solutions
réseaux

FRANCIS IA
OLIVIER MENAGER

Optimiser et sécuriser son trafic IP

Avec la contribution de
Jean-Marc Barozet,
Pascal Delprat et
Olivier Seznec,
de Cisco Systems,
et la collaboration de
Olivier Salvatori

ISBN : 2-212-11274-2
EYROLLES

10

Méthodologie de mise en œuvre d'une politique de gestion de trafic IP

Plusieurs justifications plaident en faveur de la mise en œuvre d'une réelle politique de gestion de trafic IP, ou GTI :

- Elle permet de structurer totalement l'architecture globale du système d'information sous l'angle premier de l'optimisation de la performance et de la sécurité.
- Elle offre une souplesse et une extensibilité pratiquement sans limite au système d'information lui permettant d'accompagner de manière optimale la montée en charge du trafic ainsi que l'évolution en nouvelles fonctionnalités et services, sans compromis avec la sécurité.
- Elle délivre au système d'information global une robustesse à toute épreuve, à la fois pour la continuité de service et pour la sécurité applicative.
- Elle fournit enfin une visibilité de l'ensemble du système d'information selon les critères de qualité de service et de performance de bout en bout.

Le coût de mise en œuvre d'une telle politique de GTI peut être lissé dans le temps. Son implémentation peut se faire par palier, avec pour chaque étape de transit des résultats significatifs d'amélioration des performances du système global. Le retour sur investissement est rapide, soutenant positivement la suite des déploiements.

La politique de GTI est un gage de pérennité des investissements réalisés, chaque brique posée dans l'infrastructure étant conservée dans les étapes suivantes de déploiement. De plus, dans les postes de coût d'un projet GTI, les coûts les plus importants restent ceux de construction. Il est dès lors essentiel de conserver de bas coûts de fonctionnement, qui sont pour la plupart récurrents.

Tableau 10.1 Coûts et ventilation d'une politique de GTI

Coût	Ventilation
Coût de construction	– Étude amont et spécification des besoins – Audit de l'existant et recommandations d'optimisation – Coût d'acquisition des équipements et logiciels – Intégration de l'infrastructure – Test en pilote, recettes et mise en production.
Coûts récurrents de fonctionnement	– Contrats de maintenance logicielle et matérielle – Coût d'exploitation de l'infrastructure (gestion, supervision, maintenance des versions, etc.) – Gestion des compétences d'exploitation et maintenance

L'infrastructure technique

La conception de l'architecture technique est un point clé dans le projet de mise en œuvre d'une politique de GTI. Elle doit prendre en compte les contraintes de l'environnement existant, les soucis d'interopérabilité avec les autres éléments existants et futurs, ainsi que sa propre sécurité et sa haute disponibilité.

Encore au stade émergent, le domaine de la GTI souffre d'un manque de retour sur expérience de mises en œuvre d'infrastructures techniques qui fonctionnent. Il est donc fortement conseillé de faire appel à une assistance d'expertise pour s'assurer de la bonne réussite de ce genre de projet. En particulier, la compétence d'architecte se révèle cruciale du fait de la multiplicité des briques matérielles et logicielles à assembler de façon correcte au sein de l'infrastructure.

Toute intégration d'éléments actifs permettant d'optimiser les performances ou la sécurité impose au préalable une analyse fine de l'environnement technique. La qualité de ce travail préalable est le garant du bon déroulement de l'intégration dans le système d'information de l'optimisation de performance et de sécurité désirée.

La qualification technique de l'existant se fait selon deux axes :

- la nature des éléments, réseau ou application ;
- l'analyse qualitative et quantitative.

Qualification du réseau

La qualification qualitative du réseau passe par la détermination des points techniques relatifs à l'architecture physique et à sa nature de connectivité réseau, d'une part, et à l'architecture logique (pontage, routage, interconnexion), d'autre part.

La qualification quantitative du réseau consiste à mesurer la volumétrie globale en terme de débit réseau, de temps de latence, de capacités de commutation du réseau local et de largeur de la bande passante d'accès au réseau WAN (Wide Area Network).

Qualification des applications

Une méthode d'analyse qualitative des applications consiste à les distinguer selon deux groupes, les applications interactives et les applications transactionnelles, et selon leur comportement réseau associé.

Il est important de bien identifier ces applications selon leur fonctionnement client-serveur. Il existe deux types d'applications client-serveur. Le premier emploie un numéro de port TCP ou UDP constant et unique. Par exemple, le flux HTTP/Web utilise le port TCP 80. Le second utilise des numéros de port dynamiques. En d'autres termes, les numéros de port sont générés dynamiquement pour chaque session et ne sont pas les mêmes d'une session à une autre, et il est impossible d'anticiper les numéros qui vont être utilisés par une session d'un utilisateur.

