《openEuler操作系统》

内核实验手册

Taishan物理服务器版

v1.0

华为技术有限公司

中科院软件所

目录

[前言 1](#_Toc59480131)

[1 实验一 openEuler的安装及内核编译 3](#_Toc59480132)

[1.1 实验介绍 3](#_Toc59480133)

[1.1.1 任务描述 3](#_Toc59480134)

[1.2 实验目的 3](#_Toc59480135)

[1.3 实验任务 3](#_Toc59480136)

[1.3.1 openEuler操作系统安装 3](#_Toc59480137)

[1.3.2 openEuler内核编译与安装 21](#_Toc59480138)

[1.3.3 Hello, world! 22](#_Toc59480139)

[2 实验二 内存管理 24](#_Toc59480140)

[2.1 实验介绍 24](#_Toc59480141)

[2.1.1 相关知识 24](#_Toc59480142)

[2.1.2 任务描述 25](#_Toc59480143)

[2.2 实验目的 25](#_Toc59480144)

[2.3 实验任务 25](#_Toc59480145)

[2.3.1 使用kmalloc分配1KB，8KB的内存，打印指针地址 25](#_Toc59480146)

[2.3.2 使用vmalloc分配8KB、1MB、64MB的内存，打印指针地址 26](#_Toc59480147)

[2.3.3 实验结果分析 27](#_Toc59480148)

[2.4 思考题 28](#_Toc59480149)

[3 实验三 进程管理 30](#_Toc59480150)

[3.1 实验介绍 30](#_Toc59480151)

[3.1.1 任务描述 30](#_Toc59480152)

[3.2 实验目的 30](#_Toc59480153)

[3.3 实验任务 30](#_Toc59480154)

[3.3.1 创建内核进程 30](#_Toc59480155)

[3.3.2 打印输出当前系统CPU负载情况 33](#_Toc59480156)

[3.3.3 打印输出当前处于运行状态的进程的PID和名字 35](#_Toc59480157)

[3.3.4 使用cgroup实现限制CPU核数 36](#_Toc59480158)

[3.3.5 使用cgroup实现不允许访问U盘 41](#_Toc59480159)

[4 实验四 中断和异常管理 44](#_Toc59480160)

[4.1 实验介绍 44](#_Toc59480161)

[4.1.1 相关知识 44](#_Toc59480162)

[4.1.2 任务描述 45](#_Toc59480163)

[4.2 实验目的 45](#_Toc59480164)

[4.3 实验任务 46](#_Toc59480165)

[4.3.1 使用tasklet实现打印helloworld 46](#_Toc59480166)

[4.3.2 用工作队列实现周期打印helloworld 47](#_Toc59480167)

[4.3.3 编写一个信号捕获程序，捕获终端按键信号 50](#_Toc59480168)

[5 实验五 内核时间管理 55](#_Toc59480169)

[5.1 实验介绍 55](#_Toc59480170)

[5.1.1 任务描述 55](#_Toc59480171)

[5.2 实验目的 55](#_Toc59480172)

[5.3 实验任务 55](#_Toc59480173)

[5.3.1 调用内核时钟接口打印当前时间 55](#_Toc59480174)

[5.3.2 编写timer，在特定时刻打印 hello,world 57](#_Toc59480175)

[5.3.3 调用内核时钟接口，监控累加计算代码的运行时间 59](#_Toc59480176)

[6 实验六 设备管理 61](#_Toc59480177)

[6.1 USB设备驱动程序实验 61](#_Toc59480178)

[6.1.1 相关知识 61](#_Toc59480179)

[6.1.2 任务描述 66](#_Toc59480180)

[6.1.3 实验步骤 66](#_Toc59480181)

[6.2 内核模块测试硬盘的读写速率实验 68](#_Toc59480182)

[6.2.1 相关知识 68](#_Toc59480183)

[6.2.2 任务描述 70](#_Toc59480184)

[6.2.3 实验目的 70](#_Toc59480185)

[6.2.4 实验步骤 70](#_Toc59480186)

[7 实验七 文件系统 72](#_Toc59480187)

[7.1 实验介绍 72](#_Toc59480188)

[7.1.1 任务描述 72](#_Toc59480189)

[7.2 实验目的 72](#_Toc59480190)

[7.3 实验任务 72](#_Toc59480191)

[7.3.1 为Ext4文件系统添加扩展属性 72](#_Toc59480192)

[7.3.2 注册一个自定义的文件系统类型 78](#_Toc59480193)

[7.3.3 在/proc下创建目录 80](#_Toc59480194)

[7.3.4 使用sysfs文件系统传递内核模块参数 82](#_Toc59480195)

[8 实验八 网络管理 86](#_Toc59480196)

[8.1 实验介绍 86](#_Toc59480197)

[8.1.1 任务描述 86](#_Toc59480198)

[8.2 实验目的 86](#_Toc59480199)

[8.3 实验任务 86](#_Toc59480200)

[8.3.1 编写基于socket的UDP发送接收程序 86](#_Toc59480201)

[8.3.2 使用tshark抓包 91](#_Toc59480202)

[8.3.3 使用setsockopt发送记录路由选项 92](#_Toc59480203)

[9 内核虚拟化 95](#_Toc59480204)

[9.1 实验介绍 95](#_Toc59480205)

[9.1.1 相关知识 95](#_Toc59480206)

[9.2 实验目的 97](#_Toc59480207)

[9.3 实验任务 97](#_Toc59480208)

[9.3.1 在Taishan服务器中搭建QEMU虚拟机运行环境 97](#_Toc59480209)

[9.3.2 使用VNC登录虚拟机 104](#_Toc59480210)

前言

本实验手册是基于Taishan服务器物理机平台的openEuler操作系统内核编程实验手册，若无特别说明，术语“Taishan服务器”在本手册中都是指Taishan物理服务器。

就实验内容而言，本实验手册包括了openEuler操作系统的安装、内核编译和安装、内存管理、进程管理、中断异常管理、内核时间管理、设备管理、文件管理以及网络管理等内核相关实验。

这些内容不仅包含实验步骤，还在每章简要介绍了实验相关的原理和背景知识，以便读者能更好地理解操作系统内核的原理并进行实验。

一、实验网络环境介绍

PC: 192.168.101.X

Switch

Router 192.168.137.1 (Gateway)

Taishan Server

br0: 192.168.101.X

br1: 192.168.137.X

openEuler操作系统内核编程实验拓扑图

实验中Taishan物理服务器所在网络环境包含两个网段，其中网段192.168.101.X供终端PC机与Taishan服务器连接用；网段192.168.137.X使得Taishan服务器可以访问公网。您应该根据实际的实验环境配置网络，但必须保证终端能通过VNC及ssh连接服务器，且服务器能访问公网。

图中的网桥br0和br1是为在Taishan物理服务器中创建虚拟机使用的。

二、实验设备介绍

实验设备如下表所示：

|  |  |  |
| --- | --- | --- |
| 设备名称 | 设备型号 | 软件版本 |
| 泰山服务器 | 2280 V2 | openEuler 20.03 LTS 服务器版安装 |
| 3层交换机 | 不限制 | 不限制 |
| 路由器 | 不限制 | 不限制 |
| PC终端 | 不限制 | Windows 10 / macOS / Linux / Unix  TigerVNC客户端 |

三、本手册用到的参考源代码



# 实验一 openEuler的安装及内核编译

## 实验介绍

本实验通过安装openEuler操作系统、编译安装openEuler操作系统新内核以及简单的内核模块编程任务操作带领大家了解操作系统及内核编程。

### 任务描述

* 安装openEuler操作系统
* 编译安装openEuler操作系统新内核
* 简单的内核模块编程实验，在内核模块中打印“Hello, world!”

## 实验目的

* 学习掌握如何安装操作系统
* 学习掌握如何编译操作系统内核
* 了解内核模块编程。

## 实验任务

### openEuler操作系统安装

准备安装镜像

您可以访问openEuler官方网站（[openeuler.org](https://openeuler.org/)）下载AArch64平台的openEuler 20.03 LTS安装包：[openEuler-20.03-LTS-aarch64-dvd.iso](https://repo.openeuler.org/openEuler-20.03-LTS/ISO/aarch64/openEuler-20.03-LTS-aarch64-dvd.iso)（文件大小约4.3 GiB）。

安装openEuler操作系统

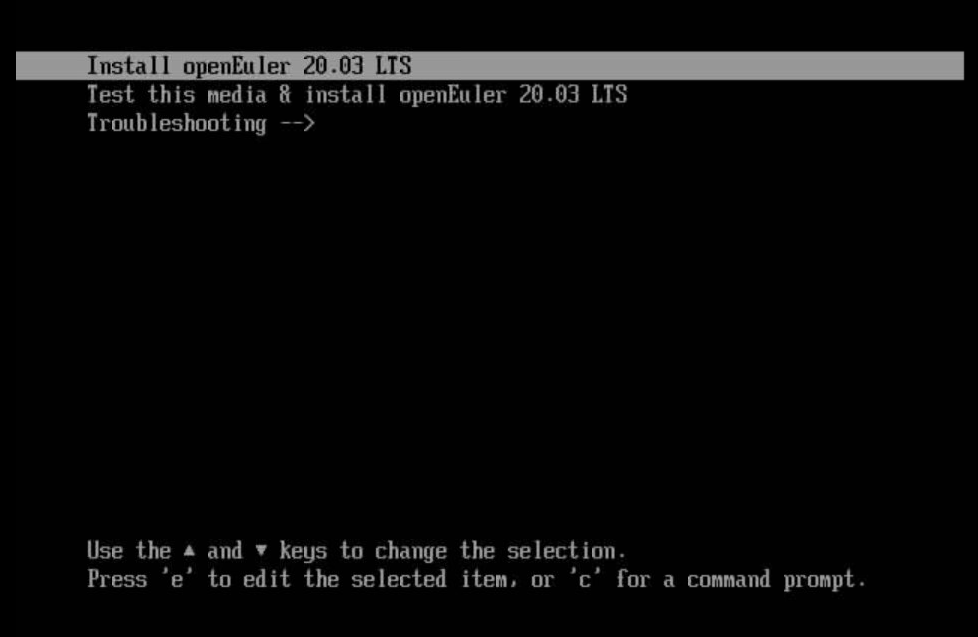
根据自己的实际情况，将安装介质与Taishan服务器连接，然后开机开始安装openEuler操作系统：



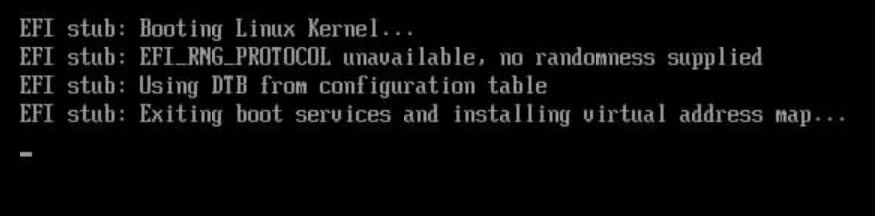
启动过程：



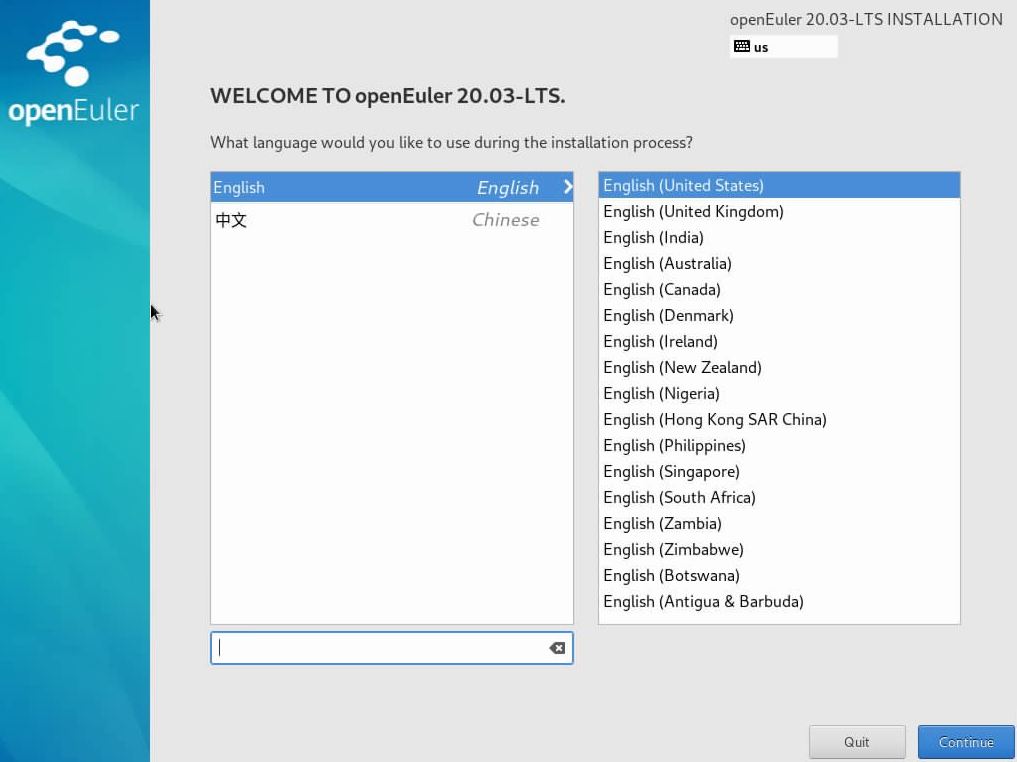
选择Install openEuler 20.03 LTS进行安装：



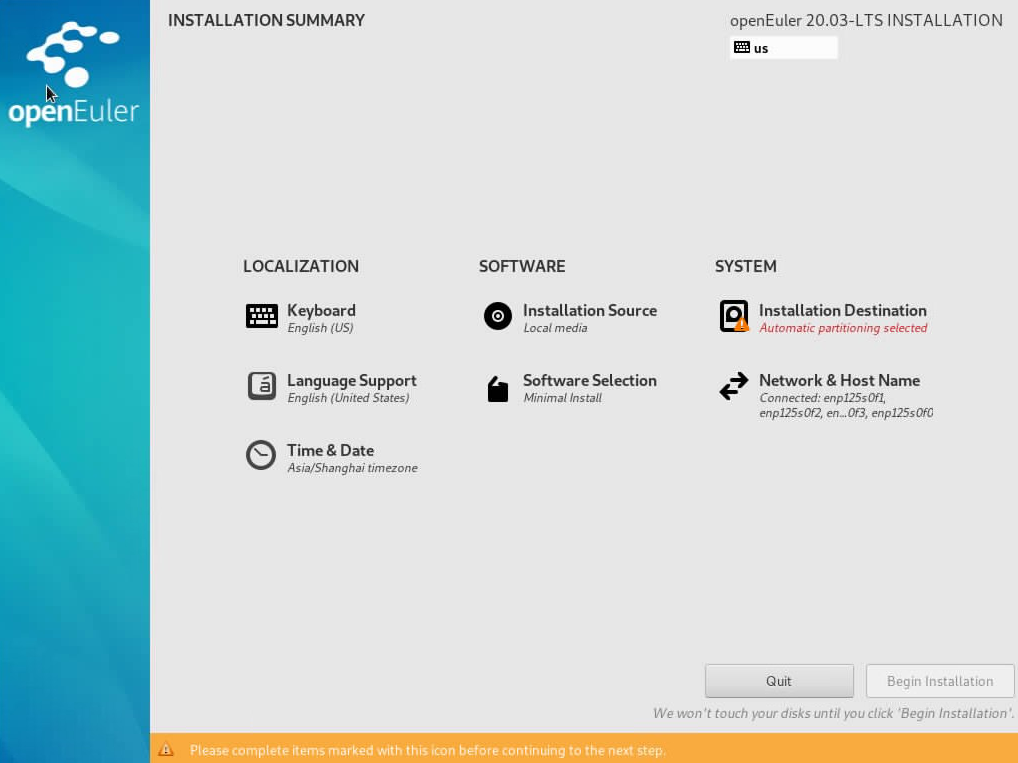
加载过程：



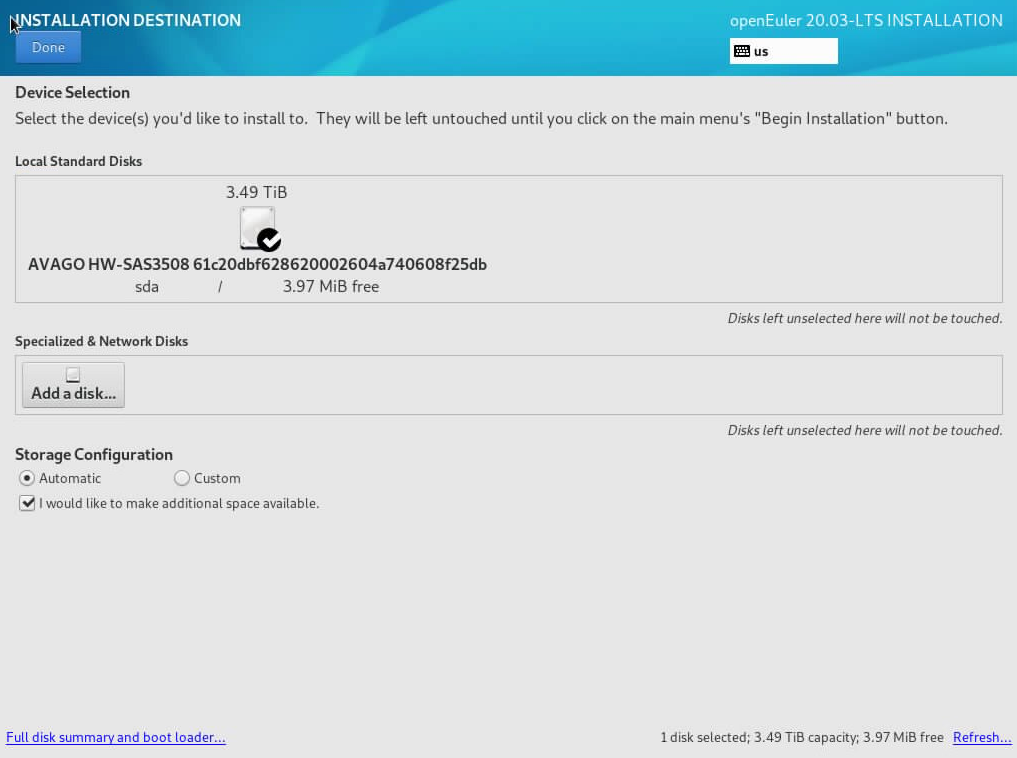
保持默认的英文安装：



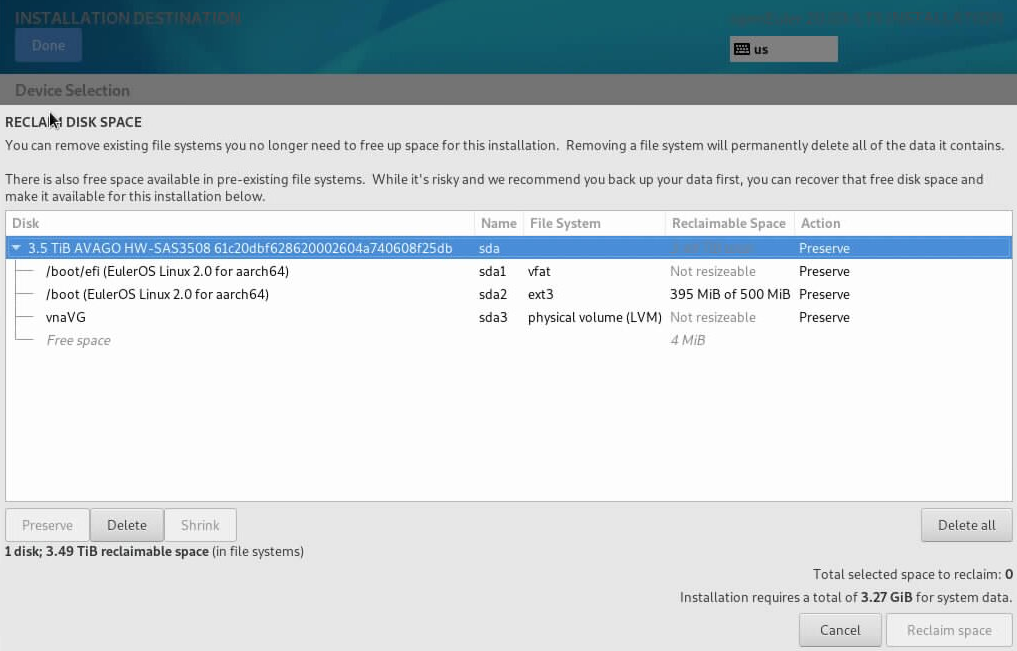
请在“SYSTEM”栏目的“Installation Destination”里确认安装位置：



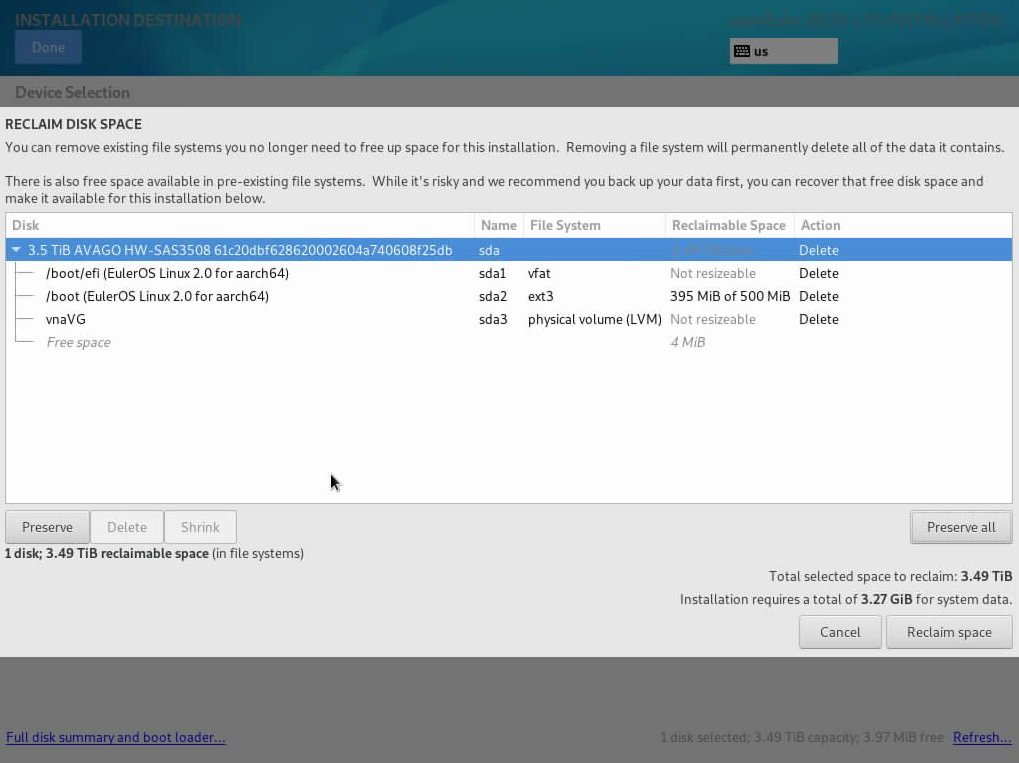
一般保持默认值即可，如果以前安装过操作系统导致空间不够可以选中“I would like to make additional space available”选项，然后点击左上角的“Done”按钮进行设置：



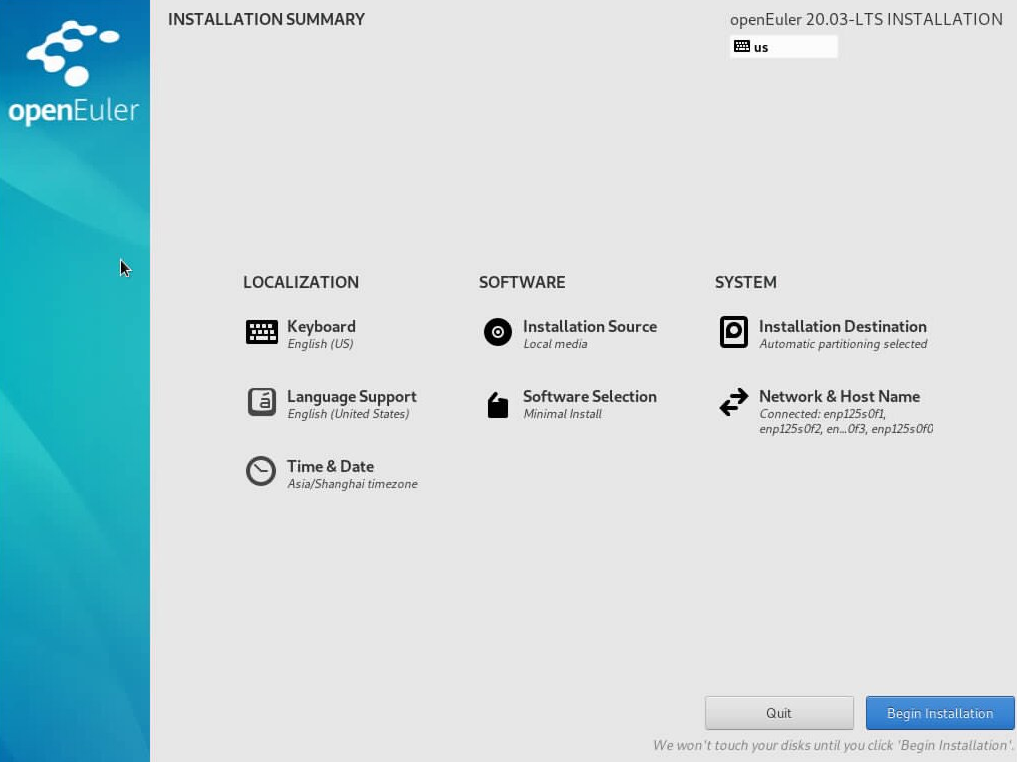
点击“Delete all”按钮：



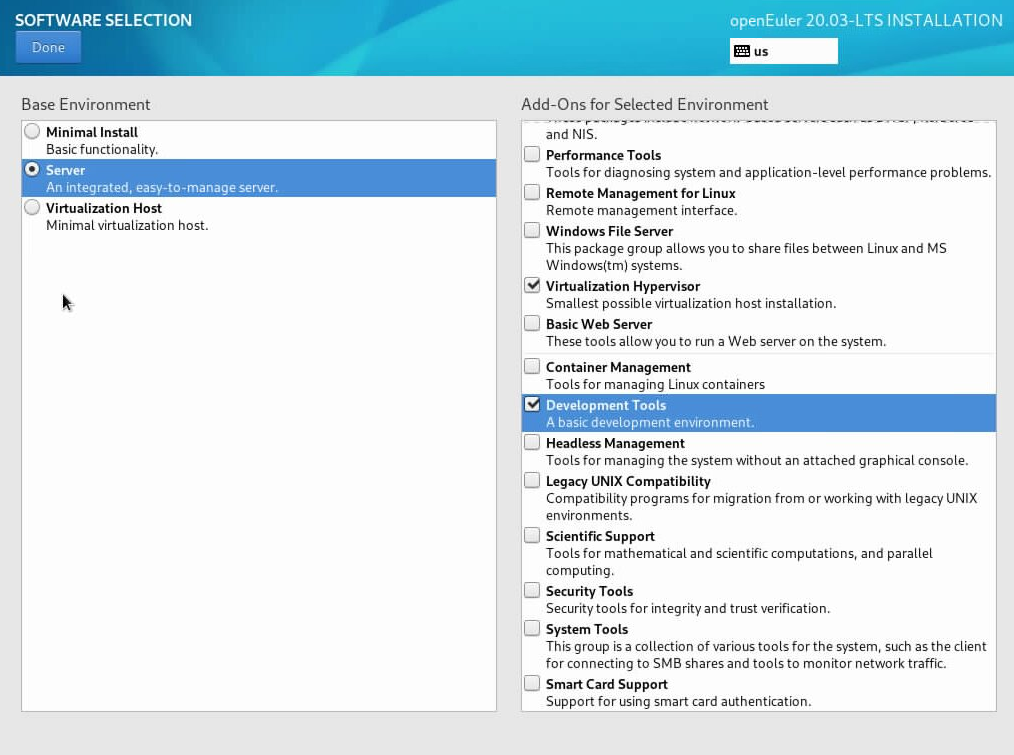
然后点击“Reclaim space”按钮：



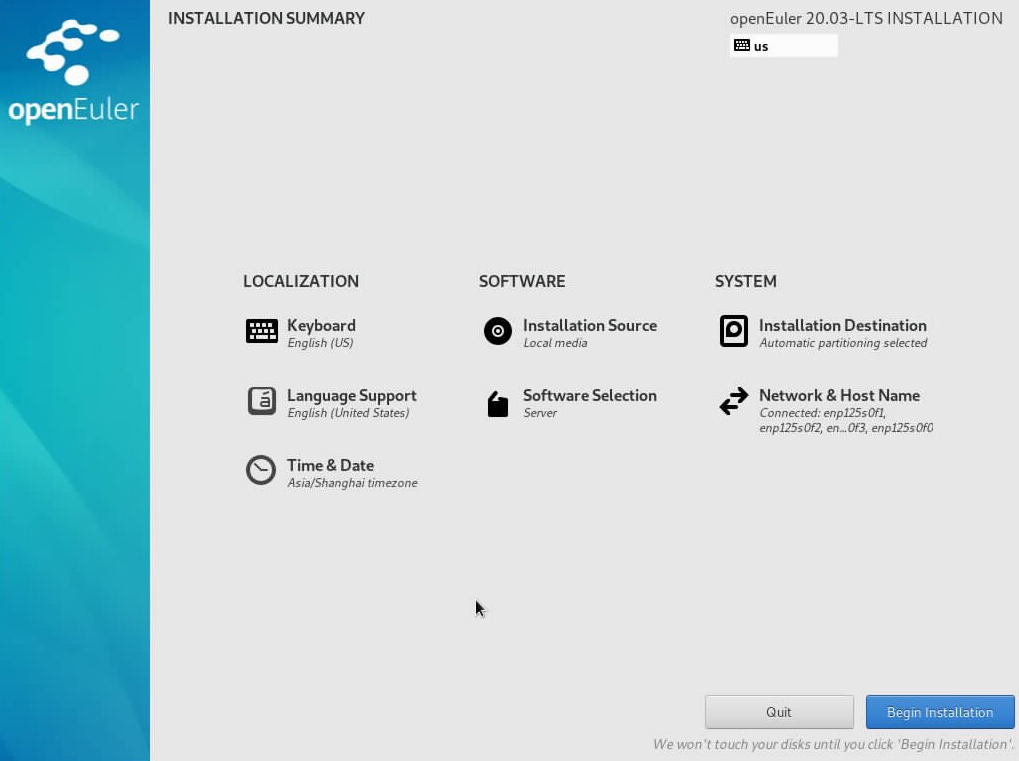
回到主界面后，点击“SOFTWARE”栏目下面的“Software Selection”按钮：

