Contrôle de congestion et gestion du trafic à partir de mesures

Nicolas LARRIEU — Philippe OWEZARSKI

LAAS – CNRS
7, avenue du Colonel ROCHE
31077 TOULOUSE Cedex 4
{nlarrieu, owe}@laas.fr

RÉSUMÉ. *Ce papier présente une nouvelle approche pour l'Internet, dont l'objectif est d'améliorer la gestion du trafic, la QdS*[1] *et plus généralement, les services réseaux. Cette approche, appelée MBN*[2]*, repose principalement sur l'utilisation de techniques de métrologie actives et passives qui permettent de mesurer en temps réel différents paramètres du réseau et de son trafic pour ainsi réagir très rapidement et très précisément à des évènements spécifiques se produisant dans le réseau (apparition de congestion par exemple). Ce papier illustre l'approche MBN au travers du développement d'un mécanisme de contrôle de congestion orienté mesures intitulé MBCC*[3] *et évalué par des simulations NS-2. Elles montrent, en particulier, comment ce nouveau mécanisme permet d'améliorer les caractéristiques du trafic ainsi que la QdS dans l'Internet, malgré la complexité et la variabilité du trafic actuel.*

ABSTRACT. *This paper deals with a new approach for the Internet which aims at improving traffic and quality of service management and more generally network services. This approach, called MBN, mainly relies on the use of active and passive monitoring techniques which measure in real time network and traffic parameters and makes possible to react quickly and suitably to specific events arising in the network (congestion for example). This paper illustrates the MBN approach with a measurement based congestion control mechanism (MBCC) and evaluates it thanks to NS-2 simulation. They particularly show that MBCC improves traffic characteristics and QoS despite the complexity and variability of current traffic.*

MOTS-CLÉS : *Métrologie réseau, caractérisation de trafic, "Measurement Based Networking", contrôle de congestion, TFRC, QdS Internet*

KEYWORDS: *Network monitoring, traffic characterization, Measurement Based Networking, congestion control, TFRC, QoS*

1. QdS : Qualité de Service
2. MBN : Measurement Based Networking
3. MBCC : Measurement Based Congestion Control

1. Introduction

Les travaux de recherche récents en réseau, basés sur l'utilisation de techniques de métrologie, ont permis d'améliorer les connaissances sur le trafic Internet. En effet, ces études ont montré que ce trafic est loin d'être régulier et qu'il est possible d'observer des caractéristiques oscillatoires à toutes les échelles temporelles i.e. pour des échelles de temps fines (inférieures à la seconde) et importantes (supérieures à l'heure). Les oscillations pour les petites échelles de temps ne sont pas surprenantes. A l'inverse, la mise en évidence d'oscillations à long terme sont assez inattendues et problématiques pour la stabilité et les performances du réseau [PAR 97]. En particulier, les récentes études ont montré que les applications pair à pair (P2P), utilisées massivement pour échanger des volumes d'information très importants (albums musicaux ou films) sont en train de modifier les caractéristiques du trafic Internet [OWE 04a]. Ces applications, par les oscillations à long terme qu'elles induisent, créent de la dépendance longue mémoire (LRD) dans le réseau. Cette LRD ainsi que ces oscillations sont très néfastes pour la QdS du réseau étant donné qu'elles provoquent des phénomènes importants de congestion et un niveau de service très instable pour les utilisateurs [WIL 95]. De plus, les oscillations à long terme créent des propriétés de non-stationnarité dans le trafic, la valeur moyenne du trafic changeant fréquemment de façon significative.

Ce papier propose donc d'exploiter les techniques de métrologie réseau en temps réel afin de définir une architecture orientée mesure (appelée dans la suite MBA pour Measurement Based Architecture) et l'ensemble des mécanismes nécessaires permettant de mieux adapter les mécanismes du réseau aux fréquents changements mesurés dans le trafic. Ainsi, la partie 2 débute ce papier en analysant les caractéristiques de l'Internet qui rendent si difficile une amélioration de la QdS. Dans un deuxième temps, la partie 3 présente notre approche orientée mesure (appelée dans la suite MBN pour Measurement Based Networking) et l'architecture MBA qui y est associée (section 3.2). Un composant essentiel de cette architecture est le protocole MSP (Measurement Signaling Protocol) qui permet d'informer en temps réel tous les acteurs du réseau de l'évolution des caractéristiques du trafic. Ce protocole devra réaliser un compromis entre le besoin en rapidité pour permettre de délivrer des informations sur l'état du réseau en temps réel et le soucis d'éviter sa surcharge afin de lui permettre ainsi de fonctionner dans des réseaux de grande taile (section 3.2). Les exemples d'applications de l'approche MBN sont nombreuses mais nous avons décidé de l'illustrer par un nouveau mécanisme de contrôle de congestion orienté mesures (appelé dans la suite MBCC pour Measurement Based Congestion Control). Il devra être capable de limiter le nombre de congestions et de pertes dans le réseau mais aussi de pouvoir améliorer la régularité du trafic et l'utilisation des ressources du réseau. Dès lors, la partie 4 présente les principes de fonctionnement de MBCC et de MSP et leur évaluation conjointe en simulation (NS-2). Cette dernière se déroule en deux parties. La première étape (section 5.2) permettra de définir quels sont les paramètres optimaux de MSP pour concilier performance (vitesse), passage à l'échelle et absence de surcharge du réseau. La deuxième (section 5.3) démontre les avantages de MBCC par rapport aux mécanismes de contrôle de congestion traditionnels comme ceux de TCP. Au final, la partie 6 conclut ce papier et introduit les évolutions à venir pour l'approche MBN.

2. Problématique de la QdS dans l'Internet

Garantir la QdS consiste à fournir le service demandé dans toutes les circonstances, y compris les plus difficiles. Parmi ces dernières, le niveau de QdS dans l'Internet est

particulièrement sensible à un grand nombre de «ruptures» dues à un volume important de trafic inattendu qui peut être légitime dans le cas de la diffusion d'un évènement populaire sur le réseau, une défaillance technique ou encore un comportement illicite d'un utilisateur du réseau (attaque de déni de service par exemple). Les ruptures de trafic incluent plus généralement tous les évènements qui peuvent provoquer un changement important dans les caractéristiques du trafic et qui peuvent affecter la QdS dans le réseau. Dans ce contexte, le développement de méthodologies pour permettre une mesure globale du réseau est devenue un enjeu important. Ces méthodes sont essentielles pour permettre de détecter et de réagir à ces ruptures. Ce sont les conclusions qui ont été avancées par le projet MetroPoLIS [MET] ainsi que de nombreux autres projets récents de recherche qui ont montré que le trafic Internet est très loin de présenter un comportement régulier et mettent en évidence de larges variations de son débit mesurables à toutes les échelles de temps [PAR 96]. Tous ces projets ont observé des propriétés d'auto-similarité [PAR 00], de (multi-)fractalité [FEL 98] ou encore de LRD [ERR 96] dans le trafic actuel qui sont des paramètres qui mesurent le degré important de variabilité du trafic Internet. Cette dernière est causée, en particulier par les mécanismes de contrôle de congestion, spécifiquement ceux de TCP qui reste le protocole dominant de l'Internet [PAR 96]. Evidemment, le mode d'émission en rafale de TCP induit les caractéristiques oscillatoires qui sont de plus en plus observées. Ces oscillations sont très néfastes pour l'utilisation globale des ressources du réseau étant donné que la capacité libérée par un flux TCP à la suite d'une perte, par exemple, ne peut pas immédiatement être récupérée par un autre flux. Ce constat se traduit par un gaspillage des ressources et induit une diminution de la QdS globale du trafic et du réseau [PAR 97].