Une illustration d'une application client-serveur à numéro de port dynamique est le NFS (Network File System), le système de partage réseau de fichiers qui s'appuie sur les appels RPC (Remote Procedure Call), gérant des numéros de port TCP ou UDP alloués dynamiquement.

Le dernier point à identifier est l'existence ou non d'un contexte applicatif géré par l'application sur le serveur. Une application peut nécessiter de gérer un contexte applicatif par session utilisateur. Cela entraîne une contrainte sur l'élément d'optimisation, contrainte qui consiste à gérer soigneusement la persistance de session. Cette dernière consiste à garantir que tous les paquets d'une même session sont acheminés vers le même serveur.

Le tableau 10.2 dresse une typologie des applications.

Tableau 10.2 Typologie des applications

Application	Description	Exemple
Interactive	Peu de volume de trafic échangé entre le client et le serveur mais énormément de sessions courtes	Le Web est une application interactive. À chaque requête de contenu ou à chaque clic de souris de l'utilisateur, une connexion HTTP, et, dans le cas de HTTP 1.0, une connexion TCP associée est ouverte.
Transactionnelle	Beaucoup de volume transmis entre le client et le serveur mais peu de sessions ouvertes	FTP (File Transfer Protocol) est le type même d'une application transactionnelle. Chaque fois qu'un utilisateur se connecte pour transférer un fichier, cela se caractérise sur le réseau par des sessions qui sont maintenues assez longtemps (la durée du transfert des fichiers demandée). Par ailleurs, le volume de données qui transitent à travers le réseau est souvent relativement important.
À port statique	Caractérisée et donc identifiable par un numéro de port TCP ou UDP unique	La majorité des applications standards TCP/IP (Web, SMTP, Telnet, etc.) sont à numéro de port statique.

Tableau 10.2 Typologie des applications *(suite)*

Application	Description	Exemple
À port dynamique	Caractérisée par un ensemble de numéros de port TCP ou UDP utilisables pour les sessions utilisateur. Ces applications présentent souvent une session de contrôle (établissement-maintien-fermeture des sessions utilisateur) qui fonctionne sur un numéro de port TCP ou UDP statique et donc connu. La négociation du numéro de port pour la session se fait à travers cette session de contrôle.	Toutes les applications client-serveur reposant sur les appels de fonctions RPC (Remote Procedure Call) fonctionnent ainsi (exemple NFS). Les applications multimédias sont souvent dans ce mode. Par exemple, le standard de flux en streaming RTSP (Real-time Transport Session Protocol) présente une session de contrôle sur un port TCP unique et autant de sessions utilisateur (ouvertes en UDP avec un numéro alloué dynamiquement) que de requêtes d'objets multimédias à consulter.
À gestion ou non du contexte applicatif par session utilisateur	Certaines applications génèrent plusieurs sessions entre un client et un serveur. Dans le cas où il y a un contexte global applicatif à maintenir sur le même serveur physique, il est primordial que toutes les sessions du client soient redirigées vers le même serveur.	Pour illustration, la vidéo sur IP de type RSTP (streaming) crée une session de contrôle de communication client-serveur puis autant de sessions d'échange de données que de fichiers à transmettre et le même nombre de sessions de contrôle de transmission associées. Dans ce cas précis, il faut garder la totalité des sessions (de contrôle et de données) acheminées vers le même serveur pour un client donné.

Même s'il est toujours délicat de généraliser le profil de trafic des applications, chaque système d'information possédant ses spécificités de profil de trafic, il existe quelques similitudes remarquables.

Pour illustrer les nuances de caractérisation quantitative des applications, trois applications sont analysées au tableau 10.3, le Web, le téléchargement multimédia et l'application métier critique.

Tableau 10.3 Profil quantitatif des applications

Application	Description
Web	Généralement gourmande en bande passante, le trafic Web générant souvent un grand nombre de connexions par seconde. Avec HTTP 1.0, par exemple, chaque objet déclenche l'ouverture d'une session TCP, ce qui n'est pas le cas avec HTTP 1.1. Les objets d'une page Web sont habituellement de petite taille, de façon à optimiser leur téléchargement. Cela donne généralement des pages d'un très faible volume de données transmises sur le réseau, d'environ 70-80 Ko.
Téléchargement de vidéo	Le téléchargement en général et celui de vidéo en particulier est caractérisé par un établissement de session puis un volume important de données échangées. Cela a comme conséquence de saturer rapidement le réseau.
Application métier critique	L'application métier critique étant par essence vitale pour l'entreprise, le trafic doit être manipulé avec beaucoup de soin. La rapidité est le point crucial. C'est l'adoption de l'application par les utilisateurs qui en dépend. En règle générale, le nombre de connexions par seconde, et donc le débit total associé, est faible. La capacité et la qualité d'accueil de l'application sont les critères les plus importants car elles sont garantes du succès de l'application.

