RFC 5475 : Sampling and Filtering Techniques for IP Packet Selection

Stéphane Bortzmeyer
<stephane+blog@bortzmeyer.org>

Première rédaction de cet article le 1 avril 2009

Date de publication du RFC : Mars 2009

https://www.bortzmeyer.org/5475.html

Ce RFC fait partie de la série des normes produites par le groupe de travail psamp <http://www.ietf.org/html.charters/psamp-charter.html> sur l'échantillonage (ou, plus exactement, la **sélection**) des paquets IP. Les explications générales sur ce mécanisme figurent dans le RFC 5474 [1] et notre RFC 5475, lui, détaille les techniques de sélection des paquets.

Il est donc très recommandé de lire le RFC 5474 avant puisque ce RFC 5475 en reprend les concepts et le vocabulaire (sections 2 et 3, ainsi que la section 4, qui reprend la classification des techniques de sélection en **filtrage** (où on sélectionne certains paquets selon leurs propriétés, par exemple tous les paquets UDP) et **échantillonage** où on veut juste sélectionner un sous-ensemble représentatif des paquets. L'échantillonage peut se faire via une fonction de hachage (avec certaines de ces fonctions, cela peut faire un échantillonage quasi-aléatoire) mais, à part ce cas, le contenu du paquet n'est pas pris en compte. La même section 4 rappelle les différentes techniques de sélection, déjà présentées dans le RFC 5474 comme par exemple l'échantillonnage fondé sur le comptage des paquets ou celui fondé sur le temps.

La section 5 est consacrée à l'étude détaillée des techniques d'échantillonage. En fonction du phénomène qu'on veut étudier, on utilisera telle ou telle technique. Par exemple, l'échantillonage systématique (section 5.1) est déterministe et se décline en échantillonage systématique fondé sur le rang des paquets (on sélectionne un paquet sur N) et en échantillonage systématique fondé sur le temps (on prend tous les paquets, mais pendant une courte période).

L'échantillonage aléatoire est décrit dans la section 5.2 et lui aussi se décline en de nombreuses variantes, par exemple selon la distribution utilisée.

1. Pour voir le RFC de numéro NNN, https://www.ietf.org/rfc/rfcNNN.txt, par exemple https://www.ietf.org/rfc/rfc5474.txt

La section 6 parle, elle, du filtrage. Contrairement à l'échantillonage, le filtrage considère que le contenu du paquet est important, mais pas sa place dans le temps. Dans le cas du filtrage fondé sur les propriétés (section 6.1), le RFC 5475 permet d'utiliser comme critères de filtrage les attributs d'IPFIX (RFC 7012). Ces attributs (un sous-ensemble de ce qui est disponible, par exemple dans le BPF) sont combinables avec un opérateur ET (pas de OU en standard, par contre). Dans le filtrage fondé sur une fonction de hachage (section 6.2), le contenu du paquet n'est pas utilisé directement mais via cette fonction. Un des buts est, par exemple, de coordonner la sélection des mêmes paquets en différents points d'observation. Le RFC détaille ce filtrage, les différentes utilisations possibles et les propriétés souhaitables pour ls fonctions de hachage (comme la vitesse). Les programmeurs noteront avec plaisir que l'annexe A contient plusieurs mises en œuvre d'une « bonne » fonction de hachage, dont une optimisée pour IPv4, où le résultat n'a pas de lien évident avec le contenu du paquet. (Il existe aussi une étude comparative de ces fonctions dans "*A Comparative Experimental Study of Hash Functions Applied to Packet Sampling*" <http://www.research.att.com/~duffield/papers/31-085A.pdf>.)

Les différentes techniques de sélection sont résumées par un tableau synthétique dans la section 7.

Notez, hélas, que certains prétendent <https://datatracker.ietf.org/ipr/558/> avoir breveté certaines de ces techniques. Il s'agit probablement de brevets futiles, comme 99,9 % des brevets logiciels mais on ne sait jamais...