



このコンテンツは公開から3年以上経過しており内容が古い可能性があります
最新情報については[サービス別資料](#)もしくはサービスのドキュメントをご確認ください

[AWS Black Belt Online Seminar]

Amazon EC2 Auto Scaling & AWS Auto Scaling

サービスカットシリーズ
ソリューションアーキテクト
滝口 開資
2019-10-02

AWS 公式 Webinar
<https://amzn.to/JPWebinar>



過去資料
<https://amzn.to/JPArchive>



自己紹介

滝口 開資 (たきぐちはるよし)

ソリューションアーキテクト - EC2スポットインスタンススペシャリスト

日本市場でのEC2スポットインスタンス技術担当

好きなAWSサービス

- Amazon EC2 Auto Scaling
- AWS Auto Scaling
- AWSサポート



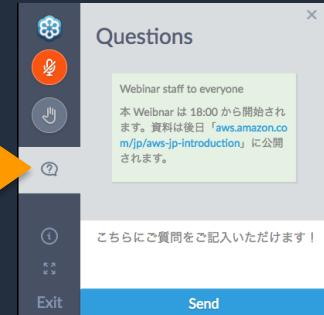
AWS Black Belt Online Seminar とは

「サービス別」「ソリューション別」「業種別」のそれぞれのテーマに分かれて、Amazon ウェブ サービス ジャパン株式会社が主催するオンラインセミナーシリーズです。

質問を投げることができます！

- 書き込んだ質問は、主催者にしか見えません
- 今後のロードマップに関するご質問はお答えできませんのでご了承下さい

- ①吹き出しをクリック
- ②質問を入力
- ③Sendをクリック



Twitter ハッシュタグは以下をご利用ください
#awsblackbelt

内容についての注意点

- 本資料では2018年x月x日時点のサービス内容および価格についてご説明しています。最新の情報はAWS公式ウェブサイト(<http://aws.amazon.com>)にてご確認ください。
- 資料作成には十分注意しておりますが、資料内の価格とAWS公式ウェブサイト記載の価格に相違があった場合、AWS公式ウェブサイトの価格を優先とさせていただきます。
- 価格は税抜表記となっています。日本居住者のお客様が東京リージョンを使用する場合、別途消費税をご請求させていただきます。
- AWS does not offer binding price quotes. AWS pricing is publicly available and is subject to change in accordance with the AWS Customer Agreement available at <http://aws.amazon.com/agreement/>. Any pricing information included in this document is provided only as an estimate of usage charges for AWS services based on certain information that you have provided. Monthly charges will be based on your actual use of AWS services, and may vary from the estimates provided.

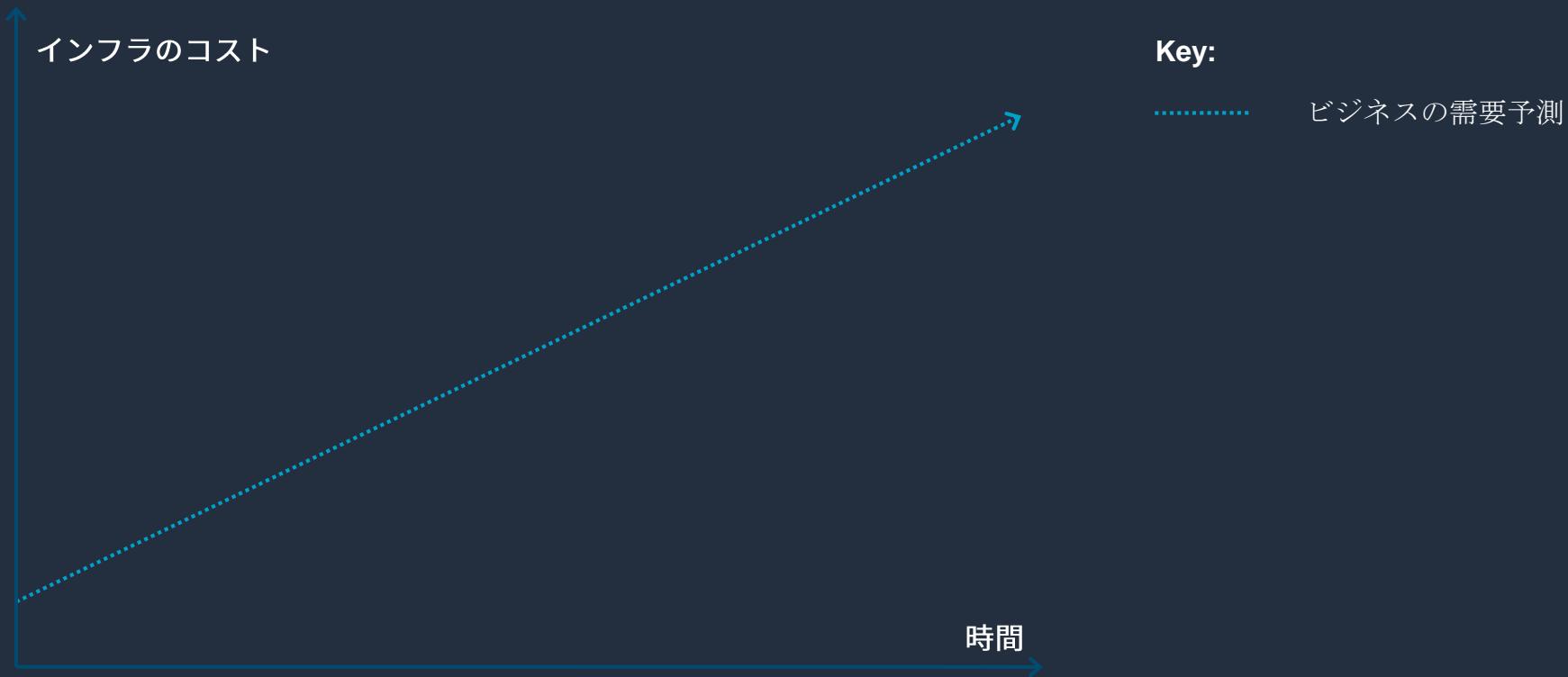
本日のアジェンダ

- Auto Scalingサービスのコンセプト
- Auto Scalingの基礎知識
- 主要機能：スケーリングの整理
- Auto Scalingを使ってみる
- こんなときどうする？ - 各種機能の紹介
- まとめ・参考資料

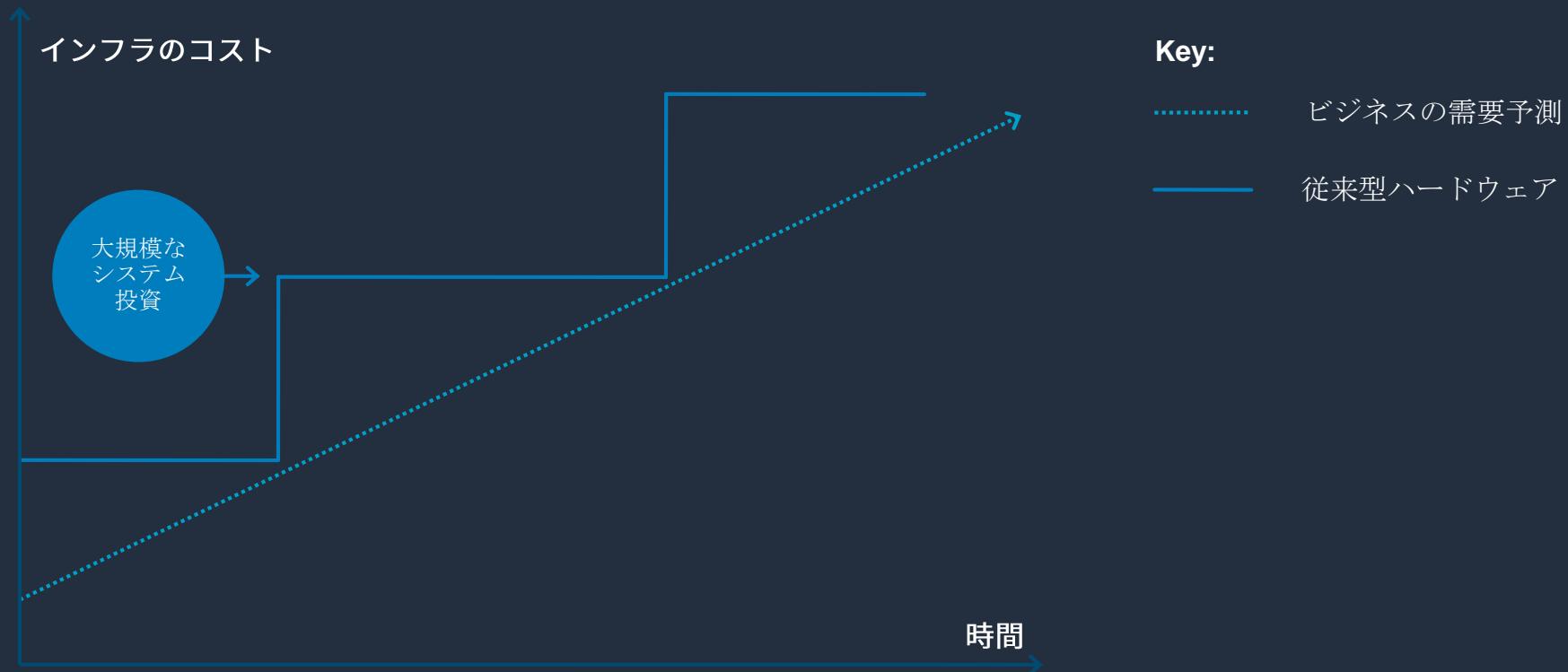
本日のアジェンダ

- Auto Scalingサービスのコンセプト
- Auto Scalingの基礎知識
- 主要機能：スケーリングの整理
- Auto Scalingを使ってみる
- こんなときどうする？ - 各種機能の紹介
- まとめ・参考資料

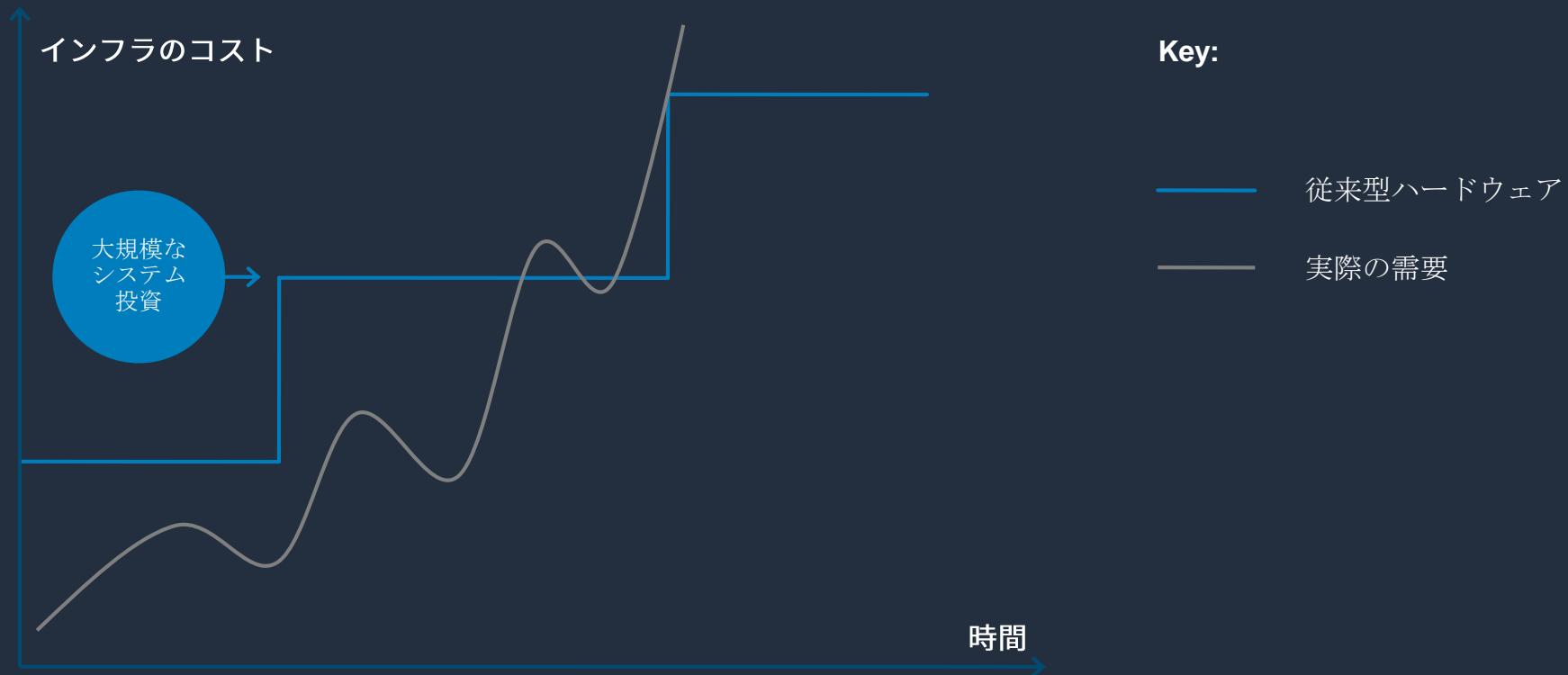
ビジネス需要に応じたキャパシティ準備



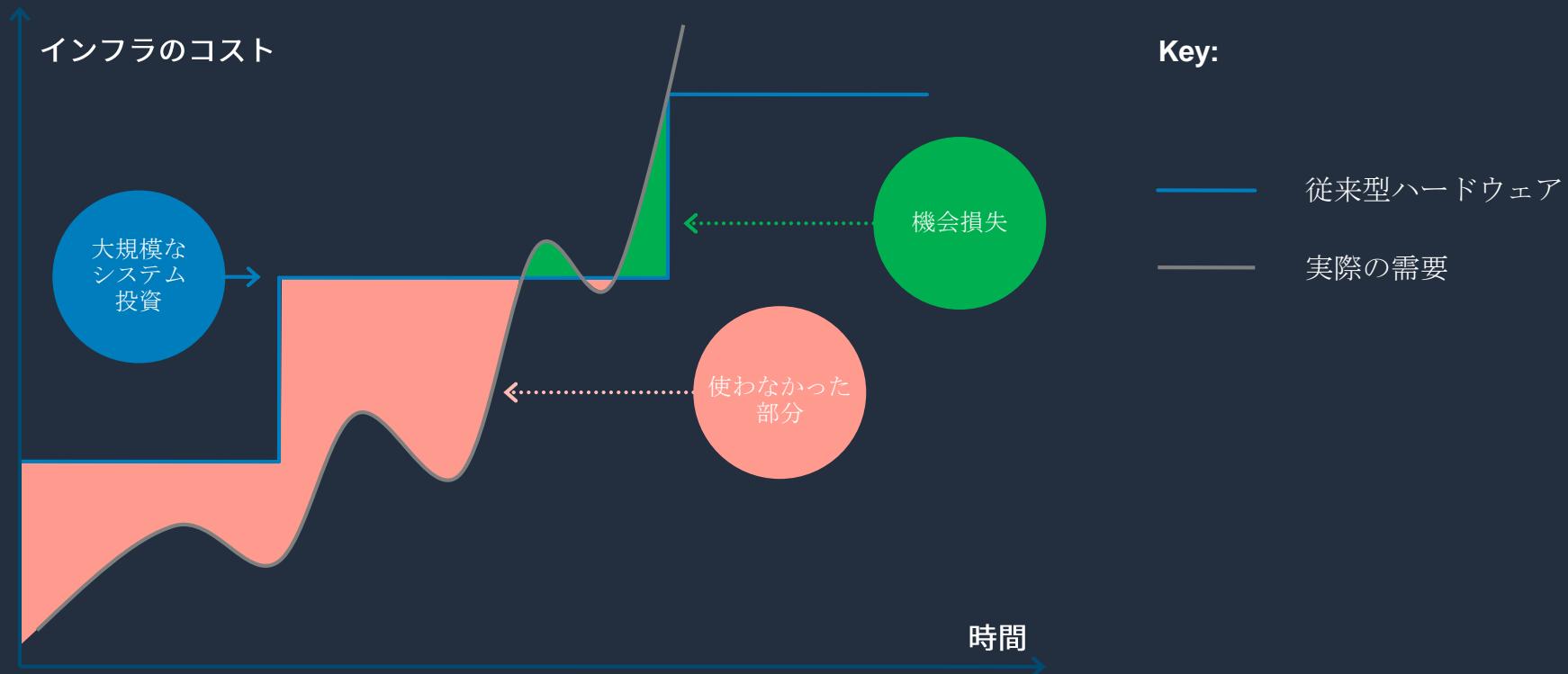
ビジネス需要に応じたキャパシティ準備



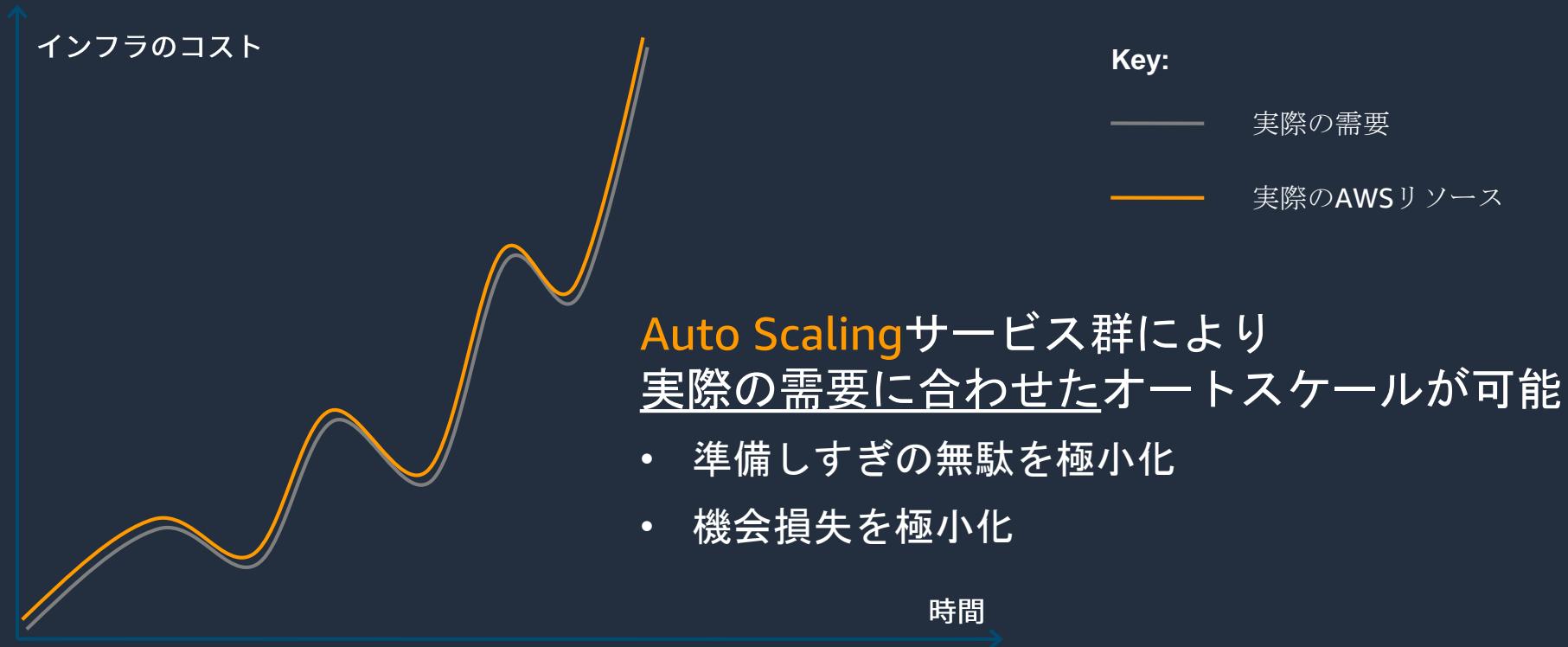
ビジネス需要に応じたキャパシティ準備



ビジネス需要に応じたキャパシティ準備



ビジネス需要に応じたキャパシティ準備



本日のアジェンダ

- Auto Scalingサービスのコンセプト
- Auto Scalingの基礎知識
- 主要機能：スケーリングの整理
- Auto Scalingを使ってみる
- こんなときどうする？ - 各種機能の紹介
- まとめ・参考資料

Auto Scalingの基礎知識

- 動作原理 - 希望する容量 (Desired Capacity, 以下「希望容量」) を目標に
- インスタンスの分散
- 均質性 - 「名前をつけてかわいがらない」

Auto Scalingの基礎知識

- 動作原理 - 希望する容量 (Desired Capacity, 以下「希望容量」) を目標に
- インスタンスの分散
- 均質性 - 「名前をつけてかわいがらない」

動作原理

Auto Scalingは、

1. 希望容量と現実の起動台数との差を監視し、
2. 常に希望容量に合致するようにリソース(EC2インスタンスなど)を増減する

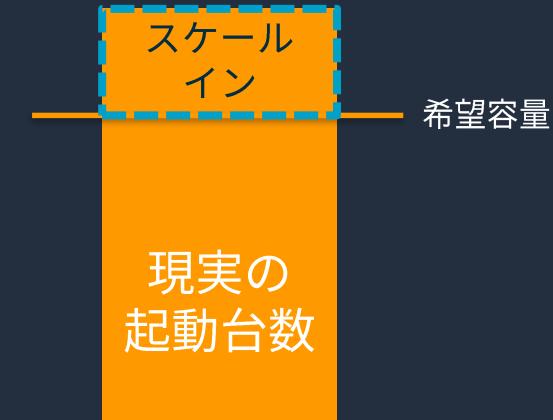
1) 静観



2) スケールアウト



3) スケールイン



希望容量の使われ方

- サイズの維持
- 手動スケーリング
- 自動スケーリング

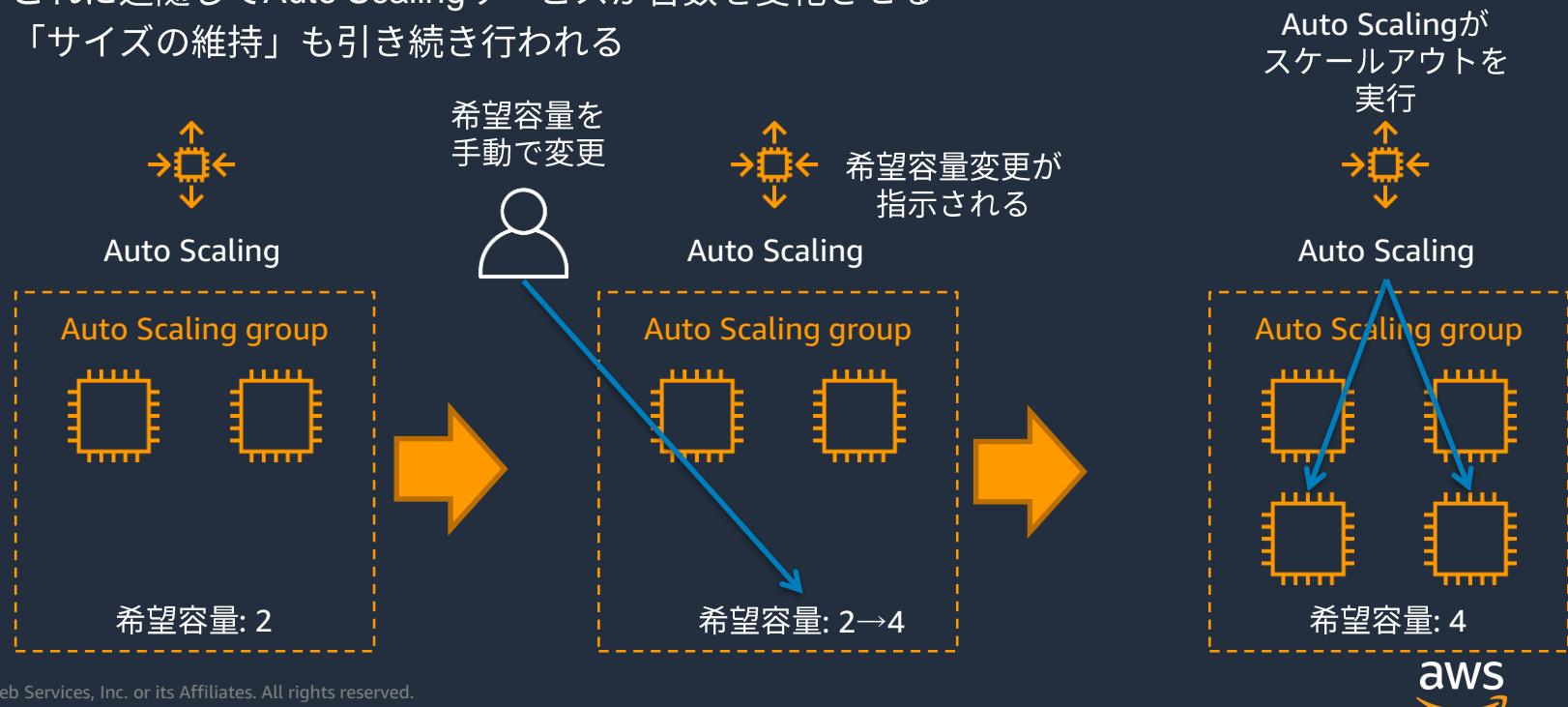
希望容量の使われ方

- ・ サイズの維持
 - ・ 希望容量は固定
 - ・ 現実の台数が減るとその差分を検知して1台追加する
 - ・ 一番シンプルな使い方



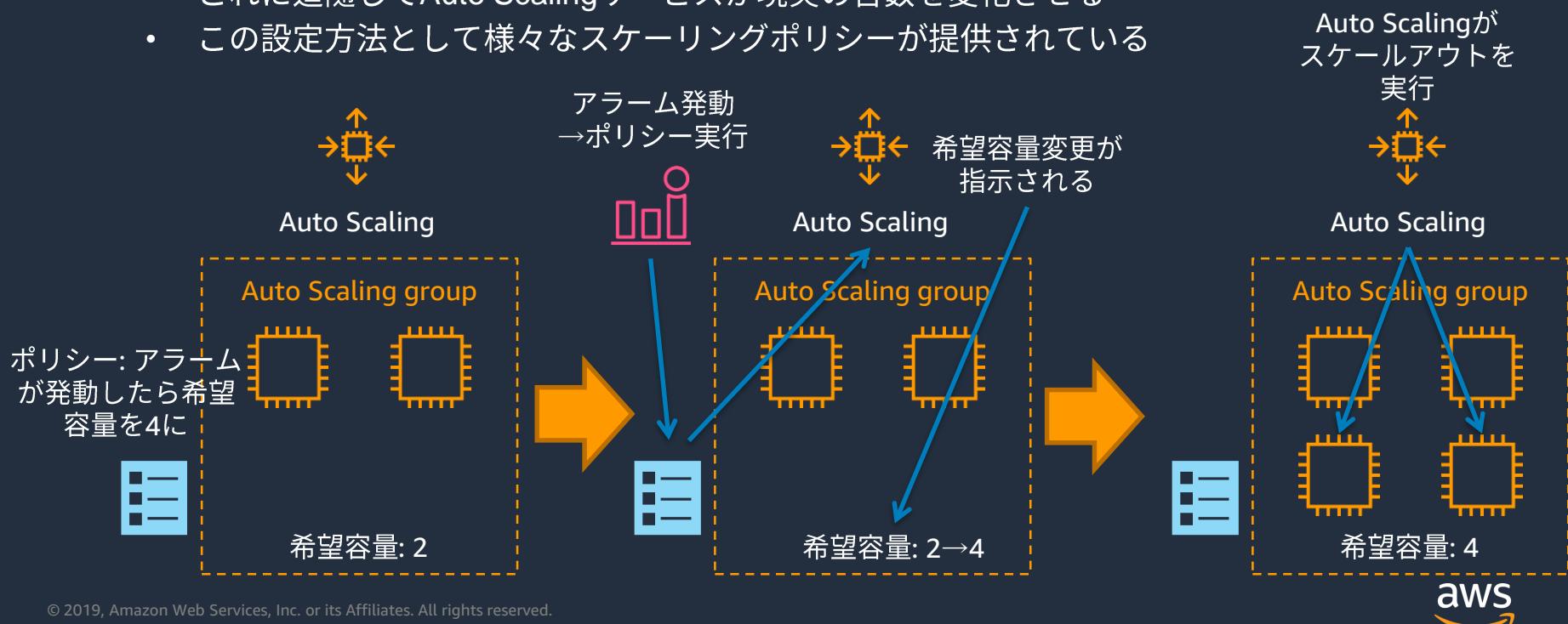
希望容量の使われ方

- 手動スケーリング
 - 希望容量を手動で変更する
 - これに追随してAuto Scalingサービスが台数を変化させる
 - 「サイズの維持」も引き続き行われる



希望容量の使われ方

- 自動スケーリング
 - 様々な条件に応じて希望容量が動的に変化する
 - これに追随してAuto Scalingサービスが現実の台数を変化させる
 - この設定方法として様々なスケーリングポリシーが提供されている



Auto Scalingの基礎知識

- 動作原理 - 希望する容量 (Desired Capacity, 以下「希望容量」) を目標に
- インスタンスの分散
- 均質性 - 「名前をつけてかわいがらない」

インスタンスの分散

- 使用できるアベイラビリティゾーンの間で、均等にインスタンスを配置しようとする
 - スケールアウトするとき：インスタンス数が最も少ないアベイラビリティゾーンに新規起動
 - これに失敗する場合、別のアベイラビリティゾーンを選択



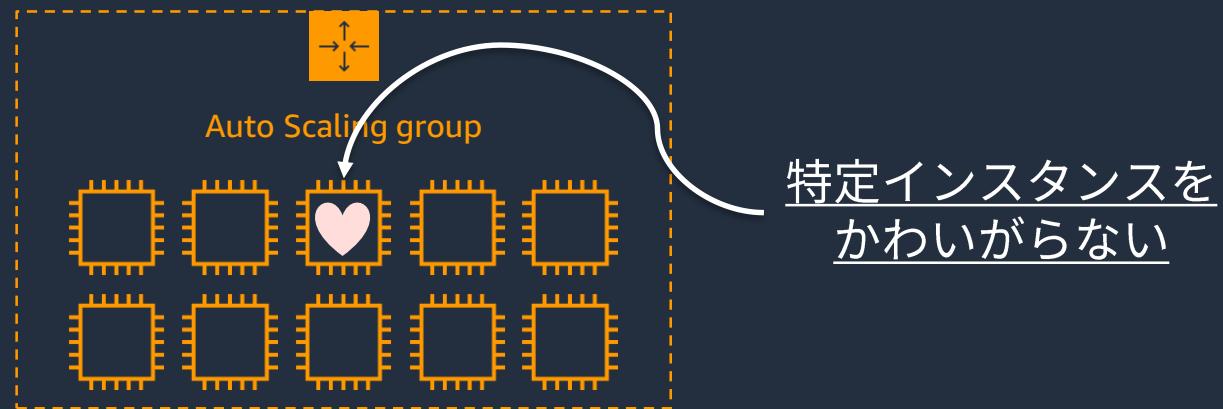
EC2 Auto Scalingはスケール動作時に
「インスタンスの分散」を最も重視する

Auto Scalingの基礎知識

- 動作原理 - 希望する容量 (Desired Capacity, 以下「希望容量」) を目標に
- インスタンスの分散
- 均質性 - 「名前をつけてかわいがらない」

均質性 - 「名前をつけてかわいがらない」

- Auto Scalingグループ内のインスタンスは原則として全て均一で、同一の価値を持つ
- 名前を付けるのはバッドプラクティス。スケールインはいつでも発生しうるものとして、置き換え可能にしておくのが重要



Auto Scalingの世界の整理

EC2 Auto
Scaling

EC2インスタンス

Auto Scalingの世界の整理

EC2 Auto Scaling

EC2インスタンス

Application Auto Scaling

ECSクラスター、スポットフリート、
EMRクラスター、AppStream 2.0フリー
ト、DynamoDBテーブル、Auroraレプリ
カ、SageMakerエンドポイントバリアン
ト、カスタムリソース

Auto Scalingの世界の整理

AWS Auto Scaling

様々なリソース

スケーリングプラン

(動的スケーリング+予測スケーリング)

EC2 Auto Scaling

EC2インスタンス

Application Auto Scaling

ECSクラスター、スポットフリート、
EMRクラスター、AppStream 2.0フリー
ト、DynamoDBテーブル、Auroraレプリ
カ、SageMakerエンドポイントバリアン
ト、カスタムリソース

Auto Scalingの世界の整理

AWS Auto Scaling

様々なリソース

スケーリングプラン

(動的スケーリング+予測スケーリング)

予測スケーリングの管理
(EC2のみ)

EC2の
管理

EC2 Auto Scaling

EC2インスタンス

Application Auto Scaling

ECSクラスター、スポットフリート、
EMRクラスター、AppStream 2.0フリー
ト、DynamoDBテーブル、Auroraレプリ
カ、SageMakerエンドポイントバリアン
ト、カスタムリソース

その他
リソース
の管理

本日のアジェンダ

- Auto Scalingサービスのコンセプト
- Auto Scalingの基礎知識
- 主要機能：スケーリングの整理
- Auto Scalingを使ってみる
- こんなときどうする？ - 各種機能の紹介
- まとめ・参考資料

主要機能：スケーリングの整理

- 動的なスケーリング
 - 簡易スケーリング
 - ステップスケーリング
 - ターゲット追跡スケーリング
- 予測スケーリング
- スケジュールスケーリング
- スケーリングオプションの選択指針

動的なスケーリング – 簡易スケーリング

- EC2 Auto Scalingのみ
- 1つのメトリクスに対して1種類だけのスケーリング調整値を指定
 - 例: CPUUtilizationが50%にならば1台追加

mysimplescalingpolicy

操作 ▾

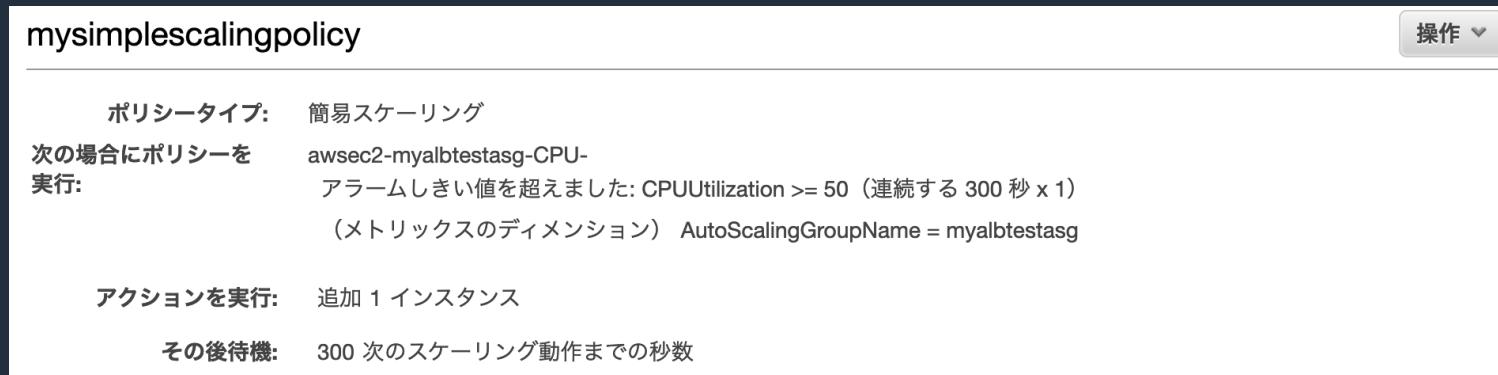
ポリシータイプ: 簡易スケーリング

次の場合にポリシーを 実行:

awsec2-myalbtestasg-CPU-
アラームしきい値を超えると: CPUUtilization >= 50 (連続する 300 秒 x 1)
(メトリックスのディメンション) AutoScalingGroupName = myalbtestasg

アクションを実行: 追加 1 インスタンス

その後待機: 300 次のスケーリング動作までの秒数



- 現在は非推奨であり、ステップスケーリングを推奨[1]
 - 「スケーリング調整が 1 つの場合でも、簡易スケーリングポリシーではなくステップスケーリングポリシーを使用することをお勧めします。」
 - 「ほとんどの場合、ステップスケーリングポリシーは簡易スケーリングポリシーよりも適しています。」

[1] https://docs.aws.amazon.com/ja_jp/autoscaling/ec2/userguide/as-scaling-simple-step.html

動的なスケーリング - ステップスケーリング (1)

- EC2 Auto Scaling, Application Auto Scaling
- 1つのメトリクスに対して複数のスケーリング調整値を指定可能
- きめ細やかな設定が可能

mystepscalingpolicy

操作 ▾

ポリシータイプ: ステップスケーリング

次の場合にポリシーを実行: awsec2-myalbtestasg-CPU-アラームしきい値を超えるました: CPUUtilization >= 50 (連続する 300 秒 x 1)
(メトリックスのディメンション) AutoScalingGroupName = myalbtestasg

アクションを実行:

- 追加 1 インスタンス 次の条件の場合 50 <= CPUUtilization < 60
- 追加 2 インスタンス 次の条件の場合 60 <= CPUUtilization < 70
- 追加 3 インスタンス 次の条件の場合 70 <= CPUUtilization < 80
- 追加 4 インスタンス 次の条件の場合 80 <= CPUUtilization < +無限大

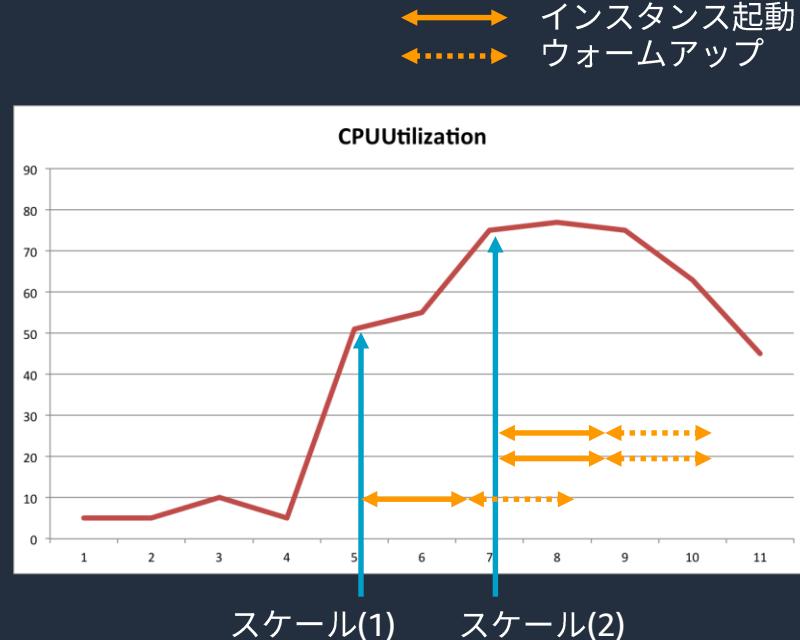
インスタンスは: 300 秒のウォームアップが各ステップ後に必要です

https://docs.aws.amazon.com/ja_jp/autoscaling/ec2/userguide/as-scaling-simple-step.html

https://docs.aws.amazon.com/ja_jp/autoscaling/application/userguide/application-auto-scaling-step-scaling-policies.html

動的なスケーリング - ステップスケーリング (2)

- ウォームアップ期間：新しいインスタンスがサービス開始できるようになるまでに何秒を要するかを設定する値
 - スケール(1)のウォームアップ期間中に次のアラームが来てスケール(2)が開始される。このとき、3台追加ではなく、「今1台追加中」とみなし、差し引き2台を追加する
 - これにより追加しすぎ問題を解決できる
- スケールアウトのタイミングで、一つ前のスケールアウトが進行中かどうかを判断してくれる、と考えても良い
- デフォルト値は300秒



ステップスケーリングポリシー定義
1台追加: $50 \leq \text{CPUUtil} < 60$
2台追加: $60 \leq \text{CPUUtil} < 70$
3台追加: $70 \leq \text{CPUUtil} < 80$

動的なスケーリング – ターゲット追跡スケーリング (1/3)

- EC2 Auto Scaling, Application Auto Scaling
- 1つのメトリクスに対し、単に目標値を指定するのみで良い
 - CPUUtilizationを40%に維持して欲しい。ただこれだけ

AutoScaling-albtest1-58558132-caa3-4ee0-a475-a340c4dcf26d 操作 ▾

ポリシータイプ: ターゲットの追跡スケーリング

次の場合にポリシーを実行:

CPU の平均使用率 を 40 に維持するために必要な場合

アクションを実行:

必要に応じてインスタンスを追加または削除

インスタンスは:

300 スケーリング後にウォームアップする秒数

スケールインの無効化

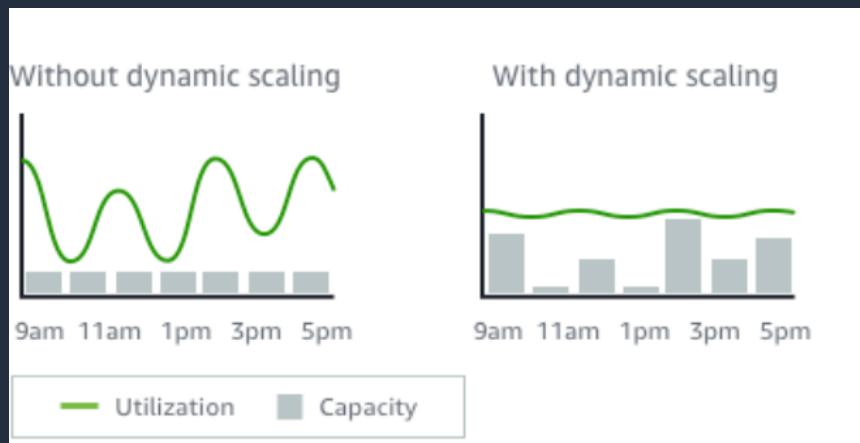
いいえ

https://docs.aws.amazon.com/ja_jp/autoscaling/ec2/userguide/as-scaling-target-tracking.html

https://docs.aws.amazon.com/ja_jp/autoscaling/application/userguide/application-auto-scaling-target-tracking.html

動的なスケーリング – ターゲット追跡スケーリング (2/3)

- 目標値を満たすように自動的にリソースが調整される
 - 何も設定しない場合、キャパシティ (灰色) が一定のため負荷 (緑色) が変動する
 - ターゲット追跡スケーリングを設定すると、負荷に応じてキャパシティが増減する。その結果、負荷が一定の値におさまる



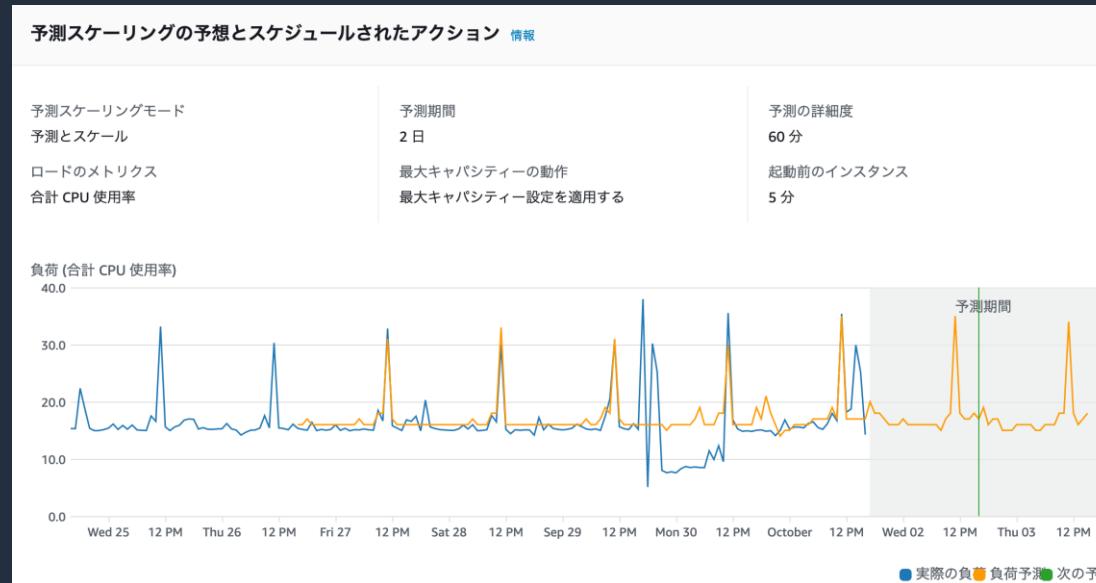
動的なスケーリング - ターゲット追跡スケーリング (3/3)

- スケールアウト用・スケールイン用の2本のアラームが自動的に作成される
 - TargetTracking-xxx-AlarmLow-UUID : スケールイン条件
 - TargetTracking-xxx-AlarmHigh-UUID : スケールアウト条件
- Highは敏感(3分など)、Lowはゆっくり(15分など)
- 基本的に、これらのアラームがアラーム状態になったときスケール



予測スケーリング (1/3)

- EC2 Auto Scalingのみ (2019-10現在)
- 2週間分のメトリクスを分析し、次の2日の今後の需要を予測
使用可能なメトリクス : CPUUtilization, NetworkIn, NetworkOut, および任意のメトリクス



https://docs.aws.amazon.com/ja_jp/autoscaling/plans/userguide/how-it-works.html

© 2019, Amazon Web Services, Inc. or its Affiliates. All rights reserved.



予測スケーリング (2/3)

- 24時間ごとに、次の48時間の予測値を作成し、キャパシティの増減をスケジュールする



- 予測の基準時刻は毎時0分
 - 「インスタンスの事前起動」設定により、スケール動作を前もって実行させることができる
 - デフォルトは5分(300秒)前
 - 午前10時に負荷が予測されている場合、対応するスケールアウトは午前9時55分に実行される

予測スケーリング (3/3)

- 考慮点とベストプラクティス
 - AWS Auto Scalingマネジメントコンソールを使う
 - いきなり使い始めず、「予測のみ」モードでどのような予測値が評価されるかを確認できる
 - ASGの作成後、24時間待つ。予測の開始には最低24時間分のデータポイントが必要

スケジュールスケーリング (1/2)

- EC2 Auto Scaling, Application Auto Scaling
- 一度限り、もしくは定期的なスケジュールを指定可能

The screenshot shows the AWS Auto Scaling Groups console for the group 'myalbtestasg'. The 'Schedule Actions' tab is selected. A search bar at the top allows filtering by action name, start time, end time, repeat frequency, desired capacity, and minimum and maximum values. One scheduled action is listed: 'mytestsched' starting at 2019 October 1 22:55:00 UTC+9.

