NFS Active Active Deployment Overview

Author: David Vossel <dvossel@redhat.com>
Version: 5

An automated deployment script outlining the specifics of how to deploy HA NFS active-active with Pacemaker can be found at the link below.

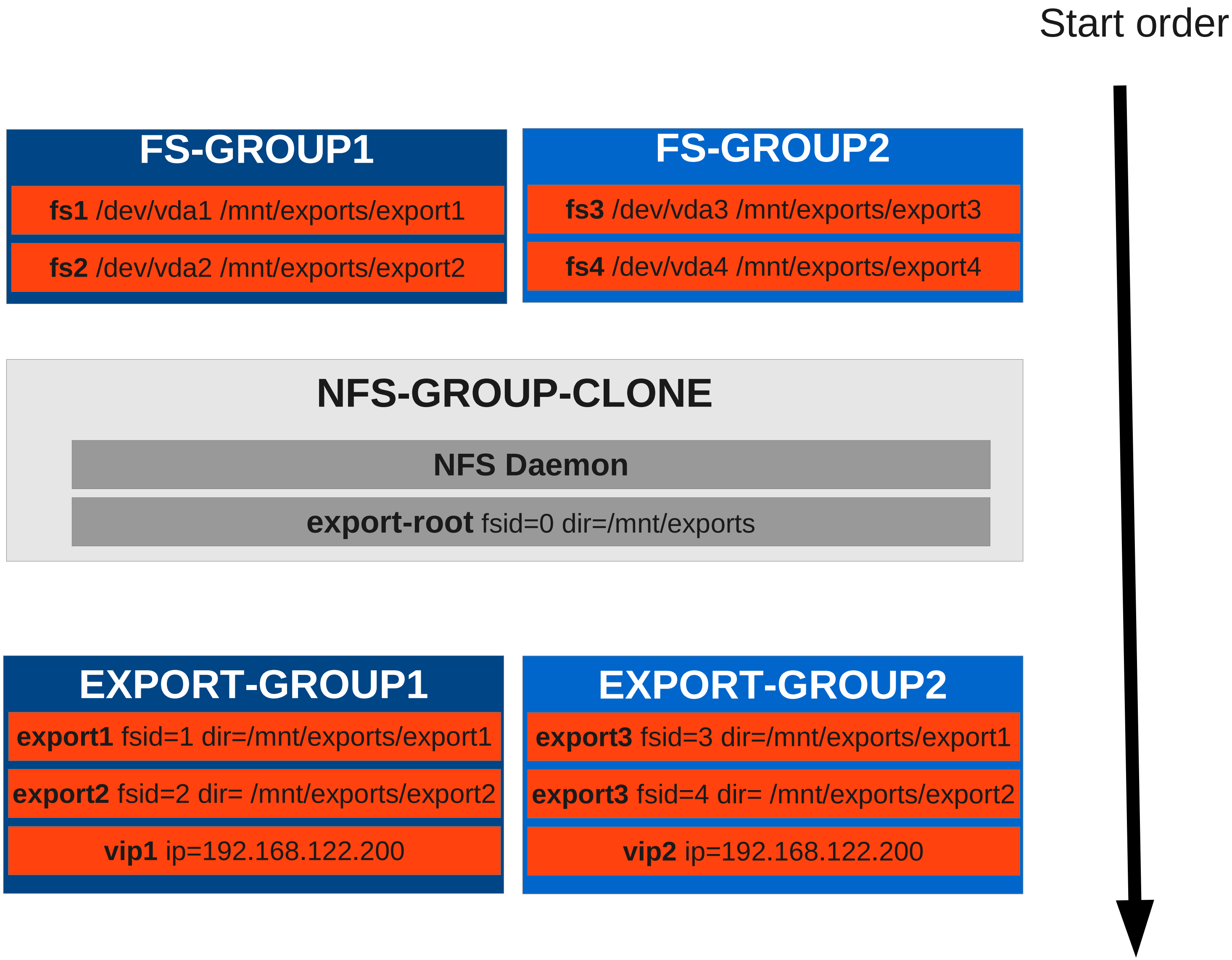
https://github.com/davidvossel/phd/blob/master/scenarios/nfs-active-active.scenario

NFSv4 Active Active

The nfs resource stack consists of shared filesystems

Followed by a cloned instance of the nfs daemons

And lastly a set of export groups that define how the shared filesystems should be exported.



