## C++开源tesseract语言文字识别程序

### 实验介绍

本实验对C++语言编写的文字识别tesseract程序，在ECS主机上进行编译安装，验证对图片文件进行文字识别功能，包括中文、英文两种语言。

**tesseract依赖的图像处理库**源码包**leptonica是采用C语言编写**的，需**要首先编译安装**，同时对文字识别程序需要的**中文、英文语言包**要下载到主机。

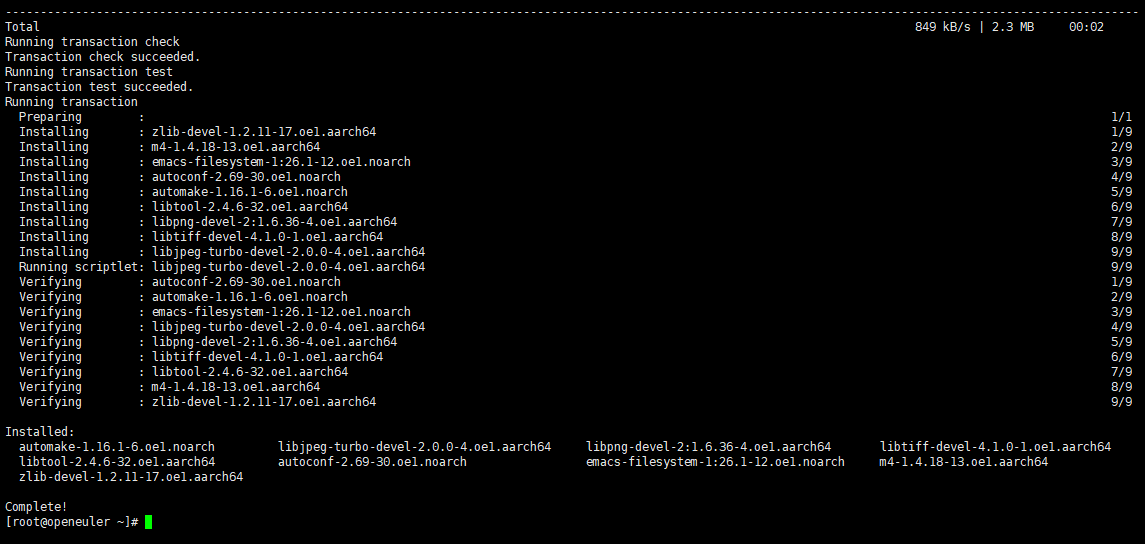
通过实验，可以学习在鲲鹏ECS，openEuler操作系统下，对linux c/c++语言编写的程序，进行编译部署的综合场景。

### 安装依赖基础组件

配置tesseract所依赖的环境，安装构建系统的工具automake、创建库的工具libtool、C++编译器，png图片库、jpg图片库、tiff图片库等。

登录主机，执行下面命令，安装上述依赖组件：

yum install --nogpgcheck automake libtool gcc-c++ libjpeg-devel libpng-devel libtiff-devel -y



