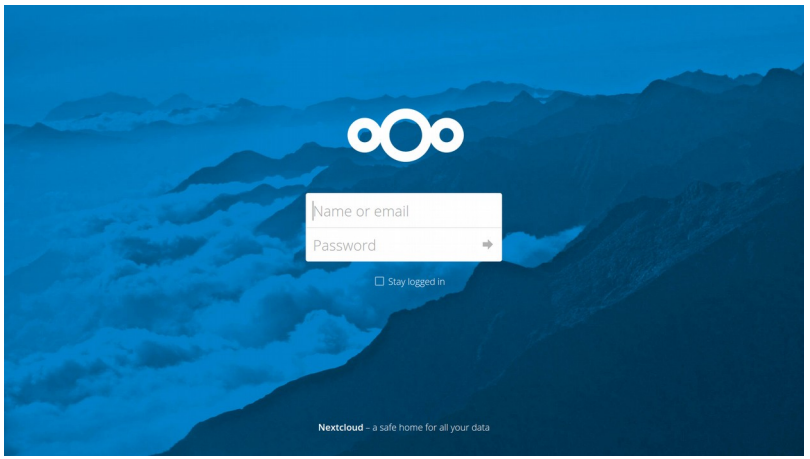

360 ICT
Owncloud scalability and a Nextcloud design for 10.000-20.000 users.

Introduction

- Dennis Pennings
- 360 ICT (.nl)

The goals

Design a 20.000 user NC implementation.



Documentation
(docs.nextcloud.com)

Recommended System Requirements

4 to 20 application/Web servers.

A cluster of two or more database servers.

Storage is an NFS server, or an object store that is S3 compatible.

Cloud federation for a distributed setup over several data centers.

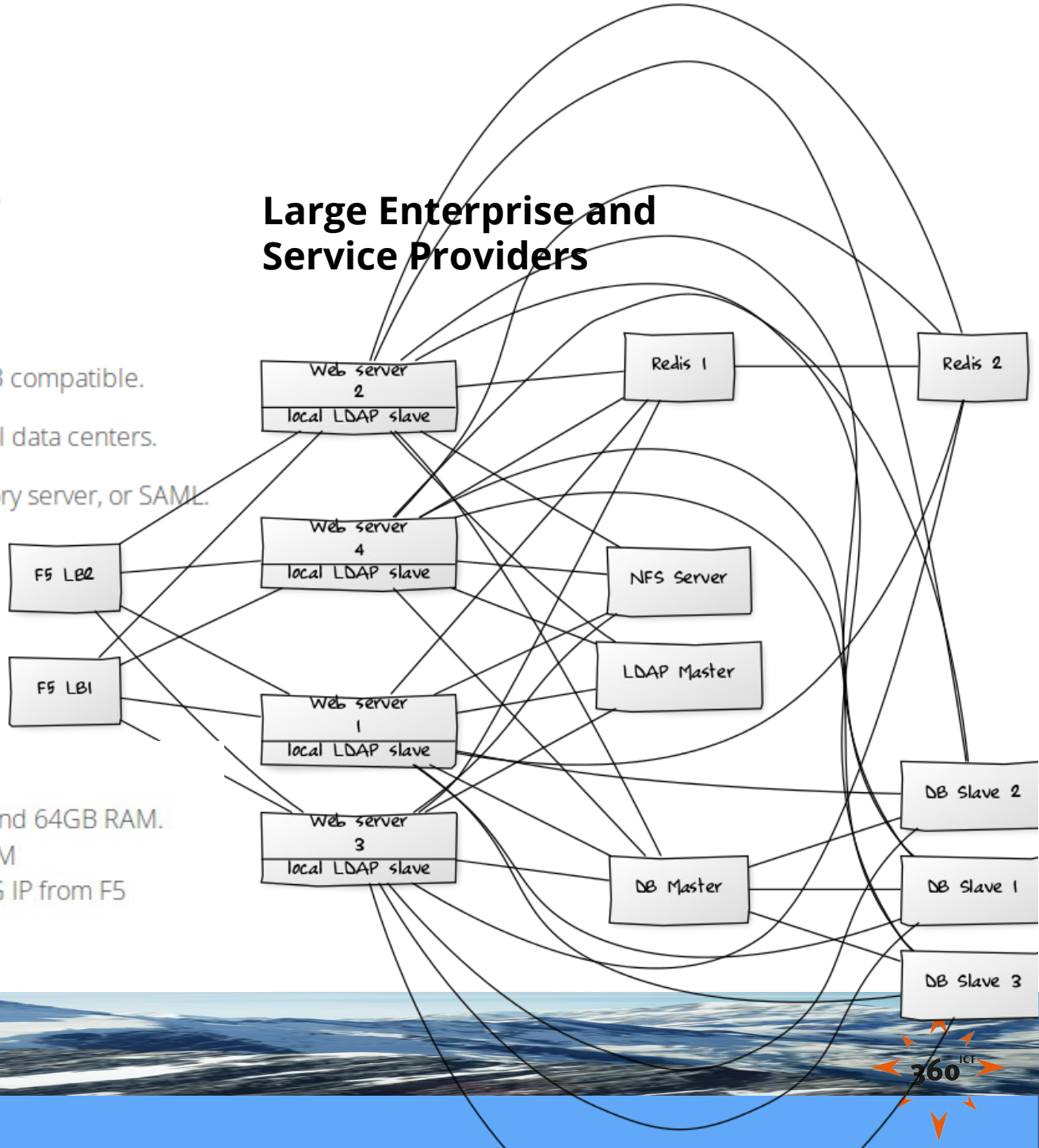
Authentication via an existing LDAP or Active Directory server, or SAML.

