RFC 1995 : Incremental Zone Transfer in DNS

Stéphane Bortzmeyer
<stephane+blog@bortzmeyer.org>

Première rédaction de cet article le 28 septembre 2010

Date de publication du RFC : Août 1996

http://www.bortzmeyer.org/1995.html

Le mécanisme standard de transfert d'une zone DNS entre deux serveurs faisant autorité (depuis le maître vers l'esclave) est normalement le transfert de zones, dit AXFR, aujourd'hui normalisé dans le RFC 5936[1]. Ce mécanisme convient parfaitement aux « petites zones » (de quelques centaines d'enregistrement au plus) mais achoppe, par exemple, lorsqu'un « gros » TLD veut pousser les changements du jour (ou de l'heure) vers tous ses esclaves. En raison, entre autre, de l'"*anycast*", certains de ces esclaves sont situés dans des endroits pas très bien connectés (comme l'île de la Réunion) et l'envoi d'un fichier de la taille de celui de .fr (aujourd'hui, 1,8 millions de domaines <http://www.afnic.fr/actu/stats> et 190 Mo) peut prendre du temps. On peut transférer les zones par d'autres moyens que le DNS, par exemple rsync. Mais il existe une solution DNS standard (qui est celle utilisée par .fr), IXFR ("*Incremental Zone Transfer*"), normalisée dans ce RFC 1995.

Le principe est simple et résumé en section 2 du RFC. Lorsqu'un client veut un transfert (typiquement, parce qu'il est esclave et a appris que sa version de la zone était en retard), il envoie un message DNS de type IXFR (code 251, AXFR étant 252, cf. <https://www.iana.org/assignments/dns-parameters>) au maître qui lui transmet alors uniquement les nouveaux enregistrements. Le serveur IXFR (le maître) doit donc garder trace des changements entre les différentes versions, pour ne transmettre que ces changements (BIND ne le fait, par défaut, que si on utilise les mises à jour dynamiques du RFC 2136 ; autrement, il faut ajouter l'option ixfr-from-differences yes;). À noter qu'un serveur IXFR a toujours le droit de renvoyer la zone complète, par exemple si le client a un numéro de version trop vieux, ne correspondant plus à une version qui avait été gardée (voir la section 5).

Comme la réponse IXFR a des chances d'être de petite taille, le serveur a même le droit de répondre en UDP. Autrement, on utilise TCP, comme avec AXFR. Donc, l'algorithme recommandé est, pour le client, d'essayer en UDP d'abord, puis TCP.

1. Pour voir le RFC de numéro NNN, https://www.ietf.org/rfc/rfcNNN.txt, par exemple https://www.ietf.org/rfc/rfc5936.txt

La requête IXFR a le format DNS standard (section 3), la section Autorité contenant le SOA de la version courante chez le client. La réponse (section 4) ressemble beaucoup à une réponse AXFR. Elle est composée de séquences, chaque séquence commençant par l'ancien SOA, puis comportant les enregistrements supprimés, puis le nouvel enregistrement SOA, indiquant la version actuelle sur le maître, puis enfin les enregistrements ajoutés. Les séquences sont classés chronologiquement donc on peut voir la réponse IXFR comme un historique des changements. À noter que ce sont bien des enregistrements qui sont transmis, pas des "*RRsets*". Si on a :

```
foobar  IN   NS   ns1.example.net.
        IN   NS   ns3.example.net.
```

et qu'on ajoute un serveur `ns2.example.net`, seul l'enregistrement NS de ce serveur aura besoin d'être transmis, pas les trois enregistrements du nouveau "*RRset*".

La réponse commence, comme pour AXFR, par le SOA de la version locale du serveur mais le client peut savoir si sa demande IXFR a reçu une réponse IXFR ou AXFR en examinant le second enregistrement : si la réponse est incrémentale (IXFR), ce second enregistrement est un SOA.

Naturellement, le client IXFR ne doit mettre à jour sa copie locale qu'une fois qu'il a reçu tous les enregistrements. (Dit autrement, IXFR doit être atomique.)

La section 5 spécifie le comportement d'un serveur IXFR pour ce qui concerne les vieilles versions : il n'est évidemment pas obligé de les garder toutes et peut donc supprimer les plus anciennes, pour gagner de la place. C'est d'autant plus important qu'au bout d'un moment, les changements s'accumulant, la réponse IXFR deviendra plus longue que la réponse AXFR ! Remarquons aussi (section 6) que la réponse incrémentale n'est pas forcée de refléter l'histoire exacte des changements : un serveur a le droit de condenser plusieurs changements successifs en un seul. (Les exemples de la section 7 incluent une telle condensation.)

Voyons maintenant un exemple de mise en œuvre, entre deux BIND 9.7. Deux serveurs font autorité pour `.fr`. Sans IXFR, le transfert prend une demi-minute sur un Ethernet à 100 M/s (et bien plus longtemps avec des serveurs mal connectés au bout du monde ; il ne faut pas oublier que `.fr` a un serveur à Katmandou, un à Manille, etc). Sur le maître, on voit :

```
28-Sep-2010 10:54:35.388 client 192.0.2.97#56385: transfer of 'fr/IN': AXFR started
28-Sep-2010 10:55:00.162 client 192.0.2.97#56385: transfer of 'fr/IN': AXFR ended
```

Et sur l'esclave (tous les serveurs utilisent le port 9053, pour des tests, et pas le port standard 53 ; ainsi, le serveur esclave est configuré avec `masters { 192.0.2.69 port 9053; };`) :

```
28-Sep-2010 10:55:00.182 transfer of 'fr/IN' from 192.0.2.69#9053: \
                   Transfer completed: 2699 messages, 3965064 records, \
                   106151552 bytes, 24.857 secs (4270489 bytes/sec)
```

(Notez au passage que ce sont des enregistrements binaires DNS qui sont transférés, pas un fichier texte, ce qui explique la taille totale plus petite.)

