DRBD

Network Raid, High Availability and General Awesomeness

The Problem

- Moving VMs from one datacenter to another datacenter several hundred miles away with minimum downtime.

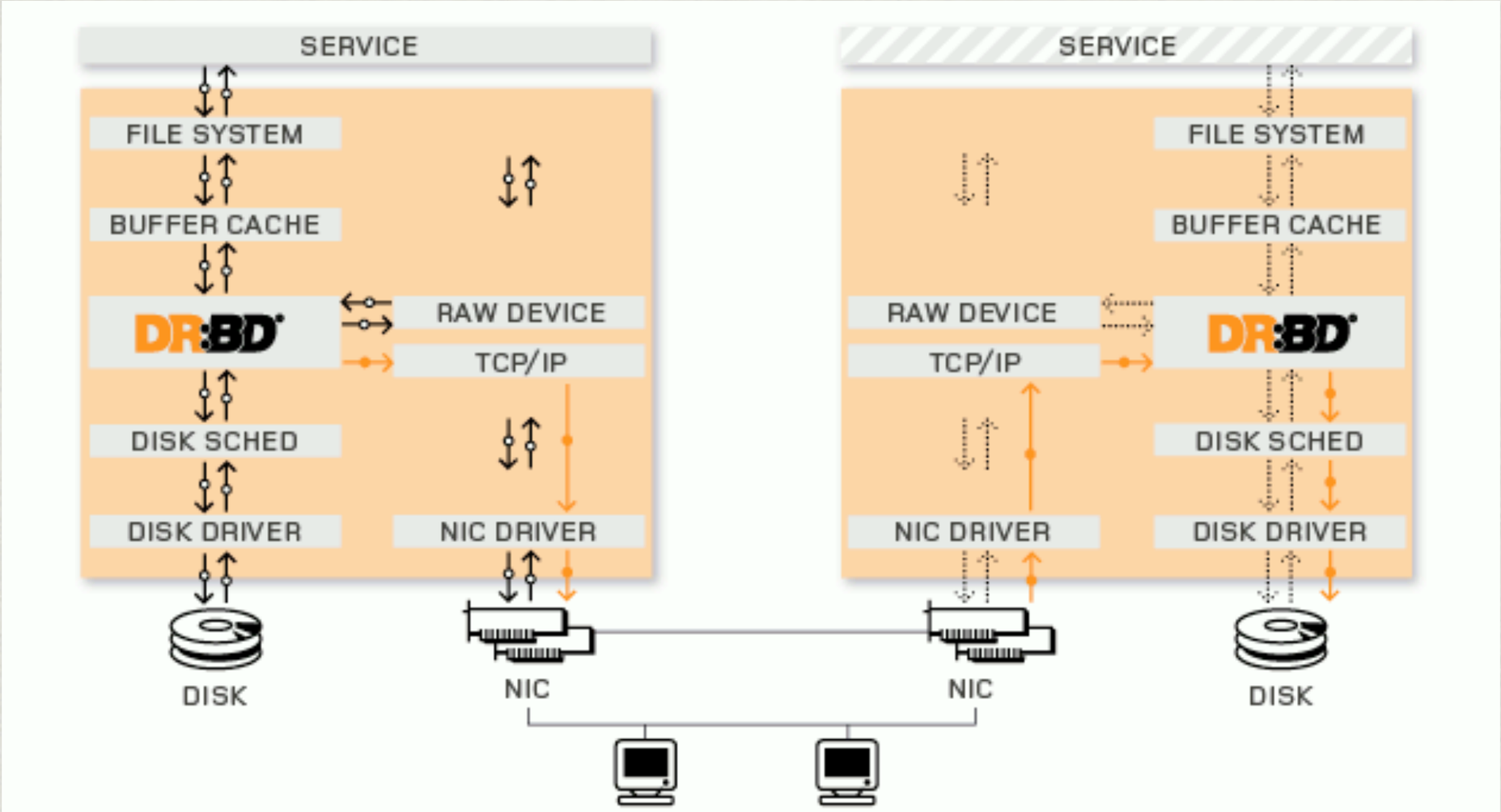
RAID1 Explained

- Redundant Arrays Of {Inexpensive|Independent} Disks
  - Disk to disk mirroring without parity
  - Can be done in hardware or software
  - Usually done at the disk controller level
  - Why hardware RAID is sometimes bad

