[AWS Hands-on for Beginners]

Amazon EC2 Auto Scaling スケーリング基礎編

アマゾンウェブサービスジャパン株式会社
Solutions Architect
上野　涼平 / Ryohei Ueno
(収録日：2021/03/12)

自己紹介

- 名前

  上野　涼平　Ryohei Ueno

- ロール

  テクニカルソリューションアーキテクト

- 経歴

  ユーザ系企業にてクラウドへの移行から運用面までを横断的に支援

- 好きなAWSサービス

  AWS Systems Manager

AWS Hands-on for Beginnersとは

実際に手を動かしながら AWS の各サービスを学んでいただきます

初めてそのサービスをご利用される方がメインターゲットです

お好きな時間、お好きな場所でご受講いただけるオンデマンド形式です

テーマごとに合計1~2時間の内容 & 細かい動画に分けて公開
スキマ時間の学習や、興味のある部分だけの聴講も可能

内容についての注意点

- 資料では2021年03月12日収録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、資料とAWS公式ウェブサイトとで記載内容に相違があった場合、AWS公式ウェブサイトの記載を優先させていただきます。

- マネージメントコンソールについても、収録時点のものとなります。
  差異がある場合がございますので、ご注意ください。

- ハンズオンでは AWS の各種サービスの利用、リソースの作成を行います。
  無料枠を超えるハンズオンもございますが、その場合はご利用料金が発生することを
  あらかじめご認識ください。

- 学習後のリソースの削除についても、お客様の責任でご実施いただくようお願いいたします。

本ハンズオンのゴール

- Amazon EC2 Auto Scalingを使ったスケーリングについて理解する
- ユースケースに応じたスケーリングの使い分けを理解する
- 実際に手を動かし、スケーリングの設定方法を理解する

