RFC 6057 : Comcast's Protocol-Agnostic Congestion Management System

Stéphane Bortzmeyer
<stephane+blog@bortzmeyer.org>

Première rédaction de cet article le 8 décembre 2010

Date de publication du RFC : Décembre 2010

https://www.bortzmeyer.org/6057.html

Pour tout FAI, la gestion de la congestion est un problème récurrent. Même avec de très gros tuyaux, les utilisateurs, avides de profiter de leur abonnement à l'Internet (et poussés par des évolutions comme le remplacement de cours sous forme texte par des conneries multimédia genre "*webinar*"), envoient de plus en plus de paquets. Beaucoup de FAI, lors de débats sur la neutralité du réseau défendent leur droit à « gérer la congestion » par tous les moyens, même les plus inavouables. C'est ainsi que l'un des plus gros FAI états-uniens, Comcast, s'est rendu célèbre en usurpant l'adresse IP de ses propres abonnés pour couper leurs sessions BitTorrent fin 2007 <http://www.eff.org/wp/packet-forgery-isps-report-comcast-af Ces méthodes de cow-boys ont suscité beaucoup de protestation et, dans ce RFC, Comcast annonce un changement de cap et l'adoption d'un nouveau système de gestion de la congestion, neutre ("*protocol agnostic*" car « neutre » est un gros mot pour les FAI) et fondé sur des mesures objectives (le nombre d'octets envoyés ou reçus). Ce système est mis en œuvre dans le réseau de Comcast depuis la fin de 2008.

Comcast, qui connecte des dizaines de millions de foyers états-uniens, notamment par le câble (des détails sur leur réseau figurent en section 7), s'est donc fait allumer par la FCC au sujet de ses curieuses pratiques de gestion du réseau, qui consistaient à générer, depuis un boîtier Sandvine, des faux paquets TCP de type `RST` ("*Reset*") prétendant venir des pairs, de manière à couper les sessions pair-à-pair (voir le rapport très détaillé de l'EFF <http://www.eff.org/files/eff_comcast_report.pdf> ou son résumé à Ars Technica <http://arstechnica.com/old/content/2007/11/eff-study-reveals-evidence-of-comcasts-bittorrent-interference.ars>). Ce mécanisme visait, de manière complètement arbitraire, un protocole réseau particulier, BitTorrent.

La section 1 du RFC resitue le problème de la congestion. TCP est un protocole gourmand : tant qu'il n'y a pas de perte (cf. RFC 7680 [1]) de paquets, il va augmenter le débit jusqu'à ce que le réseau ne puisse

1. Pour voir le RFC de numéro NNN, https://www.ietf.org/rfc/rfcNNN.txt, par exemple https://www.ietf.org/rfc/rfc7680.txt

plus suivre et le signale en laissant tomber certains paquets (TCP réduira alors le débit). La congestion est donc un état normal du réseau (sinon, cela veut dire qu'il est sous-utilisé).

Le mécanisme déployé étant spécifique au réseau de Comcast, il est utile, pour suivre le RFC, de bien apprendre la section 3 sur la terminologie, notamment si on n'est pas habitué aux réseaux par câble TV coaxial. Ainsi, un modem câble est le CPE, l'engin placé chez le client, et un CMTS ("*Cable Modem Termination System*", cf. section 2.6 du RFC 3083) est l'équipement situé dans les locaux du FAI où sont connectés les clients (à peu près l'équivalent d'un DSLAM pour les FAI ADSL). Le tout fonctionne grâce à la norme DOCSIS dont les documents sont disponibles en <http://www.cablelabs.com/>. D'autre part, la rubrique terminologique en section 3 référence également les RFC sur la qualité de service, les RFC 1633 et RFC 2475.

Le nouveau système de contrôle de la congestion a évolué, comme le résume la section 4, en partie en suivant les débats à l'IETF comme l'atelier de 2008, qui a été documenté dans le RFC 5594. La contribution de Comcast à cet atelier consistait en un document « "*Service Provider Perspective*" <http://trac.tools.ietf.org/area/rai/trac/raw-attachment/wiki/PeerToPeerInfrastructure/02-Comcast-I pdf> ». Le système décrit par Comcast à l'occasion de cet atelier (et qui avait été également présenté dans les réponses de Comcast à la FCC, qui était mécontente du système de piratage des sessions BitTorrent) a été progressivement déployé à partir de la fin de 2008.

