown accustomed to, while still being compliant with local, national, and international requirements. This is achieved through a federated global scale architecture that implements a co-management between Sunet and the local university.

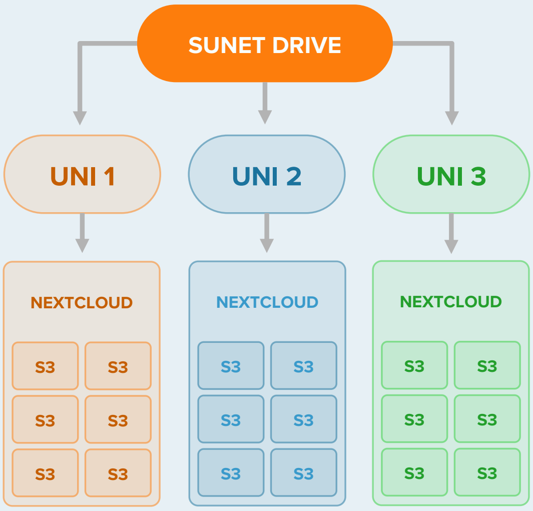
Each organization joining the solution will be able to manage their own Nextcloud node in accordance with commonly agreed standards, while still being able to support local processes and procedures. The federated login through a global site selector binds a user to their respective Nextcloud nodes. This guarantees that users from one organization will only operate on their own node.

When an international user logs on to Sunet Drive, they will be delegated to an external Nextcloud node where a user account will be provisioned, and they can be invited by other users to collaborate.
The main storage for each Nextcloud node is S3-storage running on the same infrastructure as the Nextcloud node. S3 comprises of buckets, which are flexible storage entities that can be handled like virtual hard disks.

Data managers and data access units at the universities can flexibly create S3-buckets and assign them to researchers, projects, institutions, or other logical organizations. This way it is possible to keep data that belongs to a specific project or research group separate. Existing data can easily be integrated and indexed by Nextcloud, which can be used for applications where researchers need to upload the data directly to S3 instead of using Nextcloud. On the other hand provides Nextcloud synchronization clients for all platforms, but also supports file transfer protocols such as WebDAV to act as a local file server.

By using S3 as backend it is always possible to reach the core data directly from the storage solution. This makes it easy to migrate the data to another solution should that need arise. Nextcloud adds user mapping, local synchronization, sharing and collaboration features to the solution. Generally, it acts as a user-friendly frontend to the data stored in the S3-solution.

Sunet Drive is built with separation of access and ownership of the data in mind. That means that it is possible for a researcher to move to another university with another Nextcloud node without complex migration procedures. Even after changing their affiliation, researchers will be able to access their data with minimal administrative effort.



Building Blocks

There are a number of components that combined make Sunet Drive.

Collaboration platform - Nextcloud

Nextcloud is an on-premises collaboration platform. It uniquely combines the convenience and ease of use of consumer-grade solutions like Dropbox, OneDrive and Google Drive with the security, privacy and control large organizations needs.

Users gain access to their documents and can share them with others within and outside their organization with an easy to use web interface or clients for Windows, Mac, Linux, Android and iOS.

Since Nextcloud is an open source project it is possible to integrate with other solutions to tailor to specific needs. Sunet Drive includes a federated login integration with SWAMID which makes it easier to collaborate between different universities but also helps to solve the problem with researchers migrating to another university and how the data should be able to follow them to their new location.

Nextcloud also supports the Global Site Selector functionality which in Sunet Drive is used to provide single sign on to the whole solution. All logins start at a common login page and the user then gets redirected to the correct node, which could be at Sunet or at the users university. The location of the users data is stored in metadata in the SWAMID-solution which makes it easier to track which users reside where.

S3 can be used both as the backing storage for the whole solution as well as connected as separate external drives to the Nextcloud node. The latter is practical to cater for storage needs in a specific research project since all the belonging files could be stored in a S3 bucket which in turn is presented as a separate drive in Nextcloud.

With advanced functionality to handle access to files, directories or external drives (whole S3 buckets) the users can control who gets access to what. It is even possible to give access to external users which is very useful when dealing with projects that spans over both private companies and the academia.

If the sharing settings allow, all files in Nextcloud could be shared with standardized protocols such as WebDAV. This adds a flexibility that standard S3 does not have and makes it easier to build archiving piplines or other kind of data flows.

Object Storage - Ceph

Primary storage requirements for Sunet have been defined as:

- Availability
- Consistency
- Resilience
- Cost

It is significantly cheaper to provide guarantees that meet these requirements per object instead of across a whole filesystem. The fact the S3 has become a de-facto standard as an interface to object storage the requirements for long-term operability and the ability to migrate the data elsewhere are also met.

Ceph, with the RadosGW implementation, is a proven solution for high scale object storage solutions. By providing the same underlying solution to different universities in Sweden possibilities for federation or even a boundless data lake for research opens up.

