

# AWS – Auto Scaling

lundi 20 février 2023 09:29

Formateur : Mohamed AIJOU

<https://aws.amazon.com/fr/autoscaling/>

[https://fr.wikipedia.org/wiki/Point\\_de\\_d%C3%A9faillance\\_unique](https://fr.wikipedia.org/wiki/Point_de_d%C3%A9faillance_unique)

Définition :

**Auto-scaling horizontal : augmenter les performances de l'instance (passer de 2 CPU à 4 CPU ou de 16 Go de RAM à 32 Go de RAM)**

**Auto-scaling vertical : augmenter le nombre d'instance pour encaisser la montée en charge**

**SPOF** : Un point individuel de défaillance est un point qui peut être identifié dans une infrastructure ou une architecture donnée comme **étant critique pour ce où celui-ci vient à défaillir**.

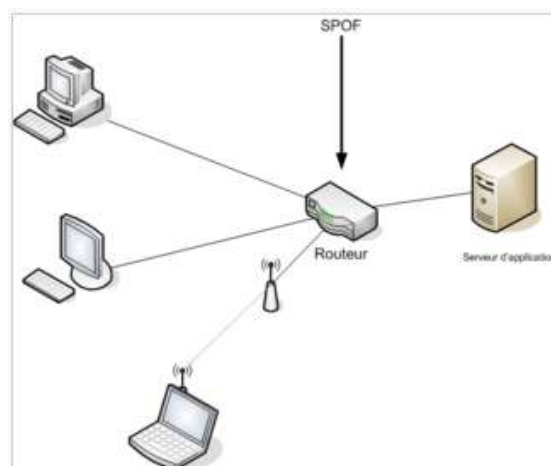
Un **point de défaillance unique** (*single point of failure* ou *SPOF* en anglais) est un point d'un **système informatique** dont le reste du système est dépendant et dont une panne entraîne l'arrêt complet du système.

Le point de défaillance unique a comme principale caractéristique de ne pas être protégé (redondant). Il est donc un risque pour la disponibilité du système. Dans la définition « *single point of failure* », le mot anglais *single* souligne le caractère unique et donc fragile du « composant ».

La notion de point de défaillance unique est fortement liée à celle de service, dans la mesure où un problème sur le point concerné entraîne une **interruption de service**.

La présence d'un point de défaillance unique dans un système augmentant la probabilité d'apparition d'un **déni de service**, elle entraîne un **risque** sur la **qualité de service**.

Dans un cadre de **haute disponibilité**, il est impossible de laisser des points individuels de défaillance dans un système.



Il y a ici deux points individuels de défaillance, le routeur et le serveur d'application qui sont tous les deux seuls et pourraient bloquer le service de production pour les 3 postes.

