1.squid服务搭建

squid的服务搭建，前面的视频中实操已经掌握了的，这里补充一个文档

准备工作：

- 公网服务器一台，Linux系统
- ssh远程工具，远程操作公网服务器
- 服务器安装Docker服务

准备好了，开始搭建环境

1. 安装squid镜像

squid有现成的Docker镜像，使用命令直接拉取

```
sudo docker pull sameersbn/squid
```

2. 准备配置文件

squid的默认是只允许内网访问的，不允许公网访问。

所以在最开始的时候，就需要把配置文件给准备好，一会启动的时候，直接给挂载到容器上，一步到位

在任意位置，新建一个文件，名称是 squid.conf，内容如下：

```
acl all src 0.0.0.0/0.0.0.0
acl SSL_ports port 443
acl Safe_ports port 80 # http
acl Safe_ports port 443 # https
acl CONNECT method CONNECT
http_access allow all
http_port 3128
visible_hostname proxy
```

3. 启动squid容器

一切准备就绪，把squid的镜像启动起来，如下命令：

```
sudo docker run -d --name squid -p 3128:3128 -v /path_to/squid.conf:/etc/squid/squid.conf   sameersbn/squid
```

简单介绍一下命令：

- sudo docker run 是启动镜像的命令
- -d是指后台运行
- –name squid是容器的名称
- -p 3128:3128 是宿主和容器的端口绑定，前宿主后容器
- -v /path_to/squid.conf/etc/squid/squid.conf 是文件的配置，前宿主后容器
- sameersbn/squid 最后这个就是指定镜像了

启动后，没报错，并会返回一个长字符串，就是成功了

4. 测试代理服务

在本地电脑测试，这里的测试方法是wget命令，适用Linux和MacOS系统，Windows请自行找方法测试。

首先打开终端，输入如下命令：

```
export http_proxy=http://xxx.xxx.xxx.xxx:3128

wget http://www.baidu.com
```

简单介绍：

1. 第一行命令是设置终端代理,xx.xx.xx.xx是你的公网IP，3128是squid的端口号
2. 第二行命令是下载一个网址，这里写的是百度

如果代理工作正常，是这样的输出结果：

如果代理异常，则肯定是失败的，贴个示意图：

没有安全验证的代理，是“裸贷”，非常的不安全，所以下面贴上带有安全加密的配置文件和启动命令。

5. 加密配置

```
acl localnet src 10.0.0.0/8     # RFC1918 possible internal network
acl localnet src 172.16.0.0/12    # RFC1918 possible internal network
```

```
acl localnet src 192.168.0.0/16    # RFC1918 possible internal network
acl localnet src fc00::/7       # RFC 4193 local private network range
acl localnet src fe80::/10      # RFC 4291 link-local (directly plugged) machines
acl localnet src 0.0.0.0/0.0.0.0
acl localnet src 0.0.0.0/8

acl SSL_ports port 443
acl Safe_ports port 80        # http
acl Safe_ports port 21        # ftp
acl Safe_ports port 443        # https
acl Safe_ports port 70        # gopher
acl Safe_ports port 210        # wais
acl Safe_ports port 1025-65535    # unregistered ports
acl Safe_ports port 280        # http-mgmt
acl Safe_ports port 488        # gss-http
acl Safe_ports port 591        # filemaker
acl Safe_ports port 777        # multiling http
acl CONNECT method CONNECT

# username&password auth config
auth_param basic program /usr/lib/squid/basic_ncsa_auth /etc/squid/squid_passwd
acl ncsa_users proxy_auth REQUIRED
http_access allow ncsa_users


http_access deny !Safe_ports
http_access deny CONNECT !SSL_ports
http_access allow localhost manager
http_access deny manager
http_access deny to_localhost
http_access allow localnet
http_access allow localhost
http_access deny all
http_port 3128

refresh_pattern ^ftp:        1440    20%    10080
refresh_pattern ^gopher:    1440    0%    1440
refresh_pattern -i (/cgi-bin/|\?) 0    0%    0
refresh_pattern (Release|Packages(.gz)*)$      0       20%     2880
refresh_pattern .        0    20%    4320
```

这个是配置文件，就是前面的squid.conf，用这个更安全

但是这里并未提及账号密码，因为账号密码需要额外的配置，所以下一步

6. 生成账号密码

这里的账号密码是需要用命令去生成，该命令是 `htpasswd` 。

如果你直接使用，系统报错，则需要做个安装，这里以ubuntu为例，安装命令是：

```
$ sudo apt install apache2-utils
```

确保系统中有htpasswd命令，然后开始生成

```
$ sudo htpasswd -c squid_passwd username
# 这里会提示你输入密码
# 然后再输入一次密码
# 最后提示成功，完事
```

