



Compte-rendu des tests pour le déploiement des classes de services sur gw-rap

Description : Ce document présente la validation des mécanismes de qualité de service sur un Juniper M7i à l'aide d'une maquette en vue du déploiement des classes de services sur gw-rap.

Version actuelle :1.4

Date : 11/07/07

Auteur : ld

Version	Dates	Remarques
1.0	27/04/05	Création du document
1.1	15/06/05	Définition QoS Juniper
1.2	07/07/05	Validation des tests
1.4	11/07/07	Avertissement modifs DSCP

Table des matières

Avertissement	3
Introduction	3
1. Schéma général de la maquette	3
2. Classification et mode de traitement :	4
2.1. Description :	4
2.1.1. Modes de classification :	4
2.1.2. Mode de traitement du paquet (Per-Hop Behavior - PHB) :	5
2.2. Validation des mécanismes de classification :	5
2.2.1. Classification par DiffServ (DiffServ Code Point) :	5
2.2.2. Classification par 802.1P :	7
2.2.3. Classification par access-list :	8
3. Control du flux entrant : rate-limiting	9
3.1. Description :	9
3.2. Validation :	9
4. Ordonnancement :	10
4.1. Description :	10
4.1.1. Priorité :	11
4.1.2. Contrôle de transmission : Bande passante et taille de buffer	11
4.1.3. Contrôle de congestion WRED et profile de rejet :	11
4.2. Validation :	11
5. Remarquage et Transport de l'information de CoS :	13
5.1. Définition :	13
5.2. Validation :	13
6. Conclusion :	15

Avertissement

Les valeurs DSCP utilisées dans ce document ont été par la suite modifiées. Merci de vous référer à la page web <http://www.rap.prd.fr/services/qos.php> et au document « Spécifications des classes de service sur RAP » (sur la même page) pour obtenir la bonne mise à jour.

Introduction

Ce document traite des mécanismes à mettre en œuvre pour le déploiement des classes de service sur les équipements d'accès de RAP (Juniper M5 et M7i). Le Juniper M5 est le routeur d'accès actuel de RAP. Le Juniper M7i est le routeur dédié à la mise en place d'un second accès à RENATER pour un partage des charges.

Le traitement de ces classes de service (Less than Best Effort, Best Effort, Better than Best Effort et Premium IP) dans JUNOS comporte de nombreuses possibilités qui ne seront pas toutes abordées dans ce document. Ici, nous avons testé et validé les quatre mécanismes qui permettent de traiter le trafic en fonction de la Classe de service dont les spécifications sont propres à RAP : La classification, la gestion des flux entrant (rate-limiting), l'ordonnancement (prioritisation, shaping) et le marquage.

Ces tests et validations ont été réalisés à l'aide d'une maquette comportant le Juniper M7i. Il a été placé de telle manière à ce que la problématique soit la même que pour le routeur Juniper M5.