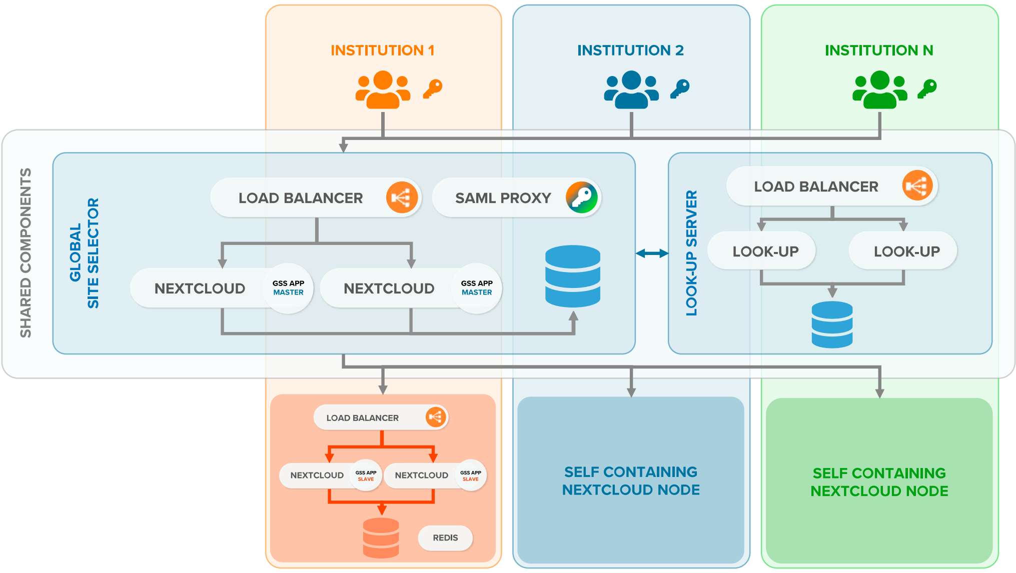
Compute – OpenStack

The OpenStack installation is small in comparison with the Ceph installation since it is primarily used to be able to run the virtual machines needed for Nextcloud and the Galera Cluster. If the university needs further resources for data processing the capacity can easily be increased.

By having compute power close to the data, researchers will have the ability to process and do calculations on the data. The resources will be assigned with flavors, where some will be regular CPU resources and some GPU resources if the university has need for it and has GPU-cards installed in the instance. The instances will also come with fast, local NVME storage for high performance when processing the data.

Hardware Infrastructure

The hardware infrastructure is based on standard i386 nodes placed in the university's IT-environment. Most of the management is done remotely but in the case of physical work such as drive swaps and replacements the local IT-staff perform these assignments.



Implementation

There are a number of components that combined make Sunet Drive.

There are several components that have been integrated to create Sunet Drive. At the bottom we have the self-containing nodes which are complete, redundant installations of Nextcloud with frontends with application logic and database clusters. Container-technology streamlines all DevOps aspects, from automated deployment to maintenance. In the middle of the picture are the shared components operated by Sunet.

In the middle of the picture, we see the shared components running centrally at Sunet.

A Global Site Selector, a redundant Nextcloud node with the GSS master role enabled. This node handles log ins and redirects users to their respective node.

The SAML2 Proxy (for integration with existing federated ID platforms). The GSS uses the SAML2 Proxy node to forward login request which gets forwarded to each university's IDP for authorization. When a successful login attempt is made the information is passed back to the GSS node through the SAML2 Proxy.

The look-up server which makes it possible for the self-containing Nextcloud nodes to look-up users residing in other Nextcloud nodes. This is to make it possible for users in different Nextcloud nodes to find each other for collaboration between different universities and institutions.
A load-balancer service to ensure high resilience and uptime of the solution.

Not in the picture is the test-environment, which is a replicate of the architecture for verification and validation of new features and upgrades.
At the top of the picture, we see the users at the different institutions accessing the service by logging in through the Global Site Selector and being redirected to their university's Nextcloud node, respectively.

Operations

Most of the components of Sunet Drive are managed as code (infrastructure as code), enabling immutable infrastructure divided between stateful and stateless components. Nextcloud frontend servers are a good example for stateless components, they can easily be scaled out by adding more instances. But even the database instances, which are clustered, stateful instances, are mostly decoupled from the underlying S3-storage, containing only general information about operations on files, rather than actual research data.

Compliance

Data compliance is achieved on multiple levels, incorporating processes that support, but are not limited to, processes frameworks like ISO 27001, ITIL, or the Swedish MSBFS 2020:7, which determines regulations on security measures in information systems for state authorities. This is done through a clear distinction of responsibilities of the involved parties. Users and operators gain access to their respective parts of the federated architecture, which then can be aligned with local processes.

User-perspective

The perspective of an end user is as simple as it can get while using an Enterprise File Sync and Share (EFSS) solution. A user logs on through their institutional account and will be delegated to the federated node which is co-managed between Sunet/Safespring and their institution. This means for example that a user can apply for project-specific storage (an S3-bucket), which will then show as a regular folder in Nextcloud and can be synchronized using Nextcloud's standard applications.