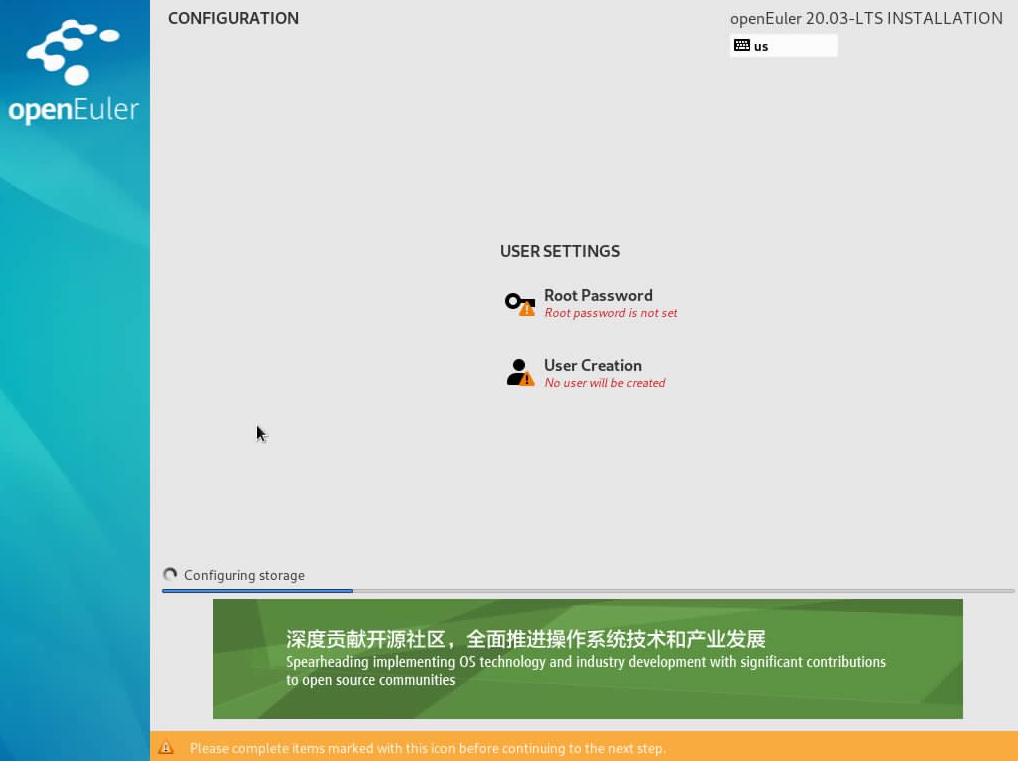


选中左边的“Server”，勾选右边的“Development Tools”和“Virtulaization Hypervisor”项：

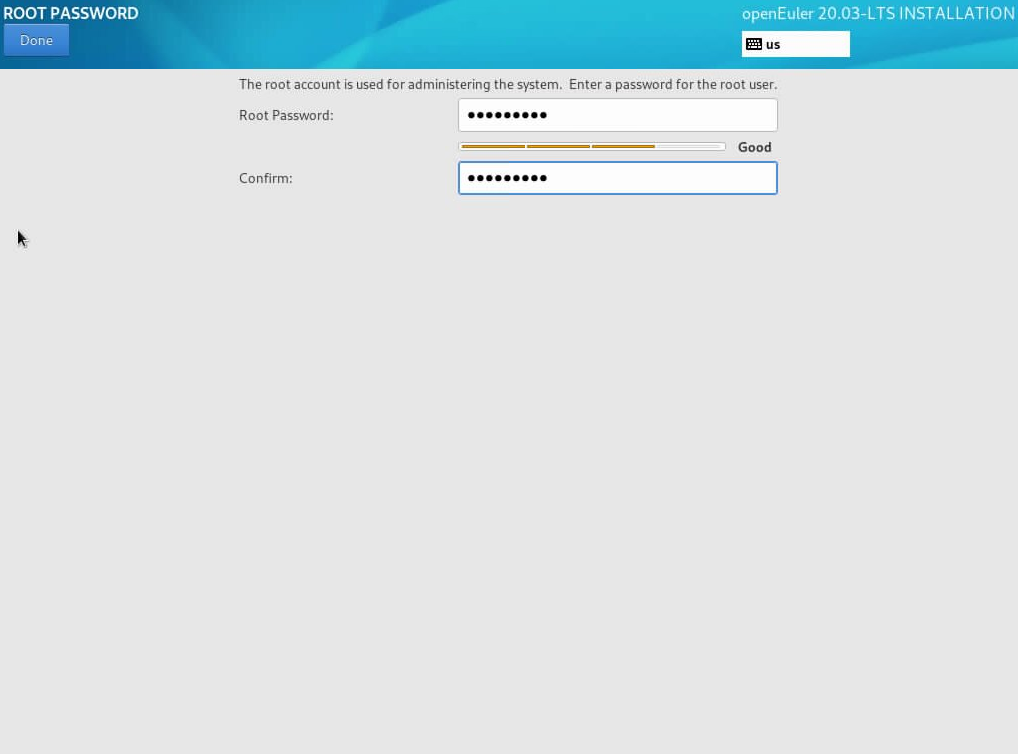


然后点击左上角的“Done”按钮回到主界面，并点击右下角的“Begin Installation”按钮进行安装：

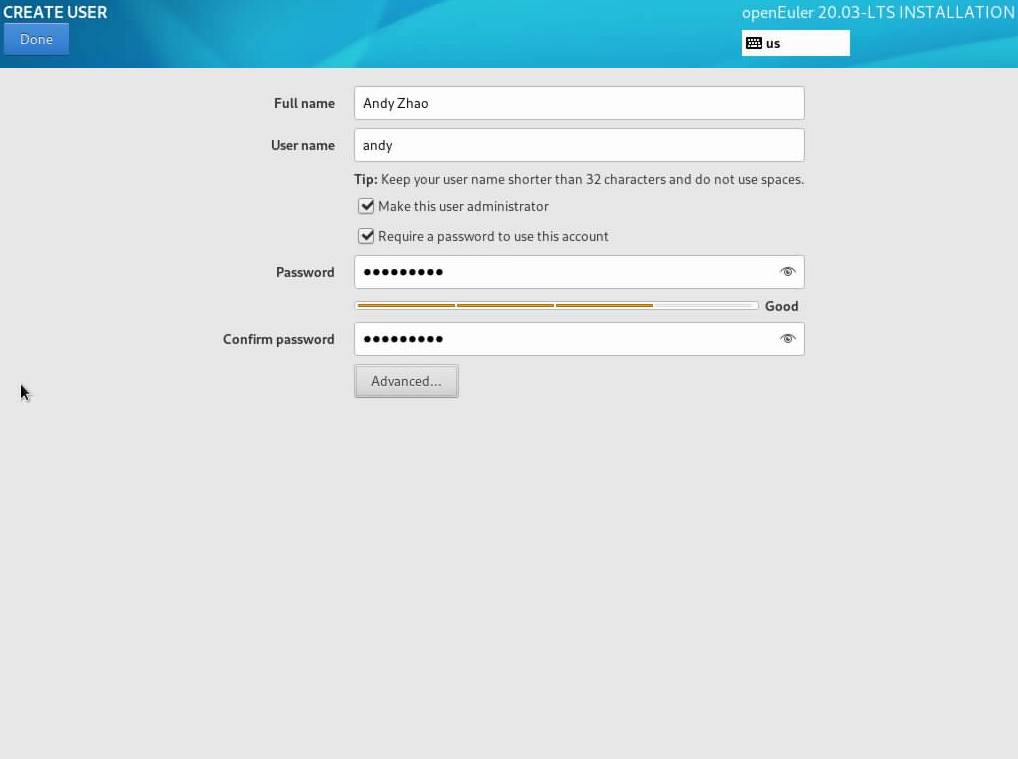




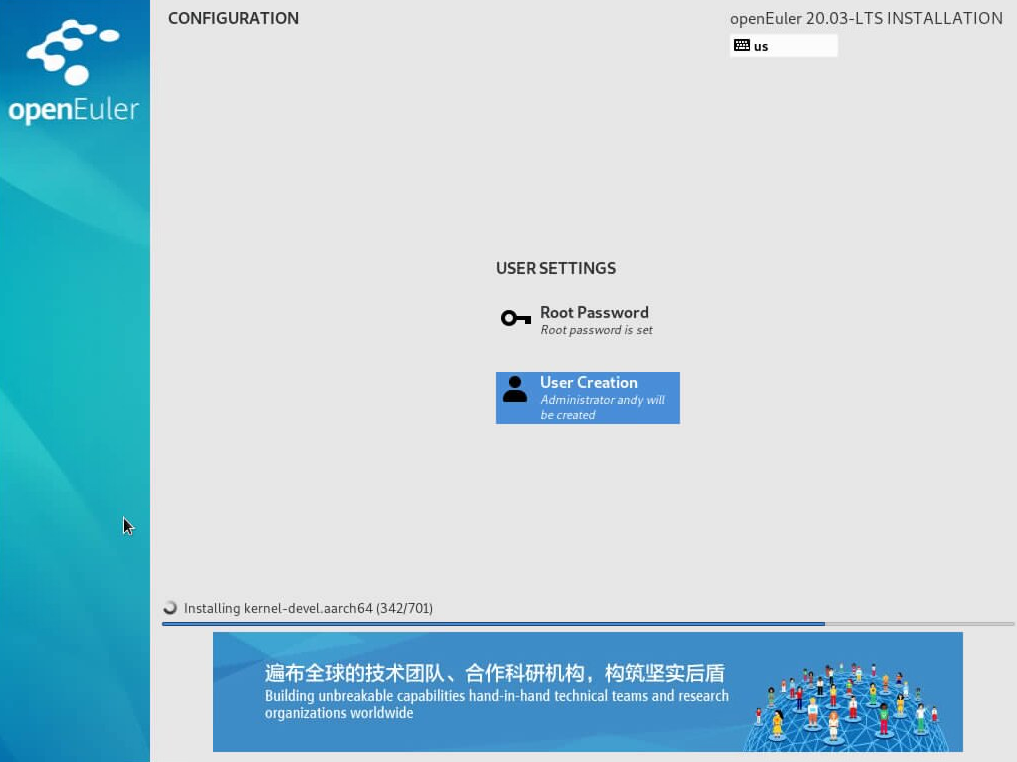
在安装期间请按照提示设置root用户的密码（必需）：



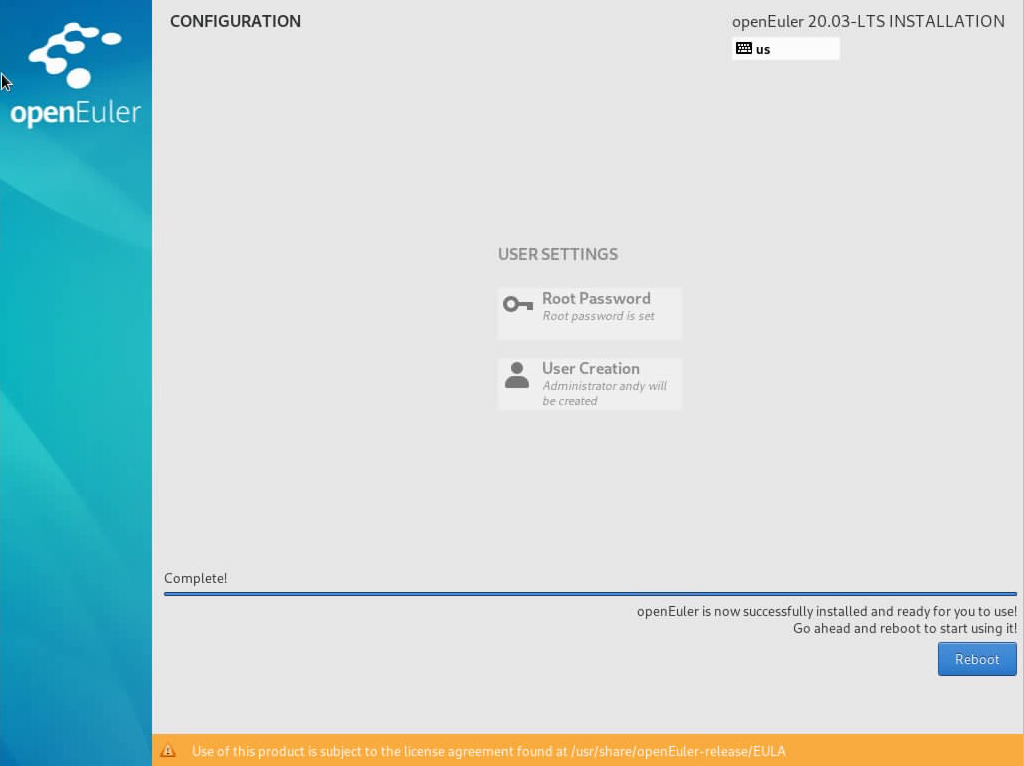
您也可以设置普通用户的密码（非必需）：



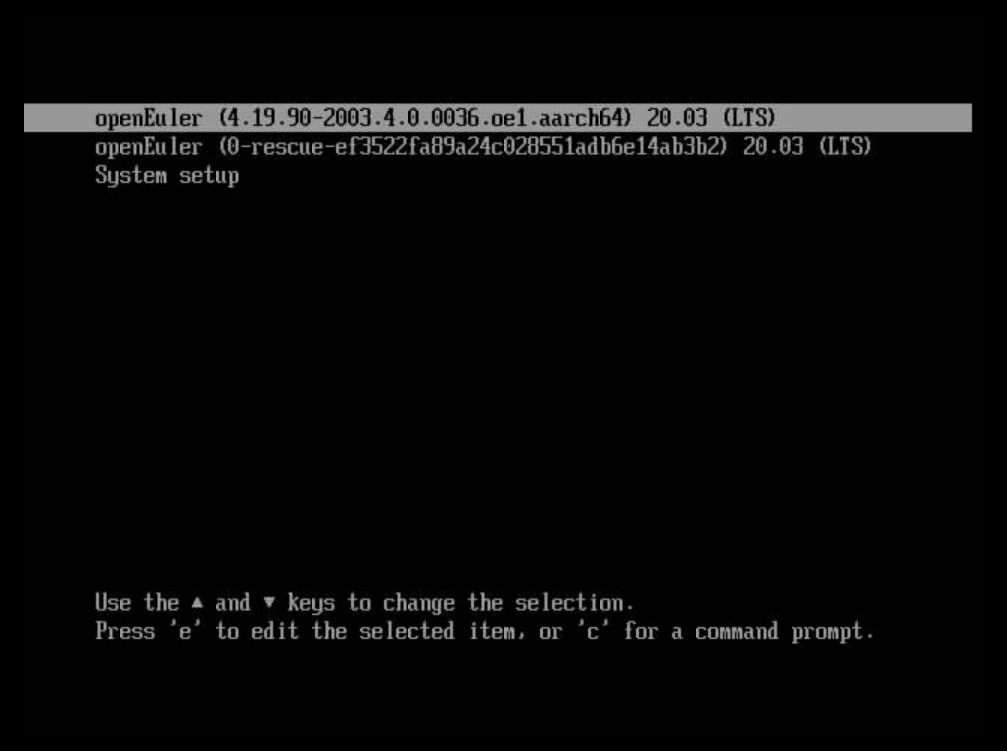
静等安装完成：



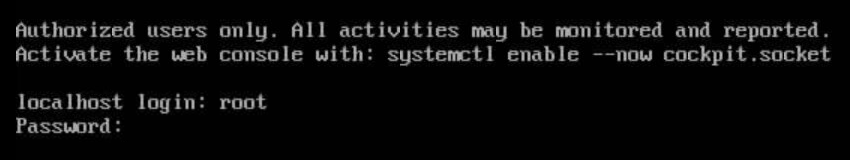
安装完成后点击“Reboot”按钮重启系统：

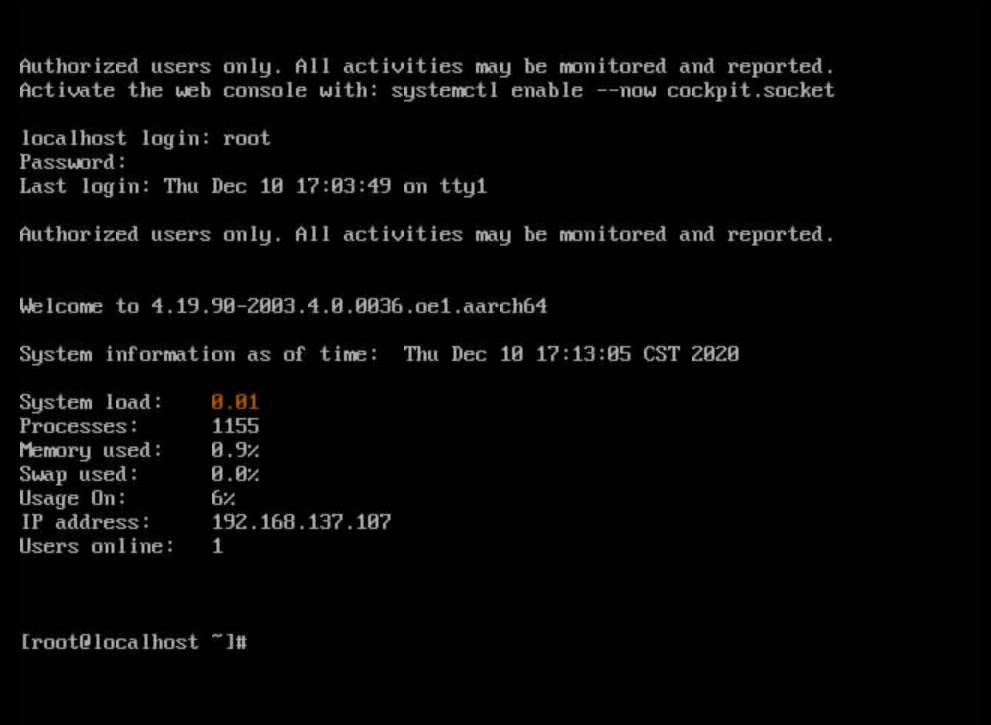


出现如下图所示界面时请按照其默认的启动项“openEuler (4.19.90-2003.4.0.0036.oel.aarch64) 20.03 (LTS)”启动（您的子版本号可能有所不同，上图中的第二个启动项为救援模式）：



在登录界面中以root用户登录（输入密码时不会有任何显示）：





这里是登录成功后的界面。

注意：您的系统显示的信息可能与此不同。

验证网络

本机以太网接口及IP地址如下所示：

[root@localhost ~]# **ifconfig**

enp125s0f0: flags=4163<UP,BROADCAST,RUNNING,MULTICAST> mtu 1500

inet 192.168.137.62 netmask 255.255.255.0 broadcast 192.168.137.255

inet6 fe80::aace:2b25:8cb0:30bd prefixlen 64 scopeid 0x20<link>

ether 1c:20:db:fc:4f:c9 txqueuelen 1000 (Ethernet)

RX packets 128 bytes 30602 (29.8 KiB)

RX errors 0 dropped 0 overruns 0 frame 0

TX packets 28 bytes 6068 (5.9 KiB)

TX errors 0 dropped 0 overruns 0 carrier 0 collisions 0

enp125s0f1: flags=4163<UP,BROADCAST,RUNNING,MULTICAST> mtu 1500

inet 192.168.137.217 netmask 255.255.255.0 broadcast 192.168.137.255

inet6 fe80::a213:9f60:6ccd:7203 prefixlen 64 scopeid 0x20<link>

ether 1c:20:db:fc:4f:ca txqueuelen 1000 (Ethernet)

RX packets 237 bytes 43930 (42.9 KiB)

RX errors 0 dropped 0 overruns 0 frame 0

TX packets 215 bytes 21240 (20.7 KiB)

TX errors 0 dropped 0 overruns 0 carrier 0 collisions 0

enp125s0f2: flags=4163<UP,BROADCAST,RUNNING,MULTICAST> mtu 1500

inet 192.168.101.244 netmask 255.255.255.0 broadcast 192.168.101.255

inet6 fe80::9de:da16:b7a6:d13c prefixlen 64 scopeid 0x20<link>

ether 1c:20:db:fc:4f:cb txqueuelen 1000 (Ethernet)

RX packets 265 bytes 33786 (32.9 KiB)

RX errors 0 dropped 0 overruns 0 frame 0

TX packets 18 bytes 3548 (3.4 KiB)

TX errors 0 dropped 0 overruns 0 carrier 0 collisions 0

enp125s0f3: flags=4163<UP,BROADCAST,RUNNING,MULTICAST> mtu 1500

inet 192.168.101.242 netmask 255.255.255.0 broadcast 192.168.101.255

inet6 fe80::15c3:e57d:200a:74bd prefixlen 64 scopeid 0x20<link>

ether 1c:20:db:fc:4f:cc txqueuelen 1000 (Ethernet)

RX packets 405 bytes 46541 (45.4 KiB)

RX errors 0 dropped 0 overruns 0 frame 0

TX packets 131 bytes 20023 (19.5 KiB)

TX errors 0 dropped 0 overruns 0 carrier 0 collisions 0

enp3s0: flags=4099<UP,BROADCAST,MULTICAST> mtu 1500

ether 18:cf:24:cf:9e:b3 txqueuelen 1000 (Ethernet)

RX packets 0 bytes 0 (0.0 B)

RX errors 0 dropped 0 overruns 0 frame 0

TX packets 0 bytes 0 (0.0 B)

TX errors 0 dropped 0 overruns 0 carrier 0 collisions 0

enp4s0: flags=4099<UP,BROADCAST,MULTICAST> mtu 1500

ether 18:cf:24:cf:9e:b4 txqueuelen 1000 (Ethernet)

RX packets 0 bytes 0 (0.0 B)

RX errors 0 dropped 0 overruns 0 frame 0

TX packets 0 bytes 0 (0.0 B)

TX errors 0 dropped 0 overruns 0 carrier 0 collisions 0

enp5s0: flags=4099<UP,BROADCAST,MULTICAST> mtu 1500

ether 18:cf:24:cf:9e:b5 txqueuelen 1000 (Ethernet)

RX packets 0 bytes 0 (0.0 B)

RX errors 0 dropped 0 overruns 0 frame 0

TX packets 0 bytes 0 (0.0 B)

TX errors 0 dropped 0 overruns 0 carrier 0 collisions 0

enp6s0: flags=4099<UP,BROADCAST,MULTICAST> mtu 1500

ether 18:cf:24:cf:9e:b6 txqueuelen 1000 (Ethernet)

RX packets 0 bytes 0 (0.0 B)

RX errors 0 dropped 0 overruns 0 frame 0

TX packets 0 bytes 0 (0.0 B)

TX errors 0 dropped 0 overruns 0 carrier 0 collisions 0

lo: flags=73<UP,LOOPBACK,RUNNING> mtu 65536

inet 127.0.0.1 netmask 255.0.0.0

inet6 ::1 prefixlen 128 scopeid 0x10<host>

loop txqueuelen 1000 (Local Loopback)

RX packets 0 bytes 0 (0.0 B)

RX errors 0 dropped 0 overruns 0 frame 0

TX packets 0 bytes 0 (0.0 B)

TX errors 0 dropped 0 overruns 0 carrier 0 collisions 0

virbr0: flags=4099<UP,BROADCAST,MULTICAST> mtu 1500

inet 192.168.122.1 netmask 255.255.255.0 broadcast 192.168.122.255

ether 52:54:00:f6:a5:21 txqueuelen 1000 (Ethernet)

RX packets 0 bytes 0 (0.0 B)

RX errors 0 dropped 0 overruns 0 frame 0

TX packets 0 bytes 0 (0.0 B)

TX errors 0 dropped 0 overruns 0 carrier 0 collisions 0

[root@localhost ~]#

本文Taishan服务器的共有4张网卡，两个属于192.168.101.X网段，用于终端登录；两个属于192.168.137.X网段，用于连接公网。它们的IP地址都是由DHCP协议自动分配的。

注意：您的网络信息与此会有不同。

若没有访问公网的默认路由，需要自行添加：

[root@localhost ~]# **ip route add default via *192.168.137.1***

注意：要将上述命令中ip地址替换为您所在网络的可以访问公网的网关地址。

若没有DNS，可按如下内容配置DNS：

[root@localhost ~]# **vi /etc/resolv.conf**

[root@localhost ~]# **cat /etc/resolv.conf**

# Generated by NetworkManager

##search mshome.net

nameserver 8.8.8.8

nameserver 114.114.114.114

验证网络是否可以访问公网：

[root@localhost ~]# **ping www.huawei.com**

PING www.huawei.com.lxdns.com (101.227.102.169) 56(84) bytes of data.

64 bytes from 101.227.102.169 (101.227.102.169): icmp\_seq=1 ttl=52 time=22.2 ms

64 bytes from 101.227.102.169 (101.227.102.169): icmp\_seq=2 ttl=52 time=22.9 ms

64 bytes from 101.227.102.169 (101.227.102.169): icmp\_seq=3 ttl=52 time=21.8 ms

注意：若不能访问，请按自己的实际情况配置网络。

设置yum仓库。请以如下文件的内容替换**/etc/yum.repos.d/openEuler\_aarch64.repo**文件内容：



更新后的repolist如下所示：

[root@localhost ~]# **yum repolist**

repo id repo name

EPOL EPOL

OS OS

debuginfo debuginfo

everything everything

source source

update update

构建开发环境：

[root@localhost ~]# **yum install openssl-devel -y**

### openEuler内核编译与安装

备份boot目录以防后续步骤更新内核失败

[root@localhost ~]# **tar czvf boot\_origin.tgz /boot/**

保存当前内核版本信息

[root@localhost ~]# **uname –r > uname\_r.log**

获取内核源代码并解压

[root@localhost ~]# **wget https://gitee.com/openeuler/kernel/repository/archive/kernel-4.19.zip**

[root@localhost ~]# **unzip kernel-4.19.zip**

编译内核

[root@localhost ~]# **cd kernel**

[root@localhost kernel]# **make openeuler\_defconfig**

在这里，我们按源代码文件kernel/arch/arm64/configs/openeuler\_defconfig的配置配置内核。

[root@localhost kernel]# **make help | grep Image**

\* Image.gz - Compressed kernel image (arch/arm64/boot/Image.gz)

Image - Uncompressed kernel image (arch/arm64/boot/Image)

这一步查看了可编译的Image。

[root@localhost kernel]# **make -j16 Image modules dtbs**

这一步是编译内核的Image、modules和dtbs。

安装内核

[root@localhost kernel]# **make modules\_install**

由于kernel/Makefile把获取的架构字符串“aarch64”直接用于寻找“kernel/arch/arm64”里面的内容导致失败，这里简单一个应对方法是将“kernel/arch/arm64”文件夹复制一份为“kernel/arch/aarch64”即可：

[root@localhost kernel]# **cp -r ./arch/arm64/ ./arch/aarch64 /**

[root@localhost kernel]# **make install**

重启机器

[root@localhost kernel]# **reboot**

登录新内核

在启动菜单上会出现新编译的内核版本，我们选择这个新版本内核启动系统：



查看版本

登录后用uname命令查看内核版本：

[root@localhost ~]# **uname -r**

4.19.162

以上文字表明内核版本已更新（注意：以上显示的内核版本可能和您下载的内核版本不同）。

### Hello, world!

正确编写满足功能的源文件，包括.c源文件和Makefile文件。

[root@localhost ~]# **cd tasks\_k/1/task3**

[root@localhost task1]# **ls**

helloworld.c Makefile

示例源文件请见源码压缩包中tasks\_k/1/task3目录下。

编译源文件

[root@localhost task3]# **make**

make -C /root/kernel M=/root/tasks\_k/1/task3 modules

make[1]: Entering directory '/root/kernel'

CC [M] /root/tasks\_k/1/task3/helloworld.o

Building modules, stage 2.

MODPOST 1 modules

CC /root/tasks\_k/1/task3/helloworld.mod.o

LD [M] /root/tasks\_k/1/task3/helloworld.ko

make[1]: Leaving directory '/root/kernel'

加载编译完成的内核模块，并查看加载结果。

[root@localhost task3]# **insmod helloworld.ko**

[root@localhost task3]# **lsmod | grep helloworld**

helloworld 262144 0

卸载内核模块，并查看结果。

[root@localhost task3]# **rmmod helloworld**

[root@localhost task3]# **dmesg | tail -n5**

[ 808.740327] helloworld: loading out-of-tree module taints kernel.

[ 808.740382] helloworld: module verification failed: signature and/or required key missing - tainting kernel

[ 808.740989] hello\_init

[ 808.740990] Hello, world!

[ 829.767482] hello\_exit

在这里可以忽略前两行错误信息。

# 实验二 内存管理

## 实验介绍

本实验通过在内核态分配内存的任务操作，让学生们了解并掌握操作系统中内存管理的布局，内核态内存分配的实现，以及内核模块的加载、卸载。

### 相关知识

一、kmalloc()和vmalloc()分配的是内核态的内存分配函数。

kmalloc()

功能：在设备驱动程序或者内核模块中动态分配内存。

函数原型：static \_\_always\_inline void \*kmalloc(size\_t size, gfp\_t flags)

头文件：#include <linux/slab.h>

参数说明：

size：要分配内存的大小，以字节为单位。

flags：要分配的内存类型。如：GFP\_USER（代表用户分配内存）、GFP\_KERNEL（分配内核内存）、GFP\_ATOMIC等（更多请参考 linux/gfp.h）

返回值：分配成功时，返回分配的虚拟地址；分配失败时，返回NULL。

特点：

分配的内存在物理上是连续的（这对于要进行DMA的设备十分重要），用于小内存分配。

最多只能分配 32\*PAGESIZE大小的内存。

最小处理 32 字节或者 64 字节的内存块。

分配速度较快，内核中主要的内存分配方法。

使用完之后，用 kfree() 释放内存：void kfree(const void \*);

vmalloc()

功能：在设备驱动程序或者内核模块中动态分配内存。

函数原型：void \*vmalloc(unsigned long size)

头文件：#include <linux/vmalloc.h>

参数说明：

size：要分配内存的大小，以字节为单位。

返回值：分配成功时，返回分配的虚拟地址；分配失败时，返回 NULL。

特点：

分配的内存：虚拟地址连续，物理地址不连续。

最小处理 4KB 的内存块。

分配速度较慢，一般用于大块内存的分配。

使用完之后，用 vfree() 释放内存：void vfree(const void \*addr)

二、内存布局

不同的体系架构，内存布局各不相同，在内核源码的Document目录下，有部分架构关于内核布局的详细描述，如：

arm64: Documentation/arm64/memory.txt

三、内核模块编程

源码编写—— .c源文件

Makefile文件编写

编译模块——make

模块加载进内核——insmod

查看加载的内容——dmesg

查看内核模块——lsmod

卸载内核模块——rmmod

### 任务描述

* 使用kmalloc分配1KB，8KB的内存，打印指针地址；
* 使用vmalloc分配8KB、1MB、64MB的内存，打印指针地址；
* 查看已分配的内存，根据机器是32位或64位的情况，分析地址落在的区域。

## 实验目的

* 掌握正确编写满足功能的源文件，正确编译。
* 掌握正常加载、卸载内核模块；且内核模块功能满足任务所述。
* 了解操作系统的内存管理。

## 实验任务

### 使用kmalloc分配1KB，8KB的内存，打印指针地址

正确编写满足功能的源文件，包括.c源文件和Makefile文件。在这里我们的示例源文件存放在tasks\_k/2/task1目录下。

[root@localhost ~]# **cd tasks\_k/2/task1**

[root@localhost task1]# **ls**

kmalloc.c Makefile

编译源文件

[root@localhost task1]# **make**

make -C /root/kernel M=/root/tasks\_k/2/task1 modules

make[1]: Entering directory '/root/kernel'

CC [M] /root/tasks\_k/2/task1/kmalloc.o

Building modules, stage 2.

