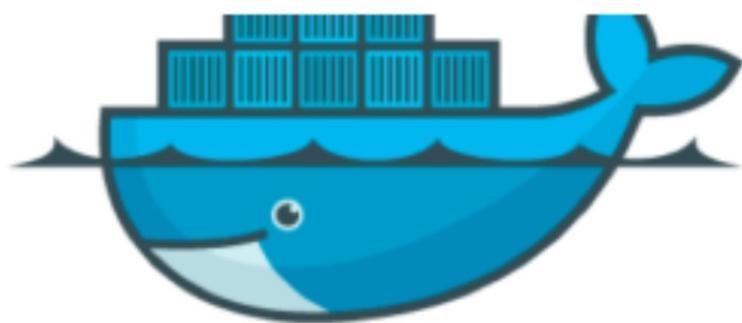


Docker

——从入门到实践



目錄

1. 前言
2. Docker 简介
 - i. 什么是 Docker
 - ii. 为什么要用 Docker
3. 基本概念
 - i. 镜像
 - ii. 容器
 - iii. 仓库
4. 安装
 - i. Ubuntu
 - ii. CentOS
5. 镜像
 - i. 获取镜像
 - ii. 列出
 - iii. 创建
 - iv. 存出和载入
 - v. 移除
 - vi. 实现原理
6. 容器
 - i. 启动
 - ii. 守护态运行
 - iii. 终止
 - iv. 进入容器
 - v. 导出和导入
 - vi. 删除
7. 仓库
 - i. Docker Hub
 - ii. 私有仓库
 - iii. 配置文件
8. 数据管理
 - i. 数据卷
 - ii. 数据卷容器
 - iii. 备份、恢复、迁移数据卷
9. 使用网络
 - i. 外部访问容器
 - ii. 容器互联
10. 高级网络配置
 - i. 快速配置指南
 - ii. 配置 DNS
 - iii. 容器访问控制
 - iv. 端口映射实现
 - v. 配置 docker0 网桥

- vi. 自定义网桥
- vii. 工具和示例
- viii. 编辑网络配置文件
- ix. 实例：创建一个点到点连接
- 11. 实战案例
 - i. 使用 Supervisor 来管理进程
 - ii. 创建 tomcat/weblogic 集群
 - iii. 多台物理主机之间的容器互联
 - iv. 标准化开发测试和生产环境
- 12. 安全
 - i. 内核名字空间
 - ii. 控制组
 - iii. 服务端防护
 - iv. 内核能力机制
 - v. 其它安全特性
 - vi. 总结
- 13. Dockerfile
 - i. 基本结构
 - ii. 指令
 - iii. 创建镜像
- 14. 底层实现
 - i. 基本架构
 - ii. 名字空间
 - iii. 控制组
 - iv. 联合文件系统
 - v. 容器格式
 - vi. 网络
- 15. Docker Compose 项目
 - i. 简介
 - ii. 安装
 - iii. 使用
 - iv. 命令说明
 - v. YAML 模板文件
- 16. Docker Machine 项目
 - i. 简介
 - ii. 安装
 - iii. 使用
- 17. Docker Swarm 项目
 - i. 简介
 - ii. 安装
 - iii. 使用
 - iv. 调度器
 - v. 过滤器
- 18. Etcd 项目
 - i. 简介
 - ii. 安装

- iii. [使用 etcdctl](#)
- 19. [Fig 项目](#)
 - i. [简介](#)
 - ii. [安装](#)
 - iii. [命令参考](#)
 - iv. [fig.yml参考](#)
 - v. [环境变量参考](#)
 - vi. [实战 Django](#)
 - vii. [实战 Rails](#)
 - viii. [实战 wordpress](#)
- 20. [CoreOS 项目](#)
- 21. [Kubernetes 项目](#)
 - i. [简介](#)
 - ii. [快速上手](#)
 - iii. [基本概念](#)
 - iv. [kubectl 使用](#)
 - v. [架构设计](#)
- 22. [Mesos 项目](#)
 - i. [简介](#)
 - ii. [安装与使用](#)
 - iii. [原理与架构](#)
 - iv. [配置项解析](#)
 - v. [常见框架](#)
- 23. [附录一：命令查询](#)
- 24. [附录二：常见仓库介绍](#)
 - i. [Ubuntu](#)
 - ii. [CentOS](#)
 - iii. [MySQL](#)
 - iv. [MongoDB](#)
 - v. [Redis](#)
 - vi. [Nginx](#)
 - vii. [WordPress](#)
 - viii. [Node.js](#)
- 25. [附录三：有用的资源](#)

Docker —— 从入门到实践

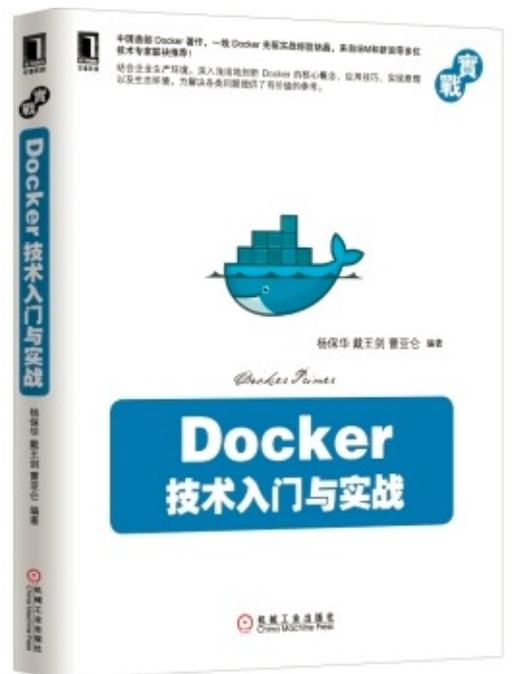
v0.5.1

Docker 是个伟大的项目，它彻底释放了虚拟化的威力，极大降低了云计算资源供应的成本，同时让应用的部署、测试和分发都变得前所未有的高效和轻松！

本书既适用于具备基础 Linux 知识的 Docker 初学者，也希望可供理解原理和实现的高级用户参考。同时，书中给出的实践案例，可供在进行实际部署时借鉴。前六章为基础内容，供用户理解 Docker 的基本概念和操作；7~9 章介绍一些高级操作；第 10 章给出典型的应用场景和实践案例；11~13 章介绍关于 Docker 实现的相关细节技术。后续章节则分别介绍一些相关的热门开源项目。

在线阅读：[GitBook](#) 或 [DockerPool](#)。

欢迎关注 DockerPool 社区微博 @dockerpool，或加入 DockerPool QQ 群（419042067），分享 Docker 资源，交流 Docker 技术。



《Docker 技术入门与实战》一书已经正式出版，包含大量第一手实战案例和更为深入的技术剖析，欢迎大家阅读使用。

- [China-Pub](#)
- [京东图书](#)
- [当当图书](#)
- [亚马逊图书](#)

主要版本历史

- 0.6: 2015-07-?
 - 补充 Machine 项目

- 0.5: 2015-06-29
 - 添加 Compose 项目
 - 添加 Machine 项目
 - 添加 Swarm 项目
 - 完善 Kubernetes 项目内容
 - 添加 Mesos 项目内容
- 0.4: 2015-05-08
 - 添加 Etcd 项目
 - 添加 Fig 项目
 - 添加 CoreOS 项目
 - 添加 Kubernetes 项目
- 0.3: 2014-11-25
 - 完成仓库章节；
 - 重写安全章节；
 - 修正底层实现章节的架构、名字空间、控制组、文件系统、容器格式等内容；
 - 添加对常见仓库和镜像的介绍；
 - 添加 Dockerfile 的介绍；
 - 重新校订中英文混排格式。
 - 修订文字表达。
 - 发布繁体版本分支：zh-Hant。
- 0.2: 2014-09-18
 - 对照官方文档重写介绍、基本概念、安装、镜像、容器、仓库、数据管理、网络等章节；
 - 添加底层实现章节；
 - 添加命令查询和资源链接章节；
 - 其它修正。
- 0.1: 2014-09-05
 - 添加基本内容；
 - 修正错别字和表达不通顺的地方。

Docker 自身仍在快速发展中，生态环境也在蓬勃成长。源码开源托管在 Github 上，欢迎参与维护：https://github.com/yeasy/docker_practice。贡献者 [名单](#)。

参加步骤

- 在 GitHub 上 `fork` 到自己的仓库，如 `docker_user/docker_practice`，然后 `clone` 到本地，并设置用户信息。

```
$ git clone git@github.com:docker_user/docker_practice.git
$ cd docker_practice
$ git config user.name "yourname"
$ git config user.email "your_email"
```

- 修改代码后提交，并推送到自己的仓库。

```
$ #do some change on the content
$ git commit -am "Fix issue #1: change helo to hello"
$ git push
```

- 在 GitHub 网站上提交 pull request。
- 定期使用项目仓库内容更新自己仓库内容。

```
$ git remote add upstream https://github.com/yeasy/docker_practice
$ git fetch upstream
$ git checkout master
$ git rebase upstream/master
$ git push -f origin master
```

简介

本章将带领你进入 Docker 的世界。

什么是 Docker？

用它会带来什么样的好处？

好吧，让我们带着问题开始这神奇之旅。

什么是 Docker

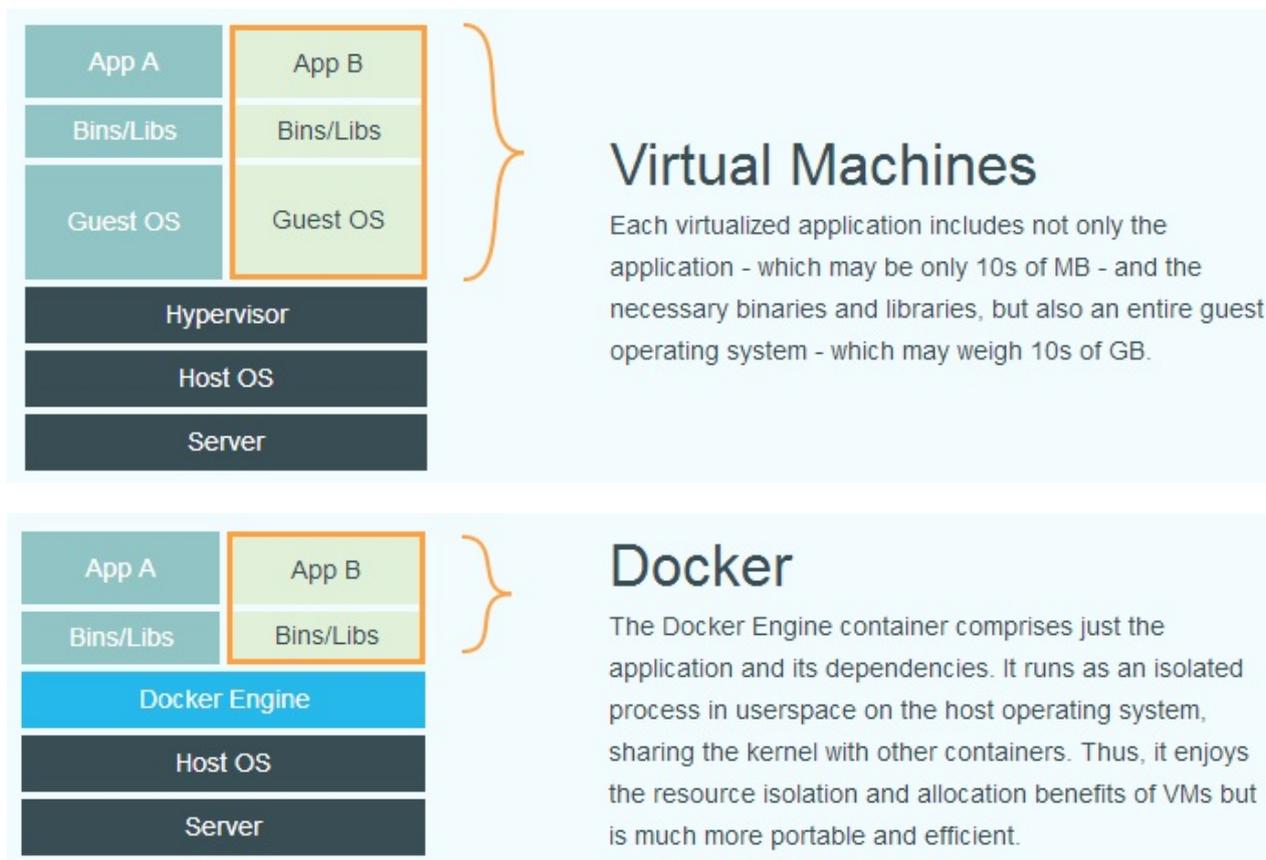
Docker 是一个开源项目，诞生于 2013 年初，最初是 dotCloud 公司内部的一个业余项目。它基于 Google 公司推出的 Go 语言实现。项目后来加入了 Linux 基金会，遵从了 Apache 2.0 协议，项目代码在 [GitHub](#) 上进行维护。

Docker 自开源后受到广泛的关注和讨论，以至于 dotCloud 公司后来都改名为 Docker Inc。Redhat 已经在其 RHEL6.5 中集中支持 Docker；Google 也在其 PaaS 产品中广泛应用。

Docker 项目的目标是实现轻量级的操作系统虚拟化解决方案。Docker 的基础是 Linux 容器（LXC）等技术。

在 LXC 的基础上 Docker 进行了进一步的封装，让用户不需要去关心容器的管理，使得操作更为简便。用户操作 Docker 的容器就像操作一个快速轻量级的虚拟机一样简单。

下面的图片比较了 Docker 和传统虚拟化方式的不同之处，可见容器是在操作系统层面上实现虚拟化，直接复用本地主机的操作系统，而传统方式则是在硬件层面实现。



为什么要使用 Docker ?

作为一种新兴的虚拟化方式，Docker 跟传统的虚拟化方式相比具有众多的优势。

首先，Docker 容器的启动可以在秒级实现，这相比传统的虚拟机方式要快得多。其次，Docker 对系统资源的利用率很高，一台主机上可以同时运行数千个 Docker 容器。

容器除了运行其中应用外，基本不消耗额外的系统资源，使得应用的性能很高，同时系统的开销尽量小。传统虚拟机方式运行 10 个不同的应用就要起 10 个虚拟机，而 Docker 只需要启动 10 个隔离的应用即可。

具体说来，Docker 在如下几个方面具有较大的优势。

更快速的交付和部署

对开发和运维（devop）人员来说，最希望的就是一次创建或配置，可以在任意地方正常运行。

开发者可以使用一个标准的镜像来构建一套开发容器，开发完成之后，运维人员可以直接使用这个容器来部署代码。Docker 可以快速创建容器，快速迭代应用程序，并让整个过程全程可见，使团队中的其他成员更容易理解应用程序是如何创建和工作的。Docker 容器很轻很快！容器的启动时间是秒级的，大量地节约开发、测试、部署的时间。

更高效的虚拟化

Docker 容器的运行不需要额外的 hypervisor 支持，它是内核级的虚拟化，因此可以实现更高的性能和效率。

更轻松的迁移和扩展

Docker 容器几乎可以在任意的平台上运行，包括物理机、虚拟机、公有云、私有云、个人电脑、服务器等。这种兼容性可以让用户把一个应用程序从一个平台直接迁移到另外一个。

更简单的管理

使用 Docker，只需要小小的修改，就可以替代以往大量的更新工作。所有的修改都以增量的方式被分发和更新，从而实现自动化并且高效的管理。

对比传统虚拟机总结

特性	容器	虚拟机
启动	秒级	分钟级
硬盘使用	一般为 MB	一般为 GB
性能	接近原生	弱于
系统支持量	单机支持上千个容器	一般几十个

基本概念

Docker 包括三个基本概念

- 镜像 (Image)
- 容器 (Container)
- 仓库 (Repository)

理解了这三个概念，就理解了 Docker 的整个生命周期。

Docker 镜像

Docker 镜像就是一个只读的模板。

例如：一个镜像可以包含一个完整的 ubuntu 操作系统环境，里面仅安装了 Apache 或用户需要的其它应用程序。

镜像可以用来创建 Docker 容器。

Docker 提供了一个很简单的机制来创建镜像或者更新现有的镜像，用户甚至可以直接从其他人那里下载一个已经做好的镜像来直接使用。

Docker 容器

Docker 利用容器来运行应用。

容器是从镜像创建的运行实例。它可以被启动、开始、停止、删除。每个容器都是相互隔离的、保证安全的平台。

可以把容器看做是一个简易版的 Linux 环境（包括root用户权限、进程空间、用户空间和网络空间等）和运行在其中的应用程序。

*注：镜像是只读的，容器在启动的时候创建一层可写层作为最上层。

Docker 仓库

仓库是集中存放镜像文件的场所。有时候会把仓库和仓库注册服务器（Registry）混为一谈，并不严格区分。实际上，仓库注册服务器上往往存放着多个仓库，每个仓库中又包含了多个镜像，每个镜像有不同的标签（tag）。

仓库分为公开仓库（Public）和私有仓库（Private）两种形式。

最大的公开仓库是 [Docker Hub](#)，存放了数量庞大的镜像供用户下载。国内的公开仓库包括 [Docker Pool](#) 等，可以提供大陆用户更稳定快速的访问。

当然，用户也可以在本地网络内创建一个私有仓库。

当用户创建了自己的镜像之后就可以使用 `push` 命令将它上传到公有或者私有仓库，这样下次在另外一台机器上使用这个镜像时候，只需要从仓库上 `pull` 下来就可以了。

*注：Docker 仓库的概念跟 [Git](#) 类似，注册服务器可以理解为 [GitHub](#) 这样的托管服务。

安装

官方网站上有各种环境下的 [安装指南](#)，这里主要介绍下Ubuntu和CentOS系列的安装。

Ubuntu 系列安装 Docker

通过系统自带包安装

Ubuntu 14.04 版本系统中已经自带了 Docker 包，可以直接安装。

```
$ sudo apt-get update
$ sudo apt-get install -y docker.io
$ sudo ln -sf /usr/bin/docker.io /usr/local/bin/docker
$ sudo sed -i '$acomplete -F _docker docker' /etc/bash_completion.d/docker.io
```

如果使用操作系统自带包安装 Docker，目前安装的版本是比较旧的 0.9.1。要安装更新的版本，可以通过使用 Docker 源的方式。

通过 Docker 源安装最新版本

要安装最新的 Docker 版本，首先需要安装 apt-transport-https 支持，之后通过添加源来安装。

```
$ sudo apt-get install apt-transport-https
$ sudo apt-key adv --keyserver hkp://keyserver.ubuntu.com:80 --recv-keys 36A1D7869245C895
$ sudo bash -c "echo deb https://get.docker.io/ubuntu docker main > /etc/apt/sources.list"
$ sudo apt-get update
$ sudo apt-get install lxc-docker
```

14.04 之前版本

如果是较低版本的 Ubuntu 系统，需要先更新内核。

```
$ sudo apt-get update
$ sudo apt-get install linux-image-generic-lts-raring linux-headers-generic-lts-raring
$ sudo reboot
```

然后重复上面的步骤即可。

安装之后启动 Docker 服务。

```
$ sudo service docker start
```

CentOS 系列安装 Docker

Docker 支持 CentOS6 及以后的版本。

CentOS6

对于 CentOS6，可以使用 [EPEL](#) 库安装 Docker，命令如下

```
$ sudo yum install http://mirrors.yun-idc.com/epel/6/i386/epel-release-6-8.noarch.rpm
$ sudo yum install docker-io
```

CentOS7

CentOS7 系统 `CentOS-Extras` 库中已带 Docker，可以直接安装：

```
$ sudo yum install docker
```

安装之后启动 Docker 服务，并让它随系统启动自动加载。

```
$ sudo service docker start
$ sudo chkconfig docker on
```

Docker 镜像

在之前的介绍中，我们知道镜像是 Docker 的三大组件之一。

Docker 运行容器前需要本地存在对应的镜像，如果镜像不存在本地，Docker 会从镜像仓库下载（默认是 Docker Hub 公共注册服务器中的仓库）。

本章将介绍更多关于镜像的内容，包括：

- 从仓库获取镜像；
- 管理本地主机上的镜像；
- 介绍镜像实现的基本原理。

获取镜像

可以使用 `docker pull` 命令来从仓库获取所需要的镜像。

下面的例子将从 Docker Hub 仓库下载一个 Ubuntu 12.04 操作系统的镜像。

```
$ sudo docker pull ubuntu:12.04
Pulling repository ubuntu
ab8e2728644c: Pulling dependent layers
511136ea3c5a: Download complete
5f0ffaa9455e: Download complete
a300658979be: Download complete
904483ae0c30: Download complete
ffdaafd1ca50: Download complete
d047ae21eeaf: Download complete
```

下载过程中，会输出获取镜像的每一层信息。

该命令实际上相当于 `$ sudo docker pull registry.hub.docker.com/ubuntu:12.04` 命令，即从注册服务器 `registry.hub.docker.com` 中的 `ubuntu` 仓库来下载标记为 `12.04` 的镜像。

有时候官方仓库注册服务器下载较慢，可以从其他仓库下载。从其它仓库下载时需要指定完整的仓库注册服务器地址。例如

```
$ sudo docker pull dl.dockerpool.com:5000/ubuntu:12.04
Pulling dl.dockerpool.com:5000/ubuntu
ab8e2728644c: Pulling dependent layers
511136ea3c5a: Download complete
5f0ffaa9455e: Download complete
a300658979be: Download complete
904483ae0c30: Download complete
ffdaafd1ca50: Download complete
d047ae21eeaf: Download complete
```

完成后，即可随时使用该镜像了，例如创建一个容器，让其中运行 `bash` 应用。

```
$ sudo docker run -t -i ubuntu:12.04 /bin/bash
root@fe7fc4bd8fc9:/#
```

列出本地镜像

使用 `docker images` 显示本地已有的镜像。

```
$ sudo docker images
REPOSITORY      TAG          IMAGE ID      CREATED       VIRTUAL SIZE
ubuntu          12.04       74fe38d11401 4 weeks ago   209.6 MB
ubuntu          precise     74fe38d11401 4 weeks ago   209.6 MB
ubuntu          14.04       99ec81b80c55 4 weeks ago   266 MB
ubuntu          latest      99ec81b80c55 4 weeks ago   266 MB
ubuntu          trusty      99ec81b80c55 4 weeks ago   266 MB
...
```

在列出信息中，可以看到几个字段信息

- 来自于哪个仓库，比如 `ubuntu`
- 镜像的标记，比如 `14.04`
- 它的 `ID` 号（唯一）
- 创建时间
- 镜像大小

其中镜像的 `ID` 唯一标识了镜像，注意到 `ubuntu:14.04` 和 `ubuntu:trusty` 具有相同的镜像 `ID`，说明它们实际上是同一镜像。

`TAG` 信息用来标记来自同一个仓库的不同镜像。例如 `ubuntu` 仓库中有多个镜像，通过 `TAG` 信息来区分发行版本，例如 `10.04`、`12.04`、`12.10`、`13.04`、`14.04` 等。例如下面的命令指定使用镜像 `ubuntu:14.04` 来启动一个容器。

```
$ sudo docker run -t -i ubuntu:14.04 /bin/bash
```

如果不指定具体的标记，则默认使用 `latest` 标记信息。

创建镜像

创建镜像有很多方法，用户可以从 Docker Hub 获取已有镜像并更新，也可以利用本地文件系统创建一个。

修改已有镜像

先使用下载的镜像启动容器。

```
$ sudo docker run -t -i training/sinatra /bin/bash
root@0b2616b0e5a8:/#
```

注意：记住容器的 ID，稍后还会用到。

在容器中添加 json 和 gem 两个应用。

```
root@0b2616b0e5a8:/# gem install json
```

当结束后，我们使用 exit 来退出，现在我们的容器已经被我们改变了，使用 `docker commit` 命令来提交更新后的副本。

```
$ sudo docker commit -m "Added json gem" -a "Docker Newbee" 0b2616b0e5a8 ouruser/sinatra:4f177bd27a9ff0f6dc2a830403925b5360bfe0b93d476f7fc3231110e7f71b1c
```

其中，`-m` 来指定提交的说明信息，跟我们使用的版本控制工具一样；`-a` 可以指定更新的用户信息；之后是用来创建镜像的容器的 ID；最后指定目标镜像的仓库名和 tag 信息。创建成功后会返回这个镜像的 ID 信息。

使用 `docker images` 来查看新创建的镜像。

```
$ sudo docker images
REPOSITORY          TAG         IMAGE ID      CREATED       VIRTUAL SIZE
training/sinatra    latest     5bc342fa0b91  10 hours ago  446.7 MB
ouruser/sinatra     v2        3c59e02ddd1a  10 hours ago  446.7 MB
ouruser/sinatra     latest     5db5f8471261  10 hours ago  446.7 MB
```

之后，可以使用新的镜像来启动容器

```
$ sudo docker run -t -i ouruser/sinatra:v2 /bin/bash
root@78e82f680994:/#
```

利用 Dockerfile 来创建镜像

使用 `docker commit` 来扩展一个镜像比较简单，但是不方便在一个团队中分享。我们可以使用 `docker build` 来创建一个新的镜像。为此，首先需要创建一个 Dockerfile，包含一些如何创建镜像的指令。

新建一个目录和一个 Dockerfile

```
$ mkdir sinatra
$ cd sinatra
$ touch Dockerfile
```

Dockerfile 中每一条指令都创建镜像的一层，例如：

```
# This is a comment
FROM ubuntu:14.04
MAINTAINER Docker Newbee <newbee@docker.com>
RUN apt-get -qq update
RUN apt-get -qqy install ruby ruby-dev
RUN gem install sinatra
```

Dockerfile 基本的语法是

- 使用 `#` 来注释
- `FROM` 指令告诉 Docker 使用哪个镜像作为基础
- 接着是维护者的信息
- `RUN` 开头的指令会在创建中运行，比如安装一个软件包，在这里使用 `apt-get` 来安装了一些软件

编写完成 Dockerfile 后可以使用 `docker build` 来生成镜像。

```
$ sudo docker build -t="ouruser/sinatra:v2" .
Uploading context 2.56 kB
Uploading context
Step 0 : FROM ubuntu:14.04
---> 99ec81b80c55
Step 1 : MAINTAINER Newbee <newbee@docker.com>
---> Running in 7c5664a8a0c1
---> 2fa8ca4e2a13
Removing intermediate container 7c5664a8a0c1
Step 2 : RUN apt-get -qq update
---> Running in b07cc3fb4256
---> 50d21070ec0c
Removing intermediate container b07cc3fb4256
Step 3 : RUN apt-get -qqy install ruby ruby-dev
---> Running in a5b038dd127e
Selecting previously unselected package libasan0:amd64.
(Reading database ... 11518 files and directories currently installed.)
Preparing to unpack .../libasan0_4.8.2-19ubuntu1_amd64.deb ...
Setting up ruby (1:1.9.3.4) ...
Setting up ruby1.9.1 (1.9.3.484-2ubuntu1) ...
Processing triggers for libc-bin (2.19-0ubuntu6) ...
---> 2acb20f17878
Removing intermediate container a5b038dd127e
Step 4 : RUN gem install sinatra
---> Running in 5e9d0065c1f7
```

```

. . .
Successfully installed rack-protection-1.5.3
Successfully installed sinatra-1.4.5
4 gems installed
---> 324104cde6ad
Removing intermediate container 5e9d0065c1f7
Successfully built 324104cde6ad

```

其中 `-t` 标记来添加 tag，指定新的镜像的用户信息。“.”是 Dockerfile 所在的路径（当前目录），也可以替换为一个具体的 Dockerfile 的路径。

可以看到 build 进程在执行操作。它要做的第一件事情就是上传这个 Dockerfile 内容，因为所有的操作都要依据 Dockerfile 来进行。然后，Dockerfile 中的指令被一条一条的执行。每一步都创建了一个新的容器，在容器中执行指令并提交修改（就跟之前介绍过的 `docker commit` 一样）。当所有的指令都执行完毕之后，返回了最终的镜像 id。所有的中间步骤所产生的容器都被删除和清理了。

*注意一个镜像不能超过 127 层

此外，还可以利用 `ADD` 命令复制本地文件到镜像；用 `EXPOSE` 命令来向外部开放端口；用 `CMD` 命令来描述容器启动后运行的程序等。例如

```

# put my local web site in myApp folder to /var/www
ADD myApp /var/www
# expose httpd port
EXPOSE 80
# the command to run
CMD ["/usr/sbin/apachectl", "-D", "FOREGROUND"]

```

现在可以利用新创建的镜像来启动一个容器。

```

$ sudo docker run -t -i ouruser/sinatra:v2 /bin/bash
root@8196968dac35:/#

```

还可以用 `docker tag` 命令来修改镜像的标签。

```

$ sudo docker tag 5db5f8471261 ouruser/sinatra:devel
$ sudo docker images ouruser/sinatra

```

REPOSITORY	TAG	IMAGE ID	CREATED	VIRTUAL SIZE
ouruser/sinatra	latest	5db5f8471261	11 hours ago	446.7 MB
ouruser/sinatra	devel	5db5f8471261	11 hours ago	446.7 MB
ouruser/sinatra	v2	5db5f8471261	11 hours ago	446.7 MB

*注：更多用法，请参考 [Dockerfile](#) 章节。

从本地文件系统导入

要从本地文件系统导入一个镜像，可以使用 `openvz`（容器虚拟化的先锋技术）的模板来创建：`openvz` 的模板下载地址为 [templates](#)。

比如，先下载了一个 ubuntu-14.04 的镜像，之后使用以下命令导入：

```
sudo cat ubuntu-14.04-x86_64-minimal.tar.gz |docker import - ubuntu:14.04
```

然后查看新导入的镜像。

```
docker images
```

REPOSITORY	TAG	IMAGE ID	CREATED	VIRTUAL S
ubuntu	14.04	05ac7c0b9383	17 seconds ago	215.5 MB

上传镜像

用户可以通过 `docker push` 命令，把自己创建的镜像上传到仓库中来共享。例如，用户在 Docker Hub 上完成注册后，可以推送自己的镜像到仓库中。

```
$ sudo docker push ouruser/sinatra
The push refers to a repository [ouruser/sinatra] (len: 1)
Sending image list
Pushing repository ouruser/sinatra (3 tags)
```

存出和载入镜像

存出镜像

如果要导出镜像到本地文件，可以使用 `docker save` 命令。

```
$ sudo docker images
REPOSITORY          TAG                 IMAGE ID            CREATED             VIRTUAL S
ubuntu              14.04              c4ff7513909d       5 weeks ago        225.4 MB
...
$ sudo docker save -o ubuntu_14.04.tar ubuntu:14.04
```

载入镜像

可以使用 `docker load` 从导出的本地文件中再导入到本地镜像库，例如

```
$ sudo docker load --input ubuntu_14.04.tar
```

或

```
$ sudo docker load < ubuntu_14.04.tar
```

这将导入镜像以及其相关的元数据信息（包括标签等）。

移除本地镜像

如果要移除本地的镜像，可以使用 `docker rmi` 命令。注意 `docker rm` 命令是移除容器。

```
$ sudo docker rmi training/sinatra
Untagged: training/sinatra:latest
Deleted: 5bc342fa0b91cabf65246837015197eecfa24b2213ed6a51a8974ae250fedd8d
Deleted: ed0ffffdcdae5eb2c3a55549857a8be7fc8bc4241fb19ad714364cbfd7a56b22f
Deleted: 5c58979d73ae448df5af1d8142436d81116187a7633082650549c52c3a2418f0
```

*注意：在删除镜像之前要先用 `docker rm` 删掉依赖于这个镜像的所有容器。

清理所有未打过标签的本地镜像

`docker images` 可以列出本地所有的镜像，其中很可能会包含有很多中间状态的未打过标签的镜像，大量占据着磁盘空间。

使用下面的命令可以清理所有未打过标签的本地镜像

```
$ sudo docker rmi $(docker images -q -f "dangling=true")
```

其中 `-q` 和 `-f` 是缩写，完整的命令其实可以写着下面这样，是不是更容易理解一点？

```
$ sudo docker rmi $(docker images --quiet --filter "dangling=true")
```

镜像的实现原理

Docker 镜像是怎么实现增量的修改和维护的？每个镜像都由很多层次构成，Docker 使用 [Union FS](#) 将这些不同的层结合到一个镜像中去。

通常 Union FS 有两个用途，一方面可以实现不借助 LVM、RAID 将多个 disk 挂到同一个目录下，另一个更常用的就是将一个只读的分支和一个可写的分支联合在一起，Live CD 正是基于此方法可以允许在镜像不变的基础上允许用户在其上进行一些写操作。Docker 在 AUFS 上构建的容器也是利用了类似的原理。

Docker 容器

容器是 Docker 又一核心概念。

简单的说，容器是独立运行的一个或一组应用，以及它们的运行态环境。对应的，虚拟机可以理解为模拟运行的一整套操作系统（提供了运行态环境和其他系统环境）和跑在上面的应用。

本章将具体介绍如何来管理一个容器，包括创建、启动和停止等。

启动容器

启动容器有两种方式，一种是基于镜像新建一个容器并启动，另外一个是在终止状态（stopped）的容器重新启动。

因为 Docker 的容器实在太轻量级了，很多时候用户都是随时删除和新创建容器。

新建并启动

所需要的命令主要为 `docker run`。

例如，下面的命令输出一个“Hello World”，之后终止容器。

```
$ sudo docker run ubuntu:14.04 /bin/echo 'Hello world'
Hello world
```

这跟在本地直接执行 `/bin/echo 'hello world'` 几乎感觉不出任何区别。

下面的命令则启动一个 `bash` 终端，允许用户进行交互。

```
$ sudo docker run -t -i ubuntu:14.04 /bin/bash
root@af8bae53bdd3:/#
```

其中，`-t` 选项让 Docker 分配一个伪终端（pseudo-tty）并绑定到容器的标准输入上，`-i` 则让容器的标准输入保持打开。

在交互模式下，用户可以通过所创建的终端来输入命令，例如

```
root@af8bae53bdd3:/# pwd
/
root@af8bae53bdd3:/# ls
bin boot dev etc home lib lib64 media mnt opt proc root run sbin srv sys tmp usr var
```

当利用 `docker run` 来创建容器时，Docker 在后台运行的标准操作包括：

- 检查本地是否存在指定的镜像，不存在就从公有仓库下载
- 利用镜像创建并启动一个容器
- 分配一个文件系统，并在只读的镜像层外面挂载一层可读写层
- 从宿主主机配置的网桥接口中桥接一个虚拟接口到容器中去
- 从地址池配置一个 ip 地址给容器
- 执行用户指定的应用程序
- 执行完毕后容器被终止

