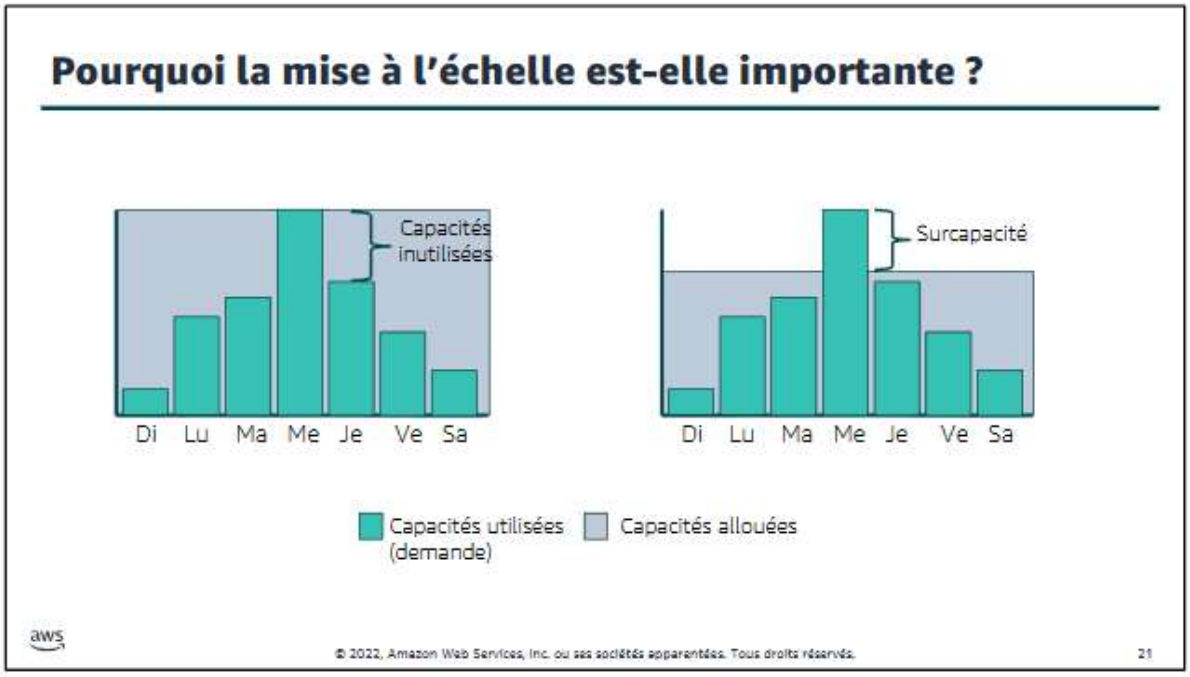
Amazon EC2 Auto Scaling

Lorsque vous exécutez vos applications sur AWS, vous voulez vous assurer que votre architecture peut être mise à l'échelle pour faire face aux changements de la demande ? Amazon EC2 Auto Scaling est pour vous.