DRBD Explained

- Distributed Replicated Block Device
  - Open Source from LINBIT
  - Equivalent of RAID1, but with the mirror being another system on the network
  - Works at the block device level
  - Linux kernel module
  - Works with LVM, MD, dm-crypt and all filesystems
  - Replicates across IPv4, IPv6, SuperSockets, IPoIB
  - Linux NIC Bonding is also supported

DRBD Architecture



- this image blatantly stolen from DRBD.ORG

DRBD vs RAID1

- RAID1 protects from single disk failure
  - Single Node Solution
- DRBD protects from whole system failure
  - Part of a clustering eco-system, with heartbeat
  - Can participate in a three node (stacked) configuration

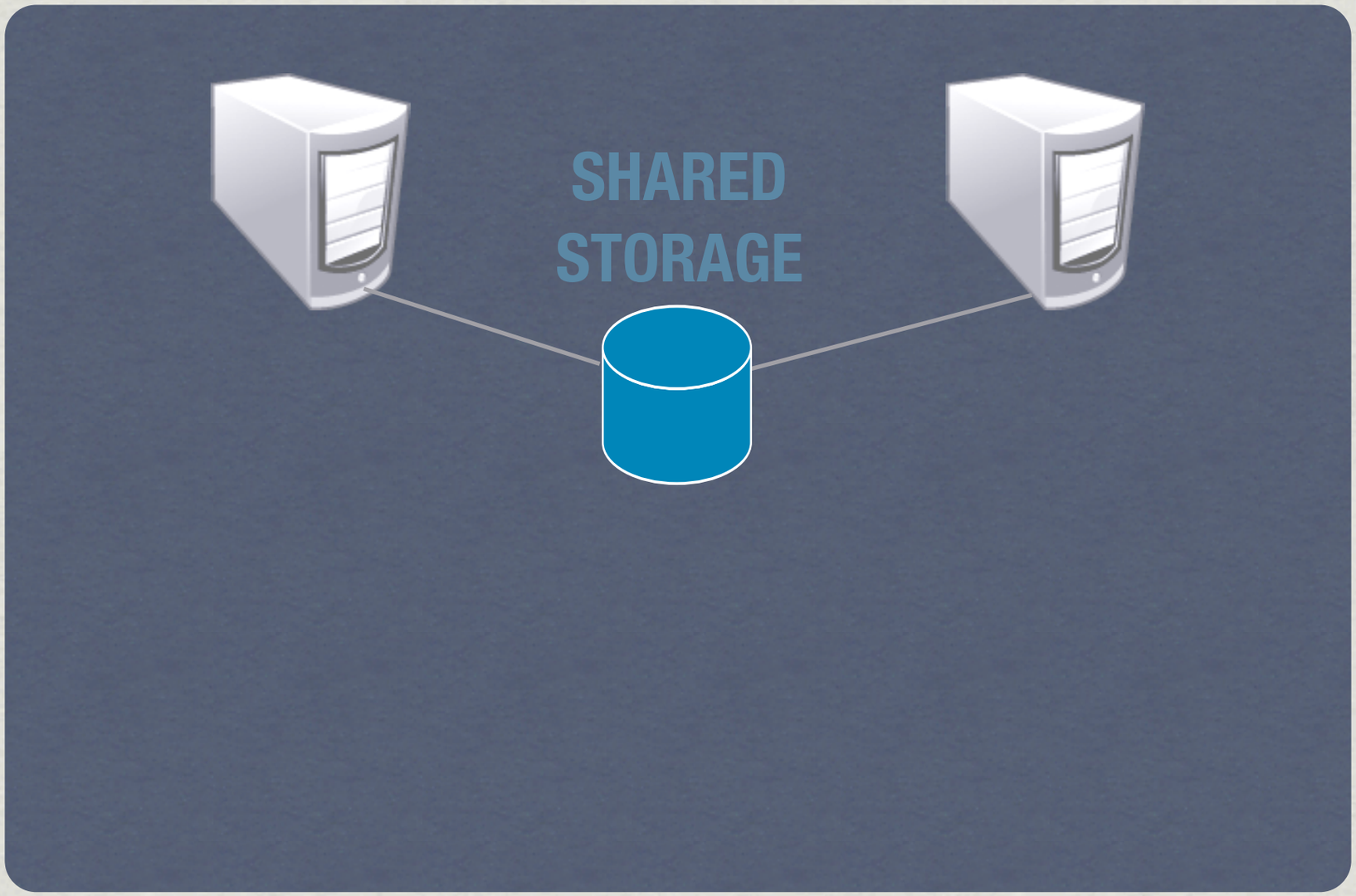
DRBD Protocols

- Protocol A: (asynchronous) write IO operations are reported as completed if they have been committed to the local device and the local TCP send buffer
- Protocol B: (semi-synchronous) write IO operations are reported as completed if they have been committed to the local device and the remote buffer cache
- Protocol C: (fully synchronous) write IO operations are reported as completed if they have reached both the local and remote devices
- “Truck” based replication - shortens initial sync time

