Nextcloud

Kronenstr. 22A
70173 Stuttgart
Germany

Tel: +49 711 89 66 56 - 0
Fax: +49 711 89 66 56 -10

Web: nextcloud.com
E-Mail: sales@nextcloud.com

Nextcloud Global Scale

Architecture Whitepaper

Credits:

Frederik Orellana (DeiC)

Frank Karlitschek (Nextcloud)

Nextcloud Engineers

Summary

In this whitepaper we describe the current web application architecture Nextcloud uses and its limitations; and introduce Global Scale (GS), an architecture we designed to lift these limitations. GS spreads users and data over independent nodes and introduces 3 components to manage them: the Global Site Selector which enables users to log in from one place; the Lookup Server which mediates sharing and stores certain metadata; and the Balancer which monitors and manages the nodes. We address some other elements that need to be considered (sharing, comments and other metadata and more) and conclude with limitations and side benefits.

Nextcloud

Kronenstr. 22A
70173 Stuttgart
Germany

Tel: +49 711 89 66 56 - 0
Fax: +49 711 89 66 56 -10

Web: nextcloud.com
E-Mail: sales@nextcloud.com

Table of Contents

Nextcloud

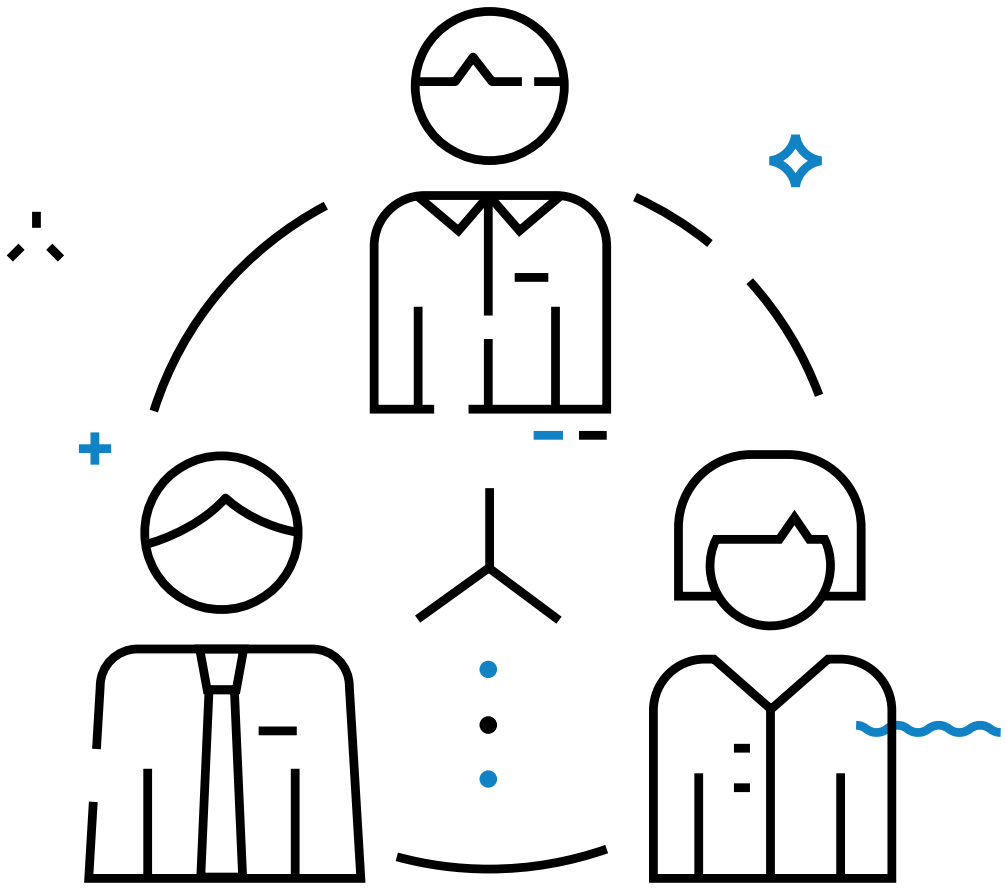
Kronenstr. 22A
70173 Stuttgart
Germany

Tel: +49 711 89 66 56 - 0
Fax: +49 711 89 66 56 -10

Web: nextcloud.com
E-Mail: sales@nextcloud.com

Standard Nextcloud architecture

Currently, Nextcloud is using a standard web application architecture. The incoming user traffic is distributed through load balancers to a number of application servers that run independent Nextcloud instances. These application servers access a shared storage, a shared database and a shared cache. This architecture scales well up to a 6 figure number of users. The application servers are easy to scale because doubling the number of servers doubles the performance. But scalability limitations can be found in the shared components. These are the load balancers, the database, the storage, and the cache. The datacenter uplink also becomes a limitation at large scale. Let us examine the problems with very large installations.



Nextcloud

Kronenstr. 22A
70173 Stuttgart
Germany

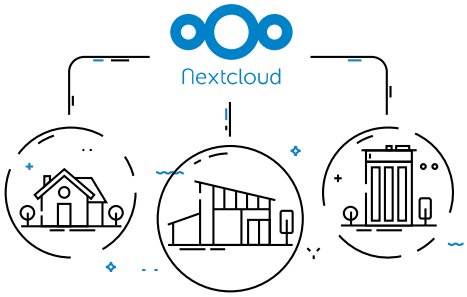
Tel: +49 711 89 66 56 - 0
Fax: +49 711 89 66 56 -10

Web: nextcloud.com
E-Mail: sales@nextcloud.com

Limitations of the standard architecture:

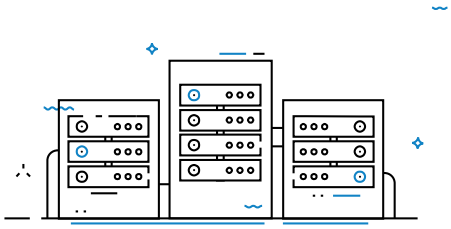
The standard Nextcloud architecture works well for a lot of use-cases but at large scale, three issues become apparent: scalability, costs and the need for more flexibility.