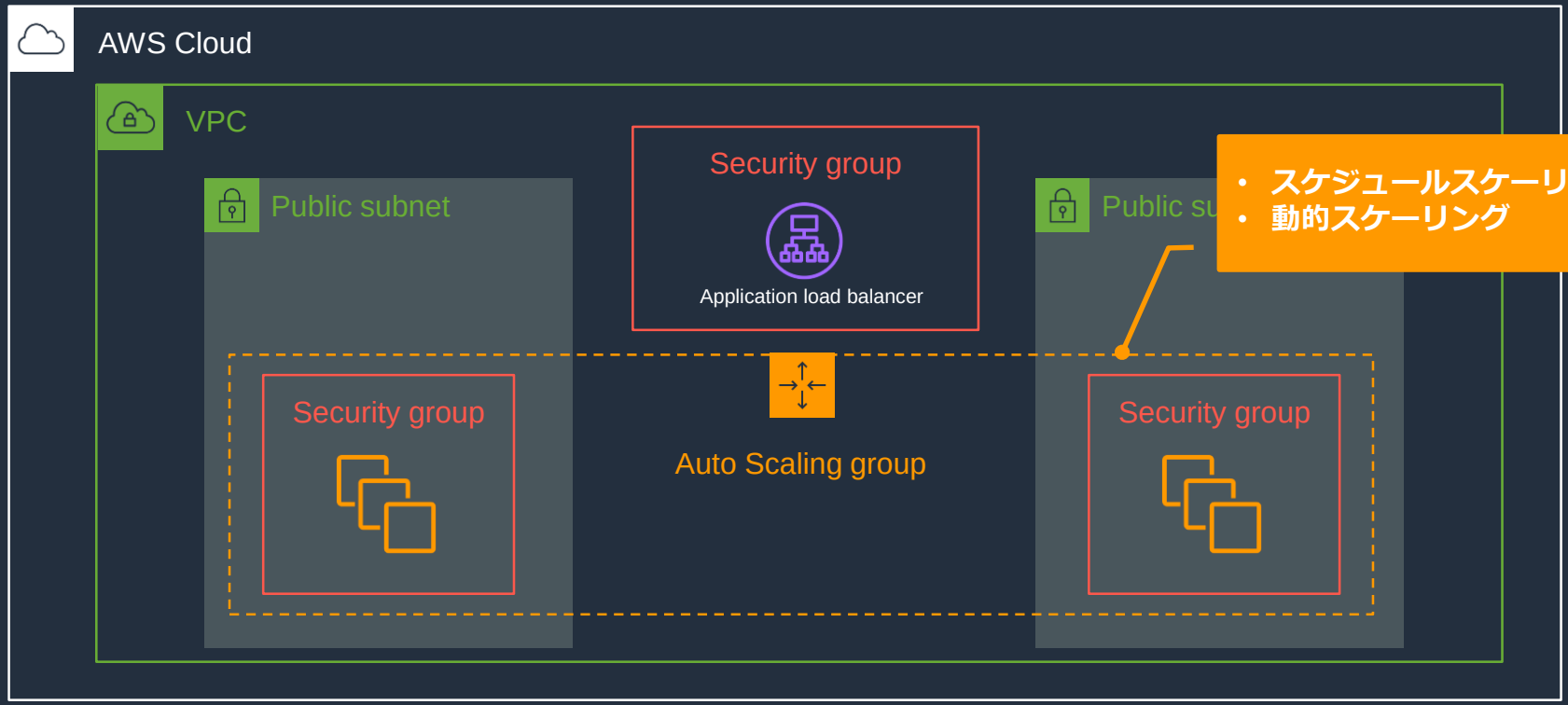
前提条件

- 事前にハンズオンのためのAWSアカウントのご用意をお願いします
- AdministratorAccess ポリシーのついた IAM ユーザーで作業を進めています
  - ※アカウントの作成方法、IAMユーザの作成方法がわからない方は、”ハンズオンはじめの一歩編”(※1)をご覧ください
- Amazon VPC、Amazon EC2の基礎知識をお持ちの方
- AWS Hands-on for Beginners ～スケーラブルウェブサイト構築編～(※2)を事前にご受講いただいていると、より理解を深めやすいと思います

※1 AWS Hands-on for Beginners ハンズオンはじめの一歩: AWS アカウントの作り方 & IAM 基本のキ
https://pages.awscloud.com/event_JAPAN_Ondemand_Hands-on-for-Beginners-1st-Step_LP.html?trk=aws_introduction_page

※2 AWS Hands-on for Beginners ～スケーラブルウェブサイト構築編～
https://pages.awscloud.com/event_JAPAN_Hands-on-for-Beginners-Scalable_LP.html?trk=aws_introduction_page

本ハンズオンで構築する最終構成



※ハンズオン利用するサービスの大部分は無料利用枠の対象になりますが、無料利用枠の適用範囲・料金については、サービス別の料金ページをご確認ください

アジェンダ

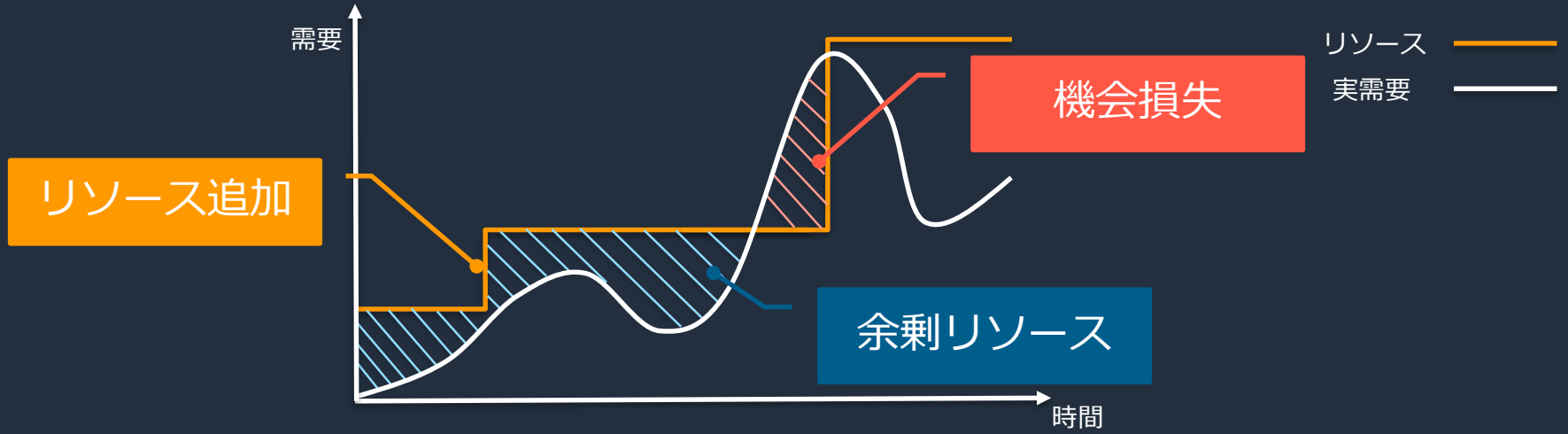
- はじめに
    - スケーリングにまつわる従来の課題
    - Amazon EC2 Auto Scalingとは？
- Hands-on
    - 事前準備
    - スケジュールスケーリング
    - 動的スケーリング - ターゲット追跡スケーリング
    - 異常なインスタンスの置き換え
- 本ハンズオンのまとめ、リソースの削除

