Cœur de Commutation/Routage Gigabit Ethernet

■ Didier CHASSIGNOL, Didier.Chassignol@inrialpes.fr
INRIA, Rhône-Alpes

A l'occasion de l'extension du bâtiment de l'INRIA Rhône-Alpes, le réseau est profondément remanié de manière à pouvoir supporter les flux de données et les contraintes d'administration pour les trois années à venir. L'augmentation exponentielle du trafic avec en particulier les flux multicast ainsi que les besoins de qualité de service implicites *conduisent à basculer d'un backbone 100 Mb vers un backbone Gigabit Ethernet.*
Un second grand principe est mis en œuvre : La commutation Ethernet 10/100 Mb pour l'ensemble du parc, afin d'envisager sereinement les futurs débits, et surtout de simplifier grandement l'administration du réseau : les re-cablage seront quasiment nuls. A noter une conséquence bénéfique : Il ne sera plus possible « d'écouter » sur le réseau.

■ Descriptif global du projet

Intégration de l'existant

Les matériels existants sur le site peuvent difficilement évoluer vers les nouveaux standards ou à des coûts prohibitifs. Un routeur Cisco 7507 sera réutilisé comme routeur d'entrée de site (deux pattes 100Mb). Les commutateurs Catalyst 5000 (première génération) ne seront réutilisé qu'en cas de nécessité et uniquement en commutateur d'extrémité et sans gestion de trunks sauf si les nouveaux équipements permettent les trunks ISL.

Topologie du nouveau réseau : Objectifs principaux

Le futur réseau est architecturé autour d'un commutateur/routeur Gigabit Ethernet disposant de 8 à 24 ports. L'architecture de ce cœur de réseau est non bloquante et le routage se fait à « la vitesse du câble » et ceci même si des fonctions telles que Access Lists, RMON... sont activées.

Chaque VLAN est routé afin « d'isoler/protéger » chaque projet de recherche.

L'essentiel des ports du cœur de réseau est du type 1000BASE-SX.Des ports du type 1000BASE-LX pourront être ajoutés à la configuration de base, et dans une moindre mesure des ports 10/100 Mb.

Ce commutateur/routeur Gigabit Ethernet dessert les commutateurs d'extrémités repartis dans 6 locaux techniques :

- Locaux B, D, E, H : 150 ports/local,
- Locaux C, F : 80 ports/local,
- Soit un total de 760 ports commutés 10/100 Mb.

Dans un premier temps, chaque local technique est raccordé au cœur de réseau par un seul lien Gigabit Ethernet. Par la suite la solution proposée permettra d'évoluer vers un ratio (nombre de liens Gigabit vers le cœur de réseau)/ (nombre de ports 10/100) égal a 1/ 24.

Les ports 10/100 Mbs supportent l'auto-négociation telle que définie dans la clause 28 du standard 802.3 de l'IEEE.

Conditions de l'appel d'offre

Afin de pouvoir sélectionner le(s) produits correspondant au mieux aux besoins, il fut décider de procéder à un appel d'offre sur performance. Cette procédure quoique un peu lourde donne suffisamment de souplesse pour tendre vers le seuil de performance requis en permettant que s'instaure un véritable dialogue avec les candidats. L'appel d'offre a été découpé en deux lots :

- Lot 1 : cœur de commutation/routage.
- Lot 2 : équipements d'extrémité.

Une préférence est toutefois donnée si un constructeur sait proposer une solution cohérente pour les deux lots.

Les candidats devront proposer une solution répondant au mieux aux critères énumérés dans le programme fonctionnel. Celui-ci peut donc être particulièrement riche. Le programme fonctionnel stipule également que le matériel proposé devra être mis à disposition de l'INRIA Rhône-Alpes pour évaluation de ses performances.

■ Le Programme fonctionnel (extrait)

> On trouvera dans ce chapitre les principaux éléments du programme fonctionnel utilisés par l'INRIA Rhône-Alpes, ainsi que des annotations destinées à aider les futurs rédacteurs. Nous avons volontairement laissé les critères forcément très spécifiques à l'INRIA Rhône-Alpes.