启动已终止容器

可以利用 `docker start` 命令，直接将一个已经终止的容器启动运行。

容器的核心为所执行的应用程序，所需要的资源都是应用程序运行所必需的。除此之外，并没有其它的资源。可以在伪终端中利用 `ps` 或 `top` 来查看进程信息。

```
root@ba267838cc1b:/# ps
PID TTY          TIME CMD
  1 ?            00:00:00 bash
 11 ?            00:00:00 ps
```

可见，容器中仅运行了指定的 `bash` 应用。这种特点使得 Docker 对资源的利用率极高，是货真价实的轻量级虚拟化。

后台(background)运行

更多的时候，需要让 Docker 在后台运行而不是直接把执行命令的结果输出在当前宿主机下。此时，可以通过添加 `-d` 参数来实现。

下面举两个例子来说明一下。

如果不使用 `-d` 参数运行容器。

```
$ sudo docker run ubuntu:14.04 /bin/sh -c "while true; do echo hello world; sleep 1; done
hello world
hello world
hello world
hello world
```

容器会把输出的结果(STDOUT)打印到宿主机上面

如果使用了 `-d` 参数运行容器。

```
$ sudo docker run ubuntu:14.04 /bin/sh -c "while true; do echo hello world; sleep 1; done
77b2dc01fe0f3f1265df143181e7b9af5e05279a884f4776ee75350ea9d8017a
```

此时容器会在后台运行并不会把输出的结果(STDOUT)打印到宿主机上面(输出结果可以用 `docker logs` 查看)。

注：容器是否会会长久运行，是和 `docker run` 指定的命令有关，和 `-d` 参数无关。

使用 `-d` 参数启动后会返回一个唯一的 id，也可以通过 `docker ps` 命令来查看容器信息。

```
$ sudo docker ps
CONTAINER ID   IMAGE          COMMAND                  CREATED        STATUS        PORTS   NAMES
77b2dc01fe0f   ubuntu:14.04  /bin/sh -c 'while tr    2 minutes ago Up 1 minute          agita
```

要获取容器的输出信息，可以通过 `docker logs` 命令。

```
$ sudo docker logs [container ID or NAMES]
hello world
hello world
hello world
. . .
```

终止容器

可以使用 `docker stop` 来终止一个运行中的容器。

此外，当Docker容器中指定的应用终结时，容器也自动终止。例如对于上一章节中只启动了一个终端的容器，用户通过 `exit` 命令或 `Ctrl+d` 来退出终端时，所创建的容器立刻终止。

终止状态的容器可以用 `docker ps -a` 命令看到。例如

```
sudo docker ps -a
CONTAINER ID        IMAGE               COMMAND             CREATED            STATUS
ba267838cc1b      ubuntu:14.04       "/bin/bash"        30 minutes ago    Exited
98e5efa7d997      training/webapp:latest "python app.py"    About an hour ago Exited
```

处于终止状态的容器，可以通过 `docker start` 命令来重新启动。

此外，`docker restart` 命令会将一个运行态的容器终止，然后再重新启动它。

进入容器

在使用 `-d` 参数时，容器启动后会进入后台。某些时候需要进入容器进行操作，有很多种方法，包括使用 `docker attach` 命令或 `nsenter` 工具等。

attach 命令

`docker attach` 是 Docker 自带的命令。下面示例如何使用该命令。

```
$ sudo docker run -idt ubuntu
243c32535da7d142fb0e6df616a3c3ada0b8ab417937c853a9e1c251f499f550
$ sudo docker ps
CONTAINER ID        IMAGE               COMMAND             CREATED             STATUS
243c32535da7        ubuntu:latest      "/bin/bash"        18 seconds ago     Up 17 sec
$ sudo docker attach nostalgic_hypatia
root@243c32535da7:/#
```

但是使用 `attach` 命令有时候并不方便。当多个窗口同时 `attach` 到同一个容器的时候，所有窗口都会同步显示。当某个窗口因命令阻塞时，其他窗口也无法执行操作了。

nsenter 命令

安装

`nsenter` 工具在 `util-linux` 包 2.23 版本后包含。如果系统中 `util-linux` 包没有该命令，可以按照下面的方法从源码安装。

```
$ cd /tmp; curl https://www.kernel.org/pub/linux/utils/util-linux/v2.24/util-linux-2.24.t
$ ./configure --without-ncurses
$ make nsenter && sudo cp nsenter /usr/local/bin
```

使用

`nsenter` 可以访问另一个进程的名字空间。`nsenter` 要正常工作需要有 `root` 权限。很不幸，Ubuntu 14.04 仍然使用的是 `util-linux 2.20`。安装最新版本的 `util-linux (2.24)` 版，请按照以下步骤：

```
$ wget https://www.kernel.org/pub/linux/utils/util-linux/v2.24/util-linux-2.24.tar.gz; ta
$ cd util-linux-2.24
$ ./configure --without-ncurses && make nsenter
$ sudo cp nsenter /usr/local/bin
```

为了连接到容器，你还需要找到容器的第一个进程的 PID，可以通过下面的命令获取。

```
PID=$(docker inspect --format "{{ .State.Pid }}" <container>)
```

通过这个 PID，就可以连接到这个容器：

```
$ nsenter --target $PID --mount --uts --ipc --net --pid
```

下面给出一个完整的例子。

```
$ sudo docker run -idt ubuntu
243c32535da7d142fb0e6df616a3c3ada0b8ab417937c853a9e1c251f499f550
$ sudo docker ps
CONTAINER ID        IMAGE               COMMAND             CREATED             STATUS
243c32535da7        ubuntu:latest      "/bin/bash"        18 seconds ago     Up 17 sec
$ PID=$(docker-pid 243c32535da7)
10981
$ sudo nsenter --target 10981 --mount --uts --ipc --net --pid
root@243c32535da7:/#
```

更简单的，建议大家下载 [.bashrc_docker](#)，并将内容放到 `.bashrc` 中。

```
$ wget -P ~ https://github.com/yeasy/docker_practice/raw/master/_local/.bashrc_docker;
$ echo "[ -f ~/.bashrc_docker ] && . ~/.bashrc_docker" >> ~/.bashrc; source ~/.bashrc
```

这个文件中定义了很多方便使用 Docker 的命令，例如 `docker-pid` 可以获取某个容器的 PID；而 `docker-enter` 可以进入容器或直接在容器内执行命令。

```
$ echo $(docker-pid <container>)
$ docker-enter <container> ls
```

导出和导入容器

导出容器

如果要导出本地某个容器，可以使用 `docker export` 命令。

```
$ sudo docker ps -a
CONTAINER ID        IMAGE               COMMAND             CREATED             STATUS
7691a814370e      ubuntu:14.04       "/bin/bash"        36 hours ago       Exited (0)
$ sudo docker export 7691a814370e > ubuntu.tar
```

这样将导出容器快照到本地文件。

导入容器快照

可以使用 `docker import` 从容器快照文件中再导入为镜像，例如

```
$ cat ubuntu.tar | sudo docker import - test/ubuntu:v1.0
$ sudo docker images
REPOSITORY          TAG                 IMAGE ID            CREATED             VIRTUAL
test/ubuntu         v1.0               9d37a6082e97      About a minute ago 171.3 MB
```

此外，也可以通过指定 URL 或者某个目录来导入，例如

```
$ sudo docker import http://example.com/exampleimage.tgz example/imagerepo
```

*注：用户既可以使用 `docker load` 来导入镜像存储文件到本地镜像库，也可以使用 `docker import` 来导入一个容器快照到本地镜像库。这两者的区别在于容器快照文件将丢弃所有的历史记录和元数据信息（即仅保存容器当时的快照状态），而镜像存储文件将保存完整记录，体积也要大。此外，从容器快照文件导入时可以重新指定标签等元数据信息。

删除容器

可以使用 `docker rm` 来删除一个处于终止状态的容器。例如

```
$sudo docker rm trusting_newton
trusting_newton
```

如果要删除一个运行中的容器，可以添加 `-f` 参数。Docker 会发送 `SIGKILL` 信号给容器。

清理所有处于终止状态的容器

用 `docker ps -a` 命令可以查看所有已经创建的包括终止状态的容器，如果数量太多要一个个删除可能会很麻烦，用 `docker rm $(docker ps -a -q)` 可以全部清理掉。

*注意：这个命令其实会试图删除所有的包括还在运行中的容器，不过就像上面提过的 `docker rm` 默认并不会删除运行中的容器。

仓库

仓库（Repository）是集中存放镜像的地方。

一个容易混淆的概念是注册服务器（Registry）。实际上注册服务器是管理仓库的具体服务器，每个服务器上可以有多个仓库，而每个仓库下面有多个镜像。从这方面来说，仓库可以被认为是一个具体的项目或目录。例如对于仓库地址 `dl.dockerpool.com/ubuntu` 来说，`dl.dockerpool.com` 是注册服务器地址，`ubuntu` 是仓库名。

大部分时候，并不需要严格区分这两者的概念。

Docker Hub

目前 Docker 官方维护了一个公共仓库 [Docker Hub](#)，其中已经包括了超过 15,000 的镜像。大部分需求，都可以通过在 Docker Hub 中直接下载镜像来实现。

登录

可以通过执行 `docker login` 命令来输入用户名、密码和邮箱来完成注册和登录。注册成功后，本地用户目录的 `.dockercfg` 中将保存用户的认证信息。

基本操作

用户无需登录即可通过 `docker search` 命令来查找官方仓库中的镜像，并利用 `docker pull` 命令来将它下载到本地。

例如以 `centos` 为关键词进行搜索：

```
$ sudo docker search centos
NAME                DESCRIPTION
centos              The official build of CentOS.
tianon/centos       CentOS 5 and 6, created using rinse inste
blalor/centos       Bare-bones base CentOS 6.5 image
saltstack/centos-6-minimal
tutum/centos-6.4    DEPRECATED. Use tutum/centos:6.4 instead.
...
```

可以看到返回了很多包含关键字的镜像，其中包括镜像名字、描述、星级（表示该镜像的受欢迎程度）、是否官方创建、是否自动创建。官方的镜像说明是官方项目组创建和维护的，`automated` 资源允许用户验证镜像的来源和内容。

根据是否是官方提供，可将镜像资源分为两类。一种是类似 `centos` 这样的基础镜像，被称为基础或根镜像。这些基础镜像是由 Docker 公司创建、验证、支持、提供。这样的镜像往往使用单个单词作为名字。还有一种类型，比如 `tianon/centos` 镜像，它是由 Docker 的用户创建并维护的，往往带有用户名称前缀。可以通过前缀 `user_name/` 来指定使用某个用户提供的镜像，比如 `tianon` 用户。

另外，在查找的时候通过 `-s N` 参数可以指定仅显示评价为 `N` 星以上的镜像。

下载官方 `centos` 镜像到本地。

```
$ sudo docker pull centos
Pulling repository centos
0b443ba03958: Download complete
539c0211cd76: Download complete
511136ea3c5a: Download complete
7064731afe90: Download complete
```

用户也可以在登录后通过 `docker push` 命令来将镜像推送到 Docker Hub。

自动创建

自动创建 (Automated Builds) 功能对于需要经常升级镜像内程序来说，十分方便。有时候，用户创建了镜像，安装了某个软件，如果软件发布新版本则需要手动更新镜像。。

而自动创建允许用户通过 Docker Hub 指定跟踪一个目标网站（目前支持 [GitHub](#) 或 [BitBucket](#)）上的项目，一旦项目发生新的提交，则自动执行创建。

要配置自动创建，包括如下的步骤：

- 创建并登录 Docker Hub，以及目标网站；
- 在目标网站中连接帐户到 Docker Hub；
- 在 Docker Hub 中 [配置一个自动创建](#)；
- 选取一个目标网站中的项目（需要含 Dockerfile）和分支；
- 指定 Dockerfile 的位置，并提交创建。

之后，可以在 Docker Hub 的 [自动创建页面](#) 中跟踪每次创建的状态。

私有仓库

有时候使用 Docker Hub 这样的公共仓库可能不方便，用户可以创建一个本地仓库供私人使用。

本节介绍如何使用本地仓库。

`docker-registry` 是官方提供的工具，可以用于构建私有的镜像仓库。

安装运行 `docker-registry`

容器运行

在安装了 Docker 后，可以通过获取官方 registry 镜像来运行。

```
$ sudo docker run -d -p 5000:5000 registry
```

这将使用官方的 registry 镜像来启动本地的私有仓库。用户可以通过指定参数来配置私有仓库位置，例如配置镜像存储到 Amazon S3 服务。

```
$ sudo docker run \
  -e SETTINGS_FLAVOR=s3 \
  -e AWS_BUCKET=acme-docker \
  -e STORAGE_PATH=/registry \
  -e AWS_KEY=AKIAHSHB43HS3J92MXZ \
  -e AWS_SECRET=xdDoww1K7TJajV1Y7Eo0ZrmuPEJlHYcNP2k4j49T \
  -e SEARCH_BACKEND=sqlalchemy \
  -p 5000:5000 \
  registry
```

此外，还可以指定本地路径（如 `/home/user/registry-conf`）下的配置文件。

```
$ sudo docker run -d -p 5000:5000 -v /home/user/registry-conf:/registry-conf -e DOCKER_RE
```

默认情况下，仓库会被创建在容器的 `/tmp/registry` 下。可以通过 `-v` 参数来将镜像文件存放在本地的指定路径。例如下面的例子将上传的镜像放到 `/opt/data/registry` 目录。

```
$ sudo docker run -d -p 5000:5000 -v /opt/data/registry:/tmp/registry registry
```

本地安装

对于 Ubuntu 或 CentOS 等发行版，可以直接通过源安装。

- Ubuntu

```
$ sudo apt-get install -y build-essential python-dev libevent-dev python-pip liblzma-dev
$ sudo pip install docker-registry
```

- CentOS

```
$ sudo yum install -y python-devel libevent-devel python-pip gcc xz-devel
$ sudo python-pip install docker-registry
```

也可以从 [docker-registry](#) 项目下载源码进行安装。

```
$ sudo apt-get install build-essential python-dev libevent-dev python-pip libssl-dev libl
$ git clone https://github.com/docker/docker-registry.git
$ cd docker-registry
$ sudo python setup.py install
```

然后修改配置文件，主要修改 dev 模板段的 `storage_path` 到本地的存储仓库的路径。

```
$ cp config/config_sample.yml config/config.yml
```

之后启动 Web 服务。

```
$ sudo gunicorn -c contrib/gunicorn.py docker_registry.wsgi:application
```

或者

```
$ sudo gunicorn --access-logfile - --error-logfile - -k gevent -b 0.0.0.0:5000 -w 4 --max
```

此时使用 `curl` 访问本地的 5000 端口，看到输出 `docker-registry` 的版本信息说明运行成功。

*注：`config/config_sample.yml` 文件是示例配置文件。

在私有仓库上传、下载、搜索镜像

创建好私有仓库之后，就可以使用 `docker tag` 来标记一个镜像，然后推送它到仓库，别的机器上就可以下载下来了。例如私有仓库地址为 `192.168.7.26:5000`。

先在本机查看已有的镜像。

```
$ sudo docker images
REPOSITORY          TAG          IMAGE ID          CREATED
ubuntu              latest      ba5877dc9bec     6 weeks ago
ubuntu              14.04      ba5877dc9bec     6 weeks ago
```



使用 `docker tag` 将 `ba58` 这个镜像标记为 `192.168.7.26:5000/test`（格式为 `docker tag IMAGE[:TAG] [REGISTRYHOST/][USERNAME/]NAME[:TAG]`）。

```
$ sudo docker tag ba58 192.168.7.26:5000/test
root ~ # docker images
REPOSITORY          TAG          IMAGE ID          CREATED
ubuntu              14.04       ba5877dc9bec     6 weeks ago
ubuntu              latest      ba5877dc9bec     6 weeks ago
192.168.7.26:5000/test latest      ba5877dc9bec     6 weeks ago
```

使用 `docker push` 上传标记的镜像。

```
$ sudo docker push 192.168.7.26:5000/test
The push refers to a repository [192.168.7.26:5000/test] (len: 1)
Sending image list
Pushing repository 192.168.7.26:5000/test (1 tags)
Image 511136ea3c5a already pushed, skipping
Image 9bad880da3d2 already pushed, skipping
Image 25f11f5fb0cb already pushed, skipping
Image ebc34468f71d already pushed, skipping
Image 2318d26665ef already pushed, skipping
Image ba5877dc9bec already pushed, skipping
Pushing tag for rev [ba5877dc9bec] on {http://192.168.7.26:5000/v1/repositories/test/tags}
```

用 `curl` 查看仓库中的镜像。

```
$ curl http://192.168.7.26:5000/v1/search
{"num_results": 7, "query": "", "results": [{"description": "", "name": "library/miaxis_j"}
```

这里可以看到 `{"description": "", "name": "library/test"}`，表明镜像已经被成功上传了。

现在可以到另外一台机器去下载这个镜像。

```
$ sudo docker pull 192.168.7.26:5000/test
Pulling repository 192.168.7.26:5000/test
ba5877dc9bec: Download complete
511136ea3c5a: Download complete
9bad880da3d2: Download complete
25f11f5fb0cb: Download complete
ebc34468f71d: Download complete
2318d26665ef: Download complete
$ sudo docker images
REPOSITORY          TAG          IMAGE ID          CREATED
192.168.7.26:5000/test latest      ba5877dc9bec     6 weeks ago
```

可以使用 [这个脚本](#) 批量上传本地的镜像到注册服务器中，默认是本地注册服务器 `127.0.0.1:5000`。例如：

```
$ wget https://github.com/yeasy/docker_practice/raw/master/_local/push_images.sh; sudo ch
$ ./push_images.sh ubuntu:latest centos:centos7
The registry server is 127.0.0.1
Uploading ubuntu:latest...
The push refers to a repository [127.0.0.1:5000/ubuntu] (len: 1)
Sending image list
Pushing repository 127.0.0.1:5000/ubuntu (1 tags)
Image 511136ea3c5a already pushed, skipping
Image bfb8b5a2ad34 already pushed, skipping
Image c1f3bdbd8355 already pushed, skipping
Image 897578f527ae already pushed, skipping
Image 9387bcc9826e already pushed, skipping
Image 809ed259f845 already pushed, skipping
Image 96864a7d2df3 already pushed, skipping
Pushing tag for rev [96864a7d2df3] on {http://127.0.0.1:5000/v1/repositories/ubuntu/tags/
Untagged: 127.0.0.1:5000/ubuntu:latest
Done
Uploading centos:centos7...
The push refers to a repository [127.0.0.1:5000/centos] (len: 1)
Sending image list
Pushing repository 127.0.0.1:5000/centos (1 tags)
Image 511136ea3c5a already pushed, skipping
34e94e67e63a: Image successfully pushed
70214e5d0a90: Image successfully pushed
Pushing tag for rev [70214e5d0a90] on {http://127.0.0.1:5000/v1/repositories/centos/tags/
Untagged: 127.0.0.1:5000/centos:centos7
Done
```

仓库配置文件

Docker 的 Registry 利用配置文件提供了一些仓库的模板（flavor），用户可以直接使用它们来进行开发或生产部署。

模板

在 `config_sample.yml` 文件中，可以看到一些现成的模板段：

- `common`：基础配置
- `local`：存储数据到本地文件系统
- `s3`：存储数据到 AWS S3 中
- `dev`：使用 `local` 模板的基本配置
- `test`：单元测试使用
- `prod`：生产环境配置（基本上跟s3配置类似）
- `gcs`：存储数据到 Google 的云存储
- `swift`：存储数据到 OpenStack Swift 服务
- `glance`：存储数据到 OpenStack Glance 服务，本地文件系统为后备
- `glance-swift`：存储数据到 OpenStack Glance 服务，Swift 为后备
- `elliptics`：存储数据到 Elliptics key/value 存储

用户也可以添加自定义的模板段。

默认情况下使用的模板是 `dev`，要使用某个模板作为默认值，可以添加 `SETTINGS_FLAVOR` 到环境变量中，例如

```
export SETTINGS_FLAVOR=dev
```

另外，配置文件中支持从环境变量中加载值，语法格式为 `_env:VARIABLENAME[:DEFAULT]`。

示例配置

```
common:
  loglevel: info
  search_backend: "_env:SEARCH_BACKEND:"
  sqlalchemy_index_database:
    "_env:SQLALCHEMY_INDEX_DATABASE:sqlite:///tmp/docker-registry.db"

prod:
  loglevel: warn
  storage: s3
  s3_access_key: _env:AWS_S3_ACCESS_KEY
  s3_secret_key: _env:AWS_S3_SECRET_KEY
  s3_bucket: _env:AWS_S3_BUCKET
  boto_bucket: _env:AWS_S3_BUCKET
  storage_path: /srv/docker
  smtp_host: localhost
```

```
from_addr: docker@myself.com
to_addr: my@myself.com

dev:
  loglevel: debug
  storage: local
  storage_path: /home/myself/docker

test:
  storage: local
  storage_path: /tmp/tmpdockertmp
```

选项

Docker 数据管理

这一章介绍如何在 Docker 内部以及容器之间管理数据，在容器中管理数据主要有两种方式：

- 数据卷 (Data volumes)
- 数据卷容器 (Data volume containers)

数据卷

数据卷是一个可供一个或多个容器使用的特殊目录，它绕过 UFS，可以提供很多有用的特性：

- 数据卷可以在容器之间共享和重用
- 对数据卷的修改会立马生效
- 对数据卷的更新，不会影响镜像
- 数据卷默认会一直存在，即使容器被删除

*注意：数据卷的使用，类似于 Linux 下对目录或文件进行 mount，镜像中的被指定为挂载点的目录中的文件会隐藏掉，能显示看的是挂载的数据卷。

创建一个数据卷

在用 `docker run` 命令的时候，使用 `-v` 标记来创建一个数据卷并挂载到容器里。在一次 run 中多次使用可以挂载多个数据卷。

下面创建一个名为 web 的容器，并加载一个数据卷到容器的 `/webapp` 目录。

```
$ sudo docker run -d -P --name web -v /webapp training/webapp python app.py
```

*注意：也可以在 Dockerfile 中使用 `VOLUME` 来添加一个或者多个新的卷到由该镜像创建的任意容器。

删除数据卷

数据卷是被设计用来持久化数据的，它的生命周期独立于容器，Docker 不会在容器被删除后自动删除数据卷，并且也不存在垃圾回收这样的机制来处理没有任何容器引用的数据卷。如果需要在删除容器的同时移除数据卷。可以在删除容器的时候使用 `docker rm -v` 这个命令。无主的数据卷可能会占据很多空间，要清理会很麻烦。Docker 官方正在试图解决这个问题，相关工作的进度可以查看这个 [PR](#)

挂载一个主机目录作为数据卷

使用 `-v` 标记也可以指定挂载一个本地主机的目录到容器中去。

```
$ sudo docker run -d -P --name web -v /src/webapp:/opt/webapp training/webapp python app.
```

上面的命令加载主机的 `/src/webapp` 目录到容器的 `/opt/webapp` 目录。这个功能在进行测试的时候十分方便，比如用户可以放置一些程序到本地目录中，来查看容器是否正常工作。本地目录的路径必须是绝对路径，如果目录不存在 Docker 会自动为你创建它。

*注意：Dockerfile 中不支持这种用法，这是因为 Dockerfile 是为了移植和分享用的。然而，不同操作系统的路径格式不一样，所以目前还不能支持。

Docker 挂载数据卷的默认权限是读写，用户也可以通过 `:ro` 指定为只读。

```
$ sudo docker run -d -P --name web -v /src/webapp:/opt/webapp:ro
training/webapp python app.py
```

加了 `:ro` 之后，就挂载为只读了。

查看数据卷的具体信息

在主机里使用以下命令可以查看指定容器的信息

```
$ docker inspect web
...
```

在输出的内容中找到其中和数据卷相关的部分，可以看到所有的数据卷都是创建在主机的 `/var/lib/docker/volumes/` 下面的

```
"Volumes": {
  "/webapp": "/var/lib/docker/volumes/fac362...80535"
},
"VolumesRW": {
  "/webapp": true
}
...
```

挂载一个本地主机文件作为数据卷

`-v` 标记也可以从主机挂载单个文件到容器中

```
$ sudo docker run --rm -it -v ~/.bash_history:/.bash_history ubuntu /bin/bash
```

这样就可以记录在容器输入过的命令了。

*注意：如果直接挂载一个文件，很多文件编辑工具，包括 `vi` 或者 `sed --in-place`，可能会造成文件 inode 的改变，从 Docker 1.1.0 起，这会导致报错误信息。所以最简单的办法就直接挂载文件的父目录。

数据卷容器

如果你有一些持续更新的数据需要在容器之间共享，最好创建数据卷容器。

数据卷容器，其实就是一个正常的容器，专门用来提供数据卷供其它容器挂载的。

首先，创建一个名为 `dbdata` 的数据卷容器：

```
$ sudo docker run -d -v /dbdata --name dbdata training/postgres echo Data-only container
```

然后，在其他容器中使用 `--volumes-from` 来挂载 `dbdata` 容器中的数据卷。

```
$ sudo docker run -d --volumes-from dbdata --name db1 training/postgres
$ sudo docker run -d --volumes-from dbdata --name db2 training/postgres
```

可以使用超过一个的 `--volumes-from` 参数来指定从多个容器挂载不同的数据卷。也可以从其他已经挂载了数据卷的容器来级联挂载数据卷。

```
$ sudo docker run -d --name db3 --volumes-from db1 training/postgres
```

*注意：使用 `--volumes-from` 参数所挂载数据卷的容器自己并不需要保持在运行状态。

如果删除了挂载的容器（包括 `dbdata`、`db1` 和 `db2`），数据卷并不会被自动删除。如果要删除一个数据卷，必须在删除最后一个还挂载着它的容器时使用 `docker rm -v` 命令来指定同时删除关联的容器。这可以让用户在容器之间升级和移动数据卷。具体的操作将在下一节中进行讲解。

利用数据卷容器来备份、恢复、迁移数据卷

可以利用数据卷对其中的数据进行备份、恢复和迁移。

备份

首先使用 `--volumes-from` 标记来创建一个加载 `dbdata` 容器卷的容器，并从主机挂载当前目录到容器的 `/backup` 目录。命令如下：

```
$ sudo docker run --volumes-from dbdata -v $(pwd):/backup ubuntu tar cvf /backup/backup.t
```

容器启动后，使用了 `tar` 命令来将 `dbdata` 卷备份为容器中 `/backup/backup.tar` 文件，也就是主机当前目录下的名为 `backup.tar` 的文件。

恢复

如果要恢复数据到一个容器，首先创建一个带有空数据卷的容器 `dbdata2`。

```
$ sudo docker run -v /dbdata --name dbdata2 ubuntu /bin/bash
```

然后创建另一个容器，挂载 `dbdata2` 容器卷中的数据卷，并使用 `untar` 解压备份文件到挂载的容器卷中。

```
$ sudo docker run --volumes-from dbdata2 -v $(pwd):/backup busybox tar xvf /backup/backup.tar
```

为了查看/验证恢复的数据，可以再启动一个容器挂载同样的容器卷来查看

```
$ sudo docker run --volumes-from dbdata2 busybox /bin/ls /dbdata
```

Docker 中的网络功能介绍

Docker 允许通过外部访问容器或容器互联的方式来提供网络服务。

外部访问容器

容器中可以运行一些网络应用，要让外部也可以访问这些应用，可以通过 `-P` 或 `-p` 参数来指定端口映射。

当使用 `-P` 标记时，Docker 会随机映射一个 `49000~49900` 的端口到内部容器开放的网络端口。

使用 `docker ps` 可以看到，本地主机的 49155 被映射到了容器的 5000 端口。此时访问本机的 49155 端口即可访问容器内 web 应用提供的界面。

```
$ sudo docker run -d -P training/webapp python app.py
$ sudo docker ps -l
CONTAINER ID   IMAGE                COMMAND             CREATED          STATUS            PORTS
bc533791f3f5   training/webapp:latest  python app.py      5 seconds ago   Up 2 seconds     0.0.0.0:
```

同样的，可以通过 `docker logs` 命令来查看应用的信息。

```
$ sudo docker logs -f nostalgic_morse
* Running on http://0.0.0.0:5000/
10.0.2.2 - - [23/May/2014 20:16:31] "GET / HTTP/1.1" 200 -
10.0.2.2 - - [23/May/2014 20:16:31] "GET /favicon.ico HTTP/1.1" 404 -
```

`-p` (小写的) 则可以指定要映射的端口，并且，在一个指定端口上只可以绑定一个容器。支持的格式有 `ip:hostPort:containerPort` | `ip::containerPort` | `hostPort:containerPort`。

映射所有接口地址

使用 `hostPort:containerPort` 格式本地的 5000 端口映射到容器的 5000 端口，可以执行

```
$ sudo docker run -d -p 5000:5000 training/webapp python app.py
```

此时默认会绑定本地所有接口上的所有地址。

映射到指定地址的指定端口

可以使用 `ip:hostPort:containerPort` 格式指定映射使用一个特定地址，比如 localhost 地址 127.0.0.1

```
$ sudo docker run -d -p 127.0.0.1:5000:5000 training/webapp python app.py
```

映射到指定地址的任意端口

使用 `ip::containerPort` 绑定 localhost 的任意端口到容器的 5000 端口，本地主机会自动分配一个端

口。

```
$ sudo docker run -d -p 127.0.0.1::5000 training/webapp python app.py
```

还可以使用 `udp` 标记来指定 `udp` 端口

```
$ sudo docker run -d -p 127.0.0.1:5000:5000/udp training/webapp python app.py
```

查看映射端口配置

使用 `docker port` 来查看当前映射的端口配置，也可以查看到绑定的地址

```
$ docker port nostalgic_morse 5000
127.0.0.1:49155.
```

注意：

- 容器有自己的内部网络和 ip 地址（使用 `docker inspect` 可以获取所有的变量，Docker 还可以有一个可变的网络配置。）
- `-p` 标记可以多次使用来绑定多个端口

例如

```
$ sudo docker run -d -p 5000:5000 -p 3000:80 training/webapp python app.py
```

容器互联

容器的连接（linking）系统是除了端口映射外，另一种跟容器中应用交互的方式。

该系统会在源和接收容器之间创建一个隧道，接收容器可以看到源容器指定的信息。

自定义容器命名

连接系统依据容器的名称来执行。因此，首先需要自定义一个好记的容器命名。

虽然当创建容器的时候，系统默认会分配一个名字。自定义命名容器有2个好处：

- 自定义的命名，比较好记，比如一个web应用容器我们可以给它起名叫web
- 当要连接其他容器时候，可以作为一个有用的参考点，比如连接web容器到db容器

使用 `--name` 标记可以为容器自定义命名。

```
$ sudo docker run -d -P --name web training/webapp python app.py
```

使用 `docker ps` 来验证设定的命名。

```
$ sudo docker ps -l
CONTAINER ID   IMAGE                                COMMAND                  CREATED        STATUS        PORTS
aed84ee21bde   training/webapp:latest              python app.py           12 hours ago  Up 2 seconds  0.0.0.0:49
```

也可以使用 `docker inspect` 来查看容器的名字

```
$ sudo docker inspect -f "{{ .Name }}" aed84ee21bde
/web
```

注意：容器的名称是唯一的。如果已经命名了一个叫 web 的容器，当你要再次使用 web 这个名称的时候，需要先用 `docker rm` 来删除之前创建的同名容器。

在执行 `docker run` 的时候如果添加 `--rm` 标记，则容器在终止后会立刻删除。注意，`--rm` 和 `-d` 参数不能同时使用。

容器互联

使用 `--link` 参数可以让容器之间安全的进行交互。

下面先创建一个新的数据库容器。

```
$ sudo docker run -d --name db training/postgres
```

删除之前创建的 web 容器

```
$ docker rm -f web
```

然后创建一个新的 web 容器，并将它连接到 db 容器

```
$ sudo docker run -d -P --name web --link db:db training/webapp python app.py
```

此时，db 容器和 web 容器建立互联关系。

--link 参数的格式为 --link name:alias，其中 name 是要链接的容器的名称，alias 是这个连接的别名。

使用 docker ps 来查看容器的连接

```
$ docker ps
CONTAINER ID   IMAGE                                COMMAND                  CREATED        STATUS
349169744e49   training/postgres:latest            su postgres -c '/usr    About a minute ago   Up About
aed84ee21bde   training/webapp:latest              python app.py           16 hours ago     Up 2 mi
```

可以看到自定义命名的容器，db 和 web，db 容器的 names 列有 db 也有 web/db。这表示 web 容器链接到 db 容器，web 容器将被允许访问 db 容器的信息。

Docker 在两个互联的容器之间创建了一个安全隧道，而且不用映射它们的端口到宿主主机上。在启动 db 容器的时候并没有使用 -p 和 -P 标记，从而避免了暴露数据库端口到外部网络上。

Docker 通过 2 种方式为容器公开连接信息：

- 环境变量
- 更新 /etc/hosts 文件

使用 env 命令来查看 web 容器的环境变量

```
$ sudo docker run --rm --name web2 --link db:db training/webapp env
. . .
DB_NAME=/web2/db
DB_PORT=tcp://172.17.0.5:5432
DB_PORT_5000_TCP=tcp://172.17.0.5:5432
DB_PORT_5000_TCP_PROTO=tcp
DB_PORT_5000_TCP_PORT=5432
DB_PORT_5000_TCP_ADDR=172.17.0.5
. . .
```

其中 DB_ 开头的环境变量是供 web 容器连接 db 容器使用，前缀采用大写的连接别名。

除了环境变量，Docker 还添加 host 信息到父容器的 /etc/hosts 的文件。下面是父容器 web 的 hosts 文

件

```
$ sudo docker run -t -i --rm --link db:db training/webapp /bin/bash
root@aed84ee21bde:/opt/webapp# cat /etc/hosts
172.17.0.7  aed84ee21bde
. . .
172.17.0.5  db
```

这里有 2 个 hosts，第一个是 web 容器，web 容器用 id 作为他的主机名，第二个是 db 容器的 ip 和主机名。可以在 web 容器中安装 ping 命令来测试跟db容器的连通。

```
root@aed84ee21bde:/opt/webapp# apt-get install -yqq inetutils-ping
root@aed84ee21bde:/opt/webapp# ping db
PING db (172.17.0.5): 48 data bytes
56 bytes from 172.17.0.5: icmp_seq=0 ttl=64 time=0.267 ms
56 bytes from 172.17.0.5: icmp_seq=1 ttl=64 time=0.250 ms
56 bytes from 172.17.0.5: icmp_seq=2 ttl=64 time=0.256 ms
```

