UbuntuClusters

- **Launchpad Entry**: https://launchpad.net/distros/ubuntu/+spec/ubuntu-**cluster**
- **Created**: 2005/10/20 by FabioMassimoDiNitto
- **Contributors**: FabioMassimoDiNitto, IvanKrstic
- **Packages affected**: ocfs2-tools, redhat-**cluster**-suite, kernel

Summary

This specification defines the **cluster** support and commitments we will make in Dapper.

Rationale

Breezy was the first Ubuntu release with server CDs. Already, people are deploying Ubuntu in server environments. There is some demand for a free Linux platform for HA and HPC. This is an area where most non-commercial distributions are quite lacking, and a good way to attract a new crowd to Ubuntu, as well as to help dispell the 'desktop-only' distribution image by showing Ubuntu in a cutting-edge server light.

Use cases

- Company alpha wants to throw 250 of its Apache web application servers in a load-balanced pool, achieving 100% availability as a side benefit.
- Company beta is developing a huge oracle DB based application. They clearly want Ubuntu as the distribution for their HA **cluster** with an OCFS2 backend.
- University gamma is developing a new, extremely complex algorithm to break new science frontiers and they need a HPC **cluster**. Of course, they want Ubuntu as their development and computation platform.
- Hospital delta is deploying hundreds of thin clients with Ubuntu and LTSP, and wants complete failover for the terminal services.

Implementation

For Dapper, we will specifically support the following individual **cluster**ing components:

- the SLURM job queue/resource manager (HPC)
- the DRBD shared device solution (HA)
- the ganglia monitoring solution (HA, HPC)
- the LVS highly-available, highly-scalable network service system
- the GFS and OCFS2 filesystems

This means we will:

- update GFS and OCFS2 (already part of Breezy and already updated in Dapper).
- package SLURM for main (+ libraries).
- include drbd in the kernel package.
- promote ganglia, drbd0.7-utils from universe to main
- promote the packages that accompany LVS (keepalived, ipvsadm) from universe to main

Outstanding issues

- Configuration tools: OCFS2 and GFS provide GUIs and some configuration facilities. This does not seem to be the case for other tools. Due to the nature of **cluster**s, it is simply not possible to provide a default configuration that just works. In both HA and HPC environments, configurations are so environment-specific that it is impossible to ship a sane default.

NOTES

- The sections below are informational, and there is clearly no need to review them.

Possible test bed

A number of groups at Harvard Medical School are interested in an "out-of-the-box" **cluster** install for commodity **cluster** hardware. Sasha Wait offered to run a test site for Ubuntu-based **cluster**ing solutions. Production **cluster**s at the Medical School have 100s of nodes, and could possibly switch to Ubuntu for **cluster**ing. I am especially interested in an OpenSSI based (Single System Image) solution. Is there another specification for this?

Software Evaluated in the process (and comments)

SLURM is the job queue/resource manager from the Lawrence Livermore National Laboratory, and is widely used for resource management in high-end **cluster**s, scaling up to BlueGene/L's 65 thousand nodes with twice as many processors. Will integrate.

Oscar and gridengine are extremely big and complex and they do not come from a Debian/Ubuntu world. Probably part of the code will need porting from RedHat/Solaris. Will not integrate.

Lustre's public (GPL) version is a year behind the commercial version, and is known-broken and not available for our latest kernel. Will not integrate.

DDRAID is alpha software, considered far from production quality. Not available for our kernel. Will not integrate.

Ganglia is a widely used **cluster** monitoring tool. Will promote from main.

DRBD is a production-quality, widely used network shared storage device, providing network RAID 1 functionality. Will promote from main.

Keepalived is the de facto LVS failover and monitoring solution. ipvsadm is the LVS administration tool. Both will be promoted to main.

LVS itself is already in the mainline kernel starting with 2.6.10.

Comments

Questions that were raised before the BoF:

- How to make sure all nodes are up to date?
  - This falls under the NetworkWideUpdates spec.
- How to make sure only the user that are currently running jobs on the node have access to it?
- How to distribute /etc/passwd and/or any other user related information? Or should ldap+kerberos be used?
  - These questions fall under NetworkAuthentication.
- How to handle optimized mathlibs?
- How to handle various versions of MPI?
  - These are 'just another package'. We package them up, and throw them in the archive.

Reviewer Comments

OK, this is in good shape now, thank you very much for the effort and the review of code out there.

CategorySpec

Edgy Crack

All moved to [UbuntuEdgy**Cluster**s]

l'ultima modifica è del 2006-10-20 05:55:49, fatta da FabioMassimoDiNitto