名前	開始時刻	終了時刻	繰り返し	希望するキャパシティ	最小	最大
mytestsched	2019 October 1 22:55:00 UTC+9			1	5	

https://docs.aws.amazon.com/ja_jp/autoscaling/ec2/userguide/schedule_time.html

https://docs.aws.amazon.com/ja_jp/autoscaling/application/userguide/application-auto-scaling-scheduled-scaling.html

スケジュールスケーリング (2/2)

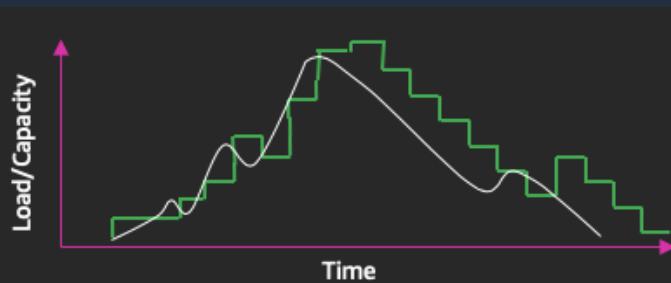
- MinCapacity(最小キャパシティ)とMaxCapacity(最大キャパシティ)のいずれか、あるいは両方を指定可能
 - 設定時刻時点の容量がMinCapacityに満たない→MinCapacityまでスケールアウト
 - 設定時刻時点の容量がMaxCapacityを超している→MaxCapacityまでスケールイン
- (EC2 ASのみ) MinCapacity, MaxCapacity, DesiredCapacity(希望キャパシティ)を指定可能



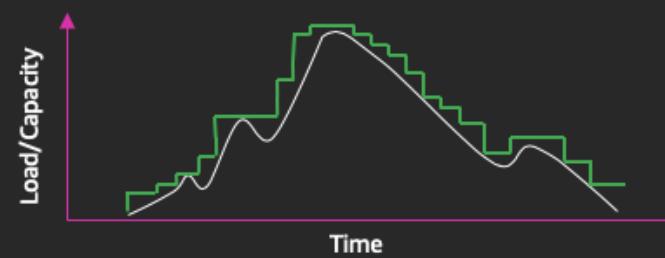
スケーリングオプションの選択指針 (1/2)

EC2でのお勧めオプション：予測スケーリングを使い、同時にターゲット追跡スケーリングも有効にする

- 1) 大まかなキャパシティ増減は予測スケーリングに任せ、前もってスケールしておく
- 2) 実際の負荷に対して不足した分をターゲット追跡で補充する



予測スケーリングによる
キャパシティの事前準備



予測スケーリング+ターゲット追跡スケーリング
によるキャパシティ準備

スケーリングオプションの選択指針 (2/2)

ステップスケーリングおよびスケジュールスケーリングを使う

- ・ 個々の条件下でのスケーリングを細かく制御したいときの考え方として引き続き有効
- ・ EC2以外のリソースについてはこちらを選択
- ・ キャパシティ設計時に、個別のパターンを下から積み上げて設定していく場合に向いている

本日のアジェンダ

- Auto Scalingサービスのコンセプト
- Auto Scalingの基礎知識
- 主要機能：スケーリングの整理
- Auto Scalingを使ってみる
- こんなときどうする？ - 各種機能の紹介
- まとめ・参考資料

Auto Scalingを使ってみる

- EC2 Auto Scalingを使ってみる
- Application Auto Scalingを使ってみる
- AWS Auto Scaling – 予測スケーリングを使ってみる

Auto Scalingを使ってみる

- EC2 Auto Scalingを使ってみる
- Application Auto Scalingを使ってみる
- AWS Auto Scaling – 予測スケーリングを使ってみる

EC2 Auto Scalingの新機能 – ミックスインスタンスグループ

- ・ オンデマンドインスタンスとスポットインスタンスをひとつのAuto Scaling グループで管理
 - (オンデマンド:スポット) = (9:1)といった指定ができる
 - インスタンスタイプを複数指定できる
 - インスタンスタイプを分散できる



EC2 Auto Scaling Groupの作成 (1)

The screenshot shows the AWS Auto Scaling Groups creation interface. On the left, a sidebar menu lists various services: Key-Pair, Network Interface, Load Balancing, Target Groups, and Auto Scaling. Under Auto Scaling, there are sub-options for Launch Configuration and Auto Scaling Groups, with the latter being highlighted by a yellow oval. In the main content area, a blue box contains a message about the start of support for launch templates. Below this, the 'Auto Scaling へようこそ' (Welcome to Auto Scaling) section provides an overview of what Auto Scaling does and includes a link to detailed information. A prominent blue button labeled 'Auto Scaling グループの作成' (Create Auto Scaling Group) is centered in the middle of the page. To the right, a sidebar titled '追加情報' (Additional Information) lists links to the User Guide, Documentation, All EC2 Resources, Forum, Pricing, and Contact Us.

キーペア
ネットワークインターフェイス
ロードバランシング
ロードバランサー
ターゲットグループ
AUTO SCALING
起動設定
Auto Scaling グループ

SYSTEMS MANAGER SERVICES
コマンドの実行
ステートマネージャー

● 起動テンプレートの提供が開始されました。

EC2 Auto Scaling コンソールで、EC2 起動テンプレートのフルサポートが開始されました。新しい Auto Scaling グループには起動テンプレートを使用することをお勧めします。起動テンプレートにより、Amazon EC2 の最新機能を活用することができます。Auto Scaling グループを作成して開始するか、[詳細はこちら](#)を参照してください。

Auto Scaling へようこそ

Auto Scaling を使用すると、Amazon EC2 キャパシティの自動的な管理、適切な数のアプリケーションインスタンスの維持、インスタンスの正常なグループの運用、必要に応じたスケーリングを行うことができます。
[詳細はこちら](#)

Auto Scaling グループの作成

注意: 別のリージョンで Auto Scaling グループを作成するには、ナビゲーションバーからリージョンを選択します。

追加情報

入门ガイド
ドキュメント
すべての EC2 リソース
フォーラム
料金
お問い合わせ

EC2 Auto Scaling Groupの作成 (2)

Auto Scaling グループの作成

キャンセルして終了

このウィザードを終了して Auto Scaling グループを作成します。最初に、起動設定または起動テンプレートを選択して、インスタンスの起動に Auto Scaling グループが使用するパラメータを指定します。

起動設定
必要な Amazon EC2 の機能をサポートしている場合は、引き続き起動設定を使用できます。[詳細ははこちら](#)

起動テンプレート [新規](#)
起動テンプレートにより、1つの種類のインスタンスを起動するか、複数のインスタンスタイプと購入オプションの組み合わせを起動するかのオプションを利用できます。起動テンプレートには Amazon EC2 の最新機能が含まれていて、更新とバージョニングができます。[詳細ははこちら](#)
[新しい起動テンプレートの作成](#)



ミックスインスタンスグループ機能を使うには
「起動テンプレート」を用いる必要がある

EC2 Auto Scaling Groupの作成 (3)

1. Auto Scaling グループの詳細設定 2. スケーリングポリシーの設定 3. 通知の設定 4. タグを設定 5. 確認

Auto Scaling グループの作成

グループ名

起動テンプレート lt-0757b443c586fc262

起動テンプレートのバージョン 1 (デフォルト)

起動テンプレートの説明

フリートの構築

起動テンプレートに従う
起動テンプレートにより、インスタンスタイプと購入オプション(オンデマンドまたはスポット)が決まります。

購入オプションとインスタンスを組み合わせる
オンデマンドインスタンスとスポットインスタンスの組み合わせ、および複数のインスタンスタイプを選択します。スポットインスタンスは、利用できる最も安い料金で自動的に起動されます。

グループサイズ 開始時 1 インスタンス

「購入オプションとインスタンスを組み合わせる」を選択

EC2 Auto Scaling Groupの作成 (4)

1. Auto Scaling グループの詳細設定 2. スケーリングポリシーの設定 3. 通知の設定 4. タグを設定 5. 確認

Auto Scaling グループの作成

グループ名

起動テンプレート lt-0757b443c586fc262

起動テンプレートのバージョン

起動テンプレートの説明 -

フリートの構築

起動テンプレートに従う
起動テンプレートにより、インスタンスタイプと購入オプション(オンデマンドまたはスポット)が決まります。

購入オプションとインスタンスを組み合わせる
オンデマンドインスタンスとスポットインスタンスの組み合わせ、および複数のインスタンスタイプを選択します。スポットインスタンスは、利用できる最も安い料金で自動的に起動されます。

インスタンスタイプ

許容できるインスタンスタイプをフリートに追加します。順序を変更し、オンデマンドインスタンスの起動の優先度を設定します。この順序によるスポットインスタンスへの影響はありません。

最低2つのインスタンスタイプを追加してください

インスタンスの分散 次のデフォルト設定を使用し、すぐに開始します。

EC2 Auto Scaling Groupの作成 (5)

1. Auto Scaling グループの詳細設定 2. スケーリングポリシーの設定 3. 通知の設定 4. タグを設定 5. 確認

Auto Scaling グループの作成

グループ名

起動テンプレート lt-0757b443c586fc262

起動テンプレートのバージョン 1 (デフォルト)

起動テンプレートの説明 -

フリートの構築

○ 起動テンプレートに従う
起動テンプレートにより、インスタンスタイプと購入オプション(オンデマンドまたはスポット)が決まります。

◎ 購入オプションとインスタンスを組み合わせる
オンデマンドインスタンスとスポットインスタンスの組み合わせ、および複数のインスタンスタイプを選択します。スポットインスタンスは、利用できる最も安い料金で自動的に起動されます。

インスタンスタイプ

許容できるインスタンスタイプをフリートに追加します。順序を変更し、オンデマンドインスタンスの起動の優先度を設定します。この順序によるスポットインスタンスへの影響はありません。

インスタンスタイプの選択 最低2つのインスタンスタイプを追加してください

インスタンスの分散 次のデフォルト設定を使用し、すぐに開始します。

EC2 Auto Scaling Groupの作成 (6)

フリートの構築

起動テンプレートに従う
起動テンプレートにより、インスタンスタイプと購入オプション（オンデマンドまたはスポット）が決まります。

購入オプションとインスタンスを組み合わせる
オンデマンドインスタンスとスポットインスタンスの組み合わせ、および複数のインスタンスタイプを選択します。スポットインスタンスは、利用できる最も安い料金で自動的に起動されます。

インスタンスタイプ

許容できるインスタンスタイプをフリートに追加します。順序を変更し、オンデマンドインスタンスの起動の優先度を設定します。この順序によるスポットインスタンスへの影響はありません。

m4.large (2vCPU、8GiB)

c4.large (2vCPU、3.75GiB)

[インスタンスタイプの追加](#)

- 要件に合うインスタンスタイプを複数選択する
- 起動テンプレートに指定しておくことも可能

EC2 Auto Scaling Groupの作成 (7)

インスタンスの分散



次のデフォルト設定を使用し、すぐに開始します。

- 上記の優先度に基づき、インテーマントインスタンスを起動します。
- アベイラビリティーゾーンごとに 2 つの最低価格インスタンスタイプ間でスポットインスタンスを多様化します。
- 各インスタンスタイプの最大スポット料金を、オンデマンド料金と同じに設定します。
- 70% のオンデマンドインスタンスと 30% のスポットインスタンスを組み合わせて維持します。

グループサイズ



開始時

1

インスタンス

「インスタンスの分散」のチェックを外す

EC2 Auto Scaling Groupの作成 (9)

インスタンスの分散 次のデフォルト設定を使用し、すぐに開始します。

オンデマンドの割り当て戦略 優先順位付け

最大スポット料金 デフォルトを使用 (推奨)
デフォルトでは現在のスポット料金が使用されますが、オンデマンド価格に上限が設定されます。
 上限価格を設定 (1 インスタンス/時間あたり)

スポットの配分戦略 スpotトインスタンスを アベイラビリティーゾーンごとに最も価格の安いインスタンスタイプ間で多様化する

オプションのオンデマンドベース 最初のインスタンスを オンデマンドとして指定します

ベースを超えるオンデマンド割合 % オンデマンドおよび 30% スpot

グループサイズ 開始時 インスタンス

台数の考え方について次のスライドで説明

EC2 Auto Scaling Groupの作成 (10)

インスタンスの分散 次のデフォルト設定を使用し、すぐに開始します。

オンデマンドの割り当て戦略 優先順位付け

最大スポット料金 デフォルトを使用 (推奨)
デフォルトでは現在のスポット料金が使用されますが、オンデマンド価格に上限が設定されます。
 上限価格を設定 (1 インスタンス/時間あたり)

スポットの配分戦略 スpotトインスタンスを アベイラビリティーゾーンごとに最も価格の安いインスタンスタイプ間で多様化する

オプションのオンデマンドベース 最初のインスタンスを オンデマンドとして指定します

ベースを超えるオンデマンド割合 % オンデマンドおよび 30% スpot

グループサイズ 開始時 インスタンス

台数の考え方の例

- 「グループサイズ」: 12

グループサイズ = 12

EC2 Auto Scaling Groupの作成 (10)

インスタンスの分散	<input type="checkbox"/> 次のデフォルト設定を使用し、すぐに開始します。
オンデマンドの割り当て戦略	<input type="checkbox"/> 優先順位付け
最大スポット料金	<input checked="" type="radio"/> デフォルトを使用 (推奨) デフォルトでは現在のスポット料金が使用されますが、オンデマンド価格に上限が設定されます。 <input type="radio"/> 上限価格を設定 (1 インスタンス/時間あたり)
スポットの配分戦略	スポットインスタンスを <input type="text" value="2"/> アベイラビリティーゾーンごとに最も価格の安いインスタンスタイプ間で多様化する
オプションのオンデマンドベース	<input type="checkbox"/> 最初のインスタンスを <input type="text" value="2"/> オンデマンドとして指定します
ベースを超えるオンデマンド割合	<input type="text" value="70"/> % オンデマンドおよび 30% スpot
グループサイズ	開始時 <input type="text" value="12"/> インスタンス

台数の考え方の例

- 「グループサイズ」: 12
- 「オプションのオンデマンドベース」: 2

グループサイズ = 12

オンデマンド = 2

EC2 Auto Scaling Groupの作成 (10)

インスタンスの分散	<input type="checkbox"/> 次のデフォルト設定を使用し、すぐに開始します。
オンデマンドの割り当て戦略	<input type="checkbox"/> 優先順位付け
最大スポット料金	<input checked="" type="radio"/> デフォルトを使用 (推奨) デフォルトでは現在のスポット料金が使用されますが、オンデマンド価格に上限が設定されます。 <input type="radio"/> 上限価格を設定 (1 インスタンス/時間あたり)
スポットの配分戦略	スポットインスタンスを <input type="text" value="2"/> アベイラビリティーゾーンごとに最も価格の安いインスタンスタイプ間で多様化する
オプションのオンデマンドベース	<input type="checkbox"/> 最初のインスタンスを <input type="text" value="2"/> オンデマンドとして指定します
ベースを超えるオンデマンド割合	<input type="text" value="70"/> % オンデマンドおよび 30% スpot
グループサイズ	開始時 <input type="text" value="12"/> インスタンス

台数の考え方の例

- 「グループサイズ」: 12
- 「オプションのオンデマンドベース」: 2
- 「ベースを超えるオンデマンド割合」: 70:30

$$\text{グループサイズ} = 12$$

オンデマンド = 2

オンデマンド = 7

スポート = 3

EC2 Auto Scaling Groupの作成 (10)

インスタンスの分散 ① 次のデフォルト設定を使用し、すぐに開始します。

オンデマンドの割り当て戦略 ① 優先順位付け

最大スポット料金 ① デフォルトを使用 (推奨)
デフォルトでは現在のスポット料金が使用されますが、オンデマンド価格に上限が設定されます。
 上限価格を設定 (1 インスタンス/時間あたり)

スポットの配分戦略 ① スpotインスタンスを アベイラビリティーゾーンごとに最も価格の安いインスタンスタイプ間で多様化する

オプションのオンデマンドベース ① 最初のインスタンスを オンデマンドとして指定します

ベースを超えるオンデマンド割合 ① % オンデマンドおよび 30% スpot

グループサイズ ① 開始時 インスタンス

台数の考え方の例

- 「グループサイズ」: 12
- 「オプションのオンデマンドベース」: 2
- 「ベースを超えるオンデマンド割合」: 70:30

結果

オンデマンド 9台
スポット 3台

$$\text{グループサイズ} = 12$$

オンデマンド = 2

オンデマンド = 7

スポット = 3

Auto Scalingを使ってみる

- EC2 Auto Scalingを使ってみる
- Application Auto Scalingを使ってみる
- AWS Auto Scaling – 予測スケーリングを使ってみる

スポットフリートでのApplication Auto Scaling の活用

(1)

The screenshot shows the AWS EC2 Spot Instances Request interface. The left sidebar has a navigation menu with the following items: EC2 ダッシュボード, イベント, タグ, レポート, 制限, インスタンス, インスタンス, 起動テンプレート, **スポットリクエスト** (highlighted with an orange box), リザーブドインスタンス, 専有ホスト, and スケジュール済みインスタンス. The main content area has a title bar with tabs: 'スポットインスタンスのリクエスト' (highlighted with an orange box), 'アクション', '価格設定履歴', and 'Savings Summary'. Below the title bar is a search bar with dropdowns for 'リクエストタイプ: all' and '状態: all', and a 'キーワードによる検索' field. To the right of the search bar are navigation arrows: '< リクエストなし >'. Underneath the search bar is a table header with columns: 'リクエスト ID', 'リクエストタ...', 'インスタンスタ...', '状態', '容量', and 'ステータス'. A message in the center says: '現在、このリージョンにスポットリクエストはありません。' Below it, instructions say: 'EC2 スpotトインスタンスを初めて使用する場合は、「開始方法」ページにアクセスしてください。' and 'スポットインスタンスを起動するには、スポットインスタンスのリクエスト ボタンをクリックします。' At the bottom of the main content area is a large blue button labeled 'スポットインスタンスのリクエスト'. A note at the bottom of the page says: '詳細を表示するには、上記から 1 つのスポットリクエストを選択します。'

「スポットリクエスト」 → 「スポットインスタンスのリクエスト」

詳細は以下の「EC2スポットインスタンスのすべて」資料をご参照ください

<https://aws.amazon.com/jp/summits/tokyo-osaka-2019-report/>

スポットフリートでのApplication Auto Scaling の活用

(2)

必要な容量をお知らせください

起動するターゲット容量(インスタンス数または vCPU 数)を設定します。起動テンプレートを指定した場合、ターゲット容量の一部をオンデマンドとして割り当てることができます。オンデマンドインスタンスの数は常に保持されますが、スポットインスタンスはスケールできます。

合計ターゲット容量 ⓘ

1

インスタンス▼

オプションのオンデマンド部分 [詳細はこち](#)

ら↗

0

インスタンス

起動テンプレートを指定するリクエストのみが
オンデマンドの対象です

ターゲット容量を維持する

中断動作 ⓘ

終了

- 作成時の注意点：「ターゲット容量を維持する」にチェックを入れる(maintainモードを指定する)
https://docs.aws.amazon.com/ja_jp/AWSEC2/latest/UserGuide/spot-fleet-target-tracking.html
 - 「スポットフリートリクエストには、タイプが maintain のリクエストが必要です。」
- 中断などで指定容量を下回った場合にスポットフリートが自動で新しいインスタンスを起動する

スポットフリートでのApplication Auto Scaling の活用

(3)

The screenshot shows the AWS Spot Fleet Requests interface. At the top, there are tabs for 'Spotインスタンスのリクエスト' (selected), 'アクション', '価格設定履歴', and 'Savings Summary'. On the right are refresh and settings icons. Below the tabs is a search bar with filters for 'リクエストタイプ: all' and '状態: all', and a keyword search field. A pagination bar indicates '1 リクエスト中 1 から 1 を表示'. The main table lists one request:

リクエスト ID	リクエストタ...	インスタンスタ...	状態	容量	ステータス	永続性	作成日
sfr-d8948be1-dc6b...	fleet	t3.medium,t2.m...	active	3 of 3	fulfilled	maintain	a day

Below the table, a note says 'リクエスト ID: sfr-d8948be1-dc6b-47d6-b1e7-b20212525f78'. Underneath, there are tabs: '説明', 'インスタンス', '履歴', '削減額', 'Auto Scaling' (which is highlighted with an orange box), and 'スケジュールに基づくスケーリング'. A message states 'このフリートに対して Auto Scaling は設定されていません' (Auto Scaling is not set for this fleet). A large blue '設定' button is prominently displayed, also outlined in orange. A note below it says 'CloudWatch アラームに応じてフリートのターゲット容量を指定範囲内で自動的に調整します' (Automatically adjusts the fleet's target capacity within the specified range based on CloudWatch Alarms).

“Auto Scaling”タブ→「設定」

スポットフリートでのApplication Auto Scaling の活用

(4)



- ターゲット追跡スケーリングポリシーの目標値を設定する
- もしくはステップスケーリングポリシーを選択することもできる

Application Auto Scaling の設定ポイント

- マネジメントコンソールを使う場合と CLI(API, SDK)を使う場合の違い
 - マネジメントコンソールの場合
各サービスのマネジメントコンソールから設定する
 - CLI(API, SDK)の場合
使用するAPIはすべてApplication Auto ScalingサービスのAPIである
 - (CLIの例)
`aws application-autoscaling register-scalable-target ¥
--service-namespace サービス名 ¥
--resource-id リソースID ¥`
...

Auto Scalingを使ってみる

- EC2 Auto Scalingを使ってみる
- Application Auto Scalingを使ってみる
- AWS Auto Scaling – 予測スケーリングを使ってみる

下準備 - EC2 Auto Scalingグループの設定変更

Auto Scaling グループ: myasgforssm

詳細 アクティビティ履歴 スケーリングポリシー インスタンス モニタリング 通知 タグ スケジュールされたアクション ラ...

Auto Scaling メトリックス: グループメトリックコレクションを有効にする 次のデータを表示: 過去 1 時間

表示:Auto Scaling または EC2

警告 以下の Auto Scaling グループでは、グループメトリックコレクションが有効になっていません: myasgforssm

以下は、選択されたリソースの CloudWatch メトリックスです (最大 10)。画面を拡大するには、グラフをクリックします。すべての時刻は協定世界時 (UTC) で表示されています。> すべての CloudWatch メトリックスを表示

myasgforssm (有効でない)

最小グループサイズ (カウント)	最大グループサイズ (カウント)	希望するキャパシティ (カウント)
1 0.75	1 0.75	1 0.75



EC2 Auto Scalingマネジメントコンソールから
「モニタリング」→「グループメトリックコレクションを有効にする」

AWS Auto ScalingによるEC2 Auto Scalingグループの予測スケーリング (1)



管理ツール

AWS Auto Scaling 複数のリソースを迅速かつ 簡単にスケールできるよう 支援します

AWS Auto Scaling では、アプリケーションの基になるすべてのスケーラブルなリソースを
すばやく検出し、組み込みのスケーリング推奨項目を使用して数分でアプリケーションの
スケーリングをセットアップできます。

この機能の説明

スケーリングプランの作成

わずか数ステップでアプリケーションを最適化します

今すぐ始める

料金表

AWS Auto Scaling は無料です。

AWS Auto Scaling は、Amazon CloudWatch で有効
にすることができますが、追加料金は発生しません。
アプリケーションリソースおよび Amazon
CloudWatch のサービス料金が適用されます。

AWS Auto ScalingによるEC2 Auto Scalingグループの予測スケーリング (2)

AWS Auto Scaling > スケーリングプラン > スケーリングプランの作成

ステップ 1
スケーラブルなリソースの検索

自動検出または手動で、スケーリングプランに追加するリソースを選択します。 [情報](#)

ステップ 2
スケーリング戦略を指定します。

ステップ 3
詳細設定の設定 (オプション)

ステップ 4
確認と作成

メソッドの選択

- CloudFormation スタックによる検索
AWS CloudFormation によってプロビジョニングされたリソースを検索します。
- タグによる検索
適用されたタグを使用してリソースを検索します。
- Amazon EC2 Auto Scaling グループの選択
スケーリング計画に含めるための、Auto Scaling グループを 1 つ以上選択します。

「EC2 Auto Scalingグループの選択」を選択

AWS Auto ScalingによるEC2 Auto Scalingグループの予測スケーリング (3)

AWS Auto Scaling > スケーリングプラン > スケーリングプランの作成

ステップ 1
スケーラブルなリソースの検索

自動検出または手動で、スケーリングプランに追加するリソースを選択します。 [情報](#)

メソッドの選択

- CloudFormation スタックによる検索
AWS CloudFormationによってプロビジョニングされたリソースを検索します。
- タグによる検索
適用されたタグを使用してリソースを検索します。
- Amazon EC2 Auto Scaling グループの選択
スケーリング計画に含めるため、Auto Scaling グループを 1 つ以上選択します。

Auto Scaling グループの選択 [情報](#)

Auto Scaling グループ

Auto Scaling グループを選択します。

myalbttestasg
myasgforssm

キャンセル 次へ



既存のAuto Scalingグループを選択して「次へ」

AWS Auto ScalingによるEC2 Auto Scalingグループの予測スケーリング (4)

AWS Auto Scaling > スケーリングプラン > スケーリングプランの作成

ステップ 1
スケーラブルなリソースの検索

ステップ 2
スケーリング戦略を指定します。

ステップ 3
詳細設定の設定 (オプション)

ステップ 4
確認と作成

スケーリング戦略を指定します。

スケーリング戦略を使用して、アプリケーションのスケーラブルなリソースを最適化する方法を定義します。 [情報](#)

スケーリングプランの詳細

名前 長さは 1~128 文字にする必要があり、パイプ文字「|」、コロン「:」、およびスラッシュ「/」を含めることはできません。

リソース
1 Auto Scaling グループ 選択されました。

Auto Scaling グループ (1)
1 Auto Scaling グループ のためにスケーリング戦略を指定します。 スケーリングプランに含める

スケーリングプラン名を入力

AWS Auto ScalingによるEC2 Auto Scalingグループの予測スケーリング (5)

Auto Scaling グループ (1)
1 Auto Scaling グループのためにスケーリング戦略を指定します。

スケーリングプランに含める

スケーリング戦略
その戦略では、リソースの拡張に使用するスケーリングメトリックとターゲット値を定義します。

可用性を考えた最適化
高い可用性を提供し需要の急増に対応できるキャパシティーを確保するため、Auto Scaling グループの平均 CPU 使用率を常に 40% に維持します。

可用性とコストのバランスを取ります
最適な可用性を提供しこストを削減するため Auto Scaling グループの平均 CPU 使用率を常に 50% に維持します。

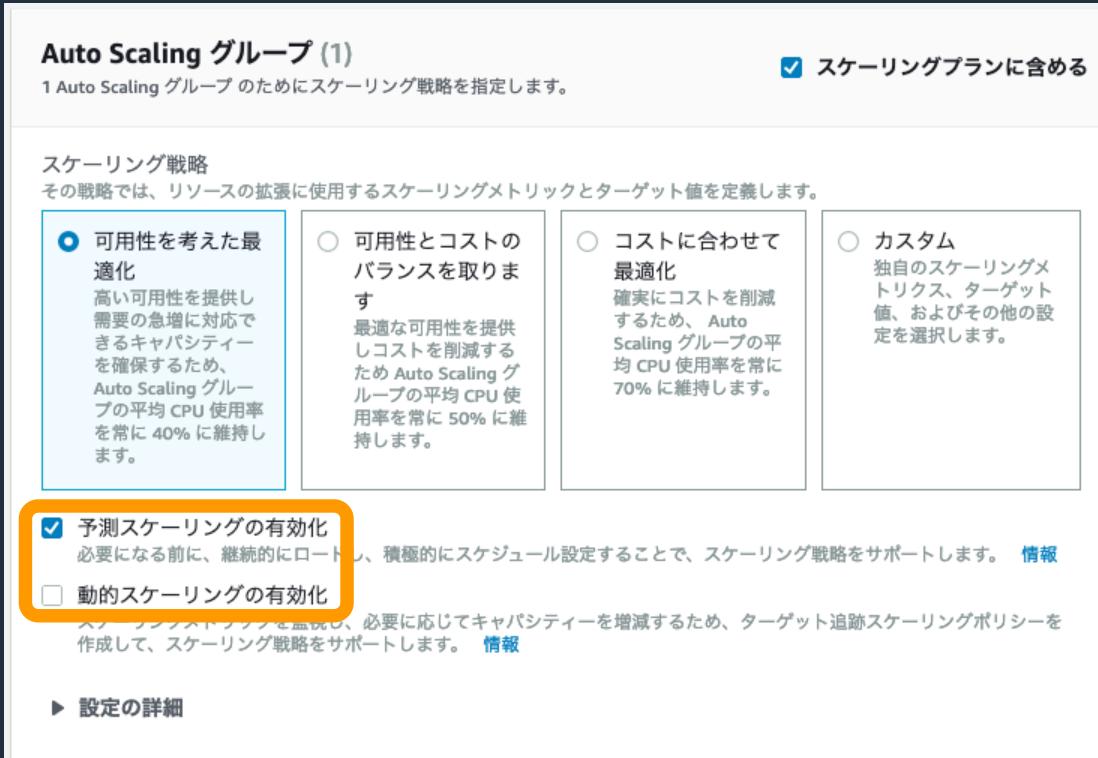
コストに合わせて最適化
確実にコストを削減するため、Auto Scaling グループの平均 CPU 使用率を常に 70% に維持します。

カスタム
独自のスケーリングメトリクス、ターゲット値、およびその他の設定を選択します。

予測スケーリングの有効化
必要になる前に、継続的にロートンし、積極的にスケジュール設定することで、スケーリング戦略をサポートします。 [情報](#)

動的スケーリングの有効化
ヘイブンメントメントを監視し、必要に応じてキャパシティーを増減するため、ターゲット追跡スケーリングポリシーを作成して、スケーリング戦略をサポートします。 [情報](#)

▶ 設定の詳細



- 「予測スケーリングの有効化」にチェックが入っていることを確認
- 「動的スケーリングの有効化」のチェックを外す

AWS Auto ScalingによるEC2 Auto Scalingグループの予測スケーリング (6)

ステップ 1
スケーラブルなリソースの検索

ステップ 2
スケーリング戦略を指定します。

ステップ 3
詳細設定の設定 (オプション)

ステップ 4
確認と作成

詳細設定の設定 (オプション)

個々のリソースまたは複数のリソースの設定を、同時にカスタマイズします。 [情報](#)

▶ Auto Scaling グループ (1)

Auto Scaling グループでは、カスタム設定が使用されます。予測スケーリングは有効です。

キャンセル 戻る 次へ



“Auto Scalingグループ”をクリックして展開

AWS Auto ScalingによるEC2 Auto Scalingグループの予測スケーリング (7)

詳細設定の設定 (オプション)
個々のリソースまたは複数のリソースの設定を、同時にカスタマイズします。 [情報](#)

▼ Auto Scaling グループ (1 個中 1 個を選択) [元に戻す](#)

カスタム設定を指定する 1 つ以上の Auto Scaling グループを選択します。

<input checked="" type="checkbox"/> リソース	▲	プランに含める	外部のスケーリングポリシーの置き換え	既存のポリシー
<input checked="" type="checkbox"/> myasgforssm		はい	なし	なし

1 個のリソースが選択されました

スケーリングプランに含める

▶ 全般設定

▶ 動的スケーリングの設定

▶ **予測スケーリング設定**

[キャンセル](#) [戻る](#) [次へ](#)

対象Auto Scalingグループを選択すると詳細設定メニューが展開されるので
「予測スケーリング設定」を展開する

AWS Auto ScalingによるEC2 Auto Scalingグループの予測スケーリング (8)

▼ 予測スケーリング設定

予測スケーリングモード
予測の実行にスケーリングを使用するかどうかを決定します。これはいつでも変更できます。 [情報](#)

予測のみ
予測とスケール
予測のみ

予測のみ

予測とスケール

予測のみ

最大キャパシティーの動作
予測キャパシティーが最大キャパシティーに近づいたか、それを超えたときに使用するルールを選択します。 [情報](#)

最大キャパシティー設... ▾

予測期間
事前予測する日数。 [情報](#)

2 日

予測の詳細度
予測とキャパシティーの計算間隔。 [情報](#)

60 分

予測頻度
予測更新の頻度。 [情報](#)

毎日

キャンセル 戻る 次へ

「予測のみ」を選択して「次へ」

AWS Auto ScalingによるEC2 Auto Scalingグループの予測スケーリング (9)

ステップ1
スケーラブルなリソースの検索

ステップ2
スケーリング戦略を指定します。

ステップ3
詳細設定の設定 (オプション)

ステップ4
確認と作成

確認と作成

スケーリングプランの詳細

名前
myfirstscalingplan

リソース
1 Auto Scaling グループ選択されました。1 個のスケーリングポリシーが作成され、0 個の外部ポリシーが維持されます。

Auto Scaling グループ

スケーリング戦略	スケーリングメトリクス	ターゲット値
可用性を考えた最適化	CPU の平均使用率	40 %

概要
お客様のスケーリングプランは、40 % で CPU の平均使用率 メトリクスを保持することで、1 Auto Scaling グループを最適化するように設定されています。

動的スケーリング
動的スケーリングは有効です。40 % で CPU の平均使用率 メトリクスを保持する必要に応じて、インスタンスを追加または削除するため、1 ターゲット追跡スケーリングが適用されます。

予測スケーリング
予測スケーリングは有効です。40 % で CPU の平均使用率 メトリクスを保持するために必要なインスタンスの最小数を維持するため、合計 CPU 使用率 メトリクスの予測に基づいて、スケジュールされたスケーリングアクションが生成されます。

▶ 詳細

キャンセル 戻る スケーリングプランの作成

AWS Auto ScalingによるEC2 Auto Scalingグループの予測スケーリング(10)



The screenshot shows the AWS Auto Scaling Scaling Plan page. At the top, there is a breadcrumb navigation: AWS Auto Scaling > スケーリングプラン. Below the navigation, a table displays a single scaling plan entry:

スケーリングプラン (1) 情報		名前	ステータス	スケーリングポリシー	作成時刻
<input type="checkbox"/>	<input type="checkbox"/> myfirstscalingplan	Active	<input checked="" type="checkbox"/> 1		2019-10-02 01:33:59 UTC+0900

The "myfirstscalingplan" row is highlighted with an orange border. The "Active" status is indicated by a green checkmark icon.

「ステータス」が"Active"になつたらスケーリングプラン名をクリック

AWS Auto ScalingによるEC2 Auto Scalingグループの予測スケーリング(11)

AWS Auto Scaling > スケーリングプラン > myfirstscalingplan

myfirstscalingplan

スケーリングプランの詳細

ステータス
④ Active

ステータスの説明
Scaling plan has been created and applied to all resources.

Auto Scaling グループ (1)

autoScalingGroup/myasgforssm

ステータス
④ アクティブ

スケーリングメトリクス
CPU の平均使用率

ターゲット値
40 %

1 時間 3 時間 12 時間 1 日 3 日 1 週

ダッシュボードに追加

合計 CPU 使用率 (%)

2.67
1.33
0

14:00 14:30 15:00 15:30 16:00 16:30

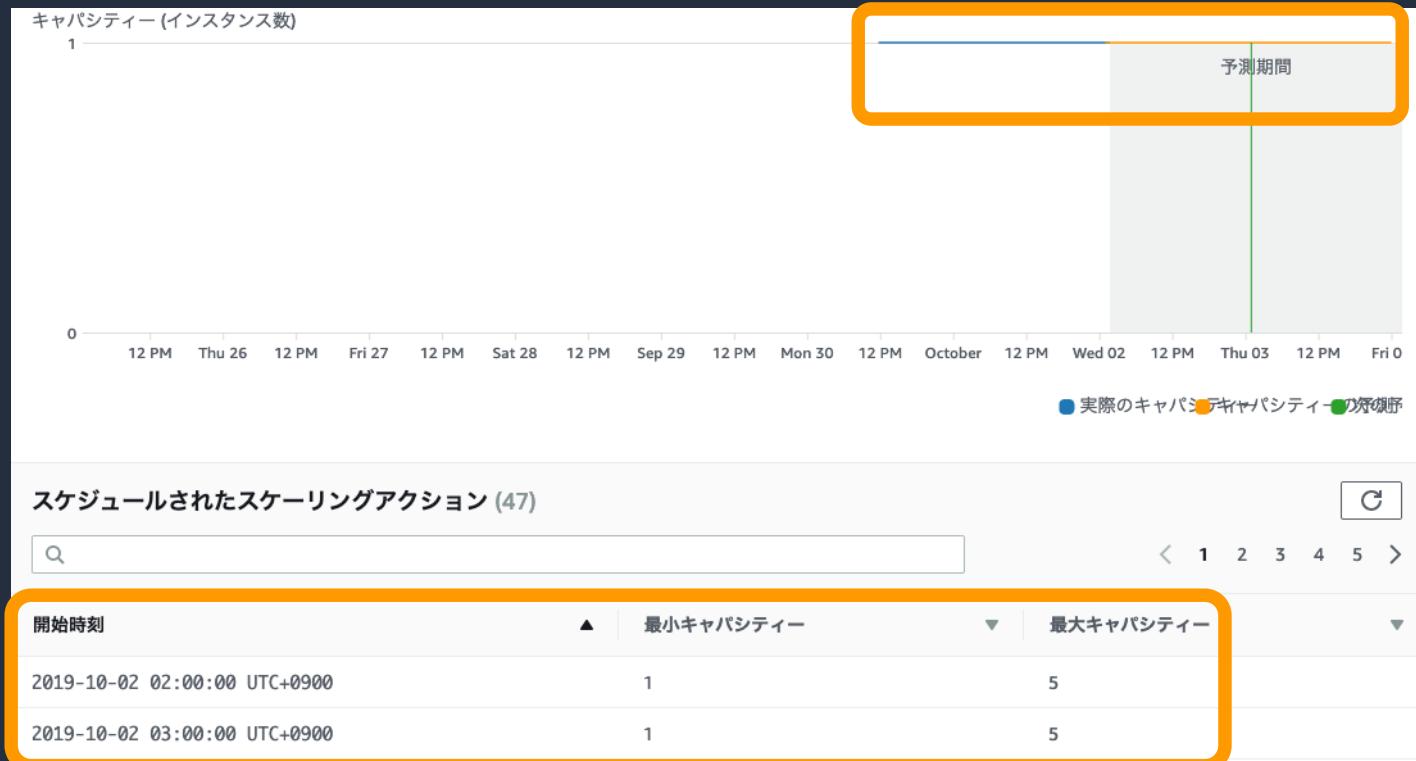
“Auto Scalingグループ”の下の対象Auto Scalingグループリソース名をクリック

AWS Auto ScalingによるEC2 Auto Scalingグループの予測スケーリング (12)



画面下にスクロールすると、向こう48時間の負荷の予測結果が表示される

AWS Auto ScalingによるEC2 Auto Scalingグループの予測スケーリング (13)



さらに下にスクロールすると、スケール予定のインスタンス数とそのためのスケジュールスケーリング設定の計画を確認できる

AWS Auto ScalingによるEC2 Auto Scalingグループの予測スケーリング (14)

AWS Auto Scaling > スケーリングプラン > myfirstscalingplan

myfirstscalingplan

編集

削除

スケーリングプランの詳細

ステータス

Active

ステータスの説明

Scaling plan has been created and applied to all resources.

- 「予測のみ」モードから「予測とスケーリング」モードへ変更
- 対象スケーリングプランを選択して「編集」

AWS Auto ScalingによるEC2 Auto Scalingグループの予測スケーリング(15)

AWS Auto Scaling > スケーリングプラン > myfirstscalingplan > 編集

myfirstscalingplan の編集

▶ Auto Scaling グループ (1)

すべての Auto Scaling グループでは、「可用性を考えた最適化」スケーリング戦略が使用されます。予測スケーリングは有効です。

キャンセル

次へ

“Auto Scalingグループ”をクリックして展開

AWS Auto ScalingによるEC2 Auto Scalingグループの予測スケーリング (16)

myfirstscalingplan の編集

▼ Auto Scaling グループ (1 個中 1 個を選択)
カスタム設定を指定する 1 つ以上の Auto Scaling グループを選択します。

<input checked="" type="checkbox"/>	リソース	▲	プランに含める	外部のスケーリングポリシーの置き換え
<input checked="" type="checkbox"/>	myalbtestasg	はい	なし	

1 個のリソースが選択されました

スケーリングプランに含める

▶ 全般設定

▶ 動的スケーリングの設定

▶ **予測スケーリング設定**

対象Auto Scalingグループを選択すると詳細設定メニューが展開されるので
「予測スケーリング設定」を展開する

AWS Auto ScalingによるEC2 Auto Scalingグループの予測スケーリング (17)

The screenshot shows the 'Predictive Scaling Settings' section of the AWS Auto Scaling configuration interface. A dropdown menu is open, with the 'Predict and Scale' option highlighted and surrounded by an orange rectangle. Other options in the dropdown are 'Predict and Scale' and 'Predict only'. Below the dropdown, there is a field set with a value of '5' and a unit of '分' (minutes). To the right, there are sections for 'Maximum Capacity Action' (describing how it applies rules when capacity approaches or exceeds maximum), 'Prediction Detail' (set to 60 minutes), 'Prediction Frequency' (set to daily), and two buttons at the bottom: 'Cancel' and an orange 'Next Step' button.

▶ 全般設定

▶ 動的スケーリングの設定

▼ 予測スケーリング設定

予測スケーリングモード
予測の実行にスケーリングを使用するかどうかを決定します。これはいつでも変更できます。 [情報](#)

予測とスケール

予測とスケール

予測のみ

最大キャパシティーの動作
予測キャパシティーが最大キャパシティーに近づいたか、それを超えたときに使用するルールを選択します。 [情報](#)

予測の詳細度
予測とキャパシティーの計算間隔。 [情報](#)

60 分

最大キャパシティー設定を適用する

予測期間
事前予測する日数。 [情報](#)

毎日

5 分

キャンセル

次へ

「予測とスケール」を選択して「次へ」

AWS Auto ScalingによるEC2 Auto Scalingグループの予測スケーリング (18)

AWS Auto Scaling > スケーリングプラン > myfirstscalingplan > 編集

myfirstscalingplan の編集

Auto Scaling グループ

スケーリング戦略 可用性を考えた最適化	スケーリングメトリクス CPU の平均使用率	ターゲット値 40 %
------------------------	---------------------------	----------------

概要
動的スケーリングが有効になっている Auto Scaling グループがありません。スケーリングプランで 1 Auto Scaling グループ のターゲット追跡スケーリングポリシーを作成できるようにするには、先に動的スケーリングを有効にする必要があります。

動的スケーリング
動的スケーリングは無効です。

予測スケーリング
予測スケーリングは有効です。 **40 %** で **CPU の平均使用率** メトリクスを保持するために必要なインスタンスの最小数を維持するため、**合計 CPU 使用率** メトリクスの予測に基づいて、スケジュールされたスケーリングアクションが生成されます。

▶ 詳細

キャンセル 戻る **変更の保存**

「変更の保存」

AWS Auto ScalingによるEC2 Auto Scalingグループの予測スケーリング(19)

The screenshot shows the AWS Auto Scaling Groups console for the group 'myalbtestasg'. The 'Scheduled Actions' tab is selected. A yellow box highlights four scheduled actions listed in the table below:

名前	開始時刻	終了時刻	繰り返し	希望するキャパシティ	最小	最大
AutoScaling-myfirstscalingplan-1-201910011800	2019 October 2 03:00:00 UTC+9			1	5	
AutoScaling-myfirstscalingplan-1-201910011900	2019 October 2 04:00:00 UTC+9			1	5	
AutoScaling-myfirstscalingplan-1-201910012000	2019 October 2 05:00:00 UTC+9			1	5	
AutoScaling-myfirstscalingplan-1-201910012100	2019 October 2 06:00:00 UTC+9			1	5	

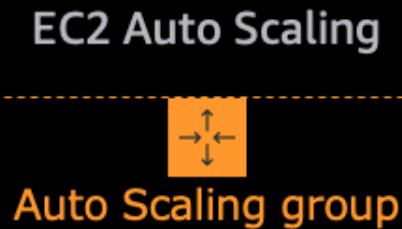
EC2 Auto Scalingマネジメントコンソールから
「スケジュールされたアクション」に毎時のアクションが設定されたことを確認

本日のアジェンダ

- Auto Scalingサービスのコンセプト
- Auto Scalingの基礎知識
- 主要機能：スケーリングの整理
- Auto Scalingを使ってみる
- こんなときどうする？ - 各種機能の紹介
- まとめ・参考資料

こんなときどうする？

- (EC2 Auto Scaling) スポットインスタンスを活用したいです
 - →ミックスインスタンスグループを活用してください



こんなときどうする？

- (EC2 Auto Scaling) 「起動設定」と「起動テンプレート」のどちらを使えば良いか
 - → 「起動テンプレート」を強く推奨します！

こんなときどうする？

- (EC2 Auto Scaling) 速やかにスケールアウト(スケールイン)してくれません
 - →インスタンスの詳細モニタリングを有効にしてください
- CloudWatch Metricsを1分粒度にする。5分粒度では速やかにスケールできない
- 有料オプションながらAuto Scalingを使用する際のベストプラクティス

https://docs.aws.amazon.com/ja_jp/AWSEC2/latest/UserGuide/using-cloudwatch-new.html

こんなときどうする？

- (EC2 Auto Scaling) 正常に動作しないインスタンスを自動的に置き換える
• →ヘルスチェックを活用します
- 特に指定しない場合、EC2ヘルスチェックが有効になっている
 - 2/2以外のステータスが続くとAuto Scalingが置き換える
- ELB配下のASGの場合、ELBヘルスチェックを有効にする
 - EC2ヘルスチェックに加え、ELBからのヘルスチェックに応答しない場合の速やかな入れ替えが可能になる

https://docs.aws.amazon.com/ja_jp/autoscaling/ec2/userguide/healthcheck.html

こんなときどうする？

- (EC2 Auto Scaling)スケールイン・スケールアウトを繰り返してしまい、いつまでたってもインスタンスが追加されない
 - → 「ヘルスチェックの猶予期間」の設定を見直す
- ヘルスチェックの猶予期間：起動したばかりでヘルスチェックに応答できないインスタンスを保護する期間
 - /index.html などは速やかに返せるようになるが、S3からのコンテンツ配備やDB接続などが整った前提のヘルスチェックパスを指定している場合は準備期間が必要
 - 特にELBヘルスチェックにアプリケーションのパスを採用している場合に有効
- デフォルトは5分(300秒)

https://docs.aws.amazon.com/ja_jp/autoscaling/ec2/userguide/healthcheck.html

こんなときどうする？

- (EC2 Auto Scaling) 次にどのインスタンスがスケールイン対象になるか知りたい
 - →デフォルトの終了ポリシー
- おおまかには次の流れで決まる
 1. インスタンスが最も多いアベイラビリティゾーンを選択
 2. (そのアベイラビリティゾーンに候補が複数あるなら) 最も古い起動設定・起動テンプレートから起動されたインスタンスを選択
 3. (複数候補が残っている場合) 次のインスタンス時間に近いものを選択
 4. まだ複数いるならランダム
- カスタマイズも可能

https://docs.aws.amazon.com/ja_jp/autoscaling/ec2/userguide/as-instance-termination.html

こんなときどうする？

- (EC2 Auto Scaling) 特定のインスタンスをスケールインから保護したい
 - →インスタンスの保護
- ASG単位、もしくはインスタンス単位で設定。スケールインされなくなる
- 次の条件からは保護できないことに注意
 - 手動でのインスタンス削除(Terminate)
 - ヘルスチェックによる置き換え
 - スポットインスタンスの中止
- すべてのインスタンスが終了保護された状態でスケールインイベントが発生した場合、希望容量だけが減少し、スケールイン(インスタンス削除)は行われない

https://docs.aws.amazon.com/ja_jp/autoscaling/ec2/userguide/as-instance-termination.html#instance-protection

こんなときどうする？

- (EC2 Auto Scaling) 一時的にスケールインやスケールアウトを止めたい
 - →スケーリングプロセスの中斷
- 一時的にスケール動作を停止できる
- ASG単位で設定
- 中断できるプロセス一覧：Launch, Terminate, AddToLoadBalancer, AlarmNotification, AZRebalance, HealthCheck, ReplaceUnhealthy, ScheduledActions
- 使いどころ：機能テストなど、一時的にAuto Scalingグループの特定プロセスの動作を止めてテスト条件を整えたい場合
 - LaunchとTerminateの両方のプロセスを中断することで、「何もしない」Auto Scalingグループを作り出せる
- 動作のおかしいインスタンスがいるのでスケールイン・スケールアウトを止めたい
 - →プロセスの中斷ではなく次の項目を参照

https://docs.aws.amazon.com/ja_jp/autoscaling/ec2/userguide/as-suspend-resume-processes.html

こんなときどうする？

- (EC2 Auto Scaling) このインスタンスをAuto Scalingグループから外したい
 - →スタンバイ、もしくはデタッチ
- スタンバイ(「一時的なインスタンスの削除」)
 - インスタンス単位で設定
 - そのインスタンスはAuto Scalingグループにいながら「スタンバイ」状態に入る
 - 具体的にはそのインスタンスはELBから登録解除され、ヘルスチェック対象から外される。
そのAuto Scalingグループの希望容量は1つ減少する
 - その間にインスタンスのトラブルシューティングなどを行う
- デタッチ
 - インスタンス単位で設定
 - そのインスタンスはそのAuto Scalingグループのメンバーから外れる
 - スタンバイと実質的な効果は同一。インスタンスはそのままRunning状態で保持される。ただしデタッチの場合、Auto Scalingグループとして与えていたタグも除去される

https://docs.aws.amazon.com/ja_jp/autoscaling/ec2/userguide/as-enter-exit-standby.html
https://docs.aws.amazon.com/ja_jp/autoscaling/ec2/userguide/detach-instance-asg.html

こんなときどうする？

- (EC2 Auto Scaling) スケールアウトした後、サービス開始前にインスタンスに準備させたい / スケールインの前にログ退避させたいのでTerminateを少し待って欲しい
 - →ライフサイクルフック
- ライフサイクルフック：インスタンス起動時・削除時にインスタンスを一時停止し、カスタムアクションを実行できる
- ライフサイクルフックはAuto Scalingグループ単位に設定
- 実際のライフサイクルフックによる待機はインスタンスごと
- 実装例：CloudWatch Eventからライフサイクル通知を受け取り、Lambdaがカスタムアクションを実行する

https://docs.aws.amazon.com/ja_jp/autoscaling/ec2/userguide/lifecycle-hooks.html

こんなときどうする？

- (EC2 Auto Scaling) スケールアウトを素早くしたい
 - →ユーザーデータでのyum updateやyum installなどを用いず、なるべくミドルウェアや必要設定などを済ませた状態のAMIを起動テンプレートに指定する(いわゆるゴールデンイメージ)
 - ただし考慮点があるので次のスライドで説明

こんなときどうする？

- (EC2 Auto Scaling) WindowsやRed Hat Enterprise Linuxなどの考慮点は？
 - 1時間単位の課金になるため、終了ポリシーはデフォルトでお使いいただくのをお勧めする
 - 起動時間を短縮する際、ゴールデンイメージの起動時間と、標準AMIからの起動+ユーザーデータでセットアップした場合とを比較すると良い
 - 2019年現在、標準AMIはカスタマイズしたAMIより素早く起動できるようにチューニングされている
 - 場合によってはユーザーデータの方が速い可能性も

本日のアジェンダ

- Auto Scalingサービスのコンセプト
- Auto Scalingの基礎知識
- 主要機能：スケーリングの整理
- Auto Scalingを使ってみる
- こんなときどうする？ - 各種機能の紹介
- まとめ・参考資料

本日のまとめ

- Auto Scalingの価値
 - アプリケーションの可用性の維持
 - アベイラビリティゾーン間でのインスタンスの分散、異常なインスタンスの自動置き換え
 - 自動的なキャパシティの増減
 - 動的なスケーリング、予測スケーリング、スケジューリングスケーリング
 - 予測スケーリングとターゲット追跡スケーリングの組み合わせは2019年におススメする推奨セット
 - 様々なユースケースをカバーする機能群
 - コスト最適化のためのミックスインスタンスグループ、ライフサイクルフック
- Auto Scalingを使いこなし、クラウドの世界の本質をぜひ実感してください

参考資料

よくある質問

- よくある質問 - Amazon EC2 Auto Scaling | AWS — <https://aws.amazon.com/jp/ec2/autoscaling/faqs/>
- よくある質問 - AWS Auto Scaling | AWS — <https://aws.amazon.com/jp/autoscaling/faqs/>

ユーザーガイド

- Amazon EC2 Auto Scaling とは - Amazon EC2 Auto Scaling (日本語) —
https://docs.aws.amazon.com/ja_jp/autoscaling/ec2/userguide/what-is-amazon-ec2-auto-scaling.html
- Application Auto Scaling とは - Application Auto Scaling —
https://docs.aws.amazon.com/ja_jp/autoscaling/application/userguide/what-is-application-auto-scaling.html
 - 各サービスでのApplication Auto Scalingの使い方・考慮点は以下のリンクから
 - ご利用開始にあたって - Application Auto Scaling —
https://docs.aws.amazon.com/ja_jp/autoscaling/application/userguide/what-is-application-auto-scaling.html#getting-started
- AWS Auto Scaling とは - AWS Auto Scaling —
https://docs.aws.amazon.com/ja_jp/autoscaling/plans/userguide/what-is-aws-auto-scaling.html

Q&A

お答えできなかったご質問については
AWS Japan Blog 「<https://aws.amazon.com/jp/blogs/news/>」にて
後日掲載します。

AWS の日本語資料の場所「AWS 資料」で検索



The screenshot shows the AWS Japan Language Resources page. At the top, there's a navigation bar with the AWS logo, search bar, and links for "日本担当チームへお問い合わせ", "サポート", "日本語", "アカウント", and "コンソールにサインイン". Below the navigation is a horizontal menu with links for "製品", "ソリューション", "料金", "ドキュメント", "学習", "パートナー", "AWS Marketplace", "その他", and a search icon. The main content area features a large title "AWS クラウドサービス活用資料集トップ" and a descriptive paragraph about the service. At the bottom, there are four call-to-action buttons: "AWS Webinar お申込", "AWS 初心者向け", "業種・ソリューション別資料", and "サービス別資料".