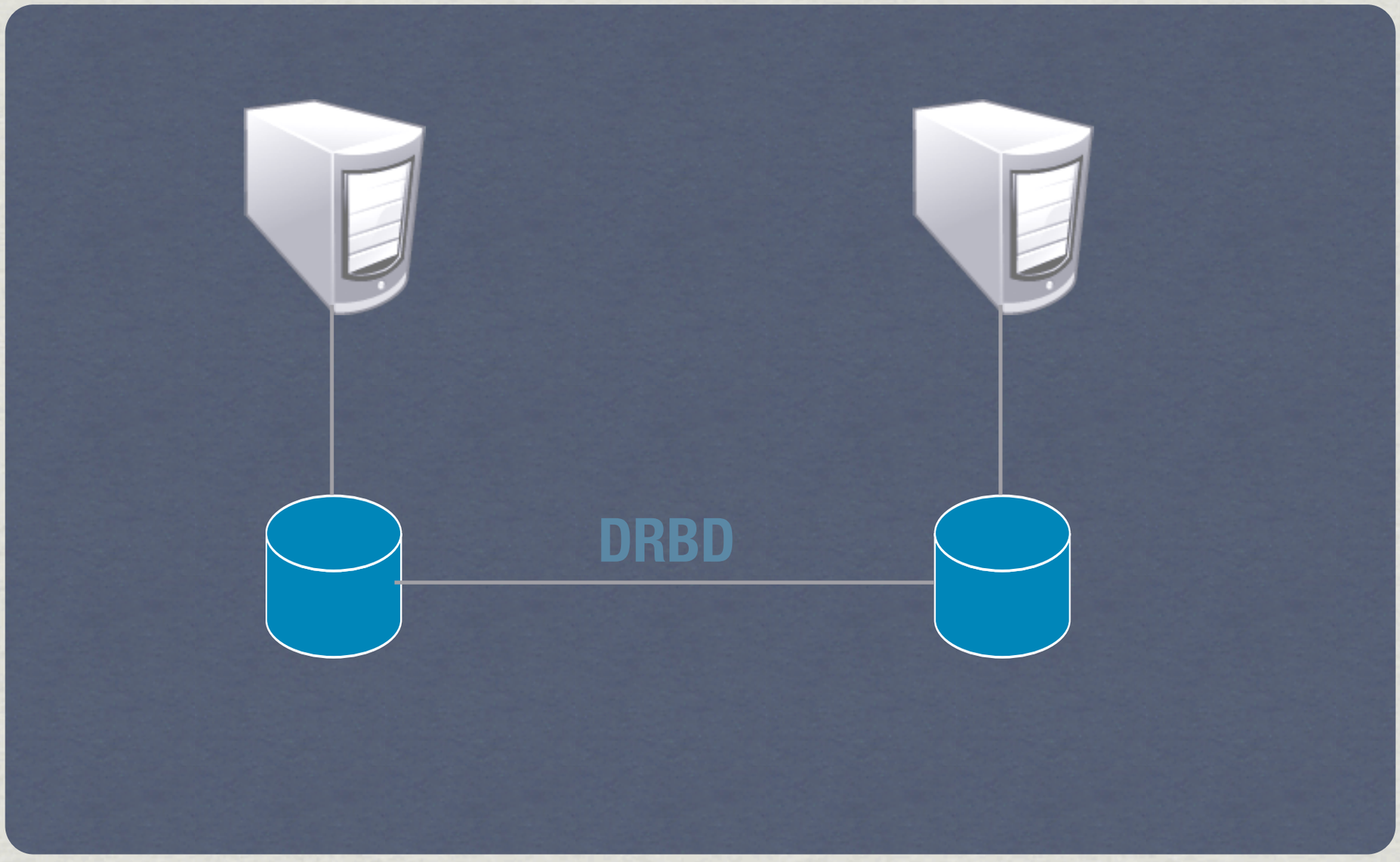
Applications Of DRBD

- Active/Passive Clusters
- Active/Active Clusters
- Disaster Recovery
- Read Only Data Replication To Remote Sites

