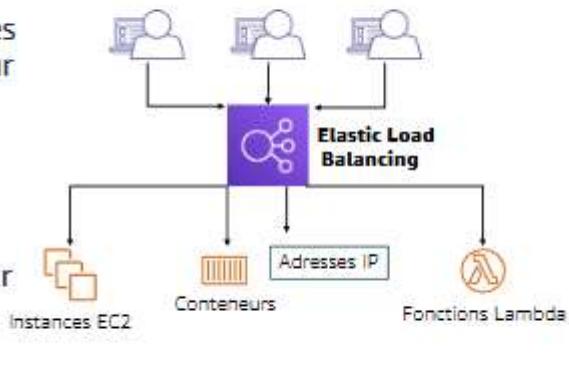


Elastic Load Balancing

Elastic Load Balancing

- Distribue le trafic entrant des applications ou du réseau sur plusieurs cibles dans une seule zone de disponibilité ou dans plusieurs zones de disponibilité.
- Fait évoluer votre équilibreur de charge en fonction de l'évolution du trafic vers votre application.



© 2022, Amazon Web Services, Inc. ou ses sociétés affiliées. Tous droits réservés.

5

Les sites Web modernes à trafic élevé doivent servir des centaines, voire des millions, de requêtes simultanées d'utilisateurs ou de clients, puis renvoient le texte, les images, la vidéo ou les données d'application correcte, de manière rapide et fiable. Des serveurs supplémentaires sont généralement nécessaires pour répondre à ces volumes élevés.

Elastic Load Balancing est un service AWS qui distribue le **trafic entrant** d'une application ou d'un réseau entre plusieurs cibles, telles que des instances Amazon EC2, des conteneurs, des adresses de protocole Internet (IP) et des fonctions Lambda, dans une seule zone de disponibilité ou dans plusieurs zones de disponibilité.

Elastic Load Balancing met à l'échelle votre **équilibrEUR** de charge en fonction de l'évolution du trafic vers votre application. Il peut s'adapter automatiquement à la plupart des charges de travail.

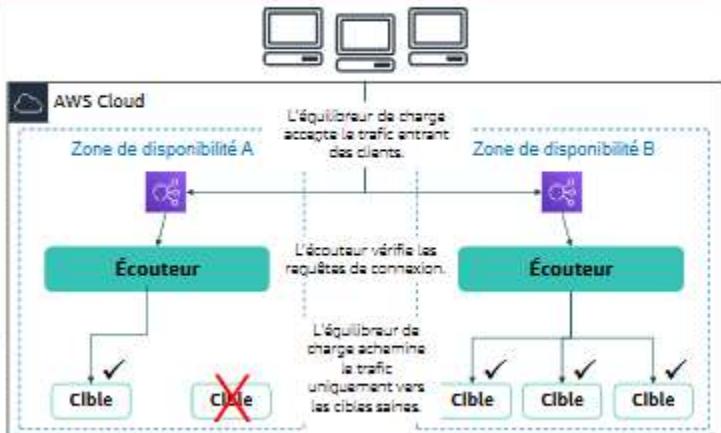
Elastic Load Balancing est disponible en trois types :

