European Journal of Computer Science and Information Technology
Vol.10, No.5, pp.1-23, 2022
Print ISSN: 2054-0957 (Print),
Online ISSN: 2054-0965 (Online)

De l'optimisation de la performance du signal dans la gestion du trafic des réseaux locaux d'entreprise en RDC : Modèles pour les réseaux de transmission

Signal performance optimization in the local area network trafic management in the DRC : Models for transmission networks

[1]Dr. RaphaeL Grevisse Yende, [2]Dr. Sr Tshiela Marie-Alice Nkuna , [3]Kazadi Pamphile Mulumba , [4]Ntumba Freddy Katayi , [5]Kaseka Viviane Katadi , [6]Musubao Patient Swambi & [7]Muamba Bernard. Tshiasuma

Citation : RaphaeL Grevisse Yende ; Sr Tshiela Marie-Alice Nkuna ; Kazadi Pamphile Mulumba ; Ntumba Freddy Katayi ; Kaseka Viviane Katadi ; Musubao Patient Swambi ; Muamba Bernard. Tshiasuma (2022) Signal performance optimization in the local area network trafic management in the DRC : Models for transmission networks. *European Journal of Computer Science and Information Technology*, Vol.10, No.5, pp.1-23

RESUME La haute disponibilité des réseaux informatiques est une condition indispensable dans les grandes entreprises et des fournisseurs de services. Par conséquent, les administrateurs de réseaux informatiques sont appelés à faire face aux différents défis croissants liés aux temps d'arrêt non programmés des services ; au manque d'expertise ; à l'insuffisance des outils ; à la complexité des technologies ; à la consolidation du marché et à la concurrence, quant à fournir une meilleure qualité des services. Il sied de rappeler que, l'organisation des réseaux de téléinformatiques vise à donner la maîtrise des phénomènes qui s'y produisent, à l'occasion du traitement des communications, et plus généralement des services qu'ils offrent. Ces phénomènes sont gouvernés par le hasard de l'apparition des requêtes, et s'étudient indépendamment des choix de la technologie mis en œuvre. Ils sont justiciables du formalisme du calcul des probabilités, et donnent naissance à la notion de trafic, qui joue un rôle central dans leur appréhension, étant donnée, qu'ils conditionnent pour une large part la structure effective des réseaux modernes. La présente recherche s'interroge spécialement sur les problèmes d'optimisation de la performance du signal dans la gestion du trafic des réseaux locaux d'entreprises en République Démocratique du Congo, en présentant les différents éléments de la théorie du trafic réseau et de la qualité des services démontrant ainsi, leur impact sur les architectures des réseaux et des systèmes informatiques, et introduisant par la même occasion les grands problèmes que doit affronter quotidiennement les administrateurs des réseaux informatiques. L'objectif capital visé par la présente recherche est d'évaluer la performance de développement des réseaux informatiques et de télécommunications selon leurs applications dans les domaines de la conception des réseaux, du dimensionnement et de la caractérisation des équipements, de la mesure de la Qualité de Service, et des techniques de la planification des réseaux.
MOTS CLES : *Optimisation, QoS, Performance, Réseau local, Administration, Trafic, Signal, Téléinformatique, Transmission, Modèle, Entreprise, RDC.*

ABSTRACT

The high availability of computer networks is a prerequisite in large companies and service providers. Thus, computer network administrators are called upon to face the various growing challenges related to unscheduled downtime of services; lack of expertise; lack of tools; the complexity of technologies; market consolidation and competition to provide better quality services. It should be remembered that the organization of the networks of telecommunication aims at giving control of the phenomena which occur there, during the treatment of the communications, and more generally of the services which they offer. These phenomena are governed by the randomness of the appearance of the requests, and are studied independently of the choices of technology implemented. They are amenable to the formalism of the calculation of

[1]*Ingénieur Docteur ès sciences en Systèmes Informatiques (Télécommunications et Réseaux Informatique) de l'Université Ouverte du Commonwealth (COU-UK), et Enseignant-chercheur (Professeur des Universités) attaché à l'Université de Bas-Uélé (RDC).*
[2]*Ingénieur Docteur ès sciences Informatique (Machine Informatique & Informatique Décisionnelle) de l'Université de Kinshasa (UNIKIN), et Enseignante-chercheure (Professeur des Universités) attachée à l'Université Notre Dame du Kasayi (RDC).*
[3, 5,6] *Ingénieurs (Chefs des travaux) respectivement en Réseaux Informatiques et Conception des systèmes informatiques attachés à la Faculté de l'Informatique de l'Université Notre Dame du Kasayi (RDC).*
[4]*Ingénieur en réseaux informatiques et Télécommunications de l'Université Notre Dame du Kasayi et Chef des travaux attaché à la Faculté de l'Informatique de l'Université de KANANGA (RDC).*
[7]*Ingénieur en Réseaux Informatiques et Assistant de Recherche attaché à la Faculté de l'Informatique de l'Université Notre Dame du Kasayi (RDC).*

probabilities, and give birth to the notion of traffic, which will play a central role in their apprehension, given that they condition for a large part the effective structure of modern networks. This research specifically examines the problems of optimization of signal performance in traffic management of local area networks of companies in the Democratic Republic of Congo, by presenting the different elements of network traffic theory and quality. Thus demonstrating their impact on the architectures of networks and computer systems, and at the same time introducing the major problems faced by IT network administrators on a daily basis; The main objective of this research is to evaluate the development performance of computer and telecommunications networks according to their applications in the areas of network design, equipment dimensioning, and characterization and quality measurement. Service, and network planning techniques.

KEYWORDS: Optimization, QoS, Performance, Local network, Administration, Traffic, Signal, Teletransmission, Transmission, Model, Company RDC.

INTRODUCTION

L'évolution technologique de ces dernières années a conduit les sociétés modernes à adopter de nouvelles habitudes face au monde du travail dans le souci d'augmenter la productivité et de rendre plus performant les activités humaines. Toutefois, nous ne sommes sans ignorer que les premiers outils de communications à distance utilisés par les hommes furent la fumée, le feu ainsi que la lumière, qui ont évolués surtout dans le domaine militaire permettant la communication plus rapide et le déplacement le moins possible. Aujourd'hui, ils ont abouti à toutes les technologies informatiques et dans les entreprises, le partage de données est devenu une des tâches capitales pour devenir plus concurrentielle par rapport autres. Ce qui rend la répartition des tâches via les réseaux informatiques, un outil majeur pouvant permettre d'accomplir plus efficacement les différentes missions assignées de l'entreprise moderne.

La notion de la gestion des flux partagés et du trafic réseau reste une activité importante et délicate de l'administration des réseaux informatiques. Le nombre croissant d'utilisateurs et la nature même des informations à partager dégradent du jour au lendemain la performance des réseaux informatiques ; qui se heurte désormais, à la mauvaise distribution des flux des données, influant directement sur la qualité de services des réseaux informatiques qui suscite l'apport continu des techniques pouvant ainsi permettre l'amélioration à la fois des signaux, des supports ainsi que des équipements réseaux [27]. Dans la plupart de pays en développement, le problème de connexion reste une réelle difficulté, vu que tout le monde veut se connecter en temps réel pour effectuer différentes recherches sur l'internet en créant ainsi le problème de la congestion[2] dans les réseaux informatiques. Cette difficulté est une réalité qui se fait sentir dans les entreprises congolaises qui utilisent la connexion internet comme moyen ultime de partage rapide et sécurisé des informations. C'est dans cette optique, que la présente recherche s'inscrit afin de répondre aux problèmes de l'optimisation de la performance du signal dans la gestion du trafic dans les réseaux locaux d'entreprises congolais. Ainsi, après une minutieuse analyse, nous nous sommes posée de la question de savoir quels peuvent être les facteurs influant sur l'optimisation de la performance du signal dans la distribution de flux des données dans les entreprises congolaises et quel mécanisme efficace devra-t-il répondre aux divers besoins des utilisateurs congolais dans la gestion du trafic des réseaux informatiques.

Ainsi, pour pallier les problèmes supputés ci-haut, la présente recherche estime que les facteurs influant directement sur l'optimisation de la performance du signal dans le gestion du trafic des flux de données dans les réseaux informatiques seraient la bonne application dans les domaines de la conception des réseaux, du dimensionnement et de la caractérisation des équipements, de la mesure de la Qualité de Service et des techniques de la planification des réseaux. Par l'emmagasinage des données non spécifique, le mauvais classement et partage de la bande passante allouée et la sécurité non adaptée contribueraient autant à la mauvaise gestion du trafic des données dans les réseaux informatiques. Et le mécanisme efficace pour répondre aux divers besoins des utilisateurs congolais serait l*a* mise en place des modèles de développement des réseaux informatiques ainsi les

[2]*En informatique, il est question de la congestion lorsque les réseaux informatiques ont des ressources insuffisantes pour face à toutes les demandes de transfert qui sont adressées à ces derniers. On parle également de l'encombrement ou de l'embouteillage dans les réseaux informatiques.*

différentes techniques telles que : Le filtrage des données afin d'assurer une haute disponibilité des données sans interruption des services.

METHODES ET MATERIELS

METHODES

S'agissant de la présente rubrique ; il est important de signaler que le vocable « *Méthode* » [21] est défini suivant plusieurs auteurs interprétant de déférentes manières, mais tous, dans le but d'atteindre les objectifs poursuivis par chacun d'eux. Ainsi, Tout chercheur doit se focaliser sur une démarche susceptible de l'orienter à atteindre son objectif et à résoudre le problème qu'il tente d'élucider dans son observation scientifique. Quant à notre travail, nous avons usé des méthodes analytiques et systémiques qui nous ont permis d'étudier rationnellement les grandeurs manipulables pouvant influer dans la mauvaise gestion du trafic des flux des données dans les réseaux locaux d'entreprises congolaises en schématisant un ensemble d'éléments complexes en relation dans la gestion de la performance et la qualité de service des réseaux informatiques afin d'aboutir à une modélisation optimisée, simplifiée et efficace. La technique d'interview et celle de l'observation ont été les moyens par excellence qui nous permis de collecter les informations indispensables concourant à l'élaboration de la dite recherche. Cependant, la technique de la documentation a également suppléé aux techniques précédemment supputées ci-haut.

Il sied autant de rappeler que 52 entreprises congolaises ont été sélectionnées aléatoirement dans le cadre d'échantillon sur l'étendue du territoire congolais, en l'occurrence de deux entreprises par province en utilisant les réseaux sociaux comme le champ de rencontre. Toutefois, les différents entretiens n'ont pas été aisés à cause du caractère virtuel de notre mode d'investigation. C'est alors, que notre échantillon a été réduit de plus de la moitié c'est-à-dire 37 entreprises. Mais les autres entreprises restantes nous ont fourni les informations nécessaires à la réalisation de notre recherche. Ces informations ont conduit à la formulation des recommandations et des pistes de solutions proposées à la fin de cette recherche ainsi que les différents modèles des réseaux de transmission envisagés.

MATERIELS

GESTION DE QUALITE DE SERVICE

La qualité de service (QoS) peut se définir comme étant l'*ensemble des phénomènes pouvant influencer les performances du service qui détermine le degré de satisfaction de l'utilisateur de ce service*. La qualité de service peut être appréhendée différemment selon les acteurs (clients, operateurs,) et le rôle (demandeurs, fournisseurs, consommateurs). Si la qualité de service se résume habituellement à l'écoulement du trafic dans un réseau, beaucoup d'autres facteurs sont à prendre en compte. Ainsi, de nombreux facteurs comme la perte d'alimentation, la surcharge ou la panne de plate-forme de service, la rupture d'un lien de transmission, ... sont plus souvent la cause d'une mauvaise qualité perçue par l'utilisateur qu'une déficience du réseau au niveau de l'acheminement des données.

