Latence dans les réseaux, c'est quoi ?

Stéphane Bortzmeyer
<stephane+blog@bortzmeyer.org>

Première rédaction de cet article le 25 avril 2013

http://www.bortzmeyer.org/latence.html

Lorsqu'il s'agit de performances des réseaux informatiques, le vocabulaire utilisé, que ce soit sur les forums, sur les réseaux sociaux ou même dans les livres est en général catastrophique, flou et mélangeant tout. D'où deux articles pour préciser rigoureusement ce que sont la **latence** et la capacité <http://www.bortzmeyer.org/capacite.html>.

D'habitude, lorsque je parle d'un concept comme latence ou capacité, je mets juste un lien vers l'article de Wikipédia. Mais, ici, Wikipédia a curieusement décidé de nommer ce concept d'un terme anglais, "*lag*", un terme que je n'ai jamais entendu en français en dehors du monde du jeu vidéo en ligne. Revenons donc au terme correct, **latence**.

Pourquoi pinailler sur le vocabulaire ? Parce que, comme je l'indique plus haut, le monde de la mesure de performances est particulièrement mauvais de ce point de vue, préférant des termes flous comme « vitesse ». Ainsi, l'article de Wikipédia sur la latence explique qu'une latence élevée peut être causée par une bande passante insuffisante, ce qui n'est que très approximativement vrai. Et que le vocabulaire flou ou incorrect n'est pas innocent : cela sert par exemple aux commerciaux à vendre des produits inadaptés.

Donc, la **latence**, c'est simplement le temps que met un message à accomplir un certain trajet. On parle aussi de délai. Elle peut concerner l'aller-simple (c'est ce qu'utilise le RFC 7679 [1]) ou l'aller-retour (RFC 2681). Elle n'a pas de lien direct avec la capacité du réseau (ce que certains nomment « bande passante »).

(Avec certains protocoles comme TCP, il y a toutefois une relation subtile entre latence et capacité, voir le RFC 7323.)

1. Pour voir le RFC de numéro NNN, https://www.ietf.org/rfc/rfcNNN.txt, par exemple https://www.ietf.org/rfc/rfc7679.txt

Pourquoi est-elle une métrique importante ? Car certaines utilisations du réseau dépendent beaucoup de la latence. Par exemple, des applications interactives (SSH ou XMPP) envoient relativement peu d'octets et sont donc peu sensibles à la capacité <http://www.bortzmeyer.org/capacite.html> du réseau mais le sont beaucoup plus à la latence : lorsque j'envoie un message en XMPP, je voudrais qu'il arrive le plus vite possible, sous peine d'avoir une conversation hachée. D'autres cas d'usage, comme les transferts de gros fichiers, sont relativement insensibles à la latence (si les paramètres de TCP sont corrects).

Certains types de liens réseaux ont une latence particulièrement mauvaise (particulièrement élevée). Ainsi, un satellite géostationnaire ajoute forcément une latence de 238 ms (aller-retour vers le satellite, qui est à 36 000 km d'latitude, divisée par la vitesse de la lumière, qu'on ne peut pas augmenter). C'est une durée très longue et qui explique pourquoi il ne faut utiliser le satellite pour les liaisons Internet que lorsqu'on n'a absolument pas le choix (contrairement à ce que tentent de faire croire les commerciaux qui jouent sur le côté romantique d'une communication spatiale, voir par exemple cette publicité <http://www.topnetpro.com/vsat.php> qui parle d'un accès « rapide »).

La définition que j'ai donnée parlait de « message » et, en effet, la latence peut se mesurer dans différentes couches. Le célèbre outil ping mesure une latence aller-retour pour la couche 3 :

```
% ping6 -c 10 f.root-servers.net
...
10 packets transmitted, 10 received, 0% packet loss, time 9002ms
rtt min/avg/max/mdev = 177.097/183.951/192.332/5.940 ms
```

Ici, la latence moyenne (ping ne calcule hélas pas la médiane <http://www.bortzmeyer.org/mediane-et-moyenne.html>) est de 184 millisecondes. Notez que ping n'est pas un outil très précis : il dépend de la charge des deux machines (celle qui pingue et l'amer <http://www.bortzmeyer.org/amer-mire.html>), il ne connait pas l'instant exact où le premier bit du paquet touche le réseau (cf. RFC 7679), etc. C'est un outil approximatif, mais suffisant dans la plupart des cas. Un exemple de latence mesurée au niveau 7 est fournie par l'outil check-soa <http://www.bortzmeyer.org/check-soa-go.html> pour le DNS. Dans ce cas, la latence mesurée ne dépend pas que du réseau mais aussi de l'application distante (le serveur DNS) :

```
% check-soa -i societegenerale.com
tigdns01.socgen.com.
193.178.155.113: OK: 2013042301 (30 ms)
tigdns02.socgen.com.
193.178.155.114: OK: 2013042301 (34 ms)
```

Dernier exemple, avec HTTP, curl a une option peu connue pour afficher au format qu'on souhaite certains résultats de la connexion :

```
% curl --silent   --write-out "Delay: %{time_total} seconds\n" \
       --output /dev/null http://www.societegenerale.com/
Delay: 0.124 seconds
```

La documentation de curl indique plusieurs autres variables pour étudier les différents facteurs qui, ensemble, composent la latence :

```
% curl --silent   --write-out "Delay: %{time_total} s, TCP connection delay: %{time_connect},  Negociation delay
       --output /dev/null http://www.societegenerale.com/
Delay: 0.274 s, TCP connection delay: 0.095,  Negociation delay: 0.178
```

Ici, on voit que le transfert effectif des données ne faisait même pas la moitié de la latence : la majorité du temps était passé à faire la connexion TCP puis à envoyer la commande `GET`. Pour une page Web de plus grande taille, les résultats seraient bien sûrs différents.

Un bon article en français expliquant clairement la latence et son importance est « Bande passante et temps de latence réseau <`http://alainfaure.net/2011/03/03/bande-passante-et-temps-de-latence-res`> » d'Alain Faure. Sinon, le grand classique en anglais est « "*It's the Latency, Stupid*" <`http://rescomp.stanford.edu/~cheshire/rants/Latency.html`> » de Stuart Cheshire.