Westfälische
Wilhelms-Universität
Münster

The Infrastructure of sciebo, the NRW academic Sync and Share service

30.8.2015

Holger Angenent
Röntgenstr. 7-13, 48149 Münster

wissen.leben
WWU Münster

Zentrum für
Informationsverarbeitung

> sciebo

- sciebo is the brand name, short for “science box”
- Sync and share service for universities and universities of applied sciences in North Rhine Westphalia
- Up to 500,000 users in NRW (if all 30 universities and applied universities participate)
- Organized as a consortium of all participants with the University of Münster as the head

> Infrastructure for large sync and share installation

What components are needed:

- Filesystem
- Database
- Loadbalancers
- Webserver
- Management
- Identity Management/Additional Tools

> Filesystem

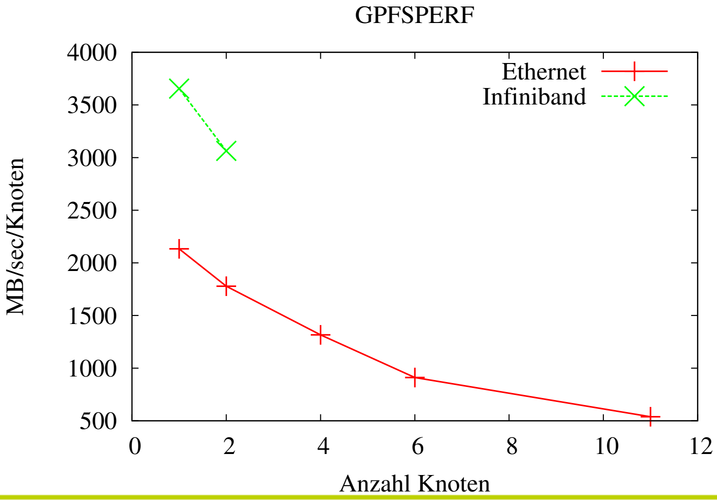
- Store lot’s of data
- Performance: Throughput and IOPS
- Reliability: Harddrives will fail (all the time), maybe even fileservers

> Filesystem Benchmarks

We made some benchmarks on storage cluster in advance

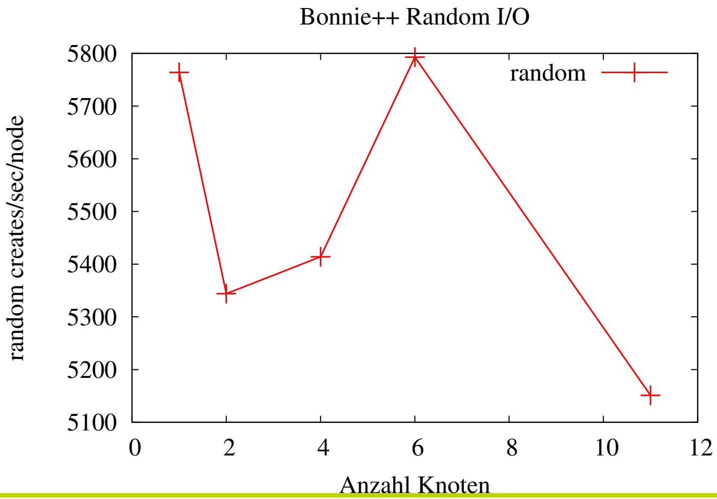
- 11 nodes
- 16 Xeon-cores 2,9 GHz
- min 8 GB RAM
- 2 with Infiniband
- 9 with 10 GBit Ethernet

> Throughput



wissen.leben WWU Münster

> Random I/O



wissen.leben
WWU Münster

> Database

Important for databases:

- Redundancy ⇒ cluster
- Performance ⇒ more RAM, tuning (mysqltuner), cluster, spread data to different databases (federated cloud sharing needed for this)
- Reliability ⇒ robust code

> Database

For example: This might become a problem?

```
\OC_DB::beginTransaction();
$query = \OC_DB::prepare(’UPDATE ‘*PREFIX*filecache‘ SET ‘path‘ = ?, ‘
  foreach ($childEntries as $child) {
$targetPath = $target . substr($child[’path’], $sourceLength);
\OC_DB::executeAudited($query, array($targetPath, md5($targetPath), $c
}
\OC_DB::commit();
```

