Red Hat Advanced Server 2.1

Advanced Level

Red Hat's latest flagship goes by the name of Advanced Server. Approved by major software publishers and equipped with enterprise features such as high availability and clustering the Advanced Server targets the more demanding customer, but despite the version number 2.1 it is quite obviously a newcomer.

BY MIRKO DÖLLE, ULRICH WOLF & ACHIM LEITNER

Remember Red Hat for Oracle or Red Hat for SAP? These were both available as separate products and certified by the appropriate ISVs (Independent Service Vendors). To prevent this list from running into the middle of next week, the Marketing guys in Raleigh have come up with the Advanced Server. At least 20 ISVs have given the go ahead with the list including major players like the IBM Software Group, application server specialist BEA and SAP. Additionally, Red Hat is asking the hardware manufacturers to climb on board.

As the SuSE Linux Enterprise Server proves, certification justifies a much higher selling price in its own right. But in contrast to their competitor, Red Hat have added some technical enhancements and are pushing the products scalability on SMP machines, cluster support, load balancing via Piranha and high availability. In order to do justice to Red Hat's technical claims we decided to focus our activities on setting up a cluster to provide high availability with two node failover.

The availability of the distribution itself was not too hot. Although we waited until well after our editorial deadline, Red Hat was unable to deliver a boxed product to our test lab. So our test is based on the CDs we created from the ISO images that Red Hat finally managed to upload to our FTP server.

Installation

The installation procedure for Advanced Server is very similar to the procedure already used in Professional 7.3. Red Hat uses the same GUI installation program for both distributions. The welcome page now additionally offers the Advanced Server option, in contrast to the various server and workstation variants available in the Professional edition.

Advanced Server is an international version providing multi-lingual support, although the documentation is entirely in English.

The character based installation does not seem to be any different from the

Tourismus, Frutigen, visipix.com

Figure 1: Two mode failover cluster configurations require a dual channel SCSI RAID or fiberchannel solution that can be accessed simultaneously by both nodes.

Professional 7.3 Red Hat distributions, although you may discover one or two issues (as we did), if you need special keyboard layouts.

Setting up a firewall on a cluster is more complex than on a single machine. You cannot perform the installation just using the defaults (medium security level, allow no services or just DHCP) because the defaults will interfere with the cluster configuration. We recommend omitting the firewall installation at this step and manually adding customized rules for the cluster at a later stage.

Red Hat distributions still use the Gnome desktop, although a KDE option is available. But production systems will tend to be managed remotely, and that makes the GUI redundant. To install a text-based environment, you simply disable the Gnome package during the installation.

Hardware en masse

The documentation describes a two node failover as a typical setup for Advanced Server 2.1, so we decided to base our test on this scenario. The cluster for our test system comprised two Dual Athlon machines running at a clock speed of 1.533 and 1.666 GHz respectively, both equipped with an Adaptec 29160 U160 SCSI controller. We installed the Red Hat system on the internal hard disks of both machines.

There were no complaints regarding hardware, although a Promise Fasttrack 100 RAID 0 system was recognized as two separate disks. This meant having to break up an existing RAID array or replace it with a software RAID array. And there was a slight APIC issue with the Asus A7M266-D board in the first machine. The kernel kept on crashing during initialization, but the "noapic" boot parameter soon sorted that out.

Two Channel SCSI

We stored data for the cluster services on an Easy-RAID X12 by Starline Computers [2]. This SCSI / IDE RAID system (see Figure 1) features a dual channel SCSI host controller and twelve 120 GB drives, although we used only the first four. When we attempted to mount the total capacity of 1.44 TB, we could not access the device. Linux complained about read errors on "/dev/sda". To allow both machines simultaneous access to all the partitions on the RAID system we then configured the four disks as a large share.

Red Hat supports fiberchannel systems, which you would need to configure for parallel access. NAS systems are not currently supported and the cluster configuration will not talk to network drives.

Network Power Switch

Red Hat's "Cluster Manager Installation and Administration Guide" [3] recommends the use of a power switch, to completely power down a faulty machine in case of node failure. The idea is to prevent the common RAID system from freezing. APC kindly provided us with a Master Switch AP9212 (Figure 2), which featured eight switchable outlets. We attached the power switch to the network leaving the serial port unused.