- Un équilibrer de charge d'application fonctionne au niveau de l'application (couche 7 du modèle Open Systems Interconnection, ou OSI). **Il achemine le trafic vers les cibles** : instances Amazon Elastic Compute Cloud (Amazon EC2), conteneurs, adresses IP (Internet Protocol) et fonctions Lambda, en fonction du contenu de la requête. Il est **idéal pour la répartition de charge avancée du trafic HyperText Transfer Protocol (HTTP) et Secure HTTP (HTTPS)**. Un équilibrer de charge d'application fournit un routage avancé des requêtes qui est destiné aux architectures d'application modernes, notamment des micros services et des applications conteneurisées. Un équilibrer de charge d'application simplifie et améliore la sécurité de votre application en garantissant que les derniers chiffres et protocoles de la couche de sockets sécurisés/du protocole TLS (SSL/TLS) sont utilisés à tout moment.
- Un Network Load Balancer fonctionne au niveau du transport réseau (couche 4 du modèle OSI), **en acheminant les connexions vers les cibles** : instances EC2, micro services et conteneurs, sur la base des données du protocole IP. **Il est particulièrement utile pour la répartition de charge du trafic TCP (Transmission Control Protocol) et UDP (User Datagram Protocol)**. Un Network Load Balancer est capable de traiter des millions de requêtes par seconde tout en maintenant des latences extrêmement faibles. Un Network Load Balancer est optimisé pour traiter les modèles de trafic soudains ou volatiles du réseau.
- Un Classic Load Balancer assure une répartition de charge de base entre plusieurs instances EC2. Il fonctionne à la fois au niveau des applications et du transport réseau. Un Classic Load Balancer prend en charge la répartition de charge des applications qui utilisent HTTP, HTTPS, TCP et SSL. Le Classic Load Balancer est une **implémentation plus ancienne**. **Dans la mesure du possible, AWS vous recommande d'utiliser un équilibrer de charge d'application ou un Network Load Balancer dédié**.

Fonctionnement d'AWS Elastic Load Balancing

- Avec les équilibreurs de charge d'application et les Network Load Balancers, vous enregistrez les cibles dans des groupes de cibles et acheminez le trafic vers les groupes de cibles.
- Avec les Classic Load Balancers, vous enregistrez les instances auprès de l'équilibreur de charge.

L'équilibreur de charge effectue des vérifications de l'état pour surveiller l'état des cibles enregistrées.



© 2022, Amazon Web Services, Inc. ou ses sociétés apparentées. Tous droits réservés.

7

Un équilibreur de charge accepte le trafic entrant des clients et achemine les requêtes vers ses cibles enregistrées (telles que les instances EC2) dans une ou plusieurs zones de disponibilité.

Vous configurez votre équilibreur de charge pour qu'il accepte le trafic entrant en spécifiant un ou plusieurs écouteurs.

Un écouteur est un processus qui vérifie l'existence de requêtes de connexion. Il est configuré avec un protocole et un numéro de port pour les connexions des clients à l'équilibreur de charge. De même, il est configuré avec un protocole et un numéro de port pour les connexions entre l'équilibreur de charge et les cibles.

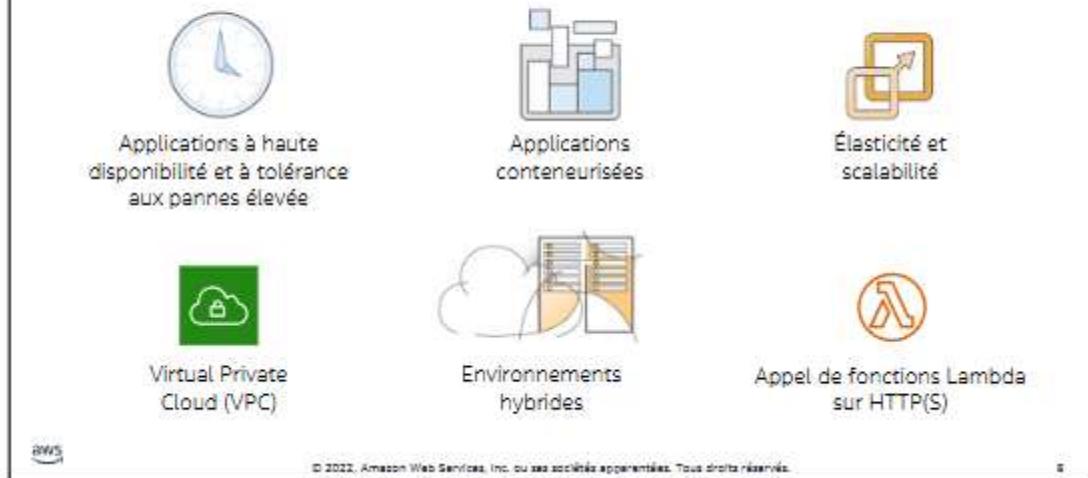
Vous pouvez également configurer votre équilibreur de charge pour qu'il effectue des vérifications de l'état, qui servent à surveiller l'état des cibles enregistrées afin que l'équilibreur de charge n'envoie des requêtes qu'aux instances saines. Lorsque l'équilibreur de charge détecte une cible malsaine, il arrête d'acheminer le trafic vers cette cible. Il reprend ensuite l'acheminement du trafic vers cette cible lorsqu'il détecte que la cible est à nouveau saine.