1. Schéma général de la maquette

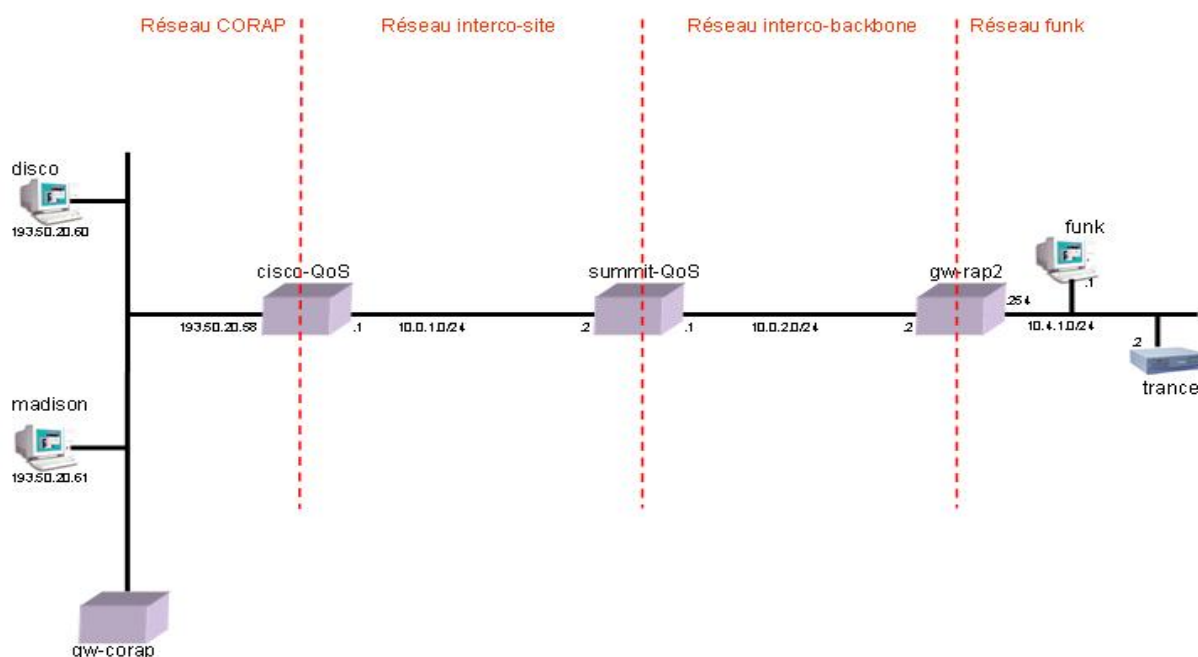


Figure 1 : Schéma général de la maquette QoS

Cette maquette est composée de deux commutateurs :

- cisco-QoS : Cisco Catalyst 3550
- summit-QoS : ExtremeNetworks Summit 48si

D'un routeur :

- gw-rap2 : juniper M7i (interface ge-0/0/0 sur le réseau funk, ge-1/3/0 sur le réseau interco-backbone)

Et de quatre stations :

- madison : PC Linux Fedora 3
- disco : PC Linux Mandrake 10
- funk : PC Linux Fedora 3
- trance : station Sparc Solaris 8.

Les mécanismes de QoS ont été testés et validés sur l'équipement gw-rap2 (Juniper M7i). La commande *show interfaces queue <interface>* est utilisée pour montrer le trafic sortant d'une interface par file matérielle.

La transmission de l'information de QoS est validée par la commande *show <port> qosmonitor* sur le summit-QoS.

Les tests sont effectués à l'aide de l'outil IPerf v.2.0.1 qui génère du trafic marqué (d'un champ DSCP correspondant à une classe de service) et qui donne la perte de paquets et la gigue entre un client et un serveur.

2. Classification et mode de traitement :

2.1. Description :

Un flux entrant dans le routeur est dans un premier temps classifié pour être soumis à un mode de traitement qui peut être assimilé à une classe de service interne au routeur.

2.1.1. Modes de classification :

2.1.1.1. Classification par l'information de QoS : «Behavior Aggregate » BA

Le routeur peut classifier le trafic en fonction du champ DSCP (IPv4, IPv6), du champ IP Precedence, du champ EXP en MPLS ou du champ 802.1p de la trame Ethernet.

2.1.1.2. Classification « multi-champ » : « Multifield Classifier » MF

Le routeur classifie le trafic en fonction d'un ou plusieurs champs du paquet. Cela peut s'effectuer en fonction de l'adresse destination, l'adresse source, le protocole IP, le port source, le port destination et le champ DSCP.

S'il existe un conflit entre la classification en BA et la classification en MF sur un même flux, c'est la classification en MF qui est prise en compte.

2.1.2. Mode de traitement du paquet (Per-Hop Behavior - PHB) :

Un paquet entrant sur le routeur est classifié (méthode BA, MF ou les deux) et intégré dans une « Forwarding Class » (FC). Cette FC permet au routeur de déterminer la file matérielle en sortie, l'ordonnancement et le marquage du paquet.

Chaque FC correspond à une file matérielle en sortie et le routeur Juniper M7i en comporte quatre par interface. Quatre FCs sont donc disponibles (Les FCs par défaut sont : Expedited Forwarding (EF), Assured Forwarding (AF), Best Effort (BE) et Network Control (NC)).