D'autre part, les capacités grandissantes de l'Internet permettant aux utilisateurs de transmettre des flux de plus en plus gros (appelés éléphants) comme de la musique ou des films par exemple entrainent une augmentation de l'échelle de la LRD observable et avec elle, du niveau d'oscillation du trafic [WIL 95]. Ces variations d'amplitude causées par l'augmentation du nombre de flux éléphants dans le réseau sont bien plus importantes qu'avec les flux souris (les flux courts) [BEN 03]. En effet, les éléphants, à cause de leur durée de vie très importante dans le réseau, ont le temps d'atteindre d'importantes valeurs pour la fenêtre de congestion (CWND) de TCP et ainsi, une simple séquence de pertes peut provoquer une importante réduction suivie d'une importante augmentation du débit d'émission. Ce phénomène est aujourd'hui exacerbé dans les réseaux commerciaux. En effet, le très grand succès des applications P2P utilisées pour échanger des fichiers volumineux (albums musicaux ou films), modifie les propriétés du trafic Internet qui devient un mélange de trafic Web et P2P où les éléphants sont de plus en plus nombreux et de plus en plus volumineux (grâce en particulier à la démocratisation des accès haut-débit : ADSL, cable...) ! Nos études sur cette évolution du trafic démontrent que les flux éléphants représentent environ 5 % du nombre de flux dans l'Internet (ils n'étaient que 2 à 3 % il y a cinq ans) et que ces 5 % représentent plus de 60 % de la totalité du trafic Internet [LAR 04]. Evidemment, ces variations importantes dans le profil du trafic ont un impact sur ses propriétés de stationnarité. Cette remarque justifie le développement de mécanismes efficaces permettant de garantir une QdS stable au cours du temps pour tous les utilisateurs. La non-stationnarité du trafic, définie comme un changement dans la moyenne du débit, a donc des répercutions importantes sur la QdS du réseau. De plus, les études récentes sur le trafic Internet, ont mis en évidence une versatilité forte du trafic dont les caractéristiques sont à la fois très différentes d'un lien à l'autre et évoluent aussi très rapidement au cours du temps [SOU 04]. Notre proposition qui est présentée dans la partie 3 introduit ainsi une nouvelle solution pour gérer le réseau en prenant en compte les propriétés de non-stationnarité du trafic, les fortes variations que l'on peut mesurer

sur un même lien ou au contraire les caractéristiques très différentes du trafic d'un lien à l'autre dans le réseau.

Un autre problème important est relatif à la difficulté de mise en œuvre des mécanismes de QdS de bout en bout face à l'hétérogénéïté de la topologie et de la structure administrative de l'Internet. Ce point est illustré par la figure 1. L'Internet est généralement défini comme une interconnexion de réseaux. C'est évidemment vrai mais incomplet. En effet, l'Internet doit être de plus en plus vu comme un réseau global découpé en différents domaines ou Systèmes Autonomes (SA), indépendant administrativement et gérés de façon autonome. Chaque réseau de chaque SA propose donc différents niveaux de service et de QdS à ses usagers. Ce phénomène devient de plus en plus important avec la prolifération de nouveaux types de réseaux reposant sur des technologies sans-fil (WIFI, GPRS, UMTS, etc.) ou satellite qui proposent des niveaux de QdS très différents. Dans un tel contexte, assurer de la QdS de bout en bout est particulièrement ardu étant donné que le niveau de service obtenu par un utilisateur sera minoré par le domaine ayant les prestations les plus basses parmi tous les domaines traversés sur le chemin entre la source et la destination. En particulier, les liens de "peering" sont souvent sous-dimensionnés et à l'origine d'une diminution importante de la QdS et des performances dans la communication de bout en bout [NOR 04]. Dans un tel contexte, améliorer la QdS de bout en bout pourrait nécessiter la mise en place d'une infrastructure globale et de procédures de gestion centralisée pour éviter les différences entre SA. Une telle hypothèse ne peut que rester utopique tant la compétition «économique» entre opérateurs et fournisseurs d'accès est importante. Ainsi, trouver un accord global entre tous ces acteurs pour définir comment échanger le trafic Internet est totalement invraissemblable. La QdS de bout en bout est donc un problème à considérer avec une vision multi-domaine.

3. Principes de l'approche MBN

3.1. *Définition de l'architecture MBA s'appuyant sur des mesures*

Conscients des différents aspects relatifs à la problématique de l'amélioration de la QdS dans l'Internet, il est aisé de comprendre qu'une solution statique optimale pour l'ensemble des connexions n'est pas possible à établir. Cette constatation nous a amené à proposer l'approche MBN qui permet de réagir en temps réel, globalement, localement et ponctuellement à différents évènements se produisant dans le réseau. Ainsi, l'approche MBN nécessite de guetter des changements qui se produisent dans le réseau ou le trafic par l'intermédiaire de la mesure des paramètres de QdS et du trafic. La figure 1 décrit comment ces outils de mesure doivent être déployés dans le réseau. Elle détaille le cas plus spécifique d'une connexion MBCC régie par l'approche MBN entre une source et une destination, traversant deux SA Internet ainsi que les routeurs de bordure et de coeur. Ces routeurs intègrent le mécanisme MSP (permettant de mesurer et de signaler aux équipements réseaux concernés ces résultats de mesure). Il est à noter que les équipements de mesure sont de plus en plus déployés au sein des différents SA Internet à l'heure actuelle. Néanmoins, même si tous les nœuds du réseau ne seront sans doute jamais tous équipés d'outils de mesure, nous pensons qu'en collectant et en utilisant les résultats de mesure des sondes effectivement déployées dans l'Internet, nous pouvons améliorer considérablement la gestion du réseau et de son trafic. Ainsi, MBN est pensé selon l'idée suivante : les performances et la QdS peuvent être grandement améliorées et même devenir optimales en utilisant des informations de mesure sur le réseau mais même si en certains

points du réseau l'information de mesure n'est pas disponible, le réseau doit continuer à fonctionner avec de bonnes performances et une bonne QdS.