Une application offrant une excellente qualité de service se doit d'être fiable, robuste et tolérant aux pannes avant même de se préoccuper du bon acheminement des données. En règle générale, un opérateur exige que la plate-forme qui rend le service d'afficher un MTBF (*Mean Time Between Failure*) supérieur ou égale à 99,999% soit une interruption de service inférieure à 5 minutes par an. Enfin, il est aussi nécessaire de définir qui quantifie la qualité de service et quels sont les indicateurs mesurables a même de la vérifier. Dans le cas où ; c'est un utilisateur final qui a demandé un service, c'est la qualité perçue ou *Quality of Experience* qui sera mise en avant. Et dans le cas où, c'est un opérateur de service ou de réseau, ce sont des objectifs mesurables correspondant aux paramètres de la qualité de service qui serviront à évaluer le bon respect des engagements de qualité de service.

La gestion de la QoS [5], et donc sa garantie, ne peut en aucun cas se restreindre à une technologie ou une partition de réseau mais doit être abordée dans son intégralité. Elle doit former un tout et être considérée de bout en bout même si elle est ensuite subdivisée afin de rendre le problème solvable. La qualité de service finale est

égale à la qualité de service rendue par l'élément le plus faible dans toute la chaine de distribution du service. Ainsi, il ne sert à rien de garantir un niveau de qualité de service excellent dans le réseau de l'opérateur si la liaison Wifi du réseau domestique entre le terminal de l'utilisateur et son modem ADSL présente un taux de perte de 2%.

A l'inverse, les paramètres utilisent entre fournisseur de service et opérateur de réseau et entre opérateurs de réseaux eux-mêmes se doivent de répondre à des exigences techniques. Les paramètres de qualité de service associés à un flux de données sont principalement :

— *Le débit* (*bandwidth*) du flux qui désigne la quantité d'informations écoulée par unité de temps exprimée en bit/s. Il faudra également déterminer le point de mesure auquel correspond le débit. En effet, le débit indiqué n'aura pas la même valeur suivant la couche (*au sens modèle ISO*) où il est mesuré, du fait des diverses encapsulations. Le débit est également associé à la notion de bande passante. Dans la suite du mémoire, la bande passante sera utilisée pour designer la capacité d'écoulement d'un lien ou d'un équipement et le débit pour exprimer le volume de données émise ou nécessaire par une application ou un service par unité de temps ;

— *Le taux de perte* (l*oss rate / error rate*) qui désigne la probabilité maximale de perte de données ou de paquets. Ce paramètre, sans unité, est bien entendu très inférieur à 1. nous chercherons toujours à se rapprocher d'un taux de perte égal à 0 qui désigne une qualité de service excellente.

— *Le délai* (*delay*) de transmission, exprime en ms, désigne le temps nécessaire pour acheminer un volume élémentaire de données de la source jusqu'à la destination. Ce paramètre peut correspondre à une valeur maximale à ne pas dépasser, une mesure moyenne ou minimale ... mais en aucun cas ne désigne le temps total de transfert des données. Il est mesuré de bout en bout ou entre deux points de référence comme étant le temps nécessaire à l'acheminement d'une unité de volume, en général, un paquet, entre les deux points de mesure. Il correspond au temps de transport proprement dit (*ex. 5 µs/km pour la vitesse de propagation de la lumière dans une fibre optique*) plus au temps de traversée des différents équipements.

— *La gigue* (*jitter / delay variation*) qui désigne la variation du délai de transmission, exprimée en ms. Si le débit et le taux de perte concernent toutes les applications, le délai et la gigue affectent plus particulièrement les applications à temps réel ou requérant une grande synchronisation. Par exemple une gigue trop élevée va affecter la synchronisation de l'annulation d'échos pour un service conversationnel.

A ces paramètres, s'ajoutent des critères de disponibilité du transfert des données dans le réseau :

→ *La disponibilité du réseau* qui se traduit par la probabilité qu'un élément tombe en panne. Le MTBF pour l'ensemble de la chaine de service peut être calculé sur la base des MTBF des éléments qui la constituent,

→ *La durée d'interruption de service* qui complète la disponibilité. Le MTTR (*Mean Time To Repair*) pour l'ensemble de la chaine de service peut être également calculé sur la base des MTTR des éléments qui la constituent ;

→ *La probabilité de refus de transfert* qui est lié à la notion *d'admission d'appel*. L'opérateur peut en effet décider, en fonction du trafic présent, si son réseau à les capacités suffisantes ou non pour acheminer ce nouveau service. Cette probabilité sera maintenue la plus faible possible. A titre d'exemple, le réseau téléphonique commuté est dimensionné pour garantir un taux de rejet inférieur à 0,05 %.

Il est clair que tout d'abord qu'un service ne peut être utilisé que s'il est fourni, et à cette fourniture doit être associée une description de la qualité offerte. Du point de vue du fournisseur, le concept de performance du réseau est un concept par lequel des caractéristiques réseaux peuvent être définies, mesurées et contrôlées en vue d'atteindre un niveau de qualité de service donné[12]. Il relève de la responsabilité des fournisseurs de réseau de combiner adéquatement différents paramètres de performance, de façon à atteindre à la fois leurs objectifs économiques et les objectifs de satisfaction de l'usager. Le degré de satisfaction perçu lors de la fourniture d'un service peut s'évaluer par la performance présentée dans les différents domaines suivants :

— *La qualité du support logistique* - Il s'agit ici de la capacité d'une organisation à fournir le service dans des délais adéquats, à assurer la bonne gestion de ce service en termes de facturation par exemple, d'assistance à l'usager. Ceci est particulièrement opportun dans le cas des accès à Internet, de la gestion des abonnements (*abonnement et résiliation*) mais aussi dans le cas du mobile avec les multiples forfaits et modes de facturation.

— *La facilité d'utilisation* - Il s'agit de rendre aisée l'utilisation du service par l'usager, de lui éviter des erreurs de manipulation, de l'aider à « *naviguer* » aisément à travers toutes les possibilités des sites. On voit encore l'intérêt immédiat d'une interface conviviale pour le terminal mobile avec ses multiples menus, mais aussi pour Internet et ses innombrables sites et offres de services.

— *La sécurité* - Il s'agit de protéger l'usager et le réseau contre des utilisations frauduleuses et malveillantes des services offerts. L'exemple le plus simple est celui de la confidentialité des informations transportées, mais les protections apportées par les mots de passe aux comptes informatiques ou aux terminaux mobiles procèdent aussi de ce domaine. La protection des moyens de communication contre des événements catastrophiques tels que les tremblements de terre ou toute autre destruction est aussi un élément fondamental de la sécurité.

— *La capacité d'utilisation du service* - Il s'agit de la capacité du service à pouvoir être utilisé lorsque celui-ci est requis par l'usager, c'est-à-dire d'une part, la capacité à pouvoir l'obtenir puis, une fois obtenu, la capacité à continuer à fournir le service pendant la durée requise. Un exemple un peu caricatural mais explicite de cette performance est celui du voyage en avion : il s'agit d'abord d'obtenir un billet et de partir à l'heure, puis bien sûr de voyager sans incident jusqu'à destination. Le parallèle avec les services de télécommunication est immédiat, par exemple pour un appel téléphonique mobile ou fixe, il faut d'abord obtenir le réseau puis le destinataire et enfin ne pas être coupé pendant la communication. De même pour le Web, il faut d'abord accéder à son fournisseur et au serveur puis pouvoir ou transférer un fichier sans perte d'information, ou communiquer en audio ou vidéo sans problème notable de qualité. De manière générale, nous parlerons pour ces deux aspects d'*accessibilité* et de *continuité* du service, et ce sont eux que nous allons maintenant traiter en détail *via* les paramètres de performance des réseaux.

LES CLASSES DE SERVICES

Les paramètres de la qualité de service élucidés ci-haut peuvent ensuite regroupés entre eux en fonction des besoins des applications et des services. Ces groupes forment alors des *Classes de Services* (*Class of Services*). Les requêtes de QoS des applications ou des services seront toujours affectées à une classe de service donnée. A chaque classe, correspond un ensemble de paramètres de QoS avec des objectifs quantifies.

Plusieurs modèles de Classes de Service ont été standardisés et peuvent être utilisés indifféremment. Ici, Chaque operateur définit également ses propres classes de services avec des objectifs quantifiés différents. Il est très complexe de synthétiser les différentes propositions. En effet, même le nombre de classes de services est sujet à des débats intensifs dans les instances de standardisations[11]. Dans la suite du mémoire, il s'agira plus de se référer au concept de classes de services qui consiste à regrouper entre-elles les applications ayant des exigences de qualité de service similaires, plutôt qu'à se référer explicitement à tel ou tel modèle. Cependant, une classification des principales applications est fournie :

— *Voix :* Regroupe toutes les applications du type conversationnel (*Voix, Visio, Conférence, ...)* ayant pour contrainte forte des objectifs sur le délai et la gigue. Elles sont également sensibles au taux de perte bien qu'il ne soit pas possible de retransmettre les données et requièrent des débits assez faibles ;

— *Vidéo :* Regroupe toutes les applications multimédia diffusées ou non (*Video à la Demande, la télévision sur IP – IP TV, ...)* ayant pour contrainte forte le taux de perte et le débit et dans une moindre mesure le délai et la gigue ;

— *Données* **:** Regroupe toutes les applications de transfert de données ayant pour seule contrainte, un taux de perte nul et qui s'accommodent d'un délai et d'une gigue quelconque. Un débit garanti caractérise cette classe sans toutefois en faire une contrainte stricte ;

— *Défaut :* Désigne toutes les applications n'exigeant aucune garantie de la qualité de service. Bien connu sous l'anglicisme *« Best-Effort »,* c'est le mode de transport du protocole IP.

LES RESSOURCES DE LA QUALITE DE SERVICE

Les ressources sont étroitement liées aux paramètres de la qualité de service. En effet, une qualité de service donnée sera respectée si et seulement si les ressources influant sur le débit, la perte, le délai et la gigue sont bien disponibles. Si la ressource liée au débit est simple à définir[3], il n'en est pas de même pour les autres paramètres. Le délai et la gigue sont étroitement liés aux capacités d'acheminement du réseau en termes de longueur de connexion, de temps de transit dans les équipements ... Le taux de perte est lui aussi lié aux capacités d'écoulement du trafic par le réseau. Cependant, il existe beaucoup d'autres ressources comme celles liées aux plates-formes de service, aux terminaux ... Qu'il s'agisse de puissance de calcul, de capacité mémoire ou disque, de capacité d'affichage ou de reproduction sonore, toutes ces ressources nécessitent d'être vérifiées avant de pouvoir délivrer le service au client. Enfin, et à l'aube de l'extinction du plan d'adressage IPv4, les adresses sont également une ressource qu'il est nécessaire de gérer au fur et à mesure que le nombre de clients raccordés progresse[23].

PERFORMANCE DES RESEAUX

La performance des réseaux est mesurée en terme de paramètres qui sont significatifs du point de vue des fournisseurs de réseaux et de services, et qui sont utilisés pour la conception des systèmes, la configuration des équipements, le dimensionnement du réseau, la maintenance... de manière à obtenir la satisfaction du fournisseur et de l'usager.

L'usager moyen ne s'intéresse pas à la façon dont un service particulier est assuré, ni aux problèmes de conception interne des systèmes. Par contre, comme nous l'avons vu, il est intéressé par un certain nombre de paramètres traduisant sa perception de la qualité du service[24]. Ces paramètres intéressant l'usager ne peuvent cependant pas être utilisés directement pour établir les spécifications de qualité pour les réseaux. Il faut donc aussi définir une qualité réseau exprimée de façon qualitative et quantitative de manière à donner au fournisseur des informations sur :

— les performances exigées des systèmes ;
— la planification des réseaux, le dimensionnement des équipements ;
— l'exploitation et la maintenance.

