# Que nous dit la métrologie sur le futur d'Internet ?

**Philippe Owezarski**

LAAS-CNRS
7, Avenue du Colonel Roche
31077 Toulouse cedex 4
e-mail : owe@laas.fr

***Résumé*** *: Cet article a pour objectif de faire un tour d'horizon du domaine de la métrologie, en démarrant par une introduction exposant la problématique et dressant un état de l'art des principaux projets actuels et de leurs principales contributions. Ensuite, en exploitant les premiers résultats du projet IPMON de Sprint, cet article se propose de montrer les contributions possibles de la métrologie et surtout de montrer que ces découvertes vont parfois à l'encontre des croyances collectives. A terme, et cet article essaie de le montrer, c'est la façon de concevoir le réseau et ses mécanismes qui est remise en cause.*

## 1. Introduction

La métrologie, ou science des mesures, est une activité en plein essors dans le domaine des réseaux IP. Les opérateurs réseaux utilisent des techniques de métrologie depuis la mise en place des premiers réseaux de communication, mais cette discipline n'a jusqu'à présent jamais été utilisée comme elle aurait dû l'être. Pour l'instant, les opérateurs utilisaient la métrologie, souvent passive et en ligne, pour faire de la supervision du réseau et du trafic qui circulait dessus. Avec l'essors du réseau Internet, la métrologie devient la pierre angulaire de nombreuses activités autant au niveau de la recherche en réseau que de la conception, la mise en place, la gestion et l'opération des réseaux. Ainsi, la métrologie recouvre maintenant des domaines d'étude comme :

- La classification des flux et du trafic, soit pour pouvoir trier les flux en fonction de la qualité de service (QoS) qu'ils requièrent, soit, par rapport à des problèmes de routage, pour pouvoir les encapsuler dans des « trunks » de trafic qui empruntent tous la même route.
- Le dimensionnement des réseaux qui permet de mettre en place des capacités suffisantes pour assurer en permanence un service adéquat à tous les utilisateurs.
- L'analyse des mécanismes du réseau, et ce autant au niveaux des mécanismes des routeurs, les algorithmes de routages et les mécanismes du transport assurant le contrôle de flux, d'erreur, de congestion, etc. Cette analyse permet de comprendre comment tous ces mécanismes interagissent entre eux, et de régler de façon fine les différents paramètres mis en jeu. Au niveau de la recherche, les résultats d'analyse permettent la conception de nouveaux mécanismes et protocoles.
- L'échantillonnage qui consiste à déterminer les moments et les endroits stratégiques à observer et donnant une vision globale (à partir d'un ensemble d'informations partielles) du réseau et du trafic.
- La modélisation de trafic qui permet de représenter et comprendre le trafic actuel, et ensuite d'adapter les paramètres du réseau aux caractéristiques du trafic.

- La tarification et les SLA qui permettent de définir des coûts de service en relation avec les ressources consommées, par exemple.
- Etc.

Cet article se propose donc de faire une introduction à la métrologie, aux différentes techniques existantes en détaillant les aptitudes de chacune d'elles dans les différents domaines énoncés ci-dessus. Il se propose également de dresser un état de l'art des différentes expériences en métrologie qui ont été conduites de par le monde. Enfin, en utilisant les résultats d'un des projets de métrologie les plus ambitieux du moment – le projet IPMON de Sprint aux Etats-Unis [FRA01] – cet article se propose de donner quelques premiers résultats (décrits dans la partie 4) et de montrer quelle sera leur influence sur l'architecture du réseau et de ses mécanismes. Enfin, cet article présente le projet METROPOLIS[1] qui est le premier projet français en matière de métrologie. Les objectifs de METROPOLIS sont présentés ainsi que ses plus par rapport aux autres projets en cours.

## 2. Etat de l'art

Vue l'importance de la métrologie des réseaux IP, beaucoup de projets de recherche sur ce sujet sont en cours, en particulier conduits par des opérateurs Internet et des laboratoires d'étude et de recherche Internet aux Etats-Unis. Ces projets peuvent être répartis en deux grandes classes : ceux fondés sur les mesures actives et ceux reposant sur les mesures passives décomposées elles-mêmes en mesures passives en ligne et hors ligne. Chacune de ces deux classes permet de mieux comprendre le comportement à la fois du réseau (observation passive des taux de perte, des délais, etc.) et des applications (réaction en temps réel des applications aux pertes dans le réseau, du taux de transmission utile, etc.) et de mettre en lumière les interactions entre les applications et le réseau.

