用 Kubernetes 部署超级账本 Fabric 的区块链即服务

张海宁、陈家豪
VMware 中国研发中心

本文读者需要对 Docker，Kubernetes 比较熟悉，同时对 Fabric 架构有一定了解。

1. 概述

盼望着，盼望着，超级账本 Fabric 1.0 正式来了，社区用户为之欢呼雀跃：终于等到一个企业级区块链应用平台了。然而在激动过后，回归平静之时，人们却往往发现，搭建 Fabric 平台是个相当披荆斩棘的历程。不仅要具备密码学、分布式计算、共识算法等区块链理论基础，而且要熟悉容器、Golang / Node.js 这些企业用户不常用的工程技术，这常常是很多人把区块链放弃在起跑线的原因。降低使用门槛，提高易用性，将是今后一段时间内推广企业区块链应用的重要工作。

之前我们的文章描述过安装部署多节点 Fabric 集群的步骤，旨在说明 Fabric 基本的运作原理，因此部署过程是手 (cao) 动 (gen) 的安装方式。在实际的开发测试中，需要自动化部署来提高效率，本文介绍如何利用容器平台 Kubernetes（K8s）来自动部署 Fabric，实现区块链即服务 (Blockchain as a Service, BaaS)的原型。

需要指出的是，BaaS 目前多用于开发测试，即在同一个 BaaS 平台，部署多个区块链节点，每个节点代表不同组织机构。这样显然是中心化的部署方式，所以只能用于开发测试用途。真实环境的部署需要分布在网络中多个 BaaS 协同工作才能完成，这是另外一个尚待完善的工作。

我们选择把 Fabric 部署在 K8s 上有几个原因。首先 Fabric 的组件都经过容器封装好的，很方便部署在 K8s 这类容器平台上，并借助平台实现高可用、监控管理、自动化运维等目的。其次，Fabric 是分布式系统，根据应用的具体需求，集群的各个组件数会有不同，需要灵活地配置和调整。而 K8s 是面向微服务架构的容器平台，扩展方便，能够很好地满足 Fabric 这方面的要求。还有就是 K8s 具备了多租户的能力，可在同一个平台中运行多个互相隔离的 Fabric 实例，方便开发测试，比如一个实例作为开发用，另一个实例作为测试用。

在超级账本中有个子项目叫 Cello，其目的是提供 Hyperledger 的 BaaS。据了解，目前已经支持把 Fabric 部署在 Docker 和 Swarm 上，有关 K8s 的支持还在开发中。由

