

# Qualité de service

Avez-vous déjà pris part à un appel téléphonique longue distance nécessitant une connexion par satellite? La conversation peut être interrompue par de brèves pauses, mais perceptibles, à des intervalles irréguliers. Ces pauses correspondent au délai, appelé latence, entre l'envoi et la réception des paquets sur le réseau. Certains trafics réseau, comme la voix et la vidéo, ne peuvent tolérer des temps de latence élevés. La qualité de service (QoS) est une fonctionnalité qui vous permet de donner la priorité au trafic critique, d'empêcher la saturation de la bande passante et de gérer les goulots d'étranglement du réseau pour éviter les pertes de paquets.

Les rubriques suivantes décrivent comment appliquer des politiques QoS.


- À propos de QoS, à la page 1
- Lignes directrices pour la QoS, à la page 3
- Configurer QoS, à la page 4
- Surveiller la QoS, à la page 8
- Historique QoS, à la page 10


# À propos de QoS

Vous devez considérer que dans un environnement réseau en constante évolution, la QoS n'est pas un déploiement unique, mais un élément continu et essentiel de la conception du réseau.

Cette section décrit les caractéristiques de QoS disponibles sur l'ASA.

## Fonctionnalités QoS prises en charge

L'ASA prend en charge les fonctionnalités QoS suivantes :

- Régulation : pour éviter que le trafic classé ne monopolise la bande passante du réseau, vous pouvez limiter la bande passante maximale utilisée par classe. Consultez Police, à la page 2 pour de plus amples renseignements.
- Mise en file prioritaire : pour le trafic critique qui ne peut pas tolérer la latence, comme la voix sur IP (VoIP), vous pouvez identifier le trafic pour la mise en file d'attente à faible latence (LLQ) afin qu'il soit toujours transmis avant les autres trafics. Consultez File d'attente prioritaire, à la page 2.

# Qu'est-ce qu'un compartiment de jetons?

Un compartiment à jetons est utilisé pour gérer un mécanisme qui régule les données dans un flux, par exemple un contrôleur de trafic. Un compartiment de jetons n'a pas de politique de suppression ou de priorité. u contraire, un compartiment à jetons rejette les jetons et laisse au flux le soin de gérer sa file d'attente de transmission si le flux surcharge le régulateur.

Un compartiment à jetons est une définition formelle d'un débit de transfert. Il comporte trois composants : une taille de rafale, un débit moyen et un intervalle de temps. Bien que le débit moyen soit généralement représenté sous forme de bits par seconde, deux valeurs peuvent être dérivées de la troisième par la relation indiquée comme suit :

débit moyen = taille de rafale/intervalle de temps

Voici quelques définitions de ces termes :

- Débit moyen : également appelé débit d'information validé (CIR), il spécifie la quantité de données qui peut être envoyée ou transmise par unité de temps en moyenne.
- Taille de la rafale : également appelée taille de rafale validée (Bc), elle spécifie en octets par rafale la quantité du trafic qui peut être envoyée dans une unité de temps donnée pour ne pas créer de problèmes de planification.
- Intervalle de temps : également appelé intervalle de mesure, il spécifie le quantum de temps en secondes par rafale.

Dans la métaphore du compartiment à jetons, les jetons sont placés dans le compartiment à un certain débit. Le compartiment lui-même a une capacité spécifiée. Si le compartiment se remplit à sa capacité, les jetons nouvellement entrants sont rejetés. Chaque jeton est l'autorisation pour la source d'envoyer un certain nombre de bits dans le réseau. Pour envoyer un paquet, le régulateur doit supprimer du compartiment un nombre de jetons égal en représentation à la taille du paquet.

S'il n'y a pas assez de jetons dans le compartiment pour envoyer un paquet, le paquet attend jusqu'à ce que le paquet soit rejeté ou marqué. Si le compartiment est déjà plein de jetons, les jetons entrants débordent et ne sont pas disponibles pour les futurs paquets. Ainsi, à tout moment, la plus grande rafale qu'une source peut envoyer dans le réseau est à peu près proportionnelle à la taille du compartiment.