### 2.1. Projets de mesures actives

Le principe des mesures actives consiste à générer du trafic dans le réseau à étudier et à observer les effets des composants et protocoles – réseaux et transport – sur le trafic : taux de perte, délai, RTT, etc. Cette première approche possède l'avantage de prendre un positionnement orienté utilisateur. Les mesures actives restent le seul moyen pour un utilisateur de mesurer les paramètres du service dont il pourra bénéficier. En revanche, l'inconvénient majeur de cette approche est la perturbation introduite par le trafic de mesure qui peut faire évoluer l'état du réseau et ainsi fausser la mesure. De nombreux travaux menés actuellement abordent ce problème en essayant de trouver les profils de trafics de mesures qui minimisent les effets du trafic supplémentaire sur l'état du réseau. C'est par exemple le travail en cours au sein du groupe IPPM[2] de l'IETF[3] [PAX98] [ALM99a] [ALM99b] [ALM99c].
Les mesures actives simples restent tout de même monnaie courante dans l'Internet pour lequel de nombreux outils de test, validation et / ou mesure sont disponibles. Parmi eux, on peut citer les très célèbres *ping* et *traceroute*. *Ping* permet de vérifier qu'un chemin est valide entre deux stations et de mesurer certains paramètres comme le RTT[4] ou le taux de perte. *Traceroute* permet de voir apparaître l'ensemble des routeurs traversés par les paquets émis

[1] METROPOLIS : Metrologie pour l'Internet et ses Services
[2] IPPM : IP Performance Metrics
[3] IETF : Internet Engineering Task Force
[4] RTT : Round Trip Time

jusqu'à leur destination et donne une indication sur les temps de passage en chacun de ces nœuds.

L'un des projets les plus simples en théorie était le projet ***Surveyor*** [KAL99] de la NSF[5] aux Etats-Unis qui reposait sur l'utilisation de *ping*, amélioré par la présence d'horloges GPS[6] sur les machines de mesure. L'objectif de ce projet était donc d'étudier les délais de bout en bout et les pertes dans l'Internet.

Plusieurs projets ont actuellement pour sujet les mesures actives.

- Le projet NIMI[7] (initié par Vern Paxson aux Etats-Unis) [PAX00] a pour objectif le déploiement d'une infrastructure nationale (au niveau des Etats-Unis) de mesures actives. Cette infrastructure est flexible et permet le recueil de diverses mesures actives. Cette infrastructure a été utilisée durant les deux ou trois années passées pour plusieurs campagnes de mesures, dont la détermination d'une matrice de distance dans Internet. L'infrastructure NIMI s'est aussi étendue en Europe, notamment en Suisse.
- En Europe, le projet RIPE (Réseaux IP Européens) [RIP01], tente de déployer une infrastructure semblable à l'infrastructure de NIMI en Europe. Par rapport à NIMI, RIPE fournit des services à des clients : RIPE se propose de réaliser toutes les études qui peuvent être demandées par des clients, en plus des services classiques d'accès à des statistiques globales d'utilisation des liens du réseau Européen de la recherche surveillés. A noter que des boîtes de mesure du projet RIPE ont été ou seraient sur le point d'être déployées dans le cadre de la plate-forme VTHD[8] par l'ENST[9], pour mesurer la qualité de service sur cette plate-forme.
- Le projet MINC[10] [ADA00] [CAC99] est un client du projet NIMI. Il utilise la diffusion de sondes actives par le biais du multicast pour inférer sur la structure interne du réseau et les propriétés sur tous les liens d'interconnexion ainsi traversés. En allant plus loin, c'est le sujet de la tomographie qui est au centre de ce projet qui se focalise sur certains aspects dynamiques du trafic, comme les propriétés du routage, les pertes et les délais. Toutefois, comme le multicast n'est pas un service disponible partout, et comme nous allons montrer que le trafic dans l'Internet n'est pas symétrique, l'intérêt du multicast dans cette tâche n'est concrètement pas évident. Aussi, le projet UINC (Unicast INC) a vu le jour et tente de reproduire le travail de MINC en unicast.
- Le projet *Netsizer* [NET01] de Telcordia (ex Bellcore) a pour objectifs de mesurer la croissance de l'Internet, les points durs de congestion, les délais, *etc*. Pour cela, depuis une ensemble de stations situées chez Telcordia, un programme teste la présence sur le réseau de toutes les adresses IP existantes et met à jour suivant les résultats une carte de l'Internet. Un des gros problèmes de ce projet reste ainsi un problème de représentation.
- Le projet américain AMP[11] de NLANR[12] [NLA01] [MCG00] est l'un des plus récent à avoir démarré. De ce fait, peu de résultats sont aujourd'hui disponibles, mais l'objectif de ce projet est de faire de « l'active probing ».