Il existe une différence essentielle dans la façon dont les types d'équilibreurs de charge sont configurés.

Avec les équilibreurs de charge d'application et les Network Load Balancer, vous enregistrez les cibles dans des groupes de cibles et acheminez le trafic vers les groupes de cibles.

Avec les Classic Load Balancers, vous enregistrez les instances auprès de l'équilibreur de charge.

Cas d'utilisation d'Amazon Elastic Load Balancing



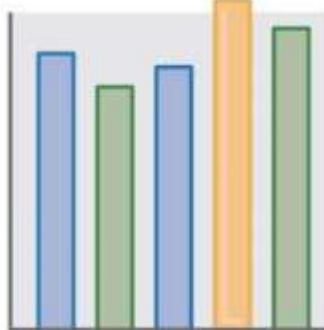
Il existe de nombreuses raisons d'utiliser un équilibrEUR de charge :

- *Obtenez une haute disponibilité et une meilleure tolérance aux pannes pour vos applications.* Elastic Load Balancing équilibre le trafic entre des cibles saines dans plusieurs zones de disponibilité. Si une ou plusieurs de vos cibles d'une seule zone de disponibilité ne sont pas saines, Elastic Load Balancing acheminera le trafic vers des cibles saines dans d'autres zones de disponibilité. Une fois les cibles revenues à un état sain, la répartition de charge reprendra automatiquement le trafic vers elles.
- *Répartissez automatiquement la charge de vos applications conteneurisées.* Grâce à la prise en charge améliorée des conteneurs pour Elastic Load Balancing, vous pouvez désormais équilibrer la charge entre plusieurs ports sur la même instance EC2. Vous pouvez également profiter de l'intégration profonde avec Amazon Elastic Container Service (Amazon ECS), qui fournit une offre de conteneurs entièrement gérés. Il vous suffit d'enregistrer un service auprès d'un équilibrEUR de charge, et Amazon ECS gère de manière transparente l'enregistrement et le désenregistrement des conteneurs Docker. L'équilibrEUR de charge détecte automatiquement le port et se reconfigure dynamiquement.

- *Mettez automatiquement vos applications à l'échelle.* Elastic Load Balancing fonctionne avec Amazon Cloud Watch et Amazon EC2 Auto Scaling pour vous aider à mettre vos applications à l'échelle en fonction des demandes de vos clients. Les alarmes Amazon Cloud Watch peuvent déclencher la scalabilité automatique de votre flotte d'instances EC2 lorsque la latence de l'une de vos instances EC2 dépasse un seuil préconfiguré. Amazon EC2 Auto Scaling fournit alors **de nouvelles instances** et vos applications seront prêtes à servir la prochaine requête du client. L'équilibrEUR de charge enregistre l'instance EC2 et dirige le trafic vers elle selon les besoins.
- *Utilisez Elastic Load Balancing dans votre cloud privé virtuel (VPC)* Vous pouvez utiliser Elastic Load Balancing pour créer un point d'entrée public dans votre VPC, ou pour acheminer le trafic de requête entre les niveaux de votre application au sein de votre VPC. Vous pouvez affecter des groupes de sécurité à votre programme de répartition de charge pour contrôler quels ports sont ouverts par rapport à une liste de sources autorisées. Comme Elastic Load Balancing fonctionne avec votre VPC, toutes vos listes de contrôle d'accès au réseau (ACL réseau) et vos tables de routage existantes continuent de fournir des contrôles supplémentaires du réseau. Lorsque vous créez un équilibrEUR de charge dans votre VPC, vous pouvez préciser si l'équilibrEUR de charge est public (par défaut) ou interne. Si vous optez pour un équilibrEUR interne, vous n'avez pas besoin de disposer d'une passerelle Internet par laquelle accéder à l'équilibrEUR de charge. Les adresses IP privées de l'équilibrEUR de charge seront utilisées dans l'enregistrement du système de noms de domaine (DNS) de l'équilibrEUR.
- *Activation de la répartition de charge hybride* Elastic Load Balancing vous permet de répartir la charge entre les ressources AWS et les ressources sur site en utilisant le même équilibrEUR de charge. Par exemple, si vous devez répartir le trafic des applications entre les ressources AWS et les ressources sur site, vous pouvez enregistrer toutes les ressources dans le même groupe cible et associer le groupe cible à un équilibrEUR de charge. Vous pouvez également utiliser la répartition de charge pondérée basée sur le DNS pour les ressources AWS et sur site en utilisant deux équilibrEURS de charge, l'un pour AWS et l'autre pour les ressources sur site. Vous pouvez également utiliser la répartition de charge hybride au profit d'applications distinctes lorsqu'une application se trouve dans un VPC et l'autre dans un emplacement sur site. Placez les cibles VPC dans un groupe cible et les cibles sur site dans un autre groupe cible, puis utilisez le routage basé sur le contenu pour acheminer le trafic vers chaque groupe cible.