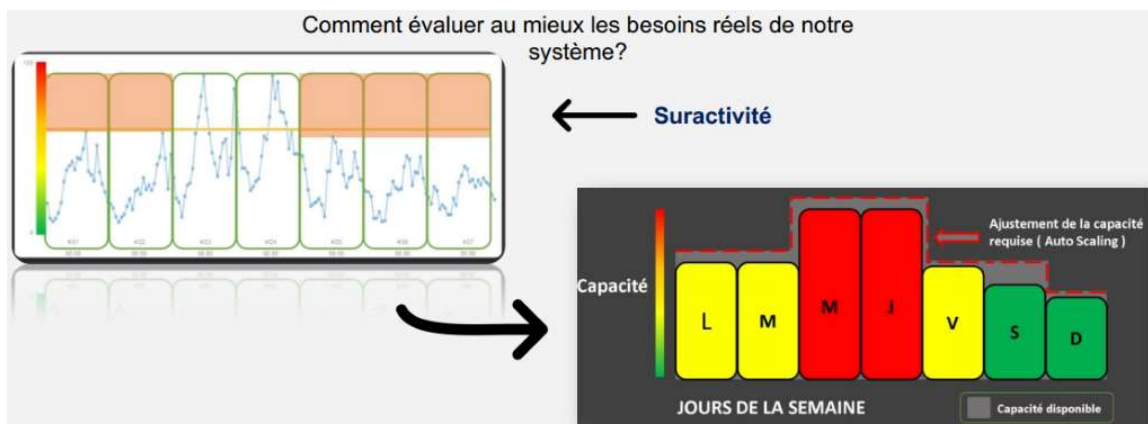
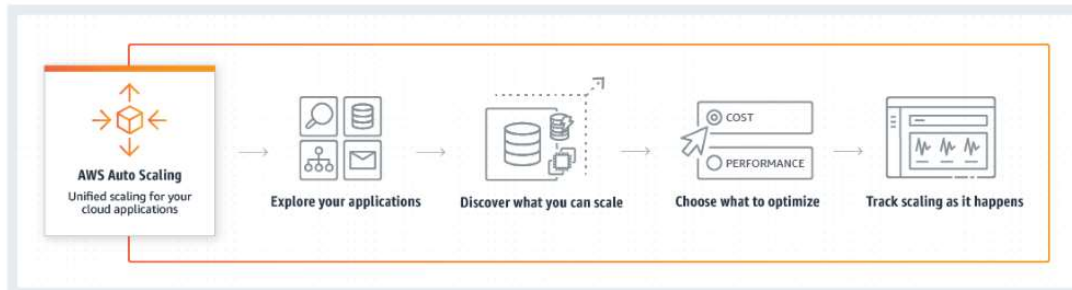
## 14 – Auto-scaling (Mise à l'échelle)



AWS Auto Scaling contrôle vos applications et ajuste automatiquement la capacité à maintenir des performances constantes et prévisibles de la manière la plus rentable possible. Grâce à AWS Auto Scaling, il est facile de configurer le dimensionnement de l'application pour diverses ressources réparties entre de multiples services en quelques minutes seulement. AWS Auto Scaling est doté d'une interface utilisateur à la fois simple et performante qui vous permet de mettre en place des plans de dimensionnement pour les ressources, notamment les instances [Amazon EC2](#) et celles du parc d'instances Spot, les tâches [Amazon ECS](#), les tables et

indices [Amazon DynamoDB](#), ainsi que les répliques [Amazon Aurora](#). AWS Auto Scaling simplifie le dimensionnement tout en apportant des recommandations qui vous permettent d'optimiser vos performances et vos coûts, mais aussi de maintenir l'équilibre entre ces valeurs. Si vous utilisez déjà [Amazon EC2 Auto Scaling](#) pour mettre dynamiquement à l'échelle vos instances Amazon EC2, vous pouvez désormais le combiner avec AWS Auto Scaling afin d'effectuer le dimensionnement de ressources supplémentaires pour d'autres services AWS. Avec AWS Auto Scaling, vos applications disposent toujours des bonnes ressources, au bon moment.

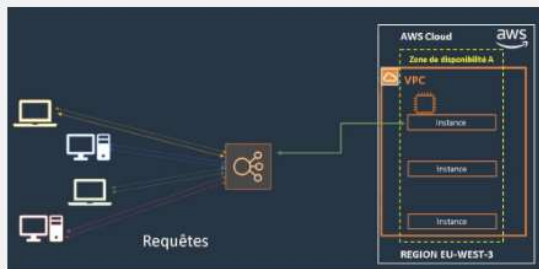
## Fonctionnement



### Auto-scaling :

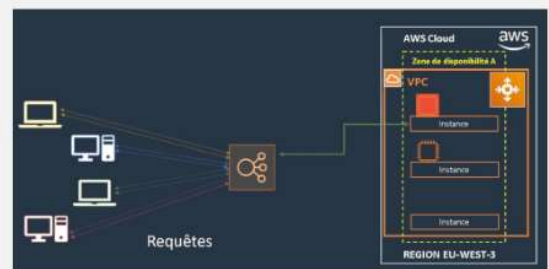
- **Scalable :**
  - *Echelonnable* : On va pouvoir anticiper dans le temps le besoin en instance.
  - *Extensible* : Il n'y a pas limite dans le provisionnement.
  - *Evolutif* : Instance plus performante. On peut provisionner des instances différentes car plus adaptées à un pic d'activité.
  - *Distribuable* : Les instances créées vont pouvoir être manipuler par ELB.
- **Automatiser le plus possible :**
  - *Un système automatisé est un ensemble d'éléments qui effectue des actions sans intervention de l'utilisateur*
- **Analyser régulièrement les données :**
  - *Issues journaux et des données Cloudwatch, pour corriger et améliorer votre système et réduire les pannes ou défaillances récurrentes*