用 ping 来测试db容器，它会解析成 172.17.0.5。*注意：官方的 ubuntu 镜像默认没有安装 ping，需要自行安装。

用户可以链接多个父容器到子容器，比如可以链接多个 web 到 db 容器上。

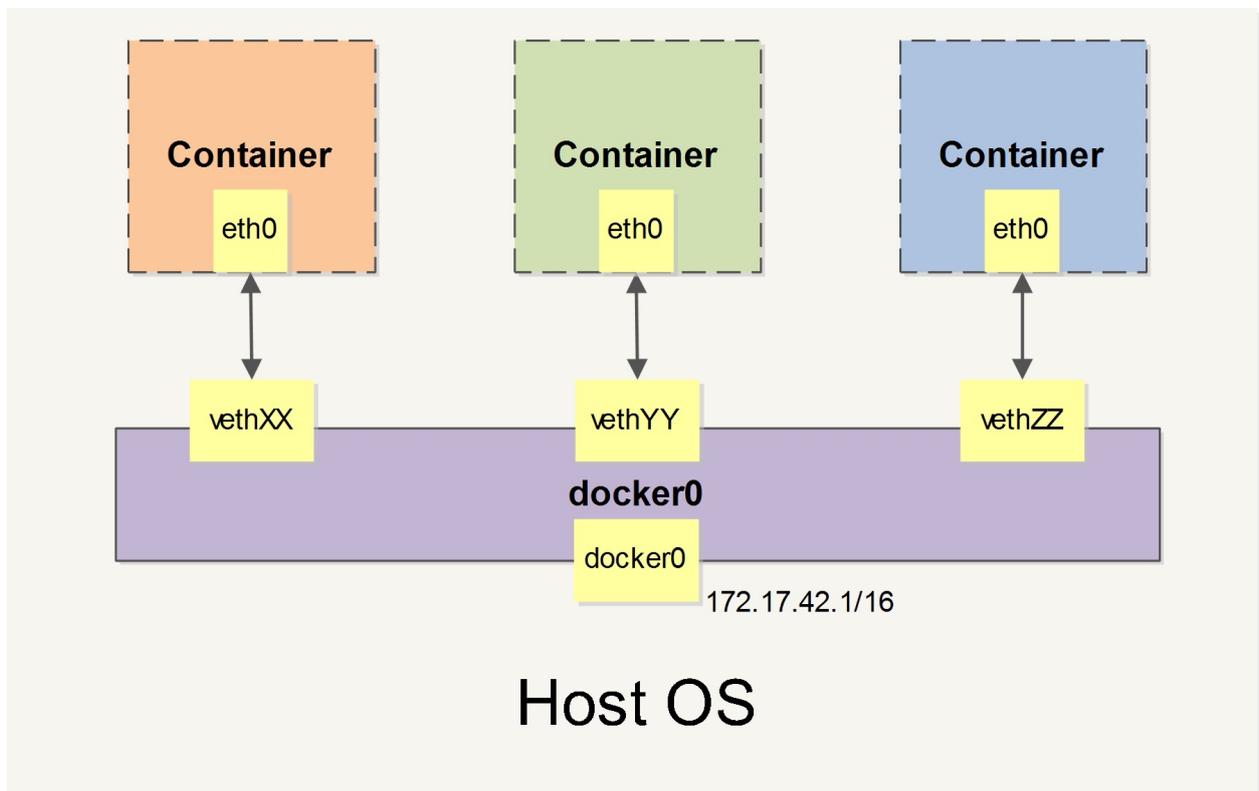
高级网络配置

本章将介绍 Docker 的一些高级网络配置和选项。

当 Docker 启动时，会自动在主机上创建一个 `docker0` 虚拟网桥，实际上是 Linux 的一个 bridge，可以理解为一个软件交换机。它会在挂载到它的网口之间进行转发。

同时，Docker 随机分配一个本地未占用的私有网段（在 RFC1918 中定义）中的一个地址给 `docker0` 接口。比如典型的 `172.17.42.1`，掩码为 `255.255.0.0`。此后启动的容器内的网口也会自动分配一个同一网段（`172.17.0.0/16`）的地址。

当创建一个 Docker 容器的时候，同时会创建了一对 `veth pair` 接口（当数据包发送到一个接口时，另外一个接口也可以收到相同的数据包）。这对接口一端在容器内，即 `eth0`；另一端在本地并被挂载到 `docker0` 网桥，名称以 `veth` 开头（例如 `vethAQI2QT`）。通过这种方式，主机可以跟容器通信，容器之间也可以相互通信。Docker 就创建了在主机和所有容器之间一个虚拟共享网络。



接下来的部分将介绍在一些场景中，Docker 所有的网络定制配置。以及通过 Linux 命令来调整、补充、甚至替换 Docker 默认的网络配置。

快速配置指南

下面是一个跟 Docker 网络相关的命令列表。

其中有些命令选项只有在 Docker 服务启动的时候才能配置，而且不能马上生效。

- `-b BRIDGE` or `--bridge=BRIDGE` --指定容器挂载的网桥
- `--bip=CIDR` --定制 docker0 的掩码
- `-H SOCKET...` or `--host=SOCKET...` --Docker 服务端接收命令的通道
- `--icc=true|false` --是否支持容器之间进行通信
- `--ip-forward=true|false` --请看下文容器之间的通信
- `--iptables=true|false` --禁止 Docker 添加 iptables 规则
- `--mtu=BYTES` --容器网络中的 MTU

下面2个命令选项既可以在启动服务时指定，也可以 Docker 容器启动（`docker run`）时候指定。在 Docker 服务启动的时候指定则会成为默认值，后面执行 `docker run` 时可以覆盖设置的默认值。

- `--dns=IP_ADDRESS...` --使用指定的DNS服务器
- `--dns-search=DOMAIN...` --指定DNS搜索域

最后这些选项只有在 `docker run` 执行时使用，因为它是针对容器的特性内容。

- `-h HOSTNAME` or `--hostname=HOSTNAME` --配置容器主机名
- `--link=CONTAINER_NAME:ALIAS` --添加到另一个容器的连接
- `--net=bridge|none|container:NAME_or_ID|host` --配置容器的桥接模式
- `-p SPEC` or `--publish=SPEC` --映射容器端口到宿主主机
- `-P` or `--publish-all=true|false` --映射容器所有端口到宿主主机

配置 DNS

Docker 没有为每个容器专门定制镜像，那么怎么自定义配置容器的主机名和 DNS 配置呢？秘诀就是它利用虚拟文件来挂载到来容器的 3 个相关配置文件。

在容器中使用 `mount` 命令可以看到挂载信息：

```
$ mount
...
/dev/disk/by-uuid/1fec...ebdf on /etc/hostname type ext4 ...
/dev/disk/by-uuid/1fec...ebdf on /etc/hosts type ext4 ...
tmpfs on /etc/resolv.conf type tmpfs ...
...
```

这种机制可以让宿主主机 DNS 信息发生更新后，所有 Docker 容器的 dns 配置通过 `/etc/resolv.conf` 文件立刻得到更新。

如果用户想要手动指定容器的配置，可以利用下面的选项。

`-h HOSTNAME` or `--hostname=HOSTNAME` 设定容器的主机名，它会被写到容器内的 `/etc/hostname` 和 `/etc/hosts`。但它在容器外部看不到，既不会在 `docker ps` 中显示，也不会其他的容器的 `/etc/hosts` 看到。

`--link=CONTAINER_NAME:ALIAS` 选项会在创建容器的时候，添加一个其他容器的主机名到 `/etc/hosts` 文件中，让新容器的进程可以使用主机名 ALIAS 就可以连接它。

`--dns=IP_ADDRESS` 添加 DNS 服务器到容器的 `/etc/resolv.conf` 中，让容器用这个服务器来解析所有不在 `/etc/hosts` 中的主机名。

`--dns-search=DOMAIN` 设定容器的搜索域，当设定搜索域为 `.example.com` 时，在搜索一个名为 host 的主机时，DNS 不仅搜索 host，还会搜索 `host.example.com`。注意：如果没有上述最后 2 个选项，Docker 会默认用主机上的 `/etc/resolv.conf` 来配置容器。

容器访问控制

容器的访问控制，主要通过 Linux 上的 `iptables` 防火墙来进行管理和实现。`iptables` 是 Linux 上默认的防火墙软件，在大部分发行版中都自带。

容器访问外部网络

容器要想访问外部网络，需要本地系统的转发支持。在 Linux 系统中，检查转发是否打开。

```
$sysctl net.ipv4.ip_forward
net.ipv4.ip_forward = 1
```

如果为 0，说明没有开启转发，则需要手动打开。

```
$sysctl -w net.ipv4.ip_forward=1
```

如果在启动 Docker 服务的时候设定 `--ip-forward=true`，Docker 就会自动设定系统的 `ip_forward` 参数为 1。

容器之间访问

容器之间相互访问，需要两方面的支持。

- 容器的网络拓扑是否已经互联。默认情况下，所有容器都会被连接到 `docker0` 网桥上。
- 本地系统的防火墙软件 -- `iptables` 是否允许通过。

访问所有端口

当启动 Docker 服务时候，默认会添加一条转发策略到 `iptables` 的 `FORWARD` 链上。策略为通过（`ACCEPT`）还是禁止（`DROP`）取决于配置 `--icc=true`（缺省值）还是 `--icc=false`。当然，如果手动指定 `--iptables=false` 则不会添加 `iptables` 规则。

可见，默认情况下，不同容器之间是允许网络互通的。如果为了安全考虑，可以在 `/etc/default/docker` 文件中配置 `DOCKER_OPTS="--icc=false` 来禁止它。

访问指定端口

在通过 `-icc=false` 关闭网络访问后，还可以通过 `--link=CONTAINER_NAME:ALIAS` 选项来访问容器的开放端口。

例如，在启动 Docker 服务时，可以同时使用 `icc=false --iptables=true` 参数来关闭允许相互的网络访问，并让 Docker 可以修改系统中的 `iptables` 规则。

此时，系统中的 `iptables` 规则可能是类似

```
$ sudo iptables -nL
...
Chain FORWARD (policy ACCEPT)
target     prot opt source                destination
DROP      all  --  0.0.0.0/0             0.0.0.0/0
...
```

之后，启动容器（`docker run`）时使用 `--link=CONTAINER_NAME:ALIAS` 选项。Docker 会在 iptable 中为两个容器分别添加一条 `ACCEPT` 规则，允许相互访问开放的端口（取决于 Dockerfile 中的 `EXPOSE` 行）。

当添加了 `--link=CONTAINER_NAME:ALIAS` 选项后，添加了 `iptables` 规则。

```
$ sudo iptables -nL
...
Chain FORWARD (policy ACCEPT)
target     prot opt source                destination          tcp spt:80
ACCEPT     tcp  --  172.17.0.2            172.17.0.3
ACCEPT     tcp  --  172.17.0.3            172.17.0.2          tcp dpt:80
DROP      all  --  0.0.0.0/0            0.0.0.0/0
```

注意：`--link=CONTAINER_NAME:ALIAS` 中的 `CONTAINER_NAME` 目前必须是 Docker 分配的名字，或使用 `--name` 参数指定的名字。主机名则不会被识别。

映射容器端口到宿主主机的实现

默认情况下，容器可以主动访问到外部网络的连接，但是外部网络无法访问到容器。

容器访问外部实现

容器所有到外部网络的连接，源地址都会被NAT成本地系统的IP地址。这是使用 `iptables` 的源地址伪装操作实现的。

查看主机的 NAT 规则。

```
$ sudo iptables -t nat -nL
...
Chain POSTROUTING (policy ACCEPT)
target     prot opt source                destination
MASQUERADE all  --  172.17.0.0/16          !172.17.0.0/16
...
```

其中，上述规则将所有源地址在 `172.17.0.0/16` 网段，目标地址为其他网段（外部网络）的流量动态伪装为从系统网卡发出。MASQUERADE 跟传统 SNAT 的好处是它能动态从网卡获取地址。

外部访问容器实现

容器允许外部访问，可以在 `docker run` 时候通过 `-p` 或 `-P` 参数来启用。

不管用那种办法，其实也是在本地的 `iptables` 的 `nat` 表中添加相应的规则。

使用 `-P` 时：

```
$ iptables -t nat -nL
...
Chain DOCKER (2 references)
target     prot opt source                destination
DNAT       tcp  --  0.0.0.0/0             0.0.0.0/0             tcp dpt:49153 to:172.17.0.2
```

使用 `-p 80:80` 时：

```
$ iptables -t nat -nL
Chain DOCKER (2 references)
target     prot opt source                destination
DNAT       tcp  --  0.0.0.0/0             0.0.0.0/0             tcp dpt:80 to:172.17.0.2:80
```

注意：

- 这里的规则映射了 `0.0.0.0`，意味着将接受主机来自所有接口的流量。用户可以通过 `-p`

`IP:host_port:container_port` 或 `-p IP::port` 来指定允许访问容器的主机上的 IP、接口等，以制定更严格的规则。

- 如果希望永久绑定到某个固定的 IP 地址，可以在 Docker 配置文件 `/etc/default/docker` 中指定 `DOCKER_OPTS="--ip=IP_ADDRESS"`，之后重启 Docker 服务即可生效。

配置 docker0 网桥

Docker 服务默认会创建一个 `docker0` 网桥（其上有一个 `docker0` 内部接口），它在内核层连通了其他的物理或虚拟网卡，这就将所有容器和本地主机都放到同一个物理网络。

Docker 默认指定了 `docker0` 接口的 IP 地址和子网掩码，让主机和容器之间可以通过网桥相互通信，它还给出了 MTU（接口允许接收的最大传输单元），通常是 1500 Bytes，或宿主主机网络路由上支持的默认值。这些值都可以在服务启动的时候进行配置。

- `--bip=CIDR` -- IP 地址加掩码格式，例如 192.168.1.5/24
- `--mtu=BYTES` -- 覆盖默认的 Docker mtu 配置

也可以在配置文件中配置 `DOCKER_OPTS`，然后重启服务。由于目前 Docker 网桥是 Linux 网桥，用户可以使用 `brctl show` 来查看网桥和端口连接信息。

```
$ sudo brctl show
bridge name      bridge id                STP enabled    interfaces
docker0          8000.3a1d7362b4ee       no             veth65f9
                                     vethdda6
```

*注：`brctl` 命令在 Debian、Ubuntu 中可以使用 `sudo apt-get install bridge-utils` 来安装。

每次创建一个新容器的时候，Docker 从可用的地址段中选择一个空闲的 IP 地址分配给容器的 `eth0` 端口。使用本地主机上 `docker0` 接口的 IP 作为所有容器的默认网关。

```
$ sudo docker run -i -t --rm base /bin/bash
$ ip addr show eth0
24: eth0: <BROADCAST,UP,LOWER_UP> mtu 1500 qdisc pfifo_fast state UP group default qlen 1
    link/ether 32:6f:e0:35:57:91 brd ff:ff:ff:ff:ff:ff
    inet 172.17.0.3/16 scope global eth0
        valid_lft forever preferred_lft forever
    inet6 fe80::306f:e0ff:fe35:5791/64 scope link
        valid_lft forever preferred_lft forever
$ ip route
default via 172.17.42.1 dev eth0
172.17.0.0/16 dev eth0 proto kernel scope link src 172.17.0.3
$ exit
```

自定义网桥

除了默认的 `docker0` 网桥，用户也可以指定网桥来连接各个容器。

在启动 Docker 服务的时候，使用 `-b BRIDGE` 或 `--bridge=BRIDGE` 来指定使用的网桥。

如果服务已经运行，那需要先停止服务，并删除旧的网桥。

```
$ sudo service docker stop
$ sudo ip link set dev docker0 down
$ sudo brctl delbr docker0
```

然后创建一个网桥 `bridge0`。

```
$ sudo brctl addbr bridge0
$ sudo ip addr add 192.168.5.1/24 dev bridge0
$ sudo ip link set dev bridge0 up
```

查看确认网桥创建并启动。

```
$ ip addr show bridge0
4: bridge0: <BROADCAST,MULTICAST> mtu 1500 qdisc noop state UP group default
    link/ether 66:38:d0:0d:76:18 brd ff:ff:ff:ff:ff:ff
    inet 192.168.5.1/24 scope global bridge0
        valid_lft forever preferred_lft forever
```

配置 Docker 服务，默认桥接到创建的网桥上。

```
$ echo 'DOCKER_OPTS="-b=bridge0"' >> /etc/default/docker
$ sudo service docker start
```

启动 Docker 服务。新建一个容器，可以看到它已经桥接到了 `bridge0` 上。

可以继续用 `brctl show` 命令查看桥接的信息。另外，在容器中使用 `ip addr` 和 `ip route` 命令来查看 IP 地址配置和路由信息。

工具和示例

在介绍自定义网络拓扑之前，你可能会对一些外部工具和例子感兴趣：

pipework

Jérôme Petazzoni 编写了一个叫 [pipework](#) 的 shell 脚本，可以帮助用户在比较复杂的场景中完成容器的连接。

playground

Brandon Rhodes 创建了一个提供完整的 Docker 容器网络拓扑管理的 [Python库](#)，包括路由、NAT 防火墙；以及一些提供 HTTP, SMTP, POP, IMAP, Telnet, SSH, FTP 的服务器。

编辑网络配置文件

Docker 1.2.0 开始支持在运行中的容器里编辑 `/etc/hosts`，`/etc/hostname` 和 `/etc/resolve.conf` 文件。

但是这些修改是临时的，只在运行的容器中保留，容器终止或重启后并不会被保存下来。也不会被 `docker commit` 提交。

示例：创建一个点到点连接

默认情况下，Docker 会将所有容器连接到由 `docker0` 提供的虚拟子网中。

用户有时候需要两个容器之间可以直连通信，而不用通过主机网桥进行桥接。

解决办法很简单：创建一对 `peer` 接口，分别放到两个容器中，配置成点到点链路类型即可。

首先启动 2 个容器：

```
$ sudo docker run -i -t --rm --net=none base /bin/bash
root@1f1f4c1f931a:/#
$ sudo docker run -i -t --rm --net=none base /bin/bash
root@12e343489d2f:/#
```

找到进程号，然后创建网络名字空间的跟踪文件。

```
$ sudo docker inspect -f '{{.State.Pid}}' 1f1f4c1f931a
2989
$ sudo docker inspect -f '{{.State.Pid}}' 12e343489d2f
3004
$ sudo mkdir -p /var/run/netns
$ sudo ln -s /proc/2989/ns/net /var/run/netns/2989
$ sudo ln -s /proc/3004/ns/net /var/run/netns/3004
```

创建一对 `peer` 接口，然后配置路由

```
$ sudo ip link add A type veth peer name B

$ sudo ip link set A netns 2989
$ sudo ip netns exec 2989 ip addr add 10.1.1.1/32 dev A
$ sudo ip netns exec 2989 ip link set A up
$ sudo ip netns exec 2989 ip route add 10.1.1.2/32 dev A

$ sudo ip link set B netns 3004
$ sudo ip netns exec 3004 ip addr add 10.1.1.2/32 dev B
$ sudo ip netns exec 3004 ip link set B up
$ sudo ip netns exec 3004 ip route add 10.1.1.1/32 dev B
```

现在这 2 个容器就可以相互 ping 通，并成功建立连接。点到点链路不需要子网和子网掩码。

此外，也可以不指定 `--net=none` 来创建点到点链路。这样容器还可以通过原先的网络来通信。

利用类似的办法，可以创建一个只跟主机通信的容器。但是一般情况下，更推荐使用 `--icc=false` 来关闭容器之间的通信。

实战案例

介绍一些典型的应用场景和案例。

使用 Supervisor 来管理进程

Docker 容器在启动的时候开启单个进程，比如，一个 ssh 或者 apache 的 daemon 服务。但我们经常需要在同一台机器上开启多个服务，这可以有多种方法，最简单的就是把多个启动命令放到一个启动脚本里面，启动的时候直接启动这个脚本，另外就是安装进程管理工具。

本节将使用进程管理工具 supervisor 来管理容器中的多个进程。使用 Supervisor 可以更好的控制、管理、重启我们希望运行的进程。在这里我们演示一下如何同时使用 ssh 和 apache 服务。

配置

首先创建一个 Dockerfile，内容和各部分的解释如下。

```
FROM ubuntu:13.04
MAINTAINER examples@docker.com
RUN echo "deb http://archive.ubuntu.com/ubuntu precise main universe" > /etc/apt/sources.
RUN apt-get update
RUN apt-get upgrade -y
```

安装 ssh、apache 和 supervisor

```
RUN apt-get install -y openssh-server apache2 supervisor
RUN mkdir -p /var/run/sshd
RUN mkdir -p /var/log/supervisor
```

这里安装 3 个软件，还创建了 2 个 ssh 和 supervisor 服务正常运行所需要的目录。

```
COPY supervisord.conf /etc/supervisor/conf.d/supervisord.conf
```

添加 supervisord 的配置文件，并复制配置文件到对应目录下面。

```
EXPOSE 22 80
CMD ["/usr/bin/supervisord"]
```

这里我们映射了 22 和 80 端口，使用 supervisord 的可执行路径启动服务。

supervisor 配置文件内容

```
[supervisord]
nodaemon=true
[program:sshd]
command=/usr/sbin/sshd -D
```

```
[program:apache2]
command=/bin/bash -c "source /etc/apache2/envvars && exec /usr/sbin/apache2 -DFOREGROUND"
```

配置文件包含目录和进程，第一段 `supervisord` 配置软件本身，使用 `nodaemon` 参数来运行。第二段包含要控制的 2 个服务。每一段包含一个服务的目录和启动这个服务的命令。

使用方法

创建镜像。

```
$ sudo docker build -t test/supervisord .
```

启动 `supervisor` 容器。

```
$ sudo docker run -p 22 -p 80 -t -i test/supervisord
2013-11-25 18:53:22,312 CRIT Supervisor running as root (no user in config file)
2013-11-25 18:53:22,312 WARN Included extra file "/etc/supervisor/conf.d/supervisord.conf"
2013-11-25 18:53:22,342 INFO supervisord started with pid 1
2013-11-25 18:53:23,346 INFO spawned: 'sshd' with pid 6
2013-11-25 18:53:23,349 INFO spawned: 'apache2' with pid 7
```

使用 `docker run` 来启动我们创建的容器。使用多个 `-p` 来映射多个端口，这样我们就能同时访问 `ssh` 和 `apache` 服务了。

可以使用这个方法创建一个只有 `ssh` 服务的基础镜像，之后创建镜像可以使用这个镜像为基础来创建

创建 tomcat/weblogic 集群

安装 tomcat 镜像

准备好需要的 jdk、tomcat 等软件放到 home 目录下面，启动一个容器

```
docker run -t -i -v /home:/opt/data --name mk_tomcat ubuntu /bin/bash
```

这条命令挂载本地 home 目录到容器的 /opt/data 目录，容器内目录若不存在，则会自动创建。接下来就是 tomcat 的基本配置，jdk 环境变量设置好之后，将 tomcat 程序放到 /opt/apache-tomcat 下面 编辑 /etc/supervisor/conf.d/supervisor.conf 文件，添加 tomcat 项

```
[supervisord]
nodaemon=true

[program:tomcat]
command=/opt/apache-tomcat/bin/startup.sh

[program:sshd]
command=/usr/sbin/sshd -D
```

```
docker commit ac6474aeb31d tomcat
```

新建 tomcat 文件夹，新建 Dockerfile。

```
FROM mk_tomcat
EXPOSE 22 8080
CMD ["/usr/bin/supervisord"]
```

根据 Dockerfile 创建镜像。

```
docker build tomcat tomcat
```

安装 weblogic 镜像

步骤和 tomcat 基本一致，这里贴一下配置文件

```
supervisor.conf
[supervisord]
nodaemon=true

[program:weblogic]
command=/opt/Middleware/user_projects/domains/base_domain/bin/startWebLogic.sh
```

```
[program:sshd]
command=/usr/sbin/sshd -D
dockerfile
FROM weblogic
EXPOSE 22 7001
CMD ["/usr/bin/supervisord"]
```

tomcat/weblogic 镜像的使用

存储的使用

在启动的时候，使用 `-v` 参数

```
-v, --volume=[]          Bind mount a volume (e.g. from the host: -v /host:/container,
```

将本地磁盘映射到容器内部，它在主机和容器之间是实时变化的，所以我们更新程序、上传代码只需要更新物理主机的目录就可以了

tomcat 和 weblogic 集群的实现

tomcat 只要开启多个容器即可

```
docker run -d -v -p 204:22 -p 7003:8080 -v /home/data:/opt/data --name tm1 tomcat /usr/bi
docker run -d -v -p 205:22 -p 7004:8080 -v /home/data:/opt/data --name tm2 tomcat /usr/bi
docker run -d -v -p 206:22 -p 7005:8080 -v /home/data:/opt/data --name tm3 tomcat /usr/bi
```

这里说一下 weblogic 的配置，大家知道 weblogic 有一个域的概念。如果要使用常规的 administrator +node 的方式部署，就需要在 supervisord 中分别写出 administrartor server 和 node server 的启动脚本，这样做的优点是：

- 可以使用 weblogic 的集群，同步等概念
- 部署一个集群应用程序，只需要安装一次应用到集群上即可

缺点是：

- Docker 配置复杂了
- 没办法自动扩展集群的计算容量，如需添加节点，需要在 administrator 上先创建节点，然后再配置新的容器 supervisor 启动脚本，然后再启动容器

另外种方法是把所有的程序都安装在 adminiserver 上面，需要扩展的时候，启动多个节点即可，它的优点和缺点和上一种方法恰恰相反。（建议使用这种方式来部署开发和测试环境）

```
docker run -d -v -p 204:22 -p 7001:7001 -v /home/data:/opt/data --name node1 weblogic /us
docker run -d -v -p 205:22 -p 7002:7001 -v /home/data:/opt/data --name node2 weblogic /us
docker run -d -v -p 206:22 -p 7003:7001 -v /home/data:/opt/data --name node3 weblogic /us
```



这样在前端使用 nginx 来做负载均衡就可以完成配置了

多台物理主机之间的容器互联（暴露容器到真实网络中）

Docker 默认的桥接网卡是 docker0。它只会在本机桥接所有的容器网卡，举例来说容器的虚拟网卡在主机上看一般叫做 veth* 而 Docker 只是把所有这些网卡桥接在一起，如下：

```
[root@opnvz ~]# brctl show
bridge name      bridge id                STP enabled    interfaces
docker0          8000.56847afe9799        no             veth0889
                                     veth3c7b
                                     veth4061
```

在容器中看到的地址一般是像下面这样的地址：

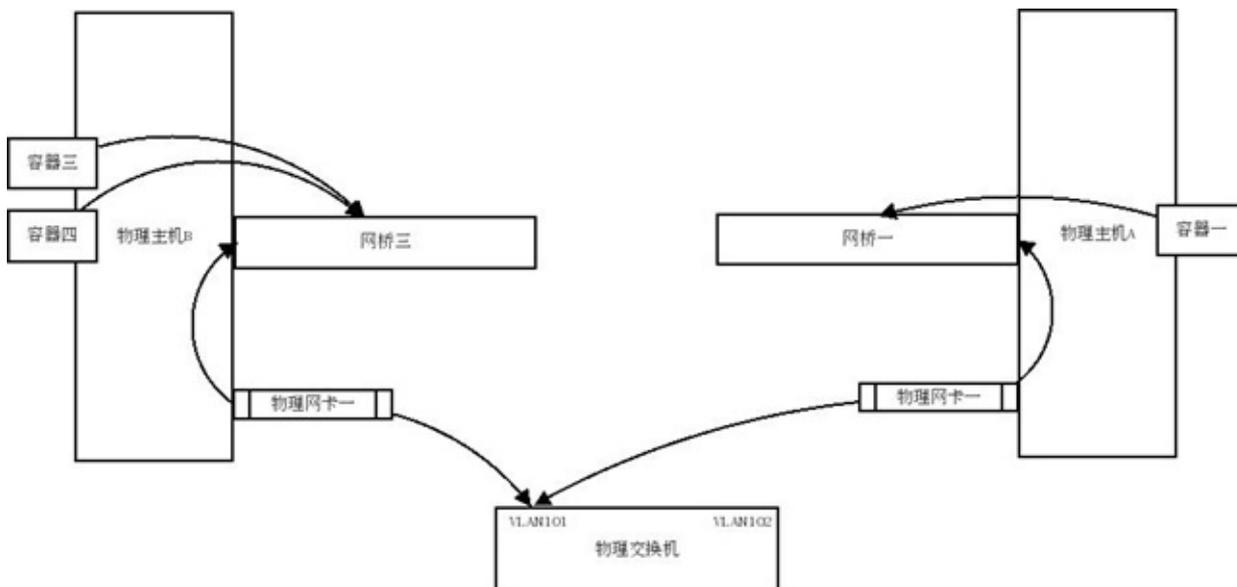
```
root@ac6474aeb31d:~# ip a
1: lo: <LOOPBACK,UP,LOWER_UP> mtu 1500 qdisc noqueue state UNKNOWN group default
   link/loopback 00:00:00:00:00:00 brd 00:00:00:00:00:00
   inet 127.0.0.1/8 scope host lo
       valid_lft forever preferred_lft forever
   inet6 ::1/128 scope host
       valid_lft forever preferred_lft forever
11: eth0: <BROADCAST,UP,LOWER_UP> mtu 1500 qdisc pfifo_fast state UP group default qlen 1
   link/ether 4a:7d:68:da:09:cf brd ff:ff:ff:ff:ff:ff
   inet 172.17.0.3/16 scope global eth0
       valid_lft forever preferred_lft forever
   inet6 fe80::487d:68ff:feda:9cf/64 scope link
       valid_lft forever preferred_lft forever
```

这样就可以把这个网络看成是一个私有的网络，通过 nat 连接外网，如果要让外网连接到容器中，就需要做端口映射，即 -p 参数。

如果在企业内部应用，或者做多个物理主机的集群，可能需要将多个物理主机的容器组到一个物理网络中来，那么就需要将这个网桥桥接到我们指定的网卡上。

拓扑图

主机 A 和主机 B 的网卡一都连着物理交换机的同一个 vlan 101,这样网桥一和网桥三就相当于在同一个物理网络中了，而容器一、容器三、容器四也在同一物理网络中了，他们之间可以相互通信，而且可以跟同一 vlan 中的其他物理机器互联。



ubuntu 示例

下面以 ubuntu 为例创建多个主机的容器联网: 创建自己的网桥,编辑 /etc/network/interface 文件

```

auto br0
iface br0 inet static
address 192.168.7.31
netmask 255.255.240.0
gateway 192.168.7.254
bridge_ports em1
bridge_stp off
dns-nameservers 8.8.8.8 192.168.6.1
    
```

将 Docker 的默认网桥绑定到这个新建的 br0 上面, 这样就将这台机器上容器绑定到 em1 这个网卡所对应的物理网络上了。

ubuntu 修改 /etc/default/docker 文件, 添加最后一行内容

```

# Docker Upstart and SysVinit configuration file
# Customize location of Docker binary (especially for development testing).
#DOCKER="/usr/local/bin/docker"
# Use DOCKER_OPTS to modify the daemon startup options.
#DOCKER_OPTS="--dns 8.8.8.8 --dns 8.8.4.4"

# If you need Docker to use an HTTP proxy, it can also be specified here.
#export http_proxy="http://127.0.0.1:3128/"

# This is also a handy place to tweak where Docker's temporary files go.
#export TMPDIR="/mnt/bigdrive/docker-tmp"

DOCKER_OPTS="-b=br0"
    
```

在启动 Docker 的时候使用 -b 参数 将容器绑定到物理网络上。重启 Docker 服务后, 再进入容器可以看到它已经绑定到你的物理网络上了。

```

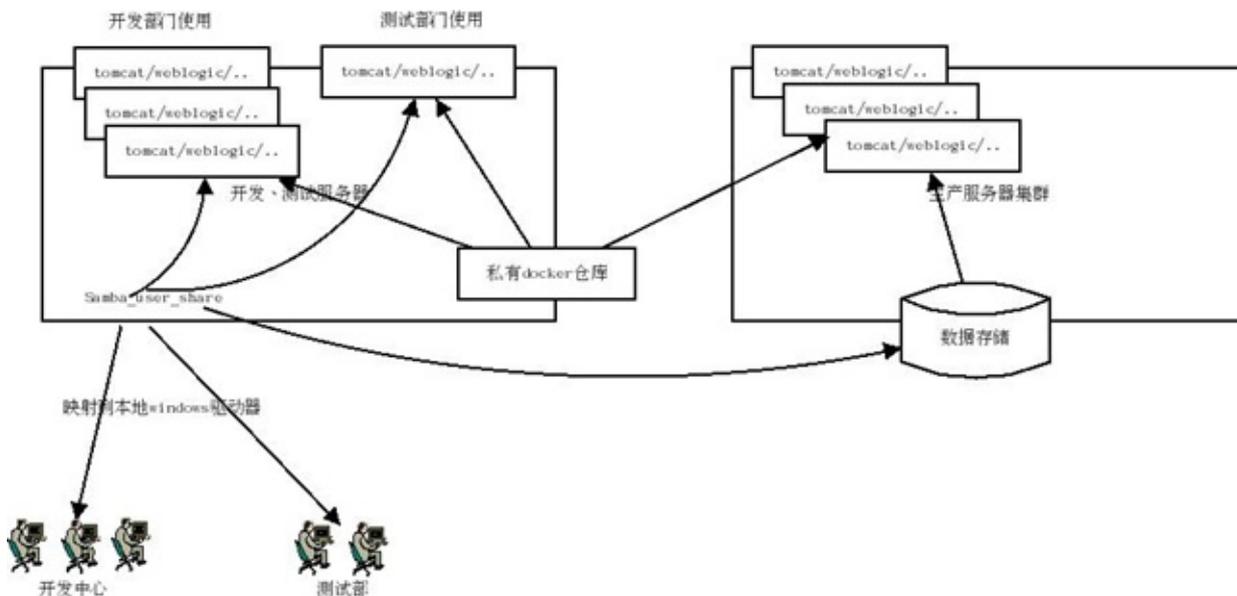
root@ubuntudocker:~# docker ps
CONTAINER ID        IMAGE               COMMAND             CREATED             STATUS
58b043aa05eb       desk_hz:v1         "/startup.sh"      5 days ago         Up 2 seco
root@ubuntudocker:~# brctl show
bridge name      bridge id          STP enabled        interfaces
br0              8000.7e6e617c8d53 no                  em1
                  vethe6e5
    
```