Caractéristiques détaillées du commutateur routeur (cœur de réseau)

Nombres d'interfaces

Dans un premier temps 6 interfaces Gigabit Ethernet sont nécessaires pour desservir les 6 locaux techniques. Toutefois 2 interfaces supplémentaires sont demandées immédiatement en réserve. La solution proposée devra donc disposer d'au minimum 8 interfaces Gigabit Ethernet.

Seize ports supplémentaires devront pouvoir être installés dans le même chassi portant le nombre total d'interfaces Gigabit Ethernet (IEEE 802.3z) a vingt-quatre.

L'essentiel des ports du cœur de réseau sera du type 1000BASE-SX car les distances maximales entre les locaux techniques n'excèdent pas 220 mètres. Des connexions 1000BASE-LX devront être possible en vue de raccordements de futurs bâtiments situés à moins de 550 mètres.

La modularité sera la plus grande possible.

> Une solution proposant des cartes Gigabit avec modules GBIC apporte une grande souplesse, par contre pensez à faire valider quel type est fournit en standard pour les interfaces en réserve.

Performances

L'architecture du cœur de réseau devra être non bloquante et le routage devra se faire à « la vitesse du câble » et ceci même si des fonctions telles que filtrage, RMON... sont activées.

> Ne pas hésiter à demander un niveau de performances élevé afin d'anticiper les futurs besoins !

Les chiffres de performances mesurés par un organisme indépendant seront fournis : capacité de commutation, temps de latence, taille des tables d'adresses… Et cela dans diverses conditions (commutation niveau 3, présence de filtres ou non…).

> Il devient de plus en plus difficile de réaliser avec des équipements courants des tests de stress. Autant s'appuyer sur le travail d'organismes spécialisés.

Fiabilisation (redondance)

L'équipement central devra permettre une redondance de ses éléments les plus importants (alimentation, matrice de connexion, carte de supervision…) avec reprise automatique lors de défaillance et changement à chaud possible sans arrêter le système.

Fonctionnalités

- **Routage**

Le routeur devra au minimum pouvoir router les protocoles IP unicasts et multicasts (RFC 1812 et RFC 1122). Le routeur IP devra permettre la création de routes IP unicast statiques. Au minimum 64 routes pourront être créées.

Les protocoles de routage suivants devront être disponibles : RIP v1 et v2, OSPF v2, DVMRP. L'implémentation de PIM (tel que décrit dans le RFC 2362) serait un plus. Des tunnels DVMRP ainsi que des tunnels AppleTalk et éventuellement des tunnels PIM devront pouvoir être tirés.

> Demander PIM v2 Dense mode et Sparse mode permettait à l'INRIA Rhône-Alpes non seulement de s'assurer d'un bon routage du multicast mais aussi de vérifier l'aptitude du constructeur à implémenter des protocoles récents ou en cours de standardisation.

Le routage d'AppleTalk phase 2 est nécessaire, éventuellement sur un routeur séparé et à un niveau de performance plus faible. Dans l'hypothèse d'un routeur AppleTalk externe, il devra supporter le trunk 802.1Q et disposer au moins d'un port 100Mb/s.

- **Filtrage**

Dans un premier temps le nouveau cœur de commutation/routage n'aura pas à faire de filtrage, toutefois cette fonctionnalité sera présente pour une utilisation future. Elle permettra d'interdire (ou d'autoriser) des flux en fonction :

- Des protocoles (IP, IPX, AppleTalk…).
- Des adresses sources et destinations avec la possibilité de grouper ces adresses sous la forme d'un préfixe et d'un masque.
- Des ports sources et destinations.
- Du type de flux (udp, tcp).
- On précisera également :
  - Les limitations éventuelles (nombre de critères par liste, nombre de liste…),
  - L'impact de l'utilisation de ces filtres sur les performances du matériel.

S'il existe la possibilité de télécharger les filtres, et si la prise en compte se fait dynamiquement ou nécessite un reboot.

