Caractérisation de l'autosimilarité de trafics WEB et FTP par un outil de simulation

H.HASSAN (1) JM. GARCIA (1) O. BRUN (1) D. GAUCHARD (1)

(1) LAAS-CNRS, 7 avenue du Colonel Roche, 31077 Toulouse, France.

Résumé

L'autosimilarité est une caractéristique importante du trafic Internet. Les études effectuées sur les effets de l'autosimilarité du trafic sur les réseaux ont montré une dégradation importante de la performance des réseaux par rapport aux trafics à courte dépendance (i.e Poisson). Cet article est consacré à l'étude de l'influence des différents paramètres du trafic HTTP/FTP sur la génération d'un trafic autosimilaire. Un éditeur générique des Sources de Trafic a été développé afin de reproduire le comportement de populations d'utilisateurs et d'un ensemble d'applications : sources aléatoires quelconques, sources audio et vidéo avec la plupart des codecs, et des sources HTTP et FTP . Les simulations sont effectuées avec un outil de simulation hybride DHS (Distributed Hybrid Simulation) développé au LAAS-CNRS.

Keywords: Sources de trafic,Autosimilarité, HTTP, FTP, simulation événementielle.

1 Introduction

Les différentes études effectuées sur des réseaux LAN ou WAN, ont mis en évidence le caractère autosimilaire et LRD (long range dependency) de leur trafic. Les traces étudiées représentent le nombre de paquets ou de bits entrant dans le réseau en fonction du temps. La propriété d'autosimilarité du trafic signifie qu'une trace générée a le même comportement statistique , quel que soit l'échelle de temps choisie. L'autosimilarité (et la longue dépendance) est un caractère important du trafic. Les études sur les effets de l'autosimilarité sur les réseaux ont montré [2] que cette propriété cause la dégradation de la performance des réseaux par rapport aux trafics à courte dépendance (i.e Poisson).

Dans cette étude nous rappelons la définition de l'autosimilarité et sa mesure par le paramètre de Hurst, ainsi que la méthode R/S plot pour l'évaluation de ce dernier. Nous présentons brièvement l'Editeur des Sources de Trafic et les modèles étudiés. Nous décrivons ensuite la génération des traces Web, basées sur des modèles conçus avec l'Editeur de Source de Trafic développé à cet effet, et simulés avec l'outil de simulation DHS (Distributed Hybrid Simulation). Dans un premier temps, nous étudions les paramètres du modèle HTTP : distribution des tailles des fichiers et périodes de lectures et leur influence sur l'autosimilarité. Ensuite nous nous intéressons à l'influence des paramètres réseaux : bande passante et débit des sources. Enfin nous évaluons l'impact des gros transferts de fichiers via FTP sur l'autosimilarité du trafic résultant.

2 La propriété d'autosimilarité

Définition : Soit ($Xn, n \in \mathrm{N}$) un processus stationnaire. On note $(X_k^m)_{k\in N}$ son processus agrégé d'ordre m défini comme suit :

$$X_k^{(m)} = \frac{1}{m}(X_{(k-1)m+1} + ... + X_{km}) \tag{1}$$

Le processus $(X_n)_{n\geq 0}$ est dit autosimilaire de paramètre H si $\forall m$ entier :

$$(X_k^{(m)} = m^{1-H}(X_{(k-1)m+1} + ... + X_{km}))_{k\in N} \overset{\ell}{=} (X_n)_{n\in N} \tag{2}$$

Cette égalité est au sens de la distribution. Le paramètre de Hurst H associé à un processus autosimilaire, représente intuitivement le « Zoom ». La corrélation d'un processus autosimilaire s'exprime aussi par rapport au paramètre de Hurst. En effet, la fonction d'autocorrélation associée à un processus autosimilaire $(X_n)_{n\geq 0}$ de paramètre H est de la forme :

$$\rho_{X_n}(k) \approx_{k\to\infty} H(2H-1)k^{2H-2} \tag{3}$$

On en déduit qu'un processus $(X_n)_{n\geq 0}$ est à dépendance longue si 0.5<H<1. Plus le paramètre de Hurst est grand, plus le processus est autosimilaire. Plusieurs méthodes statistiques existent pour estimer le paramètre de Hurst. Hurst [4] a proposé