- *Appel des fonctions Lambda sur HTTP(S)* Elastic Load Balancing prend en charge l'appel de fonctions Lambda pour répondre aux requêtes HTTP(S). Cela permet aux utilisateurs d'accéder à leurs applications sans serveur à partir de n'importe quel client HTTP, y compris des navigateurs web. Vous pouvez enregistrer des fonctions Lambda en tant que cibles et utiliser la prise en charge des règles de routage basé sur le contenu dans les équilibriseurs de charge d'application pour acheminer les requêtes vers différentes fonctions Lambda. Vous pouvez utiliser un équilibrleur de charge d'application comme point de terminaison HTTP commun pour les applications qui utilisent des serveurs et le calcul sans serveur. Vous pouvez créer la totalité d'un site web à l'aide de fonctions Lambda, ou combiner des instances EC2, des conteneurs, des serveurs sur site et des fonctions Lambda pour générer des applications.

Surveillance de l'équilibrer de charge



- **Métriques Amazon CloudWatch** – utilisées pour vérifier que le système fonctionne comme prévu et crée une alarme pour déclencher une action si une métrique sort d'une plage acceptable.
- **Journaux d'accès** – capturez des informations détaillées sur les requêtes envoyées à votre équilibrer de charge.
- **Journaux AWS CloudTrail** – capturez le qui, quoi, quand et où des interactions API dans les services AWS.

© 2022, Amazon Web Services, Inc. ou ses sociétés appartenantes. Tous droits réservés.

Vous pouvez utiliser les fonctions suivantes pour surveiller vos équilibreurs de charge, analyser les modèles de trafic et résoudre les problèmes liés à vos équilibreurs de charge et à vos cibles :

- *Métriques Amazon CloudWatch* Elastic Load Balancing publie des points de données vers Amazon Cloud Watch pour vos équilibreurs de charge et vos cibles. Cloud Watch vous permet de récupérer des statistiques relatives à ces points de données sous la forme d'un ensemble classé de données en séries chronologiques, appelées métriques. Vous pouvez utiliser les métriques pour vérifier que le système fonctionne comme prévu. Par exemple, vous pouvez créer une alarme Cloud Watch pour surveiller une métrique spécifiée et déclencher une action (comme l'envoi d'une notification à une adresse électronique) si la métrique sort de ce que vous considérez comme une plage acceptable.
- *Journaux d'accès* Vous pouvez utiliser les journaux d'accès pour capturer des informations détaillées sur les requêtes qui ont été émises à votre équilibrer de charge et les stocker sous forme de fichiers journaux dans Amazon Simple Storage Service (Amazon S3). Vous pouvez utiliser ces journaux d'accès pour analyser les modèles de trafic et pour résoudre les problèmes avec vos cibles ou vos applications backend.
- *Journaux AWS Cloud Trail* Vous pouvez utiliser AWS Cloud Trail pour capturer des informations détaillées sur les appels effectués à l'interface de programmation d'application (API) d'Elastic Load Balancing et les stocker sous forme de fichiers journaux dans Amazon S3. Vous pouvez utiliser ces journaux CloudTrail pour déterminer qui a passé l'appel, quels appels ont été passés, quand l'appel a été passé, l'adresse IP source d'où provient l'appel, etc.

