pacemaker overview

Friday, March 27, 2020 3:40 PM

Redhat Cluster Core Components:

1.Resource Agents

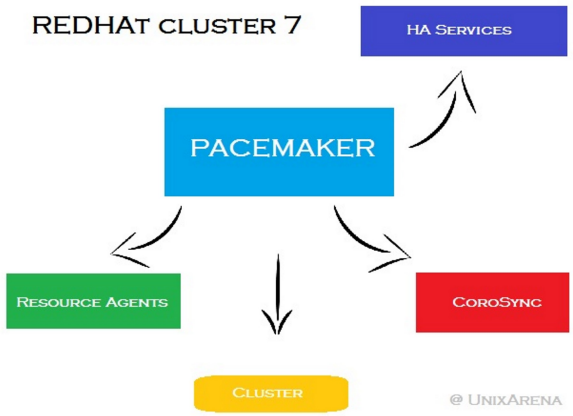
Resource agents are nothing but a scripts that start, stop and monitor them.

2.Resource Manager

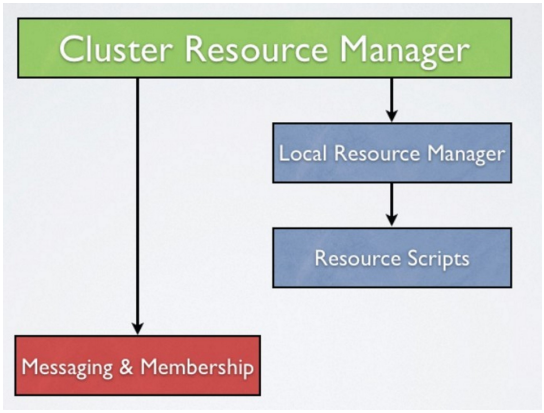
Pacemaker provides the brain that processes and reacts to events regarding the cluster. These events include nodes joining or leaving the cluster. Resource events caused by failures, maintenance and scheduled activities and other administrative actions. Pacemaker will compute the ideal state of the cluster and plot a path to achieve it after any of these events. This may include moving resources, stopping nodes and even forcing them offline with remote power switches.

3. Low-level infrastructure:

Corosync provide reliable messaging, membership and quorum information about the cluster.



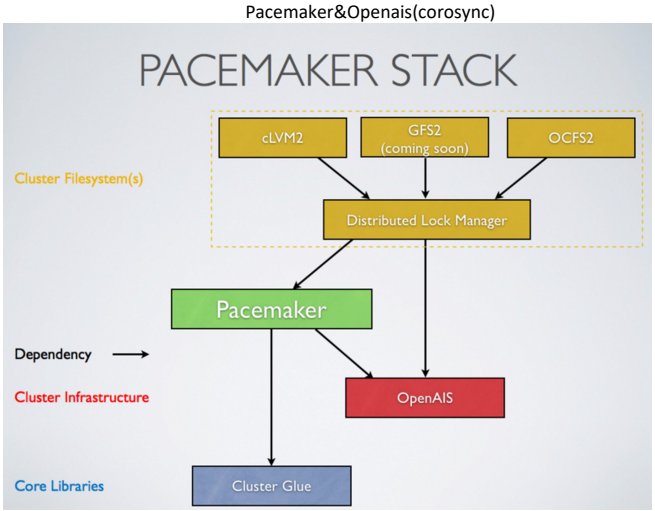
HA集群的基本组成部分：



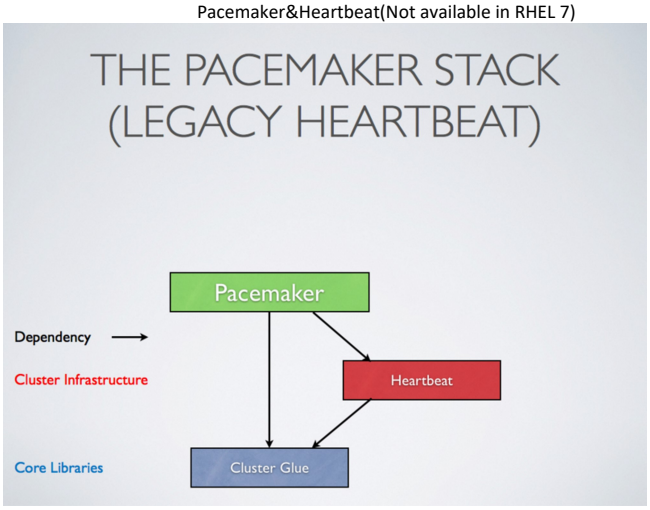
1. Messaging & Membership部分是基础核心，负责消息传递以及集群中的成员关系管理；
2. CRM部分是集群的大脑，负责对集群（节点的加入或退出）和资源（失效监测）的各种事件做出反应和决策；
3. LRM及RS是直接和服务相关的底层组件，CRM调用LRM来管理资源，而LRM调用RS来和具体的服务打交道。