une méthode appelée analyse des étendues normalisées, ou *Rescaled Range Analysis* (*R/S Analysis*), pour évaluer la valeur du paramètre H.
Soit une série temporelle de *N* nombres, notés *x(t).* La méthode préconisée par Hurst consiste à prendre en considération une série de subdivisions indépendantes, d'effectif τ. Pratiquement, on part de séries d'effectif 10, puis 11, 12, 13, jusqu'à l'effectif le plus élevé permettant de distinguer deux subdivisions (*N*/2 ou (*N*/2)-1). Pour chaque subdivision considérée, la moyenne des τ données est :

$$E(x)_\tau = \frac{1}{\tau}\sum_{t=1}^{\tau} x(t) \tag{4}$$

Pour chaque subdivision, on centralise les données, en leur soustrayant la moyenne locale. Puis on établit une série de valeurs cumulées à l'intérieur de chaque période: on ajoute à chaque donnée la somme des valeurs centrées qui la précèdent:

$$X(t,\tau) = \sum_{u=1}^{t}\{x(u) - E(x)_\tau\} \tag{5}$$

Donc pour chaque valeur de t ($1 \le t \le \tau$), on a une valeur de X(t, τ). X(t, τ) est une distribution intégrée. On peut noter que si le signal original est un bruit blanc, X(t, τ) est un mouvement brownien. L'étendue *R* (*range*) est la différence entre le minimum et le maximum de *X(t, τ) :*

$$R = \max_{1\le t\le\tau} X(t,\tau) - \min_{1\le t\le\tau} X(t,\tau) \tag{6}$$

Cette étendue est ensuite normalisée au moyen d'une division par l'écart-type local (S(t,τ)). Enfin on procède à la pondération moyenne, par niveau d'effectif, des étendues normalisées R/S. Hurst montre que l'étendue normalisée R/S, où S représente l'écart type de la série des x(t) est liée à la grandeur de l'intervalle considéré par la relation suivante:

$$R/S = (a\tau)^H \tag{7}$$

Où « a » est une constante et H est l'exposant de Hurst. H peut être estimé par la méthode des moindres carrés, ou par l'estimation de la pente de la régression log/log de R/S sur τ.

3 Editeur des Sources de Trafic et DHS

La modélisation fiable des flux Internet est confrontée à la diversité des applications déployées sur Internet. Dans le but de modéliser ces nouvelles applications nous avons développé une bibliothèque de sources de trafic qui permet une représentation hiérarchique et générique des sources de trafic. L'Editeur des Sources de Trafic généralise le modèle ON-OFF pour décrire les périodes d'activité et inactivité des applications. La modélisation se fait en trois niveaux : Session, Application et Paquets, avec un niveau Application qui peut comporter des sous niveaux Applications pour représenter l'activité quelque soit son degré de complexité. L'Editeur permet aussi de représenter plusieurs flux en parallèle pour modéliser le trafic des protocoles de signalisation liés aux applications multimédia.
L'Editeur des Sources de Trafic génère des modèles destinés à l'outil de simulation DHS. DHS est un simulateur de réseaux IP/MPLS, basé sur la théorie de du trafic différentiel [3] et de la simulation hybride. Il permet de combiner des modèles continus avec des modèles discrets événementiels.

