High-Availability Software for Linux

Status and future of the Linux-HA project

Alan Robertson
IBM Linux Technology Center
alanr@unix.sh
Purpose: To give an overview of Open Source HA software for Linux.
The Desire for HA Systems

Who wants low-availability systems?

♦ Why are so few systems High-Availability?
Barriers to HA systems

- Hardware costs
- Software costs
- Complexity
- Standards
Potential User Community

Linux

HA
What would be the result?

🔹 Drastically multiplying customers multiplies experience - products mature faster
🔹 OSS developers grow with customers
🔹 OSS Clustering is a disruptive technology (both HA and HPC)
Open Source HA Software for Linux

- Cluster Management (Linux-HA)
- Filesystem Replication (DRBD)
- Load Balancing (LVS)

NOTE: These are all available on all Linux platforms, as well as FreeBSD and Solaris
Provides basic HA failover capabilities

Very simple to install, manage, understand

Active, open development community

Thousands of production sites worldwide

In production for mission-critical sites > 3 years

Wide variety of industries, applications

Shipped with most Linux distributions

Supports DB2, WAS, many others

Readily extensible to support other products
IP Failover

www.linux-ha.org = 135.9.216.110

"Normal" Configuration

"just.linux-ha.org"
135.9.216.100
135.9.216.110

"doit.linux-ha.org"
135.9.216.101

"Backup" Configuration

139.9.216.100

"doit.linux-ha.org"
135.9.216.101
135.9.216.110

Requests

High Availability
IBM/Linux-HA Successes

- **Emageon** – medical imaging services
- **Cisco** – bases a product off Linux-HA and DB2
- **ADC** - telco provisioning manager product (w/ x330/335)
- **Contraloria General de la Republica** (Colombian government)
- **Incredimail** bases their mail service on Linux-HA on IBM hardware
- The **Rose F. Kennedy Center for Research in Mental Retardation and Developmental Disabilities** uses Linux-HA on x340s with 2 TB of storage
- **Bavarian Radio Station** (Munich) used Linux-HA and xSeries for coverage of 2002 Olympics in Salt Lake City
- **Citysavings Bank** in Munich (infrastructure)
- **Banco do Brasil** – 7500 (!) clusters planned
- **University of Toledo** – 20K user Computer Aided Education System
Linux-HA Successes

- The Weather Channel (weather.com)
- Sony
- MSC Software – uses it for OSCAR “head node”
- Intuit (Quicken, TurboTax, etc.)
- ISO New England manages the New England power grid using 12 Linux-HA clusters
- Many more listed at: http://linux-ha.org/heartbeat/users.html
"Setting up heartbeat was cake and pie (piece of cake, easy as pie). I can't believe how easy it is to set up and use. Oh yeah, it works perfectly too."

"... it works very well indeed! Fast and simple, it is a far cry from the lumbering cluster software I was using under [my old OS]"

"...I aggressively removed the power cables of an in-production x330 running DB2. Then I watched customer's smile getting bigger and bigger seeing DB2 and all resources becoming available in the second machine." (IBM Brazil)

"... just yesterday I got a chance to test it when someone pulled the power cable from the active box - worked flawlessly."
Complexity is the Enemy of Reliability...

As Proposed by Marketing

As Architected

Beta Version
Complexity (cont'd)

Production Version
As Installed
What The Customer Really Wanted

Linux-HA is great tire-swing
HA ;-)
Near-Future Linux-HA Developments

- Fast failover
- Larger (more than two-node) clusters (~ 2004)
- Addition of features needed for Carrier Grade Linux
- Integration with OpenGFS
- Implementation of OCF standard APIs
- Basis for OCF reference implementation
Linux-HA Resources

♦ Article in November 2003 Linux Magazine (Linux-HA, DRBD and LVS)
♦ Article in December 2002 LinuxJournal on HA OpenLDAP servers
♦ Interview with Alan Robertson in September 2002 LinuxJournal
♦ IBM Red Book on Open Source Solutions
♦ IBM Red Paper on Linux-HA on zSeries
♦ Article in September 2001 SysAdmin magazine on HA file servers
DRBD - RAID over the LAN

- Block-device (filesystem) level replication
- Clever synchronization methods make resyncs faster, decrease latency, preserve integrity
- Useful for both HA and Disaster Recovery
- **NO** single point of failure
- Extremely cost-effective
  - $200 (max) instead of $20,000 ($USD)
- Probably not suitable for high-end write-intensive applications, or very large filesystems
DRBD Futures

- Faster full-syncs (using md5 checksums)
- Allow active/active partitions
- Allow primary/secondary switch during syncs
- Become part of standard kernels
LVS - The Linux Virtual Server Project