这样就直接把容器暴露到物理网络上了，多台物理主机的容器也可以相互联网了。需要注意的是，这就需要自己来保证容器的网络安全了。

标准化开发测试和生产环境

对于大部分企业来说，搭建 PaaS 既没有那个精力，也没那个必要，用 Docker 做个人的 sandbox 用处又小了点。

可以用 Docker 来标准化开发、测试、生产环境。



Docker 占用资源小，在一台 E5 128 G 内存的服务器上部署 100 个容器都绰绰有余，可以单独抽一个容器或者直接在宿主物理主机上部署 samba，利用 samba 的 home 分享方案将每个用户的 home 目录映射到开发中心和测试部门的 Windows 机器上。

针对某个项目组，由架构师搭建好一个标准的容器环境供项目组 and 测试部门使用，每个开发工程师可以拥有自己单独的容器，通过 `docker run -v` 将用户的 home 目录映射到容器中。需要提交测试时，只需要将代码移交给测试部门，然后分配一个容器使用 `-v` 加载测试部门的 home 目录启动即可。这样，在公司内部的开发、测试基本就统一了，不会出现开发部门提交的代码，测试部门部署不了的问题。

测试部门发布测试通过的报告后，架构师再一次检测容器环境，就可以直接交由部署工程师将代码和容器分别部署到生产环境中了。这种方式的部署横向性能的扩展性也极好。

安全

评估 Docker 的安全性时，主要考虑三个方面：

- 由内核的名字空间和控制组机制提供的容器内在安全
- Docker程序（特别是服务端）本身的抗攻击性
- 内核安全性的加强机制对容器安全性的影响

内核名字空间

Docker 容器和 LXC 容器很相似，所提供的安全特性也差不多。当用 `docker run` 启动一个容器时，在后台 Docker 为容器创建了一个独立的名字空间和控制组集合。

名字空间提供了最基础也是最直接的隔离，在容器中运行的进程不会被运行在主机上的进程和其它容器发现和作用。

每个容器都有自己独有的网络栈，意味着它们不能访问其他容器的 sockets 或接口。不过，如果主机系统上做了相应的设置，容器可以像跟主机交互一样的和其他容器交互。当指定公共端口或使用 links 来连接 2 个容器时，容器就可以相互通信了（可以根据配置来限制通信的策略）。

从网络架构的角度来看，所有的容器通过本地主机的网桥接口相互通信，就像物理机器通过物理交换机通信一样。

那么，内核中实现名字空间和私有网络的代码是否足够成熟？

内核名字空间从 2.6.15 版本（2008 年 7 月发布）之后被引入，数年间，这些机制的可靠性在诸多大型生产系统中被实践验证。

实际上，名字空间的想法和设计提出的时间要更早，最初是为了在内核中引入一种机制来实现 OpenVZ 的特性。而 OpenVZ 项目早在 2005 年就发布了，其设计和实现都已经十分成熟。

控制组

控制组是 Linux 容器机制的另外一个关键组件，负责实现资源的审计和限制。

它提供了很多有用的特性；以及确保各个容器可以公平地分享主机的内存、CPU、磁盘 IO 等资源；当然，更重要的是，控制组确保了当容器内的资源使用产生压力时不会连累主机系统。

尽管控制组不负责隔离容器之间相互访问、处理数据和进程，它在防止拒绝服务（DDOS）攻击方面是必不可少的。尤其是在多用户的平台（比如公有或私有的 PaaS）上，控制组十分重要。例如，当某些应用程序表现异常的时候，可以保证一致地正常运行和性能。

控制组机制始于 2006 年，内核从 2.6.24 版本开始被引入。

Docker服务端的防护

运行一个容器或应用程序的核心是通过 Docker 服务端。Docker 服务的运行目前需要 root 权限，因此其安全性十分关键。

首先，确保只有可信的用户才可以访问 Docker 服务。Docker 允许用户在主机和容器间共享文件夹，同时不需要限制容器的访问权限，这就容易让容器突破资源限制。例如，恶意用户启动容器的时候将主机的根目录 / 映射到容器的 /host 目录中，那么容器理论上就可以对主机的文件系统进行任意修改了。这听起来很疯狂？但是事实上几乎所有虚拟化系统都允许类似的资源共享，而没法禁止用户共享主机根文件系统到虚拟机系统。

这将会造成很严重的安全后果。因此，当提供容器创建服务时（例如通过一个 web 服务器），要更加注意进行参数的安全检查，防止恶意的用户用特定参数来创建一些破坏性的容器

为了加强对服务端的保护，Docker 的 REST API（客户端用来跟服务端通信）在 0.5.2 之后使用本地的 Unix 套接字机制替代了原先绑定在 127.0.0.1 上的 TCP 套接字，因为后者容易遭受跨站脚本攻击。现在用户使用 Unix 权限检查来加强套接字的访问安全。

用户仍可以利用 HTTP 提供 REST API 访问。建议使用安全机制，确保只有可信的网络或 VPN，或证书保护机制（例如受保护的 stunnel 和 ssl 认证）下的访问可以进行。此外，还可以使用 HTTPS 和证书来加强保护。

最近改进的 Linux 名字空间机制将可以实现使用非 root 用户来运行全功能的容器。这将从根本上解决了容器和主机之间共享文件系统而引起的安全问题。

终极目标是改进 2 个重要的安全特性：

- 将容器的 root 用户映射到本地主机上的非 root 用户，减轻容器和主机之间因权限提升而引起的安全问题；
- 允许 Docker 服务端在非 root 权限下运行，利用安全可靠的子进程来代理执行需要特权权限的操作。这些子进程将只允许在限定范围内进行操作，例如仅仅负责虚拟网络设定或文件系统管理、配置操作等。

最后，建议采用专用的服务器来运行 Docker 和相关的管理服务（例如管理服务比如 ssh 监控和进程监控、管理工具 nrpe、collectd 等）。其它的业务服务都放到容器中去运行。

内核能力机制

能力机制（Capability）是 Linux 内核一个强大的特性，可以提供细粒度的权限访问控制。Linux 内核自 2.2 版本起就支持能力机制，它将权限划分为更加细粒度的操作能力，既可以作用在进程上，也可以作用在文件上。

例如，一个 Web 服务进程只需要绑定一个低于 1024 的端口的权限，并不需要 root 权限。那么它只需要被授予 `net_bind_service` 能力即可。此外，还有很多其他的类似能力来避免进程获取 root 权限。

默认情况下，Docker 启动的容器被严格限制只允许使用内核的一部分能力。

使用能力机制对加强 Docker 容器的安全有很多好处。通常，在服务器上会运行一堆需要特权权限的进程，包括有 ssh、cron、syslogd、硬件管理工具模块（例如负载模块）、网络配置工具等等。容器跟这些进程是不同的，因为几乎所有的特权进程都由容器以外的支持系统来进行管理。

- ssh 访问被主机上 ssh 服务来管理；
- cron 通常应该作为用户进程执行，权限交给使用它服务的应用来处理；
- 日志系统可由 Docker 或第三方服务管理；
- 硬件管理无关紧要，容器中也就不需执行 udevd 以及类似服务；
- 网络管理也都在主机上设置，除非特殊需求，容器不需要对网络进行配置。

从上面的例子可以看出，大部分情况下，容器并不需要“真正的” root 权限，容器只需要少数的能力即可。为了加强安全，容器可以禁用一些没必要的权限。

- 完全禁止任何 mount 操作；
- 禁止直接访问本地主机的套接字；
- 禁止访问一些文件系统的操作，比如创建新的设备、修改文件属性等；
- 禁止模块加载。

这样，就算攻击者在容器中取得了 root 权限，也不能获得本地主机的较高权限，能进行的破坏也有限。

默认情况下，Docker 采用 **白名单** 机制，禁用 **必需功能** 之外的其它权限。当然，用户也可以根据自身需求来为 Docker 容器启用额外的权限。

其它安全特性

除了能力机制之外，还可以利用一些现有的安全机制来增强使用 Docker 的安全性，例如 TOMOYO, AppArmor, SELinux, GRSEC 等。

Docker 当前默认只启用了能力机制。用户可以采用多种方案来加强 Docker 主机的安全，例如：

- 在内核中启用 GRSEC 和 PAX，这将增加很多编译和运行时的安全检查；通过地址随机化避免恶意探测等。并且，启用该特性不需要 Docker 进行任何配置。
- 使用一些有增强安全特性的容器模板，比如带 AppArmor 的模板和 Redhat 带 SELinux 策略的模板。这些模板提供了额外的安全特性。
- 用户可以自定义访问控制机制来定制安全策略。

跟其它添加到 Docker 容器的第三方工具一样（比如网络拓扑和文件系统共享），有很多类似的机制，在不改变 Docker 内核情况下就可以加固现有的容器。

总结

总体来看，Docker 容器还是十分安全的，特别是在容器内不使用 root 权限来运行进程的话。

另外，用户可以使用现有工具，比如 Apparmor, SELinux, GRSEC 来增强安全性；甚至自己在内核中实现更复杂的安全机制。

Dockerfile

使用 Dockerfile 可以允许用户创建自定义的镜像。

基本结构

Dockerfile 由一行行命令语句组成，并且支持以 `#` 开头的注释行。

一般的，Dockerfile 分为四部分：基础镜像信息、维护者信息、镜像操作指令和容器启动时执行指令。

例如

```
# This dockerfile uses the ubuntu image
# VERSION 2 - EDITION 1
# Author: docker_user
# Command format: Instruction [arguments / command] ..

# Base image to use, this must be set as the first line
FROM ubuntu

# Maintainer: docker_user <docker_user at email.com> (@docker_user)
MAINTAINER docker_user docker_user@email.com

# Commands to update the image
RUN echo "deb http://archive.ubuntu.com/ubuntu/ raring main universe" >> /etc/apt/sources
RUN apt-get update && apt-get install -y nginx
RUN echo "\ndaemon off;" >> /etc/nginx/nginx.conf

# Commands when creating a new container
CMD /usr/sbin/nginx
```

其中，一开始必须指明所基于的镜像名称，接下来推荐说明维护者信息。

后面则是镜像操作指令，例如 `RUN` 指令，`RUN` 指令将对镜像执行跟随的命令。每运行一条 `RUN` 指令，镜像添加新的一层，并提交。

最后是 `CMD` 指令，来指定运行容器时的操作命令。

下面是一个更复杂的例子

```
# Nginx
#
# VERSION                0.0.1

FROM    ubuntu
MAINTAINER Victor Vieux <victor@docker.com>

RUN apt-get update && apt-get install -y inotify-tools nginx apache2 openssh-server

# Firefox over VNC
#
# VERSION                0.3

FROM ubuntu
```

```
# Install vnc, xvfb in order to create a 'fake' display and firefox
RUN apt-get update && apt-get install -y x11vnc xvfb firefox
RUN mkdir /.vnc
# Setup a password
RUN x11vnc -storepasswd 1234 ~/.vnc/passwd
# Autostart firefox (might not be the best way, but it does the trick)
RUN bash -c 'echo "firefox" >> /.bashrc'

EXPOSE 5900
CMD ["x11vnc", "-forever", "-usepw", "-create"]

# Multiple images example
#
# VERSION          0.1

FROM ubuntu
RUN echo foo > bar
# Will output something like ==> 907ad6c2736f

FROM ubuntu
RUN echo moo > oink
# Will output something like ==> 695d7793cbe4

# You'll now have two images, 907ad6c2736f with /bar, and 695d7793cbe4 with
# /oink.
```

指令

指令的一般格式为 `INSTRUCTION arguments`，指令包括 `FROM`、`MAINTAINER`、`RUN` 等。

FROM

格式为 `FROM <image>` 或 `FROM <image>:<tag>`。

第一条指令必须为 `FROM` 指令。并且，如果在同一个 Dockerfile 中创建多个镜像时，可以使用多个 `FROM` 指令（每个镜像一次）。

MAINTAINER

格式为 `MAINTAINER <name>`，指定维护者信息。

RUN

格式为 `RUN <command>` 或 `RUN ["executable", "param1", "param2"]`。

前者将在 shell 终端中运行命令，即 `/bin/sh -c`；后者则使用 `exec` 执行。指定使用其它终端可以通过第二种方式实现，例如 `RUN ["/bin/bash", "-c", "echo hello"]`。

每条 `RUN` 指令将在当前镜像基础上执行指定命令，并提交为新的镜像。当命令较长时可以使用 `\` 来换行。

CMD

支持三种格式

- `CMD ["executable", "param1", "param2"]` 使用 `exec` 执行，推荐方式；
- `CMD command param1 param2` 在 `/bin/sh` 中执行，提供给需要交互的应用；
- `CMD ["param1", "param2"]` 提供给 `ENTRYPOINT` 的默认参数；

指定启动容器时执行的命令，每个 Dockerfile 只能有一条 `CMD` 命令。如果指定了多条命令，只有最后一条会被执行。

如果用户启动容器时候指定了运行的命令，则会覆盖掉 `CMD` 指定的命令。

EXPOSE

格式为 `EXPOSE <port> [<port>...]`。

告诉 Docker 服务端容器暴露的端口号，供互联系统使用。在启动容器时需要通过 `-P`，Docker 主机会自动分配一个端口转发到指定的端口。

ENV

指令

格式为 `ENV <key> <value>`。指定一个环境变量，会被后续 `RUN` 指令使用，并在容器运行时保持。

例如

```
ENV PG_MAJOR 9.3
ENV PG_VERSION 9.3.4
RUN curl -SL http://example.com/postgres-$PG_VERSION.tar.xz | tar -xJC /usr/src/postgress
ENV PATH /usr/local/postgres-$PG_MAJOR/bin:$PATH
```

ADD

格式为 `ADD <src> <dest>`。

该命令将复制指定的 `<src>` 到容器中的 `<dest>`。其中 `<src>` 可以是 Dockerfile 所在目录的一个相对路径；也可以是一个 URL；还可以是一个 tar 文件（自动解压为目录）。

COPY

格式为 `COPY <src> <dest>`。

复制本地主机的 `<src>`（为 Dockerfile 所在目录的相对路径）到容器中的 `<dest>`。

当使用本地目录为源目录时，推荐使用 `COPY`。

ENTRYPOINT

两种格式：

- `ENTRYPOINT ["executable", "param1", "param2"]`
- `ENTRYPOINT command param1 param2`（shell 中执行）。

配置容器启动后执行的命令，并且不可被 `docker run` 提供的参数覆盖。

每个 Dockerfile 中只能有一个 `ENTRYPOINT`，当指定多个时，只有最后一个起效。

VOLUME

格式为 `VOLUME ["/data"]`。

创建一个可以从本地主机或其他容器挂载的挂载点，一般用来存放数据库和需要保持的数据等。

USER

格式为 `USER daemon`。

指定运行容器时的用户名或 UID，后续的 `RUN` 也会使用指定用户。

当服务不需要管理员权限时，可以通过该命令指定运行用户。并且可以在之前创建所需要的用户，例如：`RUN groupadd -r postgres && useradd -r -g postgres postgres`。要临时获取管理员权限可以使用 `gosu`，而不推荐 `sudo`。

WORKDIR

格式为 `WORKDIR /path/to/workdir`。

为后续的 `RUN`、`CMD`、`ENTRYPOINT` 指令配置工作目录。

可以使用多个 `WORKDIR` 指令，后续命令如果参数是相对路径，则会基于之前命令指定的路径。例如

```
WORKDIR /a
WORKDIR b
WORKDIR c
RUN pwd
```

则最终路径为 `/a/b/c`。

ONBUILD

格式为 `ONBUILD [INSTRUCTION]`。

配置当所创建的镜像作为其它新创建镜像的基础镜像时，所执行的操作指令。

例如，Dockerfile 使用如下的内容创建了镜像 `image-A`。

```
[...]
ONBUILD ADD . /app/src
ONBUILD RUN /usr/local/bin/python-build --dir /app/src
[...]
```

如果基于 `image-A` 创建新的镜像时，新的Dockerfile中使用 `FROM image-A` 指定基础镜像时，会自动执行 `ONBUILD` 指令内容，等价于在后面添加了两条指令。

```
FROM image-A

#Automatically run the following
ADD . /app/src
RUN /usr/local/bin/python-build --dir /app/src
```

使用 `ONBUILD` 指令的镜像，推荐在标签中注明，例如 `ruby:1.9-onbuild`。

创建镜像

编写完成 Dockerfile 之后，可以通过 `docker build` 命令来创建镜像。

基本的格式为 `docker build [选项] 路径`，该命令将读取指定路径下（包括子目录）的 Dockerfile，并将该路径下所有内容发送给 Docker 服务端，由服务端来创建镜像。因此一般建议放置 Dockerfile 的目录为空目录。也可以通过 `.dockerignore` 文件（每一行添加一条匹配模式）来让 Docker 忽略路径下的目录和文件。

要指定镜像的标签信息，可以通过 `-t` 选项，例如

```
$ sudo docker build -t myrepo/myapp /tmp/test1/
```

底层实现

Docker 底层的核心技术包括 Linux 上的名字空间（Namespaces）、控制组（Control groups）、Union 文件系统（Union file systems）和容器格式（Container format）。

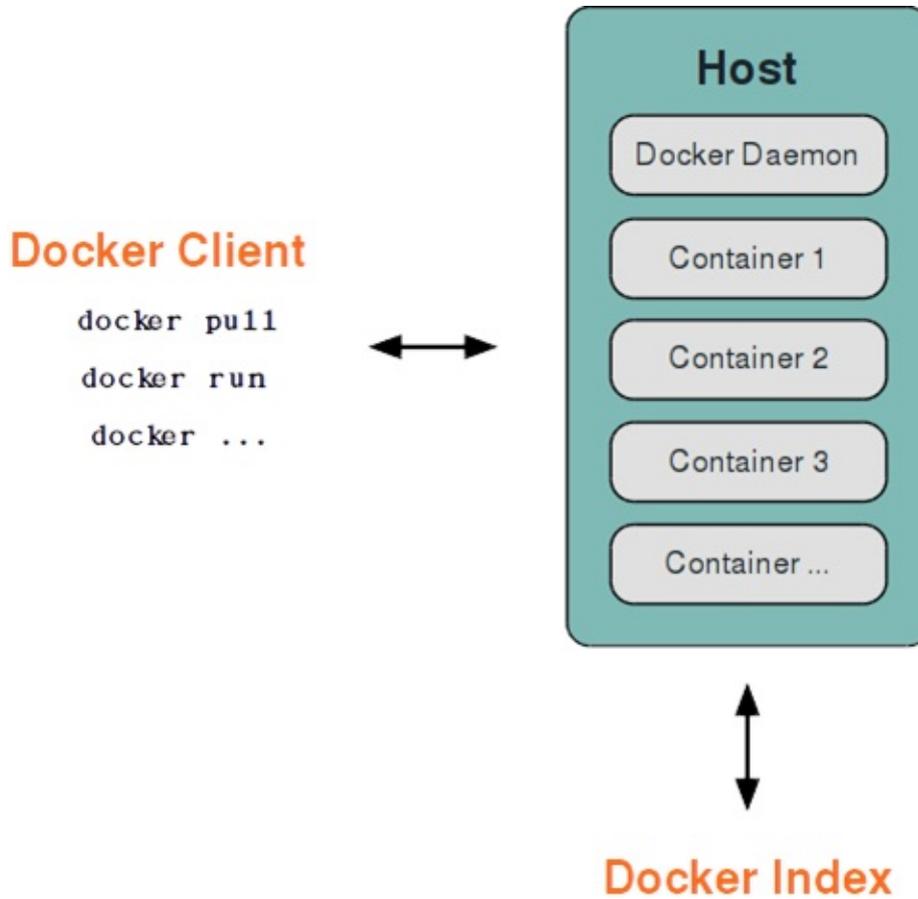
我们知道，传统的虚拟机通过在宿主主机中运行 hypervisor 来模拟一整套完整的硬件环境提供给虚拟机的操作系统。虚拟机系统看到的环境是可限制的，也是彼此隔离的。这种直接的做法实现了对资源最完整的封装，但很多时候往往意味着系统资源的浪费。例如，以宿主机和虚拟机系统都为 Linux 系统为例，虚拟机中运行的应用其实可以利用宿主机系统中的运行环境。

我们知道，在操作系统中，包括内核、文件系统、网络、PID、UID、IPC、内存、硬盘、CPU 等等，所有的资源都是应用进程直接共享的。要想实现虚拟化，除了要实现对内存、CPU、网络IO、硬盘IO、存储空间等的限制外，还要实现文件系统、网络、PID、UID、IPC等等的相互隔离。前者相对容易实现一些，后者则需要宿主机系统的深入支持。

随着 Linux 系统对于名字空间功能的完善实现，程序员已经可以实现上面的所有需求，让某些进程在彼此隔离的名字空间中运行。大家虽然都共用一个内核和某些运行时环境（例如一些系统命令和系统库），但是彼此却看不到，都以为系统中只有自己的存在。这种机制就是容器（Container），利用名字空间来做权限的隔离控制，利用 cgroups 来做资源分配。

基本架构

Docker 采用了 C/S架构，包括客户端和服务端。Docker daemon 作为服务端接受来自客户的请求，并处理这些请求（创建、运行、分发容器）。客户端和服务端既可以运行在一个机器上，也可通过 socket 或者 RESTful API 来进行通信。



Docker daemon 一般在宿主主机后台运行，等待接收来自客户端的消息。Docker 客户端则为用户提供一系列可执行命令，用户用这些命令实现跟 Docker daemon 交互。

名字空间

名字空间是 Linux 内核一个强大的特性。每个容器都有自己单独的名字空间，运行在其中的应用都像是在独立的操作系统中运行一样。名字空间保证了容器之间彼此互不影响。

pid 名字空间

不同用户的进程就是通过 pid 名字空间隔离开的，且不同名字空间中可以有相同 pid。所有的 LXC 进程在 Docker 中的父进程为 Docker 进程，每个 LXC 进程具有不同的名字空间。同时由于允许嵌套，因此可以很方便的实现嵌套的 Docker 容器。

net 名字空间

有了 pid 名字空间，每个名字空间中的 pid 能够相互隔离，但是网络端口还是共享 host 的端口。网络隔离是通过 net 名字空间实现的，每个 net 名字空间有独立的网络设备，IP 地址，路由表，/proc/net 目录。这样每个容器的网络就能隔离开来。Docker 默认采用 veth 的方式，将容器中的虚拟网卡同 host 上的一个 Docker 网桥 docker0 连接在一起。

ipc 名字空间

容器中进程交互还是采用了 Linux 常见的进程间交互方法(interprocess communication - IPC)，包括信号量、消息队列和共享内存等。然而同 VM 不同的是，容器的进程间交互实际上还是 host 上具有相同 pid 名字空间中的进程间交互，因此需要在 IPC 资源申请时加入名字空间信息，每个 IPC 资源有一个唯一的 32 位 id。

mnt 名字空间

类似 chroot，将一个进程放到一个特定的目录执行。mnt 名字空间允许不同名字空间的进程看到的文件结构不同，这样每个名字空间中的进程所看到的文件目录就被隔离开了。同 chroot 不同，每个名字空间中的容器在 /proc/mounts 的信息只包含所在名字空间的 mount point。

uts 名字空间

UTS("UNIX Time-sharing System") 名字空间允许每个容器拥有独立的 hostname 和 domain name，使其在网络上可以被视作一个独立的节点而非主机上的一个进程。

user 名字空间

每个容器可以有不同的用户和组 id，也就是说可以在容器内用容器内部的用户执行程序而非主机上的用户。

*注：关于 Linux 上的名字空间，[这篇文章](#)介绍的很好。

控制组

控制组（[cgroups](#)）是 Linux 内核的一个特性，主要用来对共享资源进行隔离、限制、审计等。只有能控制分配到容器的资源，才能避免当多个容器同时运行时的对系统资源的竞争。

控制组技术最早是由 Google 的程序员 2006 年起提出，Linux 内核自 2.6.24 开始支持。

控制组可以提供对容器的内存、CPU、磁盘 IO 等资源的限制和审计管理。

联合文件系统

联合文件系统（**UnionFS**）是一种分层、轻量级并且高性能的文件系统，它支持对文件系统的修改作为一次提交来一层层的叠加，同时可以将不同目录挂载到同一个虚拟文件系统下(unite several directories into a single virtual filesystem)。

联合文件系统是 Docker 镜像的基础。镜像可以通过分层来进行继承，基于基础镜像（没有父镜像），可以制作各种具体的应用镜像。

另外，不同 Docker 容器就可以共享一些基础的文件系统层，同时再加上自己独有的改动层，大大提高了存储的效率。

Docker 中使用的 AUFS（AnotherUnionFS）就是一种联合文件系统。AUFS 支持为每一个成员目录（类似 Git 的分支）设定只读（readonly）、读写（readwrite）和写出（whiteout-able）权限，同时 AUFS 里有一个类似分层的概念，对只读权限的分支可以逻辑上进行增量地修改(不影响只读部分的)。

Docker 目前支持的联合文件系统种类包括 AUFS, btrfs, vfs 和 DeviceMapper。

容器格式

最初，Docker 采用了 LXC 中的容器格式。自 1.20 版本开始，Docker 也开始支持新的 [libcontainer](#) 格式，并作为默认选项。

对更多容器格式的支持，还在进一步的发展中。

Docker 网络实现

Docker 的网络实现其实就是利用了 Linux 上的网络名字空间和虚拟网络设备（特别是 veth pair）。建议先熟悉了解这两部分的基本概念再阅读本章。

基本原理

首先，要实现网络通信，机器需要至少一个网络接口（物理接口或虚拟接口）来收发数据包；此外，如果不同子网之间要进行通信，需要路由机制。

Docker 中的网络接口默认都是虚拟的接口。虚拟接口的优势之一是转发效率较高。Linux 通过在内核中进行数据复制来实现虚拟接口之间的数据转发，发送接口的发送缓存中的数据包被直接复制到接收接口的接收缓存中。对于本地系统和容器内系统看来就像是一个正常的以太网卡，只是它不需要真正同外部网络设备通信，速度要快很多。

Docker 容器网络就利用了这项技术。它在本地主机和容器内分别创建一个虚拟接口，并让它们彼此连通（这样的一对接口叫做 veth pair）。

创建网络参数

Docker 创建一个容器的时候，会执行如下操作：

- 创建一对虚拟接口，分别放到本地主机和新容器中；
- 本地主机一端桥接到默认的 docker0 或指定网桥上，并具有一个唯一的名字，如 veth65f9；
- 容器一端放到新容器中，并修改名字作为 eth0，这个接口只在容器的名字空间可见；
- 从网桥可用地址段中获取一个空闲地址分配给容器的 eth0，并配置默认路由到桥接网卡 veth65f9。

完成这些之后，容器就可以使用 eth0 虚拟网卡来连接其他容器和其他网络。

可以在 `docker run` 的时候通过 `--net` 参数来指定容器的网络配置，有4个可选值：

- `--net=bridge` 这个是默认值，连接到默认的网桥。
- `--net=host` 告诉 Docker 不要将容器网络放到隔离的名字空间中，即不要容器化容器内的网络。此时容器使用本地主机的网络，它拥有完全的本地主机接口访问权限。容器进程可以跟主机其它 root 进程一样可以打开低范围的端口，可以访问本地网络服务比如 D-bus，还可以让容器做一些影响整个主机系统的事情，比如重启主机。因此使用这个选项的时候要非常小心。如果进一步的使用 `--privileged=true`，容器会被允许直接配置主机的网络堆栈。
- `--net=container:NAME_or_ID` 让 Docker 将新建容器的进程放到一个已存在容器的网络栈中，新容器进程有自己的文件系统、进程列表和资源限制，但会和已存在的容器共享 IP 地址和端口等网络资源，两者进程可以直接通过 `lo` 环回接口通信。
- `--net=none` 让 Docker 将新容器放到隔离的网络栈中，但是不进行网络配置。之后，用户可以自己进行配置。

网络配置细节

用户使用 `--net=none` 后，可以自行配置网络，让容器达到跟平常一样具有访问网络的权限。通过这个过

程，可以了解 Docker 配置网络的细节。

首先，启动一个 `/bin/bash` 容器，指定 `--net=none` 参数。

```
$ sudo docker run -i -t --rm --net=none base /bin/bash
root@63f36fc01b5f:/#
```

在本地主机查找容器的进程 id，并为它创建网络命名空间。

```
$ sudo docker inspect -f '{{.State.Pid}}' 63f36fc01b5f
2778
$ pid=2778
$ sudo mkdir -p /var/run/netns
$ sudo ln -s /proc/$pid/ns/net /var/run/netns/$pid
```

检查桥接网卡的 IP 和子网掩码信息。

```
$ ip addr show docker0
21: docker0: ...
inet 172.17.42.1/16 scope global docker0
...
```

创建一对“veth pair”接口 A 和 B，绑定 A 到网桥 `docker0`，并启用它

```
$ sudo ip link add A type veth peer name B
$ sudo brctl addif docker0 A
$ sudo ip link set A up
```

将 B 放到容器的网络命名空间，命名为 `eth0`，启动它并配置一个可用 IP（桥接网段）和默认网关。

```
$ sudo ip link set B netns $pid
$ sudo ip netns exec $pid ip link set dev B name eth0
$ sudo ip netns exec $pid ip link set eth0 up
$ sudo ip netns exec $pid ip addr add 172.17.42.99/16 dev eth0
$ sudo ip netns exec $pid ip route add default via 172.17.42.1
```

以上，就是 Docker 配置网络的具体过程。

当容器结束后，Docker 会清空容器，容器内的 `eth0` 会随网络命名空间一起被清除，A 接口也被自动从 `docker0` 卸载。

此外，用户可以使用 `ip netns exec` 命令来在指定网络名字空间中进行配置，从而配置容器内的网络。

Docker Compose 项目

Docker Compose 是 Docker 官方编排（Orchestration）项目之一，负责快速在集群中部署分布式应用。

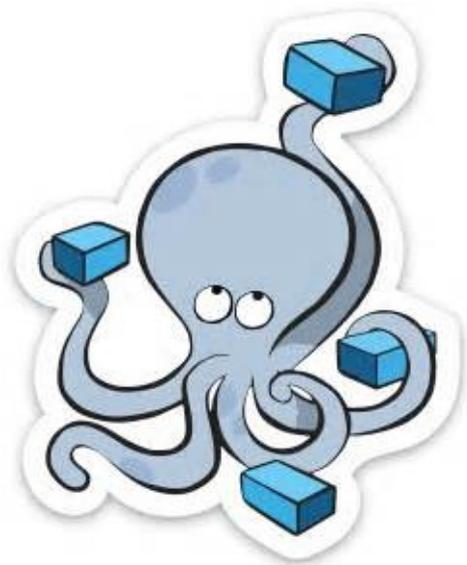
本章将介绍 Compose 项目情况以及安装和使用。

简介

Compose 项目目前在 [Github](#) 上进行维护，目前最新版本是 1.2.0。

Compose 定位是“defining and running complex applications with Docker”，前身是 Fig，兼容 Fig 的模板文件。

Dockerfile 可以让用户管理一个单独的应用容器；而 Compose 则允许用户在一个模板（YAML 格式）中定义一组相关联的应用容器（被称为一个 `project`，即项目），例如一个 Web 服务容器再加上后端的数据库服务容器等。



该项目由 Python 编写，实际上调用了 Docker 提供的 API 来实现。

安装

安装 Compose 之前，要先安装 Docker，在此不再赘述。

PIP 安装

这种方式最为推荐。

执行命令。

```
$ sudo pip install -U docker-compose
```

安装成功后，可以查看 `docker-compose` 命令的用法。

```
$ docker-compose -h
Fast, isolated development environments using Docker.

Usage:
  docker-compose [options] [COMMAND] [ARGS...]
  docker-compose -h|--help

Options:
  --verbose          Show more output
  --version          Print version and exit
  -f, --file FILE   Specify an alternate compose file (default: docker-compose.yml)
  -p, --project-name NAME Specify an alternate project name (default: directory name)

Commands:
  build      Build or rebuild services
  help      Get help on a command
  kill      Kill containers
  logs      View output from containers
  port      Print the public port for a port binding
  ps        List containers
  pull      Pulls service images
  rm        Remove stopped containers
  run       Run a one-off command
  scale     Set number of containers for a service
  start     Start services
  stop      Stop services
  restart   Restart services
  up        Create and start containers
```

之后，可以添加 `bash` 补全命令。

```
$ curl -L https://raw.githubusercontent.com/docker/compose/1.2.0/contrib/completion/bash/
```

二进制包

发布的二进制包可以在 <https://github.com/docker/compose/releases> 找到。

下载后直接放到执行路径即可。

例如，在常见的 Linux 平台上。

```
$ sudo curl -L https://github.com/docker/compose/releases/download/1.2.0/docker-compose-`uname -s`-`uname -m`>/usr/local/bin/docker-compose
$ sudo chmod a+x /usr/local/bin/docker-compose
```

使用

术语

首先介绍几个术语。

- 服务 (service) : 一个应用容器, 实际上可以运行多个相同镜像的实例。
- 项目(project) : 由一组关联的应用容器组成的一个完整业务单元。