AWS クラウドサービス活用資料集トップ

アマゾン ウェブ サービス (AWS) は安全なクラウドサービスプラットフォームで、ビジネスのスケールと成長をサポートする処理能力、データベースストレージ、およびその他多種多様な機能を提供します。お客様は必要なサービスを選択し、必要な分だけご利用いただけます。それらを活用するために役立つ日本語資料、動画コンテンツを多数ご提供しております。(本サイトは主に、AWS Webinar で使用した資料およびオンデマンドセミナー情報を掲載しています。)

AWS Webinar お申込 »

AWS 初心者向け »

業種・ソリューション別資料 »

サービス別資料 »

<https://amzn.to/JPArchive>

AWS Well-Architected 個別技術相談会

毎週”W-A個別技術相談会”を実施中

- AWSのソリューションアーキテクト(SA)に
対策などを相談することも可能

• 申込みはイベント告知サイトから

(<https://aws.amazon.com/jp/about-aws/events/>)

AWS イベント で[検索]

AWS Well-Architected



ご視聴ありがとうございました

AWS 公式 Webinar
<https://amzn.to/JPWebinar>



過去資料
<https://amzn.to/JPArchive>





このコンテンツは公開から3年以上経過しており内容が古い可能性があります
最新情報については[サービス別資料](#)もしくはサービスのドキュメントをご確認ください

[AWS Black Belt Online Seminar]

Amazon EC2 Auto Scaling & AWS Auto Scaling

サービスカットシリーズ
ソリューションアーキテクト
滝口 開資
2019-10-02

AWS 公式 Webinar
<https://amzn.to/JPWebinar>



過去資料
<https://amzn.to/JPArchive>



自己紹介

滝口 開資 (たきぐちはるよし)

ソリューションアーキテクト - EC2スポットインスタンススペシャリスト

日本市場でのEC2スポットインスタンス技術担当

好きなAWSサービス

- Amazon EC2 Auto Scaling
- AWS Auto Scaling
- AWSサポート



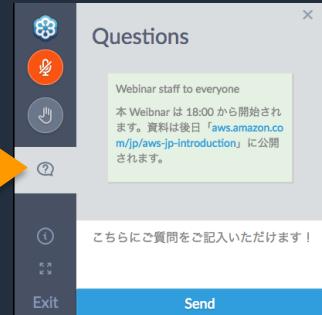
AWS Black Belt Online Seminar とは

「サービス別」「ソリューション別」「業種別」のそれぞれのテーマに分かれて、Amazon ウェブ サービス ジャパン株式会社が主催するオンラインセミナーシリーズです。

質問を投げることができます！

- 書き込んだ質問は、主催者にしか見えません
- 今後のロードマップに関するご質問はお答えできませんのでご了承下さい

- ①吹き出しをクリック
- ②質問を入力
- ③Sendをクリック



Twitter ハッシュタグは以下をご利用ください
#awsblackbelt

内容についての注意点

- 本資料では2018年x月x日時点のサービス内容および価格についてご説明しています。最新の情報はAWS公式ウェブサイト(<http://aws.amazon.com>)にてご確認ください。
- 資料作成には十分注意しておりますが、資料内の価格とAWS公式ウェブサイト記載の価格に相違があった場合、AWS公式ウェブサイトの価格を優先とさせていただきます。
- 価格は税抜表記となっています。日本居住者のお客様が東京リージョンを使用する場合、別途消費税をご請求させていただきます。
- AWS does not offer binding price quotes. AWS pricing is publicly available and is subject to change in accordance with the AWS Customer Agreement available at <http://aws.amazon.com/agreement/>. Any pricing information included in this document is provided only as an estimate of usage charges for AWS services based on certain information that you have provided. Monthly charges will be based on your actual use of AWS services, and may vary from the estimates provided.

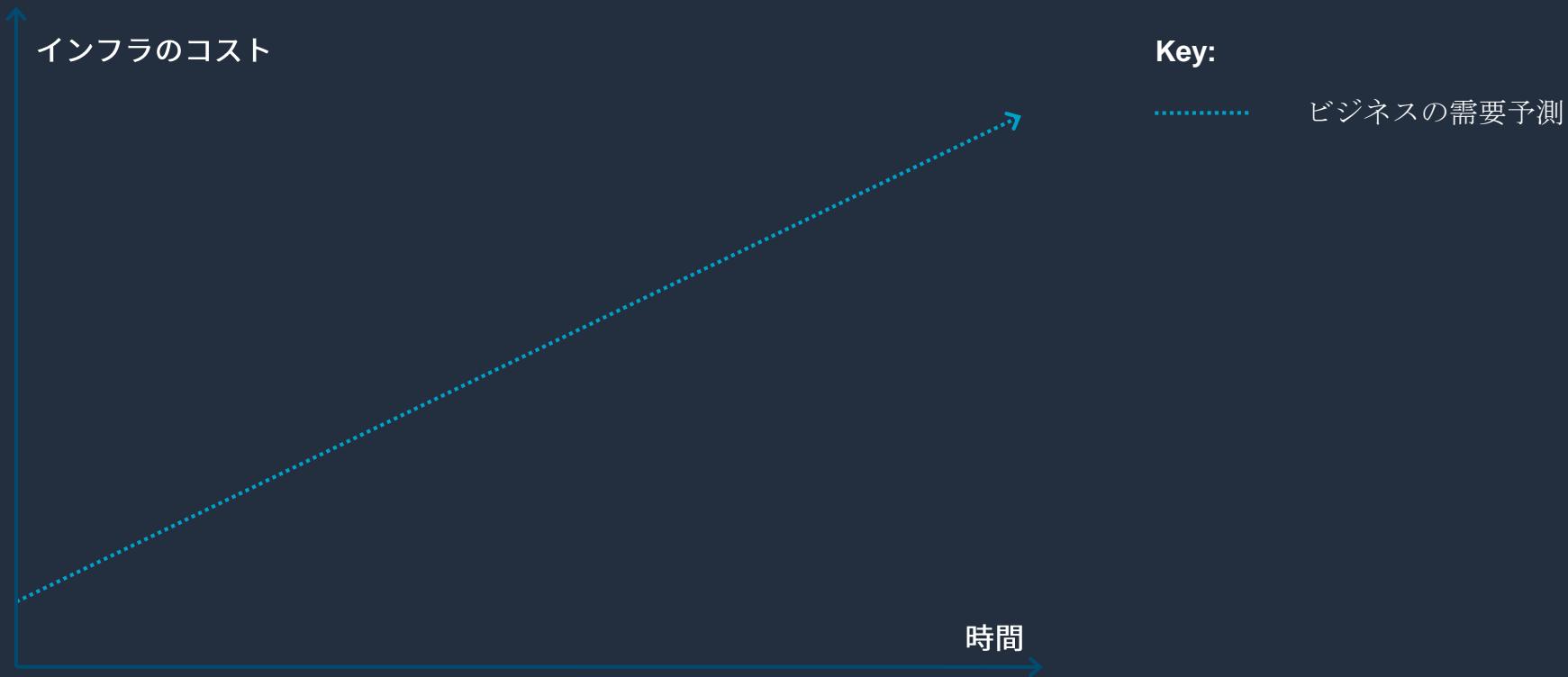
本日のアジェンダ

- Auto Scalingサービスのコンセプト
- Auto Scalingの基礎知識
- 主要機能：スケーリングの整理
- Auto Scalingを使ってみる
- こんなときどうする？ - 各種機能の紹介
- まとめ・参考資料

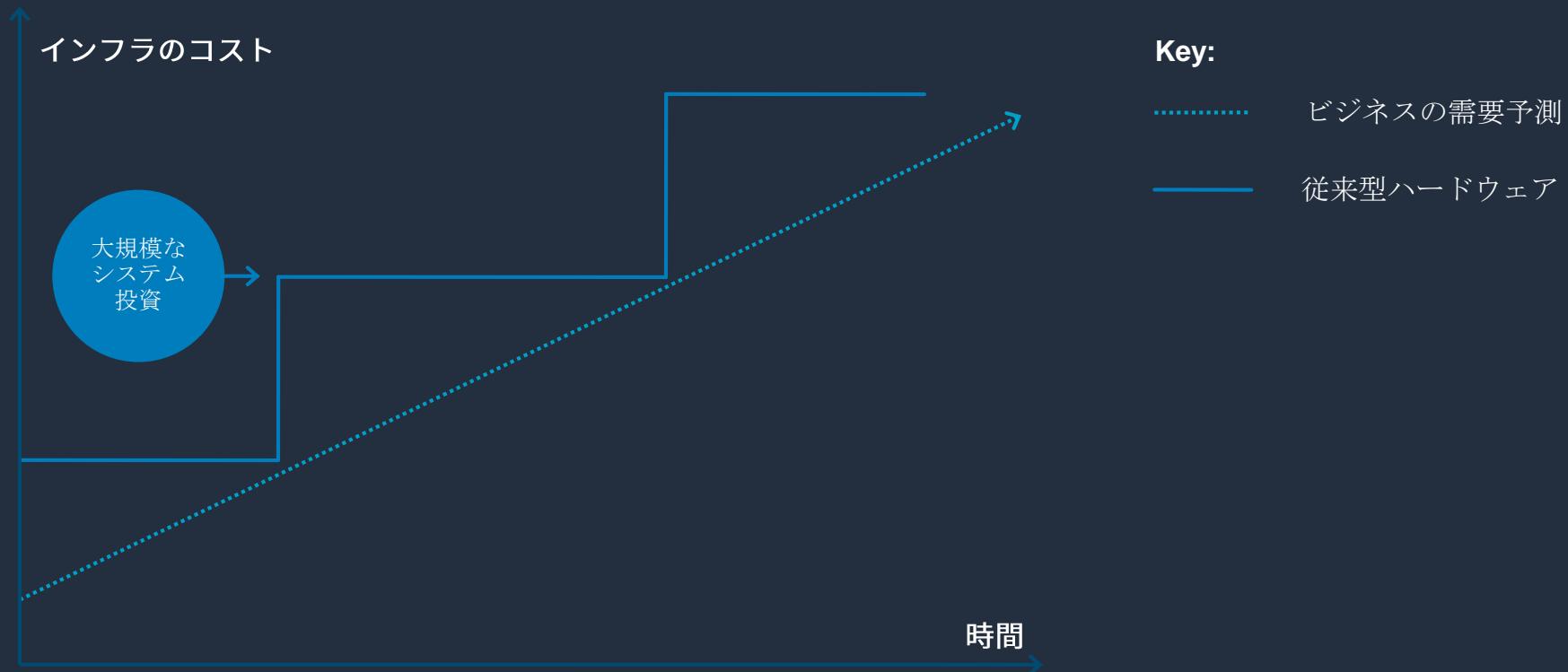
本日のアジェンダ

- Auto Scalingサービスのコンセプト
- Auto Scalingの基礎知識
- 主要機能：スケーリングの整理
- Auto Scalingを使ってみる
- こんなときどうする？ - 各種機能の紹介
- まとめ・参考資料

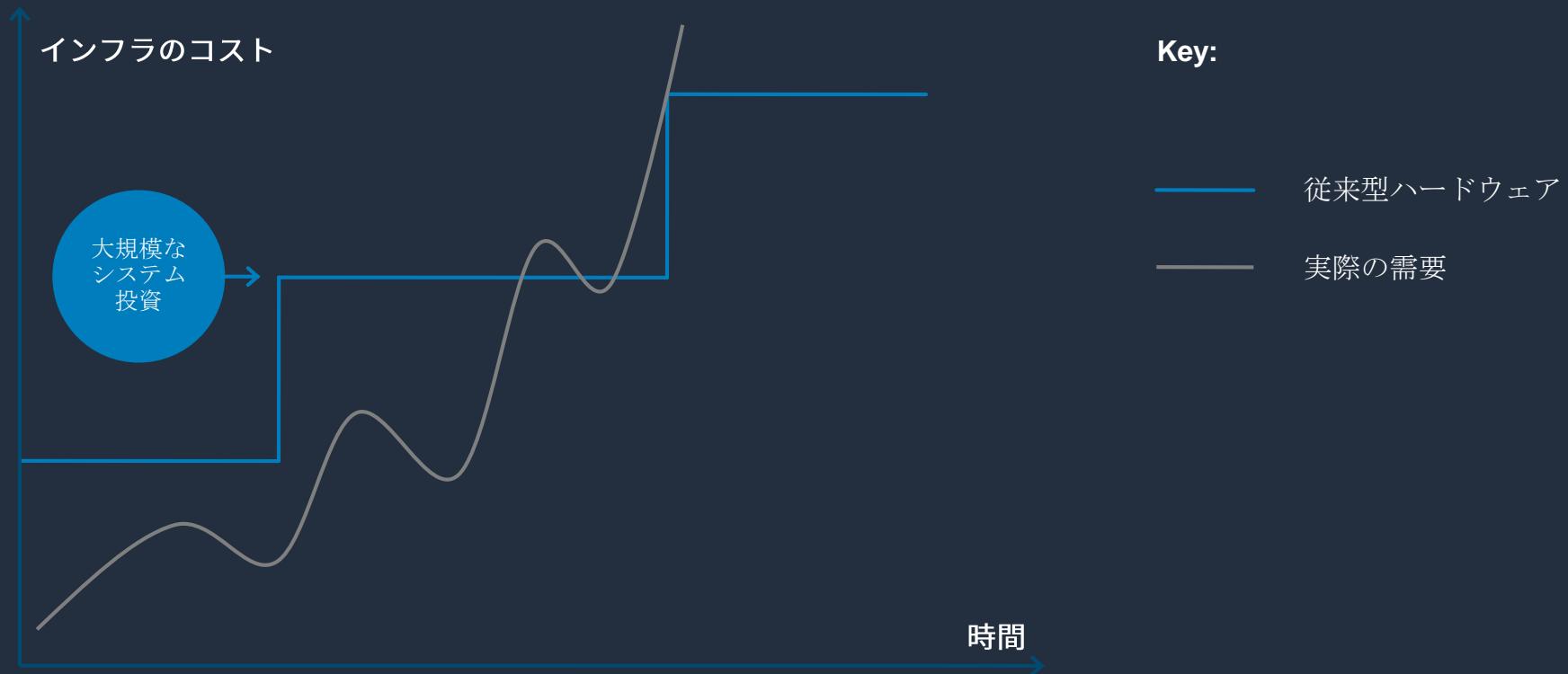
ビジネス需要に応じたキャパシティ準備



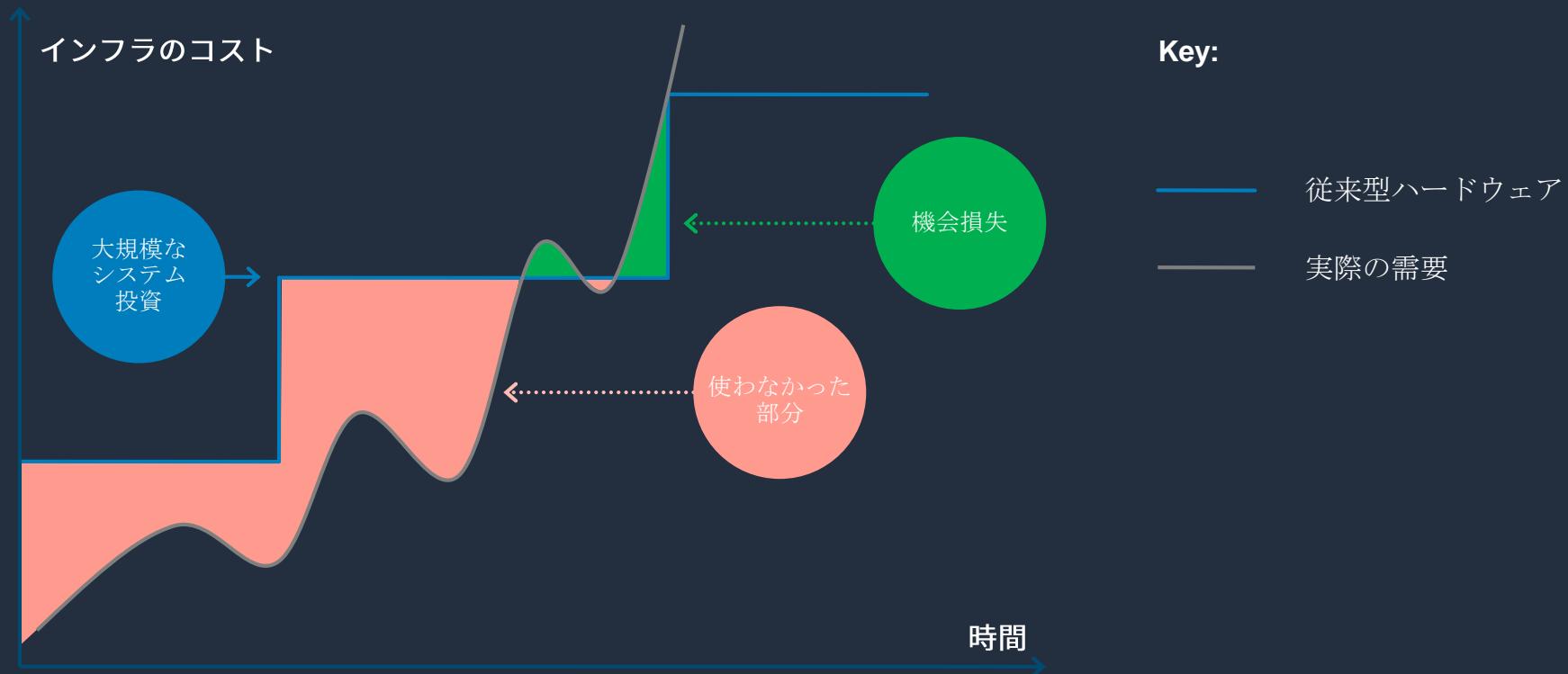
ビジネス需要に応じたキャパシティ準備



ビジネス需要に応じたキャパシティ準備



ビジネス需要に応じたキャパシティ準備



ビジネス需要に応じたキャパシティ準備



本日のアジェンダ

- Auto Scalingサービスのコンセプト
- Auto Scalingの基礎知識
- 主要機能：スケーリングの整理
- Auto Scalingを使ってみる
- こんなときどうする？ - 各種機能の紹介
- まとめ・参考資料

Auto Scalingの基礎知識

- 動作原理 - 希望する容量 (Desired Capacity, 以下「希望容量」) を目標に
- インスタンスの分散
- 均質性 - 「名前をつけてかわいがらない」

Auto Scalingの基礎知識

- 動作原理 - 希望する容量 (Desired Capacity, 以下「希望容量」) を目標に
- インスタンスの分散
- 均質性 - 「名前をつけてかわいがらない」

動作原理

Auto Scalingは、

1. 希望容量と現実の起動台数との差を監視し、
2. 常に希望容量に合致するようにリソース(EC2インスタンスなど)を増減する

1) 静観



2) スケールアウト



3) スケールイン



希望容量の使われ方

- サイズの維持
- 手動スケーリング
- 自動スケーリング

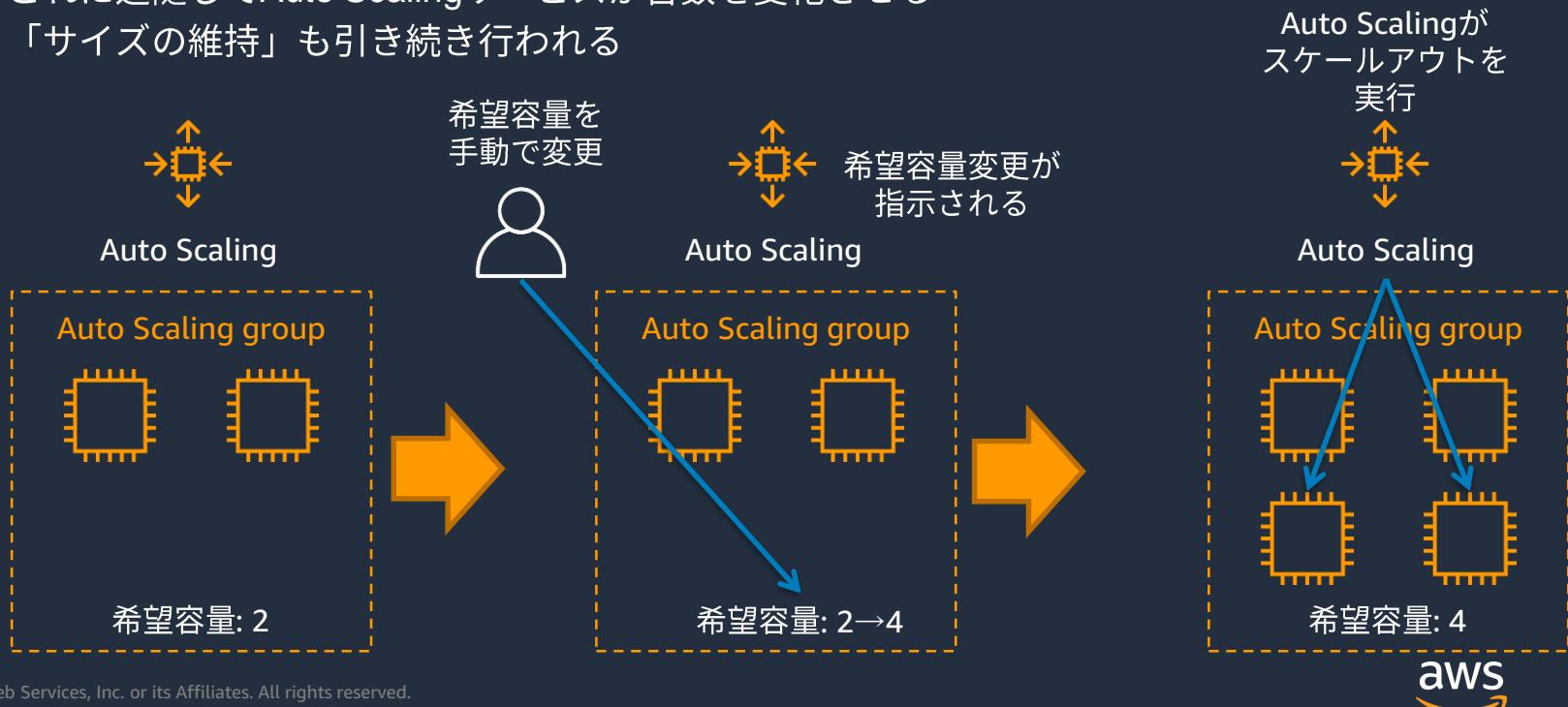
希望容量の使われ方

- ・ サイズの維持
 - ・ 希望容量は固定
 - ・ 現実の台数が減るとその差分を検知して1台追加する
 - ・ 一番シンプルな使い方



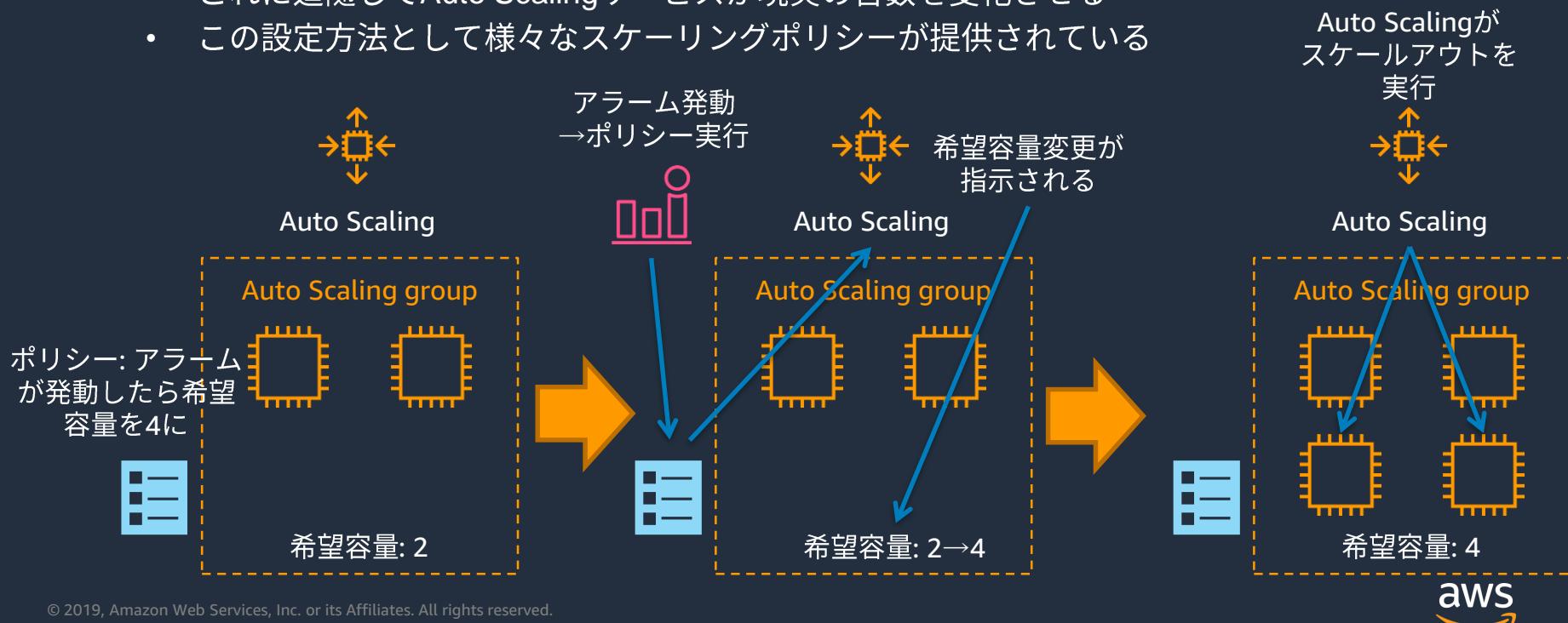
希望容量の使われ方

- 手動スケーリング
 - 希望容量を手動で変更する
 - これに追随してAuto Scalingサービスが台数を変化させる
 - 「サイズの維持」も引き続き行われる



希望容量の使われ方

- 自動スケーリング
 - 様々な条件に応じて希望容量が動的に変化する
 - これに追随してAuto Scalingサービスが現実の台数を変化させる
 - この設定方法として様々なスケーリングポリシーが提供されている



Auto Scalingの基礎知識

- 動作原理 - 希望する容量 (Desired Capacity, 以下「希望容量」) を目標に
- インスタンスの分散
- 均質性 - 「名前をつけてかわいがらない」

インスタンスの分散

- 使用できるアベイラビリティゾーンの間で、均等にインスタンスを配置しようとする
 - スケールアウトするとき：インスタンス数が最も少ないアベイラビリティゾーンに新規起動
 - これに失敗する場合、別のアベイラビリティゾーンを選択



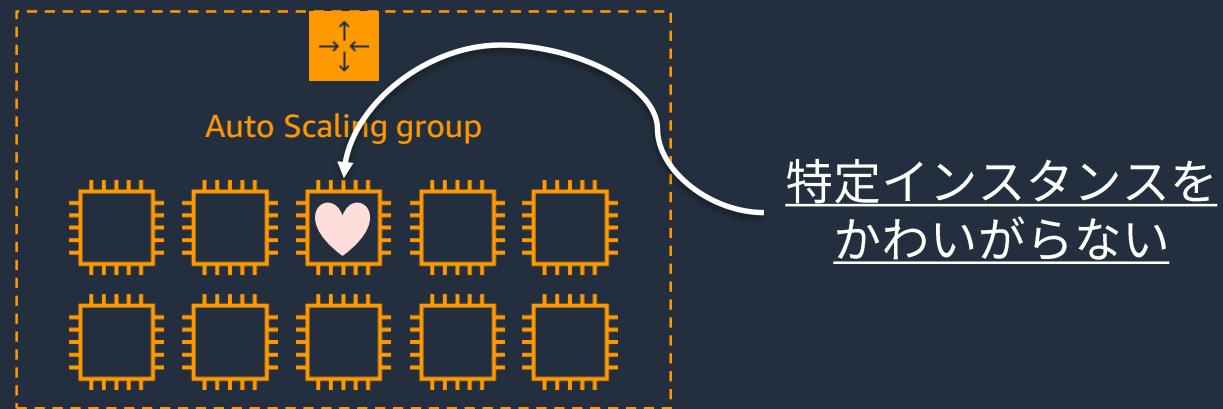
EC2 Auto Scalingはスケール動作時に
「インスタンスの分散」を最も重視する

Auto Scalingの基礎知識

- 動作原理 - 希望する容量 (Desired Capacity, 以下「希望容量」) を目標に
- インスタンスの分散
- 均質性 - 「名前をつけてかわいがらない」

均質性 - 「名前をつけてかわいがらない」

- Auto Scalingグループ内のインスタンスは原則として全て均一で、同一の価値を持つ
- 名前を付けるのはバッドプラクティス。スケールインはいつでも発生しうるものとして、置き換え可能にしておくのが重要



Auto Scalingの世界の整理

EC2 Auto
Scaling

EC2インスタンス

Auto Scalingの世界の整理

EC2 Auto Scaling

EC2インスタンス

Application Auto Scaling

ECSクラスター、スポットフリート、
EMRクラスター、AppStream 2.0フリー
ト、DynamoDBテーブル、Auroraレプリ
カ、SageMakerエンドポイントバリアン
ト、カスタムリソース

Auto Scalingの世界の整理

AWS Auto Scaling

様々なリソース

スケーリングプラン

(動的スケーリング+予測スケーリング)

EC2 Auto Scaling

EC2インスタンス

Application Auto Scaling

ECSクラスター、スポットフリート、
EMRクラスター、AppStream 2.0フリー
ト、DynamoDBテーブル、Auroraレプリ
カ、SageMakerエンドポイントバリアン
ト、カスタムリソース

Auto Scalingの世界の整理

AWS Auto Scaling

様々なリソース

スケーリングプラン

(動的スケーリング+予測スケーリング)

予測スケーリングの管理
(EC2のみ)

EC2の
管理

EC2 Auto Scaling

EC2インスタンス

Application Auto Scaling

ECSクラスター、スポットフリート、
EMRクラスター、AppStream 2.0フリー
ト、DynamoDBテーブル、Auroraレプリ
カ、SageMakerエンドポイントバリアン
ト、カスタムリソース

その他
リソース
の管理

本日のアジェンダ

- Auto Scalingサービスのコンセプト
- Auto Scalingの基礎知識
- 主要機能：スケーリングの整理
- Auto Scalingを使ってみる
- こんなときどうする？ - 各種機能の紹介
- まとめ・参考資料

主要機能：スケーリングの整理

- 動的なスケーリング
 - 簡易スケーリング
 - ステップスケーリング
 - ターゲット追跡スケーリング
- 予測スケーリング
- スケジュールスケーリング
- スケーリングオプションの選択指針

動的なスケーリング – 簡易スケーリング

- EC2 Auto Scalingのみ
- 1つのメトリクスに対して1種類だけのスケーリング調整値を指定
 - 例: CPUUtilizationが50%にならば1台追加

mysimplescalingpolicy

操作 ▾

ポリシータイプ: 簡易スケーリング

次の場合にポリシーを 実行:

awsec2-myalbtestasg-CPU-
アラームしきい値を超えると: CPUUtilization >= 50 (連続する 300 秒 x 1)
(メトリックスのディメンション) AutoScalingGroupName = myalbtestasg

アクションを実行: 追加 1 インスタンス

その後待機: 300 次のスケーリング動作までの秒数

- 現在は非推奨であり、ステップスケーリングを推奨[1]
 - 「スケーリング調整が 1 つの場合でも、簡易スケーリングポリシーではなくステップスケーリングポリシーを使用することをお勧めします。」
 - 「ほとんどの場合、ステップスケーリングポリシーは簡易スケーリングポリシーよりも適しています。」

[1] https://docs.aws.amazon.com/ja_jp/autoscaling/ec2/userguide/as-scaling-simple-step.html

動的なスケーリング - ステップスケーリング (1)

- EC2 Auto Scaling, Application Auto Scaling
- 1つのメトリクスに対して複数のスケーリング調整値を指定可能
- きめ細やかな設定が可能

mystepscalingpolicy

操作 ▾

ポリシータイプ: ステップスケーリング

次の場合にポリシーを実行: awsec2-myalbtestasg-CPU-アラームしきい値を超えるました: CPUUtilization >= 50 (連続する 300 秒 x 1)
(メトリックスのディメンション) AutoScalingGroupName = myalbtestasg

アクションを実行:

- 追加 1 インスタンス 次の条件の場合 50 <= CPUUtilization < 60
- 追加 2 インスタンス 次の条件の場合 60 <= CPUUtilization < 70
- 追加 3 インスタンス 次の条件の場合 70 <= CPUUtilization < 80
- 追加 4 インスタンス 次の条件の場合 80 <= CPUUtilization < +無限大

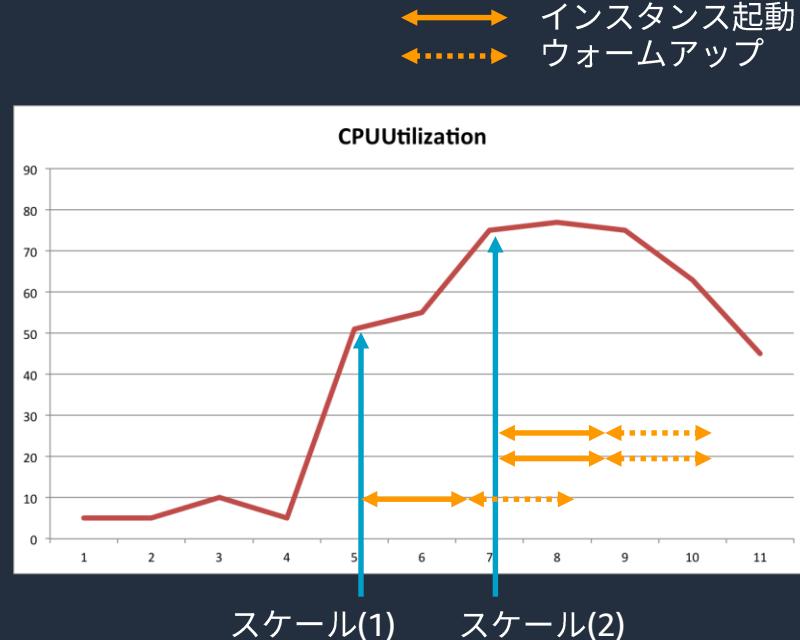
インスタンスは: 300 秒のウォームアップが各ステップ後に必要です

https://docs.aws.amazon.com/ja_jp/autoscaling/ec2/userguide/as-scaling-simple-step.html

https://docs.aws.amazon.com/ja_jp/autoscaling/application/userguide/application-auto-scaling-step-scaling-policies.html

動的なスケーリング - ステップスケーリング (2)

- ウォームアップ期間：新しいインスタンスがサービス開始できるようになるまでに何秒を要するかを設定する値
 - スケール(1)のウォームアップ期間中に次のアラームが来てスケール(2)が開始される。このとき、3台追加ではなく、「今1台追加中」とみなし、差し引き2台を追加する
 - これにより追加しすぎ問題を解決できる
- スケールアウトのタイミングで、一つ前のスケールアウトが進行中かどうかを判断してくれる、と考えても良い
- デフォルト値は300秒



ステップスケーリングポリシー定義
1台追加: $50 \leq \text{CPUUtil} < 60$
2台追加: $60 \leq \text{CPUUtil} < 70$
3台追加: $70 \leq \text{CPUUtil} < 80$

動的なスケーリング – ターゲット追跡スケーリング (1/3)

- EC2 Auto Scaling, Application Auto Scaling
- 1つのメトリクスに対し、単に目標値を指定するのみで良い
 - CPUUtilizationを40%に維持して欲しい。ただこれだけ

AutoScaling-albtest1-58558132-caa3-4ee0-a475-a340c4dcf26d 操作 ▾

ポリシータイプ: ターゲットの追跡スケーリング

次の場合にポリシーを実行:

CPU の平均使用率 を 40 に維持するために必要な場合

アクションを実行:

必要に応じてインスタンスを追加または削除

インスタンスは:

300 スケーリング後にウォームアップする秒数

スケールインの無効化

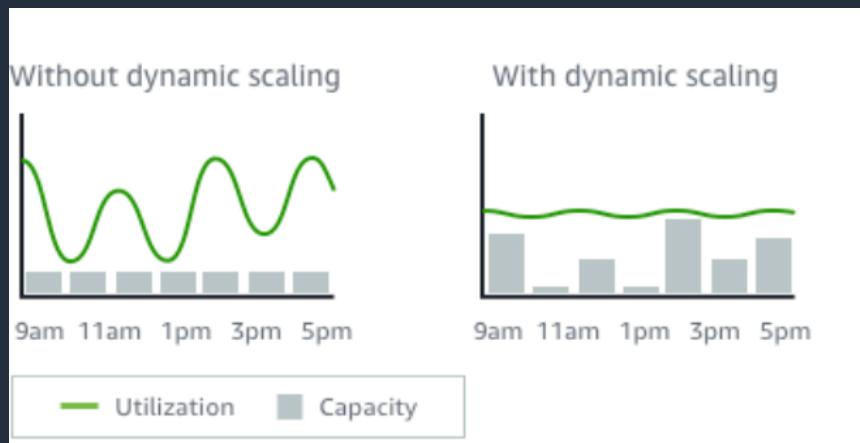
いいえ

https://docs.aws.amazon.com/ja_jp/autoscaling/ec2/userguide/as-scaling-target-tracking.html

https://docs.aws.amazon.com/ja_jp/autoscaling/application/userguide/application-auto-scaling-target-tracking.html

動的なスケーリング – ターゲット追跡スケーリング (2/3)

- 目標値を満たすように自動的にリソースが調整される
 - 何も設定しない場合、キャパシティ (灰色) が一定のため負荷 (緑色) が変動する
 - ターゲット追跡スケーリングを設定すると、負荷に応じてキャパシティが増減する。その結果、負荷が一定の値におさまる



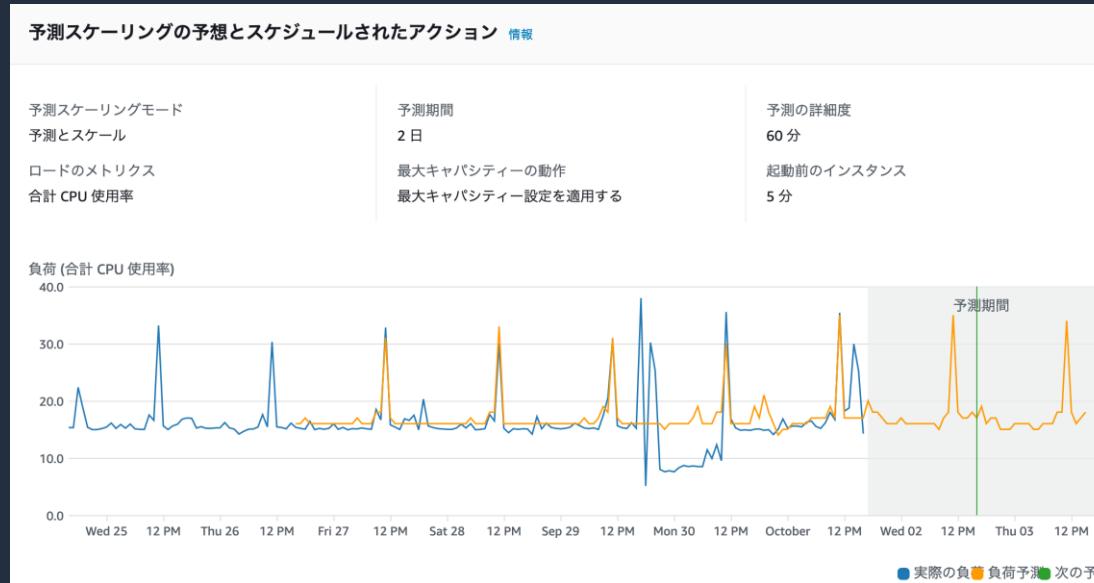
動的なスケーリング - ターゲット追跡スケーリング (3/3)

- スケールアウト用・スケールイン用の2本のアラームが自動的に作成される
 - TargetTracking-xxx-AlarmLow-UUID : スケールイン条件
 - TargetTracking-xxx-AlarmHigh-UUID : スケールアウト条件
- Highは敏感(3分など)、Lowはゆっくり(15分など)
- 基本的に、これらのアラームがアラーム状態になったときスケール



予測スケーリング (1/3)

- EC2 Auto Scalingのみ (2019-10現在)
- 2週間分のメトリクスを分析し、次の2日の今後の需要を予測
使用可能なメトリクス : CPUUtilization, NetworkIn, NetworkOut, および任意のメトリクス



https://docs.aws.amazon.com/ja_jp/autoscaling/plans/userguide/how-it-works.html

© 2019, Amazon Web Services, Inc. or its Affiliates. All rights reserved.



予測スケーリング (2/3)

- 24時間ごとに、次の48時間の予測値を作成し、キャパシティの増減をスケジュールする



- 予測の基準時刻は毎時0分
 - 「インスタンスの事前起動」設定により、スケール動作を前もって実行させることができる
 - デフォルトは5分(300秒)前
 - 午前10時に負荷が予測されている場合、対応するスケールアウトは午前9時55分に実行される

予測スケーリング (3/3)

- 考慮点とベストプラクティス
 - AWS Auto Scalingマネジメントコンソールを使う
 - いきなり使い始めず、「予測のみ」モードでどのような予測値が評価されるかを確認できる
 - ASGの作成後、24時間待つ。予測の開始には最低24時間分のデータポイントが必要

スケジュールスケーリング (1/2)

- EC2 Auto Scaling, Application Auto Scaling
- 一度限り、もしくは定期的なスケジュールを指定可能

The screenshot shows the AWS Auto Scaling Groups console for the group 'myalbtestasg'. The 'Schedule Actions' tab is selected. A search bar at the top allows filtering by action name, start time, end time, repeat frequency, desired capacity, and minimum and maximum values. One scheduled action is listed: 'mytestsched' starting at 2019 October 1 22:55:00 UTC+9.