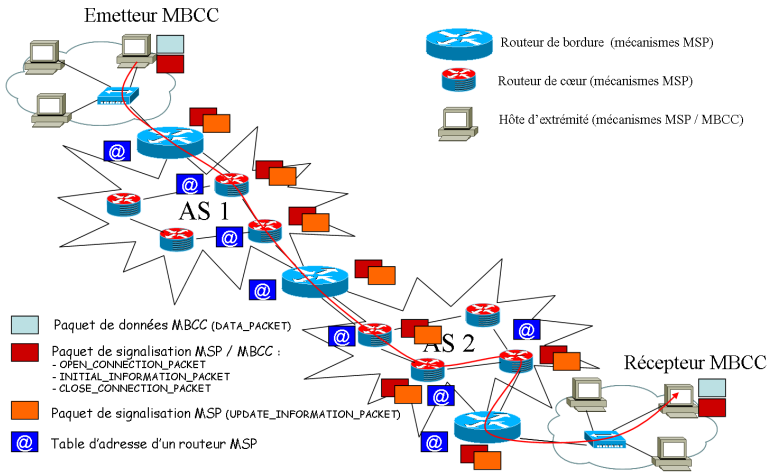


Figure 1. *Déploiement architectural de MBN : exemple du contrôle de congestion MBCC*

La structure administrative de l'Internet nous amène à considérer différentes techniques de mesure. En effet, les mesures intra-domaines peuvent être réalisées par des équipements passifs (systèmes basés sur SNMP, NetFlow, DAG...) déployés par l'opérateur qui souhaite réaliser une gestion du réseau plus pertinente. Il pourra ainsi disposer d'information sur le niveau de bande passante utilisée et disponible, le nombre de flux actifs dans son réseau, le taux de perte... A l'inverse la mesure du délai sera plus facile en employant des techniques actives. Pour ce qui est de la mesure inter-domaine, le problème est différent. En effet, l'opérateur considéré (qui peut être réduit à un simple utilisateur du réseau) ne disposera pas facilement d'information fiable sur un SA concurrent. Dans ce cas, il est nécessaire d'utiliser des techniques de mesures actives, le principe étant d'émettre des paquets sondes à travers les domaines pour lesquels on souhaite estimer les ressources ou la QdS en observant ce qui arrive à ces paquets sondes.

Ainsi, l'ensemble de ces mesures réalisées en temps réel et signalées à l'ensemble des équipements réseaux concernés (les sources de trafic par exemple), donne une connaissance précise de l'état du réseau et du trafic et permet ainsi de parfaitement adapter leur débit d'émission (dans le cas de MBCC par exemple) aux ressources disponibles. Il est important de noter qu'un aspect primordial de MBN à trait à la définition d'un protocole permettant de signaler les informations de mesure dans le réseau à la fois en intra-domaine mais aussi éventuellement entre différents opérateurs ou FAI (Fournisseurs d'Accès Internet) si ceux-ci décident de coopérer étroitement pour un objectif commun de fourniture de QdS.

3.2. *Principes du protocole de signalisation des mesures (MSP)*

MSP est un composant clé de l'architecture MBA mais il nécessite de trouver pour lui le compromis entre efficacité et capacité à fonctionner à large échelle. En effet, il est nécessaire de fournir des informations sur le trafic le plus rapidement possible pour permettre aux composants du réseau de réagir suite à la réception d'informations très récentes sur l'état du réseau tout en évitant de saturer le réseau avec des informations de signalisation. Ce bon fonctionnement à grande échelle nécessite aussi que les composants MSP ne stockent pas une trop grande quantité d'informations : les tables de correspondances à rajouter pour MBN dans les routeurs doivent être aussi petites que possible (avec un nombre limité d'entrées) pour permettre de minimiser le temps de recherche d'une information.

La figure 1 présente le mode de fonctionnement de MSP dans ses grandes lignes permettant d'atteindre ces objectifs d'efficacité et de passage à l'échelle (plus de détails seront fournis dans la section 4.1 relative à l'étude de cas de MBCC). Tout d'abord, MSP est orienté connexion, c'est à dire que le chemin signalé doit être le même pour tous les paquets d'une connexion choisie. Pour cela, nous avons utilisé le principe de RSVP [BRA 97] avec un premier paquet qui découvre le chemin de la source à la destination et ensuite un paquet de retour qui revient à la source. Les différences existent pour le paquet de retour qui est un paquet de réservation dans RSVP mais un paquet de signalisation dans MSP : il transporte des informations de mesure. D'autre part, les paquets de signalisation sont envoyés à chaque fois que nécessaire, alors que dans RSVP ils sont juste envoyés lorsque la connexion est ouverte. MSP utilise simplement le principe de RSVP qui consiste à trouver un chemin et de revenir le long de ce chemin. Cette méthode permet à MSP de parfaitement identifier quels sont les composants réseaux (les routeurs) rencontrés sur le chemin et de limiter le nombre de sources et de destinations pour les messages de mesure.

Pour permettre de prendre en compte le problème du facteur d'échelle rencontré par RSVP dans son déploiement Internet, nous avons choisi :

– De ne considérer que les flux éléphants. En fait, les souris ne créent pas de réel problème dans le trafic et tous les dommages sont générés par des éléphants [BEN 03]. Ainsi, les routeurs MSP conservent juste une information sur les flux éléphants les traversant. Cette technique permet de limiter le nombre d'entrées dans la table de connexion étant donné que les éléphants représentent une très petite proportion du nombre total de flux [LAR 04] ;

– D'envoyer les informations de mesure seulement lorsqu'une rupture est détectée dans le trafic. Cette technique permet de générer du trafic de signalisation seulement quand les conditions du réseau changent[1]. Ce principe va donc limiter la quantité de données de signalisation et permettre aux émetteurs et aux routeurs de disposer très rapidement d'informations importantes sur l'évolution du réseau et du trafic : rappelons que les mesures sont réalisées tout au long du chemin entre la source et la destination et potentiellement très prêt de la source.

En procédant de la sorte, nous souhaitons résoudre le problème du facteur d'échelle qui a été précédemment rencontré dans l'Internet dans les tentatives d'amélioration et

1. Evidemment, un envoi périodique d'information de mesure est aussi intégré dans MSP pour détecter les variations très douces dans les fluctuations du trafic (c'est à dire un trafic sans rupture forte mais avec une composante de non-stationarité). Etant donné son principe de réaction basé sur la détection de rupture, la période pourra être très grande, induisant ainsi un faible trafic de signalisation.

de garantie de la QdS. Les performances de MSP seront précisément évaluées dans la section 5.2.

4. Principes du contrôle de congestion orienté mesures (MBCC)

Etant donné le niveau d'oscillation et de non-stationnarité du trafic qui cause tant de difficultés pour fournir une QdS stable voire garantie, ce papier propose une illustration du concept MBN appliqué à la mise en place d'un nouveau contrôle de congestion. Les objectifs de MBCC sont conjointement d'améliorer les caractéristiques du trafic et la performance du réseau en lissant le trafic (de façon à limiter les effets de variabilité du trafic) et d'optimiser l'utilisation des ressources (la bande passante disponible) en utilisant l'infrastructure de mesure MBA / MSP. De plus, MBCC sera capable d'assurer une certaine équité à des flux concurrents et de continuer à fonctionner avec de bonnes performances même si certaines mesures manquent.