Missing exception handling (try and catch):

http://php.net/manual/en/pdo.transactions.php

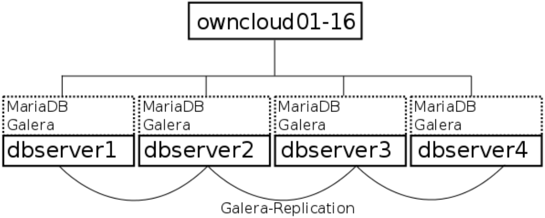
wissen.leben
WWU Münster

> Cluster setup

- ownCloud recommended us to use a master-master setup
- Every database node has the same data as the others, kept in sync by Galera replication
- If data is written to one of the nodes, it is transferred internally to the other nodes ⇒ Write actions (almost) cannot be scaled via adding more cluster nodes!
- On a database cluster with 20 million files we see: Reads: 70%, Writes: 30 %

> Problems with database clusters

- What happens, if you use this for ownCloud:
  `Serialization failure, Deadlock found when trying to get lock #14757`
- Different webservers try to write to different database servers, but database might be locked and you get a conflict
- Workaround: Tell your database loadbalancers to write to a single database node ⇒ the error vanishes



> Loadbalancers

- Fast failover necessary
- Even more important: Only put load to servers that are really online and not just pretend to be online

In /etc/keepalived/keepalived.con

```
MISC_CHECK {
    misc_path /some_path/check_sciebo40.sh
    misc_timeout 6
}
```

Use for example curl in the script

```
curl --resolve $SRVURL:443:$SRVIP https://$SRVURL 2>/dev/null
   | grep -E -q $SUCHE && GEFUNDEN=1
```

wissen.leben
WWU Münster

> Webserver

Apache tuning

- php-fpm
- Threading instead of processes
- Use recommendations from owncloud.org

Other components

- Use a recent version of php (and find a repository for your distro)
- Increase php timeouts
- Use some kind of monitoring for your server load

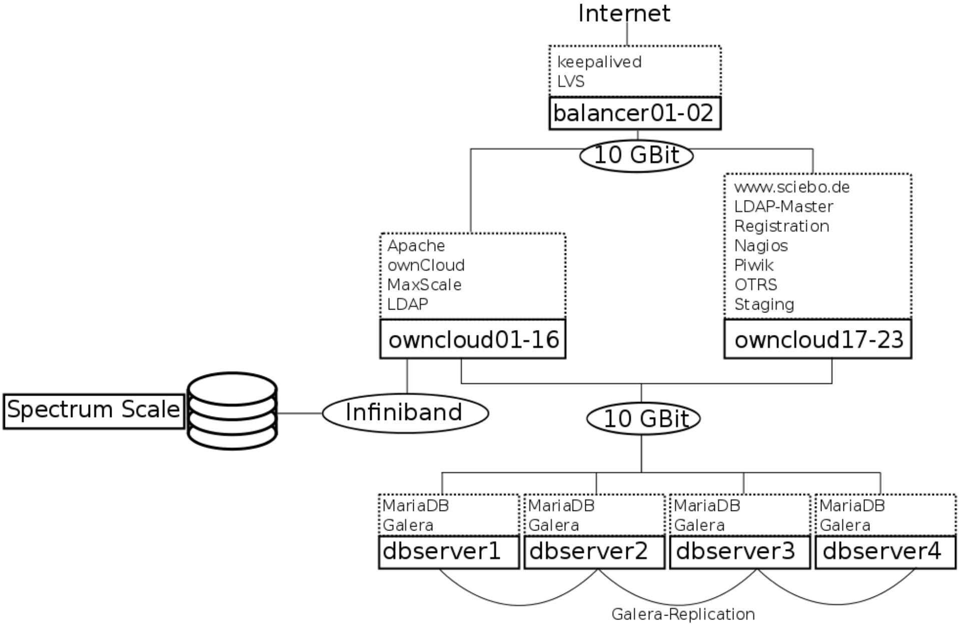
Horizontal Scaling: If performance is not sufficient, use more webservers

