学习情况

2014-4-22

周永强

六院八隊

Contents

说明

1 学习情况

课程已经进入中期，有许多课程任务要做，主要是前期拖延积攒下来的任务。实验室那边去的事情不多，所以这段时间去的比较少，于是便以课程为主。我选的课普遍几乎都是应用性比较强的课程，Linux内核要做源码分析，高级并行程序要做并行程序设计，分布式系统要结合分布式做一个android应用，路由器原理又要做netmagic平台上的SDN计数器。前三个任务已经完成过半，最后一个还没有开始做，因为我的硬件基础几乎等于0，所以比较吃力。

后面的内容是有关于内核网卡驱动中napi技术的介绍和分析，也是为选的课程作业，结合了网上别人的分析和自己阅读源代码后写出来的，由于表达能力不是太好，写得可能不是太明白，同时限于水平问题，可能会有错的地方。

NAPI

1 NAPI简介

NAPI(”New API”) 是对设备驱动程序报文处理的一个框架，其设计的初衷是为了提高高速网络设备的性能。 NAPI 通过两种机制来提高性能，分别是：中断减负和报文节流。

中断减负：典型的报文处理流程是当报文到达网卡时，设备产生中断通知 CPU，然后 CPU 响应中断，从网卡读取报文数据交由上层协议栈。但是对于高速网卡设备来说，单位时间内会产生数以千计的中断通知系统响应报文处理，频繁的中断会严重影响系统的性能。 NAPI 允许驱动在大流量的网络环境时禁用中断，从而降低中断次数，降低系统因中断造成的负担。

报文节流：当大量报文涌向系统时，系统必然会产生丢包，相比交由系统处理时才丢弃报文，能够尽早地丢弃报文更好。使用 NAPI 的驱动能够在报文到达网络适配器时就将其丢弃，内核根本无法看到它们，这里丢弃报文指的是指因系统处理不过而丢弃的报文。

新的网卡在开发驱动时应该使用 NAPI 技术以获得更好的性能。

2 NAPI驱动设计

NAPI的设计思想就一句话，综合了中断和轮询。当网络数据到达网卡时，产生中断，此时中断程序响应，但是中断程序并不处理数据，而是禁用中断，并且通过NAPI的poll函数以轮询的方式对数据进行处理，而且在适当的时候重新启用中断，具体什么时候，后面还会讲。使用NAPI共分为四步骤，分别为：定义，注册与初始化，调度和删除。

下面以e1000驱动为例子，详细说明使用NAPI技术的驱动程序接收数据的全过程。

3 NAPI定义

NAPI使用结构体napi_struct描述，如果驱动程序要使用该技术的话，可以在自己的结构体中内嵌实现。在e1000驱动adapter的定义中可以看到napi_struct的存在。

Program 1: e1000.h

```
/*e1000.h*/
struct e1000_adapter {
  /*...*/
  /*嵌入napi结构*/
  struct napi_struct napi;
  /*...*/
}
```

4 NAPI注册与初始化

napi_struct在定义以后，必须初始化和注册才能够使用。其初始化与注册是在e1000_probe函数中完成，该函数是探测函数，会在。。（PCI routine）。。执行。注册函数为：

```
netif_napi_add(dev, &napi, poll, weight)
```

前两个参数是设备和napi结构指针，重要的是第三个参数，第三个参数是函数指针，指向使用napi时要使用的轮询函数，在注册的时候必须指明该参数，然后当使用轮询接收数据时，会回调该函数。在e1000驱动中，轮询函数为e1000_clean。

Program 2: e1000_probe

```
/*e1000_main.c*/
static int e1000_probe(struct pci_dev *pdev, const struct pci_device_id *ent)
{
  struct net_device *netdev;
  struct e1000_adapter *adapter;
```

```
  /*...*/
  SET_NETDEV_DEV(netdev, &pdev->dev);
  pci_set_drvdata(pdev, netdev);
  adapter = netdev_priv(netdev);
  adapter->netdev = netdev;
  adapter->pdev = pdev;
  /*...*/
  //初始化网卡驱动的相关操作，包括open，close等
  netdev->netdev_ops = &e1000_netdev_ops;
  e1000_set_ethtool_ops(netdev);
  /*...*/
  //napi初始化与注册，这里面要指明轮询函数为e1000_clean
  netif_napi_add(netdev, &adapter->napi, e1000_clean, 64);
  /*...*/
}
```

5 调度NAPI

在讲napi的调度之前必须先说明一下中断，中断在e1000_open中注册完成，注册完成后，每当网卡有数据到达，便会产生中断，然后调用中断处理程序，在中断处理程序中禁用中断，并且调度napi轮询接收数据。napi的调度函数为：

```
void napi_schedule(struct napi_struct *napi);
```