MODPOST 1 modules

CC /root/tasks\_k/2/task1/kmalloc.mod.o

LD [M] /root/tasks\_k/2/task1/kmalloc.ko

make[1]: Leaving directory '/root/kernel'

[root@localhost task1]# **ls**

kmalloc.c kmalloc.ko kmalloc.mod.c kmalloc.mod.o kmalloc.o Makefile modules.order Module.symvers

加载编译完成的内核模块，并查看加载结果。

[root@localhost task1]# **insmod kmalloc.ko**

[root@localhost task1]# **dmesg | tail -n3**

[12892.541517] Start kmalloc!

[12892.541688] kmallocmem1 addr = ffffa0bf78165400

[12892.541920] kmallocmem2 addr = ffff80bfcda82000

卸载内核模块，并查看结果。

[root@localhost task1]# **rmmod kmalloc**

[root@localhost task1]# **dmesg | tail -n4**

[12892.541517] Start kmalloc!

[12892.541688] kmallocmem1 addr = ffffa0bf78165400

[12892.541920] kmallocmem2 addr = ffff80bfcda82000

[12994.315305] Exit kmalloc!

[root@localhost task1]#

### 使用vmalloc分配8KB、1MB、64MB的内存，打印指针地址

正确编写满足功能的源文件，包括.c源文件和Makefile文件。在这里我们的示例源文件存放在tasks\_k/2/task1目录下。

[root@localhost ~]# **cd tasks\_k/2/task2**

[root@localhost task2]# **ls**

vmalloc.c Makefile

编译源文件

[root@localhost task2]# **make**

make -C /root/kernel M=/root/tasks\_k/2/task2 modules

make[1]: Entering directory '/root/kernel'

CC [M] /root/tasks\_k/2/task2/vmalloc.o

Building modules, stage 2.

MODPOST 1 modules

CC /root/tasks\_k/2/task2/vmalloc.mod.o

LD [M] /root/tasks\_k/2/task2/vmalloc.ko

make[1]: Leaving directory '/root/kernel'

[root@localhost task2]# **ls**

Makefile modules.order Module.symvers vmalloc.c vmalloc.ko vmalloc.mod.c vmalloc.mod.o vmalloc.o

加载编译完成的内核模块，并查看加载结果。

[root@localhost task2]# **insmod vmalloc.ko**

[root@localhost task2]# **dmesg | tail -n4**

[20608.083498] Start vmalloc!

[20608.083718] vmallocmem1 addr = ffff00003af10000

[20608.083953] vmallocmem2 addr = ffff00017e3a0000

[20608.084370] vmallocmem3 addr = ffff000185840000

卸载内核模块，并查看结果。

[root@localhost task2]# **rmmod vmalloc**

[root@localhost task2]# **dmesg | tail -n5**

[20608.083498] Start vmalloc!

[20608.083718] vmallocmem1 addr = ffff00003af10000

[20608.083953] vmallocmem2 addr = ffff00017e3a0000

[20608.084370] vmallocmem3 addr = ffff000185840000

[20759.537586] Exit vmalloc!

### 实验结果分析

通过uname -a或arch命令查看当前是ARM64位的机器。使用“getconf PAGE\_SIZE”查看系统的页表大小；或者通过查看内核配置选项，内核配置文件是：arch/arm64/configs/openeuler\_defconfig，其中与内存管理相关的配置内容是（可通过vi打开配置文件，关键词搜索定位）：

CONFIG\_ARM64=y

CONFIG\_64BIT=y

CONFIG\_PGTABLE\_LEVELS=3

CONFIG\_ARM64\_PAGE\_SHIFT=16

CONFIG\_ARM64\_64K\_PAGES=y

CONFIG\_ARM64\_VA\_BITS\_48=y

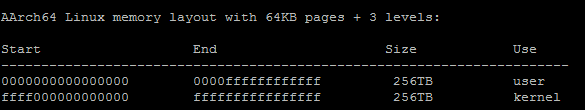
也就是说，目前的配置是：

虚拟地址位数为48位（CONFIG\_ARM64\_VA\_BITS\_48=y、CONFIG\_ARM64\_VA\_BITS=48）；

页表的大小是64K（CONFIG\_ARM64\_PAGE\_SHIFT=16、CONFIG\_ARM64\_64K\_PAGES=y）；

页表转化是3级（CONFIG\_PGTABLE\_LEVELS=3）。

因此查看内核源码的文档Documentation/arm64/memory.txt可见，对应的内存布局如下：



由运行结果可知，kmalloc和vmalloc分配的内存地址，都位于内核空间。

## 思考题

一、什么是内存泄漏、内存溢出、内存越界？

内存泄漏（memory leak）：程序中己动态分配的内存未释放或无法释放，就产生了内存泄露。

内存溢出（out of memory）：程序在申请内存时，没有足够的内存空间供其使用。

内存越界：是指程序向系统申请一块内存后，使用时超出申请范围。

二、分析程序

1、分析：下面这个程序是否会产生内存泄露、内存溢出或内存越界？

#include <stdlib.h>

void function\_which\_allocates(void)

{

float \*a = malloc(sizeof(float) \* 45); // 分配包含45个浮点数的数组

/\* 使用数组a的代码 \*/

}

int main(void)

{

function\_which\_allocates();

/\* 指针a已经不存在了，所以包含45个浮点数的数组就不能被释放了，但这个数组还存在于内存中，因此就造成了内存泄露\*/

}

分析：内存泄漏，己动态分配的内存未释放。

2、分析：下面这个程序是否会产生内存泄露、内存溢出或内存越界？

void func(char \* input)

{

char buffer[16];

strcpy(buffer, input);

}

void main()

{

char longstring[256];

int i;

for( i = 0; i < 255; i++)

longstring [i] = 'B';

func(longstring);

}

分析：内存越界：上述代码中，strcpy()直接将input中的内容copy到buffer中。而main函数中，传入input的长度是256，大于buffer的长度，造成buffer越界，使程序运行出错。

# 实验三 进程管理

## 实验介绍

本实验通过在内核态创建进程，读取系统CPU负载，打印系统当前运行进程PID以及使用cgroup限制CPU核数等任务操作，让学生们了解并掌握操作系统中的进程管理。

### 任务描述

* 编写内核模块，创建一个内核线程；并在模块退出时杀死该线程。
* 编写一个内核模块，实现读取系统一分钟内的CPU负载。
* 编写一个内核模块，打印当前系统处于运行状态的进程的PID和名字。
* 使用cgroup实现限制CPU核数。

使用cgroup实现不允许访问U盘。

## 实验目的

* 掌握正确编写满足功能的源文件，正确编译。
* 掌握正常加载、卸载内核模块；且内核模块功能满足任务所述。
* 了解操作系统的进程管理。
* 了解cgroup的功能。

## 实验任务

### 创建内核进程

#### 相关知识

一、内核线程介绍

内核经常需要在后台执行一些操作，这种任务就可以通过内核线程（kernel thread）完成，内核线程是指独立运行在内核空间的标准进程。内核线程和普通的进程间的区别在于：内核线程没有独立的地址空间，mm指针被设置为NULL；它只在内核空间运行，从来不切换到用户空间去；并且和普通进程一样，可以被调度，也可以被抢占。

内核线程只能由其它的内核线程创建，Linux内核通过给出的函数接口与系统中的初始内核线程kthreadd交互，由kthreadd衍生出其它的内核线程。

二、相关接口函数

1、kthread\_create()：

函数返回一个task\_struct指针，指向代表新建内核线程的task\_struct结构体。注意：使用该函数创建的内核线程处于不可运行状态，需要将kthread\_create返回的task\_struct传递给wake\_up\_process函数，通过此函数唤醒新建的内核线程。

2、kthread\_run()

头文件：<linux/kthread.h>

函数原型：struct task\_struct \*kthread\_run(int (\*threadfn)(void \*data),void \*data,const char \*namefmt, ...);

功能：创建并启动一个线程。

参数：int (\*threadfn)(void \*data)------>线程函数，指定该线程要完成的任务。这个函数会一直运行直到接收到终止信号。

void \*data-------------------------->线程函数的参数。

const char \*namefmt-------------->线程名字。

3、线程函数

用户在线程函数中指定要让该线程完成的任务。该函数会一直运行，直到接收到结束信号。因此函数中需要有判断是否收到信号的语句。

static int func(void \*data){

while(!kthread\_should\_stop()){

/\* 在此处做一些工作 \*/

msleep(2000);

}

return 0;

}

注意在线程函数中需要在每一轮迭代之后休眠一定时间，让出CPU给其他的任务，否则创建的这个线程会一直占用CPU，使得其他任务均瘫痪。更严重的是，使线程终止的命令也无法执行，导致这种状态一直持续下去。

4、kthread\_stop()：

头文件：<linux/kthread.h>

函数原型：int kthread\_stop(struct task\_struct \*k);

功能：在模块卸载时，发送信号给k指向的线程，使之退出。

线程一旦启动起来之后，会一直运行，除非该线程主动调用do\_exit函数，或者其他的进程调用kthread\_stop函数，结束线程的运行。当然，如果线程函数永远不返回，并且不检查信号，它将永远不会停止。因此线程函数信号检查语句以及返回值非常重要。

注意，在调用kthread\_stop函数时，线程不能已经结束运行，否则，kthread\_stop函数会一直等待。

5、kthread\_should\_stop()：

头文件：<linux/kthread.h>

函数原型：bool kthread\_should\_stop(void);

功能：该函数位于内核线程函数体内，与kthread\_stop配合使用，用于接收kthread\_stop传递的结束线程信号，如果内核线程中未用此函数，则kthread\_stop使其结束。

#### 实验步骤

正确编写满足功能的源文件，包括.c源文件和Makefile文件。在这里我们的示例源文件存放在tasks\_k/3/task1目录下。

[root@localhost ~]# **cd tasks\_k/3/task1**

[root@localhost task1]# **ls**

kthread.c Makefile

编译源文件

[root@localhost task1]# **make**

make -C /root/kernel M=/root/tasks\_k/3/task1 modules

make[1]: Entering directory '/root/kernel'

CC [M] /root/tasks\_k/3/task1/kthread.o

Building modules, stage 2.

MODPOST 1 modules

CC /root/tasks\_k/3/task1/kthread.mod.o

LD [M] /root/tasks\_k/3/task1/kthread.ko

make[1]: Leaving directory '/root/kernel'

[root@localhost task1]# **ls**

kthread.c kthread.ko kthread.mod.c kthread.mod.o kthread.o Makefile modules.order Module.symvers

加载编译完成的内核模块，并查看加载结果。

[root@localhost task1]# **insmod kthread.ko**

[root@localhost task1]# **dmesg | tail –n5**

[834853.034010] Create kernel thread!

[834853.034072] New kthread is running.

[834855.060481] New kthread is running.

[834857.076456] New kthread is running.

[834859.092413] New kthread is running.

卸载内核模块，并查看结果。

[root@localhost task1]# **rmmod kthread**

[root@localhost task1]# **dmesg | tail –n5**

[834901.427791] New kthread is running.

[834903.443756] New kthread is running.

[834905.459734] New kthread is running.

[834907.475699] New kthread is running.

[834908.407177] Kill new kthread.

### 打印输出当前系统CPU负载情况

#### 相关知识

一、proc文件系统

1、proc文件简介

proc文件系统是Linux中的特殊文件系统，提供给用户一个可以了解内核内部工作过程的可读窗口，在运行时访问内核内部数据结构、改变内核设置的机制。

proc文件系统能够保存系统当前工作的特殊数据，但并不存在于任何物理设备中，对其进行读写时，才根据系统中的相关信息即时生成。所有proc文件挂载在/proc目录下。

/proc的文件可以用于访问有关内核的状态、计算机的属性、正在运行的进程的状态等信息。大部分 /proc 中的文件和目录提供系统物理环境最新的信息。它们实际上并不存在磁盘上，也不占用任何空间。（用ls –l 可以显示它们的大小）当查看这些文件时，实际上是在访问存在内存中的信息，这些信息用于访问系统。

尽管/proc中的文件是虚拟的，但它们仍可以使用任何文件编辑器或像 'more'、'less' 或 'cat' 这样的程序来查看。当编辑程序试图打开一个虚拟文件时，这个文件就通过内核中的信息被凭空地创建了。

2、proc文件组成

（1）有用内核信息

proc文件系统可以被用于收集有用的关于系统和运行中的内核的信息。下面是一些重要的文件：

/proc/cpuinfo ------------------------------------- CPU的信息（型号，家族，缓存大小等）

/proc/meminfo ----------------------------------- 物理内存、交换空间等的信息

/proc/loadavg ------------------------------------ 查看系统1分钟、5分钟、15分钟的平均负载情况

/proc/mounts ------------------------------------- 已加载的文件系统的列表

/proc/devices ------------------------------------- 可用设备的列表

/proc/filesystems --------------------------------- 被支持的文件系统

/proc/modules ------------------------------------ 已加载的模块

/proc/version ------------------------------------- 内核版本

/proc/cmdline ------------------------------------ 系统启动时输入的内核命令行参数

proc中的文件远不止上面列出的这么多。想要进一步了解，可以对/proc的每一个文件都‘more’一下。

（2）进程相关信息

/proc文件系统可以用于获取运行中的进程的信息。在/proc中有一些编号的子目录。每个编号的目录对应一个进程id(PID)。这样，每一个运行中的进程/proc中都有一个用它的PID命名的目录。这些子目录中包含可以提供有关进程的状态和环境的重要细节信息的文件。

（3）通过proc文件与内核交互

上面讨论的大部分/proc的文件是只读的。而实际上/proc文件系统通过/proc中可读写的文件提供了对内核的交互机制。写这些文件可以改变内核的状态，因而要慎重改动这些文件。

/proc/sys ---------------------------目录存放所有可读写的文件的目录，可以被用于改变内核行为。

/proc/sys/kernel ------------------这个目录包含通用内核行为的信息。/proc/sys/kernel/{domainname, hostname} 存放着机器/网络的域名和主机名。

二、内核中读写文件数据的方法

有时候需要在Linux kernel中读写文件数据，如调试程序的时候，或者内核与用户空间交换数据的时候。在kernel中操作文件没有标准库可用，需要利用kernel的一些函数，这些函数主要有：

filp\_open() 在kernel中打开指定文件。

filp\_close() kernel中文件的读操作。

kernel\_read() kernel中文件的写操作。

kernel\_write() 关闭指定文件。

这些函数在 <linux/fs.h> 头文件中声明。具体读写接口说明如下：

#### 实验步骤

正确编写满足功能的源文件，包括.c源文件和Makefile文件。在这里我们的示例源文件存放在tasks\_k/3/task2目录下。

[root@localhost ~]# **cd tasks\_k/3/task2**

[root@localhost task2]# **ls**

cpu\_loadavg.c Makefile

编译源文件

[root@localhost task2]# **make**

make -C /root/kernel M=/root/tasks\_k/3/task2 modules

make[1]: Entering directory '/root/kernel'

CC [M] /root/tasks\_k/3/task2/cpu\_loadavg.o

Building modules, stage 2.

MODPOST 1 modules

CC /root/tasks\_k/3/task2/cpu\_loadavg.mod.o

LD [M] /root/tasks\_k/3/task2/cpu\_loadavg.ko

make[1]: Leaving directory '/root/kernel'

[root@localhost task2]# **ls**

cpu\_loadavg.c cpu\_loadavg.ko cpu\_loadavg.mod.c cpu\_loadavg.mod.o cpu\_loadavg.o Makefile modules.order Module.symvers

加载编译完成的内核模块，并查看加载结果。

[root@localhost task2]# **insmod cpu\_loadavg.ko**

[root@localhost task2]# **dmesg | tail -n2**

[27644.911012] Start cpu\_loadavg!

[27644.911209] The cpu loadavg in one minute is: 0.01

卸载内核模块，并查看结果。

[root@localhost task2]# **rmmod cpu\_loadavg**

[root@localhost task2]# **dmesg | tail -n3**

[27644.911012] Start cpu\_loadavg!

[27644.911209] The cpu loadavg in one minute is: 0.01

[27686.382949] Exit cpu\_loadavg!

### 打印输出当前处于运行状态的进程的PID和名字

#### 相关知识

当前进程在在/proc文件系统也有保存，只不过需要遍历所有进程文件夹，从stat文件中读取状态，来判定是否为当前运行进程。而内核中可用进程遍历函数来遍历所有进程，且进程描述符task\_struct结构里边有state状态，state为0的进程就是当前进程。

1、进程描述符task\_struct

系统中存放进程的管理和控制信息的数据结构称为进程控制块PCB（Process Control Block），是进程管理和控制的最重要的数据结构。

每一个进程均有一个PCB，在创建进程时，建立PCB，伴随进程运行的全过程，直到进程撤消而撤消。

在Linux中，每一个进程都有一个进程描述符task\_struct，也就是PCB；task\_struct结构体是Linux内核的一种数据结构，它会被装载到RAM里并包含每个进程所需的所有信息。是对进程控制的唯一手段也是最有效的手段。

task\_struct定义在<linux/ sched.h>头文件中。

2、for\_each\_process

for\_each\_process是一个宏，定义在<linux/sched/signal.h>文件中，提供了依次访问整个任务队列的能力。

#### 实验步骤

正确编写满足功能的源文件，包括.c源文件和Makefile文件。在这里我们的示例源文件存放在tasks\_k/3/task3目录下。

[root@localhost ~]# **cd tasks\_k/3/task3**

[root@localhost task3]# **ls**

Makefile process\_info.c

编译源文件

[root@localhost task3]# **make**

make -C /root/kernel M=/root/tasks\_k/3/task3 modules

make[1]: Entering directory '/root/kernel'

CC [M] /root/tasks\_k/3/task3/process\_info.o

Building modules, stage 2.

MODPOST 1 modules

CC /root/tasks\_k/3/task3/process\_info.mod.o

LD [M] /root/tasks\_k/3/task3/process\_info.ko

make[1]: Leaving directory '/root/kernel'

[root@localhost task3]# **ls**

Makefile modules.order Module.symvers process\_info.c process\_info.ko process\_info.mod.c process\_info.mod.o process\_info.o

加载编译完成的内核模块，并查看加载结果。

[root@localhost task3]# **insmod process\_info.ko**

[root@localhost task3]# **dmesg | tail -n4**

[835268.267301] Start process\_info!

[835268.267412] 1)name:systemd-udevd 2)pid:1480 3)state:0

[835268.267440] 1)name:kworker/38:1 2)pid:173023 3)state:0

[835268.267443] 1)name:insmod 2)pid:180488 3)state:0

卸载内核模块，并查看结果。

[root@localhost task3]# **rmmod process\_info**

[root@localhost task3]# **dmesg | tail -n5**

[835268.267301] Start process\_info!

[835268.267412] 1)name:systemd-udevd 2)pid:1480 3)state:0

[835268.267440] 1)name:kworker/38:1 2)pid:173023 3)state:0

[835268.267443] 1)name:insmod 2)pid:180488 3)state:0

[835315.787362] Exit process\_info!

### 使用cgroup实现限制CPU核数

#### 相关知识

一、cgroup介绍

cgroup（Control Groups）是将任意进程进行分组化管理的Linux内核功能，提供将进程进行分组化管理的功能和接口的基础结构。I/O或内存的分配控制等具体的资源管理功能是通过这个功能来实现的，这些具体的资源管理功能称为cgroup子系统或控制器。

cgroup的机制是：它以分组的形式对进程使用系统资源的行为进行管理和控制。也就是说，用户通过cgroup对所有进程进行分组，再对该分组整体进行资源的分配和控制。

cgroup中的每个分组称为进程组，它包含多个进程。最初情况下，系统内的所有进程形成一个进程组（根进程组），根据系统对资源的需求，这个根进程组将被进一步细分为子进程组，子进程组内的进程是根进程组内进程的子集。而这些子进程组很有可能继续被进一步细分，最终，系统内所有的进程组形成一颗具有层次等级（hierarchy）关系的进程组树。

如果某个进程组内的进程创建了子进程，那么该子进程默认与父进程处于同一进程组中。也就是说，cgroup对该进程组的资源控制同样作用于子进程。比如，我们限制进程的CPU使用为20%，我们就可以建一个cpu占用为20%的cgroup，然后将进程添加到这个cgroup中。当然，一个cgroup可以有多个进程。

cgroup提供了一个cgroup虚拟文件系统，作为进行分组管理和各子系统设置的用户接口。要使用cgroup，必须挂载cgroup文件系统。这时通过挂载选项指定使用哪个子系统。

cgroup为每种可以控制的资源定义了一个子系统。典型的子系统介绍如下：

1）cpu子系统：该子系统为每个进程组设置一个使用CPU的权重值，以此来管理进程对CPU的访问，限制进程的CPU使用率。

2）cpuset子系统：对于多核CPU，该子系统可以设置进程组只能在指定的核上运行，并且还可以设置进程组在指定的内存节点上申请内存。如果要使用cpuset控制器，需要同时配置 cpuset.cpus和cpuset.mems两个文件（参数）。cpuset.mems用来设置内存节点的；cpuset.cpus用来限制进程可以使用的CPU核心；这两个参数中CPU核心、内存节点都用id表示，之间用 “,” 分隔，比如0,1,2 ；也可以用 “-” 表示范围，如0-3 ；两者可以结合起来用。如“0-2,6,7”。在添加进程前，cpuset.cpus、cpuset.mems必须同时设置，而且必须是兼容的，否则会出错。

3）cpuacct子系统：该子系统只用于生成当前进程组内的进程对CPU的使用报告。

4）memory子系统：该子系统提供了以页面为单位对内存的访问，比如对进程组设置内存使用上限等，同时可以生成内存资源报告。

5）blkio子系统：该子系统用于限制每个块设备的输入输出（比如物理设备（磁盘，固态硬盘，USB等等）。首先，与CPU子系统类似，该系统通过为每个进程组设置权重来控制块设备对其的I/O时间；其次，该子系统也可以限制进程组的I/O带宽以及IOPS。

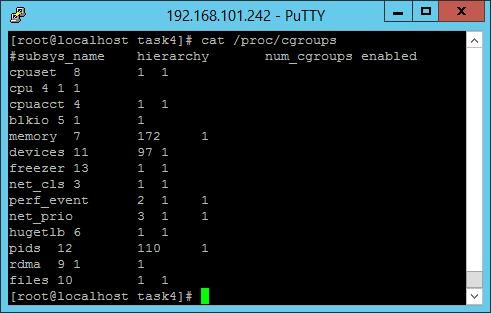
6）devices子系统：通过该子系统可以限制进程组对设备的访问，即该允许或禁止进程组对某设备的访问。

7）freezer子系统：该子系统可以使得进程组中的所有进程挂起或恢复。

8）net-cls子系统：该子系统使用等级识别符标记网络数据包，可允许Linux流量控制程序识别从具体cgroup中生成的数据包，提供对网络带宽的访问限制，比如对发送带宽和接收带宽进程限制。

9）ns子系统：名称空间子系统，可以使不同cgroups下面的进程使用不同的namespace。

针对运行中的内核而言，可以使用的cgroup子系统由/proc/cgroup来确认。



cgroups最初的目标是为资源管理提供的一个统一的框架，既整合现有的cpuset等子系统，也为未来开发新的子系统提供接口。现在的cgroups适用于多种应用场景，从单个进程的资源控制，到实现操作系统层次的虚拟化。

cgroups提供了以下功能：

限制进程组可以使用的资源数量。比如：memory子系统可以为进程组设定一个memory使用上限，一旦进程组使用的内存达到限额再申请内存，就会出发OOM。

进程组的优先级控制。比如：可以使用cpu子系统为某个进程组分配特定cpu share。

记录进程组使用的资源数量。比如：可以使用cpuacct子系统记录某个进程组使用的cpu时间。

进程组隔离。比如：使用ns子系统可以使不同的进程组使用不同的namespace，以达到隔离的目的，不同的进程组有各自的进程、网络、文件系统挂载空间。

进程组控制。比如：使用freezer子系统可以将进程组挂起和恢复。

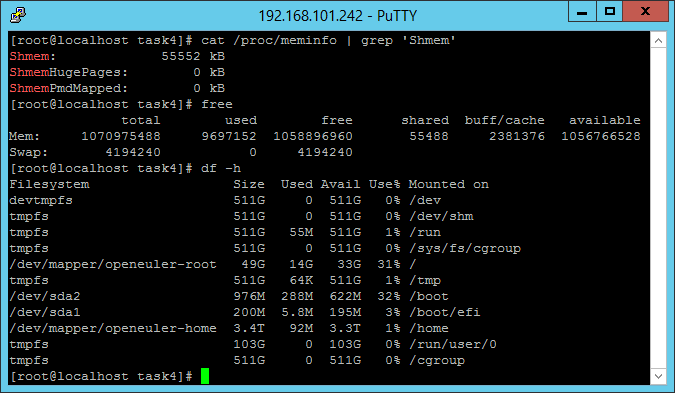
二、tmpfs文件系统

tmpfs即临时文件系统，是一种基于内存的文件系统，也称之为虚拟内存文档系统。它不同于传统的用块设备形式来实现的ramdisk，也不同于针对物理内存的ramfs。tmpfs能够使用物理内存，也能够使用交换分区。

在Linux内核中，虚拟内存资源由物理内存（RAM）和交换分区swap组成，这些资源是由内核中的虚拟内存子系统来负责分配和管理。tmpfs就是和虚拟内存子系统来"打交道"的，他向虚拟内存子系统请求页来存储文档，他同Linux的其他请求页的部分相同，不知道分配给自己的页是在内存中还是在交换分区中。tmpfs同ramfs相同，其大小也不是固定的，而是随着所需要的空间而动态的增减。

所有在tmpfs上储存的资料在理论上都是临时存放的，也就是说，档案不会建立在硬盘上面。一旦重新开机，所有在tmpfs里面的资料都会消失不见。理论上，内存使用量会随着tmpfs的使用而时有增长或消减。tmpfs将所有内容放入内核内部高速缓存中，并进行扩展和收缩以容纳其中包含的文件，并且能够将不需要的页面交换出来以交换空间。它具有最大大小限制，可以通过“mount -o remount”即时调整。

由于tmpfs完全存在于页面缓存和交换中，因此所有tmpfs页面将在/proc/meminfo中显示为“Shmem”，在free命令后中显示为“Shared”。请注意，这些计数器还包括共享内存（shmem）。获取计数的最可靠方法是使用df和du。



使用tmpfs，首先您编译内核时得选择“虚拟内存文档系统支持（Virtual memory filesystem support）”或设置CONFIG\_TMPFS=y，然后就能够加载tmpfs文档系统了。

详细的介绍，可参见内核源码中的官方文档：Documentation/filesystems/tmpfs.txt。

可挂载tmpfs格式的cgroup文件夹进行cgroup的相关操作。

三、相关命令

mount命令：加载指定的文件系统。

echo命令：显示文字。

taskset命令：依据线程PID（TID）查询或设置线程的CPU亲和性（即与哪个CPU核心绑定）。

cgexec 命令：在指定的cgroup中运行任务。

#### 实验步骤

安装libcgroup：dnf install libcgroup -y

libcgroup包含cgroup用户空间工具套件（如lscgroup，lssubsys等）以及静态或者动态库，以供其他程序调用，并且包含debug套件。

挂载tmpfs格式的cgroup文件夹

在root权限下执行以下命令：

mkdir /cgroup

mount -t tmpfs tmpfs /cgroup

cd /cgroup

挂载tmpfs文件类型：tmpfs是直接建立在VM之上的，用一个简单的mount命令就可以创建tmpfs文件系统了。速度快，可以动态分配文件系统大小。

挂载cpuset管理子系统

挂载某一个cgroups子系统到挂载点之后，就可以通过在挂载点下面建立文件夹或者使用cgcreate命令的方法创建cgroups层级结构中的节点/控制组；对应的删除则使用rmdir删除文件夹，或使用cgdelete命令删除。

mkdir cpuset

mount -t cgroup -o cpuset cpuset /cgroup/cpuset #挂载cpuset子系统

cd cpuset

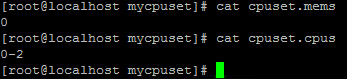
mkdir mycpuset #创建一个控制组，删除用 rmdir 命令

cd mycpuset

设置cpu核数

echo 0 > cpuset.mems #设置0号内存结点。mems默认为空，因此需要填入值。

echo 0-2 > cpuset.cpus #这里的0-2指的是使用cpu的0、1、2三个核。实现了只是用这三个核。



简单的死循环C源文件while\_long.c

#include <stdio.h>

#include <stdlib.h>

int main(int argc, char \*argv[])

{

while (1){}

printf("Over");

exit(0);

}

测试验证

1. 打开一个终端，执行以下命令：

gcc while\_long.c -o while\_long # 编译上述C源文件

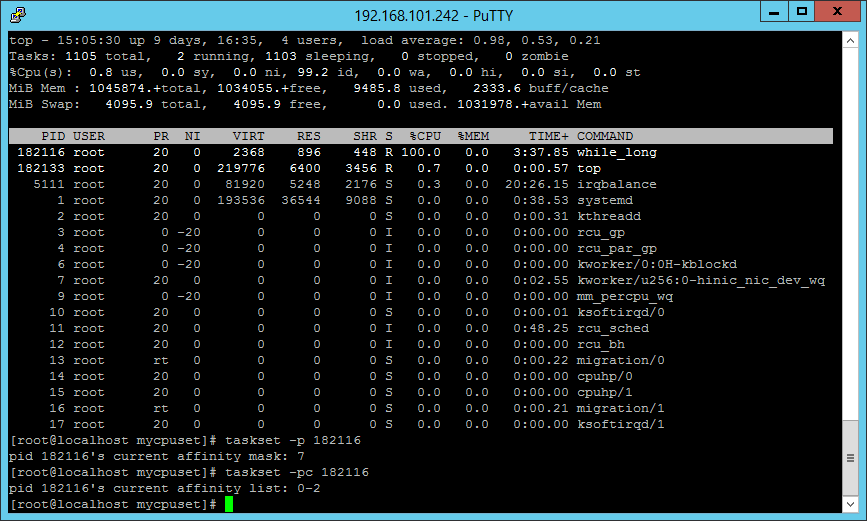
cgexec -g cpuset:mycpuset ./while\_long # 指定在cpuset子系统的mycpuset控制组中运行

1. 不要关闭上述终端，另打开一个终端，执行以下命令：

top #查看程序while\_long的PID, 假设为182116。输入q 退出当前查看状态

taskset -p 182116 #显示的如果是7（111），则测试限制cpu核数成功。

执行如下：



### 使用cgroup实现不允许访问U盘

#### 相关知识

有关cgroup的知识请参照3.3.4.1的内容。与本节有关的知识介绍如下：

1. devices子系统

使用devices 子系统可以允许或者拒绝cgroup中的进程访问设备。

devices子系统有三个控制文件：devices.allow，devices.deny，devices.list。

* devices.allow用于指定cgroup中的进程可以访问的设备；
* devices.deny用于指定cgroup中的进程不能访问的设备；
* devices.list用于报告cgroup中的进程访问的设备。

devices.allow文件中包含若干条目，每个条目有四个字段：type、major、minor 和 access。type、major 和 minor 字段中使用的值对应 Linux 分配的设备。

（1）type指定设备类型：

a -------- 应用所有设备，可以是字符设备，也可以是块设备

b -------- 指定块设备

c -------- 指定字符设备

（2）major和minor指定设备的主次设备号。如下述打印结果中8为主设备号，17为次设备号。

[root@localhost ~]# ls -l /dev/sdb1

brw-rw----. 1 root disk 8, 17 Dec 11 13:15 /dev/sdb1

（3）access 则指定相应的权限：

r -------- 允许任务从指定设备中读取

w ------- 允许任务写入指定设备

m ------- 允许任务生成还不存在的设备文件

写入示例：

# echo 'a 8:17 rwm' > /cgroup/devices/mydevices/devices.deny

2. 请了解以下Shell命令

* fdisk命令
* dd命令
* ls命令

#### 实验步骤

本实验在3.3.4实验基础之上进行，请参照3.3.4.2的步骤一和步骤二，确认已经安装了libcgroup组件和挂载了/cgroup文件夹。

获取U盘插入后的磁盘盘符

在Taishan物理服务器上插入U盘之前运行如下命令：

[root@localhost ~]# fdisk -l > fdisk-list0.log

插入U盘后：

[root@localhost ~]# fdisk -l > fdisk-list1.log

[root@localhost ~]# diff fdisk-list0.log fdisk-list1.log

30a31,42

>

>

> Disk /dev/sdb: 58.43 GiB, 62724374528 bytes, 122508544 sectors

> Disk model: Cruzer Blade

> Units: sectors of 1 \* 512 = 512 bytes

> Sector size (logical/physical): 512 bytes / 512 bytes

> I/O size (minimum/optimal): 512 bytes / 512 bytes

> Disklabel type: dos

> Disk identifier: 0xd505c9fd

>

> Device Boot Start End Sectors Size Id Type

> **/dev/sdb1** 2048 122508543 122506496 58.4G 7 HPFS/NTFS/exFAT

[root@localhost ~]#

可见U盘的盘符为/dev/sdb1。

获取U盘的设备号

使用shell命令 “ls -l”获取设备号：

[root@localhost ~]# ls -l /dev/sdb1

brw-rw----. 1 root disk **8, 17** Dec 11 13:15 /dev/sdb1

可见其设备号为8, 17.

