RFC 7713 : Congestion Exposure (ConEx) Concepts, Abstract Mechanism and Requirements

Stéphane Bortzmeyer
<stephane+blog@bortzmeyer.org>

Première rédaction de cet article le 6 janvier 2016

Date de publication du RFC : Janvier 2016

https://www.bortzmeyer.org/7713.html

Le projet **ConEx** de l'IETF vise à développer des mécanismes permettant à l'émetteur de paquets IP de prévenir le réseau situé en aval que ce flot de données rencontre de la congestion. Les routeurs pourront alors éventuellement prendre des décisions concernant ce flot. Ce nouveau RFC expose le mécanisme abstrait de signalisation (le mécanisme concret, dans des protocoles comme TCP, a été normalisé plus tard).

La spécification se fait donc en trois temps, décrire le problème et les scénarios d'utilisation (RFC 6789 [1]), décrire un mécanisme de signalisation de la congestion abstrait, sans se soucier des détails techniques (ce RFC 7713), puis enfin normaliser le protocole concret (RFC 7786 et RFC 7837), avec les inévitables compromis que cela implique.

Aujourd'hui, les équipements réseau comme les routeurs signalent la congestion vers l'aval, en utilisant ECN (RFC 3168), en retardant les paquets (car les files d'attente mettent du temps à se vider) ou, tout simplement, en laissant tomber des paquets. Ce signalement arrive au récepteur qui peut alors informer l'émetteur, par exemple en réduisant la fenêtre, dans TCP. Cette boucle de rétroaction préserve l'Internet de la congestion (RFC 5681). (De ces trois signaux, le temps d'acheminement est le moins utilisé, car il n'est pas un indicateur univoque de la congestion.)

Mais ce signalement n'est pas visible de tous les équipements réseau traversés par ce flot de paquets. Ceux situés en amont du premier routeur qui détecte la congestion ne sont pas prévenus, alors qu'ils auraient pu jouer un rôle. Le rôle de ConEx est justement que tout le monde soit informé. Le principe

1. Pour voir le RFC de numéro NNN, https://www.ietf.org/rfc/rfcNNN.txt, par exemple https://www.ietf.org/rfc/rfc6789.txt

de base est que l'émetteur, une fois prévenu de la congestion, mette dans les paquets qu'il envoie un indicateur que ce flot rencontre de la congestion quelque part sur le trajet.

Pour que ConEx soit déployé, il faudra que les gens y trouvent un avantage, surtout au début où le déploiement sera limité. Ce problème est commun à tous les nouveaux protocoles mais il sera encore plus crucial pour un protocole qui permettra de signaler les gros consommateurs... Ceux-ci pourraient se retrouver pénalisés (« vos flots signalent tous de la congestion, vous devriez ne les lancer qu'aux heures creuses ») ou poussés à utiliser des protocoles moins gourmands comme LEDBAT (RFC 6817).

ConEx semblait moins nécessaire autrefois car les protocoles comme TCP étaient conçus pour limiter la congestion, en se calmant dès qu'ils la détectaient. Un problème jamais résolu était celui de machines qui ignorent délibérement cet objectif et, par exemple, utilisent des protocoles de transport qui maximisent le débit individuel et tant pis pour le reste de l'Internet. Difficile de lutter contre cet égoïsme. Il pourrait être contagieux puisque les bons citoyens se trouveraient alors défavorisés, et seraient donc tentés de suivre la voie égoïste à leur tour, poussant les gros consommateurs à faire des protocoles encore plus agressifs, etc. L'IETF a toujours découragé ces pratiques (comme celle d'ouvrir plusieurs connexions TCP simultanées pour avoir une plus grosse part du gâteau) avec un succès variable.

ConEx est une approche complémentaire : les gens peuvent consommer de la capacité réseau mais ils doivent le signaler.

Comme une des conséquences de ce signalement pourrait être un traitement différencié du flot en question (« il encombre le réseau, je le limite »), une partie importante du projet ConEx est de s'assurer qu'il n'y aura pas de triche (par exemple d'émetteur « négligeant » d'informer le réseau des problèmes qu'il pose). Les signaux ConEx doivent donc être auditables, pour détecter les tricheurs. (Voir Briscoe, « "*Re-feedback : Freedom with Accountability for Causing Congestion in a Connectionless Internetwork*" `<http://discovery.ucl.ac.uk/16274/>` »).

Une dernière chose avant d'attaquer le cahier des charges exact de ConEx, le vocabulaire :

— Transport qui ne peut **pas** faire du ConEx ("*Not-ConEx*") : un mécanisme de transport des données qui ne connait pas du tout ConEx, ce qui est le cas de la totalité des mécanismes actuels.
— Transport qui peut faire du ConEx ("*ConEx-capable*") : les futurs mécanismes de transport qui sauront faire du ConEx.
— Signal ConEx : quelque chose dans un paquet (transporté par un mécanisme capable de faire du ConEx) qui indique soit une perte de paquets (signal `Re-Echo-Loss`), soit une marque ECN (signal `Re-Echo-ECN`), soit que l'émetteur s'attend à déclencher de la congestion bientôt, par exemple parce qu'il démarre et teste la capacité du réseau (signal `Credit`), soit qu'il n'a rien à signaler, à part qu'il sait faire du ConEx (signal `ConEx-Not-Marked`).