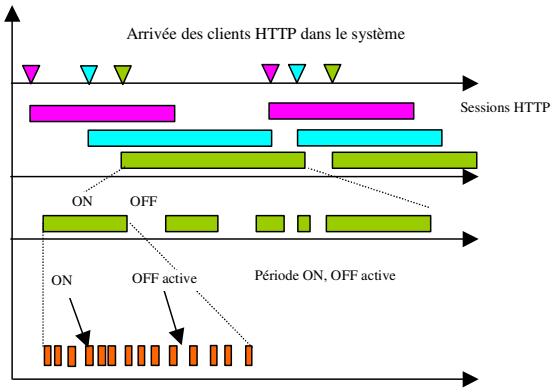

Fig. 1 – *Modèle HTTP*

Le même noyau de simulation permet de simuler tout un réseau en mode analytique, événementiel ou hybride. Dans ce dernier mode certains flux sont analytiques et d'autres sont événementiels. Dans l'ensemble des tests qui suit, nous avons utilisé DHS dans le mode événementiel. Les sources HTTP utilisées dans nos simulations comportent un niveau ON-OFF. Une période ON qui représente le téléchargement d'une page web, avec une période OFF qui représente le temps de la lecture passé par l'utilisateur. Plusieurs modèles de sources HTTP, notés W1, W2 et W3 ont été créés avec les paramètres suivants :

Modèle de source HTTP	**Distribution du nombre de pages**	**Moyenne nombre pages**	**Variance nombre de pages**	**Distribution de la taille d'une page**	**Taille Moyenne d'une page**	**Variance taille page**	**Distribution période OFF**	**Moyenne OFF**	**Variance OFF**
W1	Normale	10	5	Pareto	15 Ko	60 Ko	Pareto	30 sec	60 sec
W2	Normale	10	5	Normale	15 Ko	60 Ko	Normale	30 sec	60 sec
W3	Normale	10	5	Exponentielle	15 Ko	15 Ko	Exponentielle	30 sec	30 sec

Le modèle de source utilisé pour FTP est plus simple et comporte une seule période ON qui représente le téléchargement d'un fichier. Plusieurs sources FTP, notées F1 et F2 ont été créées avec les paramètre suivants :

Modèle de source FTP	Distribution de la taille du fichier	Moyenne	Variance
F1	Pareto	2 Mo	4 Mo
F2	Constant	1 Mo	0

- Les distributions de probabilité associées aux tailles de fichiers de FTP, ou du nombre de pages de HTTP ou de la taille des pages HTTP, ou des périodes OFF peuvent être choisies parmi de nombreuses lois : Normale, Lognormale, Gamma, Pareto Weibull, Exponentielle et constante.
- Le transport des paquets se fait par TCP new Reno implémenté de bout en bout dans le simulateur. Le débit max est limité à 1024 Kbps par connexion TCP, sauf indication contraire.
- La distribution de la taille des paquets peut aussi être paramétrée de manière quelconque. Les distributions utilisées dans la simulation sont généralement des distributions tronquées avec une valeur inférieure et supérieure ceci afin de refléter plus précisément le trafic IP. Dans nos simulations nous avons généré des paquets TCP de taille nominale de 984 octets et des ack d'une taille de 40 octets.

4 Expérimentations

Le réseau étudié est composé de deux routeurs, un routeur Source et un routeur Destination. Les deux routeurs sont reliés par un lien ayant un débit nominal de 10 Mb/sec, et un délai nul. La taille du buffer de sortie du routeur Source est de 64 paquets. L'activité des utilisateurs (clients) est représentée par des sessions WEB ou FTP selon les cas. Ces sessions démarrent au cours du temps suivant un processus de Poisson (processus de naissance de la session utilisateur). Chaque session se termine lorsque l'ensemble des données liées à cette session a été transmise (ensemble de pages http ou fichier transmis par FTP. L'ensemble des pages et la taille des pages http est aléatoire comme cela a été décrit précédemment et la taille des fichiers sous FTP est elle aussi aléatoire.
Les simulations sont réalisées sur 12000 secondes avec une granularité de 10 ms pour avoir des échantillons représentatifs du trafic.