名前	開始時刻	終了時刻	繰り返し	希望するキャパシティ	最小	最大
mytestsched	2019 October 1 22:55:00 UTC+9				1	5

https://docs.aws.amazon.com/ja_jp/autoscaling/ec2/userguide/schedule_time.html

https://docs.aws.amazon.com/ja_jp/autoscaling/application/userguide/application-auto-scaling-scheduled-scaling.html

スケジュールスケーリング (2/2)

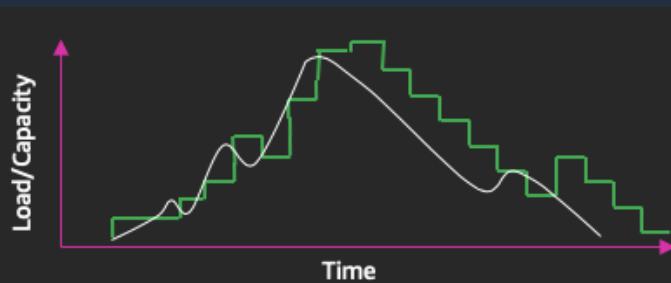
- MinCapacity(最小キャパシティ)とMaxCapacity(最大キャパシティ)のいずれか、あるいは両方を指定可能
 - 設定時刻時点の容量がMinCapacityに満たない→MinCapacityまでスケールアウト
 - 設定時刻時点の容量がMaxCapacityを超している→MaxCapacityまでスケールイン
- (EC2 ASのみ) MinCapacity, MaxCapacity, DesiredCapacity(希望キャパシティ)を指定可能



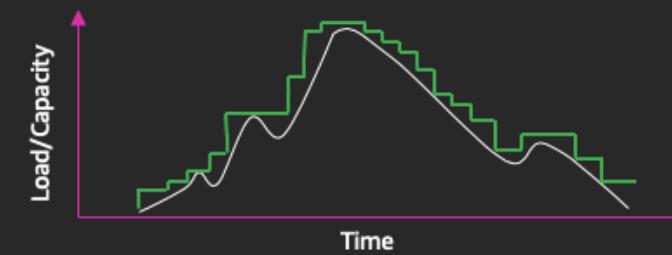
スケーリングオプションの選択指針 (1/2)

EC2でのお勧めオプション：予測スケーリングを使い、同時にターゲット追跡スケーリングも有効にする

- 1) 大まかなキャパシティ増減は予測スケーリングに任せ、前もってスケールしておく
- 2) 実際の負荷に対して不足した分をターゲット追跡で補充する



予測スケーリングによる
キャパシティの事前準備



予測スケーリング+ターゲット追跡スケーリング
によるキャパシティ準備

—— 準備されたキャパシティ量

—— 実際の負荷

スケーリングオプションの選択指針 (2/2)

ステップスケーリングおよびスケジュールスケーリングを使う

- ・ 個々の条件下でのスケーリングを細かく制御したいときの考え方として引き続き有効
- ・ EC2以外のリソースについてはこちらを選択
- ・ キャパシティ設計時に、個別のパターンを下から積み上げて設定していく場合に向いている

本日のアジェンダ

- Auto Scalingサービスのコンセプト
- Auto Scalingの基礎知識
- 主要機能：スケーリングの整理
- Auto Scalingを使ってみる
- こんなときどうする？ - 各種機能の紹介
- まとめ・参考資料

Auto Scalingを使ってみる

- EC2 Auto Scalingを使ってみる
- Application Auto Scalingを使ってみる
- AWS Auto Scaling – 予測スケーリングを使ってみる

Auto Scalingを使ってみる

- EC2 Auto Scalingを使ってみる
- Application Auto Scalingを使ってみる
- AWS Auto Scaling – 予測スケーリングを使ってみる

EC2 Auto Scalingの新機能 – ミックスインスタンスグループ

- ・ オンデマンドインスタンスとスポットインスタンスをひとつのAuto Scaling グループで管理
 - (オンデマンド:スポット) = (9:1)といった指定ができる
 - インスタンスタイプを複数指定できる
 - インスタンスタイプを分散できる



EC2 Auto Scaling Groupの作成 (1)

The screenshot shows the AWS Auto Scaling Groups creation interface. On the left, a sidebar menu lists various services: Key-Pair, Network Interface, Load Balancing, Target Groups, and Auto Scaling. Under Auto Scaling, there are sub-options for Launch Configuration and Auto Scaling Groups, with the latter being highlighted by a yellow oval. In the main content area, a blue box contains a message about the start of support for launch templates. Below this, the 'Auto Scaling へようこそ' (Welcome to Auto Scaling) section provides an overview of what Auto Scaling does and includes a link to detailed information. A prominent blue button labeled 'Auto Scaling グループの作成' (Create Auto Scaling Group) is centered in the middle of the page. To the right, a sidebar titled '追加情報' (Additional Information) lists links to the User Guide, Documentation, All EC2 Resources, Forum, Pricing, and Contact Us.

キーペア
ネットワークインターフェイス
ロードバランシング
ロードバランサー
ターゲットグループ
AUTO SCALING
起動設定
Auto Scaling グループ

SYSTEMS MANAGER SERVICES
コマンドの実行
ステートマネージャー

● 起動テンプレートの提供が開始されました。

EC2 Auto Scaling コンソールで、EC2 起動テンプレートのフルサポートが開始されました。新しい Auto Scaling グループには起動テンプレートを使用することをお勧めします。起動テンプレートにより、Amazon EC2 の最新機能を活用することができます。Auto Scaling グループを作成して開始するか、[詳細はこちら](#)を参照してください。

Auto Scaling へようこそ

Auto Scaling を使用すると、Amazon EC2 キャパシティの自動的な管理、適切な数のアプリケーションインスタンスの維持、インスタンスの正常なグループの運用、必要に応じたスケーリングを行うことができます。
[詳細はこちら](#)

Auto Scaling グループの作成

注意: 別のリージョンで Auto Scaling グループを作成するには、ナビゲーションバーからリージョンを選択します。

追加情報

入门ガイド
ドキュメント
すべての EC2 リソース
フォーラム
料金
お問い合わせ

EC2 Auto Scaling Groupの作成 (2)

Auto Scaling グループの作成

キャンセルして終了

このウィザードを終了して Auto Scaling グループを作成します。最初に、起動設定または起動テンプレートを選択して、インスタンスの起動に Auto Scaling グループが使用するパラメータを指定します。

起動設定
必要な Amazon EC2 の機能をサポートしている場合は、引き続き起動設定を使用できます。[詳細ははこちら](#)

起動テンプレート [新規](#)
起動テンプレートにより、1つの種類のインスタンスを起動するか、複数のインスタンスタイプと購入オプションの組み合わせを起動するかのオプションを利用できます。起動テンプレートには Amazon EC2 の最新機能が含まれていて、更新とバージョニングができます。[詳細ははこちら](#)
[新しい起動テンプレートの作成](#)



ミックスインスタンスグループ機能を使うには
「起動テンプレート」を用いる必要がある

EC2 Auto Scaling Groupの作成 (3)

1. Auto Scaling グループの詳細設定 2. スケーリングポリシーの設定 3. 通知の設定 4. タグを設定 5. 確認

Auto Scaling グループの作成

グループ名

起動テンプレート lt-0757b443c586fc262

起動テンプレートのバージョン 1 (デフォルト)

起動テンプレートの説明

フリートの構築

起動テンプレートに従う
起動テンプレートにより、インスタンスタイプと購入オプション(オンデマンドまたはスポット)が決まります。

購入オプションとインスタンスを組み合わせる

購入オプションとインスタンスを組み合わせる場合、および複数のインスタンスタイプを選択します。スポットインスタンスは、利用できる最も安い料金で自動的に起動されます。

グループサイズ 開始時 1 インスタンス

「購入オプションとインスタンスを組み合わせる」を選択

EC2 Auto Scaling Groupの作成 (4)

1. Auto Scaling グループの詳細設定 2. スケーリングポリシーの設定 3. 通知の設定 4. タグを設定 5. 確認

Auto Scaling グループの作成

グループ名

起動テンプレート lt-0757b443c586fc262

起動テンプレートのバージョン

起動テンプレートの説明 -

フリートの構築

起動テンプレートに従う
起動テンプレートにより、インスタンスタイプと購入オプション(オンデマンドまたはスポット)が決まります。

購入オプションとインスタンスを組み合わせる
オンデマンドインスタンスとスポットインスタンスの組み合わせ、および複数のインスタンスタイプを選択します。スポットインスタンスは、利用できる最も安い料金で自動的に起動されます。

インスタンスタイプ

許容できるインスタンスタイプをフリートに追加します。順序を変更し、オンデマンドインスタンスの起動の優先度を設定します。この順序によるスポットインスタンスへの影響はありません。

最低2つのインスタンスタイプを追加してください

インスタンスの分散 次のデフォルト設定を使用し、すぐに開始します。

EC2 Auto Scaling Groupの作成 (5)

1. Auto Scaling グループの詳細設定 2. スケーリングポリシーの設定 3. 通知の設定 4. タグを設定 5. 確認

Auto Scaling グループの作成

グループ名

起動テンプレート lt-0757b443c586fc262

起動テンプレートのバージョン

起動テンプレートの説明 -

フリートの構築

○ 起動テンプレートに従う
起動テンプレートにより、インスタンスタイプと購入オプション(オンデマンドまたはスポット)が決まります。

◎ 購入オプションとインスタンスを組み合わせる
オンデマンドインスタンスとスポットインスタンスの組み合わせ、および複数のインスタンスタイプを選択します。スポットインスタンスは、利用できる最も安い料金で自動的に起動されます。

インスタンスタイプ

許容できるインスタンスタイプをフリートに追加します。順序を変更し、オンデマンドインスタンスの起動の優先度を設定します。この順序によるスポットインスタンスへの影響はありません。

インスタンスタイプの選択

最低2つのインスタンスタイプを追加してください

インスタンスタイプの追加

インスタンスの分散 次のデフォルト設定を使用し、すぐに開始します。

EC2 Auto Scaling Groupの作成 (6)

フリートの構築

起動テンプレートに従う
起動テンプレートにより、インスタンスタイプと購入オプション（オンデマンドまたはスポット）が決まります。

購入オプションとインスタンスを組み合わせる
オンデマンドインスタンスとスポットインスタンスの組み合わせ、および複数のインスタンスタイプを選択します。スポットインスタンスは、利用できる最も安い料金で自動的に起動されます。

インスタンスタイプ

許容できるインスタンスタイプをフリートに追加します。順序を変更し、オンデマンドインスタンスの起動の優先度を設定します。この順序によるスポットインスタンスへの影響はありません。

m4.large (2vCPU、8GiB)

c4.large (2vCPU、3.75GiB)

[インスタンスタイプの追加](#)

- 要件に合うインスタンスタイプを複数選択する
- 起動テンプレートに指定しておくことも可能

EC2 Auto Scaling Groupの作成 (7)

インスタンスの分散



次のデフォルト設定を使用し、すぐに開始します。

- 上記の優先度に基づき、インテーマントインスタンスを起動します。
- アベイラビリティーゾーンごとに 2 つの最低価格インスタンスタイプ間でスポットインスタンスを多様化します。
- 各インスタンスタイプの最大スポット料金を、オンデマンド料金と同じに設定します。
- 70% のオンデマンドインスタンスと 30% のスポットインスタンスを組み合わせて維持します。

グループサイズ



開始時

1

インスタンス

「インスタンスの分散」のチェックを外す

EC2 Auto Scaling Groupの作成 (9)

インスタンスの分散 ① 次のデフォルト設定を使用し、すぐに開始します。

オンデマンドの割り当て戦略 ① 優先順位付け

最大スポット料金 ① デフォルトを使用 (推奨)
デフォルトでは現在のスポット料金が使用されますが、オンデマンド価格に上限が設定されます。
 上限価格を設定 (1 インスタンス/時間あたり)

スポットの配分戦略 ① スpotトインスタンスを アベイラビリティーゾーンごとに最も価格の安いインスタンスタイプ間で多様化する

オプションのオンデマンドベース ① 最初のインスタンスを オンデマンドとして指定します

ベースを超えるオンデマンド割合 ① % オンデマンドおよび 30% スpot

グループサイズ ① 開始時 インスタンス

台数の考え方について次のスライドで説明

EC2 Auto Scaling Groupの作成 (10)

インスタンスの分散 次のデフォルト設定を使用し、すぐに開始します。

オンデマンドの割り当て戦略 優先順位付け

最大スポット料金 デフォルトを使用 (推奨)
デフォルトでは現在のスポット料金が使用されますが、オンデマンド価格に上限が設定されます。
 上限価格を設定 (1 インスタンス/時間あたり)

スポットの配分戦略 スpotトインスタンスを アベイラビリティーゾーンごとに最も価格の安いインスタンスタイプ間で多様化する

オプションのオンデマンドベース 最初のインスタンスを オンデマンドとして指定します

ベースを超えるオンデマンド割合 % オンデマンドおよび 30% スpot

グループサイズ 開始時 インスタンス

台数の考え方の例

- 「グループサイズ」: 12

グループサイズ = 12

EC2 Auto Scaling Groupの作成 (10)

インスタンスの分散	<input type="checkbox"/> 次のデフォルト設定を使用し、すぐに開始します。
オンデマンドの割り当て戦略	<input type="checkbox"/> 優先順位付け
最大スポット料金	<input checked="" type="radio"/> デフォルトを使用 (推奨) デフォルトでは現在のスポット料金が使用されますが、オンデマンド価格に上限が設定されます。 <input type="radio"/> 上限価格を設定 (1 インスタンス/時間あたり)
スポットの配分戦略	スポットインスタンスを <input type="text" value="2"/> アベイラビリティーゾーンごとに最も価格の安いインスタンスタイプ間で多様化する
オプションのオンデマンドベース	最初のインスタンスを <input type="text" value="2"/> オンデマンドとして指定します
ベースを超えるオンデマンド割合	<input type="text" value="70"/> % オンデマンドおよび 30% スpot
グループサイズ	開始時 <input type="text" value="12"/> インスタンス

台数の考え方の例

- 「グループサイズ」: 12
- 「オプションのオンデマンドベース」: 2

グループサイズ = 12

オンデマンド = 2

EC2 Auto Scaling Groupの作成 (10)

インスタンスの分散	<input type="checkbox"/> 次のデフォルト設定を使用し、すぐに開始します。
オンデマンドの割り当て戦略	<input type="checkbox"/> 優先順位付け
最大スポット料金	<input checked="" type="radio"/> デフォルトを使用 (推奨) デフォルトでは現在のスポット料金が使用されますが、オンデマンド価格に上限が設定されます。 <input type="radio"/> 上限価格を設定 (1 インスタンス/時間あたり)
スポットの配分戦略	スポットインスタンスを <input type="text" value="2"/> アベイラビリティーゾーンごとに最も価格の安いインスタンスタイプ間で多様化する
オプションのオンデマンドベース	<input type="checkbox"/> 最初のインスタンスを <input type="text" value="2"/> オンデマンドとして指定します
ベースを超えるオンデマンド割合	<input type="text" value="70"/> % オンデマンドおよび 30% スpot
グループサイズ	開始時 <input type="text" value="12"/> インスタンス

台数の考え方の例

- 「グループサイズ」: 12
- 「オプションのオンデマンドベース」: 2
- 「ベースを超えるオンデマンド割合」: 70:30

$$\text{グループサイズ} = 12$$

オンデマンド = 2

オンデマンド = 7

スポート = 3

EC2 Auto Scaling Groupの作成 (10)

インスタンスの分散 ① 次のデフォルト設定を使用し、すぐに開始します。

オンデマンドの割り当て戦略 ① 優先順位付け

最大スポット料金 ① デフォルトを使用 (推奨)
デフォルトでは現在のスポット料金が使用されますが、オンデマンド価格に上限が設定されます。
 上限価格を設定 (1 インスタンス/時間あたり)

スポットの配分戦略 ① スpotインスタンスを アベイラビリティーゾーンごとに最も価格の安いインスタンスタイプ間で多様化する

オプションのオンデマンドベース ① 最初のインスタンスを オンデマンドとして指定します

ベースを超えるオンデマンド割合 ① % オンデマンドおよび 30% スpot

グループサイズ ① 開始時 インスタンス

台数の考え方の例

- 「グループサイズ」: 12
- 「オプションのオンデマンドベース」: 2
- 「ベースを超えるオンデマンド割合」: 70:30

結果

オンデマンド 9台
スポット 3台

$$\text{グループサイズ} = 12$$

オンデマンド = 2

オンデマンド = 7

スポット = 3

Auto Scalingを使ってみる

- EC2 Auto Scalingを使ってみる
- Application Auto Scalingを使ってみる
- AWS Auto Scaling – 予測スケーリングを使ってみる

スポットフリートでのApplication Auto Scaling の活用

(1)

The screenshot shows the AWS EC2 Spot Instances Request interface. On the left, a sidebar menu lists several options: EC2 ダッシュボード, イベント, タグ, レポート, 制限, インスタンス, インスタンス, 起動テンプレート, and **スポットリクエスト**. The last item, 'スポットリクエスト', is highlighted with an orange border. At the top, there's a navigation bar with tabs: 'スポットインスタンスのリクエスト' (which is selected and highlighted with an orange border), 'アクション', '価格設定履歴', and 'Savings Summary'. To the right of the navigation are refresh and settings icons. Below the navigation is a search bar with dropdown menus for 'リクエストタイプ: all' and '状態: all', and a 'キーワードによる検索' field. To the right of the search bar are navigation arrows: '<' (left), '< リクエストなし >', and '>' (right). Underneath the search bar is a table header with columns: 'リクエスト ID', 'リクエストタ...', 'インスタンスタ...', '状態', '容量', and 'ステータス'. A message in the center of the page states: '現在、このリージョンにスポットリクエストはありません。' (Currently, there are no spot requests in this region.) Below this message, two instructions are provided: 'EC2 スpot インスタンスを初めて使用する場合は、「開始方法」ページにアクセスしてください。' (If you are using EC2 Spot Instances for the first time, please access the 'Getting Started' page.) and 'スポットインスタンスを起動するには、スポットインスタンスのリクエスト ボタンをクリックします。' (To start a spot instance, click the 'Spot Instances Request' button.) At the bottom of the main content area, a note says: '詳細を表示するには、上記から 1 つのスポットリクエストを選択します。' (To view details, select one of the spot requests listed above.)

「スポットリクエスト」 → 「スポットインスタンスのリクエスト」

詳細は以下の「EC2スポットインスタンスのすべて」資料をご参照ください

<https://aws.amazon.com/jp/summits/tokyo-osaka-2019-report/>

スポットフリートでのApplication Auto Scaling の活用

(2)

必要な容量をお知らせください

起動するターゲット容量 (インスタンス数または vCPU 数) を設定します。起動テンプレートを指定した場合、ターゲット容量の一部をオンデマンドとして割り当てることができます。オンデマンドインスタンスの数は常に保持されますが、スポットインスタンスはスケールできます。

合計ターゲット容量 ?

1

インスタンス▼

オプションのオンデマンド部分 [詳細はこち](#)

ら▼

0

インスタンス

起動テンプレートを指定するリクエストのみが
オンデマンドの対象です

ターゲット容量を維持する

中断動作 ?

終了

- 作成時の注意点：「ターゲット容量を維持する」にチェックを入れる(maintainモードを指定する)
https://docs.aws.amazon.com/ja_jp/AWSEC2/latest/UserGuide/spot-fleet-target-tracking.html
 - 「スポットフリート リクエストには、タイプが maintain のリクエストが必要です。」
- 中断などで指定容量を下回った場合にスポットフリートが自動で新しいインスタンスを起動する

スポットフリートでのApplication Auto Scaling の活用

(3)

The screenshot shows the AWS Spot Fleet Requests interface. At the top, there are tabs for 'Spotインスタンスのリクエスト' (selected), 'アクション', '価格設定履歴', and 'Savings Summary'. On the right are refresh and settings icons. Below the tabs is a search bar with filters for 'リクエストタイプ: all' and '状態: all', and a keyword search field. A pagination bar indicates '1 リクエスト中 1 から 1 を表示'. The main table lists one request:

リクエスト ID	リクエストタ...	インスタンスタ...	状態	容量	ステータス	永続性	作成日
sfr-d8948be1-dc6b...	fleet	t3.medium,t2.m...	active	3 of 3	fulfilled	maintain	a day

Below the table, a note says 'リクエスト ID: sfr-d8948be1-dc6b-47d6-b1e7-b20212525f78'. Underneath, there are tabs: '説明', 'インスタンス', '履歴', '削減額', 'Auto Scaling' (which is highlighted with an orange box), and 'スケジュールに基づくスケーリング'. A message states 'このフリートに対して Auto Scaling は設定されていません' (Auto Scaling is not set for this fleet). A large blue '設定' button is prominently displayed, also outlined in orange. A note below it says 'CloudWatch アラームに応じてフリートのターゲット容量を指定範囲内で自動的に調整します' (Automatically adjusts the fleet's target capacity within the specified range based on CloudWatch Alarms).

“Auto Scaling”タブ→「設定」

スポットフリートでのApplication Auto Scaling の活用

(4)



- ターゲット追跡スケーリングポリシーの目標値を設定する
- もしくはステップスケーリングポリシーを選択することもできる

Application Auto Scaling の設定ポイント

- マネジメントコンソールを使う場合と CLI(API, SDK)を使う場合の違い
 - マネジメントコンソールの場合
各サービスのマネジメントコンソールから設定する
 - CLI(API, SDK)の場合
使用するAPIはすべてApplication Auto ScalingサービスのAPIである
 - (CLIの例)
`aws application-autoscaling register-scalable-target ¥
--service-namespace サービス名 ¥
--resource-id リソースID ¥`
...

Auto Scalingを使ってみる

- EC2 Auto Scalingを使ってみる
- Application Auto Scalingを使ってみる
- AWS Auto Scaling – 予測スケーリングを使ってみる

下準備 - EC2 Auto Scalingグループの設定変更

Auto Scaling グループ: myasgforssm

詳細 アクティビティ履歴 スケーリングポリシー インスタンス モニタリング 通知 タグ スケジュールされたアクション ラ...

Auto Scaling メトリックス: グループメトリックコレクションを有効にする 次のデータを表示: 過去 1 時間

表示:Auto Scaling または EC2

警告 以下の Auto Scaling グループでは、グループメトリックコレクションが有効になっていません: myasgforssm

以下は、選択されたリソースの CloudWatch メトリックスです (最大 10)。画面を拡大するには、グラフをクリックします。すべての時刻は協定世界時 (UTC) で表示されています。> すべての CloudWatch メトリックスを表示

myasgforssm (有効でない)

最小グループサイズ (カウント)	最大グループサイズ (カウント)	希望するキャパシティ (カウント)
1 0.75	1 0.75	1 0.75



EC2 Auto Scalingマネジメントコンソールから
「モニタリング」→「グループメトリックコレクションを有効にする」

AWS Auto ScalingによるEC2 Auto Scalingグループの予測スケーリング (1)



管理ツール

AWS Auto Scaling 複数のリソースを迅速かつ 簡単にスケールできるよう 支援します

AWS Auto Scaling では、アプリケーションの基になるすべてのスケーラブルなリソースを
すばやく検出し、組み込みのスケーリング推奨項目を使用して数分でアプリケーションの
スケーリングをセットアップできます。

この機能の説明

スケーリングプランの作成

わずか数ステップでアプリケーションを最適化します

今すぐ始める

料金表

AWS Auto Scaling は無料です。

AWS Auto Scaling は、Amazon CloudWatch で有効
にすることができますが、追加料金は発生しません。
アプリケーションリソースおよび Amazon
CloudWatch のサービス料金が適用されます。

AWS Auto ScalingによるEC2 Auto Scalingグループの予測スケーリング (2)

AWS Auto Scaling > スケーリングプラン > スケーリングプランの作成

ステップ 1
スケーラブルなリソースの検索

自動検出または手動で、スケーリングプランに追加するリソースを選択します。 [情報](#)

ステップ 2
スケーリング戦略を指定します。

ステップ 3
詳細設定の設定 (オプション)

ステップ 4
確認と作成

メソッドの選択

- CloudFormation スタックによる検索
AWS CloudFormation によってプロビジョニングされたリソースを検索します。
- タグによる検索
適用されたタグを使用してリソースを検索します。
- Amazon EC2 Auto Scaling グループの選択
スケーリング計画に含めるための、Auto Scaling グループを 1 つ以上選択します。

「EC2 Auto Scalingグループの選択」を選択

AWS Auto ScalingによるEC2 Auto Scalingグループの予測スケーリング (3)

AWS Auto Scaling > スケーリングプラン > スケーリングプランの作成

ステップ 1
スケーラブルなリソースの検索

自動検出または手動で、スケーリングプランに追加するリソースを選択します。 [情報](#)

メソッドの選択

- CloudFormation スタックによる検索
AWS CloudFormationによってプロビジョニングされたリソースを検索します。
- タグによる検索
適用されたタグを使用してリソースを検索します。
- Amazon EC2 Auto Scaling グループの選択
スケーリング計画に含めるため、Auto Scaling グループを 1 つ以上選択します。

Auto Scaling グループの選択 [情報](#)

Auto Scaling グループ

Auto Scaling グループを選択します。

myalbttestasg
myasgforssm

キャンセル 次へ



既存のAuto Scalingグループを選択して「次へ」

AWS Auto ScalingによるEC2 Auto Scalingグループの予測スケーリング (4)

AWS Auto Scaling > スケーリングプラン > スケーリングプランの作成

ステップ 1
スケーラブルなリソースの検索

ステップ 2
スケーリング戦略を指定します。

ステップ 3
詳細設定の設定 (オプション)

ステップ 4
確認と作成

スケーリング戦略を指定します。

スケーリング戦略を使用して、アプリケーションのスケーラブルなリソースを最適化する方法を定義します。 [情報](#)

スケーリングプランの詳細

名前 長さは 1~128 文字にする必要があり、パイプ文字「|」、コロン「:」、およびスラッシュ「/」を含めることはできません。

リソース
1 Auto Scaling グループ 選択されました。

Auto Scaling グループ (1)
1 Auto Scaling グループ のためにスケーリング戦略を指定します。 スケーリングプランに含める

スケーリングプラン名を入力

AWS Auto ScalingによるEC2 Auto Scalingグループの予測スケーリング (5)

Auto Scaling グループ (1)
1 Auto Scaling グループのためにスケーリング戦略を指定します。

スケーリングプランに含める

スケーリング戦略
その戦略では、リソースの拡張に使用するスケーリングメトリックとターゲット値を定義します。

可用性を考えた最適化
高い可用性を提供し需要の急増に対応できるキャパシティーを確保するため、Auto Scaling グループの平均 CPU 使用率を常に 40% に維持します。

可用性とコストのバランスを取ります
最適な可用性を提供しこストを削減するため Auto Scaling グループの平均 CPU 使用率を常に 50% に維持します。

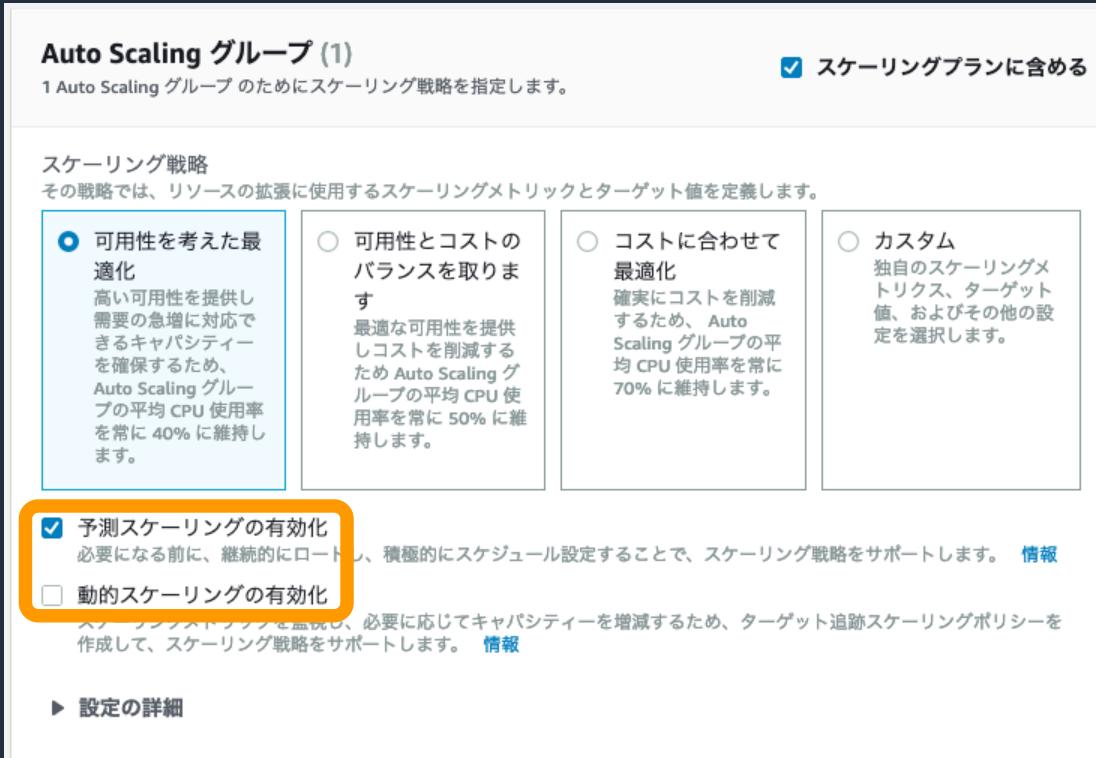
コストに合わせて最適化
確実にコストを削減するため、Auto Scaling グループの平均 CPU 使用率を常に 70% に維持します。

カスタム
独自のスケーリングメトリクス、ターゲット値、およびその他の設定を選択します。

予測スケーリングの有効化
必要になる前に、継続的にロートンし、積極的にスケジュール設定することで、スケーリング戦略をサポートします。 [情報](#)

動的スケーリングの有効化
ヘイブンメントメントを監視し、必要に応じてキャパシティーを増減するため、ターゲット追跡スケーリングポリシーを作成して、スケーリング戦略をサポートします。 [情報](#)

▶ 設定の詳細



- 「予測スケーリングの有効化」にチェックが入っていることを確認
- 「動的スケーリングの有効化」のチェックを外す

AWS Auto ScalingによるEC2 Auto Scalingグループの予測スケーリング (6)

ステップ 1
スケーラブルなリソースの検索

ステップ 2
スケーリング戦略を指定します。

ステップ 3
詳細設定の設定 (オプション)

ステップ 4
確認と作成

詳細設定の設定 (オプション)

個々のリソースまたは複数のリソースの設定を、同時にカスタマイズします。 [情報](#)

▶ Auto Scaling グループ (1)

Auto Scaling グループでは、カスタム設定が使用されます。予測スケーリングは有効です。

キャンセル 戻る 次へ

“Auto Scalingグループ”をクリックして展開

AWS Auto ScalingによるEC2 Auto Scalingグループの予測スケーリング (7)

詳細設定の設定 (オプション)
個々のリソースまたは複数のリソースの設定を、同時にカスタマイズします。 [情報](#)

▼ Auto Scaling グループ (1 個中 1 個を選択) [元に戻す](#)

カスタム設定を指定する 1 つ以上の Auto Scaling グループを選択します。

<input checked="" type="checkbox"/> リソース	▲	プランに含める	外部のスケーリングポリシーの置き換え	既存のポリシー
<input checked="" type="checkbox"/> myasgforssm		はい	なし	なし

1 個のリソースが選択されました

スケーリングプランに含める

▶ 全般設定

▶ 動的スケーリングの設定

▶ **予測スケーリング設定**

[キャンセル](#) [戻る](#) [次へ](#)

対象Auto Scalingグループを選択すると詳細設定メニューが展開されるので
「予測スケーリング設定」を展開する

AWS Auto ScalingによるEC2 Auto Scalingグループの予測スケーリング (8)

▼ 予測スケーリング設定

予測スケーリングモード
予測の実行にスケーリングを使用するかどうかを決定します。これはいつでも変更できます。 [情報](#)

予測のみ
予測とスケール
予測のみ

予測のみ

予測とスケール

予測のみ

最大キャパシティーの動作
予測キャパシティーが最大キャパシティーに近づいたか、それを超えたときに使用するルールを選択します。 [情報](#)

最大キャパシティー設... ▾

予測期間
事前予測する日数。 [情報](#)

2 日

予測の詳細度
予測とキャパシティーの計算間隔。 [情報](#)

60 分

予測頻度
予測更新の頻度。 [情報](#)

毎日

キャンセル 戻る 次へ

「予測のみ」を選択して「次へ」

AWS Auto ScalingによるEC2 Auto Scalingグループの予測スケーリング (9)

ステップ1
スケーラブルなリソースの検索

ステップ2
スケーリング戦略を指定します。

ステップ3
詳細設定の設定 (オプション)

ステップ4
確認と作成

確認と作成

スケーリングプランの詳細

名前
myfirstscalingplan

リソース
1 Auto Scaling グループ選択されました。1 個のスケーリングポリシーが作成され、0 個の外部ポリシーが維持されます。

Auto Scaling グループ

スケーリング戦略	スケーリングメトリクス	ターゲット値
可用性を考えた最適化	CPU の平均使用率	40 %

概要
お客様のスケーリングプランは、40 % で CPU の平均使用率 メトリクスを保持することで、1 Auto Scaling グループを最適化するように設定されています。

動的スケーリング
動的スケーリングは有効です。40 % で CPU の平均使用率 メトリクスを保持する必要に応じて、インスタンスを追加または削除するため、1 ターゲット追跡スケーリングが適用されます。

予測スケーリング
予測スケーリングは有効です。40 % で CPU の平均使用率 メトリクスを保持するために必要なインスタンスの最小数を維持するため、合計 CPU 使用率 メトリクスの予測に基づいて、スケジュールされたスケーリングアクションが生成されます。

▶ 詳細

キャンセル 戻る スケーリングプランの作成

AWS Auto ScalingによるEC2 Auto Scalingグループの予測スケーリング(10)



The screenshot shows the AWS Auto Scaling Scaling Plan page. At the top, there is a breadcrumb navigation: AWS Auto Scaling > スケーリングプラン. Below the navigation, a table displays the scaling plan details. The table has columns: 削除 (Delete) button, 名前 (Name), ボックス (checkbox), ステータス (Status), ボックス (checkbox), スケーリングポリシー (Scaling Policy), ボックス (checkbox), and 作成時刻 (Creation Time). There is also a C (Create) button and a 編集 (Edit) button. The table contains one row for a scaling plan named "myfirstscalingplan". The status is "Active" with a green checkmark icon. The scaling policy count is 1. The creation time is 2019-10-02 01:33:59 UTC+0900. The "myfirstscalingplan" cell is highlighted with an orange border.

	C	編集	削除	スケーリングプランの作成
名前	myfirstscalingplan	Active	1	2019-10-02 01:33:59 UTC+0900

「ステータス」が"Active"になつたらスケーリングプラン名をクリック

AWS Auto ScalingによるEC2 Auto Scalingグループの予測スケーリング(11)

AWS Auto Scaling > スケーリングプラン > myfirstscalingplan

myfirstscalingplan

スケーリングプランの詳細

ステータス
④ Active

ステータスの説明
Scaling plan has been created and applied to all resources.

Auto Scaling グループ (1)

autoScalingGroup/myasgforssm

ステータス
④ アクティブ

スケーリングメトリクス
CPU の平均使用率

ターゲット値
40 %

1 時間 3 時間 12 時間 1 日 3 日 1 週

ダッシュボードに追加

合計 CPU 使用率 (%)

2.67
1.33
0

14:00 14:30 15:00 15:30 16:00 16:30

“Auto Scalingグループ”の下の対象Auto Scalingグループリソース名をクリック

AWS Auto ScalingによるEC2 Auto Scalingグループの予測スケーリング (12)



画面下にスクロールすると、向こう48時間の負荷の予測結果が表示される

AWS Auto ScalingによるEC2 Auto Scalingグループの予測スケーリング (13)



さらに下にスクロールすると、スケール予定のインスタンス数とそのためのスケジュールスケーリング設定の計画を確認できる

AWS Auto ScalingによるEC2 Auto Scalingグループの予測スケーリング (14)

AWS Auto Scaling > スケーリングプラン > myfirstscalingplan

myfirstscalingplan

編集

削除

スケーリングプランの詳細

ステータス

Active

ステータスの説明

Scaling plan has been created and applied to all resources.

- 「予測のみ」モードから「予測とスケーリング」モードへ変更
- 対象スケーリングプランを選択して「編集」

AWS Auto ScalingによるEC2 Auto Scalingグループの予測スケーリング(15)

AWS Auto Scaling > スケーリングプラン > myfirstscalingplan > 編集

myfirstscalingplan の編集

▶ Auto Scaling グループ (1)

すべての Auto Scaling グループでは、「可用性を考えた最適化」スケーリング戦略が使用されます。予測スケーリングは有効です。

キャンセル

次へ

“Auto Scalingグループ”をクリックして展開

AWS Auto ScalingによるEC2 Auto Scalingグループの予測スケーリング (16)

myfirstscalingplan の編集

▼ Auto Scaling グループ (1 個中 1 個を選択)
カスタム設定を指定する 1 つ以上の Auto Scaling グループを選択します。

<input checked="" type="checkbox"/> リソース ▲	プランに含める	外部のスケーリングポリシーの置き換え
<input checked="" type="checkbox"/> myalbtestasg	はい	なし

1 個のリソースが選択されました

スケーリングプランに含める

▶ 全般設定

▶ 動的スケーリングの設定

▶ **予測スケーリング設定**

対象Auto Scalingグループを選択すると詳細設定メニューが展開されるので
「予測スケーリング設定」を展開する

AWS Auto ScalingによるEC2 Auto Scalingグループの予測スケーリング (17)

The screenshot shows the 'Predictive Scaling Settings' section of the AWS Auto Scaling configuration interface. A dropdown menu is open, showing three options: 'Predict and Scale', 'Predict and Scale (using Maximum Capacity)', and 'Predict only'. The 'Predict and Scale' option is highlighted with an orange border. Below the dropdown, there is a field set for the prediction interval, currently set to '5 分' (5 minutes). To the right, there are sections for 'Maximum Capacity Behavior' (describing how it applies rules when capacity approaches or exceeds maximum), 'Prediction Detail' (set to 60 minutes), 'Prediction Frequency' (set to daily), and two buttons at the bottom: 'Cancel' and a large orange 'Next Step' button.

▶ 全般設定

▶ 動的スケーリングの設定

▼ 予測スケーリング設定

予測スケーリングモード

予測の実行にスケーリングを使用するかどうかを決定します。これはいつでも変更できます。 [情報](#)

予測とスケール

予測とスケール

予測のみ

最大キャパシティーの動作

予測キャパシティーが最大キャパシティーに近づいたか、それを超えたときに使用するルールを選択します。 [情報](#)

最大キャパシティー設定を適用する ▾

予測期間

事前予測する日数。 [情報](#)

2 日

予測の詳細度

予測とキャパシティーの計算間隔。 [情報](#)

60 分

予測頻度

予測更新の頻度。 [情報](#)

毎日

キャンセル

次へ

「予測とスケール」を選択して「次へ」

AWS Auto ScalingによるEC2 Auto Scalingグループの予測スケーリング (18)

AWS Auto Scaling > スケーリングプラン > myfirstscalingplan > 編集

myfirstscalingplan の編集

Auto Scaling グループ

スケーリング戦略 可用性を考えた最適化	スケーリングメトリクス CPU の平均使用率	ターゲット値 40 %
------------------------	---------------------------	----------------

概要
動的スケーリングが有効になっている Auto Scaling グループがありません。スケーリングプランで 1 Auto Scaling グループ のターゲット追跡スケーリングポリシーを作成できるようにするには、先に動的スケーリングを有効にする必要があります。

動的スケーリング
動的スケーリングは無効です。

予測スケーリング
予測スケーリングは有効です。 **40 %** で **CPU の平均使用率** メトリクスを保持するために必要なインスタンスの最小数を維持するため、**合計 CPU 使用率** メトリクスの予測に基づいて、スケジュールされたスケーリングアクションが生成されます。

▶ 詳細

キャンセル 戻る **変更の保存**

「変更の保存」

AWS Auto ScalingによるEC2 Auto Scalingグループの予測スケーリング(19)

The screenshot shows the AWS Auto Scaling Groups console for the group 'myalbtestasg'. The 'Scheduled Actions' tab is selected. A yellow box highlights four scheduled actions listed in the table below:

名前	開始時刻	終了時刻	繰り返し	希望するキャパシティ	最小	最大
AutoScaling-myfirstscalingplan-1-201910011800	2019 October 2 03:00:00 UTC+9			1	5	
AutoScaling-myfirstscalingplan-1-201910011900	2019 October 2 04:00:00 UTC+9			1	5	
AutoScaling-myfirstscalingplan-1-201910012000	2019 October 2 05:00:00 UTC+9			1	5	
AutoScaling-myfirstscalingplan-1-201910012100	2019 October 2 06:00:00 UTC+9			1	5	

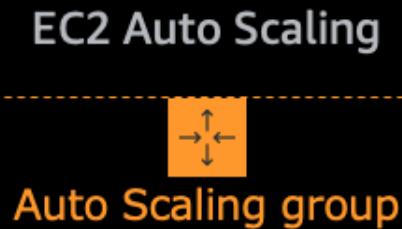
EC2 Auto Scalingマネジメントコンソールから
「スケジュールされたアクション」に毎時のアクションが設定されたことを確認

本日のアジェンダ

- Auto Scalingサービスのコンセプト
- Auto Scalingの基礎知識
- 主要機能：スケーリングの整理
- Auto Scalingを使ってみる
- こんなときどうする？ - 各種機能の紹介
- まとめ・参考資料

こんなときどうする？

- (EC2 Auto Scaling) スポットインスタンスを活用したいです
 - →ミックスインスタンスグループを活用してください



こんなときどうする？

- (EC2 Auto Scaling) 「起動設定」と「起動テンプレート」のどちらを使えば良いか
 - → 「起動テンプレート」を強く推奨します！

こんなときどうする？

- (EC2 Auto Scaling) 速やかにスケールアウト(スケールイン)してくれません
 - →インスタンスの詳細モニタリングを有効にしてください
- CloudWatch Metricsを1分粒度にする。5分粒度では速やかにスケールできない
- 有料オプションながらAuto Scalingを使用する際のベストプラクティス

https://docs.aws.amazon.com/ja_jp/AWSEC2/latest/UserGuide/using-cloudwatch-new.html

こんなときどうする？

- (EC2 Auto Scaling) 正常に動作しないインスタンスを自動的に置き換える
• →ヘルスチェックを活用します
- 特に指定しない場合、EC2ヘルスチェックが有効になっている
 - 2/2以外のステータスが続くとAuto Scalingが置き換える
- ELB配下のASGの場合、ELBヘルスチェックを有効にする
 - EC2ヘルスチェックに加え、ELBからのヘルスチェックに応答しない場合の速やかな入れ替えが可能になる

https://docs.aws.amazon.com/ja_jp/autoscaling/ec2/userguide/healthcheck.html

こんなときどうする？

- (EC2 Auto Scaling)スケールイン・スケールアウトを繰り返してしまい、いつまでたってもインスタンスが追加されない
 - → 「ヘルスチェックの猶予期間」の設定を見直す
- ヘルスチェックの猶予期間：起動したばかりでヘルスチェックに応答できないインスタンスを保護する期間
 - /index.html などは速やかに返せるようになるが、S3からのコンテンツ配備やDB接続などが整った前提のヘルスチェックパスを指定している場合は準備期間が必要
 - 特にELBヘルスチェックにアプリケーションのパスを採用している場合に有効
- デフォルトは5分(300秒)

https://docs.aws.amazon.com/ja_jp/autoscaling/ec2/userguide/healthcheck.html

こんなときどうする？

- (EC2 Auto Scaling) 次にどのインスタンスがスケールイン対象になるか知りたい
 - →デフォルトの終了ポリシー
- おおまかには次の流れで決まる
 1. インスタンスが最も多いアベイラビリティゾーンを選択
 2. (そのアベイラビリティゾーンに候補が複数あるなら) 最も古い起動設定・起動テンプレートから起動されたインスタンスを選択
 3. (複数候補が残っている場合) 次のインスタンス時間に近いものを選択
 4. まだ複数いるならランダム
- カスタマイズも可能

https://docs.aws.amazon.com/ja_jp/autoscaling/ec2/userguide/as-instance-termination.html

こんなときどうする？

- (EC2 Auto Scaling) 特定のインスタンスをスケールインから保護したい
 - →インスタンスの保護
- ASG単位、もしくはインスタンス単位で設定。スケールインされなくなる
- 次の条件からは保護できないことに注意
 - 手動でのインスタンス削除(Terminate)
 - ヘルスチェックによる置き換え
 - スポットインスタンスの中止
- すべてのインスタンスが終了保護された状態でスケールインイベントが発生した場合、希望容量だけが減少し、スケールイン(インスタンス削除)は行われない

https://docs.aws.amazon.com/ja_jp/autoscaling/ec2/userguide/as-instance-termination.html#instance-protection

こんなときどうする？

- (EC2 Auto Scaling) 一時的にスケールインやスケールアウトを止めたい
 - →スケーリングプロセスの中斷
- 一時的にスケール動作を停止できる
- ASG単位で設定
- 中断できるプロセス一覧：Launch, Terminate, AddToLoadBalancer, AlarmNotification, AZRebalance, HealthCheck, ReplaceUnhealthy, ScheduledActions
- 使いどころ：機能テストなど、一時的にAuto Scalingグループの特定プロセスの動作を止めてテスト条件を整えたい場合
 - LaunchとTerminateの両方のプロセスを中断することで、「何もしない」Auto Scalingグループを作り出せる
- 動作のおかしいインスタンスがいるのでスケールイン・スケールアウトを止めたい
 - →プロセスの中斷ではなく次の項目を参照

https://docs.aws.amazon.com/ja_jp/autoscaling/ec2/userguide/as-suspend-resume-processes.html

こんなときどうする？

- (EC2 Auto Scaling) このインスタンスをAuto Scalingグループから外したい
 - →スタンバイ、もしくはデタッチ
- スタンバイ(「一時的なインスタンスの削除」)
 - インスタンス単位で設定
 - そのインスタンスはAuto Scalingグループにいながら「スタンバイ」状態に入る
 - 具体的にはそのインスタンスはELBから登録解除され、ヘルスチェック対象から外される。
そのAuto Scalingグループの希望容量は1つ減少する
 - その間にインスタンスのトラブルシューティングなどを行う
- デタッチ
 - インスタンス単位で設定
 - そのインスタンスはそのAuto Scalingグループのメンバーから外れる
 - スタンバイと実質的な効果は同一。インスタンスはそのままRunning状態で保持される。ただしデタッチの場合、Auto Scalingグループとして与えていたタグも除去される

https://docs.aws.amazon.com/ja_jp/autoscaling/ec2/userguide/as-enter-exit-standby.html
https://docs.aws.amazon.com/ja_jp/autoscaling/ec2/userguide/detach-instance-asg.html

こんなときどうする？

- (EC2 Auto Scaling) スケールアウトした後、サービス開始前にインスタンスに準備させたい / スケールインの前にログ退避させたいのでTerminateを少し待って欲しい
 - →ライフサイクルフック
- ライフサイクルフック：インスタンス起動時・削除時にインスタンスを一時停止し、カスタムアクションを実行できる
- ライフサイクルフックはAuto Scalingグループ単位に設定
- 実際のライフサイクルフックによる待機はインスタンスごと
- 実装例：CloudWatch Eventからライフサイクル通知を受け取り、Lambdaがカスタムアクションを実行する

https://docs.aws.amazon.com/ja_jp/autoscaling/ec2/userguide/lifecycle-hooks.html

こんなときどうする？

- (EC2 Auto Scaling) スケールアウトを素早くしたい
 - →ユーザーデータでのyum updateやyum installなどを用いず、なるべくミドルウェアや必要設定などを済ませた状態のAMIを起動テンプレートに指定する(いわゆるゴールデンイメージ)
 - ただし考慮点があるので次のスライドで説明

こんなときどうする？

- (EC2 Auto Scaling) WindowsやRed Hat Enterprise Linuxなどの考慮点は？
 - 1時間単位の課金になるため、終了ポリシーはデフォルトでお使いいただくのをお勧めする
 - 起動時間を短縮する際、ゴールデンイメージの起動時間と、標準AMIからの起動+ユーザーデータでセットアップした場合とを比較すると良い
 - 2019年現在、標準AMIはカスタマイズしたAMIより素早く起動できるようにチューニングされている
 - 場合によってはユーザーデータの方が速い可能性も

本日のアジェンダ

- Auto Scalingサービスのコンセプト
- Auto Scalingの基礎知識
- 主要機能：スケーリングの整理
- Auto Scalingを使ってみる
- こんなときどうする？ - 各種機能の紹介
- まとめ・参考資料

本日のまとめ

- Auto Scalingの価値
 - アプリケーションの可用性の維持
 - アベイラビリティゾーン間でのインスタンスの分散、異常なインスタンスの自動置き換え
 - 自動的なキャパシティの増減
 - 動的なスケーリング、予測スケーリング、スケジューリングスケーリング
 - 予測スケーリングとターゲット追跡スケーリングの組み合わせは2019年におススメする推奨セット
 - 様々なユースケースをカバーする機能群
 - コスト最適化のためのミックスインスタンスグループ、ライフサイクルフック
- Auto Scalingを使いこなし、クラウドの世界の本質をぜひ実感してください

参考資料

よくある質問

- よくある質問 - Amazon EC2 Auto Scaling | AWS — <https://aws.amazon.com/jp/ec2/autoscaling/faqs/>
- よくある質問 - AWS Auto Scaling | AWS — <https://aws.amazon.com/jp/autoscaling/faqs/>

ユーザーガイド

- Amazon EC2 Auto Scaling とは - Amazon EC2 Auto Scaling (日本語) —
https://docs.aws.amazon.com/ja_jp/autoscaling/ec2/userguide/what-is-amazon-ec2-auto-scaling.html
- Application Auto Scaling とは - Application Auto Scaling —
https://docs.aws.amazon.com/ja_jp/autoscaling/application/userguide/what-is-application-auto-scaling.html
 - 各サービスでのApplication Auto Scalingの使い方・考慮点は以下のリンクから
 - ご利用開始にあたって - Application Auto Scaling —
https://docs.aws.amazon.com/ja_jp/autoscaling/application/userguide/what-is-application-auto-scaling.html#getting-started
- AWS Auto Scaling とは - AWS Auto Scaling —
https://docs.aws.amazon.com/ja_jp/autoscaling/plans/userguide/what-is-aws-auto-scaling.html

Q&A

お答えできなかったご質問については
AWS Japan Blog 「<https://aws.amazon.com/jp/blogs/news/>」にて
後日掲載します。

AWS の日本語資料の場所「AWS 資料」で検索



The screenshot shows the AWS Japan Language Resources page. At the top, there's a navigation bar with the AWS logo, search bar, and links for '日本担当チームへお問い合わせ' (Contact Support), 'サポート' (Support), '日本語' (Japanese), 'アカウント' (Account), and 'コンソールにサインイン' (Sign In). Below the navigation bar is a secondary navigation menu with links for '製品' (Products), 'ソリューション' (Solutions), '料金' (Pricing), 'ドキュメント' (Documentation), '学習' (Learning), 'パートナー' (Partners), 'AWS Marketplace' (AWS Marketplace), 'その他' (Other), and a search icon. The main content area features a large title 'AWS クラウドサービス活用資料集トップ' (Top of the AWS Cloud Service Utilization Document Collection) in white. Below the title is a paragraph of Japanese text about AWS services. At the bottom, there are four call-to-action buttons: 'AWS Webinar お申込 »' (AWS Webinar Application), 'AWS 初心者向け »' (AWS Beginner), '業種・ソリューション別資料 »' (Industry-Solution Specific Documentation), and 'サービス別資料 »' (Service-Specific Documentation).

