

Création d'un framework pour des routeurs avec support pour notification de vitesse explicite (kernel space)

Encadrant: Dino Lopez

Calypso Petit, François Chapuis, Sophie Valentin, Mathieu Bivert

31 janvier 2012

Table des matières

1	Introduction	1
2	Description du problème	2
2.1	Élaboration de statistiques sur les paquets TCP	2
2.2	Netfilter	2
3	Solutions	3
3.1	Environnement de tests	3
3.2	Test de connectivité	4
3.3	Détermination de la taille de la pile avec Netfilter	5
3.3.1	<i>NF_IP_FORWARD</i> et <i>NF_IP_POST_ROUTING</i>	5
3.3.2	<i>NF_IP_PRE_ROUTING</i> et <i>NF_IP_POST_ROUTING</i>	6
3.4	Lecture des données sur les cartes réseaux	6
3.4.1	Statistiques de la carte	6
3.4.2	npkt++	6
3.4.3	Wrapper pour les drivers ethernet (npkt-)	6
4	Bibliographie	7
A	Compléments sur la programmation de modules Linux	8
A.1	Organisation générale	8
A.2	Mise en place d'un hook netfilter	8
A.3	Synchronisation (spinlock)	8
A.4	Compilation d'un noyau Linux	8

1 Introduction

L'essor de réseaux de grande envergure nécessite la mise en place de protocoles de contrôle de congestion. En effet, l'augmentation du trafic, et la distribution non-équitable des ressources entre les usagers, peuvent conduire à un ralentissement des communications.

Une famille de protocoles de contrôle de congestion est l'IP-ERN (Explicit Rate Notification). Ils reposent sur les informations envoyées/calculées par des routeurs, comme le taux d'émission optimal. Dans le cadre de notre projet, nous devons élaborer un module noyau à destination d'un routeur supportant l'IP-ERN. Un tel protocole modifie les paquets TCP transitant par le routeur. Dans un premier temps, nous allons élaborer des statistiques sur les paquets TCP qui traversent le routeur. Puis dans un second temps, nous nous attarderons sur le protocole (ou pas).

2 Description du problème

2.1 Élaboration de statistiques sur les paquets TCP

Un routeur est une machine qui assure l'acheminement des données, d'un réseau informatique à un autre. Son rôle consiste à diriger les paquets d'une interface réseau vers une autre par le chemin le plus rapide et selon les règles définies dans la table de routage. Les interfaces réseaux peuvent être assimilées à des portes par lesquelles les données entrent et sortent du routeur, et la table de routage à une façon de diriger les paquets à travers ces portes. Il s'agit ici dans un premier temps de calculer des statistiques nécessaires pour le protocole IP-ERN.

2.2 Netfilter

Netfilter est un framework permettant l'interception des paquets grâce à des hooks : ce sont des points d'accroche dans le noyau. Lorsqu'un événement réseau se produit (entrée d'un paquet par exemple), une fonction de callback associée à ce type d'événement est appelée. A chaque hook, Netfilter permet d'accepter les paquets ou de s'en débarrasser. Pour le protocole IPv4, on compte cinq hooks. Ces derniers sont organisés comme sur 1.

