Implémentation des classes de services dans RENATER-3

Franck SIMON
GIP RENATER
ENSAM – 151 Boulevard de l'Hôpital – 75013 PARIS
Franck.Simon@renater.fr
15 octobre 2003

Résumé

Présentation des classes de service qui vont être proposées dans RENATER-3 (classe « Less than Best Effort » ou LBE, classe « Best Effort » ou BE, classe « Better than Best Effort» ou BBE, classe « IP Premium », classe « supervision »), principes d'implémentation de ces classes dans le backbone RENATER-3, possibilités d'extension de ces classes vers d'autres réseaux (les réseaux de collecte – réseaux régionaux et métropolitains – connectés à RENATER, et extensions à l'international), et inter-opérabilité des classes de service avec des services avancés (IPv6, Multicast, MPLS).
Cette présentation, qui s'adresse à un public possédant déjà de bonnes connaissance théoriques sur les classes de service et sur les files d'attente dans les routeurs, sera en outre illustrée par des exemples de configuration (pour des équipements Cisco uniquement).

Mots clefs

RENATER, classes de service.

Introduction

Des classes de services vont être déployées dans le backbone RENATER-3 début 2004. La phase actuelle (dernier trimestre 2003) est réservée pour la validation des règles d'ingénierie, lesquelles règles seront validés par des tests avancés sur maquettes (les tests pour vérifier tous les grands principes et les différentes classes de service ont déjà été effectués). Des tests vont ensuite être étendus à des sites pilotes, avant un passage à l'échelle.

Le dimensionnement du backbone actuel RENATER-3 est basé sur une politique d'« over-provisionning » pour garantir un service « Best-Effort » de bonne qualité. Mais ce surdimensionnement n'est pas toujours une réponse aux problèmes posés et surtout n'est pas toujours possible dès que l'on intègre la notion de services de bout en bout (avec possibilité de goulet d'étranglement aux extrémités du réseau). Certains services ou applications nécessitent de la bande passante garantie, d'autres des temps de réponse stables (visio, Vo/IP..), et d'autres encore n'ont pas ces contraintes mais peuvent vouloir utiliser toute la bande disponible ou du moins le maximum possible (gros transferts de fichiers, miroirs FTP, applications de type GRID – grilles de calcul -…).

Dans ce contexte, la mise en place de classes de service s'avère indispensable, mais cela ne présente un intérêt que si la notion de service de bout en bout est respectée, d'où la nécessité d'étendre ces classes de service jusqu'aux sites RENATER via les réseaux de collecte. L'intérêt des classes de service peut être encore accru si ces classes sont étendues à l'international (notamment vers d'autres réseaux nationaux de la recherche avec lesquels RENATER est interconnecté via le réseau de la recherche GEANT).

1. Les classes de service dans RENATER-3

RENATER prévoit de déployer 5 classes de services, toutes basées sur l'utilisation du champ DSCP (Differenciated Service Code Point).

Plusieurs valeurs DSCP seront utilisées/réservées :

LBE (Less than Best Effort)	DSCP-LBE = 8
BE (Best Effort)	DSCP-BE = 0
BBE (Better than Best Effort)	DSCP-BBE-INC (In-Contract) = à définir DSCP-BBE-OOC (Out-Of-Contract) = à définir
IP Premium	DSCP-PREMIUM
Supervision/contrôle du réseau RENATER-3	DSCP-SUP = à définir

En dehors de ces valeurs DSCP réservées, si des paquets entrent sur le backbone RENATER marqués avec une valeur DSCP différente, alors la valeur DSCP de ces paquets ne sera pas modifiée par les routeurs RENATER (« DSCP transparency »). Si des paquets entrent dans le backbone avec des valeurs DSCP réservées, et qu'il s'agit de paquets non autorisés, alors ces paquets seront remarqués avec une valeur DSCP 0 (valeur DSCP-BE).

A l'exception de la classe IP Premium où le trafic sera traité en EF (Expedited Forwarding), les autres classes de service seront traités en AF (Assured Forwarding).