アジェンダ

- はじめに
  - スケーリングにまつわる従来の課題
  - Amazon EC2 Auto Scalingとは？
- Hands-on
  - 事前準備
  - スケジュールスケーリング
  - 動的スケーリング - ターゲット追跡スケーリング
  - 異常なインスタンスの置き換え
- 本ハンズオンのまとめ、リソースの削除

スケーリングにまつわる従来の課題

- ビジネスの拡大とともに**マシンリソースの追加**が必要
- 正確な需要予測は難しく、**余剰リソース**を見込んだ調達が必要
- 突発的な需要拡大に対応できずビジネスの**機会損失**につながる



スケーリングにまつわる従来の課題

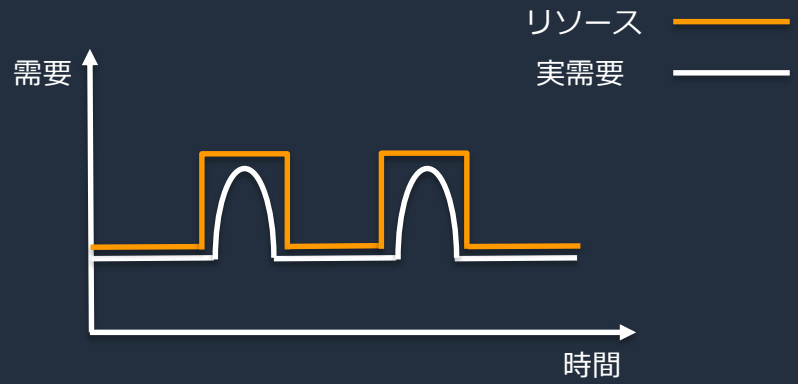
スケーリングにまつわる課題		求められること
リソース追加	✓ リソース追加作業が都度発生	自動でリソース追加
余剰リソース	✓ 需要予測作業の負荷 ✓ 余剰リソース分のコストが発生	需要に基づいた動的なスケーリング
機会損失	✓ ビジネス的な機会損失 ✓ サービス停止(可用性の低下)	スケーリングによる可用性の維持

アジェンダ

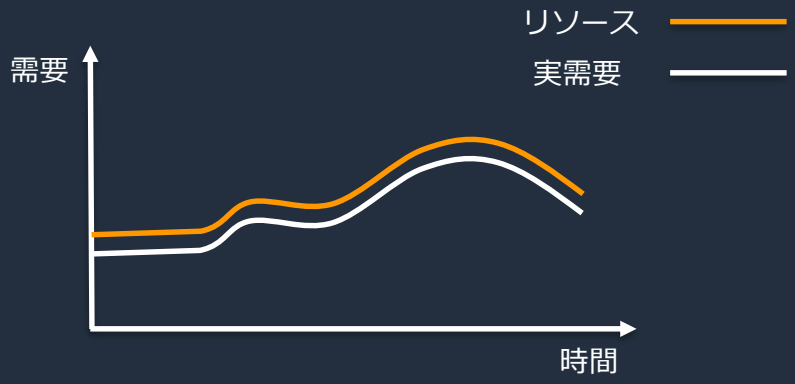
- はじめに
    - スケーリングにまつわる従来の課題
    - Amazon EC2 Auto Scalingとは？
- Hands-on
    - 事前準備
    - スケジュールスケーリング
    - 動的スケーリング - ターゲット追跡スケーリング
    - 異常なインスタンスの置き換え
- 本ハンズオンのまとめ、リソースの削除

Amazon EC2 Auto Scalingとは

- Amazon EC2 Auto ScalingはEC2インスタンスを自動的に追加または、削除することでアプリケーションの**可用性維持**、**コスト最適化**に役立ちます
- スケジュールに基づく**スケジュールスケーリング**、需要の変化に動的に対応する**動的スケーリング**を用いて柔軟なスケーリングを実現