于 Fabric 的设计中没有考虑到 K8s 等平台的特点，因此把 Fabric 部署在 K8s 上还需要一些变通的处理方法，后文相关部分会提到 。

2. 整体架构

2.1 基础设施

1) 网络

Kubernetes 集群由多个节点组成，为使得集群上的容器正常通信，需要创建一个 overlay 网络，并把集群上的容器都连接到这个网络上。

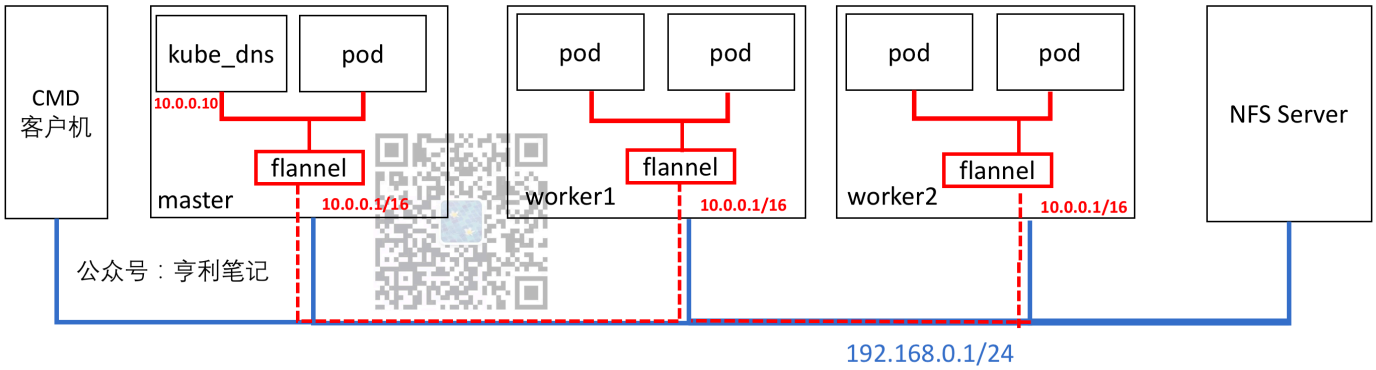

图 2- 1

如图 2-1 所示，宿主机网络由蓝色线标记，节点有 cmd 客户机, Kubernetes 的 master 和 worker，还有 NFS 服务器。其中，cmd 客户机作为操作 K8S 和 Fabric 集群的命令行客户机。NFS 服务器在各个节点间用于共享 Fabric 和 K8s 的各种配置文件，也可以用其他 K8s 支持的共享存储代替。

通过 Kubernetes 的 cluster add-ons 可以很方便地创建 overlay 网络，如 Flannel。如图 2-1 的红色线所示(为说明 Flannel 作用省去部分细节)，Kubernetes 把所有 pod 加入到 Flannel 网络中，因此 pod 中的容器可以相互通信。Flannel 网络的地址段可以在 add-on 的配置文件中指定，同时 kube_dns 的 IP 地址也可以通过配置修改，但该 IP 地址必须处于指定地址段中。如图 2-1 中 Flannel 的网络为 10.0.0.1/16，kube_dns 的 IP 地址为 10.0.0.10。

在 Fabric 设计中， chaincode 目前是以 Docker 容器的方式运行在 peer 容器所在的宿主机上，peer 容器需要调用 Docker 引擎的接口来构建和创建 chaincode 容器，调用接口是通过这个连接：

unix:///var/run/docker.sock

这种方式其实是有安全隐患的，这里不作过多讨论。正确的姿势应该是调用 chaincode 专用的运行环境，如新起一个 Docker Host，用 TCP 接口远程调用。

通过 docker.sock 创建的容器脱离在 Kubernetes 的体系之外，虽然它仍在 Flannel 网络上，但却无法获得 peer 节点的 IP 地址。这是因为创建该容器的 Docker 引擎使用宿主机默认的 DNS 解析来 peer 的域名，所以无法找到。

为了解决解析域名的问题， 需要在每个 worker 的 DOCKER_OPTS 中加入相关参数，以图 2-1 为例，kube_dns 的 IP 为 10.0.0.10，宿主机网络 DNS 的 IP 地址假设为 192.168.0.1，为使得 chaincode 的容器可以解析到 peer 节点，修改步骤如下：

编辑 /etc/default/docker，在 DOCKER_OPTS 中添加以下参数，设置 Kubernetes 使用的 DNS （很重要！）

1. ：

```
"--dns=10.0.0.10 --dns=192.168.0.1  --dns-search \
 default.svc.cluster.local --dns-search svc.cluster.local \
--dns-opt ndots:2 --dns-opt timeout:2 --dns-opt attempts:2 "
```

2. 运行以下命令重启 Docker (注: 不同的 Linux 环境中命令可能会有不同)：

```
systemctl daemon-reload
systemctl restart docker
systemctl restart docker.service
```

2) 共享存储

K8s 和 Fabric 集群需要较多的配置文件，为方便管理，可通过 NFS 服务器来统一储存这些文件，如图 2-1 所示。

cmd 客户机可通过 cryptogen 工具生成 crypto-config 目录，该用于储存 Fabric 集群中节点的配置文件，如 peer0.org1 所用到的 msp 存放在以下目录：

crypto-config/PeerOrganization/org1/peers/peer0/msp

cmd 客户机只需要把 NFS 的共享目录挂载到本地/opt/share/，再把 crypto-config 写到该目录，即可让各个节点访问到配置文件。

在 Kubernetes 中，通过 PV 和 PVC 来把 NFS 上的文件挂载到容器中，除了创建相应的 PV 和 PVC 外，还需在节点的配置文件中把正确的路径挂载进去。若 NFS 服务器的共享目录为/opt/share， 创建 PV 时可指定 peer.org1 的挂载点为：