Start order

FS-GROUP1

fs1 /dev/vda1 /mnt/exports/export1

fs2 /dev/vda2 /mnt/exports/export2

FS-GROUP2

fs3 /dev/vda3 /mnt/exports/export3

fs4 /dev/vda4 /mnt/exports/export4

NFS-GROUP-CLONE

NFS Daemon

export-root fsid=0 dir=/mnt/exports

EXPORT-GROUP1

export1 fsid=1 dir=/mnt/exports/export1

export2 fsid=2 dir= /mnt/exports/export2

vip1 ip=192.168.122.200

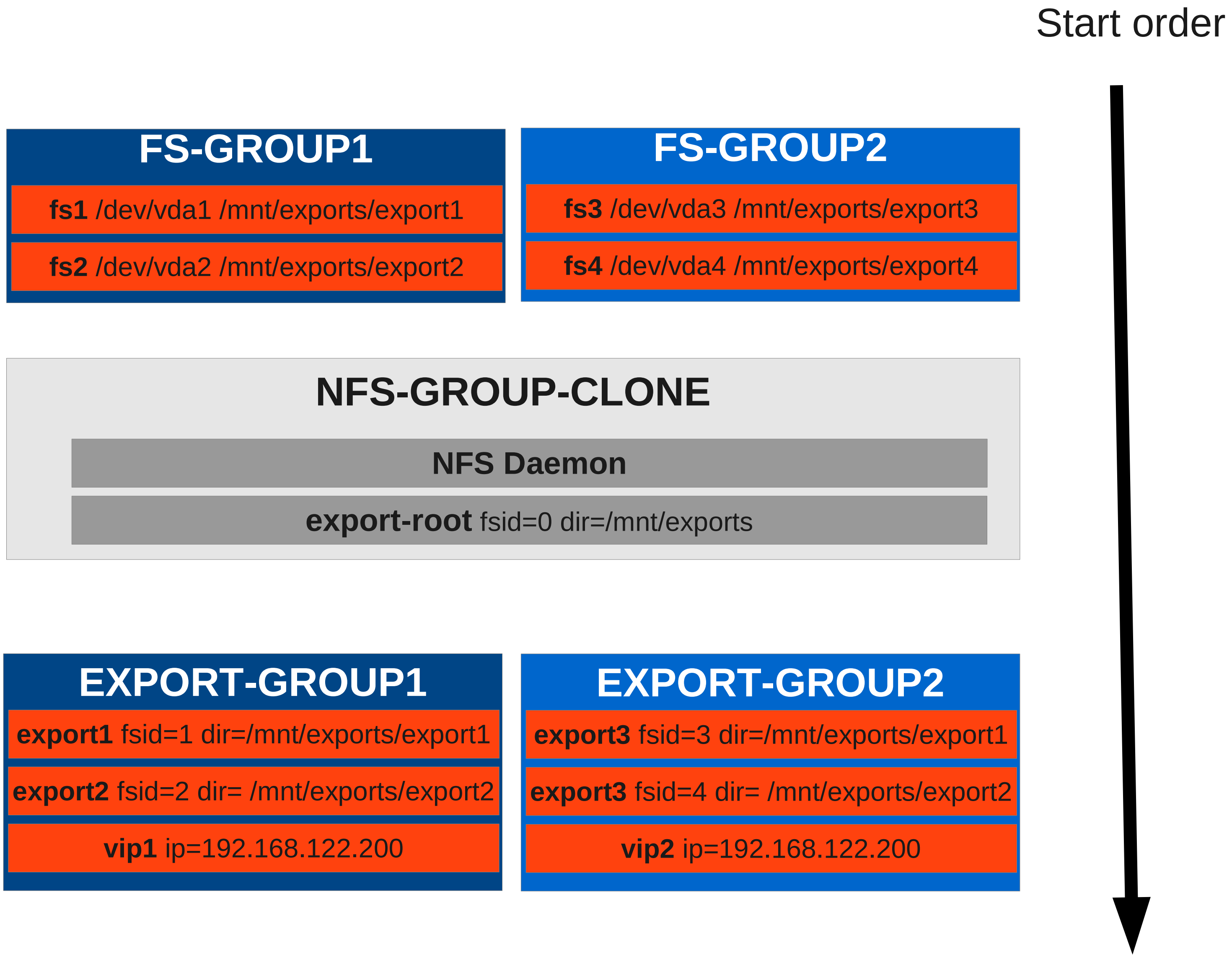
EXPORT-GROUP2

export3 fsid=3 dir=/mnt/exports/export1

export3 fsid=4 dir= /mnt/exports/export2

vip2 ip=192.168.122.200

NFSv4 Active Active



Start order

The nfs resource stack consists of shared filesystems

FS-GROUP1

fs1 /dev/vda1 /mnt/exports/export1

fs2 /dev/vda2 /mnt/exports/export2

FS-GROUP2

fs3 /dev/vda3 /mnt/exports/export3

fs4 /dev/vda4 /mnt/exports/export4

Followed by a cloned instance of the nfs daemons

NFS-GROUP-CLONE

NFS Daemon

export-root fsid=0 dir=/mnt/exports

And lastly a set of export groups that define how the shared filesystems should be exported.