将U盘挂载到当前系统中

[root@localhost ~]# mkdir /usb

[root@localhost ~]# mount /dev/sdb1 /usb

mount: /usb: unknown filesystem type 'ntfs'.

如果U盘是NTFS文件系统则会出现以上提示，在这里可以忽略以上提示信息。

挂载设备管理devices子系统

[root@localhost ~]# cd /cgroup/

[root@localhost cgroup]# mkdir devices

[root@localhost cgroup]# mount -t cgroup -o devices devices /cgroup/devices

[root@localhost cgroup]# cd /cgroup/devices

[root@localhost devices]# mkdir mydevices

[root@localhost devices]# cd mydevices

上述命令挂载了devices子系统并创建了mydevices控制组。

正常访问U盘

[root@localhost mydevices]# cgexec -g devices:mydevices dd if=/dev/sdb1 of=/home/temfile bs=1 count=12800

12800+0 records in

12800+0 records out

12800 bytes (13 kB, 12 KiB) copied, 0.0361779 s, 354 kB/s

其中/dev/sdb1就是前文读到的U盘的设备编号，以上打印信息表明，U盘是可以正常访问的。

设置拒绝U盘访问并验证

[root@localhost mydevices]# echo 'a 8:17 rwm' > /cgroup/devices/mydevices/devices.deny

[root@localhost mydevices]# cgexec -g devices:mydevices dd if=/dev/sdb1 of=/home/temfile bs=1 count=12800

dd: failed to open '/dev/sdb1': Operation not permitted

命令行中的8:17就是我们上文中得到的设备号，将相关设置写入devices.deny文件实现拒绝U盘访问。如果终端显示Operation not permitted。就说明不允许访问U盘设置成功。

# 实验四 中断和异常管理

## 实验介绍

本实验通过在使用软中断延迟机制tasklet实现打印helloworld，用工作队列实现周期打印helloworld以及在用户态下捕获终端按键信号等任务操作，让学生们了解并掌握操作系统中的中断和异常管理机制。

### 相关知识

1、中断

在计算机科学中，中断（Interrupt）是指处理器接收到来自硬件或软件的信号，提示发生了某个事件，应该被注意，这种情况就称为中断。中断通常分为同步中断（synchronous）和异步中断（asynchronous）。

同步中断：是当指令执行时，由CPU控制单元产生的，只有在一条指令终止执行后CPU才会发生中断。

异步中断：是由其他硬件设备依照CPU时钟信号随机产生的。

Intel微处理器手册中，把同步和异步中断分别称为异常（exception）和中断（interrupt）。

中断是由间隔定时器和I/O设备产生的，而异常是由程序的错误产生的，或是由内核必须处理的异常条件产生的。

2、中断的作用

每个中断或者异常都对应着它的中断或者异常处理程序，中断或异常处理程序不是一个进程，而是一个内核控制路径，代表中断发生时正在运行的进程执行。中断处理是内核执行的最敏感的任务之一，必须满足以下约束：

1) 内核相应中断后的操作分两部分：关键的紧急的部分，内核立即执行；其余的推迟随后执行。

2) 中断程序必须使内核控制路径能以嵌套的方式执行，当最后一个内核控制路径终止时，内核必须能恢复被中断进程的执行，或者如果中断信号已经导致了重新调度，内核能切换到另外的进程。

3) 尽管中断可以嵌套，但在临界区中，中断必须禁止。但内核必须尽可能限制这样的临界区，大部分时间应该以开中断的方式运行。

3、中断请求（Interrupt Request，IRQ）

IRQ（Interrupt Request）的作用就是在我们所用的电脑中，执行硬件中断请求的动作，用来停止其相关硬件的工作状态。比如我们要打印一份文件，在打印结束时就需要由系统对打印机提出相应的中断请求，来以此结束这个打印的操作。IRQ具有以下特点：

1) 硬件设备控制器通过IRQ线向CPU发出中断，可以通过禁用某条IRQ线来屏蔽中断。

2) 被禁止的中断不会丢失，激活IRQ后，中断还会被发到CPU

3) 激活/禁止IRQ线 ！= 可屏蔽中断的全局屏蔽/非屏蔽

4、中断的上半部和下半部

中断服务程序一般都是在中断请求关闭的条件下执行的，以避免嵌套而使中断控制复杂化。但是，中断是一个随机事件，它随时会到来，如果关中断的时间太长，CPU就不能及时响应其他的中断请求，从而造成中断的丢失。因此，Linux内核的目标就是尽可能快的处理完中断请求，尽其所能把更多的处理向后推迟。例如，假设一个数据块已经达到了网线，当中断控制器接受到这个中断请求信号时，Linux内核只是简单地标志数据到来了，然后让处理器恢复到它以前运行的状态，其余的处理稍后再进行（如把数据移入一个缓冲区，接受数据的进程就可以在缓冲区找到数据）。因此，内核把中断处理分为两部分：上半部（tophalf）和下半部（bottomhalf），上半部（就是中断服务程序）内核立即执行，而下半部（就是一些内核函数）留着稍后处理，

首先，一个快速的“上半部”来处理硬件发出的请求，它必须在一个新的中断产生之前终止。通常，除了在设备和一些内存缓冲区（如果你的设备用到了DMA，就不止这些）之间移动或传送数据，确定硬件是否处于健全的状态之外，这一部分做的工作很少。

下半部运行时是允许中断请求的，而上半部运行时是关中断的，这是二者之间的主要区别。但是，内核到底什时候执行下半部，以何种方式组织下半部？这就是我们要讨论的下半部实现机制，这种机制在内核的演变过程中不断得到改进，在以前的内核中，这个机制叫做bottomhalf（简称bh），在2.4以后的版本中有了新的发展和改进，改进的目标使下半部可以在多处理机上并行执行，并有助于驱动程序的开发者进行驱动程序的开发。

### 任务描述

* 编写内核模块，使用tasklet实现打印helloworld。
* 编写一个内核模块程序，用工作队列实现周期打印helloworld。
* 在用户态编写一个信号捕获程序，捕获终端按键信号（包括ctrl+c、ctrl+z、ctrl+\）。

## 实验目的

* 正确编写满足功能的源文件，正确编译。
* 正常加载、卸载内核模块；且内核模块功能满足任务所述。
* 了解操作系统的中断与异常管理。

## 实验任务

### 使用tasklet实现打印helloworld

#### 相关知识

tasklet

tasklet是Linux中断处理机制中的软中断延迟机制。引入tasklet，最主要的是考虑支持SMP（多处理，Symmetrical Multi-Processing），提高SMP多个CPU的利用率；不同的tasklet可以在不同的cpu上运行。tasklet可以理解为softirq（软中断）的派生，所以它的调度时机和软中断一样。对于内核中需要延迟执行的多数任务都可以用tasklet来完成，由于同类tasklet本身已经进行了同步保护，所以使用tasklet比软中断要简单的多，tasklet不需要考虑SMP下的并行问题，而又比workqueues有着更好的性能。tasklet通常作为中断下半部来使用，它在性能和易用性之间有着很好的平衡。

1、定义

tasklet由tasklet\_struct结构体来表示，定义在头文件<linux/interrupt.h>中。在使用tasklet前，必须首先创建一个tasklet\_struct类型的变量。tasklet的结构体中包含处理函数的函数指针func，它指向的是这样的一个函数：

void tasklet\_handler(unsigned long data);

如同上半部分的中断处理程序一样，这个函数需要我们自己来实现。

2、tasklet常用接口

创建好之后，我们还要通过如下的方法对tasklet进行初始化与调度：

void tasklet\_init(struct tasklet\_struct \*t, void (\*func)(unsigned long), unsigned long data);   
/\* 初始化tasklet，func指向要执行的函数，data为传递给函数func的参数 \*/

tasklet\_schedule(&my\_tasklet) /\*调度执行指定的tasklet\*/

void tasklet\_kill(struct tasklet\_struct \*t) /\*移除指定tasklet\*/

void tasklet\_disable(struct tasklet\_struct \*t) /\*禁用指定tasklet\*/

void tasklet\_enable(struct tasklet\_struct \*t) /\*启用先前被禁用的tasklet\*/

#### 实验步骤

正确编写满足功能的源文件，包括.c源文件和Makefile文件。在这里我们的示例源文件存放在tasks\_k/4/task1目录下。

[root@localhost ~]# **cd tasks\_k/4/task1**

[root@localhost task1]# **ls**

Makefile tasklet\_intertupt.c

编译源文件

[root@localhost task1]# **make**

make -C /root/kernel M=/root/tasks\_k/4/task1 modules

make[1]: Entering directory '/root/kernel'

CC [M] /root/tasks\_k/4/task1/tasklet\_intertupt.o

Building modules, stage 2.

MODPOST 1 modules

CC /root/tasks\_k/4/task1/tasklet\_intertupt.mod.o

LD [M] /root/tasks\_k/4/task1/tasklet\_intertupt.ko

make[1]: Leaving directory '/root/kernel'

[root@localhost task1]# **ls**

Makefile Module.symvers tasklet\_intertupt.ko tasklet\_intertupt.mod.o

modules.order tasklet\_intertupt.c tasklet\_intertupt.mod.c tasklet\_intertupt.o

加载编译完成的内核模块，并查看加载结果。

[root@localhost task1]# **insmod tasklet\_intertupt.ko**

[root@localhost task1]# **dmesg | tail -n2**

[837684.470110] Start tasklet module...

[837684.470195] Hello World! tasklet is working...

卸载内核模块，并查看结果。

[root@localhost task1]# **rmmod tasklet\_intertupt**

[root@localhost task1]# **dmesg | tail -n3**

[837684.470110] Start tasklet module...

[837684.470195] Hello World! tasklet is working...

[837699.706366] Exit tasklet module...

### 用工作队列实现周期打印helloworld

#### 相关知识

一、工作队列

我们把推后执行的任务叫做工作（work），描述它的数据结构为work\_struct；这些工作以队列结构组织成工作队列（workqueue），其数据结构为workqueue\_struct，而工作线程就是负责执行工作队列中的工作。系统默认的工作者线程为events，自己也可以创建自己的工作者线程。

工作队列是实现延迟的新机制，从2.5版本Linux内核开始提供该功能。工作队列可以把工作延迟，交由一个内核线程去执行，也就是说，这个下半部分可以在进程上下文中执行。这样，通过工作队列执行的代码能占尽进程上下文的所有优势，且工作队列实现了内核线程的封装，不易出错。最重要的就是工作队列允许被重新调度甚至是睡眠。

那么，什么情况下使用工作队列，什么情况下使用tasklet：如果推后执行的任务需要睡眠，那么就选择工作队列；如果推后执行的任务不需要睡眠，那么就选择tasklet。如果需要用一个可以重新调度的实体来执行你的下半部处理，也应该使用工作队列。它是唯一能在进程上下文运行的下半部实现的机制，也只有它才可以睡眠。如果推后执行的任务需要延时指定的时间再触发，那么使用工作队列，因其可以利用timer延时；如果推后执行的任务需要在一个tick之内处理，则只有软中断或tasklet，因其可以抢占普通进程和内核线程；如果推后执行的任务对延迟的时间没有任何要求，则使用工作队列，此时通常为无关紧要的任务。

工作队列允许内核代码来请求在将来某个时间调用一个函数；用来处理不是很紧急事件的回调方式处理方法。

二、工作队列的数据结构与编程接口API

1、表示工作的数据结构（定义在内核源码：include/linux/workqueue.h）

（1）正常的工作用 <linux/workqueue.h> 中定义的work\_struct结构表示。这些结构被连接成链表。当一个工作者线程被唤醒时，它会执行它的链表上的所有工作。工作被执行完毕，它就将相应的work\_struct对象从链表上移去。当链表上不再有对象的时候，它就会继续休眠。

（2）延迟的工作用delayed\_work数据结构，可直接使用delay\_work将任务推迟执行。

2、工作队列中待执行的函数（定义在内核源码：include/linux/workqueue.h）

work\_struct结构中包含工作队列待执行的函数定义 work\_func\_t func；该工作队列待执行的函数原型是：typedef void (\*work\_func\_t)(struct work\_struct \*work)

这个函数会由一个工作者线程执行，因此，函数会运行在进程上下文中。默认情况下，允许响应中断，并且不持有任何锁。如果需要，函数可以睡眠。需要注意的是，尽管该函数运行在进程上下文中，但它不能访问用户空间，因为内核线程在用户空间没有相关的内存映射。通常在系统调用发生时，内核会代表用户空间的进程运行，此时它才能访问用户空间，也只有在此时它才会映射用户空间的内存。

三、工作队列的使用

1、工作队列的创建

要使用工作队列，需要先创建工作项，有以下两种方式：

（1）静态创建

DECLARE\_WORK(n, f); #定义正常执行的工作项

DECLARE\_DELAYED\_WORK(n, f); #定义延后执行的工作项

其中，n表示工作项的名字，f表示工作项执行的函数。

这样就会静态地创建一个名为n，待执行函数为f的work\_struct结构。

（2）动态创建、运行时创建：

通常在内核模块函数中执行以下函数：

INIT\_WORK(\_work, \_func); #初始化正常执行的工作项

INIT\_DELAYED\_WORK(\_work, \_func); #初始化延后执行的工作项

其中，\_work表示work\_struct的任务对象；\_func表示工作项执行的函数。

这会动态地初始化一个由work指向的工作。

2、工作项与工作队列的调度运行

（1）工作项的调度运行

工作成功创建后，我们可以调度它了。想要把给定工作的待处理函数提交给缺省的events工作线程，只需调用schedule\_work(&work)；work马上就会被调度，一旦其所在的处理器上的工作者线程被唤醒，它就会被执行。有时候并不希望工作马上就被执行，而是希望它经过一段延迟以后再执行。在这种情况下，可以调度它在指定的时间执行：

schedule\_delayed\_work(&work,delay);

这时，&work指向的work\_struct直到delay指定的时钟节拍用完以后才会执行。

（2）工作队列的调度运行

对于工作队列的调度，则使用以下两个函数：

bool queue\_work(struct workqueue\_struct \*wq, struct work\_struct \*work)

#调度执行一个指定workqueue中的任务。

bool queue\_delayed\_work(struct workqueue\_struct \*wq, struct delayed\_work \*dwork, unsigned long delay)

#延迟调度执行一个指定workqueue中的任务，功能与queue\_work类似，输入参数多了一个delay。  
3、工作队列的释放

void flush\_workqueue(struct workqueue\_struct \*wq); #刷新工作队列，等待指定列队中的任务全部执行完毕。

void destroy\_workqueue(struct workqueue\_struct \*wq); #释放工作队列所占的资源

#### 实验步骤

正确编写满足功能的源文件，包括.c源文件和Makefile文件。在这里我们的示例源文件存放在tasks\_k/4/task2目录下。

[root@localhost ~]# **cd tasks\_k/4/task2**

[root@localhost task2]# **ls**

Makefile workqueue\_test.c

编译源文件。

[root@localhost task2]# **make**

make -C /root/kernel M=/root/tasks\_k/4/task2 modules

make[1]: Entering directory '/root/kernel'

CC [M] /root/tasks\_k/4/task2/workqueue\_test.o

Building modules, stage 2.

MODPOST 1 modules

CC /root/tasks\_k/4/task2/workqueue\_test.mod.o

LD [M] /root/tasks\_k/4/task2/workqueue\_test.ko

make[1]: Leaving directory '/root/kernel'

[root@localhost task2]# **ls**

Makefile Module.symvers workqueue\_test.ko workqueue\_test.mod.o

modules.order workqueue\_test.c workqueue\_test.mod.c workqueue\_test.o

加载编译完成的内核模块，并查看加载结果。

[root@localhost task2]# **insmod workqueue\_test.ko**

[root@localhost task2]# **dmesg | tail -n5**

[837850.772405] Start workqueue\_test module.

[837855.783670] Hello World!

[837871.143413] Hello World!

[837886.247192] Hello World!

[837901.351008] Hello World!

卸载内核模块，并查看结果。

[root@localhost task2]# **rmmod workqueue\_test**

[root@localhost task2]# **dmesg | tail -n6**

[837850.772405] Start workqueue\_test module.

[837855.783670] Hello World!

[837871.143413] Hello World!

[837886.247192] Hello World!

[837901.351008] Hello World!

[837940.645987] Exit workqueue\_test module.

由代码逻辑知：工作队列延时 5\*HZ（5秒）开始执行，模块加载后5秒才打印HelloWorld!；而后每次执行工作队列中间休眠15秒。

### 编写一个信号捕获程序，捕获终端按键信号

#### 相关知识

一、Linux信号处理机制

1、基本概念

Linux提供的信号机制是一种进程间异步的通信机制，每个进程在运行时，都要通过信号机制来检查是否有信号到达，若有，便中断正在执行的程序，转向与该信号相对应的处理程序，以完成对该事件的处理；处理结束后再返回到原来的断点继续执行。实质上，信号机制是对中断机制的一种模拟，在实现上是一种软中断。

2、信号的产生

信号的生成来自内核，让内核生成信号的请求来自3个地方：

1）用户：用户能够通过终端按键产生信号，例如；

ctrl+c ----> 2) SIGINT（终止、中断）

ctrl+\ ----> 3)SIGQUIT（退出）

ctrl+z ----> 20)SIGTSTP（暂时、停止）

或者是终端驱动程序分配给信号控制字符的其他任何键来请求内核产生信号

2）内核：当进程执行出错时，内核会给进程发送一个信号，例如非法段存取(内存访问违规)、浮点数溢出等；

3）进程：一个进程可以通过系统调用kill给另一个进程发送信号，一个进程可以通过信号和另外一个进程进行通信。

当信号发送到某个进程中时，操作系统会中断该进程的正常流程，并进入相应的信号处理函数执行操作，完成后再回到中断的地方继续执行。需要说明的是，信号只是用于通知进程发生了某个事件，除了信号本身的信息之外，并不具备传递用户数据的功能。

3、信号的响应动作/处理

每个信号都有自己的响应动作，当接收到信号时，进程会根据信号的响应动作执行相应的操作，信号的响应动作有以下几种：

1) 中止进程Term

2) 忽略信号Ign

3) 中止进程并保存内存信息Core

4) 停止进程Stop

5) 继续运行进程Cont

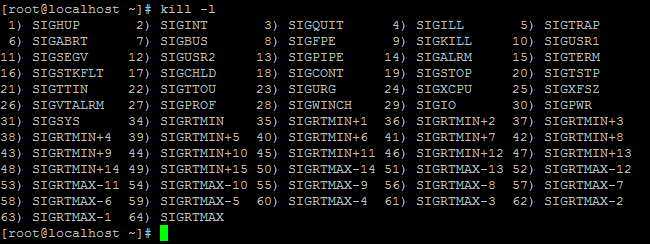
用户可以通过signal或sigaction函数修改信号的响应动作（也就是常说的“注册信号”）。另外，在多线程中，各线程的信号响应动作都是相同的，不能对某个线程设置独立的响应动作。

4、信号类型

Linux支持的信号类型可以参考下面给出的列表。

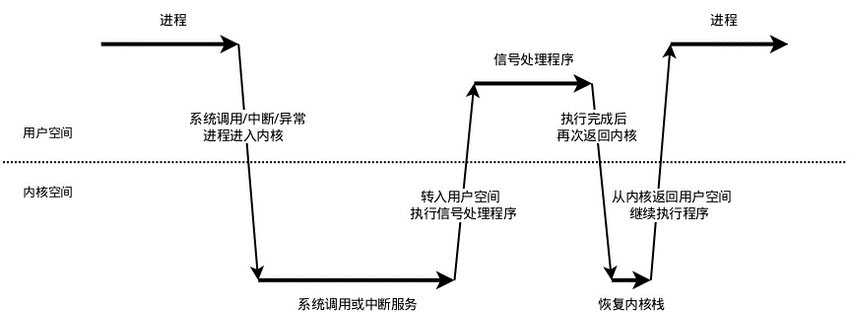


也可使用kill -l查看当前系统的信号编号列表，其中1~31为常规信号，34~64为实时信号。



5、信号机制

前面提到过，信号是异步的，这就涉及信号何时接收、何时处理的问题。我们知道，函数运行在用户态，当遇到系统调用、中断或是异常的情况时，程序会进入内核态。信号涉及到了这两种状态之间的转换，过程可以先看一下下面的示意图：



（1）信号的接收

接收信号的任务是由内核代理的，当内核接收到信号后，会将其放到对应进程的信号队列中，同时向进程发送一个中断，使其陷入内核态。注意，此时信号还只是在队列中，对进程来说暂时是不知道有信号到来的。

（2）信号的检测

进程陷入内核态后，有两种场景会对信号进行检测：

1）进程从内核态返回到用户态前进行信号检测；

2）进程在内核态中，从睡眠状态被唤醒的时候进行信号检测；

当发现有新信号时，便会进入下一步，信号的处理。

（3）信号的处理

信号处理函数是运行在用户态的，调用处理函数前，内核会将当前内核栈的内容备份拷贝到用户栈上，并且修改指令寄存器（eip）将其指向信号处理函数。

接下来进程返回到用户态中，执行相应的信号处理函数。

信号处理函数执行完成后，还需要返回内核态，检查是否还有其它信号未处理。如果所有信号都处理完成，就会将内核栈恢复（从用户栈的备份拷贝回来），同时恢复指令寄存器（eip）将其指向中断前的运行位置，最后回到用户态继续执行进程。

至此，一个完整的信号处理流程便结束了，如果同时有多个信号到达，上面的处理流程会在第2步和第3步骤间重复进行。

二、信号处理函数signal()

函数原型：void (\*signal (int signum ,void (\*handler)(int))) (int) ;

功 能：设置捕捉某一信号后，对应的处理函数。

头文件：#include <signal.h>

参数说明：

signum：指定的信号的编号（或捕捉的信号），可以使用头文件中规定的宏；

handle：函数指针，是信号到来时需要运行的处理函数，参数是signal()的第一个参数signum。

对于第二个参数，可以设置为SIG\_IG，表示忽略第一个参数的信；可以设置为SIG\_DFL，表示采用默认的方式处理信号；也可以指定一个函数地址，自己实现处理方式。

返回值：运行成功，返回原信号处理函数的指针；失败则返回SIG\_ERR。

三、信号与中断的异同点

1、信号与中断的相似点

（1）采用了相同的异步通信方式；

（2）当检测出有信号或中断请求是，都暂停正在执行的程序，转而去执行相应的处理程序；

（3）都在处理完毕后返回到原来的断点；

（4）对信号或中断都可进行评比。

2、信号与中断的区别

（1）中断有优先级，而信号没有优先级，所有信号都是平等的；

（2）信号处理程序是在用户态下运行的；而中断处理程序是在内核态下运行的；

（3）中断响应是及时的，而信号响应通常有较大的时间延迟。

#### 实验步骤

正确编写满足功能的源文件。在这里我们的示例源文件存放在tasks\_k/4/task3目录下。

[root@localhost ~]# **cd tasks\_k/4/task3**

[root@localhost task2]# **ls**

catch\_signal.c

编译源文件。

[root@localhost task3]# **gcc catch\_signal.c -o catch\_signal**

[root@localhost task3]# **ls**

catch\_signal catch\_signal.c

执行并验证。

[root@localhost task3]# **./catch\_signal**

Current process ID is 21128

**^C**

Get a signal:SIGINT. You pressed ctrl+c.

[root@localhost task3]# **./catch\_signal**

Current process ID is 21141

**^\**

Get a signal:SIGQUIT. You pressed ctrl+\.

[root@localhost task3]# **./catch\_signal**

Current process ID is 21154

**^Z**

Get a signal:SIGHUP. You pressed ctrl+z.

[root@localhost task3]#

# 实验五 内核时间管理

## 实验介绍

本实验通过调用内核时钟接口打印系统当前时间，实现一个timer延时打印以及监控实现累加计算所花时间等任务操作，让学生们了解并掌握操作系统中的内核时间管理。

### 任务描述

* 编写内核模块，调用内核时钟接口，打印出系统当前时间。格式示例：2020-03-09 11:54:31；
* 编写内核模块程序，实现一个timer，该定时器延时10秒后打印“hello,world”。
* 调用内核时钟接口，编写内核模块，监控实现累加计算sum=1+2+3+...+100000所花时间。

## 实验目的

* 正确编写满足功能的源文件，正确编译。
* 正常加载、卸载内核模块；且内核模块功能满足任务所述。
* 了解操作系统的内核时间管理。

## 实验任务

### 调用内核时钟接口打印当前时间

#### 相关知识

一、内核时钟相关定义

1. timeval结构体

头文件：<linux/time.h>

struct timeval {

\_\_kernel\_time\_t tv\_sec; /\* seconds \*/

\_\_kernel\_suseconds\_t tv\_usec; /\* microseconds \*/

};

其中tv\_sec是自1970年1月1日 00:00:00 起到现在的秒数。而tv\_usec是当前秒数已经经过的微秒数。

2. do\_gettimeofday()

头文件：<linux/time.h>

函数原型：void do\_gettimeofday(struct timeval \*tv);

功能：返回自1970-01-01 00:00:00到现在的秒数，及当前秒经过的毫秒数，保存在tv指向的timeval结构体中。

3. rtc\_time结构体

头文件<linux/rtc.h>

struct rtc\_time {

int tm\_sec; // 表「秒」数，在[0,61]之间，多出来的两秒是用来处理跳秒问题用的。

int tm\_min; // 表「分」数，在[0,59]之间。

int tm\_hour; // 表「时」数，在[0,23]之间。

int tm\_mday; // 表「本月第几日」，在[1,31]之间。

int tm\_mon; // 表「本年第几月」，在[0,11]之间。

int tm\_year; // 要加1900表示那一年。

int tm\_wday; // 表「本周第几日」，在[0,6]之间。

int tm\_yday; // 表「本年第几日」，在[0,365]之间，闰年有366日。

int tm\_isdst; // 表是否为「日光节约时间」。

};

年份加上1900，月份加上1，小时加上8。

注意：可以自己先思考一下这样的加上x操作如果导致溢出，应该怎么处理。例如，得到的UTC小时数为22，加上8之后30明显溢出（超过24），并且由于小时数溢出到了第二天，因此天数也应该加1。

4. rtc\_time\_to\_tm()

头文件：<linux/rtc.h>

函数原型：void rtc\_time\_to\_tm(unsigned long time, struct rtc\_time \*tm)；

功能：将time存储的秒数转换为年月日时分秒等信息保存在rtc\_time结构体中。

参数：time为秒数，可以是do\_gettimeofday()函数获取的秒数。tm是rtc\_time结构体指针，结构体中存放了年月日时分秒等信息。

#### 实验步骤

正确编写满足功能的源文件，包括.c源文件和Makefile文件。在这里我们的示例源文件存放在tasks\_k/5/task1目录下。

[root@localhost ~]# **cd tasks\_k/5/task1**

[root@localhost task1]# **ls**

current\_time.c Makefile

编译源文件。

[root@localhost task1]# **make**

make -C /root/kernel M=/root/tasks\_k/5/task1 modules

make[1]: Entering directory '/root/kernel'

CC [M] /root/tasks\_k/5/task1/current\_time.o

Building modules, stage 2.

MODPOST 1 modules

CC /root/tasks\_k/5/task1/current\_time.mod.o

LD [M] /root/tasks\_k/5/task1/current\_time.ko

make[1]: Leaving directory '/root/kernel'

[root@localhost task1]# **ls**

current\_time.c current\_time.ko current\_time.mod.c current\_time.mod.o current\_time.o Makefile modules.order Module.symvers

加载编译完成的内核模块，并查看加载结果。

[root@localhost task1]# **insmod current\_time.ko**

[root@localhost task1]# **dmesg | tail -n2**

[838375.008715] Start current\_time module...

[838375.008718] Current time: 2020-12-20 15:23:05

卸载内核模块，并查看结果。

[root@localhost task1]# **rmmod current\_time**

[root@localhost task1]# **dmesg | tail -n3**

[838375.008715] Start current\_time module...

[838375.008718] Current time: 2020-12-20 15:23:05

[838405.862727] Exit current\_time module...