A cet effet, et en relation avec les critères de perception de l'usager, un certain nombre de paramètres de performance réseau ont été définis. L'approche suivie consiste, de manière logique, tout d'abord définir des *paramètres globaux*, appelés aussi *paramètres de performance de bout en bout*, puis des paramètres au niveau de chaque segment, élément du réseau, que nous qualifierons de *paramètres intermédiaires*.

PARAMETRES GLOBAUX ET INTERMEDIAIRES

Les paramètres de performance sont dits « *paramètres de performance de bout en bout ou global »* car ils caractérisent des délais à respecter entre deux interfaces usagers ou extrémités de réseaux et sont ceux que perçoivent le mieux les usagers c' est-à-dire, qu'ils sont basés sur le principe traduisant au plus près de l'usager la perception globale qu'il a de la performance du réseau indépendamment de la constitution exacte de celui-ci. Ainsi, les éléments ci-après constituent ces paramètres :

— *Temps d'établissement* - C'est le temps compris entre la tentative d'appel ou d'ouverture de session et l'indication que la communication est établie (*ou ne peut l'être).* Les événements pris en référence pour évaluer les délais seront des occurrences de messages (*message initial d'adresse, IAM, et message de réponse, ANM, par exemple*) ou de paquets (*Call Request packet*, CR, et *Call Connected packet*, CC).

[3] *Elle correspond à la capacité disponible sur les liens et les équipements.*

— *Temps de libération (relâchement) de la communication* - C'est le temps compris entre l'émission par le terminal qui arrête la communication du signal de demande de libération et la réception par l'autre terminal de l'indication du réseau de cette demande. On retrouve des événements de référence semblables à ceux de l'établissement (*messages Release et Disconnect, clear request packet, clear indication packet*).

— *Probabilité d'échec d'établissement* – C'est le taux de demandes d'établissement de communication « *bloquées* », c'est-à-dire refusées par les mécanismes d'acceptation pour manque de ressources, ou temps de traitement excessifs (*conduisant par exemple à abandon puis renouvellement par l'usager*), ou perte de message, etc.

Alors quels sont « *les paramètres intermédiaires* », des paramètres de niveau inférieur, plus proches des équipements constituant les réseaux (*liens, nœuds, etc.),* qui assurent la contribution individuelle de la performance de chaque élément du réseau permettra d'atteindre l'objectif global de performance. Ainsi, les éléments ci-après constituent ces paramètres :

— *Temps d'établissement (ou de sélection) à l'accès* – C'est le temps compris entre le moment où les informations requises pour déterminer la direction sortante sont disponibles, jusqu'au moment où l'information est retransmise au nœud suivant.

— *Temps de libération à l'accès* - C'est le temps compris entre l'émission par le terminal qui arrête la communication du signal de demande de libération et la libération ou la réception par ce même terminal de l'indication locale du réseau confirmant la libération, et ainsi que le terminal puisse réinitialiser une nouvelle demande de communication.

— *Temps de traversée d'un message d'ouverture de communication* – C'est le temps nécessaire pour traiter et retransmettre un message d'ouverture dans un nœud intermédiaire. C'est typiquement le cas pour un centre de transit recevant un message initial d'adresse destinatrice (IAM) [14].

— *Temps de transfert d'un message* - Il s'agit cette fois de durées de transfert de messages sans traitement particulier associé, une simple fonction de routage est effectuée.

— *Temps de transfert de paquet* - Il s'agit du même concept, c'est le délai nécessaire au bon transfert du paquet, qui peut être par exemple dans le cas d'une connexion virtuelle un paquet de type *Call Request*.

— *Délai d'émission d'indication d'arrivée d'appel* – C'est l'intervalle de temps qui s'écoule entre l'instant où l'identification du demandé est disponible dans le nœud de terminaison et l'instant où le signal d'arrivée d'appel est envoyé au demandé. En téléphonie, ceci correspond dans le centre arrivé à la réception du numéro du demandé et à l'envoi de la sonnerie d'appel au terminal (*fixe ou mobile*), comme spécifié dans la recommandation Q.543 de l'UIT.

— *Temps de traversée d'un message de réponse* - C'est le temps nécessaire pour traiter et retransmettre un message de réponse au nœud suivant ; il indique que la connexion a bien été établie et que le transfert des informations de niveau usager peut commencer. C'est typiquement le cas d'un message ANS (*answer*) pour un centre de transit. C'est en quelque sorte le symétrique du message d'ouverture.

— *Temps de transfert de signalisation* - C'est le temps que met un nœud de commutation pour transférer un message d'un système de signalisation à un autre. L'intervalle est le temps qui s'écoule entre le moment où le message est reçu en provenance d'un système de signalisation, et celui où le message correspondant est transmis à un autre système de signalisation.

— *Temps de transmission* – C'est le temps nécessaire à la propagation du message, ou paquet, sur le support physique (*il peut être terrestre, sous-marin, par câble coaxial ou par fibre optique ou aérien par satellite*). C'est bien sûr, une fonction de la distance parcourue par l'information.

— *Taux de perte de message de signalisation et de paquet* – C'est le taux de messages de signalisations perdus pour cause interne (*délai, erreur, défaillance*) dans un nœud du réseau.

— *Probabilité d'échec d'établissement à un nœud* – C'est le même paramètre que le paramètre global mais spécifié au niveau d'un nœud. C'est donc encore le taux de demandes d'établissement de communication « *bloquées* », c'est-à-dire refusées par les mécanismes d'acceptation pour cause de manque de ressources, ou de temps de traitement excessifs (*conduisant par exemple à abandon puis renouvellement par l'usager*), ou perte de message, etc.

— *Délai d'authentification, délai d'obtention des informations de routage* - Il s'agit de paramètres spécifiques à la téléphonie mobile. Le délai d'authentification spécifie le temps requis pour cette opération (*accès éventuel à une base de données et traitement*). *Le délai d'obtention* des informations de routage correspond au temps d'interrogation du HLR (*Home Location Register*), plus, en cas de *roaming*, celui du VLR (*Visitor Location Register*).

PARAMETRES EN PHASE D'ACCES ET DE TRANSFERT

Des paramètres précédents se dégagent aussi deux autres notions : celles d'établissement ou rupture de connexion et celle de communication établie. Nous parlerons plus généralement de *phase d'accès (ou de désengagement)* et de *phase de transfert* [1]. Deux grandes familles de paramètres de performance sont définies selon ces deux notions. Il existerait deux raisons majeures pour distinguer ces deux familles de paramètres : l'aspect *service* et l'aspect *support*.

L'aspect service a déjà été présenté en introduction à ce chapitre. Il s'agit d'abord d'accéder au réseau et à l'usager ou fournisseur de service destinataire de la demande, et ceci dans des délais raisonnables. Ceci correspond pour un service avec connexion à l'établissement de la connexion, mais ce peut être aussi la phase d'accès à une « *gateway* », ou à son fournisseur de service. Puis, la communication étant établie, le transfert des informations entre les deux (ou plus) usagers devra s'effectuer avec une qualité suffisante pour garantir sa bonne exploitation (*intégrité sémantique, spatiale, temporelle, etc*.). Mais il y a aussi que le support physique utilisé pour chacune de ces phases peut être différent. C'est bien sûr le cas lorsqu.il y a séparation des réseaux de *commande* et de *transport*. De manière très simple, comme présenté au chapitre 1, on a déjà en téléphonie une séparation claire entre le réseau de signalisation n° 7 avec ses points sémaphores, qui sert à l'établissement et la rupture des communications, et le réseau transport de la parole avec ses liaisons et ses centres de commutation à 64 kbit/s. Cette séparation est encore plus distincte avec le concept de NGN (*Next generation Network*).

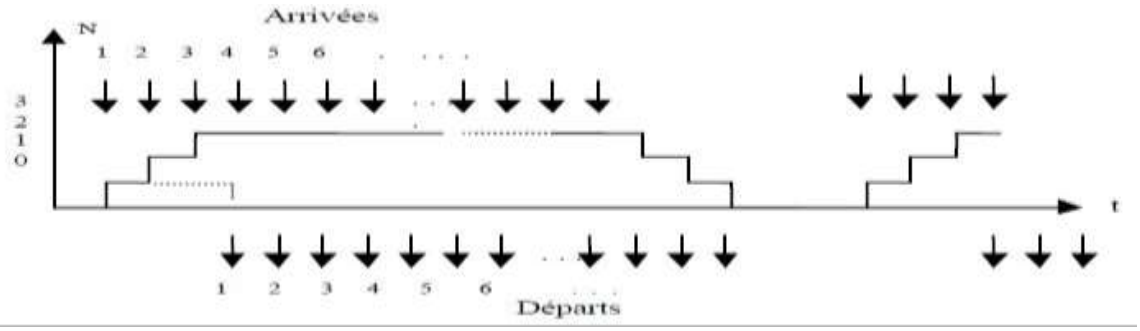
D'autres aspects liés à des notions comme des notions de durée différentes des phases plaident aussi en faveur de cette distinction. Ainsi, la durée d'établissement d'un appel, ou d'ouverture d'une session sera bien plus courte que la durée de l'appel ou de la session elle-même. Enfin, c'est fortement en relation avec ces deux phases que l'on parlera de trafic dans le *plan de commande* (*établissement et rupture*) et de trafic dans le *plan usager* (*information « utile » de niveau utilisateur*). Même s'il faut apporter des nuances dans l'interprétation détaillée de ces concepts en fonction du type de réseau et de la technologie employée, leur usage reste indispensable par l'aspect générique qu'ils présentent.

NOTION DU TRAFIC RESEAU

Le trafic réseau correspond au volume d'informations transportées ou traitées par réseau donné (*Réseau de télécommunication, informatique...).* Il pourra s'agir de données relatives aux échanges d'informations entre usagers (*voix, images, e-mails, fichiers...),* mais aussi des données relatives aux échanges d'informations entre *machines de commande* du réseau[7] (*données de signalisation dans un réseau de circuits, informations de routage dans un réseau IP, données d'exploitation...).*

Il est clair que plus les échanges entre les usagers ou les machines ne sont fréquents et de longues durées, plus les ressources nécessaires à l'écoulement de ce trafic seront importantes. Par exemple, si un réseau reçoit sur une période donnée, une demande permanente d'une communication par seconde telle que chaque communication a une durée de trois secondes, alors le réseau verra en permanence $n = 3$ communications de coexister. En effet, après un régime transitoire, dit « *montée en charge* », chaque fin de communication (*correspondant au processus de départ*) sera remplacée par un nouveau début de communication (correspondant au processus d'arrivée),

maintenant ainsi le niveau de charge du réseau pendant la période considérée. La figure ci-dessous décrit le phénomène.



Modélisation du trafic d'un réseau de transmission.

