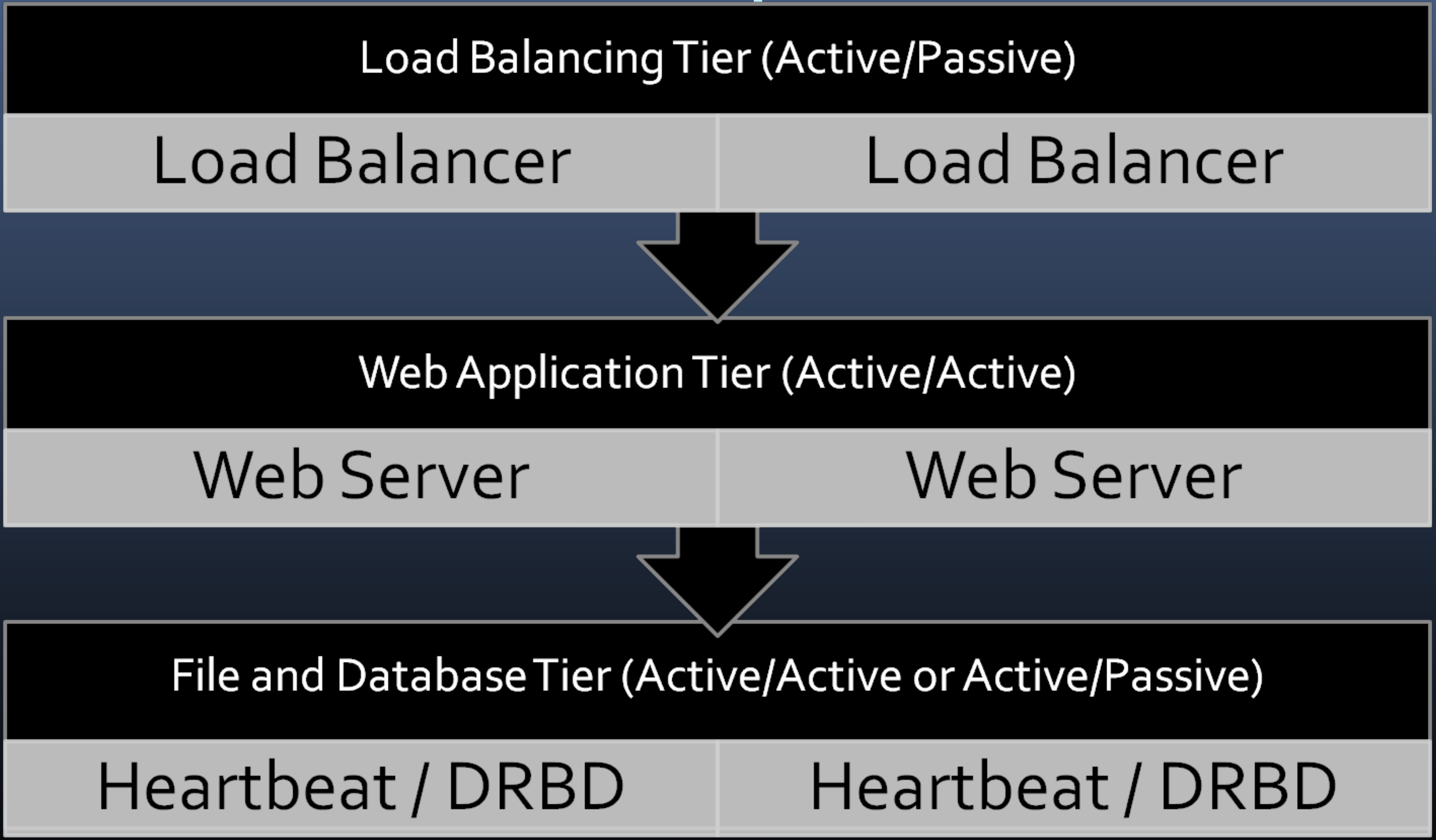
High Availability Low Dollar Clustered Storage

Simon Karpen

Karpen Internet Systems

skarpen@karpeninternet.com

Web Infrastructure Example



Load Balancing Tier (Active/Passive)

Load Balancer

Load Balancer

Web Application Tier (Active/Active)

Web Server

Web Server

File and Database Tier (Active/Active or Active/Passive)

Heartbeat / DRBD

Heartbeat / DRBD

Overview

- Shared storage with commodity hardware
- 100% Open Source software stack
- Minimal barrier to entry
- Scales down to laptop-sized demonstration
- Scales up to hundreds of TB, possibly low PB

Example Sites

- Shodor – A National Resource for Computational Science Education
  - http://www.shodor.org/

- VoiceThread –A Powerful New Way to Talk About and Share your Images, Documents and Videos
  - http://voicethread.com/

What Can This Do?