# Police

La régulation est un moyen de s'assurer que le trafic ne dépasse pas le débit maximal (en bits/seconde) que vous avez configuré, garantissant ainsi qu'aucune classe de trafic ne peut prendre le contrôle de toute la ressource. Lorsque le trafic dépasse le débit maximal, l'ASA abandonne le trafic excédentaire. La régulation définit également la plus grande rafale de trafic autorisée.

# File d'attente prioritaire

La mise en file d'attente prioritaire LLQ vous permet de hiérarchiser certains flux de trafic (comme le trafic sensible à la latence comme la voix et la vidéo) avant d'autres trafics. La mise en file d'attente prioritaire utilise une file d'attente de priorité LLQ sur une interface (voir Configurer la file d'attente prioritaire pour une interface, à la page 6), tandis que tout autre trafic passe dans la file d'attente « au mieux ». Comme les files d'attente ne sont pas de capacité infinie, elles peuvent atteindre leur capacité maximale et déborder. Lorsqu'une file d'attente est remplie, tous les paquets supplémentaires ne peuvent pas entrer dans la file d'attente et sont abandonnés. C'est ce qu'on appelle *abandon de file d'attente*. Pour éviter que la file d'attente

ne se remplisse, vous pouvez augmenter la taille du tampon de file d'attente. Vous pouvez également affiner le nombre maximal de paquets autorisés dans la file d'attente de transmission. Ces options vous permettent de contrôler la latence et la robustesse de la mise en file d'attente prioritaire. Les paquets dans la file d'attente LLQ sont toujours transmis avant les paquets dans la file d'attente au mieux.

## Comment les fonctionnalités QoS interagissent

Vous pouvez configurer chacune des fonctionnalités QoS seule si vous le souhaitez pour l'ASA. Souvent, cependant, vous configurez plusieurs fonctionnalités QoS sur l'ASA afin de pouvoir hiérarchiser certains trafics, par exemple, et empêcher d'autres trafics de provoquer des problèmes de bande passante. Vous pouvez configurer :

Mise en file prioritaire (pour un trafic spécifique) + régulation (pour le reste du trafic).

Vous ne pouvez pas configurer la mise en file d'attente prioritaire et la régulation pour le même ensemble de trafic.

## Préservation DSCP (DiffServ)

Les marquages DSCP (DiffServ) sont conservés sur tout le trafic passant par l'ASA. L'ASA ne marque/remarque pas localement le trafic classé. Par exemple, vous pourriez désactiver les bits DSCP de transfert accéléré (EF) de chaque paquet pour déterminer s'il nécessite un traitement « priorité » et faire en sorte que l'ASA dirige ces paquets vers le LLQ.

# Lignes directrices pour la QoS

### Directives relatives au mode contextuel

Pris en charge en mode contexte unique uniquement Ne prend pas en charge le mode contexte multiple.

### Directives sur le mode pare-feu

Pris en charge uniquement en mode de pare-feu routé. Le mode pare-feu transparent n'est pas pris en charge.

### Directives IPv6

Ne prend pas en charge IPv6.

### Directives et limites additionnelles

- La QoS est appliquée de manière unidirectionnelle; seul le trafic qui entre (ou en sort, selon la fonctionnalité QoS) de l'interface à laquelle vous appliquez la liste des politiques est touché.
- Pour le trafic prioritaire, vous ne pouvez pas utiliser la carte de trafic **par défaut**.
- Pour la mise en file d'attente prioritaire, la file d'attente prioritaire doit être configurée pour une interface physique.
- Pour la régulation, le trafic vers le boîtier n'est pas pris en charge.
- Pour la régulation, le trafic vers et depuis un tunnel VPN contourne la régulation d'interface.

- Pour la régulation, lorsque vous faites correspondre une carte de trafic de groupe de tunnels, seule la régulation sortante est prise en charge.

# Configurer QoS

Utilisez la séquence suivante pour mettre en œuvre la QoS sur l'ASA.

**Procédure**

**Étape 1** Déterminer les limites de la file d'attente et de l'anneau TX pour une file d'attente prioritaire, à la page 4.