Pour simplifier, nous avons représenté des arrivées régulières et des durées de communications constantes, mais le phénomène reste bien sûr le même avec des arrivées et des durées de service variables autour de valeurs moyennes. Le nombre moyen N de communications en cours simultanément est appelé « *l'intensité de trafic* ». Et l'unité de mesure est l'*Erlang*, notée *E*, du nom du célèbre ingénieur danois A.K. Erlang[15] (1878-1929) qui établit les premières lois fondamentales de la théorie du trafic. Le concept « *trafic* » est fondamental en transmission car il définit la base du dimensionnement du réseau. Ainsi, si une ressource (circuit radio ou numérique, ou circuit virtuel, débit, etc.) est associée à chacune des *N* communications, il faudra pour écouler ce trafic un réseau d'une capacité d'au moins *N* ressources. Le nombre exact de ressources à provisionner dépendra de la loi d'arrivée et de la loi de service. Et c'est justement ce que permet de calculer la fameuse loi d'Erlang dans le cas d'arrivées dites « *poissoniennes* », c'est-à-dire suivant une loi de Poisson. De manière formelle, nous appelons « *A* » le trafic en Erlang et, si nous désignons par *n(t)* le nombre de ressources occupées, nous avons pour une période d'observation T :

$$A = \frac{1}{T}\int_0^T n(t)dt$$

Plus concrètement, si nous supposons un nombre de ressources suffisant pour écouler toutes les demandes présentées, et que nous appelons λ le nombre moyen, constant, de demandes par unité de temps, et *tm* la durée moyenne d'occupation de la ressource par chaque demande, nous avons :

$$A = \lambda t_m$$

A.K. Erlang a démontré le résultat fondamental suivant, dit *formule de perte d'Erlang*, qui donne la probabilité de rejet (*B*) d.une nouvelle demande, du fait de manque de ressources, pour un trafic *A* offert à *N* ressources :

$$E(N,A) = B = \frac{\frac{A^N}{N!}}{\sum_{j=0}^{N}\frac{A^j}{j!}}$$

Le trafic écoulé est alors :

$$A_e = A(1-B)$$

Cette formule exprime donc aussi la capacité du système considéré à écouler le trafic qui lui est offert. La réalité d'un réseau est bien plus complexe que ce modèle de base, nous aurons notamment à traiter aussi des phénomènes d'attente, de gigue, etc. Mais il s'agira toujours d'évaluer les ressources nécessaires pour écouler un trafic offert dans des conditions acceptables (perte, délai, etc.). Un réseau de téléinformatique est par quintessence soumis à des conditions de trafic variables dans son transport. Non seulement le trafic varie aux différentes heures de la journée mais il varie également en fonction de dates ou d'événements particuliers, et de la nature du trafic considéré. On distinguera cependant trois grandes conditions d'environnement en trafic:

- les conditions *normales* qui correspondent aux conditions de trafic de tous les jours ;
- mais avec aussi des jours *plus chargés* (jours de fin d'année par exemple) et ;
- les conditions *exceptionnelles* qui correspondent à des événements totalement imprévisibles (*catastrophes par exemple*).

[20] C'est afin de tenir compte de ces différentes situations et pour y associer des exigences de performance différentes qu'ont été définies les notions de *charges* dites *normale* (ou charge A), *haute* (ou charge B) et de *surcharge* (ou exceptionnelle). Il faut noter que ces principes s'appliquent à tous les types de trafic, que ce soit du trafic de signalisation, ou du trafic de niveau usager et quelle que soit la technologie utilisée. Les états de charge pouvant en outre être différents à un même moment pour les différents types de trafic. Par ailleurs, les situations peuvent différer, bien évidemment, d'un pays à l'autre et d'une partie du réseau à une autre. C'est l'observation du trafic dans les réseaux qui permettra de définir les niveaux de charge pour les différents éléments de ces réseaux.

Enfin, les caractéristiques de trafic évoluant en permanence, soit du fait de l'évolution du nombre d'usagers, de leur activité, soit à cause de l'évolution des services et des technologies (*par exemple, abonné mobile versus abonné fixe, technologie IP versus technologie circuit*), il sera nécessaire de réévaluer régulièrement les volumes et la nature des trafics offerts. Comme évoqué précédemment, ces volumes seront mesurés durant des périodes pendant lesquelles le trafic est dit stationnaire [3]. Cela signifie concrètement que, sur de telles périodes, il sera possible de caractériser le processus réel d'arrivée du trafic par un modèle stationnaire donné avec sa moyenne, sa variance, etc.… Notons aussi, que ce modèle reste particulièrement bien adapté à tous les types de trafic. En effet, d'une part, il traduit de manière évidente le comportement naturel des usagers qui expriment leurs demandes d'appels, de sessions, de manière aléatoire et indépendante, et d'autre part, il induit des propriétés importantes pour la caractérisation des trafics subséquents aux niveaux flots et paquets.

CHARGE NORMALE (CHARGE A)

Ce niveau de charge, dite *normale*, correspond aux conditions les plus fréquentes d'occupation du réseau, pour lesquelles le niveau normal de qualité de service attendu par l'usager doit être atteint. Dans ses recommandations relatives à l'exploitation du réseau. L'UIT préconise d'effectuer des mesures sur des durées d'un mois, pour avoir un échantillon significatif et pour tenir compte des variations saisonnières. Ayant classé l'ensemble des mesures journalières obtenues en fonction de l'intensité de trafic offert, nous choisirons la valeur de la quatrième période (*l'heure par exemple*) de mesure la plus élevée.

Les jours particuliers tels que les jours de fêtes de fin d'année sont exclus. Puis l'on retiendra parmi ces valeurs mensuelles, la valeur la plus élevée des douze mois de l'année *(éventuellement la seconde si la dispersion n'est pas trop grande d'un mois à l'autr*e). Ce sera la charge normale de référence pour le dimensionnement. En fait, ce qui est important, ce n'est pas tant la valeur absolue que l'esprit de la norme qui consiste à différencier une situation normale d'une situation plus rare. C'est en réalité un problème d'optimisation du nombre d'équipements, nous ne souhaitons pas dimensionner toujours pour le pire cas.

Ainsi, dans cette optique, des exploitants peuvent admettre de ne pouvoir écouler que *la charge A* en situation dite dégradée (*matériel partiellement et momentanément en panne*). Cet entendement est repris dans les recommandations de l'UIT (voir Rec. Q.543) ou dans les cahiers des charges des opérateurs qui traitent des équipements, avec la notion de *charge A* ainsi définie : *la charge A* représente la limite supérieure de la charge moyenne normale de travail que les fournisseurs de réseau souhaitent assurer pour leurs usagers, la charge B représentant un niveau plus élevé que le niveau d'activité normale prévue. Il s'agit bien de situations relatives. Ceci nous amène à préciser maintenant les notions de charge élevée et de charge B [26].

CHARGE ELEVEE, CHARGE B

Ce niveau de charge correspond à des situations peu fréquentes d'occupation du réseau, pour lesquelles le niveau normal de qualité de service attendu par l'usager ne sera pas nécessairement atteint, mais cependant suffisamment élevé pour éviter une perception très négative par l'usager. Dans la recommandation E.500, il est préconisé de retenir, toujours sur des durées d'observation d'un mois, la deuxième période ayant la mesure la plus

élevée. Puis, comme pour la détermination de la charge normale, on retiendra parmi les valeurs de charge élevée mensuelles la valeur la plus élevée des douze mois de l'année (*éventuellement la seconde, si la dispersion n.est pas trop grande d'un mois à l'autre*), pour obtenir la charge élevée de référence.

La recommandation Q 543 stipule simplement que le niveau de charge B représente un niveau plus élevé que les niveaux d'activité normaux prévus. Nous retiendrons surtout les ordres de grandeur proposés : la charge de référence B en intensité d'appels correspond à environ 1,2 fois la valeur correspondant à la charge de référence A en arrivée et 1,4 fois en départ, donc de l'ordre de 1,3 en moyenne ; la charge de référence B en Erlang correspond à environ 1,25 fois la valeur de la charge A de référence.

Ces valeurs sont données ici à titre indicatif car le développement de nouveaux services peut les modifier de manière significative. Cependant, elles mettent clairement en évidence la nécessité de distinguer les différents flux de trafic, départ et arrivée. En effet, en période très chargée, il y aura un nombre non négligeable de tentatives d'appels qui échoueront et ne créeront que peu de trafic de niveau usager dans le réseau. De même, dans notre travail d'évaluation des performances, il nous faudra tenir compte de conditions de trafic différentes dans les différents domaines : signalisation, commande, transport, etc. En pratique, ceci nous contraindra à déterminer les conditions les plus sévères pour chaque domaine.

A cet égard, il est intéressant de considérer les ordres de grandeur donnés dans Q.543 pour les niveaux de qualité de service attendus respectivement en charge A et en charge B : des probabilités de rejet d'appels de l'ordre de 10.3 en charge A passent à environ 10.2 en charge B ; des délais d'établissement à l'accès de l'ordre de 600 ms en charge A passent à environ 800 ms en charge B. Ces valeurs sont données ici à titre indicatif mais il est très vraisemblable que les ordres de grandeur resteront les mêmes quels que soient les services et les technologies car ils correspondent à une perception fondamentale de l'usager. Notons surtout le fait que les valeurs de charge B ne sont pas si « *relâchées* » que cela par rapport à celles de charge A. Le lecteur pourra en effet ultérieurement vérifier par application de la théorie des files d'attente sur un simple serveur (*M/M/1 par exemple*) que le rapport de 1,5 des délais *T* correspond à un rapport des charges ρ du serveur de la forme :

$$\frac{T_B}{T_A} = \frac{1-\rho_A}{1-\rho_B}$$

Ainsi pour = 0,7 *A* ρ, nous trouvons alors = 0,8 *B* ρ soit une charge B correspondant à environ 1,15 fois la charge A. Il est clair dans ces conditions que les objectifs de performance définis pour la charge B seront les plus contraignants pour la conception et le dimensionnement des systèmes, assez logiquement d'ailleurs. Ceci est très généralement vérifié, mais il faudra cependant à chaque fois vérifier le respect de toutes les conditions car les modèles des systèmes obéissent rarement à des types de services aussi simples que celui de la file M/M/1. Enfin, il est nécessaire de rappeler que la charge élevée est une notion à considérer avec précaution à un niveau global réseau, car les périodes de charge élevée ne sont pas forcément les mêmes sur les différentes portions du réseau. Ainsi, certains groupes de normalisation n'ont-ils spécifiés des objectifs globaux que pour le seul niveau de charge normale. Par contre, les deux niveaux de charges sont généralement pris en compte au niveau local.

SURCHARGE

Nous abordons ici des situations que nous pouvons qualifié d'exceptionnelles. Fondamentalement, elles se caractérisent par un niveau de trafic offert largement supérieur à la capacité installée des équipements du réseau. On pourra cependant y distinguer deux principaux types : d'une part, les événements prévisibles tels que par exemple toutes les fêtes et tous les événements programmés[19] (*fêtes religieuses ou sociales, jeux télévisés, etc.*), et d'autre part, les événements totalement imprévus tels que accidents ou catastrophes, etc. Dans les premiers cas l'augmentation du trafic offert pourrait à la rigueur être prévisible (*grâce à l'expérience et à l'observation de situations passées*). Mais les conséquences sur le dimensionnement et les investissements risquant d'être très coûteuses, nous préférerons se contenter de mesures limitées et d'écouler le trafic avec un certain délai.

Notons à ce sujet que du fait des phénomènes de renouvellement (l'usager dont les demandes échouent à l'accès ou dans le réseau renouvelle ses demandes un grand nombre de fois), nous pourrons ainsi constater des

augmentations très importantes du trafic offert en tentatives d'appels, alors qu'un léger surdimensionnement aurait peut-être pu éviter cet effet d'avalanche. Il sera donc toujours très important d'essayer d'évaluer le trafic dit « *frais* » (*hors renouvellement*), dans ces circonstances particulières. Dans le cas d'événements totalement imprévisibles (catastrophes, etc.), il est impossible de prévoir un quelconque surdimensionnement. Non seulement le volume de trafic ne peut être prévu, mais aussi la partie de réseau concernée est totalement imprévue, par opposition par exemple à un jeu télévisé.