支持的两种集群架构:

Pacemaker&Openais(corosync)



Pacemaker&Heartbeat(Not available in RHEL 7)



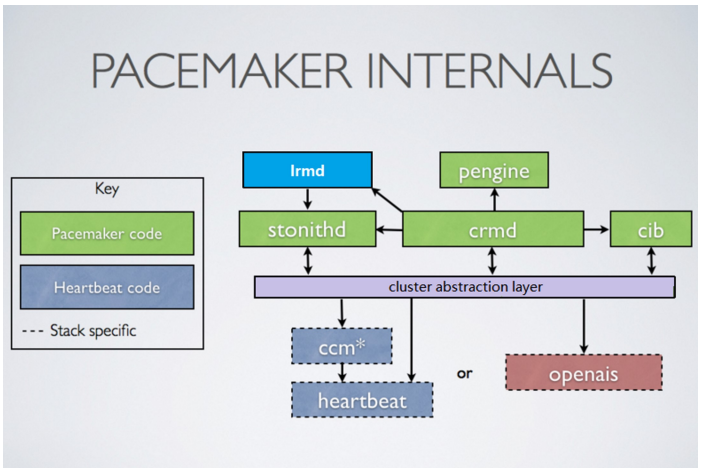
Pacemaker:

Pacemaker is responsible to provide maximum availability for your cluster services/resources by detecting and recovering from node and resource-level failures. It uses messaging and membership capabilities provided by Corosync to keep the resource available on any of the cluster nodes.

- Detection and recovery of node and service-level failures
- Storage agnostic, no requirement for shared storage
- Resource agnostic, anything that can be scripted can be clustered
- Supports *fencing* (*STONITH*) for ensuring data integrity
- Supports large (32 node) and small clusters (2 node)
- Supports both quorate and resource-driven clusters
  关于quorate and resource-driven clusters区别，可参考
  (https://www.usenix.org/legacy/publications/library/proceedings/usenix04/tech/sigs/full_papers/bottomley/bottomley_html/node7.html)
- Supports practically any redundancy configuration
- Automatically replicated configuration that can be updated from any node
- Ability to specify cluster-wide service ordering, colocation and anti-colocation
- Support for advanced service types
  - Clones: for services which need to be active on multiple nodes
  - Multi-state: for services with multiple modes (e.g. master/slave, primary/secondary)
- Unified, scriptable cluster management tools

Pacemaker's key components:

Cluster Information Base (CIB)	It uses XML format file (cib.xml) to represent the cluster configuration and current state of cluster to all the nodes. This file be kept in sync across all the nodes and used by PEngine to compute ideal state of the cluster and how it should be achieved.
Cluster Resource Management daemon (CRMd)	List of instruction will feed to the designated controller (DC).Pacemaker centralizes all cluster decision making by electing one of the *CRMd instances* to act as a **master**. If one CRMd instance fails, automatically new one will establish.
Local Resource Management daemon (LRMd)	*LRMd* is responsible to hear the instruction from *PEngine*.
Policy Engine (PEngine or PE)	*PEngine* uses the *CIB XML file* to determine the cluster state and recalculate the ideal cluster state based on the unexpected results.
Fencing daemon (STONITHd)	If any node misbehaves , it better to turned off instead of corrupting the data on shared storage. Shoot-The- Other-Node-In-The-Head (STONITHd) offers fencing mechanism in RHEL 7.