Considérations : le marquage initial des flux ne sera pas fait par les équipements RENATER, mais par les équipements des sites RENATER (ou les applications clientes, si toutefois certaines applications savent le faire), voire les équipements des réseaux de collecte, mais pour une réelle efficacité le marquage et le traitement spécifique des flux marqués doivent se faire au plus près de la source. **Les routeurs RENATER ne feront donc pas de marquage des flux mais uniquement un re-marquage si nécessaire** (effectué lors de la vérification/contrôle des flux, au niveau des interfaces de raccordement avec les éléments extérieurs - réseaux régionaux, métropolitains, sites …-). Si un élément extérieur souhaite recevoir de RENATER des flux marqués uniquement des classes de service proposées par RENATER, voire uniquement des flux systématiquement re-marqués à 0, il devra alors effectuer ce re-marquage au niveau de ses propres équipements réseaux.

S'il dispose d'équipements Cisco, l'élément extérieur pourra effectuer la classification de ses flux sur la base d'ACL (Access-List), de « class-map » (équivalent d'une « route-map » mais adaptée à la classification de flux) et NBAR (Network Based Application Recognition) (NBAR [1] permettant notamment de reconnaître les applications dont les ports changent ou sont négociés dynamiquement, NBAR se base sur l'utilisation de profils appelés PDLM [1]- Packet Description Language Module -). Note : NBAR est disponible sur les gammes Cisco 2600, 3600, 3700, 6500, 7200, 7300, 7500, 7600. NBAR est traité en software sauf sur les gammes 6500 et 7600 où la fonctionnalité est traitée via la MSFC, et sur la gamme 7500 où la fonctionnalité est traitée via la VIP (dNBAR – Distributed NBAR -). Cette fonctionnalité est intéressante à l'échelle d'un site RENATER.

L'élément extérieur pourra également effectuer à son niveau un « policing » ou un « shaping » pour la gestion de la bande passante associée aux flux marqués.

1.1. Les files d'attente

Pour rappel, voici les principaux mécanismes pour la gestion des files d'attente :

- Files d'attente orientées EF :
  - SPQ (Strict Priority Queuing) & LLQ (Low Latency Queuing) : les paquets peuvent être confinés dans une file d'attente qui est traitée avec une priorité absolue vis-à-vis des autres files d'attente.
- Files d'attente orientées AF :
  - WFQ (Weighted Fair Queuing) & CBWFQ (Class Based Weighted Fair Queuing) : les paquets peuvent être confinés dans une file d'attente spécifique qui bénéficie d'une priorité relativement importante vis-à-vis des autres files d'attente ;
  - WRR (Weighted Round Robin), DRR (Deficit Round Robin) & MDRR (Modified Deficit Round Robin) : toutes les autres files d'attente sont réduites à se partager un petit espace de temps « round robin », ce qui

revient à accorder plus de temps au traitement d'une file d'attente spécifique, et donc un traitement plus rapide des paquets.

Pour les routeurs Cisco, les files d'attente disponibles, et leur implémentation, varient en fonction de la gamme de l'équipement (c'est la raison pour laquelle, dans le reste du document, seront plutôt utilisés les termes EF et AF, plus génériques, plutôt que des noms explicites de type de file d'attente) :

- routeurs 12xxx (12000, 12400) : MDRR (Engine 0 et Engine 2, avec implémentation spécifique pour traiter néanmoins des flux EF), LLQ (Engine 3)
- routeurs 2600, 3600, 3700, 7200, 7500 : LLQ = PQ (Priority Queuing) + CBWFQ
- switchs et switchs-routeurs Catalyst (3550, 4500, 6500) : WRR, SPQ (au niveau de chaque port physique, possibilité d'utiliser 4 files d'attente dont 1 peut être traitée en SPQ)

Pour chacune des classes de service, une file d'attente spécifique est paramétrée sur les routeurs RENATER.

2. Les algorithmes des classes de service dans RENATER-3

2.1 Classe "Less than Best Effort" (LBE)

Les paquets LBE sont marqués avec la valeur DSCP-LBE (8 : 001000). C'est la même valeur que celle utilisée par « Scavenger » dans Internet2 (service QBSS : QBone Scavenger Service [2]).

Avec le service LBE :

- possibilité de disposer d'une bande passante très importante, mais non garantie
- délais ou temps de réponse variables
- pertes de paquets possibles

Cette classe de service permet d'utiliser la bande passante non utilisée par des classes de services supérieures.

