NFS Active Active Deployment Overview

Author: David Vossel <dvossel@redhat.com>
Version: 3

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

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

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

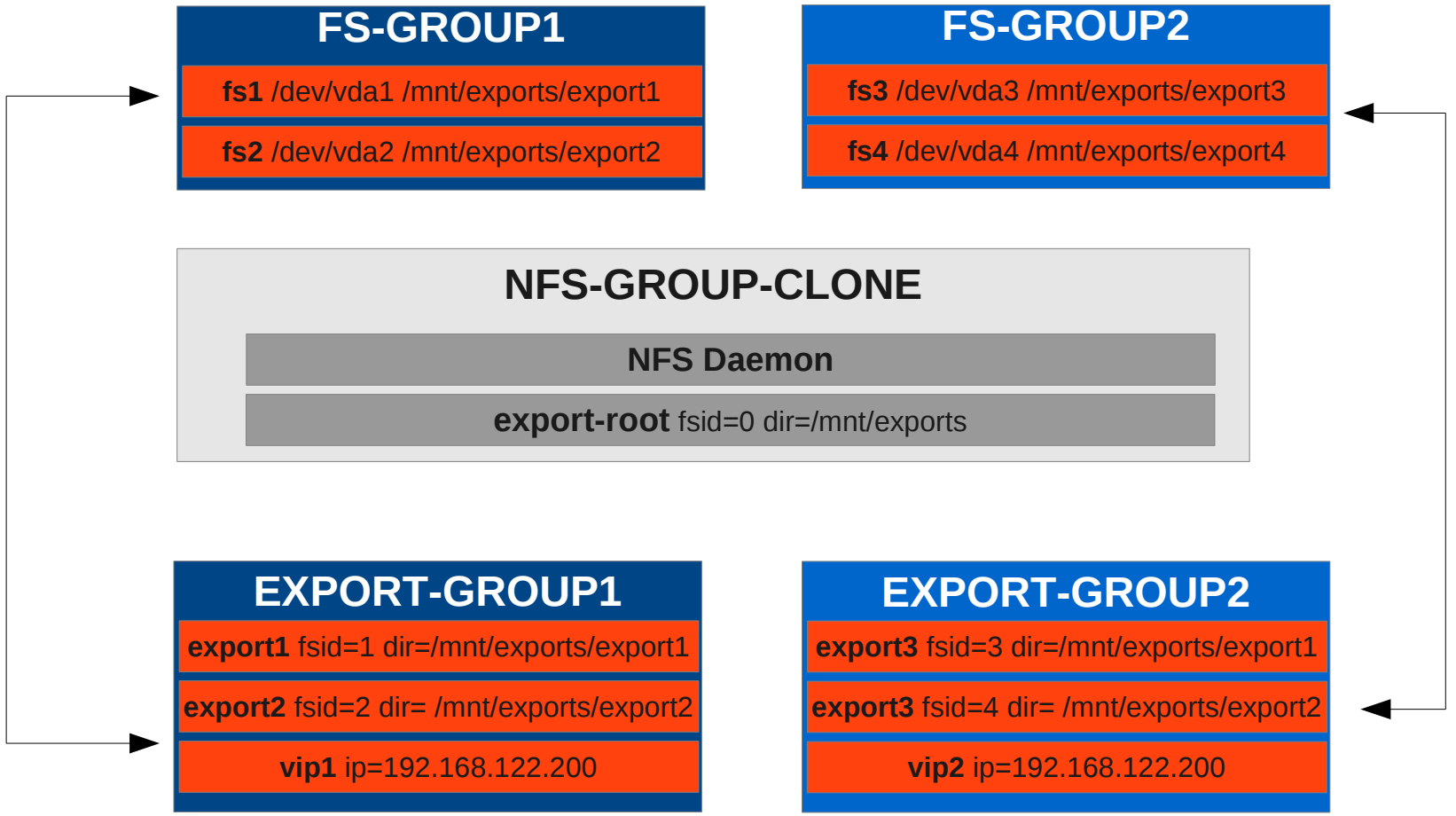
export3 fsid=4 dir= /mnt/exports/export2