Dans les travaux [LAR 03] [OWE 04a] sur l'analyse des caractéristiques du trafic (dont les résultats ont été résumés dans la section 2), la nature oscillatoire du trafic Internet a été mise en évidence. En particulier, il a été montré que ces oscillations persistantes dans le temps (sources de la LRD observée dans le trafic) étaient dues à l'inadéquation de TCP pour la transmission des fichiers très volumineux sur des réseaux à haut débit. Ainsi, le problème le plus immédiat concerne la réduction des oscillations et plus précisément la régulation des oscillations persistantes qui ont un impact dramatique sur la QdS du trafic et les performances du réseau. C'est pour cela qu'un des objectifs de MBCC est d'offrir plus de stabilité aux flux éléphants. Pour supprimer les comportements oscillatoires observables à toutes les échelles de temps, le mécanisme de contrôle de congestion TFRC (TCP Friendly Rate Control) est capable d'apporter une contribution importante. TFRC a été défini pour fournir un service adapté aux applications orientées flux qui ont besoin d'un débit lisse et régulier. Il essaie donc, d'éviter au maximum les variations brutables de débit qui apparaissent avec TCP dans le cas d'une reprise d'émission qui suit la détection d'une séquence de pertes. En associant un tel mécanisme au transfert des flux éléphants, représentant la majorité en volume du trafic, nous souhaitons contrôler les oscillations du trafic et augmenter la QdS et les performances globales du réseau. Le débit d'émission de chaque source TFRC est calculé, une fois par RTT, grâce à un algorithme orienté récepteur qui se base sur le taux d'évènements de pertes p estimé par le récepteur [FLO 00] et selon l'équation suivante :

$$X_{TFRC} = \frac{s}{R * \sqrt{2 * b * \frac{p}{3}} + (t_{RTO} * (3 * \sqrt{3 * b * \frac{p}{8}}) * p * (1 + 32 * p^2)))} \quad [1]$$

où :

– X est le débit de transmission en octets par seconde,

– s est la taille du paquet en octet,

– R est le temps aller-retour en seconde,

– p est le taux d'évènement de perte (entre 0 et 1), il s'agit de la fraction du nombre d'évènements de pertes sur le nombre de paquets transmis,

– t_{RTO} est la valeur du timer de retransmission TCP en seconde et normallement égal à $4 * R$,

– b est le nombre de paquet acquittés par un simple acquittement TCP.

Les bénéfices de l'utilisation de TFRC à la place de TCP ont été démontrés dans [LAR 03]. Cependant, si TFRC est capable de réduire les oscillations de TCP, il n'est pas capable de s'adapter aux ruptures brutales du trafic (pannes sur des liens impliquant un rééquilibrage du trafic, pics de trafic dus à un trafic légitime lié à la diffusion d'un évènement très populaire par exemple. L'approche MBN est proposée comme une solution pour faire face à ces ruptures. Ainsi, nous souhaitons que MBCC soit une solution optimale permettant d'améliorer TFRC, qui en moyenne est un petit peu moins efficace que TCP New Reno with SACK[2] (Selective ACKnowledgement [MAT 96]) [LAR 03]. Pour bénéficier des avantages de TFRC, nous avons défini MBCC comme une de ses extensions en le dotant d'une capacité à utiliser les résultats de mesure qui émanent des équipements de métrologie déployés dans le réseau. En faisant ce choix, nous sommes capables de produire de bons résultats (meilleurs que ceux de l'Internet actuel) même si les informations de mesure sont temporairement indisponibles.

Le principe de MBCC consiste à utiliser l'algorithme de TFRC pour calculer le taux d'émission nominal de chaque connexion et de corriger cette valeur grâce à la connaissance du niveau de bande passante disponible et consommée dans le réseau. Ainsi, si une fraction de la bande passante est disponible, les sources pourront générer plus de trafic qu'indiqué dans l'équation 1 (qui correspond au débit d'un flux TCP [ALT 00]) sans pour autant créer des pertes et des congestions dans le réseau. Ainsi, le niveau de congestion du réseau devrait être significativement réduit en déployant des sources de trafic «pro-actives», capables d'adapter en temps réel leur débit d'émission en fonction des ressources disponibles. Un tel mécanisme devrait aussi aider à augmenter l'équité entre les flux, étant donné que la correction réalisée sur le débit d'émission ne devrait pas dépendre de la valeur du RTT mais de la réelle fraction de bande passante disponible équitablement partagée entre les flux concurrents.

Comme dans [LAR 03], MBCC sera uniquement utilisé pour les flux éléphants qui sont les flux qui génèrent le plus de perturbations dans le réseau. A l'opposé, comme le trafic «souris» représente un bruit blanc Gaussien [BEN 03], il n'induit pas de problème de transfert important et il n'est donc pas nécessaire de modifier leur protocole de transport. Ainsi, pour une période normale (quand les informations de mesure sont correctement reçues, qu'il n'y a pas de congestion et que de la bande passante est disponible dans le réseau), chaque flux éléphant peut utiliser une fraction supplémentaire des ressources qui sont disponibles. Cette fraction est calculée en divisant la bande passante totale disponible par le nombre de flux moyens éléphants dans le réseau à ce moment (ces informations étant fournies par les équipements de mesure rencontrés tout au long du chemin). Il est logique de diviser la bande passante disponible par le nombre moyen de flux actifs (N) traversant ce lien car il a été démontré que les arrivées de flux éléphants sont proches d'un processus Poissonien [BEN 03]. En effet, pour un processus de Poisson, comme la moyenne est égale à la variance, le nombre moyen est significatif car les valeurs du processus ne seront jamais très éloignées de cette valeur moyenne. A l'inverse pour une période de congestion, les émetteurs MBCC devront réduire leur débit d'émission pour résorber la congestion et ceci en essayant d'être aussi équitable que possible. Dès lors, les sources MBCC envoient la valeur minimale entre le débit TFRC et le débit effectif obtenu par un flux à ce moment au niveau du goulot d'étranglement sur son chemin.

Ainsi, les équations de cet algorithme peuvent être résumées de la façon suivante :

2. Cette version de TCP a été choisie comme référence car elle est considérée comme la version la plus performante de ce protocole de transport.