<https://amzn.to/JPArchive>

AWS Well-Architected 個別技術相談会

毎週”W-A個別技術相談会”を実施中

- AWSのソリューションアーキテクト(SA)に
対策などを相談することも可能

• 申込みはイベント告知サイトから

(<https://aws.amazon.com/jp/about-aws/events/>)

AWS イベント で[検索]

AWS Well-Architected



ご視聴ありがとうございました

AWS 公式 Webinar
<https://amzn.to/JPWebinar>



過去資料
<https://amzn.to/JPArchive>





このコンテンツは公開から3年以上経過しており内容が古い可能性があります
最新情報については[サービス別資料](#)もしくはサービスのドキュメントをご確認ください

Amazon ECS Deep Dive

~ Capacity Providers ~

Amazon ウェブ サービス ジャパン株式会社
落水 恭介

2021.09.07

© 2021, Amazon Web Services, Inc. or its Affiliates.



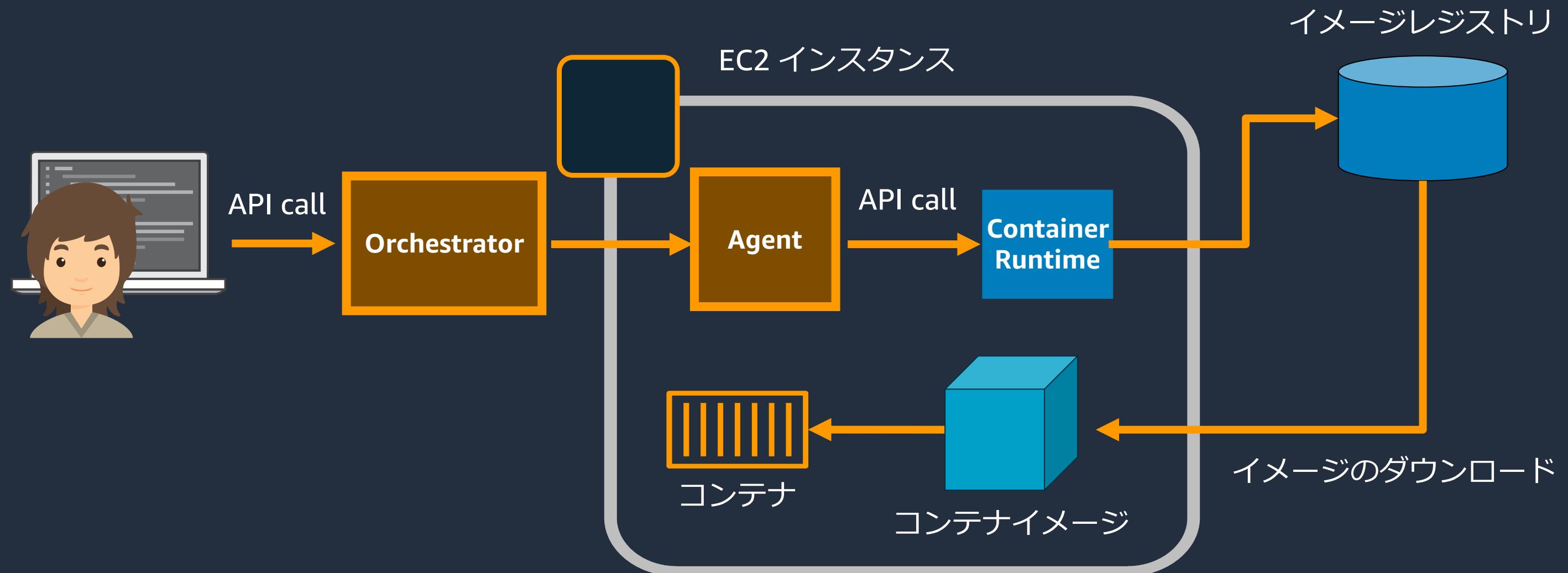
本日お話すこと

- Amazon ECS での Auto Scaling とは
 - コンテナインスタンス (EC2 インスタンスレイヤ)
 - ECS サービス (タスクレイヤ)
- ECS Cluster Auto Scaling
 - ECS Cluster Auto Scaling とは
 - Capacity Providers の概要

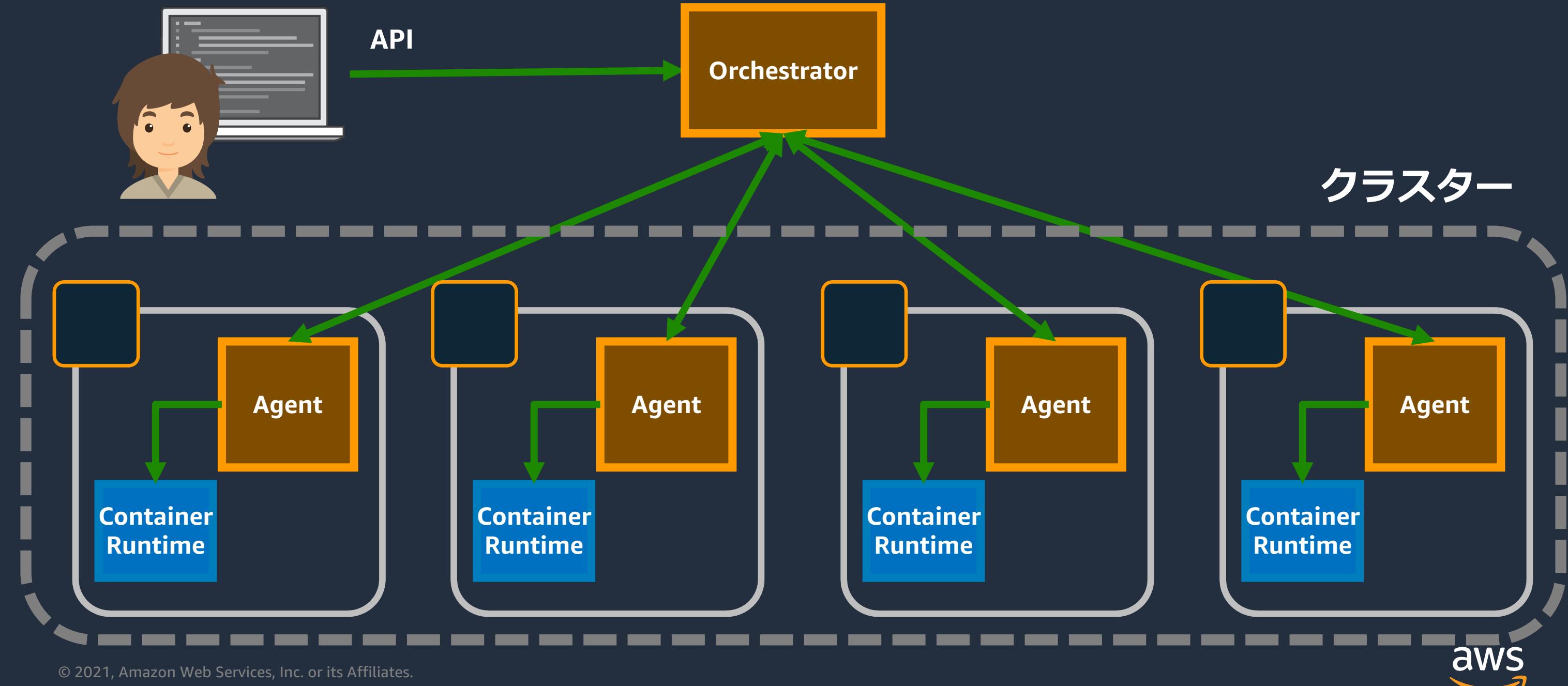
Amazon ECS での Auto Scaling とは



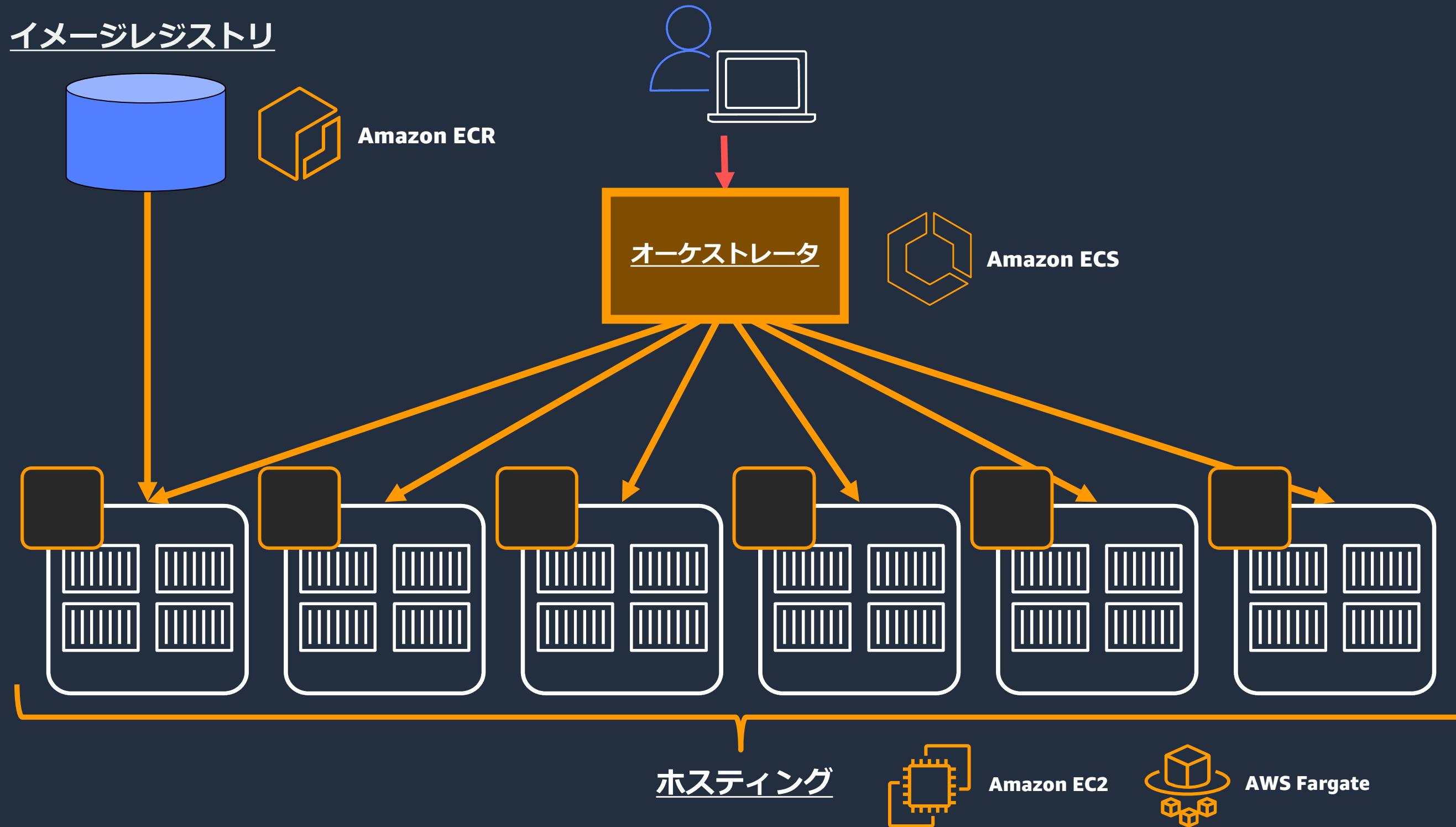
コンテナオーケストレーションの仕組み



コンテナオーケストレーションによるクラスター管理



イメージレジストリ



Amazon ECS の動作イメージ (on EC2)



Amazon ECS の動作イメージ (on Fargate)

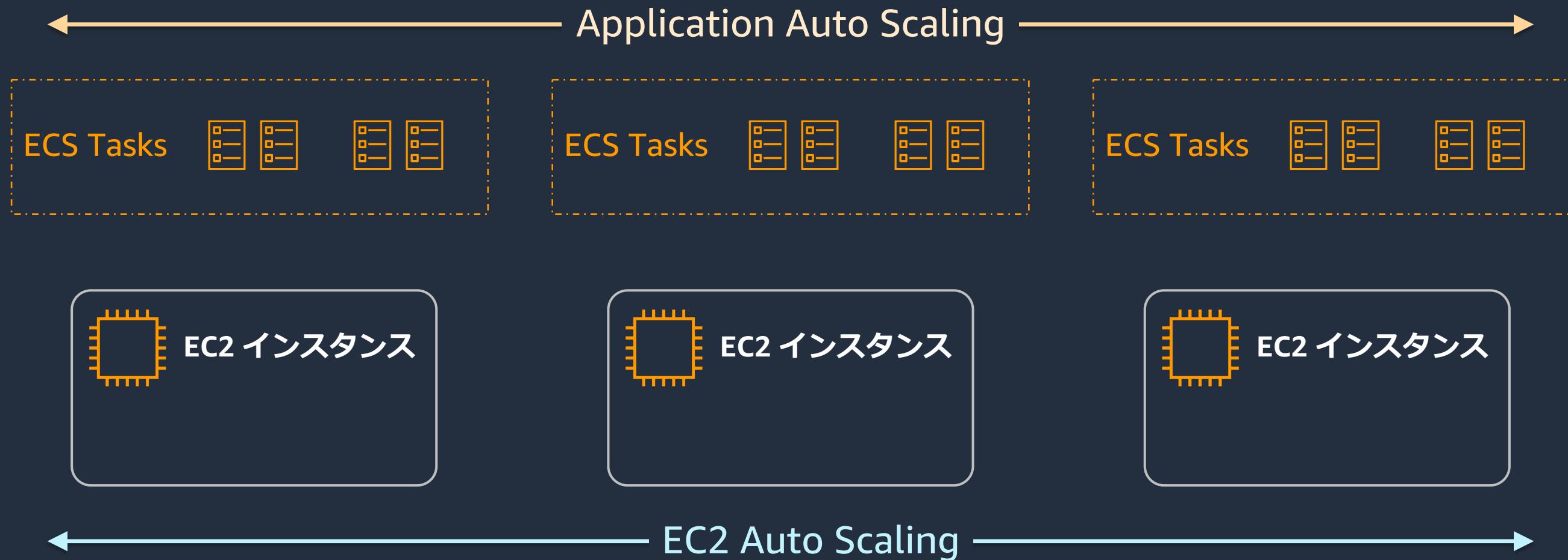


クラスター



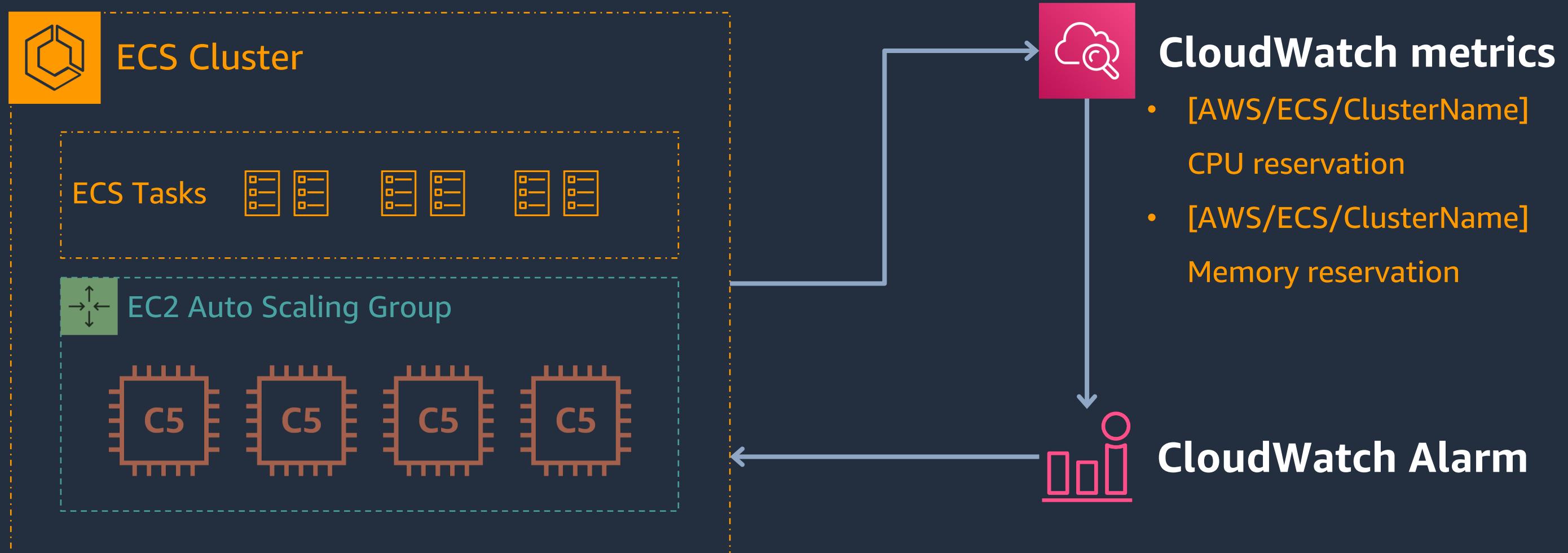
ECS on EC2 における Auto Scaling とは

EC2 インスタンス / ECS タスクのそれぞれのレイヤで Auto Scaling の設定が必要



EC2 インスタンスレイヤの Auto Scaling

ECS タスクの状況と連動するメトリクスをターゲットに設定することで、
EC2 インスタンスのキャパシティをコンテナワークロードの需要に追従させる



ECS タスクレイヤの Auto Scaling

Application Auto Scaling を活用して ECS サービスの必要タスク数を自動的に増減

ターゲット追跡スケーリングポリシー

- 指定したメトリクスがターゲットの値に近づくように自動的に調整

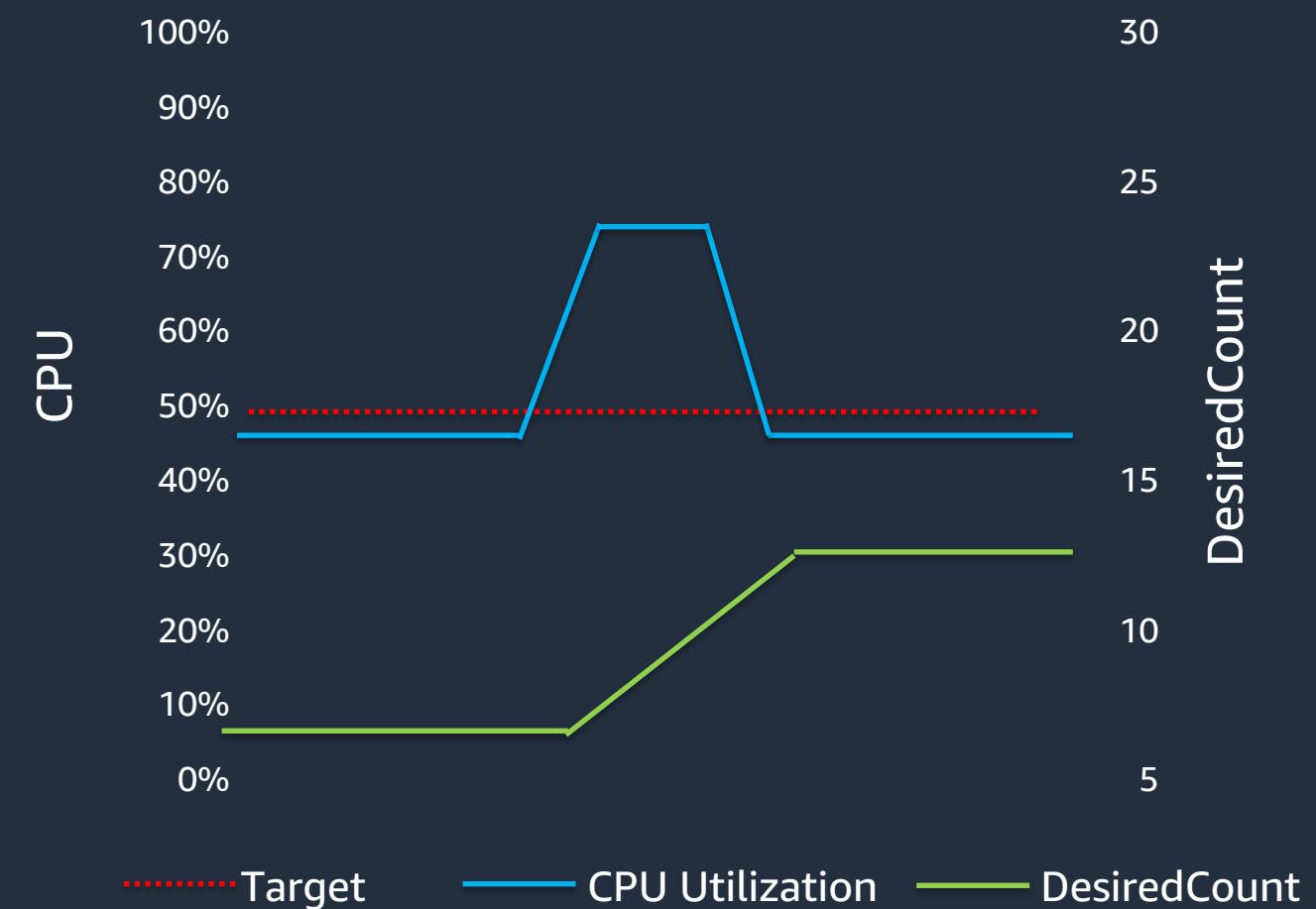
ステップスケーリングポリシー

- 設定されたスケーリングアクションに基づいて増減

スケジュールに基づくスケーリング

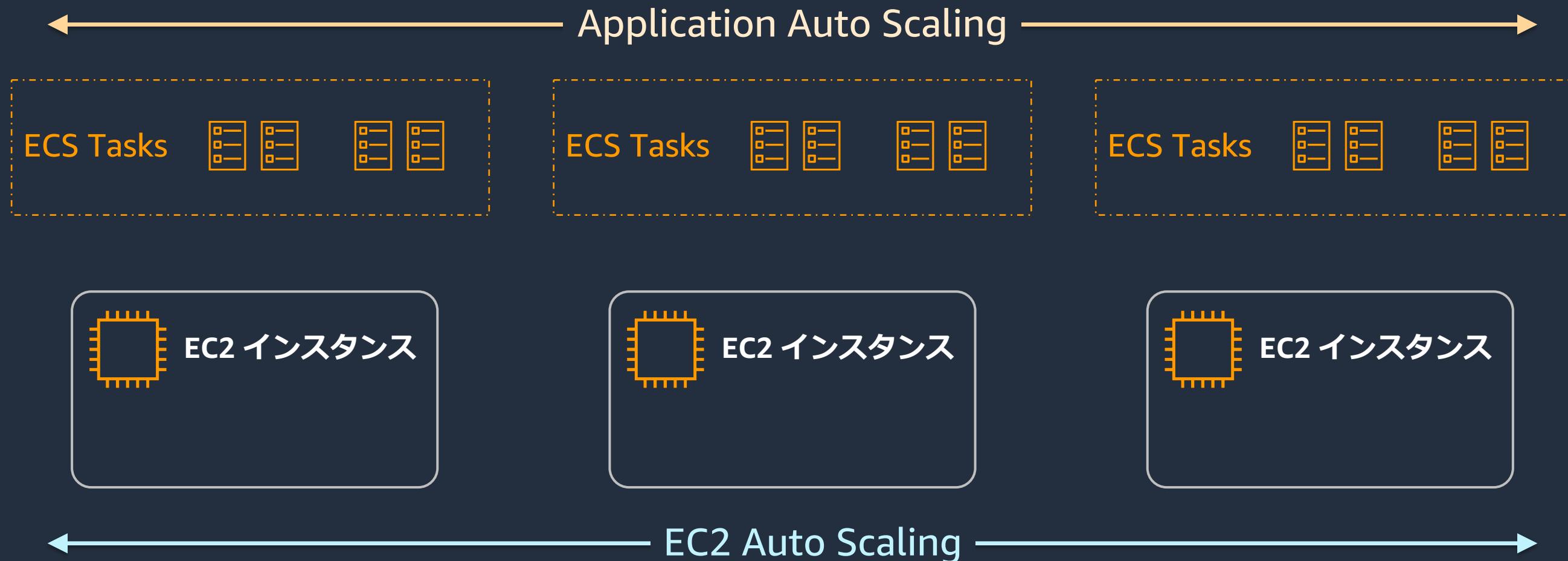
- 日付と時刻に基づいてタスク数を増減

例) ターゲット追跡スケーリングポリシー



ECS on EC2 における Auto Scaling の課題

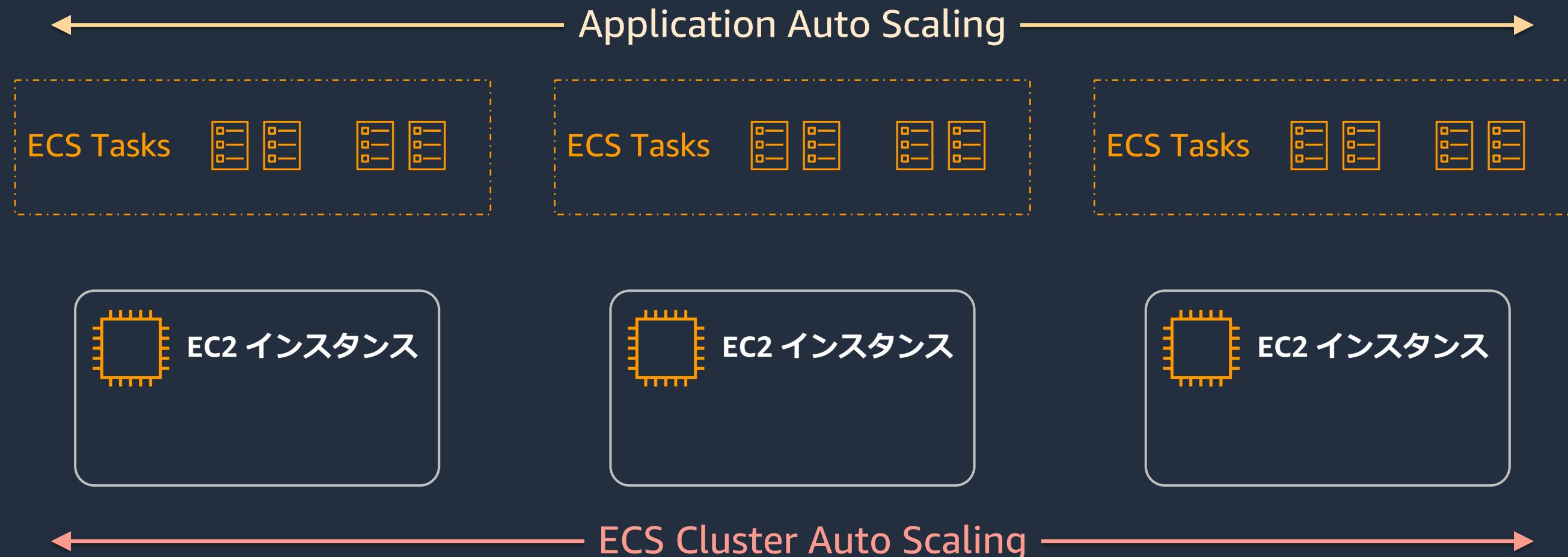
複数レイヤの Auto Scaling の管理は複雑になりやすく、運用コストの増加を招く



ECS on EC2 における Auto Scaling の課題

複数レイヤの Auto Scaling の管理は複雑になりやすく、運用コストの増加を招く

- ECS Cluster Auto Scaling により EC2 レイヤのマネージドスケーリングが可能に



Thank you!



Amazon EC2 Auto Scaling 入門編

AWS Black Belt Online Seminar

滝口 開資 (はるよし)
シニアソリューションアーキテクト
EC2 フレキシブルコンピュートスペシャリスト
2023/04

AWS Black Belt Online Seminarとは

- ・ 「サービス別」「ソリューション別」「業種別」などのテーマに分け、
アマゾン ウェブ サービス ジャパン合同会社が提供するオンラインセミナー
シリーズです
- ・ AWS の技術担当者が、AWS の各サービスやソリューションについてテーマ
ごとに動画を公開します
- ・ 動画を一時停止・スキップすることで、興味がある分野・項目だけの聴講も
可能、スキマ時間の学習にもお役立ていただけます
- ・ 以下の URL より、過去のセミナー含めた資料などをダウンロードすることができます
 - ・ <https://aws.amazon.com/jp/aws-jp-introduction/aws-jp-webinar-service-cut/>
 - ・ <https://www.youtube.com/playlist?list=PLzWGOASvSx6FIwIC2X1nObr1KcMCBBlqY>

内容についての注意点

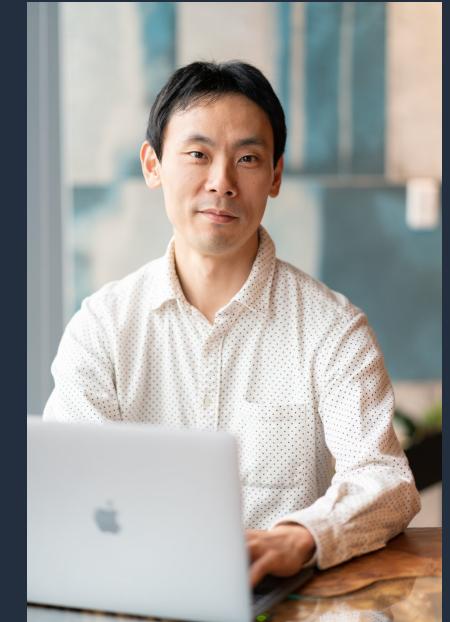
- ・ 本資料では 2023 年 4 月時点のサービス内容および価格についてご説明しています。最新の情報は AWS 公式ウェブサイト (<https://aws.amazon.com/>) にてご確認ください
- ・ 資料作成には十分注意しておりますが、資料内の価格と AWS 公式ウェブサイト記載の価格に相違があった場合、AWS 公式ウェブサイトの価格を優先とさせていただきます
- ・ 価格は税抜表記となっています。日本居住者のお客様には別途消費税をご請求させていただきます

自己紹介

名前：滝口 開資 (はるよし)

所属：アマゾンウェブサービスジャパン合同会社 コンピュート事業本部
シニアソリューションアーキテクト
EC2 フレキシブルコンピュートスペシャリスト

経歴：銀行様担当メインフレーム SE (外資ベンダー)
→クラウドサポートエンジニア (AWS)
→クラウドサポートチームリード (AWS)
→ソリューションアーキテクト (AWS)



好きなAWSサービス：Amazon EC2 Auto Scaling, AWSサポート

本セミナーの対象者

AWS 基盤環境のインフラを担当されている方

EC2 インスタンスの自動スケールの基礎を知りたい方

本セミナーの前提知識：Amazon EC2 の概要

→ Blackbelt Amazon EC2 入門を参照してください。

動画：<https://www.youtube.com/watch?v=1ALvDtb2ziM>

資料：https://pages.awscloud.com/rs/112-TZM-766/images/202111_AWS_Black_Belt_AWS_EC2_introduction.pdf

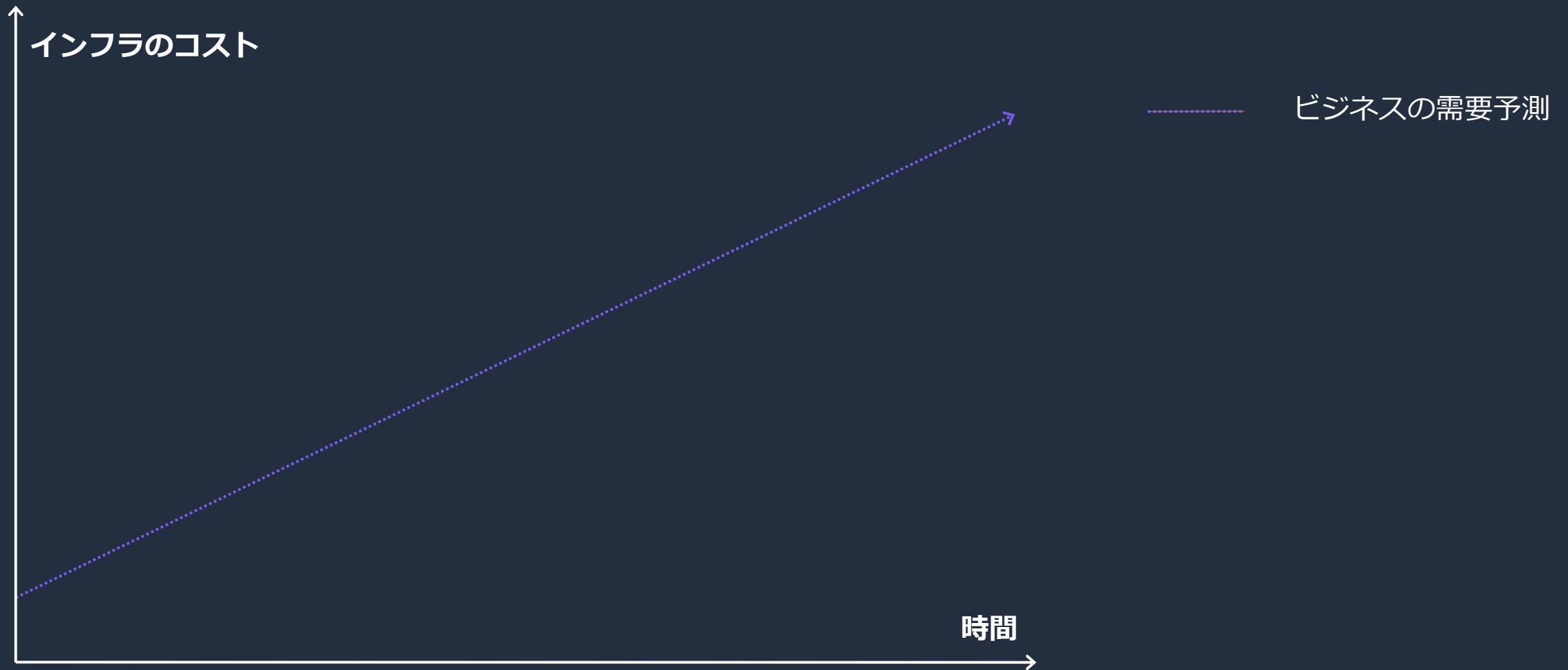
アジェンダ

- Auto Scaling サービスのコンセプト
- EC2 Auto Scaling の動作原理
- EC2 Auto Scaling を使ってみる

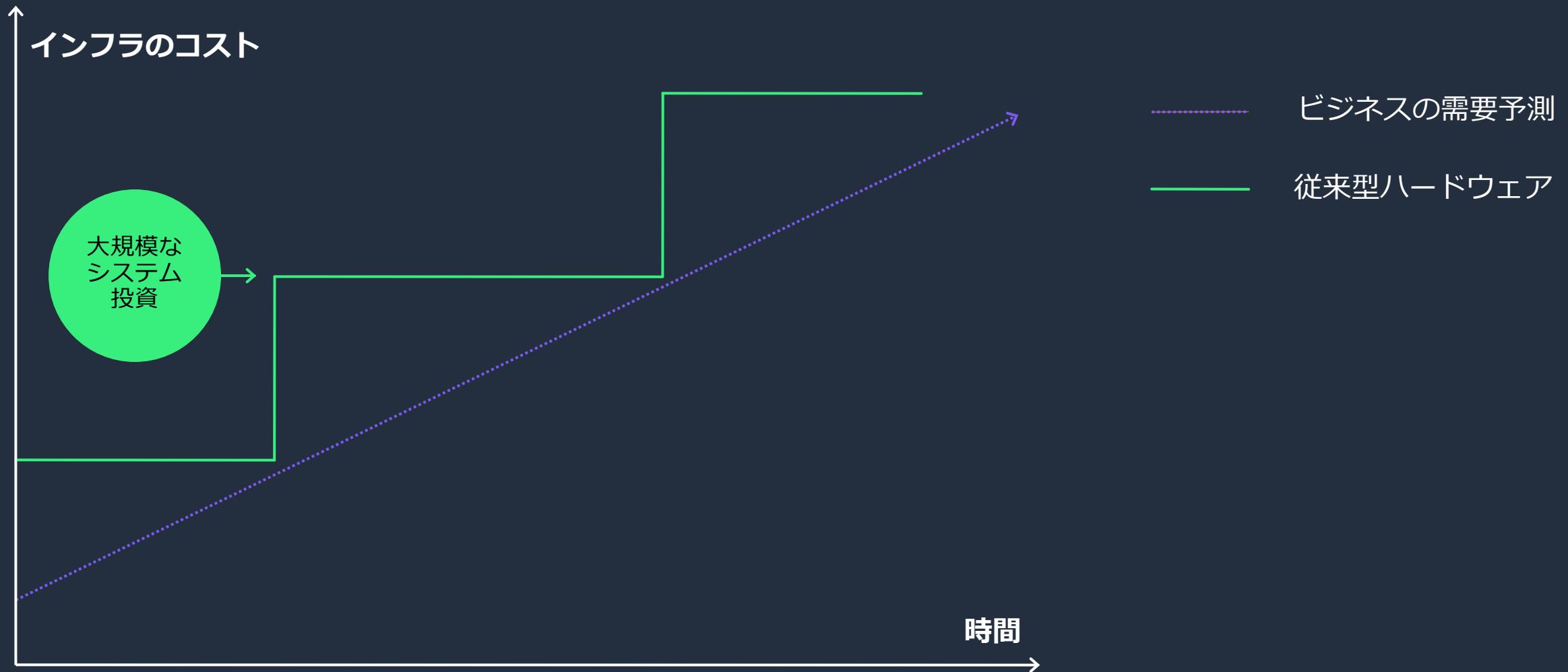
Auto Scaling サービスの コンセプト



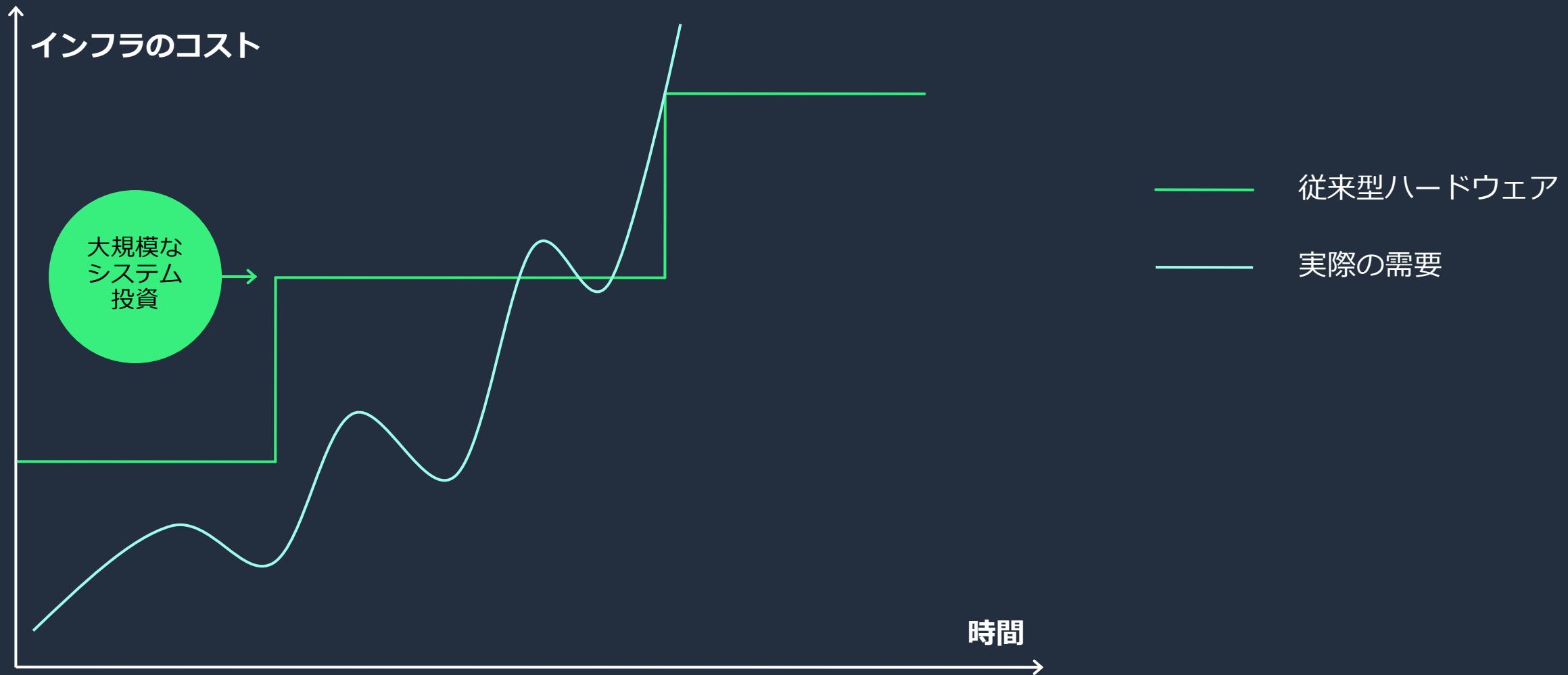
ビジネス需要に応じたキャパシティ準備



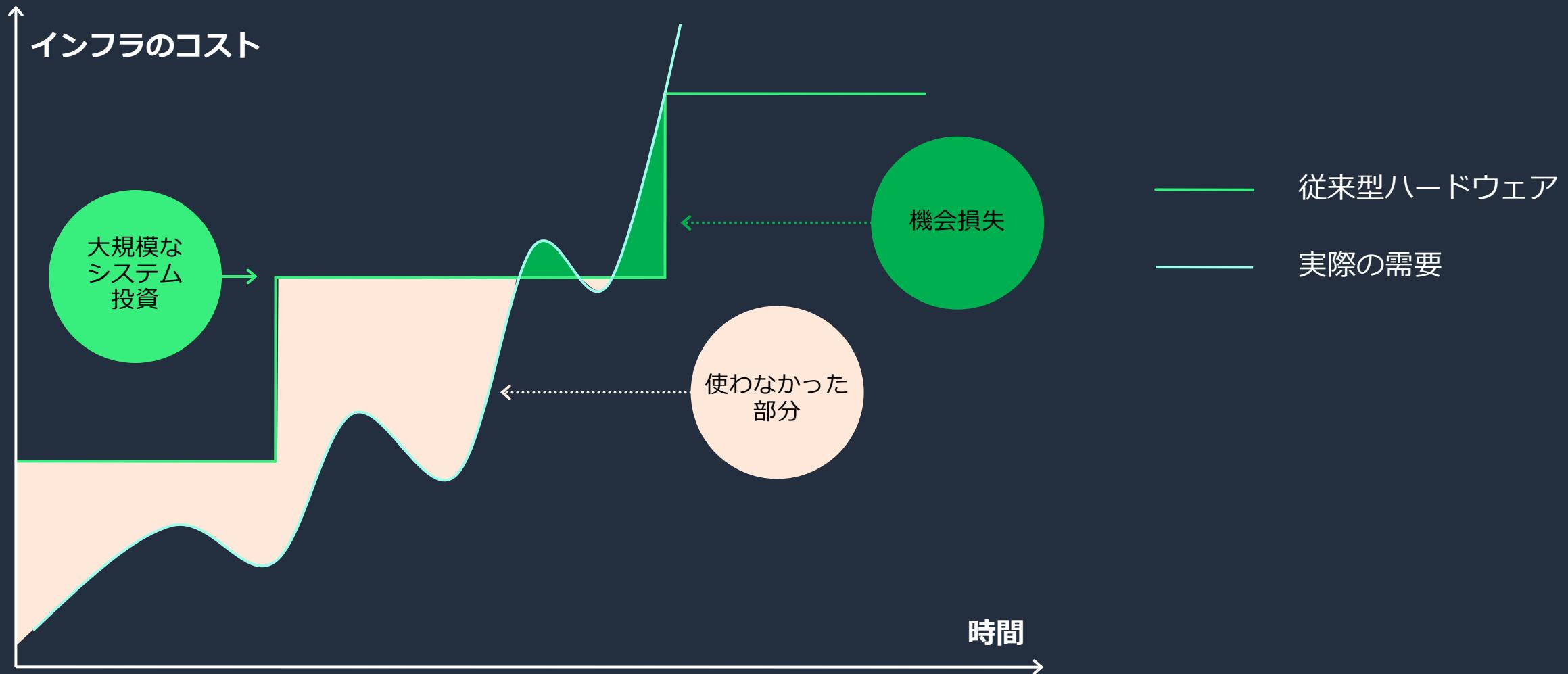
ビジネス需要に応じたキャパシティ準備



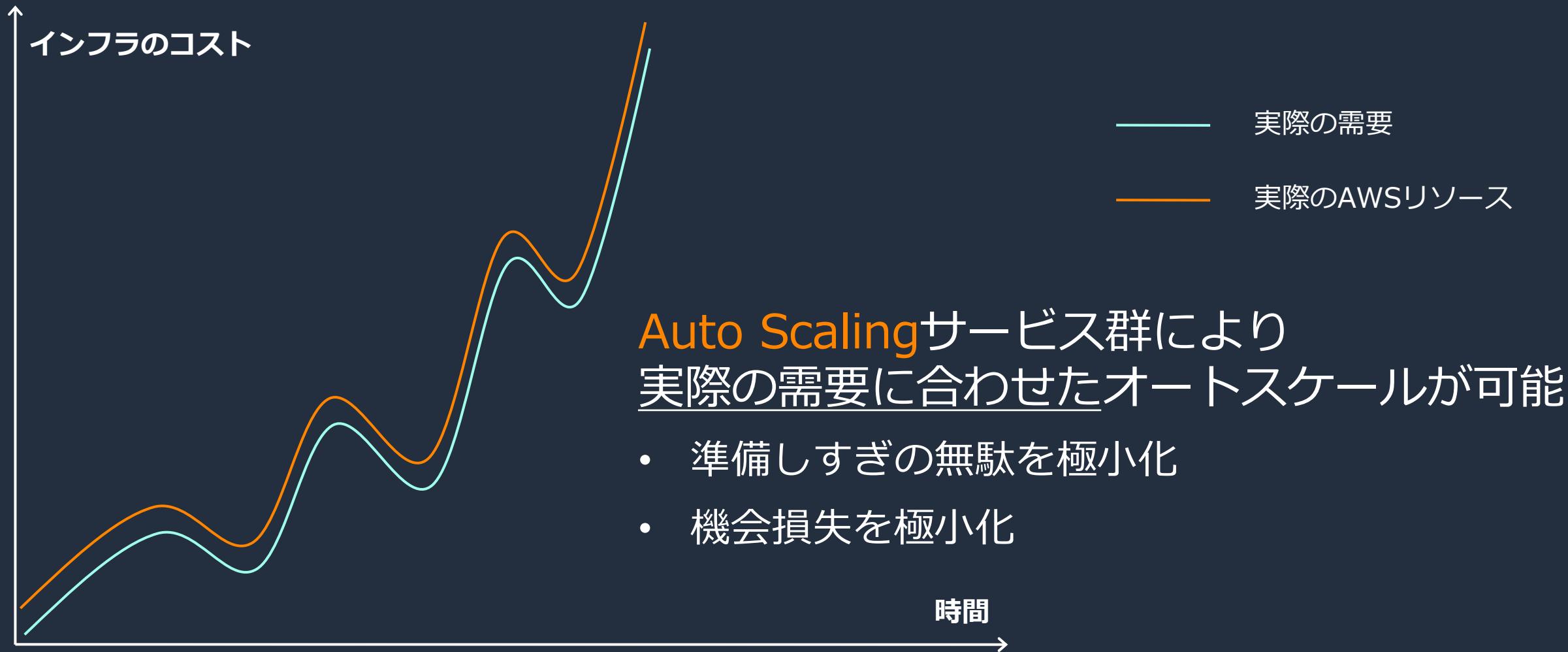
ビジネス需要に応じたキャパシティ準備



ビジネス需要に応じたキャパシティ準備



ビジネス需要に応じたキャパシティ準備



EC2 Auto Scaling の動作原理



EC2 Auto Scalingの動作原理

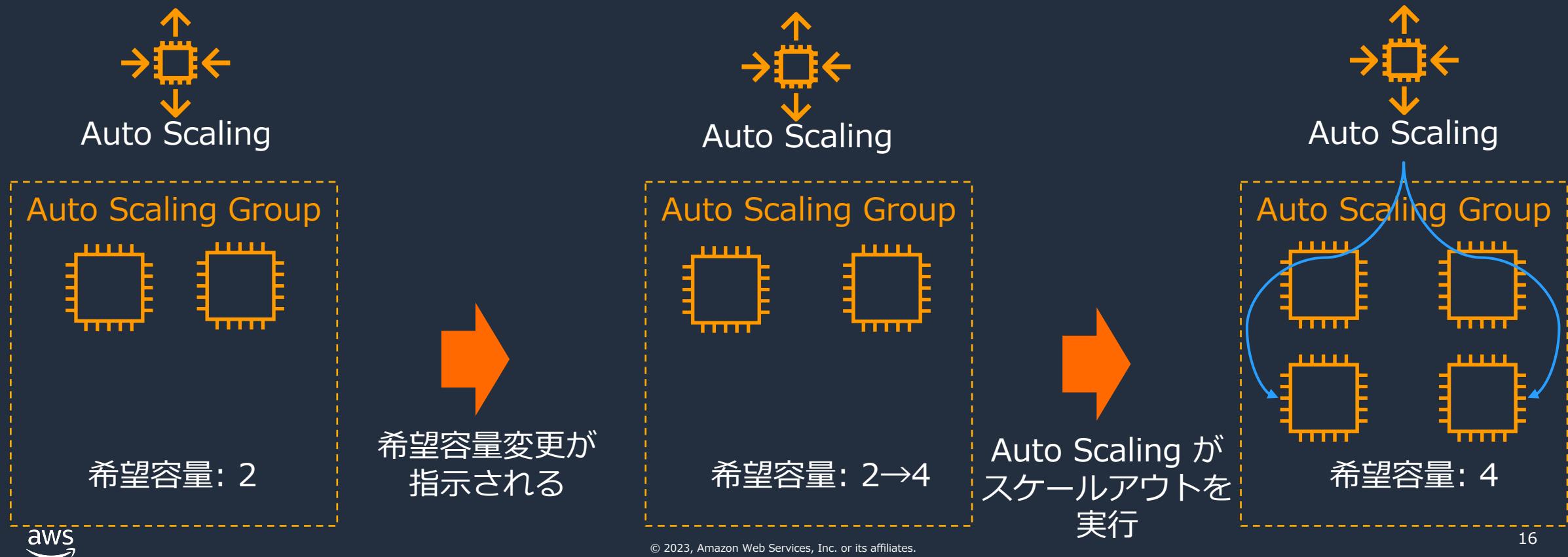
- 指定キャパシティの維持(台数維持) + 条件に応じた自動スケール
- インスタンスの分散(AZ間リバランス)