[5] NSF : National Science Fondation
[6] GPS : General Positioning System
[7] NIMI : National Internet Measurement Infrastructure
[8] VTHD: Vraiment Très Haut Débit
[9] ENST: Ecole Nationale Supérieure des Télécommunications
[10] MINC : Multicast-based Inference of Network-internal Characteristics
[11] AMP : Active Measurement Project
[12] NLANR : National Laboratory for Applied Network Research

## 2.2. Projets de mesures passives

Les projets de mesures passives sont apparus beaucoup plus récemment que les projets de mesures actives car ils nécessitent des systèmes de capture ou d'analyse du trafic en transit relativement avancés, et développés très récemment. Le principe des mesures passives est de regarder le trafic et d'étudier ses propriétés en un ou plusieurs points du réseau. L'avantage des mesures passives et qu'elles ne sont absolument pas intrusives et ne changent rien à l'état du réseau. De plus, elles permettent des analyses très avancées. En revanche, il est très difficile de déterminer le service qui pourra être offert à un client en fonction des informations obtenues en métrologie passive. Les systèmes de métrologie, actifs comme passifs, peuvent aussi se différencier en fonction du mode d'analyse des traces. Ainsi, le système peut faire une analyse en-ligne ou hors-ligne. Dans le cadre d'une analyse en-ligne, toute l'analyse doit être effectuée dans le laps de temps correspondant au passage du paquet dans la sonde de mesure. Une telle approche, temps-réel, permet de faire des analyses sur de très longues périodes et donc d'avoir des statistiques significatives. Par contre, la complexité maximale pour ces analyses reste très limitée à cause du faible temps de calcul autorisé. Une analyse hors-ligne par contre oblige la sonde à sauvegarder une trace du trafic pour analyse ultérieure. Une telle approche demande ainsi d'énormes ressources ce qui représente une limitation pour des traces de très longues durées. Par contre, une analyse hors-ligne permet des analyses extrêmement complètes et difficiles, permettant d'étudier des propriétés non triviales du trafic. De plus, comme les traces sont sauvegardées, il est possible de faire plusieurs analyses différentes sur les traces, et de corréler les résultats obtenus pour une meilleure compréhension des mécanismes complexes du réseau.

L'endroit idéal pour positionner des sondes de mesures passives est indéniablement dans les routeurs. CISCO a ainsi développé le module *Netflow* [CIS01] dans ces routeurs, qui scrute le trafic en transit, et génère régulièrement des informations statistiques sur ce trafic. *Netflow* a ainsi été utilisé dans de nombreux projets. L'expérience montre toutefois que les performances de netflow restent limitées (code écrit en Java et interprété), et que l'influence sur les performances du routeurs sont non négligeables.

Toutefois, le premier projet connu – le projet AT&T *Netscope* [FEL00] – a débuté il y a environ 5 ans et repose sur ce système *Netflow* de CISCO. Ce projet de mesure passive en ligne a pour but d'étudier les relations entre le trafic transitant en chaque nœud du réseau et les tables de routage des routeurs. L'objectif final est d'utiliser ces résultats pour améliorer les politiques et décisions de routage, afin d'équilibrer au mieux la charge dans les différents liens du réseau, et ainsi améliorer la qualité de service perçue par chaque utilisateur. C'est de la tomographie afin de trouver ensuite des politiques d'ingénierie des trafics adéquates.

Vern Paxson et al. d'ACIRI ont également conduit un projet de mesure passive en ligne dont l'objectif était de proposer un modèle pour les arrivées de flux et de paquets sur les liens de l'Internet. Ce travail [PAX95] achevé depuis 1995 a été et reste une référence dans le milieu de la recherche Internet. Cependant, aujourd'hui, avec l'apparition de nouvelles applications qui n'existaient pas à l'époque et avec les changements dans la façon d'utiliser l'Internet, ce travail doit être reconduit. Les résultats ne sont certainement plus valables aujourd'hui, et il n'existe aucune technique sûre d'extrapolation de ces résultats pour essayer de modéliser le trafic d'aujourd'hui.

De façon plus générale, le laboratoire CAIDA[13] [CAI01] à San Diego, Californie, est spécialisé dans l'étude du trafic Internet et mène un projet dont l'objectif est d'étudier sur le long terme l'évolution du trafic, avec l'apparition des nouvelles applications comme les jeux, le commerce électronique, etc. D'autre part, ce projet étudie aussi les variations en fonction du moment de la journée, du jour de la semaine, de la période de l'année, etc [CLA98] [MCC00]. Pour ce faire, le système de métrologie repose sur les modules OC3MON [APS97] et OC12MON qui permettent de traiter le trafic de liens IP/ATM dont les capacités respectent respectivement les normes OC3 (155 Mbps) et OC12 (622 Mbps). D'autre part, pour l'analyse statistique, CAIDA a développé la suite logicielle *CoralReef* [COR01] [KEY01] qui est complémentaire des systèmes OCxMON. D'autres études sont en cours, notamment sur les problèmes de représentation de l'Internet [HUF01], ou d'étude des délais. Pour plus d'information, le lecteur pourra consulter [CAI01].
A noter que les systèmes OCxMON sont aussi utilisés par Wordcom pour faire de la métrologie sur le réseau vBNS [VBN01].