可见, 一个项目可以由多个服务 (容器) 关联而成, Compose 面向项目进行管理。

场景

下面, 我们创建一个经典的 Web 项目: 一个 [Haproxy](#), 挂载三个 Web 容器。

创建一个 `compose-haproxy-web` 目录, 作为项目工作目录, 并在其中分别创建两个子目录: `haproxy` 和 `web`。

Web 子目录

这里用 Python 程序来提供一个简单的 HTTP 服务, 打印出访问者的 IP 和 实际的本地 IP。

index.py

编写一个 `index.py` 作为服务器文件, 代码为

```
#!/usr/bin/python
#authors: yeasy.github.com
#date: 2013-07-05

import sys
import BaseHTTPServer
from SimpleHTTPServer import SimpleHTTPRequestHandler
import socket
import fcntl
import struct
import pickle
from datetime import datetime
from collections import OrderedDict

class HandlerClass(SimpleHTTPRequestHandler):
    def get_ip_address(self, ifname):
        s = socket.socket(socket.AF_INET, socket.SOCK_DGRAM)
        return socket.inet_ntoa(fcntl.ioctl(
            s.fileno(),
            0x8915, # SIOCGIFADDR
            struct.pack('256s', ifname[:15])
        )[20:24])
    def log_message(self, format, *args):
        if len(args) < 3 or "200" not in args[1]:
            return
```

```

try:
    request = pickle.load(open("pickle_data.txt", "r"))
except:
    request=OrderedDict()
time_now = datetime.now()
ts = time_now.strftime('%Y-%m-%d %H:%M:%S')
server = self.get_ip_address('eth0')
host=self.address_string()
addr_pair = (host, server)
if addr_pair not in request:
    request[addr_pair]=[1, ts]
else:
    num = request[addr_pair][0]+1
    del request[addr_pair]
    request[addr_pair]=[num, ts]
file=open("index.html", "w")
file.write("<!DOCTYPE html> <html> <body><center><h1><font color=\"blue\" face=\"
for pair in request:
    if pair[0] == host:
        guest = "LOCAL: "+pair[0]
    else:
        guest = pair[0]
    if (time_now-datetime.strptime(request[pair][1], '%Y-%m-%d %H:%M:%S')).seconds
        file.write("<p style=\"font-size:150%\" >#"+ str(request[pair][1]) +": <f
    else:
        file.write("<p style=\"font-size:150%\" >#"+ str(request[pair][1]) +": <f
file.write("</body> </html>");
file.close()
pickle.dump(request,open("pickle_data.txt", "w"))

if __name__ == '__main__':
    try:
        ServerClass = BaseHTTPServer.HTTPServer
        Protocol = "HTTP/1.0"
        addr = len(sys.argv) < 2 and "0.0.0.0" or sys.argv[1]
        port = len(sys.argv) < 3 and 80 or int(sys.argv[2])
        HandlerClass.protocol_version = Protocol
        httpd = ServerClass((addr, port), HandlerClass)
        sa = httpd.socket.getsockname()
        print "Serving HTTP on", sa[0], "port", sa[1], "..."
        httpd.serve_forever()
    except:
        exit()

```

index.html

生成一个临时的 index.html 文件，其内容会被 index.py 更新。

```
$ touch index.html
```

Dockerfile

生成一个 Dockerfile，内容为

```
FROM python:2.7
WORKDIR /code
ADD . /code
EXPOSE 80
CMD python index.py
```

haproxy 目录

在其中生成一个 `haproxy.cfg` 文件，内容为

```
global
    log 127.0.0.1 local0
    log 127.0.0.1 local1 notice

defaults
    log global
    mode http
    option httplog
    option dontlognull
    timeout connect 5000ms
    timeout client 50000ms
    timeout server 50000ms

listen stats :70
    stats enable
    stats uri /

frontend balancer
    bind 0.0.0.0:80
    mode http
    default_backend web_backends

backend web_backends
    mode http
    option forwardfor
    balance roundrobin
    server weba weba:80 check
    server webb webb:80 check
    server webc webc:80 check
    option httpchk GET /
    http-check expect status 200
```

docker-compose.yml

编写 `docker-compose.yml` 文件，这个是 Compose 使用的主模板文件。内容十分简单，指定 3 个 web 容器，以及 1 个 haproxy 容器。

```
weba:
  build: ./web
  expose:
    - 80

webb:
```

```

    build: ./web
    expose:
      - 80

webc:
  build: ./web
  expose:
    - 80

haproxy:
  image: haproxy:latest
  volumes:
    - haproxy:/haproxy-override
    - haproxy/haproxy.cfg:/usr/local/etc/haproxy/haproxy.cfg:ro
  links:
    - weba
    - webb
    - webc
  ports:
    - "80:80"
    - "70:70"
  expose:
    - "80"
    - "70"

```

运行 compose 项目

现在 compose-haproxy-web 目录长成下面的样子。

```

compose-haproxy-web
├── docker-compose.yml
├── haproxy
│   └── haproxy.cfg
└── web
    ├── Dockerfile
    ├── index.html
    └── index.py

```

在该目录下执行 `docker-compose up` 命令，会整合输出所有容器的输出。

```

$sudo docker-compose up
Recreating composehaproxyweb_webb_1...
Recreating composehaproxyweb_webc_1...
Recreating composehaproxyweb_weba_1...
Recreating composehaproxyweb_haproxy_1...
Attaching to composehaproxyweb_webb_1, composehaproxyweb_webc_1, composehaproxyweb_weba_1

```

此时访问本地的 80 端口，会经过 haproxy 自动转发到后端的某个 web 容器上，刷新页面，可以观察到访问的容器地址的变化。

访问本地 70 端口，可以查看到 haproxy 的统计信息。

当然，还可以使用 consul、etcd 等实现服务发现，这样就可以避免手动指定后端的 web 容器了，更为灵活。

Compose 命令说明

大部分命令都可以运行在一个或多个服务上。如果没有特别的说明，命令则应用在项目所有的服务上。

执行 `docker-compose [COMMAND] --help` 查看具体某个命令的使用说明。

基本的使用格式是

```
docker-compose [options] [COMMAND] [ARGS...]
```

选项

- `--verbose` 输出更多调试信息。
- `--version` 打印版本并退出。
- `-f, --file FILE` 使用特定的 compose 模板文件，默认为 `docker-compose.yml`。
- `-p, --project-name NAME` 指定项目名称，默认使用目录名称。

命令

build

构建或重新构建服务。

服务一旦构建后，将会带上一个标记名，例如 `web_db`。

可以随时在项目目录下运行 `docker-compose build` 来重新构建服务。

help

获得一个命令的帮助。

kill

通过发送 `SIGKILL` 信号来强制停止服务容器。支持通过参数来指定发送的信号，例如

```
$ docker-compose kill -s SIGINT
```

logs

查看服务的输出。

port

打印绑定的公共端口。

ps

列出所有容器。

pull

拉取服务镜像。

rm

删除停止的服务容器。

run

在一个服务上执行一个命令。

例如：

```
$ docker-compose run ubuntu ping docker.com
```

将会启动一个 `ubuntu` 服务，执行 `ping docker.com` 命令。

默认情况下，所有关联的服务将会自动被启动，除非这些服务已经在运行中。

该命令类似启动容器后运行指定的命令，相关卷、链接等等都将会按照期望创建。

两个不同点：

- 给定命令将会覆盖原有的自动运行命令；
- 不会自动创建端口，以避免冲突。

如果不希望自动启动关联的容器，可以使用 `--no-deps` 选项，例如

```
$ docker-compose run --no-deps web python manage.py shell
```

将不会启动 `web` 容器所关联的其它容器。

scale

设置同一个服务运行的容器个数。

通过 `service=num` 的参数来设置数量。例如：

```
$ docker-compose scale web=2 worker=3
```

start

命令说明

启动一个已经存在的服务容器。

stop

停止一个已经运行的容器，但不删除它。通过 `docker-compose start` 可以再次启动这些容器。

up

构建，（重新）创建，启动，链接一个服务相关的容器。

链接的服务都将会启动，除非他们已经运行。

默认情况，`docker-compose up` 将会整合所有容器的输出，并且退出时，所有容器将会停止。

如果使用 `docker-compose up -d`，将会在后台启动并运行所有的容器。

默认情况，如果该服务的容器已经存在，`docker-compose up` 将会停止并尝试重新创建他们（保持使用 `volumes-from` 挂载的卷），以保证 `docker-compose.yml` 的修改生效。如果你不想容器被停止并重新创建，可以使用 `docker-compose up --no-recreate`。如果需要的话，这样将会启动已经停止的容器。

环境变量

环境变量可以用来配置 Compose 的行为。

以 `DOCKER_` 开头的变量和用来配置 Docker 命令行客户端的使用一样。如果使用 `boot2docker`，`$(boot2docker shellinit)` 将会设置它们为正确的值。

COMPOSE_PROJECT_NAME

设置通过 Compose 启动的每一个容器前添加的项目名称，默认是当前工作目录的名字。

COMPOSE_FILE

设置要使用的 `docker-compose.yml` 的路径。默认路径是当前工作目录。

DOCKER_HOST

设置 Docker daemon 的地址。默认使用 `unix:///var/run/docker.sock`，与 Docker 客户端采用的默认值一致。

DOCKER_TLS_VERIFY

如果设置不为空，则与 Docker daemon 交互通过 TLS 进行。

DOCKER_CERT_PATH

配置 TLS 通信所需要的验证（`ca.pem`、`cert.pem` 和 `key.pem`）文件的路径，默认是 `~/.docker`。

YAML 模板文件

默认的模板文件是 `docker-compose.yml`，其中定义的每个服务都必须通过 `image` 指令指定镜像或 `build` 指令（需要 `Dockerfile`）来自动构建。

其它大部分指令都跟 `docker run` 中的类似。

如果使用 `build` 指令，在 `Dockerfile` 中设置的选项(例如：`CMD`，`EXPOSE`，`VOLUME`，`ENV` 等)将会自动被获取，无需在 `docker-compose.yml` 中再次设置。

image

指定为镜像名称或镜像 ID。如果镜像在本地不存在，`Compose` 将会尝试拉去这个镜像。

例如：

```
image: ubuntu
image: orchardup/postgresql
image: a4bc65fd
```

build

指定 `Dockerfile` 所在文件夹的路径。`Compose` 将会利用它自动构建这个镜像，然后使用这个镜像。

```
build: /path/to/build/dir
```

command

覆盖容器启动后默认执行的命令。

```
command: bundle exec thin -p 3000
```

links

链接到其它服务中的容器。使用服务名称（同时作为别名）或服务名称：服务别名（`SERVICE:ALIAS`）格式都可以。

```
links:
- db
- db:database
- redis
```

使用的别名将会自动在服务容器中的 `/etc/hosts` 里创建。例如：

```
172.17.2.186 db
172.17.2.186 database
172.17.2.187 redis
```

相应环境变量也将被创建。

external_links

链接到 docker-compose.yml 外部的容器，甚至并非 Compose 管理的容器。参数格式跟 links 类似。

```
external_links:
- redis_1
- project_db_1:mysql
- project_db_1:postgresql
```

ports

暴露端口信息。

使用宿主：容器（HOST:CONTAINER）格式或者仅仅指定容器的端口（宿主将会随机选择端口）都可以。

```
ports:
- "3000"
- "8000:8000"
- "49100:22"
- "127.0.0.1:8001:8001"
```

注：当使用 HOST:CONTAINER 格式来映射端口时，如果你使用的容器端口小于 60 你可能会得到错误得结果，因为 YAML 将会解析 xx:yy 这种数字格式为 60 进制。所以建议采用字符串格式。

expose

暴露端口，但不映射到宿主机，只被连接的服务访问。

仅可以指定内部端口为参数

```
expose:
- "3000"
- "8000"
```

volumes

卷挂载路径设置。可以设置宿主机路径（HOST:CONTAINER）或加上访问模式（HOST:CONTAINER:ro）。

```
volumes:
```

```
- /var/lib/mysql
- cache:/tmp/cache
- ~/configs:/etc/configs/:ro
```

volumes_from

从另一个服务或容器挂载它的所有卷。

```
volumes_from:
- service_name
- container_name
```

environment

设置环境变量。你可以使用数组或字典两种格式。

只给定名称的变量会自动获取它在 Compose 主机上的值，可以用来防止泄露不必要的数据。

```
environment:
  RACK_ENV: development
  SESSION_SECRET:

environment:
- RACK_ENV=development
- SESSION_SECRET
```

env_file

从文件中获取环境变量，可以为单独的文件路径或列表。

如果通过 `docker-compose -f FILE` 指定了模板文件，则 `env_file` 中路径会基于模板文件路径。

如果有变量名称与 `environment` 指令冲突，则以后者为准。

```
env_file: .env

env_file:
- ./common.env
- ./apps/web.env
- /opt/secrets.env
```

环境变量文件中每一行必须符合格式，支持 `#` 开头的注释行。

```
# common.env: Set Rails/Rack environment
RACK_ENV=development
```

extends

基于已有的服务进行扩展。例如我们已经有了一个 webapp 服务，模板文件为 `common.yml`。

```
# common.yml
webapp:
  build: ./webapp
  environment:
    - DEBUG=false
    - SEND_EMAILS=false
```

编写一个新的 `development.yml` 文件，使用 `common.yml` 中的 webapp 服务进行扩展。

```
# development.yml
web:
  extends:
    file: common.yml
    service: webapp
  ports:
    - "8000:8000"
  links:
    - db
  environment:
    - DEBUG=true
db:
  image: postgres
```

后者会自动继承 `common.yml` 中的 webapp 服务及相关环境变量。

net

设置网络模式。使用 `docker client` 的 `--net` 参数一样的值。

```
net: "bridge"
net: "none"
net: "container:[name or id]"
net: "host"
```

pid

跟主机系统共享进程命名空间。打开该选项的容器可以相互通过进程 ID 来访问和操作。

```
pid: "host"
```

dns

配置 DNS 服务器。可以是一个值，也可以是一个列表。

```
dns: 8.8.8.8
dns:
```

- 8.8.8.8
- 9.9.9.9

cap_add, cap_drop

添加或放弃容器的 Linux 能力（Capability）。

```
cap_add:  
- ALL  
  
cap_drop:  
- NET_ADMIN  
- SYS_ADMIN
```

dns_search

配置 DNS 搜索域。可以是一个值，也可以是一个列表。

```
dns_search: example.com  
dns_search:  
- domain1.example.com  
- domain2.example.com
```

working_dir, entrypoint, user, hostname, domainname, mem_limit, privileged, restart, stdin_open, tty, cpu_shares

这些都是和 `docker run` 支持的选项类似。

```
cpu_shares: 73  
  
working_dir: /code  
entrypoint: /code/entrypoint.sh  
user: postgresql  
  
hostname: foo  
domainname: foo.com  
  
mem_limit: 1000000000  
privileged: true  
  
restart: always  
  
stdin_open: true  
tty: true
```

Docker Machine 项目

Docker Machine 是 Docker 官方编排（Orchestration）项目之一，负责在多种平台上快速安装 Docker 环境。

本章将介绍 Machine 项目情况以及安装和使用。

简介



Docker Machine 项目基于 Go 语言实现，目前在 [Github](#) 上进行维护。

技术讨论 IRC 频道为 `#docker-machine`。

安装

Docker Machine 可以在多种操作系统平台上安装，包括 Linux、Mac OS，以及 Windows。

Linux/Mac OS

在 Linux/Mac OS 上的安装十分简单，推荐从 [官方 Release 库](#) 直接下载编译好的二进制文件即可。

例如，在 Linux 64 位系统上直接下载对应的二进制包。

```
$ sudo curl -L https://github.com/docker/machine/releases/download/v0.3.1-rc1/docker-machine_
$ chmod +x /usr/local/bin/docker-machine
```

完成后，查看版本信息，验证运行正常。

```
$ docker-machine -v
docker-machine version 0.3.1-rc1 (993f2db)
```

Windows

Windows 下面要复杂一些，首先需要安装 [msysgit](#)。

msysgit 是 Windows 下的 git 客户端软件包，会提供类似 Linux 下的一些基本的工具，例如 ssh 等。

安装之后，启动 msysgit 的命令行界面，仍然通过下载二进制包进行安装，需要下载 docker 客户端和 docker-machine。

```
$ curl -L https://get.docker.com/builds/Windows/x86_64/docker-latest.exe > /bin/docker
$ curl -L https://github.com/docker/machine/releases/download/v0.3.1-rc1/docker-machine_w
```

使用

Docker Machine 支持多种后端驱动，包括虚拟机、本地主机和云平台等。

本地主机实例

首先确保本地主机可以通过 user 账号的 key 直接 ssh 到目标主机。

使用 generic 类型的驱动，创建一台 Docker 主机，命名为 test。

```
$ docker-machine create -d generic --generic-ip-address=10.0.100.101 --generic-ssh-user=u
```

创建主机成功后，可以通过 env 命令来让后续操作对象都是目标主机。

```
$ docker-machine env test
```

支持驱动

通过 `-d` 选项可以选择支持的驱动类型。

- amazonec2
- azure
- digitalocean
- exoscale
- generic
- google
- none
- openstack
- rackspace
- softlayer
- virtualbox
- vmwarevcloudair
- vmwarevsphere

操作命令

- `active` 查看活跃的 Docker 主机
- `config` 输出连接的配置信息
- `create` 创建一个 Docker 主机
- `env` 显示连接到某个主机需要的环境变量
- `inspect` 输出主机更多信息
- `ip` 获取主机地址

- `kill` 停止某个主机
- `ls` 列出所有管理的主机
- `regenerate-certs` 为某个主机重新生成 TLS 认证信息
- `restart` 重启主机
- `rm` 删除某台主机
- `ssh` SSH 到主机上执行命令
- `scp` 在主机之间复制文件
- `start` 启动一个主机
- `stop` 停止一个主机
- `upgrade` 更新主机 Docker 版本为最新
- `url` 获取主机的 URL
- `help, h` 输出帮助信息

每个命令，又带有不同的参数，可以通过

```
docker-machine <COMMAND> -h
```

来查看具体的用法。

Docker Swarm 项目

Docker Swarm 是 Docker 官方编排（Orchestration）项目之一，负责对 Docker 集群进行管理。

本章将介绍 Swarm 项目情况以及安装和使用。

简介

Docker Swarm 是 Docker公司官方在 2014 年 12月初发布的一套管理 Docker 集群的工具。它将一群 Docker 宿主机变成一个单一的，虚拟的主机。

Swarm 使用标准的 Docker API 接口作为其前端访问入口，换言之，各种形式的 Docker 工具比如 Dokku, Compose, Krane, Deis, docker-py, Docker 本身等都可以很容易的与 Swarm 进行集成。



在使用 Swarm 管理docker 集群时，会有一个 swarm manager 以及若干的 swarm node，swarm manager 上运行 swarm daemon，用户只需要跟 swarm manager 通信，然后 swarm manager 再根据discovery service的信息选择一个swarm node 来运行container。

值得注意的是 swarm daemon 只是一个任务调度器(scheduler)和路由器(router),它本身不运行容器，它只接受 Docker client 发送过来的请求，调度合适的 swarm node 来运行 container。这意味着，即使 swarm daemon 由于某些原因挂掉了，已经运行起来的容器也不会有任何影响。

有以下两点需要注意：

- 集群中的每台节点上面的 Docker 的版本都不能小于1.4
- 为了让 swarm manager 能够跟每台 swarm node 进行通信，集群中的每台节点的 Docker daemon 都必须监听同一个网络接口。

安装

安装swarm的最简单的方式是使用Docker官方的swarm镜像

```
$ sudo docker pull swarm
```

可以使用下面的命令来查看swarm是否成功安装。

```
$ sudo docker run --rm swarm -v
```

输出下面的形式则表示成功安装(具体输出根据swarm的版本变化)

```
swarm version 0.2.0 (48fd993)
```

使用

在使用 swarm 管理集群前，需要把集群中所有的节点的 docker daemon 的监听方式更改为 `0.0.0.0:2375`。

可以有两种方式达到这个目的，第一种是在启动 docker daemon 的时候指定

```
sudo docker -H 0.0.0.0:2375&
```

第二种方式是直接修改 Docker 的配置文件(Ubuntu 上是 `/etc/default/docker`，其他版本的 Linux 上略有不同)

在文件的最后添加下面这句代码：

```
DOCKER_OPTS="-H 0.0.0.0:2375 -H unix:///var/run/docker.sock"
```

需要注意的是，一定要在所有希望被 Swarm 管理的节点上进行的。修改之后要重启 Docker

```
sudo service docker restart
```

Docker 集群管理需要使用服务发现(Discovery service backend)功能，Swarm支持以下的几种方式：DockerHub 提供的服务发现功能，本地的文件，etcd，counsel，zookeeper 和 IP 列表，本文会详细讲解前两种方式，其他的用法都是大同小异的。

先说一下本次试验的环境，本次试验包括三台机器，IP地址分别为192.168.1.84,192.168.1.83和192.168.1.124.利用这三台机器组成一个docker集群，其中83这台机器同时充当swarm manager节点。

使用 DockerHub 提供的服务发现功能

创建集群 token

在上面三台机器中的任何一台机器上面执行 `swarm create` 命令来获取一个集群标志。这条命令执行完毕后，Swarm 会前往 DockerHub 上内置的发现服务中获取一个全球唯一的 token，用来标识要管理的集群。

```
sudo docker run --rm swarm create
```

我们在84这台机器上执行这条命令，输出如下：

```
rio@084:~$ sudo docker run --rm swarm create
b7625e5a7a2dc7f8c4faacf2b510078e
```

可以看到我们返回的 token 是 `b7625e5a7a2dc7f8c4faacf2b510078e`，每次返回的结果都是不一样的。这

个 token 一定要记住，后面的操作都会用到这个 token。

加入集群

在所有要加入集群的节点上面执行 `swarm join` 命令，表示要把这台机器加入这个集群当中。在本次试验中，就是要在 83、84 和 124 这三台机器上执行下面的这条命令：

```
sudo docker run --rm swarm join addr=ip_address:2375 token://token_id
```

其中的 `ip_address` 换成执行这条命令的机器的 IP，`token_id` 换成上一步执行 `swarm create` 返回的 token。

在 83 这台机器上面的执行结果如下：

```
rio@083:~$ sudo docker run --rm swarm join --addr=192.168.1.83:2375 token://b7625e5a7a2dc
time="2015-05-19T11:48:03Z" level=info msg="Registering on the discovery service every 2
```

这条命令不会自动返回，要我们自己执行 `Ctrl+C` 返回。

启动 swarm manager

因为我们要使用 83 这台机器充当 swarm 管理节点，所以需要在 83 这台机器上面执行 `swarm manage` 命令：

```
sudo docker run -d -p 2376:2375 swarm manage token://b7625e5a7a2dc7f8c4faacf2b510078e
```

执行结果如下：

```
rio@083:~$ sudo docker run -d -p 2376:2375 swarm manage token://b7625e5a7a2dc7f8c4faacf2b
83de3e9149b7a0ef49916d1dbe073e44e8c31c2fcbe98d962a4f85380ef25f76
```

这条命令如果执行成功会返回已经启动的 Swarm 的容器的 ID，此时整个集群已经启动起来了。

现在通过 `docker ps` 命令来看下有没有启动成功。

```
rio@083:~$ sudo docker ps
CONTAINER ID        IMAGE               COMMAND             CREATED             STATUS
83de3e9149b7       swarm:latest       "/swarm manage token  4 minutes ago      Up 4 m
```

可以看到，Swarm 已经成功启动。在执行 `Swarm manage` 这条命令的时候，有几点需要注意的：

- 这条命令需要在充当 swarm 管理者的机器上执行

- Swarm 要以 daemon 的形式执行
- 映射的端口可以使任意的除了 2375 以外的并且是未被占用的端口，但一定不能是 2375 这个端口，因为 2375 已经被 Docker 本身给占用了。

集群启动成功以后，现在我们可以任何一台节点上使用 `swarm list` 命令查看集群中的节点了，本实验在 124 这台机器上执行 `swarm list` 命令：

```
rio@124:~$ sudo docker run --rm swarm list token://b7625e5a7a2dc7f8c4faacf2b510078e
192.168.1.84:2375
192.168.1.124:2375
192.168.1.83:2375
```

输出结果列出的IP地址正是我们使用 `swarm join` 命令加入集群的机器的IP地址。

现在我们可以任何一台安装了 Docker 的机器上面通过命令(命令中要指明swarm manager机器的IP地址)来在集群中运行container了。本次试验，我们在 192.168.1.85 这台机器上使用 `docker info` 命令来查看集群中的节点的信息。

其中 info 也可以换成其他的 Docker 支持的命令。

```
rio@085:~$ sudo docker -H 192.168.1.83:2376 info
Containers: 8
Strategy: spread
Filters: affinity, health, constraint, port, dependency
Nodes: 2
sclu083: 192.168.1.83:2375
├ Containers: 1
├ Reserved CPUs: 0 / 2
├ Reserved Memory: 0 B / 4.054 GiB
sclu084: 192.168.1.84:2375
├ Containers: 7
├ Reserved CPUs: 0 / 2
├ Reserved Memory: 0 B / 4.053 GiB
```

结果输出显示这个集群中只有两个节点，IP地址分别是 192.168.1.83 和 192.168.1.84，结果不对呀，我们明明把三台机器加入了这个集群，还有 124 这一台机器呢？经过排查，发现是忘了修改 124 这台机器上面改 `docker daemon` 的监听方式，只要按照上面的步骤修改写 `docker daemon` 的监听方式就可以了。

在使用这个方法的时候，使用 `swarm create` 可能会因为网络的原因会出现类似于下面的这个问题：

```
rio@227:~$ sudo docker run --rm swarm create
[sudo] password for rio:
time="2015-05-19T12:59:26Z" level=fatal msg="Post https://discovery-stage.hub.docker.com/
```

使用文件

第二种方法相对于第一种方法要简单得多，也不会出现类似于上面的问题。

第一步：在 swarm 管理节点上新建一个文件，把要加入集群的机器 IP 地址和端口号写入文件中，本次试验就是要在83这台机器上面操作：

```
rio@083:~$ echo 192.168.1.83:2375 >> cluster
rio@083:~$ echo 192.168.1.84:2375 >> cluster
rio@083:~$ echo 192.168.1.124:2375 >> cluster
rio@083:~$ cat cluster
192.168.1.83:2375
192.168.1.84:2375
192.168.1.124:2375
```

第二步：在083这台机器上面执行 `swarm manage` 这条命令：

```
rio@083:~$ sudo docker run -d -p 2376:2375 -v $(pwd)/cluster:/tmp/cluster swarm manage fi
364af1f25b776f99927b8ae26ca8db5a6fe8ab8cc1e4629a5a68b48951f598ad
```

使用 `docker ps` 来查看有没有启动成功：

```
rio@083:~$ sudo docker ps
CONTAINER ID        IMAGE               COMMAND             CREATED             STATUS
364af1f25b77       swarm:latest       "/swarm manage fi  About a minute ago Up Ab
```

可以看到，此时整个集群已经启动成功。

在使用这条命令的时候需要注意的是注意：这里一定要使用 `-v` 命令，因为 `cluster` 文件是在本机上面，启动的容器默认是访问不到的，所以要通过 `-v` 命令共享。

接下来的就可以在任何一台安装了 `docker` 的机器上面通过命令使用集群，同样的，在85这台机器上执行 `docker info` 命令查看集群的节点信息：

```
rio@s085:~$ sudo docker -H 192.168.1.83:2376 info
Containers: 9
Strategy: spread
Filters: affinity, health, constraint, port, dependency
Nodes: 3
  atsgxxx: 192.168.1.227:2375
    ↳ Containers: 0
    ↳ Reserved CPUs: 0 / 4
    ↳ Reserved Memory: 0 B / 2.052 GiB
  sclu083: 192.168.1.83:2375
    ↳ Containers: 2
    ↳ Reserved CPUs: 0 / 2
    ↳ Reserved Memory: 0 B / 4.054 GiB
  sclu084: 192.168.1.84:2375
    ↳ Containers: 7
    ↳ Reserved CPUs: 0 / 2
    ↳ Reserved Memory: 0 B / 4.053 GiB
```

swarm 调度策略

swarm支持多种调度策略来选择节点。每次在swarm启动container的时候，swarm会根据选择的调度策略来选择节点运行container。目前支持的有:spread,binpack和random。

在执行 `swarm manage` 命令启动 swarm 集群的时候可以通过 `--strategy` 参数来指定，默认的是spread。

spread和binpack策略会根据每台节点的可用CPU，内存以及正在运行的containers的数量来给各个节点分级，而random策略，顾名思义，他不会做任何的计算，只是单纯的随机选择一个节点来启动container。这种策略一般只做调试用。

使用spread策略，swarm会选择正在运行的container的数量最少的那个节点来运行container。这种情况会导致启动的container会尽可能的分布在不同的机器上运行，这样的好处就是如果有节点坏掉的时候不会损失太多的container。

binpack 则相反，这种情况下，swarm会尽可能的把所有的容器放在一台节点上面运行。这种策略会避免容器碎片化，因为他会把未使用的机器分配给更大的容器，带来的好处就是swarm会使用最少的节点运行最多的容器。

spread 策略

先来演示下 spread 策略的情况。

```

rio@083:~$ sudo docker run -d -p 2376:2375 -v $(pwd)/cluster:/tmp/cluster swarm manage --
7609ac2e463f435c271d17887b7d1db223a5d696bf3f47f86925c781c000cb60
ats@sclu083:~$ sudo docker ps
CONTAINER ID        IMAGE               COMMAND             CREATED             STATUS
7609ac2e463f       swarm:latest       "/swarm manage --str 6 seconds ago      Up 5 s

```

三台机器除了83运行了 Swarm之外，其他的都没有运行任何一个容器，现在在85这台节点上面在swarm集群上启动一个容器

```

rio@085:~$ sudo docker -H 192.168.1.83:2376 run --name node-1 -d -P redis
2553799f1372b432e9b3311b73e327915d996b6b095a30de3c91a47ff06ce981
rio@085:~$ sudo docker -H 192.168.1.83:2376 ps
CONTAINER ID        IMAGE               COMMAND             CREATED             STATUS
2553799f1372       redis:latest       /entrypoint.sh redis 24 minutes ago     Up Les

```

启动一个 redis 容器，查看结果

```

rio@085:~$ sudo docker -H 192.168.1.83:2376 run --name node-2 -d -P redis
7965a17fb943dc6404e2c14fb8585967e114addca068f233fc60c13bcf2190
rio@085:~$ sudo docker -H 192.168.1.83:2376 ps
CONTAINER ID        IMAGE               COMMAND             CREATED

```

```
7965a17fb943      redis:latest      /entrypoint.sh redis      Less than a second ago      Up 1 s
2553799f1372      redis:latest      /entrypoint.sh redis      29 minutes ago
```

再次启动一个 redis 容器，查看结果

```
rio@085:~$ sudo docker -H 192.168.1.83:2376 run --name node-3 -d -P redis
65e1ed758b53fbf441433a6cb47d288c51235257cf1bf92e04a63a8079e76bee
rio@085:~$ sudo docker -H 192.168.1.83:2376 ps
CONTAINER ID      IMAGE      COMMAND      CREATED
7965a17fb943      redis:latest      /entrypoint.sh redis      Less than a s
65e1ed758b53      redis:latest      /entrypoint.sh redis      25 minutes ago
2553799f1372      redis:latest      /entrypoint.sh redis      33 minutes ago
```

可以看到三个容器都是分布在不同的节点上面的。

binpack 策略

现在来看看binpack策略下的情况。在083上面执行命令：

```
rio@083:~$ sudo docker run -d -p 2376:2375 -v $(pwd)/cluster:/tmp/cluster swarm manage --
f1c9affd5a0567870a45a8eae57fec7c78f3825f3a53fd324157011aa0111ac5
```

现在在集群中启动三个 redis 容器，查看分布情况：

```
rio@085:~$ sudo docker -H 192.168.1.83:2376 run --name node-1 -d -P redis
18ceefa5e86f06025cf7c15919fa64a417a9d865c27d97a0ab4c7315118e348c
rio@085:~$ sudo docker -H 192.168.1.83:2376 run --name node-2 -d -P redis
7e778bde1a99c5cbe4701e06935157a6572fb8093fe21517845f5296c1a91bb2
rio@085:~$ sudo docker -H 192.168.1.83:2376 run --name node-3 -d -P redis
2195086965a783f0c2b2f8af65083c770f8bd454d98b7a94d0f670e73eea05f8
rio@085:~$ sudo docker -H 192.168.1.83:2376 ps
CONTAINER ID      IMAGE      COMMAND      CREATED      STATUS
2195086965a7      redis:latest      /entrypoint.sh redis      24 minutes ago      Up Les
7e778bde1a99      redis:latest      /entrypoint.sh redis      24 minutes ago      Up Les
18ceefa5e86f      redis:latest      /entrypoint.sh redis      25 minutes ago      Up 22
```

可以看到，所有的容器都是分布在同一个节点上运行的。

Swarm 过滤器

swarm 的调度器(scheduler)在选择节点运行容器的时候支持几种过滤器 (filter) :

Constraint,Affinity,Port,Dependency,Health

可以在执行 `swarm manage` 命令的时候通过 `--filter` 选项来设置。

Constraint Filter

constraint 是一个跟具体节点相关联的键值对，可以看做是每个节点的标签，这个标签可以在启动docker daemon的时候指定，比如

```
sudo docker -d --label label_name=label01
```

也可以写在docker的配置文件里面（在ubuntu上面是 `/etc/default/docker`）。

在本次试验中，给083添加标签`--label label_name=083`,084添加标签`--label label_name=084`,124添加标签`--label label_name=084`,

以083为例，打开`/etc/default/docker`文件，修改`DOCKER_OPTS`：

```
DOCKER_OPTS="-H 0.0.0.0:2375 -H unix:///var/run/docker.sock --label label_name=083"
```

在使用`docker run`命令启动容器的时候使用 `-e constraint:key=value` 的形式，可以指定container运行的节点。

比如我们想在84上面启动一个 redis 容器。

```
rio@085:~$ sudo docker -H 192.168.1.83:2376 run --name redis_1 -d -e constraint:label_name=084 fee1b7b9dde13d64690344c1f1a4c3f5556835be46b41b969e4090a083a6382d
```

注意，是两个等号，不是一个等号，这一点会经常被忽略。

接下来再在084这台机器上启动一个redis 容器。

```
rio@085:~$ sudo docker -H 192.168.1.83:2376 run --name redis_2 -d -e constraint:label_name=084
```

然后再在083这台机器上启动另外一个 redis 容器。

```
rio@085:~$ sudo docker -H 192.168.1.83:2376 run --name redis_3 -d -e constraint:label_name=083
```

现在来看下执行情况：

```

rio@085:~$ sudo docker -H 192.168.1.83:2376 ps
CONTAINER ID        IMAGE               COMMAND                  CREATED             STATUS
7786300b8d22       redis:latest       "/entrypoint.sh redi   15 minutes ago     Up 53
4968d617d9cd       redis:latest       "/entrypoint.sh redi   16 minutes ago     Up 2 m
fee1b7b9dde1       redis:latest       "/entrypoint.sh redi   19 minutes ago     Up 5 m
    
```

可以看到，执行结果跟预期的一样。

但是如果指定一个不存在的标签的话来运行容器会报错。

```

rio@085:~$ sudo docker -H 192.168.1.83:2376 run --name redis_0 -d -e constraint:label_name=0
FATA[0000] Error response from daemon: unable to find a node that satisfies label_name==0
    
```