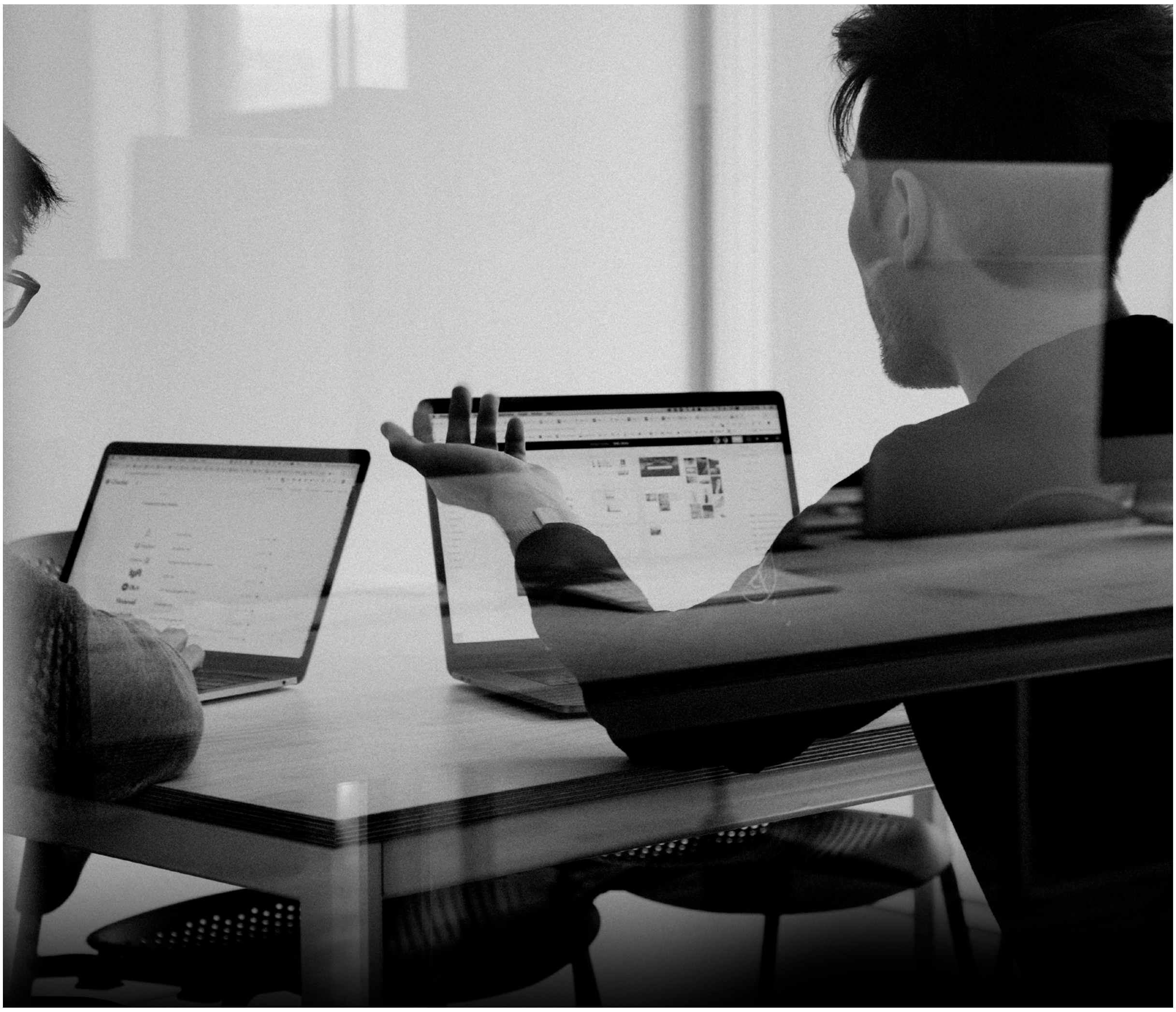
Collaborations with other researchers can easily be established by inviting them into Sunet Drive. Once a project has finished, the properties of the S3 bucket can be changed, and the ownership of the data can be transferred. This also includes the integration of metadata for publications, e.g., through the DORIS from the Swedish National Data service (SND).

Conclusion

Increased international cooperation has become one of the driving factors to influencing the success of our innovation ecosystem.

By using standard hardware and established open-source projects, Sunet Drive provides a storage solution designed to solve the ever-increasing storage needs in the academic sector.

By connecting it to already existing solutions such as SWAMID for identity handling, the researchers will be able to work and collaborate with large datasets. This is combined with predictable cost and the ability to let the data follow the researcher instead of the other way around.



Safespring is the sustainable platform for compliant cloud services

Safespring is a European provider of cloud and infrastructure as a service. Our platform is based on Open Source and open standards. Let us give you the benefits of the cloud where you retain control and digital sovereignty.

www.safespring.com/en

safespring

+46 (0)8-55 10 73 70 | info@safespring.com
Smidesvägen 12, 171 41 Solna, Sweden

www.safespring.com