EXPORT-GROUP1

export1 fsid=1 dir=/mnt/exports/export1

export2 fsid=2 dir= /mnt/exports/export2

vip1 ip=192.168.122.200

EXPORT-GROUP2

export3 fsid=3 dir=/mnt/exports/export1

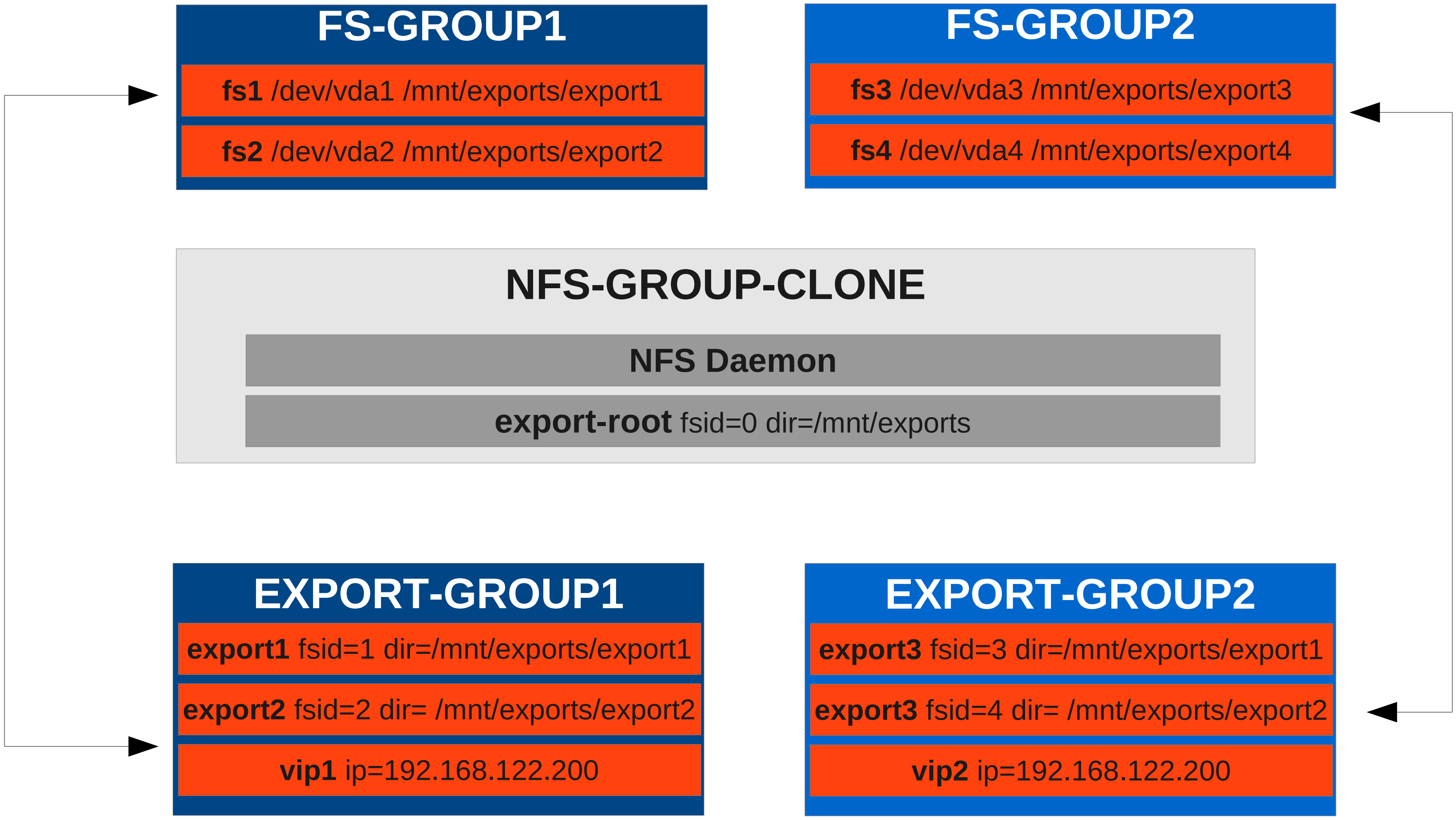
export3 fsid=4 dir= /mnt/exports/export2

vip2 ip=192.168.122.200

NFSv4 Active Active

These move as a single unit

Each filesystem group has a export group it is tied to.