**Étape 2** Configurer la file d'attente prioritaire pour une interface, à la page 6.

**Étape 3** Configurer une règle de service pour la mise en file d'attente prioritaire et la régulation, à la page 7.

# Déterminer les limites de la file d'attente et de l'anneau TX pour une file d'attente prioritaire

Utilisez les feuilles de travail suivantes pour déterminer les limites de la file d'attente prioritaire et de l'anneau TX.

## Feuille de travail sur la limite de la file d'attente

La feuille de travail suivante montre comment calculer la taille de la file d'attente prioritaire. Comme les files d'attente ne sont pas de taille infinie, elles peuvent se remplir et déborder. Lorsqu'une file d'attente est remplie, tous les paquets supplémentaires ne peuvent pas entrer dans la file d'attente et sont abandonnés (appelés *rejet de file d'attente*). Pour éviter que la file d'attente ne soit remplie, vous pouvez ajuster la taille du tampon de file d'attente en fonction de Configurer la file d'attente prioritaire pour une interface, à la page 6.

Conseils sur la feuille de travail :

- Bande passante sortante : par exemple, une connexion DSL peut avoir une vitesse de liaison ascendante de 768 kbit/s. Vérifiez auprès de votre fournisseur.
- Taille moyenne des paquets : déterminez cette valeur à partir d'un codec ou d'une taille d'échantillonnage. Par exemple, pour le protocole VoIP sur VPN, vous pouvez utiliser 160 octets. Nous vous recommandons d'utiliser 256 octets si vous ne savez pas quelle taille utiliser.
- Délai : le délai dépend de votre application. Par exemple, le délai maximal conseillé pour le protocole VoIP est de 200 ms. Nous vous recommandons d'utiliser 500 ms si vous ne savez pas quel délai utiliser.

***Tableau 1 : Feuille de travail sur la limite de la file d'attente***

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **1** | _____<br>*Bande passante sortante (Mbit/s ou kbit/s)* | Mbit/s | x | **125** | = | _____<br>*Nombre d'octets/ms* | | |
| | | kbit/s | x | **0,125** | = | _____<br>*Nombre d'octets/ms* | | |
| **2** | _____<br>*Nombre d'octets/ms de l'étape 1* | | ÷ | _____<br>*Taille moyenne des paquets (octets)* | x | _____<br>*Retard (ms)* | = | _____<br>*Limite de la file d'attente (Nombre de paquets)* |

## Feuille de calcul des limites de l'anneau TX

La feuille de travail suivante montre comment calculer la limite de la boucle TX. Cette limite détermine le nombre maximal de paquets autorisés dans le pilote de transmission Ethernet avant que le pilote ne repousse les files d'attente sur l'interface pour leur permettre de mettre les paquets en mémoire tampon jusqu'à ce que la congestion se résorbe. Ce paramètre garantit que l'anneau de transmission matériel impose une quantité limitée de latence supplémentaire pour un paquet de priorité élevée.

Conseils sur la feuille de travail :

- Bande passante sortante : par exemple, le protocole DSL peut avoir une vitesse de liaison ascendante de 768 kbit/s. Vérifiez auprès de votre fournisseur.
- Taille maximale du paquet : en général, la taille maximale est de 1 538 octets ou de 1 542 octets pour Ethernet étiqueté. Si vous autorisez les trames étendues (si elles sont prises en charge pour votre plateforme), la taille des paquets pourrait être plus grande.
- Délai : le délai dépend de votre application. Par exemple, pour contrôler la gigue pour le VoIP, vous devez utiliser 20 ms.

***Tableau 2 : Feuille de travail sur la limite de la boucle TX***

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **1** | _____<br>*Bande passante sortante (Mbit/s ou kbit/s)* | Mbit/s | x | **125** | = | _____<br>*Nombre d'octets/ms* | | |
| | | kbit/s | x | **0,125** | = | _____<br>*Nombre d'octets/ms* | | |
| **2** | _____<br>*Nombre d'octets/ms de l'étape 1* | | ÷ | _____<br>*Taille maximale des paquets (octets)* | x | _____<br>*Retard (ms)* | = | _____<br>*Limite de l'anneau TX (nombre de paquets)* |

# Configurer la file d'attente prioritaire pour une interface

Si vous activez la mise en file d'attente prioritaire pour le trafic sur une interface physique, vous devez également créer la file d'attente prioritaire sur chaque interface. Chaque interface physique utilise deux files d'attente : l'une pour le trafic prioritaire et l'autre pour tout autre trafic. Pour l'autre trafic, vous pouvez éventuellement configurer la régulation.