- LVS is the standard Linux Load Balancer
- Called "ipvs" in the standard Linux kernel
- Stable, fast, flexible
- Especially suitable for large "server farms"
LVS in action

Linux Virtual Server

Real Server #1
Real Server #2
Real Server #3
Real Server #4
Real Server #5

User

Internet or Intranet

Load Balancer (LVS Director)

High Availability
These Pieces Work Together Nicely

♦ Each of these independent services can work together to scale to large systems
♦ All single points of failure can be eliminated
♦ High-Availability, Load Balancing work together nicely
Linux-HA, LVS and DRBD

High Availability
In HA systems – sharing information is the key difficulty

How best to share up-to-date information

Options include shared storage, replication

NOTE: Some HA packages require shared storage (Red Hat, Mission Critical)

Linux-HA doesn't care ;-)

HA Solutions
Sharing by Replication

♦ Use “application-native” method (DNS, LDAP, DB2, etc.)
♦ Use DRBD
♦ Use rsync (for some cases)
Replication Pros/Cons

- Can eliminate **ALL** SPOFs
- Very inexpensive (basically free)
- Typically easy to install (no special hardware)
- Can separate cluster nodes by some distance
- Can add latency to application
- Can result in slightly out-of-date data
- Resyncing large amounts of data can be slow
- Doubles storage needs
Ideal Replication Setup

Disks can be External or Internal

External Network

Hardware or Software Mirroring

Primary Mirror-A

Primary Mirror-B

"Primary" Host

Serial Link

Replication Link (100-1000mbit)

"Secondary" Host

Secondary Mirror-A

Secondary Mirror-B

High Availability
HA Shared Storage Options

- EXP300 + IBM ServeRAID clustering
- FASTT and other FC storage systems
- ESS storage systems
- Non-IBM hardware
  - Winchester Systems external SCSI/FC RAID boxes, and other similar boxes
Shared Storage Pros/Cons

- Resyncs not required after host outages
- Very little change in application behavior
- Protection constant across host outages
- Expensive, needs special hardware
- Eliminating all SPOFs is typically difficult and expensive
- Inter-host Distance is sometimes limited
Other Technologies

♦ Journalling file systems
  • EXT3 (can also journal data)
  • ReiserFS
  • JFS
  • XFS

♦ Logical Volume Managers
  • LVM
  • EVMS
  • LVM-2
Volume Management Capabilities

- Online volume growth
- Online volume movement
- Volume Snapshots
Why Clustering Standards?

- At least 5 OSS HA "products"
- 10-20 Proprietary HA Products
- A VERY large number of HPC "offerings"
- Each has its own API and expectations
- No "900 pound gorillas" in the Linux market to create/dictate "standards"
Who is hurt?

♦ OSS projects and customers
  • Effort duplication
  • Developer, tester dilution
♦ Proprietary vendors - duplicate efforts with each other and OSS projects, and have trouble recruiting middleware partners
♦ Middleware vendors - effort duplication, and lots of workload from HA vendors
♦ End users - incomplete, incompatible products, confusion, fear, uncertainty, doubt
♦ Linux community - weakened clustering offerings
Open Clustering Framework

- Two-pronged approach
  - Define standard cluster APIs
  - Create Component-based Reference implementation

- Both proceed together

- The standards will be a 900-pound penguin...
Project Structure

API Definition

- Select Areas of Interest
- Create Subteams
- Define APIs
- Reach agreement

Reference Implementation

- Create
- Plumbing/Infrastructure
- Coordinate with API definition
- Define Framework components
- Implement components
Standard Clustering APIs

- Neutral (agnostic)
- Royalty-Free
- For OSS or proprietary software
- Creates opportunities for interoperability
- OCF is a working group of Free Standards Group
Current Status

- Mailing list created for discussions
- API drafts out for three areas
  - Resource agent API
  - Event API
  - Membership API
- Resource Agent API ready for public comments
Conclusions

♦ Open Source HA is alive and well on Linux
♦ Open Source HA will have VERY significant volumes
♦ Open Clustering Standards will benefit Linux as a whole
♦ IBM has a leadership position in both of these areas
♦ Let us know about your deployments!
References

http://linux-ha.org/
http://linux-ha.org/download/
http://linux-ha.org/heartbeat/users.html
http://opencf.org/
http://www.drbd.org/
http://www.LinuxVirtualServer.org/

Alan Robertson  alanr@unix.sh
Legal Statements

♦ IBM is a trademark of International Business Machines Corporation.

♦ Linux is a registered trademark of Linus Torvalds.

♦ Other company, product, and service names may be trademarks or service marks of others.

♦ This work represents the views of the author and does not necessarily reflect the views of the IBM Corporation.