Possibilité d'enregistrer par syslog ou tout autre mécanisme des informations sur les paquets filtrés, ainsi que les éventuels outils fournis pour analyser ces logs.

- **Forward des broadcasts**

Le routeur devra être capable au minimum de faire du forward des broadcast BOOTP et DHCP. Pouvoir étendre cette fonctionnalité à d'autre type de broadcast serait un plus.

On précisera :

- Les types de broadcast pouvant être redirigés (BOOTP, DHCP…).
- Les éventuelles limitations (redirection vers une seule adresse…).
- L'impact sur les performances du matériel.

- **VLAN**

Le commutateur devra savoir gérer les VLAN par port. La gestion des VLAN par protocole ou tout autre mécanisme serait un plus.

Le commutateur gérera une table d'adresse MAC par VLAN.

> Si un commutateur ne gère pas une table d'adresse MAC par VLAN une usurpation d'adresse MAC, ou l'utilisation d'un mode trunk sur un routeur pourra avoir des conséquences graves.

Le commutateur sera capable de gérer des trunks au format 802.1Q, et implémentera le protocole GARP VLAN Registration Protocol (GVRP) tel que définit dans la norme 802.1Q. La gestion des trunks au format ISL (protocole Cisco) serait un plus.

> Le protocole GVRP permet la reconnaissance automatique des VLANs sur un port trunk. Cette fonctionnalité est non seulement intéressante techniquement, mais son apparition récente dans le standard 802.1Q permet de tester la capacité et la volonté du constructeur à implémenter les standards.

- **QoS**

La volonté de l'INRIA Rhône-Alpes est de pas à avoir à gérer de QoS, toutefois en vue d'une possible utilisation future on précisera comment l'architecture du système prend en compte les problématiques de Qualités (ou Classes) de Service :

- Files d'attente internes dans les commutateurs.
- Codage des informations dans les trames (IEEE 802.1p, autres…).
- Nombre de niveaux de priorités gérés.
- Comment définit-on ces critères (adresse source, destination, protocole, numéro de port…).
- Limitations comme obligation d'utiliser le module de routage (même si on ne change pas de VLAN) pour disposer de fonctions de qualité de service…

- **Multicast**

De façon à prendre en compte les flux multicasts, le commutateur devra interpréter les paquets liés au protocole IGMP v1 et v2 (ce que certains constructeurs appellent IGMP snooping). Il s'appuiera dessus pour

limiter, au sein d'un même VLAN, la diffusion des paquets multicasts aux seuls ports sur lesquels des machines sont toujours abonnées.

Par ailleurs le protocole IGMP (v1 et v2) devra être implémenté de façon à permettre la gestion des groupes multicasts sur le routeur.

- **Agrégation de liens**

Il est intéressant de disposer de la possibilité d'agréger des liens pour disposer d'une bande passante plus grande. Si cette fonctionnalité est disponible, on précisera :

- Quelles interfaces offrent cette possibilité.
- Le nombre de liens pouvant être agrégés.
- Jusqu'à quels débits.
- Les mécanismes mis en œuvre et leur interopérabilité avec des matériels d'autres constructeurs.
- Limitations éventuelles, comme l'appartenance obligatoire à la même carte des liens agrégés.

Le standard LAG est en cours d'élaboration d'où l'importance de l'inter-opérabilité.

- **Possibilité de déport de trafic**

Il est intéressant de pouvoir déporter tout trafic (entrée et sortie) sur un port donné à des fins d'analyse réalisée via un analyseur externe ou un logiciel spécifique. On précisera :

- Quel type de déport est peut-être déporté (tout le trafic sur un port physique donné, le trafic entre équipements informatiques identifiés par des adresses MAC, etc.)
- Limitations éventuelles (nombre de port « mirrorés », uniquement les données entrantes ou sortantes…),
- Si les deux ports (miroir et mirroré) peuvent être situés sur deux équipements différents.
- L'éventuel impact sur les performances lorsque ce type de fonctionnalité est activé.