Le routeur associe un bit supplémentaire à chaque paquet intégré dans une FC, la PLP (Packet's Loss Priority), qui peut être en low (bit à 0) ou en high (bit à 1). C'est donc une information, qui, en fonction de sa valeur, permet au routeur de traiter les paquets différemment dans une même FC. C'est notamment le cas lorsque l'algorithme de contrôle de congestion WRED entre en action, il pourra être plus ou moins agressif en fonction de la PLP.

L'association de la FC et du PLP définit le mode de traitement PHB (Per-Hop Behavior) du paquet en transit dans le routeur. Le PHB est codé sur 3 bits, les 2 premiers définissent la FC et le dernier la PLP.

Une classe de service est donc caractérisée par un PHB que le paquet doit suivre dans le routeur. Ce PHB détermine une FC et une PLP. Grâce à la PLP, on peut définir deux types de flux à l'intérieur d'une même file matérielle et donc 2 classes de service. Il est préférable de traiter chaque classe de service dans une file matérielle différente, mais les capacités hardware du Juniper ne le permettent pas.

Remarque : Les mécanismes de classification, d'ordonnancement, de gestion des flux en entrée et de marquage ne sont configurables qu'en fonction de la FC à laquelle appartient un paquet. Si l'on configure deux classes de service dans une même FC, la priorité, la taille de buffer, le volume maximum de trafic de la file matérielle en sortie et le rate-limiting en entrée seront les mêmes pour les deux classes de service. Seule le comportement de l'algorithme RED peut être configuré en fonction de la PLP et donc de la classe de service.

2.2. Validation des mécanismes de classification :

2.2.1. Classification par DiffServ (DiffServ Code Point) :

On valide cette fonctionnalité en reprenant les mêmes valeurs utilisées sur le Backbone de RAP (Cf. document « Compte-rendu des tests pour le déploiement de classes de services sur RAP »). La configuration résultante est la suivante :

Compte-rendu des tests pour le déploiement de classes de services sur gw-rap

```
[class-of-service]
classifiers {
  dscp rap-dscp {
    forwarding-class lbe {
      loss-priority low code-points 001110;
    }
    forwarding-class be+bbe {
      loss-priority low code-points [ 000000 000001 000011 000100 000101 000110
000111 001001 001010 001011 001100 001101 001110 001111 010000 010001 010010
010011 010100 010101 010110 010111 011000 011001 011010 011011 011100 011101
011110 011111 100000 100001 100010 100011 100100 100101 100110 100111 101000
101001 101010 101011 101100 101101 101110 110000 110001 110010 110011 110100
110101 110110 110111 111000 111001 111010 111011 111100 111101 111110 111111 ];
      loss-priority high code-points 100010;
    }
    forwarding-class premium {
      loss-priority high code-points 101110;
    }
  }
}
```

N'importe quelle méthode de classification, ici la méthode diffserv, associe le trafic à une FC. Nous classifions le trafic entrant en fonction de la valeur DSCP du champ ToS : DSCP = 14 dans la FC LBE pour le trafic LBE, DSCP = 46 dans la FC premium pour le trafic Premium IP, DSCP = 34 dans la FC BE+BBE en PLP high pour le trafic BBE, DSCP = « reste » dans la FC BE+BBE en PLP low pour le trafic BE

La quatrième FC est par défaut la network-control, elle est utilisée pour le trafic de contrôle du réseau et de management. La FC be+bbe traite donc le trafic en Best Effort si la PLP est à 0 (low) et en Better than Best Effort si la PLP est à 1 (high). On configure les 4 files matérielles en fonction de la FC :

```
[class-of-service]
forwarding-classes {
  queue 0 lbe;
  queue 1 be+bbe;
  queue 2 premium;
  queue 3 network-control;
}
```

- Activation de la classification sur une interface :

```
[class-of-service]
interfaces {
  ge-1/3/0 {
    unit 0 {
      classifiers {
        dscp rap-dscp-classifier;
      }
    }
  }
}
```

- Résultats :

On marque le trafic entre funk et madison avec iperf. En fonction du DSCP, on constate à l'aide de la commande *show interfaces queue <interface>* que le trafic emprunte la bonne file matérielle en sortie.