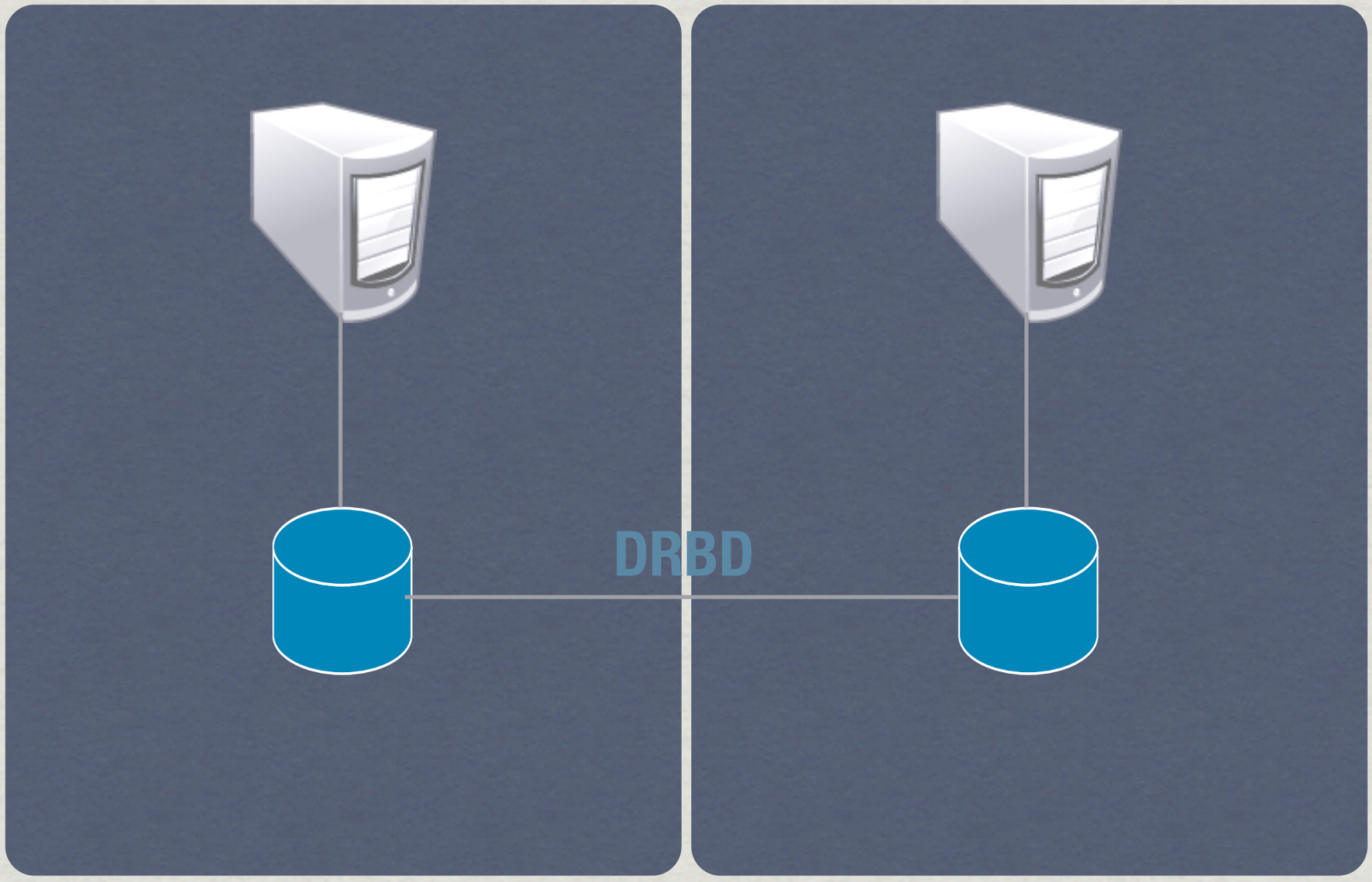
Traditional Cluster



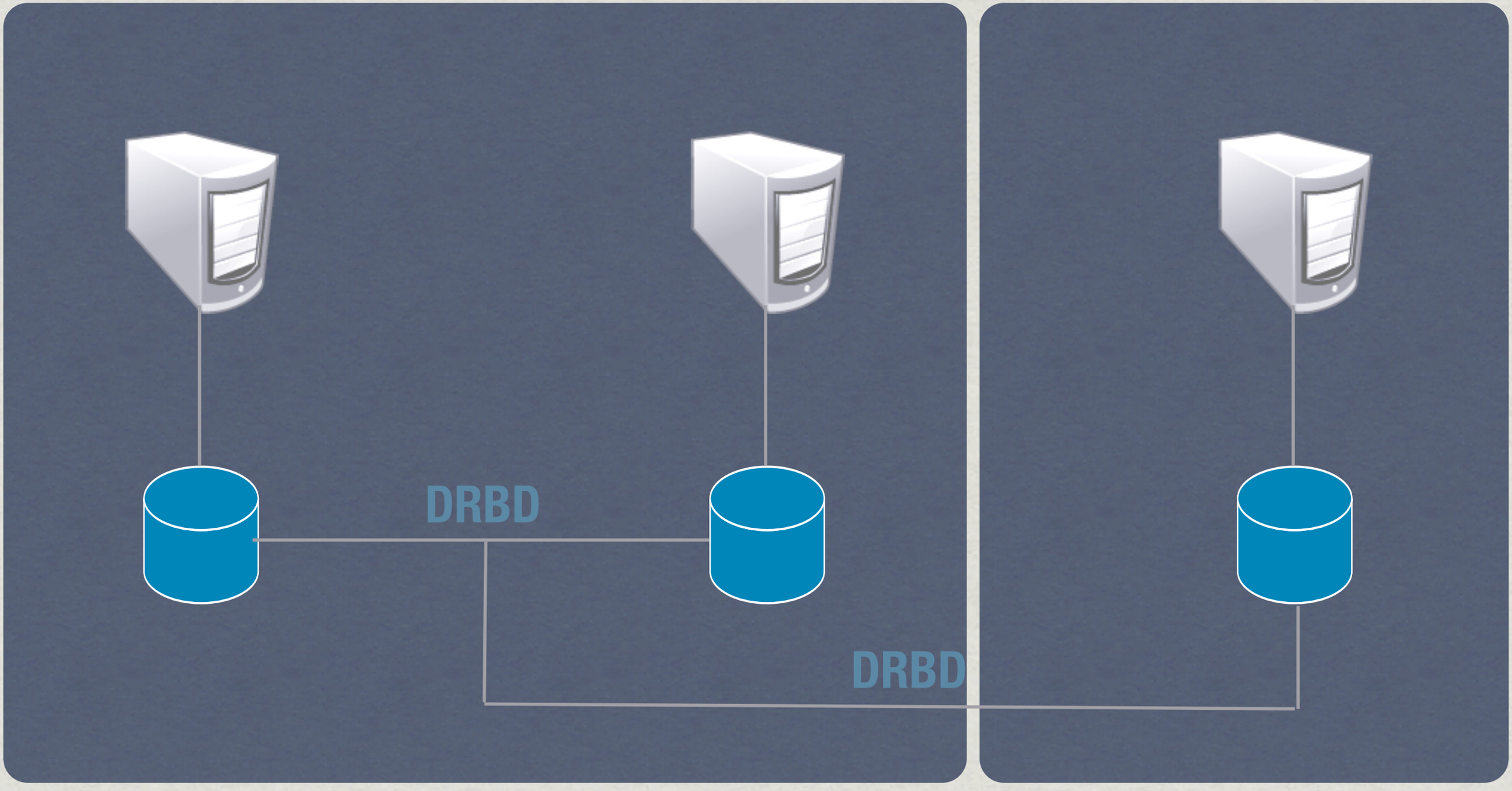
DRBD Cluster



DRBD Cluster



Three Node Cluster



BDP

- Bandwidth Delay Product
- Number of Packets in Flight
- Large Fat Network (LFN)
- BDP (bytes) = total_available_bandwidth (KBytes/sec) x round_trip_time (ms)
- 20,000 KByte/sec line x 40ms RTT = 800 KB Buffer (min)