100.000users / 1Pb
Database: MySQL/MariaDB Galera Cluster with 4x master – master replication.

Multisite?

- **Components**
  - 4 to 20 application servers with 4 sockets and 64GB RAM.
  - 4 DB servers with 4 sockets and 128GB RAM
  - 2 Hardware load balancer, for example BIG IP from F5
  - NFS storage server as needed.

Large Enterprise and Service Providers



Documentation
(docs.nextcloud.com)

Large Enterprise and Service Providers

- Additional notes
  - Use LDAP slaves on webservers.
  - Use SSL offloading on load balancers.
  - Redis for session management storage.
  - Redis for in-memory caching.
  - Redis: provides persistence, nice graphical inspection tools available, supports Nextcloud high-level file locking.
  - Memcached if shibboleth is used.

(Somewhat) Large Deployments

- Info collected from CS3 2016 cs3.ethx.ch
- Presented on Nextcloud conference with a Q&A from Frank.
  - 360ict.nl/blog
  - youtube.com/nextcloud
- All data in a sheet on 360ict.nl/blog and I will update it as I receive new information.

(Somewhat) Large Deployments

Name	CERNbox	SURFdrive	u:cloud	MyCore	SWITCHdrive
Date of inquiry	Jan 2016	Jan 2016	Jan 2017	Jan 2016	Dec 2016
Users					
total users	5k	13k	10k	4k-15k (jan-dec)	23k
max number of users					
concurrent per day		4000	3600 seen, 2000 average		10k per week
number of files	55m (500m on EOS)	75m	5M		60m
<10Mb		95%	98%		
<1Mb		90%	85%		
data over multiple users devided			2000 users have 10Tb		
quota per user	1Tb	8Gb	50Gb	20Gb	50Gb
Linux:mac:win %				20-20-40	
costs		18eu per user per year			
Database					
type	none	MySQL	MySQL	MariaDB	PostGress → mariadb
setup	none	Galera		Galera-clustercontrol	Galera
number of nodes	none	4 physical nodes	1	3 VMs	4 VMs
cpu/mem	none	32 cores, 256Gb	2 cores, 12 Gb	8 cores, 16Gb	46 vCPU, 250Gb
storage	none		200GB	local SSD storage	local SSD
size of database					
Network					
LB engine webservers					
LB engine webservers			F5	2 VMs, 4 cores, 2Gb	Haproxy, 4 vCPU / 4GB
LB engine database		Haproxy → maxscale		Haproxy	
Bandwith Inet/DB/storage					
Inet bandwith used					extremely variable, peaks at 8Gb/s