– Pour une période sans congestion ($p = 0$) : $X_{MBCC} = X_{TFRC} + BPd_{flux}$;

– Pour une période de congestion ($p \neq 0$) : $X_{MBCC} = min(X_{TFRC}; BPc_{flux})$;

Avec :

– BPd_{flux} qui correspond à la bande passante disponible dans le(s) goulot(s) d'étranglement rencontré(s) sur le chemin. Il est calculé par l'intermédiaire du rapport $\frac{\text{bande passante totale disponible}}{N}$, cette information étant fournie par les routeurs MSP rencontrés sur le chemin ;

– BPc_{flux} qui correspond à la bande passante consommée par le flux au travers du goulot d'étranglement, cette information est fournie par le récepteur MBCC avec les autres informations de bout-en-bout comme le RTT ou le taux de perte (cf. équation 1).

4.1. *Détails de MSP : illustration dans le cas du déploiement de MBCC*

Etant donné les principes de MBCC qui viennent d'être présentés, cette section va décrire comment les routeurs MSP se comportent pour transmetre les informations de signalisation aux sources MBCC que sont la bande passante disponible et le nombre moyen de flux éléphants mesurés sur le chemin emprunté par les flux MBCC. Le comportement de MSP est décrit sur la figure 1 et ses quatre étapes de fonctionnement détaillées ci-après :

1 - Ouverture d'une connexion éléphant. Un paquet de signalisation spécifique (OPEN_CONNECTION_PACKET) est émis par l'émetteur MBCC et indique à chaque routeur rencontré sur le chemin qu'une connexion éléphant va être initiée. Dans les différents routeurs traversés par le paquet de signalisation, une table d'adresse est mise à jour avec l'adresse de l'émetteur MBCC.

2 - L'agent récepteur MBCC envoie une information de mesure initiale à l'émetteur MBCC en utilisant un paquet de signalisation (INITIAL_INFORMATION_PACKET). Ce paquet est analysé par chaque routeur sur le chemin et mis à jour avec ces informations de mesure locale (voir 3 - pour les détails de la mise à jour). Ainsi, quand ce paquet arrive à l'émetteur MBCC, il peut ouvrir la connexion car il dispose ainsi d'informations de mesure sur l'état du réseau.

3 - Les paquets de données sont échangés entre l'émetteur MBCC et le récepteur (MBCC_DATA_PACKET). Dans le même temps, les routeurs envoient régulièrement des informatoins de mesure en utilisant des paquets de signalisation spécifiques (UPDATE_INFORMATION_PACKET). Cette information est envoyée quand une rupture est détectée par les routeurs dans l'ensemble des paramètres que l'on peut mesurer en temps réel (dans le cas des agents MBCC : nombre moyen de flux éléphants (N) et bande passante disponible). C'est un principe important car il limite le nombre d'informations de mesure transitant dans le réseau[3]. Les principes de l'algorithme sont les suivants :

- Si BP totale d_{Ri} (i.e. calculée par le routeur i) < BP totale d_{sig} (i.e. incluse dans le paquet de signalisation) alors BP totale $d_{sig} = BP$ totale d_{Ri} ;

3. Il faut noter que les informations de mesure sont aussi émises périodiquement de façon à informer les agents MBCC de la tendance des fluctuations même si elles sont douces. Mais dans ce cas, la période peut être très importante, ceci impliquant très peu de trafic additionnel pour la signalisation.

- Si $N_{Ri} > N_{sig}$ alors $N_{sig} = N_{Ri}$;

4 - Fermeture de la connexion éléphant. Les émetteurs MBCC envoient un paquet de signalisation spécifique (CLOSE_CONNECTION_PACKET) informant tous les routeurs sur le chemin de la fin de la connexion. Ces derniers suppriment donc de leur table d'adresse l'agent MBCC émetteur.

5. Validation expérimentale de l'approche MBN appliquée au contrôle de congestion

Dans cette section, nous présentons les résultats expérimentaux qui valident les mécanismes MSP et MBCC. En particulier, la section 5.2 quantifie les valeurs optimales pour les paramètres de MSP et MBCC de façon à trouver le meilleur compromis entre faible surcharge du réseau par les informations de mesure, temps de réponse faible et réactions précises des agents MSP et MBCC. En s'appuyant sur ces valeurs optimales, la section 5.3 étudie en détails les avantages de MBCC pour la stabilité du réseau et l'utilisation des ressources en le comparant aux mécanismes de contrôle de congestion traditionnels de TCP.

5.1. *Principes des simulations*

Les mécanismes MBCC et MSP ont été implémentés et évalués en utilisant NS-2. Il a été nécessaire de développer un ensemble d'outils pour mesurer la bande passante disponible et consommée dans le réseau simulé et pour échanger les résultats de mesure entre les routeurs et les sources de trafic.

La topologie utilisée est décrite dans la figure 2. Dans ces simulations, nous avons créé une topologie multi-domaines avec plusieurs goulôts d'étranglement. Les flux éléphants, utilisant soit MBCC, soit TCP SACK ainsi que le trafic de fond utilisant TCP New Reno sont échangés de façon à entrer en compétition dans ces goulôts d'étranglement. L'objectif est donc de mesurer l'impact réciproque des flux MBCC en théorie réguliers sur ceux TCP beaucoup plus variables. De plus, les liens de coeurs (ceux des SA 1 et 3) représentent les liens les plus «congestionnés» sur le chemin considéré. Ils induieront ainsi des périodes de congestion importantes où les capacités d'adaptabilité de MBCC seront estimées et son niveau de performance comparé avec les autres mécanismes de contrôle de congestion (ceux de TCP SACK). Chaque simulation s'appuie sur des traces de trafic collectées sur le réseau Renater. Elles sont rejouées dans le simulateur NS-2 avec une méthodologie spécifique détaillée dans [OWE 04b] dont l'objectif est de produire des simulations réalistes[4].

L'objectif principal de cette étude est de comparer les capacités d'adaptation de MBCC face à une augmentation (ou une diminution) de la charge du réseau et face aux autres mécanismes de contrôle de congestion pour évaluer l'équité qu'il offre aux différents flux. Pour réaliser son évaluation, plusieurs paramètres ont été étudiés en simulation :

– l'impact du trafic de signalisation en calculant le pourcentage entre le débit moyen nécessaire pour la signalisation et le débit moyen global sur le réseau ;

4. Il s'agit de rejouer en simulation des échantillons de trafic Internet pour reproduire toutes les caractéristiques statistiques du trafic réel.

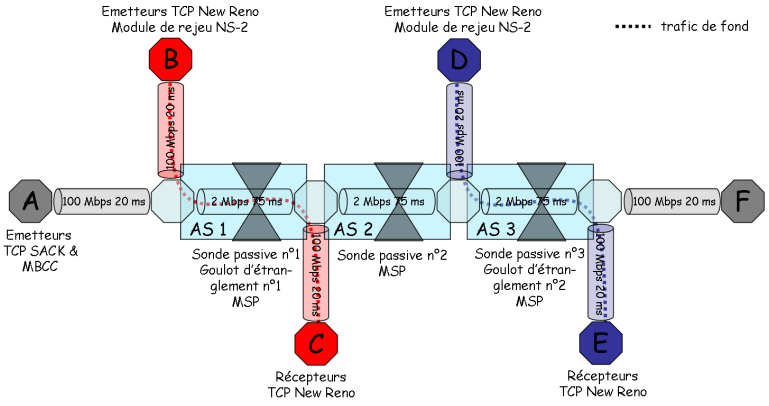


Figure 2. *Topologie du réseau utilisée pour les simulations NS-2*

– l'évolution du débit par type de trafic (TCP ou MBCC) en étudiant la variabilité du trafic. Pour cela, nous calculons le débit moyen (D), l'écart-type (σ) et un coefficient de stabilité défini de la façon suivante : CS $= \frac{D}{\sigma}$;

– l'évolution du processus de perte permettant d'évaluer les capacités d'adaptation de MBCC et de les comparer à TCP ;

– la persistance des oscillations du trafic en calculant le facteur de Hurst[5].