/opt/share/crypto-config/PeerOrganization/org1/

然后创建与该 PV 相对应的 PVC，这样在节点的配置文件中就可以使用 PVC 来挂载这个目录。节点需要根据自己的 ID 在挂载点后面加上相应的路径来保证挂载的配置文件无误，如 peer0.org1 应在路径后加上 peers/peer0/msp，则其挂载目录的完整路径如下：

/opt/share/crypto-config/PeerOrganization/org1/peers/peer0/msp

2.2 组件划分

在 Kubernetes 中， Namespace 是一个很重要的概念，它用于划分不同的虚拟集群，而 Fabric 中 organization 基于证书签发机构划分，把 K8S 的 namespace 与 Fabric 的 organization 对应，既使得 organization 之间相互独立，又充分利用了 K8S 的 DNS 服务，各个 organization 可以通过域名区分。采用 namespace 分隔各个组织的组件，还可实现网络策略(network policy)来隔离不同组织，实现多租户的能力(本文未涉及)。

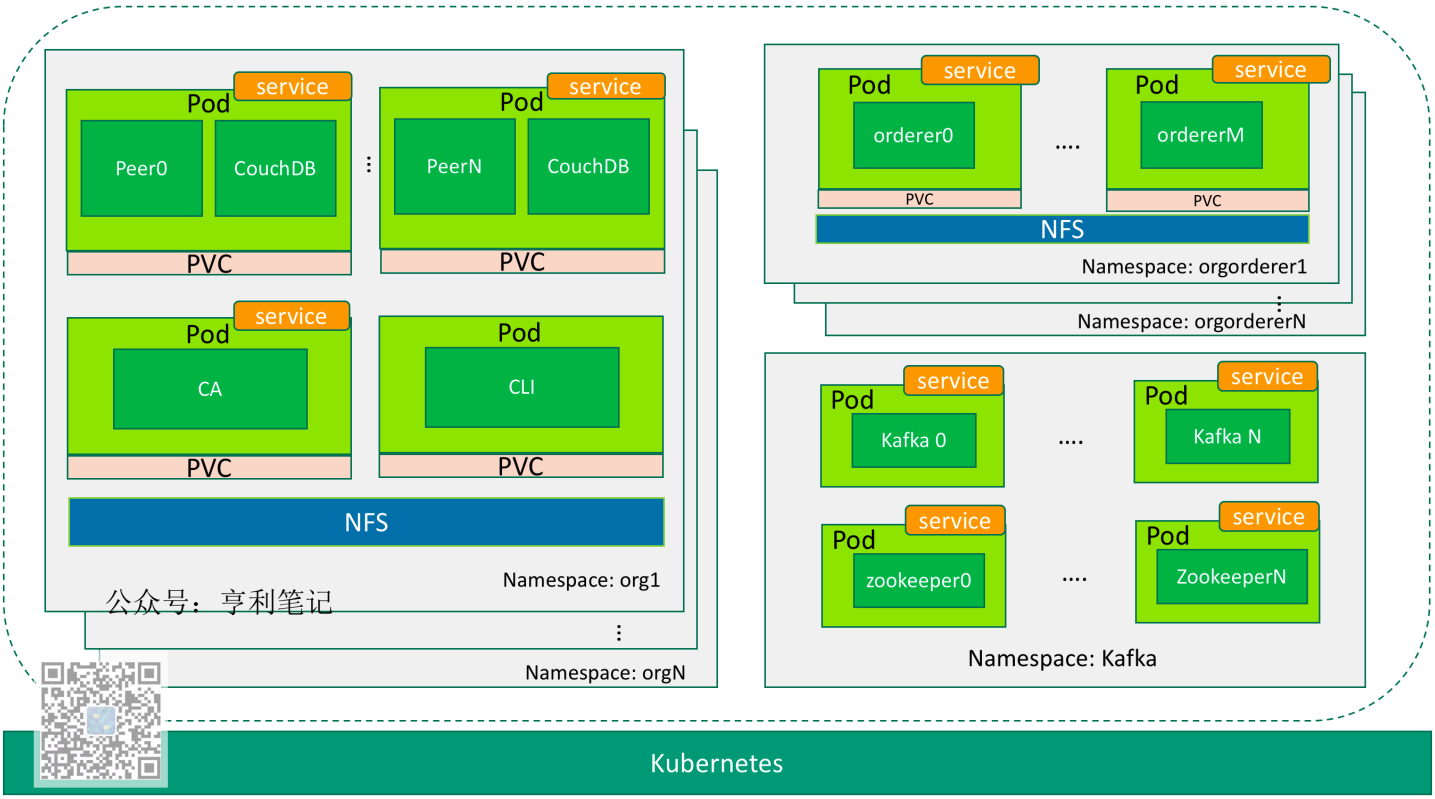

图 2- 2

如图 2-2 所示，假设 Fabric 网络中有多个 peer organization 和 orderer organization，下面阐述如何在 Kubernetes 进行划分和对应：

a. 若第 N 个 Fabric 的 peer organization 的域名为 orgN，则其在 Kubernetes 上对应的 namespace 设置为 orgN，在该 namespace 下有多个 pod，pod 的类型如下：
   1) Peer Pod：包括 Fabric peer，couchDB（可选），代表每个组织的 peer 节点。
   2) CA Server Pod： Fabric CA Server。
   3) CLI Pod：（可选）提供命令行工具的环境，用于操作本组织的节点、channel 或 chaincode。

b. 若第 N 个 Fabric 的 orderer organization 的域名为 orgordererN，则其在 Kubernetes 上对应的 namespace 为 orgordererN，该 namespace 下只有一种 Pod，它用于运行 orderer 节点。