(Somewhat) Large Deployments

Name	Cloudstor	Sciebo	UniPiBox	Polybox
Date of inquiry	Jan 2017	Dec 2016	Jan 2016	Jan 2017
Users				
total users	30k	25k		21000
max number of users		500k		
concurrent per day	3k per day	7000	75	5000-6000
number of files	43m	20m	350k	35m
<10Mb	96%			99%
<1Mb	82%			
data over multiple users devided				
quota per user	100gb, 1Tb per project (change on request	30Gb users, 500Gb for employees, 30Gb-2Tb for projects	10Gb/3Gb	
Linux:mac:win %	53-7-40			
costs				
Database				
type	Mariadb	MariaDB		Percona XtraDB Cluster (MySQL)
setup	Galera	Galera		Galera
number of nodes	12	4 physical nodes		3 VM's active, 4th on standby
cpu/mem	32 cores, 256Gb	20 cores, 256Gb		16 core, 40Gb
storage	local SSD	800Gb SSD with raid10		100Gb SSD per node
size of database				35Gb
Network				
LB engine webservers		LVS with keepalive		LVS with keepalive
LB engine webservers	HA proxy	8 core Linux machines per site		
LB engine database	maxscale	Maxscale		Maxscale
Bandwith Inet/DB/storage	120Gb L3VPN	10Gb/10Gb/56Gb		10Gb/10Gb/10Gb
Inet bandwith used		80MB/sec		6.2MB/s (3.5Tb per week)

(Somewhat) Large Deployments

Name	CERNbox	SURFdrive	u:cloud	MyCore	SWITCHdrive
Date of inquiry	Jan 2016	Jan 2016	Jan 2017	Jan 2016	Dec 2016
Storage					
number of nodes	1300	9		2 arrays, each with 6 nodes	
number of disks	40k disks				
cpu/mem				8 cores, 12Gb, 800Gb SSD	
brand/type				Dell PowerEdge R620/630/MD3420	No-name
software	EOS	GlusterFS (distr-repl)	Object storage, scality, Fuse	Scality	nfs on top of Ceph
Total space	1,3PB (64PB on EOS)	100Tb	1.8Pb		1.6Pb
space in use	104TB		10Tb (OC)		58Tb
Issues		add server takes 2 months to rebalance. Backup issues			after 4k users, Db to local SSD, ceph volumes from 100Tb lazy to 2Tb
Futureplans		GPFS, EOS, dCache			object storage, Owncloud sharding with Federation between instances
Webserver					
number of nodes		12	1	6 VMs	10 sync / 4 web
cpu/mem			4 cores, 20Gb	8 cores, 24Gb	Sync: 16GB, 16vcpu / Web: 4Gb, 4vcpu
webengine		Apache	Apache		Apache
php version			5.6		5
OC version	8	8.2	8.23	8.2	8.1
other info					
Issues				Version app causes too much load	
OS			CentOS		
identity			Shibboleth + LDAP	shibboleth	ldap
Uses docker	no	no	no	no	yes
management tools					ansible
Other					

(Somewhat) Large Deployments

Name	Cloudstor	Sciebo	UniPiBox	Polybox
Date of inquiry	Jan 2017	Dec 2016	Jan 2016	Jan 2017
Storage				
number of nodes	12 nodes	10 nodes	2X RADOS-GW, 8 CORES, 16gB, 2X KEYSTONE, 2 CORES 1gB	
number of disks	504			
cpu/mem	Xeon E5, 128GB	20 cores, 256Gb		
brand/type	supermicro	IBM GSS appliance		IBM SoNAS
software	EOS	GPFS	Ceph, radosgw+keystone for swift	NFS
Total space	1.9PB	5PB	2PB	
space in use		800Tb (including snapshots)		62TB
Issues		none		none
Futureplans	Cernbox		30 SITES, DNS-ha AS lb, Ceph, , xtradb cluster, swift storage	
Webserver				
number of nodes		16 physical nodes	2 VMs	8 VMs
cpu/mem	48	16 cores, 128Gb	4 cores, 16GB	16 core, 16Gb
webengine	Apache	Apache		Apache
php version	5,6	7		5.4
OC version	8.2	9.06		8
other info				
Issues	Database design is rubbish, inefficient query and query caching, redundancy of metadata, poor group sharing design, flawed assumptions and key selections, not built to scale			
OS	RHEL7.3	Redhat 6/7		RHEL7
identity	SAML	Shibboleth for registration, LDAP to log in to OC		
Uses docker	yes	no	no	no
management tools	ansible, rancher, jenkins	Ansible	puppet	Zabbix, Splunk
Other	Cluster spans a 65ms network end to end. Looking at moving to Cernbox to avoid the database issue.	Users distributed to three sites, backup, but no load balancing between sites. One OC instance per institution		

(Somewhat) Large Deployments

- A summary (mostly cs3-2016 info)
  - *8 different implementations. 5k-30k unique users. 2k-10k concurrent users.*
  - *Mostly 4 node MySQL-MariaDB Galera with maxscale. 2-46 cores, 12-256Gb mem. One node which handles writes (with failover).*
  - *Webservers, Apache, php5, OC8/OC9, 1-48 cores, 16-128Gb mem.*
  - *100Tb-800Tb in use, 1.3-5Pb allocated.*
  - *A few use docker.*
  - *Scaling problems/limit with DB.*

Not much differences here.