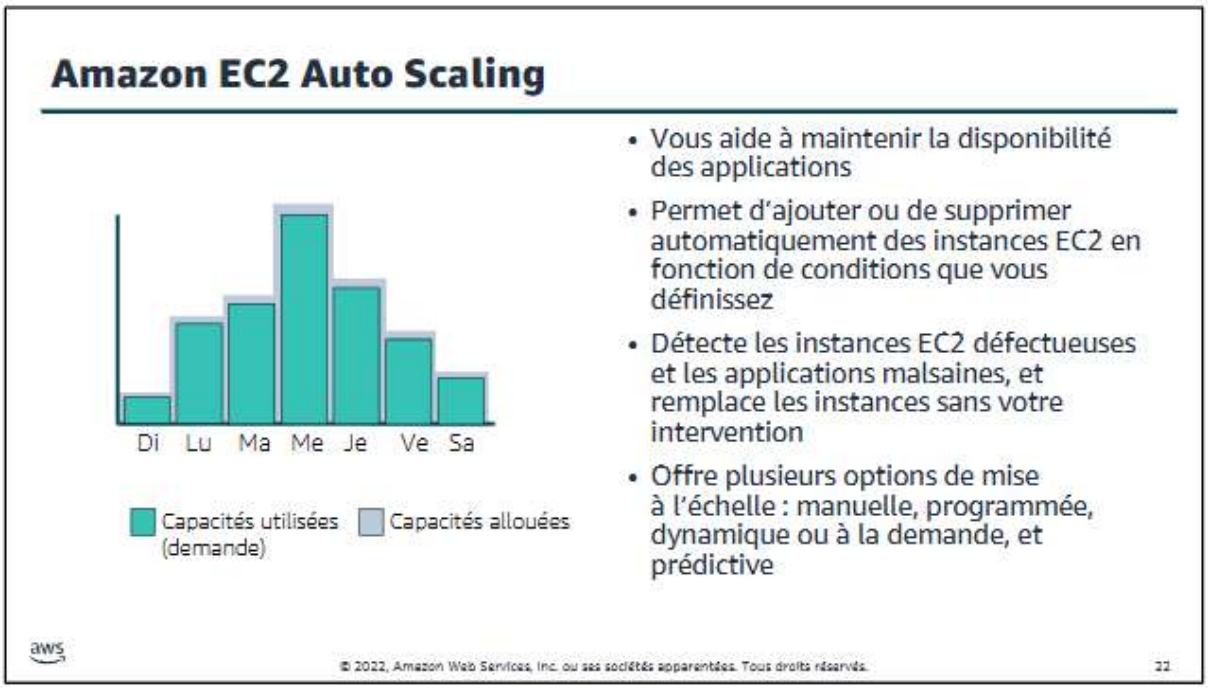


La mise à l'échelle est la capacité à augmenter ou diminuer la capacité de calcul de votre application. Pour comprendre pourquoi la mise à l'échelle est importante, prenons l'exemple d'une charge de travail dont les besoins en ressources varient. Dans cet exemple, la capacité de ressources la plus élevée est requise le **mercredi**, et la capacité de ressources la plus faible est requise le **dimanche**.

L'une des options consiste à allouer une capacité **plus que suffisante** afin de pouvoir toujours répondre à la demande la plus élevée: dans ce cas, le mercredi. Toutefois, cette situation signifie que vous utilisez des ressources qui seront **sous-utilisées la plupart des jours de la semaine**. Avec cette option, **vos coûts ne sont pas optimisés**.

Une autre option consiste à allouer **moins de capacité pour réduire les coûts**. Cette situation signifie que vous êtes en **sous-capacité certains jours**. Si vous ne résolvez pas votre problème de capacité, votre application risque de ne pas fonctionner correctement, voire de devenir indisponible pour les utilisateurs.

C'est la que Amazon EC2 Auto Scaling rentre en scène.



Dans le cloud, la puissance de calcul étant une ressource programmatique, vous pouvez adopter une approche flexible de la mise à l'échelle. Amazon EC2 Auto Scaling est un service AWS qui vous aide à maintenir la disponibilité des applications et vous permet d'ajouter ou de supprimer automatiquement des instances EC2 en fonction des conditions que vous définissez.

Amazon EC2 Auto Scaling offre plusieurs façons d'ajuster la mise à l'échelle pour répondre au mieux aux besoins de vos applications. Vous pouvez ajouter ou supprimer des instances EC2 manuellement, selon un calendrier, en réponse à l'évolution de la demande, ou en combinaison avec AWS Auto Scaling pour une scalabilité prédictive. Vous pouvez utiliser la mise à l'échelle dynamique et la mise à l'échelle prédictive en combinaison afin d'accélérer votre mise à l'échelle.

Mais pourquoi ajouter / supprimer des instances au lieu d'augmenter la puissance d'une seule ?

Parce que le cloud privilégie le "**scale horizontal**" plutôt que le "**scale vertical**".