Synthèse ELB (à retenir)

Elastic Load Balancing (ELB) est un service AWS qui **répartit automatiquement le trafic entrant** d'une application vers **plusieurs cibles** (EC2, conteneurs, IP, Lambda), dans une ou plusieurs zones de disponibilité, afin d'assurer **haute disponibilité, tolérance aux pannes et performances**.

ELB s'adapte automatiquement au volume de trafic et n'achemine les requêtes qu'aux cibles saines grâce aux **health checks**.

Il existe **trois types d'équilibreurs (ELB)** :

- **ALB (couche 7)** : routage HTTP/HTTPS avancé, microservices, conteneurs, Lambda
- **NLB (couche 4)** : TCP/UDP, très haute performance et faible latence
- **Classic ELB** : ancien, à éviter

ELB fonctionne via des **listeners** (port + protocole) et des **groupes de cibles** vers lesquels le trafic est dirigé.

Le listener, c'est la porte d'entrée de l'ELB.

Il dit : *Sur quel port ? Avec quel protocole ? Quoi faire quand une requête arrive ?*

Métaphore simple :

- ELB = le hall d'entrée de l'immeuble
- Listener = le vigile à une porte précise
- Port 80 / 443 = numéro de la porte
- Règles = instructions du vigile
- Target group = étages / bureaux

Quand quelqu'un arrive :

Je viens par la porte 443 (HTTPS)

Le listener :

OK, je t'envoie au groupe de cibles Web

Exemple : site web HTTPS

1. L'utilisateur envoie une requête HTTPS
2. ELB reçoit la requête sur le port 443
3. Le listener HTTPS (443) intercepte la requête
4. Il applique ses règles
 - /api → target group API
 - /images → target group Images
5. Il transmet la requête au groupe de cibles
6. Une cible saine répond

Un utilisateur envoie une requête vers l'ELB sur un port et un protocole donnés. Le listener correspondant intercepte la requête, applique ses règles et l'achemine vers le groupe de cibles approprié, en sélectionnant une cible saine.

ELB s'intègre avec **Auto Scaling**, **Cloud Watch**, **VPC**, et permet des architectures **scalables, sécurisées et résilientes**.

ELB = répartit le trafic

Auto Scaling = ajoute ou supprime des serveurs automatiquement

Ensemble : L'application s'adapte toute seule à la charge

Scénario simple avec :

- 1 ELB
- 2 instances EC2 derrière
- Auto Scaling activé

Situation normale

- Peu de trafic
- 2 EC2 suffisent
- ELB répartit les requêtes

Situation imprévu pic de trafic

Le trafic augmente

Cloud Watch détecte :

- CPU > 70 %
- ou latence élevée

L'**Auto Scaling** se déclenche **+** une nouvelle EC2 est créée. Elle est par la suite **automatiquement enregistrée dans le target group** et ELB commence à lui envoyer du trafic

Une fois que le **Trafic redescend**, **Cloud Watch** détecte un faible charge l'auto Scaling supprime une EC2 et ELB arrête d'envoyer du trafic à cette instance.

L'instance est terminée proprement

ELB aura servit à répartir le trafic, fait les **health checks et n'envoie du trafic qu'aux instances saines.**

L'auto Scaling à décidé quand ajouter/supprimer des instances pour répondre aux besoins se basant sur les métriques Cloud Watch surveillant en temps réel les ressources exécutées sur AWS

Pour résumé Elastic Load Balancing distribue le trafic vers les instances EC2 tandis qu'Auto Scaling ajuste automatiquement le nombre d'instances en fonction de la demande, assurant ainsi performance et haute disponibilité.