### 编写timer，在特定时刻打印 hello,world

#### 相关知识

一、内核定时器

时钟中断对于管理操作系统尤为重要，大量内核函数的生命周期都离不开流逝的时间的控制。正如我们所看到的，时钟中断能处理许多内核任务，所以它对内核来说极为重要。

定时器（有时也称为动态定时器或内核定时器）是管理内核流逝的时间的基础。内核经常需要推后执行某些代码，我们需要的是一种工具，能够使工作在指定时间点上执行——不长不短，正好在希望的时间点上。内核定时器正是解决这个问题的理想工具。

定时器的使用很简单。你只需要执行一些初始化工作，设置一个超时时间，指定超时发生后执行的函数，然后激活定时器就可以了。指定的函数在定时器到期时自动执行。注意定时器并不周期运行，它在超时后就自行撤销，这也正是这种定时器被称为动态定时器的一个原因：动态定时器不断地创建和撤销，而且它的运行次数也不受限制。

二、定时器的使用

1、定时器结构

定时器由结构timer\_list表示，定义在文件<linux/timer.h>中。

struct timer\_list

{

struct hlist\_node entry; /\* 定时器链表的入口 \*/

unsigned long expires; /\* 以 jiffies 为单位的定时值（32位） \*/

void (\*function)(struct timer\_list \*); /\* 定时器处理函数 \*/

u32 flags; /\* 定时器标志位\*/

#ifdef CONFIG\_LOCKDEP /\* 内核配置项——死锁检测模块，默认未配置\*/

struct lockdep\_map lockdep\_map; /\* 数据结构定义在 <linux/lockdep.h>中 \*/

#endif

KABI\_RESERVE(1) /\* kernel abi 的保留值\*/

KABI\_RESERVE(2)

KABI\_RESERVE(3)

KABI\_RESERVE(4)

}

注意：

（1）内核定时器用于控制某个函数（定时器处理函数）在未来的某个特定时间执行。

（2）CONFIG\_LOCKDEP 的设置依赖于 CONFIG\_DEBUG\_KERNEL，而 CONFIG\_DEBUG\_KERNEL 设置为Y时，表示正在开发驱动程序或尝试调试和确定内核问题，因此通常情况下，CONFIG\_LOCKDEP是未配置的状态。

（3）expires 的值是32位的，因为内核定时器并不适用于长的未来时间点。

2、创建定时器

创建定时器时需要先定义定时器结构：一个timer\_list结构体的实例对应一个定时器。也就是说，创建定时器就是实例化定时器数据结构。

3、初始化定时器

初始化 struct timer\_list 数据结构，只需正确地设置 expires 与 处理函数 function。其他参数无需设置。

timer.expires = jiffies + delay; /\* 定时器超时时的节拍数 \*/

timer.function = my\_function; /\* 定时器超时时调用的函数 \*/

（1）定时的数值 delay 如何设置：

因为内核定时器是基于jiffies，所以设置的时间要在jiffies之上加上我们的定时值；例如：定时2秒，则delay=2\*HZ，HZ是每秒中jiffies这个计数器会计多少值。

（2）若当前jiffies计数≥timer.expires，timer.function指向的处理函数就会开始执行。

（3）定时器处理函数必须符合下面的函数原型：

void my\_function(struct timer\_list \*timer);

4、激活定时器

函数：add\_timer(struct timer\_list \*timer)

功能：将定时器注册到内核中（将定时器连接到内核专门的链表中），使之生效。

add\_timer(&timer);

在注册之后，定时器就开始计时，在到达时间expires时，执行初始化时指定的处理函数。当激活定时器后，它只会执行一次处理函数，然后将定时器从内核中移除。

#### 实验步骤

正确编写满足功能的源文件，包括.c源文件和Makefile文件。在这里我们的示例源文件存放在tasks\_k/5/task2目录下。

[root@localhost ~]# **cd tasks\_k/5/task2**

[root@localhost task2]# **ls**

Makefile timer\_example.c

编译源文件。

[root@localhost task2]# **make**

make -C /root/kernel M=/root/tasks\_k/5/task2 modules

make[1]: Entering directory '/root/kernel'

CC [M] /root/tasks\_k/5/task2/timer\_example.o

Building modules, stage 2.

MODPOST 1 modules

CC /root/tasks\_k/5/task2/timer\_example.mod.o

LD [M] /root/tasks\_k/5/task2/timer\_example.ko

make[1]: Leaving directory '/root/kernel'

[root@localhost task2]# **ls**

Makefile modules.order Module.symvers timer\_example.c timer\_example.ko timer\_example.mod.c timer\_example.mod.o timer\_example.o

加载编译完成的内核模块，并查看加载结果。

[root@localhost task2]# **insmod timer\_example.ko**

[root@localhost task2]# **dmesg -T | tail -n2**

[Sun Dec 20 15:25:25 2020] Start timer\_example module...

[Sun Dec 20 15:25:35 2020] hello,world!

卸载内核模块，并查看结果。

[root@localhost task2]# **rmmod timer\_example**

[root@localhost task2]# **dmesg -T | tail -n3**

[Sun Dec 20 15:25:25 2020] Start timer\_example module...

[Sun Dec 20 15:25:35 2020] hello,world!

[Sun Dec 20 15:25:52 2020] Exit timer\_example module...

代码中设置定时器超时时间为10\*HZ（10秒），对应模块加载后10秒才打印hello，world!。

### 调用内核时钟接口，监控累加计算代码的运行时间

#### 实验步骤

正确编写满足功能的源文件，包括.c源文件和Makefile文件。在这里我们的示例源文件存放在tasks\_k/5/task3目录下。

[root@localhost ~]# **cd tasks\_k/5/task3**

[root@localhost task3]# **ls**

Makefile sum\_time.c

编译源文件。

[root@localhost task3]# **make**

make -C /root/kernel M=/root/tasks\_k/5/task3 modules

make[1]: Entering directory '/root/kernel'

CC [M] /root/tasks\_k/5/task3/sum\_time.o

Building modules, stage 2.

MODPOST 1 modules

CC /root/tasks\_k/5/task3/sum\_time.mod.o

LD [M] /root/tasks\_k/5/task3/sum\_time.ko

make[1]: Leaving directory '/root/kernel'

加载编译完成的内核模块，并查看加载结果。

[root@localhost task3]# **insmod sum\_time.ko**

[root@localhost task3]# **dmesg | tail -n5**

[838636.082885] Start sum\_time module...

[838636.082888] The start time is: 1608449246 s 863700 us

[838636.082889] The sum of 1 to 100000 is: 5000050000

[838636.082890] The end time is: 1608449246 s 863702 us

[838636.082891] The cost time of sum from 1 to 100000 is: 2 us

卸载内核模块，并查看结果。

[root@localhost task3]# **rmmod sum\_time**

[root@localhost task3]# **dmesg | tail -n6**

[838636.082885] Start sum\_time module...

[838636.082888] The start time is: 1608449246 s 863700 us

[838636.082889] The sum of 1 to 100000 is: 5000050000

[838636.082890] The end time is: 1608449246 s 863702 us

[838636.082891] The cost time of sum from 1 to 100000 is: 2 us

[838662.886065] Exit sum\_time module...

运行结果可以看出，从1到100000的累加和所花时间是2us。

# 实验六 设备管理

## USB设备驱动程序实验

### 相关知识

一、USB骨架（usb-skeleton）

1、简介

USB骨架程序（usb-skeleton）是USB驱动程序的基础，通过对它源码的学习和理解，可以使我们迅速地了解USB驱动架构，迅速地开发我们自己的USB硬件的驱动。

Linux USB子系统已经从仅支持2.2.7内核中的两种不同类型的设备（鼠标和键盘）发展到了2.4内核中的20多种不同类型的设备。 Linux当前支持几乎所有USB类设备（标准类型的设备，如键盘，鼠标，调制解调器，打印机和扬声器），以及越来越多的特定于供应商的设备（如USB到串行转换器，数码相机和以太网设备等）。

Linux上不支持的其余USB设备几乎是所有特定于供应商的设备。如果生产厂商在他们的产品中使用自己定义的协议，他们就需要为此设备创建特定的驱动程序。当然我们知道，有些生产厂商公开他们的USB协议，并帮助Linux驱动程序的开发，然而有些生产厂商却根本不公开他们的USB协议。因为每一个不同的协议都会产生一个新的驱动程序，所以就有了这个通用的USB驱动骨架程序，它是以pci 骨架为模板的。

因为每种不同的协议都会导致创建新的驱动程序，如果你准备写一个linux驱动程序，首先要熟悉USB协议规范。这里，通用的USB驱动程序框架以pci-skeleton.c文件为模板，内核源代码树的drivers/usb/usb-skeleton.c文件为我们提供了一个最基础的USB驱动程序，我们称之为USB骨架。USB骨架程序可以被看做一个最简单的USB设备驱动的实例，通过它我们仅需要修改极少的部分，就可以完成一个USB设备的驱动。

对于如何编写USB驱动程序，内核源码文档中，Documentation/driver-api/usb/writing\_usb\_driver.rst文件有详细介绍。

2、Linux USB基础

Linux USB 驱动程序需要做的第一件事情就是在Linux USB 子系统里注册，并提供一些相关信息，例如：这个驱动程序支持哪种设备，当被支持的设备从系统插入或拔出时，会有哪些动作。所有这些信息都传送到USB 子系统中，在USB骨架驱动程序中是这样来表示的：

static struct usb\_driver skel\_driver = {

.name = "skeleton", //名字

.probe = skel\_probe, //探测函数

.disconnect = skel\_disconnect, //抽调函数

.suspend = skel\_suspend,

.resume = skel\_resume,

.pre\_reset = skel\_pre\_reset,

.post\_reset = skel\_post\_reset,

.id\_table = skel\_table, //支持设备列表

.supports\_autosuspend = 1,

};

module\_usb\_driver(skel\_driver)

变量name是一个字符串，它对驱动程序进行描述。probe 和disconnect 是函数指针，当插入/拔出的USB设备与在id\_table 中变量信息匹配时，此函数被调用。

3、USB驱动的注册和注销

（1）注册USB驱动

通常在驱动程序的init函数中通过调用usb\_register来注册USB驱动程序，

USB驱动程序在注册时会发送一个命令给usb\_register，通常在驱动程序的初始化函数里。

static int \_\_init usb\_skel\_init(void)

{

int result;

/\* register this driver with the USB subsystem \*/

result = usb\_register(&skel\_driver);

if (result < 0) {

err("usb\_register failed for the "\_\_FILE\_\_ "driver."

"Error number %d", result);

return -1;

}

return 0;

}

module\_init(usb\_skel\_init);

（2）注销USB驱动

当要从系统卸载驱动程序时，它需要向USB子系统注销自身。即需要usb\_deregister() 函数处理：

static void \_\_exit usb\_skel\_exit(void)

{

/\* deregister this driver with the USB subsystem \*/

usb\_deregister(&skel\_driver); //注销设备

}

module\_exit(usb\_skel\_exit);

当USB设备插入时，为了使linux-hotplug（Linux中PCI、USB等设备热插拔支持）系统自动装载驱动程序，你需要创建一个MODULE\_DEVICE\_TABLE。代码如下（这个模块仅支持某一特定设备）：

以下代码告诉热插拔脚本，该模块支持具有特定供应商和产品ID的单个设备：

/\* Define these values to match your devices \*/

#define USB\_DETECT\_VENDOR\_ID 0xfff0

#define USB\_DETECT\_PRODUCT\_ID 0xfff0

/\* table of devices that work with this driver \*/

static struct usb\_device\_id skel\_table [] = { //在系统中注册设备

{ USB\_DEVICE(USB\_SKEL\_VENDOR\_ID, USB\_SKEL\_PRODUCT\_ID) }, //生产商号和设备号

{ } /\* Terminating entry \*/

};

MODULE\_DEVICE\_TABLE (usb, skel\_table);

其中，USB\_DETECT\_VENDOR\_ID 与 USB\_DETECT\_PRODUCT\_ID对应 lsusb 输出列表中的ID信息。（见后文具体描述）

4、设备操作接口说明

（1）探测函数probe

USB\_DEVICE宏利用厂商ID和产品ID为我们提供了一个设备的唯一标识。当系统插入一个ID匹配的USB设备到USB总线时，驱动会在USB core中注册。驱动程序中probe 函数也就会被调用。usb\_device 结构指针、接口号和接口ID都会被传递到函数中。

// 参见 drivers/usb/usb-skeleton.c

static int skel\_probe(struct usb\_interface \*interface, const struct usb\_device\_id \*id)

驱动程序需要确认插入的设备是否可以被接受，如果是，则返回0；如果不是，或者初始化期间发生任何错误，则从探测函数返回错误代码（例如-ENOMEM或-ENODEV）。

在骨架驱动程序中，我们确定将哪些端点标记为大进大出。我们创建了缓冲区来保存将要从设备发送和接收的数据，并且已初始化了将数据写入设备的USB urb（URB，USB Request Block，USB请求块）。

相反，将设备从USB总线中拔出时，设备指针会调用disconnect 函数。驱动程序需要清除此时已分配的所有私有数据，并关闭USB系统中所有待定的urb。

（2）抽调函数disconnect

当用户程序释放它用来与设备通信的文件句柄时，将调用驱动程序中的skel\_disconnect函数。

// 参见 drivers/usb/usb-skeleton.c

static void skel\_disconnect(struct usb\_interface \*interface)

（3）open函数

现在，skeleton驱动就已经和设备绑定上了，任何用户态程序要操作此设备都可以通过file\_operations结构所定义的函数进行了。当程序尝试为I/O打开设备时，首先被调用的函数是open()。

// 参见 drivers/usb/usb-skeleton.c

static int skel\_open(struct inode \*inode, struct file \*file)

当open完设备后，skel的read、write函数就可以收、发数据了，他们是完成驱动对读写等操作的响应。

（4）write函数

// 参见 drivers/usb/usb-skeleton.c

static ssize\_t skel\_write(struct file \*file, const char \*user\_buffer, size\_t count, loff\_t \*ppos)

参数说明：

struct file \*file 待写入的文件指针

const char \*user\_buffer 指向用户想要发送到设备的数据的指针

size\_t count 发送数据的大小

loff\_t \*ppos 在当前文件中进行写入操作的位置/偏移量

该函数根据已创建的写入urb的大小来确定可以发送给设备的数据量（此大小取决于设备具有的批量输出端点的大小）。 然后，它将数据从用户空间复制到内核空间，将urb指向数据，并将urb提交给USB子系统。

（5）read函数

// 参见 drivers/usb/usb-skeleton.c

static ssize\_t skel\_read(struct file \*file, char \*buffer, size\_t count, loff\_t \*ppos)

参数说明：

struct file \*file 待写入的文件指针

const char \*user\_buffer 指向用户想要接收的设备的数据的指针

size\_t count 接收数据的大小

loff\_t \*ppos 在当前文件中进行读取操作的位置/偏移量

read 函数与write 函数稍有不同在于：程序并没有用urb 将数据从设备传送到驱动程序，而是我们用usb\_bulk\_msg 函数代替，这个函数能够在不需要创建urbs 和操作urb函数的情况下，来发送数据给设备，或者从设备来接收数据。我们调用usb\_bulk\_msg函数，并为它提供一个缓冲区，该缓冲区用于放置从设备接收到的所有数据以及超时值。 如果超时期限未收到设备的任何数据，则该功能将失败并返回错误消息。

（6）delete函数

由于USB设备可以在任何时间点从系统中取走，即使程序目前正在访问它。USB驱动程序必须要能够很好地处理解决此问题，它需要能够切断任何当前的读写，同时通知用户空间程序：USB设备已不存在。此操作由skel\_delete函数执行。

如果程序有一个打开的设备句柄，在当前结构里，我们只要把它赋值为空，就像它已经消失了。对于每一次设备读写等其它函数操作，我们都要检查usb\_device结构是否存在。如果不存在，就表明设备已经消失，并返回一个-ENODEV错误给用户程序。当最终我们调用release函数时，在没有文件打开这个设备时，无论usb\_device结构是否存在、它都会清空skel\_disconnect函数所作工作。

二、查看USB设备的命令

1. lspci命令

[root@localhost Log]# lspci | grep USB

7a:00.0 USB controller: Huawei Technologies Co., Ltd. HiSilicon USB 1.1 Host Controller (rev 21)

7a:01.0 USB controller: Huawei Technologies Co., Ltd. HiSilicon USB 2.0 2-port Host Controller (rev 21)

7a:02.0 USB controller: Huawei Technologies Co., Ltd. HiSilicon USB 3.0 Host Controller (rev 21)

ba:00.0 USB controller: Huawei Technologies Co., Ltd. HiSilicon USB 1.1 Host Controller (rev 21)

ba:01.0 USB controller: Huawei Technologies Co., Ltd. HiSilicon USB 2.0 2-port Host Controller (rev 21)

ba:02.0 USB controller: Huawei Technologies Co., Ltd. HiSilicon USB 3.0 Host Controller (rev 21)

可以通过命令lspci | grep USB查看机器上的USB主控制器，可以通过命令dnf install -y pciutils安装lspci。

2. lsusb命令

在Linux中我们可以使用lsusb来列出USB设备和它的属性，lsusb会显示驱动和内部连接到你系统的设备。直接在控制台输入lsusb即可。

用法： lsusb [options]...

功能：查看/列出USB设备

参数说明：（可通过 lsusb --help 查看详细使用说明）

-v, --verbose 增加详细程度（显示描述符），列出所有USB的详细

-s [[bus]:][devnum] 仅显示具有指定设备和/或总线号的设备（十进制）

-d vendor:[product] 仅显示具有指定供应商和产品ID号（十六进制）的设备

-D device 选择lsusb将检查的设备

-t, --tree 将物理USB设备层次结构转储为树

-V, --version 显示程序版本

-h, --help 显示用法和帮助

如果无法运行 lsusb，使用以下命令安装

dnf install -y usbutils

然后就可以运行 lsusb了：

[root@localhost task1]# lsusb

Bus 008 Device 001: ID 1d6b:0003 Linux Foundation 3.0 root hub

Bus 007 Device 001: ID 1d6b:0002 Linux Foundation 2.0 root hub

Bus 002 Device 001: ID 1d6b:0002 Linux Foundation 2.0 root hub

Bus 004 Device 001: ID 1d6b:0001 Linux Foundation 1.1 root hub

Bus 006 Device 002: ID 0bda:0411 Realtek Semiconductor Corp.

Bus 006 Device 001: ID 1d6b:0003 Linux Foundation 3.0 root hub

Bus 005 Device 003: ID 0781:5567 SanDisk Corp. Cruzer Blade

Bus 005 Device 002: ID 0bda:5411 Realtek Semiconductor Corp.

Bus 005 Device 001: ID 1d6b:0002 Linux Foundation 2.0 root hub

Bus 001 Device 003: ID 12d1:0003 Huawei Technologies Co., Ltd.

Bus 001 Device 002: ID 0bda:5411 Realtek Semiconductor Corp.

Bus 001 Device 001: ID 1d6b:0002 Linux Foundation 2.0 root hub

Bus 003 Device 001: ID 1d6b:0001 Linux Foundation 1.1 root hub

其中ID 0bda:5411

0bda：USB\_DETECT\_VENDOR\_ID，其值表示是Realtek Semiconductor Corp.。

5411：USB\_DETECT\_PRODUCT\_ID，其值表示是RTS5411 Hub。

三、内核模块编程

源码编写—— .c源文件

Makefile文件编写

编译模块——make

模块加载进内核——insmod

查看加载的内容——dmesg

查看内核模块——lsmod

卸载内核模块——rmmod

### 任务描述

1. 参考内核源码中的drivers/usb/usb-skeleton.c文件，编写一个USB探测驱动程序，能够实现以下基本功能：

（1）在插入U盘时能够探测到；

（2）在拔出U盘时能够探测到；

2. 加载、卸载模块并查看模块打印信息。

### 实验步骤

编写检测USB设备的源代码

请参见源码压缩包tasks-k\6\task1目录下的usb\_detect.c和Makefile。

编译源码

[root@localhost task1]# **make**

usb\_detect.ko会被编译出来。

加载内核模块

[root@localhost task1]# **insmod usb\_detect.ko**

[root@localhost task1]# **lsmod | grep usb\_detect**

usb\_detect 262144 0

[root@localhost task1]# **dmesg | tail -n2**

[49756.195888] Start usb\_detect module...

[49756.195919] usbcore: registered new interface driver usbdetect

此时U盘检测模块已启动。

插入U盘并查看信息

在Taishan服务器的USB接口插入U盘后，查看信息可见模块检测到U盘插入：

[root@localhost task1]# **dmesg | tail -n19**

[49756.195888] Start usb\_detect module...

[49756.195919] usbcore: registered new interface driver usbdetect

[50062.541917] usb 5-1.2: new high-speed **USB device number 3** using xhci\_hcd

[50062.642542] usb 5-1.2: New USB device found, **idVendor=0781, idProduct=5567**, bcdDevice= 1.00

[50062.642544] usb 5-1.2: New USB device strings: Mfr=1, Product=2, SerialNumber=3

[50062.642546] usb 5-1.2: Product: **Cruzer Blade**

[50062.642547] usb 5-1.2: Manufacturer: **SanDisk**

[50062.642548] usb 5-1.2: SerialNumber: 200443243016069310F7

[50062.643302] usb-storage 5-1.2:1.0: USB Mass Storage device detected

[50062.643444] scsi host6: usb-storage 5-1.2:1.0

[50062.658083] usbcore: registered new interface driver uas

[50063.667769] scsi 6:0:0:0: Direct-Access **SanDisk Cruzer Blade** 1.00 PQ: 0 ANSI: 6

[50063.668420] sd 6:0:0:0: Attached scsi generic sg2 type 0

[50063.668504] sd 6:0:0:0: [sdb] 122508544 512-byte logical blocks: (62.7 GB/58.4 GiB)

[50063.670221] sd 6:0:0:0: [sdb] Write Protect is off

[50063.670222] sd 6:0:0:0: [sdb] Mode Sense: 43 00 00 00

[50063.670512] sd 6:0:0:0: [sdb] Write cache: disabled, read cache: enabled, doesn't support DPO or FUA

[50063.688012] sdb: sdb1

[50063.690448] sd 6:0:0:0: [sdb] Attached SCSI removable disk

获取到的U盘信息与lsusb显示的一致：

[root@localhost task1]# **lsusb**

Bus 008 Device 001: ID 1d6b:0003 Linux Foundation 3.0 root hub

Bus 007 Device 001: ID 1d6b:0002 Linux Foundation 2.0 root hub

Bus 002 Device 001: ID 1d6b:0002 Linux Foundation 2.0 root hub

Bus 004 Device 001: ID 1d6b:0001 Linux Foundation 1.1 root hub

Bus 006 Device 002: ID 0bda:0411 Realtek Semiconductor Corp.

Bus 006 Device 001: ID 1d6b:0003 Linux Foundation 3.0 root hub

Bus 005 Device 003: ID 0781:5567 **SanDisk Corp. Cruzer Blade**

Bus 005 Device 002: ID 0bda:5411 Realtek Semiconductor Corp.

Bus 005 Device 001: ID 1d6b:0002 Linux Foundation 2.0 root hub

Bus 001 Device 003: ID 12d1:0003 Huawei Technologies Co., Ltd.

Bus 001 Device 002: ID 0bda:5411 Realtek Semiconductor Corp.

Bus 001 Device 001: ID 1d6b:0002 Linux Foundation 2.0 root hub

Bus 003 Device 001: ID 1d6b:0001 Linux Foundation 1.1 root hub

从Taishan服务器上拔出U盘

[root@localhost task1]# **dmesg | tail -n20**

[49756.195888] Start usb\_detect module...

[49756.195919] usbcore: registered new interface driver usbdetect

[50062.541917] usb 5-1.2: new high-speed USB device number 3 using xhci\_hcd

[50062.642542] usb 5-1.2: New USB device found, idVendor=0781, idProduct=5567, bcdDevice= 1.00

[50062.642544] usb 5-1.2: New USB device strings: Mfr=1, Product=2, SerialNumber=3

[50062.642546] usb 5-1.2: Product: Cruzer Blade

[50062.642547] usb 5-1.2: Manufacturer: SanDisk

[50062.642548] usb 5-1.2: SerialNumber: 200443243016069310F7

[50062.643302] usb-storage 5-1.2:1.0: USB Mass Storage device detected

[50062.643444] scsi host6: usb-storage 5-1.2:1.0

[50062.658083] usbcore: registered new interface driver uas

[50063.667769] scsi 6:0:0:0: Direct-Access SanDisk Cruzer Blade 1.00 PQ: 0 ANSI: 6

[50063.668420] sd 6:0:0:0: Attached scsi generic sg2 type 0

[50063.668504] sd 6:0:0:0: [sdb] 122508544 512-byte logical blocks: (62.7 GB/58.4 GiB)

[50063.670221] sd 6:0:0:0: [sdb] Write Protect is off

[50063.670222] sd 6:0:0:0: [sdb] Mode Sense: 43 00 00 00

[50063.670512] sd 6:0:0:0: [sdb] Write cache: disabled, read cache: enabled, doesn't support DPO or FUA

[50063.688012] sdb: sdb1

[50063.690448] sd 6:0:0:0: [sdb] Attached SCSI removable disk

[50961.908882] **usb 5-1.2: USB disconnect, device number 3**

可以看出设备号为3的USB设备已断开连接。

卸载内核模块

[root@localhost task1]# **rmmod usb\_detect**

[root@localhost task1]# **dmesg | tail -n2**

[51797.480839] Exit usb\_detect module...

[51797.480841] usbcore: deregistering interface driver usbdetect

## 内核模块测试硬盘的读写速率实验

本实验通过编写内核模块测试硬盘的读写速率，加载、卸载模块并查看模块打印信息的任务操作，让学生们了解并掌握操作系统中的设备管理。

### 相关知识

一、内核文件读写介绍

有时候需要在Linux kernel中读写文件数据，如调试程序的时候，或者内核与用户空间交换数据的时候。在kernel中操作文件没有标准库可用，需要利用kernel的一些函数，这些函数主要有：

filp\_open()

filp\_close()

kernel\_read()

kernel\_write()

这些函数在 <linux/fs.h> 头文件中声明。

二、内核文件读写接口

1、打开文件

函数原型：struct file\* filp\_open(const char\* filename, int open\_mode, int mode);

功能：在kernel中打开指定文件。

返回值：该函数返回 struct file\* 结构指针，供后续函数操作使用；该返回值用IS＿ERR（）来检验其有效性。

参数说明：

filename：表明要打开或创建文件的名称（包括路径部分）。

注意：在内核中打开文件时需要注意打开的时机，很容易出现需要打开文件的驱动很早就加载并打开文件，但需要打开的文件所在设备还没有挂载到文件系统中，而导致打开失败。

open\_mode：文件的打开方式，其取值与标准库中的open相应参数类似，包括：O\_RDONLY（只读打开）、O\_WRONLY（只写打开）、O\_RDWR（读写打开）、O\_CREAT（文件不存在则创建）等。

mode：创建文件时使用，设置创建文件的读写权限（如 644），其它情况可以设为0。

2、读文件

函数原型：ssize\_t kernel\_read(struct file \*file, void \*buf, size\_t count, loff\_t \*pos);

功能：kernel中文件的读操作。

参数说明：

file：进行读取信息的目标文件，即file\_open() 函数的返回值。

buf：对应放置信息的缓冲区。

count：要读取的信息长度。

pos：表示用户在当前文件中进行读取操作的位置/偏移量，即读的位置相对于文件开头的偏移。在读取信息后，这个指针一般都会移动，移动的值为要读取信息的长度值。

3、写文件

函数原型：ssize\_t kernel\_write(struct file \*file, const void \*buf, size\_t count, loff\_t \*pos);

功能：kernel中文件的写操作。

参数说明：

file：进行信息写入的目标文件，即file\_open() 函数的返回值。

buf：要写入文件的信息缓冲区。

count：要写入信息的长度。

pos：表示用户在当前文件中进行写入操作的位置/偏移量。即写的位置相对于文件开头的偏移。

4、关闭文件

函数原型：int filp\_close(struct file\*filp, fl\_owner\_t id);

功能：关闭指定文件。

参数说明：

filp：待关闭的目标文件的文件指针。

id：一般传递NULL值，也可用current->files作为实参。

### 任务描述

* 编写内核模块测试硬盘的读、写速率，加载、卸载模块并查看模块打印信息。

### 实验目的

* 正确编写满足功能的源文件，正确编译。
* 正常加载、卸载内核模块；且内核模块功能满足任务所述。
* 了解操作系统的设备管理。

### 实验步骤

#### 编写内核模块测试硬盘的写速率

正确编写满足功能的源文件，包括.c源文件和Makefile文件。在这里我们的示例源文件存放在tasks\_k/6/write\_disk目录下。

[root@localhost ~]# **cd tasks\_k/6/write\_disk/**

[root@localhost write\_disk]# **ls**

Makefile write\_to\_disk.c

编译源文件。

[root@localhost write\_disk]# **make**

make -C /root/kernel M=/root/tasks\_k/6/write\_disk modules

make[1]: Entering directory '/root/kernel'

CC [M] /root/tasks\_k/6/write\_disk/write\_to\_disk.o

Building modules, stage 2.

MODPOST 1 modules

CC /root/tasks\_k/6/write\_disk/write\_to\_disk.mod.o

LD [M] /root/tasks\_k/6/write\_disk/write\_to\_disk.ko

make[1]: Leaving directory '/root/kernel'

加载编译完成的内核模块，并查看加载结果。

[root@localhost write\_disk]# **insmod write\_to\_disk.ko**

[root@localhost write\_disk]# **dmesg | tail -n3**

[839074.277466] Start write\_to\_disk module...

[839075.129853] Writing to file costs 852104 us

[839075.134109] Writing speed is 630 M/s

[root@localhost write\_disk]# **ll /home/tmp\_file**

-rw-------. 1 root root 512M Dec 20 15:34 /home/tmp\_file

卸载内核模块，并查看结果。

[root@localhost write\_disk]# **rmmod write\_to\_disk**

[root@localhost write\_disk]# **dmesg | tail -n4**

[839074.277466] Start write\_to\_disk module...

[839075.129853] Writing to file costs 852104 us

[839075.134109] Writing speed is 630 M/s

[839170.669591] Exit write\_to\_disk module...

#### 编写内核模块测试硬盘的读速率

正确编写满足功能的源文件，包括.c源文件和Makefile文件。在这里我们的示例源文件存放在tasks\_k/6/read\_disk目录下。

[root@localhost ~]# **cd tasks\_k/6/read\_disk/**

[root@localhost read\_disk]# **ls**

Makefile read\_from\_disk.c

编译源文件。

[root@localhost read\_disk]# **make**

make -C /root/kernel M=/root/tasks\_k/6/read\_disk modules

make[1]: Entering directory '/root/kernel'

CC [M] /root/tasks\_k/6/read\_disk/read\_from\_disk.o

Building modules, stage 2.