### Procédure

**Étape 1** Choisissez **Configuration** > **Device Management (Gestion des périphériques)** > **Advanced** > **(Avancé)** > **Priority Queue** (File d'attente prioritaire), puis cliquez sur **Add** (Ajouter).

**Étape 2** Configurez les options suivantes :

- **Interface** : nom de l'interface physique sur laquelle vous souhaitez activer la file d'attente prioritaire ou, pour l'ASASM, le nom de l'interface VLAN.
- **Queue Limit** (Limite de file d'attente) : nombre moyen de paquets de 256 octets que l'interface précisée peut transmettre dans un intervalle de 500 ms. La plage est comprise entre 0 et 2048, et 2048 est la valeur par défaut.

  Un paquet qui reste plus de 500 ms dans un nœud de réseau peut déclencher un délai d'expiration dans l'application de bout en bout. Un tel paquet peut être rejeté dans chaque nœud de réseau.

  Comme les files d'attente ne sont pas de taille infinie, elles peuvent se remplir et déborder. Lorsqu'une file d'attente est remplie, tous les paquets supplémentaires ne peuvent pas entrer dans la file d'attente et sont abandonnés (appelés *« rejet de file d'attente »*). Pour éviter que la file d'attente ne soit remplie, vous pouvez utiliser cette option pour augmenter la taille du tampon de file d'attente.

  La limite supérieure de la plage de valeurs pour cette option est déterminée dynamiquement au moment de l'exécution. Les facteurs clés sont la mémoire nécessaire pour prendre en charge les files d'attente et la mémoire disponible sur le périphérique.

  La valeur Queue Limit (Limite de file d'attente) que vous spécifiez affecte à la fois la file d'attente à faible latence de priorité plus élevée et la file d'attente d'effort optimal.
- **Transmission Ring Limit** (Limite de l'anneau de transmission) : profondeur des files d'attente prioritaires, soit le nombre maximal de paquets de 1 550 octets que l'interface précisée peut transmettre dans un intervalle de 10 ms. La plage est comprise entre 3 et 511, et 511 est la valeur par défaut.

  Ce paramètre garantit que l'anneau de transmission matériel n'imposera pas plus de 10 ms de latence supplémentaire pour un paquet de priorité élevée.

  Cette option définit le nombre maximal de paquets à faible latence ou de priorité normale autorisés dans le pilote de transmission Ethernet avant que le pilote ne renvoie les paquets vers les files d'attente de l'interface pour leur mise en mémoire tampon jusqu'à ce que la congestion se résorbe.

  La limite supérieure de la plage de valeurs est déterminée dynamiquement au moment de l'exécution. Les facteurs clés sont la mémoire nécessaire pour prendre en charge les files d'attente et la mémoire disponible sur le périphérique.

  La valeur Transmission Ring Limit (Limite de l'anneau de transmission) que vous spécifiez affecte à la fois la file d'attente à faible latence de priorité plus élevée et la file d'attente d'effort optimal.

**Étape 3** Cliquez sur **OK**, puis sur **Apply**(appliquer).

# Configurer une règle de service pour la mise en file d'attente prioritaire et la régulation

Vous pouvez configurer la mise en file d'attente prioritaire et la régulation pour différentes cartes de trafic dans la même liste des politiques. Consultez Comment les fonctionnalités QoS interagissent, à la page 3 pour obtenir des renseignements sur les configurations QoS valides.

### Avant de commencer

- Vous ne pouvez pas utiliser la carte de trafic **class-default** pour le trafic prioritaire.
- Pour la régulation, le trafic vers le boîtier n'est pas pris en charge.
- Pour la régulation, le trafic vers et depuis un tunnel VPN contourne la régulation d'interface.
- Pour la régulation, lorsque vous faites correspondre une carte de trafic de groupe de tunnels, seule la régulation sortante est prise en charge.
- Pour le trafic prioritaire, identifiez uniquement le trafic sensible à la latence.
- Pour réguler le trafic, vous pouvez choisir de réguler tout autre trafic ou de limiter le trafic à certains types.