Affinity Filter

通过使用 Affinity Filter，可以让一个容器紧挨着另一个容器启动，也就是说让两个容器在同一个节点上面启动。

现在其中一台机器上面启动一个 redis 容器。

```

rio@085:~$ sudo docker -H 192.168.1.83:2376 run -d --name redis redis
ea13eddf667992c5d8296557d3c282dd8484bd262c81e2b5af061cdd6c82158d
rio@085:~$ sudo docker -H 192.168.1.83:2376 ps
CONTAINER ID        IMAGE               COMMAND                  CREATED             STATUS
ea13eddf6679       redis:latest       /entrypoint.sh redis   24 minutes ago     Up Les
    
```

然后再次启动两个 redis 容器。

```

rio@085:~$ sudo docker -H 192.168.1.83:2376 run -d --name redis_1 -e affinity:container=bac50c2e955211047a745008fd1086eaa16d7ae4f33c192f50412e8dcd0a14cd
rio@085:~$ sudo docker -H 192.168.1.83:2376 run -d --name redis_1 -e affinity:container=bac50c2e955211047a745008fd1086eaa16d7ae4f33c192f50412e8dcd0a14cd
    
```

现在来查看下运行结果,可以看到三个容器都是在同一台机器上运行

```

rio@085:~$ sudo docker -H 192.168.1.83:2376 ps
CONTAINER ID        IMAGE               COMMAND                  CREATED             STATUS
449ed25ad239       redis:latest       /entrypoint.sh redis   24 minutes ago     Up Les
bac50c2e9552       redis:latest       /entrypoint.sh redis   25 minutes ago     Up 10
ea13eddf6679       redis:latest       /entrypoint.sh redis   28 minutes ago     Up 3 m
    
```

通过 `-e affinity:image=image_name` 命令可以指定只有已经下载了 `image_name` 镜像的机器才运行容器

```
sudo docker -H 192.168.1.83:2376 run -name redis1 -d -e affinity:image==redis redis
```

redis1 这个容器只会在已经下载了 redis 镜像的节点上运行。

```
sudo docker -H 192.168.1.83:2376 run -d --name redis -e affinity:image==~redis redis
```

这条命令达到的效果是：在有 redis 镜像的节点上面启动一个名字叫做 redis 的容器，如果每个节点上面都没有 redis 容器，就按照默认的策略启动 redis 容器。

Port Filter

Port 也会被认为是一个唯一的资源

```
sudo docker -H 192.168.1.83:2376 run -d -p 80:80 nginx
```

执行完这条命令，之后任何使用 80 端口的容器都是启动失败。

etcd

etcd 是 CoreOS 团队发起的一个管理配置信息和服务发现 (service discovery) 的项目，在这一章里面，我们将介绍该项目的目标，安装和使用，以及实现的技术。

什么是 etcd



etcd 是 CoreOS 团队于 2013 年 6 月发起的开源项目，它的目标是构建一个高可用的分布式键值（key-value）数据库，基于 Go 语言实现。我们知道，在分布式系统中，各种服务的配置信息的管理分享，服务的发现是一个很基本同时也是很重要的问题。CoreOS 项目就希望基于 etcd 来解决这一问题。

etcd 目前在 github.com/coreos/etcd 进行维护，即将发布 2.0.0 版本。

受到 [Apache ZooKeeper](#) 项目和 [doozer](#) 项目的启发，etcd 在设计的时候重点考虑了下面四个要素：

- 简单：支持 REST 风格的 HTTP+JSON API
- 安全：支持 HTTPS 方式的访问
- 快速：支持并发 1k/s 的写操作
- 可靠：支持分布式结构，基于 Raft 的一致性算法

注：*Apache ZooKeeper* 是一套知名的分布式系统中进行同步和一致性管理的工具。注：*doozer* 则是一个一致性分布式数据库。注：*Raft* 是一套通过选举主节点来实现分布式系统一致性的算法，相比于大名鼎鼎的 *Paxos* 算法，它的过程更容易被人理解，由 *Stanford* 大学的 *Diego Ongaro* 和 *John Ousterhout* 提出。更多细节可以参考 raftconsensus.github.io。

一般情况下，用户使用 etcd 可以在多个节点上启动多个实例，并添加它们为一个集群。同一个集群中的 etcd 实例将会保持彼此信息的一致性。

安装

etcd 基于 Go 语言实现，因此，用户可以从 [项目主页](#) 下载源代码自行编译，也可以下载编译好的二进制文件，甚至直接使用制作好的 Docker 镜像文件来体验。

二进制文件方式下载

编译好的二进制文件都在 github.com/coreos/etcd/releases 页面，用户可以选择需要的版本，或通过下载工具下载。

例如，下面的命令使用 curl 工具下载压缩包，并解压。

```
curl -L https://github.com/coreos/etcd/releases/download/v2.0.0-rc.1/etcd-v2.0.0-rc.1-linux-amd64.tar.gz
tar xzvf etcd-v2.0.0-rc.1-linux-amd64.tar.gz
cd etcd-v2.0.0-rc.1-linux-amd64
```

解压后，可以看到文件包括

```
$ ls
etcd  etcdctl  etcd-migrate  README-etcdctl.md  README.md
```

其中 etcd 是服务主文件，etcdctl 是提供给用户的命令客户端，etcd-migrate 负责进行迁移。

推荐通过下面的命令将三个文件都放到系统可执行目录 `/usr/local/bin/` 或 `/usr/bin/`。

```
$ sudo cp etcd* /usr/local/bin/
```

运行 etcd，将默认组建一个两个节点的集群。数据库服务端默认监听在 2379 和 4001 端口，etcd 实例监听在 2380 和 7001 端口。显示类似如下的信息：

```
$ ./etcd
2014/12/31 14:52:09 no data-dir provided, using default data-dir ./default.etcd
2014/12/31 14:52:09 etcd: listening for peers on http://localhost:2380
2014/12/31 14:52:09 etcd: listening for peers on http://localhost:7001
2014/12/31 14:52:09 etcd: listening for client requests on http://localhost:2379
2014/12/31 14:52:09 etcd: listening for client requests on http://localhost:4001
2014/12/31 14:52:09 etcdserver: name = default
2014/12/31 14:52:09 etcdserver: data dir = default.etcd
2014/12/31 14:52:09 etcdserver: snapshot count = 10000
2014/12/31 14:52:09 etcdserver: advertise client URLs = http://localhost:2379,http://localhost:2380
2014/12/31 14:52:09 etcdserver: initial advertise peer URLs = http://localhost:2380,http://localhost:7001
2014/12/31 14:52:09 etcdserver: initial cluster = default=http://localhost:2380,default=http://localhost:7001
2014/12/31 14:52:10 etcdserver: start member ce2a822cea30bfca in cluster 7e27652122e8b2ae
2014/12/31 14:52:10 raft: ce2a822cea30bfca became follower at term 0
2014/12/31 14:52:10 raft: newRaft ce2a822cea30bfca [peers: [], term: 0, commit: 0, lastindex: 0, lastlogterm: 0, lastlogid: 0]
2014/12/31 14:52:10 raft: ce2a822cea30bfca became follower at term 1
```

```
2014/12/31 14:52:10 etcdserver: added local member ce2a822cea30bfca [http://localhost:238
2014/12/31 14:52:11 raft: ce2a822cea30bfca is starting a new election at term 1
2014/12/31 14:52:11 raft: ce2a822cea30bfca became candidate at term 2
2014/12/31 14:52:11 raft: ce2a822cea30bfca received vote from ce2a822cea30bfca at term 2
2014/12/31 14:52:11 raft: ce2a822cea30bfca became leader at term 2
2014/12/31 14:52:11 raft.node: ce2a822cea30bfca elected leader ce2a822cea30bfca at term 2
2014/12/31 14:52:11 etcdserver: published {Name:default ClientURLs:[http://localhost:2379
```

此时，可以使用 `etcdctl` 命令进行测试，设置和获取键值 `testkey: "hello world"`，检查 `etcd` 服务是否启动成功：

```
$ ./etcdctl set testkey "hello world"
hello world
$ ./etcdctl get testkey
hello world
```

说明 `etcd` 服务已经成功启动了。

当然，也可以通过 HTTP 访问本地 2379 或 4001 端口的方式来进行操作，例如查看 `testkey` 的值：

```
$ curl -L http://localhost:4001/v2/keys/testkey
{"action": "get", "node": {"key": "/testkey", "value": "hello world", "modifiedIndex": 3, "created
```

Docker 镜像方式下载

镜像名称为 `quay.io/coreos/etcd:v2.0.0_rc.1`，可以通过下面的命令启动 `etcd` 服务监听到 4001 端口。

```
$ sudo docker run -p 4001:4001 -v /etc/ssl/certs:/etc/ssl/certs/ quay.io/coreos/etcd:v2.
```

使用 etcdctl

etcdctl 是一个命令行客户端，它能提供一些简洁的命令，供用户直接跟 etcd 服务打交道，而无需基于 HTTP API 方式。这在某些情况下将很方便，例如用户对服务进行测试或者手动修改数据库内容。我们也推荐在刚接触 etcd 时通过 etcdctl 命令来熟悉相关的操作，这些操作跟 HTTP API 实际上是对应的。

etcd 项目二进制发行包中已经包含了 etcdctl 工具，没有的话，可以从 github.com/coreos/etcd/releases 下载。

etcdctl 支持如下的命令，大体上分为数据库操作和非数据库操作两类，后面将分别进行解释。

```
$ etcdctl -h
NAME:
  etcdctl - A simple command line client for etcd.

USAGE:
  etcdctl [global options] command [command options] [arguments...]

VERSION:
  2.0.0-rc.1

COMMANDS:
  backup      backup an etcd directory
  mk          make a new key with a given value
  mkdir       make a new directory
  rm          remove a key
  rmdir       removes the key if it is an empty directory or a key-value pair
  get         retrieve the value of a key
  ls          retrieve a directory
  set         set the value of a key
  setdir      create a new or existing directory
  update      update an existing key with a given value
  updatedir   update an existing directory
  watch       watch a key for changes
  exec-watch  watch a key for changes and exec an executable
  member      member add, remove and list subcommands
  help, h     Shows a list of commands or help for one command

GLOBAL OPTIONS:
  --debug                output cURL commands which can be used to reproduce the request
  --no-sync              don't synchronize cluster information before sending request
  --output, -o 'simple'  output response in the given format (`simple` or `json`)
  --peers, -C            a comma-delimited list of machine addresses in the cluster (default)
  --cert-file            identify HTTPS client using this SSL certificate file
  --key-file             identify HTTPS client using this SSL key file
  --ca-file              verify certificates of HTTPS-enabled servers using this CA bundle
  --help, -h            show help
  --version, -v         print the version
```

数据库操作

数据库操作围绕对键值和目录的 CRUD（符合 REST 风格的一套操作：Create）完整生命周期的管理。

etcd 在键的组织上采用了层次化的空间结构（类似于文件系统中目录的概念），用户指定的键可以为单独的名字，如 `testkey`，此时实际上放在根目录 `/` 下面，也可以为指定目录结构，如 `cluster1/node2/testkey`，则将创建相应的目录结构。

注：*CRUD* 即 *Create, Read, Update, Delete*，是符合 *REST* 风格的一套 *API* 操作。

set

指定某个键的值。例如

```
$ etcdctl set /testdir/testkey "Hello world"
Hello world
```

支持的选项包括：

```
--ttl '0'           该键值的超时时间（单位为秒），不配置（默认为 0）则永不超时
--swap-with-value value 若该键现在的值是 value，则进行设置操作
--swap-with-index '0'  若该键现在的索引值是指定索引，则进行设置操作
```

get

获取指定键的值。例如

```
$ etcdctl set testkey hello
hello
$ etcdctl update testkey world
world
```

当键不存在时，则会报错。例如

```
$ etcdctl get testkey2
Error: 100: Key not found (/testkey2) [1]
```

支持的选项为

```
--sort      对结果进行排序
--consistent 将请求发给主节点，保证获取内容的一致性
```

update

当键存在时，更新值内容。例如

```
$ etcdctl set testkey hello
hello
$ etcdctl update testkey world
world
```

当键不存在时，则会报错。例如

```
$ etcdctl update testkey2 world
Error: 100: Key not found (/testkey2) [1]
```

支持的选项为

```
--ttl '0'    超时时间（单位为秒），不配置（默认为 0）则永不超时
```

rm

删除某个键值。例如

```
$ etcdctl rm testkey
```

当键不存在时，则会报错。例如

```
$ etcdctl rm testkey2
Error: 100: Key not found (/testkey2) [8]
```

支持的选项为

```
--dir        如果键是个空目录或者键值对则删除
--recursive  删除目录和所有子键
--with-value  检查现有的值是否匹配
--with-index '0'  检查现有的 index 是否匹配
```

mk

如果给定的键不存在，则创建一个新的键值。例如

```
$ etcdctl mk /testdir/testkey "Hello world"
Hello world
```

当键存在的时候，执行该命令会报错，例如

```
$ etcdctl set testkey "Hello world"
Hello world
$ ./etcdctl mk testkey "Hello world"
Error: 105: Key already exists (/testkey) [2]
```

支持的选项为

使用 etcdctl

```
--ttl '0'  超时时间（单位为秒），不配置（默认为 0）则永不超时
```

mkdir

如果给定的键目录不存在，则创建一个新的键目录。例如

```
$ etcdctl mkdir testdir
```

当键目录存在的时候，执行该命令会报错，例如

```
$ etcdctl mkdir testdir
$ etcdctl mkdir testdir
Error: 105: Key already exists (/testdir) [7]
```

支持的选项为

```
--ttl '0'  超时时间（单位为秒），不配置（默认为 0）则永不超时
```

setdir

创建一个键目录，无论存在与否。

支持的选项为

```
--ttl '0'  超时时间（单位为秒），不配置（默认为 0）则永不超时
```

updatedir

更新一个已经存在的目录。支持的选项为

```
--ttl '0'  超时时间（单位为秒），不配置（默认为 0）则永不超时
```

rmdir

删除一个空目录，或者键值对。

若目录不空，会报错

```
$ etcdctl set /dir/testkey hi
hi
$ etcdctl rmdir /dir
Error: 108: Directory not empty (/dir) [13]
```

ls

列出目录（默认为根目录）下的键或者子目录，默认不显示子目录中内容。

例如

```
$ ./etcdctl set testkey 'hi'
hi
$ ./etcdctl set dir/test 'hello'
hello
$ ./etcdctl ls
/testkey
/dir
$ ./etcdctl ls dir
/dir/test
```

支持的选项包括

```
--sort      将输出结果排序
--recursive 如果目录下有子目录，则递归输出其中的内容
-p          对于输出为目录，在最后添加 `/` 进行区分
```

非数据库操作

backup

备份 etcd 的数据。

支持的选项包括

```
--data-dir      etcd 的数据目录
--backup-dir    备份到指定路径
```

watch

监测一个键值的变化，一旦键值发生更新，就会输出最新的值并退出。

例如，用户更新 testkey 键值为 Hello world。

```
$ etcdctl watch testkey
Hello world
```

支持的选项包括

```
--forever      一直监测，直到用户按 `CTRL+C` 退出
--after-index '0' 在指定 index 之前一直监测
--recursive    返回所有的键值和子键值
```

exec-watch

监测一个键值的变化，一旦键值发生更新，就执行给定命令。

例如，用户更新 testkey 键值。

```
$ etcdctl exec-watch testkey -- sh -c 'ls'
default.etcd
Documentation
etcd
etcdctl
etcd-migrate
README-etcdctl.md
README.md
```

支持的选项包括

```
--after-index '0'    在指定 index 之前一直监测
--recursive          返回所有的键值和子键值
```

member

通过 list、add、remove 命令列出、添加、删除 etcd 实例到 etcd 集群中。

例如本地启动一个 etcd 服务实例后，可以用如下命令进行查看。

```
$ etcdctl member list
ce2a822cea30bfca: name=default peerURLs=http://localhost:2380,http://localhost:7001 clien
```

命令选项

- `--debug` 输出 cURL 命令，显示执行命令的时候发起的请求
- `--no-sync` 发出请求之前不同步集群信息
- `--output, -o 'simple'` 输出内容的格式 (`simple` 为原始信息，`json` 为进行json格式解码，易读性好一些)
- `--peers, -c` 指定集群中的同伴信息，用逗号隔开 (默认为: "127.0.0.1:4001")
- `--cert-file` HTTPS 下客户端使用的 SSL 证书文件
- `--key-file` HTTPS 下客户端使用的 SSL 密钥文件
- `--ca-file` 服务端使用 HTTPS 时，使用 CA 文件进行验证
- `--help, -h` 显示帮助命令信息
- `--version, -v` 打印版本信息

Fig

在你的应用里面添加一个 `fig.yml` 文件，并指定一些简单的内容，执行 `fig up` 它就能帮你快速建立一个容器。目前已经正式更名为 [Compose](#)。

快速搭建基于 Docker 的隔离开发环境

使用 `Dockerfile` 文件指定你的应用环境，让它能在任意地方复制使用：

```
FROM python:2.7
ADD . /code
WORKDIR /code
RUN pip install -r requirements.txt
```

在 `fig.yml` 文件中指定应用使用的不同服务，让它们能够在一个独立的环境中一起运行：

```
web:
  build: .
  command: python app.py
  links:
    - db
  ports:
    - "8000:8000"
db:
  image: postgres
```

*注意不需要再额外安装 *Postgres* 了！

接着执行命令 `fig up`，然后 Fig 就会启动并运行你的应用了。



Fig 可用的命令有:

- 启动、停止, 和重建服务
- 查看服务的运行状态
- 查看运行中的服务的输入日志
- 对服务发送命令

快速上手

我们试着让一个基本的 Python web 应用运行在 Fig 上。这个实验假设你已经知道一些 Python 知识, 如果你不熟悉, 但清楚概念上的东西也是没有问题的。

首先, [安装 Docker 和 Fig](#)

为你的项目创建一个目录

```
$ mkdir figtest  
$ cd figtest
```

进入目录, 创建 `app.py`, 这是一个能够让 Redis 上的一个值自增的简单 web 应用, 基于 Flask 框架。

```
from flask import Flask  
from redis import Redis
```

```
import os
app = Flask(__name__)
redis = Redis(host='redis', port=6379)

@app.route('/')
def hello():
    redis.incr('hits')
    return 'Hello World! I have been seen %s times.' % redis.get('hits')

if __name__ == "__main__":
    app.run(host="0.0.0.0", debug=True)
```

在 `requirements.txt` 文件中指定应用的 Python 依赖包。

```
flask
redis
```

下一步我们要创建一个包含应用所有依赖的 Docker 镜像，这里将阐述怎么通过 `Dockerfile` 文件来创建。

```
FROM python:2.7
ADD . /code
WORKDIR /code
RUN pip install -r requirements.txt
```

以上的内容首先告诉 Docker 在容器里面安装 Python，代码的路径还有 Python 依赖包。关于 Dockerfile 的更多信息可以查看 [镜像创建](#) 和 [Dockerfile 使用](#)

接着我们通过 `fig.yml` 文件指定一系列的服务：

```
web:
  build: .
  command: python app.py
  ports:
    - "5000:5000"
  volumes:
    - ./code
  links:
    - redis
redis:
  image: redis
```

这里指定了两个服务：

- web 服务，通过当前目录的 `Dockerfile` 创建。并且说明了在容器里面执行 `python app.py` 命令，转发在容器里开放的 5000 端口到本地主机的 5000 端口，连接 Redis 服务，并且挂载当前目录到容器里面，这样我们就可以不用重建镜像也能直接使用代码。
- redis 服务，我们使用公用镜像 `redis`。* 现在如果执行 `fig up` 命令，它就会拉取 `redis` 镜像，启动所有的服务。

```
$ fig up
Pulling image redis...
Building web...
Starting figtest_redis_1...
Starting figtest_web_1...
redis_1 | [8] 02 Jan 18:43:35.576 # Server started, Redis version 2.8.3
web_1   | * Running on http://0.0.0.0:5000/
```

这个 web 应用已经开始在你的 docker 守护进程里面监听着 5000 端口了（如果你有使用 boot2docker，执行 `boot2docker ip`，就会看到它的地址）。

如果你想要在后台运行你的服务，可以在执行 `fig up` 命令的时候添加 `-d` 参数，然后使用 `fig ps` 查看有什么进程在运行。

```
$ fig up -d
Starting figtest_redis_1...
Starting figtest_web_1...
$ fig ps
-----
Name                Command                State    Ports
-----
figtest_redis_1     /usr/local/bin/run     Up
figtest_web_1       /bin/sh -c python app.py Up        5000->5000/tcp
```

`fig run` 指令可以帮你向服务发送命令。例如：查看 web 服务可以获取到的环境变量：

```
$ fig run web env
```

执行帮助命令 `fig --help` 查看其它可用的参数。

假设你使用了 `fig up -d` 启动 Fig，可以通过以下命令停止你的服务：

```
$ fig stop
```

以上内容或多或少的讲述了如何使用 Fig。通过查看下面的引用章节可以了解到关于命令、配置和环境变量的更多细节。如果你有任何想法或建议，可以在 [GitHub](#) 上提出。

安装 Fig

首先，安装 1.3 或者更新的 Docker 版本。

如果你的工作环境是 OS X，可以通过查看 [Mac 安装指南\(英文\)](#)，完成安装 Docker 和 boot2docker。一旦 boot2docker 运行后，执行以下指令设置一个环境变量，接着 Fig 就可以和它交互了。

```
$(boot2docker shellinit)
```

*如果想避免重启后重新设置，可以把上面的命令加到你的 `~/.bashrc` 文件里。

关于 `Ubuntu` 还有 `其它的平台` 的安装，可以参照 [Ubuntu 安装指南\(中文\)](#) 以及 [官方安装手册\(英文\)](#)。

下一步，安装 Fig：

```
curl -L https://github.com/docker/fig/releases/download/1.0.1/fig-`uname -s`-`uname -m` >
```

*如果你的 `Docker` 是管理员身份安装，以上命令可能也需要相同的身份。

目前 Fig 的发行版本只支持 OS X 和 64 位的 Linux 系统。但因为它是用 Python 语言写的，所以对于其它平台上的用户，可以通过 Python 安装包来完成安装（支持的系统同样适用）。

```
$ sudo pip install -U fig
```

到这里就已经完成了。执行 `fig --version`，确认能够正常运行。

Fig客户端参考

大部分命令都可以运行在一个或多个服务上。如果没有特别的说明，这个命令则可以应用在所有的服务上。

执行 `fig [COMMAND] --help` 查看所有使用说明。

选项

`--verbose`

显示更多信息。

`--version`

打印版本并退出。

`-f, --file FILE`

使用特定的Fig文件，默认使用fig.yml。

`-p, --project-name NAME`

使用特定的项目名称，默认使用文件夹名称。

命令

`build`

构建或重新构建服务。

服务一旦构建后，将会标记为`project_service`，例如`figtest_db`。如果修改服务的`Dockerfile`或构建目录信息，你可以运行`fig build`来重新构建。

`help`

获得一个命令的帮助。

`kill`

强制停止服务容器。

`logs`

查看服务的输出。

`port`

打印端口绑定的公共端口。

`ps`

命令参考

列出所有容器。

```
pull
```

拉取服务镜像。

```
rm
```

删除停止的服务容器。

```
run
```

在一个服务上执行一个命令。

例如：

```
$ fig run web python manage.py shell
```

默认情况下，链接的服务将会启动，除非这些服务已经在运行中。

一次性命令会在使用与服务的普通容器相同的配置的新容器中开始运行，然后卷、链接等等都将会按照期望创建。与普通容器唯一的不同就是，这个命令将会覆盖原有的命令，如果端口有冲突则不会创建。

链接还可以在一次性命令和那个服务的其他容器间创建，然后你可以像下面一样进行一些操作：

```
$ fig run db psql -h db -U docker
```

如果你不希望在执行一次性命令时启动链接的容器，可以指定`--no-deps`选项：

```
$ fig run --no-deps web python manage.py shell
```

```
scale
```

设置一个服务需要运行的容器个数。

通过`service=num`的参数来设置数量。例如：

```
$ fig scale web=2 worker=3
```

```
start
```

启动一个服务已经存在的容器。

```
stop
```

停止一个已经运行的容器，但不删除它。通过 `fig start` 可以再次启动这些容器。

```
up
```

命令参考

构建，（重新）创建，启动，链接一个服务的容器。

链接的服务都将会启动，除非他们已经运行。

默认情况，`fig up` 将会聚合每个容器的输出，而且如果容器已经存在，所有容器将会停止。如果你运行 `fig up -d`，将会在后台启动并运行所有的容器。

默认情况，如果这个服务的容器已经存在，`fig up` 将会停止并重新创建他们（保持使用volumes-from挂载的卷），以保证 `fig.yml` 的修改生效。如果你不想容器被停止并重新创建，可以使用 `fig up --no-recreate`。如果需要的话，这样将会启动已经停止的容器。

环境变量

环境变量可以用来配置Fig的行为。

变量以DOCKER_开头，它们和用来配置Docker命令行客户端的使用一样。如果你在使用 `boot2docker`，`$(boot2docker shellinit)` 将会设置它们为正确的值。

`FIG_PROJECT_NAME`

设置通过Fig启动的每一个容器前添加的项目名称.默认是当前工作目录的名字。

`FIG_FILE`

设置要使用的 `fig.yml` 的路径。默认路径是当前工作目录。

`DOCKER_HOST`

设置docker进程的URL。默认docker client使用 `unix:///var/run/docker.sock`。

`DOCKER_TLS_VERIFY`

如果设置不为空的字符，允许和进程进行 TLS 通信。

`DOCKER_CERT_PATH`

配置 `ca.pem` 的路径，`cert.pem` 和 `key.pem` 文件用来进行TLS验证.默认路径是 `~/.docker`。

fig.yml 参考

每个在 `fig.yml` 定义的服务都需要指定一个镜像或镜像的构建内容。像 `docker run` 的命令行一样，其它内容是可选的。

`docker run` 在 `Dockerfile` 中设置的选项(例如：`CMD`，`EXPOSE`，`VOLUME`，`ENV`)作为已经提供的默认设置 - 你不需要在 `fig.yml` 中重新设置。

image

这里可以设置为标签或镜像ID的一部分。它可以是本地的，也可以是远程的 - 如果镜像在本地不存在，`Fig` 将会尝试拉去这个镜像。

```
image: ubuntu
image: orchardup/postgresql
image: a4bc65fd
```

build

指定 `Dockerfile` 所在文件夹的路径。`Fig` 将会构建这个镜像并给它生成一个名字，然后使用这个镜像。

```
build: /path/to/build/dir
```

command

覆盖默认的命令。

```
command: bundle exec thin -p 3000
```

links

在其它的服务中连接容器。使用服务名称（经常也作为别名）或服务名称加服务别名（`SERVICE:ALIAS`）都可以。

```
links:
- db
- db:database
- redis
```

可以在服务的容器中的 `/etc/hosts` 里创建别名。例如：

```
172.17.2.186 db
172.17.2.186 database
172.17.2.187 redis
```

环境变量也将被创建 - 细节查看环境变量参考章节。

ports

暴露端口。使用宿主和容器 (HOST:CONTAINER) 或者仅仅容器的端口 (宿主将会随机选择端口) 都可以。

注：当使用 HOST:CONTAINER 格式来映射端口时，如果你使用的容器端口小于60你可能会得到错误的结果，因为 YAML 将会解析 xx:yy 这种数字格式为60进制。所以我们建议用字符指定你得端口映射。

```
ports:
  - "3000"
  - "8000:8000"
  - "49100:22"
  - "127.0.0.1:8001:8001"
```

expose

暴露不发布到宿主机的端口 - 它们只被连接的服务访问。仅仅内部的端口可以被指定。

```
expose:
  - "3000"
  - "8000"
```

volumes

卷挂载路径设置。可以设置宿主机路径 (HOST:CONTAINER) 或访问模式 (HOST:CONTAINER:ro) 。

```
volumes:
  - /var/lib/mysql
  - cache:/tmp/cache
  - ~/configs:/etc/configs/:ro
```

volumes_from

从另一个服务或容器挂载所有卷。

```
volumes_from:
  - service_name
  - container_name
```

environment

设置环境变量。你可以使用数组或字典两种格式。

环境变量在运行 Fig 的机器上被解析成一个key。它有助于安全和指定的宿主值。

```
environment:
  RACK_ENV: development
  SESSION_SECRET:

environment:
  - RACK_ENV=development
  - SESSION_SECRET
```

net

设置网络模式。使用和 `docker client` 的 `--net` 参数一样的值。

```
net: "bridge"
net: "none"
net: "container:[name or id]"
net: "host"
```

dns

配置DNS服务器。它可以是一个值，也可以是一个列表。

```
dns: 8.8.8.8
dns:
  - 8.8.8.8
  - 9.9.9.9
```

`working_dir`, `entrypoint`, `user`, `hostname`, `domainname`, `mem_limit`, `privileged`

这些都是和 `docker run` 对应的一个值。

```
working_dir: /code
entrypoint: /code/entrypoint.sh
user: postgresql

hostname: foo
domainname: foo.com

mem_limit: 1000000000
privileged: true
```

环境变量参考

*注意: 现在已经不推荐使用环境变量链接服务。替代方案是使用链接名称（默认就是被连接的服务名字）作为主机名来链接。详情查看 [fig.yml](#) 章节。

Fig 使用 Docker 链接来暴露一个服务的容器给其它容器。每一个链接的容器会注入一组以容器名称的大写字母开头得环境变量。

查看一个服务有那些有效的环境变量可以执行 `fig run SERVICE env`。

`name_PORT`

完整URL, 例如: `DB_PORT=tcp://172.17.0.5:5432`

`name_PORT_num_protocol`

完整URL, 例如: `DB_PORT_5432_TCP=tcp://172.17.0.5:5432`

`name_PORT_num_protocol_ADDR`

容器的IP地址, 例如: `DB_PORT_5432_TCP_ADDR=172.17.0.5`

`name_PORT_num_protocol_PORT`

暴露端口号, 例如: `DB_PORT_5432_TCP_PORT=5432`

`name_PORT_num_protocol_PROTO`

协议 (tcp 或 udp), 例如: `DB_PORT_5432_TCP_PROTO=tcp`

`name_NAME`

完整合格的容器名称, 例如: `DB_1_NAME=/myapp_web_1/myapp_db_1`

使用 Django 入门 Fig

我们现在将使用 Fig 配置并运行一个 Django/PostgreSQL 应用。在此之前，先确保 Fig 已经 [安装](#)。

在一切工作开始前，需要先设置好三个必要的文件。

第一步，因为应用将要运行在一个满足所有环境依赖的 Docker 容器里面，那么我们可以通过编辑 `Dockerfile` 文件来指定 Docker 容器要安装内容。内容如下：

```
FROM python:2.7
ENV PYTHONUNBUFFERED 1
RUN mkdir /code
WORKDIR /code
ADD requirements.txt /code/
RUN pip install -r requirements.txt
ADD . /code/
```

以上内容指定应用将使用安装了 Python 以及必要依赖包的镜像。更多关于如何编写 Dockerfile 文件的信息可以查看 [镜像创建](#) 和 [Dockerfile 使用](#)。

第二步，在 `requirements.txt` 文件里面写明需要安装的具体依赖包名。

```
Django
psycopg2
```

就是这么简单。

第三步，`fig.yml` 文件将把所有的东西关联起来。它描述了应用的构成（一个 web 服务和一个数据库）、使用的 Docker 镜像、镜像之间的连接、挂载到容器的卷，以及服务开放的端口。

```
db:
  image: postgres
web:
  build: .
  command: python manage.py runserver 0.0.0.0:8000
  volumes:
    - ../code
  ports:
    - "8000:8000"
  links:
    - db
```

查看 [fig.yml 章节](#) 了解更多详细的工作机制。

现在我们就可以使用 `fig run` 命令启动一个 Django 应用了。

```
$ fig run web django-admin.py startproject figexample .
```

Fig 会先使用 `Dockerfile` 为 web 服务创建一个镜像，接着使用这个镜像在容器里运行 `django-admin.py startproject figexample .` 指令。

这将在当前目录生成一个 Django 应用。

```
$ ls
Dockerfile      fig.yml          figexample      manage.py       requirements.txt
```

首先，我们要为应用设置好数据库的连接信息。用以下内容替换 `figexample/settings.py` 文件中 `DATABASES = ...` 定义的节点内容。

```
DATABASES = {
    'default': {
        'ENGINE': 'django.db.backends.postgresql_psycopg2',
        'NAME': 'postgres',
        'USER': 'postgres',
        'HOST': 'db',
        'PORT': 5432,
    }
}
```

这些信息是在 `postgres` Docker 镜像固定设置好的。

然后，运行 `fig up`：

```
Recreating myapp_db_1...
Recreating myapp_web_1...
Attaching to myapp_db_1, myapp_web_1
myapp_db_1 |
myapp_db_1 | PostgreSQL stand-alone backend 9.1.11
myapp_db_1 | 2014-01-27 12:17:03 UTC LOG:  database system is ready to accept connections
myapp_db_1 | 2014-01-27 12:17:03 UTC LOG:  autovacuum launcher started
myapp_web_1 | Validating models...
myapp_web_1 |
myapp_web_1 | 0 errors found
myapp_web_1 | January 27, 2014 - 12:12:40
myapp_web_1 | Django version 1.6.1, using settings 'figexample.settings'
myapp_web_1 | Starting development server at http://0.0.0.0:8000/
myapp_web_1 | Quit the server with CONTROL-C.
```

这个 web 应用已经开始在你的 docker 守护进程里监听 5000 端口了（如果你有使用 `boot2docker`，执行 `boot2docker ip`，就会看到它的地址）。

你还可以在 Docker 上运行其它的管理命令，例如对于同步数据库结构这种事，在运行完 `fig up` 后，在另外一个终端运行以下命令即可：

```
$ fig run web python manage.py syncdb
```

使用 Rail 入门 Fig

我们现在将使用 Fig 配置并运行一个 Rails/PostgreSQL 应用。在开始之前，先确保 Fig 已经 [安装](#)。

在一切工作开始前，需要先设置好三个必要的文件。

首先，因为应用将要运行在一个满足所有环境依赖的 Docker 容器里面，那么我们可以通过编辑 `Dockerfile` 文件来指定 Docker 容器要安装内容。内容如下：

```
FROM ruby
RUN apt-get update -qq && apt-get install -y build-essential libpq-dev
RUN mkdir /myapp
WORKDIR /myapp
ADD Gemfile /myapp/Gemfile
RUN bundle install
ADD . /myapp
```