Augmenter la puissance d'une instance existante c'est ce qu'on appelle le **scale vertical.** Par exemple passer d'une t3.micro à t3.large

Problèmes :

- Souvent besoin d'arrêter l'instance (downtime)
- Limite physique : tu ne peux pas grossir indéfiniment
- Pas adapté aux pics de charge variables

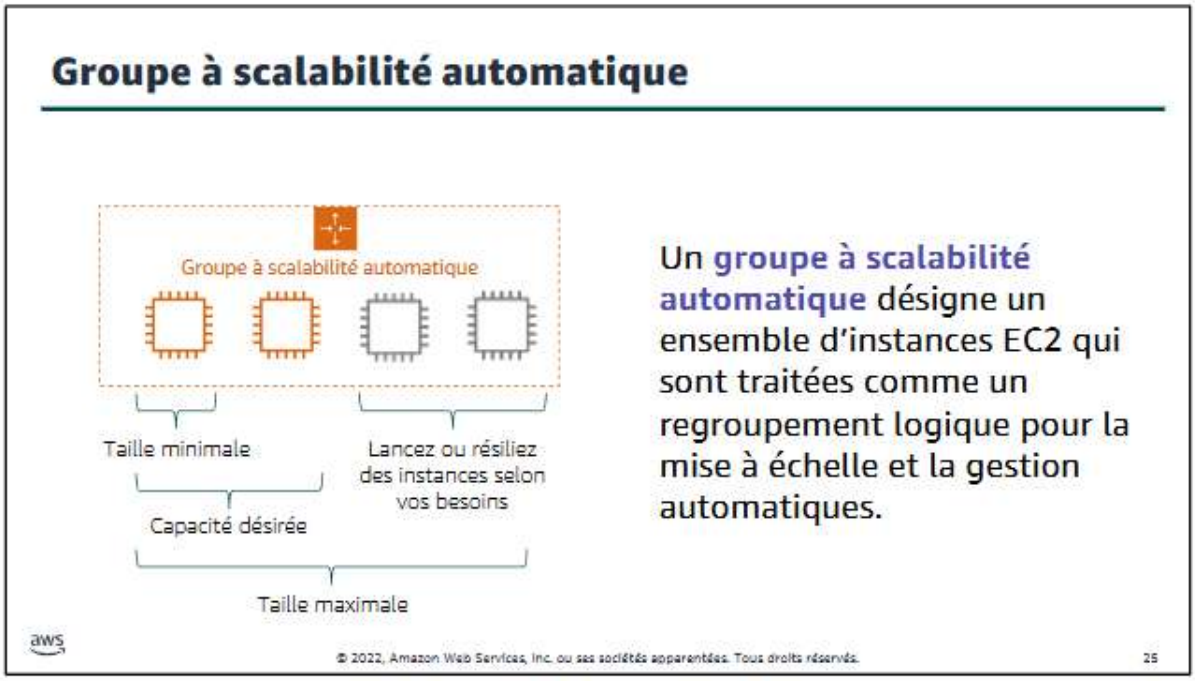
Ajouter / supprimer des instances EC2 ca c'est le **scale horizontal** (Auto Scaling)

Exemple :

- 1 instance quand il y a peu d'utilisateurs.
- 5 instances quand le trafic explose.
- Retour à 1 quand ça redescend.

Pas de downtime, Scalabilité quasi infinie et optimisation des coûts (tu ne paies que quand tu en as besoin). C'est pour ça qu'**Auto Scaling** existe.

La scalabilité automatique est également utile pour la scalabilité dynamique à la demande. Amazon.com connaît un pic saisonnier de trafic en novembre (lors du «Black Friday» et du «Cyber Monday», qui sont des jours de fin novembre où les détaillants américains organisent de grandes soldes). Si Amazon prévoit une capacité fixe pour répondre à l'utilisation la plus élevée, 76% des ressources sont inutilisées pendant la majeure partie de l'année. La mise à l'échelle de la capacité est nécessaire pour faire face aux fluctuations de la demande de services. Sans mise à l'échelle, les serveurs pourraient tomber en panne en raison de la saturation et l'entreprise perdrait la confiance de ses clients.