## 14 - 3 – cas d'usage

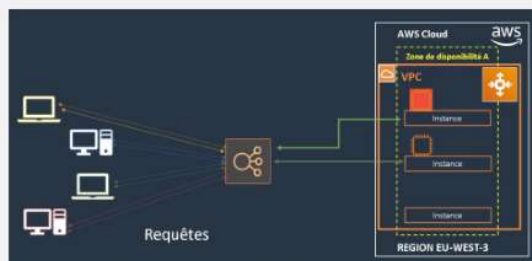


Système avec une instance d'utilisée dans un VPC avec un point d'entrée unique ELB. Activité stable en adéquation avec le matériel.

Activité augmentée

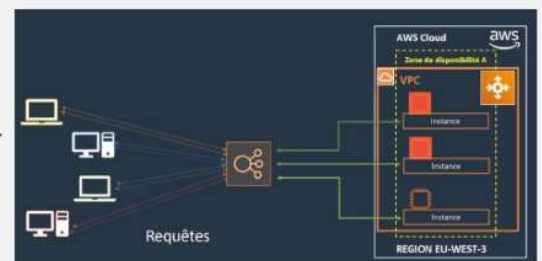


L'Activité augmente et l'instance est en phase de saturation. Détection par l'Auto-scaling.

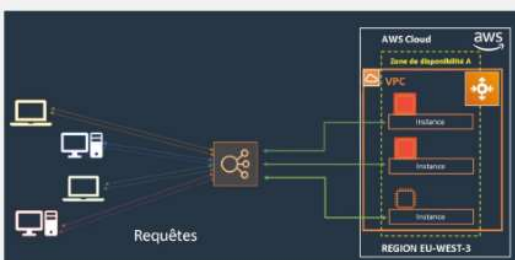


Auto-scaling va générer une seconde instance.  
ELB va pouvoir rediriger les requêtes vers cette nouvelle instance.

Activité continue de croître

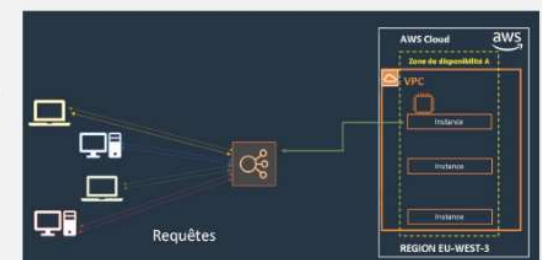


Si la seconde instance est elle aussi en surcharge.  
Auto-scaling va le détecter et créer une 3<sup>ème</sup> instance.



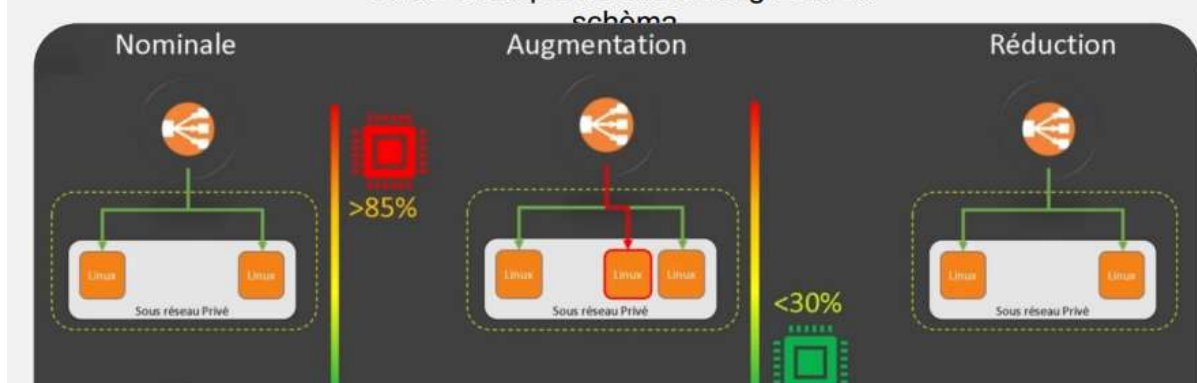
La baisse d'activité est observé par Auto-scaling.  
Il amorce la phase de « scale down »