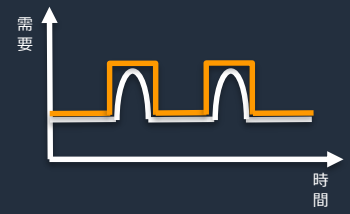


スケジュールスケーリングのイメージ図



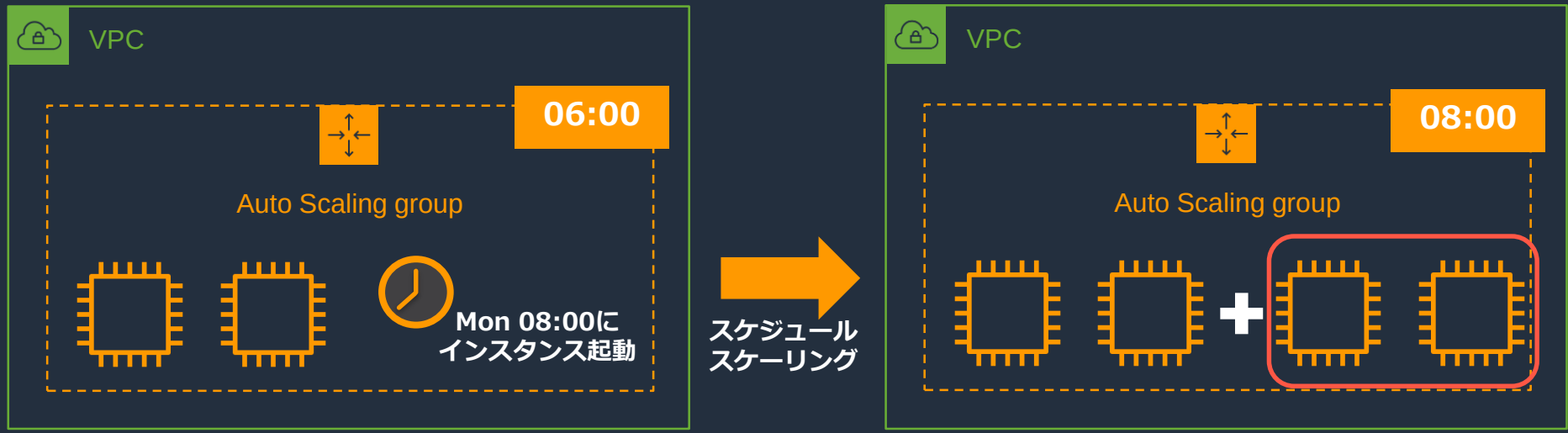
動的スケーリングのイメージ図

スケジュールスケーリング



- スケジュールスケーリングは、設定されたスケジュールに基づきスケーリングを行う

  ex. 毎週月曜日の08:00にインスタンスを4台にする
- **予測が出来る需要変化**への対策となる



スケジュールスケーリングのユースケース例

売上分析システムの例

システムの使われ方

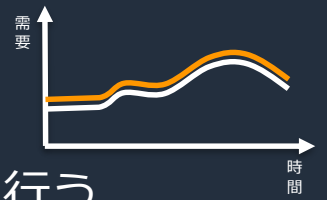
- **月曜日の午前**は、先週分の売上を分析するために**アクセスが集中**する
- 月曜日の午前以外はアクセス数が少ない