Le tableau 10.4 présente une synthèse d'une matrice d'analyse quantitative des applications qui peut servir d'exemple de qualification des applications à gérer en optimisation de performances et de sécurité.

Tableau 10.4 Matrice d'analyse quantitative des applications

Application	Métrique la plus importante	Seconde métrique importante	Métrique la moins importante
Web	Connexions par seconde	Débit	Nombre de sessions simultanées
FTP-vidéo	Débit (volume de trafic)	Nombre de sessions simultanées	Connexions par seconde
Métier	Nombre de sessions simultanées	Connexions par seconde	Débit

L'analyse de l'environnement applicatif est toujours un exercice difficile du fait de la complexité des applications. Cet exercice est cependant indispensable pour avancer dans le projet de mise en œuvre de la politique d'optimisation des performances et de la sécurité du système d'information global. C'est lui qui détermine en grande partie le choix des solutions à positionner et la conception de l'architecture (intégration des éléments) et qui sert de base à la métrologie et à la validation de l'architecture globale.

Intégration dans le système d'information

L'intégration dans le système d'information des éléments de gestion de trafic IP permettant d'optimiser les performances et la sécurité passe par la mise en adéquation des capacités d'insertion architecturale que supportent les éléments d'optimisation et des possibilités d'intégration offertes par le système existant.

Dans une seconde phase d'intégration, l'infrastructure de gestion de trafic IP (GTI) doit être vue comme un socle sur lequel viennent se poser l'ensemble des applications du système d'information. L'intégration de ces applications sur le support GTI doit se faire selon les règles de l'art.

Elle implique un projet de basculement progressif des applications vers la nouvelle infrastructure. La mise en place d'une plate-forme de test et de validation du bon fonctionnement des applications est préconisée, ainsi qu'une structure d'éducation des développeurs leur permettant d'intégrer au plus tôt la prise en compte des services d'optimisation IP disponibles dans la réalisation de leurs applications.

Capacités architecturales des éléments de GTI

Les capacités architecturales des solutions de GTI (gestion de trafic IP) des nombreuses solutions disponibles sur le marché s'adaptent au mieux, pour la plupart, aux contraintes de l'environnement existant, ainsi qu'aux besoins de services de GTI à apporter à l'infrastructure.

Les sections qui suivent présentent ces différentes possibilités d'intégration d'un élément d'optimisation de performances et de sécurité tout en faisant ressortir une méthodologie découpée en étapes de définition du mode d'intégration et de fonctionnement de l'élément GTI à intégrer.

Intégration physique

L'insertion d'un élément d'optimisation de performances et de sécurité dans l'architecture existante débute par le choix d'un type d'intégration physique et la manière de mettre en place la solution, comme le montre le tableau 10.5.

Tableau 10.5 Type d'intégration physique

Type d'intégration	Description	Avantage/inconvénient
Un seul attachement physique	L'élément consomme un port physique de l'infrastructure pour s'intégrer. En règle générale, il s'agit d'un port Ethernet (Fast ou Gigabit Ethernet). Les trafics entrant et sortant transitent par le même port physique de l'élément d'optimisation.	+ Consomme peu de port sur l'infrastructure existante. + Nécessite peu d'effort pour l'intégration physique. – Présente un débit divisé par deux pour les flux entrant et sortant.
Attachement *via* deux ports physiques	L'élément consomme deux ports physiques dans l'infrastructure. Le trafic entrant passe par un port et ressort par un autre.	+ Débit dédié et distinct pour les flux entrant et sortant. – Nécessite plus d'effort d'intégration physique (affectation des ports, choix des commutateurs d'accroche).

Le choix d'un type d'intégration physique est déterminé par les disponibilités de l'infrastructure en terme de connectique et par le débit réseau pour l'élément d'optimisation.

Configuration IP

L'élément d'optimisation se trouve généralement localisé au plus prêt des ressources qu'il optimise afin d'atteindre un niveau de performance ou de sécurité maximal.

