tags: environment

01. 初始化系统和全局变量

集群规划

- zhangjun-k8s-01：172.27.138.251
- zhangjun-k8s-02：172.27.137.229
- zhangjun-k8s-03：172.27.138.239

三台机器混合部署本文档的 etcd、master 集群和 woker 集群。

如果没有特殊说明，需要在所有节点上执行本文档的初始化操作。

设置主机名

```
hostnamectl set-hostname zhangjun-k8s-01 # 将 zhangjun-k8s-01 替换为当前主机名
```

如果 DNS 不支持主机名称解析，还需要在每台机器的 `/etc/hosts` 文件中添加主机名和 IP 的对应关系：

```
cat >> /etc/hosts <<EOF
172.27.138.251 zhangjun-k8s-01
172.27.137.229 zhangjun-k8s-02
172.27.138.239 zhangjun-k8s-03
EOF
```

退出，重新登录 root 账号，可以看到主机名生效。

添加节点信任关系

本操作只需要在 zhangjun-k8s-01 节点上进行，设置 root 账户可以无密码登录所有节点：

```
ssh-keygen -t rsa
ssh-copy-id root@zhangjun-k8s-01
ssh-copy-id root@zhangjun-k8s-02
ssh-copy-id root@zhangjun-k8s-03
```

更新 PATH 变量

```
echo 'PATH=/opt/k8s/bin:$PATH' >>/root/.bashrc
source /root/.bashrc
```

- `/opt/k8s/bin` 目录保存本文档下载安装的程序；

安装依赖包

```
yum install -y epel-release
yum install -y chrony conntrack ipvsadm ipset jq iptables curl sysstat libseccomp
wget socat git
```

- 本文档的 kube-proxy 使用 ipvs 模式，ipvsadm 为 ipvs 的管理工具；
- etcd 集群各机器需要时间同步，chrony 用于系统时间同步；

关闭防火墙

关闭防火墙，清理防火墙规则，设置默认转发策略：

```
systemctl stop firewalld
systemctl disable firewalld
iptables -F && iptables -X && iptables -F -t nat && iptables -X -t nat
iptables -P FORWARD ACCEPT
```

关闭 swap 分区

关闭 swap 分区，否则kubelet 会启动失败(可以设置 kubelet 启动参数 --fail-swap-on 为 false 关闭 swap 检查):

```
swapoff -a
sed -i '/ swap / s/^\(.*\)$/#\1/g' /etc/fstab
```

关闭 SELinux

关闭 SELinux，否则 kubelet 挂载目录时可能报错 `Permission denied`:

```
setenforce 0
sed -i 's/^SELINUX=.*/SELINUX=disabled/' /etc/selinux/config
```

优化内核参数

```
cat > kubernetes.conf <<EOF
net.bridge.bridge-nf-call-iptables=1
net.bridge.bridge-nf-call-ip6tables=1
net.ipv4.ip_forward=1
net.ipv4.tcp_tw_recycle=0
net.ipv4.neigh.default.gc_thresh1=1024
net.ipv4.neigh.default.gc_thresh1=2048
net.ipv4.neigh.default.gc_thresh1=4096
vm.swappiness=0
vm.overcommit_memory=1
vm.panic_on_oom=0
fs.inotify.max_user_instances=8192
fs.inotify.max_user_watches=1048576
fs.file-max=52706963
fs.nr_open=52706963
net.ipv6.conf.all.disable_ipv6=1
net.netfilter.nf_conntrack_max=2310720
EOF
cp kubernetes.conf  /etc/sysctl.d/kubernetes.conf
sysctl -p /etc/sysctl.d/kubernetes.conf
```

- 关闭 tcp_tw_recycle，否则与 NAT 冲突，可能导致服务不通；

设置系统时区

```
timedatectl set-timezone Asia/Shanghai
```

设置系统时钟同步

```
systemctl enable chronyd
systemctl start chronyd
```

查看同步状态：

```
timedatectl status
```

输出：

```
System clock synchronized: yes
              NTP service: active
          RTC in local TZ: no
```

- `System clock synchronized: yes`，表示时钟已同步；
- `NTP service: active`，表示开启了时钟同步服务；

```
# 将当前的 UTC 时间写入硬件时钟
timedatectl set-local-rtc 0

# 重启依赖于系统时间的服务
systemctl restart rsyslog
systemctl restart crond
```

关闭无关的服务

```
systemctl stop postfix && systemctl disable postfix
```

创建相关目录

创建目录：

```
mkdir -p /opt/k8s/{bin,work} /etc/{kubernetes,etcd}/cert
```

分发集群配置参数脚本

后续使用的环境变量都定义在文件 environment.sh 中，请根据自己的机器、网络情况修改。然后拷贝到所有节点：

```
source environment.sh # 先修改
for node_ip in ${NODE_IPS[@]}
  do
    echo ">>> ${node_ip}"
    scp environment.sh root@${node_ip}:/opt/k8s/bin/
    ssh root@${node_ip} "chmod +x /opt/k8s/bin/*"
  done
```

升级内核

CentOS 7.x 系统自带的 3.10.x 内核存在一些 Bugs，导致运行的 Docker、Kubernetes 不稳定，例如：

1. 高版本的 docker(1.13 以后) 启用了 3.10 kernel 实验支持的 kernel memory account 功能(无法关闭)，当节点压力大如频繁启动和停止容器时会导致 cgroup memory leak；
2. 网络设备引用计数泄漏，会导致类似于报错："kernel:unregister_netdevice: waiting for eth0 to become free. Usage count = 1";

解决方案如下：

1. 升级内核到 4.4.X 以上；
2. 或者，手动编译内核，disable CONFIG_MEMCG_KMEM 特性；
3. 或者，安装修复了该问题的 Docker 18.09.1 及以上的版本。但由于 kubelet 也会设置 kmem（它 vendor 了 runc），所以需要重新编译 kubelet 并指定 GOFLAGS="-tags=nokmem"；

```
  git clone --branch v1.14.1 --single-branch --depth 1
https://github.com/kubernetes/kubernetes
  cd kubernetes
  KUBE_GIT_VERSION=v1.14.1 ./build/run.sh make kubelet GOFLAGS="-tags=nokmem"
```

这里采用升级内核的解决办法：

```
rpm -Uvh http://www.elrepo.org/elrepo-release-7.0-3.el7.elrepo.noarch.rpm
# 安装完成后检查 /boot/grub2/grub.cfg 中对应内核 menuentry 中是否包含 initrd16 配置，如果没有，再安装一次！
yum --enablerepo=elrepo-kernel install -y kernel-lt
# 设置开机从新内核启动
grub2-set-default 0
```

重启机器：

```
sync
reboot
```

参考

1. 系统内核相关参数参考：https://docs.openshift.com/enterprise/3.2/admin_guide/overcommit.html
2. 3.10.x 内核 kmem bugs 相关的讨论和解决办法：
    1. https://github.com/kubernetes/kubernetes/issues/61937
    2. https://support.mesosphere.com/s/article/Critical-Issue-KMEM-MSPH-2018-0006
    3. https://pingcap.com/blog/try-to-fix-two-linux-kernel-bugs-while-testing-tidb-operator-in-k8s/