(Somewhat) Large Deployments

- Storage looks very different
  - *GlusterFS*
  - *Ceph*
  - *Ceph-NFS*
  - *NFS*
  - *Scality-Fuse*
  - *EOS*

Initial Concept Design

	NextCloud	360 ICT concept design
Users		
total users	5k	10k
max number of users	100k	20k
concurrent per day		
number of files		
quota per user		5Gb
Webserver		
number of nodes	4 to 12	4-n VM's
cpu/mem	4 sockets / 64Gb	1-4 core, 10Gb-20Gb
webengine	Apache	Apache
php version	5.5+	7
OC version	Enterprise Edition	NC11 (because of the S3 improvements)
Database		
type	MySQL/MariaDB (Oracle/PostGreSQL is supported)	MariaDB
setup	Galera	Galera
number of nodes	4	4 VM's
cpu/mem		4-24 cores / 24-256 Gb
storage	SSD	SSD

Initial Concept Design

	NextCloud	360 ICT concept design
Network		
LB engine		Kemp VLM or maxscale
LB hardware	Big IP F5	Kemp VM / appliance or VM with maxscale
Storage		
number of sites		1 to 2
number of nodes		4-n nodes
cpu/mem		4-8 cores,
brand/type		Dell PowerEdge
software	NFS/S3, Elastic Storage, Ceph, GlusterFS, GPFS, Swift	Swift (openstack or ceph)
Total space	1 Petabyte	50Tb to 100Tb
Other info		
OS	RHEL7	Ubuntu 16.04
identity	LDAP / AD / SAML	MS AD or Shibboleth
Uses docker		yes
management tools		openstack, swarm

Q&A for our initial design

- Some highlights from the Nextcloud Q&A session
  - *Multiple datacenters will be hard because the dependency of DB and storage which need to be in sync.*
  - *Docker supported by Nextcloud (OwnCloud does not)*
  - *No preference for database, mysql/mariadb are mentioned as a starting point, but others are supported (like Oracle). That's true for Loadbalancers too (haproxy/maxscale/hardware).*
  - *Nextcloud has no preference on OS, any version with (possible) enterprise support is supported*