A ceci se rajoute encore le phénomène d'impatience et de renouvellement qui va augmenter de manière considérable le trafic offert aux directions et destinations réseau concernées. Dans ces deux cas exceptionnels se pose le problème des performances du réseau et de ses équipements. La question de base est tout d'abord de savoir s'il y a un problème dans le réseau si des appels en trop échouent ? Il y a malheureusement un problème parce que, en cas d'encombrement de certaines ressources, ce sont tous les appels qui peuvent échouer. Prenons par exemple le cas simple d'appels qui utilisent des ressources réseaux (*circuits, canaux de signalisation, bande passante de liens IP...*) et qui échouent sur encombrement à l'arrivée (*cas d'une catastrophe par exemple où tout le monde veut appeler une destination donnée*).

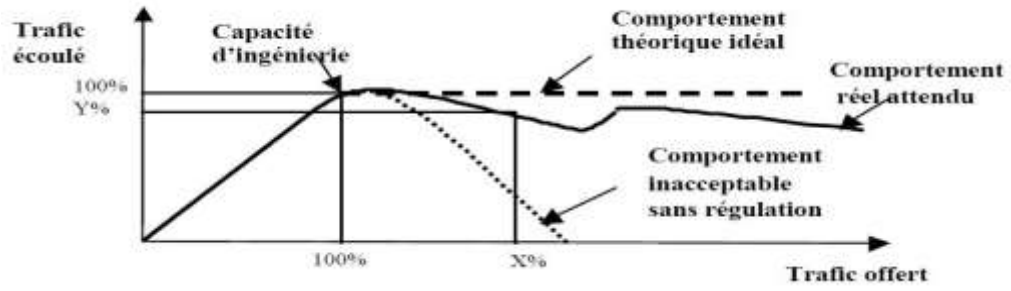
Il est évident que les ressources intermédiaires ont été utilisées inutilement, empêchant ainsi tout autre trafic de passer. Et ce peut être pire encore dans un monde IP s'il n'y avait pas de contrôle d'acceptation d'appel : les autres communications en cours seront toutes perturbées par la congestion au niveau paquets. De manière semblable, et c'est un problème fondamental, un processeur de traitement (*appel, paquet, etc.)* qui passe son temps à effectuer des traitements pour des appels qui échouent ou qu'il doit rejeter, finit par ne plus avoir de temps libre pour traiter un seul appel correctement (*imaginons un médecin qui ausculterait tout en répondant au téléphone pour refuser d'autres clients*). Il faut là aussi des règles d'acceptation. C'est le propos du *contrôle des surcharges* ou *régulation*.

Au premier abord, nous serons tentés de dire qu'il suffit de refuser les demandes à partir d'un certain seuil. Ce sera en effet la stratégie de base, mais encore faut-il que ce refus ou *rejet*, se fasse au moindre coût, et outre le problème se complique avec la nécessité de traiter en priorité certains types d'appels tels que les appels vers les numéros urgents (*ambulances, police, etc.)* [10]. Ces besoins ont été spécifiés d'une part, dans des recommandations internationales, mais aussi d'autre part, et surtout dans des spécifications propres aux différents opérateurs de réseaux. Essentiellement les exigences en cas de surcharge exceptionnelle seront :

— le système, ou le réseau ne doit pas s'écrouler. Ceci paraît trivial mais dans ces circonstances exceptionnelles les mécanismes de défense sont mis à rude épreuve et des erreurs résiduelles normalement inoffensives deviennent vite mortelles. A cet effet, nous développerons dans la présente recherche, l'aspect test de ces mécanismes ;

— le système ou réseau devrait continuer à écouler une quantité de trafic proche de la capacité pour laquelle il a été installé et équipé. Cette exigence s'atteint en rejetant les appels en excès, à condition bien sûr de détecter rapidement la surcharge, mais est à mettre en balance avec l'exigence suivante ;

— les appels acceptés doivent l'être en donnant la priorité à certains types d'appels tels que les appels arrivée (*car ils doivent aboutir et ont consommé déjà des ressources dans le réseau*), les numéros urgents (*ambulances, police...),* les lignes prioritaires, etc. Le problème est alors que la reconnaissance de l'appartenance ou non d'une nouvelle demande à l'une de ces catégories nécessite du traitement, pour que dans la majorité de cas aboutisse au rejet. Certes une indication de priorité (*ligne, circuit, appel, paquet prioritaire...*) est très vite analysée mais l'analyse d'une numérotation est plus coûteuse. Même un coût très faible peut conduire à l'inefficacité totale du système car n'oublions pas le phénomène d'avalanche qui conduit couramment à des surcharges de 1 000 %. Nous retiendrons ici qu'il est certes nécessaire de reconnaître les demandes prioritaires mais que, si la surcharge persiste, il faudra à un certain moment rejeter indistinctement tous les types de demandes (*ce sera une question de survie du système ou du réseau*).



Ecoulement du trafic en surcharge[1]

La figure ci-dessus résume assez bien le comportement possible du système ou du réseau. Il représente le trafic écoulé en fonction du trafic offert en tenant compte des différents niveaux de rejets dans la régulation. Nous développerons ultérieurement ce modèle.

RESULTATS ET DISCUSSION

EVALUATION DES PERFORMANCES DES SYSTEMES ET RESEAUX INFORMATIQUES

Dans cette partie, nous déterminons un assortiment de modèles d'évaluation des performances de systèmes et réseaux, créés à partir des résultats et outils théoriques présentés dans les parties précédentes. Notre apport scientifique est de donner un réflexion aussi incontestable que possible des problèmes très variés agrégés au travail de modélisation dans un environnement industriel (*constructeurs, exploitants, laboratoires, etc.).* Le choix de ces modèles est bien sûr arbitraire et ne prétend pas couvrir l'ensemble du sujet abordé dans la présente recherche. Il est cependant basé sur deux grands critères : l'utilité dans notre domaine d'étude (*le domaine des télécommunications*) et l'intérêt didactique du point de vue de la modélisation et de l'application pratique de la théorie.

Par la suite, pour la clarté de la recherche, nous allons différencier deux grandes catégories de modèles : ceux relatifs à la commande des systèmes et ceux relatifs au domaine du transport de l'information. Là aussi la distinction est arbitraire. Elle est avant tout basée sur l'usage le plus fréquent dans notre domaine. On notera que certains équipements tels que le lien Ethernet peuvent être considérés comme appartenant aux deux domaines : interconnexion de processeurs de commande ou réseau de transport de données de niveau usager, etc.

MODELES POUR LA COMMANDE DES SYSTEMES [7]

— *Le système simple bouclé :* Il s'agit d'un système tel que chaque demande, à sa fin de traitement, a une probabilité *p* d'être représentée au système. Ce modèle correspond bien, par exemple, à un système de transmission présentant un taux d'échec par message *p*, entraînant sa retransmission. Il peut correspondre aussi à un système de traitement ayant à exécuter des macrotâches qui vont ne pouvoir s'exécuter qu'en plusieurs tâches élémentaires. Ce peut être, par exemple, le cas de l'exécution d'un traitement dont le code n'est pas présent totalement en mémoire locale : on doit interrompre le traitement avec une probabilité *p* pour aller chercher la suite en mémoire centrale (problème de la mémoire virtuelle et des fautes de pages par exemple).

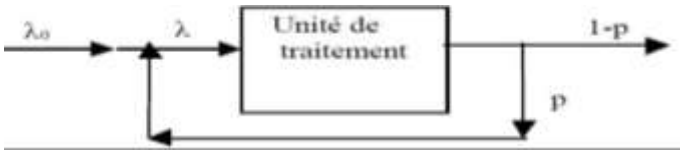


Schéma du système simple bouclé.

— *Le Centre serveur :* Il s'agit de modéliser un système tel un serveur (une unité centrale, ou un centre de réparation, d'où aussi le nom de *repairman model*), traitant les requêtes d'un nombre fini de clients identiques (terminaux, équipements en parc...). Chaque client soumet une nouvelle requête toutes les $1/\lambda$ unités de temps

en moyenne, intervalle appelé *temps de réflexion, temps de fonctionnement*... Le processeur du centre serveur traite une requête en un temps moyen $1/\mu$, appelé *temps de service*. Ce modèle est très populaire car il correspond à de très nombreux cas de réseau. Par exemple, il correspond particulièrement bien au cas d'un serveur commun à plusieurs terminaux dans un réseau local, ou à celui d'un serveur IN (*Intelligent Network*) commun à plusieurs commutateurs dans le réseau téléphonique ou dans le NGN (Next Generation Network).

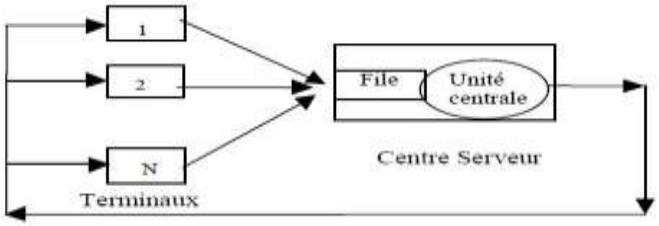


Schéma du centre serveur

— *Le Processeur à temps partagé :* Il s'agit ici de modéliser un système dit *à temps partager*, car dans un tel système le processeur de traitement partage son temps de manière égale entre toutes les tâches (clients) en cours. Cela peut être dans plusieurs contextes d'appels simultanément en cours de traitement par exemple. Il n'y a pas de file d'attente à proprement parler. Idéalement, on admet que le processeur passe un quantum de temps infiniment petit sur chaque tâche. Il s'ensuit que le temps de traitement d'une tâche est d'autant plus long qu'il y a d'autant plus de tâches en cours simultanément, mais aussi que ce même temps de traitement est d'autant plus court que la tâche est intrinsèquement courte. Un tel système favorise les tâches de courte durée. Les résultats sont relatifs aux caractéristiques (attente, temps de séjour) du temps de traitement d'une tâche donnée (de durée donnée) [6]. Nous nous intéressons ici à la détermination des caractéristiques « moyennes » du temps de traitement de l'ensemble des tâches. C'est en effet de ces paramètres dont nous aurons besoin pour l'évaluation de la qualité de service, par exemple dans les processeurs de traitement d'appels, le traitement d'un appel étant constitué de plusieurs tâches de durées différentes. Des caractéristiques essentielles sont bien sûr la moyenne et la variance du temps de traitement total d'une tâche (son temps de séjour dans le système, c'est-à-dire sa durée de traitement intrinsèque, plus son élongation ou *attente*).

— *Le polling et Token Ring :* Considérons d'abord le cas général d'un service cyclique, ou *Polling*. Ce type de service est utilisé par de nombreux systèmes, aussi bien en tant que protocole de communication entre stations qu'en tant que mode de services pour gérer le temps réel entre processus, ou entre périphériques de réception et d'émission de messages. Ici, nous cherchons à déterminer le temps d'attente et de traitement pour un paquet (ou message, ou processus). En supposant des arrivées poissonniennes, nous voyons que ce temps sera composé du temps d'attente de scrutation (pour le premier message) plus le temps d'attente devant un serveur M/G/1. Evaluons la valeur moyenne du temps de cycle de l'ensemble des stations. Ce temps est composé à chaque entité pollée d'un temps intrinsèque de scrutation à chaque station τi (libre ou occupée, ou passage de station à station), plus un temps de service *si* s'il y a un service à effectuer.