Administration du cœur de commutation/routage

On précisera :

- Pour le mode commande les mécanismes d'édition (historique, commandes à la emacs…), ainsi que l'accessibilité (console VT100, telnet…), et si la totalité des commandes est accessible via ce mode.
- Si un outil graphique est proposé. Si c'est le cas, on décrira les mécanismes (HTTP, application dédiée, JAVA…) et les plates-formes sur lesquelles cet outil peut être installé. On précisera également si toutes les commandes sont disponibles à partir de cet outil.
- Si le fichier de configuration est téléchargeable et par quel mécanisme (TFTP…) ainsi que le type du fichier (binaire, ASCII…).
- Les limitations telles que : reboot nécessaire pour prise en compte des modifications de configuration, reboot après un téléchargement de configuration, (temps maximum de reboot à préciser)…
- Les mécanismes mis en œuvre pour assurer la protection de l'équipement (niveaux de mots de passe lecture et lecture/écriture, limitations des accès telnet/HTTP à partir d'adresses IP sources…).
- Si des mécanismes sont prévus afin de propager certains aspects de la configuration sur d'autres équipements (VLANs, listes d'accès, classes de priorité…).

Administration distante

Afin de faciliter l'administration à distance, il serait souhaitable que les MIB SNMP suivantes soient implémentées :

- MIB II (RFC 1213, RFC 2011, RFC 2012, RFC 2013).
- IP Forwarding Table (RFC 2096).
- Bridge MIB (RFC 1493).
- Evolution of interfaces (RFC 2233).
- RIP2 MIB (RFC 1724).
- RMON (RFC 1757).
- OSPF2 (RFC 1850).
- RMON II (RFC 2021).
- 802.3 MAU (RFC 2239).

Les principaux événements devront pouvoir être enregistrés à la fois dans le commutateur mais également sur une machine distante par un mécanisme de SYSLOG. Parmi ces événements, on devra trouver au minimum

les tentatives de connexions échouées, les changements de configuration, les chargements de nouveaux binaires, qui s'est connecté à l'équipement, depuis quelle adresse.

Caractéristiques détaillées des commutateurs d'extrémités

Nombres d'interfaces

Le commutateur/routeur Gigabit Ethernet dessert des commutateurs d'extrémités répartis dans six (6) locaux techniques nommés B, C, D, E, F, H. Ci-dessous le nombre de prises 10/100Mb/s dans chaque local technique :

- Locaux B, D, E, H : 150 ports / local,
- Locaux C, F : 80 ports / local,
- Soit un total de 760 ports commutés 10/100 Mb/s.

Dans un premier temps, chaque local technique est raccordé au cœur de réseau par un seul lien Gigabit Ethernet. Par la suite la solution proposée devra pouvoir évoluer vers un ratio (nombre de liens Gigabit vers le cœur de réseau)/ (nombre de ports 10/100) égal à 1/ 24. Les ports Gigabit Ethernet (IEEE Std 802.3z) seront du type 1000BASE-SX. Toutefois la disponibilité d'interfaces 1000BASE-LX serait un plus.

Les ports 10/100Mb/s devront supporter l'auto-négociation (choix de la vitesse *et* du mode half ou full duplex) telle que définie dans la clause 28 du standard 802.3 de l'IEEE.

On précisera :

- Le type de chaînage entre équipements périphériques si besoin.

Performances

Les performances ne devront pas être dégradées si des fonctions telles que "mirroring"... sont activées.

Les performances mesurées par un organisme indépendant seront fournies : capacité de commutation, temps de latence, taille des tables d'adresses…

Fiabilisation (redondance)

Les équipements périphériques peuvent permettre une redondance de certains de ses éléments les plus importants (alimentation, matrice de connexion, carte de supervision…) avec ou sans reprise automatique lors de défaillance et avec ou sans changement à chaud possible sans arrêter le système.

Fonctionnalités

- **VLAN**

Idem lot 1 (§ « VLAN » précédent).

- **QoS**

Idem lot 1 (§ « QoS » précédent).