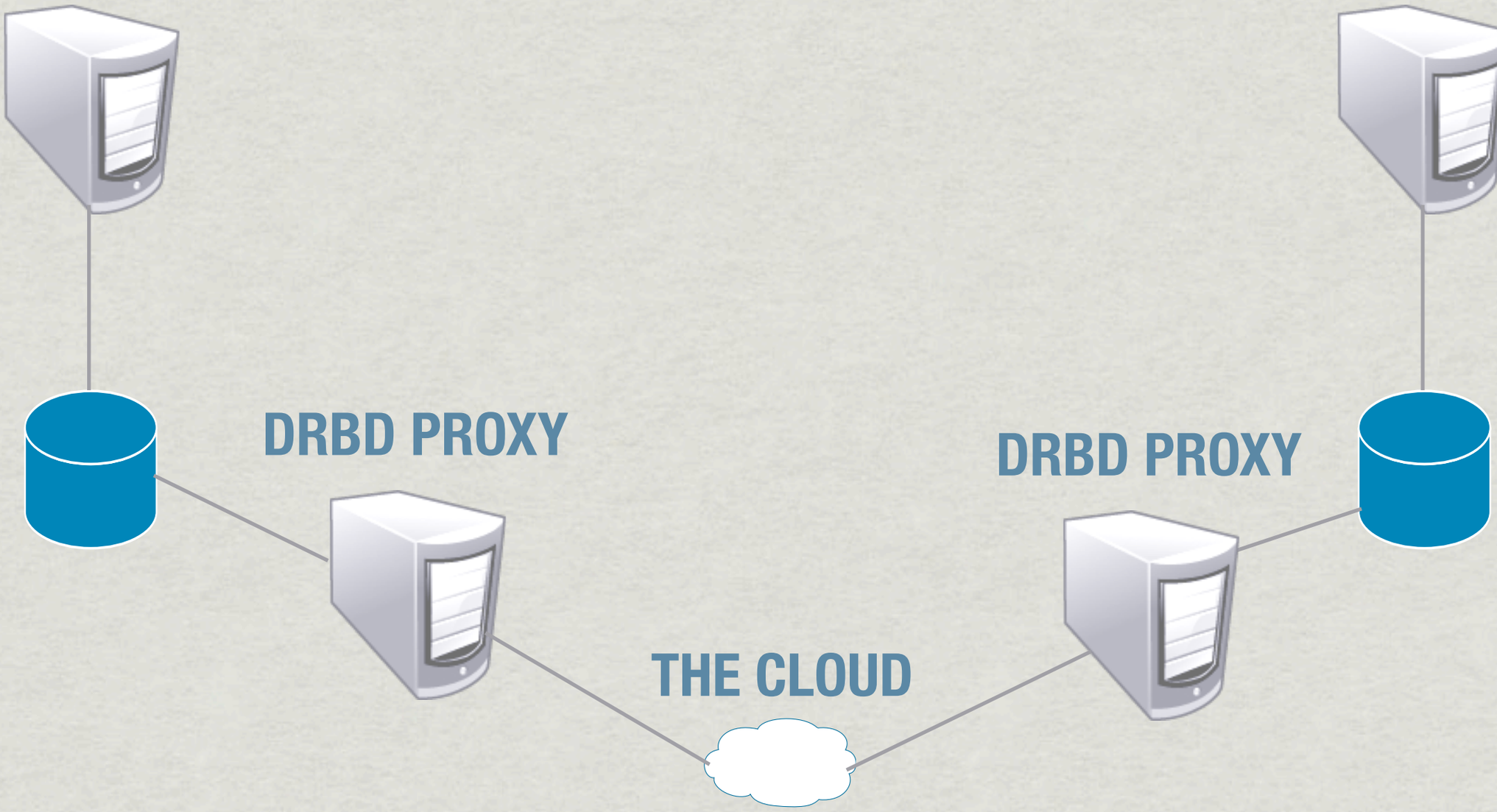
Linux Stack Tuning

```
/etc/sysctl.conf

# increase TCP max buffer size setable using setsockopt()
net.core.rmem_max = 16777216
net.core.wmem_max = 16777216

# increase Linux autotuning TCP buffer limits
# min, default, and max number of bytes to use
# set max to at least 4MB, or higher if you use very high BDP paths
net.ipv4.tcp_rmem = 4096 87380 16777216
net.ipv4.tcp_wmem = 4096 65536 16777216

# don't cache ssthresh from previous connection
net.ipv4.tcp_no_metrics_save = 1
net.ipv4.tcp_moderate_rcvbuf = 1

# recommended to increase this for 1000BT or higher
net.core.netdev_max_backlog = 3072

net.ipv4.tcp_max_syn_backlog = 8192
```