However, we found the cluster software was unable to control the power switch correctly: Instead of powering off a failed machine (Immediate Off) the cluster merely emitted an Immediate Reboot signal, causing the failed machine to power off for a few seconds before it then powered on again.

Depending on the BIOS configuration the computer may attempt to restart, and in this case a damaged SCSI controller could lead to the RAID system freezing. Since the software will not transmit a second signal, this would take the whole cluster down.

Cluster Installation

The configuration of the cluster software with the "cluconfig" console tool is detailed in the Cluster Guide. Although the software has outstripped the guide in some places, this should not give the administrator too much of a headache.

You should be cautious of following all the sample configurations given without considering your options. The Cluster Guide recommends the activating of the "Relocate when preferred member joins the cluster" option for an Apache configuration on page 126, but fails to mention that the relocating will drop any current sessions. This causes active downloads to fail when the primary node rejoins the cluster after a failure.

Figure 2: The Master Switch AP9212 can switch eight power circuits individually. The integrated web server provides administrator access to the management software via SNMP or telnet.

```
bash
[root@lab2 root]# cluadmin -- service relocate apache
Relocating apache, Error; failed; service apache not relocated,
[root@lab2 root]# cluadmin -- service relocate nfs_redaktion
Relocating nfs_redaktion, Error; failed; service nfs_redaktion not relocated,
[root@lab2 root]#
```

Figure 3: In case of interrupted network services, the services cannot be transferred and therefore fail.

The nodes use quorum partitions to transfer status information, for which no details are available. The partitions, which are about 10 MB and mounted as unbuffered raw devices, store status information on the clusters and active services. You need to use separate RAID partitions for your data to provide the redundancy for individual services. The node that owns the process will mount the partition assigned to the process.

Interrupted Connections

We used an Active-Active configuration for our Cluster comprising one machine with an NFS drive as its primary node, and the other with an Apache web server. In this constellation one machine would take over the service that had failed on the other machine. Failover means the restoring of services of the failed node as quickly as possible, but this does not mean necessarily that active connections will be kept. Our clients could only continue working unaffected by the failure if they were using connectionless protocols (such as NFS).

When a node fails over, the IP address of the cluster service is assigned to the other machine. The address is then bound to the network device responsible for the subnet by IP aliasing. This means that the hardware address of the cluster will change to match, and needs to be accounted for when configuring switches or routers, and also that redundant services will need an IP address of their own.

Hidden Heartbeat

You will need to configure at least one heartbeat channel for the cluster operations. The nodes use the heartbeat channel to check how the other nodes respond, if a node fails to update the timestamp on the quorum partition.

The heartbeat channel is unused in the current 2.1 Version of Advanced Server. Only the status output from "clustat" or "cluadmin" (Figure 5) shows you if the heartbeat channel is online or offline. You cannot define any actions for these cases, and there was no sign of scripting access. Red Hat has stated that this feature will be available in the following version.

The cluster software does not offer any options for launching customized actions on failure of a service or device. You can use the status function in the services init script only to implement a verification function. The cluster software calls the init script with the "status" flag set at predefined intervals and the Cluster Manager determines whether to restart the service based on an analysis of the return value. The administrator can specify what details the status check covers. The Apache script, for example, checks whether the daemon is running.