2.2.2. Classification par 802.1P :

On valide cette fonctionnalité en reprenant les mêmes valeurs du champ 802.1p utilisées sur le Backbone de RAP :

802.1p	CoS
000	LBE
001	BE
010	non définit
011	BBE
100	non définit
101	Premium
110	non définit
111	NC

Tableau 1: Correspondance priorité 802.1p et CoS

La configuration résultante est la suivante :

```
[class-of-service]
  classifiers {
    ieee-802.1 rap-802.1p{
      forwarding-class lbe {
        loss-priority low code-points 000;
      }
      forwarding-class be+bbe {
        loss-priority low code-points 001;
        loss-priority high code-points 011;
      }
      forwarding-class premium {
        loss-priority high code-points 101;
      }
      forwarding-class network-control {
        loss-priority high code-points 111;
      }
    }
  }
```

- Activation de la classification sur une interface :

```
[class-of-service]
  interfaces {
    ge-1/3/0 {
      unit 0 {
        classifiers {
          ieee-802.1 rap-802.1p;
        }
      }
    }
  }
```

- Résultats :

De même qu'avec la classification par DSCP, on constate bien que le trafic emprunte les files matérielles en fonction du champ 802.1p.

2.2.3. Classification par access-list :

Cette méthode de classification est utilisée pour traiter le trafic en fonction d'un champ de l'en-tête IP autre que le champ ToS. On peut donc classifier le trafic en fonction de l'adresse IP source/destination. Pour cela, on se base sur des préfixes IP :

```
[policy-options]
  prefix-list réseau-funk {
    10.4.1.0/24>;
  }
  prefix-list madison {
    193.50.20.61/32;
  }
```

On effectue la classification dans le contexte firewall. Par exemple on traite tout le trafic à destination de madison en Best Effort :

```
[firewall]
filter funk-in {
  term madison {
    from {
      destination-prefix-list {
        madison;
      }
    }
    then {
      loss-priority low;
      forwarding-class be+bbe;
    }
  }
}

term défaut {
  then accept;
}
}
```


Puis on active ce filtre de classification sur l'interface dans le contexte interface :

```
[interfaces]
ge-0/0/0 {
  unit 0 {
    description funk;
    vlan-id 201;
    family inet {
      filter {
        input funk-in;
      }
    }
    address 10.4.1.254/24
  }
}
```

- Résultats :

On envoie des requêtes iperf quelconques de funk vers madison. De la même manière que pour les classifications par DSCP et 802.1p, on constate par la commande *show interfaces queue ge-0/0/0* que le trafic emprunte bien la file matérielle dédiée à la classe de service LBE.

3. Control du flux entrant : rate-limiting

3.1. Description :

Le rate-limiting permet de contrôler le volume de trafic entrant sur un routeur. Il est possible de limiter la bande passante en entrée en fonction de la Forwarding Class. Si le volume de trafic pour une FC dépasse le volume autorisé, le surplus peut être soit jeté soit traité dans une FC moins prioritaire.

3.2. Validation :

On limite la bande passante du trafic à destination de madison en Premium IP à 2Mbit/s. Pour cela on crée un policer dans le contexte firewall :

```
[firewall]
policer premium {
  if-exceeding {
    bandwidth-limit 2m;
    burst-size-limit 40k;
  }
  then discard;
}
```

Ici, tout le trafic en surplus est rejeté. Pour traiter le surplus dans une autre FC, on peut utiliser « *then forwarding-class <forwarding-class>* ». On active ce policer dans le firewall-filter correspondant :

```
[firewall]
filter funk-in {
  term acces_premium_madison {
    from {
      destination-prefix-list {
        madison;
      }
      forwarding-class premium;
    }
    then policer premium;
  }
  term default {
    then accept;
  }
}
```