EC2 Auto Scalingの動作原理

- 指定キャパシティの維持(台数維持) + 条件に応じた自動スケール
- インスタンスの分散(AZ間リバランス)

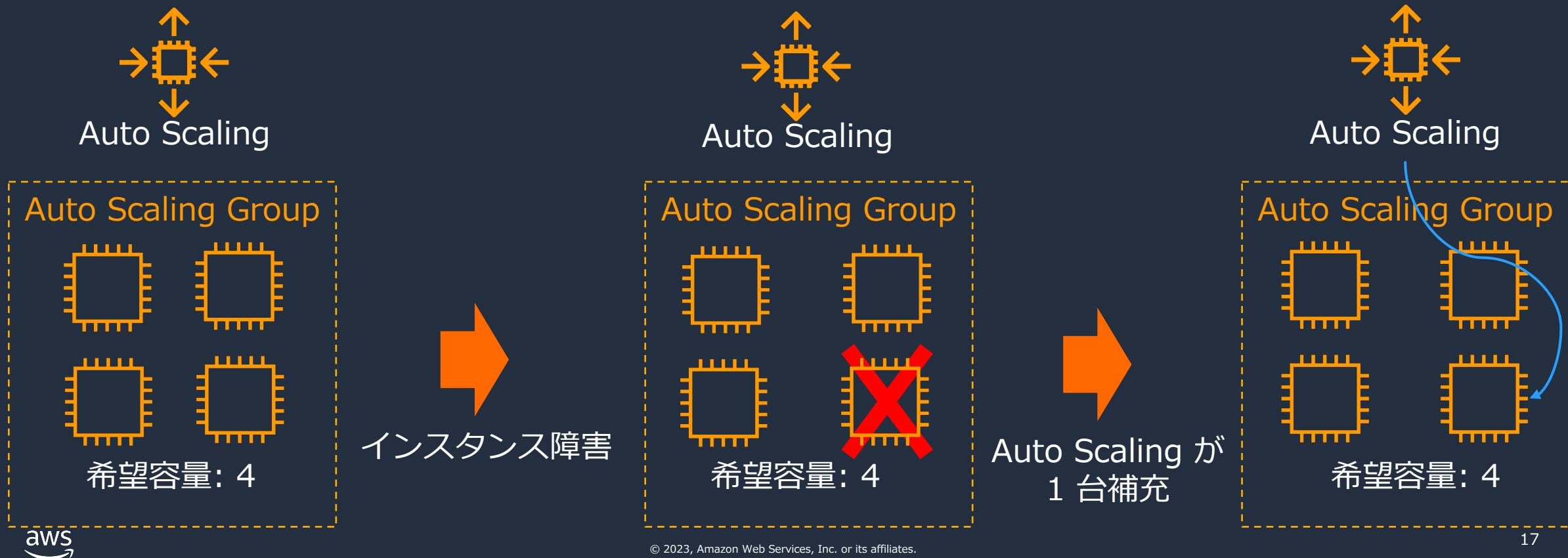
指定キャパシティの維持

EC2 Auto Scalingの動作原理(1)：指定された数の EC2 インスタンスを維持する



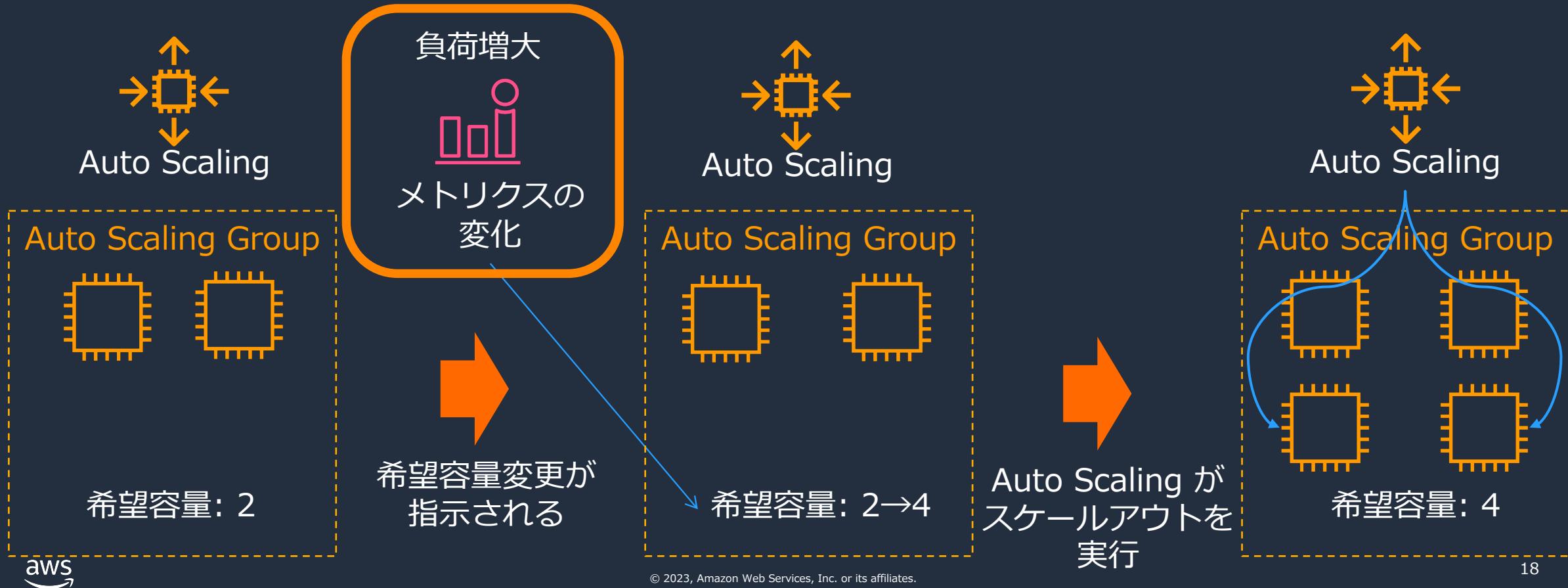
指定キャパシティの維持

EC2 Auto Scalingの動作原理(1)：指定された数の EC2 インスタンスを維持する



条件に応じた自動スケール

自動スケール：負荷に応じて自動的に増減してくれる仕組み

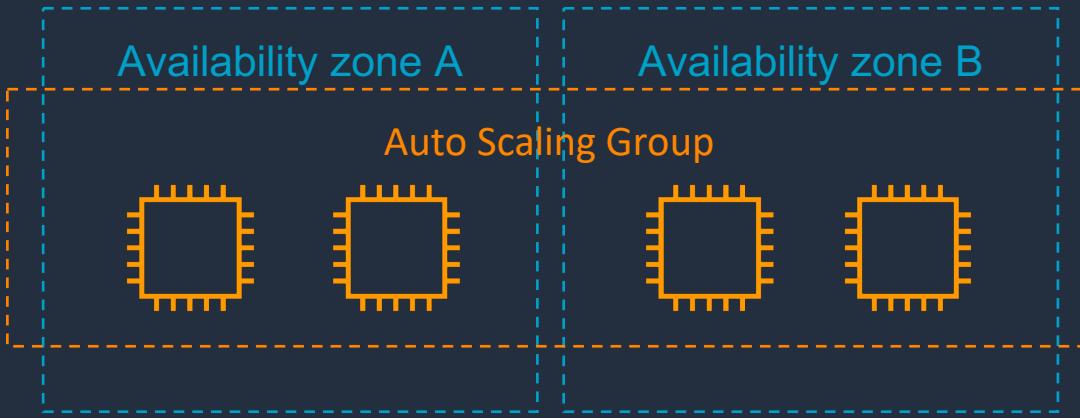


EC2 Auto Scalingの動作原理

- ・ 指定キャパシティの維持(台数維持) + 条件に応じた自動スケール
- ・ インスタンスの分散(AZ間リバランス)

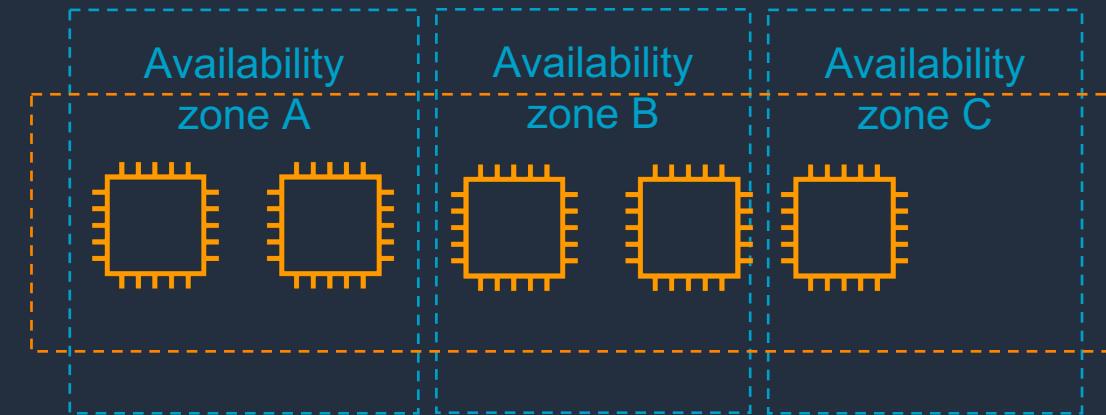
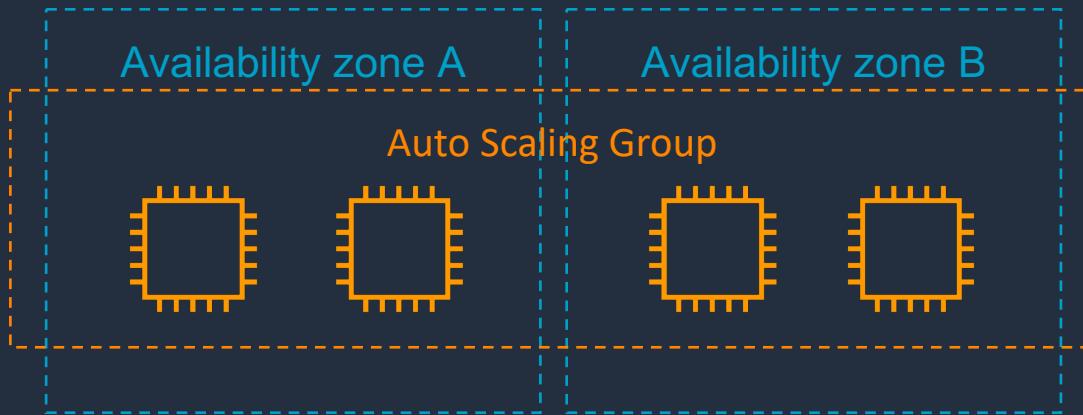
インスタンスの分散

EC2 Auto Scalingの動作原理(2)：使用できるアベイラビリティゾーンの間で均等にインスタンスを配置しようとする



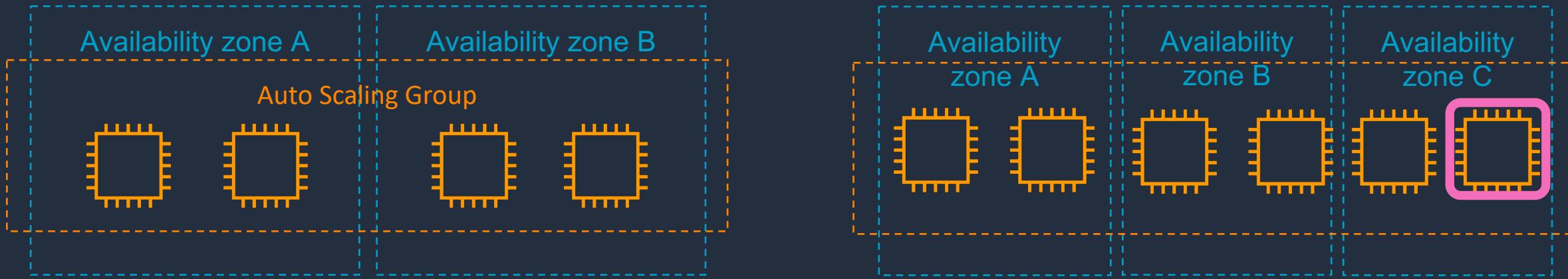
インスタンスの分散

EC2 Auto Scalingの動作原理(2)：使用できるアベイラビリティゾーンの間で均等にインスタンスを配置しようとする



インスタンスの分散

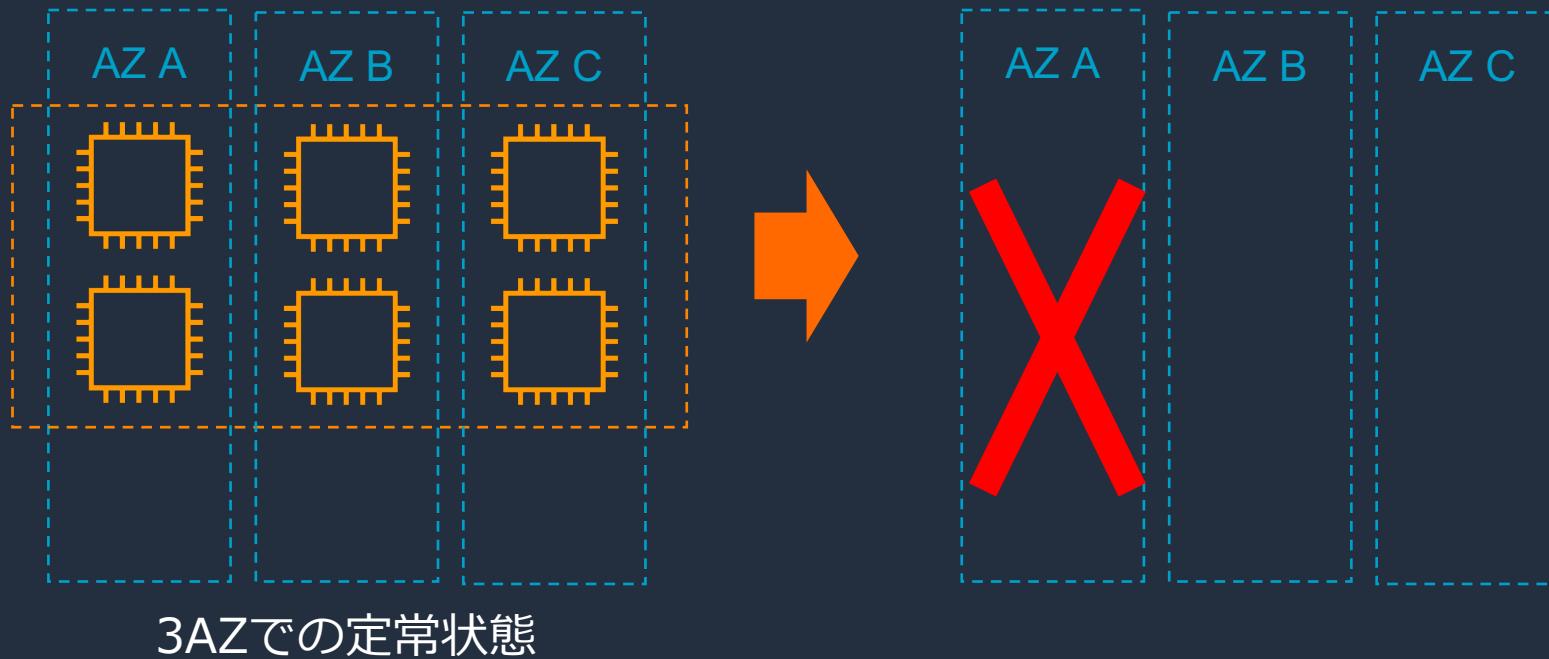
EC2 Auto Scalingの動作原理(2)：使用できるアベイラビリティゾーンの間で均等にインスタンスを配置しようとする



- スケールアウトするとき、インスタンス数が最も少ないアベイラビリティゾーンに新規起動
- これに失敗する場合、別のアベイラビリティゾーンが選択される

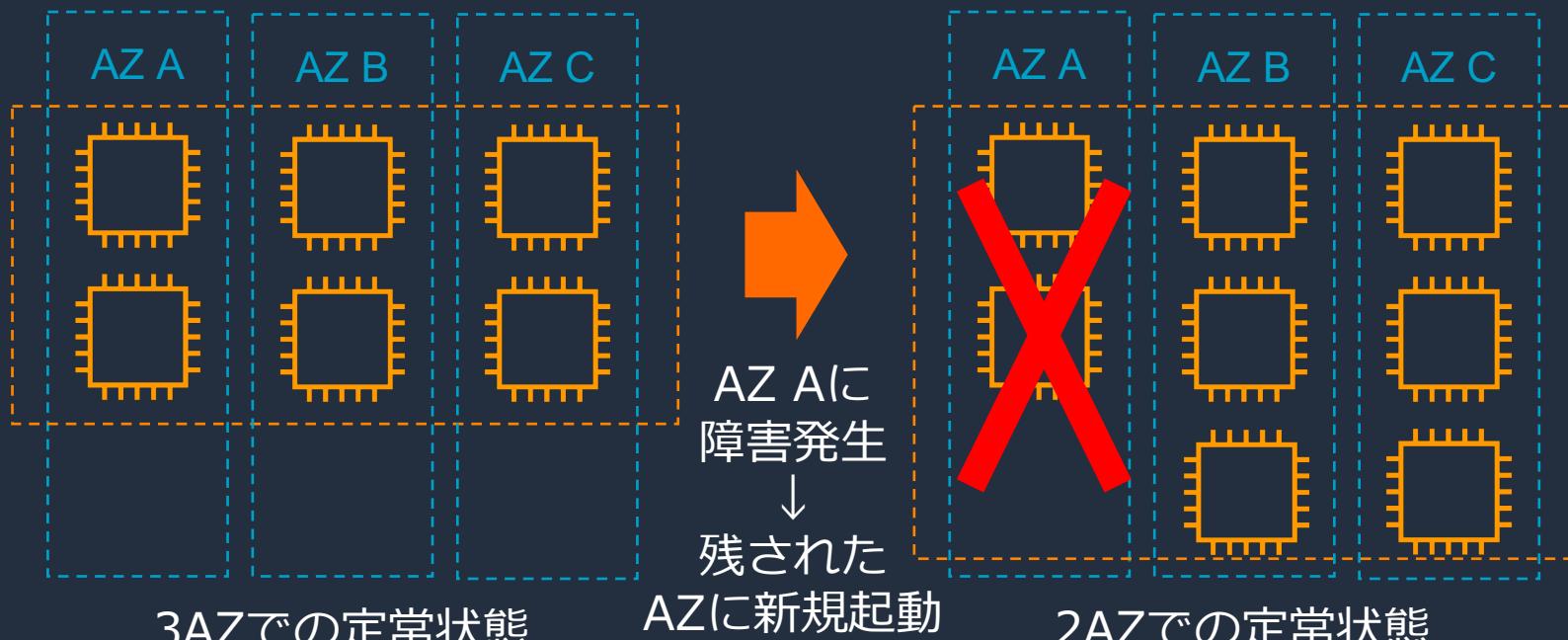
よくある質問：アベイラビリティゾーン (AZ) 障害時の動作

- 3AZ 構成で 6 台を起動しているとき、1AZ が障害になつたらどのような動作になるか？



よくある質問：アベイラビリティゾーン (AZ) 障害時の動作

- 3AZ 構成で 6 台を起動しているとき、1AZ が障害になつたらどのような動作になるか？
 - 残された AZ の間で均等に希望容量を維持する
 - サブネット定義が残っている限り、障害になつた AZ での起動をゆるやかに試みる

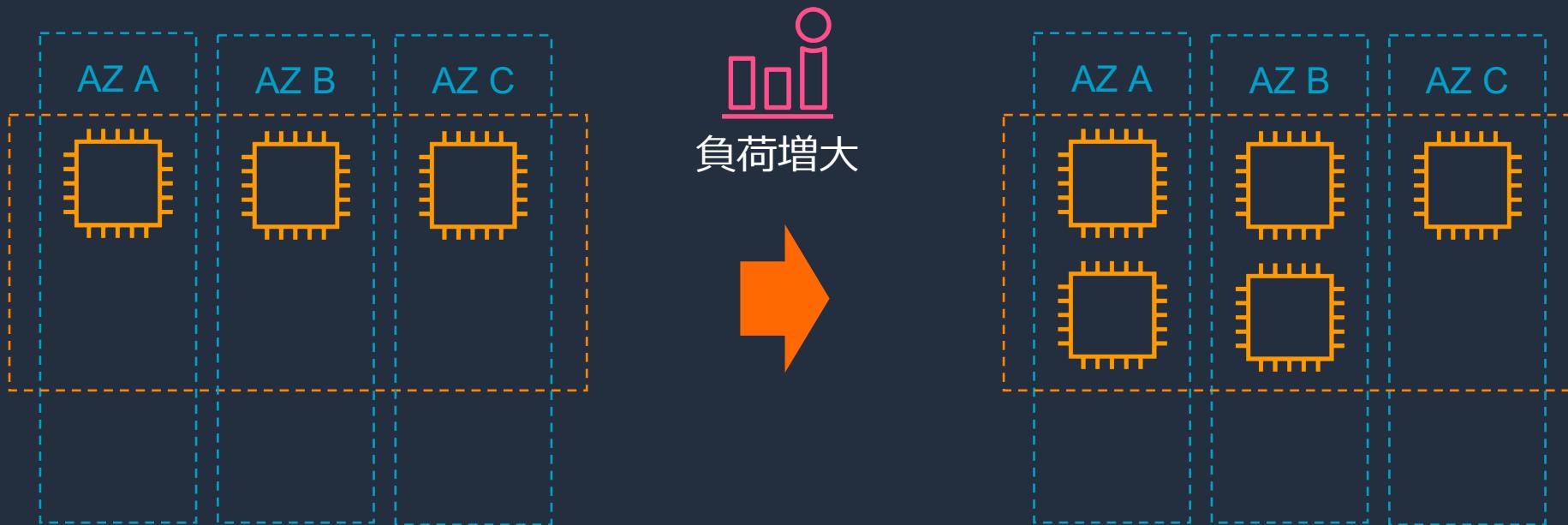


EC2 Auto Scaling を使ってみる



実現する構成のイメージ図

- ・3つのアベイラビリティーゾーン (AZ) 構成
- ・合計 3 台のオンデマンドインスタンスを起動する
- ・インスタンスタイプは t3.micro を選択する
- ・全体の CPU 使用率が 50% となるよう自動スケールさせる



EC2 Auto Scaling を使ってみる

- ・起動テンプレートの準備
- ・Auto Scaling グループの作成
- ・自動スケールの設定 (スケーリングポリシーの作成)

EC2 Auto Scaling を使ってみる

- 起動テンプレートの準備
- Auto Scaling グループの作成
- 自動スケールの設定 (スケーリングポリシーの作成)

起動テンプレートの準備 (1)

▼ インスタンス

インスタンス

インスタンスタイプ

起動テンプレート

スポットリクエスト

Savings Plans

リザーブドインスタンス

起動テンプレートを使用して、インスタンスの作成の自動化、アクセス権限ポリシーの簡素化、および組織全体のベストプラクティスの強化を行うことができます。起動パラメータをテンプレートに保存して、オンデマンドでの起動や、EC2 Auto Scaling や EC2 フリートなどのマネージドサービスで使用できます。新しい起動テンプレートのバージョンを作成することで、起動パラメータを簡単に更新できます。

新しい起動テンプレート

起動テンプレートを作成

- 「起動テンプレートを作成」を押す

起動テンプレートの準備 (2)

EC2 > 起動テンプレート > 起動テンプレートを作成

起動テンプレートを作成

起動テンプレートを作成することで、後で再利用、共有、起動できる保存済みインスタンス設定を作成できます。テンプレートには複数のバージョンを含めることができます。

起動テンプレート名と説明

起動テンプレート名 - 必須

このアカウントに固有である必要があります。最大 128 文字です。スペースや「&」、「*」、「@」などの特殊文字は使用できません。

テンプレートバージョンの説明

最大 255 文字

Auto Scaling のガイダンス 情報
EC2 Auto Scaling でこのテンプレートを使用する場合は、これを選択します
 EC2 Auto Scaling で使用できるテンプレートをセットアップする際に役立つガイダンスを提供

▶ テンプレートタグ
▶ ソーステンプレート

起動テンプレートのコンテンツ

起動テンプレートの詳細を以下で指定します。フィールドを空白のままにすると、フィールドが起動テンプレートに含まれません。

▼ アプリケーションおよび OS イメージ (Amazon マシンイメージ) **必須 情報**

AMI は、インスタンスの起動に必要なソフトウェア設定 (オペレーティングシステム、アプリケーションなど) を含むテンプレートです。お探しのものが以下に表示されない場合は、AMI を検索または参照してください。

- 起動テンプレートの名前を入力
- 「EC2 Auto Scaling で使用できるテンプレートをセットアップする際に役立つガイダンスを提供」にチェック
- これ以降の設定項目に「必須」表示が追加されるようになる



起動テンプレートの準備 (3)

起動テンプレートのコンテンツ

起動テンプレートの詳細を以下で指定します。フィールドを空白のままにすると、フィールドが起動テンプレートに含まれません。

▼ アプリケーションおよび OS イメージ (Amazon マシンイメージ) - 必須 [情報](#)

AMI は、インスタンスの起動に必要なソフトウェア設定 (オペレーティングシステム、アプリケーションサーバー、アプリケーション) を含むテンプレートです。お探しのものが以下に表示されない場合は、AMI を検索または参照してください。

Q 何千ものアプリケーションイメージと OS イメージを含むカタログ全体を検索します。

クイックスタート

Amazon Linux macOS Ubuntu Windows Red Hat S > その他の AMI を見る 検索 AWS、Marketplace、コミュニティからの AMI を含む

Amazon マシンイメージ (AMI)

Amazon Linux 2023 AMI 無料利用枠の対象 ami-0779c326801d5a843 (64 ビット (x86), uefi-preferred) / ami-020155f62513c0fc1 (64 ビット (Arm), uefi)
仮想化: hvm ENA 有効: true ルートデバイストラップ: ebs

説明

Amazon Linux 2023 AMI 2023.0.20230315.0 x86_64 HVM kernel-6.1

アーキテクチャ ブートモード AMI ID
64 ビット (x86) uefi-preferred ami-0779c326801d5a843 検証済みプロバイダー



- AMIを選択

起動テンプレートの準備 (4)



- 「起動テンプレートを作成」を押す

EC2 Auto Scaling を使ってみる

- ・起動テンプレートの準備
- ・Auto Scaling グループの作成
- ・自動スケールの設定 (スケーリングポリシーの作成)

Auto Scaling グループの作成 (1)

The screenshot shows the AWS CloudFormation console with the following interface elements:

- Sidebar:** A navigation menu on the left side with sections: Elastic Block Store, ネットワーク & セキュリティ, ロードバランシング, and Auto Scaling. The "Auto Scaling" section is highlighted with a purple box, and "Auto Scaling グループ" is selected.
- Main Content Area:** A large central area with the heading "Amazon EC2 Auto Scaling は、アプリケーションの可用性を維持するために役立ちます". Below it is a description: "Auto Scaling グループは、オートスケーリングとフリート管理機能を可能にする Amazon EC2 インスタンスのコレクションです。これらの機能は、アプリケーションの正常性と可用性を維持するために役立ちます。".
- Call-to-Action:** A button labeled "Auto Scaling グループを作成する" located in a white box with a purple border.
- Side Panels:** Two panels on the right side:
 - 仕組み:** An icon showing two boxes connected by arrows, labeled "Auto Scaling group".
 - 料金:** A description of the costs associated with Auto Scaling, mentioning Amazon EC2, CloudWatch, and other AWS services.

- 「Auto Scaling グループを作成する」を押す

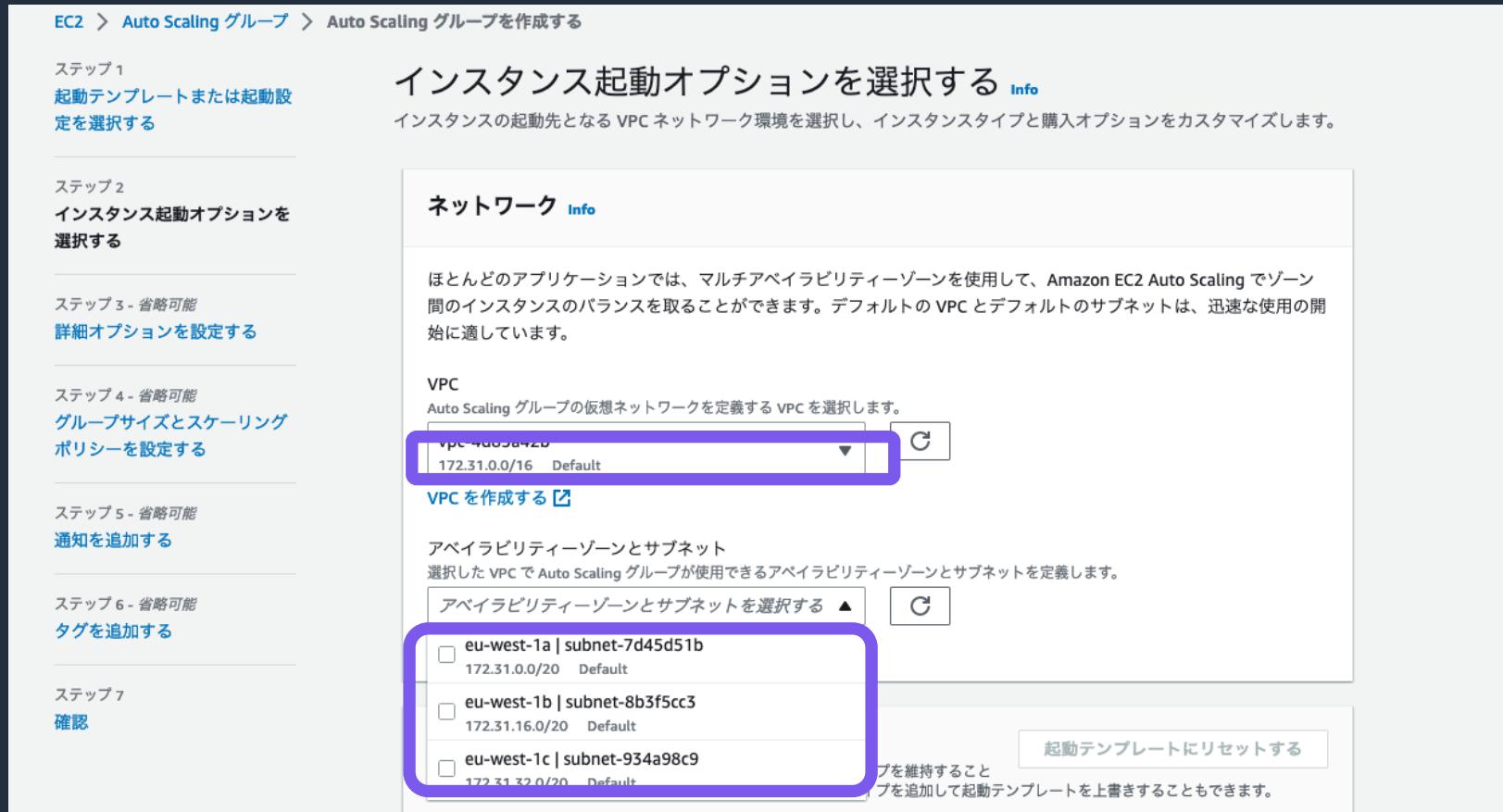
Auto Scaling グループの作成 (2)

The screenshot shows the 'Auto Scaling グループを作成する' (Create Auto Scaling Group) wizard. On the left, a sidebar lists steps 1 through 7. Step 1 is active, titled '起動テンプレートまたは起動設定を選択する'. The main area has a heading '起動テンプレートまたは起動設定を選択する' with an 'Info' link. It contains instructions about selecting a launch template for all instances in the group. Below this is a 'Name' input field containing 'my_first_asg', which is highlighted with a purple rectangle. A note below says '現在のリージョンにあるこのアカウントに固有で、255 文字以内にする必要があります。' To the right is a 'Launch Template' section with a dropdown menu showing 'my_first_launch_template' selected, also highlighted with a purple rectangle. Buttons for '次へ' (Next) and 'キャンセル' (Cancel) are at the bottom.

- Auto Scaling グループ名を入力
- 先ほど作成した起動テンプレート名を選択

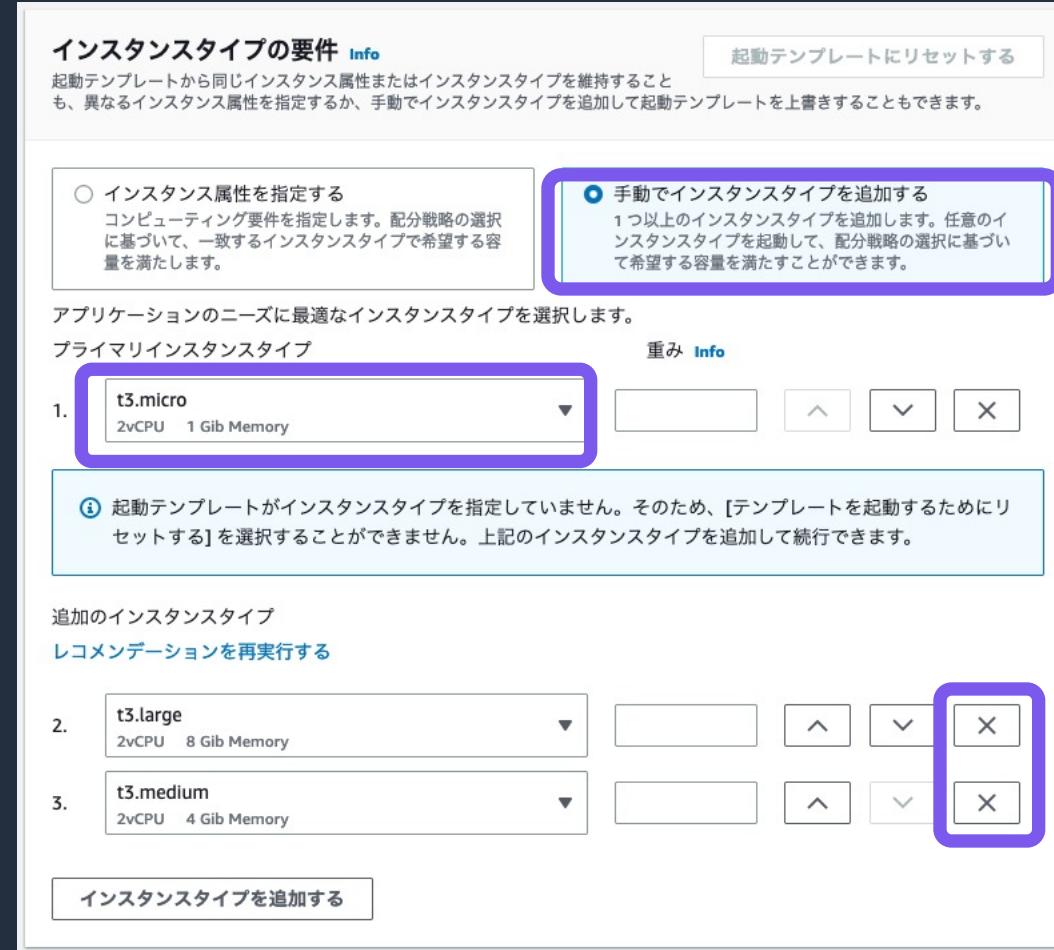
Auto Scaling グループの作成 (3)

ヒント
特別な理由がない限り
そのリージョンで使用
できるすべての
アベイラビリティー^{ゾーン}が含まれる
ようにしましょう



- VPC を選択
- その VPC 内に定義されたサブネットを選択

Auto Scaling グループの作成 (4)



ヒント

複数インスタンスタイプの活用について
は後続の「複数のインスタンスタイプと購入オプションの活用編」で解説します

- 「手動でインスタンスタイプを追加」を選択
- 要件に合ラインスタンスタイプを選択
- 自動推奨されるインスタンスタイプを使用しない場合は削除

Auto Scaling グループの作成 (5)

インスタンスの購入オプション [Info](#)

インスタンスの分散
耐障害性のあるワークロードを低成本で実行するには、スポットインスタンスとなるインスタンスの割合を定義します。スポットインスタンスは、AWS が 2 分前に通知することで変更できるオンデマンド料金に比べて大幅な割引を提供する予備の EC2 容量です。

100 % オンデマンド

0 % スポット

オンデマンドベース容量を含める
パーセンテージでスケールする前に、Auto Scaling グループがそのベース部分のために使用するオンデマンド容量を指定します。最大グループサイズはこの値まで増加します(減少することはできません)。

0 オンデマンドインスタンス

配分戦略 [Info](#)

オンデマンド配分戦略
オンデマンドインスタンスの起動時に適用する配分戦略を選択します。

高い優先順位で設定済み
上記で設定したインスタンスタイプの優先順位に基づいて、オンデマンドインスタンスをリクエストします。この戦略は、属性ベースのインスタンスタイプの選択では使用できません。

最低料金
アベイラビリティゾーン内で最低料金のプールからオンデマンドインスタンスをリクエストします。

① [インスタンスタイプの要件] のインスタンスタイプが 1 つのみであるため、このセクションは使用できません。スポットインスタンスまたはオンデマンドインスタンスで配分戦略を使用するには、少なくとも 2 つのインスタンスタイプまたは一連のインスタンス属性を指定する必要があります。

キャンセル スキップして確認 戻る 次へ

- 残りの項目はそのままにし、「次へ」を押す

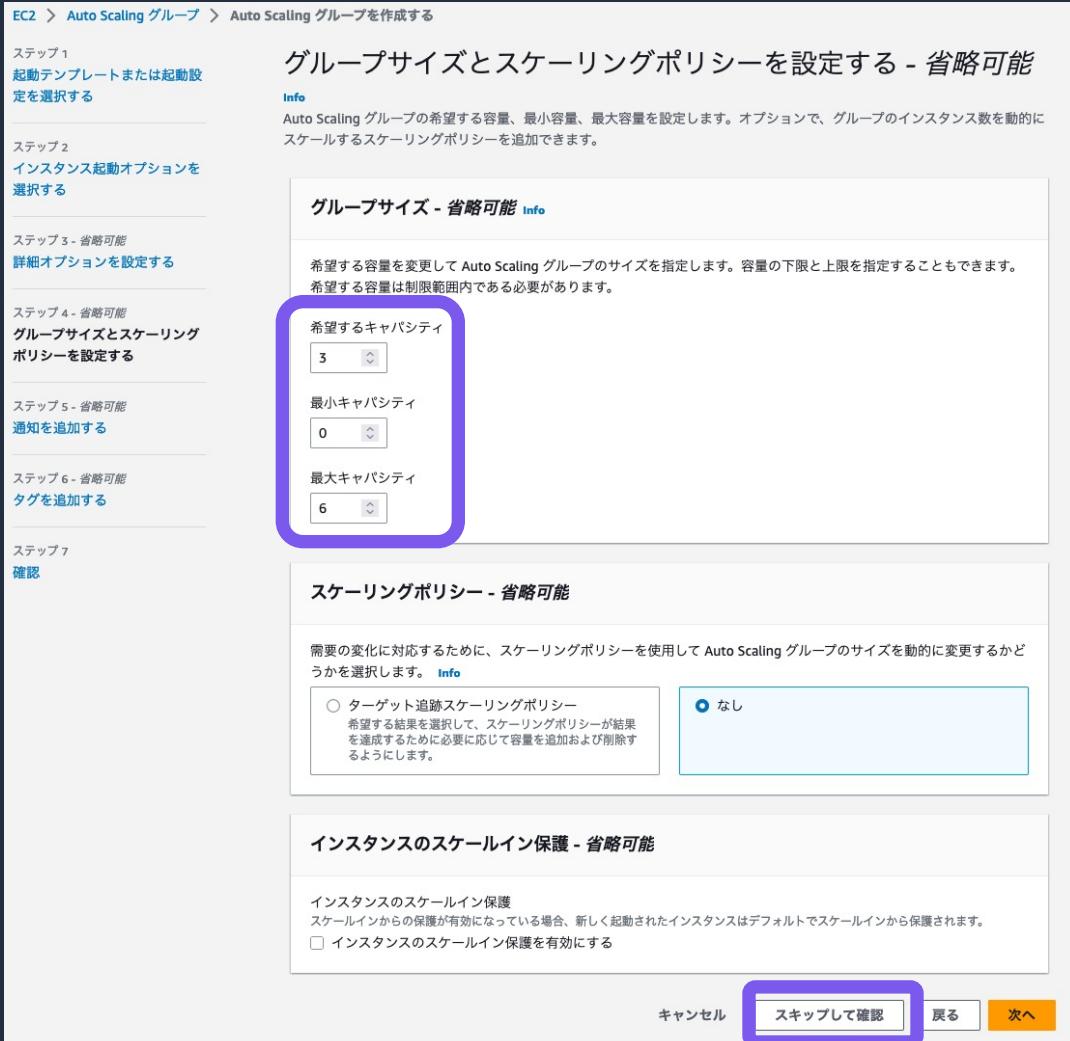
Auto Scaling グループの作成 (6)



- 残りの項目はそのままにし、「次へ」を押す

- グループメトリクスを有効にするのがオススメ。ASG内の台数などをCloudWatchから確認できる

Auto Scaling グループの作成 (7)



- 最初は 3 台起動する
- 0 - 6 台の間でスケールできるようにする

- 残りの項目はそのままにし、「スキップして確認」を押す

Auto Scaling グループの作成 (8)

EC2 > Auto Scaling グループ > Auto Scaling グループを作成する

確認 [Info](#)

ステップ 1
起動テンプレートまたは起動設定を選択する

ステップ 2
インスタンス起動オプションを選択する

ステップ 3 - 省略可能
[詳細オプションを設定する](#)

ステップ 4 - 省略可能
グループサイズとスケーリングポリシーを設定する

ステップ 5 - 省略可能
通知を追加する

ステップ 6 - 省略可能
[タグを追加する](#)

ステップ 7
確認

確認 Info

ステップ 1: 起動テンプレートまたは起動設定を選択する

編集

グループの詳細

Auto Scaling グループ名
my_first_asg

起動テンプレート

起動テンプレート	バージョン	説明
my_first_launch_template	Default	lt-0e5c87a9fe91678ef

ステップ 2: インスタンスの起動オプションを選択する

編集

ネットワーク

ネットワーク

ステップ 6: タグを追加する

編集

タグ (0)

キー	▼	値	▼	新しいインスタンスをタグ付けする	▼
タグなし					

キャンセル 戻る [Auto Scaling グループを作成する](#)

- 内容を確認し、「Auto Scaling グループを作成する」を押す

Auto Scaling グループの作成確認と詳細画面への遷移



The screenshot shows the AWS Auto Scaling Groups page. At the top, there is a breadcrumb navigation: EC2 > Auto Scaling グループ. Below the breadcrumb, a search bar contains the placeholder text "Auto Scaling グループを検索する". To the right of the search bar are buttons for "C" (Create), "編集" (Edit), "削除" (Delete), and "Auto Scaling グループを作成する" (Create Auto Scaling Group). A yellow box highlights the "Auto Scaling グループを作成する" button. Below these buttons is a pagination control with arrows and the number "1".