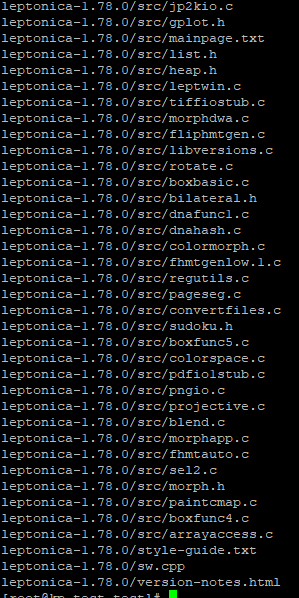
### 上传实验所需的文件

将 实验用文件 文件夹中的所有文件一个个的上传到/root文件夹下，/root文件夹是root用户的home文件夹，每次用cd命令可以直接回到这个文件夹

### 安装leptonica

**执行cd命令回到/root文件夹下，解压leptonica源码包**，**执行命令：**

tar zxvf leptonica-1.78.0.tar.gz



**编译和安装leptonica**：

进入解压后的目录：

cd leptonica-1.78.0

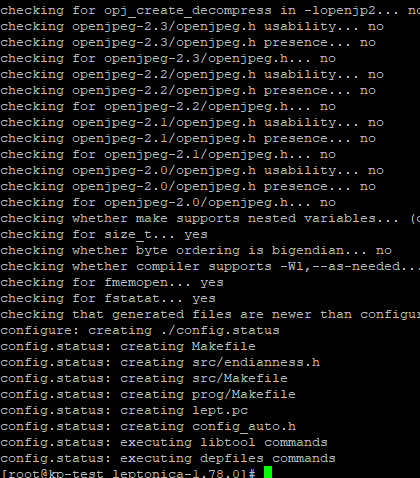
执行命令：

./configure --prefix=/usr/

或

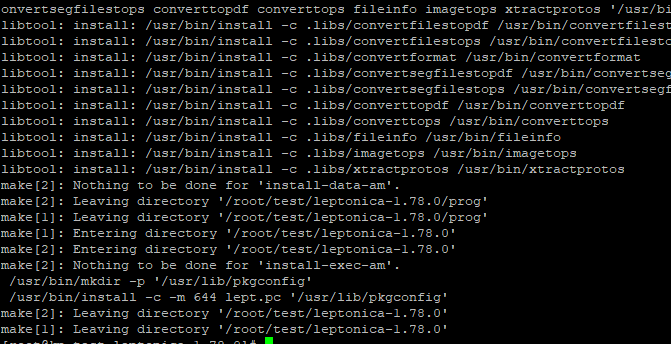
[root@kp-test test]# cd leptonica-1.78.0

[root@kp-test leptonica-1.78.0]# ./configure --prefix=/usr/



编译安装：make -j4 && make install，参数j4根据实际cpu内核数调整。

make -j4 && make install



**配置leptonica环境。**

编辑文件：

/etc/profile

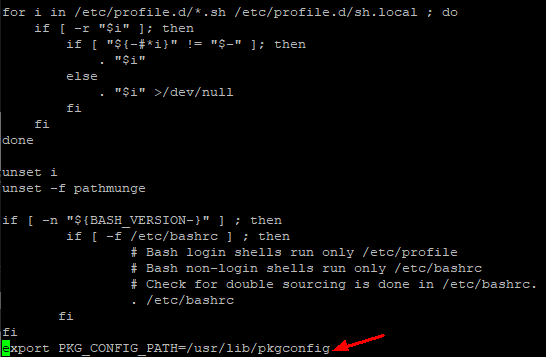
输入命令：

vi /etc/profile

vi是命令行下常用的文本编辑器。执行上述命令进入vi之后移动光标到最后一行。之后点击键盘上的i字母进入insert模式，这个时候就可以编辑这个文本了。

增加环境变量：

export PKG\_CONFIG\_PATH=/usr/lib/pkgconfig



编辑结束后，要保存vi编辑后的文件，先按ESC按键退出insert模式，然后再同时按下shift键和:键，也就是输入一个冒号，这个时候左下角会出现一个冒号，进入命令模式：



输入wq回车就可以了，这样文件的修改就生效了，回到命令行了。

使环境变量生效：

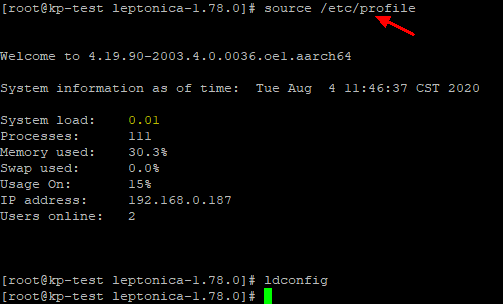
source /etc/profile

\*如果你实在学不会上面用vi编辑文件，你也可以在左侧文件操作栏里把/etc/profile下载下来，用本地机器的编辑器编辑一下，然后再上传覆盖掉。或者实在不行就每次启动一个命令行窗口的时候执行一下如下命令就行了。系统启动的时候会自动从/etc/profile里面获取所有的系统环境变量信息，你每次自己把PKG\_CONFIG\_PATH添加到系统环境变量里面也是一样的。