L'activité diminue jusqu'à la situation initiale



Les instances inutilisées sont stoppées par Auto-scaling.

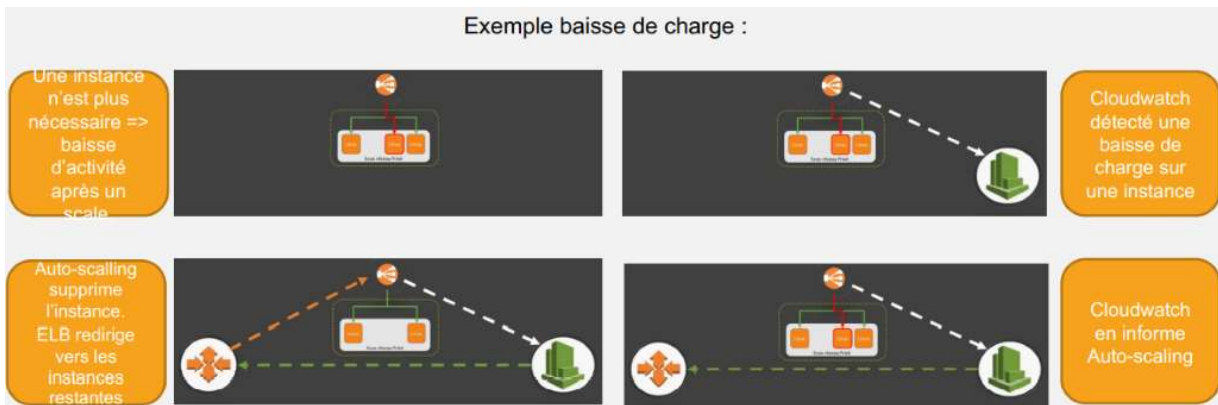
## Autre exemple de cas d'usage sur un schéma







Configuration de lancement	Groupe Auto-scaling	Politique d'auto-scaling
<b>Quoi?</b> <ul style="list-style-type: none"> <li>Image AMI ( Linux / Windows).</li> <li>Type d'instance.</li> <li>Groupe de sécurité.</li> <li>Rôles.</li> </ul>	<b>Où?</b> <ul style="list-style-type: none"> <li>VPC ou sous-réseaux.</li> <li>Equilibreur de charge.</li> <li>Minimum d'instance.</li> <li>Maximum d'instance.</li> <li>Capacité désirée.</li> </ul>	<b>Quand?</b> <ul style="list-style-type: none"> <li>Planifié.</li> <li>Politique d'augmentation.</li> <li>Politique de diminution.</li> </ul>



Dans EC2 : nous allons intégrer notre autoscaling avec la création d'un **modèle de lancement** :

EC2 > Modèles de lancement > Créer un modèle de lancement

## Créer un modèle de lancement

Créer un modèle de lancement vous permet de créer une configuration d'instance enregistrée qui peut être réutilisée, partagée et lancée ultérieurement. Les modèles peuvent avoir plusieurs versions.

### Nom et description du modèle de lancement

Nom du modèle de lancement - *obligatoire*

Doit être propre à ce compte. 128 caractères maximum. Aucun espace ni caractère spécial tel que « & », « \* » et « @ ».

Description de la version du modèle

255 caractères maximum

Conseils Auto Scaling [Informations](#)

Sélectionnez cette option si vous avez l'intention d'utiliser ce modèle avec EC2 Auto Scaling.

☐ Fournit des conseils pour vous aider à configurer un modèle que vous pouvez utiliser avec EC2 Auto Scaling.

