Cluster-aware raid in Linux

Guoqing Jiang
gqjiang@suse.com
May 15, 2018

Contents

- Motivation
- Introduction
- Current status
- Comparison
- Further readings

Motivation

Previously, there are two data replication solutions for HA storage, such as DRBD and CLVM/cmirrord, but:

.For DRBD, the storage is dedicated to each node. And it can only supports two nodes with Active/Active mode.

.CLVM/cmirrord can support shared storage and more node num (limited by pacemaker/corosync), but, it has severe performance issue.

So people want a better solution to provide data replication within cluster, and support shared storage and as more nodes as possible.

Introduction

.Cluster-aware raid keeps write-intent-bitmap for each cluster node, and DLM is used for inter-communication.

.During normal I/O access, we assume the clustered filesystem ensures that only one node writes to any given block at a time.

.With each node have it's own bitmap, there would be no locking and no need to keep sync array during normal operation.

.Cluster-aware raid would only handle the bitmaps when resync and recovery happened.

Introduction

- **Node change**

- 1. node failure
- If a node left cluster or it is crashed, other nodes know about the change. So a node will copy the failed node's bitmap into its own bitmap, and this node would launch resync thread based on the change.

- 2. node join
- When a node joins a cluster, it will get an index number assigned by DLM, so the node know where it needs to store its intent to write. Also the new node doesn't write to the resyncing area.

Introduction

- **Device change**

- 1. device failure
- When a node finds there is problem with a device, it would record it as faulty in metadata, and broadcast a message for it, then all nodes acknowledge the device failure.

- 2. device add
- For adding a new device, it is necessary that all nodes "see" the new device to be added, so the master node (issues add device cmd) can’t add the new device immediately until other nodes acknowledge it.

Introduction

- **Resync**

- To avoid corruption, only one node could perform resync at a time. When a node is resyncing a given range, other nodes also need to be informed by the RESYNCING message, so other nodes will block new writes from the resyncing range, which is similar as traditional raid.

Current status

- **The first personality supported by cluster aware raid is raid1, and this feature should be matured now.**

- How about other raid levels?
- 1. raid10 has limited support from last year, there is still some work to be done.

- 2. raid5? It is not possible now, because cluster fs can’t guarantee two nodes don't write to the same stripe simultaneously.

Comparison

	Active/Active mode	Suitable for Geo	Shared Storage
DRBD	Supported (limited to two nodes)	Yes	No, storage is dedicated to each node
CLVM (cmirrord)	Supported, the node num is limited by pacemaker/corosync	No	Yes
Cluster-aware raid	Supported, the node num is limited by pacemaker/corosync	No	Yes

Further readings

- Kernel document – Documentation/md/md-cluster.txt

- LWN article - https://lwn.net/Articles/674085/

- Read code if you want to know more details :)

Question?


SUSE
We adapt. You succeed.