En quoi consiste ce système de contrôle de la congestion ? La section 5 résume ses principes. Plusieurs éléments du réseau étant partagés, le but du système est de partager « équitablement » ces éléments (le terme « équitablement » est entre guillemets dans le RFC car il n'est pas évident d'en donner une définition rigoureuse). D'autre part, le système actuel est agnostique (le RFC aurait pu dire « neutre » mais la plupart des FAI ne veulent pas entendre parler du concept de neutralité), c'est-à-dire qu'il ne cible pas un protocole particulier, il tient juste compte d'un facteur objectif : le débit de la ligne de l'abonné. L'algorithme (simplifié) est donc le suivant : le logiciel surveille en permanence l'usage des ressources réseau. Si l'une de celles-ci devient congestionnée, le logiciel regarde quel client sur cette ressource consommait plus que les autres et lui affecte une priorité plus basse. Cela n'aura pas de conséquence pratique si les lignes ont de la capacité suffisante mais, si la congestion empêche de faire passer tous les paquets, ceux de ce(s) client(s) seront retardés, voire jetés (cela peut sembler violent mais c'est le mécanisme normal d'IP pour gérer la congestion ; les protocoles de transport savent réagir à ce problème, par exemple TCP réduit automatiquement son débit). Si le client diminue son usage, sa priorité redevient normale.

Ne serait-ce pas préférable, en cas de congestion, d'augmenter la capacité du réseau ? Certes, dit la section 6, mais le problème est plus compliqué que cela. D'abord, l'ajout de capacité nécessite souvent des travaux matériels, qui prennent du temps. Ensuite, quelle que soit la capacité du réseau, il y aura des pics de trafic imprévus et aucun réseau ne peut être assez dimensionné pour éviter complètement la congestion. (De même qu'aucun autoroute ne peut être assez large pour un samedi 1er août.) Pouvoir gérer la congestion est donc une nécessité.

Quelles sont les valeurs numériques exactes utilisées pour des opérations comme « abaisser la priorité » ? L'abaisser de combien ? Et que veut dire exactement « proche de la congestion » ? 90 % 95 % ? La section 7 du RFC traite de l'implémentation concrète des principes de la section 5. Elle commence par un excellent résumé de l'architecture du réseau de Comcast, où 3 200 CMTS servent environ 15 millions d'utilisateurs. Le réseau, comme illustré dans la figure 1 du RFC, est en fait mixte, les CMTS étant connectés en fibre optique, seul le dernier "*mile*" étant en câble de cuivre coaxial. La section 7.1 définit ensuite rigoureusement la métrique utilisée pour déclarer qu'un CMTS approche de la congestion. Ensuite, la valeur numérique a été déterminée par des essais en laboratoire : aujourd'hui 70 % d'utilisation en montée et 80 % en descente pendant 5 minutes de suite. La section 7.2 traite de la seconde partie de

l'algorithme : définir ce que signifie, pour un utilisateur, « consommer plus que sa part ». (Les chiffres exacts dépendent de l'abonnement qu'il a souscrit : les riches ont plus que les pauvres.)

Ce RFC n'hésite pas devant les chiffres précis. Ainsi, un exemple donné est celui d'un abonné à un service à 22 Mb/s descendants. Malgré plusieurs usages simultanés (un flux vidéo HD depuis Hulu à 2,5 Mb/s, un appel Skype à 131 kb/s, et un flux musical à 128 kb/s), il reste en dessous de la limite. L'idée est donc que beaucoup d'utilisations, même multimédia, n'amèneront pas à la limite. Lors de tests avec de vrais utilisateurs, par exemple à Colorado Springs en 2008, 22 utilisateurs sur 6 016 ont été « déprioritisés ». Lors de tels tests, quel a été le ressenti des utilisateurs ? La section 7.3 note simplement que personne ne s'est plaint <https://www.bortzmeyer.org/personne-ne-s-est-plaint.html>.