Q&A for our initial design

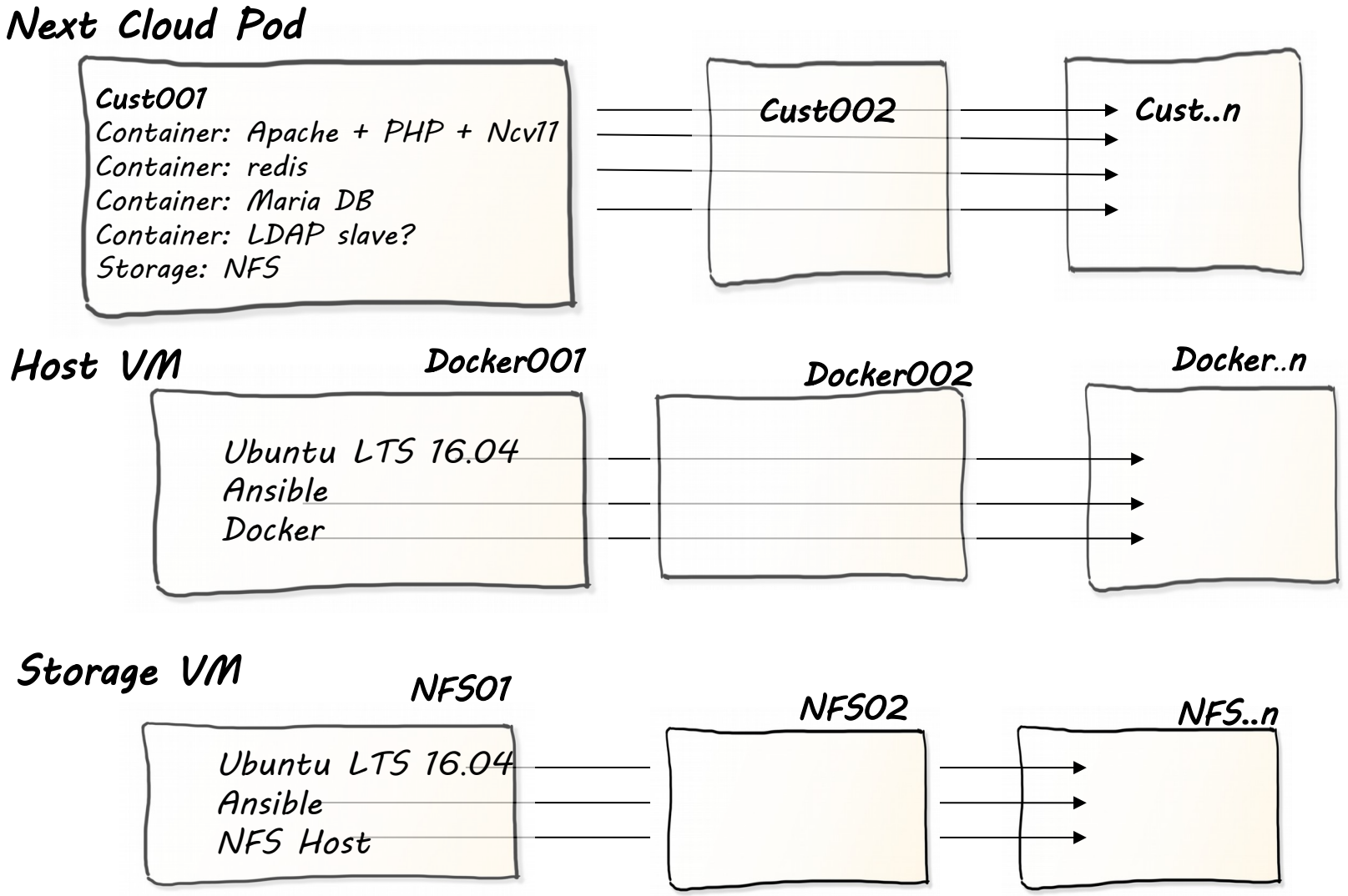
- Some highlights from the Nextcloud Q&A session
  - *Object storage is supported, multibucket storage available in NC11/OC9.*
  - *Use php v7 if possible if your distro supports it.*
  - *There's definitely a scale limit to the database. NC11 claims better database scalability (bold statements like 80% less queries)*
  - *Storage is a concern, but in all fairness this is a client specific issue. OC/NC support lots of solutions.*
  - *The full session from the Q&A session (1 hour) can be found on 360ict.nl/blog and youtube.com/nextcloud.*