Au niveau IP, la configuration se décline selon deux possibilités. L'élément d'optimisation et les ressources optimisées appartiennent soit au même sous-réseau IP, soit à des sous-réseaux différents. Dans ce dernier cas, l'élément d'optimisation doit assurer une fonction de routage IP.

Proxy ou non

Le mode proxy signifie que l'élément d'optimisation est le destinataire explicite des trafics qu'il doit gérer. En d'autres termes, tous les paquets portent comme adresse IP destination celle de l'élément d'optimisation, et non celle du destinataire réel, autrement dit la ressource optimisée.

À l'inverse, le mode non proxy se traduit par un élément d'optimisation qui n'est pas destinataire de son flux. Cela signifie qu'il doit être soit en coupure physique, sur le chemin physique de parcours, afin de voir transiter les flux qu'il doit traiter, soit en coupure logique, en fonctionnant comme routeur participant à l'acheminement des flux.

Un cas particulier de fonctionnement en mode non proxy est en architecture dite non intrusive.

Transparent ou non

La distinction entre transparent et non transparent porte sur la façon de gérer l'adressage source du trafic à traiter par l'élément d'optimisation.

En mode transparent, l'élément d'optimisation ne modifie pas l'adresse IP source des paquets qu'il relaie vers la ressource optimisée destinataire. Cela signifie que le destinataire du flux possède l'adresse source réelle du client. Pour des destinataires serveur, par exemple, il est possible d'extraire des statistiques de fréquentation en se reposant sur cette information d'adresse IP source.

Dans le mode non transparent, l'élément d'optimisation met son adresse IP en source des paquets qu'il relaie vers le destinataire. Il est alors impossible de connaître, au niveau de la ressource optimisée, l'identité réelle du client. Ce dernier ne voit que l'adresse de l'élément d'optimisation.

Ce mode est employé dans les cas particuliers où l'on cherche à contraindre le chemin de retour à repasser par l'élément d'optimisation, sans être en coupure physique ou logique (routage) du flux.

Mode intrusif ou non

Un élément d'optimisation peut être positionné complètement à l'extérieur du chemin de parcours du trafic. Grâce à un mécanisme de copie de trafic, tel que la duplication de port sur un commutateur ou la copie de trafic *via* un connecteur à trois branches, l'élément d'optimisation reçoit une copie du trafic à traiter. Ce mode est appelé non intrusif. Une sonde de détection d'intrusion fonctionne généralement de cette façon.

L'avantage premier de ce mode est sa complète transparence vis-à-vis du réseau opérationnel. En particulier, une défaillance n'engendre aucun impact sur le fonctionnement du réseau opérationnel. La seule conséquence est l'absence de la fonction d'optimisation qu'il est censé apporter au système global.

Le mode intrusif est le mode opposé. La plupart des solutions d'optimisation fonctionnent plutôt en mode intrusif. Se pose toutefois la question du point de défaillance unique que sous-tend ce mode. Il est alors important de considérer une redondance de la solution afin d'assurer la continuité de service.

Étapes d'implémentation

L'intégration dans le système d'information d'éléments d'optimisation de la performance et de la sécurité se décompose en quatre étapes :

1. Analyse du contexte existant (réseau et applications).
2. Expression des besoins.
3. Selon les capacités architecturales des solutions, détermination de la meilleure façon d'intégrer l'élément d'optimisation.
4. Configuration progressive de la prise en compte des applications par les nouveaux éléments d'optimisation mis en place.

Métrologie et validation

Cette section détaille les différentes facettes de la performance associée à la gestion de trafic IP. Il existe plusieurs façons de mesurer la performance dans une architecture intégrant l'optimisation de la performance et de la sécurité. Chaque métrique possède un niveau d'importance distinct, dépendant des besoins spécifiques des applications à considérer.

Métrologie

Les trois métriques à analyser sont :

- le nombre de connexions par seconde ;
- le nombre total de connexions simultanées (concourantes) ;
- le débit, en bit par seconde.

La maîtrise de ces métriques est primordiale car elles jaugent les limites du système implémenté.

Nombre de connexions par seconde

Avec la montée en charge, c'est probablement la métrique la plus importante, surtout lorsqu'il s'agit de trafic HTTP.

Le nombre de connexions par seconde correspond au total des sessions entrantes acceptées par l'architecture pendant une seconde. Parfois nommé nombre de sessions ou de transactions par seconde, il est représentatif de la limite de performance de l'équipement. La gestion des sessions applicatives de type HTTP sollicite en effet énormément de ressources matérielles du fait du grand nombre d'ouvertures et de fermetures de sessions applicatives à gérer au niveau de la pile protocolaire TCP/IP.