5.2. *Evaluation de la configuration optimale de MSP*

Plusieurs simulations ont été menées, chacune divisée en deux scénarios différents : dans le premier, les flux éléphants (trafic entre les noeuds A et F) sont tranmis en utilisant MBCC tandis que dans le second, ils utilisent TCP SACK. Le second scénario est utilisé comme référence expérimentale pour évaluer les avantages de MBCC (en termes de stabilité et de congestion dans le réseau). Dans ces deux scénarios, le trafic de fond (trafics entre les noeuds B à C et D à E), qui est constitué d'un trafic Internet normal, mélange de flux souris et éléphants, est émis en utilisant TCP version New Reno qui est la plus utilisée à l'heure actuelle dans l'Internet.

Chaque simulation dure 300 secondes. 100 éléphants et 2000 souris ont été rejouées. Un des principaux objectifs de ces expérimentations a été d'étudier l'impact du trafic de signalisation généré par MSP sur la congestion du réseau et l'efficacité des réactions de MBCC. Pour cela, il a été nécessaire de trouver les valeurs optimales pour les différents paramètres de fonctionnement de MSP :

– Le premier paramètre est relatif à la granularité du système de mesure. En fait, la mesure doit être réalisée sur des intervales courts et par exemple le débit instantané

5. Ce paramètre, noté H, quantifie la LRD du trafic et représente aussi une bonne évaluation de son degré d'oscillation [OWE 04a]. Pour le calculer nous utilisons une analyse basée sur une décomposition en ondelettes [ABR 98].

est ainsi calculé comme le débit moyen sur de courte période de temps ($Periode$). Ce paramètre a un impact fort étant donné que plus la granularité est importante, plus le débit semble lisse. En conséquence cette granularité agit sur le volume de trafic généré par MSP[6]. Plusieurs périodes de 0,2 à 5 secondes ont donc été testées.

– Le deuxième paramètre permet de fixer un seuil de détection pour les ruptures qui sont analysées dans le réseau. Il s'agit de déterminer la variation minimale ($Seuil$) entre deux mesures consécutives pour lesquelles nous pouvons considérer que les conditions du réseau ont changé et qu'il est nécessaire de le signaler aux sources de trafic pour leur permettre de s'adapter à ce changement. Ce seuil est exprimé en pourcentage de la capacité totale du lien.

– Pour finir, la valeur "Time Out" (TO) correspond au comportement périodique de MSP nécessaire dans le cas où aucune rupture ne se produit mais si l'évolution du réseau bien que très lente génère une tendance non-stationnaire. Cette valeur est définie par rapport au paramètre $Periode$, elle ne doit pas être beaucoup plus importante de façon à informer régulièrement les sources MBCC des variations même lentes de débit. En suivant ce principe, nous avons empiriquement sélectionné les couples ($Periode$, TO) : ($Periode = 0,2s$ et $TO = 2s$) ou ($Periode = 0,5s$ et $TO = 4s$) ou ($Periode = 1s$ et $TO = 5s$) ou ($Periode = 2s$ et $TO = 8s$) ou ($Periode = 5s$ et $TO = 10s$).

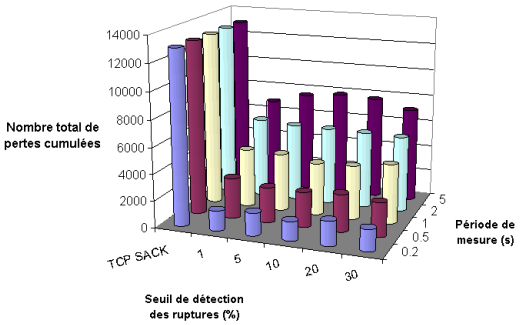
Nous avons fait plusieurs simulations en utilisant différentes traces pour trouver le couple optimal ($Periode_{optimale}$, $Seuil_{optimal}$). Les résultats sont présentés dans les figures 3 qui représentent le nombre cumulé de pertes dans le réseau, la surcharge de trafic induite par MSP (en pourcentage du trafic total) et le coefficient de stabilité mesuré pour le trafic échangé.

Tout d'abord, nous allons inférer la valeur optimale pour la $Periode$ de mesure. Un des objectifs principaux de MBCC est d'optimiser au mieux l'utilisation des ressources du réseau en générant le moins de pertes possibles. Dans la figure 3(a), seuls les résultats avec une $Periode \leq 1s$ sont acceptables[7] (quel que soit la valeur du seuil) : $pertes_{MBCC} \leq \frac{pertes_{TCPSACK}}{3}$. Un autre objectif de MBCC est de transférer les données avec un débit régulier et d'éviter les comportements oscillants qui induisent une mauvaise utilisation des ressources du réseau. Ainsi, dans la figure 3(c), seuls les résultats avec une $Periode \geq 1s$ sont acceptables (quel que soit la valeur du seuil) : CS(MBCC) $\geq CS$(TCP SACK). Lorsqu'on considère les résultats sur l'impact du trafic de signalisation, ils ne nécessitent pas de considération supplémentaire. Ainsi, en croisant les résultats des trois paramètres précédents (congestion, stabilité et trafic de signalisation), uniquement les résultats obtenus avec une $Periode = 1s$ respectent tous les critères de choix.

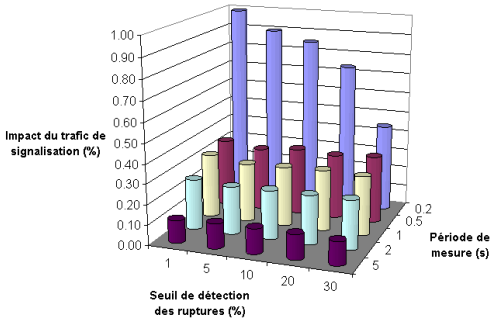
Dans un deuxième temps, nous allons inférer la valeur optimale pour le $Seuil$. Dans ce cas, seuls les résultats expérimentaux s'appuyant sur la stabilité du trafic peuvent nous apporter de l'information pour choisir le seuil optimal. Pour les deux paramètres restants (signalisation et perte), les résultats sont vraiment trop proches pour nous apporter une information utile. Ainsi, nous prennons en compte le seuil où le CS est maximum, il s'agit du cas où $Seuil = 1\%$. En conclusion, la paire de valeurs optimales est ($Periode_{optimale} = 1s$, $Seuil_{optimal} = 1\%$). Elles seront utilisées dans la section suivante pour quantifier précisément les avantages de MBCC par rapport à TCP SACK.