Our current design

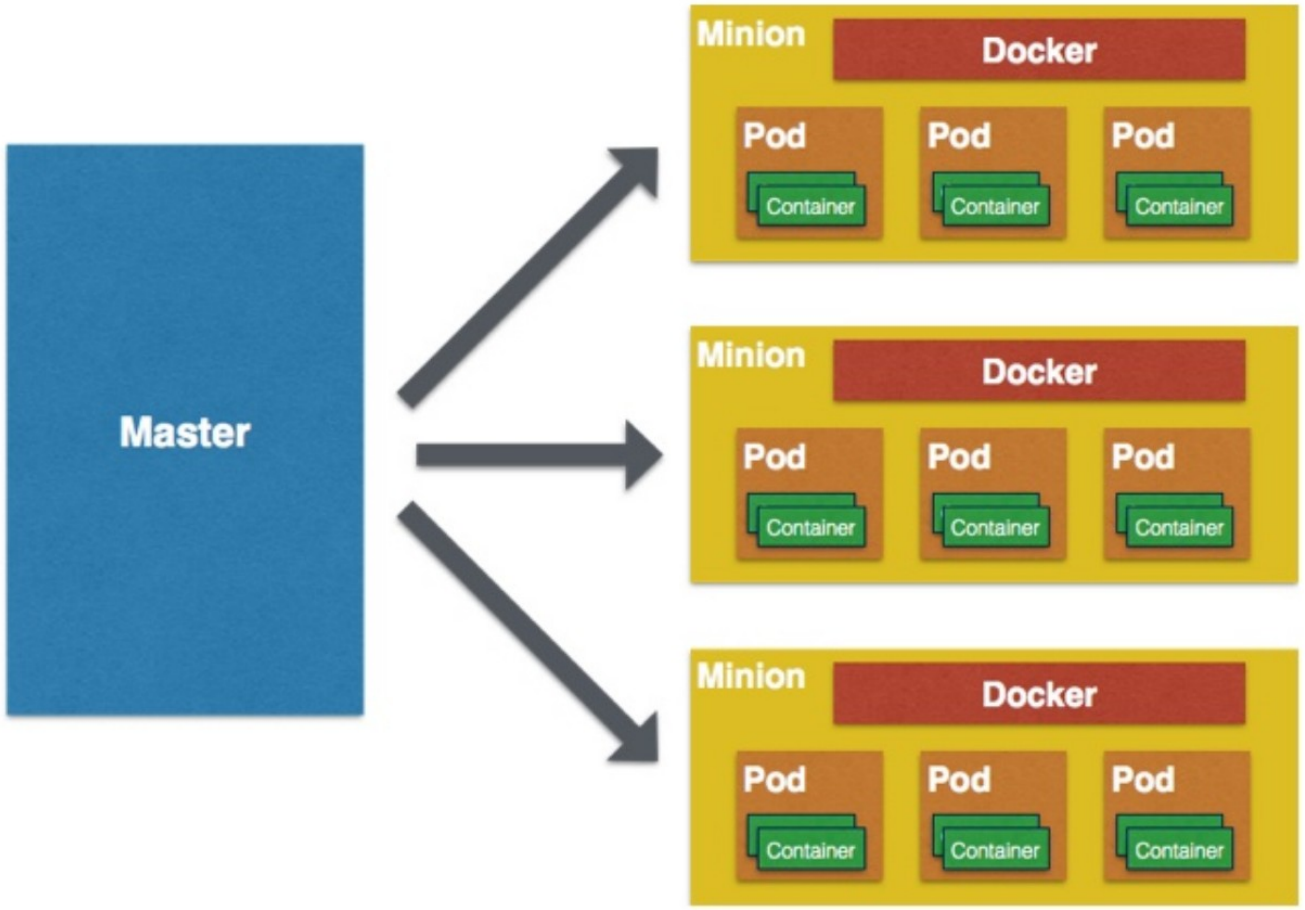
- Lots of small instances instead of 1 (or more) monolithic approach.
- Based on Docker containers.
- Kubernetes as container orchestration
- Distributed storage abstracted in the hypervisor. NFS storage offered through VMs on the hypervisor.