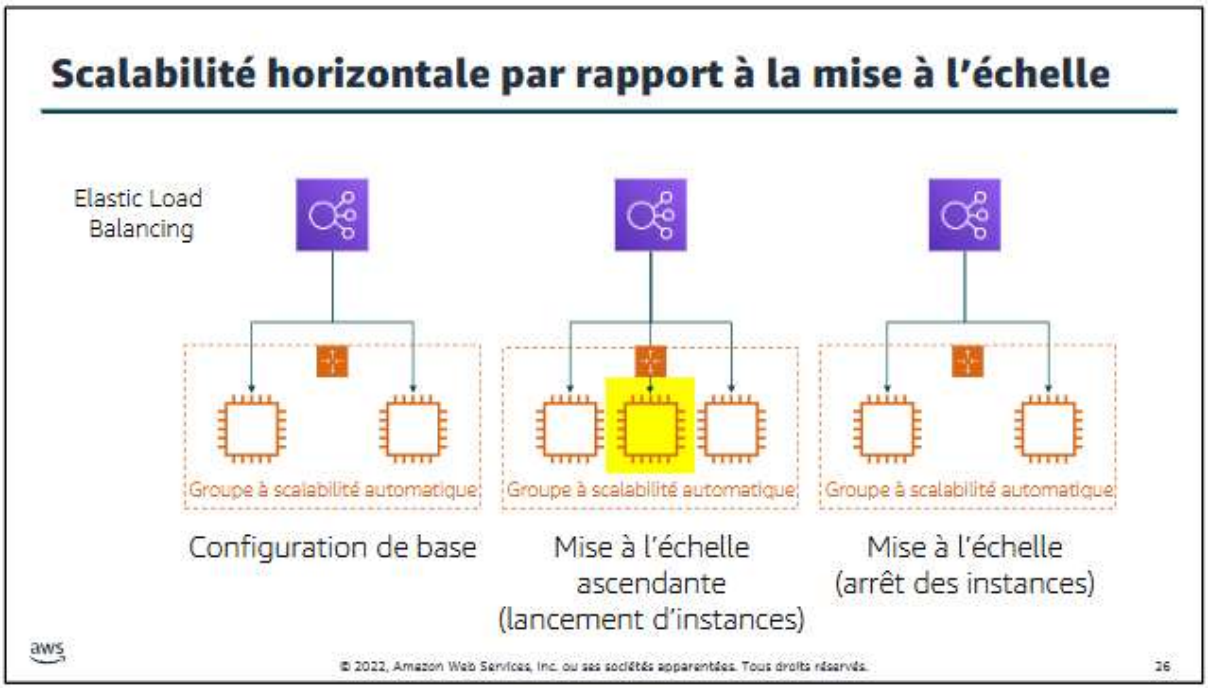
Un **groupe à scalabilité automatique** désigne une collection d'instances Amazon EC2 qui sont traitées comme un groupe logique à des fins de scalabilité automatique et de gestion. La taille d'un groupe à scalabilité automatique dépend du nombre d'instances que vous définissez comme la capacité souhaitée. Vous pouvez ajuster sa taille afin de répondre à la demande, **manuellement** ou à l'aide de la mise à l'échelle **automatique**.

Vous pouvez spécifier le nombre minimum d'instances dans chaque groupe à scalabilité automatique, et Amazon EC2 Auto Scaling est conçu pour empêcher votre groupe de **passer en dessous** de ce nombre.