La section 3 de notre RFC décrit les exigences précises de ConEx. D'abord, pour les signaux :

— Le signal ConEx doit être visible par les éléments intermédiaires du réseau comme les routeurs. Il doit donc être mis dans l'en-tête IP, pas enfoui au fin fond du paquet. Il ne doit **pas** être modifié par ces éléments intermédiaires, l'émetteur met le signal, le reste du réseau ne fait que regarder.
— Comme le déploiement ne sera pas instantané (et qu'il est très probable, quand on voit le déploiement très limité d'ECN, qu'il ne sera jamais complet), il est important que ConEx soit utilisable, et ait des bénéfices, même en cas de déploiement partiel.
— Le signal doit être rapidement émis, pour bien rendre compte de la situation actuelle. Le RFC reconnait qu'il faudra évidemment au moins un RTT avant l'émission du signal (et davantage avec RTP, cf. RFC 3550 et RFC 6679).
— Le signal doit être exact (évidemment...) **et** auditable, ce qui veut dire qu'un observateur extérieur doit pouvoir le vérifier, afin de détecter d'éventuels tricheurs.

Ces exigences sont parfois contradictoires et le RFC note qu'il faudra sans doute, en passant de ces abstractions à un protocole concret, accepter quelques compromis.

J'ai parlé de l'importance d'auditer les signaux ConEx. La fonction d'audition n'est pas normalisée dans ce RFC mais il pose des exigences qu'elle doit suivre, notamment :

— Le moins de faux positifs possibles (flots honnêtes notés à tort comme tricheurs).
— Le moins de faux négatifs possibles (tricheurs non détectés). Je note personnellement que ces deux premières exigences sont souvent en conflit...
— Des sanctions suffisantes pour être dissuasives mais proportionnelle à la triche : en raison des faux positifs, des sanctions excessives n'encourageront pas à déployer le protocole (et pourraient représenter un vecteur d'attaque par déni de service).

Ce RFC décrit des mécanismes abstraits. Le futur travail du groupe de travail ConEx `<https://tools.ietf.org/wg/conex>` devra en faire un ensemble de mécanismes concrets, traitant entre autres de ces points :

— L'encodage des signaux.
— Les mécanismes d'authentification et d'intégrité.
— Les techniques de coexistence entre des machines ConEx et non-ConEx.
— L'extensibilité du protocole.

Là encore, il ne sera sans doute pas possible de tout satisfaire. Les futurs RFC ConEx devront donc noter quelles exigences ont été délibérement affaiblies, voire abandonnées.(Cf. les RFC 7837 et RFC 7786.)

Place à l'encodage, maintenant, pour satisfaire mes lecteurs qui préfèrent savoir à quoi vont ressembler les bits sur le câble. Je rappelle que le protocole précis n'est pas encore normalisé, cette section 4 du RFC ne fait que pointer les problèmes liés à l'encodage des signaux. Prenons en effet un encodage naïf : on décide d'un bit dans l'en-tête IP qu'on met à 1 dès qu'il y a eu une retransmission TCP (des données sans accusé de réception, qu'il a fallu réémettre), ou dès qu'on a réduit la fenêtre TCP en réponse à ECN. Cet encodage semble satisfaire les premières exigences de ConEx : il est trivial à réaliser (sans doute une seule ligne de code dans la plupart des mises en œuvre de TCP/IP), il est compatible avec les matériels et logiciels existants (les non-ConEx ignoreront tout simplement ce bit), tout routeur ConEx sur le trajet peut lire ce bit et agir en fonction de sa valeur.

Mais cet encodage trop simple a des défauts : comme dans le cas du RFC 3514, il ne permet pas un audit. On n'a aucun moyen de vérifier si l'émetteur triche ou pas. Bon, ne soyons pas trop critique : un tel encodage est utile pour aider à comprendre ConEx et à se représenter à quoi cela peut bien ressembler. Dans des environnements fermés où tout le monde est honnête, il peut suffire.

D'autres encodages sont imaginables. Par exemple, on pourrait ne rien encoder du tout et dire aux routeurs ConEx de regarder les flots TCP et d'en déduire s'il y a eu rencontre avec la congestion en aval ; en effet, si on voit des retransmissions, ou la fenêtre TCP se réduire, on peut en déduire que l'émetteur a vu de la congestion et a réagi. Cela implique que les routeurs analysent TCP (ce qu'ils ne font normalement pas) et gardent un état (au moins un RTT de données), ce qui serait coûteux. Mais cela dispenserait de toute modification des émetteurs. Et cela ne permet pas de triche.