1. Scalability



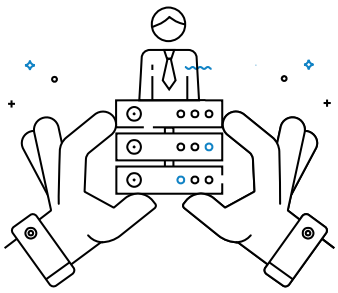
It is hard to scale the standard architecture to instances with a 7, 8 or 9 digit number of users. The shared components, the load balancers, database, storage and cache, as well as the datacenter uplink, will sooner or later become bottlenecks. The database is particularly hard to scale beyond a 4 Node Galera Cluster which limits the number of users and files.

2. Storage Cost



Several big Nextcloud users raised the issue that scaling the storage becomes exponentially more expensive when dealing with many petabytes. Software-defined storage and object stores are unfortunately not a practical solution. Dr. Hildmann from the Technical University of Berlin estimated that 60% to 80% of the cost of running a file sync and share service is caused by the storage subsystem alone. It would significantly lower the total cost of ownership if the storage cost could be reduced by not being forced to rely on expensive, high-end large scale storage systems.

3. Global distribution



Some Nextcloud users have the need to run a Nextcloud instance that is distributed over several hosting centers, countries or continents. The reasons include the need for data locality for legal reasons, the need to leverage fast local networking speed or the fact that some parts of the instance have to be operated by a different group of administrators or that different auditing or security rules apply. These requirements are hard to

implement with the standard Nextcloud architecture. There are solutions to replicate storage across continents and the same is true for some databases or caches. But there is no technology available that can replicate this in a synchronous way so that file system, database, and cache changes are bound together atomically in replication transactions. And even if this could be done then it would be very slow and require massive bandwidth.

Goals:

The goal of the Nextcloud GS architecture is to solve these three architecture limitations. Specifically, the goals are to build an architecture which:

- Is highly scalable up to a 9 figure number of users.
- Makes it possible to distribute an instance over different hosting centers and continents
- Reduces hardware costs by leveraging commodity hardware
- Reduces maintenance costs
- Has no significant software license costs besides Nextcloud
- Gives the option of starting with a small installation and scaling up over time.