► Balises de modèle

► Modèle source

### ▼ Images d'applications et de systèmes d'exploitation (Amazon Machine Image) [Informations](#)

Une AMI est un modèle contenant la configuration logicielle (système d'exploitation, serveur d'applications et applications) requise pour lancer votre instance. Parcourez ou recherchez des AMI si vous ne voyez pas ce que vous recherchez ci-dessous.

## Démarrage rapide

Ne pas inclure dans le modèle de lancement

**Amazon Linux**  
aws

macOS  
Mac

Ubuntu  
ubuntu

Windows  
Microsoft

>

  
**Browse more AMIs**  
Including AMIs from AWS, Marketplace and the Community

Amazon Machine Image (AMI)

Amazon Linux 2 AMI (HVM) - Kernel 5.10, SSD Volume Type  
ami-0dfcb1ef8550277af (64 bits (x86)) / ami-0cd7323ab3e63805f (64 bits (Arm))  
Virtualisation: hvm ENA activée: true Type d'appareil racine: ebs

Éligible à l'offre gratuite

### Description

Amazon Linux 2 Kernel 5.10 AMI 2.0.20230207.0 x86\_64 HVM gp2

Architecture

64 bits (x86)

AMI ID

ami-0dfcb1ef8550277af

Fournisseur vérifié

### Type d'instance Informations

Simple

#### ☒ Sélectionner manuellement le type d'instance

Sélectionnez un type d'instance qui correspond à vos besoins de calcul, de mémoire, de mise en réseau et de stockage.

#### ☐ Spécifier les attributs du type d'instance

Spécifiez les attributs d'instances qui correspondent à vos besoins de calcul.

Type d'instance

t2.micro

Éligible à l'offre gratuite

Famille: t2 1 vCPU 1 GiB Mémoire

À la demande RHEL tarification: 0.0726 USD par heure

À la demande Linux tarification: 0.0126 USD par heure

À la demande SUSE tarification: 0.0126 USD par heure

À la demande Windows tarification: 0.0172 USD par heure

Comparer les types d'instance

### Paramètres réseau Informations

Sous-réseau Informations

subnet-0e44b3fcf0989d6ac

VPC: vpc-03ca2332efe290131 Propriétaire: 639962416620

Zone de disponibilité: eu-west-1c Adresses IP disponibles: 4089 CIDR: 172.31.0.0/20

Create new subnet

When you specify a subnet, a network interface is automatically added to your template.

#### Pare-feu (groupes de sécurité) Informations

Un groupe de sécurité est un ensemble de règles de pare-feu qui contrôlent le trafic de votre instance. Ajoutez des règles pour autoriser un trafic spécifique à atteindre votre instance.

#### ☒ Sélectionner un groupe de sécurité existant

#### ☐ Créer un groupe de sécurité

Groupes de sécurité courants Informations

Sélectionner les groupes de sécurité

groupe\_securite\_irlande\_benoit\_m2i sg-03a0cb920d2a9770d

VPC: vpc-03ca2332efe290131

Comparer les règles de groupe de sécurité

Les groupes de sécurité que vous ajoutez ou supprimez ici seront ajoutés ou supprimés de toutes vos interfaces réseau.

► Configuration réseau avancée

### Résumé

#### Image logicielle (AMI)


Amazon Linux 2 Kernel 5.10 AMI...en savoir plus


ami-06e0ce9d3339cb039

#### Type de serveur virtuel (type d'instance)

Pare-feu (groupe de sécurité)  
groupe\_securite\_irlande\_benoit\_m2i

Stockage (volumes)  
1 volume(s) - 8 Gio

 **Offre gratuite :** La première année inclut 750 heures d'utilisation mensuelle des instances t2.micro (ou t3.micro dans les régions où t2.micro n'est pas disponible) sur les AMI de l'offre gratuite, 30 Gio de stockage EBS, 2 millions d'I/O, 1 Go d'instantanés et 100 Go de bande passante vers Internet



## Étapes suivantes

### Lancer une instance

Avec les instances à la demande, vous payez la capacité de calcul à la sec partir de votre modèle de lancement.