vip2 ip=192.168.122.200

NFSv4 Active Active

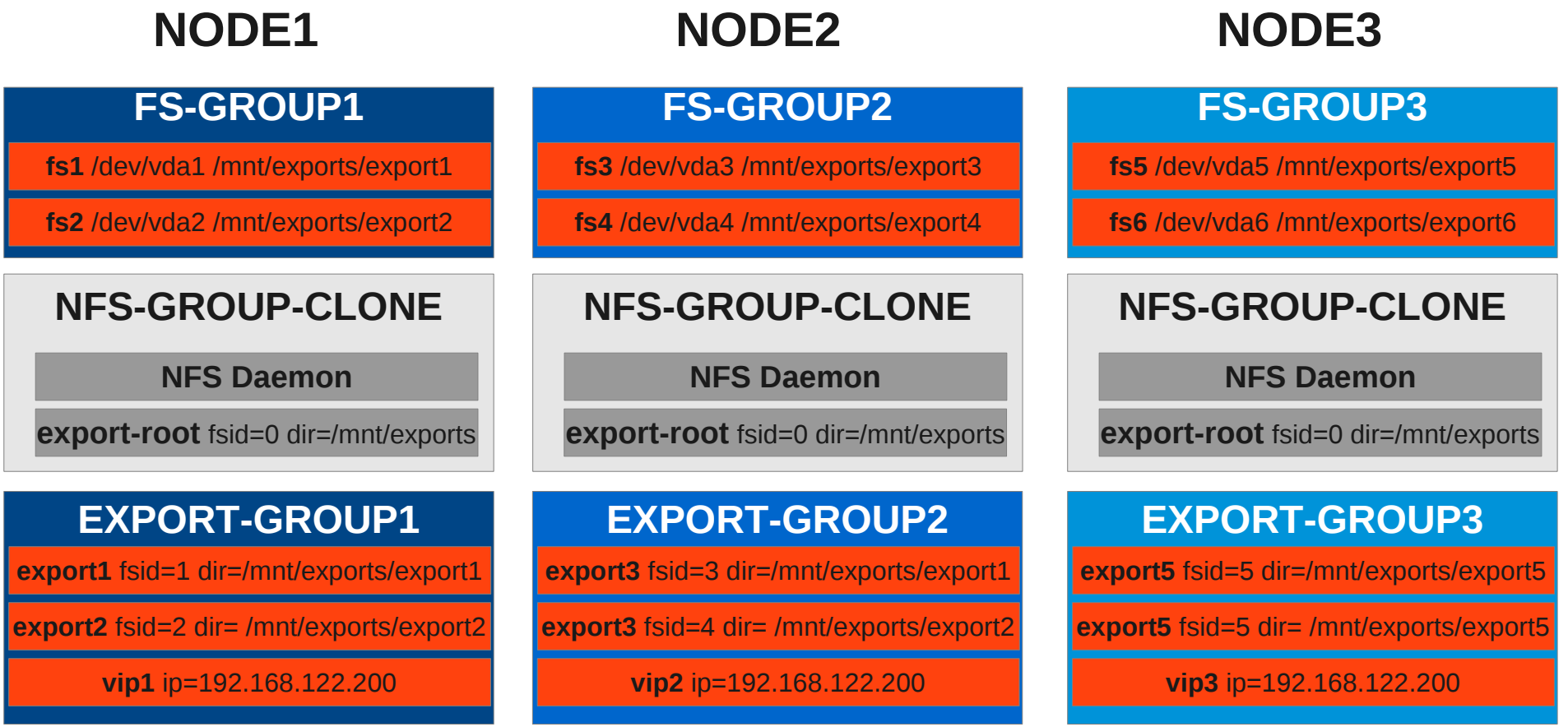
These move as a single unit

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