Enfin, SPRINT a démarré il y a presque 3 ans un des projets les plus ambitieux du moment basé sur des mesures actives hors ligne. Ainsi, Sprint enregistre des traces complètes de tous les entêtes de tous les paquets qui transitent en certains point de son réseau IP. Cette granularité microscopique permet d'approfondir les analyses que l'on peut faire dans la compréhension des interactions qui existent entre tous les flux, les mécanismes des routeurs, etc. A noter que le système IPMON de Sprint, décrit plus en détail dans la suite, repose sur la carte DAG [DAG01a] conçue par l'université de Waikato en Nouvelle Zélande et qui se charge d'extraire les entêtes des paquets, de les estampiller suivant une horloge GPS [DAG01b] et de les stocker sur un disque dur [CLE00].

## 3. Le système de métrologie de Sprint

Le but du projet IPMON de Sprint est de collecter des données sur son Backbone Internet afin de fournir des bases fiables à différents projets de recherche. En particulier, des études sur le comportement de TCP, l'évaluation des performances du réseau, et le développement d'outils d'analyse du réseau.

L'approche métrologique dans IPMON comporte trois phases :

- la première consiste à capturer des traces sur le réseau
- la seconde partie concerne l'analyse des traces (selon divers paramètres)
- la troisième consiste à déduire à partir des analyses des informations pertinentes sur le réseau et sur les services qu'il doit assurer ; cette étape permet aussi de comparer des modèles de trafic au cas réel.

Le système de capture consiste à insérer des *splitters* optiques sur les liens au niveau des POP. Ainsi, on collecte et on date (à l'aide d'une horloge GPS) toutes les entêtes TCP/IP (44 octets). L'utilisation d'une horloge globale permet, entre autre, de déterminer les délais de bout en bout, des temps de passage dans les routeurs, d'avoir une idée sur le temps passé dans les files d'attentes des routeurs, etc. Enfin, ces informations sont transférées sur la plate-forme d'analyse.

Les avantages d'un tel système sont nombreux : en particulier, on peut citer :

- la transparence et la non intrusivité pour le réseau

[13] CAIDA : Cooperative Association for Internet Data Analysis

- les données collectées sont réalistes puisque provenant d'un réseau IP opérationnel
- capture de l'entête TCP/IP complète
- archivage des traces pour des exploitations futures
- analyse différée qui permet une exploitation très avancée

Mais un tel système a aussi des inconvénients :

- il requiert un développement sur un réseau opérationnel
- limitations technologiques (PC, espace disques)
- ne donne lieu qu'à des analyses différées
- 44 octets sont quelquefois insuffisants

Le système IPMON supporte des liens allant de OC3 (155 MBits/s) à OC48 (2480 MBits/s). Il est composé de PC Linux, d'une interface réseau POS/PCI (Packets Over Sonet) ainsi que de deux disques durs de taille importante (en effet une trace sur un lien OC3 de 24h, en supposant que la ligne est occupé à 40% au maximum, occupe 54 Go).

## 4. Résultats et conséquences

Cette partie présente quelques uns des paramètres qui ont été mesurés lors des premières études. Ce résumé, qui sera détaillé lors de la présentation, se contente de donner dans les grandes lignes quelques résultats importants, et surtout indique les tendances d'évolutions du réseau qui se dégagent.

### 4.1. Caractérisation du trafic

L'un des premiers travaux à faire pour commencer une étude métrologique d'un lien d'un réseau consiste à trouver une caractérisation du trafic qui passe sur ce lien. Le nombre de paramètres à mesurer semble sans limite. Toutefois, ce paragraphe va détailler quelques uns d'entre eux qui avaient parus importants dans le cadre de l'étude métrologique du réseau Sprint. Parmi eux, on trouve :

- le débit total sur chaque lien, qui permet d'avoir une idée du trafic transmis, de sa variabilité en fonction des différents moments de la journée, de la semaine ou de l'année. Cette donnée est une première information importante en vue de dimensionner un réseau ;
- la répartition du trafic par protocoles (UDP, TCP, ICMP, etc.) et par applications (Web, telnet, mail, etc.) ainsi que son évolution. En particulier, on peut voir au cours de la journée des pics d'utilisation de certains types d'applications, comme le mail en début de matinée, le web en fin de matinée et les applications orientées *stream* en fin d'après-midi. De même, on voit beaucoup plus d'achats électroniques au mois de décembre que le reste de l'année. Cette information présente donc une importance substantielle pour un opérateur qui pourra avoir à adapter le dimensionnement de son réseau et la QoS fournie en fonction des différentes périodes ;
- la distribution des tailles des paquets qui sont majoritairement de petits paquets d'une quarantaine d'octets (signalisation et acquittements TCP). Sur les différents liens étudiés, la répartition des petits paquets varie entre 40 et 75 %. Cette information remet peut être en cause les principes de conception des routeurs qui sont optimisés pour les gros paquets très minoritaires, et pas pour les petits paquets en nombre plus importants ;