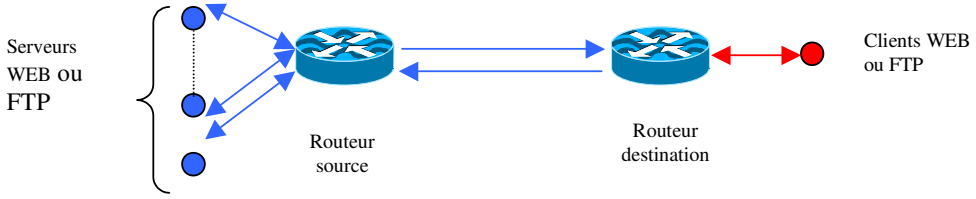

Fig. 2 – *Réseau de simulation*

Sources HTTP.
Nous avons généré les traces correspondantes aux trois modèles W1, W2, W3 séparément, avec une moyenne de 40 sessions simultanées. Comme le montre le tableau ci-contre, le modèle W1 (ON-OFF de type Pareto) a généré un trafic autosimilaire. Ce résultat est cohérent avec les études qui mettent en cause les distributions à queue lourde de taille de page [1]. Toutefois, les résultats de la simulation montrent aussi que les autres types de distribution génèrent un trafic autosimilaire bien que la taille des pages ne soit plus à queue lourde. La distribution exponentielle génère un trafic corrélé, mais moins autosimilaire que les autres distributions. Il est clair que la nature ON-OFF du trafic WEB est une caractéristique essentielle qui provoque la longue dépendance du trafic, bien plus que la distribution de taille de page ou de la période de lecture. Le réseau dans le cas précèdent est très peu chargé et les pertes sont nulles. Nous allons étudier dans la suite l'influence des pertes sur l'autosimilarité du trafic.

Modèle	Débit moyen du trafic agrégé	Nombre moyen de session	Taux d'arrivée des sessions	Durée moyenne des sessions	Perte	Hurst
W1	131 Kbps	40	0.107 sessions/sec	372.2 sec	0%	0.87
W2	128 Kbps	40	0.105 sessions/sec	378.5 sec	0%	0.85
W3	129 Kbps	40	0.092 sessions/sec	434.5 sec	0%	0.74

Sources FTP.
1) Pour analyser l'influence des pertes et du mécanisme de retransmission de TCP sur l'autosimilarité, nous avons généré la trace de la source F2 avec une moyenne d'une session ; le débit de la source TCP est de 1024 Kb/sec. Le trafic généré est très peu corrélé avec une perte nulle. Pour évaluer l'impact de la perte, nous avons réduit la bande passante du lien entre les deux routeurs (de 0.1 à 1 Mbps). Le courbe 3 trace l'évolution du paramètre de Hurst en fonction de la bande passante du lien. Nous observons une augmentation de la corrélation avec l'augmentation des pertes. Ce résultat a été vérifié également par l'augmentation du nombre moyen de session pour une bande passante fixe. Les pertes mesurées ne dépassent pas les 4% dans

le cas d'une bande passante de 0.1 Mb/sec (H=0.98), mais la corrélation du trafic devient très importante. Ce résultat montre bien que le mécanisme de retransmission de TCP et d'adaptation du débit est un facteur prédominant dans la longue dépendance du trafic.

2) Nous avons évalué l'influence des gros transferts de fichiers par FTP. Nous avons généré une source de type F2 avec une moyenne d'une session. Le trafic généré est très peu corrélé avec une perte nulle. Nous avons additionné à ce trafic l'activité de la source F1 qui représente un gros transfert de fichier. Nous avons calculé le paramètre H en augmentant le nombre moyen de session F1. La figure 4 montre que la corrélation du trafic augmente avec l'augmentation de l'activité des gros transferts. L'activité de FTP ressemble à un énorme trafic de type ON-OFF. Ce trafic influence rapidement les propriétés statistiques du trafic global pour le rendre de plus en plus autosimilaire.

Pour minimiser l'effet des gros transferts, Nous avons baissé le débit max de TCP pour les sources F1 (de 1 Mb/sec à 128 Kb/sec). On constate alors que la corrélation du trafic baisse avec la diminution du débit max. Cette corrélation reste tout de même plus importante que la corrélation du trafic sans les sources F1. Cette limitation différenciée des débits permet de limiter l'influence des gros transferts de fichier sur l'autosimilarité du trafic.