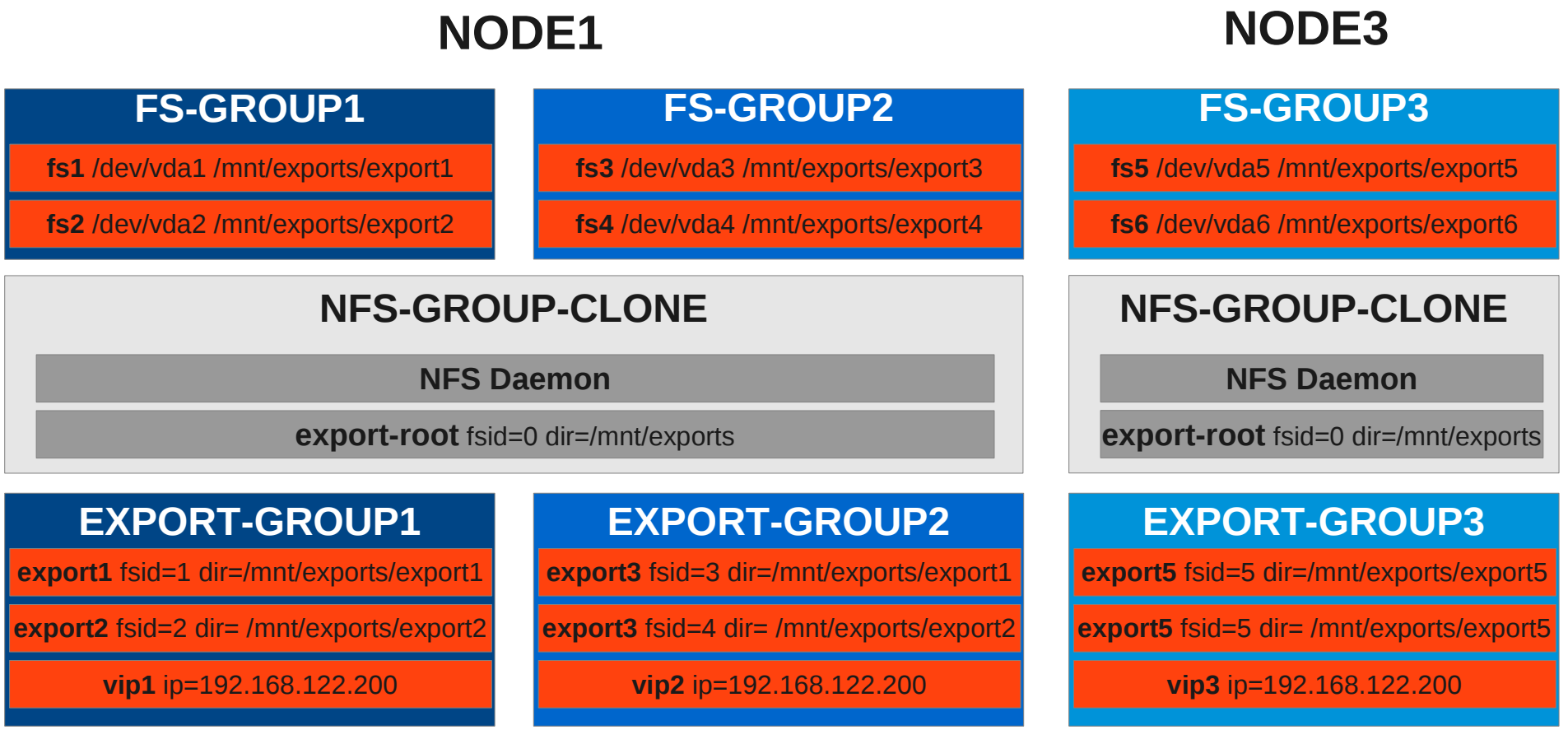
NFSv4 Active Active

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



Node Failure

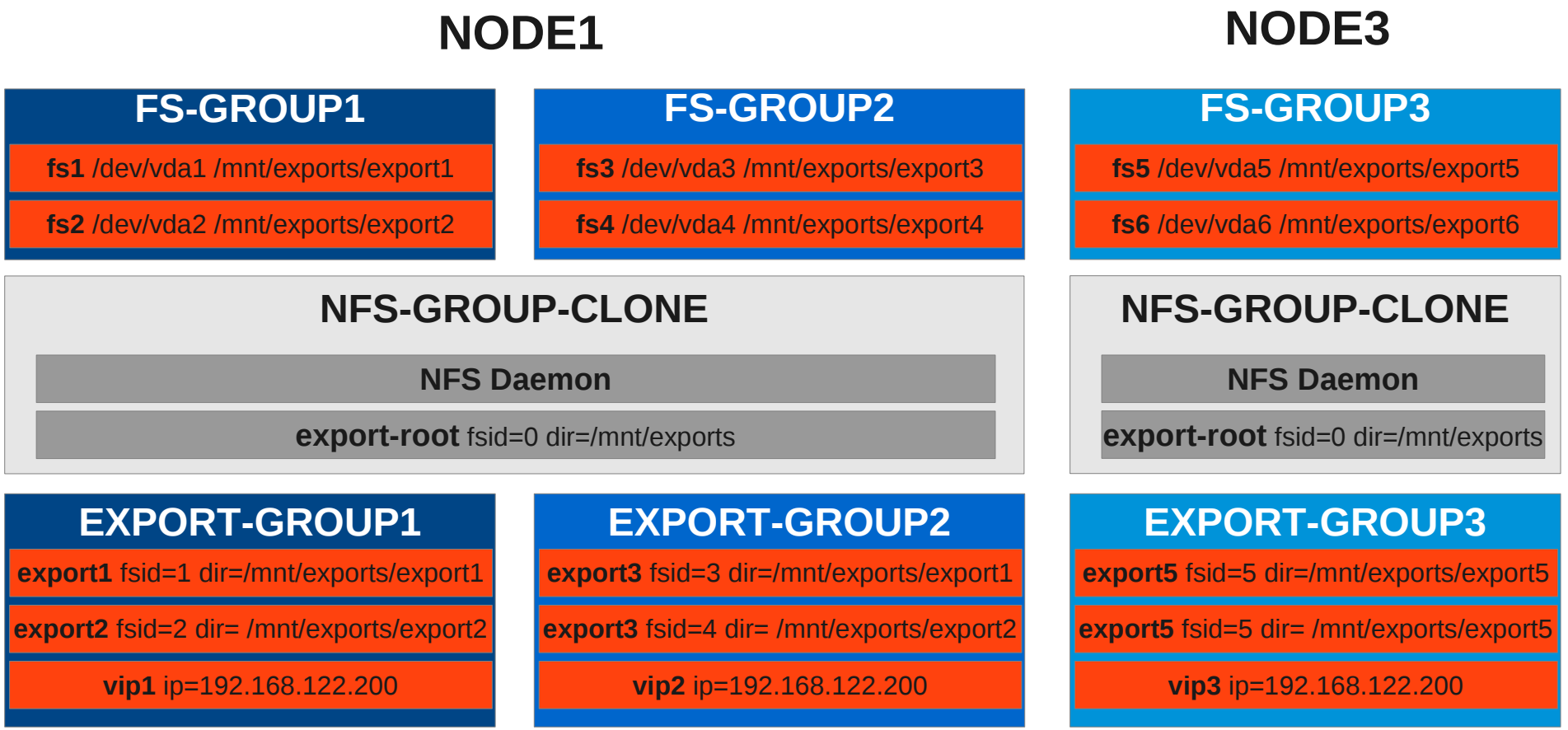
After node failure, the unallocated export groups are distributed across the remaining nodes.



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.