以上内容指定应用将使用安装了 Ruby、Bundler 以及其依赖件的镜像。更多关于如何编写 Dockerfile 文件的信息可以查看 [镜像创建](#) 和 [Dockerfile 使用](#)。下一步，我们需要一个引导加载 Rails 的文件 `Gemfile`。等一会儿它还会被 `rails new` 命令覆盖重写。

```
source 'https://rubygems.org'
gem 'rails', '4.0.2'
```

最后，`fig.yml` 文件才是最神奇的地方。`fig.yml` 文件将把所有的东西关联起来。它描述了应用的构成（一个 web 服务和一个数据库）、每个镜像的来源（数据库运行在使用预定义的 PostgreSQL 镜像，web 应用侧将从本地目录创建）、镜像之间的连接，以及服务开放的端口。

```
db:
  image: postgres
  ports:
    - "5432"
web:
  build: .
  command: bundle exec rackup -p 3000
  volumes:
    - ./myapp
  ports:
    - "3000:3000"
  links:
    - db
```

所有文件就绪后，我们就可以通过使用 `fig run` 命令生成应用的骨架了。

```
$ fig run web rails new . --force --database=postgresql --skip-bundle
```

Fig 会先使用 `Dockerfile` 为 web 服务创建一个镜像，接着使用这个镜像在容器里运行 `rails new` 和它

之后的命令。一旦这个命令运行完后，应该就可以看一个崭新的应用已经生成了。

```
$ ls
Dockerfile  app          fig.yml      tmp
Gemfile     bin          lib          vendor
Gemfile.lock config       log
README.rdoc config.ru    public
Rakefile    db          test
```

在新的 `Gemfile` 文件去掉加载 `therubyracer` 的行的注释，这样我们便可以使用 Javascript 运行环境：

```
gem 'therubyracer', platforms: :ruby
```

现在我们已经有一个新的 `Gemfile` 文件，需要再重新创建镜像。（这个步骤会改变 `Dockerfile` 文件本身，仅仅需要重建一次）。

```
$ fig build
```

应用现在就可以启动了，但配置还未完成。Rails 默认读取的数据库目标是 `localhost`，我们需要手动指定容器的 `db`。同样的，还需要把用户名修改成和 `postgres` 镜像预定的一致。打开最新生成的 `database.yml` 文件。用以下内容替换：

```
development: &default
  adapter: postgresql
  encoding: unicode
  database: postgres
  pool: 5
  username: postgres
  password:
  host: db

test:
  <<: *default
  database: myapp_test
```

现在就可以启动应用了。

```
$ fig up
```

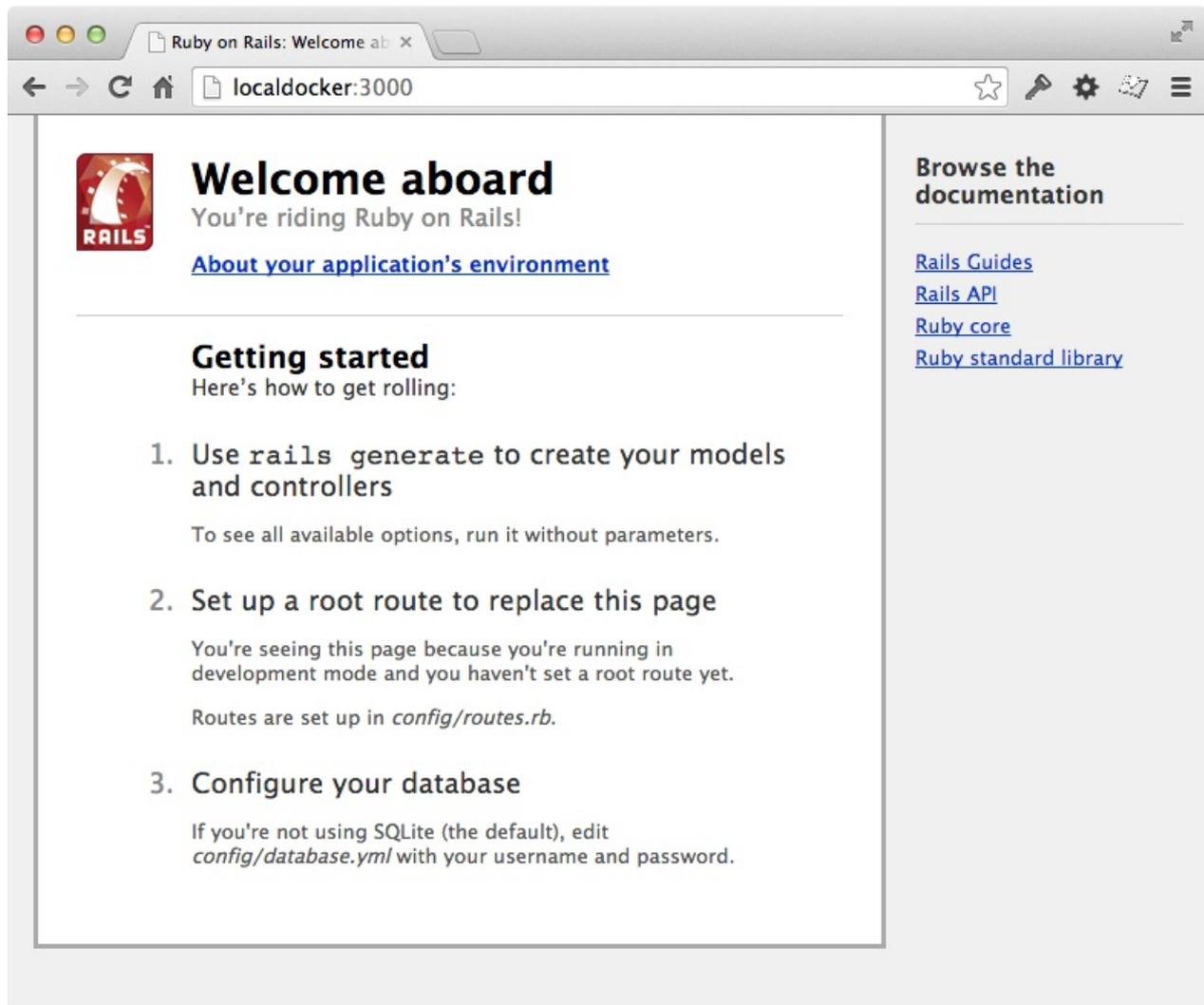
如果一切正常，你应该可以看到 PostgreSQL 的输出，几秒后可以看到这样的重复信息：

```
myapp_web_1 | [2014-01-17 17:16:29] INFO WEBrick 1.3.1
myapp_web_1 | [2014-01-17 17:16:29] INFO ruby 2.0.0 (2013-11-22) [x86_64-linux-gnu]
myapp_web_1 | [2014-01-17 17:16:29] INFO WEBrick::HTTPServer#start: pid=1 port=3000
```

最后，我们需要做的是创建数据库，打开另一个终端，运行：

```
$ fig run web rake db:create
```

这个 web 应用已经开始在你的 docker 守护进程里面监听着 3000 端口了（如果你有使用 boot2docker，执行 `boot2docker ip`，就会看到它的地址）。



使用 Wordpress 入门 Fig

Fig 让 Wordpress 运行在一个独立的环境中很简易。 [安装 Fig](#)，然后下载 Wordpress 到当前目录：

```
wordpress.org/latest.tar.gz | tar -xvzf -
```

这将会创建一个叫 wordpress 目录，你也可以重命名成你想要的名字。在目录里面，创建一个 `Dockerfile` 文件，定义应用的运行环境：

```
FROM orchardup/php5
ADD . /code
```

以上内容告诉 Docker 创建一个包含 PHP 和 Wordpress 的镜像。更多关于如何编写 Dockerfile 文件的信息可以查看 [镜像创建](#) 和 [Dockerfile 使用](#)。

下一步， `fig.yml` 文件将开启一个 web 服务和一个独立的 MySQL 实例：

```
web:
  build: .
  command: php -S 0.0.0.0:8000 -t /code
  ports:
    - "8000:8000"
  links:
    - db
  volumes:
    - ./code
db:
  image: orchardup/mysql
  environment:
    MYSQL_DATABASE: wordpress
```

要让这个应用跑起来还需要两个文件。第一个， `wp-config.php`，它是一个标准的 Wordpress 配置文件，有一点需要修改的是把数据库的配置指向 `db` 容器。

```
<?php
define('DB_NAME', 'wordpress');
define('DB_USER', 'root');
define('DB_PASSWORD', '');
define('DB_HOST', "db:3306");
define('DB_CHARSET', 'utf8');
define('DB_COLLATE', '');

define('AUTH_KEY',          'put your unique phrase here');
define('SECURE_AUTH_KEY',  'put your unique phrase here');
define('LOGGED_IN_KEY',    'put your unique phrase here');
define('NONCE_KEY',        'put your unique phrase here');
define('AUTH_SALT',        'put your unique phrase here');
define('SECURE_AUTH_SALT', 'put your unique phrase here');
```

```

define('LOGGED_IN_SALT', 'put your unique phrase here');
define('NONCE_SALT', 'put your unique phrase here');

$table_prefix = 'wp_';
define('WPLANG', '');
define('WP_DEBUG', false);

if ( !defined('ABSPATH') )
    define('ABSPATH', dirname(__FILE__) . '/');

require_once(ABSPATH . 'wp-settings.php');

```

第二个， `router.php` ，它告诉 PHP 内置的服务器怎么运行 Wordpress:

```

<?php

$root = $_SERVER['DOCUMENT_ROOT'];
chdir($root);
$path = '/'.ltrim(parse_url($_SERVER['REQUEST_URI'])['path'],'/');
set_include_path(get_include_path().':'.__DIR__);
if(file_exists($root.$path))
{
    if(is_dir($root.$path) && substr($path,strlen($path) - 1, 1) !== '/')
        $path = rtrim($path, '/').'/index.php';
    if(strpos($path, '.php') === false) return false;
    else {
        chdir(dirname($root.$path));
        require_once $root.$path;
    }
}
}else include_once 'index.php';

```

这些配置文件就绪后，在你的 Wordpress 目录里面执行 `fig up` 指令，Fig 就会拉取镜像再创建我们所需要的镜像，然后启动 web 和数据库容器。接着访问 docker 守护进程监听的 8000 端口就能看你的 Wordpress 网站了。（如果你有使用 `boot2docker`，执行 `boot2docker ip`，就会看到它的地址）。

CoreOS

CoreOS的设计是为你提供能够像谷歌一样的大型互联网公司一样的基础设施管理能力来动态扩展和管理的计算能力。

CoreOS的安装文件和运行依赖非常小,它提供了精简的Linux系统。它使用Linux容器在更高的抽象层来管理你的服务,而不是通过常规的YUM和APT来安装包。

同时, CoreOS几乎可以运行在任何平台: Vagrant, Amazon EC2, QEMU/KVM, VMware 和 OpenStack 等等,甚至你所使用的硬件环境。

Kubernetes

注意：目前 *Kubernetes* 还处于 *beta* 状态，不推荐生产环境使用。部分概念和设计还可能会有后续调整。

Kubernetes 是 Google 团队发起并维护的基于 Docker 的开源容器集群管理系统，它不仅支持常见的云平台，而且支持内部数据中心。

建于 Docker 之上的 Kubernetes 可以构建一个容器的调度服务，其目的是让用户透过 Kubernetes 集群来进行云端容器集群的管理，而无需用户进行复杂的设置工作。系统会自动选取合适的工作节点来执行具体的容器集群调度处理工作。其核心概念是 Container Pod（容器仓）。一个 Pod 是有一组工作于同一物理工作节点的容器构成的。这些组容器拥有相同的网络命名空间/IP 以及存储配额，可以根据实际情况对每一个 Pod 进行端口映射。此外，Kubernetes 工作节点会由主系统进行管理，节点包含了能够运行 Docker 容器所用到的服务。

本章将分为 5 节介绍 Kubernetes。包括

- 项目简介
- 快速入门
- 基本概念
- 实践例子
- 架构分析等高级话题

项目简介



Kubernetes 是 Google 团队发起的开源项目，它的目标是管理跨多个主机的容器，提供基本的部署，维护以及运用伸缩，主要实现语言为Go语言。Kubernetes是：

- 易学：轻量级，简单，容易理解
- 便携：支持公有云，私有云，混合云，以及多种云平台
- 可拓展：模块化，可插拔，支持钩子，可任意组合
- 自修复：自动重调度，自动重启，自动复制

Kubernetes构建于Google数十年经验，一大半来源于Google生产环境规模的经验。结合了社区最佳的想法和实践。

在分布式系统中，部署，调度，伸缩一直是最为重要的也最为基础的功能。Kubernetes就是希望解决这一序列问题的。

Kubernetes 目前在github.com/GoogleCloudPlatform/kubernetes进行维护，截至定稿最新版本为 0.7.2 版本。

Kubernetes 能够运行在任何地方！

虽然Kubernetes最初是为GCE定制的，但是在后续版本中陆续增加了其他云平台的支持，以及本地数据中心的支持。

快速上手

目前，Kubernetes 支持在多种环境下的安装，包括本地主机（Fedora）、云服务（Google GAE、AWS 等）。然而最快速体验 Kubernetes 的方式显然是本地通过 Docker 的方式来启动相关进程。

下图展示了在单节点使用 Docker 快速部署一套 Kubernetes 的拓扑。



Kubernetes 依赖 Etcd 服务来维护所有主节点的状态。

启动 Etcd 服务。

```
docker run --net=host -d gcr.io/google_containers/etcd:2.0.9 /usr/local/bin/etcd --addr=1
```

启动主节点

启动 kubelet。

```
docker run --net=host -d -v /var/run/docker.sock:/var/run/docker.sock gcr.io/google_cont
```

启动服务代理

```
docker run -d --net=host --privileged gcr.io/google_containers/hyperkube:v0.17.0 /hyperku
```

测试状态

在本地访问 8080 端口，应该获取到类似如下的结果：

```
$ curl 127.0.0.1:8080
{
  "paths": [
    "/api",
    "/api/v1beta1",
    "/api/v1beta2",
    "/api/v1beta3",
    "/healthz",
    "/healthz/ping",
    "/logs/",
    "/metrics",
    "/static/",
```

```

    "/swagger-ui/",
    "/swaggerapi/",
    "/validate",
    "/version"
  ]
}

```

查看服务

所有服务启动后过一会，查看本地实际运行的 Docker 容器，应该有如下几个。

CONTAINER ID	IMAGE	COMMAND	C
ee054db2516c	gcr.io/google_containers/hyperkube:v0.17.0	"/hyperkube schedule	2
3b0f28de07a2	gcr.io/google_containers/hyperkube:v0.17.0	"/hyperkube apiserve	2
2eaa44ecd8e	gcr.io/google_containers/hyperkube:v0.17.0	"/hyperkube controll	2
30aa7163cbef	gcr.io/google_containers/hyperkube:v0.17.0	"/hyperkube proxy --	2
a2f282976d91	gcr.io/google_containers/pause:0.8.0	"/pause"	2
c060c52acc36	gcr.io/google_containers/hyperkube:v0.17.0	"/hyperkube kubelet	2
cc3cd263c581	gcr.io/google_containers/etcd:2.0.9	"/usr/local/bin/etcd	2

这些服务大概分为三类：主节点服务、工作节点服务和其它服务。

主节点服务

- apiserver 是整个系统的对外接口，提供 RESTful 方式供客户端和其它组件调用；
- scheduler 负责对资源进行调度，分配某个 pod 到某个节点上；
- controller-manager 负责管理控制器，包括 endpoint-controller（刷新服务和 pod 的关联信息）和 replication-controller（维护某个 pod 的复制为配置的数值）。

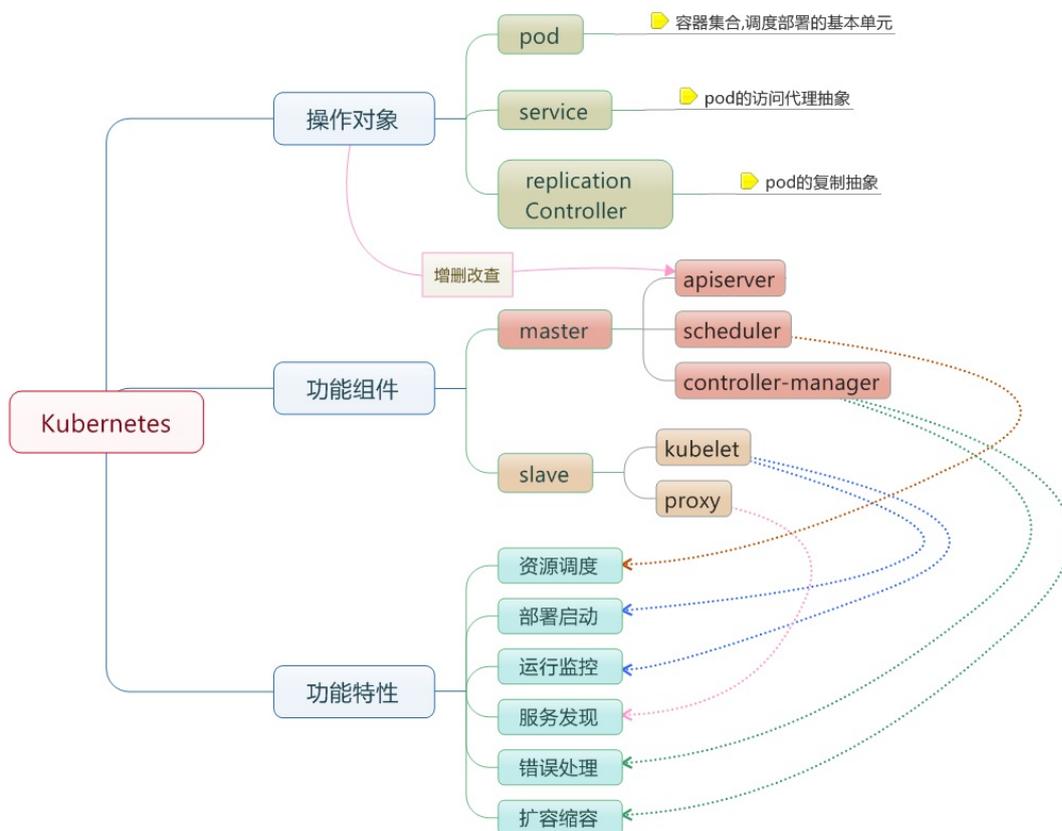
工作节点服务

- kubelet 是工作节点执行操作的 agent，负责具体的容器生命周期管理，根据从数据库中获取的信息来管理容器，并上报 pod 运行状态等；
- proxy 为 pod 上的服务提供访问的代理。

其它服务

- etcd 是所有状态的存储数据库；
- gcr.io/google_containers/pause:0.8.0 是 Kubernetes 启动后自动 pull 下来的测试镜像。

基本概念



- 节点 (Node) : 一个节点是一个运行 Kubernetes 中的主机。
- 容器组 (Pod) : 一个 Pod 对应于由若干容器组成的一个容器组, 同个组内的容器共享一个存储卷 (volume)。
- 容器组生命周期 (pod-states) : 包含所有容器状态集合, 包括容器组状态类型, 容器组生命周期, 事件, 重启策略, 以及replication controllers。
- Replication Controllers (replication-controllers) : 主要负责指定数量的pod在同一时间一起运行。
- 服务 (services) : 一个Kubernetes服务是容器组逻辑的高级抽象, 同时也对外提供访问容器组的策略。
- 卷 (volumes) : 一个卷就是一个目录, 容器对其有访问权限。
- 标签 (labels) : 标签是用来连接一组对象的, 比如容器组。标签可以被用来组织和选择子对象。
- 接口权限 (accessing_the_api) : 端口, ip地址和代理的防火墙规则。
- web 界面 (ux) : 用户可以通过 web 界面操作Kubernetes。
- 命令行操作 (cli) : `kubectfg` 命令。

节点

在 Kubernetes 中, 节点是实际工作的点, 以前叫做 Minion。节点可以是虚拟机或者物理机器, 依赖于一个集群环境。每个节点都有一些必要的服务以运行容器组, 并且它们都可以通过主节点来管理。必要服务包括 Docker, kubelet 和代理服务。

容器状态

容器状态用来描述节点的当前状态。现在，其中包含三个信息：

主机IP

主机IP需要云平台来查询，Kubernetes把它作为状态的一部分来保存。如果Kubernetes没有运行在云平台上，节点ID就是必需的。IP地址可以变化，并且可以包含多种类型的IP地址，如公共IP，私有IP，动态IP，ipv6等等。

节点周期

通常来说节点有 `Pending`，`Running`，`Terminated` 三个周期，如果Kubernetes发现了一个节点并且其可用，那么Kubernetes就把它标记为 `Pending`。然后在某个时刻，Kubernetes将会标记其为 `Running`。节点的结束周期称为 `Terminated`。一个已经terminated的节点不会接受和调度任何请求，并且已经在其上运行的容器组也会删除。

节点状态

节点的状态主要是用来描述处于 `Running` 的节点。当前可用的有 `NodeReachable` 和 `NodeReady`。以后可能会增加其他状态。`NodeReachable` 表示集群可达。`NodeReady` 表示kubelet返回 `StatusOk`并且HTTP状态检查健康。

节点管理

节点并非Kubernetes创建，而是由云平台创建，或者就是物理机器、虚拟机。在Kubernetes中，节点仅仅是一条记录，节点创建之后，Kubernetes会检查其是否可用。在Kubernetes中，节点用如下结构保存：

```
{
  "id": "10.1.2.3",
  "kind": "Minion",
  "apiVersion": "v1beta1",
  "resources": {
    "capacity": {
      "cpu": 1000,
      "memory": 1073741824
    },
  },
  "labels": {
    "name": "my-first-k8s-node",
  },
}
```

Kubernetes校验节点可用依赖于id。在当前的版本中，有两个接口可以用来管理节点：节点控制和Kube管理。

节点控制

在Kubernetes主节点中，节点控制器是用来管理节点的组件。主要包含：

- 集群范围内节点同步
- 单节点生命周期管理

节点控制有一个同步轮寻，主要监听所有云平台的虚拟实例，会根据节点状态创建和删除。可以通过 `--node_sync_period` 标志来控制该轮寻。如果一个实例已经创建，节点控制将会为其创建一个结构。同样的，如果一个节点被删除，节点控制也会删除该结构。在Kubernetes启动时可用通过 `--machines` 标记来显示指定节点。同样可以使用 `kubect1` 来一条一条的添加节点，两者是相同的。通过设置 `--sync_nodes=false` 标记来禁止集群之间的节点同步，你也可以使用`api/kubect1` 命令行来增删节点。

容器组

在Kubernetes中，使用的最小单位是容器组，容器组是创建，调度，管理的最小单位。一个容器组使用相同的Dokcer容器并共享卷（挂载点）。一个容器组是一个特定运用的打包集合，包含一个或多个容器。

和运行的容器类似，一个容器组被认为只有很短的运行周期。容器组被调度到一组节点运行，知道容器的生命周期结束或者其被删除。如果节点死掉，运行在其上的容器组将会被删除而不是重新调度。（也许在将来的版本中会添加容器组的移动）。

容器组设计的初衷

资源共享和通信

容器组主要是为了数据共享和它们之间的通信。

在一个容器组中，容器都使用相同的网络地址和端口，可以通过本地网络来相互通信。每个容器组都有独立的ip，可用通过网络来和其他物理主机或者容器通信。

容器组有一组存储卷（挂载点），主要是为了让容器在重启之后可以不丢失数据。

容器组管理

容器组是一个运用管理和部署的高层次抽象，同时也是一组容器的接口。容器组是部署、水平放缩的最小单位。

容器组的使用

容器组可以通过组合来构建复杂的运用，其本来的意义包含：

- 内容管理，文件和数据加载以及本地缓存管理等。
- 日志和检查点备份，压缩，快照等。
- 监听数据变化，跟踪日志，日志和监控代理，消息发布等。
- 代理，网桥
- 控制器，管理，配置以及更新

替代方案

为什么不在一个单一的容器里运行多个程序？

- 1.透明化。为了使容器组中的容器保持一致的基础设施和服务，比如进程管理和资源监控。这样设计是为了用户的便利性。
- 2.解耦软件之间的依赖。每个容器都可能重新构建和发布，Kubernetes必须支持热发布和热更新（将来）。
- 3.方便使用。用户不必运行独立的程序管理，也不用担心每个运用程序的退出状态。
- 4.高效。考虑到基础设施有更多的职责，容器必须要轻量化。

容器组的生命状态

包括若干状态值：pending、running、succeeded、failed。

pending

容器组已经被节点接受，但有一个或多个容器还没有运行起来。这将包含某些节点正在下载镜像的时间，这种情形会依赖于网络情况。

running

容器组已经被调度到节点，并且所有的容器都已经启动。至少有一个容器处于运行状态（或者处于重启状态）。

succeeded

所有的容器都正常退出。

failed

容器组中所有容器都意外中断了。

容器组生命周期

通常来说，如果容器组被创建了就不会自动销毁，除非被某种行为出发，而触发此种情况可能是人为，或者复制控制器所为。唯一例外的是容器组由 succeeded状态成功退出，或者在一定时间内重试多次依然失败。

如果某个节点死掉或者不能连接，那么节点控制器将会标记其上的容器组的状态为 failed。

举例如下。

- 容器组状态 running，有 1 容器，容器正常退出
 - 记录完成事件
 - 如果重启策略为：
 - 始终：重启容器，容器组保持 running
 - 失败时：容器组变为 succeeded
 - 从不：容器组变为 succeeded
- 容器组状态 running，有1容器，容器异常退出
 - 记录失败事件
 - 如果重启策略为：

- 始终：重启容器，容器组保持 `running`
 - 失败时：重启容器，容器组保持 `running`
 - 从不：容器组变为 `failed`
- 容器组状态 `running`，有2容器，有1容器异常退出
 - 记录失败事件
 - 如果重启策略为：
 - 始终：重启容器，容器组保持 `running`
 - 失败时：重启容器，容器组保持 `running`
 - 从不：容器组保持 `running`
 - 当有2容器退出
 - 记录失败事件
 - 如果重启策略为：
 - 始终：重启容器，容器组保持 `running`
 - 失败时：重启容器，容器组保持 `running`
 - 从不：容器组变为 `failed`
- 容器组状态 `running`，容器内存不足
 - 标记容器错误中断
 - 记录内存不足事件
 - 如果重启策略为：
 - 始终：重启容器，容器组保持 `running`
 - 失败时：重启容器，容器组保持 `running`
 - 从不：记录错误事件，容器组变为 `failed`
- 容器组状态 `running`，一块磁盘死掉
 - 杀死所有容器
 - 记录事件
 - 容器组变为 `failed`
 - 如果容器组运行在一个控制器下，容器组将会在其他地方重新创建
- 容器组状态 `running`，对应的节点段溢出
 - 节点控制器等到超时
 - 节点控制器标记容器组 `failed`
 - 如果容器组运行在一个控制器下，容器组将会在其他地方重新创建

Replication Controllers

服务

卷

标签

接口权限

web界面

命令行操作

kubectl 使用

`kubectl` 是 Kubernetes 自带的客户端，可以用它来直接操作 Kubernetes。

使用格式有两种：

```
kubectl [flags]  
kubectl [command]
```

get

Display one or many resources

describe

Show details of a specific resource

create

Create a resource by filename or stdin

update

Update a resource by filename or stdin.

delete

Delete a resource by filename, stdin, resource and ID, or by resources and label selector.

namespace

SUPERCEDED: Set and view the current Kubernetes namespace

log

Print the logs for a container in a pod.

rolling-update

Perform a rolling update of the given ReplicationController.

resize

Set a new size for a Replication Controller.

exec

Execute a command in a container.

port-forward

Forward one or more local ports to a pod.

proxy

Run a proxy to the Kubernetes API server

run-container

Run a particular image on the cluster.

stop

Gracefully shut down a resource by id or filename.

expose

Take a replicated application and expose it as Kubernetes Service

label

Update the labels on a resource

config

config modifies kubeconfig files

cluster-info

Display cluster info

api-versions

Print available API versions.

version

Print the client and server version information.

help

Help about any command

基本架构

任何优秀的项目都离不开优秀的架构设计。本小节将介绍 Kubernetes 在架构方面的设计考虑。

基本考虑

如果让我们自己从头设计一套容器管理平台，有如下几个方面是很容易想到的：

- 分布式架构，保证扩展性；
- 逻辑集中式的控制平面 + 物理分布式的运行平面；
- 一套资源调度系统，管理哪个容器该分配到哪个节点上；
- 一套对容器内服务进行抽象和 HA 的系统。

运行原理

下面这张图完整展示了 Kubernetes 的运行原理。



可见，Kubernetes 首先是一套分布式系统，由多个节点组成，节点分为两类：一类是属于管理平面的主节点/控制节点（Master Node）；一类是属于运行平面的工作节点（Worker Node）。

显然，复杂的工作肯定都交给控制节点去做了，工作节点负责提供稳定的操作接口和能力抽象即可。

从这张图上，我们没有能发现 Kubernetes 中对于控制平面的分布式实现，但是由于数据后端自身就是一套分布式的数据库（Etcd），因此可以很容易扩展到分布式实现。

控制平面

主节点服务

主节点上需要提供如下的管理服务：

- apiserver 是整个系统的对外接口，提供一套 RESTful 的 [Kubernetes API](#)，供客户端和其它组件调用；
- scheduler 负责对资源进行调度，分配某个 pod 到某个节点上。是 pluggable 的，意味着很容易选择其它实现方式；
- controller-manager 负责管理控制器，包括 endpoint-controller（刷新服务和 pod 的关联信息）和 replication-controller（维护某个 pod 的复制为配置的数值）。

Etcd

这里 Etcd 即作为数据后端，又作为消息中间件。

通过 Etcd 来存储所有的主节点上的状态信息，很容易实现主节点的分布式扩展。

组件可以自动的去侦测 Etcid 中的数值变化来获得通知，并且获得更新后的数据来执行相应的操作。

工作节点

- kubelet 是工作节点执行操作的 agent，负责具体的容器生命周期管理，根据从数据库中获取的信息来管理容器，并上报 pod 运行状态等；
- kube-proxy 是一个简单的网络访问代理，同时也是一个 Load Balancer。它负责将访问到某个服务的请求具体分配给工作节点上的 Pod（同一类标签）。



Mesos 项目

简介

Mesos 是一个集群资源的自动调度平台，Apache 开源项目，它的定位是要做数据中心操作系统的内核。目前由 Mesosphere 公司维护，更多信息可以自行查阅 [Mesos 项目地址](#) 或 [Mesosphere](#)。

Mesos + Marathon 安装与使用

Marathon 是可以跟 Mesos 一起协作的一个 framework，用来运行持久性的应用。

安装

一共需要安装四种组件，mesos-master、marathon、zookeeper 需要安装到所有的主节点，mesos-slave 需要安装到从节点。

mesos 利用 zookeeper 来进行主节点的同步，以及从节点发现主节点的过程。

源码编译

下载源码

```
git clone https://git-wip-us.apache.org/repos/asf/mesos.git
```

安装依赖

```
#jdk-7
sudo apt-get update && sudo apt-get install -y openjdk-7-jdk
#autotools
sudo apt-get install -y autoconf libtool
#Mesos dependencies.
sudo apt-get -y install build-essential python-dev python-boto libcurl4-nss-dev libsasl2-
```

编译&安装

```
$ cd mesos

# Bootstrap (Only required if building from git repository).
$ ./bootstrap

$ mkdir build
$ cd build && ../configure
$ make
$ make check && make install
```

软件源安装

以 ubuntu 系统为例。

安装 Docker，不再赘述，可以参考 [这里](#)。

```
# Setup
sudo apt-key adv --keyserver keyserver.ubuntu.com --recv E56151BF
```

```
DISTRO=$(lsb_release -is | tr '[:upper:]' '[:lower:]')
CODENAME=$(lsb_release -cs)

# Add the repository
echo "deb http://repos.mesosphere.io/${DISTRO} ${CODENAME} main" | \
  sudo tee /etc/apt/sources.list.d/mesosphere.list

sudo apt-get -y update && sudo apt-get -y install zookeeper mesos marathon
```

基于 Docker

将基于如下镜像：

- ZooKeeper : <https://registry.hub.docker.com/u/garland/zookeeper/>
- Mesos : <https://registry.hub.docker.com/u/garland/mesosphere-docker-mesos-master/>
- Marathon : <https://registry.hub.docker.com/u/garland/mesosphere-docker-marathon/>

其中 mesos-master 镜像将作为 master 和 slave 容器使用。

导出本地机器的地址到环境变量。

```
HOST_IP=10.11.31.7
```

启动 Zookeeper 容器。

```
docker run -d \
-p 2181:2181 \
-p 2888:2888 \
-p 3888:3888 \
garland/zookeeper
```

启动 Mesos Master 容器。

```
docker run --net="host" \
-p 5050:5050 \
-e "MESOS_HOSTNAME=${HOST_IP}" \
-e "MESOS_IP=${HOST_IP}" \
-e "MESOS_ZK=zk://${HOST_IP}:2181/mesos" \
-e "MESOS_PORT=5050" \
-e "MESOS_LOG_DIR=/var/log/mesos" \
-e "MESOS_QUORUM=1" \
-e "MESOS_REGISTRY=in_memory" \
-e "MESOS_WORK_DIR=/var/lib/mesos" \
-d \
garland/mesosphere-docker-mesos-master
```

启动 Marathon。

```
docker run \
```

```
-d \
-p 8080:8080 \
garland/mesosphere-docker-marathon --master zk://{HOST_IP}:2181/mesos --zk zk://{HOST_I
```

启动 Mesos slave 容器。

```
docker run -d \
--name mesos_slave_1 \
--entrypoint="mesos-slave" \
-e "MESOS_MASTER=zk://{HOST_IP}:2181/mesos" \
-e "MESOS_LOG_DIR=/var/log/mesos" \
-e "MESOS_LOGGING_LEVEL=INFO" \
garland/mesosphere-docker-mesos-master:latest
```

接下来，可以通过访问本地 8080 端口来使用 Marathon 启动任务了。

配置说明

ZooKeeper

ZooKeeper 是一个分布式应用的协调工具，用来管理多个 Master 节点的选举和冗余，监听在 2181 端口。

配置文件在 /etc/zookeeper/conf/ 目录下。

首先，要修改 myid，手动为每一个节点分配一个自己的 id（1-255 之间）。

zoo.cfg 是主配置文件，主要修改如下的三行（如果你启动三个 zk 节点）。

```
server.1=zookeeper1:2888:3888
server.2=zookeeper2:2888:3888
server.3=zookeeper3:2888:3888
```

主机名需要自己替换，并在 /etc/hosts 中更新。

第一个端口负责从节点连接到主节点的；第二个端口负责主节点的选举通信。

Mesos

Mesos 的默认配置目录分别为：

- /etc/mesos：共同的配置文件，最关键的是 zk 文件；
- /etc/mesos-master：主节点的配置，等价于启动 mesos-master 时候的默认选项；
- /etc/mesos-slave：从节点的配置，等价于启动 mesos-master 时候的默认选项。

主节点

首先在所有节点上修改 /etc/mesos/zk，为主节点的 zookeeper 地址列表，例如：

```
zk://ip1:2181,ip2:2181/mesos
```

创建 /etc/mesos-master/ip 文件，写入主节点监听的地址。

还可以创建 /etc/mesos-master/cluster 文件，写入集群的别名。

之后，启动服务：

```
sudo service mesos-master start
```

更多选项可以参考[这里](#)。

从节点

在从节点上，修改 /etc/mesos-slave/ip 文件，写入跟主节点通信的地址。

之后，启动服务。

```
sudo service mesos-slave start
```

更多选项可以参考[这里](#)。

此时，通过浏览器访问本地 5050 端口，可以看到节点信息。

The screenshot shows the Mesos web interface for a cluster named 'MyCluster'. The top navigation bar includes 'Mesos', 'Frameworks', 'Slaves', 'Offers', and 'MyCluster'. Below the navigation bar, the master node information is displayed: 'Master 20150619-192346-1298446857-5050-14153'. The main content area is divided into two columns. The left column contains a summary of cluster information: 'Cluster: MyCluster', 'Server: [redacted]:5050', 'Version: 0.22.1', 'Built: a month ago by root', 'Started: 35 minutes ago', and 'Elected: 35 minutes ago'. Below this is a 'LOG' section, followed by 'Slaves' (2 Activated, 0 Deactivated), 'Tasks' (0 Staged, 0 Started, 0 Finished, 0 Killed, 0 Failed, 0 Lost), and 'Resources' (Total: 8 CPUs, 9.7 GB Mem; Used: 0 CPUs, 0 B Mem; Offered: 0 CPUs, 0 B Mem; Idle: 8 CPUs, 9.7 GB Mem). The right column contains two tables: 'Active Tasks' and 'Completed Tasks'. Both tables have columns for ID, Name, State, Started, Stopped, and Host, and both show 'No active tasks.' and 'No completed tasks.' respectively.