c. Kafka namespace 与 Fabric 的 organization 并无关系，它只用来运行和管理 Zookeeper 和 Kafka 容器，实现共识算法。

2.3 Pod 之间通信

Kubernetes 中的每个 Pod 都有独立的 IP 地址，然而在各个 Pod 之间直接通过 IP:port 的方式来通信会带来很多麻烦，因此有必要给每一个 Pod 绑定一个的 service，以便用 service 名称来访问。

service 的命名方式应当遵循以下原则，彰显与其绑定的 Pod 信息：

1) service 与 pod 的 namespace 应当一致。
2) service 的 name 应与 Pod 中容器的 id 一致。

例如，Fabric 中属于 org1 的 peer0 节点，在 K8S 中用 namespace 为 org1、名字为 peer0 的 Pod 来运行，与该 Pod 绑定的 service 全称应为 peer0.org1。其中 peer0 为 service 的名称，org1 为 service 的 namespace。这样的映射关系已经和主机域名很接近了。

3. 源码的说明与使用

3.1 环境准备

假定 K8s 平台已经成功部署，并且在各个 worker 节点已经预先下载相应的 Fabric v1.0.0 镜像，如表 3.1。（预先下载镜像是由于国内网速较慢，在一台机器预先下载后，可用 Docker 命令导出并导入其他机器。）

IMAGE	TAG	ID
hyperledger/fabric-tools	x86_64-1.0.0	0403fd1c72c7
hyperledger/fabric-orderer	x86_64-1.0.0	e317ca5638ba
hyperledger/fabric-peer	x86_64-1.0.0	6830dcd7b9b5
hyperledger/fabric-ccenv	x86_64-1.0.0	7182c260a5ca
hyperledger/fabric-ca	x86_64-1.0.0	a15c59ecda5b
hyperledger/fabric-baseimage	x86_64-0.3.1	9f2e9ec7c527
hyperledger/fabric-baseos	x86_64-0.3.1	4b0cab202084