MODPOST 1 modules

CC /root/tasks\_k/6/read\_disk/read\_from\_disk.mod.o

LD [M] /root/tasks\_k/6/read\_disk/read\_from\_disk.ko

make[1]: Leaving directory '/root/kernel'

加载编译完成的内核模块，并查看加载结果。

[root@localhost read\_disk]# **insmod read\_from\_disk.ko**

[root@localhost read\_disk]# **dmesg | tail -n3**

[839397.876094] Start read\_from\_disk module...

[839398.137164] Read file costs 261064 us

[839398.140903] Reading speed is 2056 M/s

卸载内核模块，并查看结果。

[root@localhost read\_disk]# **rmmod read\_from\_disk**

[root@localhost read\_disk]# **dmesg | tail -n4**

[839397.876094] Start read\_from\_disk module...

[839398.137164] Read file costs 261064 us

[839398.140903] Reading speed is 2056 M/s

[839419.751503] Exit read\_from\_disk module...

# 实验七 文件系统

## 实验介绍

本实验通过为Ext4文件系统添加扩展属性，注册自定义文件系统类型，以及使用内核模块操作文件系统等任务，让学生们了解并掌握操作系统中的文件系统。

### 任务描述

* 熟悉文件系统扩展属性EA，为Ext4文件系统添加扩展属性；
* 使用文件系统注册/注销函数，注册一个自定义文件系统类型；
* 编写一个模块，在加载模块时，在/proc目录下创建一个名称为myproc的目录；
* 编写一个模块，该模块有三个参数：一个为字符串型，两个为整型。两个整型中，一个在/sys下不可见。加载模块后，使用echo向模块传递参数值来改变指定参数的值。

## 实验目的

* 正确编写满足功能的源文件，正确编译。
* 正常加载、卸载内核模块；且内核模块功能满足任务所述。
* 了解操作系统的内核文件系统管理。

## 实验任务

### 为Ext4文件系统添加扩展属性

#### 相关知识

1、基本概念

文件扩展属性（xattr-Extended attributes）提供了一种机制，用来将key-value键值对永久地关联到文件；让现有的文件系统得以支持在原始设计中未提供的功能。扩展属性是目前流行的POSIX文件系统具有的一项特殊的功能，可以给文件、文件夹添加额外的Key-value的键值对，键和值都是字符串并且有一定长度的限制——定义于include/uapi/linux/limits.h 文件中。在保存xattr时，key的长度不能超过255byte，value不能超过64k，总的配对数不能超过64k。

xattr扩展函数，仅仅作用于“支持扩展属性的文件系统”，并在挂载文件系统时，需要开启xattr。因为扩展属性需要底层文件系统的支持，在使用扩展属性的时候，需要查看文件系统说明文章，看此文件系统是否支持扩展属性，以及对扩展属性命名空间等相关的支持。支持扩展属性常见的文件系统有：ext2、ext3、reiserfs、jfs和xfs，这些文件系统对于扩展属性的支持都是可选项。

2、扩展属性名称空间

扩展属性名称的格式是namespace.attribute，名称空间namespace是用来定义不同的扩展属性的类。目前有security，system，trusted，user四种扩展属性类。

（1）扩展的安全属性——security

安全属性名称空间被内核用于安全模块，例如SELinux。对安全属性的读和写权限依赖于策略的设定。这策略是由安全模块载入的。如果没有载入安全模块，所有的进程都对安全属性有读权限，写权限只有那些有CAP\_SYS\_ADMIN（允许执行系统管理任务，如加载或卸载文件系统、设置磁盘配额等）的进程才有。

（2）扩展的系统属性——system

扩展的系统属性被内核用来存储系统对象，比如说ACL。对系统属性的读和写权限依赖于策略的设定。

（3）受信任的扩展属性——trusted

受信任的扩展属性只对那些有CAP\_SYS\_ADMIN的进程可见和可获得。这个类中的属性被用来在用户空间中保存一些普通进程无法得到的信息。

（4）扩展的用户属性——user

扩展的用户属性被分配给文件和目录用来存储任意的附加信息，比如mime type、字符集或是文件的编码。用户属性的权限由文件权限位来定义。对于普通文件和目录，文件权限位定义文件内容的访问，对于设备文件来说，它们定义对设备的访问。扩展的用户属性只被用于普通的文件和目录，对用户属性的访问被限定于属主和那些对目录有sticky位设置的用户。

3、文件系统特殊要求

对于ext文件系统，为了能使用扩展用户属性，要求文件系统挂载时有user\_xattr选项。

在ext文件系统中，每一个扩展属性必须占用一个单独的文件系统块，块大小取决于创建文件系统时的设置。

4、文件扩展属性的设置命令

（1）setfattr：设置文件系统对象的扩展属性（无文件系统限制）

语法：setfattr [-h] -n name [-v value] pathname...

setfattr [-h] -x name pathname...

setfattr [-h] --restore=file

-n name：指定属性名称

-v value：设置属性值。可以使用三种方法对value进行编码：如果给定的字符串用双引号引起来，则将其中的字符串视为文本。在这种情况下，其中的反斜线和双引号需转义。任何控制字符都可以编码为“反斜杠后跟三个数字”作为八进制的ASCII码。如果给定的字符串以0x开头，则表示一个十六进制数。如果给定的字符串以0s开头，则需要base64编码。

-x name：删除属性

-h：不要遵循符号链接。 如果路径名是符号链接，则不会遵循它，而是将其本身修改为inode。

--version：打印setfattr的版本并退出。

--help：打印帮助以解释命令行选项。

（2）getfattr：获取文件系统对象的扩展属性

-n name：显示指定名称的属性。

-d：显示所有属性的值。

-e en：在提取属性之后进行编码，en的值为text、hex和base64。

en= "text"时，编码为文本的字符串的值用双引号（“）引起来；

en= "hex"时，编码为十六进制的字符串以0x为前缀；

en= "base64"时，编码为base64的字符串以0s为前缀。

-m pattern：正则表达式匹配的属性显示。默认匹配为'^user\\.'，即匹配user名称空间的属性，可以指定为 '-' 或 '.' 匹配所有属性。

-R：递归显示。

#### 实验步骤

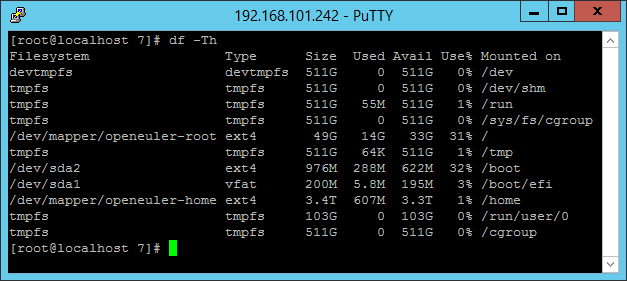
环境准备：为了使用扩展属性，需要安装libattr：

dnf install -y libattr

查看当前文件系统类型

df -Th

其中，-T 用于显示文件系统类型；-h表示以1024的幂为单位显示文件系统大小。

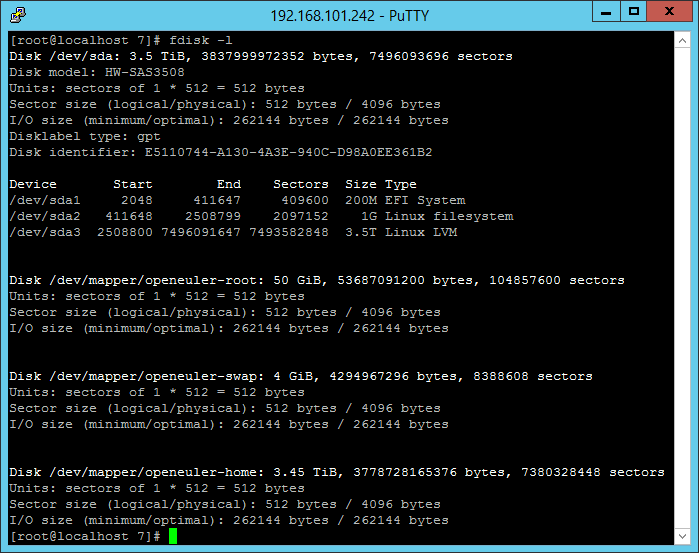


可使用df --help查看df的具体参数及用法。

检查当前文件系统是否支持文件扩展属性

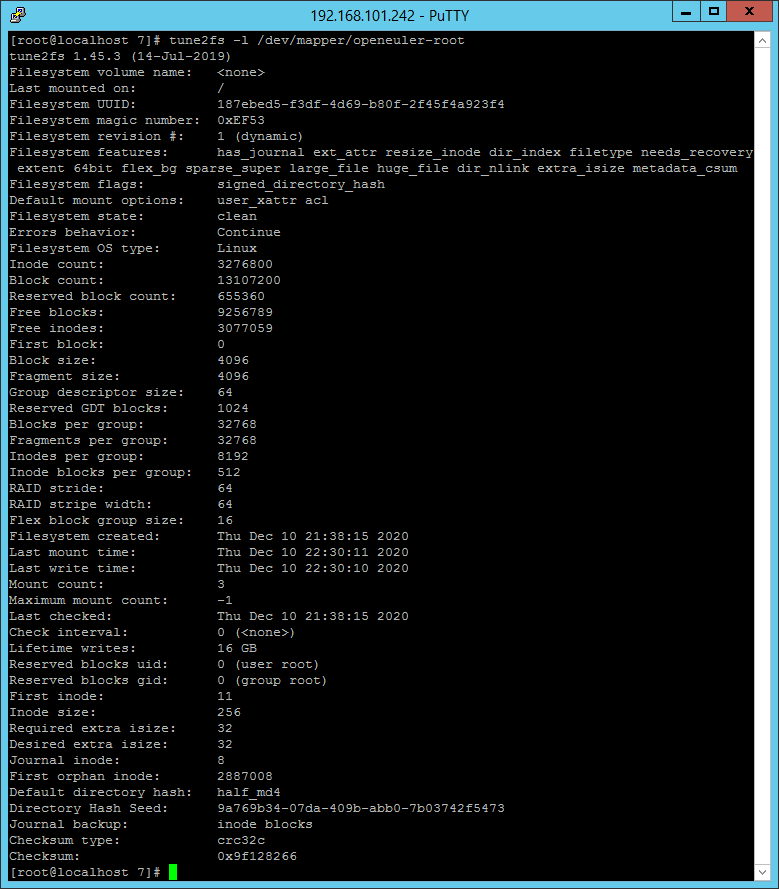
在ext3和ext4文件系统上，可通过命令tune2fs -l来检查是否支持扩展属性。

（1）用 fdisk -l 查看硬盘及分区信息



（2）用 tune2fs -l 显示设备的详细信息

tune2fs命令允许系统管理员在ext2、ext3或ext4文件系统上调整各种可调的文件系统参数。这些选项的当前值可以使用-l选项显示。



通过检查/dev/mapper/openeuler-root文件系统参数中的默认挂载选项“Default mount options”是否有对应内容；来确定分区设备的文件系统是否支持扩展属性——上图所示说明，该文件系统支持扩展用户属性user\_xattr、与ACL权限。

创建文件file.txt，用setfattr设置文件系统对象的扩展属性

[root@localhost task1]# **touch file.txt**

[root@localhost task1]# **setfattr -n user.name -v xattr\_test file.txt**

[root@localhost task1]# **setfattr -n user.city -v "Beijing" file.txt**

[root@localhost task1]# **getfattr -d -m . file.txt**

# file: file.txt

security.selinux="unconfined\_u:object\_r:user\_home\_t:s0"

user.city="Beijing"

user.name="xattr\_test"

对于不含转义字符 \ 的纯文本属性值，有无双引号限定效果一样。

设置八进制数属性值“\012”，最终以八进制数的base64编码存储。

[root@localhost task1]# **setfattr -n user.age -v "\012" file.txt**

[root@localhost task1]# **getfattr -d -m . file.txt**

# file: file.txt

security.selinux="unconfined\_u:object\_r:user\_home\_t:s0"

user.age=0sCg==

user.city="Beijing"

user.name="xattr\_test"

对于包含转义字符 \ 的文本属性值，无双引号则不对转义符 \ 进行转义；有双引号则对其进行转义。

设置十六进制数属性值，所设置的数的位数必须为偶数，即0x 或 0X后的数字必须为偶数位，否则出错。 若设置成功，最终以十六进制数的base64编码存储。

[root@localhost task1]# **setfattr -n user.hex -v 0x0123 file.txt**

[root@localhost task1]# **getfattr -d -m . file.txt**

# file: file.txt

security.selinux="unconfined\_u:object\_r:user\_home\_t:s0"

user.age=0sCg==

user.city="Beijing"

user.hex=0sASM=

user.name="xattr\_test"

设置base64编码属性值，所设置的编码必须符合base64编码，即0s后的编码字符串必须符合base64编码，否则出错。 若设置成功，最终以base64编码对应的文本信息存储。

[root@localhost task1]# **setfattr -n user.base64 -v 0sSGVsbG8gV29ybGQh file.txt**

[root@localhost task1]# **getfattr -d -m . file.txt**

# file: file.txt

security.selinux="unconfined\_u:object\_r:user\_home\_t:s0"

user.age=0sCg==

user.base64="Hello World!"

user.city="Beijing"

user.hex=0sASM=

user.name="xattr\_test"

Tips：可在<https://base64.us/> 中，将需要设置的属性值进行base64编码后，再使用setfattr命令设置，注意设置时需在base64编码前加0s前缀。

用getfattr编码设置。

保持原编码设置：

[root@localhost task1]# **getfattr -d -m . file.txt**

# file: file.txt

security.selinux="unconfined\_u:object\_r:user\_home\_t:s0"

user.age=0sCg==

user.base64="Hello World!"

user.city="Beijing"

user.hex=0sASM=

user.name="xattr\_test"

对属性设置text编码时，结果如下：

[root@localhost task1]# **getfattr -d -e text file.txt**

# file: file.txt

user.age="\012"

user.base64="Hello World!"

user.city="Beijing"

user.hex="#"

user.name="xattr\_test"

对属性设置hex编码：

[root@localhost task1]# **getfattr -d -e hex file.txt**

# file: file.txt

user.age=0x0a

user.base64=0x48656c6c6f20576f726c6421

user.city=0x4265696a696e67

user.hex=0x0123

user.name=0x78617474725f74657374

user.age 的属性值由八进制变成了十六进制；user.hex 的属性值还原成了最初设置的原值；user.name、user.city、user.base64的属性值都转化为对应的十六进制值。

对属性设置base64编码：

[root@localhost task1]# **getfattr -d -e base64 file.txt**

# file: file.txt

user.age=0sCg==

user.base64=0sSGVsbG8gV29ybGQh

user.city=0sQmVpamluZw==

user.hex=0sASM=

user.name=0seGF0dHJfdGVzdA==

user.age 与user.hex的属性值都保留最初的 base64编码存储；user.name、user.city、user.base64的属性值都转化为对应的base64编码值。

### 注册一个自定义的文件系统类型

#### 相关知识

通常，用户在为自己的系统编译内核时可以把Linux配置为能够识别所有需要的文件系统。但是，文件系统的源代码实际上要么包含在内核映像中，要么作为一个模块被动态装入。VFS必须对目前已在内核中的所有文件系统类型进行跟踪，这是通过进行文件系统类型注册来实现的。

注册函数register\_filesystem/注销函数unregister\_filesystem：

两个函数分别用于文件系统类型的注册和注销，都只有一个参数，即代表文件系统类型的一个指向file\_system\_type对象的指针。

1、头文件：<linux/fs.h>

2、函数原型：两个函数的原型分别为：

int register\_filesystem(struct file\_system\_type \*)；

int unregister\_filesystem(struct file\_system\_type \* )；

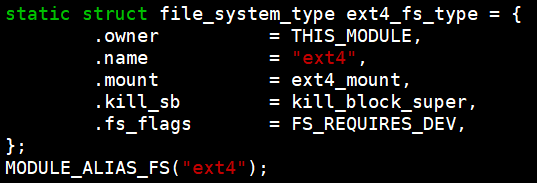
3、返回值：函数执行成功时，返回0；否则，返回一个错误值。

4、参数说明：

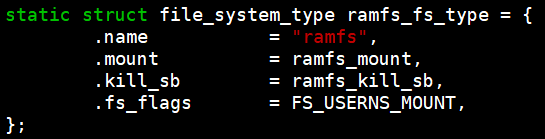
每个注册的文件系统都用一个类型为file\_system\_type的对象来表示。

文件系统对file\_system\_type结构体的填充：

1）ext4fs文件系统对这个结构体的填充（fs/ext4/super.c）：



2）ramfs文件系统对这个结构体的填充（fs/ramfs/inode.c）：



可见，一般文件系统的实现，实现这几个字段即可；其它的由内核利用或者填充。

#### 实验步骤

查看系统中已经注册的文件系统类型

cat /proc/filesystems

正确编写满足功能的源文件，包括.c源文件和Makefile文件。在这里我们的示例源文件存放在tasks\_k/7/task2目录下。

[root@localhost ~]# **cd tasks\_k/7/task2**

[root@localhost task2]# **ls**

Makefile register\_newfs.c

编译源文件。

[root@localhost task2]# **make**

make -C /root/kernel M=/root/tasks\_k/7/task2 modules

make[1]: Entering directory '/root/kernel'

CC [M] /root/tasks\_k/7/task2/register\_newfs.o

Building modules, stage 2.

MODPOST 1 modules

CC /root/tasks\_k/7/task2/register\_newfs.mod.o

LD [M] /root/tasks\_k/7/task2/register\_newfs.ko

make[1]: Leaving directory '/root/kernel'

对比加载内核模块前后的文件系统结果。

[root@localhost task2]# **cat /proc/filesystems | grep myfs**

[root@localhost task2]# **insmod register\_newfs.ko**

[root@localhost task2]# **cat /proc/filesystems | grep myfs**

nodev myfs

卸载内核模块，并查看结果。

[root@localhost task2]# **rmmod register\_newfs**

[root@localhost task2]# **cat /proc/filesystems | grep myfs**

[root@localhost task2]# **dmesg | tail -n2**

[843003.176924] Start register\_newfs module...

[843054.390144] Exit register\_newfs module...

由实验结果可见：

当未加载内核模块时，当前系统中无自定义的文件系统“myfs”；当加载内核模块时，当前系统中可打印出自定义的文件系统“myfs”；当卸载内核模块时，当前系统中无自定义的文件系统“myfs”。

### 在/proc下创建目录

#### 相关知识

一、proc文件系统

Linux系统上的/proc目录是一种文件系统，即proc文件系统。与其它常见的文件系统不同的是，/proc是一种伪文件系统（也即虚拟文件系统），存储的是当前内核运行状态的一系列特殊文件，用户可以通过这些文件查看有关系统硬件及当前正在运行进程的信息，甚至可以通过更改其中某些文件来改变内核的运行状态。

基于/proc文件系统如上所述的特殊性，其内的文件也常被称作虚拟文件，并具有一些独特的特点。例如，其中有些文件虽然使用查看命令查看时会返回大量信息，但文件本身的大小却会显示为0字节。此外，这些特殊文件中大多数文件的时间及日期属性通常为当前系统时间和日期，这跟它们随时会被刷新（存储于RAM中）有关。

为了查看及使用上的方便，这些文件通常会按照相关性进行分类存储于不同的目录甚至子目录中，如/proc/scsi目录中存储的就是当前系统上所有SCSI设备的相关信息，/proc/N中存储的则是系统当前正在运行的进程的相关信息，其中N为正在运行的进程PID（在某进程结束后其相关目录则会消失）。

大多数虚拟文件可以使用文件查看命令如cat、more或者less进行查看，有些文件信息表述的内容可以一目了然，但也有文件的信息却不怎么具有可读性。不过，这些可读性较差的文件在使用一些命令如apm、free、lspci或top查看时却可以有着不错的表现。

二、/proc 目录中的目录操作函数

1、proc\_mkdir()

功能：用于在/proc目录中添加一个目录。

头文件：linux/proc\_fs.h

函数原型为：static inline struct proc\_dir\_entry \*proc\_mkdir(const char \*name, struct proc\_dir\_entry \*parent)

参数说明：

name 建立的目录的名称

parent 父目录指针。如果为NULL，则在 /proc目录下建立。

返回值：成功时返回创建的目录的指针；失败时返回NULL。

2、proc\_dir\_entry结构体

头文件：定义于fs/proc/internal.h，使用时添加linux/proc\_fs.h即可。

结构体对proc文件系统目录项的对象做了定义。分为三个部分：

proc目录项的属性，例如：low\_ino此目录项对应的索引节点的值；mode 对应的索引节点的类型和权限模式；size对应的索引节点的大小；name目录项的名称；

proc目录项的方法，例如：proc\_iops索引节点的操作的集合；proc\_fops文件操作的集合；

linux内核使用的属性，例如in\_use，pde\_unload\_lock等等。

3、proc\_remove()

功能：用于移除 /proc目录下的一个目录及其子目录

函数原型为：static inline void proc\_remove(struct proc\_dir\_entry \*de)

参数为要删除的目录的指针（即proc\_mkdir()函数的返回值）。

#### 实验步骤

正确编写满足功能的源文件，包括.c源文件和Makefile文件。在这里我们的示例源文件存放在tasks\_k/7/task3目录下。

[root@localhost ~]# **cd tasks\_k/7/task3**

[root@localhost task3]# **ls**

Makefile proc\_mkdir.c

编译源文件.

[root@localhost task3]# **make**

make -C /root/kernel M=/root/tasks\_k/7/task3 modules

make[1]: Entering directory '/root/kernel'

CC [M] /root/tasks\_k/7/task3/proc\_mkdir.o

Building modules, stage 2.

MODPOST 1 modules

CC /root/tasks\_k/7/task3/proc\_mkdir.mod.o

LD [M] /root/tasks\_k/7/task3/proc\_mkdir.ko

make[1]: Leaving directory '/root/kernel'

对比加载内核模块前后的文件系统结果.

[root@localhost task3]# **find /proc/ -name myproc**

[root@localhost task3]# **insmod proc\_mkdir.ko**

[root@localhost task3]# **find /proc/ -name myproc**

/proc/myproc

卸载内核模块，并查看结果.

[root@localhost task3]# **rmmod proc\_mkdir**

[root@localhost task3]# **find /proc/ -name myproc**

[root@localhost task3]# **dmesg | tail -n2**

[843158.420990] Start proc\_mkdir module...

[843171.384731] Exit proc\_mkdir module...

由实验结果可见：

当未加载内核模块时，/proc下无myproc目录；当加载内核模块后，/proc下可查找到myproc目录。当卸载内核模块后，/proc下无myproc目录。

### 使用sysfs文件系统传递内核模块参数

#### 相关知识

一、sysfs文件系统

1、基本概念

内核子系统或设备驱动可以直接编译到内核，也可以编译成模块。如果编译到内核，可以通过内核启动参数来向它们传递参数；如果编译成模块，则可以通过命令行在插入模块时传递参数，或者在运行时，通过sysfs来设置或读取模块数据。

sysfs是一个基于内存的虚拟文件系统，可以看成与proc、devfs和devpty同类别的文件系统；它的作用是将内核信息以文件的方式提供给用户程序使用。sysfs文件系统要求总是被挂载在/sys挂载点上，这个文件系统不仅可以把设备（devices）和驱动程序（drivers）的信息从内核输出到用户空间，也可以用来对设备和驱动程序做设置。

sysfs提供一种机制，使得可以显式地描述内核对象、对象属性及对象间关系。sysfs有两组接口，一组针对内核，用于将设备映射到文件系统中；另一组针对用户程序，用于读取或操作这些设备。下表描述了内核中的sysfs要素及其在用户空间的表现：

|  |  |
| --- | --- |
| sysfs在内核中的组成要素 | 在用户空间的显示 |
| 内核对象（kobject） | 目录 |
| 对象属性（attribute） | 文件 |
| 对象关系（relationship） | 链接（Symbolic Link） |

2、sysfs 与 sysctl区别

sysctl：是内核的一些控制参数，其目的是方便用户对内核的行为进行控制；

sysfs：仅仅是把内核的kobject对象的层次关系与属性开放给用户查看，因此sysfs的绝大部分是只读的，模块作为一个kobject也被出口到sysfs，模块参数则是作为模块属性出口的，内核实现者为模块的使用提供了更灵活的方式，允许用户设置模块参数在sysfs的可见性并允许用户在编写模块时设置这些参数在sysfs下的访问权限，然后用户就可以通过sysfs来查看和设置模块参数，从而使得用户能在模块运行时控制模块行为。

二、module\_param宏

对于模块而言，声明为static的变量都可以通过命令行来设置，但要想在sysfs下可见，必须通过宏module\_param来显式声明。

module\_param宏有三个参数：

第一个为参数名，即已经定义的变量名；

第二个参数则为变量类型，可用的类型有byte，short，ushort，int，uint，long，ulong，charp和bool或invbool，分别对应于c语言类型char，short，unsigned short，int，unsigned int，long，unsigned long，char \*和int，用户也可以自定义类型XXX（如果用户自己定义了param\_get\_XXX，param\_set\_XXX和param\_check\_XXX）。

第三个参数用于指定访问权限，如果为0，该参数将不出现在sysfs文件系统中，允许的访问权限为S\_IRUSR，S\_IWUSR，S\_IRGRP，S\_IWGRP，S\_IROTH和 S\_IWOTH的组合，它们分别对应于用户读，用户写，用户组读，用户组写，其他用户读和其他用户写，因此用文件的访问权限设置是一致的。

module\_param宏使用示例：

module\_param(a, int, 0);

MODULE\_PARM\_DESC(a, "An invisible int under sysfs");

module\_param(b, int, S\_IRUSR | S\_IWUSR | S\_IRGRP | S\_IROTH);

MODULE\_PARM\_DESC(b, "An visible int under sysfs");

module\_param(c, charp, S\_IRUSR | S\_IWUSR | S\_IRGRP | S\_IROTH);

MODULE\_PARM\_DESC(c, "An visible string under sysfs");注：MODULE\_PARM\_DESC()是对模块的参数进行描述，使用modinfo查看模块信息时可见。

三、使用示例

当模块加载后，可通过/sys目录与模块进行交互，通常的做法是使用echo向内核传递参数。例如：

echo -n 'example' > /sys/module/module\_exam/parameters/c

说明：用echo传递字符串时，使用-n选项；指定传递参数时，在/sys/module/目录下找到自定义的模块名（如此处示例的module\_exam），模块名目录下的parameters目录，即为该模块的参数列表，选定需要传递的参数名即可（如此处示例的参数c）；必须使用su切换到root模式下，使用sudo命令会出现权限受限的问题。

#### 参考答案

正确编写满足功能的源文件，包括.c源文件和Makefile文件。在这里我们的示例源文件存放在tasks\_k/7/task4目录下。

[root@localhost ~]# **cd tasks\_k/7/task4**

[root@localhost task4]# **ls**

Makefile sysfs\_exam.c

编译源文件。

[root@localhost task4]# **make**

make -C /root/kernel M=/root/tasks\_k/7/task4 modules

make[1]: Entering directory '/root/kernel'

CC [M] /root/tasks\_k/7/task4/sysfs\_exam.o

Building modules, stage 2.

MODPOST 1 modules

CC /root/tasks\_k/7/task4/sysfs\_exam.mod.o

LD [M] /root/tasks\_k/7/task4/sysfs\_exam.ko

make[1]: Leaving directory '/root/kernel'

加载编译完成的内核模块，并查看加载结果。

[root@localhost task4]# **find /sys/module/ -name 'sysfs\_exam'**

[root@localhost task4]# **insmod sysfs\_exam.ko**

[root@localhost task4]# **find /sys/module/ -name 'sysfs\_exam'**

/sys/module/sysfs\_exam

[root@localhost task4]# **modinfo sysfs\_exam.ko**

filename: /home/Root/tasks\_k/7/task4/sysfs\_exam.ko

license: GPL

srcversion: A24FE62F33A5A6268FDE97A

depends:

name: sysfs\_exam

vermagic: 4.19.162 SMP mod\_unload modversions aarch64

parm: a:An invisible int under sysfs (int)

parm: b:An visible int under sysfs (int)

parm: c:An visible string under sysfs (charp)

[root@localhost task4]# **dmesg | tail -n4**

[843265.590489] Start sysfs\_exam module...

[843265.590491] a = 0

[843265.590491] b = 0

[843265.590492] c = 'Hello, World'

[root@localhost task4]# **ls -al /sys/module/sysfs\_exam/parameters/**

total 0

drwxr-xr-x. 2 root root 0 Dec 20 16:44 .

drwxr-xr-x. 6 root root 0 Dec 20 16:44 ..

-rw-r--r--. 1 root root 65536 Dec 20 16:46 b

-rw-r--r--. 1 root root 65536 Dec 20 16:46 c

[root@localhost task4]# **cat /sys/module/sysfs\_exam/parameters/c**

Hello, World

[root@localhost task4]# **echo -n '*Happy new year!*' > /sys/module/sysfs\_exam//parameters/c**

[root@localhost task4]# **cat /sys/module/sysfs\_exam/parameters/c**

Happy new year!

卸载内核模块，并查看结果。

[root@localhost task4]# **rmmod sysfs\_exam**

[root@localhost task4]# **dmesg | tail -n4**

[843429.858532] Exit sysfs\_exam module...

[843429.858533] a = 0

[843429.858534] b = 0

[843429.858535] c = 'Happy new year!'