[Lancement d'une instance à partir de ce modèle](#)

### Création d'un groupe Auto Scaling à partir de votre modèle

Amazon EC2 Auto Scaling vous permet de gérer la disponibilité des appli exécutez le nombre d'instances Amazon EC2 souhaité pendant les pics d

[Créer le groupe Auto Scaling](#)

### Créer un parc d'instances Spot

Les instances Spot sont des instances EC2 inutilisées qui sont disponibles considérablement vos coûts Amazon EC2. Le tarif horaire d'une instance Les instances Spot sont particulièrement adaptées à l'analyse des donn

[Créer un parc d'instances Spot](#)

Une fois créer, on se dirige dans le menu **groupe auto scaling** dans **EC2** pour effectuer la **configuration** de notre auto scaling

## ▼ Auto Scaling

### Configurations de lancement

### Groupe Auto Scaling

## Choisir un modèle ou une configuration du lancement Info

Spécifiez un modèle de lancement qui contient les paramètres communs à toutes les instances EC2 lancées par ce groupe Auto Scaling. Si vous utilisez actuellement des configurations du lancement, vous pouvez envisager de migrer vers des modèles de lancement.

### Nom

Nom du groupe Auto Scaling  
Saisissez un nom pour identifier le groupe.

Doit être unique pour ce compte dans la région actuelle et ne doit pas dépasser 255 caractères.

Modèle de lancement Info

[Basculer vers la configuration du lancement](#)

Modèle de lancement

Choisissez un modèle de lancement qui contient les paramètres au niveau de l'instance, tels que l'Amazon Machine Image (AMI), le type d'instance, la paire de clés et les groupes de sécurité.

benoit\_modele\_test

Bien faire attention à la version du modèle sélectionné

Modèle de lancement

Info

Basculer vers la configuration du lancement

Modèle de lancement

Choisissez un modèle de lancement qui contient les paramètres au niveau de l'instance, tels que l'Amazon Machine Image (AMI), le type d'instance, la paire de clés et les groupes de sécurité.

benoit\_modele\_test

Créer un modèle de lancement

Version

2

Créer une version de modèle de lancement

Description	Modèle de lancement	Type d'instance
-	<a href="#">benoit_modele_test</a> lt-0ae01601aabb0ec7	t2.micro
AMI ID	Groupes de sécurité	Demander des instances Spot
ami-06e0ce9d3339cb039	-	Non
Nom de la paire de clés	ID des groupes de sécurité	
benoit_irlande_m2i	<a href="#">sg-03a0cb920d2a9770d</a>	

Détails supplémentaires

Stockage (volumes)	Date de création
-	Mon Feb 20 2023 09:56:48 GMT+0100 (heure normale d'Europe centrale)

Réseau

Info

Pour la plupart des applications, vous pouvez utiliser plusieurs zones de disponibilité pour équilibrer vos instances entre les zones. Le VPC par défaut et les sous-réseaux peuvent être créés rapidement.

VPC

Choisissez le VPC qui définit le réseau virtuel de votre groupe Auto Scaling.

vpc-03ca2332efe290131

172.31.0.0/16

Default

Créer un VPC

Zones de disponibilité et sous-réseaux

Définissez les zones de disponibilité et les sous-réseaux que votre groupe Auto Scaling peut utiliser.

Sélectionner les zones de disponibilité et les sous-réseaux

eu-west-1a | subnet-0124dbbb88aa8d826

172.31.16.0/20

Default

Créer un sous-réseau

On laisse les paramètres de l'étape 3 par défaut, on descend juste la période de grâce à 60 secondes :

Configurer les options avancées - facultatif

Info

Choisissez un équilibreur de charge pour répartir le trafic entrant de votre application entre les instances afin de la rendre plus fiable et facile à mettre à l'échelle. Vous pouvez également définir des options qui vous permettent de mieux contrôler les remplacements et le contrôle de la surveillance de l'état.

Répartition de charge - facultatif

Info



Utilisez les options ci-dessous pour attacher votre groupe Auto Scaling à un équilibreur de charge existant ou à un nouvel équilibreur de charge que vous définissez.