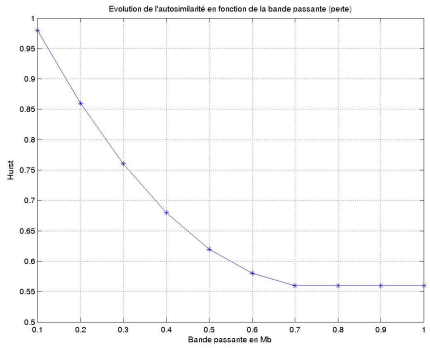

Fig. 3 – *Evolution du paramètre de Hurst avec la bande passante du lien entre les deux routeurs*

3) Nous avons appliqué cette technique aux petits flux homogènes de type W1 générés précédemment. On différencie deux cas. La limitation du débit n'a aucune influence sur l'autosimilarité dans le cas où les pertes sont nulles (les files d'attente ne se remplissent pas). Par contre, en cas de congestion (les files d'attente se remplissent), la diminution du débit des sources réduit l'autosimilarité ainsi que les pertes de paquets.

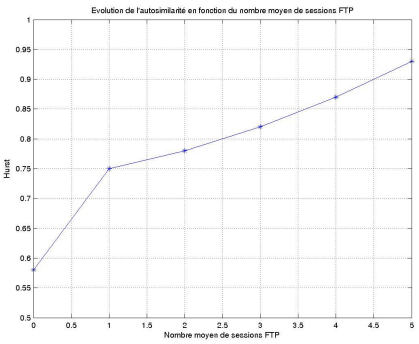

Fig. 4 – *Influence des gros transferts sur l'autosimilarité du trafic*

5 Conclusion

Nous avons étudié dans cet article l'autosimilarité du trafic Internet sur des traces générées à partir des modèles HTTP et FTP. Nous avons pu vérifier que le modèle ON-OFF avec le transport TCP produit un trafic autosimilaire. Ce résultat est valable pour toutes les distributions et non seulement pour celles à queue lourde. Le mécanisme de crédit et de retransmission de TCP génère des rafales de données qui vont être à l'origine de la dépendance longue du trafic. Dans le cas des gros transferts par FTP, nous avons montré que la diminution du débit maximum de la source TCP réduit considérablement le caractère autosimilaire du trafic. Ce résultat s'applique aussi au cas des petits flux, mais uniquement en cas de pertes. Il s'agit ici d'une diminution constante imposée à la source avant que TCP n'ai pu modifier le débit. L'autosimilarité du trafic est une caractéristique importante à prendre en compte avec la charge du réseau. Toutefois, un trafic autosimilaire peut exister sur un réseau peu chargé, sans que cela n'ait de conséquence néfaste sur les performances. Par contre la performance d'un réseau déjà chargé peut être fortement dégradée par un trafic autosimilaire. Dans ce cas, la diminution du débit des gros flux permettra de maintenir une bonne performance. Une autre alternative pourrait consister à créer plusieurs classes de flux (petits flux, moyens, gros etc.) affectées respectivement sur plusieurs files d'attente DiffServ ordonnancées par WFQ avec des taux de service appropriés.

Références

[1] S.TAQQU, W WILLINGER and R SHERMAN. Proof of a fundamental Result in Self-Similar Traffic Modeling. ACM SIGCOMM Computer Communication Review Volume 27, Issue 2. Pages: 5 – 23. 1997

[2] A. ADAS and A. MUKHERJEE. On resource management and QoS guarantees for long range dependent traffic. In Proc. IEEE INFOCOM '95, pages 779–787, 1995.

[3] J.M.GARCIA, D.GAUCHARD, O.BRUN, P.BACQUET, J.SEXTON et E.LAWLESS Modélisation différentielle du trafic et simulation hybride distribuée. Rapport LAAS No01660, 2001

[4] HURST, H.E. *Long-term storage: An experimental study*. London: Constable. 1965.

[5] D. STAEHLE, K. LEIBNITZ and P. TRAN-GIAN, Source Traffic Wireless Applications, Report N°261, June 2000, University of Würzburg

[6] R. ADDIE, M. ZUKERMAN, and T. NEAME. Fractal traffic: measurements, modelling and performance evaluation. In Proc. IEEE INFOCOM '95, pages 977–984, 1995

[7] Ian KAPLAN, Estimation of the Hurst Exponent. May 2003. http://www.bearcave.com/misl/misl_tech/wavelets/hurst/