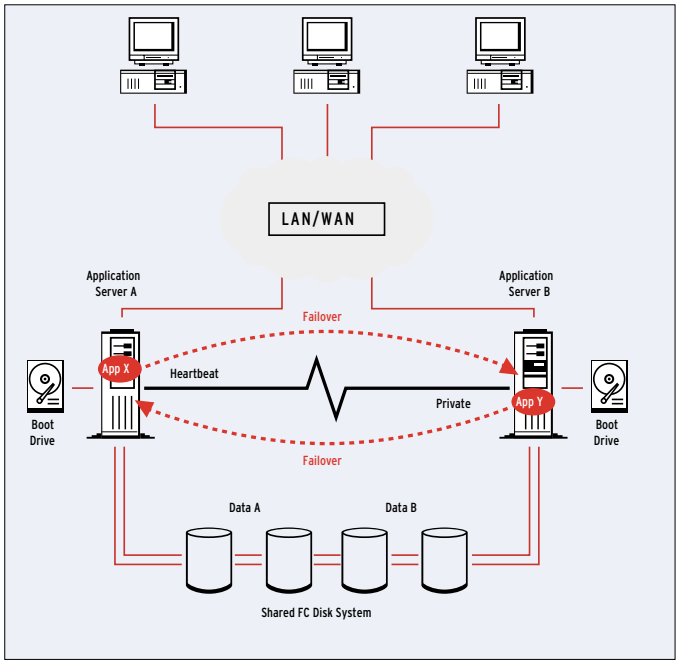


Figure 4: Both servers have redundant connections to disk system(s), but Red Hat Linux Cluster Manager controls access.

Relocate or bust

But don't expect a failed status check to launch a "relocate". If the script detects

```
bash
Cluster Status Monitor (Pensacola-Cluster)                    20:47:12

Cluster alias: lab0.linux-magazin.de

======================== M e m b e r   S t a t u s ========================

  Member          Status     Node Id    Power Switch
  --------------- ---------- ---------- ------------
  lab1            Up         0          Good
  lab2            Up         1          Good

======================== H e a r t b e a t   S t a t u s ==================

  Name                          Type       Status
  ----------------------------- ---------- ------------
  /dev/ttyS0   <--> /dev/ttyS0  serial     ONLINE

======================== S e r v i c e   S t a t u s ======================

                                                 Last                 Monitor  Restart
  Service         Status   Owner                 Transition           Interval Count
  --------------- -------- --------------------- -------------------- -------- -------
  nfs_redaktion   started  lab1                  20:36:44 Jun 14      30       0
  apache          started  lab2                  20:35:43 Jun 14      120      0
cluadmin>
```

Figure 5: The cluster status can be queried using "clustat" or interactively using the "cluster status" flag with cluadmin. The service section shows how the services are distributed across the cluster nodes.

```
bash
Cluster Status Monitor (Pensacola-Cluster)                    20:53:24

Cluster alias: lab0.linux-magazin.de

======================== M e m b e r   S t a t u s ========================

  Member          Status     Node Id    Power Switch
  --------------- ---------- ---------- ------------
  lab1            Up         0          Good
  lab2            Up         1          Unknown

======================== H e a r t b e a t   S t a t u s ==================

  Name                          Type       Status
  ----------------------------- ---------- ------------
  /dev/ttyS0   <--> /dev/ttyS0  serial     ONLINE

======================== S e r v i c e   S t a t u s ======================

                                                 Last                 Monitor  Restart
  Service         Status   Owner                 Transition           Interval Count
  --------------- -------- --------------------- -------------------- -------- -------
  nfs_redaktion   started  lab1                  20:36:44 Jun 14      30       0
  apache          started  lab2                  20:35:43 Jun 14      120      0
cluadmin>
```

Figure 6: Although only the network connection to the second cluster node has failed, the power switch status is unknown. This effect also occurs if you have not configured a power switch.

an error condition that would necessitate switching to a backup system, you have to launch this action using the "cluadmin – service relocate *service*" syntax. The cluster server handled a total node failure gracefully; depending on the service they were using, the clients simply had to repeat a file transfer process.

But a partial failure caused a whole bunch of unanticipated problems. Although the affected node did a clean reboot after disconnecting the SCSI subsystem, there seems to be no way to deal with a disconnected network cable. Although the heartbeat channel and the SCSI connection were both active, the missing network link between the two nodes meant that it was impossible to relocate a service to a backup machine: "cluadmin" kept on reporting errors (see Figure 3). Figures 7 and 8 show our attempts to relocate the service via the console.

While we were searching for the cause of this problem with "tcpdump", we noted that "clupowerd" continually talks to its neighbors via TCP/IP port 4004. The daemon seems to be responsible for power switches and that would explain the "unknown" status in Figure 6, where the network connection is down.