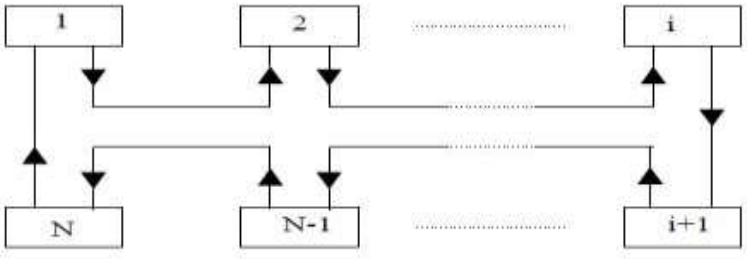


Schéma du système polling et token ring

— *Le lien Ethernet :* Les LAN de type Ethernet basés sur des protocoles de communication de type CSMA/CD sont des éléments caractéristiques de support de communication entre machines (processeurs ou stations d'un

système, serveurs, nœuds d'un réseau). La modélisation de leurs performances est donc indispensable, et surtout la mise en évidence des paramètres de performance les plus significatifs. Ainsi que nous allons le voir, les performances de ce type de medium de communication peuvent grandement différer selon les applications. Ces protocoles sont issus des problèmes d'accès à un canal radio par des stations indépendantes. Chacune ignorant si l'autre se met à émettre, il peut y avoir collision. De manière semblable, avec un LAN Ethernet des stations indépendantes sont connectées en étoile à un *hub*, ou à un *switch*, Ethernet, et plusieurs peuvent décider d'émettre en même temps, si la liaison leur paraît libre, il y a alors collision.

Pour résoudre ces conflits, les protocoles CSMA/CD (CSMA, *Carrier Sense Multiple Access*, CD, *Collision Detection*) ont été définis. On distingue trois types : le type CSMA/CD *persistant*, le type CSMA/CD *non persistant*, le CSMA/CD *p-persistant*. *Le type CSMA CD persistant.* Un terminal prêt à émettre teste si le LAN est libre, c'est-à-dire s'il n'y a pas de transmission en cours. Si le LAN est libre, il transmet le paquet avec une probabilité 1. Si le LAN est occupé il attend jusqu'à ce que le LAN soit libre et alors retransmet le paquet avec probabilité 1 (d'où le nom *persistance 1*). *Le type CSMA/CD non persistant*. Dans le cas du *non persistant* le terminal tente la rémission au bout d'un délai aléatoire déterminé selon une certaine loi de distribution (exponentielle par exemple). Dans les deux cas, le temps peut être « découpé » en périodes élémentaires de durée *T*s (reliée au temps de propagation). Tous les terminaux sont synchronisés et testent le canal en début de période. Dans le cas du persistant, on voit donc que lorsqu'on a deux terminaux ou plus qui sont prêts à émettre pendant une même période *T*s, ils émettront tous et rentreront en conflit avec une probabilité de 1. D'où l'idée de transmettre au bout d'un temps aléatoire pour minimiser les probabilités de conflit. *Le type CSMA/CD p-persistant*. C'est une évolution du non persistant ; Après une collision, l'émetteur réémet immédiatement avec une probabilité *p*, et sinon (avec la probabilité 1-*p*) attend une durée *T*s, puis recommence la même procédure.

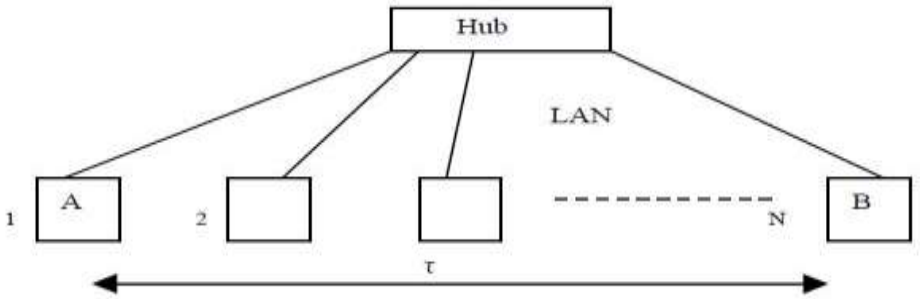


Schéma du lien Ethernet

Soit m la durée d'occupation du bus par un message ; N le nombre de stations connectées au bus ; τ la durée de test de la liaison (égale de la durée de propagation entre les stations A et B, les plus éloignées physiquement l'une de l'autre) ; Ts : le temps est divisé en intervalles de durée Ts, égale à 2 fois le temps de propagation τ. La spécification physique de IEEE 802.2 donne Ts = 512 temps bit pour 5 km, ce qui correspondrait à 51,2 µs à 10 Mbit/s (c'est le cas le pire, en pratique ce sera plus rapide car les LAN seront plus courts).

— *Le Benchmarks (modélisation processeur) : Il* s'agit ici d'évaluer la capacité de traitement d'un nouveau système, ou d'une nouvelle version de ce système par l'utilisation par exemple de processeurs ou plateformes plus rapides. S'il s'agit d'une évolution technologique simple, l'utilisation d'un *benchmark*, ou logiciel de test de performance, s'impose. Dès que possible, on fera exécuter un extrait significatif du code à exécuter sur la nouvelle machine afin de mesurer l'évolution des temps d'exécution. Dans des applications de télécommunication, il sera assez facile de définir la mesure des primitives les plus fréquentes de communication (*send*, *receive*) ainsi que la mesure de traitements typiques au niveau applicatif. Cependant, bien souvent, à l'instant du cycle de vie où nous nous plaçons, il n'est pas encore vraiment possible de faire exécuter du code significatif sur les machines, et les évolutions sont souvent complexes (processeur, mais aussi mémoire, bus, etc.). Des informations sur l'évolution possible des performances peuvent être obtenues à partir de résultats de *benchmarks* publiés dans la littérature, ou sur le Web, par les utilisateurs ou par les constructeurs. On trouve ainsi les résultats connus sous le terme *SPEC Integer*, issus de l'organisation *System*

Performance Evaluation Corporation, organisation de fabricants de l'industrie informatique, qui développe des *benchmarks* et publie les résultats périodiquement.

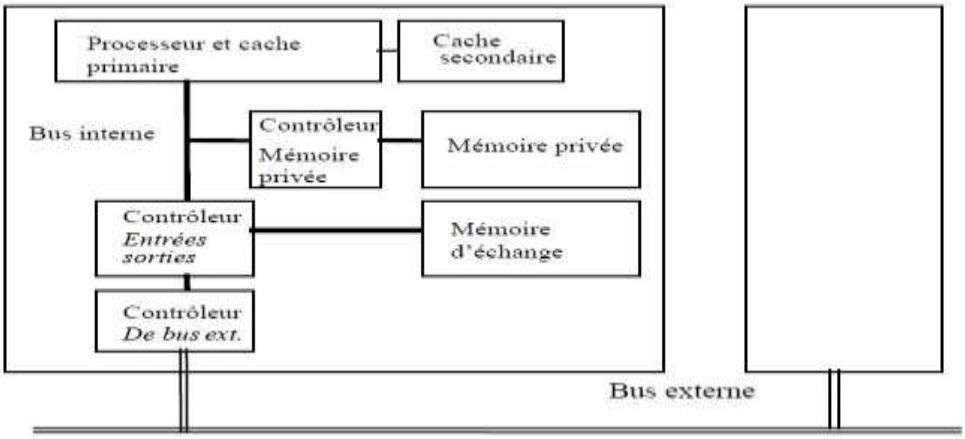


Schéma général d'une carte processeur.

— *Le système à disque :* Le disque constitue un élément essentiel des systèmes de télécommunication, du fait de son énorme capacité de stockage des informations : plusieurs dizaines de Giga-octets (9, 18, 36, 72 Go ou GBytes), par exemple. Les utilisations en sont multiples, dans les serveurs (sur le Web, dans les réseaux fixes et mobiles), comme dans les systèmes d'exploitation des réseaux de télécommunication (récolte des observations, stockage de la taxation). L'évaluation de son temps de service est donc un élément important de l'évaluation des performances des systèmes qui l'utilisent (temps de réponse d'un serveur par exemple). Et ceci, d'autant plus qu'ici interviennent des éléments mécaniques qui conduisent à des durées d'exécution d'un tout autre ordre de grandeur que celles des processeurs [9].

MODELES POUR LE TRANSPORT

— *Le concentrateur de trafic multidébit :* cela consiste à évaluer la capacité d'écoulement de trafic d'un système concentrant le trafic de sources de débits différents. Ce cas est typiquement celui de concentrateurs de lignes d'usagers ISDN, ou de liens réseaux synchrones, devant écouler des communications de débits différents (à n × 64 kbit/s par exemple), en appliquant la *méthode du peakedness factor, qui consiste* simplement d'ajuster la loi d'Erlang généralisée à une loi d'Erlang simple. On définit ainsi un débit unique équivalent, mais aussi un nombre de serveurs équivalents [11].

— *Le multiplexage* **:** permet de mettre en évidence des propriétés utiles pour simplifier le traitement du cas de sources de débits différents et de longueurs de paquets différentes.

— *La bande passante équivalente :* Le concept de bande passante équivalente est utilisé pour simplifier les mécanismes d'acceptation d'appel, de session, etc., dans les systèmes supportant des services de caractéristiques de débit différentes, ou variables. Il s'agit en effet de calculer la performance du support de communication avec le nouvel appel qui se présente, compte tenu des caractéristiques des appels déjà acceptés, et de la comparer avec le critère de qualité de service requis. On conçoit aisément que, si l'on pouvait se ramener au problème simple de la pure somme de débits identiques, comme dans le cas de communications à 64 kbit/s, l'acceptation d'appel se résumerait au test du respect de la charge maximale du lien, ou du nombre maximum d'appels en cours sur un lien.

— *La modélisation et multiplexage du trafic IP :* Dans ce modèle, nous allons considérer que le trafic IP a les caractéristiques fondamentales suivantes :

- le trafic est issu d'un grand nombre d'utilisateurs et le processus des demandes individuelles au niveau session, comme au niveau appel en téléphonie, forme un processus de Poisson. De même, la demande globale, à ces niveaux, est prévisible, résultant de l'activité journalière des utilisateurs;
- le trafic IP peut être décomposé en trois grands niveaux : *le niveau session* (les sessions arrivent selon un processus de Poisson. Soit un lien de capacité *C*, si λ est le taux d'arrivée et *v* le volume moyen par session (en bits par exemple, produit du débit moyen et de la durée), alors on note la charge du lien $\rho = \lambda v/C$) ; *le niveau flot* (chaque session est constituée d'une succession de flots et de périodes de silence. Les flots correspondent au transfert de fichiers, d'e-mails, d'images, etc. Le trafic au niveau flot est de nature sporadique, le volume des flots est, lui aussi, extrêmement variable) ; *le niveau paquet* (le trafic au niveau paquet présente une très grande sporadicité et des caractéristiques dite *d'autosimilarité,* en particulier du fait de l'interaction du trafic d'origine avec les mécanismes de contrôle de flux (TCP) et de correction d'erreurs) ;
- le trafic IP peut être structuré en deux grandes catégories de flots : *les flots temps réel* (il s'agit de la transmission en temps réel de données du type voix, vidéo, généralement sous contrôle du protocole UDP[14] (pas de contrôle de flux, pas de retransmission). La performance est caractérisée surtout par la minimisation du délai de transfert (impact sur la perception de l'interactivité par l'usager, nécessité éventuelle d'annuleurs d'échos), par la minimisation de la gigue (*jitter*) et le respect d'un débit intrinsèque) ; *les flots élastiques* (il s'agit du transfert de fichiers, d'e-mails, de pages Web, etc. généralement sous contrôle du protocole TCP (contrôle de flux, retransmission). La contrainte temps réel est moins sévère, par exemple des délais de transfert de l'ordre de la seconde restent acceptables. La performance est surtout caractérisée par la durée totale de transfert, ou encore le débit moyen effectif qui correspond au rapport du volume sur la durée). Le modèle de multiplexage du trafic IP va être basé sur ces caractéristiques, comme suit : D'une part, on distingue trafics temps réel et trafics élastiques, et surtout priorité (non préemptive) est donnée au trafic temps réel. Il s'en suit que l'évaluation de la bande passante requise peut être effectuée dans un premier temps indépendamment pour chaque catégorie de trafic, puis les gains possibles par intégration sont étudiés. D'autre part, du fait des propriétés des niveaux sessions et flots, il s'avère possible de surmonter les problèmes complexes d'autosimilarité du niveau paquet.