FS-GROUP1

fs1 /dev/vda1 /mnt/exports/export1

fs2 /dev/vda2 /mnt/exports/export2

FS-GROUP2

fs3 /dev/vda3 /mnt/exports/export3

fs4 /dev/vda4 /mnt/exports/export4

NFS-GROUP-CLONE

NFS Daemon

export-root fsid=0 dir=/mnt/exports

EXPORT-GROUP1

export1 fsid=1 dir=/mnt/exports/export1

export2 fsid=2 dir= /mnt/exports/export2

vip1 ip=192.168.122.200

EXPORT-GROUP2

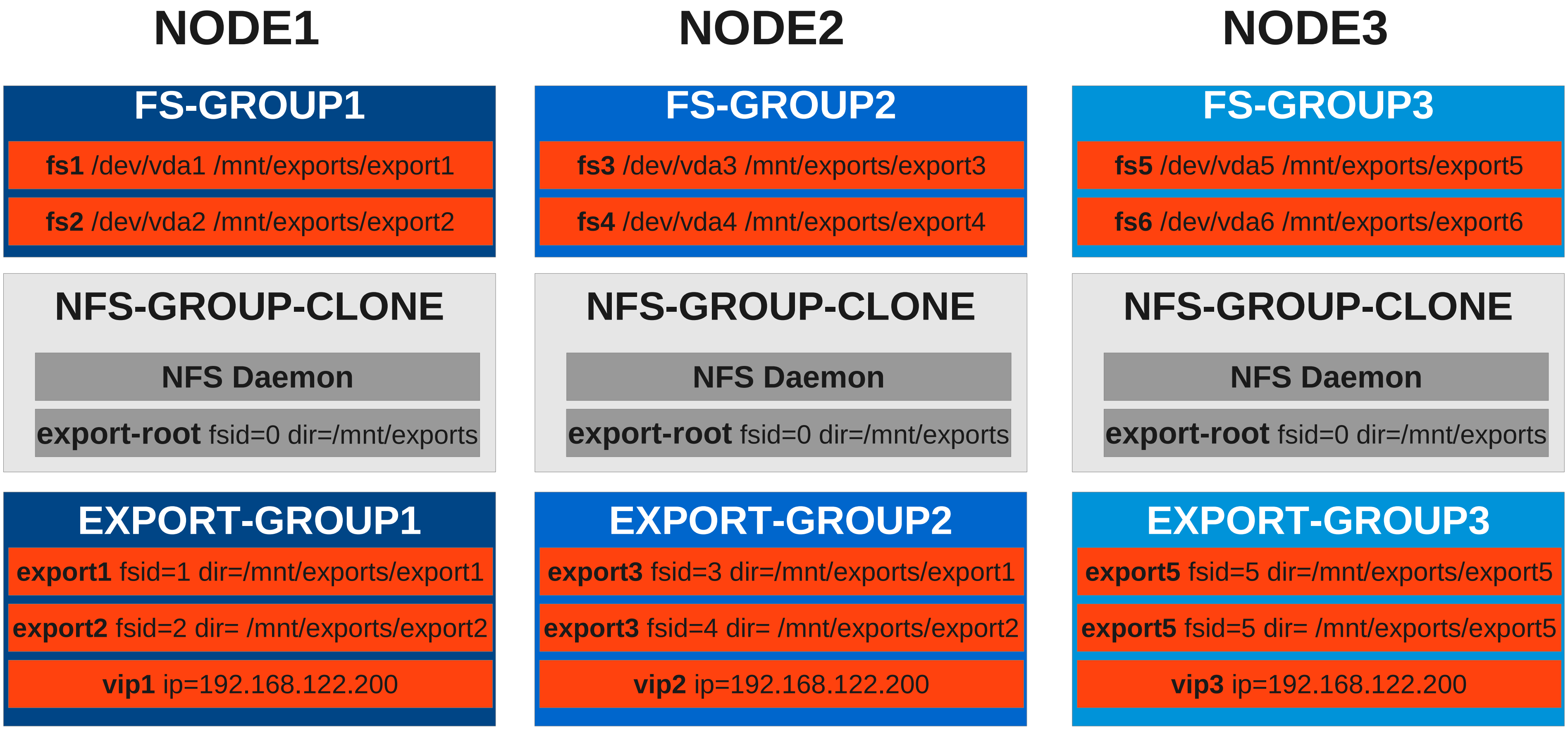
export3 fsid=3 dir=/mnt/exports/export1

export3 fsid=4 dir= /mnt/exports/export2

vip2 ip=192.168.122.200

NFSv4 Active Active

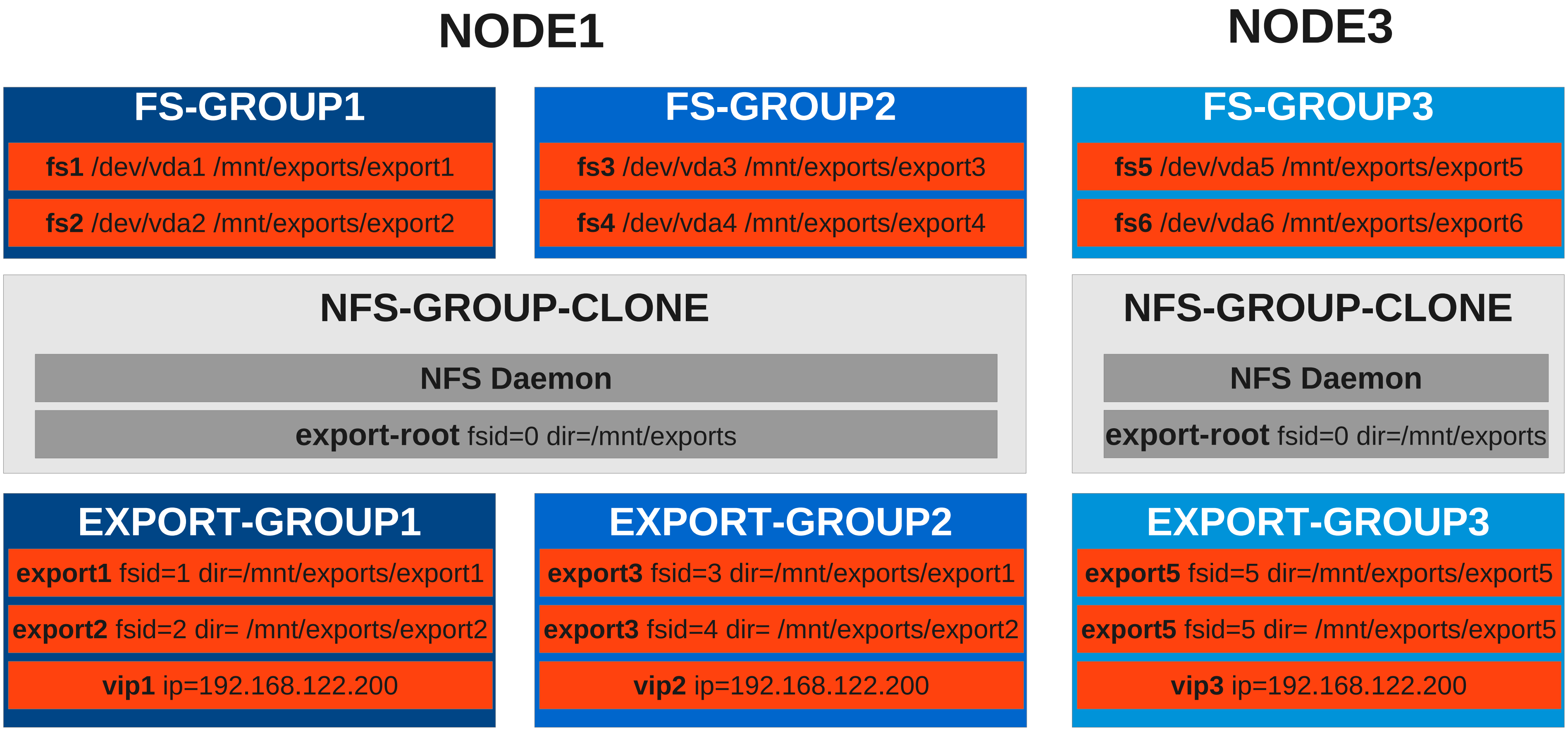
Each node gets a cloned instance of the nfs daemons. The export and filesystem groups are spread evenly across the cluster.