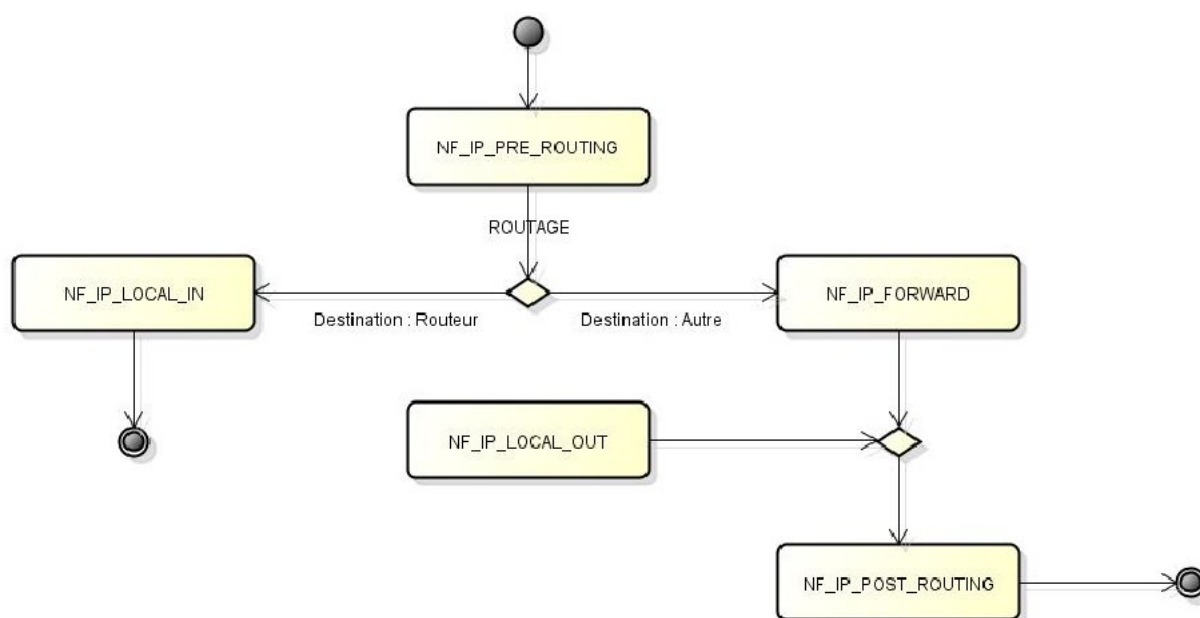


FIG. 1 – Position des Hooks netfilter

Les paquets entrent par le haut. Après avoir subi des tests au niveau de la carte réseau, ils traversent un premier hook, appelé hook `NF_IP_PRE_ROUTING`. À ce niveau, seule l'interface d'entrée du routeur est connue. Ensuite, après consultation de la table de routage, les paquets à destination du routeur sont dirigés vers un processus interne et passent par le deuxième hook, `NF_IP_LOCAL_IN`. Quant aux autres paquets, ils sont dirigés vers une interface de sortie et traversent successivement le troisième hook, `NF_IP_FORWARD` puis le quatrième, `NF_IP_POST_ROUTING`. Au niveau du troisième hook, les interfaces d'entrée et sortie sont connues. En revanche, dans le quatrième, seule l'interface de sortie est connue.

En utilisant l'architecture de Netfilter, nous devons renseigner un tableau de statistiques comportant, pour chaque interface, le nombre de paquets entrants à destination de cette interface (Input traffic Rate, I) et le nombre de paquets sortants (Output traffic Rate, O) par cette interface. La quantité de donnée pouvant circuler sera donnée par l'Output Link Capacity (C). Le tableau sera de la forme donnée par 2.

Nom de l'interface	I	O	C	Q = O-I
eth0
eth1
...
ethn

FIG. 2 – Statistiques requises par le protocole

En comptabilisant l'Input traffic Rate et l'Output traffic Rate, on pourra déterminer la taille de la file d'attente (Qsize). Cependant, ne sachant pas où ce buffer se situe, le choix des hooks va être crucial.

Afin d'utiliser Netfilter, il faut programmer en mode noyau. Les modules sont des morceaux de code écrits en langage C, pouvant être ajoutés ou retirés du noyau dynamiquement. Ils apportent de nouvelles fonctionnalités au noyau. Dans notre cas, le module fournira des statistiques sur la taille de la file d'attente. La programmation en mode noyau sous Linux est différente de celle en mode utilisateur. En effet, les bibliothèques sont plus restreintes et le débogage plus ardu, car le moindre accès mémoire invalide peut rendre le système inutilisable.

L'utilisation de machines virtuelles permet donc d'éviter de devoir redémarrer à chaque plantage, et va aussi faciliter la mise en place d'interfaces réseaux, donc les tests de façon générale.

Les détails techniques concernant la programmation de modules Linux sont évoqués en annexe.

3 Solutions

3.1 Environnement de tests

Dans un premier temps, on souhaite mettre en place une architecture réseau simple, constituée de trois machines, la figure 3 indiquant la topologie réseau suivie :

- un émetteur ;
- un routeur ;
- et un récepteur.