Ici, tout trafic à destination de madison et classifié en Premium IP ne peut dépasser un débit de 2Mbit/s. On active ce rate-limiting sur l'interface du réseau funk :

```
[interfaces]
ge-0/0/0 {
  unit 0 {
    description funk;
    vlan-id 201;
    family inet {
      filter {
        input funk-in;
      }
      address 10.4.1.254/24
    }
  }
}
```

- Résultats :

On envoie une requête iperf depuis funk vers madison avec un DSCP à 46 en UDP. On fait varier le débit de la requête pour progressivement atteindre les 2Mbits/s et les dépasser. Les résultats iperf sur madison montre une perte de paquets dès que l'on dépasse un débit de 2 Mbit/s.

4. Ordonnancement :

4.1. Description :

La Forwarding Class établit la priorité, le taux de transfert maximal, la taille de buffer pour une file matérielle en sortie. L'ordonnancement d'un paquet résulte de ces caractéristiques.

4.1.1. Priorité :

La priorité d'une file matérielle détermine l'ordre dans lequel le routeur transmet le trafic en sortie. Une priorité est configurée pour une forwarding class et donc une file matérielle en sortie. Deux classes de services dans une même FC et transitant en sortie dans une même file matérielle auront donc la même priorité. Il existe 3 degrés de priorité : low, high, strict-high. High est prioritaire à low. La particularité de strict-high est qu'il peut prendre toute la bande passante en restreignant les files matérielles en low, toutefois, le trafic dans les files matérielles en priorité high est protégé du moment qu'il ne dépasse le maximum de bande passante alloué.

4.1.2. Contrôle de transmission : Bande passante et taille de buffer

Il s'agit de configurer la bande passante maximale par file matérielle. Il est possible de traiter le trafic en surplus dans une file matérielle moins prioritaire. La taille du buffer peut être modifiée pour mieux gérer les forts afflux sporadiques possibles dans une classe de trafic sensible à la gigue et au délai.

4.1.3. Contrôle de congestion WRED et profile de rejet :

Pour prévenir le phénomène de congestion, l'algorithme WRED (Random Early Detection) régule le trafic TCP. Un profile de rejet du trafic dans une file matérielle est configurable en fonction de son taux de remplissage, de la PLP et de sa nature : TCP, non-TCP. On peut donc configurer, en fonction du taux de remplissage de la file matérielle 4 profile de rejet (« drop profile »).

4.2. Validation :

Le trafic Best Effort est acheminé dans la même file matérielle que le trafic Better than Best Effort. On configure un profile de rejet plus agressif pour le trafic BE dans la FC be+bbe, ceci afin de protéger le trafic BBE. Mais le degré d'agressivité reste modéré car le trafic BE représente la majorité du trafic. On configure le profile de rejet dans le sous contexte *drop-profiles* :

```
[class-of-service]
drop-profiles {
  drop-be {
    fill-level 70 drop-probability 0;
    fill-level 75 drop-probability 25;
    fill-level 80 drop-probability 50;
    fill-level 85 drop-probability 75;
    fill-level 90 drop-probability 100;
  }
}
```

Compte-rendu des tests pour le déploiement de classes de services sur gw-rap

Si la file matérielle de la FC be+bbe est rempli à 90%, 100% du trafic BE sera rejeté. Le trafic en BBE ne le sera pas avant que la file matérielle soit rempli à 100%. On réserve ainsi 10% de la bande passante de la FC be+bbe. Cette fonctionnalité est gérée par l'algorithme WRED qui régule notamment le trafic TCP en fonction du remplissage de la file matérielle.