命令行中，请使用你的账号，替换这里面的username。

密码也是输入你自己的。

在加密的配置文件和账号密码文件都准备好之后呢，把前面创建的squid容器停止并删除，然后使用命令行，重新创建一个，如下：

```
$ sudo docker run -d --name squid -p 3128:3128 -v /home/ubuntu/squid.conf:/etc/squid/squid.conf -v /home/ubuntu/squid_passwd:/etc/squid/squid_passwd sameersbn/squid
```

这条命令行，就是在前面的容器启动命令之上，加了一个密码文件的配置

```
/home/ubuntu/squid_passwd:/etc/squid/squid_passwd
```

最后就是再次测试，命令行如下：

```
$ export http_proxy=http://username:password@xxx.xxx.xxx.xxx:3128
$ wget http://www.baidu.com
```

2. vps介绍

搭建squid测试服务，我们用的是腾讯云的云服务器，也是vps的一种。

那vps是什么意思？它包含哪些主机呢？

下面是我从百科上拿来的介绍

> VPS（Virtual Private Server 虚拟专用服务器）技术，将一台服务器分割成多个虚拟专享服务器的优质服务。实现VPS的技术分为容器技术，和虚拟化技术在容器或虚拟机中，每个VPS都可选配独立公网IP地址、独立操作系统、实现不同VPS间磁盘空间、内存、CPU资源、进程和系统配置的隔离，为用户和应用程序模拟出“独占”使用计算资源的体验。VPS可以像独立服务器一样，重装操作系统，安装程序，单独重启服务器。VPS为使用者提供了管理配置的自由，可用于企业虚拟化，也可以用于IDC资源租用。

介绍的还算详细，但是肯定不适合小白理解，我在这里做个直播的翻译和概括。

> vps是一个创造云服务器的系统，而且特指非物理服务器。
>
> 你买的云服务商的云服务器，包含完整的真实物理机系统体验，有完备的网络和设备信息和性能。

vps的简单介绍就是创建云服务器系统的，那它包含哪些类型的主机呢？简单概括一下：

1. 固定IP的云服务器，例如演示视频中用的腾讯云服务器
2. 不固定IP的云服务器，ADSL拨号服务器
3. 共享IP的云服务器，共用一个IP因为可以指定不同的端口