或：

```
if (napi_schedule_prep(napi))
__napi_schedule(napi);
```

二者的效果一样，只不过后者是先测试当前已经在进行napi调度了。

Program 3: e1000_open

```
/*e1000_open*/
static int e1000_open(struct net_device *netdev)
{
    struct e1000_adapter *adapter = netdev_priv(netdev);
    struct e1000_hw *hw = &adapter->hw;
    /*...*/
    //网卡设置，其中apapter->clean_rx 在其中进行设置
    e1000_configure(adapter);
    /*...*/
    //申请中断
    err = e1000_request_irq(adapter);
    if (err)
        goto err_req_irq;
```

```
    /*...*/
    napi_enable(&adapter->napi);
    e1000_irq_enable(adapter);
    /*...*/
}

//中断注册
static int e1000_request_irq(struct e1000_adapter *adapter)
{
    struct net_device *netdev = adapter->netdev;
    //中断处理函数设定
    irq_handler_t handler = e1000_intr;
    int irq_flags = IRQF_SHARED;
    int err;
    //申请中断，中断处理函数为e1000_intr
    err = request_irq(adapter->pdev->irq, handler, irq_flags, netdev->name, netdev);
    /*...*/
}

static irqreturn_t e1000_intr(int irq, void *data)
{
    /*...*/
    /* disable interrupts, without the synchronize_irq bit */
    ew32(IMC, ~0);
    E1000_WRITE_FLUSH();
    //调度napi进行数据接收
    if (likely(napi_schedule_prep(&adapter->napi))) {
        adapter->total_tx_bytes = 0;
        adapter->total_tx_packets = 0;
        adapter->total_rx_bytes = 0;
        adapter->total_rx_packets = 0;
        __napi_schedule(&adapter->napi);
    }
    /*...*/
}
```

当napi调度以后，后面的工作便是轮询接收数据，轮询函数已经在e1000_probe中设定为e1000_clean。

Program 4: e1000_clean

```
static int e1000_clean(struct napi_struct *napi, int budget)
{
```

```
    struct e1000_adapter *adapter = container_of(napi, struct e1000_adapter, napi);
    int tx_clean_complete = 0, work_done = 0;
    //发送数据报文，这是发送流程里面的，暂时可以不关心
    tx_clean_complete = e1000_clean_tx_irq(adapter, &adapter->tx_ring[0]);
    //接收数据，这里clean_rx 为函数指针，指向数据接收函数
    //budget为接收数据量的限制
    adapter->clean_rx(adapter, &adapter->rx_ring[0], &work_done, budget);

    if (!tx_clean_complete)
        work_done = budget;

    /* If budget not fully consumed, exit the polling mode */
    //如果budget没有用完，也即接收的数据量没有超多限度，
    //那么说明轮询可以暂时结束了，重新启用中断，响应后续到来的数据
    if (work_done < budget) {
        if (likely(adapter->itr_setting & 3))
            e1000_set_itr(adapter);
        napi_complete(napi);
        if (!test_bit(__E1000_DOWN, &adapter->flags))
            //重新启用中断
            e1000_irq_enable(adapter);
    }
    return work_done;
}
```

最后还要看一下clean_rx函数，这个函数之所以设定成指针是为了灵活使用接收函数，不同型号的网卡可能不一样，但是原理都是一样的，该指针的初始化在e1000_open函数中的e1000_configure()中进行设定，而该函数又调用e1000_configure_rx进行设定接收函数，假定接收函数被设为e1000_clean_rx_irq（设成其他函数也一样，原理是相同的），接收函数完成功能见代码。

Program 5: e1000_clean_rx_irq

```
static bool e1000_clean_rx_irq(struct e1000_adapter *adapter,
                               struct e1000_rx_ring *rx_ring,
                               int *work_done, int work_to_do)
{
  /*...*/
  i = rx_ring->next_to_clean;
  rx_desc = E1000_RX_DESC(*rx_ring, i);
  buffer_info = &rx_ring->buffer_info[i];

  while (rx_desc->status & E1000_RXD_STAT_DD) {
```

```
    //sk_buff代表报文，下面接收完毕就上交协议栈做下一步的处理
    //数据也就从网卡驱动进入内核协议栈进行解析了
    struct sk_buff *skb;
    u8 status;

    if (*work_done >= work_to_do)
      break;
    (*work_done)++;
   /* read descriptor and rx_buffer_info after status DD */
    rmb();

    skb = buffer_info->skb;
    buffer_info->skb = NULL;

    prefetch(skb->data - NET_IP_ALIGN);
    /*...*/

    length = le16_to_cpu(rx_desc->length);
    /*...*/
    process_skb:
      total_rx_bytes += (length - 4); /* don't count FCS */
      total_rx_packets++;

    /*...*/
    skb_put(skb, length);
    /*...*/
    e1000_receive_skb(adapter, status, rx_desc->special, skb);

    next_desc:
      rx_desc->status = 0;
    /*...*/
    }
}
```

6 删除NAPI

消注册在网卡消注册的时候进行调用，调用函数为netif_napi_del()，具体的代码不再分析。

Bibliography

[1] http://www.linuxfoundation.org/node/add/wiki?gids[]=5066.

[2] Linux kernel 2.6.34.