> Management

- How to roll out configuration changes to dozen of servers? ⇒ configuration management tools like ansible, salt, puppet or chef come in handy
- Asking Google for example for “ansible switch owncloud” gives you very well done ansible scripts

> User Management

- Thousands of users cannot be managed via local accounts -> LDAP or Shibboleth needed
- We wrote our own portal to fill the LDAP server
- Are special accounts needed (guests or bigger accounts)? -> Find a method to create and manage them
- Users will leave the institution somewhen -> Accounts have to be deleted after a while
- For performance reasons, it is better to have a replica of the LDAP on each webserver


Internet
keepalived
LVS
balancer01-02
10 GBit
Apache
ownCloud
MaxScale
LDAP
owncloud01-16
www.sciebo.de
LDAP-Master
Registration
Nagios
Piwik
OTRS
Staging
owncloud17-23
Spectrum Scale
Infiniband
10 GBit
MariaDB
Galera
dbserver1
MariaDB
Galera
dbserver2
MariaDB
Galera
dbserver3
MariaDB
Galera
dbserver4
Galera-Replication

> sciebo

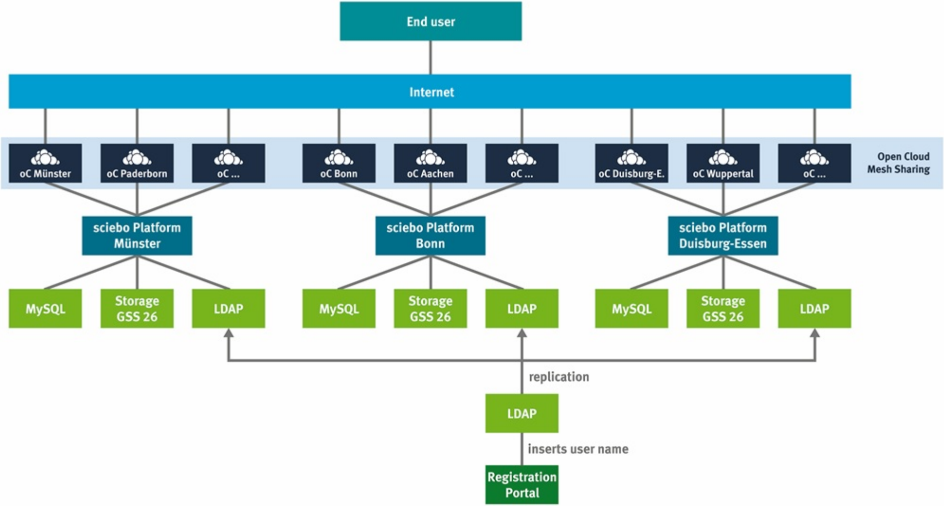
- Started in February 2015 with ∼ 12 institutions
- 30 GB/user with the possibility to increase the quota for the employees up to 500 GB
- Project boxes for groups (30 GB - 2 TB)
- ownCloud 7.0.6 Enterprise (We start to upgrade to 8.1.1 next week)
- RedHat 6/7

> Three Sites

- One ownCloud installation per institution
- (At the moment) No replication of user data between the sites
- Sharing between institutions via the server to server sharing/federated cloud sharing (ownCloud 7/8)

Reasons:

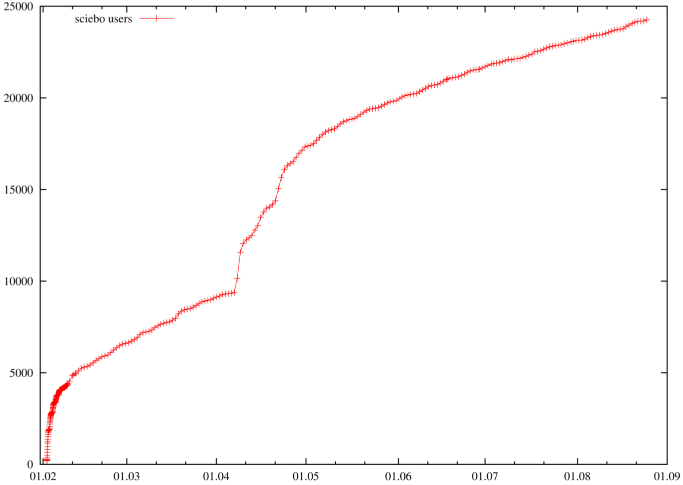
- More aggregated network bandwith
- Servers are closer to users
- If data must not be stored outside of your institution, an own server site could be easily integrated (via federated cloud sharing)