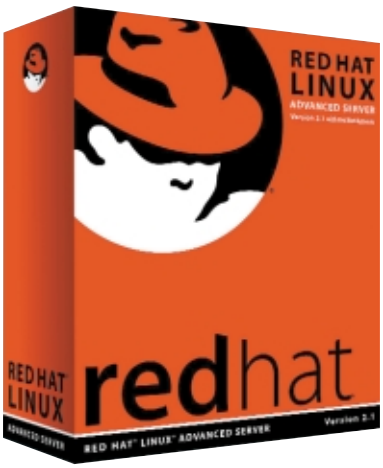
While relocating a service we noticed some traffic between the nodes on port 4002, i.e. the port the Cluster Service Manager "clusvmgr" listens on. It seems that service relocations are negotiated via this connection, and that means a failure is inevitable if the network connection is down. We will need to check the sources to be sure, though, because we could not find any man pages for the Cluster Tools, or any documentation anywhere else for that matter. Even the "--help" switch only worked on rare occasions.

So Red Hats failover solution only works as advertised in case of total system failure, and that is not our idea of a high availability solution. The remedy would seem to be a script that uses a power switch to power a node off. Or as a colleague put it "All we need is someone to watch the machine and blast it with a shotgun if something goes wrong."

Conclusion

Advanced Server 2.1 is a tried and trusted solution, that is in line for certification by hardware and software manufacturers. If you need this and are also a faithful Red Hat customer, the Red Hat flagship is your only option. However, the high availability features were not convincing. The cluster can only manage two nodes and despite the additional hardware resources required it seemed incapable of dealing with error conditions apart from the total failure of one server. ■

Red Hat Advanced Server 2.1



Scope: 4 CDs, 2 manuals

Support: 12 months Red Hat Network and maintenance

Basic: 12 months support for installation and configuration

Standard: 12 months all-in support, 4 hour response time (weekdays)

Premium: 12 months all-in support, 1 hour response time (24x7)

Price: US $800 (Basic), US $1,500 (Standard), US $2,500 (Premium)

INFO

[1] Red Hat: *http://www.redhat.com*

[2] Starline Computer: *http://www.starline.de/produkte/easyraid/easyraid_x12/easyraid_x12.htm*

[3] Easy-RAID X12: *http://www.phertron.com/products/easyraid_x16/erx16_fc.htm*

[4] Cluster-Guide: *http://www.redhat.com/docs/manuals/advserver/RHLAS-2.1-Manual/cluster-manager*

```
bash
  lab1          Up         0          Good
  lab2          Up         1          Unknown

======================== Heartbeat Status ====================

  Name                          Type       Status
  ----------------------------- ---------- ------------
  /dev/ttyS0   <--> /dev/ttyS0  serial     ONLINE

======================== Service Status ======================

                                           Last            Monitor  Restart
  Service        Status   Owner            Transition      Interval Count
  -------------- -------- ---------------- --------------- -------- -------
  nfs_redaktion  started  lab1             20:36:44 Jun 14  30       0
  apache         started  lab2             20:35:43 Jun 14  120      0
cluadmin> service relocate
  0) nfs_redaktion
  1) apache
  c) cancel

Choose service to relocate: 1
Are you sure? (yes/no/?) yes
Relocating apache. Error: failed: service apache not relocated.
cluadmin>
```

Figure 7: Following a failure of the network connection to "lab2", there was not even a manual option available for relocating Apache to a running machine.

```
bash
======================== Service Status ======================

                                           Last            Monitor  Restart
  Service        Status   Owner            Transition      Interval Count
  -------------- -------- ---------------- --------------- -------- -------
  nfs_redaktion  started  lab1             20:36:44 Jun 14  30       0
  apache         started  lab1             20:57:46 Jun 14  120      0
[root@lab2 root]# route add 192.168.1.191 lo
[root@lab2 root]# cluadmin
Fre Jun 14 20:58:17 CEST 2002

You can obtain help by entering help and one of the following commands:

cluster     service       clear
help        apropos       exit
version     quit
cluadmin> service relocate
  0) nfs_redaktion
  1) apache
  c) cancel

Choose service to relocate: 1
Are you sure? (yes/no/?) yes
Relocating apache. Error: failed: service apache not relocated.
cluadmin>
```

Figure 8: When a service needs to be relocated, the cluster software obviously attempts to contact the other node via Ethernet. A faulty route could take the cluster down.