☒ **Aucun équilibreur de charge**  
Aucun équilibreur de charge ne se trouvera devant le trafic vers votre groupe Auto Scaling.

☐ **Attacher à un équilibreur de charge existant**  
Choisissez parmi vos équilibreurs de charge existants.

☐ **Attacher à un nouvel équilibreur de charge**  
Créez rapidement un équilibreur de charge de base à attacher à votre groupe Auto Scaling.

---

**Surveillances de l'état - facultatif**

Type de surveillance de l'état [Info](#)

EC2 Auto Scaling remplace automatiquement les instances qui échouent aux surveillances de l'état. Si vous avez activé la répartition de charge, vous pouvez activer les surveillances de l'état ELB en plus des surveillances de l'état EC2 qui sont toujours activées.

☒ EC2    ☐ ELB

Période de grâce de la surveillance de l'état

Durée devant s'écouler avant qu'EC2 Auto Scaling effectue la première surveillance de l'état sur les nouvelles instances après leur mise en service.

secondes

On configure la taille du groupe auto scaling : nous souhaitons **2 instance**, au **minimum** nous en aurons **1** et au **maximum 5**

## Configurer la taille du groupe et les politiques de mise à l'échelle - facultatif [Info](#)

Définissez la capacité souhaitée, minimale et maximale de votre groupe Auto Scaling. Vous pouvez éventuellement ajouter une politique de mise à l'échelle pour mettre dynamiquement à l'échelle le nombre d'instances dans le groupe.

---

**Taille du groupe - facultatif** [Info](#)

Spécifiez la taille du groupe Auto Scaling en modifiant la capacité souhaitée. Vous pouvez également spécifier des limites de capacité minimale et maximale. La capacité que vous souhaitez doit être comprise dans la plage limite.

Capacité souhaitée

Capacité minimale

Capacité maximale

Pour effectuer notre politique de mise à l'échelle : nous pouvons utiliser différents types de métrique :

## Politiques de mise à l'échelle - facultatif

Choisissez si vous souhaitez utiliser une politique de mise à l'échelle pour mettre dynamiquement à l'échelle votre groupe Auto Scaling et répondre aux modifications à la demande. [Info](#)

☒ **Politique de suivi des objectifs et d'échelonnement**  
Choisissez un résultat souhaité et laissez la politique de mise à l'échelle ajouter et supprimer de la capacité si nécessaire afin d'atteindre ce résultat.

☐ **Aucun(e)**

Nom de politique de mise à l'échelle

Type de métrique

▲  
  
 ✓  
  
  
 usion dans la métrique

☐ Désactiver la mise à l'échelle horizontale pour créer uniquement une politique de montée en puissance

Ici nous allons sélectionner l'**utilisation moyenne du processeur** comme **politique de mise à l'échelle** avec une **valeur cible à 40%**



## Politiques de mise à l'échelle - *facultatif*

Choisissez si vous souhaitez utiliser une politique de mise à l'échelle pour mettre dynamiquement à l'échelle votre groupe Auto Scaling et répondre aux modifications à la demande. [Info](#)

☒ **Politique de suivi des objectifs et d'échelonnement**

Choisissez un résultat souhaité et laissez la politique de mise à l'échelle ajouter et supprimer de la capacité si nécessaire afin d'atteindre ce résultat.

☐ Aucun(e)

Nom de politique de mise à l'échelle

Target Tracking Policy

Type de métrique

Utilisation moyenne du processeur ▼

Valeur cible

40

Besoin d'instances

30

secondes pour effectuer la préparation avant l'inclusion dans la métrique

☐ Désactiver la mise à l'échelle horizontale pour créer uniquement une politique de montée en puissance

On ne définit pas de notification et d'identification pour ce TP :

## Ajouter des notifications - *facultatif* [Info](#)

Envoyez des notifications aux rubriques SNS chaque fois qu'Amazon EC2 Auto Scaling lance ou résilie les instances EC2 de votre groupe Auto Scaling.