由实验结果可见：

（1）当未加载内核模块时，当前系统中无自定义的内核模块“sysfs\_exam”;

当加载内核模块后，当前系统可见自定义的文件系统“sysfs\_exam”;

（2）使用modinfo查看内核模块信息，可见a、b、c三个参数的描述：

parm: a:An invisible int under sysfs (int)

parm: b:An visible int under sysfs (int)

parm: c:An visible string under sysfs (charp)

即参数a在/sys中不可见，b、c在/sys中可见，ls -al /sys/module/sysfs\_exam/parameters/的输出列表可验证；

（3）可使用echo向模块传递参数值来改变指定参数的值。

# 实验八 网络管理

## 实验介绍

本实验通过编写C源码程序实现客户端与服务端的简单通信，使用tshark抓取该通信数据包以及查看使用setsockopt发送的带IP记录路由选项的数据包中是否包含了记录路由选项等任务操作，让学生们了解并掌握操作系统中的网络管理。

### 任务描述

* 编写C源码，基于socket的UDP发送接收程序，实现客户端与服务端的简单通信。客户端从命令行输入中读取要发送的内容，服务端接收后实时显示。
* 基于任务1的服务端与客户端程序运行时，使用tshark抓取该通信数据包。
* 基于任务1的客户端与服务端，使用setsockopt发送一个带IP记录路由选项的数据包；使用tshark查看发送的数据包中是否包含了记录路由选项。

## 实验目的

* 正确编写满足功能的源文件，正确编译。
* 正常加载、卸载内核模块；且内核模块功能满足任务所述。
* 了解操作系统的网络管理。

## 实验任务

### 编写基于socket的UDP发送接收程序

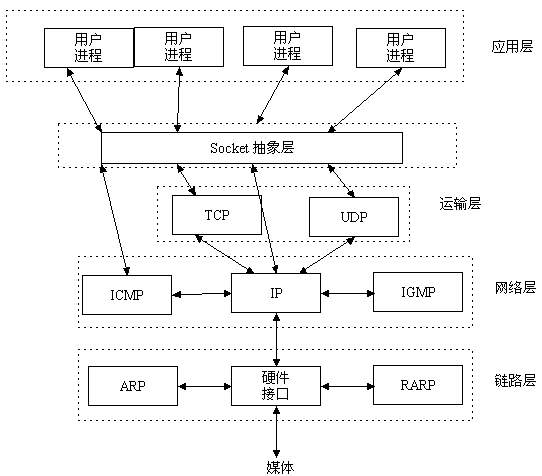
#### 相关知识

一、socket是做什么的？

socket起源于Unix，而Unix/Linux基本哲学之一就是“一切皆文件”，都可以用“打开open–>读写write/read–>关闭close”模式来操作。Socket就是该模式的一个实现，socket即是一种特殊的文件，一些socket函数就是对其进行的操作（读/写IO、打开、关闭）。

简言之，socket是应用层与TCP/IP协议族通信的中间软件抽象层，它是一组接口。在设计模式中，socket其实就是一个门面模式，它把复杂的TCP/IP协议族隐藏在socket接口后面，对用户来说，一组简单的接口就是全部，让socket去组织数据，以符合指定的协议。

TCP/IP协议族包括运输层、网络层、链路层，而socket所在位置如图，Socket是应用层与TCP/IP协议族通信的中间软件抽象层。

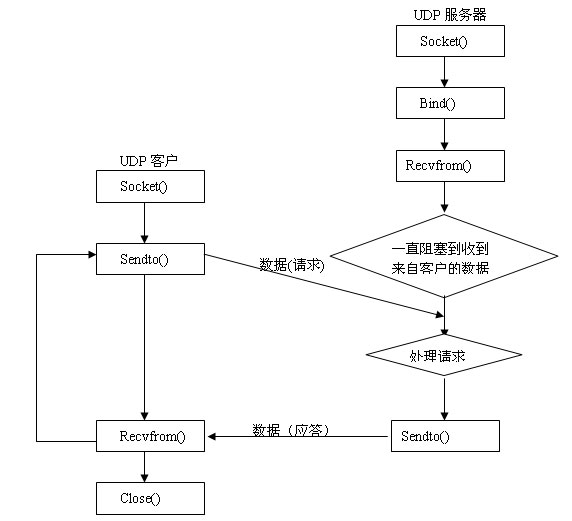


二、UDP

在TCP/IP模型中，UDP为网络层以上和应用层以下提供了一个简单的接口。UDP只提供数据的不可靠传递，它一旦把应用程序发给网络层的数据发送出去，就不保留数据备份（所以UDP有时候也被认为是不可靠的数据报协议）。UDP在IP数据报的头部仅仅加入了复用和数据校验（字段）。

UDP首部字段由4个部分组成，其中两个是可选的。各16bit的来源端口和目的端口用来标记发送和接受的应用进程。因为UDP不需要应答，所以来源端口是可选的，如果来源端口不用，那么置为零。在目的端口后面是长度固定的以字节为单位的长度域，用来指定UDP数据报包括数据部分的长度，长度最小值为8byte。首部剩下的16bit是用来对首部和数据部分一起做校验和（Checksum）的，这部分是可选的，但在实际应用中一般都使用这一功能。

由于缺乏可靠性且属于非连接导向协定，UDP应用一般必须允许一定量的丢包、出错和复制粘贴。但有些应用，比如TFTP，如果需要则必须在应用层增加根本的可靠机制。但是绝大多数UDP应用都不需要可靠机制，甚至可能因为引入可靠机制而降低性能。流媒体（串流技术）、即时多媒体游戏和IP电话（VoIP）一定就是典型的UDP应用。如果某个应用需要很高的可靠性，那么可以用传输控制协议（TCP协议）来代替UDP。



三、socket编程API

1、int socket(int domain, int type, int protocol);

（1）功能：根据指定的地址族、数据类型和协议来分配一个socket的描述字及其所用的资源。

（2）头文件（C库）：

#include <sys/types.h>

#include<sys/socket.h>

（3）参数说明：

domain：地址族，常用的有：AF\_INET、AF\_INET6、AF\_LOCAL、AF\_ROUTE，其中AF\_INET代表使用ipv4地址。

type：socket类型，常用的socket类型有：SOCK\_STREAM、SOCK\_DGRAM、SOCK\_RAW、SOCK\_PACKET、SOCK\_SEQPACKET等。SOCK\_STREAM----提供有序的、可靠的、双向的和基于连接的字节流，使用带外数据传送机制，为Internet地址族使用TCP。SOCK\_DGRAM----支持无连接的、不可靠的和使用固定大小（通常很小）缓冲区的数据报服务，为Internet地址族使用UDP。

protocol：协议。常用的协议有：IPPROTO\_IP、IPPROTO\_TCP、IPPTOTO\_UDP、IPPROTO\_SCTP、IPPROTO\_TIPC等。

（4）返回值：调用成功则返回新创建的套接字描述符；失败就返回INVALID\_SOCKET。

2、int sendto(int sockfd, const void\* msg, int len, int flags, const struct sockaddr \*to, int tolen);

（1）功能：进行无连接的UDP通讯使用，使用时，数据会在没有建立任何网络连接的网络上传输。

（2）头文件（C库）：

#include <sys/types.h>

#include<sys/socket.h>

（3）参数说明：

sockfd：指与远程程序连接的套接字，即socket()函数的返回值。

msg：是一个指针，指向发送的信息的地址。

len：指发送信息的长度。

flags：通常是0。

to：一个指向 struct sockaddr 结构的指针，里面包含了远程主机和端口数据。

tolen ：指出了struct sockaddr 的大小，通常用 sizeof(struct sockaddr)。

（4）返回值：正常时返回真正发送的数据的大小；错误时：返回-1。

3、int bind( int sockfd, struct sockaddr\* addr, socklen\_t addrlen)

（1）功能：将指定地址与指定套接口绑定。

（2）头文件（C库）：

#include <sys/types.h>

#include<sys/socket.h>

（3）参数说明：

sockfd：指定地址与哪个套接字绑定，即socket()函数调用返回的套接字。调用bind的函数之后，该套接字与一个相应的地址关联，发送到这个地址的数据可以通过这个套接字来读取与使用。

addr：指定地址。这是一个地址结构，并且是一个已经经过填写的有效的地址结构。调用bind之后这个地址与参数sockfd指定的套接字关联，从而实现上面所说的效果。

addrlen：地址的长度。正如大多数socket接口一样，内核不关心地址结构，当它复制或传递地址给驱动的时候，它依据这个值来确定需要复制多少数据。这已经成为socket接口中最常见的参数之一了。

bind函数并不是总是需要调用的，只有用户进程想与一个具体的地址或端口相关联的时候才需要调用这个函数。如果用户进程没有这个需要，那么程序可以依赖内核的自动的选址机制来完成自动地址选择，而不需要调用bind的函数。

（4）返回值：0 ──成功，-1 ──失败。

4、int recvfrom(int s,void \*buf,int len,unsigned int flags ,struct sockaddr \*from ,int \*fromlen);

（1）功能：接收远程主机经指定的socket 传来的数据，并把数据存到由参数buf 指向的内存空间。

（2）头文件（C库）：

#include<sys/types.h>

#include<sys/socket.h>

（3）参数说明：

s：表示正在监听的端口的套接字，即函数socket()的返回值；

buff：表示接收数据缓冲区，接收到的数据将放在这个指针所指向的内存空间中；

len：表示接收数据缓冲区大小/可接收数据的最大长度，系统根据这个值来确保接收缓冲区的安全，防止溢出；

flags：一般设0；

from：是一个struct sockaddr类型的变量，该变量保存发送方的IP地址及端口号；

fromlen：表示sockaddr的结构长度，可以使用sizeof(struct sockaddr\_in)来获得。

（4）返回值：成功则返回接收到的字符数；失败则返回-1，错误原因存于errno中。

错误代码：

EBADF 参数s非合法的socket处理代码

EFAULT 参数中有一指针指向无法存取的内存空间。

ENOTSOCK 参数s为一文件描述词，非socket。

EINTR 被信号所中断。

EAGAIN 此动作会令进程阻断，但参数s的socket为不可阻断。

ENOBUFS 系统的缓冲内存不足

ENOMEM 核心内存不足

EINVAL 传给系统调用的参数不正确。

#### 实验步骤

正确编写满足功能的源文件，包括服务端源文件和客户端源文件。在这里我们的示例源文件存放在tasks\_k/8/task1目录下。

[root@localhost ~]# **cd tasks\_k/8/task1**

[root@localhost task1]# **ls**

client.c server.c

编译源文件。

[root@localhost task1]# **gcc server.c -o server**

[root@localhost task1]# **gcc client.c -o client**

[root@localhost task1]# **ls**

client client.c server server.c

开启两个终端，一个运行客户端，一个运行服务端；client中输入发送的消息回车后，server端即能收到。

Client端

[root@localhost task1]# **./client**

Server端

[root@localhost task1]# **./server**

Client端

Please enter the content to be sent:

***Hello world!***

Please enter the content to be sent:

***Happy new year!***

Please enter the content to be sent:

Server端

The message received is: *Hello world!*

The message received is: *Happy new year!*

最后分别在Client端和Server端用Ctrl+C退出。

### 使用tshark抓包

#### 相关知识

tshark是wireshark中提供的Linux命令行工具，tshark不仅有抓包的功能，还带了解析各种协议的能力。由于wireshark的运行需要GUI（图形用户界面，Graphical User Interface）的环境，但是openEuler系统的缺省安装是不含GUI的，此时就可以使用tshark进行抓包。

#### 实验步骤

安装wireshark。

dnf install -y wireshark

用ifconfig查看网卡信息。

[root@localhost task2]# ifconfig

……

lo: flags=73<UP,LOOPBACK,RUNNING> mtu 65536

inet 127.0.0.1 netmask 255.0.0.0

inet6 ::1 prefixlen 128 scopeid 0x10<host>

loop txqueuelen 1000 (Local Loopback)

RX packets 38 bytes 12395 (12.1 KiB)

RX errors 0 dropped 0 overruns 0 frame 0

TX packets 38 bytes 12395 (12.1 KiB)

TX errors 0 dropped 0 overruns 0 carrier 0 collisions 0

……

由于任务1的client与server均运行于本机，因此指定的网络接口应该是回环地址接口lo。

开启三个终端，一个运行客户端，一个运行服务端，一个运行tshark；client中输入发送的消息回车后，server端即能收到。

Client端

[root@localhost task1]# **./client**

Please enter the content to be sent:

***Hello world!***

Please enter the content to be sent:

***Happy new year!***

Please enter the content to be sent:

Server端

[root@localhost task1]# **./server**

The message received is: *Hello world!*

The message received is: *Happy new year!*

tshark端，直接把抓包结果输出到命令行

[root@localhost ~]# **tshark -i lo -n -f 'udp port 40000'**

Running as user "root" and group "root". This could be dangerous.

Capturing on 'Loopback'

1 0.000000000 127.0.0.1 → 127.0.0.1 UDP 55 46360 → 40000 Len=13

2 7.174615157 127.0.0.1 → 127.0.0.1 UDP 58 46360 → 40000 Len=16

### 使用setsockopt发送记录路由选项

#### 相关知识

getsockopt/setsockopt系统调用

1、功能描述：

获取或者设置与某个套接字关联的选项。选项可能存在于多层协议中，它们总会出现在最上面的套接字层。当操作套接字选项时，选项位于的层和选项的名称必须给出。为了操作套接字层的选项，应该 将层的值指定为SOL\_SOCKET。为了操作其它层的选项，控制选项的合适协议号必须给出。例如，为了表示一个选项由IP协议解析，层应该设定为协议号IPPROTO\_IP。

2、用法：

//头文件

#include <sys/types.h>

#include <sys/socket.h>

//函数原型

int getsockopt(int sock, int level, int optname, void \*optval, socklen\_t \*optlen);

int setsockopt(int sock, int level, int optname, const void \*optval, socklen\_t optlen);

3、参数说明：

sock (套接字)：将要被设置或者获取选项的套接字。

level (级别)：指定选项代码的类型/选项所在的协议层。支持：  
SOL\_SOCKET：通用套接字选项  
IPPROTO\_IP：IP套接字选项  
IPPROTO\_TCP：TCP套接字选项

optname (选项名)：需要访问的选项名。

optval(选项值)：是一个指向变量的指针，对于getsockopt()，指向返回选项值的缓冲；对于setsockopt()，指向包含新选项值的缓冲。类型：整型；支持套接口结构或其他结构类型，如：linger{}, timeval{ }。

optlen(选项长度) ：optval 的大小。对于getsockopt()，作为入口参数时，选项值的最大长度；作为出口参数时，选项值的实际长度。对于setsockopt()，现选项的长度。

4、返回值说明：成功执行时，返回0。失败返回-1，errno被设为以下的某个值：

EBADF：sock不是有效的文件描述词

EFAULT：optval指向的内存并非有效的进程空间

EINVAL：在调用setsockopt()时，optlen无效

ENOPROTOOPT：指定的协议层不能识别选项

ENOTSOCK：sock描述的不是套接字

#### 实验步骤

正确编写满足功能的源文件，包括服务端源文件和客户端源文件（服务端同任务1，客户端添加了记录路由代码）。在这里我们的示例源文件存放在tasks\_k/8/task3目录下。

[root@localhost ~]# **cd tasks\_k/8/task3**

[root@localhost task3]# **ls**

client.c server.c setsockopt.xml

编译源文件。

[root@localhost task3]# **gcc server.c -o server**

[root@localhost task3]# **gcc client.c -o client**

[root@localhost task3]# **ls**

client client.c server server.c

开启三个终端，一个运行客户端，一个运行服务端，一个运行tshark。

Client端

[root@localhost task3]# **./client**

Please enter the content to be sent:

***Hello!***

Please enter the content to be sent:

***How are you?***

Please enter the content to be sent:

***Nice to meet you!***

Server端

[root@localhost task3]# **./server**

The message received is: *Hello!*

The message received is: *How are you?*

The message received is: *Nice to meet you!*

tshark端，直接把抓包结果输出到命令行

[root@localhost task3]# **tshark -i lo -n -f 'udp port 40000' -T pdml > ./setsockopt.xml**

Running as user "root" and group "root". This could be dangerous.

Capturing on 'Loopback'

3

最后，分别在各自的命令窗口按Ctrl+C键退出各自的程序。

思考题：请在重定向输出的setsockopt.xml文件中，查看记录下来的路由内容。client与server的每一次通信的数据包中都有路由记录吗？

提示：client与server的每一次通信的数据包中，都有路由记录。client与server有3次通信，则setsockopt.xml文件中会有3段记录路由的内容。可以在文档中搜索“*Recorded Route: 127.0.0.1*”找到相应的记录。

# 内核虚拟化

## 实验介绍

### 相关知识

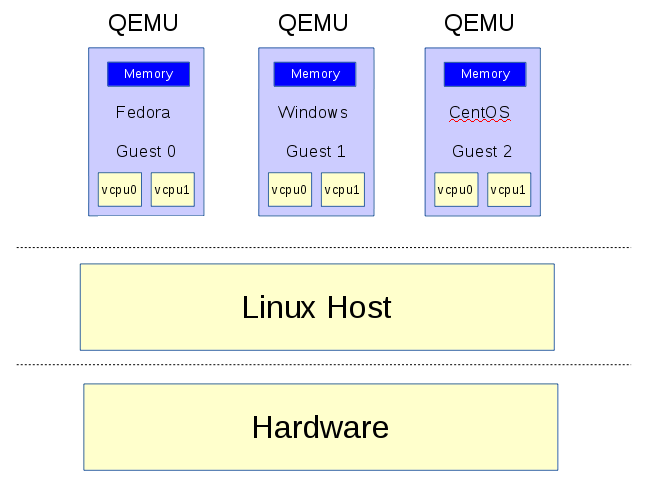
openEuler提供了支持AArch64和x86\_64处理器架构的KVM虚拟化组件。

**一、QEMU、KVM、libvirt**

1、QEMU

QEMU是一个模拟器，它向Guest OS模拟CPU和其他硬件，Guest OS认为自己和硬件直接打交道，其实是同QEMU模拟出来的硬件打交道，QEMU将这些指令转译给真正的硬件。由于所有的指令都要从QEMU里面过一手，因而性能较差。

QEMU作为系统模拟器，可以模拟出一台能够独立运行操作系统的虚拟机。如下图所示，每个虚拟机对应主机(Host)中的一个QEMU进程，而虚拟机的vCPU对应QEMU进程的一个线程。



QEMU的安装：

用户可以通过不同Linux发行版所带有的软件包管理器来安装QEMU。

如：

* 在Debian系列的发行版上可以使用下面的命令来安装：sudo apt-get install qemu
* 在红帽系列的发行版上使用如下命令安装：sudo yum install qemu -y
* 除此之外，也可以选择下载源码编译安装，例如：

# wget https://download.qemu.org/qemu-4.1.1.tar.xz

# sudo tar -xvf qemu-4.1.1.tar.xz

# cd qemu-4.1.1

# sudo ./configure --target-list=aarch64-softmmu //必须加上--target-list=aarch64-softmmu，否则无法编译出qemu-system-aarch64；configure脚本用于生成Makefile，具体选项可以用./configure --help查看

# sudo make -j8

# sudo make install

2、KVM

KVM是Linux的内核模块，它需要CPU的支持，采用硬件辅助虚拟化技术。英特尔虚拟化技术（Intel Virtualization Technology, IVT）和AMD虚拟化（AMD Virtualization, AMD-V）是硬件辅助虚拟化技术在CISC架构上的实现。鲲鹏处理器基于ARMv8-A架构，也支持硬件辅助虚拟化。KVM通过/dev/kvm暴露接口，可通过`ls /dev/kvm`或 `ls /sys/module/kvm` 查看是否存在相应目录/文件，来确定内核是否支持KVM虚拟化。若上述文件/目录不存在，则说明系统内核编译时未开启KVM虚拟化，需要更换支持KVM虚拟化的Linux内核。

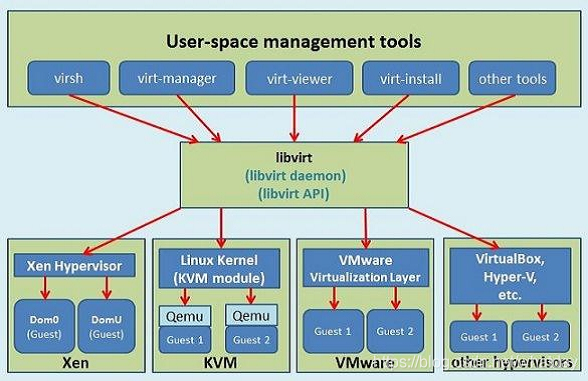
KVM内核模块本身只能提供CPU和内存的虚拟化，所以它必须结合QEMU才能构成一个完成的虚拟化技术，即qemu-kvm。

qemu-kvm：QEMU将KVM整合进来，通过ioctl调用/dev/kvm接口，将有关CPU指令的部分交由内核模块来做。KVM负责cpu虚拟化+内存虚拟化，实现了cpu和内存的虚拟化，但KVM不能模拟其他设备；而QEMU模拟IO设备（网卡，磁盘等），KVM加上QEMU之后就能实现真正意义上服务器虚拟化。因为用到了上面两个东西，所以称之为qemu-kvm。

现在的QEMU已经整合qemu-kvm，不再有qemu-kvm的说法了。一般创建x86虚拟机需要用到qemu-system-x86\_64命令、创建arm64虚拟机要用qemu-system-aarch64命令，并需要加上 --enable-kvm来支持KVM加速。有些Linux发行版的qemu-kvm命令仅仅是qemu-system-x86\_64的软链接或者简单包装。

3、libvirt

libvirt是目前使用最为广泛的对KVM虚拟机进行管理的工具和API。libvirtd是其daemon进程，可以被本地的virsh调用，也可以被远程的virsh调用，libvirtd调用qemu-kvm操作虚拟机。



Libvirt工具采用XML格式的文件描述一个虚拟机特征，包括虚拟机名称、CPU、内存、磁盘、网卡、鼠标、键盘等信息。用户可以通过修改配置文件，对虚拟机进行管理。

**二、参考资料**

1. [openEuler官方网站“虚拟化用户指南”](https://openeuler.org/zh/docs/20.03_LTS/docs/Virtualization/virtualization.html)

## 实验目的

熟悉在Taishan物理服务器上安装QEMU虚拟机的过程。

## 实验任务

### 在Taishan服务器中搭建QEMU虚拟机运行环境

以下步骤在Taishan服务器中，使用libvirt+xml配置文件的方法，搭建openEuler-20.03-aarch64系统的QEMU虚拟机运行环境。

首先安装虚拟化核心组件。

安装libvirt和QEMU

[root@localhost ~]# **yum install -y libvirt qemu**

验证是否安装成功

[root@localhost ~]# **rpm -qi libvirt**

Name : libvirt

Version : 6.2.0

Release : 8.oe1

Architecture: aarch64

Install Date: Fri 11 Dec 2020 10:44:46 PM CST

Group : Unspecified

Size : 0

License : LGPLv2+

Signature : RSA/SHA1, Mon 02 Nov 2020 05:06:20 PM CST, Key ID d557065eb25e7f66

Source RPM : libvirt-6.2.0-8.oe1.src.rpm

Build Date : Mon 02 Nov 2020 05:00:38 PM CST

Build Host : obs-worker-005

Packager : http://openeuler.org

Vendor : http://openeuler.org

URL : https://libvirt.org/

Summary : Library providing a simple virtualization API

Description :

Libvirt is a C toolkit to interact with the virtualization capabilities

of recent versions of Linux (and other OSes). The main package includes

the libvirtd server exporting the virtualization support.

[root@localhost ~]# **rpm -qi qemu**

Name : qemu

Epoch : 2

Version : 4.1.0

Release : 17.oe1

Architecture: aarch64

Install Date: Thu 24 Sep 2020 03:34:16 PM CST

Group : Unspecified

Size : 19730898

License : GPLv2 and BSD and MIT and CC-BY

Signature : RSA/SHA1, Mon 24 Aug 2020 06:02:00 PM CST, Key ID d557065eb25e7f66

Source RPM : qemu-4.1.0-17.oe1.src.rpm

Build Date : Mon 24 Aug 2020 05:59:38 PM CST

Build Host : obs-worker-008

Packager : http://openeuler.org

Vendor : http://openeuler.org

URL : http://www.qemu.org

Summary : QEMU is a generic and open source machine emulator and virtualizer

Description :

QEMU is a FAST! processor emulator using dynamic translation to achieve good emulation speed.

QEMU has two operating modes:

Full system emulation. In this mode, QEMU emulates a full system (for example a PC),

including one or several processors and various peripherals. It can be used to launch

different Operating Systems without rebooting the PC or to debug system code.

User mode emulation. In this mode, QEMU can launch processes compiled for one CPU on another CPU.

It can be used to launch the Wine Windows API emulator (https://www.winehq.org) or to ease

cross-compilation and cross-debugging.

You can refer to https://www.qemu.org for more infortmation.

若成功则可以看到libvirt、QEMU软件包的信息，如果它们已经安装则可以略过这两步。

查看内核是否支持KVM虚拟化

[root@localhost ~]# **ls /dev/kvm**

/dev/kvm

[root@localhost ~]# **ls /sys/module/kvm**

parameters uevent

上述文件存在即表示是支持KVM虚拟化的，这样我们可以进行后续的步骤。

启动libvirtd服务

[root@localhost ~]# **systemctl start libvirtd**

[root@localhost ~]# **systemctl status libvirtd**

● libvirtd.service - Virtualization daemon

Loaded: loaded (/usr/lib/systemd/system/libvirtd.service; enabled; vendor preset: enabled)

Active: active (running) since Wed 2020-10-28 10:45:50 CST; 4 days ago

若服务处于“Active”状态，说明服务启动成功。

下面步骤准备虚拟机镜像。

以root用户安装qemu-img软件包

[root@localhost ~]# **yum install qemu-img**

若该软件包已经安装请忽略这一步。

创建镜像文件

[root@localhost home]# **df -H**

Filesystem Size Used Avail Use% Mounted on

devtmpfs 549G 0 549G 0% /dev

tmpfs 549G 0 549G 0% /dev/shm

tmpfs 549G 58M 549G 1% /run

tmpfs 549G 0 549G 0% /sys/fs/cgroup

/dev/mapper/openeuler-root 53G 16G 35G 31% /

tmpfs 549G 66k 549G 1% /tmp

/dev/sda2 1.1G 301M 653M 32% /boot

/dev/sda1 210M 6.0M 204M 3% /boot/efi

/dev/mapper/openeuler-home 3.8T 638M 3.6T 1% /home

tmpfs 110G 0 110G 0% /run/user/0

[root@localhost ~]# **cd /home**

在这里，由于/home目录的空间比较大，创建虚拟机转到/home目录下进行。

[root@localhost home]# **qemu-img create -f qcow2 *openEuler-image01.qcow2* 4G**

这里***灰色斜体字***表示根据实际情况命名文件名。

我们还可以修改镜像磁盘空间大小：

[root@localhost home]# **qemu-img resize *openEuler-image01.qcow2* +20G**

查看镜像信息：

[root@localhost home]# **qemu-img info *openEuler-image01.qcow2***

接着准备虚拟机网络，我们这里以Linux网桥来准备虚拟机网络。

安装bridge-utils软件包

[root@localhost home]# **yum install bridge-utils**

若该软件工具已经被安装，请忽略此步骤。

创建网桥br0

[root@localhost home]# **brctl addbr br0**

将物理网卡绑定到br0

[root@localhost home]# **brctl addif br0 *enp125s0f3***

这里***enp125s0f3***是泰山服务器192.168.101.X网段的网口，将br0绑定到它是为了处于该网段的PC用VNC客户端和ssh客户端链接虚拟机用。

注意：这一步执行后可能导致PC终端到Taishan服务器的ssh连接中断，依赖于您的网络环境。

清除与网桥相连网口的IP地址

网口和网桥链接后不再需要IP地址，将其IP地址设置为0.0.0.0：

[root@localhost home]# **ifconfig *enp125s0f3* 0.0.0.0**

注意：以上命令会影响到路由表，如果在执行这条命令后PC终端到Taishan服务器的ssh连接终端，您需要重新构建ssh终端到Taishan服务器的路由并为br0配置IP地址。

设置br0的IP地址

[root@localhost home]# **dhclient br0**

这里是通过DHCP服务器给br0配置IP地址。如果在该网段上没有DHCP服务，则可用ifconfig命令配置静态IP地址，如：

# ifconfig br0 192.168.101.252 netmask 255.255.255.0

配置网桥br1

[root@localhost home]# **brctl addbr br1**

[root@localhost home]# **brctl addif br1 *enp125s0f1***

[root@localhost home]# **ifconfig *enp125s0f1* 0.0.0.0**

[root@localhost home]# **ifconfig br1 *192.168.137.254* netmask 255.255.255.0**

上述命令创建网桥br1，并将br1绑定到***enp125s0f1***，在本文中这是泰山服务器192.168.137.X网段的网口，绑定到它上面是为了在虚拟机内部能访问公网。由于dhclient已被br0占用，所以这里指定IP地址***192.168.137.254***到br1（注意不要引起IP地址冲突）。

现在准备引导固件。openEuler默认已安装BIOS启动对应的引导文件，不需要用户额外操作，此处使用UEFI方式引导，需要安装工具集EDK II工具集。AArch64架构对应的安装包为edk2-aarch64。

安装edk软件包

[root@localhost home]# **yum install edk2-aarch64**

查看edk软件包是否安装成功

[root@localhost home]# **rpm -qi edk2-aarch64**

Name : edk2-aarch64

Version : 202002

Release : 1.oe1

Architecture: noarch

Install Date: Thu 24 Sep 2020 04:06:22 PM CST

Group : Unspecified

Size : 137111513

License : BSD-2-Clause-Patent

Signature : RSA/SHA1, Wed 03 Jun 2020 05:19:46 PM CST, Key ID d557065eb25e7f66

Source RPM : edk2-202002-1.oe1.src.rpm

Build Date : Wed 03 Jun 2020 05:14:13 PM CST

Build Host : obs-worker-009

Packager : http://openeuler.org

Vendor : http://openeuler.org

URL : https://github.com/tianocore/edk2

Summary : AARCH64 Virtual Machine Firmware

Description :

EFI Development Kit II AARCH64 UEFI Firmware

安装完成后，/usr/share/edk2/aarch64/目录下就有所需的引导固件。若该软件包已经安装，则忽略以上步骤。

现在开始编写虚拟机xml配置文件。可按如下文件内容准备虚拟机的xml配置文件（可以使用vim命令编辑）：



关键配置如以下**蓝色粗体字**所示：

<domain type='kvm'>

<name>**openEulerVM01**</name>

<memory unit='GiB'>**8**</memory>

<vcpu>**4**</vcpu>

<os>

<type arch='aarch64' machine='virt'>hvm</type>

<loader readonly='yes' type='pflash'>/usr/share/edk2/aarch64/QEMU\_EFI-pflash.raw</loader>

</os>

<features>

<acpi/>

<gic version='3'/>

</features>

<cpu mode='host-passthrough'>

<topology sockets='2' cores='2' threads='1'/>

</cpu>

<iothreads>1</iothreads>

<clock offset='utc'/>

<on\_poweroff>destroy</on\_poweroff>

<on\_reboot>restart</on\_reboot>

<on\_crash>restart</on\_crash>

<devices>

<emulator>/usr/libexec/qemu-kvm</emulator>

<disk type='file' device='disk'>

<driver name='qemu' type='qcow2' iothread="1"/>

<source file=**'/home/openEuler-image01.qcow2**'/>

<target dev='vda' bus='virtio'/>

<boot order='1'/>

</disk>

<disk type='file' device='cdrom'>

<driver name='qemu' type='raw'/>

<source file='**/home/openEuler-20.03-LTS-aarch64-dvd.iso**'/>

<readonly/>

<target dev='sdb' bus='scsi'/>

<boot order='2'/>

</disk>

<interface type='bridge'>

<source bridge='br0'/>

<model type='virtio'/>

</interface>

<interface type='bridge'>

<source bridge='br1'/>

<model type='virtio'/>

</interface>

<console type='pty'/>

<video>

<model type='virtio'/>

</video>

<controller type='scsi' index='0' model='virtio-scsi'/>

<controller type='usb' model='ehci'/>

<input type='tablet' bus='usb'/>

<input type='keyboard' bus='usb'/>

<graphics type='vnc' listen='0.0.0.0' passwd='**Euler001**'/>

</devices>

<seclabel type='dynamic' model='dac' relabel='yes'/>

</domain>

需要注意以下几个地方：

1. 应先使用“free –g”查看树莓派宿主机的内存余量，再定义虚拟机的内存大小。若定义的内存大小，大于宿主机可使用的内存余量，则报错。

2. 注意openEuler-image01.qcow2和openEuler-20.03-LTS-aarch64-dvd.iso文件路径要设置正确，建议使用绝对路径。