---

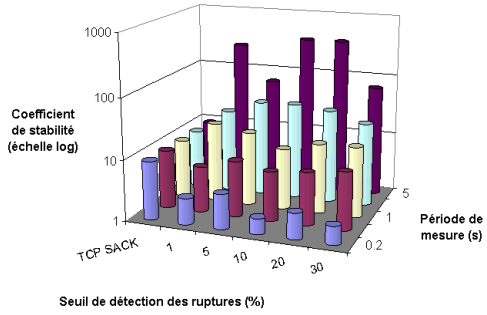
6. Plus la granularité choisie sera faible, plus la détection des variations sera de façon précise et plus le volume de signalisation sera important.
7. Pour une $Periode \geq 2s$ les niveaux de pertes entre MBCC et TCP SACK sont trop proches.



(a) Congestion du réseau



(b) Surcharge du trafic de signalisation par rapport au trafic total



(c) Stabilité du trafic

Figure 3. *Evolution des paramètres de performance en fonction des valeurs de fonctionnement de MSP*

5.3. *Contribution de MBCC à la régularité du trafic dans une configuration multi-domaine*

Cette deuxième expérience permet de comparer précisément les impacts de MBCC et de TCP SACK sur la performance du réseau, la régularité du trafic et l'optimisation des ressources. Elle va illustrer les capacités de MBCC pour améliorer la régularité du trafic pour les flux MBCC et devrait aussi montrer comment MBCC peut améliorer le profil du trafic en terme de stabilité pour les flux qui n'utilisent pas MBCC mais qui sont en concurrence avec les autres flux MBCC dans les goulôts d'étranglement de l'Internet. Pour cela, la topologie utilisée est la même que dans l'expérience précédente (cf. figure 2). Les scénarios sont les mêmes à l'exception du nombre de flux qui a augmenté (d'un facteur 10) et les simulations durent 1800 s de façon à introduire des transferts d'éléphants plus long comme c'est le cas dans l'Internet. Cette

expérience permet aussi d'évaluer les capacités à grandes échelles des mécanismes proposés étant donné que le nombre d'éléphants est augmenté de façon conséquente. D'autre part, les routeurs MSP sont configurés avec le couple de valeurs optimales inféré précédemment ($Periode = 1s$, $Seuil = 1\%$) et les paramètres étudiés dans les simulations sont les mêmes que dans la partie précédente.

Tout d'abord, le débit du trafic a été calculé. Le tableau 1 montre les valeurs résultats pour les scénarios 1 et 2. Cette expérience met donc en évidence que MBCC est plus performant que TCP SACK car le débit et l'utilisation des ressources est plus élevé et le trafic est aussi plus régulier. D'autre part, une autre information intéressante concerne le trafic de fond des goulots d'étranglement 1 et 2 quand le trafic éléphant MBCC est présent dans le réseau (scénario 1). Nous pouvons voir que dans le cas où TCP SACK est utilisé pour transmettre les éléphants entre A et F (scénario 2), le débit moyen du trafic de fond est plus bas et présente plus de variabilité que quand MBCC est utilisé dans le réseau ($CS(\text{TCP New Reno}_{\text{Scénario 2}}) < CS(\text{TCP New Reno}_{\text{Scénario 1}})$).

Tableau 1. *Analyse de la variabilité du trafic*

	Scénario 1			Scénario 2		
	MBCC	TCP New Reno Goulot d'étranglement n°1	TCP New Reno Goulot d'étranglement n°2	TCP SACK	TCP New Reno Goulot d'étranglement n°1	TCP New Reno Goulot d'étranglement n°2
Débit moyen (B / s)	109434.9	111822.5	111572.5	109420.2	101651.1	101357.3
Ecart-type du débit (σ) (B /s)	31840.4	57127.7	60299.4	44704.3	83943.6	84291.9
Coefficient de stabilité (SC)	3.437	1.957	1.850	2.448	1.211	1.202

Ce résultat est confirmé avec l'analyse du processus de perte. En effet, la figure 4 décrit un niveau de pertes plus important dans le réseau quand TCP SACK est utilisé (cf. les courbes du scénario 2) que quand MBCC est utilisé (cf. les courbes du scénario 1). Ce résultat est en plus observable à la fois pour le trafic éléphant et pour le trafic de fond global. En effet, la variabilité du trafic dans le scénario 2 est plus importante, les congestions apparaissent donc plus facilement dans le réseau et le nombre de perte est plus élevé.