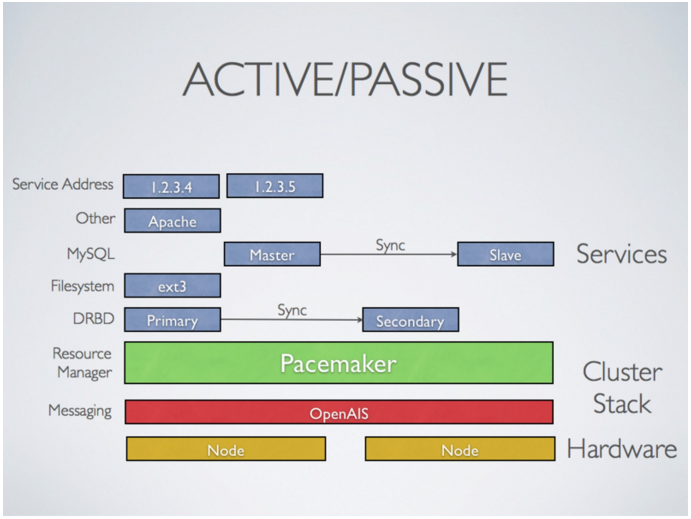


heartbeat可被corosync替换

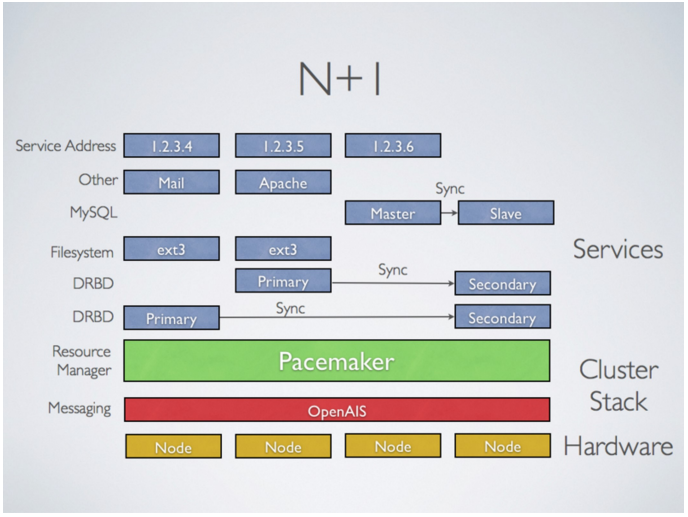
1. ccm(consensus cluster membership): 主要工作就是管理集群中各个节点的成员以及各成员之间的关系，让集群中各个节点有效的组织成一个整体，保持着稳定的连接。
2. heartbeat:心跳消息层，heartbeat模块所担当的只是一个通信工具，而CCM是通过这个通信工具来将各个成员连接到一起成为一个整体。

Types of Redhat Cluster supported with Pacemaker:

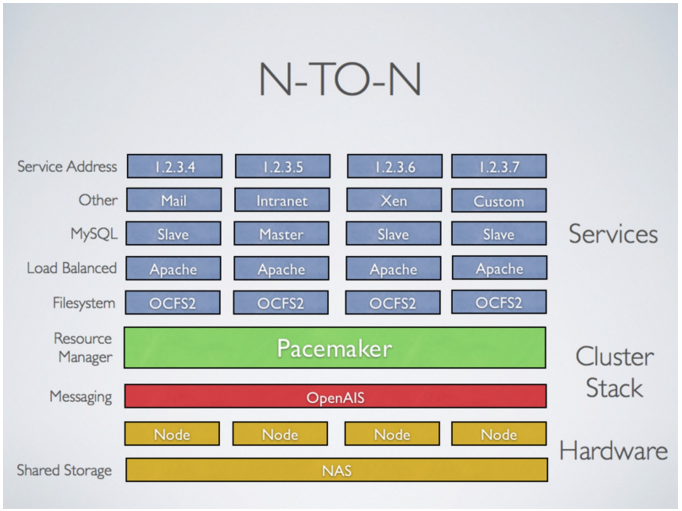
Active/Passive: 使用Pacemaker和DRBD(Distributed Replicated Block Device)的双节点主备方案作为一种经济的解决方案被很多高可用环境所采用。



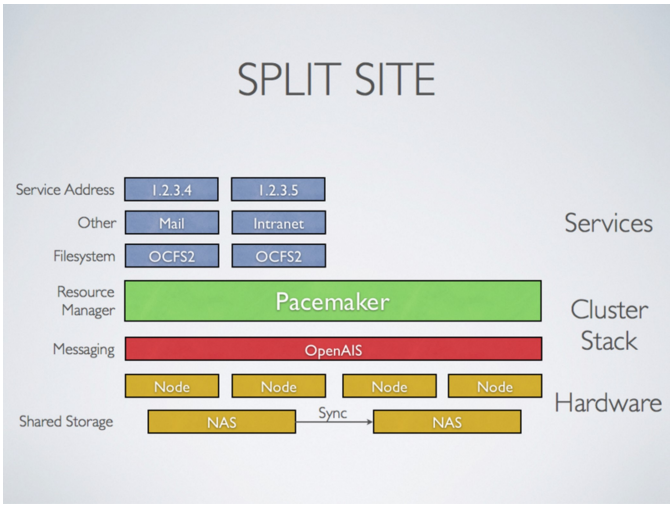
N+1: 支持多个节点，允许多个Active/Passive集群共享一个共同的备份节点， Pacemaker可以大幅降低硬件成本。



N to N: 共享存储时，每个节点都可以被用于故障切换。Pacemaker甚至可以运行服务的多个副本来展开工作量。



Split site


SPLIT SITE
Service Address
1.2.3.4
1.2.3.5
Other
Mail
Intranet
Filesystem
OCFS2
OCFS2
Services
Resource Manager
Pacemaker
Messaging
OpenAIS
Cluster Stack
Node
Node
Node
Node
Hardware
Shared Storage
NAS
Sync
NAS