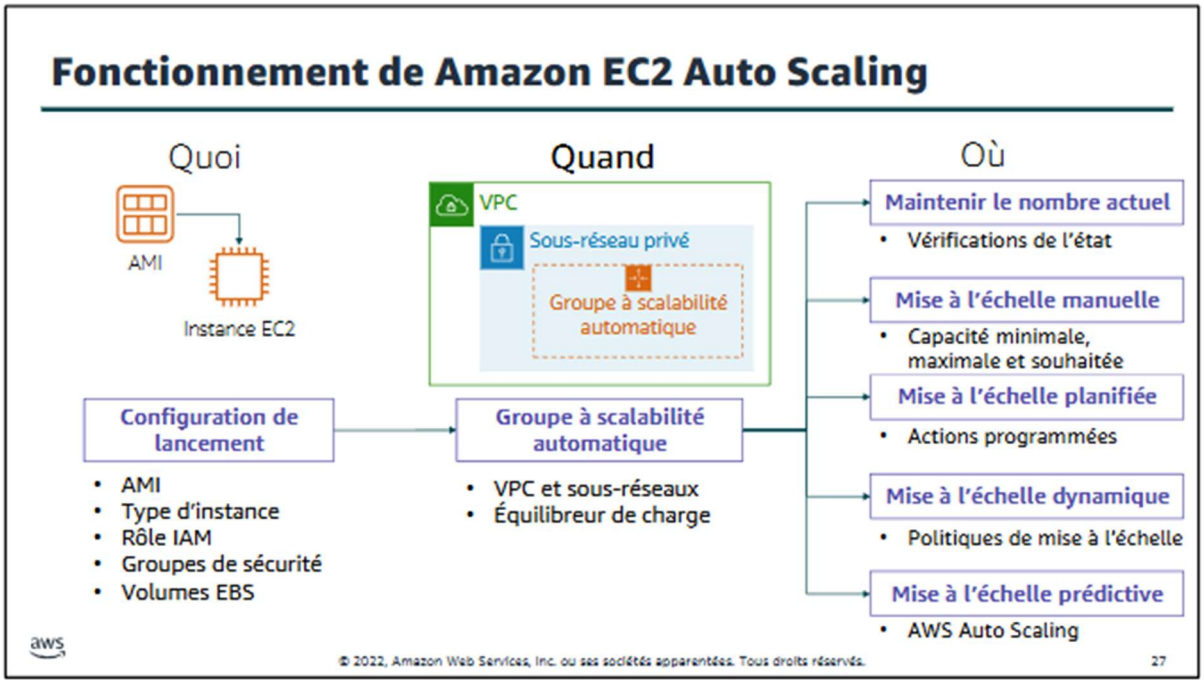
Vous pouvez spécifier le nombre maximum d'instances dans chaque groupe à scalabilité automatique, et Amazon EC2 Auto Scaling est conçu pour empêcher votre groupe de **dépasser** ce nombre.

Si vous spécifiez la capacité souhaitée, soit lorsque vous créez le groupe, soit à tout moment par la suite, Amazon EC2 Auto Scaling est conçu pour ajuster la taille de votre groupe afin qu'il dispose du nombre d'instances spécifié. Si vous spécifiez des stratégies de mise à l'échelle, Amazon EC2 Auto Scaling pourra ensuite lancer ou arrêter des instances en fonction de l'évolution des besoins de votre application.

Par exemple, ce groupe à scalabilité automatique a une taille **minimale d'une instance**, **une capacité souhaitée de deux instances** et une **taille maximale de quatre instances**. Les politiques de mise à l'échelle que vous définissez ajustent le nombre d'instances dans les limites de votre nombre minimum et maximum d'instances, en fonction des critères que vous spécifiez.



Avec Amazon EC2 Auto Scaling, le lancement des instances est appelé **scalabilité horizontale** et la résiliation des instances est appelée **diminution d'échelle**.



Pour lancer des instances EC2, un groupe à scalabilité automatique utilise une configuration de lancement, qui est un modèle de configuration d'instance. Vous pouvez considérer qu'une configuration de lancement représente ce que vous mettez à l'échelle.

Lorsque vous créez une configuration de lancement, fournissez les informations relatives aux instances.

Les informations que vous spécifiez comprennent **l'ID Amazon Machine Image (AMI)**, **le type d'instance**, **le rôle AWS Identity and Access Management (IAM)**, **le stockage supplémentaire**, un ou plusieurs **groupes de sécurité** et tout **volume Amazon Elastic Block Store (Amazon EBS)**.