NODE1	NODE2	NODE3
FS-GROUP1	FS-GROUP2	FS-GROUP3
fs1 /dev/vda1 /mnt/exports/export1	**fs3** /dev/vda3 /mnt/exports/export3	**fs5** /dev/vda5 /mnt/exports/export5
fs2 /dev/vda2 /mnt/exports/export2	**fs4** /dev/vda4 /mnt/exports/export4	**fs6** /dev/vda6 /mnt/exports/export6
NFS-GROUP-CLONE	NFS-GROUP-CLONE	NFS-GROUP-CLONE
NFS Daemon	NFS Daemon	NFS Daemon
export-root fsid=0 dir=/mnt/exports	**export-root** fsid=0 dir=/mnt/exports	**export-root** fsid=0 dir=/mnt/exports
EXPORT-GROUP1	EXPORT-GROUP2	EXPORT-GROUP3
export1 fsid=1 dir=/mnt/exports/export1	**export3** fsid=3 dir=/mnt/exports/export1	**export5** fsid=5 dir=/mnt/exports/export5
export2 fsid=2 dir= /mnt/exports/export2	**export3** fsid=4 dir= /mnt/exports/export2	**export5** fsid=5 dir= /mnt/exports/export5
vip1 ip=192.168.122.200	**vip2** ip=192.168.122.200	**vip3** ip=192.168.122.200

Node Failure

After node failure, the unallocated export groups are distributed across the remaining nodes.



NODE1

FS-GROUP1

fs1 /dev/vda1 /mnt/exports/export1

fs2 /dev/vda2 /mnt/exports/export2

FS-GROUP2

fs3 /dev/vda3 /mnt/exports/export3

fs4 /dev/vda4 /mnt/exports/export4

NFS-GROUP-CLONE

NFS Daemon

export-root fsid=0 dir=/mnt/exports

EXPORT-GROUP1

export1 fsid=1 dir=/mnt/exports/export1

export2 fsid=2 dir= /mnt/exports/export2

vip1 ip=192.168.122.200

EXPORT-GROUP2

export3 fsid=3 dir=/mnt/exports/export1

export3 fsid=4 dir= /mnt/exports/export2

vip2 ip=192.168.122.200

NODE3

FS-GROUP3

fs5 /dev/vda5 /mnt/exports/export5

fs6 /dev/vda6 /mnt/exports/export6

NFS-GROUP-CLONE

NFS Daemon

export-root fsid=0 dir=/mnt/exports

EXPORT-GROUP3

export5 fsid=5 dir=/mnt/exports/export5

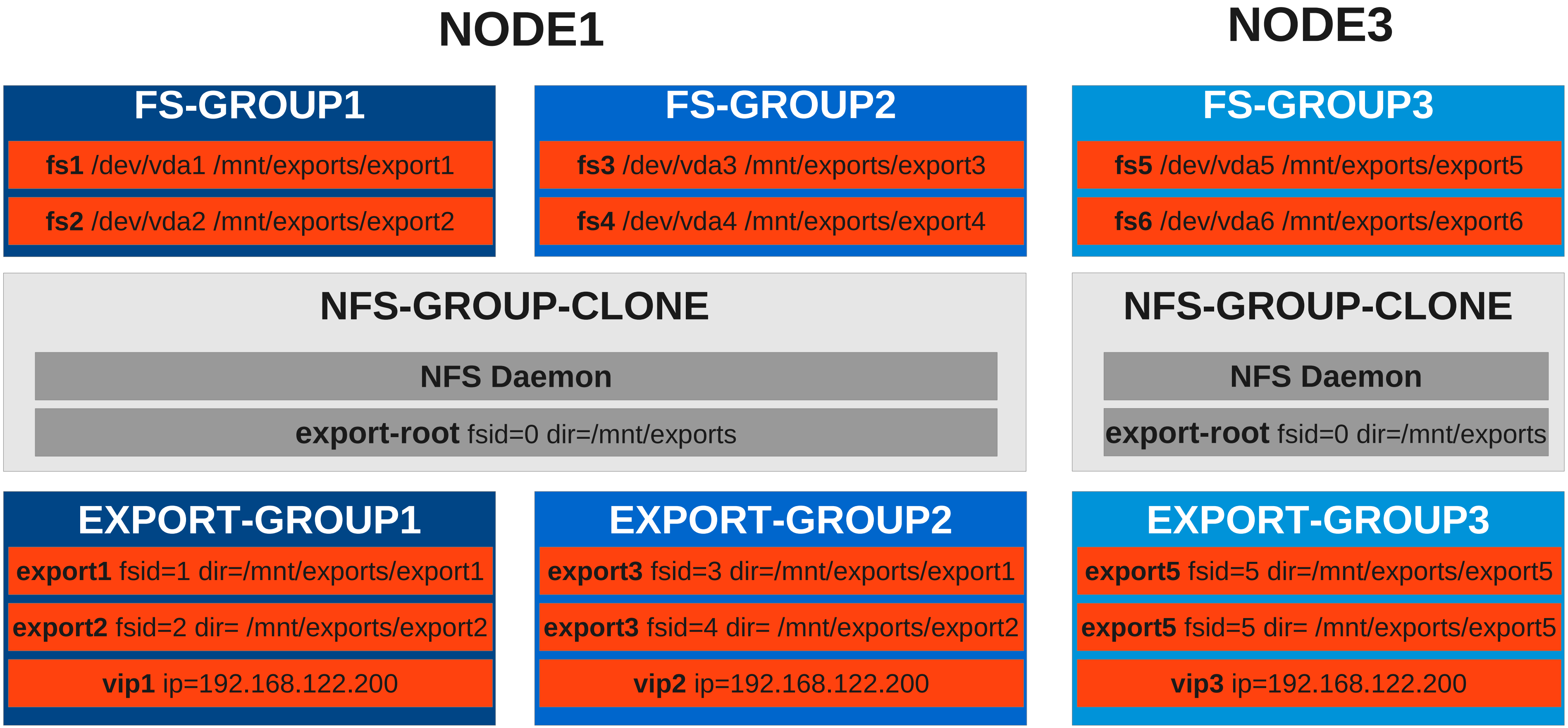
export5 fsid=5 dir= /mnt/exports/export5

vip3 ip=192.168.122.200

NFSv4 Grace and Lease Timers

The export filesystems are ordered to start before the nfs-daemons. This results in the restart of the local nfs daemons when a node acquires a new export group.