— *Le réseau de connexion :* Le réseau de connexion est un équipement qui permet d'établir un chemin physique de transport des informations entre une entrée et une sortie données pour la durée d'un appel, ou d'une session, ou de tout élément de communication dont la durée ou l'intégrité justifient l'établissement d'un tel chemin. Il peut s'agir bien sûr d'établissement de chemins (circuits) virtuels entre circuits d'entrée et de sortie virtuels, supportés par des liens physiques. Un réseau de connexion comporte par principe plusieurs sous-ensembles, généralement appelés *matrices*, disposés en *étages* et interconnectés entre eux. Le but de cet assemblage est de construire des réseaux de très grosse capacité à partir d'éléments de capacité limitée. La figure ci-dessous donne le principe d'un réseau à 3 étages. Les étages d'extrémité comportent *m* matrices à *n* entrées et *k* sorties, l'étage central comporte *k* matrices à *m* entrées et *m* sorties. Si la capacité de chaque lien d'entrée ou de sortie est *c* (par ex *c* canaux à 64 kbit/s ou *c* Mbit/s) alors le réseau est dit de capacité *nxmxc*. Mais cette capacité dépend bien sûr de la capacité d'interconnexion des matrices d'extrémité par l'étage central, donc de la valeur de *k*, et de la capacité des liens internes.

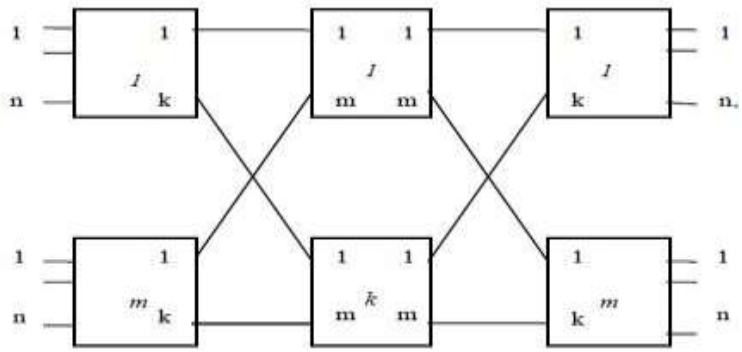


Schéma du réseau de connexion à trois étages.

De manière très générale, un réseau sera caractérisé du point de vue du trafic par trois grands paramètres :

— *son blocage*, c'est-à-dire la probabilité de ne pas trouver de chemin libre entre l'entrée et la sortie considérées (blocage point à point). Il y aura blocage par exemple si on ne trouve pas un chemin physique disposant de la bande passante minimale requise par la communication ;

— *son délai de traversée*, c'est-à-dire les caractéristiques (moyenne, dispersion, quantiles...) du total des durées d'attente et de service subies par l'information élémentaire (le paquet par exemple) à chaque étage de commutation. Dans un réseau synchrone, comme les réseaux téléphoniques classiques, il n'y aura pas de délai d'attente excepté quelques délais fixes de resynchronisation. Par contre, dans un réseau asynchrone, ATM ou paquets, chaque étage peut comporter des files d'attente et donc entraîner des délais variables.

— la probabilité de perte d'information. Dans un réseau de connexion paquet ou ATM le débordement de file d'attente entraînera la perte de l'information.

MECANISMES DE CONTROLE DE LA SURCHARGE DANS LES RESEAUX INFORMATIQUES

Les mécanismes de contrôle auront comme tâche principale de mesurer et de limiter ce trafic, attendu que la surcharge provient d'une augmentation inconsidérée du trafic offert au réseau, Une autre contrainte qu'on leur impose est d'accomplir cette limitation de la manière la plus rentable, c'est-à-dire de permettre au réseau de fonctionner au plus près de sa capacité maximale. Enfin, la contrainte d'équité [13] est aussi arguée vu que si le réseau se trouve en situation de pénurie, celle-ci doit être partagée équitablement entre les sources. Si une source voit son trafic augmenter, il ne faut pas qu'elle puisse écraser les autres sources, même si, vu du réseau, l'écroulement typique de la congestion ne se manifeste pas. La difficulté du contrôle provient en majeure partie du caractère distribué des sources de trafic : trafic émanant de sources non coordonnées, absence d'organe de contrôle central qui régulerait en fonction de l'état global du réseau. Ainsi, dans la présente recherche, il sera question du contrôle des flux dans les systèmes centralisés et des réseaux « store and forward » en utilisant des cas des réseaux à haut à débit.

CONTROLE D'UN SYSTEME CENTRALISE

Ici, nous allons décrire très brièvement l'approche mise en œuvre pour la régulation de charge dans une machine de type centralisée c'est-à-dire dans le cas le plus favorable où l'organe de commande peut avoir une vue exhaustive et instantanée de l'état des ressources. Le système temps réel va s'analyser, un réseau de files d'attentes, avec probablement des mécanismes d'ordonnancement complexes. Quoiqu'il en soit, il existera, compte tenu du profil de trafic traité, une ressource plus chargée [19]. C'est cette ressource qui va saturer en premier, quand la charge va augmenter, et c'est donc elle qu'il faut surveiller. En réalité, les profils du trafic ne sont pas immuables, et par conséquent la ressource la plus chargée pourra varier, et la surveillance sera le plus souvent multiple.

Imaginons pour simplifier que la ressource fragile est unique, et qu'il s'agit d'un processeur central (c'est un cas assez fréquent). Le principe du contrôle est fort simple : mesure de la charge de la ressource, c'est-à-dire de

son taux d'occupation. Par exemple, on mesurera le temps passé par le processeur au traitement des tâches de priorité les plus faibles. La mesure se fait cycliquement ; on définit un intervalle de mesure T, par exemple 10 secondes; selon la valeur du temps total d'inactivité I dans le cycle, le processeur sera décrété plus ou moins chargé. La charge sera estimée par 1 - I=T. En fonction du niveau détecté, les actions de correction ou de prévention seront lancées[4]. Ainsi, on peut décrire deux mécanismes de contrôle :

— *Le processus de mesure* : Sous cette description simpliste se cache de redoutables pièges, qui vont conduire à complexifier la représentation. En premier lieu, la mesure est entachée d'incertitude parce que les processus d'arrivée des requêtes et des services sont aléatoires. Si nous observons 2, 3 ou 4 secondes à plusieurs reprises, à l'état stationnaire, nous ne retirerons pas la même information. Il faut tenir compte de l'imprécision inhérente au procédé de mesure.

— *Les actions de défense* : Quelle action entreprendre, une fois la surcharge détectée? Le plus souvent, la surveillance réagira progressivement, selon le franchissement de paliers dans la charge. Par exemple, un premier palier sera défini pour une charge de 80 %, un deuxième pour 85 %, etc. A chaque palier sera attachée une action d'efficacité (et de gravité) grandissante. Les systèmes temps-réel sont assez souvent dotés de mécanismes d'auto-surveillance (exploration et test cyclique des éléments actifs). Ce test consomme une partie de la ressource. La première action consistera à désactiver ces tests (dont le système peut évidemment se passer, pourvu qu'ils soient rétablis en régime moins chargé). Les actions suivantes amèneront inévitablement à refuser l'accès du système à de nouvelles requêtes. Le plus souvent, on pourra établir une hiérarchie dans les rejets. Ainsi, dans un commutateur du réseau téléphonique on restreint d'abord l'accès des appels au départ (*afin de favoriser les appels arrivants, qui ont déjà consommé une partie des ressources du réseau*). Une faiblesse de ce type d'action se trouvera dans les contraintes posées par l'organisation du système, par exemple la nécessité de détecter et d'analyser une demande avant de pouvoir statuer sur son sort: le traitement des requêtes conduit à un gaspillage inévitable de la ressource.

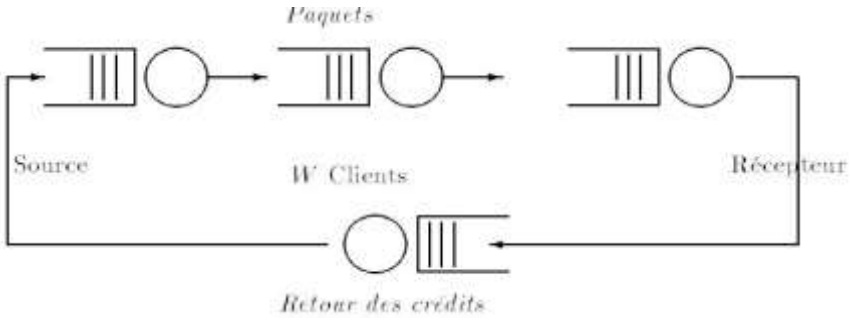
LES RESEAUX STORE-AND-FORWARD

Les réseaux de paquets actuels (Transpac, par exemple) utilisent un mode de travail dit « Store-and-Forward ». Cela signifie que l'information (structurée en « paquets ») est transmise de proche en proche, de nœud en nœud, depuis l'origine jusqu'à la destination. Les nœuds intermédiaires accomplissent les fonctions des couches de protocoles de niveau 1 à 3 : contrôle et corriger des erreurs et les acheminent par les mécanismes tels que :

— *Le contrôle de flux* : Le contrôle de flux vise à asservir le débit de la source à celui du récepteur. C'est une fonction de « bout-en-bout ». C'est, aussi, une fonction de contrôle de congestion. C'est-à-dire qu'une mauvaise régulation du flux émis va provoquer une congestion sur le chemin de la connexion : si la source émet plus vite que le récepteur ne reçoit, les paquets vont s'accumuler dans le réseau, d'abord dans le nœud de sortie, puis dans le précédent, etc.

— *Le mécanisme de Fenêtre* : Un mécanisme simple et universel est celui de la fenêtre. La source se voit allouer un certain nombre de crédits W. Elle peut émettre un paquet à condition d'avoir un crédit (et le paquet consomme le crédit correspondant). Le récepteur acquitte les paquets reçus (*et traités*). L'accusé de réception régénère les crédits de l'émetteur, l'autorisant ainsi à poursuivre. L'émetteur numérote les paquets qu'il envoie. A l'instant t, les paquets n - 2; n - 1; n sont acquittés. Il peut envoyer le numéro n+1, mais aussi anticiper en envoyant n+2; … n+W : on parle d'un mécanisme de fenêtre d'anticipation (*qu'on dit « fenêtre glissante », l'acquittement du n+ 1 faisant déplacer la fenêtre d'un cran*).