- File services – Samba, NFS
- Databases – MySQL, PgSQL, OpenLDAP, other Authentication
- Network services – DHCP
- Web services - any back-end infrastructure
- Anything app with persistent data

Limitations

- I/O rates limited by commodity hardware plus overhead
- Cross-site replication depends on available bandwidth and write rate
- Automating failover between more than two hosts is complex
- Linux support only

Components

- Linux (examples use CentOS)
- Hardware including local storage
- DRBD - Distributed Redundant Block Device
- Heartbeat - Linux-HA, manages failover
- Network - Gigabit or better strongly preferred

Operating System

- Recent Linux distribution
- Software is distribution independent
- May need software from “Extras” or equivalent
- Possible vendor support issues
- No support for FreeBSD, OSX, etc

Hardware

- Internal vs External redundancy
- Low cost: focus on external redundancy
- More 9’s: internal redundancy really helps
- RAID and network performance is key
- Two desktops or $500 special servers = proof of concept

DRBD

- Distributed Redundant Block Device
- Think “RAID-1 meets a network”
- Web site at http://www.drbd.org/
- Open source, support available from LinBit
- Supports active/passive or active/active (examples are all active/passive)

DRBD cont'd

- FAQ is at http://wiki.linux-ha.org/DRBD/FAQ
- Heartbeat plus DRBD's integrity checks work respectably as a fence
- Status in /proc/drbd
- Configuration in /etc/drbd.conf
- Configuration for each resource must match on each node

Sample DRBD Configuration

```
Resource “files” {
   protocol C;
   on drbd0 {
        device /dev/drbd0;
        disk /dev/sda4;
        address 192.168.232.10:7788;
        meta-disk internal;
   }
```

Sample DRBD Config Cont’d

```
    on drbd1 {
        device /dev/drbd0;
        disk /dev/sda4;
        address 192.168.232.11:7788;
        meta-disk internal;
    }
    syncer {
        rate 5M;
    }
}
```

Sample DRBD Command Lines

(On both hosts)

Create the metadata:

```
drbdadm create-md files
```

Bring up the DRBD itself

```
drbdadm up files
```

(One host only)

Initialize the DRBD based on one half

```
drbdadm -- --overwrite-data-of-peer primary files
```

Heartbeat

- Manages service failover
- You could substitute other cluster tools
- Part of Linux-HA project, http://www.linux-ha.org/
- Including with or readily available with most Linux distributions
- Configured in /etc/ha.d

Heartbeat

- Examples use v1 style configuration
- Controls access to DRBD devices
- Manages services that run on top of DRBD devices
- Helps prevent split-brain situation
- Not shown here, but you also need /etc/ha.d/authkeys (trivial)

Sample /etc/ha.d/ha.cf

```
ucast eth1 192.168.232.10
ucast eth1 192.168.232.11
keepalive 2
warntime 10
deadtime 30
initdead 120
udpport 694
auto_failback on
node drbd0
node drbd1
respawn hacluster /usr/lib64/heartbeat/ipfail
```

Sample /etc/ha.d/haresources

```
drbd0 192.168.232.20 drbddisk::files
  Filesystem::/dev/drbd0::/export/files::ext3::
  noatime nfs
drbd1
```

haresources Notes

- Additional services, filesystems, etc are space separated
- Centos5/RHEL5 NFS startup scripts have a bug that will break repeated failover/failback
- Patch is on the next slide; you WILL need this for reliable NFS failover
- This is a heartbeat v1 style configuration