On configure maintenant la priorité, le taux de transfert et la taille de buffer par file matérielle (par FC) dans le sous contexte *schedulers* :

```
[class-of-service]
schedulers {

    sch-lbe {
        transmit-rate percent 0;
        buffer-size percent 0;
        priority low;
    }
}
```

Ici, on configure l'ordonnancement du trafic LBE pour qu'il ne soit traité que si les autres FCs n'utilisent pas la totalité des ressources de l'interface de sortie. On accepte donc que le trafic LBE subisse des pertes en cas de forts besoins de ressources par les autres FCs. De plus, on applique une priorité basse au trafic LBE.

```
sch-be+bbe {
    transmit-rate percent 90;
    buffer-size percent 90;
    priority high;
    drop-profile-map loss-priority low protocol any drop-profile drop-be;
}
```

La FC be+bbe peut utiliser jusqu'à 90% de la bande passante totale de l'interface en sortie. On configure une priorité high afin de le rendre prioritaire au trafic LBE. Enfin, on applique le profile de rejet configuré auparavant pour le trafic TCP et non-TCP de la FC en low (BE).

```
sch-premium {
    transmit-rate percent 5 exact;
    buffer-size percent 5;
    priority high;
}
```

Le trafic Premium IP peut utiliser 5% de la bande passante totale et est strictement prioritaire, ce qui lui permet d'utiliser toute les ressources sauf celles réservées au FC en priorité high, néanmoins, le rate-limiting et le shaping configurés limitent la bande passante en Premium IP. On considère que ce type de trafic est minoritaire et est réglementé selon un SLA. Cela permet de configurer un taux de transfert en fonction des besoins sans que cela affecte le reste du trafic tout en ayant une priorité haute.

```
sch-network-ctl {
    buffer-size percent 5;
    priority strict-high;
}
}
```

Le trafic network-control doit être protégé. On lui réserve 5% de la bande passante totale avec une priorité high.

On établit maintenant la relation entre une FC et un scheduler dans le sous contexte *scheduler-maps* :

```
[class-of-service]
scheduler-maps {
  MAP-rap {
    forwarding-class lbe scheduler sch-lbe;
    forwarding-class be+bbe scheduler sch-be+bbe;
    forwarding-class premium scheduler sch-premium;
    forwarding-class network-control scheduler sch-network-ctl;
  }
}
```

Cela permet d'activer l'ordonnancement sur une interface dans le sous contexte *interfaces*:

```
[class-of-service]
interfaces {
  ge-1/3/0 {
    scheduler-map MAP-rap;
  }
}
```

5. Remarquage et Transport de l'information de CoS :

5.1. Définition :

Le marquage du trafic permet d'utiliser dans son réseau ses propres informations de CoS. Le trafic entrant sur le routeur peut être classifié par DSCP puis marqué avec un champ 802.1P correspondant à la même classe de service dont le traitement sera prolongé sur le réseau.

5.2. Validation :

On configure le marquage du trafic selon les correspondances du tableau 1. Ce marquage a lieu sur l'interface ge-1/3/0 en face du commutateur summit-QoS qui classifie de son côté le trafic en fonction du champ 802.1p.

Compte-rendu des tests pour le déploiement de classes de services sur gw-rap

```
[class-of-service]
rewrite-rules {
  ieee-802.1 rap-802.1p {
    forwarding-class lbe {
      loss-priority low code-point 000;
      loss-priority high code-point 000;
    }
    forwarding-class be+bbe {
      loss-priority low code-point 001;
      loss-priority high code-point 011;
    }
    forwarding-class premium {
      loss-priority low code-point 101;
      loss-priority high code-point 101;
    }
    forwarding-class network-control {
      loss-priority low code-point 111;
      loss-priority high code-point 111;
    }
  }
}
```

On active le marquage sur l'interface :

```
[class-of-service]
interfaces {
  ge-1/3/0 {
    unit 0 {
      rewrite-rules {
        ieee-802.1 rap-802.1p;
      }
    }
  }
}
```

- Résultats :

Cette configuration ne donne pas les bons résultats. On vérifie la valeur du champ 802.1p sur le summit-QoS (commande `show <port> qosmonitor`). Le trafic remarqué suit les correspondances suivantes :

FC	PLP	PHB	Remarquage 802.1p	summit 802.1p	CoS RAP
lbe	0	000	000	000	LBE
lbe	1	001	000	001	BE
be+bbe	0	010	001	010	non définit
be+bbe	1	011	011	011	BBE
premium	0	100	101	100	non définit
premium	1	101	101	101	premium
nc	0	110	111	110	non définit
nc	1	111	111	111	nc