3. 在graphics标签中指定了登录VNC Server时的密码。

4. 如果要批量定义虚拟机，则：(1) 虚拟机的名字为openEulerVM01、openEulerVM02、openEulerVM03等；(2) 对应的镜像文件为openEuler-image01.qcow2、openEuler-image02.qcow2、openEuler-image03.qcow2等；(3) 对应的xml文件为openEulerVM01.xml、openEulerVM02.xml、openEulerVM03.xml等。另外ISO文件最好每个虚拟机对应一份拷贝，这样安装操作系统时不至于多个虚拟机读同一个安装文件而影响安装速度。

现在，可以使用Virsh命令管理虚拟机了。

修改配置文件

[root@localhost home]# **vi /etc/libvirt/qemu.conf**

取消该配置文件中以下两处注释：

user = "root"

group = "root"

否则可能因为权限不够而报错。

创建虚拟机

[root@localhost home]# **virsh define *openEulerVM01.xml***

定义虚拟机后，可以用如下命令查看：

[root@localhost home]# **virsh list --all**

Id Name State

--------------------------------

- openEulerVM01 shut off

还可使用如下命令修改：

# **virsh edit *openEulerVM01***

启动虚拟机

[root@localhost home]# **virsh start *openEulerVM01***

这时可以看到一个和NVRAM有关的文件生成了：

[root@localhost home]# **ls /var/lib/libvirt/qemu/nvram/**

openEulerVM01\_VARS.fd

启动后若要关闭虚拟机：

# **virsh destroy *openEulerVM01***

若要删除虚拟机：

# **virsh undefine *openEulerVM01* --nvram**

注意：若没有用到nvram，则不需要添加—nvram参数。

再次查看虚拟机信息

[root@localhost home]# **virsh list**

Id Name State

-------------------------------

1 openEulerVM01 running

[root@localhost home]# **virsh domiflist *openEulerVM01***

Interface Type Source Model MAC

-----------------------------------------------------------

vnet0 bridge br0 virtio 52:54:00:4a:04:6e

vnet1 bridge br1 virtio 52:54:00:2d:f0:71

更多命令请参考“virsh --help”命令。

### 使用VNC登录虚拟机

以下步骤分别在Taishan服务器和Windows电脑终端中安装相应的VNC工具，并以此登录虚拟机并完成操作系统的安装。

查看虚拟机使用的端口号

[root@localhost home]# **virsh vncdisplay *openEulerVM01***

:0

在宿主机上安装TigerVNC

[root@localhost home]# **yum install tigervnc tigervnc-server**

启动VNC Server

先关闭防护墙：

[root@localhost home]# **systemctl stop firewalld**

注意：由于是学生实验，所以可以简单关闭防火墙。若其他安全性比较高的场景，请参照如下命令：

# firewall-cmd --zone=public --add-port=*5900*/tcp

其中5900端口是因为用“virsh vncdisplay”命令查看时端口号为0，若为1则为5901，以此类推。

若后续步骤通过TigerVNC客户端登录不了虚拟机，需要启动并配置VNC Server：

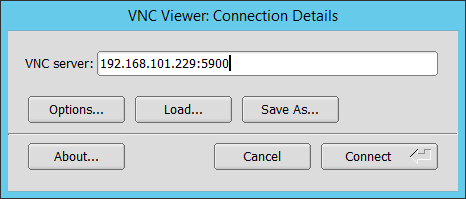
[root@localhost home]# **vncserver**

第一次启动时，VNC Server会创建其初始环境文件和用户密码文件（有管理员账户和普通账户之分，有关VNC的详细使用方法请查询相关文档，此处不作赘述）。

在另一台电脑，即Windows终端上安装TigerVNC客户端



在Windows终端打开TigerVNC Viewer进行登录

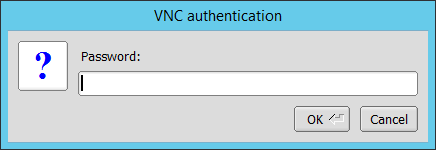


注意这里*192.168.101.229*为宿主机IP地址，*5900*是因为虚拟机运行端口为0。

如果使用两位端口号格式，则为：*192.168.101.229: 00*。

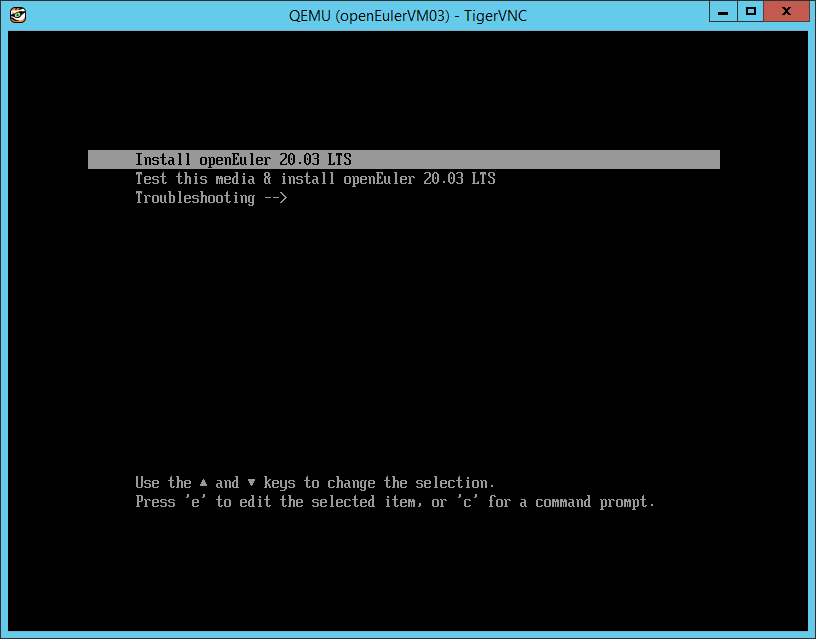
如果端口号为1、2……，则相应的vncserver的启动命令为vncserver :1、vncserver :2……

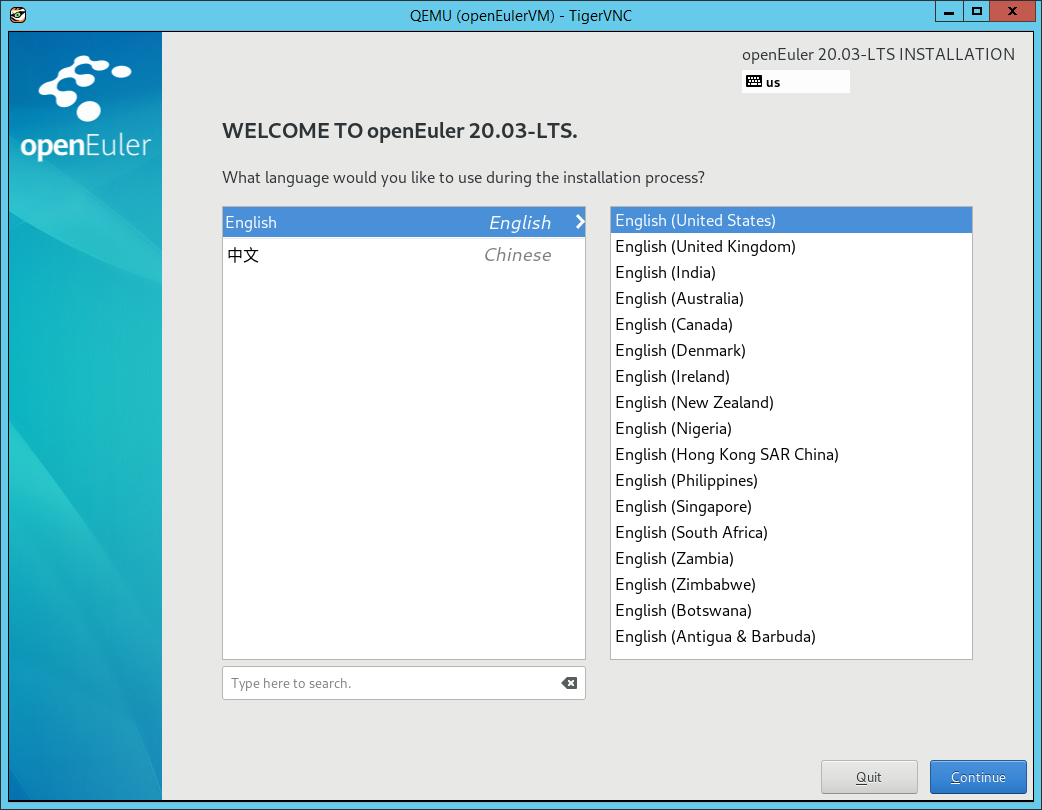
然后点击“Connect”按钮链接。



输入对应文件***openEulerVM01.xml（、openEulerVM02.xml、openEulerVM03.xml……等，如果创建了多台虚拟机）***中指定的密码后点击“OK”按钮。

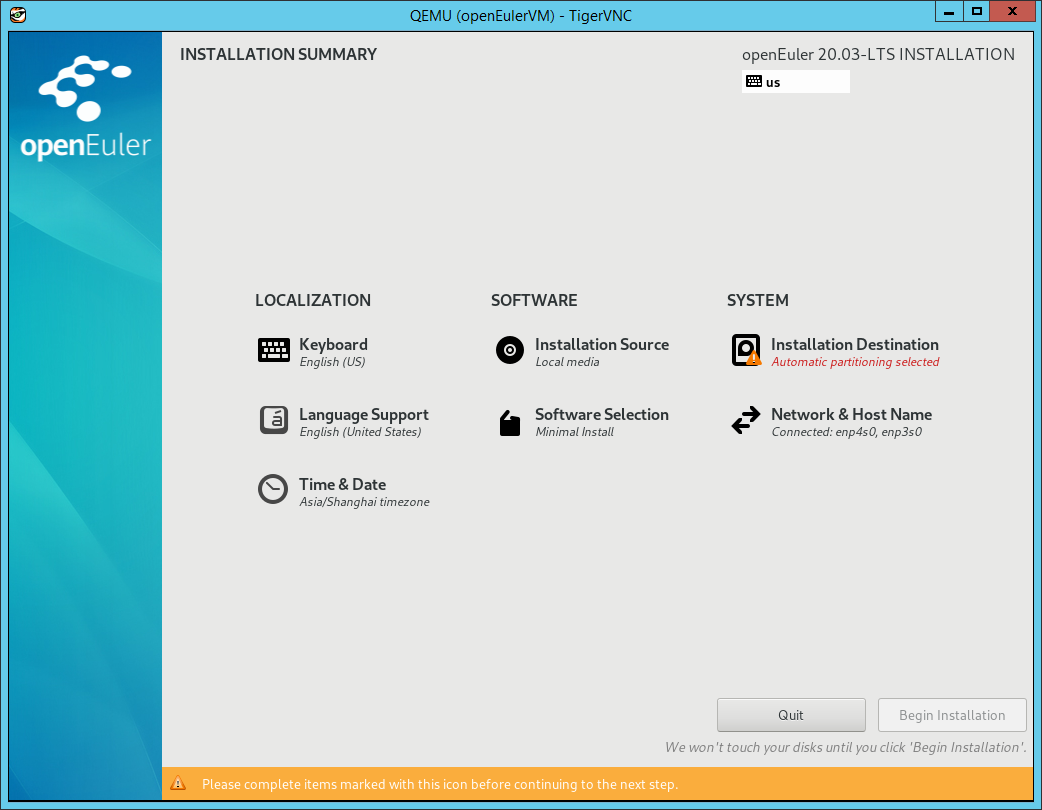
第一次进入虚拟机时会进入系统安装界面：



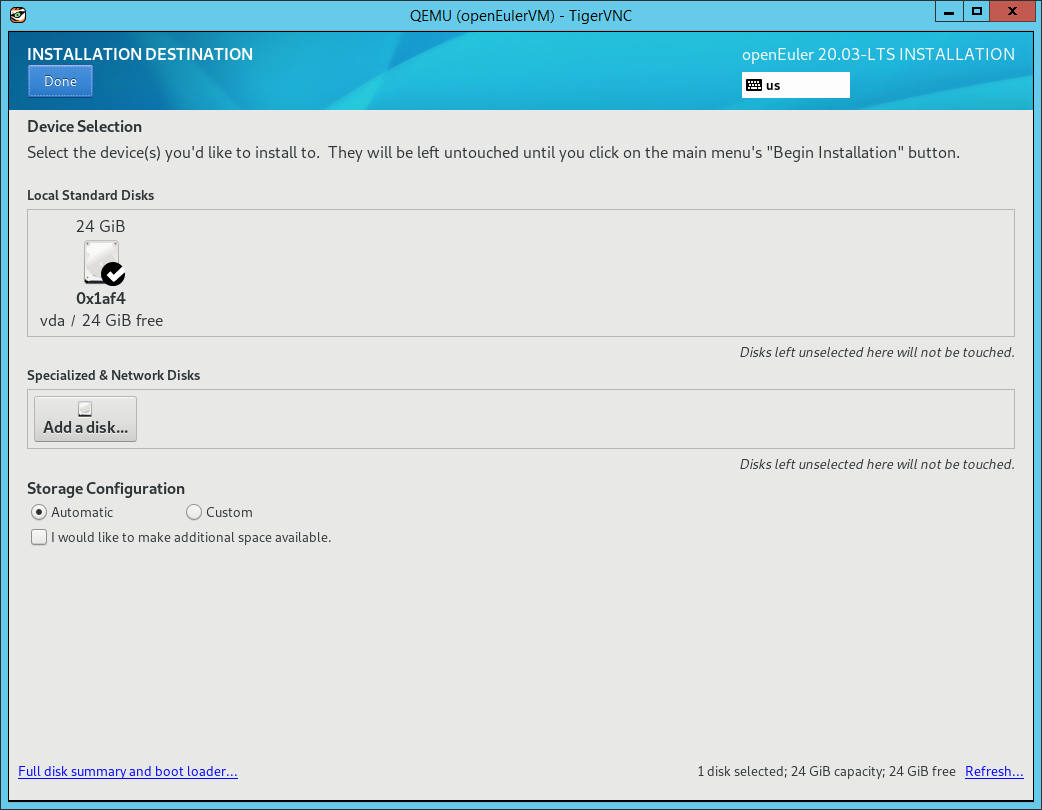


此时表明虚拟机安装成功。请保持默认的英文安装，并按下述步骤在虚拟机上安装openEuler操作系统。

请在“SYSTEM”栏目的“Installation Destination”里确认安装位置：

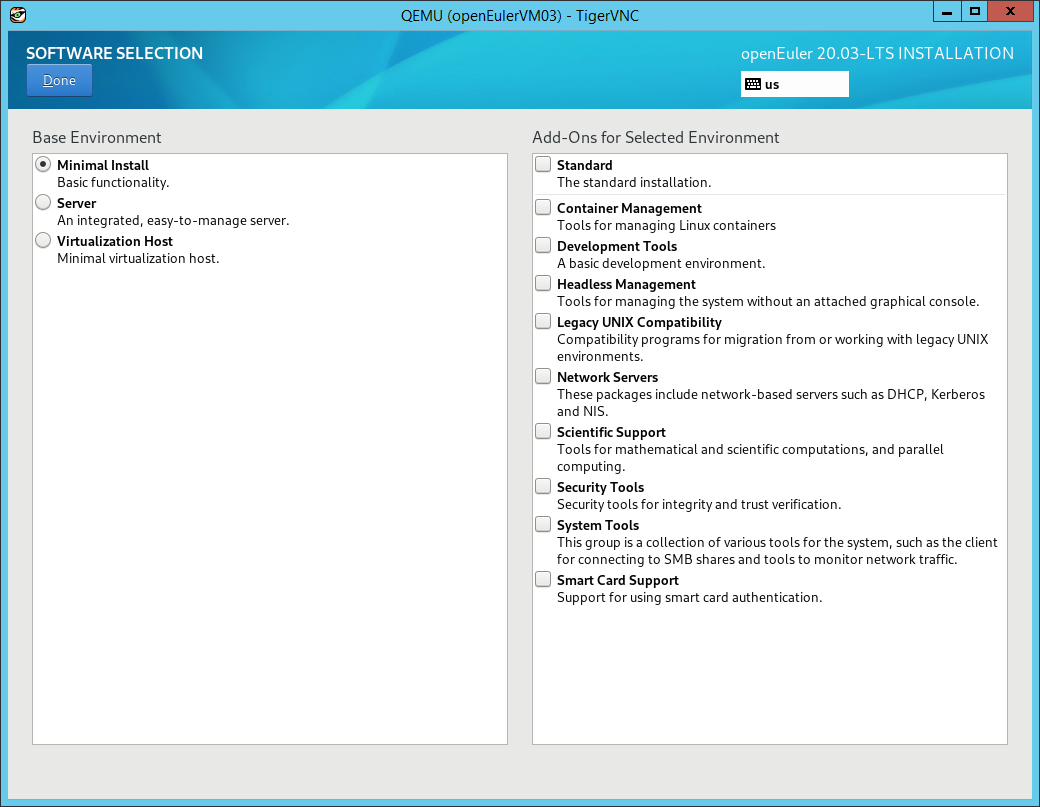


这里保持默认设置，直接点击左上角的“Done”按钮即可：



点击“SOFTWARE”栏目下面的“Software Selection”按钮：

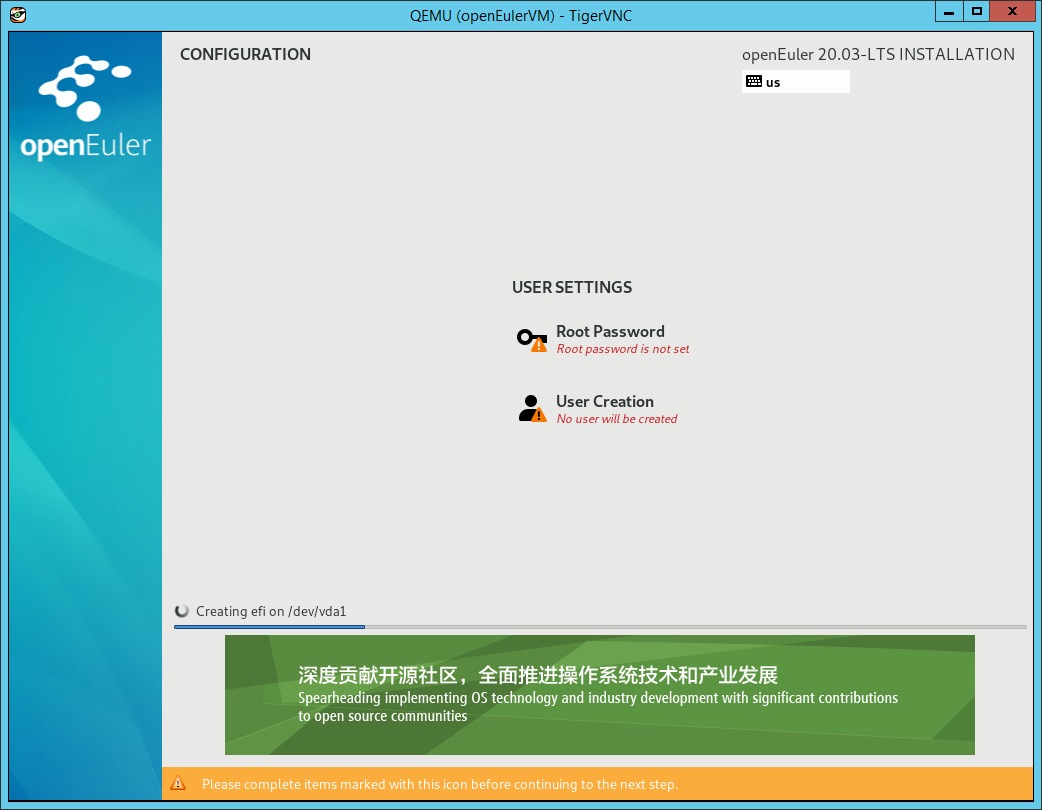




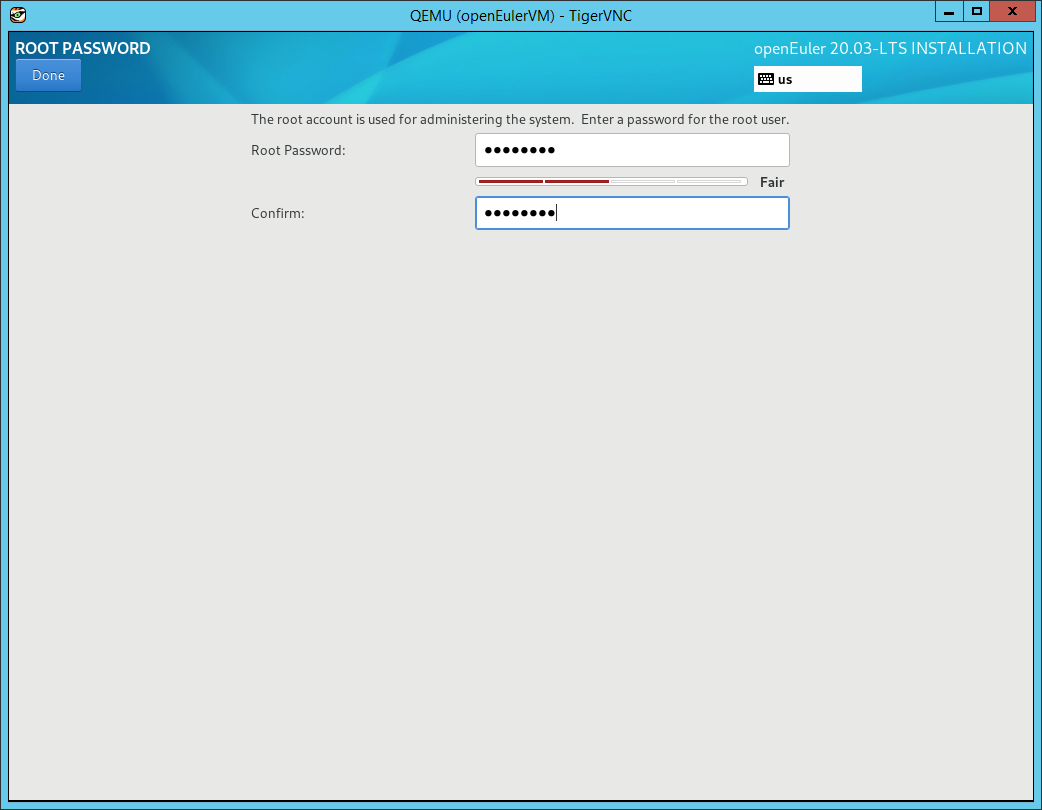
选择“Minimal Install”安装，在这里我们不需要勾选右侧的任何选项并点击左上角的“Done”按钮回到主界面。



点击右下角的“Begin Installation”开始安装。

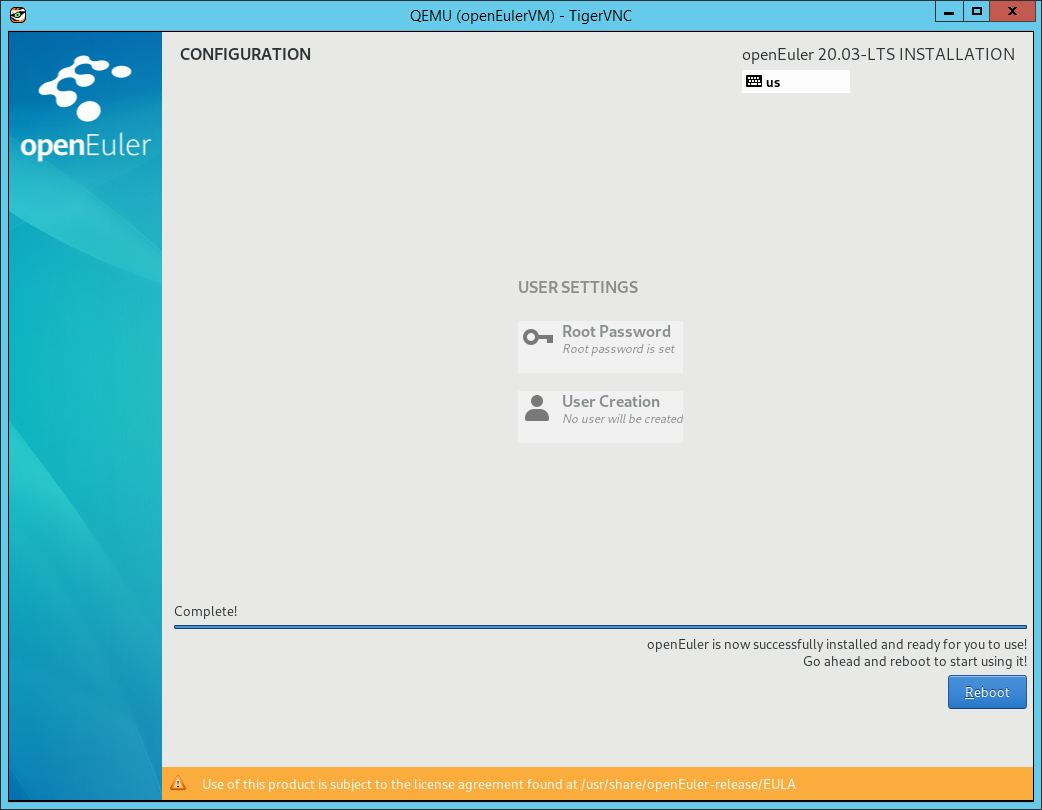


在安装期间请设置root用户的密码（必需）。

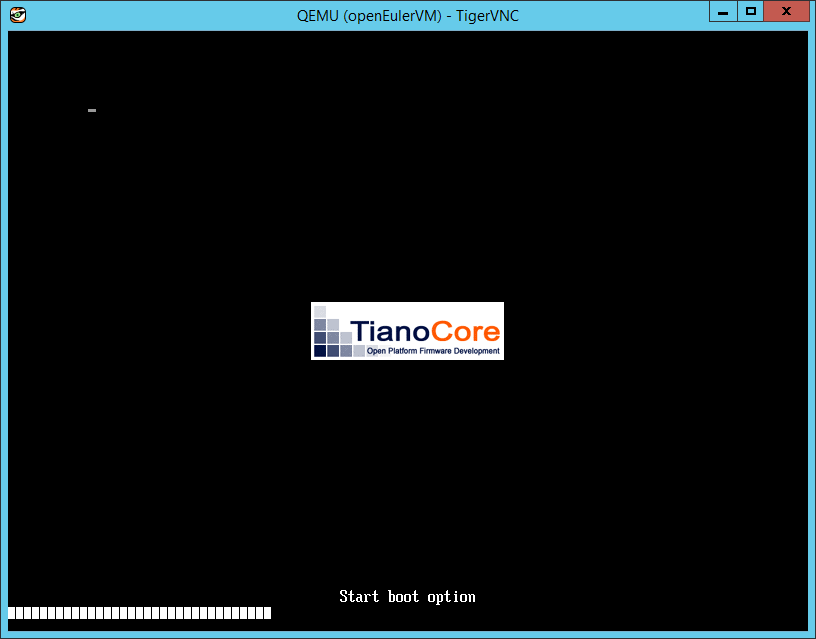


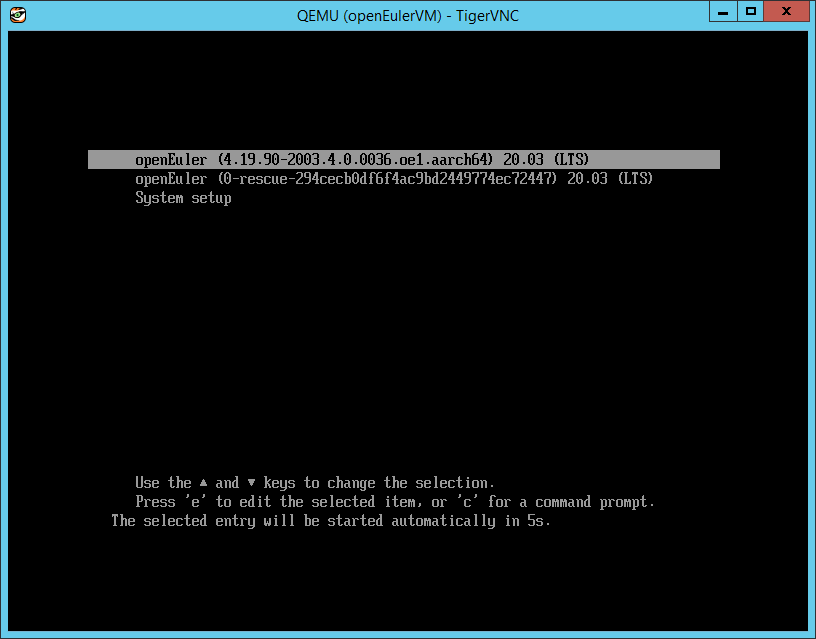
这是设置root用户密码的界面。

您也可以设置普通用户的密码（非必需）。



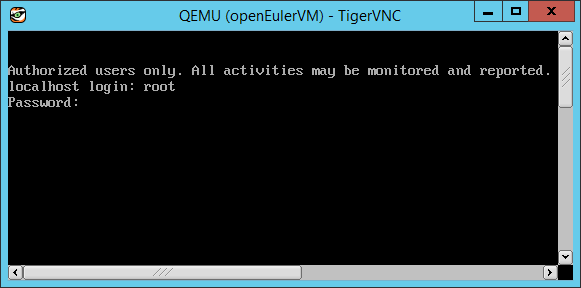
安装完成后点击“Reboot”按钮重启。



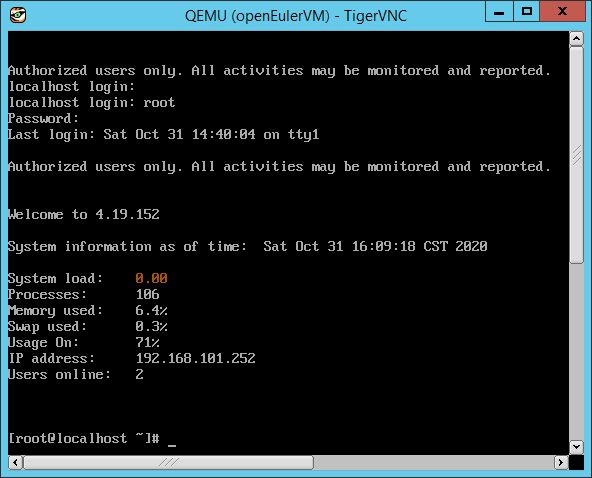


请选择“openEuler (4.19.90-2003.4.0.0036.oel.aarch64) 20.03 (LTS)”选项启动（您的子版本号可能有所不同）。

（上图中的第二个启动项为救援模式。）



在新出现的Console窗口中输入登入虚拟机openEuler操作系统的用户名（这里是root）和密码（输入时不会有任何显示）并按回车键。



这里是登录成功后的界面。至此在虚拟机上安装openEuler操作系统成功。