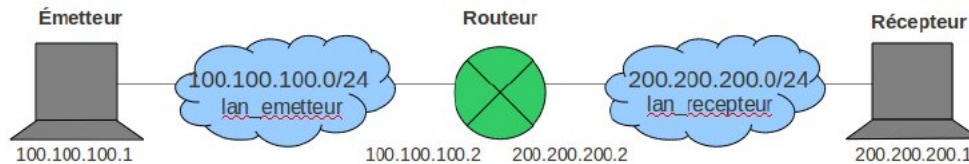


FIG. 3 – Topologie du réseau

Pour cela, nous utilisons trois machines virtuelles, lancées sur la même machine. Virtualbox met à disposition plusieurs d'interfaces réseaux virtuelles :

NAT interface connectée à un réseau local, sur lesquelles se trouvent les différentes machines virtuelles

Bridge interface connectée au réseau de la machine hôte

Un routeur servant à faire transiter des données d'un réseau à un autre, il nous en faut donc deux : un sur lequel se trouve l'émetteur, un pour le récepteur, et le routeur est connecté au deux (fig.4).

Il faut donc configurer sur chaque machine les interfaces réseaux NAT en conséquence, ainsi qu'ajouter des entrées dans la table de routage au niveau de l'émetteur et du récepteur.

Nom du réseau	CIDR	Masque de sous-réseau
lan_emetteur	100.100.100.0/24	255.255.255.0
lan_recepteur	200.200.200.0/24	255.255.255.0

FIG. 4 – Les deux réseaux

```
svalenti@svalenti-laptop:~$ ping 100.100.100.2
PING 100.100.100.2 (100.100.100.2) 56(84) bytes of data.
64 bytes from 100.100.100.2: icmp_seq=1 ttl=64 time=0.564 ms
64 bytes from 100.100.100.2: icmp_seq=2 ttl=64 time=0.592 ms
64 bytes from 100.100.100.2: icmp_seq=3 ttl=64 time=0.541 ms
^C
--- 100.100.100.2 ping statistics ---
3 packets transmitted, 3 received, 0% packet loss, time 1998ms
rtt min/avg/max/mdev = 0.541/0.565/0.592/0.034 ms
svalenti@svalenti-laptop:~$ ping 200.200.200.1
PING 200.200.200.1 (200.200.200.1) 56(84) bytes of data.
64 bytes from 200.200.200.1: icmp_seq=1 ttl=63 time=0.925 ms
64 bytes from 200.200.200.1: icmp_seq=2 ttl=63 time=1.03 ms
64 bytes from 200.200.200.1: icmp_seq=3 ttl=63 time=0.950 ms
^C
--- 200.200.200.1 ping statistics ---
3 packets transmitted, 3 received, 0% packet loss, time 2002ms
rtt min/avg/max/mdev = 0.925/0.969/1.033/0.052 ms
svalenti@svalenti-laptop:~$
```

FIG. 5 – Émetteur et récepteur peuvent communiquer

Le routeur doit simplement être connecté aux deux réseaux. Chacune des deux interfaces du routeur doit donc être connectée au bon réseau :

```
# ifconfig eth1 100.100.100.2 netmask 255.255.255.0
# ifconfig eth2 200.200.200.2 netmask 255.255.255.0
```

Enfin, il ne faut pas oublier d'activer le transfert de paquets :

```
# echo 1 > /proc/sys/net/ipv4/ip_forward
```

Du côté de l'émetteur, l'interface eth0 est reliée à *lan_emetteur*, et il faut passer par 100.100.100.2 (c'est-à-dire l'interface eth1 du routeur) pour transmettre des paquets à destination du réseau *lan_recepteur* :

```
# ifconfig eth0 100.100.100.1 netmask 255.255.255.0
# route add -net 200.200.200.0 netmask 255.255.255.0 gw 100.100.100.2 eth0
```

Côté récepteur, la configuration est analogue à celle de l'émetteur :

```
# ifconfig eth0 200.200.200.1 netmask 255.255.255.0
# route add -net 100.100.100.0 netmask 255.255.255.0 gw 200.200.200.2 eth0
```

3.2 Test de connectivité

En guise de vérification, on essaye de *pinger* le récepteur depuis l'émetteur (fig.5) :

La commande *ping* n'est pas suffisante pour envoyer des paquets dans le cadre de notre projet car elle utilise le protocole ICMP et non le protocole IP. Par conséquent, nous avons cherché un outil utilisant les protocoles IP et TCP. (Est-ce qu'il faut spécifier ce que sont les protocoles IP et TCP ?) La commande *iperf* permet d'envoyer entre un client et un serveur des paquets TCP ou UDP. Nous l'avons donc utilisé entre l'émetteur et le récepteur. Par défaut *iperf* utilise le protocole TCP.