Nextcloud

Kronenstr. 22A
70173 Stuttgart
Germany

Tel: +49 711 89 66 56 - 0
Fax: +49 711 89 66 56 -10

Web: nextcloud.com
E-Mail: sales@nextcloud.com

Approach

The approach is to remove the shared components of the standard architecture and move the logic up to the application server level. Shared components are the load balancers, the hosting center uplink, the database, the storage, and the cache. It has to be possible to run the application servers on standard, inexpensive commodity hardware to save costs and increase flexibility. Storage, database, and cache should be running locally on the application servers.

Architecture

A Nextcloud Global Scale instance consists of many independent application servers called Nodes. There are no central database, storage or caching instances. A Node is a tuple of two machines that share a cluster file system and a cluster database and a common cache. This guarantees the high availability of the Nodes. If one machine of a Node goes down then the other machine takes over. A Node could consist of more than two machines. It could even be a full standard clustered Nextcloud architecture instance if it makes sense. For the context of this document, we presume a Node to be as small as possible which is a tuple of two machines.

These Nodes can be located in different hosting centers. No fast interconnect between the sites is necessary.

Every user is local to a Node so all the user data exists only in this one location. All sharing is done via federated sharing even with users on the same Node. This is important because the location of users might change if the user is moved to a different Node.

Kronenstr. 22A
70173 Stuttgart
Germany

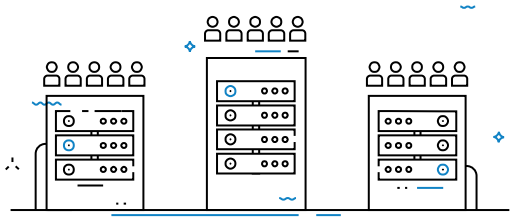
Tel: +49 711 89 66 56 - 0
Fax: +49 711 89 66 56 -10

Web: nextcloud.com
E-Mail: sales@nextcloud.com

Components

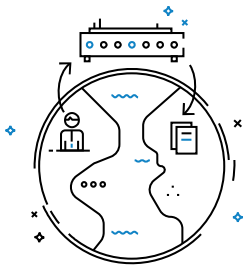
Global Scale consists of Nodes while the interaction between users, Nodes, shares and the Nodes themselves are managed by three systems: the Global Site Selector (GSS), the Lookup Server (LS) and the Balancer.

Node



This is a tuple of machines which acts as one logical Nextcloud instance. These are Linux servers using commodity hardware. They are running a web server including TLS, Nextcloud, local storage, local database and local cache. These components are shared between the machines of the Node. Possible software candidates are GlusterFS, a MySQL Cluster or Galera and Redis. The Nodes use central remote logging and a central authentication directory, for example LDAP. The software on the Nodes is packaged and distributed via, for example, a container technology like Docker to enable easy deployment of new Nodes.

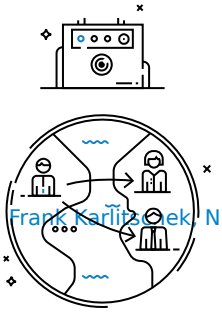
Global Site Selector



The Global Site Selector (GSS) acts as a central instance that is accessed by the user during the first login. The main URL of the services is mapped to the GSS as a publicly visible endpoint. A user accesses the global site selector via Web, WebDAV or REST. The GSS authenticates the user via the central user management system. Then it looks up the node where the user is located in the Lookup Sserver and redirects the user to the right hostname. The following calls during the same session are done directly from client to their Node. This has the benefit that the load on the GSS is low and that clients can leverage fast direct network speeds if the Node is located near the client.

The GSS component is implemented as a Nextcloud App.