La section 9 argumente même que ce mécanisme, prévu pour gérer la minorité la plus gourmande des utilisateurs, est aussi utile en cas de congestion plus globale. Par exemple, si une brusque épidémie de grippe se développe, forçant l'arrêt des transports publics et amenant de nombreuses entreprises à fermer, le télétravail depuis la maison va brusquement augmenter. Le réseau tiendra-t-il ? En tout cas, le même mécanisme de gestion de la congestion peut être utile.

Rien n'étant parfait en ce bas monde, quelles sont les limites de ce système ? La section 10 en cite quelques unes. Entre autres, le mécanisme de Comcast ne signale pas à l'utilisateur qu'il a été déprioritisé. Cela empêche ses applications de s'adapter (par exemple en basculant vers une résolution vidéo plus faible) et l'utilisateur de changer son comportement.

D'autres mécanismes de contrôle de la congestion pourraient apparaître plus tard, par exemple issus des travaux de l'IETF, que la section 11 résume. Des groupes de travail comme Conex <http://tools.ietf.org/wg/conex> (signalisation explicite de la congestion), Alto <http://tools.ietf.org/wg/alto/> (recherche du meilleur pair dans un réseau pair-à-pair) ou Ledbat <http://tools.ietf.org/wg/ledbat> (transfert de grosses quantités de données en arrière-plan, en se glissant dans les moments où le réseau est libre) produiront peut-être des solutions meilleures (voir le RFC 6817 pour Ledbat).

Enfin, la section 12, sur la sécurité, est une lecture intéressante : à chaque fois qu'on construit un tel appareillage, il faut se demander s'il ne risque pas d'être mal utilisé. Ici, la principale crainte est le risque d'injection de fausses données, pour bloquer les utilisateurs (une attaque par déni de service). Les équipements de statistiques doivent donc être protégés contre un accès non autorisé. Moins grand est le risque pour la vie privée : si le trafic par modem (donc par foyer) est stocké, cela n'inclut pas les adresses IP de destination, les ports, etc.

Arrivé au terme de ce très intéressant document, très détaillé (il faut noter qu'aucun FAI dans le monde n'a fait un tel effort de documentation, comparez par exemple avec le silence complet que maintient Free face aux accusations <http://www.universfreebox.com/article11954.html> de "*shaping*"), quel bilan tirer ? D'abord, que les protestations des clients, de l'EFF et les menaces de la FCC ont eu un effet positif. Ensuite, je pense personnellement que le système est bon dans son principe. Face à la congestion, un problème auquel peut être confronté tout réseau, quelles que soient les dimensions de ses tuyaux, il est nécessaire de prendre des mesures. Celles-ci doivent être publiquement exposées (et, on l'a dit, Comcast est le seul à le faire) et être non-discriminatoires, fondées sur des problèmes objectifs (le débit dans les tuyaux) et pas sur les intérêts financiers du FAI (comme les opérateurs 3G qui interdisent la VoIP mais qui autorisent la vidéo ; cette dernière consomme pourtant bien plus de ressources mais, elle, elle n'empiète pas sur le "*business*" traditionnel de l'opérateur...). Donc, le principe du système de gestion de la congestion de Comcast est bon. Maitenant, tout est dans l'exécution en pratique. Le système effectivement déployé est-il celui décrit ? Les évolutions futures respecteront-elles les principes posés ? Poser la question n'est pas de la paranoïa. Comcast a déjà menti, par exemple en niant la création de faux TCP resets même si la section 8 du RFC reconnait aujourd'hui leur usage (cf. RFC 3360).

Le lecteur des aventures de Thursday Next <https://www.bortzmeyer.org/jeudi-prochain.html> pensera certainement au tome 4, « "*Something rotten*" », où la tentaculaire société Goliath tente de faire croire qu'elle est devenue une église n'agissant plus que pour le bien commun...

Ce RFC a été reclassé « intérêt historique seulement » <https://datatracker.ietf.org/doc/status-change-comcast-congestion-management-rfc6057-to-historic/> en août 2020, Comcast n'utilisant plus ce système.