Pour illustrer les contraintes liées à la gestion de sessions HTTP, considérons la séquence suivante :

1. Le client initialise une connexion HTTP en envoyant un paquet TCP SYN à destination du port 80 du serveur Web.
2. Le serveur Web répond en envoyant un paquet ACK, enchaîné à un paquet SYN.
3. Le client acquitte à son tour le SYN provenant du serveur en lui envoyant un ACK.

Une fois la session établie, l'échange de données au niveau HTTP peut débuter. Cette procédure d'établissement de session, appelée Three-Way Handshake, est obligatoire pour chaque envoi de données sur le réseau, quelle que soit la taille de ce dernier. Dans le trafic Web, où les données échangées sont peu volumineuses, la gestion protocolaire est particulièrement contraignante en charge.

Nombre total de connexions simultanées

Le nombre total de connexions simultanées est la métrique qui détermine le nombre maximal de connexions TCP qu'un équipement est capable de supporter. Typiquement, ce nombre est lié à

la taille de la mémoire embarquée sur l'équipement. Ce nombre peut varier de quelques dizaines de milliers à l'infini. Il dépend de la solution considérée.

La plupart du temps, ce nombre reste théorique et n'est que rarement atteint dans la réalité.

Débit

Le débit est la troisième métrique importante à considérer. Exprimé en bit par seconde, le débit correspond au taux d'acheminement de trafic à travers l'architecture. Cette variable dépend de l'architecture interne et de la capacité du bus de l'équipement.

Même si le débit est en bit par seconde, ce paramètre découle de deux autres paramètres, la taille du paquet et le nombre de paquets par seconde.

L'unité de manipulation des données par le commutateur est le paquet. Plus la taille du paquet est importante, plus l'efficacité de débit est grande.

En résumé, les trois paramètres précédents sont nécessaires à obtenir en avance de phase, avant déploiement, puis après la mise en place des solutions d'optimisation. La comparaison des résultats des deux campagnes de mesures — avant et après — permet de démontrer l'optimisation apportée à l'infrastructure.

Ces paramètres sont également d'excellents indicateurs de la qualité de service du système global, tout au long de la vie du système.

Validation

La validation du système mis en œuvre passe par la rédaction d'un cahier des tests décrivant l'ensemble des tests de performances à effectuer, ainsi que les résultats des tests avec l'ancienne plate-forme et ceux recueillis des nouveaux tests avec les optimisations intégrées.

Les tests de performances sont exécutés par injection de trafic au moyen de générateurs de trafics. Le cahier des tests doit inclure des tests de disponibilité à des fins de détermination des temps de basculement offerts par le système, sur l'ensemble de la chaîne de communication.

Organisation et règles d'exploitation

Un partage des responsabilités techniques est souvent nécessaire afin d'expliciter les contours de compétence et d'implication de chacune des équipes d'exploitation. S'intercalant entre le réseau et l'application, le domaine de l'optimisation des performances et de la sécurité est souvent pris entre plusieurs feux. L'exploitation des éléments d'optimisation IP peut prêter à débat entre les différentes équipes, application, sécurité et réseau.

Si la ventilation des tâches d'exploitation entre les équipes ne peut se faire naturellement, il est obligatoire d'arbitrer sa définition afin de converger vers une structure organisationnelle adaptée au besoin d'exploitation GTI.

Découlent de la structure d'exploitation des règles couvrant des aspects tels que :

- les règles de mise en place d'éléments GTI ;

- les normes de supervision et les règles d'intervention en cas de problème (dépannage) ;
- les processus d'évolution des politiques de GTI ;
- les processus de suivi des performances et d'évaluation de la protection sécurité.

L'ensemble de ces règles d'exploitation est détaillé et vient alimenter une documentation technique complète de la politique de GTI de l'entreprise.

En résumé

La méthodologie proposée dans ce chapitre met l'accent sur les critères techniques à considérer dans chacune des phases importantes de la mise en œuvre d'une politique de gestion de trafic IP.

Une bonne optimisation de la performance et de la sécurité dépend de la qualification technique de l'existant, qui doit être la plus fine et juste possible. L'intégration des nouveaux éléments d'optimisation dans le système d'information doit se faire avec un minimum d'impact sur son fonctionnement, en réduisant toute modification nécessaire.

Les phases de validation et de recette clôturent le projet et permettent de s'assurer que les améliorations apportées au système d'information sont réellement significatives.