Lookup Server

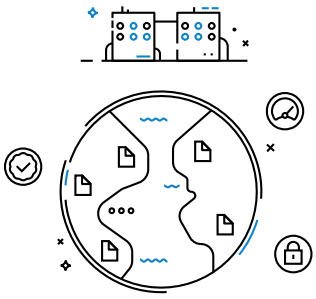


The Lookup Server is a separate server component that stores the physical location of a user. It can be queried using a valid user ID to fetch the

federated sharing ID of a user. In some situations, it is important to limit queries to a certain IP space to avoid data leaks. For the purpose of migrating users and sharing links between instances it is also necessary to store a list of old federated sharing IDs.

The Lookup Server is also responsible for storing additional policy data about the users, for example the required Quality of Service level. That would contains settings for storage/quota settings, speed class, and reliability class. The user's geo-location is also relevant and the Lookup Server also keeps track of that.

Balancer



This is a component that is implemented as a Nextcloud app. It runs once per service as dedicated machine and monitors the storage usage, CPU and RAM load, network utilization and up time of all Nodes. It can mark Nodes as online or offline and can initiate the migration of user accounts to different Nodes.

The decision to move a user between Nodes is done based on user QoS settings and Lookup Server data. Factors like the SLA of the user, physical location, and proximity of Nodes are also considered.

Kronenstr. 22A
70173 Stuttgart
Germany

Tel: +49 711 89 66 56 - 0
Fax: +49 711 89 66 56 -10

Web: nextcloud.com
E-Mail: sales@nextcloud.com

Other elements

A number of other factors need to be adjusted or taken into account when implementing Nextcloud Global Scale. We need a strategy for migration; sharing; exchanging metadata like comments and activities; and backup.

Migration

Moving users is done as follow~~sing~~:

1. Switch user offline at server 1
2. Export user at server 1 (for better performance can be done earlier)
3. Import user at server 2
4. Update lookup server
5. Switch user online at server 2
6. Delete user at server 1

These migration steps can be triggered with OCS calls from a remote server.

Public sharing links

Public sharing links are always handled by the Nodes directly and are not redirected by the Global Site Selector. Because of that, a system is required to keep sharing links working if a user is migrated to a different Node. Another new Nextcloud app is required that stores old public links and redirects to the new Node if they are accessed. The lookup server is queried to get the old and current federated sharing ID.

Federated comments and activities

To make it possible to see activities from other Nodes and enable federated comments a new API needs to be implemented. We will use a standard API for this, the ActivityPub API as recommended by W3C.

Backup and replication

If a Node fails completely no data should be lost. This means that a full backup strategy is important. Nodes should backup automatically in the background to other Nodes. This backup can happen asynchronously to prevent any performance penalty. The number of redundant copies should be configurable. A backup is the full storage of the instance and a database clone. It has to be possible to switch on a backup Node without significant downtime.

Nextcloud is designed to keep data under control of the administrator of an instance, so extensive caching between servers which share data is currently not implemented. However, in a globally distributed instance, it might make

sense to automatically replicate some data between sites for performance reasons.

Conclusion

With the above architecture, a single Nextcloud instance should scale from tens of thousands to hundreds of millions of users. It achieves the goals we have set: dramatically lowering costs by reducing complexity, and enabling administrators to control where user data is located in a fine-grained way.

Limitations

Auditing, configuration, and logging are distributed which might make operations more difficult. On the other hand specialized high load Galera Cluster and Storage Subsystems know-how is no longer needed and high performance, scalable systems to handle configuration and logging data already exist.

All operations and sysadmin tasks have to be highly automated to avoid overhead while managing the high number of commodity servers. Luckily there are plenty good orchestration tools for this purpose already available.

Side benefits

Most of the GS features are also useful for 'normal' home users. They can collaborate with others but also move their Nextcloud instance to another provider or host or device without losing shares or connections to other instances. The federation of activities, comments and more

Some of the GS features like federated activities are also a big step towards a federated social network.

Current state

As per the release of Nextcloud 12 on May 22, we introduced implementations of the Global Site Selector, Lookup Server, and federated activities. Federated comments and other metadata, the Balancer and migrations are work in progress.

The architecture as described in this whitepaper is a work in progress and it should be considered a ‘living architecture’, where we will adjust the implementation and technical details based on real world feedback from large deployments at universities and other customers we are working with.