当然还有不在此列的其他服务器，例如：

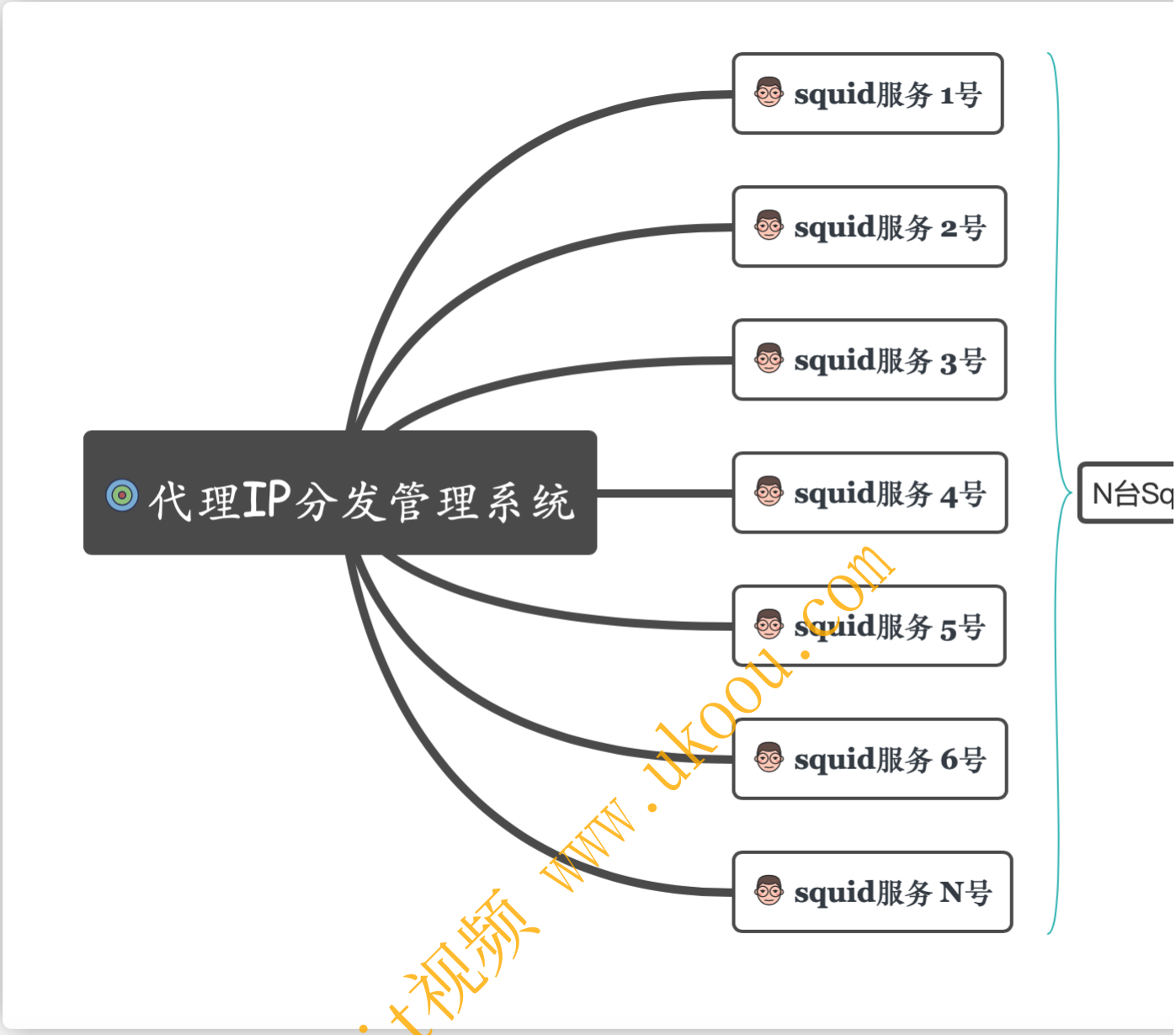
- 独立物理云服务器
- 托管物理机服务器

这里面介绍的不同种类的服务器，是因技术和环境的不同，而叫做不同的服务器，在使用来看，都是一样的。

对于爬虫来说，ADSL拨号服务器，是最适合爬虫的，因为ip来自于服务器拨号运营商随机分配的，而且带宽恒定，重点是IP可以根据网络的重启来重新分配IP。

3. 代理池架构介绍

代理池的架构，主要在于部署squid服务的服务器数量，如下截图：



因为每台squid服务，最有价值的就是服务器的代理IP，所以这里的服务器，最好统一使用ADSL拨号服务器。

特别说明一下squid的工作流程：

1. ADSL服务器启动后，自动启动squid服务，监听一个端口，这个端口要记录，也许还需要有效时间和账号密码信息等。
2. ADSL服务器服务器定期重启网络服务，例如5分钟，则IP信息每5分钟切换一个。
3. ADSL服务器还需要定期发起一个请求，例如2分钟发送一次请求，内容是端口，发送给“代理IP分发管理系统”
4. 特别说明，定期发送请求，不需要将IP发过去，因为请求过去了，管理系统就可以知道是哪个ip发过来的。

这套ADSL服务器的软件设备很简单，squid服务和定期请求服务，都是非常简单的，适合大批量的部署。

有了服务器，图中还提及了代理IP分发管理系统，这个就是一个web系统，功能如下：

- 记录收到请求ip和端口，以及请求的当前时间。
- 管理系统的账号密码管理，提取IP要识别身份。
- 收到用户的提取IP请求，直接返回爬虫需要的代理ip和端口，供爬虫使用。
- 定期查看数据库中过期的代理和新增的代理，做代理数据的清理操作。

4. 成本计算

现在的代理IP服务商的，都流行用百万IP做噱头，虽然知道大部分是假的，但我还是想计算一下，按我们这套方案，每天的百万IP，需要多少RMB。

首先是ADSL服务器的获取IP方式，每次重启网络或者重启服务器，就可以得到一个新的IP，并且这里不算会出现重复IP的概率，全部默认新IP。

爬虫使用的代理IP，最经典的还是3-5分钟的短效代理IP，虽然达不到5分钟，但是包含了重启网络或者服务器时间，直接算5分钟一个IP。

5分钟一个IP，则一天是24*60/5 = 288（个）。

一台服务器，每5分钟切换一个IP，一天就是288个全新的IP地址。

百万IP的入门，一百万个IP，则需要3473台服务器。

关于ADSL服务器的报价，大同小异，因为代理服务器不需要很高的配置，选择基础配置ADSL服务器即可。云服务器小厂商可以找到很多的ADSL服务器。

按基本的60月/月来算，一共需要3473*60 = 208380，也就是20万，一个月20W，妥妥的每天百万IP。

当然这是实打实的服务，商用肯定是不合适的，因为IP有丢失率一说，百分之十不大，那都是钱呐。

5. 技术方案

技术方案不止这一种，除了squid代理服务，还有tinyproxy等其他的代理工具。

企业内部使用，通常会选择 自建代理IP池+购买部分代理IP 的复合方案，这样才能做到有效的利用了自由的IP，且弹性的扩充少量IP，而且成本是非常可控的。