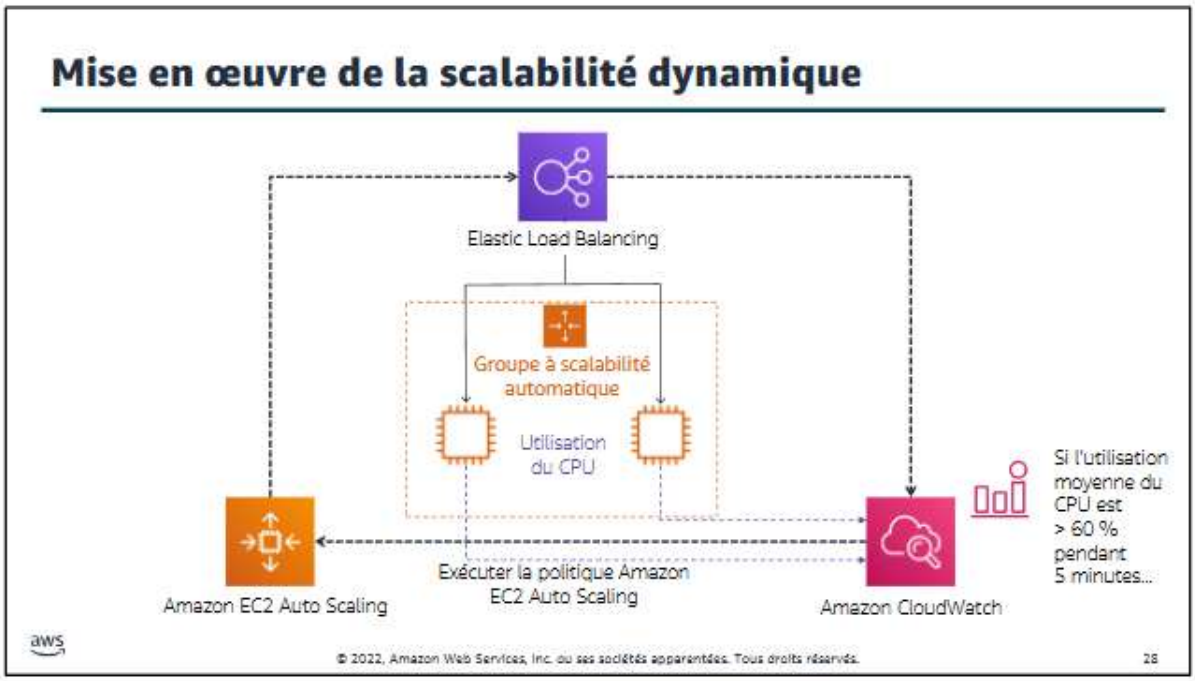
Vous définissez le nombre minimum et maximum d'instances et la capacité souhaitée de votre groupe à scalabilité automatique. Ensuite, vous le lancez dans un sous-réseau au sein d'un VPC (vous pouvez considérer que c'est l'endroit où vous effectuez la mise à l'échelle). **Amazon EC2 Auto Scaling** s'intègre à **Elastic Load Balancing (ELB)** afin de vous permettre d'attacher un ou plusieurs équilibreurs de charge à un groupe à scalabilité automatique existant.

Une fois que l'équilibreur de charge est attaché, il enregistre automatiquement les instances dans le groupe et distribue le trafic entrant aux instances.

Enfin, vous indiquez quand vous souhaitez que l'événement de mise à l'échelle se produise. Vous disposez de nombreuses possibilités de mise à l'échelle :