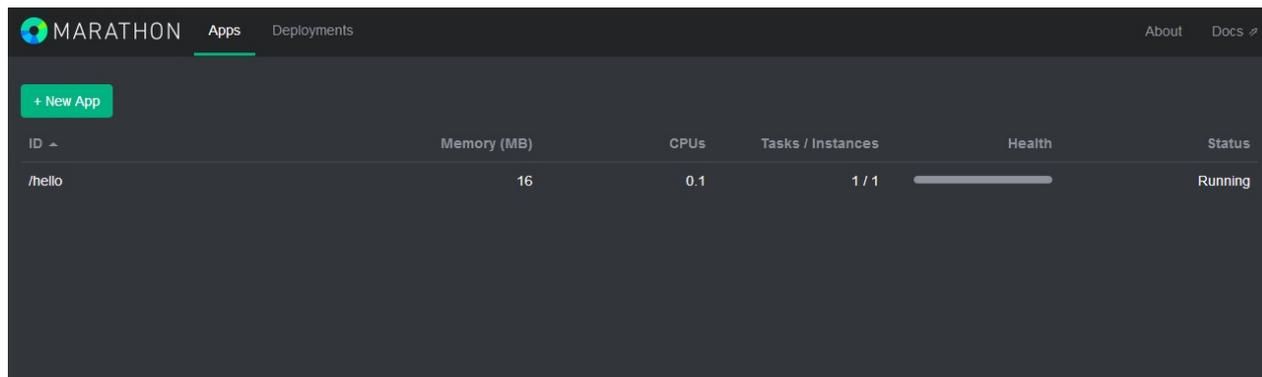
Marathon

启动 marathon 服务。

```
sudo service marathon start
```

启动成功后，在 mesos 的 web 界面的 frameworks 标签页下面将能看到名称为 marathon 的框架出现。

同时可以通过浏览器访问 8080 端口，看到 marathon 的管理界面。



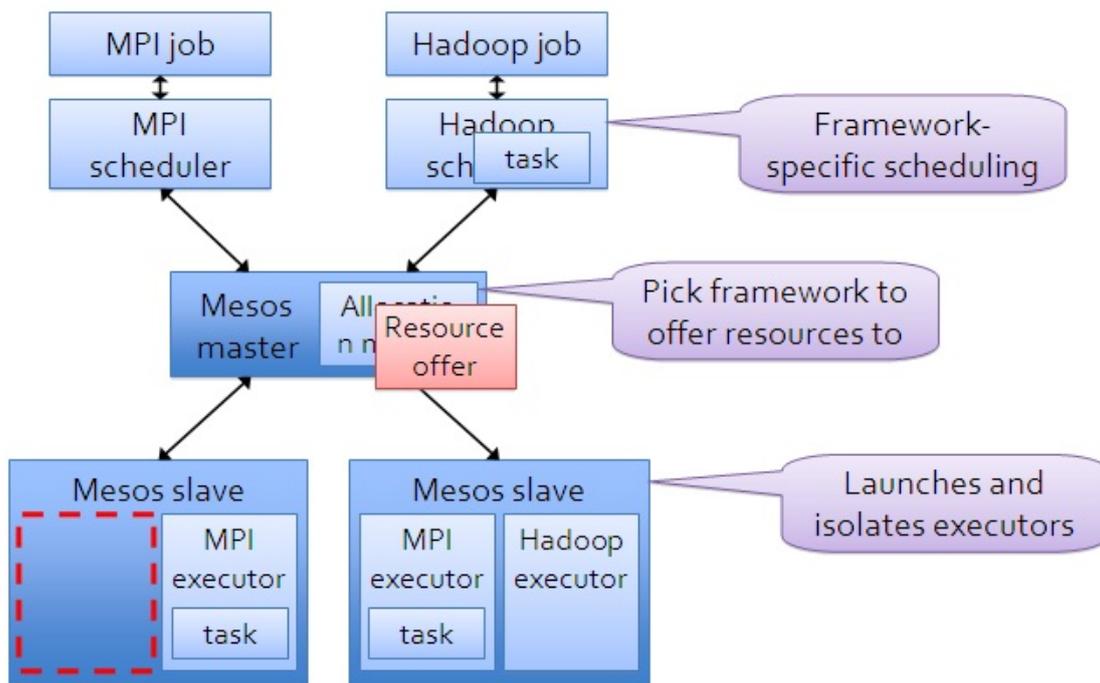
此时，可以通过界面或者 REST API 来创建一个应用，Marathon 会保持该应用的持续运行。

Mesos 基本原理与架构

首先，Mesos 自身只是一个资源调度框架，并非一整套完整的应用管理平台，本身是不能干活的。但是它比较容易的跟各种应用管理或者中间件平台整合，一起工作，提高资源使用效率。

架构

Mesos Architecture



master-slave 架构，master 使用 zookeeper 来做 HA。

master 单独运行在管理节点上，slave 运行在各个计算任务节点上。

各种具体任务的管理平台，即 framework 跟 master 交互，来申请资源。

基本单元

master

负责整体的资源调度和逻辑控制。

slave

负责汇报本节点上的资源给 master，并负责隔离资源来执行具体的任务。

隔离机制当然就是各种容器机制了。

framework

framework 是实际干活的，包括两个主要组件：

- scheduler：注册到主节点，等待分配资源；
- executor：在 slave 节点上执行本framework 的任务。

framework 分两种：一种是对资源需求可以 scale up 或者 down 的（Hadoop、Spark）；一种是对资源需求大小是固定的（MPI）。

调度

对于一个资源调度框架来说，最核心的就是调度机制，怎么能快速高效的完成对某个 framework 资源的分配（最好是能猜到它的实际需求）。

两层调度算法：

master 先调度一大块资源给某个 framework，framework 自己再实现内部的细粒度调度。

调度机制支持插件。默认是 DRF。

基本调度过程

调度通过 offer 方式交互：

- master 提供一个 offer（一组资源）给 framework；
- framework 可以决定要不要，如果接受的话，返回一个描述，说明自己希望如何使用和分配这些资源（可以说只希望使用部分资源，则多出来的会被 master 收回）；
- master 则根据 framework 的分配情况发送给 slave，以使用 framework 的 executor 来按照分配的资源策略执行任务。

过滤器

framework 可以通过过滤器机制告诉 master 它的资源偏好，比如希望分配过来的 offer 有哪个资源，或者至少有多少资源。

主要是为了加速资源分配的交互过程。

回头机制

master 可以通过回收计算节点上的任务来动态调整长期任务和短期任务的分布。

HA

master

master 节点存在单点失效问题，所以肯定要上 HA，目前主要是使用 zookeeper 来热备份。

同时 master 节点可以通过 slave 和 framework 发来的消息重建内部状态（具体能有多快呢？这里不使用数据库可能是避免引入复杂度。）。

framework 通知

framework 中相关的失效，master 将发给它的 scheduler 来通知。

Mesos 配置项解析

Mesos 的 [配置项](#) 可以通过启动时候传递参数或者配置目录下文件的方式给出（推荐方式，一目了然）。

分为三种类型：通用项（master 和 slave 都支持），只有 master 支持的，以及只有 slave 支持的。

通用项

- `--ip=VALUE` 监听的 IP 地址
- `--firewall_rules=VALUE` endpoint 防火墙规则，`VALUE` 可以是 JSON 格式或者存有 JSON 格式的文件路径。
- `--log_dir=VALUE` 日志文件路径，默认不存储日志到本地
- `--logbufsecs=VALUE` buffer 多少秒的日志，然后写入本地
- `--logging_level=VALUE` 日志记录的最低级别
- `--port=VALUE` 监听的端口，master 默认是 5050，slave 默认是 5051。

master 专属配置项

- `--quorum=VALUE` 必备项，使用基于 replicated-Log 的注册表时，复制的个数
- `--work_dir=VALUE` 必备项，注册表持久化信息存储位置
- `--zk=VALUE` 必备项，zookepr 的接口地址，支持多个地址，之间用逗号隔离，可以为文件路径
- `--acls=VALUE` ACL 规则或所在文件
- `--allocation_interval=VALUE` 执行 allocation 的间隔，默认为 1sec
- `--allocator=VALUE` 分配机制，默认为 HierarchicalDRF
- `--[no-]authenticate` 是否允许非认证过的 framework 注册
- `--[no-]authenticate_slaves` 是否允许非认证过的 slaves 注册
- `--authenticators=VALUE` 对 framework 或 slaves 进行认证时的实现机制
- `--cluster=VALUE` 集群别名
- `--credentials=VALUE` 存储加密后凭证的文件的的路径
- `--external_log_file=VALUE` 采用外部的日志文件
- `--framework_sorter=VALUE` 给定 framework 之间的资源分配策略
- `--hooks=VALUE` master 中安装的 hook 模块
- `--hostname=VALUE` master 节点使用的主机名，不配置则从系统中获取
- `--[no-]log_auto_initialize` 是否自动初始化注册表需要的 replicated 日志
- `--modules=VALUE` 要加载的模块，支持文件路径或者 JSON
- `--offer_timeout=VALUE` offer 撤销的超时
- `--rate_limits=VALUE` framework 的速率限制，比如 qps
- `--recovery_slave_removal_limit=VALUE` 限制注册表恢复后可以移除或停止的 slave 数目，超出后 master 会失败，默认是 100%
- `--slave_removal_rate_limit=VALUE` slave 没有完成健康度检查时候被移除的速率上限，例如 1/10mins 代表每十分钟最多有一个
- `--registry=VALUE` 注册表的持久化策略，默认为 `replicated_log`，还可以为 `in_memory`
- `--registry_fetch_timeout=VALUE` 访问注册表失败超时

- `--registry_store_timeout=VALUE` 存储注册表失败超时
- `--[no-]registry_strict` 是否按照注册表中持久化信息执行操作，默认为 `false`
- `--roles=VALUE` 集群中 framework 可以所属的分配角色
- `--[no-]root_submissions` root 是否可以提交 framework，默认为 `true`
- `--slave_reregister_timeout=VALUE` 新的 lead master 节点选举出来后，多久之内所有的 slave 需要注册，超时的 slave 将被移除并关闭，默认为 10mins
- `--user_sorter=VALUE` 在用户之间分配资源的策略，默认为 `drf`
- `--webui_dir=VALUE` webui 实现的文件目录所在，默认为 `/usr/local/share/mesos/webui`
- `--weights=VALUE` 各个角色的权重
- `--whitelist=VALUE` 文件路径，包括发送 offer 的 slave 名单，默认为 `None`
- `--zk_session_timeout=VALUE` session 超时，默认为 10secs
- `--max_executors_per_slave=VALUE` 配置了 `--with-network-isolator` 时可用，限制每个 slave 同时执行任务个数

slave 专属配置项

- `--master=VALUE` 必备项，master 所在地址，或 zookeeper 地址，或文件路径，可以是列表
- `--attributes=VALUE` 机器属性
- `--authenticatee=VALUE` 跟 master 进行认证时候的认证机制
- `--[no-]cgroups_enable_cfs` 采用 CFS 进行带宽限制时候对 CPU 资源进行限制，默认为 `false`
- `--cgroups_hierarchy=VALUE` cgroups 的目录根位置，默认为 `/sys/fs/cgroup`
- `--[no-]cgroups_limit_swap` 限制内存和 swap，默认为 `false`，只限制内存
- `--cgroups_root=VALUE` 根 cgroups 的名称，默认为 `mesos`
- `--container_disk_watch_interval=VALUE` 为容器进行硬盘配额查询的时间间隔
- `--containerizer_path=VALUE` 采用外部隔离机制（`--isolation=external`）时候，外部容器机制执行文件路径
- `--containerizers=VALUE` 可用的容器实现机制，包括 `mesos`、`external`、`docker`
- `--credential=VALUE` 加密后凭证，或者所在文件路径
- `--default_container_image=VALUE` 采用外部容器机制时，任务缺省使用的镜像
- `--default_container_info=VALUE` 容器信息的缺省值
- `--default_role=VALUE` 资源缺省分配的角色
- `--disk_watch_interval=VALUE` 硬盘使用情况的周期性检查间隔，默认为 1mins
- `--docker=VALUE` docker 执行文件的路径
- `--docker_remove_delay=VALUE` 删除容器之前的等待时间，默认为 6hrs
- `--[no-]docker_kill_orphans` 清除孤儿容器，默认为 `true`
- `--docker_sock=VALUE` docker sock 地址，默认为 `/var/run/docker.sock`
- `--docker_mesos_image=VALUE` 运行 slave 的 docker 镜像，如果被配置，docker 会假定 slave 运行在一个 docker 容器里
- `--docker_sandbox_directory=VALUE` sandbox 映射到容器里的哪个路径
- `--docker_stop_timeout=VALUE` 停止实例后等待多久执行 kill 操作，默认为 0secs
- `--[no-]enforce_container_disk_quota` 是否启用容器配额限制，默认为 `false`
- `--executor_registration_timeout=VALUE` 执行应用最多可以等多久再注册到 slave，否则停止它，默认为 1mins
- `--executor_shutdown_grace_period=VALUE` 执行应用停止后，等待多久，默认为 5secs

- `--external_log_file=VALUE` 外部日志文件
- `--frameworks_home=VALUE` 执行应用前添加的相对路径，默认为空
- `--gc_delay=VALUE` 多久清理一次执行应用目录，默认为 1weeks
- `--gc_disk_headroom=VALUE` 调整计算最大执行应用目录年龄的硬盘留空量，默认为 0.1
- `--hadoop_home=VALUE` hadoop 安装目录，默认为空，会自动查找 HADOOP_HOME 或者从系统路径中查找
- `--hooks=VALUE` 安装在 master 中的 hook 模块列表
- `--hostname=VALUE` slave 节点使用的主机名
- `--isolation=VALUE` 隔离机制，例如 `posix/cpu, posix/mem`（默认）或者 `cgroups/cpu, cgroups/mem`
- `--launcher_dir=VALUE` mesos 可执行文件的路径，默认为 `/usr/local/lib/mesos`
- `--modules=VALUE` 要加载的模块，支持文件路径或者 JSON
- `--perf_duration=VALUE` perf 采样时长，必须小于 `perf_interval`，默认为 10secs
- `--perf_events=VALUE` perf 采样的事件
- `--perf_interval=VALUE` perf 采样的时间间隔
- `--recover=VALUE` 回复后是否重连上旧的执行应用
- `--recovery_timeout=VALUE` slave 恢复时的超时，太久则所有相关的执行应用将自行退出，默认为 15mins
- `--registration_backoff_factor=VALUE` 跟 master 进行注册时候的重试时间间隔算法的因子，默认为 1secs，采用随机指数算法，最长 1mins
- `--resource_monitoring_interval=VALUE` 周期性监测执行应用资源使用情况的间隔，默认为 1secs
- `--resources=VALUE` 每个 slave 可用的资源
- `--slave_subsystems=VALUE` slave 运行在哪些 cgroup 子系统中，包括 memory, cpuacct 等，缺省为空
- `--[no-]strict` 是否认为所有错误都不可忽略，默认为 true
- `--[no-]switch_user` 用提交任务的用户身份来运行，默认为 true
- `--fetcher_cache_size=VALUE` fetcher 的 cache 大小，默认为 2 GB
- `--fetcher_cache_dir=VALUE` fetcher cache 文件存放目录，默认为 `/tmp/mesos/fetch`
- `--work_dir=VALUE` framework 的工作目录，默认为 `/tmp/mesos`

下面的选项需要配置 `--with-network-isolator` 一起使用

- `--ephemeral_ports_per_container=VALUE` 分配给一个容器的临时端口，默认为 1024
- `--eth0_name=VALUE` public 网络的接口名称，如果不指定，根据主机路由进行猜测
- `--lo_name=VALUE` loopback 网卡名称
- `--egress_rate_limit_per_container=VALUE` 每个容器的 egress 流量限制速率
- `--[no-]network_enable_socket_statistics` 是否采集每个容器的 socket 统计信息，默认为 false

Mesos 常见框架

framework 是实际干活的，可以理解为 mesos 上跑的 `应用`，需要注册到 master 上。

长期运行的服务

Aurora

利用 mesos 调度安排的任务，保证任务一直在运行。

提供 REST 接口，客户端和 webUI (8081 端口)

Marathon

一个 PaaS 平台。

保证任务一直在运行。如果停止了，会自动重启一个新的任务。

支持任务为任意 bash 命令，以及容器。

提供 REST 接口，客户端和 webUI (8080 端口)

Singularity

一个 PaaS 平台。

调度器，运行长期的任务和一次性任务。

提供 REST 接口，客户端和 webUI (7099、8080 端口)，支持容器。

大数据处理

Cray Chapel

支持 Chapel 并行编程语言的运行框架。

Dpark

Spark 的 Python 实现。

Hadoop

经典的 map-reduce 模型的实现。

Spark

跟 Hadoop 类似，但处理迭代类型任务会更好的使用内存做中间状态缓存，速度要快一些。

Storm

常见框架

分布式流计算，可以实时处理数据流。

批量调度

Chronos

Cron 的分布式实现，负责任务调度。

Jenkins

大名鼎鼎的 CI 引擎。使用 mesos-jenkins 插件，可以将 jenkins 的任务被 mesos 来动态调度执行。

ElasticSearch

功能十分强大的分布式数据搜索引擎。

数据存储

Cassandra

高性能分布式数据库。

Docker命令查询

基本语法

```
docker [OPTIONS] COMMAND [arg...]
```

一般来说，Docker 命令可以用来管理 daemon，或者通过 CLI 命令管理镜像和容器。可以通过 `man docker` 来查看这些命令。

选项

```
-D=true|false
    使用 debug 模式。默认为 false。

-H, --host=[unix:///var/run/docker.sock]: tcp://[host:port]来绑定或者 unix://[/path/to/socket]
    在 daemon 模式下绑定的 socket，通过一个或多个 tcp://host:port, unix:///path/to/socket, fd:

--api-enable-cors=true|false
    在远端 API 中启用 CORS 头。缺省为 false。

-b=""
    将容器挂载到一个已存在的网桥上。指定为 'none' 时则禁用容器的网络。

--bip=""
    让动态创建的 docker0 采用给定的 CIDR 地址；与 -b 选项互斥。

-d=true|false
    使用 daemon 模式。缺省为 false。

--dns=""
    让 Docker 使用给定的 DNS 服务器。

-g=""
    指定 Docker 运行时的 root 路径。缺省为 /var/lib/docker。

--icc=true|false
    启用容器间通信。默认为 true。

--ip=""
    绑定端口时候的默认 IP 地址。缺省为 0.0.0.0。

--iptables=true|false
    禁止 Docker 添加 iptables 规则。缺省为 true。

--mtu=VALUE
    指定容器网络的 mtu。缺省为 1500。

-p=""
    指定 daemon 的 PID 文件路径。缺省为 /var/run/docker.pid。

-s=""
```

强制 Docker 运行时使用给定的存储驱动。

`-v=true|false`

输出版本信息并退出。缺省值为 `false`。

`--selinux-enabled=true|false`

启用 SELinux 支持。缺省值为 `false`。SELinux 目前不支持 BTRFS 存储驱动。

命令

Docker 的命令可以采用 `docker-CMD` 或者 `docker CMD` 的方式执行。两者一致。

`docker-attach(1)`

依附到一个正在运行的容器中。

`docker-build(1)`

从一个 Dockerfile 创建一个镜像

`docker-commit(1)`

从一个容器的修改中创建一个新的镜像

`docker-cp(1)`

从容器中复制文件到宿主系统中

`docker-diff(1)`

检查一个容器文件系统的修改

`docker-events(1)`

从服务端获取实时的事件

`docker-export(1)`

导出容器内容为一个 tar 包

`docker-history(1)`

显示一个镜像的历史

`docker-images(1)`

列出存在的镜像

`docker-import(1)`

导入一个文件（典型为 tar 包）路径或目录来创建一个镜像

`docker-info(1)`

显示一些相关的系统信息

`docker-inspect(1)`

显示一个容器的底层具体信息。

`docker-kill(1)`

关闭一个运行中的容器（包括进程和所有资源）

`docker-load(1)`

从一个 tar 包中加载一个镜像

`docker-login(1)`

注册或登录到一个 Docker 的仓库服务器

`docker-logout(1)`

从 Docker 的仓库服务器登出

`docker-logs(1)`

获取容器的 log 信息

`docker-pause(1)`

暂停一个容器中的所有进程

`docker-port(1)`

查找一个 nat 到一个私有网口的公共口

`docker-ps(1)`

列出容器

`docker-pull(1)`

从一个Docker的仓库服务器下拉一个镜像或仓库

`docker-push(1)`

将一个镜像或者仓库推送到一个 Docker 的注册服务器

`docker-restart(1)`

重启一个运行中的容器

`docker-rm(1)`

删除给定的若干个容器

`docker-rmi(1)`

删除给定的若干个镜像

`docker-run(1)`

创建一个新容器，并在其中运行给定命令

`docker-save(1)`

保存一个镜像为 tar 包文件

`docker-search(1)`

在 Docker index 中搜索一个镜像

`docker-start(1)`

启动一个容器

`docker-stop(1)`

终止一个运行中的容器

`docker-tag(1)`

为一个镜像打标签

`docker-top(1)`

查看一个容器中的正在运行的进程信息

`docker-unpause(1)`

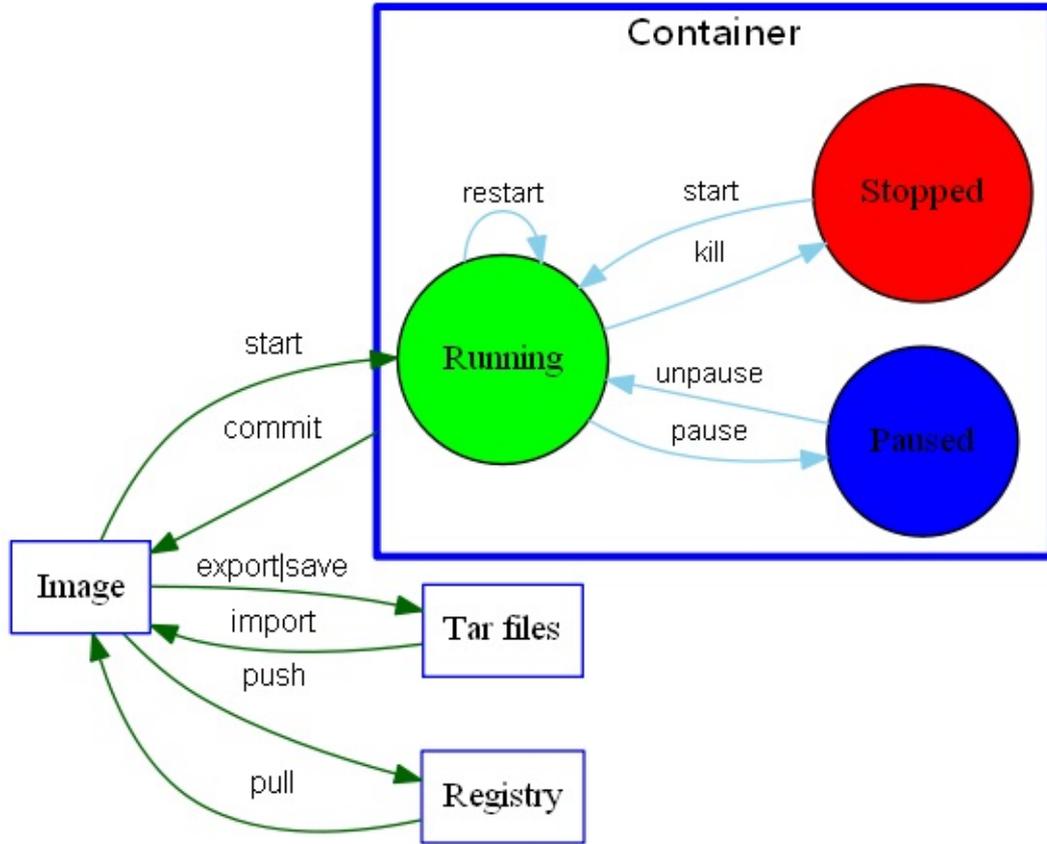
将一个容器内所有的进程从暂停状态中恢复

`docker-version(1)`

输出 Docker 的版本信息

```
docker-wait(1)  
阻塞直到一个容器终止，然后输出它的退出符
```

一张图总结 Docker 的命令



常见仓库介绍

本章将介绍常见的一些仓库和镜像的功能，使用方法和生成它们的 Dockerfile 等。包括 Ubuntu、CentOS、MySQL、MongoDB、Redis、Nginx、Wordpress、Node.js 等。

Ubuntu

基本信息

[Ubuntu](#) 是流行的 Linux 发行版，其自带软件版本往往较新一些。该仓库提供了 Ubuntu 从 12.04 ~ 14.10 各个版本的镜像。

使用方法

默认会启动一个最小化的 Ubuntu 环境。

```
$ sudo docker run --name some-ubuntu -i -t ubuntu
root@523c70904d54:/#
```

Dockerfile

- [12.04 版本](#)
- [14.04 版本](#)
- [14.10 版本](#)

CentOS

基本信息

CentOS 是流行的 Linux 发行版，其软件包大多跟 RedHat 系列保持一致。该仓库提供了 CentOS 从 5 ~ 7 各个版本的镜像。

使用方法

默认会启动一个最小化的 CentOS 环境。

```
$ sudo docker run --name some-centos -i -t centos bash
bash-4.2#
```

Dockerfile

- [CentOS 5 版本](#)
- [CentOS 6 版本](#)
- [CentOS 7 版本](#)

MySQL

基本信息

MySQL 是开源的关系数据库实现。该仓库提供了 MySQL 各个版本的镜像，包括 5.6 系列、5.7 系列等。

使用方法

默认会在 `3306` 端口启动数据库。

```
$ sudo docker run --name some-mysql -e MYSQL_ROOT_PASSWORD=mysecretpassword -d mysql
```

之后就可以使用其它应用来连接到该容器。

```
$ sudo docker run --name some-app --link some-mysql:mysql -d application-that-uses-mysql
```

或者通过 `mysql`。

```
$ sudo docker run -it --link some-mysql:mysql --rm mysql sh -c 'exec mysql -h"$MYSQL_PORT'
```

Dockerfile

- [5.6 版本](#)
- [5.7 版本](#)

MongoDB

基本信息

MongoDB 是开源的 NoSQL 数据库实现。该仓库提供了 MongoDB 2.2 ~ 2.7 各个版本的镜像。

使用方法

默认会在 27017 端口启动数据库。

```
$ sudo docker run --name some-mongo -d mongo
```

使用其他应用连接到容器，可以用

```
$ sudo docker run --name some-app --link some-mongo:mongo -d application-that-uses-mongo
```

或者通过 `mongo`

```
$ sudo docker run -it --link some-mongo:mongo --rm mongo sh -c 'exec mongo "$MONGO_PORT_2
```

Dockerfile

- [2.2 版本](#)
- [2.4 版本](#)
- [2.6 版本](#)
- [2.7 版本](#)

Redis

基本信息

Redis 是开源的内存 Key-Value 数据库实现。该仓库提供了 Redis 2.6 ~ 2.8.9 各个版本的镜像。

使用方法

默认会在 6379 端口启动数据库。

```
$ sudo docker run --name some-redis -d redis
```

另外还可以启用持久存储。

```
$ sudo docker run --name some-redis -d redis redis-server --appendonly yes
```

默认数据存储位置在 `VOLUME/data`。可以使用 `--volumes-from some-volume-container` 或 `-v /docker/host/dir:/data` 将数据存放到本地。

使用其他应用连接到容器，可以用

```
$ sudo docker run --name some-app --link some-redis:redis -d application-that-uses-redis
```

或者通过 `redis-cli`

```
$ sudo docker run -it --link some-redis:redis --rm redis sh -c 'exec redis-cli -h "$REDIS'
```

Dockerfile

- [2.6 版本](#)
- [最新 2.8 版本](#)

Nginx

基本信息

[Nginx](#) 是开源的高效的 Web 服务器实现，支持 HTTP、HTTPS、SMTP、POP3、IMAP 等协议。该仓库提供了 Nginx 1.0 ~ 1.7 各个版本的镜像。

使用方法

下面的命令将作为一个静态页面服务器启动。

```
$ sudo docker run --name some-nginx -v /some/content:/usr/share/nginx/html:ro -d nginx
```

用户也可以不使用这种映射方式，通过利用 Dockerfile 来直接将静态页面内容放到镜像中，内容为

```
FROM nginx
COPY static-html-directory /usr/share/nginx/html
```

之后生成新的镜像，并启动一个容器。

```
$ sudo docker build -t some-content-nginx .
$ sudo docker run --name some-nginx -d some-content-nginx
```

开放端口，并映射到本地的 8080 端口。

```
sudo docker run --name some-nginx -d -p 8080:80 some-content-nginx
```

Nginx 的默认配置文件路径为 `/etc/nginx/nginx.conf`，可以通过映射它来使用本地的配置文件，例如

```
docker run --name some-nginx -v /some/nginx.conf:/etc/nginx/nginx.conf:ro -d nginx
```

使用配置文件时，为了在容器中正常运行，需要保持 `daemon off;`。

Dockerfile

- [1 ~ 1.7 版本](#)

WordPress

基本信息

[WordPress](#) 是开源的 Blog 和内容管理系统框架，它基于 PHP 和 MySQL。该仓库提供了 WordPress 4.0 版本的镜像。

使用方法

启动容器需要 MySQL 的支持，默认端口为 80。

```
$ sudo docker run --name some-wordpress --link some-mysql:mysql -d wordpress
```

启动 WordPress 容器时可以指定的一些环境参数包括

- `-e WORDPRESS_DB_USER=...` 缺省为 “root”
- `-e WORDPRESS_DB_PASSWORD=...` 缺省为连接 mysql 容器的环境变量 `MYSQL_ROOT_PASSWORD` 的值
- `-e WORDPRESS_DB_NAME=...` 缺省为 “wordpress”
- `-e WORDPRESS_AUTH_KEY=...` , `-e WORDPRESS_SECURE_AUTH_KEY=...` , `-e WORDPRESS_LOGGED_IN_KEY=...` , `-e WORDPRESS_NONCE_KEY=...` , `-e WORDPRESS_AUTH_SALT=...` , `-e WORDPRESS_SECURE_AUTH_SALT=...` , `-e WORDPRESS_LOGGED_IN_SALT=...` , `-e WORDPRESS_NONCE_SALT=...` 缺省为随机 sha1 串

Dockerfile

- [4.0 版本](#)

Node.js

基本信息

[Node.js](#)是基于 JavaScript 的可扩展服务端和网络软件开发平台。该仓库提供了 Node.js 0.8 ~ 0.11 各个版本的镜像。

使用方法

在项目中创建一个 Dockerfile。

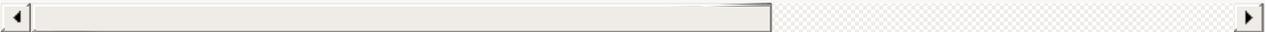
```
FROM node:0.10-onbuild
# replace this with your application's default port
EXPOSE 8888
```

然后创建镜像，并启动容器

```
$ sudo docker build -t my-nodejs-app
$ sudo docker run -it --rm --name my-running-app my-nodejs-app
```

也可以直接运行一个简单容器。

```
$ sudo docker run -it --rm --name my-running-script -v "$(pwd)":/usr/src/myapp -w /usr/src
```



Dockerfile

- [0.8 版本](#)
- [0.10 版本](#)
- [0.11 版本](#)

资源链接

- Docker 主站点: <https://www.docker.io>
- Docker 注册中心API: http://docs.docker.com/reference/api/registry_api/
- Docker Hub API: http://docs.docker.com/reference/api/docker-io_api/
- Docker 远端应用API: http://docs.docker.com/reference/api/docker_remote_api/
- Dockerfile 参考 : <https://docs.docker.com/reference/builder/>
- Dockerfile 最佳实践 : https://docs.docker.com/articles/dockerfile_best-practices/