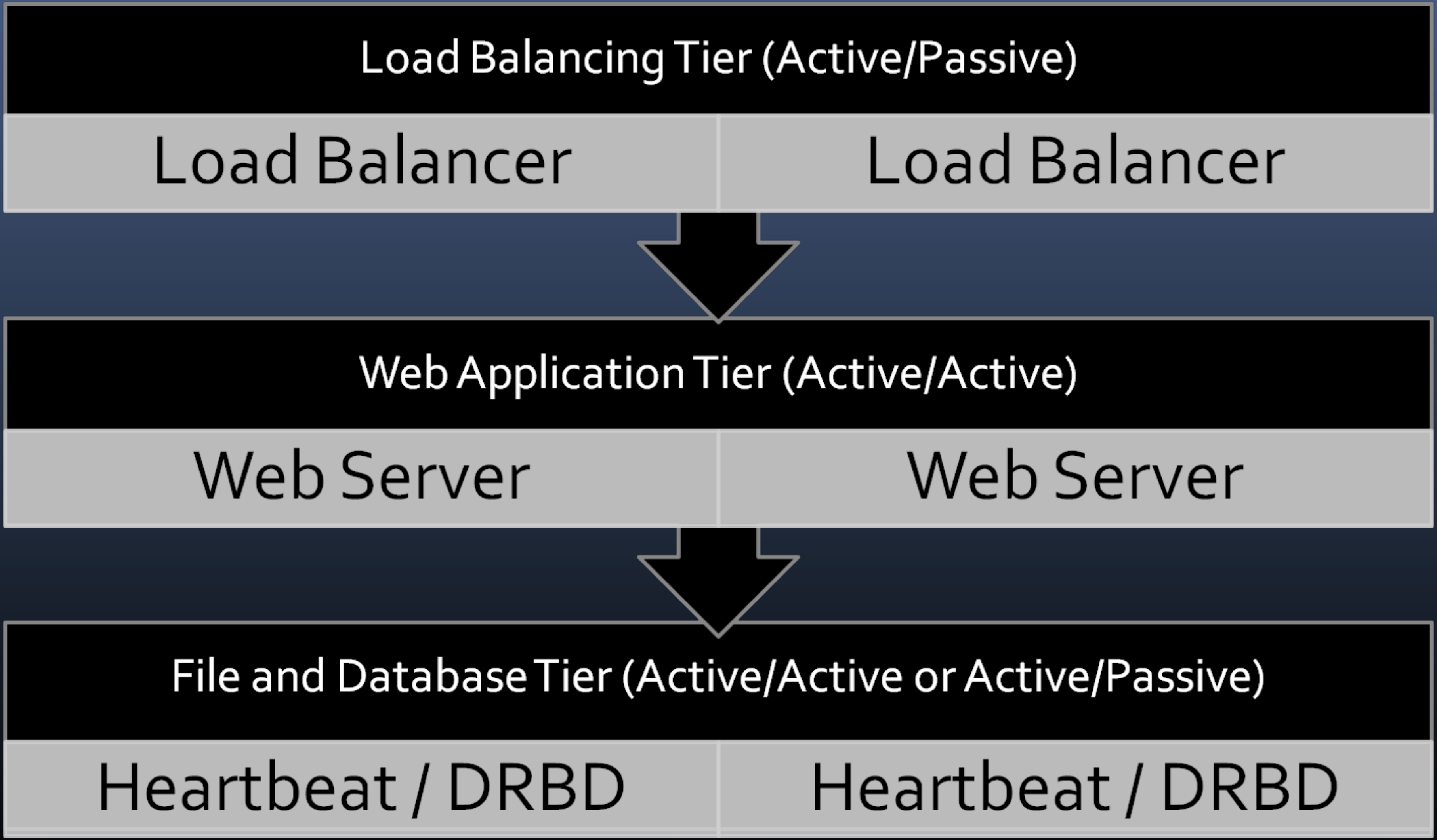
/etc/init.d/nfs patch (apply by hand)

```
@@ -134,6 +134,7 @@
  action $"Shutting down NFS services: " /bin/false
  fi
  [ -x /usr/sbin/rpc.svcgssd ] && /sbin/service rpcsvcgssd stop
+ killall -9 nfsd
  rm -f /var/lock/subsys/nfs
```

Actual Demonstration

- Three virtual machines (2 server, 1 client)
- Both running CentOS 5.3 x86_64
- VMWare Workstation
- Using the heartbeat and DRBD configuration already shown
- Simple NFS shares to CentOS client

Revisit Web Infrastructure



Load Balancing Tier (Active/Passive)

Load Balancer

Load Balancer

Web Application Tier (Active/Active)

Web Server

Web Server

File and Database Tier (Active/Active or Active/Passive)

Heartbeat / DRBD

Heartbeat / DRBD

Final Thoughts

- This is a “good enough” HA solution for many applications, at a non-HA price
- Better but not faster or cheaper than a single server.
- Cheaper but not better or faster than a replicated SAN or NAS (i.e. Netapp cluster)
- High Availability is not a replacement for backups

Questions?

- Any Questions? (Q&A and Disucssion)
- Slides will be posted on http://www.trilug.org/
- E-mail me at skarpen@karpeninternet.com