DRBD Proxy

- DRBDProxy is NOT open source
  - Requires DRBD version 8.2.7 or higher
- Protocol A is asynchronous to the point where your network buffers fill up
- DRBD Proxy is a buffer limited only by RAM size and address space
  - Can be a separate node, or run on the same nodes as DRBD devices

DRBD Proxy



Management Console

- this image shamelessly stolen from DRBD.ORG

DRBD Active/Passive

```
/etc/drbd.conf - active/passive

resource r0 {
         protocol        c;

         syncer {
                rate 110M;
                csums-alg md5;
                use-rle;
           }

          device    /dev/drbd0;
          disk      /dev/sdb1;
          meta-disk internal;

         on debian1 {
          address  192.168.128.128:7789;
          }

         on debian2 {
          address  192.168.128.129:7789;
          }
}
```

DRBD Active/Active

```
/etc/drbd.conf - active/passive

resource r1 {
        startup {
             become-primary-on both;
        }

        net {
            allow-two-primaries;
              after-sb-0pri discard-zero-changes;
              after-sb-1pri discard-secondary;
              after-sb-2pri disconnect;
        }
}

 protocol        c;

 syncer {
        rate 110M;
        csums-alg md5;
        use-rle;
 }

 device   /dev/drbd1;
 disk           /dev/sdc1;
 meta-disk      internal;

on debian1 {
 address  192.168.128.130:7790;
 }

on debian2 {
 address  192.168.128.131:7790;
 }
```

Proxy Config

```
/etc/drbd.conf - active/passive w/proxy

resource r1 {
        protocol        a;
        device          minor 1;
        disk            /dev/sdc1;
        meta-disk       internal;

        syncer {
               rate 110M;
               csums-alg md5;
        }

        proxy {
              compression on;
              memlimit 64M;
        }

        on drbda {
           address 127.0.0.1:7788;
           proxy on drbda {
                 inside 127.0.0.1:7789;
                 outside 192.168.168.11:7789;
           }
        }

        on drbdc {
           address 127.0.0.1:7788;
           proxy on drbdc {
                 inside 127.0.0.1:7789;
                 outside 71.175.110.7:7789;
           }
        }
```

Status

```
GIT-hash: 70a645ae080411c87b4482a135847d69dc90a6a2 build by root@debian1, 2009-10-15 14:15:12
 0: cs:Connected ro:Primary/Secondary ds:UpToDate/UpToDate C r----
    ns:24 nr:0 dw:24 dr:157 al:2 bm:0 lo:0 pe:0 ua:0 ap:0 ep:1 wo:b oos:0
 1: cs:Connected ro:Primary/Primary ds:UpToDate/UpToDate C r----
    ns:1030423 nr:248 dw:1030623 dr:54822 al:210 bm:7 lo:0 pe:4 ua:0 ap:4 ep:1 wo:b oos:0
```

Demo Time

- Awesome? Yup.

Where To Get It

- http://drbd.org - open source downloads, docs
- http://linbit.com - commercial support/downloads, drbd-proxy, management console
- Pre-Built
  - Debian Lenny Base Repository - apt-get
  - Centos Extras Repository - yum repo
- Roll your own from GIT Repository

The Solution

- DRBD + DRBD-Proxy + NFS + ESXi

Questions?

- gorkab@mysterons.org