The daemon restart guarantees the nfsv4grace period is observed after an export moves. This allows clients previously connected to the export to renew file leases after the failover.



NODE1

NODE3

FS-GROUP1

fs1 /dev/vda1 /mnt/exports/export1

fs2 /dev/vda2 /mnt/exports/export2

FS-GROUP2

fs3 /dev/vda3 /mnt/exports/export3

fs4 /dev/vda4 /mnt/exports/export4

FS-GROUP3

fs5 /dev/vda5 /mnt/exports/export5

fs6 /dev/vda6 /mnt/exports/export6

NFS-GROUP-CLONE

NFS Daemon

export-root fsid=0 dir=/mnt/exports

NFS-GROUP-CLONE

NFS Daemon

export-root fsid=0 dir=/mnt/exports

EXPORT-GROUP1

export1 fsid=1 dir=/mnt/exports/export1

export2 fsid=2 dir= /mnt/exports/export2

vip1 ip=192.168.122.200

EXPORT-GROUP2

export3 fsid=3 dir=/mnt/exports/export1

export3 fsid=4 dir= /mnt/exports/export2

vip2 ip=192.168.122.200

EXPORT-GROUP3

export5 fsid=5 dir=/mnt/exports/export5

export5 fsid=5 dir= /mnt/exports/export5

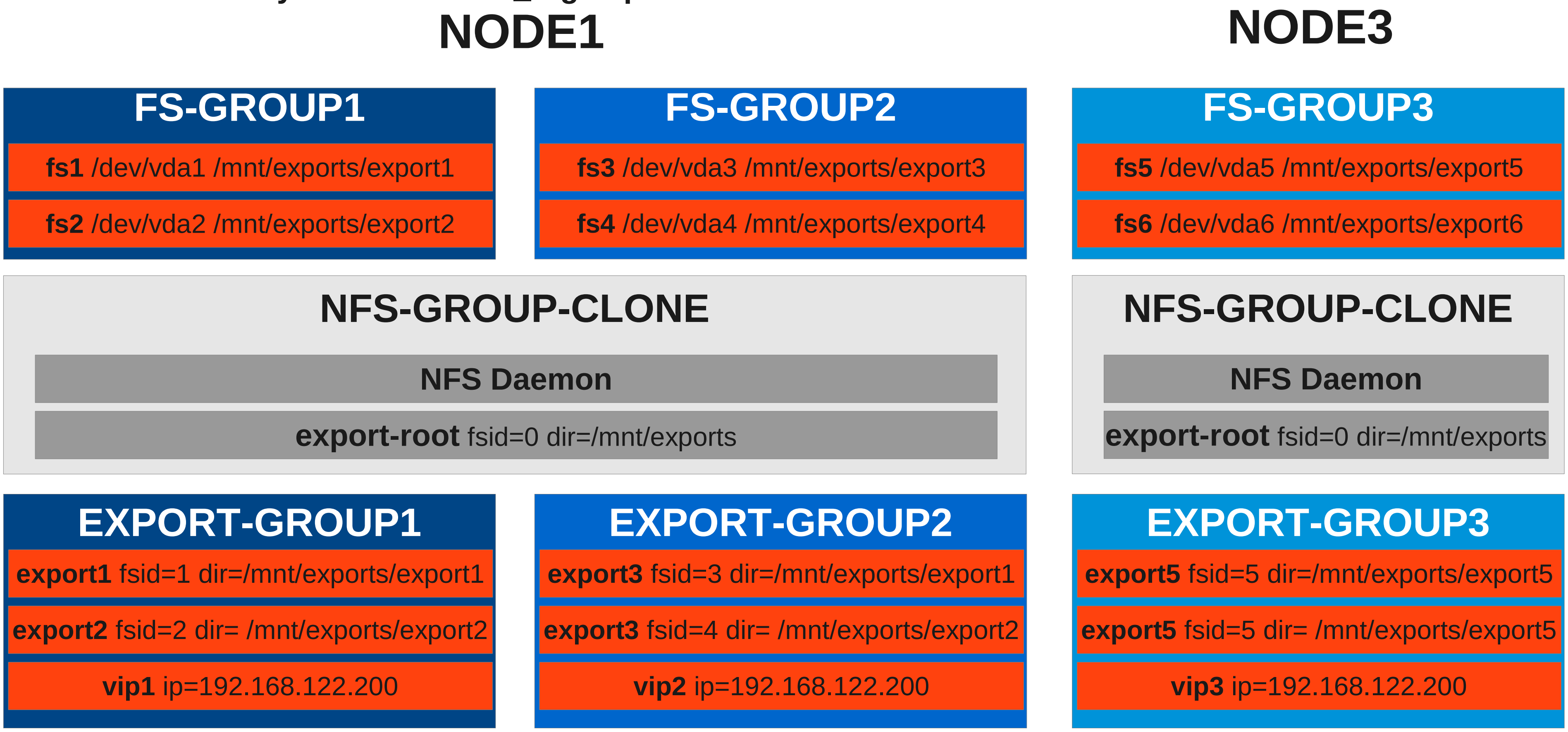
vip3 ip=192.168.122.200

Minimizing Failover Time

Each failover event will result in the grace time being observed before new clients can begin using the nfs servers exports. By default these timeouts are 90 seconds.

To reduce failover time, the nfsserver resource-agent has the ability to dynamically set the *4gracetime* and *4leasetime* values to as low as 10 seconds (nfsd_args=-G 10 -L 10). To avoid lock renewal race conditions, the grace time must always be greater than or equal to the lease time.

Make sure the -G and -L options are available for nfsd on your distro, otherwise nfsserver may fail to start when you set the nfsd_args option.



