第 6 章 Proxmox 集群文件系统（pmxcfs）

Proxmox 集群文件系统是一个数据库驱动的文件系统，用于保存配置文件，并利用 corosync 在集群节点间实现配置文件的实时同步。我们利用这个文件系统来管理 PVE 的配置文件。

该文件系统一方面将所有数据保存在磁盘上的一个数据库文件中，同时在内存中保存了一个拷贝。该设计引入了文件系统总容量的上限，目前该上限为 30MB，但仍然足以保存几千台虚拟机的配置信息。

该文件系统的优点如下：

- 在所有节点间透明地实时同步所有配置文件。
- 强一致性校验，避免虚拟机 ID 冲突。
- 节点失去多数票时自动进入只读状态。
- 自动更新所有节点上的 corosync 集群配置文件。
- 分布式锁机制。

6.1POSIX 兼容性

Pmxcfs 基于 FUSE 技术，其实现类似于 POSIX。但我们仅实现了必须的功能，因此 POSIX 标准中的部分功能并未实现。

- 仅支持普通文件和目录，不支持符号链接。
- 不能重命名非空目录（以便于确保虚拟机 ID 的独一性）。
- 不能修改文件权限（文件权限基于路径确定）。
- O_EXCL 创建不是原子操作（类似老的 NFS）。
- O_TRUNC 创建不是原子操作（FUSE 的限制）。

6.2 文件访问权限

所有的文件和目录都属于 root 用户和 www-data 用户组。只有 root 用户有写权限，www-data 用户组对大部分文件有读权限。以下路径的文件只有 root 有权访问。

/etc/pve/priv/

/etc/pve/nodes/${NAME}/priv/

6.3 技术

我们使用 Corosync 集群引擎实现集群通信，用 SQlite 管理数据库文件。文件系统用 FUSE 实现并运行在操作系统的用户空间。

6.4 文件系统布局

文件系统挂载点为：

/etc/pve

6.4.1 文件

corosync.conf	corosync 集群配置（Proxmox VE 4.x 之前为 cluster.conf）
storage.cfg	Proxmox VE 存储服务配置
datacenter.cfg	Proxmox VE 数据中心配置（键盘布局，代理…）
user.cfg	Proxmox VE 访问控制配置（users/groups/…）
domains.cfg	Proxmox VE 认证域
authkey.pub	票据签发系统的公钥
pve-root-ca.pem	集群 CA 的公共证书
priv/shadow.cfg	口令密文文件
priv/authkey.key	票据签发系统的私钥
priv/pve-root-ca.key	集群 CA 的私钥

nodes/<NAME>/pve-ssl.pem	Web 服务器的公开 SSL 证书（由集群 CA 签发）
nodes/<NAME>/pve-ssl.key	pve-ssl.pem 的私钥
nodes/<NAME>/pveproxy-ssl.pem	Web 服务器的公开 SSL 证书链（可由 pve-ssl.pem 覆盖）
nodes/<NAME>/pveproxy-ssl.key	pveproxy-ssl.pem 的私钥
nodes/<NAME>/qemu-server/<VMID>.conf	KVM 虚拟机的配置文件
nodes/<NAME>/lxc/<VMID>.conf	LXC 容器的配置文件
firewall/cluster.fw	集群级别的防火墙配置
firewall/<NAME>.fw	节点级别的防火墙配置
firewall/<VMID>.fw	虚拟机或容器级别的防火墙配置

6.4.2 符号链接

local	nodes/<LOCAL_HOST_NAME>
qemu-servcer	nodes/<LOCAL_HOST_NAME>/qemu-server
lxc	nodes/<LOCAL_HOST_NAME>/lxc/

6.4.3 用于调试的特殊状态文件（JSON）

.version	文件版本（用于检测文件内容变更）
.members	集群成员的信息
.vmlist	虚拟机列表
.clusterlog	集群日志（最近 50 条）
.rrd	RRD 数据（最近的条目）

6.4.4 启用/禁用调试

运行如下命令可以启用 syslog 调试信息：

echo "1" >/etc/pve/.debug

运行如下命令可以禁用 syslog 调试信息：

echo "0" >/etc/pve/.debug

6.5 文件系统恢复

如果你的 Proxmox VE 服务器出现故障，例如硬件故障，你可以将 pmxcfs 的数据库文件/var/lib/pve-cluster/config.db 复制到一台新的 Proxmox VE 服务器。在新服务器上（没有配置任何虚拟机或容器），停止 pve-cluster 服务，覆盖 config.db 文件（需要设置权限为 0600），然后修改/etc/hostname 和/etc/hosts 和故障服务器对应文件一致，最后重启新服务器并检查是否恢复正常（不要忘记虚拟机/容器镜像数据）。

6.5.1 删除集群配置

将一个节点从集群中删除之后，推荐的做法是重新安装 Proxmox VE。这样可以确保所有的集群/ssh 密钥和共享配置数据都被彻底清除。

某些情况下，你也许不希望重装而直接将节点恢复到单机模式运行，此时可以参考 5.5.1 节“隔离节点”给出的方法。

6.5.2 从故障节点恢复/迁移虚拟机

对于 nodes/<NAME>/qemu-server（虚拟机）和 nodes/<NAME>/lxc（容器）中的虚拟机配置文件，Proxmox VE 认为<NAME>节点是对应目录下虚拟机的拥有者。这样就可以使用本地锁来防止并发的虚拟机配置文件修改操作，而不是使用代价高昂的分布式集群锁。

但由此导致的一个负作用是，当虚拟机所属的节点停止运行时（例如，意外断电，发生集群隔离事件，...），由于不能获取该节点（已停机）上的本地锁，无法用正常方式将该节点上的虚拟机迁移到其他节点（即使相关虚拟机的磁盘镜像保存在共享存储上）。对于配置使用 HA 的虚拟机而言，则不存在这样的问题，因为 Proxmox VE 的 HA 组件已包含了必要的锁机制（集群锁）和看门狗功能，可以确保相关虚拟机能够从故障节点自动迁移到其他节点运行。

对于未配置使用 HA 的虚拟机而言，如果其磁盘镜像保存在共享存储上（并且未使用其他依赖于故障节点本地资源的配置），可以通过将虚拟机配置文件从/etc/pve 下故障节点对应目录手工移动到其他正常节点对应目录的方式（从而改变该虚拟机从属的节点），达到将虚拟机从故障节点手工迁移的目的。

例如，为将 ID 为 100 的虚拟机从故障节点 node1 迁移到正常节点 node2，可以使用 root 用户登录集群内任意正常节点，并运行如下命令：

mv /etc/pve/nodes/node1/qemu-server/100.conf /etc/pve/nodes/node2/

☒ 警告

使用以上方法迁移虚拟机之前，必须确保故障节点已经确实关机或者被隔离。否则 Proxmox VE 的锁机制将因为 mv 命令而被破坏，并导致不可预料的结果。

☒ 警告

以上方法无法迁移虚拟磁盘镜像保存在故障节点本地磁盘（或使用故障节点其他本地资源）的虚拟机。此时只能设法恢复故障节点重新加入集群，或利用之前的备份文件恢复虚拟机。