- la taille des flux en octets et nombres de paquets. Il apparaît que la grande majorité des flux (environ 80 %) sont des flux très courts de moins de 10 paquets, ce qui explique la proportion de petits paquets rencontrés. Toutefois, ces 80 % de flux ne représentent qu'un tout petit pourcentage de la quantité de trafic généré. D'un autre côté, les longs flux de plus de 100 paquets qui représentent 2 à 3 % des flux dans l'Internet transportent plus de 60 % de la quantité d'information totale. Cette information semble capitale pour la conception des routeurs : en effet, faut-il privilégier les petits flux très nombreux, ou les longs flux qui transportent l'essentiel de l'information transmise ?
- le nombre de flux actifs simultanément sur un lien. Contrairement à ce que l'on croit généralement, il n'y a pas des millions de flux simultanément actifs. Ce nombre dépend bien sûr de la définition que l'on donne à flux, mais dans le pire des cas, il n'a jamais été dénombré plus de 50 000 flux actifs simultanément. Sachant qu'un routeur actuel peut gérer 64 000 files d'attente et plus, le « *Per Flow Queuing* » devient donc une stratégie réaliste de gestion dans les routeurs, contrairement à ce qui était admis auparavant.

### 4.2. Modélisation de trafic

Pour compléter et affiner la caractérisation du trafic faite précédemment, il est utile de modéliser le trafic Internet. L'objectif avoué est de donner un modèle réaliste des arrivées des flux, des paquets et des pertes. Cette action est indispensable, car les informations sur le trafic donneront les informations nécessaires pour la conception, le dimensionnement, la gestion et l'opération d'un réseau. D'autre part, elles donneront aussi les tendances d'évolution du réseau et de ces mécanismes, et permettront de concevoir des simulateurs permettant de confronter les nouveaux protocoles de la recherche à des trafics Internet réalistes. Jusqu'à présent, les ingénieurs et chercheurs en réseau utilisaient le modèle Poissonnien pour modéliser les processus d'arrivée de trafic et le modèle de Gilbert pour modéliser les pertes. Or ces modèles, même s'ils ont été utilisés pour modéliser le trafic téléphonique, sont incapables de représenter les rafales et les relations de dépendance qui existent entre les flux, les paquets et les pertes dans l'Internet.
En fait, il est aujourd'hui établi que le trafic Internet n'est pas Poissonnien, mais qu'il possède plutôt des propriétés de dépendance à long terme (*LRD : Long Range Dependence*) et d'auto-similarité [UHL01]. Les conséquences de cette découverte sont d'une importance capitale. En effet, les processus auto-similaires, par rapport aux processus Poissonniens, présentent la caractéristique d'avoir une variance très importante. Qualitativement, cela signifie qu'il est indispensable de surdimensionner les liens et les tailles des files d'attente dans les routeurs pour prendre en compte cette variance. Quantitativement, l'évaluation du facteur de Hurst d'un processus auto-similaire permettra de quantifier le niveau de surdimensionnement nécessaire pour un fonctionnement optimal du réseau.
De même, les relations de dépendance qui existent entre les pertes remettent en cause certains mécanismes de reprise sur erreur de TCP qui ont été conçus pour des pertes individuelles et indépendantes. Ayant observé dans les traces TCP des pertes en séquence et un niveau de dépendance élevé, des mécanismes de TCP comme « *Fast Recovery* » peuvent sans aucun doute être améliorés.

### 4.3. Analyse de la congestion

Suite aux informations données par la caractérisation et la modélisation du trafic sur les liens de son réseau IP, Sprint a choisi de surdimensionner son réseau d'un facteur allant de 2,5 à 3. Ce choix a la propriété de repousser les phénomènes de congestion en bordure du réseau

Sprint, et donc chez l'utilisateur qui est ainsi libre de mettre en place la solution de son choix pour combattre ces problèmes de congestion.

D'autre part, le fait que la majorité des flux soient petits et transportent peu d'information remet en cause l'intérêt du mécanisme de « *slow-start* » de TCP. En effet, le *slow-start* pénalise (ralentit) 80 % des flux, alors qu'ils ne représentent qu'un tout petit pourcentage de la quantité de trafic qui a peu de chance de congestionner le réseau.

Finalement, il apparaît que l'essentiels des problèmes rencontrés dans l'Internet sont au moins imputables à TCP qui bénéficie pourtant de peu d'effort de recherche. C'est sans doute là un axe d'étude à développer.

### 4.4. Délais dans les routeurs

Pour expliquer les délais observés lors des transmissions – ce délai étant garanti par SLA sur le réseau Sprint – il est important de mesurer l'effet de chaque routeur individuellement. Cette étude a donc permis de mesurer le délai et la gigue introduite par un routeur du réseau, c-à-d le délai sur un saut. Il a ainsi été mis en évidence un comportement perturbant des routeurs identifié sous le terme de « *coffee break* » et qui correspond à des périodes où, vu de l'extérieur, le routeur semble inactif. Ceci se traduit en fait par un délai moyen de traversée du routeur très court (inférieur dans presque 100% des cas à 100 μs), mais qui augmente parfois jusqu'à 35 ms, une telle valeur étant évidemment inacceptable pour permettre de bonnes performances à TCP et une bonne interactivité.