Pod/Docker Image

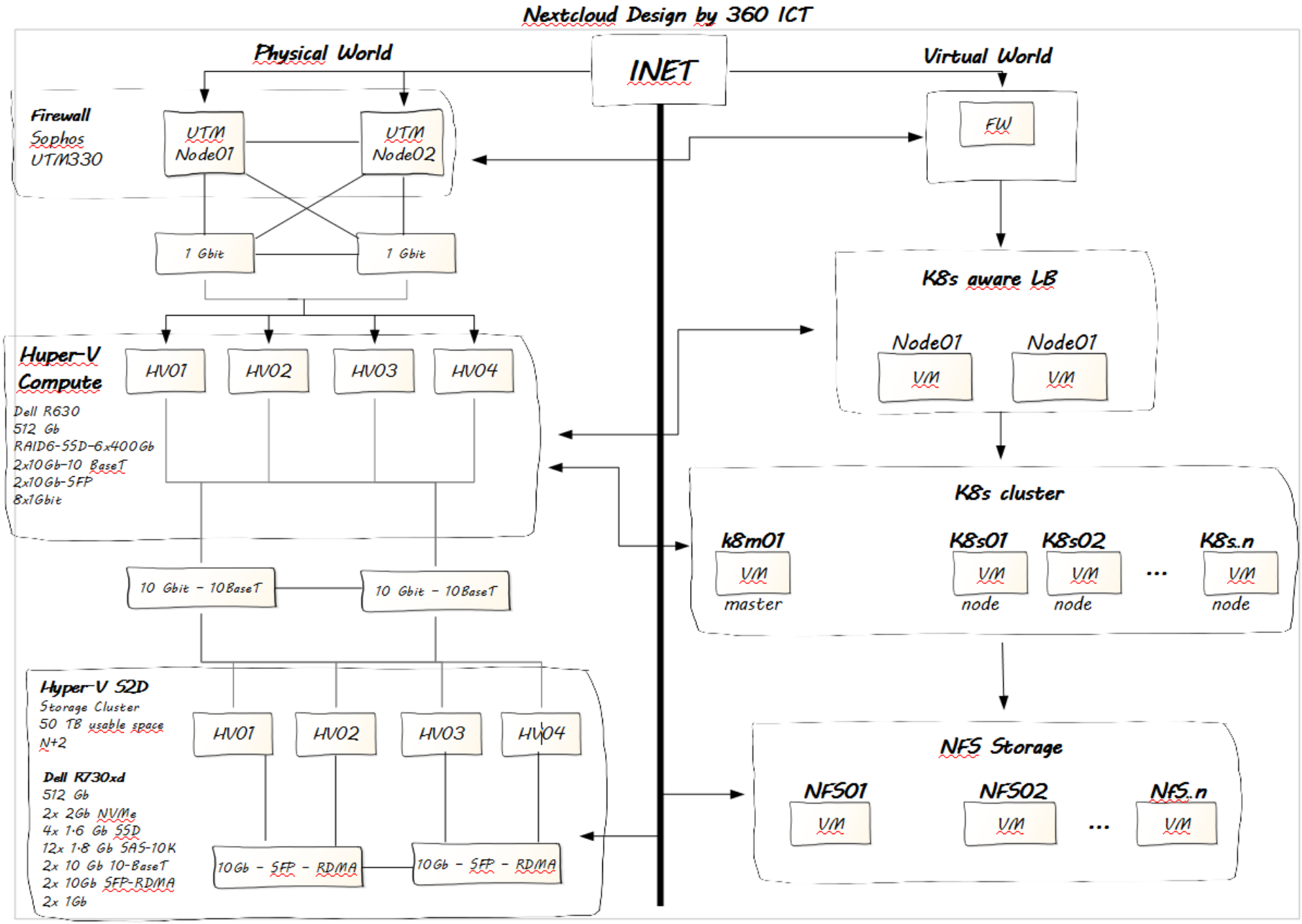
Our current design



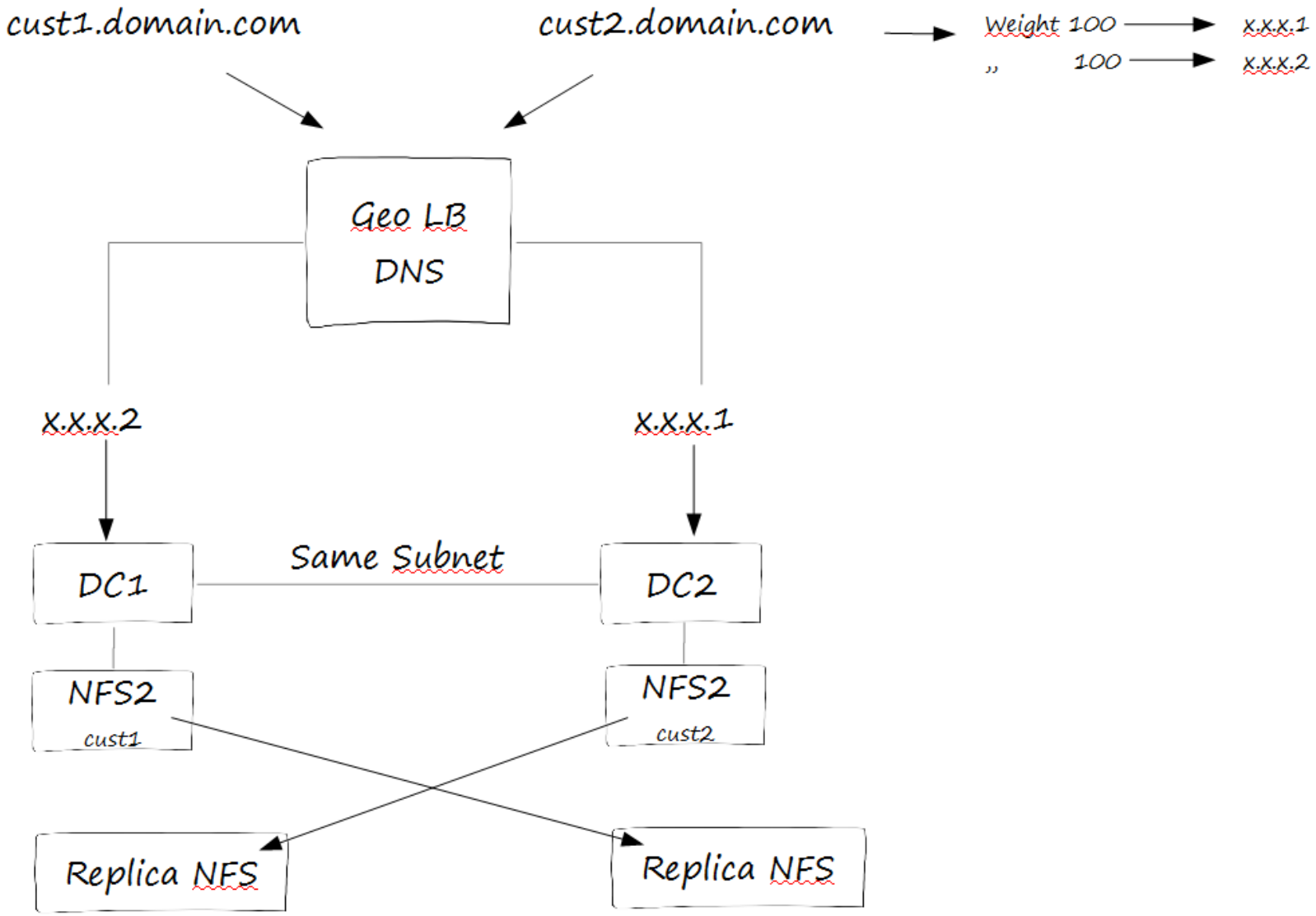
Kubernetes



http://www.slideshare.net/roland.huss/fabric8-and-docker-kubernetes-openshift


Nextcloud Design by 360 ICT
Physical World
INET
Virtual World
Firewall
Sophos
UTM330
UTM
Node01
UTM
Node02
FW
1 Gbit
1 Gbit
K8s aware LB
Node01
VM
Node01
VM
Huper-V
Compute
Dell R630
512 Gb
RAID6-SSD-6x400Gb
2x10Gb-10 BaseT
2x10Gb-SFP
8x1Gbit
HV01
HV02
HV03
HV04
K8s cluster
k8m01
VM
master
K8s01
VM
node
K8s02
VM
node
...
K8s.n
VM
node
10 Gbit - 10BaseT
10 Gbit - 10BaseT
Hyper-V S2D
Storage Cluster
50 TB usable space
N+2
Dell R730xd
512 Gb
2x 2Gb NVMe
4x 1.6 Gb SSD
12x 1.8 Gb SAS-10K
2x 10 Gb 10-BaseT
2x 10Gb SFP-RDMA
2x 1Gb
HV01
HV02
HV03
HV04
10Gb - SFP - RDMA
10Gb - SFP - RDMA
NFS Storage
NFS01
VM
NFS02
VM
...
NFS.n
VM

Multiple Datacenters



Our concept design

- Advantages
  - *Simple setup of NC. U:cloud can scale to thousands of users with 2 servers. But our typical users per instance will be 1-20 users with a few instances of 100's of users.*
  - *Customization per customer is possible (but maybe not wanted?).*
  - *Updating can be done per instance/customer. So rollout is gradually and not all or nothing.*
  - *No need for large (Galera) clusters or web/db load balancing.*

Our concept design

- Disadvantages
  - *NFS storage is unavailable during reboots. NC will be paused if NFS is unavailable and will recover. Not ideal, but good enough for first steps.*
  - *Performance of Virtualized NFS for storage? NFS will probably be an in-between step.*
  - *Will it scale? We are planning loadtests in Google Cloud Engine and our own hardware. We are curious as how far we can push the design, the max number of users in a pod and also tests on different kind of storages (and how much it would cost to do a 100k user test in GCE).*

Our concept design

- Other stuff
  - *Use federation for larger instances that don't fit in an pod?*
  - *The backend in this design is based on MS distributed storage (S2D), but this can be any storage: distributed (GlusterFS, Ceph, Netapp, etc), (small) SANs or fileclusters.*
  - *The storage can be split up in this design, which allows for more flexible solutions.*
  - *Maybe even running GlusterFS as containers within kubernetes?*
  - *Nextcloud GS?*


360 ICT
Questions?
If you have a large implementation of OC/NC and want to share your experiences, email me at **dennis@360ict.nl**.
The sheet will be shared on 360ict.nl/blog