La configuration du marquage 802.1p présente des contraintes. Pour les routeurs de série M et T, le marquage du champ 802.1p est codé sur 3 bits, le bit 0 est fixé par la valeur de la PLP, les bits 1 et 2 sont fixés par la file matérielle (Forwarding Class) utilisée. Par exemple, la file matérielle codée 00 (queue 0) dans le Juniper, ne peut marquer son trafic qu'avec un champ 802.1p commençant par 00 avec une PLP à 0 ou 1 soit un champ 802.1p à 001 ou 010, voici donc le marquage 802.1p possible sur le Juniper M7i :

CoS	FC	PLP	802.1p
LBE	00	0	000
-	00	1	001
BE	01	0	010
BBE	01	1	011
-	10	0	100
Premium	10	1	101
-	11	0	110
NC	11	1	111

6. Conclusion :

Les différents tests effectués permettent de valider un déploiement des CoS sur le Juniper M7i. Cependant, certaines caractéristiques techniques limitent ce déploiement.

Chaque interface ne dispose que de quatre files matérielles en sorties, or, il est conseillé de réserver une forwarding-class pour le trafic de contrôle du réseau. Il reste donc trois FCs pour les quatre classes de service que RAP a déployé sur son backbone.

Traiter deux classes de service dans une file matérielle n'est pas une solution confortable mais reste acceptable si l'on admet que l'on peut traiter deux classes de service avec le même ordonnancement. Pour cela on choisit celles dont les caractéristiques de traitement se rapprochent le plus, BBE et BE. Ces deux classes de service ne présentent pas les mêmes contraintes, mais les classes de services LBE et Premium IP ont des caractéristiques trop spécifiques pour être traitées de la même manière.

- Classification :

Le trafic entrant sur RAP sera classifié par la valeur DSCP du champ ToS, selon les valeurs déjà définis dans le document « Compte-rendu des tests pour le déploiement de classes de services sur RAP ». Le trafic provenant du backbone sera lui classifié grâce au champ 802.1p, méthode de classification utilisée sur le backbone de RAP. Enfin, pour les sites n'ayant pas de niveau d'accès QoS, un filtre sera configuré pour que leur trafic en entrée sur RAP soit traité en BE.

- Contrôle de flux :

Le control de flux en entrée sur RAP sera configuré pour le trafic Premium IP pour les sites ayant un niveau d'accès Premium. Ainsi, en fonction du SLA, un *policer* sera mis en place afin de limiter la bande passante du trafic Premium IP du site. Ce processus n'est pas réalisable pour le trafic BBE puisque ce dernier est traité dans la même file matérielle que le trafic en BE.

- Ordonnancement :

Le trafic en BE et en BBE disposera de 90% de bande passante totale des interfaces en sortie avec une priorité haute (high). L'utilisation d'une PLP différente permet de traiter de manière plus agressive le trafic en BE dans la même file matérielle, cela protège un minimum le trafic BBE en cas d'afflux trop important de trafic en BE. Le trafic Premium IP et NC disposeront chacun de 5% de la bande passante totale des interfaces en sortie avec respectivement une priorité strictement haute (strict-high) et haute (high). Le trafic en Premium IP faisant l'objet de SLAs auprès des sites, cette valeur pourra être modifiée. Enfin, le trafic en LBE sera traité avec une priorité basse (low) et ne disposera que de la bande passante non utilisée par les autres classes de service.

- Remarquage et transport de l'information de QoS :

La valeur du champ 802.1p que le routeur peut remarquer en sortie est fixée par la FC et la PLP. Il n'est donc pas possible de remarquer le champ 802.1p du trafic BE à 001. Le trafic Best Effort provenant du Juniper est marqué d'un champ 802.1p à 010, une petite modification sur le backbone de RAP permet de contourner ce problème.

La validation de ces mécanismes de QoS n'est pas valable pour le Juniper M5 car il ne dispose pas de carte Enhanced-FPC.