NODE1

NODE3

FS-GROUP1

fs1 /dev/vda1 /mnt/exports/export1

fs2 /dev/vda2 /mnt/exports/export2

FS-GROUP2

fs3 /dev/vda3 /mnt/exports/export3

fs4 /dev/vda4 /mnt/exports/export4

FS-GROUP3

fs5 /dev/vda5 /mnt/exports/export5

fs6 /dev/vda6 /mnt/exports/export6

NFS-GROUP-CLONE

NFS Daemon

export-root fsid=0 dir=/mnt/exports

NFS-GROUP-CLONE

NFS Daemon

export-root fsid=0 dir=/mnt/exports

EXPORT-GROUP1

export1 fsid=1 dir=/mnt/exports/export1

export2 fsid=2 dir= /mnt/exports/export2

vip1 ip=192.168.122.200

EXPORT-GROUP2

export3 fsid=3 dir=/mnt/exports/export1

export3 fsid=4 dir= /mnt/exports/export2

vip2 ip=192.168.122.200

EXPORT-GROUP3

export5 fsid=5 dir=/mnt/exports/export5

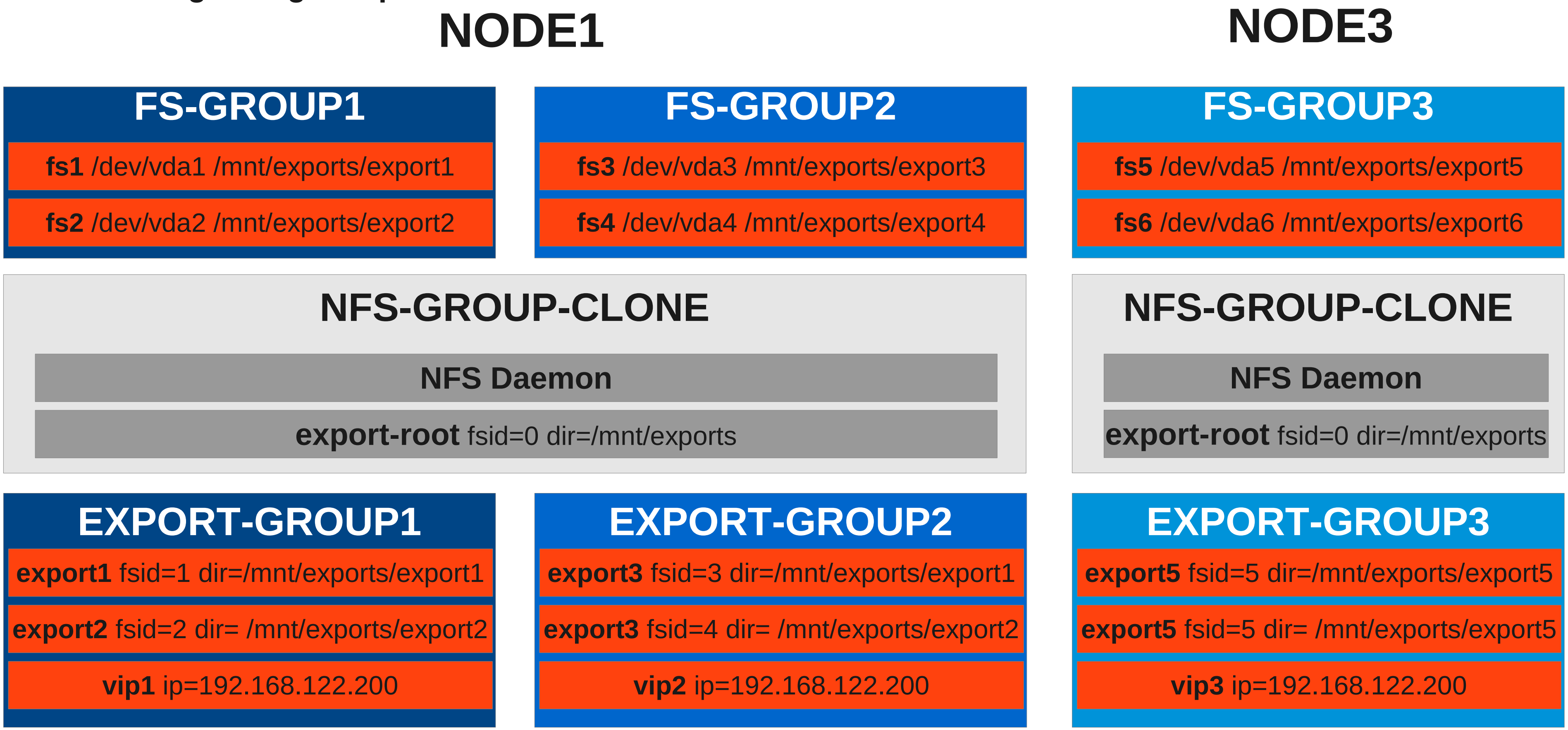
export5 fsid=5 dir= /mnt/exports/export5

vip3 ip=192.168.122.200

Changing v4 Lease/Grace period

When changing the lease and grace periods for an already running server, the procedure below must be followed. Note that changing the grace/lease times should always be done from the pacemaker configuration, never outside of the cluster.

1. Change the lease period
2. Restart server
3. Wait the grace period time (This gives a chance for all the clients to find out about the new grace period.)
4. Change the grace period.