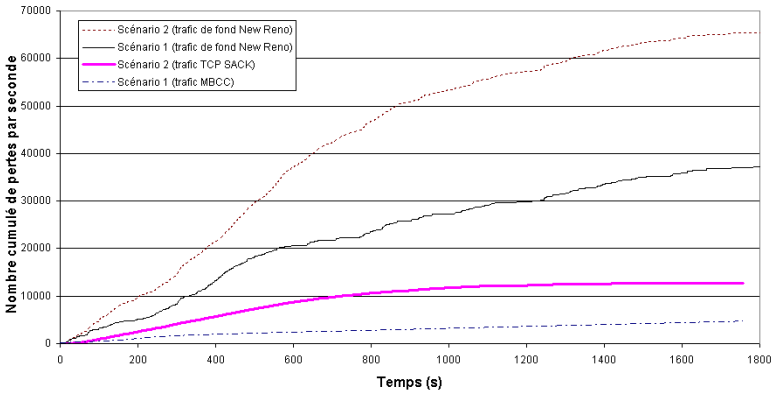


Figure 4. *Estimation du niveau de congestion*

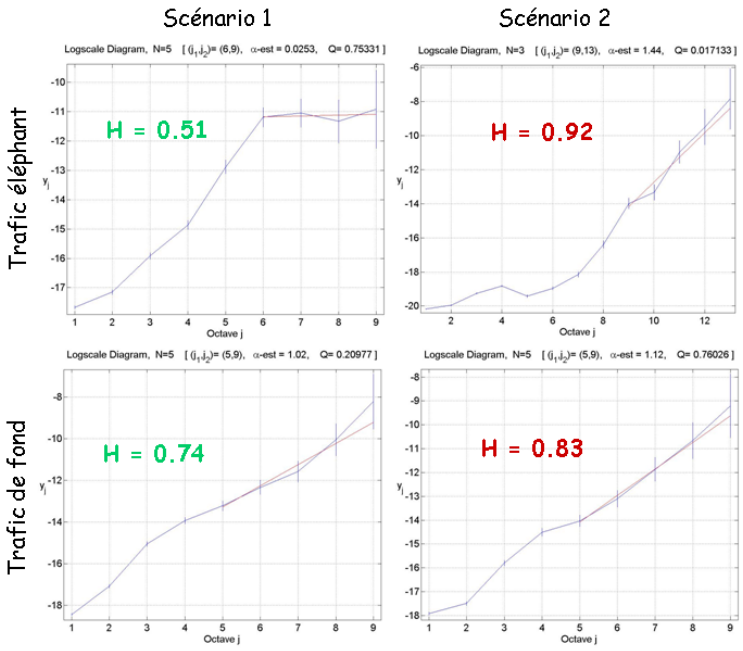


Figure 5. *Estimation de la LRD du trafic*

MBCC a aussi un impact très bénéfique sur la LRD du trafic (figure 5). En fait, grâce à MBCC, la LRD est beaucoup plus réduite dans le trafic éléphant (voir le scénario 1 où $H = 0,51$) en comparaison du trafic TCP SACK éléphant de référence où la LRD est très élevée ($H = 0,92$ dans le scénario 2). En conséquence, il y a moins d'oscillations (cf. coefficient de stabilité du tableau 1). De plus, l'analyse de la LRD du trafic de fond (cf. le bas de la figure 5) fait apparaître que le trafic global quand les éléphants sont transmis avec MBCC (cf. $H = 0,74$ dans le scénario 1) est moins dépendant à long terme qu'avec TCP SACK (cf. $H = 0,83$ dans le scénario 2), cette spécificité générant plus de stabilité dans le profil du trafic et donc moins de congestion dans le réseau.

6. Conclusion et travaux futurs

Dans ce papier, nous avons proposé une nouvelle approche qui utilise en temps réel les résultats de métrologie pour améliorer la QdS Internet avec l'objectif final de pouvoir proposer un service stable. Cette approche a été appliquée pour la conception d'un mécanisme de contrôle de congestion (MBCC) dont l'objectif est de lisser le trafic (un besoin primordial pour pouvoir fournir des services stables et garantis), limiter le nombre de pertes, optimiser l'utilisation des ressources et fournir de l'équité. Il faut noter que MBCC repose sur une approche originale qui consiste davantage à gérer de façon intelligente le trafic par rapport à ses caractéristiques complexes (oscillations et ruptures) plutôt que de réagir uniquement à des congestions. Les résultats expérimen-

taux prouvent que MBCC atteint ses objectifs : il fournit un débit optimal et régulier, il utilise toutes les ressources et il fournit aussi plus d'équité entre les flux.

Au final, il est clair que les résultats de MBCC démontrent les bénéfices de notre approche MBN appliquée au contrôle de congestion. Nous croyons aussi que MBN a une vocation plus universelle pour gérer l'Internet et son trafic. En effet, MBN peut être défini de façon à fournir une solution adaptée qui puisse faire face à différents types de réseaux, de trafic ou de conditions de fonctionnement. En particuler, MBN devrait avoir des applications dans d'autres domaines comme l'ingénierie du trafic, la tarification ou encore la sécurité réseau.

7. Bibliographie

[ABR 98] ABRY P., VEITCH D., « Wavelet Analysis of Long Range Dependent Traffic », *Trans. Info. Theory, Vol.44, No.1*, January 1998, p. 2-15.

[ALT 00] ALTMAN E., AVRACHENKOV K., BARAKAT C., « A Stochastic Model of TCP/IP with Stationary Random Losses », *Proceedings of ACM SIGCOMM*, 2000.

[BEN 03] BEN AZZOUNA N., GUILLEMIN F., « Analysis of ADSL traffic on an IP backbone link », *Proceedings of Globecom 2003*, December 2003.

[BRA 97] BRADEN R., ZHANG L., « Resource ReSerVation Protocol (RSVP) – Version 1 message processing rules », *RFC*, , n° 2209, September 1997.

[ERR 96] ERRAMILLI A., NARAYAN ., WILLINGER W., « Experimental queuing analysis with long range dependent packet traffic », *IEEE/ACM Transactions on Networking, Vol. 4, No. 2*, 1996, p. 209–223.

[FEL 98] FELDMANN A., GILBERT A., WILLINGER W., « Data networks as cascades : Investigating the multifractal nature of Internet WAN traffic », *Proceedings of ACM SIGCOMM'98*, 1998.

[FLO 00] FLOYD S., HANDLEY M., PADHYE J., WIDMER J., « Equation-based congestion control for unicast applications », *Proceedings of ACM SIGCOMM'00*, 2000.

[LAR 03] LARRIEU N., OWEZARSKI P., « TFRC contribution to Internet QoS improvement », *Proceedings of the fourth COST 263 international workshop on Quality of Future Internet Services (QoFIS'2003)*, October 2003.

[LAR 04] LARRIEU N., OWEZARSKI P., « De l'utilisation des mesures de trafic pour l'ingénierie des réseaux de l'Internet », *Techniques et Sciences Informatiques*, , n° 5-6, RSTI, volume 23, 2004.

[MAT 96] MATHIS M., MAHDAVI J., FLOYD S., ROMANOV A., « TCP Selective acknowledgement options », *RFC*, , n° 2018, October 1996.

[MET] METROPOLIS, « Site web : http ://www.laas.fr/ owe/METROPOLIS/metropolis.html ».

[NOR 04] NORTON B., « Evolution of the U.S. Peering Ecosystem », *Proceedings of the North American Network Operators' Group Workshop*, May 2004.

[OWE 04a] OWEZARSKI P., LARRIEU N., « Internet traffic characterization – An analysis of traffic oscillations », *Proceedings of the 7th IEEE International Conference on High Speed Networks and Multimedia Communications (HSNMC'2004)*, July 2004.

[OWE 04b] OWEZARSKI P., LARRIEU N., « A trace based method for realistic simulations », *Proceedings of the IEEE International Conference on Communications (ICC'2004)*, June 2004.

[PAR 96] PARK K., KIM G., CROVELLA M., « On the relationship between file sizes, transport protocols, and self-similar network traffic », *IEEE ICNP*, 1996.

[PAR 97] PARK K., KIM G., CROVELLA M., « On the Effect of Traffic Self-similarity on Network Performance », *SPIE International Conference on Performance and Control of Network Systems*, November 1997.

[PAR 00] PARK K., WILLINGER W., *Self-similar network traffic : an overview*, In Self-similar network traffic and performance evaluation, J.Wiley & Sons, 2000.

[SOU 04] SOULE A., NUCCI A., CRUZ R., LEONARDI E., TAFT N., « How to identify and estimate the largest traffic matrix elements in a dynamic environment », *Proceedings of the joint international conference on Measurement and modeling of computer systems*, ACM Press, 2004, p. 73–84.

[WIL 95] WILLINGER W., TAQQU M., SHERMAN R., , WILSON D., « Self-similarity through highvariability : statistical analysis of Ethernet LAN traffic at the source level », *ACM Sigcomm'95, pages 100–113*, 1995.