- **Multicast**

De façon à prendre en compte les flux multicasts, le commutateur devra interpréter les paquets liés au protocole IGMP v1 et v2 (ce que certains constructeurs appellent IGMP snooping). Il s'appuiera dessus pour limiter, au sein d'un même VLAN, la diffusion des paquets multicasts aux seuls ports sur lesquels des machines sont toujours abonnées.

- **Agrégation de liens**

Idem lot 1 (§ « Agrégation de liens » précédent).

- **Possibilité de déport de trafic**

Idem lot 1 (§ « Possibilité de déport de trafic » précédent).

- **Administration de chaque commutateur**

Idem lot 1 (§ « Administration du cœur de commutation/routage » précédent).

Administration distante

Afin de faciliter l'administration à distance, il serait souhaitable que les MIB SNMP suivantes soient implémentées :

- MIB II (RFC 1213, RFC 2011, RFC 2012, RFC 2013).
- Bridge MIB (RFC 1493).

- 802.3 MAU (RFC 2239).

Les principaux événements devront pouvoir être enregistrés à la fois dans le commutateur mais également sur une machine distante par un mécanisme de SYSLOG. Parmi ces événements, on devra trouver au minimum les tentatives de connexions échouées, les changements de configuration, les chargements de nouveaux binaires, qui s'est connecté à l'équipement, depuis quelle adresse.

On précisera :

- Les limitations (nombre d'événements enregistrés en NVRAM…).

■ Les tests

Que tester ?

Les performances pures ?

Nous n'avons pas la capacité pour réaliser des tests de stress. Par conséquent nous avons choisi de nous appuyer sur les tests réalisés par des organismes indépendants.

Les tests de fonctionnalités ?

Beaucoup plus faciles à réaliser, ces tests prennent toutefois beaucoup de temps. Nous avons choisi de les réaliser nous-mêmes sans l'assistance permanente du constructeur et/ou de l'installateur. Cette façon de procéder nous permettait d'évaluer l'ergonomie des interfaces utilisateurs, la qualité des documentations, et le temps de réaction des forces de supports.

Nous avions choisi de faire porter nos tests sur les points suivants :

- Interface utilisateur & Administration
  - Disponibilité d'une interface CLI avec un éditeur à la emacs, un historique, une aide en ligne.
  - Vérification des niveaux de sécurité (login, limitation des accès telnet HTTP a certaines adresses IP source).
  - Sauvegarde d'un fichier de configuration en ASCII sur un serveur, et rechargement de ce fichier après modification.
  - Tests sur différentes OID de la MIB II.
  - Essai de logguer quelques évènements significatifs sur un serveur et en NVRAM.
- Commutation
  - Création de VLANs par port.
  - Vérification de la gestion des tables d'adresses MAC par VLAN.
  - Essai de création de trunk ISL et/ou 802.1Q.
  - Essai de mise en place des mécanismes d'auto apprentissage des VLANs sur les ports trunks.
  - Vérification de l'IGMP snooping.
- Routage
  - Tests d'interopérabilité des protocoles de routage RIP, OSPF, PIM, DVMRP.
  - Vérification des mécanismes de forward des broadcasts UDP.
  - Mise en place d'Access Lists.
  - Support technique.
  - Evaluation des documentations papiers/en ligne.
  - Evaluation du niveau de qualité du support technique.

■ Conclusions

La présentation de cette étude lors des journées du JRES sera l'occasion de d'exposer en détail les résultats de notre évaluation. La très grande disparité en terme de performance pure et de richesse fonctionnelle montre à quel point ces produits sont encore émergeants en particulier chez certains « gros constructeurs ». Les effets d'annonces sont nombreux et les implémémentations parfois hasardeuses. Toutefois quelques produits sortent du lot et pas nécessairement ceux qu'on attendait. Je citerai parmi les produits les plus intéressants de cette étude :

- En cœur de commutation/routage le SSR 8600 de Cabletron, ainsi que le Black Diamond 6800 d'Extreme Networks.
- En commutateur d'extrémité le Summit 48 d'Extreme Networks.