Le service LBE proposé par RENATER est compatible avec celui proposé dans GEANT, et donc des extensions sont techniquement possibles vers les autres réseaux de la recherche (sous réserve qu'ils aient implémenté de telles classes dans leur réseau). La valeur DSCP retenue pour marquer les flux LBE est la même valeur dans RENATER-3 et dans GEANT.

Le service LBE est réservé à des projets spécifiques (exemple d'application : projets GRID – grilles de calcul - très « gourmands » en ressources).

Pour les routeurs implémentant le LBE, en cas de congestion, les paquets marqués LBE seront jetés avant tout autre type de paquets, (via du WRED - Weighted Random Early Detection-). Néanmoins, pour éviter une « famine » totale des flux TCP basés marqués avec la classe LBE, il faut réserver une petite partie de la bande passante pour le LBE (environ 1%). En outre, afin de protéger les flux non LBE des flux LBE, il faut placer le trafic LBE dans une file d'attente dédiée.

Pour les routeurs n'implémentant pas le LBE (par exemple des routeurs de concentration d'un réseau de collecte, alors même qu'un site connecté à ce réseau de collecte souhaite néanmoins bénéficier du service LBE), la préconisation minimum est de ne pas modifier la valeur DSCP des paquets marqués LBE (« DSCP transparency »). Donc, un réseau de collecte qui veut implémenter un service LBE avec les préconisations minimum, n'a en fait rien à paramétrer sur ses équipements (vérifier tout même que par défaut, l'équipement ne modifie pas la valeur de DSCP).

Algorithme LBE :

- si un paquet entre dans RENATER avec une valeur DSCP-LBE et ne correspond pas à un flux autorisé alors :
  - ne pas jeter le paquet, mais le re-marquer avec la valeur DSCP-BE et le classer dans la file d'attente BE
- si un paquet entre dans RENATER avec une valeur DSCP-LBE et correspond à un flux autorisé alors il faut ensuite vérifier que ce flux rentre dans le cadre du débit LBE accordé au site :
  - si le flux rentre dans le cadre du débit accordé alors le paquet est accepté
  - si le flux ne rentre pas dans le cadre du débit accordé (dépassement du débit) alors jeter le paquet

2.2 Best Effort (BE)

Le service BE est celui fourni par défaut à l'ensemble des sites RENATER.

Algorithme BE :

- si un paquet entre dans RENATER avec une valeur DSCP-BE et correspond à un flux autorisé (à priori le flux est forcément autorisé) alors il faut ensuite vérifier que ce flux rentre dans le cadre du débit BE accordé au site :
    - si le flux rentre dans le cadre du débit accordé alors le paquet est accepté
    - si le flux ne rentre pas dans le cadre du débit accordé (dépassement du débit) alors le paquet est jeté
- si un paquet entre dans RENATER avec une valeur DSCP qui n'est pas une des valeurs réservées (LBE, BE, BBE, IP Premium, Supervision), alors la valeur DSCP de ce flux ne sera pas modifiée (application du « DSCP transparency ») mais il sera traité exactement comme du BE.

Note : pour le BE, le débit RENATER retenu pour la configuration du « policing » (débit à partir duquel les paquets du site seront jetés), est fixé à la borne supérieure de l'intervalle dans lequel le débit RENATER du site s'inscrit, les intervalles actuellement utilisés pour la tarification étant : [0-1 Mbit/s], [1-2 Mbit/s], [2-4 Mbit/s], [4-10 Mbit/s], [10-20 Mbit/s], [20-40 Mbit/s], [40-100 Mbit/s].

2.3 Better than Best Effort (BBE)

L'idée de ce service est d'offrir un compromis entre le BE et l'IP Premium, le souci principal de l'IP Premium étant que ce service n'est pas généralisable et donc pas accessible à l'ensemble de la communauté RENATER.

Le BBE est ouvert à tous les sites RENATER. Chaque site RENATER pourra marquer ses flux en BBE à hauteur d'environ 5% (pourcentage exact pas encore validé) du débit RENATER souscrit. Comme pour les autres classes de service, une file d'attente dédiée sera utilisée.

Cette classe de service sera confinée au backbone RENATER.

