RFC 3168 : The Addition of Explicit Congestion Notification (ECN) to IP

Stéphane Bortzmeyer
<stephane+blog@bortzmeyer.org>

Première rédaction de cet article le 21 mai 2008

Date de publication du RFC : Septembre 2001

https://www.bortzmeyer.org/3168.html

De tout temps, un des plus gros problèmes de l'Internet a été de gérer la **congestion**, le fait qu'il y a plus de paquets IP qui veulent passer par les tuyaux que de capacité pour les faire passer. La réponse classique d'IP à la congestion est de laisser tomber certains paquets, les protocoles de plus haut niveau étant alors censés ralentir en réponse à ces pertes. Notre RFC normalise une autre méthode : lorsque le routeur sent que la congestion est proche, il marque les paquets IP pour indiquer aux machines terminales qu'il faudrait ralentir. Une excellente page, par une des auteures du RFC, "*ECN (Explicit Congestion Notification) in TCP/IP*" <http://www.icir.org/floyd/ecn.html> décrit ce système, qui n'a malheureusement pratiquement pas connu de déploiement.

C'est le RFC 2481 [1] qui avait proposé pour la première fois ce mécanisme (et notre RFC le remplace et passe du statut « expérimental » au statut « chemin des normes »). La méthode classique, laisser tomber des paquets, rappelée dans la section 1 du RFC, a toujours bien marché. Le réseau est traité comme une boîte noire, on y envoie des paquets et la seule signalisation reçue est le fait que certains paquets ne ressortent pas. Les protocoles de plus haut niveau, comme TCP (RFC 793) ou DCCP (RFC 4340) ralentissent alors leur débit. Mais cela fait perdre des données, juste pour envoyer une indication. D'où l'idée de prévenir **avant** que la congestion ne soit réellement installée, en modifiant deux bits de l'en-tête IP, pour indiquer que le routeur a du mal à suivre. Les machines terminales, si elles gèrent ECN, vont alors ralentir leur débit de manière plus efficace que si elles avaient attendu en vain un paquet ("*Active Queue Management*", décrit dans le RFC 7567 et dans la section 4 de notre RFC 3168).

Avec la section 5, commence la description des changements effectués dans IP. Deux bits de l'en-têtes sont réservés pour ECN. Ils peuvent prendre les valeurs 01, 10 (ces deux valeurs, dites ECT, indiquant que l'expéditeur du paquet comprend ECN, l'intéressante section 20 explique pourquoi il y en a deux),

1. Pour voir le RFC de numéro NNN, https://www.ietf.org/rfc/rfcNNN.txt, par exemple https://www.ietf.org/rfc/rfc2481.txt

00 (pas d'ECN accepté) et 11 (dite CE, cette valeur indiquant la congestion). Ces deux bits se placent juste après le DSCP (l'ancien TOS) du RFC 2474. Si le routeur détecte que la congestion est proche, par exemple parce que ses files d'attente de sortie se remplissent, et que le paquet IP indique que l'expéditeur comprend ECN, alors le routeur peut, au lieu d'attendre la congestion, mettre un bit à 1 pour indiquer le problème au destinataire.

La section 6 du RFC décrit ensuite ce que le protocole de transport au dessus d'IP devrait faire avec les bits ECN (seul TCP est couvert dans ce RFC). 6.1 couvre TCP et modifie donc la section 3.1 du RFC 793. Deux nouveaux bits sont réservés dans l'en-tête TCP, permettant aux deux machines qui se parlent en TCP de se prévenir qu'elles comprennent ECN et, si nécessaire, d'indiquer le début de congestion, détecté en regardant les paquets IP modifiés par le routeur. Une fois ECN ainsi « négocié », les paquets IP porteront le bit ECN adapté, comme on le voit avec tcpdump :

```
12:04:58.525094 IP (tos 0x2,ECT(0), ttl 64, id 2069, offset 0, flags [DF], proto TCP (6), length 142) 192.13
12:04:58.525242 IP (tos 0x0, ttl 64, id 61994, offset 0, flags [DF], proto TCP (6), length 52) 192.134.4.69.
```

ECT(0) désignant le motif 10 (« je comprends ECN »). Les paquets de pur acquittement TCP (le deuxième paquet affiché ci-dessus) ne portent pas de bits ECN.

Les sections 8 à 11 du RFC discutent de détails pratiques et d'évaluation d'ECN. Les section 18 et 19 sont consacrées au cas où une "*middlebox*" dans le réseau modifie maladroitement les bits ECN. Les sections 21 et 22 ne normalisent rien mais expliquent les choix qui ont été fait dans les en-têtes de paquets, en revenant sur les anciennes définitions de ces en-têtes.

L'Internet étant infecté de machines non gérées et non modifiables, comme certains routeurs bas de gamme (les CPE, par exemple), il est très difficile de déployer de nouveaux protocoles ou de nouveaux services. Ainsi, sept ans après sa normalisation formelle, ECN ne passe toujours pas avec certains sites <http://www.icir.org/floyd/ecnProblems.html>, où les paquets marqués sont jetés. La section 6.1.1.1 du RFC discute ce problème (voir aussi <http://www.icir.org/tbit/ecn-tbit.html>).

C'est pour cela que beaucoup de systèmes d'exploitation viennent avec ECN coupé par défaut. Par exemple, sur Linux, on peut voir si ECN est accepté avec :

```
 % sysctl net.ipv4.tcp_ecn
net.ipv4.tcp_ecn = 0
```

et les développeurs discutent régulièrement de son activation par défaut <http://lkml.org/lkml/2008/11/4/151>. FreeBSD, lui, ne semble pas avoir de support ECN du tout ? Un guide complet de configuration pour les Cisco se trouve en <http://www.cisco.com/en/US/docs/ios/12_2t/12_2t8/feature/guide/ftwrdecn.html>. Pour MPLS, on peut consulter le RFC 5129. Autre lecture possible, une proposition d'activer ECN par défaut <https://trammell.ch/2015/03/making-the-int>.

Depuis des années, des utilisateurs demandent un support de ECN dans echoping <http://sourceforge.net/tracker/index.php?func=detail&aid=670565&group_id=4581&atid=354581> mais ça reste à développer. Il n'est pas évident que ECN doive être sous le contrôle des applications.