export PKG\_CONFIG\_PATH=/usr/lib/pkgconfig

刷新默认路径动态库为系统共享：

ldconfig



### 安装tesseract

**用cd命令回到/root文件夹下，解压tesseract源码包**：

执行命令：

unzip tesseract-main.zip

ls命令可以看到tesseract-main文件夹已经解压好了



**编译和安装tesseract：**

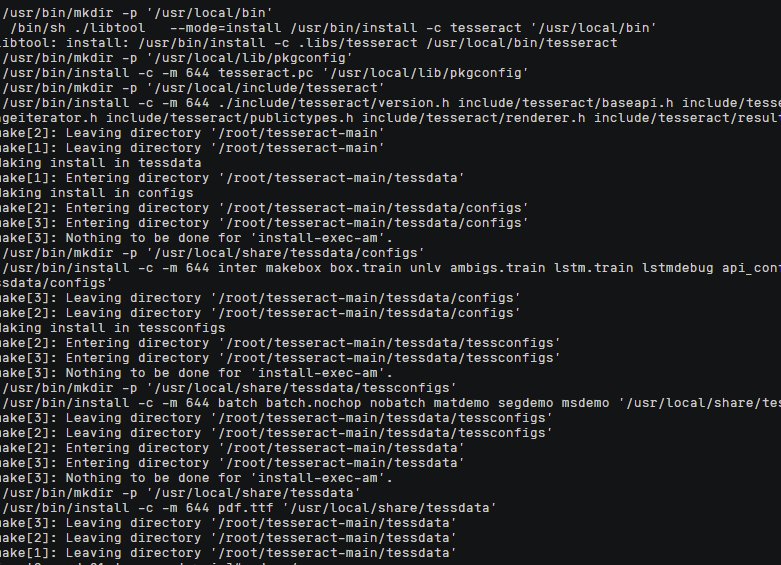
进入目录：

cd tesseract-main

执行命令：（这一步可能要花几分钟时间，可以先刷刷手机）

.**/autogen.sh && ./configure && make -j4 && make install**





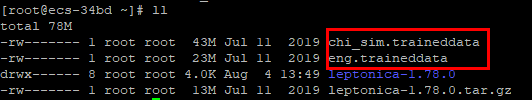
**上传语言包：**

tesseract语言包需要存放在usr/local/share/tessdata，以便于tesseract识别。

在/root文件夹下执行命令：

cp eng.traineddata /usr/local/share/tessdata -r

cp chi\_sim.traineddata /usr/local/share/tessdata -r



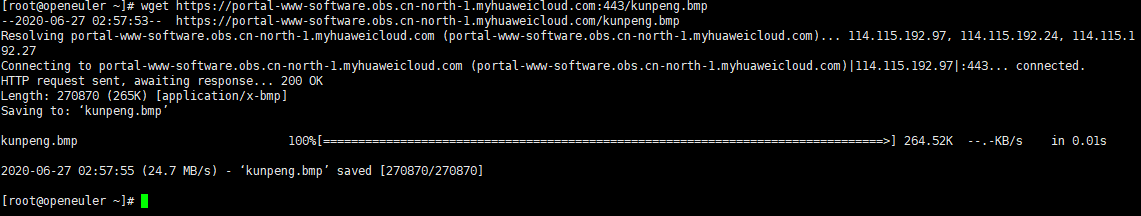
**下载kunpeng文字图片进行demo测试**：

下载英文文字，图片用例：kunpeng.bmp。

C:\Users\x84113935\Downloads\kunpeng.bmp

执行命令：

wget <https://portal-www-software.obs.cn-north-1.myhuaweicloud.com:443/kunpeng.bmp>