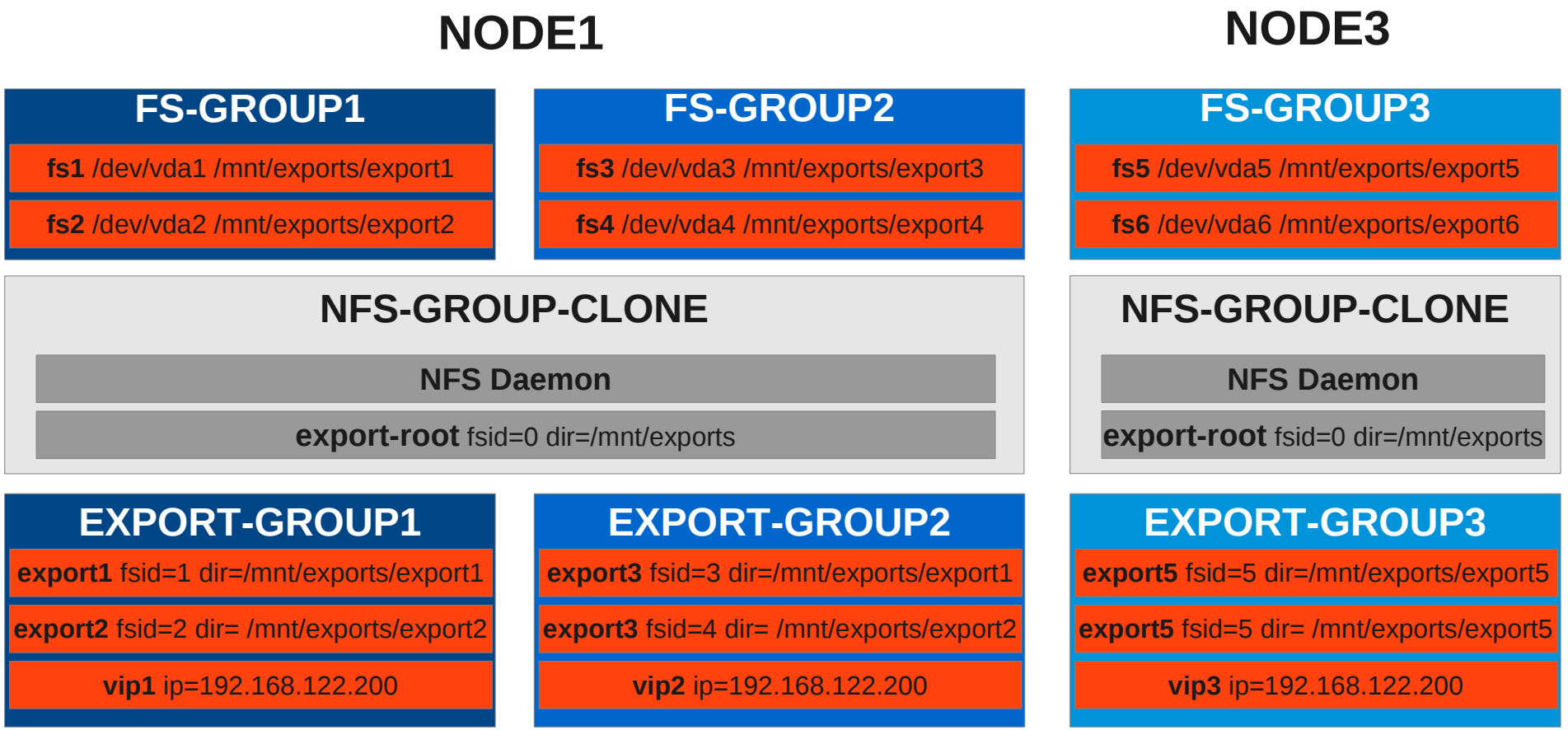
The daemon restart guarantees the nfsv4grace period is observed after an export moves. This allows clients previously connected to the export to renew file leases after the failover.



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.



NFSv3 Active Active Limitations

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