Algorithme BBE :

- si un paquet entre dans RENATER avec une valeur DSCP-BBE et ne correspond pas à un flux autorisé alors :
    - ne pas jeter le paquet, mais le re-marquer avec la valeur DSCP-BE et le mettre dans la file d'attente BE (puisque le BBE est réservé à la communauté RENATER exclusivement, il s'agit donc de trafic entrant dans le réseau RENATER et venant d'éléments extérieurs comme le SFINX, GEANT, les US…)
- si un paquet entre dans RENATER avec une valeur DSCP-BBE et correspond à un flux autorisé alors il faut ensuite vérifier que ce flux rentre dans le cadre du débit BBE accordé au site :
    - si le flux rentre dans le cadre du débit accordé alors le paquet est accepté
    - si le flux ne rentre pas dans le cadre du débit accordé (dépassement du débit) alors :
        - ne pas jeter le paquet (contrairement à l'algorithme appliqué pour le BE) mais le re-marquer avec une valeur DSCP-BBE-OOC et le laisser dans la file d'attente BBE, ce qui permet d'identifier très facilement les paquets BBE en excès et de les laisser quand même passer si finalement les ressources réseaux sont disponibles.

2.4 IP Premium

Les paquets IP Premium sont marqués avec la valeur DSCP-PREMIUM (46 : 101110).

Avec le service IP Premium :

- garantie de la bande passante
- délais ou temps de réponse stables (très faible gigue) pour un chemin donné, ceci quelle que soit la charge des liaisons
- pas de pertes de paquets (ou très négligeables, et en tout cas pas liées à une congestion du backbone)

Le service IP Premium proposé par RENATER est compatible avec celui proposé dans GEANT [4][5], et donc des extensions sont techniquement possibles vers les autres réseaux de la recherche (sous réserve qu'ils aient implémenté de telles classes dans leur réseau). La valeur DSCP retenue pour marquer les flux IP Premium est la même valeur dans

RENATER-3 et dans GEANT. Mais, et cela est d'autant plus vrai pour un service tel que l'IP Premium, il ne suffit pas d'utiliser la même valeur DSCP dans deux réseaux pour que l'interopérabilité du service soit totale. Il convient également de définir des règles d'ingénierie correctes au sein de chaque réseau afin de traiter convenablement de tels flux et garantir une qualité de bout en bout. Les règles de base à respecter sont décrites dans la suite de ce document.

Le service IP Premium est réservé à des projets spécifiques, car n'est pas extensible en terme d'exploitation. En effet, la configuration de ce service nécessite de connaître chaque couple d'adresse IP source/destination, afin que sur les routeurs la bande passante et les temps de réponse (file d'attente prioritaire) soient garantis, et ceci sur le chemin aller comme sur le chemin retour.

L'activation du service IP Premium ne doit pas générer d'effets de bords pour le reste du trafic.

Des vérifications doivent être faites en entrée sur toutes les interfaces des routeurs qui permettent l'interconnexion avec les éléments extérieurs pour vérifier que les flux marqués IP Premium correspondent bien à des flux identifiés/autorisés (la vérification peut, suivant les cas, se faire sur les adresses sources/destinations, sur les numéros d'AS…).

Algorithme IP Premium :

- si un paquet entre dans RENATER avec une valeur DSCP-PREMIUM et ne correspond pas à un flux autorisé alors :
    - ne pas jeter le paquet, mais le re-marquer avec la valeur DSCP-BE et le classer dans la file d'attente BE
- si un paquet entre dans RENATER avec une valeur DSCP-PREMIUM et correspond à un flux autorisé alors il faut ensuite vérifier que ce flux rentre dans le cadre du débit IP Premium accordé au site :
    - si le flux rentre dans le cadre du débit accordé alors le paquet est accepté
    - si le flux ne rentre pas dans le cadre du débit accordé (dépassement du débit) alors :
        - jeter les paquets en excès. Cela se justifie par le fait que contrairement aux autres classes de service, l'IP Premium est traité comme du trafic EF, et non comme de l'AF, et que placer les flux EF dans une file d'attente et y appliquer un traitement basé sur du WRED n'est pas cohérent.