- Maintenir les niveaux d'instance actuels à tout moment Vous pouvez configurer votre groupe à scalabilité automatique pour maintenir un nombre spécifié d'instances en fonctionnement à tout moment. Pour maintenir les niveaux d'instance actuels, AmazonEC2 Auto Scaling effectue une vérification de l'état périodique sur les instances en cours d'exécution au sein d'un groupe à scalabilité automatique. Lorsqu'Amazon EC2 Auto Scaling détecte une instance défectueuse, il la résilie et en lance une nouvelle.
- Mise à l'échelle manuelle Avec la mise à l'échelle manuelle, vous spécifiez uniquement la modification de la capacité maximale, minimale ou souhaitée de votre groupe à scalabilité automatique.
- Mise à l'échelle planifiée Avec la mise à l'échelle planifiée, les actions de mise à l'échelle sont exécutées automatiquement en fonction de la date et de l'heure. Ceci est utile pour les charges de travail prévisibles lorsque vous savez exactement quand augmenter ou diminuer le nombre d'instances dans votre groupe. Par exemple, partons du principe que, chaque semaine, le trafic vers l'application web commence à augmenter le mercredi, reste élevé le jeudi et amorce une baisse le vendredi. Vous pouvez planifier des activités de mise à l'échelle en fonction des modèles de trafic prévisibles de l'application web. Pour mettre en œuvre la mise à l'échelle planifiée, vous créez une action planifiée.
- Mise à l'échelle dynamique et à la demande Une façon plus avancée de mettre à l'échelle vos ressources vous permet de définir des paramètres qui contrôlent le processus de mise à l'échelle. Par exemple, vous avez une application web qui fonctionne actuellement sur deux instances et vous voulez que l'utilisation du CPU du groupe à scalabilité automatique reste proche de 50% lorsque la charge de l'application évolue. Cette option est utile pour mettre à l'échelle en réponse à des conditions changeantes, lorsque vous ne savez pas quand ces conditions vont changer. La scalabilité dynamique vous donne une capacité supplémentaire pour gérer les pics de trafic sans maintenir une quantité excessive de ressources inutilisées. Vous pouvez configurer votre groupe à scalabilité automatique pour qu'il soit mis à l'échelle automatiquement pour répondre à ce besoin. Le type de politique de mise à l'échelle détermine la façon dont l'action de mise à l'échelle est exécutée. Vous pouvez utiliser Amazon EC2 Auto Scaling avec Amazon Cloud Watch pour déclencher la politique de mise à l'échelle en réponse à une alarme.

- Mise à l'échelle prédictive Vous pouvez utiliser Amazon EC2 Auto Scaling avec AWS Auto Scaling pour mettre en œuvre une mise à l'échelle prédictive, où votre capacité évolue en fonction de la demande prévue. La mise à l'échelle prédictive utilise des données collectées à partir de votre utilisation réelle d'EC2, et ces données sont complétées par des milliards de points de données issus de nos propres observations. AWS utilise ensuite des modèles de machine learning bien entraînés pour prédire votre trafic attendu (et l'utilisation de EC2), y compris les modèles quotidiens et hebdomadaires. Le modèle a besoin d'au moins un jour de données historiques pour commencer à effectuer des prédictions. Il est réévalué toutes les 24heures pour créer une prédiction pour les 48heures suivantes. Le processus de prédiction produit un plan de mise à l'échelle qui peut piloter un ou plusieurs groupes d'instances EC2 mises à l'échelle automatiquement.



Une configuration courante pour mettre en œuvre la scalabilité dynamique consiste à créer une alarme Cloud Watch basée sur les informations de performance de vos instances EC2 ou de votre équilibreur de charge. Lorsqu'un seuil de performance est franchi, une alarme Cloud Watch déclenche un événement de scalabilité automatique qui réduit ou ajoute des instances EC2 dans le groupe à scalabilité automatique.

Pour comprendre comment cela fonctionne, prenez l'exemple suivant :

- Vous créez une alarme **Amazon Cloud Watch** pour surveiller l'utilisation du CPU **dans votre flotte d'instances EC2** et exécuter des stratégies de scalabilité automatique si l'utilisation moyenne du CPU dans la flotte dépasse 60% pendant cinq minutes.
- **Amazon EC2 Auto Scaling** instancie une **nouvelle instance EC2 dans votre groupe à scalabilité automatique** en fonction de la configuration de lancement que vous avez créée.
- Une fois la nouvelle instance ajoutée, **Amazon EC2 Auto Scaling** appelle **Elastic Load Balancing** pour enregistrer la nouvelle instance EC2 dans ce groupe à scalabilité automatique.
- **Elastic Load Balancing** effectue ensuite les vérifications d'état requises et commence à **distribuer le trafic vers cette instance**. **Elastic Load Balancing** achemine le trafic entre les instances EC2 et **transmet les métriques à Amazon Cloud Watch**.