Côté récepteur, on lance *iperf* en mode serveur :

```
# iperf -s
```

et côté émetteur, on envoie des paquets sur le récepteur :

```
# iperf -c 200.200.200.1
```

L'option *-i* d'*iperf* peut-être utilisée pour fixer l'intervalle de temps entre les rapports d'envois de paquets (mesure bande passante, ...).

```

[ 543.299399] FORWARD
[ 543.299402] PostIn: eth1 | PostOut: eth2
[ 543.299403] InputPkts: 121566
[ 543.299405] POST_ROUTING
[ 543.299406] PostIn: <NULL> | PostOut: eth2
[ 543.299407] OutPkt: 121566
[ 543.299408] Qsize: 0
[ 543.304527]
[ 543.304530] FORWARD
[ 543.304534] PostIn: eth2 | PostOut: eth1
[ 543.304536] InputPkts: 28096
[ 543.304538] POST_ROUTING
[ 543.304539] PostIn: <NULL> | PostOut: eth1
[ 543.304540] OutPkt: 28096
[ 543.304541] Qsize: 0
[ 543.304859]
[ 543.304860] FORWARD
[ 543.304862] PostIn: eth1 | PostOut: eth2
[ 543.304864] InputPkts: 121567
[ 543.304865] POST_ROUTING
[ 543.304866] PostIn: <NULL> | PostOut: eth2
[ 543.304868] OutPkt: 121567
[ 543.304869] Qsize: 0

```

FIG. 6 – Trace obtenue entre FORWARD et ROUTING

3.3 Détermination de la taille de la pile avec Netfilter

A présent, nous écrivons un module utilisant Netfilter. Les fonctions de callback données par les hooks permettent d'accéder à un *struct sk_buff*. Cette structure représente un paquet dans le noyau Linux. Il est ainsi possible de récupérer les informations du datagramme telles que le protocole ou encore l'adresse IP de destination.

Les fonctions de callback fournissent également deux structures de type *struct net_device* : une pour l'interface réseau d'entrée et une autre pour l'interface réseau de sortie. Avec une telle structure, on peut connaître le nom de l'interface grâce au champ *net_device.name*.

Cependant, ces structures ne sont pas toujours remplies par le noyau : cela dépend du hook dans lequel nous nous trouvons. Pour déterminer où se trouve la file d'attente, nous devons travailler avec deux hooks : le premier hook doit se trouver en amont de l'emplacement supposé de la file d'attente et le second en aval. Dans le premier hook, l'input trafic rate de l'interface de sortie est incrémenté et dans le second hook, l'output trafic rate de l'interface de sortie est incrémenté.

Après chaque opération, on met à jour la taille de la file d'attente de l'interface de sortie en soustrayant l'input trafic rate par l'output trafic rate.

Les deux hooks pouvant accéder aux mêmes données en même temps, il est indispensable de mettre en place un mécanisme de synchronisation. Les hooks étant déclenchés par des interruptions, il est impossible d'utiliser des mutex ou encore des sémaphores. Le noyau fournit des spinlocks, jouant le même rôle, mais dans un contexte d'interruptions.

Les hooks concernant les paquets qui traversent le routeur sont : *NF_IP_PRE_ROUTING*, *NF_IP_FORWARD* et *NF_IP_POST_ROUTING*. Parmi ces hooks, seuls *NF_IP_FORWARD* et *NF_IP_POST_ROUTING* contiennent l'information sur l'interface de sortie dans la structure *net_device*. Par conséquent, nous allons tout d'abord chercher la taille de la pile entre *NF_IP_FORWARD* et *NF_IP_POST_ROUTING*.

3.3.1 *NF_IP_FORWARD* et *NF_IP_POST_ROUTING*

Pour commencer, les vitesses maximales des interfaces eth1 et eth2 sont configurées de la même façon. Ainsi, on s'attend à ce que la sortie des paquets soit fluide : nous ne nous attendons pas à la création d'une file d'attente interne. L'exécution du module utilisant ces deux hooks donne l'extrait de trace en fig.6.

Après examination entière de la trace, on constate que Qsize pour eth1 et Qsize pour eth2 sont constamment égaux à zéro. On peut tout de même vérifier que le nombre de paquets entrants est incrémenté correctement à chaque exécution de la fonction de callback de *NF_IP_FORWARD*. Quant au nombre de paquets sortants, il est incrémenté correctement à chaque exécution de la fonction de callback du hook *NF_IP_POST_ROUTING*.