[4] *Ici, c'est la charge du processeur, qu'on va mesurer, et non le taux des arrivées. Pourquoi? Simplement, parce que dans un système à file d'attente, la charge ($p = \lambda/\mu$) est le principal paramètre. Et qu'en situation de surcharge, le temps de service n'est pas constant. La prudence commande donc d'observer p et non λ.*

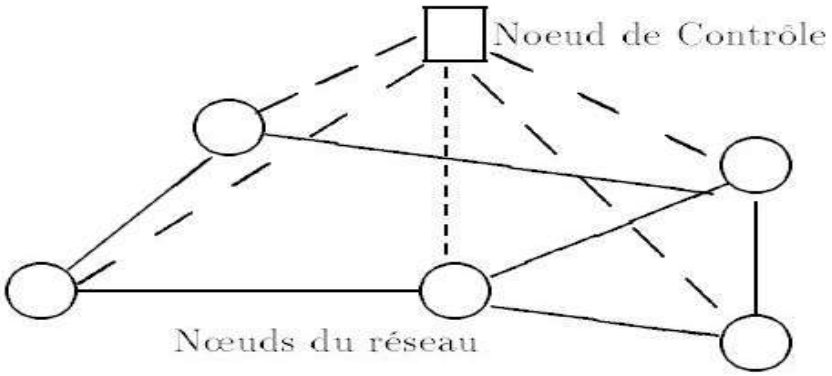


Modèle à files d'attente du contrôle par fenêtre

On peut le rendre plus réaliste, au prix d'une complication, en introduisant les flux incidents : dans chaque nœud traversé, le flux pisté doit affronter d'autres courants qui occupent les mêmes ressources. De même, chaque nœud peut être représenté par un modèle plus complexe.

— *Le contrôle de Congestion* : Le contrôle de flux s'exerce de bout en bout. Le contrôle de congestion vise à protéger le réseau, et pourra s'exercer de proche en proche. Il utilisera le cas échéant les mêmes outils. Par exemple, on protège un nœud en instaurant une régulation par fenêtre sur chaque flux incident. Si la somme de chaque allocation accordée aux flux concernés est égale à la capacité mémoire disponible, il n'y a pas de phénomène de congestion lié aux capacités limitées. Cette méthode est fiable : le nœud récepteur asservit parfaitement le débit de l'émetteur voisin à ses possibilités instantanées. C'est une heureuse conséquence du mécanisme « store-and-forward ». Malheureusement, cette méthode présente l'inconvénient majeur d'être très coûteuse en bande passante. Supposons que le nœud examiné se partage entre 100 circuits virtuels de débits identiques. Il divise sa capacité en 100 parties, et le partage entre les circuits. La variabilité des trafics issus de chaque source va rendre cette solution très défavorisant, la mémoire sera toujours vide, etc.

On a donc dû imaginer d'autres méthodes, plus avantageuses. La première idée consiste à gérer les crédits globalement pour tout le réseau (le réseau peut accommoder au total N paquets, on crée N crédits, dispersés sur tous les nœuds). Le contrôle *isarithmique* utilise ce principe. Voici une illustration simplifiée :



Modèle du réseau isarithmique [25]

On construit un réseau de files d'attentes représentant le mouvement des jetons. C'est un réseau fermé, si le nombre des jetons est fixe. Dans le nœud de contrôle, on peut mettre en place des mécanismes complexes tels que la génération ou destruction de jetons; accélération ou retard de l'envoi des jetons en fonction de l'état du réseau. Par exemple, on modélisera cette fonction par un taux de service dans le nœud de contrôle de la forme $\mu(n)$.

LE CAS DES RESEAUX HAUT DEBIT

L'étude du contrôle de la surcharge dans les réseaux haut débit prend une importance fondamentale. D'une part les conséquences des engorgements sont à la mesure des débits mis en jeu; en même temps, les mécanismes

possibles atteignent les limites de leur efficacité. D'autre part, la « surcharge » [27] devient un élément normal du fonctionnement de ces réseaux afin d'en assurer un régime optimal. C'est notamment le cas de l'Internet comme de l'ATM (mode ABR), ou du Frame Relay.

CONCLUSION

Depuis les années 1990, Internet a permis le développement massif de la connaissance et de la donnée. Outil indispensable pour les professionnels (recherche ou échange d'informations) et pour le grand public (jeux, outils de communication instantanée, achats en ligne...), il est utilisé par 3,025 milliards d'individus recensés soit 42 % de la population dans le monde. Face à la croissance des données échangées, l'optimisation des réseaux est devenue un enjeu majeur. La disponibilité des réseaux est un réel défi. L'information est parfois vitale : les données de type politique, éducatif et même culturel sont importantes car elle permet de rester en contact. Exclure certaines individualités de cette source d'information revient à les priver des avancées se produisant dans le monde. La disponibilité d'Internet n'est pas homogène à la surface du globe. 2/3 de la population en est privée. Cette différence de développement ne permet pas d'avoir une information précise et constante sur le temps et la météo notamment. Les récoltes dans certaines parties du monde pourraient être meilleures et les catastrophes naturelles prévenues. Pour couvrir l'ensemble du globe, des initiatives sont créées par différents acteurs pour offrir un accès Internet aux communautés qui en sont encore privées.

En parallèle, la problématique de la quantité d'information et des échanges se pose. En effet, la requête vers Internet ne passe pas directement d'un terminal à la page ou l'outil en question. Les différentes couches représentant l'ensemble des fonctionnalités nécessaires à la communication des systèmes informatiques sont très chargées et engendrent un temps d'attente relativement long. Avec la digitalisation, la croissance des données est exponentielle. L'enjeu est de taille pour les marques. Les outils mobiles transmettent des données relatives aux actions des utilisateurs : émission de signaux GPS, navigation Internet, de recherche, messages laissés sur les réseaux sociaux, téléchargement, utilisation d'applications, publication en ligne de photos et de vidéos, achats sur des sites de vente en ligne... L'augmentation permanente du nombre d'utilisateurs d'Internet et de téléphones mobiles engendre une croissance exponentielle du volume des données numériques. Les consommateurs s'engagent et partagent de plus en plus leurs données personnelles sur les réseaux sociaux. 46 % des propriétaires de Smartphones sont actifs sur plusieurs réseaux sociaux à la fois. En 2010, le trafic des données mobiles représentait le double de celui généré par les appels six ans plus tard : les données mobiles ont considérablement creusé l'écart puisqu'elles sont 6 fois plus importantes en termes d'utilisation. A l'échelle mondiale, 7, 3 milliards d'abonnements mobiles ont été souscrits soit plus que d'individus sur terre. Cela représente une augmentation de 3 % chaque année grâce à des forfaits mobiles comprenant plus de data.

Depuis le début des années 2000, le sujet brûlant du réchauffement climatique occupe l'inconscient collectif. L'empreinte carbone de l'informatique est de plus en plus grande. Certains centres de données comme celui de Facebook dépassent les 678 millions kWh en une seule année, soit 100 millions kWh de plus par rapport à l'année précédente. Les pays en développement sont tout aussi concernés. Bien que des initiatives soient créées pour qu'Internet soit disponible sur un rayon de plus en plus grand, les débits restent faibles et les terminaux utilisés trop peu performants. Pour ce faire, Les entreprises doivent collaborer pour créer le nouveau modèle Internet. Des acteurs tels que Facebook ou Google doivent prendre des initiatives et d'élargir la zone et la qualité d'accès à Internet. L'un des phénomènes émergent est le « mobile-first » qui priorise l'expérience de l'utilisateur mobile permettant de réserver des applications pour les utilisateurs à un périmètre restreint laissant plus de champ pour ceux qui peuvent encaisser de plus lourdes fonctionnalités.

Nous vivons dans un monde déconnecté et alimenté par la batterie, où les méthodes et bonnes pratiques proviennent du monde passé, connecté et lié au réseau électrique. Une connexion à Internet est devenue un composant qui doit pouvoir être facilement retiré des applications, tout en assurant leur bon fonctionnement. La présente recherche préconise comme la solution pour optimiser les échanges est de réduire les couches de traitement grâce à des architectures modulaires, l'agrégation de données, l'accélération des requêtes et leur réduction, la disponibilité des données et des ressources en hors-ligne et la réduction du poids des ressources.

Cependant des axes d'améliorations sont encore possibles comme l'optimisation des applications et des services dont l'appel est lancé à tout chercheur en informatique de pouvoir approfondir quant à ce.

REFERENCES

[1]. D. Goderis et al., *Service Level Specification Semantics, Parameters and negotiation requirements*, draft-tequila-sls-02.txt, February, 2002.

[2] E. Gelenbe, G. Pujolle., *Introduction aux réseaux de files d'attentes*. Eyrolles, 1982.

[3] Ehran Cinlar, *Introduction to Stochastic Processes.* Prentice Hall 1975.

[4] Eurescom, *Selected Scenarios and requirements for end-to-end IP QoSmanagement*, Project P1008, Deliverable D2, 2001.

[5] ETSI TISPAN, Quality of Service (QoS) Framework and Requirements, ETSI TS 185 001, November, 2005

[6] G. Ash, A. Brader, C. Kappler, D. Oran, *QoS NSLP QSPEC Template*, draft-ietfnsis-qspec-20.txt, April, 2008.

[7] G. Cortese, *Service Configuration and Provisioning Framework: Requirements, Architecture, Design*, Cadenus Project, Deliverable D3.2, 2001.

[8] G. Doyon, *Systèmes et Réseaux de Télécommunications en régime stochastique*. Coll Technique et Scientifique des Télécommunications, Masson 1989.

[9] G. Hébuterne, *Ecoulement du trafic dans les autocommutateurs*. Masson, 1985

[10] Gross et Harris, *Fundamentals of Queueing Theory*. J. Wiley, 2nde édition, 1985.

[11] H. Takagi, *Queueing Analysis*. Vol. 1, North Holland, 1991.

[12] IST/FP6 Project, *EuQoS: End-to-End Quality Service over Heterogeneous Networks*, http://www.euqos.eu.

[13] M. Menai and G. Fromentoux, *Outil de modélisation, de conception et d'optimisation de l'urbanisme de l'architecture de commande d'un réseau de télécom.*, INPI 05 05557, 2005

[14] M. Schwartz, *Telecommunication Networks: Protocols, Modeling and Analysis.* Addison Wesley 1987.

[15] M. van Hartskamp et al., *UPnP QoS Architecture*, UPnP Forum, October, 2006

[16] O. Dugeon, *Service Provisioning in Premium IP: Recommendations to Telecom Operators and ISPs*, Cadenus Project, Deliverable D6.2, 2001.

[17] O. Dugeon and A. Brajeul (Ed.), *Project Recommendations*, Cadenus Project, Deliverable D6.3, 2003

[18] O. Dugeon et al., *End to End Quality of Service over Heterogeneous Networks*, In Proc. Networking Control, Lannion, France, Novembre, 2005.

[19] P. Boyer, O. Dugeon et M. Servel, *Systèmes d'allocation d'intervalle de temps et multiplexeurs pourvus d'un de ces systèmes d'allocation d'intervalle de temps*, Brevet INPI n° 93 09645, Novembre, 1993.

[20] P. Levis and M. Boucadair, *Considerations of Provider-to-Provider Agreements for Internet-Scale Quality of Service (QoS)*, RFC5160, March, 2008.

[21] R. Jain, *The art of computer systems performance analysis;* J. Wiley, 1991

[22] S.S. Lavenberg, *Computer Performance Modelling Handbook*. Academic Press, 1983

[23] T. Braun, M. Diaz, J.E. Gabeiras and T. Staub, *End-to-End Quality of Service Over Heterogeneous Network*, Springer, 2008

[24] T.G. Robertazzi, *Computer Networks and Systems: Queueing Theory and Performance Evaluation.* Springer Verlag 1990

[25] YENDE Raphael Grevisse, *De l'éventuel danger des téléphones portables en RDC*, coll. Convergence – FAB N°7/Butembo, 2018

[26] W. Bux, H.L. Truong, *Mean delay approximation for cyclic service queueing systems*. Performance Evaluation, vol 3, pp 187-196, 1983

[27] W. Htira, *Découverte et Agrégation de Topologies de Réseaux. Application au Contrôle d'Admission*, Ph.D Thesis from University of Toulouse, 2008

[28] W. Htira, O. Dugeon and M. Diaz, *A Mesh Aggregation Scheme for Call Admission Control*, New-Zealand, December, 2007.