ワークロードの特性

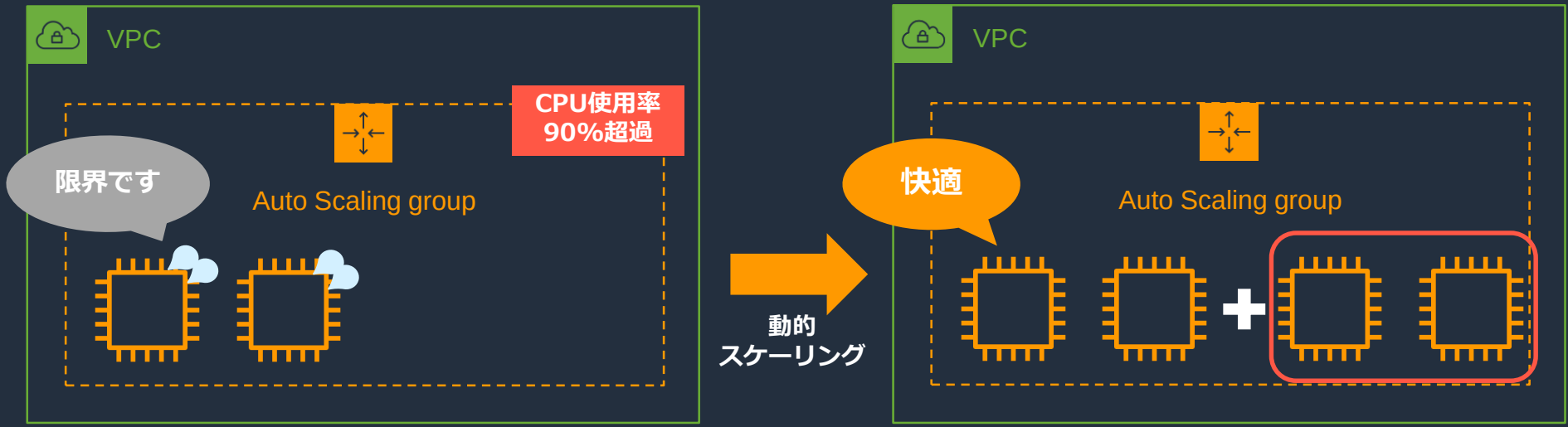
- アクセス数が増えてくるとサーバのCPU・メモリ使用率が高騰
- **マシンリソース不足でパフォーマンス劣化**

毎週月曜日の午前でスケジュールスケーリングを利用

動的スケーリング



- 動的スケーリングは、設定した閾値に基づき動的にスケーリングを行う
  ex. CPU使用率が90%を超えたらスケーリングさせる
- **予測ができない需要**への対策となるが、スケーリングには数分の時間を要するため、**スパイクアクセスにおける利用には注意**が必要



動的スケーリングのユースケース例

ECサイトの例

ECサイトの使われ方

- アクセス数におおまかな傾向はある
- SNSの普及により特定のコンテンツについて**予測できないアクセス増**がある

ワークロードの特性

- アクセス数が増えてくるとサーバのCPU使用率が高騰
- **CPU使用率が高いとパフォーマンス劣化**

CPU使用率をしきい値にした動的スケーリングを利用

アジェンダ

- はじめに
    - スケーリングにまつわる従来の課題
    - Amazon EC2 Auto Scalingとは？
- Hands-on
    - 事前準備
    - スケジュールスケーリング
    - 動的スケーリング - ターゲット追跡スケーリング
    - 異常なインスタンスの置き換え
- 本ハンズオンのまとめ、リソースの削除