### Procédure

**Étape 1** Choisissez **Configuration** > **Firewall (Pare-feu)** > **Service Policy Rules (Règles de politique de service)**, et ouvrez une règle.

Vous pouvez configurer QoS dans le cadre d'une nouvelle règle de politique de service ou modifier une politique de service existante.

**Étape 2** Poursuivez dans l'assistant jusqu'à la page Rules (Règles), en sélectionnant l'interface (ou globale) et les critères de correspondance du trafic en cours de route.

Pour réguler le trafic, vous pouvez choisir de réguler tout le trafic que vous ne priorisez pas, ou vous pouvez limiter le trafic à certains types.

**Astuces**
Si vous utilisez une liste de contrôle d'accès pour la correspondance du trafic, la régulation est appliquée dans la direction spécifiée dans la liste de contrôle d'accès uniquement. C'est-à-dire que le trafic allant de la source à la destination est régulé, mais pas l'inverse.

**Étape 3** Dans la boîte de dialogue Rule Actions (Actions de règle), cliquez sur l'onglet **QoS** (QoS).

**Étape 4** .Sélectionnez **Enable priority for this flow** (Activer la priorité pour ce flux).

Si cette règle de politique de service s'applique à une interface individuelle, ASDM crée automatiquement la file d'attente prioritaire pour l'interface (Configuration > Device Management (Gestion des périphériques) > Advanced (Avancé) > Priority Queue (File d'attente prioritaire) ; pour plus d'informations, consultez Configurer

la file d'attente prioritaire pour une interface, à la page 6). Si cette règle s'applique à la politique globale, vous devez ajouter manuellement la file d'attente prioritaire à une ou plusieurs interfaces *avant* de configurer la règle de politique de service.

**Étape 5** Sélectionnez **Enable policing** (Activer la régulation), puis cochez les cases **Input policing** (Régulation en entrée) ou **Output policing** (Régulation en sortie) (ou les deux) pour activer le type de régulation du trafic spécifié. Pour chaque type de régulation de trafic, configurez les options suivantes :

- **conform-rate** : limite de débit pour cette classe de trafic, de 8 000 à 2 000 000 000 bits par seconde. Par exemple, pour limiter le trafic à 5 Mbit/s, saisissez 5 000 000.
- **Conform Action** (Action conforme) : action à prendre lorsque le trafic est inférieur au débit de régulation et à la taille de rafale. Vous pouvez abandonner ou transmettre le trafic. La valeur par défaut est de transmettre le trafic.
- **Exceed Action** (Action de dépassement) : action à prendre lorsque le trafic dépasse le débit de régulation et la taille de rafale. Vous pouvez abandonner ou transmettre des paquets qui dépassent le débit de régulation et la taille de rafale. La valeur par défaut est d'abandonner les paquets en excédent.
- **Burst Rate** (Débit de rafale) : nombre maximal d'octets instantanés autorisés dans une rafale soutenue avant de passer à la valeur de débit conforme, entre 1 000 et 512 000 000 octets. Calculez la taille de rafale comme 1/32 du débit conforme en octets. Par exemple, la taille de rafale pour un débit de 5 Mbit/s serait de 156 250. La valeur par défaut est de 1 500, mais le système peut recalculer toute valeur que vous saisissez au besoin.

**Étape 6** Cliquez sur **Finish** (Terminer), puis sur **Apply** (Appliquer).

# Surveiller la QoS

Les rubriques suivantes expliquent comment surveiller la QoS.

Pour surveiller la QoS dans ASDM, vous pouvez entrer des commandes dans l'outil de l'interface de ligne de commande.