À présent, on configure la bande passante maximale de l'interface eth2 afin que la vitesse maximale soit de 10 Mbits/seconde. Pour cela, on utilise la commande *tc* (traffic control). Cette commande va agir sur le fonctionnement des files d'attente en sortie d'une interface donnée.

```
# tc qdisc add dev eth2 root handle 1: tbf rate 10mbit buffer 1600 limit 3000
# tc qdisc add dev eth2 parent 1: handle 10: netem delay 3ms
```

Pour limiter la bande passante, on utilise un Token Bucket Filter (tbf) en précisant qu'on ne peut pas dépasser un débit de 10Mbits/seconde.

Après avoir ainsi configuré la vitesse maximale, nous nous attendons à obtenir une file d'attente non nulle. Les résultats des tests effectués nous montrent que la taille de la file d'attente est toujours nulle, ce qui prouve que la file d'attente interne ne se trouve pas entre les hooks *NF_IP_FORWARD* et *NF_IP_POST_ROUTING*.

Il nous faut donc regarder entre deux autres hooks.

3.3.2 *NF_IP_PRE_ROUTING* et *NF_IP_POST_ROUTING*

L'interface de sortie n'étant plus disponible en PRE ROUTING, il va nous falloir effectuer le routage à la main. Notre objectif étant de chercher où se trouve la file, nous nous contentons d'un routage manuel, avec les valeurs entrées en dur dans le code.

Nous obtenons finalement le même résultats qu'entre FORWARD et POST, il semble donc qu'il soit impossible d'utiliser Netfilter pour obtenir les statistiques. On pourrait cependant empiler les paquets sur lesquelles on désire avoir des statistiques, et passer le buffer en userspace pour y faire les calculs. Il est néanmoins important pour le code d'être efficace, et sortir du mode noyau a de fortes chances d'impacter la vitesse d'exécution du code (pourquoi ?) ; on va donc chercher une autre façon d'obtenir la taille de la pile tout en restant en mode noyau.

3.4 Lecture des données sur les cartes réseaux

3.4.1 Statistiques de la carte

Grâce à une fonction du noyau, il est possible d'accéder à des statistiques concernant les interfaces. La fonction *dev_get_stats* définie dans *linux/netdevice.h* fournit :

- le nombre de paquets (ou octets) reçus sur une interface ;
- le nombre de paquets (ou octets) transmis par une interface ;
- le nombre d'erreurs ; et le nombre de paquets "dropped".

Ces statistiques pourraient nous aider à calculer Qsize. Cependant, elles ne permettent pas de différencier les paquets TCP des autres paquets. Il nous faut donc aller un peu plus loin dans le code.

3.4.2 npkt++

3.4.3 Wrapper pour les drivers ethernet (npkt-)

Les fonctions permettant de manipuler une carte ethernet sont sous linux abstraite par une structure : ajouter un driver revient donc à déclarer une structure et à remplir les différents champs, avec les fonctions adaptées à la carte. La structure *net_device_ops*, déclarée dans *linux/netdevice.h* joue ce rôle :

```
struct net_device_ops {
    int (*ndo_init)(struct net_device *dev);
    void (*ndo_uninit)(struct net_device *dev);
    int (*ndo_open)(struct net_device *dev);
    int (*ndo_stop)(struct net_device *dev);
    netdev_tx_t (*ndo_start_xmit) (struct sk_buff *skb, struct net_device *dev);
    ...
};
```

Et un driver, par exemple e1000, l'utilise ainsi :

```
static const struct net_device_ops e1000_netdev_ops = {
    .ndo_open          = e1000_open,
    .ndo_stop          = e1000_close,
    .ndo_start_xmit     = e1000_xmit_frame,
    .ndo_get_stats      = e1000_get_stats,
    ...
};
```

La fonction *dev_hard_start_xmit*, située dans *net/core/dev.c*, est chargée de transmettre le paquet à la carte. Après quelques tests à l'aide de *printk* bien placé, nous avons constaté que dans tous les cas, la fonction est appelée, plus ou moins directement, et qu'elle contient un appel à *ndo_start_xmit*. La décrementation de la taille de la pile peut donc être effectuée ici, plutôt que de l'être dans le noyau.

4 Bibliographie

A Compléments sur la programmation de modules Linux

A.1 Organisation générale

A.2 Mise en place d'un hook netfilter

A.3 Synchronisation (spinlock)

A.4 Compilation d'un noyau Linux