The main table lists one Auto Scaling group:

名前	起動テンプレート/設定	インス...	ステータス	希望するキャ...	最小	最大
my_first_asg	my_first_launch_template バージョン 3	-	3	0	6	

- 作成された Auto Scaling グループ名をクリック

EC2 インスタンス起動状況の確認

The screenshot shows the AWS Auto Scaling Groups console. The navigation path is EC2 > Auto Scaling グループ > my_first_asg. The main tab 'my_first_asg' is selected. Below it, the 'Activities' tab is highlighted with a purple rectangle. The 'Activities' section displays the following details:

グループの詳細		
Auto Scaling グループ名 my_first_asg	希望するキャパシティ 3	ステータス -
作成日 Sun Mar 19 2023 23:13:48 GMT+0900 (Japan Standard Time)	最小キャパシティ 0	
	最大キャパシティ 6	

Below this, there is a section titled '起動テンプレート'.

- 「アクティビティ」をクリック

EC2 インスタンス起動状況の確認

The screenshot shows the AWS Auto Scaling Groups console for the group 'my_first_asg'. The 'Activity' tab is selected. The 'Successful' status column is highlighted with a purple box. The table lists three successful launches of EC2 instances:

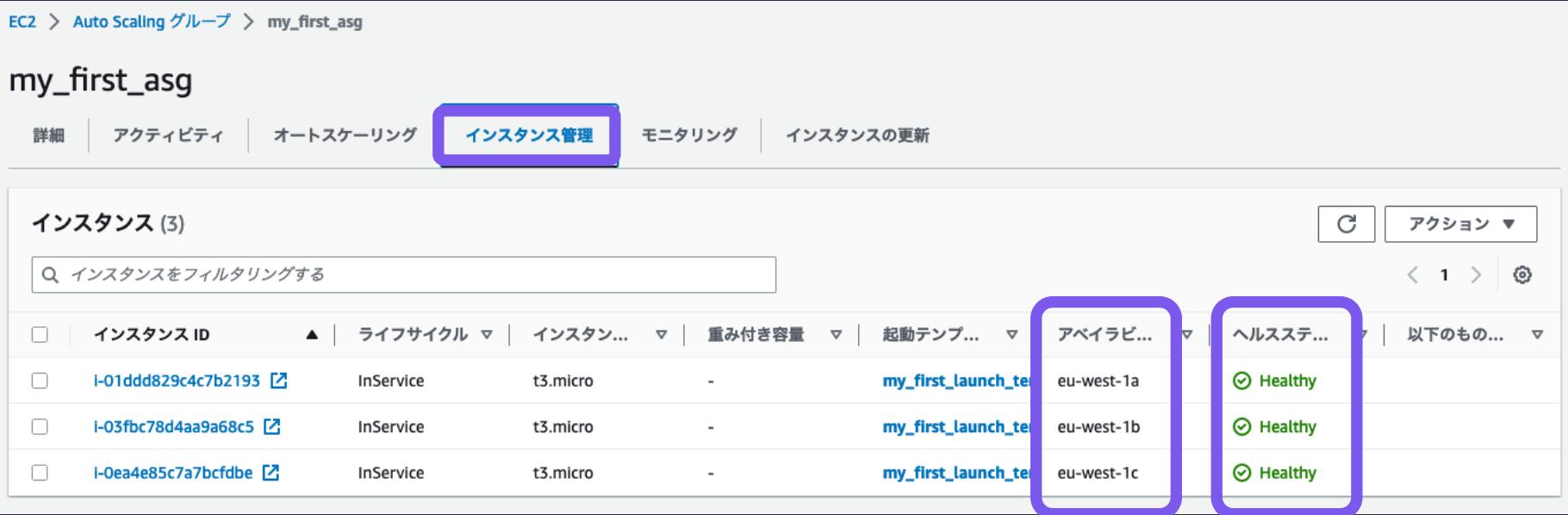
ステータス	説明	原因	開始時刻	終了時刻
Successful	Launching a new EC2 Instance: i-0ea4e85c7a7bcfdbe	At 2023-03-19T14:13:48Z a user request created an AutoScalingGroup changing the desired capacity from 0 to 3. At 2023-03-19T14:13:50Z an instance was started in response to a difference between desired and actual capacity, increasing the capacity from 0 to 3.	2023 March 19, 11:13:54 PM +09:00	2023 March 19, 11:14:00 PM +09:00
Successful	Launching a new EC2 Instance: i-01ddd829c4c7b2193	At 2023-03-19T14:13:48Z a user request created an AutoScalingGroup changing the desired capacity from 0 to 3. At 2023-03-19T14:13:50Z an instance was started in response to a difference between desired and actual capacity, increasing the capacity from 0 to 3.	2023 March 19, 11:13:54 PM +09:00	2023 March 19, 11:14:00 PM +09:00
Successful	Launching a new EC2 Instance: i-03fb78d4aa9a68c5	At 2023-03-19T14:13:48Z a user request created an AutoScalingGroup changing the desired capacity from 0 to 3. At 2023-03-19T14:13:50Z an instance was started in response to a difference between desired and actual capacity, increasing the capacity from 0 to 3.	2023 March 19, 11:13:54 PM +09:00	2023 March 19, 11:14:00 PM +09:00

ヒント

Auto Scaling グループのトラブルシューティングには「アクティビティ」タブから動作記録を確認しましょう

- ステータス列を確認
- エラーが発生している場合、Auto Scaling ドキュメントのトラブルシューティング節を参照して対処する
 - https://docs.aws.amazon.com/ja_jp/autoscaling/ec2/userguide/CHAP_Troubleshooting.html

EC2 インスタンス起動状況の確認



The screenshot shows the AWS EC2 Auto Scaling Groups console for the group 'my_first_asg'. The 'Instances Management' tab is selected. The table displays three instances:

インスタンス ID	ライフサイクル	インスタンス型	重み付き容量	起動テンプレート	アベイラビリティゾーン	ヘルステータス
i-01ddd829c4c7b2193	InService	t3.micro	-	my_first_launch_template	eu-west-1a	Healthy
i-03fbc78d4aa9a68c5	InService	t3.micro	-	my_first_launch_template	eu-west-1b	Healthy
i-0ea4e85c7a7bcfdbe	InService	t3.micro	-	my_first_launch_template	eu-west-1c	Healthy

Three instances are highlighted with purple boxes: eu-west-1a, eu-west-1b, and eu-west-1c. Each instance has a green checkmark icon next to the word 'Healthy'.

- それぞれのインスタンスがどのアベイラビリティーゾーンに起動されたかを確認
- ヘルステータスが Healthy であることを確認

EC2 Auto Scaling を使ってみる

- ・起動テンプレートの準備
- ・Auto Scaling グループの作成
- ・自動スケールの設定 (スケーリングポリシーの作成)

スケーリングポリシーの作成



- 「動的スケーリングポリシーを作成する」を押す

スケーリングポリシーの作成 - ターゲット追跡スケーリング

EC2 > Auto Scaling グループ > my_first_asg

動的スケーリングポリシーを作成する

ポリシータイプ
ターゲット追跡スケーリング

スケーリングポリシー名
Target Tracking Policy

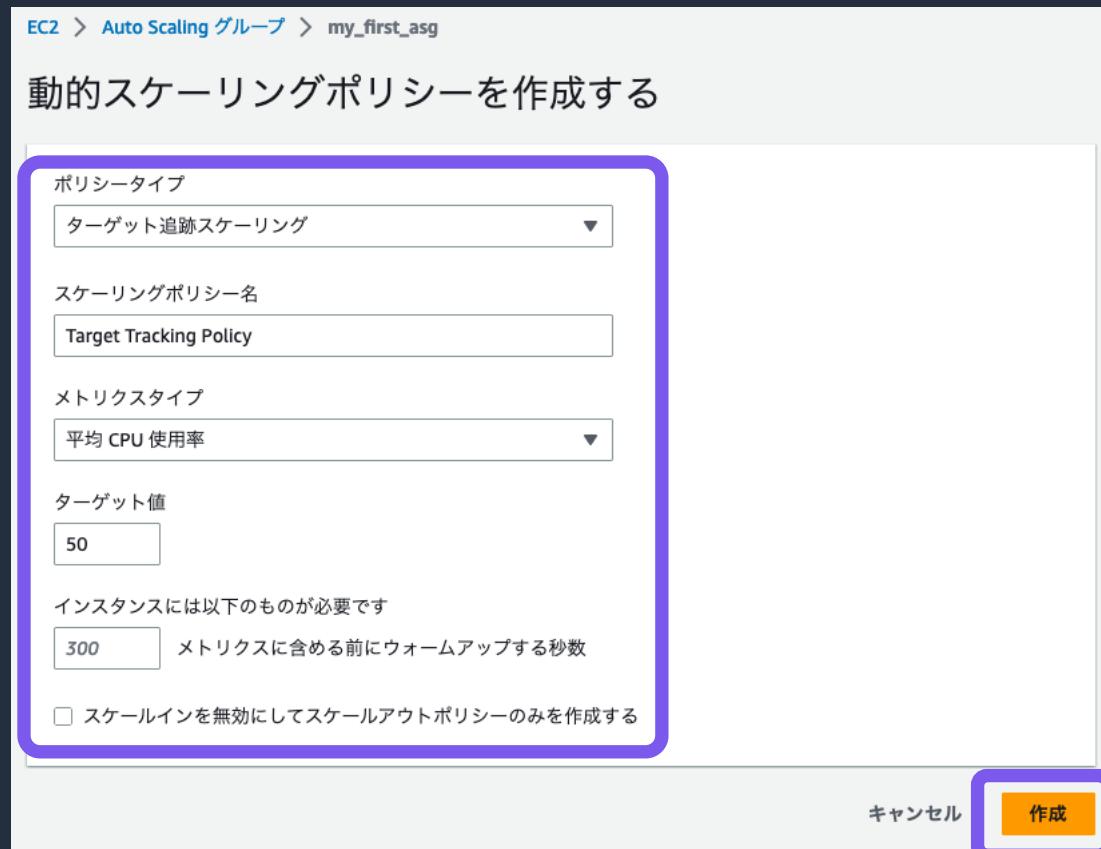
メトリクスタイプ
平均 CPU 使用率

ターゲット値
50

インスタンスには以下のものが必要です
300 メトリクスに含める前にウォームアップする秒数

スケールインを無効にしてスケールアウトポリシーのみを作成する

キャンセル 作成



スケーリングポリシーの作成 - ターゲット追跡スケーリング

EC2 > Auto Scaling グループ > my_first_asg

動的スケーリングポリシーを作成する

ポリシータイプ
ターゲット追跡スケーリング

スケーリングポリシー名
Target Tracking Policy

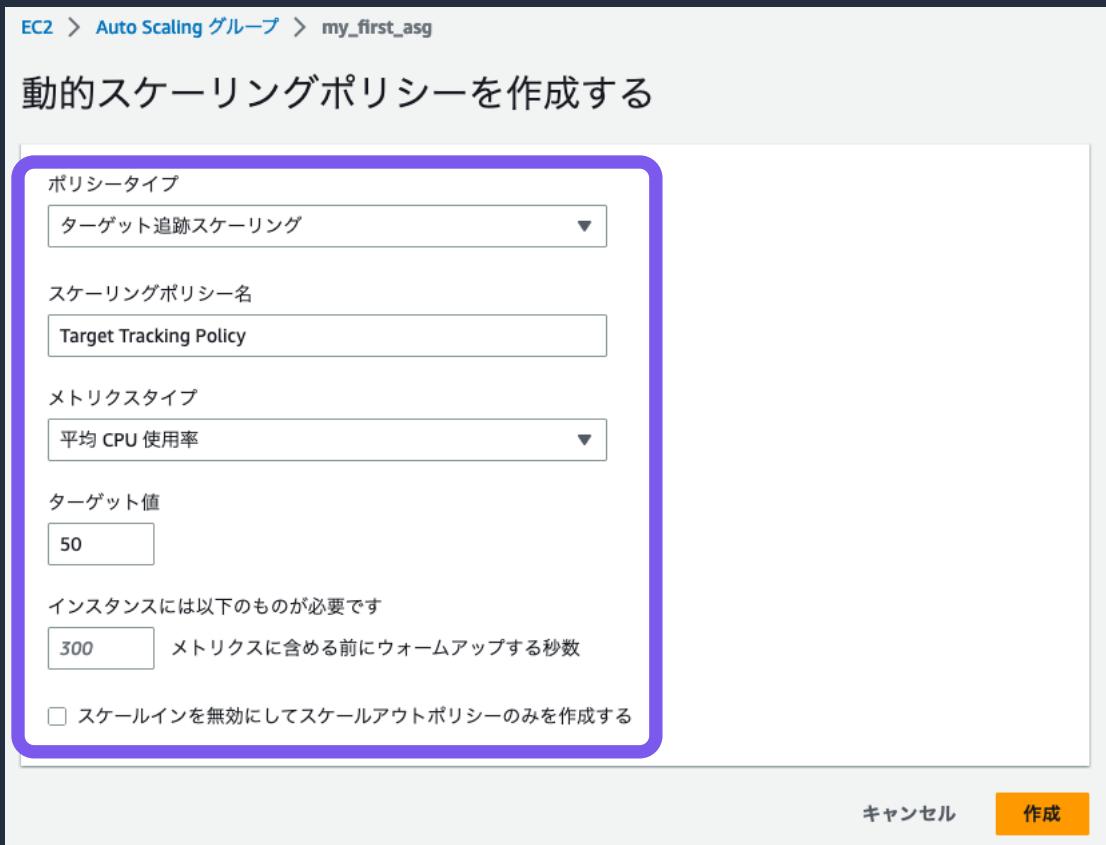
メトリクスタイプ
平均 CPU 使用率

ターゲット値
50

インスタンスには以下のものが必要です
300 メトリクスに含める前にウォームアップする秒数

スケールインを無効にしてスケールアウトポリシーのみを作成する

キャンセル 作成



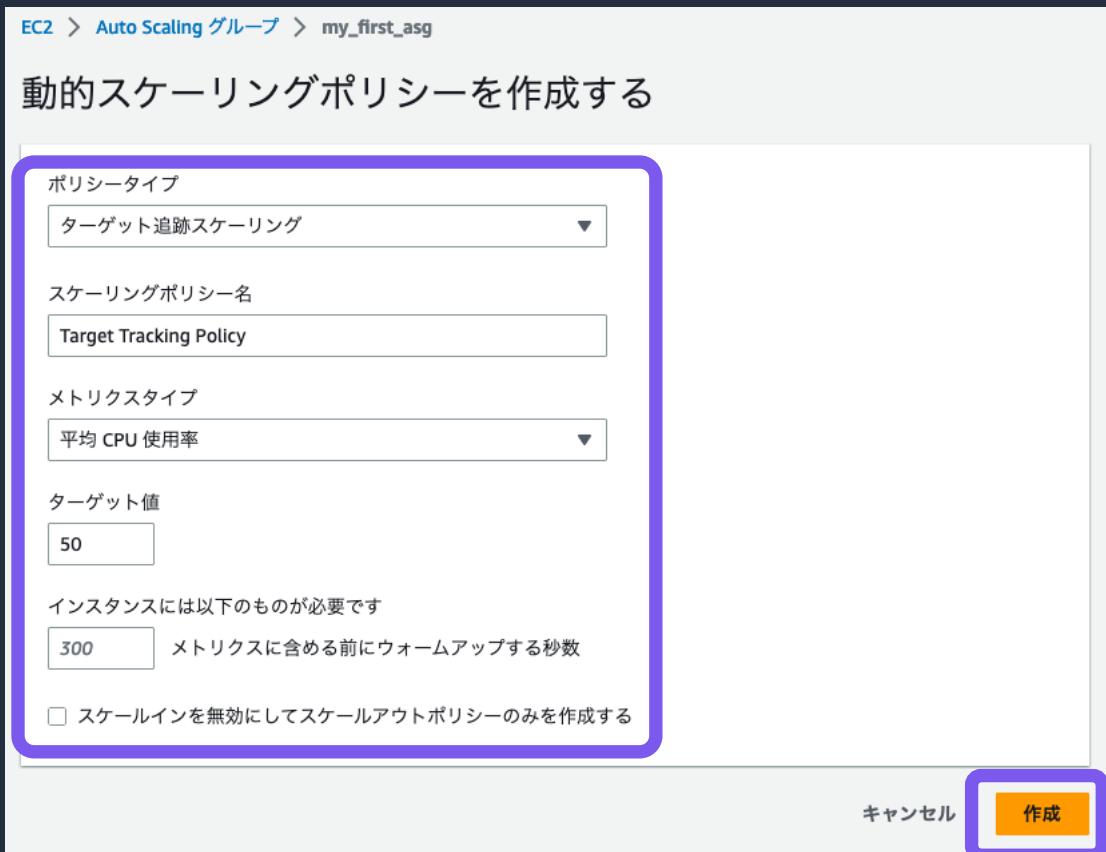
ヒント

スケーリングポリシーについては後続の「スケーリング編」で解説します



- この Auto Scaling グループではCPU 使用率を 50% に保ってほしい、と指定
- 自動スケールのためのCW Alarmが2本作られる
 - 負荷が上がり、50%を超えた期間がしばらく続くとスケールアウト（台数増加）
 - 負荷が下がり、50%以下の期間がある程度続くとスケールイン（台数減少）
- https://docs.aws.amazon.com/ja_jp/autoscaling/ec2/userguide/as-scaling-target-tracking.html

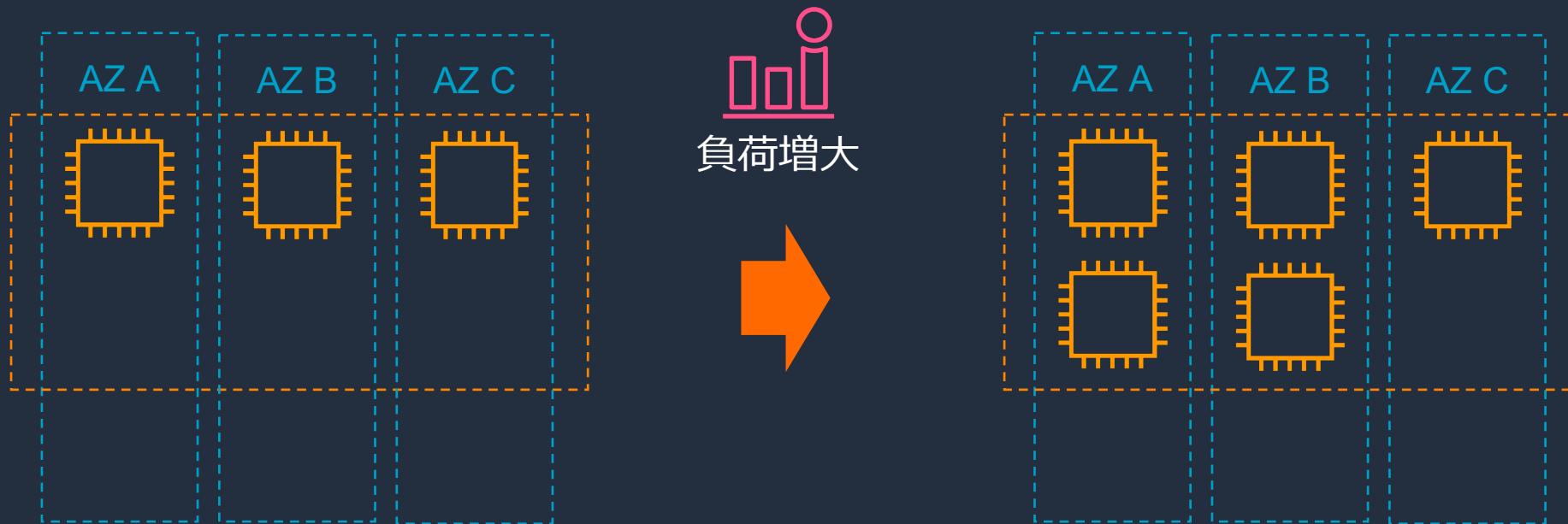
スケーリングポリシーの作成 - ターゲット追跡スケーリング



- この Auto Scaling グループではCPU 使用率を 50% に保ってほしい、と指定
 - 自動スケールのためのCW Alarmが2本作られる
 - 負荷が上がり、50%を超えた期間がしばらく続くとスケールアウト（台数増加）
 - 負荷が下がり、50%以下の期間がある程度続くとスケールイン（台数減少）
 - https://docs.aws.amazon.com/ja_jp/autoscaling/ec2/userguide/as-scaling-target-tracking.html
-
- 内容を確認し、「作成」を押す

実現する構成のイメージ図 (再掲)

- ・3つのアベイラビリティーゾーン (AZ) 構成
- ・合計3台のオンデマンドインスタンスを起動する
- ・インスタンスタイプは t3.micro を選択する
- ・全体のCPU使用率が50%となるよう自動スケールさせる



おわりに



今回お話しした内容

- Auto Scaling サービスのコンセプト
- EC2 Auto Scaling の動作原理
- EC2 Auto Scaling を使ってみる
- 実際の負荷をかけるテストを体験できます。「AWS Hands-on for Beginners - Amazon EC2 Auto Scaling スケーリング基礎編」をご覧ください
https://pages.awscloud.com/JAPAN-event-OE-Hands-on-for-Beginners-Auto_Scaling-2022-reg-event.html

本資料に関するお問い合わせ・ご感想

技術的な内容に関しましては、有料のAWSサポート窓口へ
お問い合わせください

<https://aws.amazon.com/jp/premiumsupport/>

料金面でのお問い合わせに関しましては、カスタマーサポート窓口へ
お問い合わせください（マネジメントコンソールへのログインが必要です）

<https://console.aws.amazon.com/support/home#/case/create?issueType=customer-service>

具体的な案件に対する構成相談は、後述する個別相談会をご活用ください



ご感想はTwitterへ！ハッシュタグは以下をご利用ください
#awsblackbelt



その他コンテンツのご紹介

ウェビナーなど、AWSのイベントスケジュールをご参照いただけます

<https://aws.amazon.com/jp/events/>

ハンズオンコンテンツ

<https://aws.amazon.com/jp/aws-jp-introduction/aws-jp-webinar-hands-on/>

AWS 個別相談会

AWSのソリューションアーキテクトと直接会話いただけます

<https://pages.awscloud.com/JAPAN-event-SP-Weekly-Sales-Consulting-Seminar-2021-reg-event.html>



Thank you!



Amazon EC2 Auto Scaling

複数のインスタンスタイプと購入オプションの活用編

AWS Black Belt Online Seminar

滝口 開資 (はるよし)
シニアソリューションアーキテクト
EC2 フレキシブルコンピュートスペシャリスト
2023/04

AWS Black Belt Online Seminarとは

- ・ 「サービス別」「ソリューション別」「業種別」などのテーマに分け、
アマゾン ウェブ サービス ジャパン合同会社が提供するオンラインセミナー
シリーズです
- ・ AWS の技術担当者が、AWS の各サービスやソリューションについてテーマ
ごとに動画を公開します
- ・ 動画を一時停止・スキップすることで、興味がある分野・項目だけの聴講も
可能、スキマ時間の学習にもお役立ていただけます
- ・ 以下の URL より、過去のセミナー含めた資料などをダウンロードすることができます
 - ・ <https://aws.amazon.com/jp/aws-jp-introduction/aws-jp-webinar-service-cut/>
 - ・ <https://www.youtube.com/playlist?list=PLzWGOASvSx6FIwIC2X1nObr1KcMCBBlqY>

内容についての注意点

- ・ 本資料では 2023 年 4 月時点のサービス内容および価格についてご説明しています。最新の情報は AWS 公式ウェブサイト (<https://aws.amazon.com/>) にてご確認ください
- ・ 資料作成には十分注意しておりますが、資料内の価格と AWS 公式ウェブサイト記載の価格に相違があった場合、AWS 公式ウェブサイトの価格を優先とさせていただきます
- ・ 価格は税抜表記となっています。日本居住者のお客様には別途消費税をご請求させていただきます

自己紹介

名前：滝口 開資 (はるよし)

所属：アマゾンウェブサービスジャパン合同会社 コンピュート事業本部
シニアソリューションアーキテクト
EC2 フレキシブルコンピュートスペシャリスト

経歴：銀行様担当メインフレーム SE (外資ベンダー)
→クラウドサポートエンジニア (AWS)
→クラウドサポートチームリード (AWS)
→ソリューションアーキテクト (AWS)



好きなAWSサービス：Amazon EC2 Auto Scaling, AWSサポート

本セミナーの対象者

AWS 環境のインフラを担当されている方

EC2 Auto Scaling でスポットインスタンスを活用したい方

本セミナーの前提知識

- Black Belt Online Seminar Amazon EC2 入門
 - 動画：<https://www.youtube.com/watch?v=1ALvDtb2ziM>
 - 資料：<https://pages.awscloud.com/rs/112-TZM-766/images/202111 AWS Black Belt AWS EC2 introduction.pdf>
- Black Belt Online Seminar Amazon EC2 Auto Scaling 入門編
 - 本セミナーと同時に公開されます

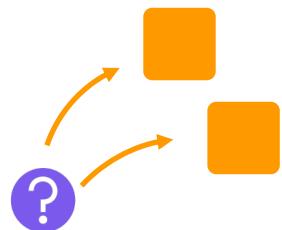
アジェンダ

- 複数購入オプションの指定
- 複数インスタンスタイプの指定
 - 属性ベースのインスタンスタイプ選択
- 配分戦略

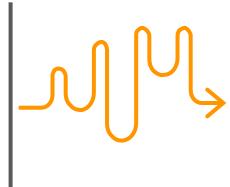
複数購入オプションの指定

Amazon EC2の購入オプション

オンデマンドインスタンス

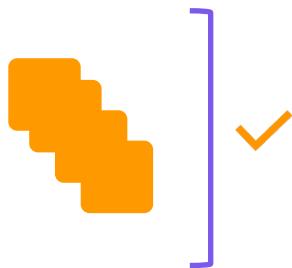


長期コミットなし、
使用分への支払い
(秒単位/時間単位)。
Amazon EC2 の定価

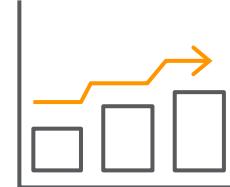


スパイクするような
ワークロードや未知
のワークロード

リザーブドインスタンス / Savings Plans

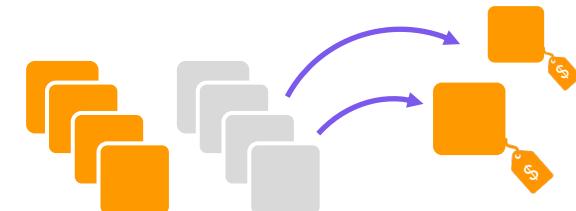


1 年 / 3 年の長期コミットに応じ
た大幅なディスカウント価格。
Savings Plans はリザーブドインス
タンスの後継で、より優れた
柔軟性を提供



一定の負荷の見通し
があり、長期間
コミットできる
ワークロード

スポットインスタンス

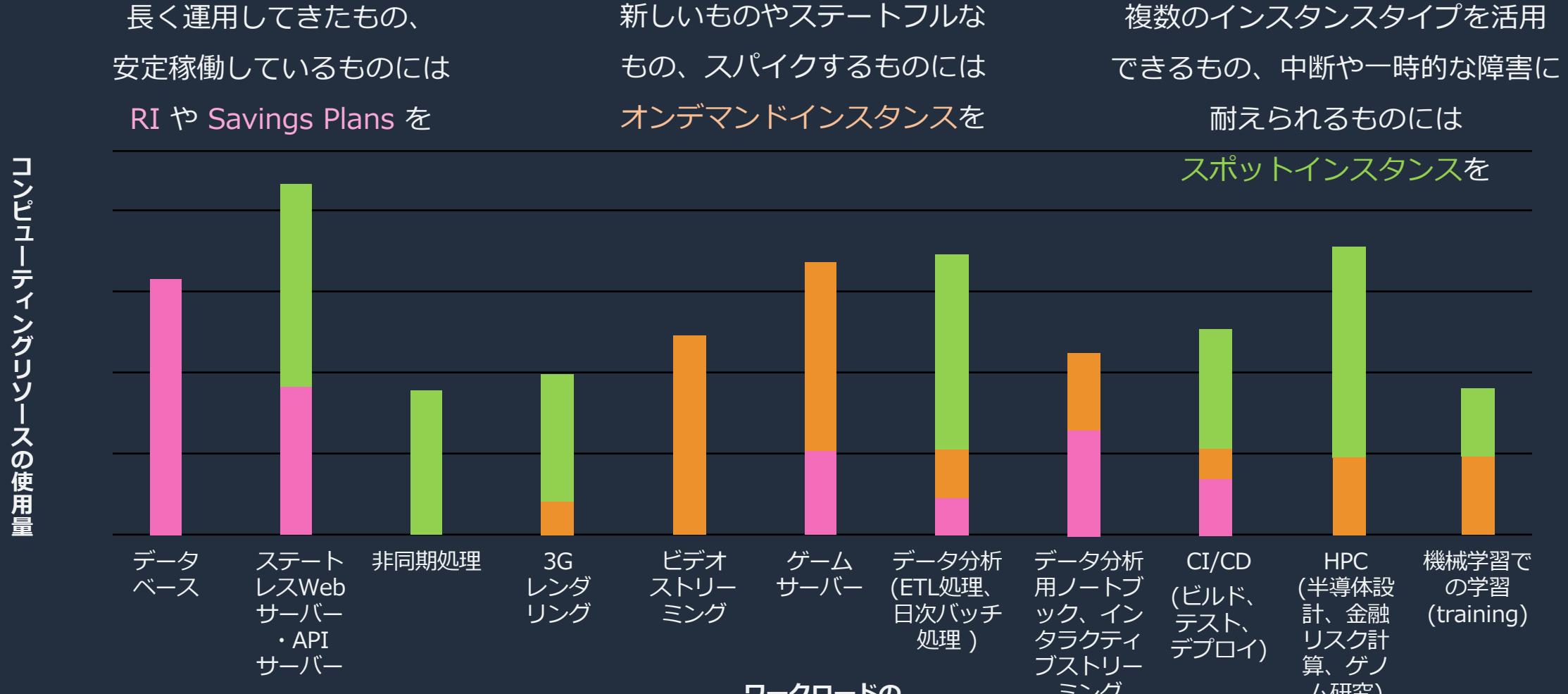


Amazon EC2 の空きキャパシ
ティを活用し、最大 90% の
値引き。中断あり

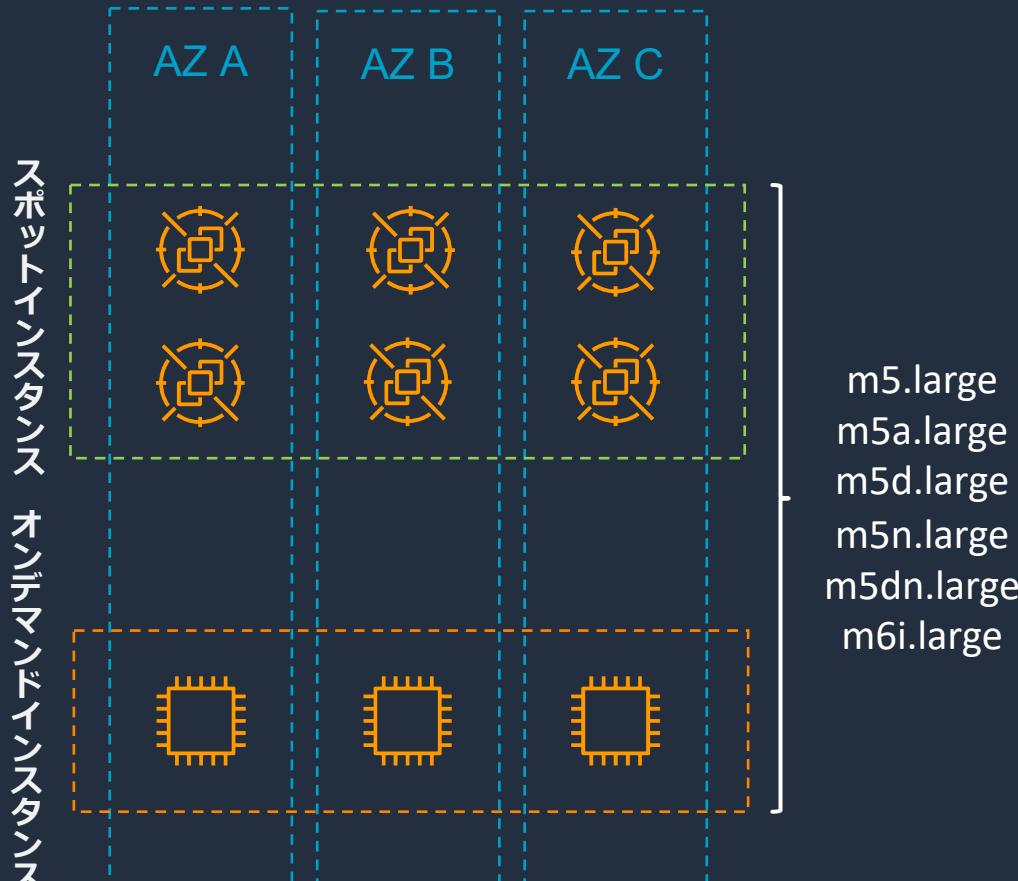


中断に強くステート
レスで、様々な
インスタンスタイプ
を活用できる
ワークロード

ワークロード別 購入オプションの選び方の例

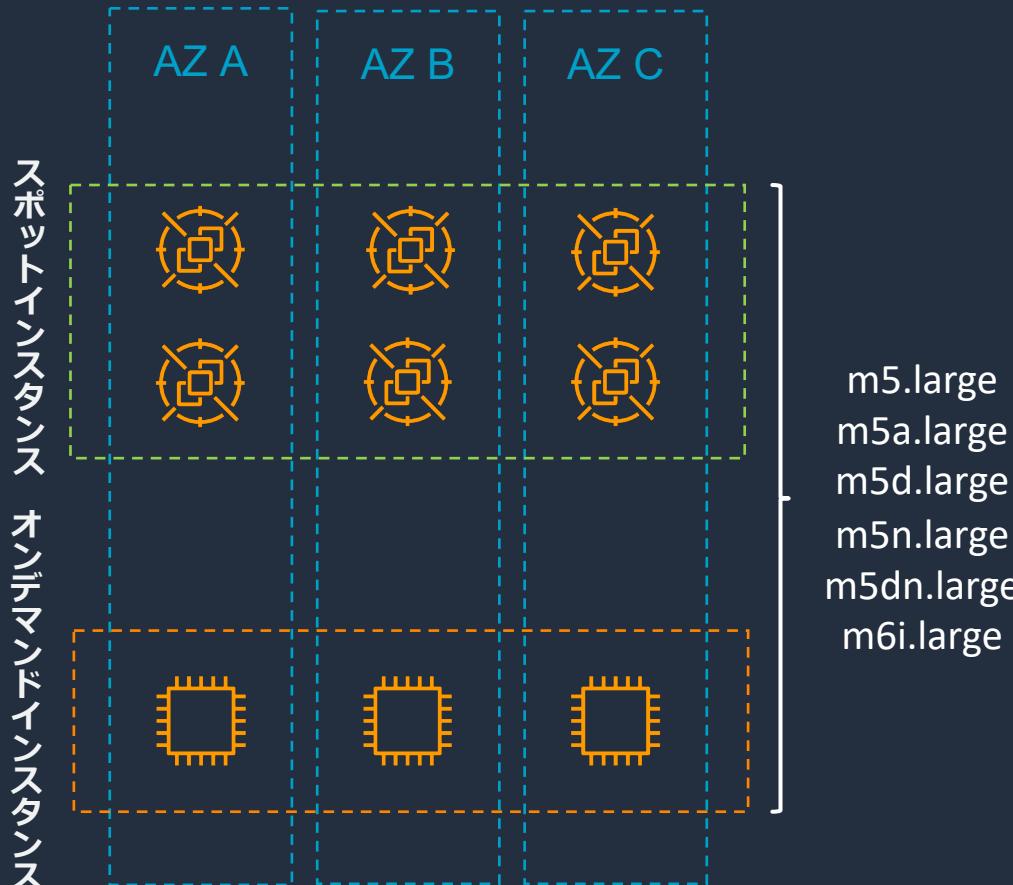


複数のインスタンスタイプと購入オプションを使用する Auto Scaling グループ



- ・ オンデマンドインスタンスとスポットインスタンスをひとつのAuto Scaling グループで管理
 - (オンデマンド:スポット) = (9:1)といった指定ができる
- ・ インスタンスタイプを複数指定できる
 - インスタンスタイプを分散できる
- ・ https://docs.aws.amazon.com/ja_jp/autoscaling/ec2/userguide/ec2-auto-scaling-mixed-instances-groups.html

複数のインスタンスタイプと購入オプションを使用する Auto Scaling グループ



- オンデマンドインスタンスとスポットインスタンスをひとつのAuto Scaling グループで管理
 - (オンデマンド:スポット) = (9:1)といった指定ができる
- インスタンスタイプを複数指定できる
 - インスタンスタイプを分散できる

複数購入オプションの指定



- 「インスタンスの分散」で指定した割合でオンデマンドインスタンスとスポットインスタンスがそれぞれ起動される

複数購入オプションの指定



- 「インスタンスの分散」で指定した割合でオンデマンドインスタンスとスポットインスタンスがそれぞれ起動される
- この例では $2/3$ がオンデマンド、 $1/3$ がスポットになる



複数購入オプションの指定

インスタンスの購入オプション [Info](#)

インスタンスの分散

耐障害性のあるワークロードを低成本で実行するには、スポットインスタンスとなるインスタンスの割合を定義します。スポットインスタンスは、AWS が 2 分前に通知することで変更できるオンデマンド料金に比べて大幅な割引を提供する予備の EC2 容量です。

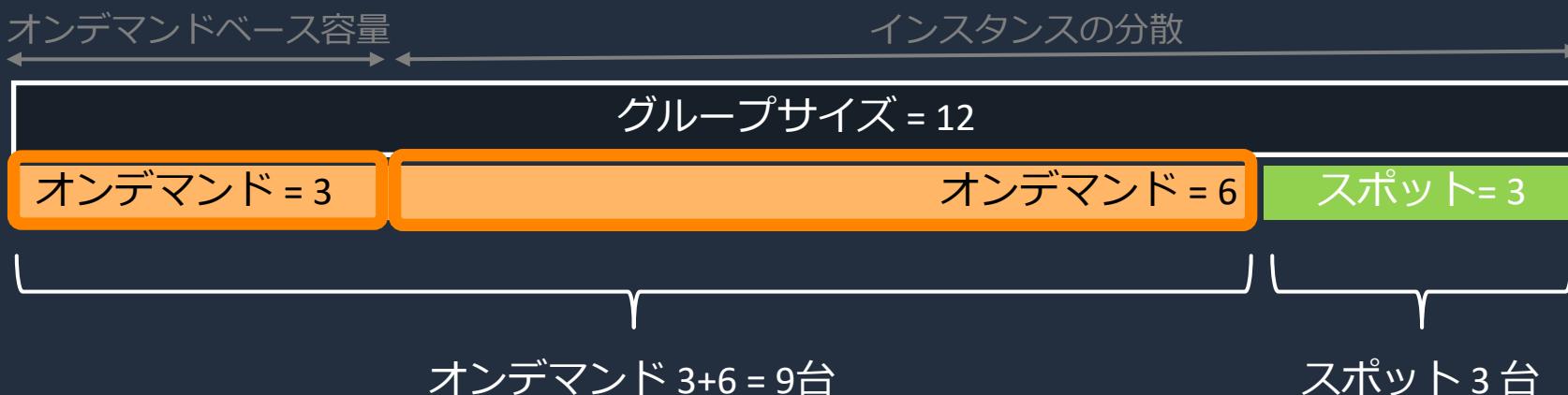
66	%	オンデマンド
34	%	スポット

オンデマンドベース容量を含める

パーセンテージでスケールする前に、Auto Scaling グループがそのベース部分のために使用するオンデマンド容量を指定します。最大グループサイズはこの値まで増加します(減少することはできません)。

3	オンデマンドインスタンス
---	--------------

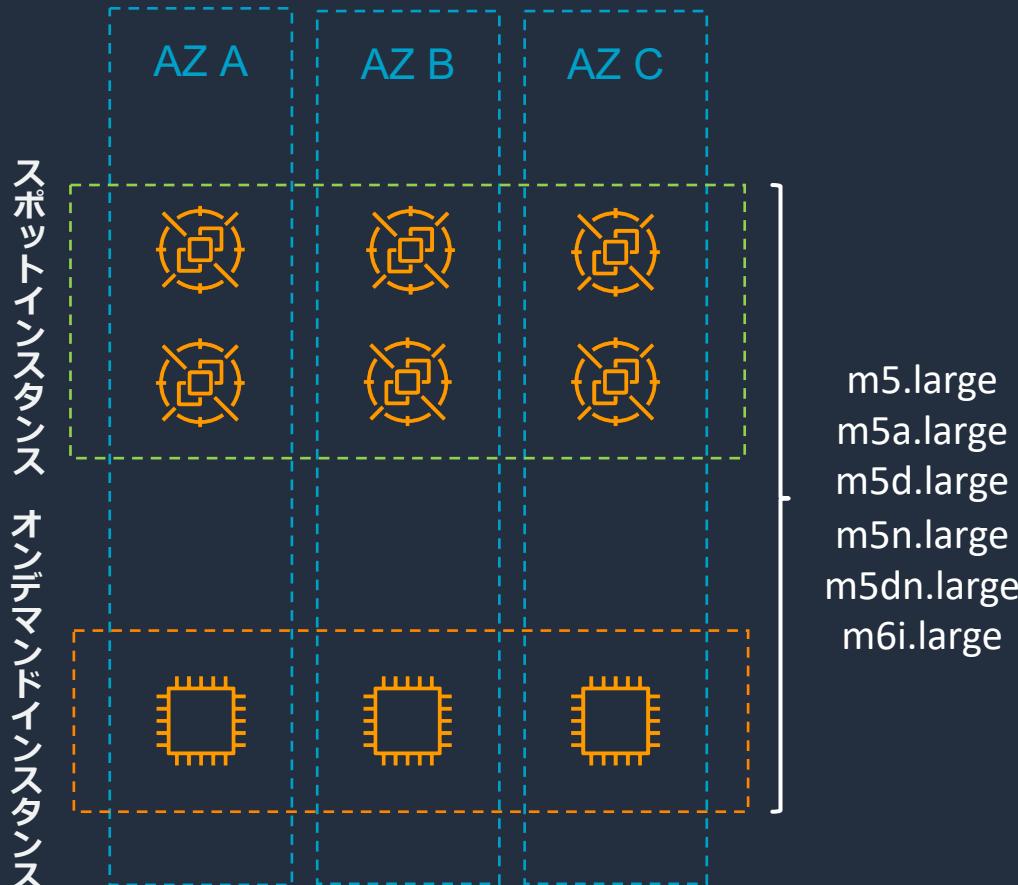
- ・ 「インスタンスの分散」で指定した割合でオンデマンドインスタンスとスポットインスタンスがそれぞれ起動される
 - ・ この例ではまず 3 台分が必ずオンデマンドで起動される
 - ・ 残りの部分の $2/3$ をオンデマンド、 $1/3$ をスポットとして分ける



複数インスタンスタイプの 指定



複数のインスタンスタイプと購入オプションを使用する Auto Scaling グループ



- ・ オンデマンドインスタンスとスポットインスタンスをひとつの Auto Scaling グループで管理
 - (オンデマンド:スポット) = (9:1)といった指定ができる
- ・ インスタンスタイプを複数指定できる
 - インスタンスタイプを分散できる

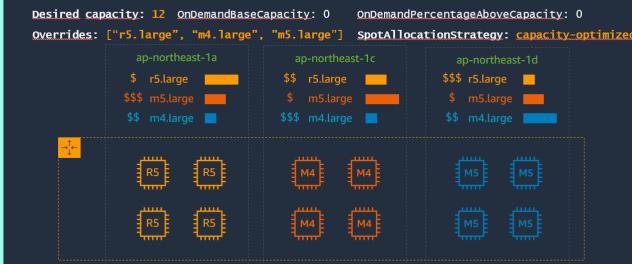
なぜ複数のインスタンスタイプを指定するのか？

- スポットインスタンスにおけるベストプラクティス - インスタンスタイプに関するもの
 - ポイント1:多様なインスタンスタイプを混ぜる
 - ポイント5: capacity-optimized 配分戦略を使う

ポイント1 : 多様なインスタンスタイプを混ぜる



ポイント5: Capacity-optimized 配分戦略を使う



- EC2 スポットを利用するうえでのベストプラクティス - Amazon Elastic Compute Cloud —
https://docs.aws.amazon.com/ja_jp/AWSEC2/latest/UserGuide/spot-best-practices.html

ヒント

詳細な解説を 2023 年 4 月公開の Blackbelt セミナー「Amazon EC2 スポットインスタンス活用のための6つのベストプラクティスと実践例」で紹介します

複数購入オプションの指定

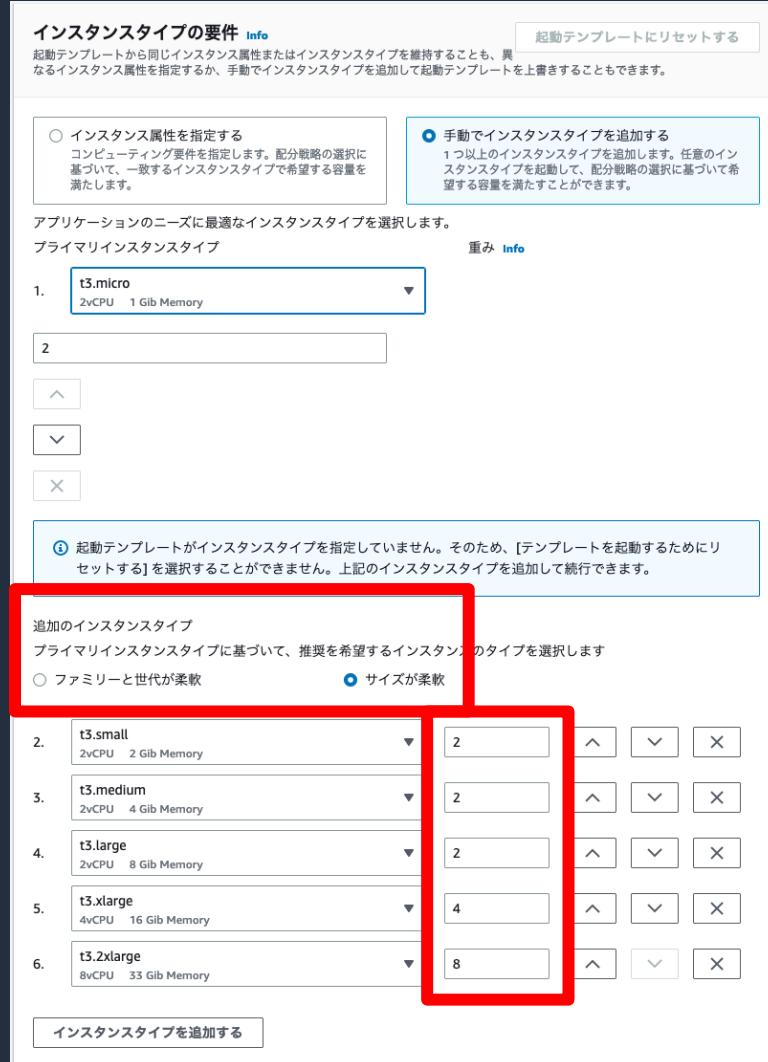


- 「インスタンスの要件」から複数インスタンスタイプを指定できる
- スポットインスタンス活用のベストプラクティスであるインスタンスタイプの分散(複数インスタンスタイプの指定)を容易に実現できる

- まず、あるインスタンスタイプを 1 つ選択する(プライマリインスタンスタイプ)
- すると「レコメンデーション」が実行され、自動的に近い性能のインスタンスタイプ群が推奨される
- 不要なインスタンスタイプは削除できる

https://docs.aws.amazon.com/ja_jp/autoscaling/ec2/userguide/ec2-auto-scaling-mixed-instances-groups.html

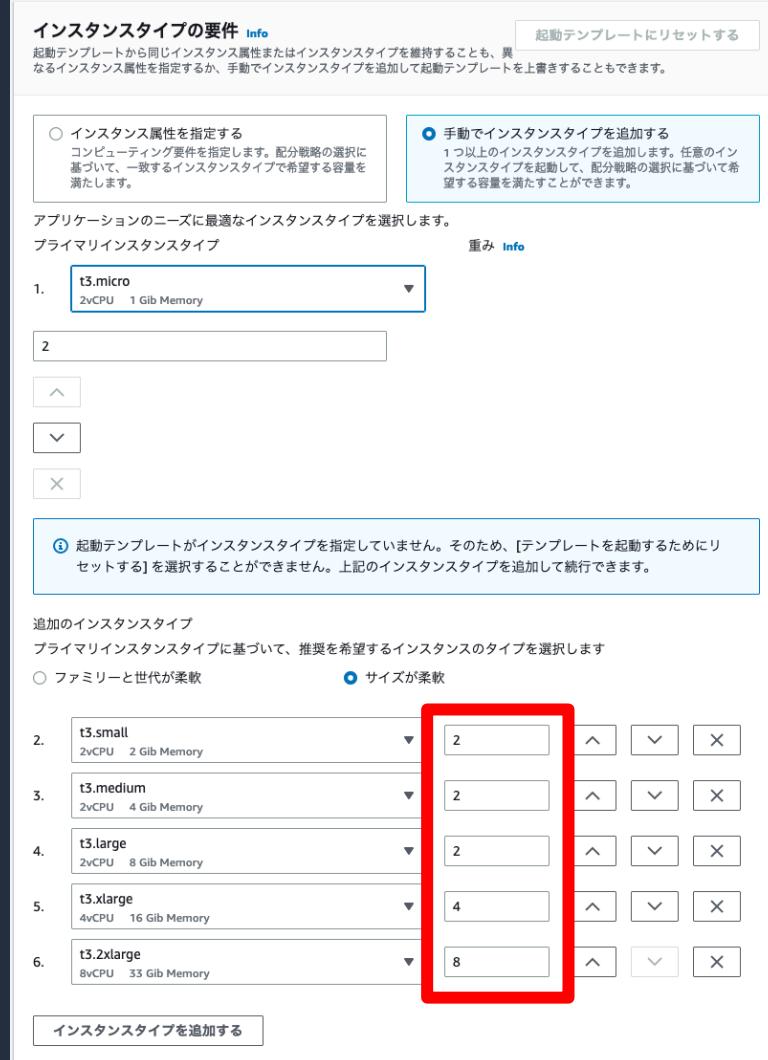
複数購入オプションの指定



- 「インスタンスの要件」から複数インスタンスタイプを指定できる
- スポットインスタンス活用のベストプラクティスであるインスタンスタイプの分散(複数インスタンスタイプの指定)を容易に実現できる
- 「追加のインスタンスタイプ」で「サイズが柔軟」を選択すると同一ファミリー内で異なるサイズのインスタンスタイプ群が推奨される
- このとき、性能比率に応じた重み(Weight)が自動的に設定される

https://docs.aws.amazon.com/ja_jp/autoscaling/ec2/userguide/ec2-auto-scaling-mixed-instances-groups.html

インスタンスタイプに重み付け (Weight) を設定する



- 重みは任意の値を指定できる
- 重みを指定した場合、グループサイズはその重みの単位で指定する必要がある
 - インスタンス台数を指定したのでは意味をなさなくなることに注意
 - 下の図は「2単位」分起動してほしい、という意味になる



https://docs.aws.amazon.com/ja_jp/autoscaling/ec2/userguide/ec2-auto-scaling-mixed-instances-groups-instance-weighting.html

属性ベースの インスタンスタイプの選択



属性ベースのインスタンスタイプの選択

- 「インスタンスの要件」から複数インスタンスタイプを指定できる

インスタンスタイプの要件 Info

起動テンプレートにリセットする

起動テンプレートから同じインスタンス属性またはインスタンスタイプを維持することも、異なるインスタンス属性を指定するか、手動でインスタンスタイプを追加して起動テンプレートを上書きすることもできます。

インスタンス属性を指定する
コンピューティング要件を指定します。配分戦略の選択に基づいて、一致するインスタンスタイプで希望する容量を満たします。

手動でインスタンスタイプを追加する
1つ以上のインスタンスタイプを追加します。任意のインスタンスタイプを起動して、配分戦略の選択に基づいて希望する容量を満たすことができます。

必須インスタンス属性

コンピューティング要件を仮想 CPU (vCPU) とメモリに入力します。

vCPU

インスタンスあたりの vCPU の最小数と最大数を入力します。

0 最小 0 最大

最小値なし 最大値なし

メモリ (GiB)

インスタンスあたりのメモリの最小量および最大量 (GiB) を入力します。

0 最小 0 最大

最小値なし 最大値なし

追加のインスタンス属性 - 省略可能

インスタンス属性を追加して、希望する容量を満たすために使用できるインスタンスタイプをさらに制限します。

属性を選択する 属性を追加する

▶ 一致するインスタンスタイプをプレビューする (0)
このリストには、コンピューティング要件に一致するすべてのインスタンスタイプが含まれます。Amazon EC2 は、これらのインスタンスタイプからプロビジョンできます。希望する容量を満たすために使用される正確なインスタンスタイプは、選択する配分戦略と使用可能な容量によって異なります。

- 「インスタンス属性を指定する」を選択

属性ベースのインスタンスタイプの選択

- vCPU やメモリ数量などの条件に応じてインスタンスタイプを自動選定する
- スポットインスタンスの活用に必須であるインスタンスタイプの多様化を簡単に実現できる

The screenshot illustrates the process of selecting instance types based on attributes. It consists of two main parts connected by a large orange arrow.