Ajouter une notification

Annuler

Passer à la vérification

Précédent

Suivant

On peut vérifier notre build avant de le valider :

## Vérifier [Info](#)

Étape 1 : choisir un modèle ou une configuration du lancement

Modifier

### Détails du groupe

Nom du groupe Auto Scaling  
modele\_benoit\_test

### Modèle de lancement

Modèle de lancement	Version	Description
benoit_modele_test <a href="#">🔗</a>	2	
lt-0ae01601aabb0ec7		

Étape 2 : choisir les options de lancement d'instance

Modifier

### Réseau

#### Réseau

VPC  
[vpc-03ca2332efe290131](#) [🔗](#)

Zone de disponibilité	Sous-réseau	
eu-west-1a	<a href="#">subnet-0124dbbb88aa8d826</a> <a href="#">🔗</a>	172.31.16.0/20

modele_benoit_test, 1 Politique de mise à l'échelle créé(e)					
EC2 > Groupes Auto Scaling					
Groupes Auto Scaling (3) Info					
Rechercher vos groupes Auto Scaling					
<input type="checkbox"/>	Nom	Modèle/configuration du lance...	Instances	Statut	Capacité souh...
<input type="checkbox"/>	modele_benoit_test	benoit_modele_test   Version 2	0	Mise à jour de la capi	2

On vérifie que nos machines se sont bien lancé :

i-055a7de4383b9fc3d	En cours d'exé	t2.micro	Initialisation en co
i-03d5f2c366ebcf877	En cours d'exé	t2.micro	Initialisation en co

Entre temps, notre seconde machine est passé en résilié car il n'y a pas eu d'activité pendant 20 minutes sur notre infrastructure

<input type="checkbox"/>	benoit_as1	i-055a7de4383b9fc3d	En cours d'exé	t2.micro
<input checked="" type="checkbox"/>	benoit_as2	i-03d5f2c366ebcf877	Résilié(e)	t2.micro

On se connecte à la **machine as1** via **SSH** pour effectuer des test de montée en charge (ne pas oublier de vérifier notre ouverture de port 22 dans notre groupe de sécurité)

```
Administrateur@LIL-JT33KN3 MINGW64 ~
$ ssh -i "benoit_irlande_m2i.pem" ec2-user@ec2-3-249-106-23.eu-west-1.compute.amazonaws.com
The authenticity of host 'ec2-3-249-106-23.eu-west-1.compute.amazonaws.com (3.249.106.23)' can't be established.
ED25519 key fingerprint is SHA256:pPveNL/fbTrjrrNY7gV5zyFafcV3wY7NCM78tD1mt9o.
This key is not known by any other names
Are you sure you want to continue connecting (yes/no/[fingerprint])? yes
Warning: Permanently added 'ec2-3-249-106-23.eu-west-1.compute.amazonaws.com' (ED25519) to the list of known hosts.

--|  _ _ |  )
 _| (  _/  Amazon Linux 2 AMI
---|_|_|_|

https://aws.amazon.com/amazon-linux-2/
[ec2-user@ip-172-31-20-117 ~]$
```

Création d'un script bash pour effectuer une boucle infinie :

```
GNU nano 2.9.8 infinity.sh

#!/bin/bash
while [ 1 ]; do echo "hello world" ;done;
```

Suite à l'exécution de mon script : il génère 3 nouvelles instances :

<input type="checkbox"/>	modele_benoit_test	benoit_modele_test   Version 2	4
--------------------------	--------------------	--------------------------------	---

<input type="checkbox"/>	benoit_as1	i-055a7de4383b9fc3d	En cours d'exé	t2.micro	2/2 vérifications réussies
<input type="checkbox"/>	-	i-0f287817dc84508d8	En cours d'exé	t2.micro	Initialisation en cours
<input type="checkbox"/>	-	i-04209082c3cd9404c	En cours d'exé	t2.micro	Initialisation en cours
<input type="checkbox"/>	-	i-0ecc87c39642bb0b9	En cours d'exé	t2.micro	2/2 vérifications réussies

Quand le service doit résilier des machines : il résilie dans un 1er temps les plus anciennes