Pour activer IXFR sur la maître, on modifie la configuration du serveur avec `ixfr-from-differences yes;` dans la directive `zone` :

```
zone "fr" {
        type master;
        file "fr";
        ixfr-from-differences yes;
};
```

On ne change rien sur le client IXFR : avec BIND, le client essaie IXFR par défaut. Sur le maître, le transfert est quasi-instantané :

```
28-Sep-2010 11:05:47.103 client 192.0.2.97#54496: transfer of 'fr/IN': IXFR started
28-Sep-2010 11:05:47.103 client 192.0.2.97#54496: transfer of 'fr/IN': IXFR ended
```

ce que confirme le journal de l'esclave, à qui il a suffi de transférer dix changements :

```
28-Sep-2010 11:05:47.049 transfer of 'fr/IN' from 192.0.2.69#9053: \
               Transfer completed: 1 messages, 10 records, \
               334 bytes, 0.004 secs (83500 bytes/sec)
```

(Note au passage : pour prévenir un serveur esclave de test, qui ne reçoit pas les `NOTIFY` du RFC 1996, qu'un changement a eu lieu chez le maître, le plus simple est d'envoyer un `NOTIFY` forcé. BIND ne permet pas de le faire facilement mais, si on a nsd, il suffit de faire un `nsd-notify -p 9053 -z fr NOM-SERVEUR`).

On peut aussi admirer le transfert incrémental avec tshark (l'option `-d` est nécessaire car on utilise un port alternatif <http://www.bortzmeyer.org/decoder-dns-port-alternatif.html>). Un domaine a été ajouté, un autre retiré (les deux domaines avaient le même jeu de serveur) :

```
% tshark -d tcp.port==9053,dns -d udp.port==9053,dns  -r ixfr.pcap
...
  3   0.001347 192.0.2.97 -> 192.0.2.69 DNS Standard query SOA fr
  4   0.001455 192.0.2.69 -> 192.0.2.97 DNS Standard query response SOA nsmaster.nic.fr
...
 10   0.002930 192.0.2.97 -> 192.0.2.69 DNS Standard query IXFR fr
...
 12   0.003089 192.0.2.69 -> 192.0.2.97 DNS Standard query response \
                       SOA nsmaster.nic.fr \
                       SOA nsmaster.nic.fr NS ns1.example.net NS ns3.example.net \
                       SOA nsmaster.nic.fr NS ns1.example.net NS ns3.example.net \
                       SOA nsmaster.nic.fr
```

On y voit bien le test initial du SOA, puis la requête du client, puis une séquence composée d'une partie « retrait » et d'une partie « ajouts ». Le fichier pcap complet est sur pcapr <http://www.bortzmeyer.org/pcapr.html>, en <http://www.pcapr.net/view/bortzmeyer+pcapr/2010/8/2/6/ixfr.pcap.html>.

Si le serveur refuse ou ne peut pas faire un transfert incrémental, le maître BIND indiquera :

```
28-Sep-2010 10:48:04.007 client 192.0.2.97#45524: transfer of 'fr/IN': AXFR-style IXFR started
28-Sep-2010 10:48:29.802 client 192.0.2.97#45524: transfer of 'fr/IN': AXFR-style IXFR ended
```

Et l'esclave recevra la totalité de la zone.

Les tests ici ont été faits avec BIND. Et avec nsd ? Il ne peut être qu'esclave : un maître nsd ne sait pas servir des transferts incrémentaux. Lorsque nsd est esclave, il essaie IXFR (en TCP par défaut mais on peut le configurer pour utiliser UDP) puis AXFR. (Il semble qu'il n'envoie pas de requête SOA avant la demande IXFR, la ligne 3 dans la trace Wireshark plus haut ; ce n'est en effet pas imposé par le RFC, puisque l'IXFR contient déjà le numéro de série connu du client.) On peut aussi lui demander de ne pas tenter IXFR :

```
zone:
        name: "langtag.net"
...
        request-xfr: AXFR 192.134.7.248 mykey
```

Ici, en raison du mot-clé `AXFR`, le serveur esclave ne tentera pas de faire de l'IXFR. Le même réglage, pour BIND, ne peut être que global au serveur (dans le bloc `options`, `request-ixfr no;`).

Le fait qu'un serveur puisse répondre à une demande IXFR par une copie complète de la zone peut être gênant dans certains cas. À la réunion IETF 75 de Stockholm en juillet 2009 a été présentée la proposition "*IXFR only*" ("*Internet-Draft*" `draft-kerr-ixfr-only`) qui normalisait un nouveau type de requête « IXFR seul » mais qui n'a pas encore été sérieusement pris en considération.

---

http://www.bortzmeyer.org/1995.html