Left Side (Configuration): A modal window titled "インスタンスタイプの要件" (Instance Type Requirements) shows two tabs: "インスタンス属性を指定する" (Specify instance attributes) and "手動でインスタンスタイプを追加する" (Add instance type manually). The first tab is selected, showing fields for "vCPU" (16) and "メモリ (GiB)" (32), both with "最小値なし" (No minimum value) checked. Below these are dropdown menus for "CPU メーカー" (CPU Manufacturer) containing "AMD" and "Intel", "世代を選択する" (Select Generation) containing "現行世代" (Current Generation), and "インスタンスファミリーを選択する" (Select Instance Family) containing several options like A, D, F, G, H, I, P, T, U, V, X, Z, and a "除外する" (Exclude) section. At the bottom are "属性を選択する" (Select attributes) and "属性を追加する" (Add attributes).

Right Side (Results): A table titled "一致するインスタンスタイプをレビューする (25)" (Review matching instance types (25)) lists 25 instance types. The columns are "インスタンスタイプ" (Instance Type), "vCPU", and "メモリ (GiB)". The listed instances include various models from the c5, c6, m4, m5, m5ad, m5d, m5dn, m5n, m6, r4, r5, r5ad, r5d, r5dn, r5n, r6, and r6id families, all meeting the specified requirements.

- 16vCPU, 最低32GBのメモリ、最新世代
- 特殊なインスタンスファミリーを除外

- 25件が自動的に選択される
- ASGやフリートの構成情報になる

実行タイミングによって結果が最適に変化

インスタンスタイプの要件 Info

起動テンプレートから同じインスタンス属性またはインスタンスタイプを維持することも、異なるインスタンス属性を指定するか、手動でインスタンスタイプを追加して起動テンプレートを上書きすることもできます。

インスタンス属性を指定する
コンピューティング要件を指定します。配分戦略の選択に基づいて、一致するインスタンスタイプで希望する容量を満たします。

手動でインスタンスタイプを追加する
1つ以上のインスタンスタイプを追加します。任意のインスタンスタイプを起動して、配分戦略の選択に基づいて希望する容量を満たすことができます。

起動テンプレート登録
コンピューティング要件を仮想 CPU (vCPU) とメモリに入力します。

vCPU
インスタンスあたりの vCPU の最小数と最大数を入力します。
16 16

最小値なし 最大値なし

メモリ (GiB)
インスタンスあたりのメモリの最小量および最大量 (GiB) を入力します。
32 最大
 最小値なし 最大値なし

追加のインスタンス属性 - 省略可能
インスタンス属性を追加して、希望する容量を満たすために使用できるインスタンスタイプをさらに制限します。

CPU メーカー
CPU メーカーを選択する AMD Intel

インスタンスの世代
世代を選択する 現行世代

インスタンスマルチ選択
インスタンスマルチを選択する A D F G
H I P T U
V X Z

属性を選択する

- 16vCPU, 最低32GBのメモリ、最新世代
- 特殊なインスタンスマルチを選択

▼ 一致するインスタンスタイプをプレビューする (25)

このリストには、コンピューティング要件に一致するすべてのインスタンスタイプが含まれます。Amazon EC2 は、これらのインスタンスタイプからプロビジョニングできます。希望する容量を満たすために使用される正確なインスタンスタイプは、選択する配分戦略と使用可能な容量によって異なります。

Q インスタンスタイプをフィルタリングする	選択したインスタンスタイプを除外する	
<input type="checkbox"/> インスタンスタイプ	vCPU	
<input type="checkbox"/> c5.4xlarge	16	32
<input type="checkbox"/> c5a.4xlarge	16	32
<input type="checkbox"/> c5d.4xlarge	16	32
<input type="checkbox"/> c5n.4xlarge	16	42
<input type="checkbox"/> c6a.4xlarge	16	32
<input type="checkbox"/> c6i.4xlarge	16	32
<input type="checkbox"/> c6id.4xlarge	16	32
<input type="checkbox"/> m4.4xlarge	16	64
<input type="checkbox"/> m5.4xlarge	16	64
<input type="checkbox"/> m5a.4xlarge	16	64
<input type="checkbox"/> m5ad.4xlarge	16	64
<input type="checkbox"/> m5d.4xlarge	16	64
<input type="checkbox"/> m5dn.4xlarge	16	64
<input type="checkbox"/> m5n.4xlarge	16	64
<input type="checkbox"/> m6a.4xlarge	16	64
<input type="checkbox"/> m6i.4xlarge	16	64
<input type="checkbox"/> m6id.4xlarge	16	64
<input type="checkbox"/> m6dn.4xlarge	16	64
<input type="checkbox"/> m6n.4xlarge	16	64
<input type="checkbox"/> r4.4xlarge	16	122
<input type="checkbox"/> r5.4xlarge	16	128
<input type="checkbox"/> r5a.4xlarge	16	128
<input type="checkbox"/> r5ad.4xlarge	16	128
<input type="checkbox"/> r5b.4xlarge	16	128
<input type="checkbox"/> r5d.4xlarge	16	128
<input type="checkbox"/> r5dn.4xlarge	16	128
<input type="checkbox"/> r5n.4xlarge	16	128
<input type="checkbox"/> r6i.4xlarge	16	128
<input type="checkbox"/> r6id.4xlarge	16	128

- 25件 (2022年12月時点)

© 2023, Amazon Web Services, Inc. or its affiliates.

▼ 一致するインスタンスタイプをプレビューする (32)

このリストには、コンピューティング要件に一致するすべてのインスタンスタイプが含まれます。Amazon EC2 は、これらのインスタンスタイプからプロビジョニングできます。希望する容量を満たすために使用される正確なインスタンスタイプは、選択する配分戦略と使用可能な容量によって異なります。

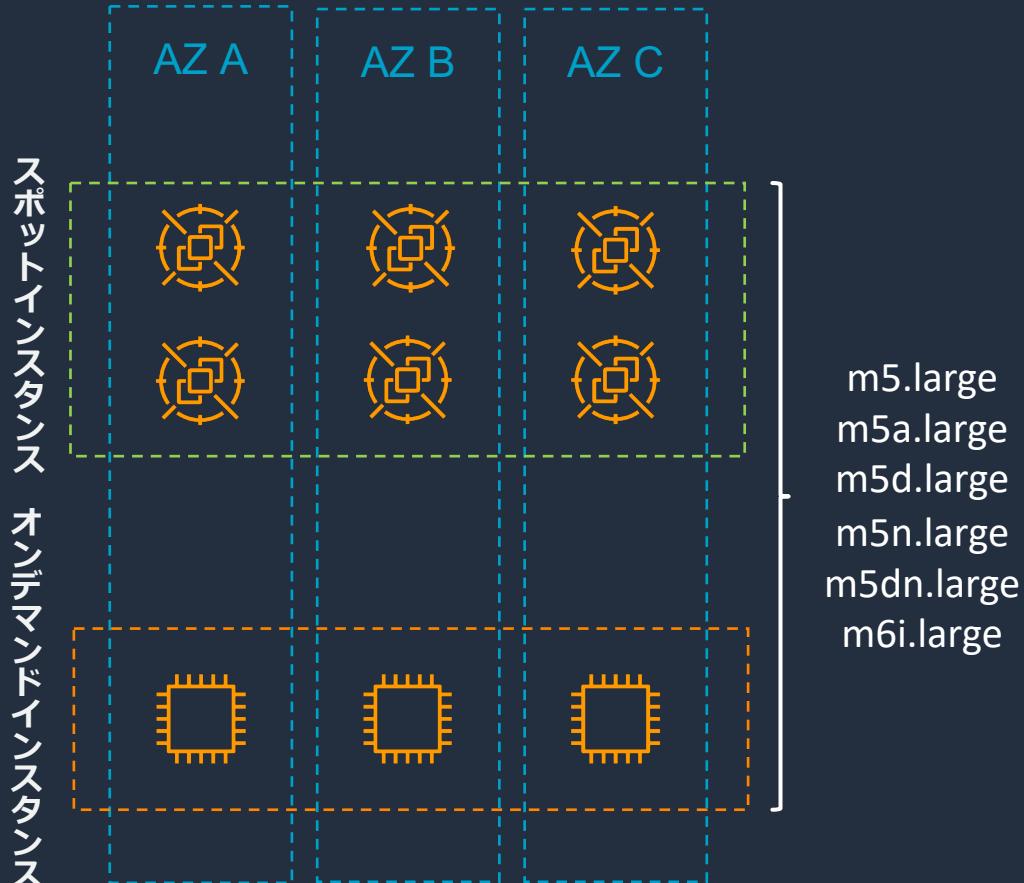
Q インスタンスタイプをフィルタリングする	選択したインスタンスタイプを除外する	
<input type="checkbox"/> インスタンスタイプ	vCPU	
<input type="checkbox"/> c5.4xlarge	16	32
<input type="checkbox"/> c5a.4xlarge	16	32
<input type="checkbox"/> c5d.4xlarge	16	32
<input type="checkbox"/> c5n.4xlarge	16	42
<input type="checkbox"/> c6a.4xlarge	16	32
<input type="checkbox"/> c6i.4xlarge	16	32
<input type="checkbox"/> c6id.4xlarge	16	32
<input type="checkbox"/> m4.4xlarge	16	64
<input type="checkbox"/> m5.4xlarge	16	64
<input type="checkbox"/> m5a.4xlarge	16	64
<input type="checkbox"/> m5ad.4xlarge	16	64
<input type="checkbox"/> m5d.4xlarge	16	64
<input type="checkbox"/> m5dn.4xlarge	16	64
<input type="checkbox"/> m5n.4xlarge	16	64
<input type="checkbox"/> m6a.4xlarge	16	64
<input type="checkbox"/> m6i.4xlarge	16	64
<input type="checkbox"/> m6id.4xlarge	16	64
<input type="checkbox"/> m6dn.4xlarge	16	64
<input type="checkbox"/> m6n.4xlarge	16	64
<input type="checkbox"/> r4.4xlarge	16	122
<input type="checkbox"/> r5.4xlarge	16	128
<input type="checkbox"/> r5a.4xlarge	16	128
<input type="checkbox"/> r5ad.4xlarge	16	128
<input type="checkbox"/> r5b.4xlarge	16	128
<input type="checkbox"/> r5d.4xlarge	16	128
<input type="checkbox"/> r5dn.4xlarge	16	128
<input type="checkbox"/> r5n.4xlarge	16	128
<input type="checkbox"/> r6i.4xlarge	16	128
<input type="checkbox"/> r6id.4xlarge	16	128
<input type="checkbox"/> r6dn.4xlarge	16	128
<input type="checkbox"/> r6n.4xlarge	16	128

- 32件 (2023年3月時点)

配分戦略



配分戦略 - どのインスタンスタイプが起動されるのか？



- 複数指定されたインスタンスタイプの中からどれを起動するかを決める設定
- オンデマンド用インスタンス用とスポットインスタンス用にそれぞれ配分戦略を指定する
- https://docs.aws.amazon.com/ja_jp/autoscaling/ec2/userguide/ec2-auto-scaling-mixed-instances-groups.html

オンデマンドインスタンスの配分戦略



- **lowest-price**
 - オンデマンド価格が最安値のものを優先して選ぶ
 - 属性ベースのインスタンスタイプの選択機能を使う場合は必須
- **prioritized**
 - インスタンスタイプリストの上から順に選ぶ
 - リザーブドインスタンスや Savings Plans で購入済みのインスタンスマリーファミリーがある場合はこちらを選択

m5.large
m5a.large
m5d.large
m5n.large
m5dn.large
m6i.large

←リザーブドインスタンス購入済みのものを最上位に記載する

スポットインスタンスの配分戦略

配分戦略 [Info](#)

オンデマンド配分戦略
オンデマンドインスタンスの起動時に適用する配分戦略を選択します。

高い優先順位で設定済み
上記で設定したインスタンスタイプの優先順位に基づいて、オンデマンドインスタンスをリクエストします。この戦略は、属性ベースのインスタンスタイプの選択では使用できません。

最低料金
アベイラビリティゾーン内で最低料金のプールからオンデマンドインスタンスをリクエストします。

スポット割り当て戦略
スポットインスタンスの起動時に適用する配分戦略を選択します。

④ 新規! [料金キャパシティ最適化] は、最低料金と利用可能なキャパシティの両方のために最適化するスポットプールを識別する新しい配分戦略です。 詳細はこちら [\[\]](#)

料金キャパシティ最適化 (推奨)
アベイラビリティゾーン内で最も利用可能なプールから最低料金のスポットインスタンスをリクエストします。これは、インスタンス料金と中断リスクのバランスを取るための最良の戦略です。

キャパシティ最適化
アベイラビリティゾーン内で最も利用可能なプールからスポットインスタンスをリクエストします。この戦略は、中断のリスクが最も低くなります。

最低料金
インスタンスタイプの要件に基づいて、アベイラビリティゾーン内の最低料金のプールからスポットインスタンスをリクエストします。この戦略は、中断のリスクが最も高くなります。

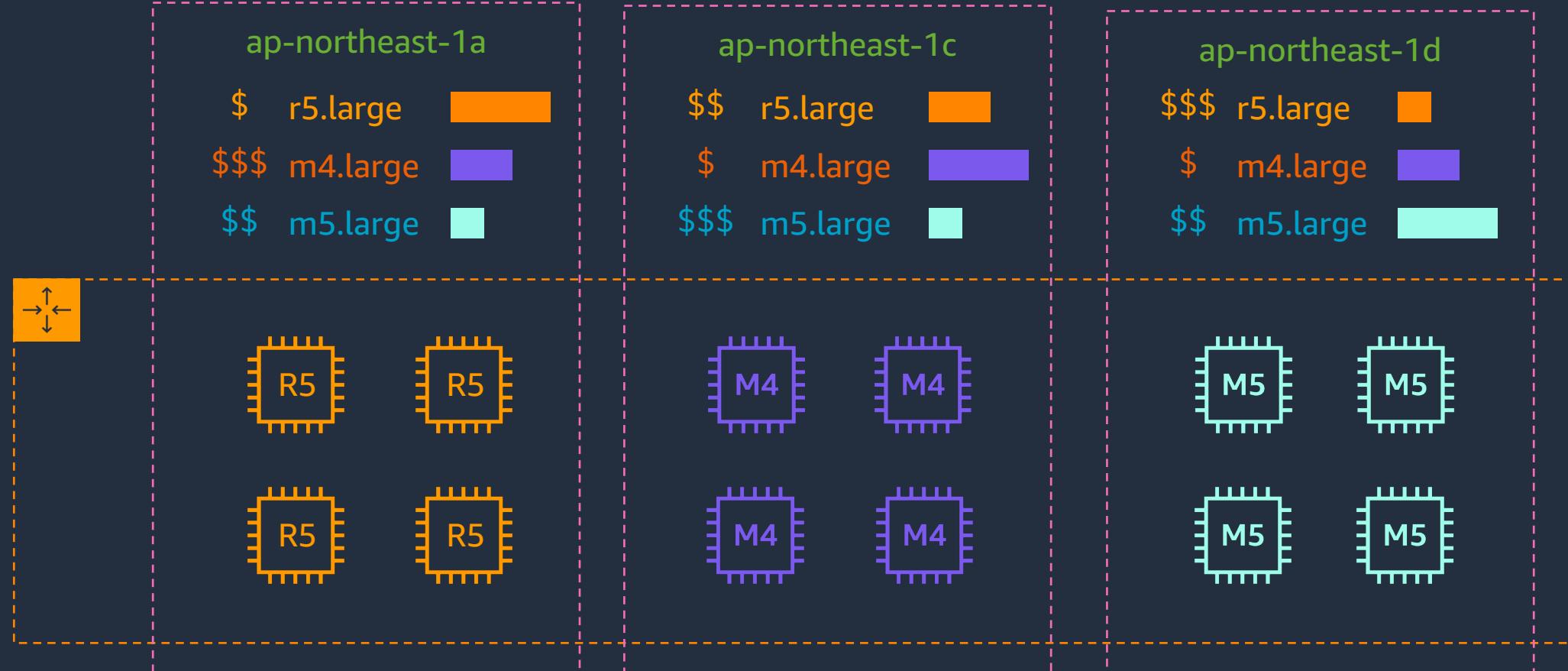
容量の再調整 [Info](#)
容量の再調整を有効にし、再調整通知がインスタンスに送信されると、EC2 Auto Scaling は中断される前に自動的にインスタンスの置き換えを試みます。

- **lowest-price**
 - スポット価格が最安値のものを優先して選ぶ
- **capacity-optimized (バリエーションあり)**
 - EC2サービスに最もキャパシティがあるものを選ぶ

capacity-optimized 配分戦略の動作

Desired capacity: 12 OnDemandBaseCapacity: 0 OnDemandPercentageAboveCapacity: 0

Overrides: ["r5.large", "m4.large", "m5.large"] SpotAllocationStrategy: capacity-optimized



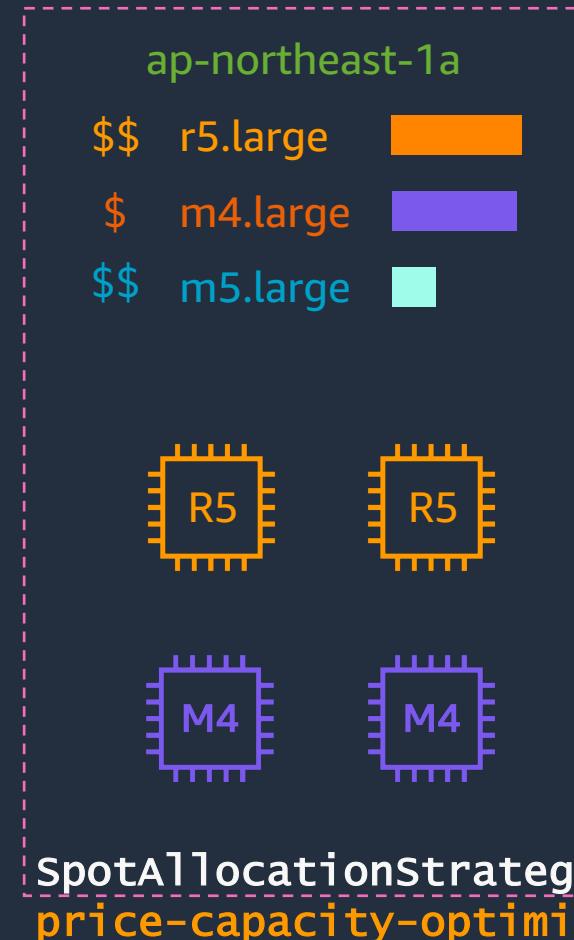
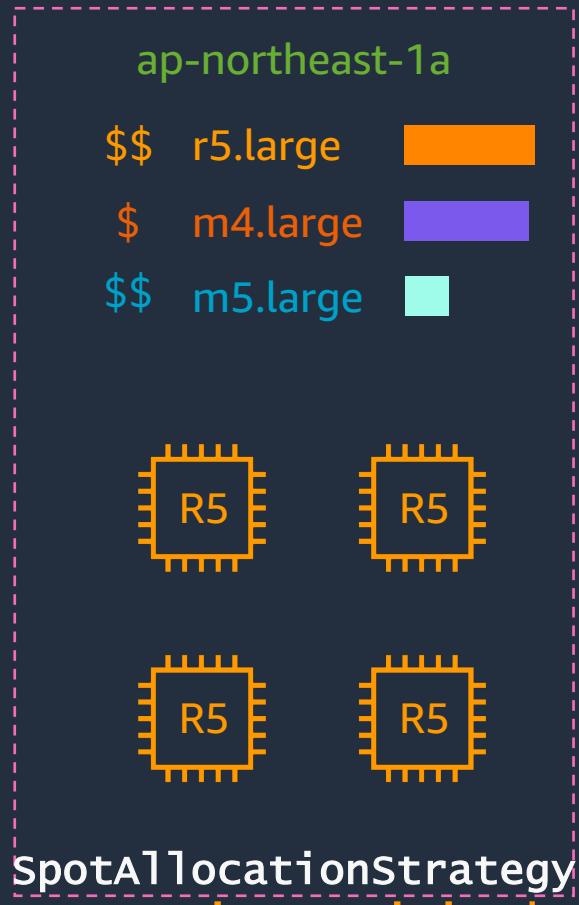
capacity-optimized 配分戦略の3つのバリエーション

- capacity-optimized
 - 起動するスポットインスタンスを起動のしやすさのみで決定
 - 最も中断しにくいスポットプールから起動
- capacity-optimized-prioritized
 - 起動するスポットインスタンスを起動のしやすさのみで決定
 - 複数のスポットプール間で起動のしやすさが同等である場合、リスト上位のものを優先
- price-capacity-optimized (新規、AWS推奨)
 - 起動するスポットインスタンスを起動のしやすさとスポット価格の組み合わせで決定
 - なるべく価格が低く、かつ起動しやすいスポットプールから起動
 - 2022年11月に発表。費用と安定性のバランスを取れるためこのオプションを推奨

price-capacity-optimized 配分戦略の動作

Desired capacity: 4 OnDemandBaseCapacity: 0 OnDemandPercentageAboveCapacity: 0

Overrides: ["r5.large", "m4.large", "m5.large"]



price-capacity-optimized 配分戦略の関連リンク集

- Amazon EC2 が Amazon EC2 スポットインスタンスをプロビジョニングするための、新しい配分戦略「料金キャパシティ最適化」を発表 –
<https://aws.amazon.com/jp/about-aws/whats-new/2022/11/amazon-ec2-price-capacity-optimized-allocation-strategy-provisioning-ec2-spot-instances/>
- EC2 スポットインスタンスの price-capacity-optimized 戰略のご紹介 | Amazon Web Services ブログ –
<https://aws.amazon.com/jp/blogs/news/introducing-price-capacity-optimized-allocation-strategy-for-ec2-spot-instances/>
- 複数のインスタンスタイプと購入オプションを使用する Auto Scaling グループ - Amazon EC2 Auto Scaling –
https://docs.aws.amazon.com/ja_jp/autoscaling/ec2/userguide/ec2-auto-scaling-mixed-instances-groups.html

おわりに



今回お話しした内容

- ・ 複数購入オプションの指定
- ・ 複数インスタンスタイプの指定
 - ・ 属性ベースのインスタンスタイプ選択
- ・ 配分戦略
 - ・ スポットインスタンスには price-capacity-optimized を！

ヒント

詳細な解説を 2023 年 4 月公開の Blackbelt セミナー「Amazon EC2
スポットインスタンス活用のための6つのベストプラクティスと実践例」で紹介します

本資料に関するお問い合わせ・ご感想

技術的な内容に関しましては、有料のAWSサポート窓口へ
お問い合わせください

<https://aws.amazon.com/jp/premiumsupport/>

料金面でのお問い合わせに関しましては、カスタマーサポート窓口へ
お問い合わせください（マネジメントコンソールへのログインが必要です）

<https://console.aws.amazon.com/support/home#/case/create?issueType=customer-service>

具体的な案件に対する構成相談は、後述する個別相談会をご活用ください



ご感想はTwitterへ！ハッシュタグは以下をご利用ください
#awsblackbelt



その他コンテンツのご紹介

ウェビナーなど、AWSのイベントスケジュールをご参照いただけます

<https://aws.amazon.com/jp/events/>

ハンズオンコンテンツ

<https://aws.amazon.com/jp/aws-jp-introduction/aws-jp-webinar-hands-on/>

AWS 個別相談会

AWSのソリューションアーキテクトと直接会話いただけます

<https://pages.awscloud.com/JAPAN-event-SP-Weekly-Sales-Consulting-Seminar-2021-reg-event.html>



Thank you!



Amazon EC2 Auto Scaling

スケーリングポリシーと おすすめ機能編

滝口 開資 (はるよし)

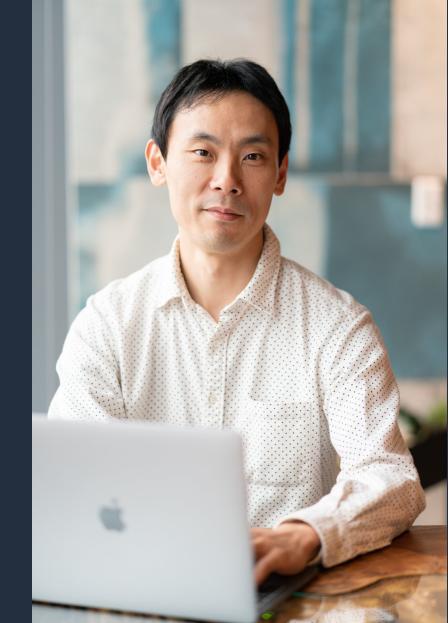
シニアソリューションアーキテクト
EC2 フレキシブルコンピュートスペシャリスト
2023/10

自己紹介

名前：滝口 開資（はるよし）

所属：アマゾンウェブサービスジャパン合同会社
コンピュート事業本部
シニアソリューションアーキテクト
EC2 フレキシブルコンピュートスペシャリスト

経歴：銀行様担当メインフレーム SE（外資ベンダー）
→クラウドサポートエンジニア（AWS）
→クラウドサポートチームリード（AWS）
→ソリューションアーキテクト（AWS）



好きなAWSサービス：Amazon EC2 Auto Scaling, AWSサポート

本セミナーの対象者

AWS 基盤環境のインフラを担当されている方

EC2 インスタンスを自動スケールさせる際に必要となる基礎知識を知りたい方

本セミナーの前提知識

- Black Belt Online Seminar Amazon EC2 入門
- Black Belt Online Seminar Amazon EC2 Auto Scaling 入門編
- Black Belt Online Seminar Amazon EC2 Auto Scaling 複数のインスタンスタイプと購入オプションの活用編

ヒント

AWS Black Belt コンピュートシリーズのあるきかた | Amazon Web Services ブログ –
<https://aws.amazon.com/jp/blogs/news/aws-black-belt-compute-series/>

アジェンダ

- Auto Scaling サービス群の整理
- EC2 Auto Scaling おすすめ機能
 - ワンタッチでできる自動スケール設定
 - ライフサイクルフック
 - インスタンスリフレッシュ
 - ウォームプール
- インスタンスの置き換えに関するよくある質問集

Auto Scaling サービス群 の整理



3 つの Auto Scaling サービス群

- Amazon EC2 Auto Scaling
 - EC2 インスタンスの自動スケール機能を提供
- Application Auto Scaling
 - EC2 インスタンス以外のリソースにも自動スケーリングを提供
 - ECS サービスタスク、スポットフリート、AppStream 2.0 フリート、DynamoDB テーブル、Aurora レプリカ、ElastiCache for Redis レプリケーショングループ、SageMaker エンドポイントバリアント、Lambda 関数プロビジョン済み同時実行数、カスタムリソースなど
- AWS Auto Scaling
 - 2種類の Auto Scaling のスケーリングを設定・管理するためのワンストップサービス

Auto Scaling サービス群

- Amazon EC2 Auto Scaling
 - EC2 インスタンスの自動スケール機能を提供
- Application Auto Scaling
 - EC2 インスタンス以外のリソースにも自動スケーリングを提供
 - ECS クラスター、スポットフリート、EMR クラスター、AppStream 2.0 フリート、DynamoDB テーブル、Aurora レプリカ、SageMaker エンドポイントバリアント、Lambda 関数プロビジョン済み同時実行数、カスタムリソースなど
- AWS Auto Scaling (新規投資予定なし)
 - 2種類の Auto Scaling のスケーリングを設定・管理するためのワンストップサービス
 - 新規適用は非推奨。各サービスの Application Auto Scaling 機能を活用してください
 - バグフィックスは引き続き提供されます

EC2 Auto Scaling の おすすめ機能

EC2 Auto Scaling のおすすめ機能

- ワンタッチでできる自動スケール設定
- ライフサイクルフック
- インスタンスリフレッシュ
- ウォームプール

EC2 Auto Scaling のおすすめ機能

- ワンタッチでできる自動スケール設定
- ライフサイクルフック
- インスタンスリフレッシュ
- ウォームプール

ワンタッチでできる自動スケール設定

- ・ターゲット追跡スケーリング + 予測スケーリングの組み合わせ
- ・スケジュールスケーリングも組み合わせられます

ターゲット追跡スケーリング

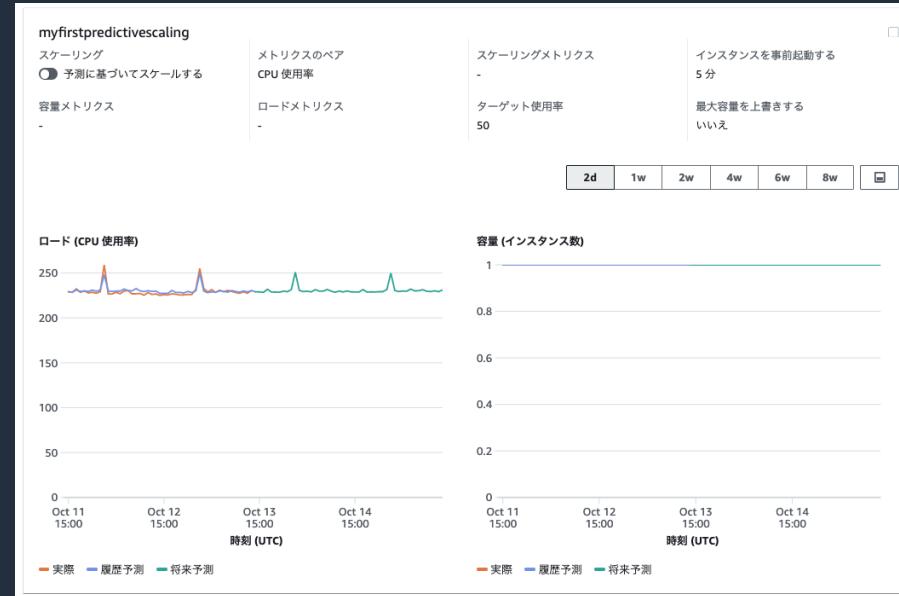
- 1つのメトリクスに対し、単に目標値を指定するのみで良い
 - CPUUtilizationを50%に維持して欲しい、ただこれだけ



https://docs.aws.amazon.com/ja_jp/autoscaling/ec2/userguide/as-scaling-target-tracking.html

予測スケーリング

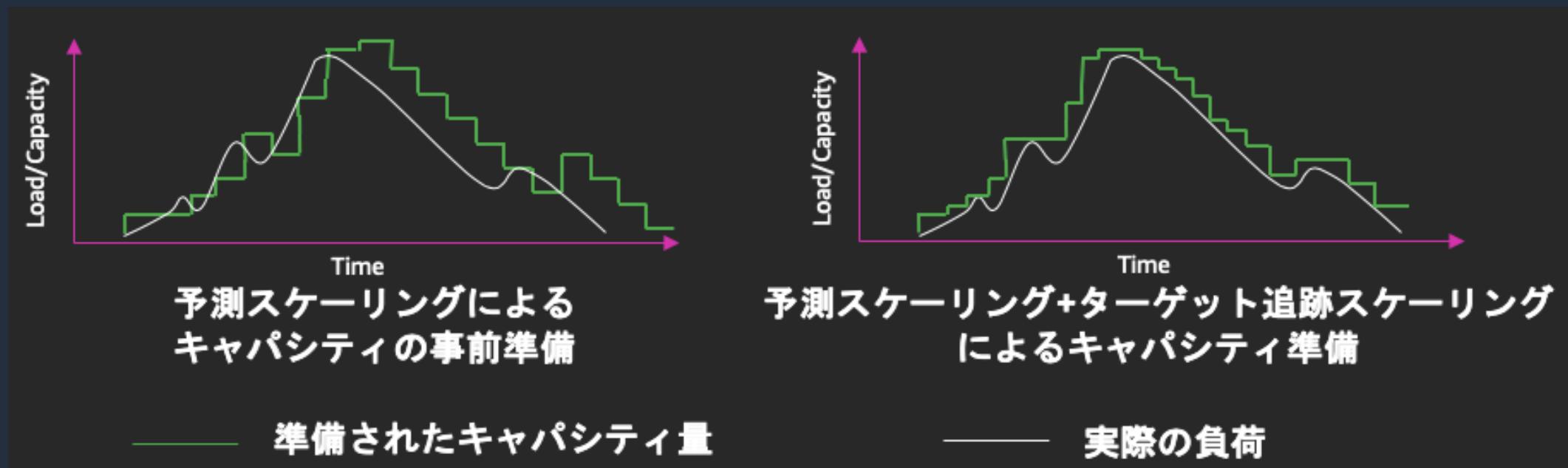
- ・2週間分のメトリクスを分析し、次の2日の今後の需要を予測
 - 最短で24時間分のメトリクスデータから始められる
- ・予測データに基づいてキャパシティの増減がスケジュールされる



https://docs.aws.amazon.com/ja_jp/autoscaling/ec2/userguide/ec2-auto-scaling-predictive-scaling.html

ターゲットトラッキング + 予測スケーリングの組み合わせ

1. 大まかなキャパシティ増減は予測スケーリングに任せ、前もってスケールしておく
2. 実際の負荷に対して不足した分をターゲット追跡で補充する
3. さらにスケジュールスケーリングを組み合わせることもできる



参考：スケーリングポリシーの整理

- ・動的なスケーリング
 - 簡易スケーリング
 - ステップスケーリング
 - ターゲット追跡スケーリング
- ・予測スケーリング
- ・スケジュールスケーリング

Amazon EC2 Auto Scaling

ユーザーガイド

- ▶ Amazon EC2 Auto Scaling とは
- セットアップする
- 開始方法
- ▶ 起動テンプレート
- ▶ 起動設定
- ▶ Auto Scaling グループ
- ▼ グループをスケールする
 - キャパシティーの制限を設定する
 - 固定数のインスタンスを維持する
- ▶ 手動スケーリング

▼ 動的なスケーリング

- ▶ ターゲット追跡スケーリングポリシー
- ステップスケーリングポリシーおよび簡易スケーリングポリシー
- ▶ デフォルトのウォームアップ値またはクールダウン値を設定する
- Amazon SQS に基づくスケーリング
- スケーリングアクティビティを検証する
- スケーリングポリシーを無効化する
- スケーリングポリシーを削除する
- AWS CLI スケーリングポリシーの例

▶ 予測スケーリング

- スケジュールされたスケーリング



参考：スケーリングポリシーの整理

- ・動的なスケーリング
 - 簡易スケーリング
 - ステップスケーリング
 - ターゲット追跡スケーリング
 - ・予測スケーリング
 - ・スケジュールスケーリング
- ・簡易スケーリングポリシーは互換性維持のために残されている。新規で作成する必要はない
 - ・ユースケースによって、きめ細やかなスケール条件を指定できるステップスケーリングポリシーを採用する場合がある。ただし無理に使う必要はない
 - ・2023年のおすすめはターゲット追跡スケーリング + 予測スケーリング + スケジュールスケーリングの組み合わせ。最小の手間で最大の効果を

EC2 Auto Scaling のおすすめ機能

- ワンタッチでできる自動スケール設定
- ライフサイクルフック
- インスタンスリフレッシュ
- ウォームプール

ライフサイクルフック

- ・インスタンスの起動時や終了時に何かしたい、を実現する仕組み
- ・起動時ライフサイクルフックが有効な場面
 - 例：ELBに登録される前にインスタンス上の様々な準備が正しく完了していることを確認したい
- ・終了時ライフサイクルフックが有効な場面
 - 例：スケールインが発生するとき、アプリケーションを安全に終了させてからのインスタンス削除を保証したい

https://docs.aws.amazon.com/ja_jp/autoscaling/ec2/userguide/lifecycle-hooks.html



インスタンスリフレッシュ

- Auto Scaling グループ内のインスタンスを自動的に更新してくれる仕組み
 - AMI更新時などの場面で、手動で入れ替える必要がなくなった
- 入れ替えは一定割合のインスタンスが稼働中(Healthy)であることを保ちながら実施される
 - デフォルトは90%

https://docs.aws.amazon.com/ja_jp/autoscaling/ec2/userguide/asg-instance-refresh.html



ウォームプール

- 起動に長い時間かかるインスタンスを事前起動できる仕組み
 - 事前に起動されたインスタンスが「ウォームプール」にStopped状態で保持
 - 事前起動することで時間を稼ぐ
 - 発生する費用はEBSボリュームとEIPのみ
 - スケールアウトが発生するとウォームプールから開始(Start)される
 - ゼロから起動(Launch)するよりも格段に速い
- 制約
 - スポットインスタンスを含むASG, 複数インスタンスタイプを指定したASGにはウォームプールを追加できない

スケール速度に問題がないケースでは
無理にウォームプールを使う必要はない

https://docs.aws.amazon.com/ja_jp/autoscaling/ec2/userguide/ec2-auto-scaling-warm-pools.html



インスタンスの置き換えに 関するよくある質問集

こんなときどうする？

- 正常に動作しないインスタンスを自動的に置き換える
• →ヘルスチェックを活用
- 特に指定しない場合、EC2 ヘルスチェックが有効になっている
 - 2/2 以外のステータスが続くとAuto Scaling サービスが置き換える
- ELB 配下の Auto Scaling グループの場合、ELB ヘルスチェックを有効にする
 - EC2 ヘルスチェックに加え、ELB からのヘルスチェックに応答しない場合の速やかな入れ替えが可能になる

https://docs.aws.amazon.com/ja_jp/autoscaling/ec2/userguide/ec2-auto-scaling-health-checks.html



こんなときどうする？

- スケールイン・スケールアウトを繰り返してしまい、いつまでたってもインスタンスが追加されない
 - → 「ヘルスチェックの猶予期間」の設定を見直す
- ヘルスチェックの猶予期間：起動したばかりでヘルスチェックに応答できないインスタンスを保護する期間
 - デフォルトは 5 分 (300 秒)
 - もしこれよりインスタンスの準備に時間がかかるとすると、ヘルスチェックがそのインスタンスを置き換えてしまう。そのままでは希望容量を満たさないので再びスケールアウトが試行され、インスタンスの起動と削除が繰り返される
 - ELB ヘルスチェックに、ユーザーデータなどで指示した S3 からのコンテンツ配備や DB 接続などを前提としたアプリケーションのパスを指定している場合に有効なことがある

https://docs.aws.amazon.com/ja_jp/autoscaling/ec2/userguide/health-check-grace-period.html



こんなときどうする？

- 特定のインスタンスをスケールインから保護したい
 - →インスタンスの保護
- ASG単位、もしくはインスタンス単位で設定。スケールインされなくなる
- 次の条件からは保護できないことに注意
 - 手動でのインスタンス削除(Terminate)
 - ヘルスチェックによる置き換え
 - スポットインスタンスの中斷
- すべてのインスタンスが終了保護された状態でスケールインイベントが発生した場合、希望容量だけが減少し、スケールイン(インスタンス削除)は行われない

https://docs.aws.amazon.com/ja_jp/autoscaling/ec2/userguide/ec2-auto-scaling-instance-protection.html



こんなときどうする？

- 一時的にスケールインやスケールアウトを止めたい
 - →スケーリングプロセスの中斷
- 一時的にスケール動作を停止できる
- ASG単位で設定
- 中断できるプロセス一覧：Launch, Terminate, AddToLoadBalancer, AlarmNotification, AZRebalance, HealthCheck, ReplaceUnhealthy, ScheduledActions
- 使いどころ：機能テストなど、一時的にAuto Scalingグループの特定プロセスの動作を止めてテスト条件を整えたい場合
 - LaunchとTerminateの両方のプロセスを中断することで、「何もしない」Auto Scaling グループを作り出せる
- 動作のおかしいインスタンスがいるのでスケールイン・スケールアウトを止めたい
 - →プロセスの中斷ではなく次の項目を参照

https://docs.aws.amazon.com/ja_jp/autoscaling/ec2/userguide/as-suspend-resume-processes.html



こんなときどうする？

- 特定のインスタンスを Auto Scaling グループから外したい
 - →スタンバイ、もしくはデタッチ
- スタンバイ(「一時的なインスタンスの削除」)
 - インスタンス単位で設定
 - そのインスタンスは Auto Scaling グループにいながら「スタンバイ」状態に入る
 - 具体的にはそのインスタンスはELBから登録解除され、ヘルスチェック対象から外され、その Auto Scaling グループの希望容量は1つ減少する
 - その間にインスタンスのトラブルシューティングなどを行う
- デタッチ
 - インスタンス単位で設定
 - そのインスタンスはその Auto Scaling グループのメンバーから外れる
 - スタンバイと実質的な効果は同一。インスタンスはそのまま Running 状態で保持される。ただしデタッチの場合、Auto Scaling グループとして与えていたタグも除去される
 - 作業後、そのまま終了予定であればデタッチが適する

https://docs.aws.amazon.com/ja_jp/autoscaling/ec2/userguide/as-enter-exit-standby.html

https://docs.aws.amazon.com/ja_jp/autoscaling/ec2/userguide/detach-instance-asg.html



おわりに



今回お話しした内容

- Auto Scaling サービス群の整理
- EC2 Auto Scaling おすすめ機能
 - ワンタッチでできる自動スケール設定
 - ライフサイクルフック
 - インスタンスリフレッシュ
 - ウオームプール
- インスタンスの置き換えに関するよくある質問集

ヒント

AWS Black Belt コンピュートシリーズのあるきかた | Amazon Web Services ブログ –
<https://aws.amazon.com/jp/blogs/news/aws-black-belt-compute-series/>

AWS Black Belt Online Seminar とは

- ・ 「サービス別」「ソリューション別」「業種別」などのテーマに分け、
アマゾン ウェブ サービス ジャパン合同会社が提供するオンラインセミナー
シリーズです
- ・ AWS の技術担当者が、AWS の各サービスやソリューションについてテーマ
ごとに動画を公開します
- ・ 以下の URL より、過去のセミナー含めた資料などをダウンロードするこ
とができます
 - ・ <https://aws.amazon.com/jp/aws-jp-introduction/aws-jp-webinar-service-cut/>
 - ・ <https://www.youtube.com/playlist?list=PLzWGOASvSx6FIwIC2X1nObr1KcMCBBlqY>



ご感想は X (Twitter) へ！ハッシュタグは以下をご利用ください
#awsblackbelt

内容についての注意点

- ・ 本資料では資料作成時点のサービス内容および価格についてご説明しています。AWS のサービスは常にアップデートを続けているため、最新の情報は AWS 公式ウェブサイト (<https://aws.amazon.com/>) にてご確認ください
- ・ 資料作成には十分注意しておりますが、資料内の価格と AWS 公式ウェブサイト記載の価格に相違があった場合、AWS 公式ウェブサイトの価格を優先とさせていただきます
- ・ 価格は税抜表記となっています。日本居住者のお客様には別途消費税をご請求させていただきます
- ・ 技術的な内容に関しましては、有料の [AWS サポート窓口](#)へお問い合わせください
- ・ 料金面でのお問い合わせに関しましては、[カスタマーサポート窓口](#)へお問い合わせください（マネジメントコンソールへのログインが必要です）



Thank you!