## Statistiques de régulation QoS

Pour afficher les statistiques QoS pour la régulation du trafic, utilisez la commande **show service-policy police**.

```
hostname# show service-policy police

Global policy:
 Service-policy: global_fw_policy

Interface outside:
 Service-policy: qos
  Class-map: browse
   police Interface outside:
    cir 56000 bps, bc 10500 bytes
    conformed 10065 packets, 12621510 bytes; actions: transmit
    exceeded 499 packets, 625146 bytes; actions: drop
    conformed 5600 bps, exceed 5016 bps
  Class-map: cmap2
```

```
   police Interface outside:
    cir 200000 bps, bc 37500 bytes
    conformed 17179 packets, 20614800 bytes; actions: transmit
    exceeded 617 packets, 770718 bytes; actions: drop
    conformed 198785 bps, exceed 2303 bps
```

# Statistiques de priorité QoS

Pour afficher les statistiques des politiques de service qui mettent en œuvre la commande **priority**, utilisez la commande **show service-policy priority**.

```
hostname# show service-policy priority
Global policy:
 Service-policy: global_fw_policy
Interface outside:
 Service-policy: qos
  Class-map: TG1-voice
   Priority:
    Interface outside: aggregate drop 0, aggregate transmit 9383
```

« Aggregate drop » (abandon agrégé) désigne l'abandon agrégé dans cette interface; « Aggregate transmit » (transmission agrégée) désigne le nombre agrégé de paquets transmis dans cette interface.

# Statistiques de file d'attente prioritaire QoS

Pour afficher les statistiques de file d'attente prioritaire pour une interface, utilisez la commande **show priority-queue statistics**. Les résultats affichent les statistiques de la file d'attente best-effort (BE) et de la file d'attente à faible latence (LLQ). L'exemple suivant montre l'utilisation de la commande **show priority-queue statistics** pour l'interface nommée test.

```
hostname# show priority-queue statistics test

Priority-Queue Statistics interface test

Queue Type         = BE
Packets Dropped    = 0
Packets Transmit   = 0
Packets Enqueued   = 0
Current Q Length   = 0
Max Q Length       = 0

Queue Type         = LLQ
Packets Dropped    = 0
Packets Transmit   = 0
Packets Enqueued   = 0
Current Q Length   = 0
Max Q Length       = 0
hostname#
```

Dans ce rapport statistique :

- « Paquets abandonnés » désigne le nombre total de paquets qui ont été abandonnés dans cette file d'attente.
- « Paquets transmis » désigne le nombre total de paquets qui ont été transmis dans cette file d'attente.

- « Paquets mis en file d'attente » désigne le nombre total de paquets qui ont été mis en file d'attente dans cette file d'attente.
- La « Longueur actuelle de la file d'attente » désigne la profondeur actuelle de cette file d'attente.
- La « Longueur maximale de la file d'attente » désigne la profondeur maximale jamais enregistrée dans cette file d'attente.

# Historique QoS

| Nom de la caractéristique | Versions de plateforme | Description |
|---|---|---|
| Mise en file et régulation prioritaires | 7.0(1) | Nous avons introduit la mise en file d'attente prioritaire et la régulation QoS.<br>Nous avons introduit les écrans suivants :<br>Configuration > Device Management (Gestion des périphériques) > Advanced (Avancé) > Priority Queue (File d'attente prioritaire) Configuration (Configuration) > Firewall (Pare-feu) > Service Policy Rules (Règles de politique de service) |
| Mise en forme et mise en file d'attente prioritaire hiérarchique | 7.2(4)/8.0(4) | Nous avons lancé la mise en forme QoS et la mise en file prioritaire hiérarchique.<br>Nous avons modifié l'écran suivant : Configuration > Firewall (Pare-feu) > Service Policy Rules (Règles de politique de service). |
| Prise en charge de dix Gigabit Ethernet pour une file d'attente de priorité standard sur l'ASA 5585-X | 8.2(3)/8.4(1) | Nous avons ajouté la prise en charge d'une file d'attente prioritaire standard sur les interfaces Ten Gigabit Ethernet pour l'ASA 5585-X. |

À propos de la traduction

Cisco peut fournir des traductions du présent contenu dans la langue locale pour certains endroits. Veuillez noter que des traductions sont fournies à titre informatif seulement et, en cas d’incohérence, la version anglaise du présent contenu prévaudra.