Ces vérifications ne sont nécessaires que sur les interfaces d'accès (pas sur les interfaces de backbone), et une fois qu'elles sont effectuées et validées, alors les paquets sont envoyés dans une file d'attente spécifique à l'IP Premium, et surtout une file d'attente prioritaire. Néanmoins, afin de minimiser la gigue, la profondeur de la file d'attente réservée à l'IP Premium doit être très petite (si des paquets sont stockés loin dans la file d'attente, alors les délais pour acheminer de tels paquets vont être plus longs).

Il est nécessaire de réserver une bande passante minimum (environ 10% de chaque liaison du backbone) pour les flux IP Premium, de façon à protéger ces flux en cas de congestion et garantir un ratio minimum de 1:1 (pour 1Mbit/s IP Premium fourni à un site RENATER, bloquer 1Mbit/s IP Premium sur les liaisons RENATER). Après allocation de cette bande passante minimum, la bande passante restante peut être répartie à peu près de la manière suivante : 60% pour le BBE, 30 % pour le BE, 10% pour le LBE, le pourcentage à réserver pour les flux de supervision étant lui négligeable (moins de 1%).

Il est nécessaire également de limiter globalement les flux IP Premium au niveau de chacune des interfaces d'interconnexion avec les éléments extérieurs : cette limitation peut être fixée à environ 5% du débit de l'interface d'interconnexion. Dans le backbone RENATER-3 le débit standard disponible au niveau de chaque NR (Nœud RENATER) est de 2,5 Gbit/s (une capacité de 10% réservée sur le lien national pour l'IP Premium représente donc 250 Mbit/s) : on part du principe qu'un NR interconnecte entre deux et cinq éléments extérieurs (sur la base d'une interconnexion en Giga-Ethernet pour chaque élément extérieur, soit 50 Mbit/s réservés pour l'IP Premium par interface de raccordement) donc cela constitue entre 2*50 Mbit/s minimum et 5*50 Mbit/s maximum de flux IP Premium par NR, et par précaution on s'aligne sur la borne supérieure de l'intervalle (soit 250 Mbit/s), d'où la valeur des 5%.

Au-delà des seuils définis pour l'IP Premium sur chacune des interfaces avec les éléments extérieurs, les paquets IP Premium seront jetés, conformément aux spécifications mentionnées dans le RFC 2598 [3] traitant de l'EF-PHB (Expedited Forwarding Per Hop Behavior), afin de se prémunir entre autres d'attaques du type « déni » de service.

Note : Toutes ces règles ne sont pas définitives, mais sont des hypothèses de démarrage pour la mise en place des classes de service.

2.5 Classe pour contrôle/supervision du réseau RENATER-3

Les flux générés pour le contrôle et la supervision du réseau RENATER-3 peuvent être traités via une classe de service spécifique. Il s'agit alors donc d'utiliser une valeur DSCP réservée pour les flux de supervision du backbone RENATER-3, flux confinés dans une file d'attente spécifique (gérée en mode AF). Les flux de supervision peuvent être traités sur la base de l'algorithme appliqué pour le BE (ou le BBE), une bande passante minimum de la bande passante restante (après allocation des ressources pour l'IP Premium) ayant de toute façon été configurée sur les liaisons du backbone pour garantir l'acheminement des flux de supervision.

3. Classes de service et services avancés dans RENATER-3

3.1 Les classes de service et IPv6 Unicast

Les critères de classification sont du même type qu'en IPv4. Il est possible de « matcher » simultanément des paquets IPv4 ou IPv6 (champ TC – Traffic Class – futur champ DSCP de l'en-tête du datagramme IPv6) et d'appliquer exactement la même politique de QoS (donc politique globale de QoS pour les flux IP Unicast IPv4 et IPv6). Au niveau des équipements Cisco, des versions d'IOS sont disponibles pour l'implémentation de la QoS IPv6, mais pour l'instant uniquement sur les équipements « edge » (>12.3.2T). Note : chez Cisco, sont considérés comme équipements « core » (gros équipements de cœur de réseau) la gamme des routeurs GSR 12000 et 12400, le reste des équipements étant considéré comme de l'« edge » (équipements de périphérie).

3.2 Les classes de service et IPv4 Multicast

Il est possible de « matcher » les flux Multicast sur des valeurs DSCP, et d'y appliquer une politique de QoS. Au niveau des équipements Cisco, les versions d'IOS sont disponibles sur les équipements « edge » (>12.3.2T) et « core » (version IOS >12.0.26S).