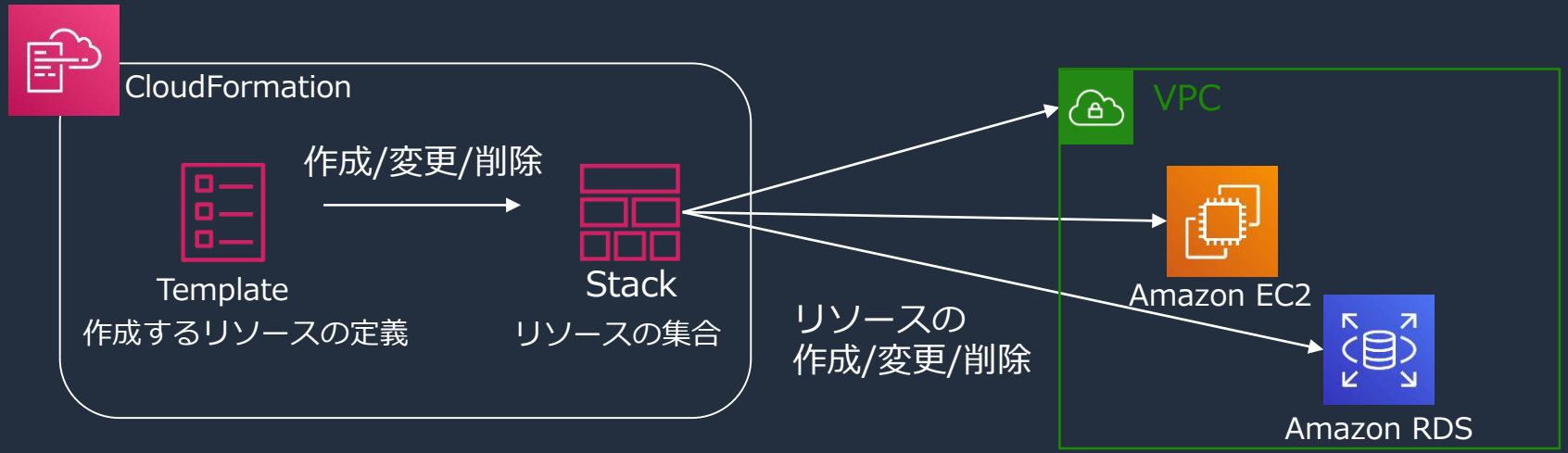
事前準備

- ハンズオンに必要なリソースは**CloudFormationテンプレート**を利用して作成
  - Amazon VPC
  - サブネット
  - インターネットゲートウェイ
  - ルートテーブル
  - セキュリティグループ
  - Application Load Balancer、ターゲットグループ

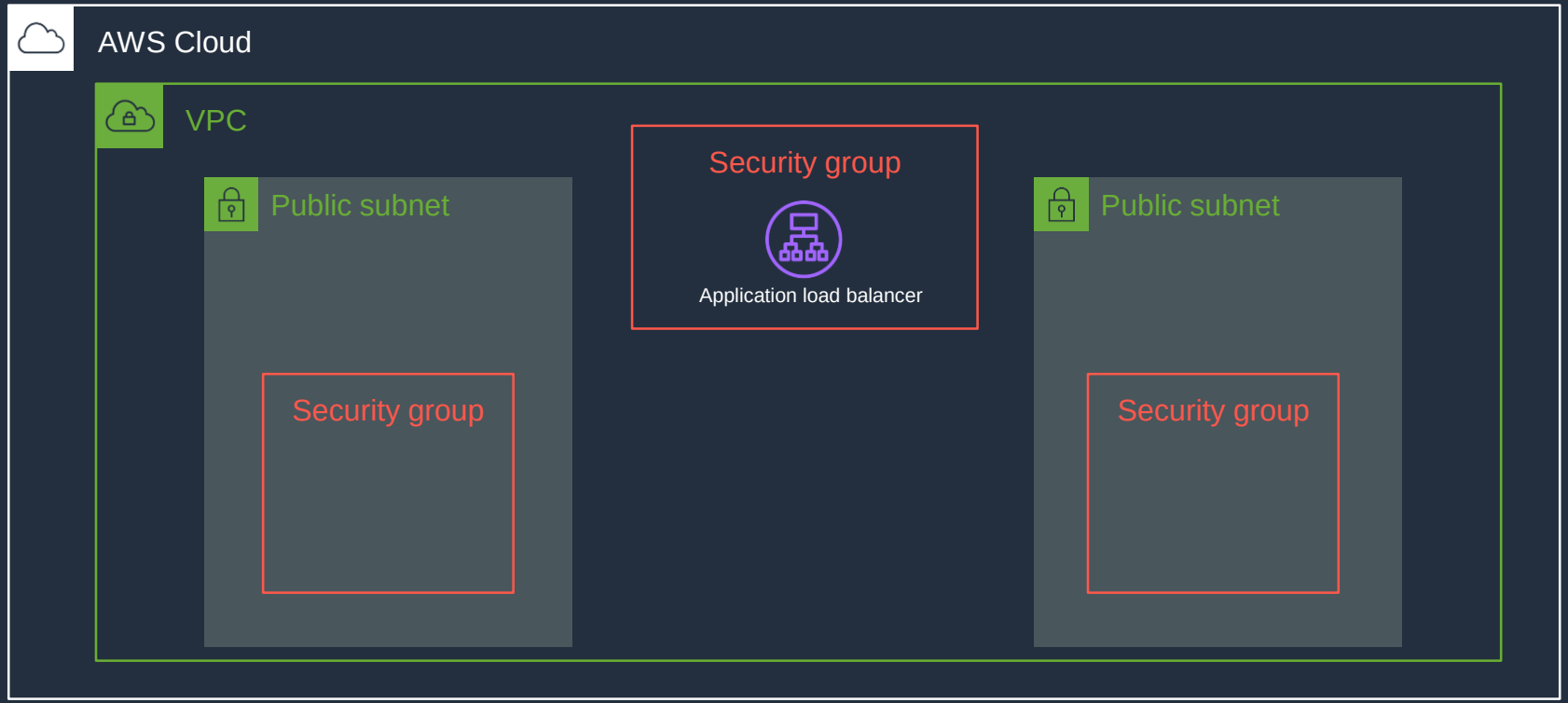
- 作成されるリソースの詳細を学ぶ際は、以下のハンズオンをご受講ください
  - AWS Hands-on for Beginners ～スケーラブルウェブサイト構築編～
    https://pages.awscloud.com/event_JAPAN_Hands-on-for-Beginners-Scalable_LP.html?trk=aws_introduction_page
  - Network編#1 AWS上にセキュアなプライベートネットワークを空間を作成する
    https://pages.awscloud.com/JAPAN-event-OE-Hands-on-for-Beginners-Network1-2020-reg-event-LP.html?trk=aws_introduction_page