### 4.5. Matrices de trafic

Cette partie traite de l'étude des routes empruntées par les différents trafics, l'étude des tables de routage et des phénomènes de « *trunking* » dans le réseau. En se reposant sur une étude qui associe pour chaque paquet entrant dans un POP[14], le POP vers lequel le paquet va être transmis, un niveau important de redondance a été observé. Cette constatation semble donc indiquer que les fonctions de routage ne sont pas aussi dynamiques et que cela semble ouvrir la voie à de nombreuses possibilités de commutation optique dans le cœur du réseau en remplaçant le routage de cœur par une simple commutation des longueurs d'ondes WDM (bien sûr avec une classification des flux adéquate au niveau des routeurs de bordure). D'autre part, si cette possibilité ce confirme, c'est toute l'activité en rapport avec l'ingénierie des trafics qui se trouvera bouleversée. En effet, faire de l'ingénierie des trafics dans un cas comme celui-ci semble inutile vus les premiers résultats de cette étude. D'ailleurs, de tels résultats confortent Sprint dans l'idée que le problème de la QoS dans le réseau peut être résolu en surdimensionnant la capacité des liens et en ayant une politique de routage performante sur un réseau équilibré.
Il faut noter que le choix du protocole de routage interne IS-IS [CAL90] et son mécanisme de répartition de charge par flux (un peu comme dans MPLS[15] [ROS01]) remplit à lui seul les fonctions de routage et d'ingénierie des trafic. De plus, en évitant de désordonner les paquets des flux, grâce à la notion de flux physique, IS-IS offre un service intéressant pour favoriser le protocole de transport TCP qui s'adapte très mal aux transmissions non ordonnées.
Enfin, il a été observé que le trafic sur le réseau de Sprint est très asymétrique. Le niveau de symétrie, c'est à dire le taux des flux dont les acquittements transitent sur le lien opposé aux données se situe aux alentours de 10 % sur un lien du cœur du réseau, et augmente

[14] POP : Point Of Presence
[15] MPLS : MultiProtocol Label Switching

péniblement jusqu'à 20 % pour les liens d'accès. Le premier chiffre est dû au maillage du réseau de cœur. Le dernier se justifie par le nombre croissant de clients de Sprint qui possèdent plusieurs accès à Internet.

### 4.6. Explosion des tables de routage

Le dernier point présenté dans cet article traite de l'analyse des raisons de l'explosion des tables de routage (BGP[16]) dans les routeurs. Ces 6 dernières années, la taille moyenne d'une table de routage est passée de 15 000 à 150 000. Depuis 6 ans, le nombre d'entrées dans les tables de routage a grandi de façon exponentielle, de part le développement et la démocratisation du web au début, l'arrivée de CIDR[17] ensuite, et aujourd'hui à cause de NAT[18] [SRI99] [SRI01] [HOL01]. Cette croissance est aujourd'hui le principal empêchement pour pouvoir fournir une qualité de service optimale aux utilisateurs à cause des baisses de performance des procédures de routage qui doivent parcourir des tables de plus en plus importantes.

L'augmentation de la taille des tables de routage est également due à la fragmentation de l'espace d'adressage IP, pour environ 75%. Le « multi-homing » (en ayant un espace d'adresses appartenant à 2 AS[19]), les mécanismes de répartition de charge et les impossibilités d'agréger les espaces d'adressage en sont les autres causes principales.

Les limites de la « *scalabilité* » du réseau Internet semblent donc atteintes avec l'approche actuelle, vues les performances actuelles dans les routeurs. Pour réduire la taille des tables de routage, Sprint force les routeurs à agréger des espaces d'adressage même si l'espace global possède des trous appelés « trous noirs ». Ces trous noirs sont des espaces d'adresses appelés par un routeur (en utilisant BGP) mais qui ne sont absolument pas déservies par ce routeur. De fait, si un paquet à destination de ce trou noir arrive sur ce mauvais routeur, il n'atteindra jamais son destinataire. Cette solution dégrade naturellement la fiabilité du réseau. Mais l'augmentation des performances de routage est aujourd'hui à ce prix.

