High Availability Low Dollar Clustered Storage

Simon Karpen
skarpen@shodor.org / simon@karpens.org

Thanks to Shodor for use of
this space for the meeting.

Overview

- Shared storage with no shared hardware
- Based entirely on an open source software stack
- Cost to try is minimal; a pair of $500 Dell special boxes will work
- Scales down enough that I can demo on a pair of virtual machines on this laptop
- Can scale up with hardware within reason

What Can This Do?

- File services – SMB, NFS, etc
- Databases – MySQL, PgSQL
- Authentication – LDAP, etc
- Network services – DHCP, etc
- Web services – better used for backend file/DB than actual web services
- Other applications? As long as persistent data is on the filesystem,

Limitations

- Not suitable for applications with very high I/O rates – limited by the commodity hardware plus overhead
- Cross-site replication depends on available bandwidth and write rate
- Automating failover between more than two hosts can be complex
- Linux support only (no BSD, etc)

Components

- Linux – Operating system. Examples are based on CentOS.
- Hardware – Just about anything that can run Linux and that has local storage
- DRBD – Distributed Redundant Block Device
- Heartbeat – Part of the Linux-HA project, manages device and service failover
- Network – You need connectivity between the hosts. Gigabit is preferred if available.

Operating System

- No point in paying for RHEL; you'd have to add DRBD yourself or from the CentOS repositories
- Redhat wouldn't support you anyway!
- Other Linux variants should be fine; Ubuntu and SuSE even ship DRBD in some versions
- Watch support cycles; in production, these clusters will be long lived. Fedora is probably a poor choice.

Other Operating Systems

- FreeBSD geom_gate/geom_mirror may be able to do something similar, but are not yet stable
- Nothing really there on Solaris yet; the Sun cluster tools in particular want shared storage
- I am not aware of anything similar on Windows or OSX

Hardware

- Think about internal redundancy versus external redundancy
- Where cost is the primary consideration, concentrate on external redundancy (two servers) plus backups
- When downtime is the primary consideration, look for 'sturdy' hardware (hardware RAID, redundant power supplies and feeds, etc)
- A pair of $500 Dell special servers is great for a proof of concept

DRBD

- Distributed Redundant Block Device
- Web site at http://www.drbd.org/
- Open source, but commercial support is available from LinBit
- Supports both traditional active/passive cluster and new support for active/active with a real cluster filesystem (OCFS2, GFS2, etc)
- Can run on LVM, or LVM on DRBD, or both

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
- Configured in /etc/drbd.conf, configuration for each resource must match on each node
- active/passive configuration has been very reliable in practice

Sample DRBD Configuration

```
resource "files" {
  protocol C;
  on c5drbd0 {
    device /dev/drbd0;
    disk /dev/vg0/voldrbd0;
    address 192.168.122.2:7788;
    meta-disk internal;
  }
  on c5drbd1 {
    device /dev/drbd0;
    disk /dev/vg0/voldrbd0;
    address 192.168.122.3:7788;
    meta-disk internal;
  }
  syncer {
    rate 5M;
  }
}
```

Sample DRBD Command Lines

By request, I have sample
DRBD command lines. All as
root on the host or hosts indicated.

<resource> is the name of the drbd, i.e. files
in the examples

On both hosts:
drbdadm create-md <resource>
drbdadm up <resource>

On one host (the primary):
drbdadm –overwrite-data-of-peer primary <resource>

wait for sync to complete
create your filesystem on the drbd device
(in our example, “mke2fs -j /dev/drbd0”)

drbdadm secondary <resource>
(assuming you will hand control to heartbeat)

What is Heartbeat?

- Tool to manage failover of services between nodes
- Web resources for more advanced configuration are mostly at http://www.linux-ha.org/
- Including with or readily available with most Linux distributions
- You could use other cluster management tools, but would have to do the integration
- ha.cf and haresources must match between nodes

Heartbeat Notes

- Examples all use Heartbeat v1 style configuration for simplicty, support of all DRBD features
- Heartbeat v2 CRM style config can work; see http://wiki.linux-ha.org/DRBD/HowTov2
- Controls access to DRBD devices and services that run on top of DRBD devices (i.e. NFS, MySQL, Samba, etc)
- Not shown here, but you also need /etc/ha.d/authkeys (trivial)

Sample Minimalist /etc/ha.d/ha.cf

```
ucast eth0 192.168.122.2
ucast eth0 192.168.122.3
keepalive 2
warntime 10
deadtime 30
initdead 120
udpport 694
auto_failback on
node c5drbd0
node c5drbd1
respawn hacluster /usr/lib64/heartbeat/ipfail
```

Sample Minimalist /etc/ha.d/haresources

```
c5drbd0 192.168.122.100 drbddisk::files
Filesystem::/dev/drbd0::/files::ext3::noatime nfs
c5drbd1
```

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
- Again, this is a v1 non-CRM configuration. You can use a v2 CRM type configuration; there are advantages and disadvantages of both.

/etc/init.d/nfs patch

```
--- nfs.orig 2008-06-08 11:56:02.000000000 -0400
+++ nfs 2008-06-08 11:56:09.000000000 -0400
@@ -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
  ;;
   status)
```

Actual Demonstration

- Two virtual machines
- Both running CentOS 5.1 x86_64
- KVM virtualization, default Fedora configuration
- Using the heartbeat and DRBD configuration already shown
- This is obviously a simple minimal setup; this is how you get started. You will need to customize for your own applications

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)
- Replication is not a replacement for backups
- Replication is not a replacement for backups
- Replication is not a replacement for backups

Questions?

- Any Questions? (Q&A and Disucssion)
- A link to the slides will be up on http://ncsysadmin.org/
- A link to the video will also eventually make its way to http://ncsysadmin.org/