表 3- 1

下载本文涉及代码的目录结构及其作用如下：

- Fabric-on-k8s
  - README.md
  - setupCluster
    - generateALL.sh // 负责生成 K8S 部署文件
    - transform // 用于启动部署文件
    - templates // 存放模板
    - cluster-config.yaml // 用于配置 Fabric 集群
    - configtx.yaml // 用于配置 channel

3.2 配置文件说明

在规划 Fabric 集群部署时，要按实际需求，编辑以下两个 Fabric 集群的定义文件：

a. cluster-config.yaml

cryptogen 工具根据 cluster-config.yaml 来生成 Fabric 成员的证书，一个简单的例子如下:

```
OrdererOrgs:
 - Name: Orderer
   Domain: orgorderer1
   Template:
     Count: 1

PeerOrgs:
 - Name: Org1
   Domain: org1
   Template:
     Count: 2

 - Name: Org2
   Domain: org2
   Template:
     Count: 2
```

其中 OrdererOrgs 和 PeerOrgs 关键字区分 organization 的类型，两种组织的内部结构如下：

1) OrdererOrgs 中定义了一个名字为 Orderer，域名为 orgorderer1 的 org，并且它指定 template 中 count 的数值为 1，则在该 org 下只有一个 orderer，其 id 为 orderer0。

2) PeerOrgs 中定义了两个 org，分别为 Org1 和 Org2，对应的域名为 org1、org2，与 orderer 类似，每个 org 生成了两个 peer，虽然 org1 中 peer0 和 org2 中 peer0 的 ID 重复，但是他不属于同一个 org，通过域名很容易就能区分出它们。

需要注意的是，由于 K8S 中的 namespace 不支持 '.' 和大写字母，因此各个组织的域名不能包含这些字符。

更多关于 cluster-config.yaml 的配置方式，请读者自行参考 Fabric 源码中的关于 cryptogen 的描述(fabric/common/tools/cryptogen/main.go.)

以上定义的 cluster-config.yaml，cryptogen 工具会生成 crypto-config 目录，该目录的结构如下：

```
crypto-config
  --- ordererOrganizations
     --- orgorderer1
        --- msp
        ---  ca
         --- tlsca
        --- users
        --- orderers
           --- orderer0.orgorderer1
              --- msp
              --- tls

  --- peerOrganizations
     --- org1
        --- msp
        ---  ca
         --- tlsca
        --- users
        --- peers
            --- peer0.org1
               --- msp
               --- tls
            --- peer1.org1
               --- msp
               --- tls
     --- org2
        --- msp
        ---  ca
        --- tlsca
```

```
--- users
--- peers
    --- peer0.org2
        --- msp
        --- tls
    --- peer1.org2
        --- msp
        --- tls
```

可以看出，每个 org 都包含了 msp、ca、tlsca 和 users 目录，然后根据 org 类型的不同，还分别有 peers 和 orderers 目录，里面存放着 org 中每个成员的 msp 和 tls 文件。

b. configtx.yaml

configtxgen 工具根据该文件生成 Orderer 初始化的时候要使用的 genesis.block，获知 organization 的各种信息，因此，用户要根据 cluster-config.yaml 中关于 organization 的定义来修改 configtx.yaml 以生成合适的 genesis.block。例如，用户在 cluster-config.yaml 中增加了一个 Org3，并且要创建一个包含 Org1，Org2，Org3 的集群，则应该通过以下两步修改 configtx.yaml：

1. 在 profile 中增加 Org3，如图 3-1。

```
Profiles:

    TwoOrgsOrdererGenesis:
        Orderer:
            <<: *OrdererDefaults
            Organizations:
                - *OrdererOrg
        Consortiums:
            SampleConsortium:
                Organizations:
                    - *Org1
                    - *Org2
                    - *Org3
    TwoOrgsChannel:
        Consortium: SampleConsortium
        Application:
            <<: *ApplicationDefaults
            Organizations:
                - *Org1
                - *Org2
                - *Org3
```