### 4.7. Bilan sur la QoS

L'objectif d'un opérateur qui veut être compétitif est incontestablement de fournir à ces utilisateurs la meilleure QoS possible. Cependant, dans le cas de Sprint, maximiser la QoS n'a rien à voir avec les architectures à fourniture de QoS traditionnelles telles IntServ [BRA94], DiffServ [BLA98] ou même MPLS. L'objectif est ici de fournir le meilleur service « best effort » possible afin de permettre aux utilisateurs d'utiliser ce service pour toutes les applications, y compris les applications de streaming temps-réel.
Vue la stratégie de Sprint qui consiste à surdimensionner la capacité du réseau, les mécanismes à différenciation de classes de service sont repoussés en bordure du réseau, chez les clients. En particulier, pour lutter contre les CDNs[20] tels Akamai, les opérateurs américains ont aussi adopté cette stratégie du surdimensionnement sur les « *peerings links* » qui représentaient jusqu'alors les goulots d'étranglement de l'Internet.

Finalement, l'optimisation de la QoS se fait grâce aux 4 mécanismes déjà décrits :

---

[16] BGP : Border Gateway Protocol
[17] CIDR : Connectionless Inter-Domain Routing
[18] NAT : Network Address Translation
[19] AS : Autonomous System
[20] CDN : Content Based Network

- le surdimensionnement de la capacité des liens ;
- la mise en place d'une topologie et d'une tomographie suffisamment statique évitant d'avoir à mettre en œuvre des mécanismes coûteux d'ingénierie des trafics ;
- l'utilisation pour garantir une tomographie favorable du trafic du mécanismes de répartition de charge d'IS-IS (qui permet aussi d'améliorer les performances de TCP en ne déséquençant pas les paquets) ;
- l'agrégation maximale des espaces d'adressage – même si cela entraîne une légère baisse de la fiabilité du réseau – pour réduire le nombre d'entrées dans les tables de routage et ainsi améliorer les performances des protocoles de routage.

## 5. Le projet Français METROPOLIS

Même si la métrologie bénéficie aujourd'hui d'un engouement considérable en France, notre pays possède entre 3 et 6 ans de retard par rapport aux principaux opérateurs américains, et 2 ans environ par rapport à certains pays européens comme la Grèce ou l'Espagne.
Pour combler ce retard, le premier projet de métrologie en France a enfin été avalisé par le RNRT et doit commencer en Octobre 2001. Les partenaires de ce projet sont : le LAAS, le LIP6, France Télécom R&D, l'INRIA, le GET, Eurécom et RENATER.

Les thèmes d'études qui seront abordés concernent :

- la classification du trafic et le dimensionnement du réseau
- l'analyse du réseau (protocoles, routeurs)
- la modélisation du trafic et de ses propriétés
- la définition de procédure de tarification et de mise en place de SLA

Le premier point fort de ce projet par rapport aux projets étrangers concerne l'utilisation des deux approches actives et passives de métrologie, ce qui permettra de corréler ces deux types de mesures. En particulier, il sera possible d'évaluer les perturbations induites par un réseau réel sur un trafic dont le profil est connu. Cette corrélation devrait apporter un plus indéniable par rapport aux projets en cours actuellement qui ne peuvent pas aller aussi loin sans cette combinaison.

Enfin, la seconde grande force de ce projet se situe dans la diversité des réseaux sur lesquels seront effectuées les mesures. En effet, 3 réseaux différents seront étudiés :

- un réseau expérimental avec le réseau VTHD ;
- un réseau public opérationnel avec le réseau Renater ;
- un réseau commercial : les plaques ADSL de France télécom.

## 6. Conclusion

La métrologie, au sens nouveau du terme, s'avère être la pierre angulaire de toute activité sur les réseaux IP grâce à la connaissance qu'elle apporte. Il est clair que même si cette discipline en est à ses premiers balbutiements, elle a déjà démontré toute sa puissance, et en dépouillant les premières traces, a mis en évidence de nombreux problèmes, et notamment de très nombreuses caractéristiques des réseaux et des trafics très éloignées des croyances collectives. L'état de l'art a également montré que les premières expérimentations étaient issues d'Amérique du Nord, et que la France était encore très en retard dans ce domaine (entre 3 et 6

ans). A noter que le premier projet français dans le domaine de la métrologie (METROPOLIS) doit démarrer en Octobre 2001.

## 7. Références


[ADA00] A. Adams, T. Bu, R. Caceres, N. Duffield, T. Friedman, J. Horowitz, F. Lo Presti, S. B. Moon, V. Paxson, D. Towsley, "The Use of End-to-end Multicast Measurements for Characterizing Internal Network Behavior", IEEE Communications, 38(5), May 2000

[ALM99a] G. Almes, S. Kalidindi, M. Zekauskas, "A One-way Delay Metric for IPPM", RFC 2679, September 1999

[ALM99b] G. Almes, S. Kalidindi, M. Zekauskas, "A One-way Packet Loss Metric for IPPM", RFC 2680, September 1999

[ALM99c] G. Almes, S. Kalidindi, M. Zekauskas, "A Round-trip Delay Metric for IPPM", RFC 2681, September 1999

[APS97] J. Apsidorf, "OC3MON: Flexible, affordable, high performance statistics collection", Proceedings of INET, June 1997