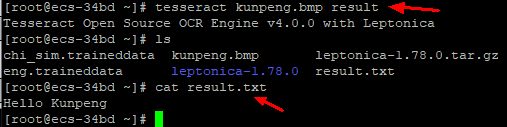


执行命令，解析信息生成到result.txt：

tesseract kunpeng.bmp result

执行命令查看结果：

cat result.txt

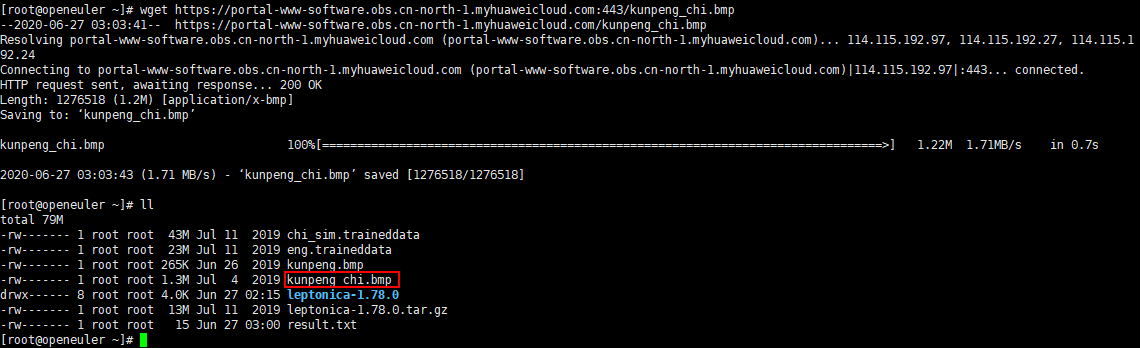


下载中文图片，图片用例：kunpeng\_chi.bmp。



执行命令下载图片：

wget <https://portal-www-software.obs.cn-north-1.myhuaweicloud.com:443/kunpeng_chi.bmp>

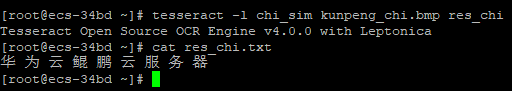


执行命令，解析信息生成到result.txt：

tesseract -l chi\_sim kunpeng\_chi.bmp res\_chi

执行命令查看结果：

cat res\_chi.txt



你也可以自己尝试截图一些中文或者英文字符，然后上传到服务器端进行测试。

## 实验环境清理

**清理ECS资源**：

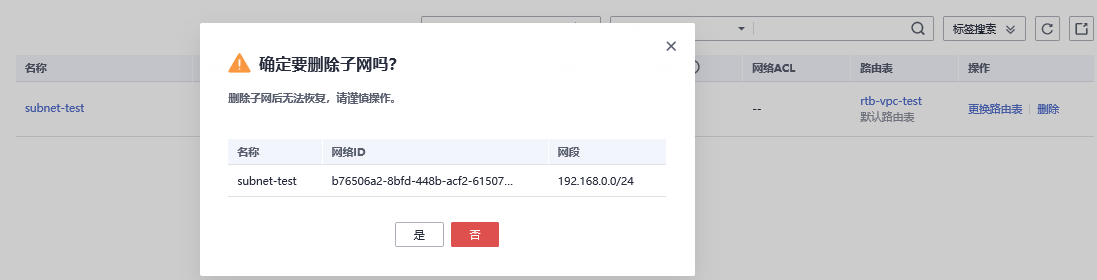
返回ECS控制台，选择kp-test云主机，然后点击“更多>删除。

在弹出的对话框中勾选“释放云服务器绑定的弹性公网IP地址”和“删除云服务器挂载的数据盘”，然后点击“是”，删除ECS。



**清理安全组、子网、EPC**：







## 实验总结与课后思考

思考1：修改实验2，调用回写方法，新增参数回写配置文件。

思考2：修改实验3学生管理系统文件版代码，支持录入多个学科成绩。

思考3：应用编程C语言与C++语言的联系和区别。