图 3- 1

2. 在 Organization 中增加 Org3 的 MSPDir，如图

```
- &Org3
    # DefaultOrg defines the organization which is used in the sampleconfig
    # of the fabric.git development environment
    Name: Org3MSP

    # ID to load the MSP definition as
    ID: Org3MSP

    MSPDir: crypto-config/peerOrganizations/org3/msp

    AnchorPeers:
        # AnchorPeers defines the location of peers which can be used
        # for cross org gossip communication.  Note, this value is only
        # encoded in the genesis block in the Application section context
        - Host: peer0.org3
          Port: 7051
```


图 3- 2

注意的是每个 organization 中的 MSPDir 的值必须是这种形式：crypto-config/{OrgType}/{OrgName}/mps

3.3 模板文件

在 Kubernetes 中部署 Fabric 时，需要为每个节点编写相应的配置文件。由于节点数可能很多，这是既复杂又易错的重复劳动。为提高效率，可通过模板自动生成配置文件。本文使用了 5 个模板文件，可用脚本替换其中的变量，均在笔者给出示例代码中的 templates 目录中，这些模板的作用如下：

a. fabric_1_0_tmeplate_namespace.yaml

定义 Fabric 集群在 K8s 中的 namespace，它对应着 organization 的域名。为了在多节点共享证书等文件，使用了 NFS 服务器作为存储。在 K8s 中通过相应的 PV 和 PVC，namespace 下的 Pod 可以通过 PVC 来获取与之相应的文件。

b. fabric_1_0_template_cli.yaml

CLI pod 模板，每个 organization 中都配备了一个 CLI pod，目的是提供命令行界面，可统一管理组织内的所有 peer，其中包括 channel 的创建，chaincode 的安装等。CLI Pod 的 CORE_PEER_ADDRESS 环境变量默认值为 org 中的第一个 peer，可以通过修改该环境变量来连接不同的 peer。

yaml 文件中的 command 是为了防止 CLI Pod 自动退出，CLI 的默认工作目录为 /opt/gopath/src/github.com/hyperledger/fabric/peer。由于该目录下的 channel-artifacts 挂载了 NFS 上/opt/share/channel-artifacts，因此把创建 channel 时返回的 xxx.block 文件放在该目录下供所有 CLI Pod 共享。

c. fabric_1_0_template_ca.yaml

Fabric 的 CA 服务的 Pod 定义模板，用于 organization 中的证书管理，其 yaml 文件除了定义 deployment 外，还定义了 service。service 通过 selector 与 deployment 绑定，其中 deployment 中的 label 是 selector 与其绑定的根据。

d. fabcric_1_0_template_orderer.yaml

Orderer 的 pod 定义模板，需要注意的是，cryptogen 并不会生成 genesis.block，然而缺少该文件时，orderer 会启动失败，因此在启动 orderer 之前需要预先生成 genesis.block，并将其放在相应的 org 目录下。

e. fabric_1_0_template_peer.yaml

每个 peer pod 的定义模板。在该 yaml 中分别定义了 peer 和 couchDB 两个 container。在实例化 chaincode (cc) 时，peer 需要连接 Docker 引擎来创建 cc 容器，因此要把 worker 宿主机的 var/run/docker.sock 映射到 peer 容器内部（参见图 2.1）。

https://github.com/hainingzhang/articles

VMware 公司招聘区块链实习生和外包开发工程师

VMware 公司为超级账本 Hyperledger 项目创始成员，中国研发中心现在招募区块链方向实习生和外包开发工程师，地点：北京知春里。

外包软件开发工程师：1-5 年软件开发经验，熟悉 Java 或 Go 开发语言，熟悉分布式系统、Docker，了解区块链技术优先。

实习生：要求在读研究生，计算机相关专业，懂 Java 或 Go 开发语言，能够实习 3 个月以上，熟悉区块链技术优先。欢迎自荐或推荐。

有兴趣者发简历到: BaaS@vmware.com