NODE1

NODE3

FS-GROUP1

fs1 /dev/vda1 /mnt/exports/export1

fs2 /dev/vda2 /mnt/exports/export2

FS-GROUP2

fs3 /dev/vda3 /mnt/exports/export3

fs4 /dev/vda4 /mnt/exports/export4

FS-GROUP3

fs5 /dev/vda5 /mnt/exports/export5

fs6 /dev/vda6 /mnt/exports/export6

NFS-GROUP-CLONE

NFS Daemon

export-root fsid=0 dir=/mnt/exports

NFS-GROUP-CLONE

NFS Daemon

export-root fsid=0 dir=/mnt/exports

EXPORT-GROUP1

export1 fsid=1 dir=/mnt/exports/export1

export2 fsid=2 dir= /mnt/exports/export2

vip1 ip=192.168.122.200

EXPORT-GROUP2

export3 fsid=3 dir=/mnt/exports/export1

export3 fsid=4 dir= /mnt/exports/export2

vip2 ip=192.168.122.200

EXPORT-GROUP3

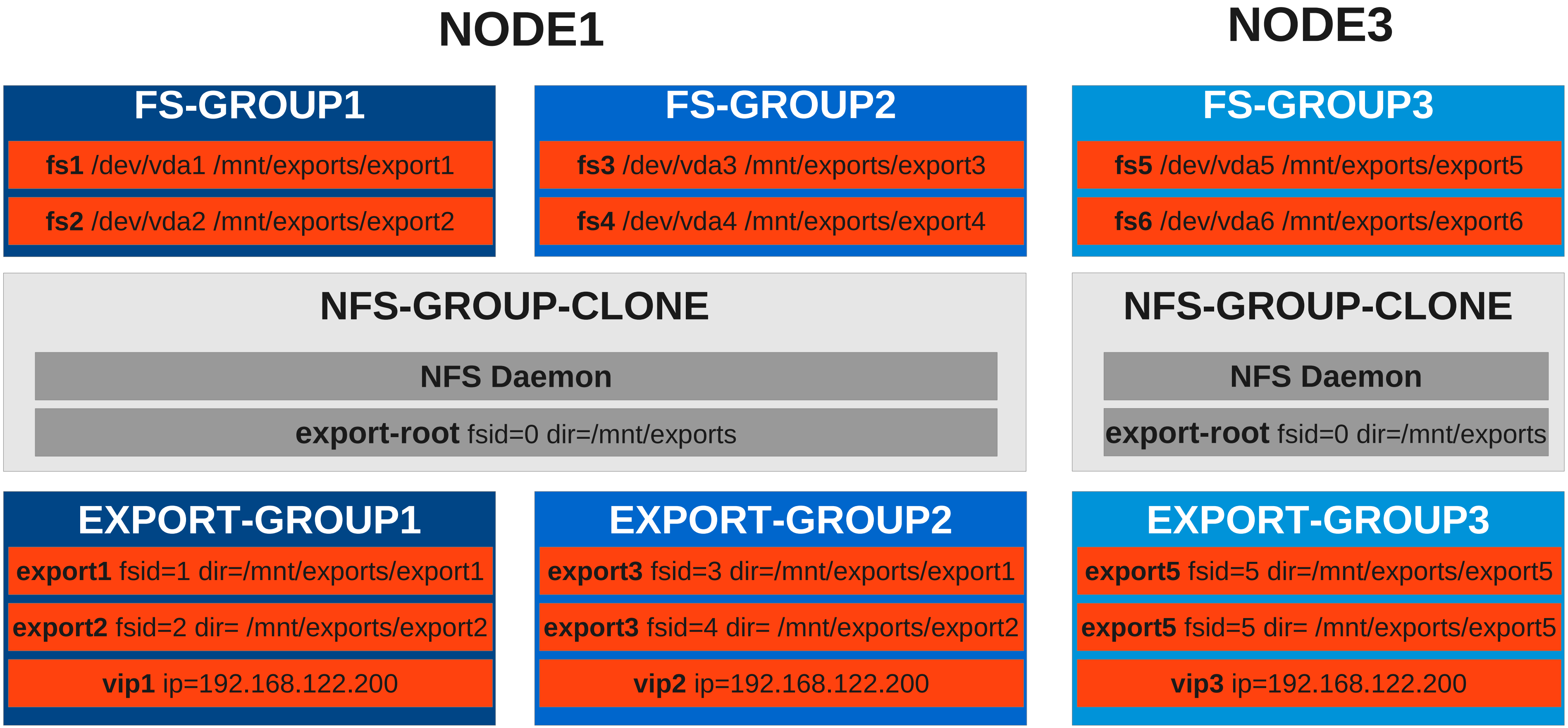
export5 fsid=5 dir=/mnt/exports/export5

export5 fsid=5 dir= /mnt/exports/export5

vip3 ip=192.168.122.200

NFSv3 Active Active Limitations

This deployment allows mixed usage of NFSv3 and NFSv4 client, but file lock recovery will only occur for NFSv4 clients.



NODE1

NODE3

FS-GROUP1

fs1 /dev/vda1 /mnt/exports/export1

fs2 /dev/vda2 /mnt/exports/export2

FS-GROUP2

fs3 /dev/vda3 /mnt/exports/export3

fs4 /dev/vda4 /mnt/exports/export4

FS-GROUP3

fs5 /dev/vda5 /mnt/exports/export5

fs6 /dev/vda6 /mnt/exports/export6

NFS-GROUP-CLONE

NFS Daemon

export-root fsid=0 dir=/mnt/exports

NFS-GROUP-CLONE

NFS Daemon

export-root fsid=0 dir=/mnt/exports

EXPORT-GROUP1

export1 fsid=1 dir=/mnt/exports/export1

export2 fsid=2 dir= /mnt/exports/export2

vip1 ip=192.168.122.200

EXPORT-GROUP2

export3 fsid=3 dir=/mnt/exports/export1

export3 fsid=4 dir= /mnt/exports/export2

vip2 ip=192.168.122.200

EXPORT-GROUP3

export5 fsid=5 dir=/mnt/exports/export5

export5 fsid=5 dir= /mnt/exports/export5

vip3 ip=192.168.122.200