Dans un routeur implémenté en logiciel, en bordure du réseau, cela pourrait être réaliste (dans un routeur de cœur traitant des milliards de paquets à la seconde, cela le serait moins). À noter que des protocoles de sécurité comme TLS ou SSH ne masquent pas TCP et ne seraient donc pas un problème pour une analyse ConEx (contrairement à IPsec avec ESP, mais qui est beaucoup moins répandu). Mais si ConEx était ainsi déployé, cela pourrait motiver certains pour tout faire passer dans un VPN qui empêcherait cette observation.

Et encoder avec ECN ? Il existe une proposition ("*Internet-Drafts*" `draft-briscoe-conex-re-ecn-tcp` et `draft-briscoe-conex-re-ecn-motiv`) d'intégration d'ECN avec ConEx, qui a l'avantage d'empêcher la triche.

L'inconvénient de l'approche précédente est de nécessiter ECN, que n'ont pas tous les récepteurs. Une approche purement ConEx à l'encodage serait d'avoir des bits dédiés à ConEx dans l'en-tête IP (ou dans un en-tête d'extension en IPv6). On peut utiliser un bit par signal ConEx (`ConEx`, `Re-Echo-Loss`, `Re-Echo-ECN` et `Credit`), ou bien essayer de profiter du fait que certains sont mutuellement exclusifs pour condenser un peu.

Un sous-problème intéressant est celui de savoir si on compte la congestion en bits ou bien en paquets ? Est-ce que la perte d'un paquet de 1 500 octets vaut celle d'un paquet de 64 octets ? La perte d'un gros paquet est-elle due à sa taille ou bien un petit paquet aurait-il été également jeté ? Certaines parties du réseau sont limitées en nombre de bits (c'est typiquement le cas des tuyaux), d'autres en nombre de paquets (c'est partiellement le cas des équipements actifs). Le RFC 7141 recommande d'utiliser plutôt les bits, le RFC 6789 suggère la même chose, au moins dans l'Internet actuel.

La section 5 du RFC décrit les composants du réseau qui auront un rôle à jouer dans ConEx (voir aussi la section 6). Les équipements intermédiaires (notamment les routeurs) d'aujourd'hui ignorent les signaux ConEx et passent les paquets tels quels (espérons qu'aucun pare-feu trop zélé ne se permettra de les bloquer <`https://www.bortzmeyer.org/options-interdites.html`>). Ceux qui seront modifiés ConEx, soit des routeurs, soit des engins spécialisés ("*congestion policer*" ?) liront les signaux ConEx, les vérifieront (cf. la partie sur l'audit) et pourront agir, par exemple en "*shapant*" les émetteurs qui déclenchent de la congestion.

Les émetteurs, typiquement des machines terminales, devront voir leur code TCP/IP modifié pour émettre des signaux ConEx. A priori, ceux qui n'ont même pas encore ECN n'utiliseront pas ConEx non plus, mais ceux qui ont ECN trouveront peut-être l'ajout de ConEx intéressant.

Les récepteurs, typiquement des machines terminales, n'ont pas à être modifiés.

L'audit des signaux ConEx, pour détecter les tricheurs (pensez au RFC 3514) est évidemment essentiel (section 5.5). Si des équipements réseaux limitent le trafic des émetteurs qui déclenchent de la congestion, afin d'épargner le réseau aval, ces émetteurs auront évidemment un bon mobile pour tricher. Ils risquent de ne pas mettre les signaux ConEx. (Voir aussi la section 8, sur la sécurité.)

Comment détecter les tricheurs ? Une solution possible est de détecter les pertes de paquets et de voir si l'émetteur envoie bien des `Re-Echo-Loss`. Mais attention : rien ne dit que l'auditeur voit tous les paquets d'un flot. Si le trafic passe par deux liens différents, l'auditeur situé sur un des liens risque de ne pas voir certains paquets et croire à tort qu'ils sont perdus.

Les signaux ECN peuvent aussi servir à l'audit. Pour être sûr de les voir, il faut être en aval des points où il y a congestion.

Une autre solution est d'utiliser les retransmissions de TCP. Si l'émetteur réémet, c'est que que paquets ont été perdus. Là, il vaut mieux être proche de l'émetteur (donc le plus en amont possible), pour être sûr de voir tous ses paquets, même en cas de chemins multiples dans le réseau.

Le routeur qui doit jeter des paquets (ou les marquer avec ECN) car il n'arrive plus à tout transmettre est aussi un bon endroit pour faire de l'audit : il peut directement comparer ce qu'il fait (jeter les paquets d'un flot donné) avec l'apparition, un RTT plus tard, des signaux ConEx.

La thèse de B. Briscoe (un des auteurs du RFC), « "*Re-feedback : Freedom with Accountability for Causing Congestion in a Connectionless Internetwork*" <`http://discovery.ucl.ac.uk/16274/`> » contient une analyse détaillée de toutes les techniques de triche, et d'audit, connues.