3.3 Les classes de service et MPLS

Dans RENATER-3, le service MPLS-VPN est disponible :

- MPLS-VPN de niveau 2 (via ATOM – Any Transport Over MPLS -)
- MPLS-VPN de niveau 3 (manipulation de tables VRF – VPN Routing/Forwarding -).

Dans le cadre de la fourniture d'un service MPLS-VPN de niveau 2, appliquer une politique de CoS (champ 802.1p) est techniquement possible (mais uniquement si Ethernet/MPLS).

Dans le cadre de la fourniture d'un service MPLS-VPN de niveau 3, appliquer une politique de QoS est techniquement possible. Néanmoins, au niveau des routeurs RENATER-3, l'identification des flux marqués ne se ferait alors pas au niveau du champ DSCP (codé sur 6 bits), mais au niveau du champ MPLS-EXP (champ EXPérimental, codé sur 3 bits).

Dans les deux cas, le minimum est alors de garantir au site/organisme une bande passante associée aux LSP configurés (LSP : Label Switched Path = tunnel MPLS).

Note sur MPLS : La mise en place de solutions MPLS-VPN dans le backbone RENATER-3 se fait au cas par cas, après étude de la demande par le GIP RENATER. Dans l'immédiat, il n'est pas prévu d'étendre les CoS dans MPLS.

Conclusion

Comme présenté dans ce document, certaines classes de service peuvent déjà être prolongées vers le réseau GEANT : c'est le cas des classes LBE et IP Premium. L'interopérabilité et l'interconnexion de réseaux distincts pour parvenir à des services de bout en bout ne doivent pas constituer un obstacle (pour exemple : le réseau actuel RENATER-3 est composé de routeurs Cisco, alors que le backbone GEANT est composé de routeurs Juniper). Avec le déploiement des classes de service dans son backbone, RENATER sera certes en avance sur beaucoup d'autres réseaux de la recherche connectés à GEANT, mais **il est important que l'ensemble des acteurs (réseaux métropolitains, réseaux régionaux et réseaux de sites) entrent dans cette démarche, et sachent proposer de tels services, car il prévu avec le successeur du réseau GEANT le déploiement de services de bout en bout entre réseaux de la recherche**.

Glossaire

AF : Assured Forwarding
ASIC : Application Specific Integrated Circuit
ATOM : Any Transport Over MPLS
BBE : Better than Best Effort
BE : Best Effort
CBWDQ : Class Based Weighted Fair Queing
CoS : Class of Service
DNBAR : Distributed Network Based Application Recognition
DRR : Deficit Round Robin
DSCP : Differenciated Service Code Point
EF : Expedited Forwarding
EF-PHB : Expedited Forwarding Per Hob Behavior
GEANT : réseau européen de la recherche
IOS : Internetwork Operating System
LBE : Less Than Best Effort
LLQ : Low Latency Queuing
LSP : Label Switched Path
MDRR : Modified Deficit Round Robin
MPLS : Multi-Protocol Label Switching
MSFC : Multi Layer Switching network Card
MTU : Maximum Transmit Unit
NBAR : Network Based Application Recognition
NR : Noeud RENATER
PDLM : Packet Description Language Module
PQ : Priority Queuing
QBSS : Qbone Scavenger Service
QoS : Quality of Service
SLA : Service Level Agreement
SFINX : Service for French INternet eXchange
SPQ : Strict Priority Queuing
TC : Traffic Class
VIP : Versatile Interface Processor
Vo/IP : Voix sur IP
VPN : Virtual Private Network
VRF : VPN Routing/Forwarding
WFQ : Weighted Fair Queuing
WRED : Weighted Random Early Detection
WRR : Weighted Round Robin.

Bibliographie

[1] NBAR, PDLM : http://www.cisco.com
[2] QBone Scavenger Service (QBSS) : http://qbone.internet2.edu/qbss/
[3] RFC 2598 – An Expedited Forwarding PHB : http//www.ietf.org/rfc/rfc2598.txt
[4] Specification and Implementation plan for a Premium IP service : http://www.dante.net/geant/public-deliverables/GEA-01-032.pdf
[5] Implementation architecture specification for the Premium IP service : http://www.dante.net/geant/public-deliverables/GEA-01-032av2.pdf