[BLA98] Blake, S., Black, D., Carlson, M., Davies, E., Wang, Z., Weiss, W., "An Architecture for Differentiated Services", RFC 2475, December 1998

[BRA94] Braden, R.,Clark, D.,Shenker, S., "Integrated Services in the Internet Architecture : An overview", RFC 1633, June 1994

[CAC99] R. Caceres, N. Duffield, D. Towsley, J. Horowitz, "Multicast-based Inference of Network-internal loss characteristics", IEEE Transactions on Information Theory, vol. 45, no. 7, November 1999

[CAI01] "CAIDA web site", http://www.caida.org

[CAL90] R. Callon, "Use of OSI IS-IS for Routing in TCP/IP and Dual Environments", RFC 1195, December 1990

[CIS01] "NetFlow Services Solutions Guide", http://www.cisco.com/univercd/cc/td/doc/-cisintwk/intsolns/netflsol/

[CLA98] K.C. Claffy, G. Miller, K. Thompson, "The nature of the beast: Recent traffic measurements from an Internet backbone", Proceedings of INET'98, Geneva, Switzerland, July 1998

[CLE00] J. Cleary, S. Donnelly, I. Graham, A. McGregor, M. Pearson, "Design principles for accurate passive measurement", PAM 2000, Hamilton, New Zealand, April 2000

[COR01] "CoralReef website", http://www.caida.org/tools/measurement/coralreef

[DAG01a] "Dag 4 SONET network interface", http://dag.cs.waikato.ac.nz/dag/dag4-arch.html

[DAG01b] "Dag synchronization and timestamping", http://dag.cs.waikato.ac.nz/-dag/docs/dagduck_v2.1.pdf

[FEL00] A. Feldmann, A. Greenberg, C. Lund, N. Reingold, J. Rexford, F. True, "Deriving traffic demands for operational IP networks: Methodology and Experience", ACM SIGCOMM Conference, Stockholm, 2000

[FRA01] C. Fraleigh, C. Diot, S. Moon, P. Owezarski, D. Papagiannaki, F. Tobagi, "Experiences Monitoring Backbone IP Networks", PAM (Passive and Active Measurements) Workshop, Amsterdam, The Netherlands, April 23-24, 2001

[HOL01] M. Holdrege, P. Srisuresh, "Protocol Complications with the IP Network Address Translator", RFC 3027, January 2001

[HUF01] B. Huffaker, M. Fomenkov, D. Moore, K. Claffy, "Macroscopic Analyses of the Infrastructure: Measurement and Visualization of Internet Connectivity and Performance", PAM'2001 (Passive and Active Measurements) workshop, Amsterdam, The Netherlands, April 2001

[KAL99] S. Kalidindi, M.J. Zekauskas, "Surveyor: An infrastructure for Internet performance measurements", Proceedings of INET'99, June 1999

[KEY01] K. Keys, D. Moore, R. Koga, E. Lagache, M. Tesch, K. Claffy, "The Architecture of the CoralReef: An Internet Traffic Monitoring Software Suite", PAM'2001 (Passive and Active Measurements) workshop, Amsterdam, The Netherlands, April 2001

[MCC00] S. McCreary, K. C. Claffy, "Trends in wide area IP traffic patterns", in ITC Specialist Seminar, Monterey, California, May 2000

[MCG00] T. McGregor, H.-W. Braun, J. Brown, "The NLANR network analysis infrastructure", IEEE Communications, vol. 38, no. 5, May 2000

[NET01] "Netsizer web site", http://www.netsizer.com

[NLA01] "AMP web site", http://watt.nlanr.net

[PAX95] V. Paxson, and S. Floyd, "Wide-Area Traffic: The Failure of Poisson Modeling", IEEE/ACM Transactions on Networking, Vol. 3 No. 3, June 1995

[PAX98] V. Paxson, G. Almes, J. Mahdavi, M. Mathis, "Framework for IP Performance Metrics", RFC 2330, May 1998

[PAX00] V. Paxson, A. Adams, M. Mathis, "Experiences with NIMI", PAM (Passive and Active Measurements) Workshop, 2000

[RIP01] "RIPE NCC web site", http://www.ripe.net

[ROS01] E. Rosen, A. Viswanathan, R. Callon, "Multiprotocol Label Switching Architecture", RFC 3031, January 2001

[SRI99] P. Srisuresh, M. Holdrege, "IP Network Address Translator (NAT) Terminology and Considerations", RFC 2663, August 1999

[SRI01] P. Srisuresh, K. Egevang, "Traditional IP Network Address Translator (Traditional NAT) ", RFC 3022, January 2001

[UHL01] S. Uhlig, O. Bonaventure, "Understanding the Long-Term Self-Similarity of Internet Traffic", Technical Report Infonet-2001-04, April 2001

[VBN01] "vBNS web site", http://www.vbns.net