End user
Internet
oC Münster
oC Paderborn
oC ...
oC Bonn
oC Aachen
oC ...
oC Duisburg-E.
oC Wuppertal
oC ...
Open Cloud Mesh Sharing
sciebo Platform Münster
sciebo Platform Bonn
sciebo Platform Duisburg-Essen
MySQL
Storage GSS 26
LDAP
MySQL
Storage GSS 26
LDAP
MySQL
Storage GSS 26
LDAP
replication
LDAP
inserts user name
Registration Portal

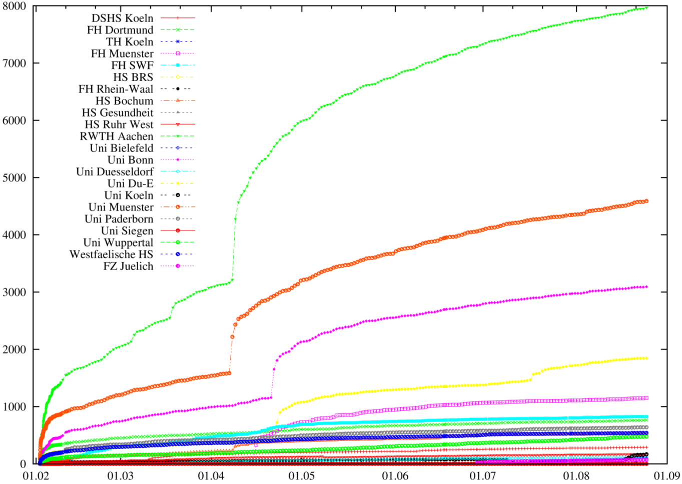
> Hardware

- Storage: 5 PB at three sites (3 PB in Muenster, 1 PB in Essen and 1 PB in Bonn)
- Sizing of server hardware bases on POC of IBM and ownCloud
- Filesystem: IBM Spectrum Scale with declustered RAID and triple parity
- Application servers: 16 IBM servers (16 cores, 128 GB RAM) for ownCloud per site
- Database: 4 IBM servers (20 cores, 256 GB RAM, 6 800 GB SSD [RAID 10]) per site
- Loadbalancers: 2 Linux machines per site (LVS with keepalive)
- Management: 1 server per site

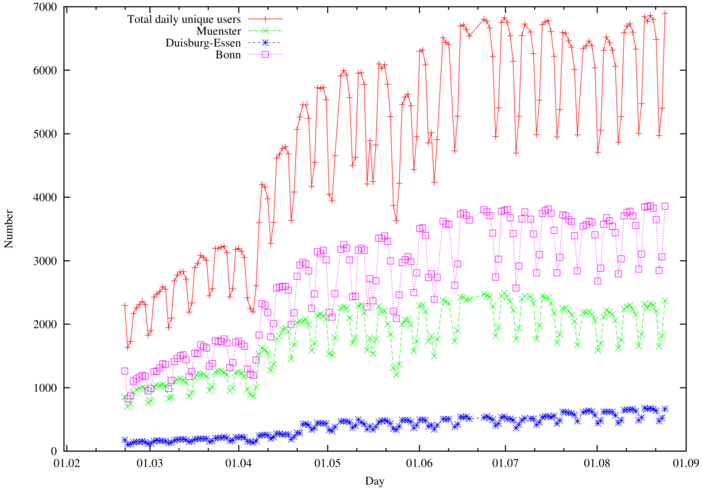
> Registered users



> Where do the users come from



> Actual users



> Summary

- Clustered Infrastructure necessary
- Many challenges
- Reliability, stability and performance more important than new features
- Very positive user feedback

Thank you for your attention!

wissen.leben
WWU Münster