AWS CloudFormation

- AWS環境のコードによる管理を実現
- テンプレートで定義した環境を 作成 / 変更 / 削除



事前準備リソース



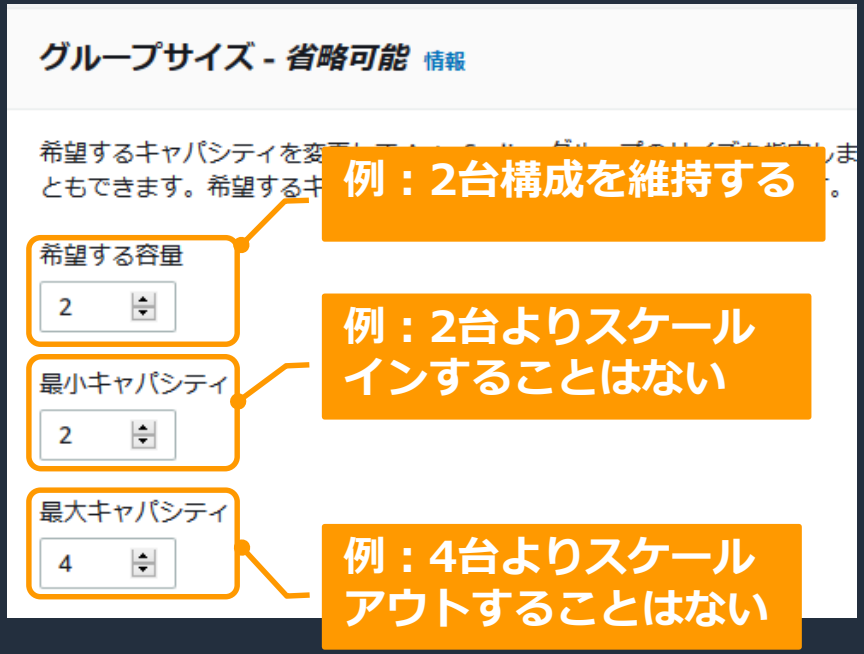
Amazon EC2 Auto Scalingのコンポーネント

Amazon EC2 Auto Scalingのコンポーネント
起動テンプレート

- Amazon EC2 Auto Scalingによって起動されるインスタンスの起動情報を設定するテンプレート

- 起動情報とは
  - Amazonマシンイメージ(AMI)
  - インスタンスタイプ
  - ネットワーク設定　等

Amazon EC2 Auto Scalingのコンポーネント
Auto Scalingグループ



- スケーリングの設定を行うコンポーネント
- 設定内容
  - どの起動テンプレートを利用するか
  - 何台構成を維持するか
  - スケールアウトする最大の台数
  - スケールインする最小の台数
  - 何をトリガーにスケールするか

アジェンダ

- はじめに
    - スケーリングにまつわる従来の課題
    - Amazon EC2 Auto Scalingとは？
- Hands-on
    - 事前準備
    - スケジュールスケーリング
    - 動的スケーリング - ターゲット追跡スケーリング
    - 異常なインスタンスの置き換え
- 本ハンズオンのまとめ、リソースの削除

スケジュールスケーリングの設定

- 一度きりの実行や毎日、毎週といったスケジュール実行が要件に合わせて可能
- マネージメントコンソールから時刻の設定を行う場合、UTCで指定する必要がある

  ex. JST 15:30 → UTC 06:30
- 周期的なスケジュールを設定する場合は、終了時刻も設定が必要

スケジュールスケーリング
ハンズオンの流れ

起動テンプレート
の作成

Auto Scaling
グループの作成

スケジュール
の設定

スケジュールスケーリング
ハンズオンの流れ

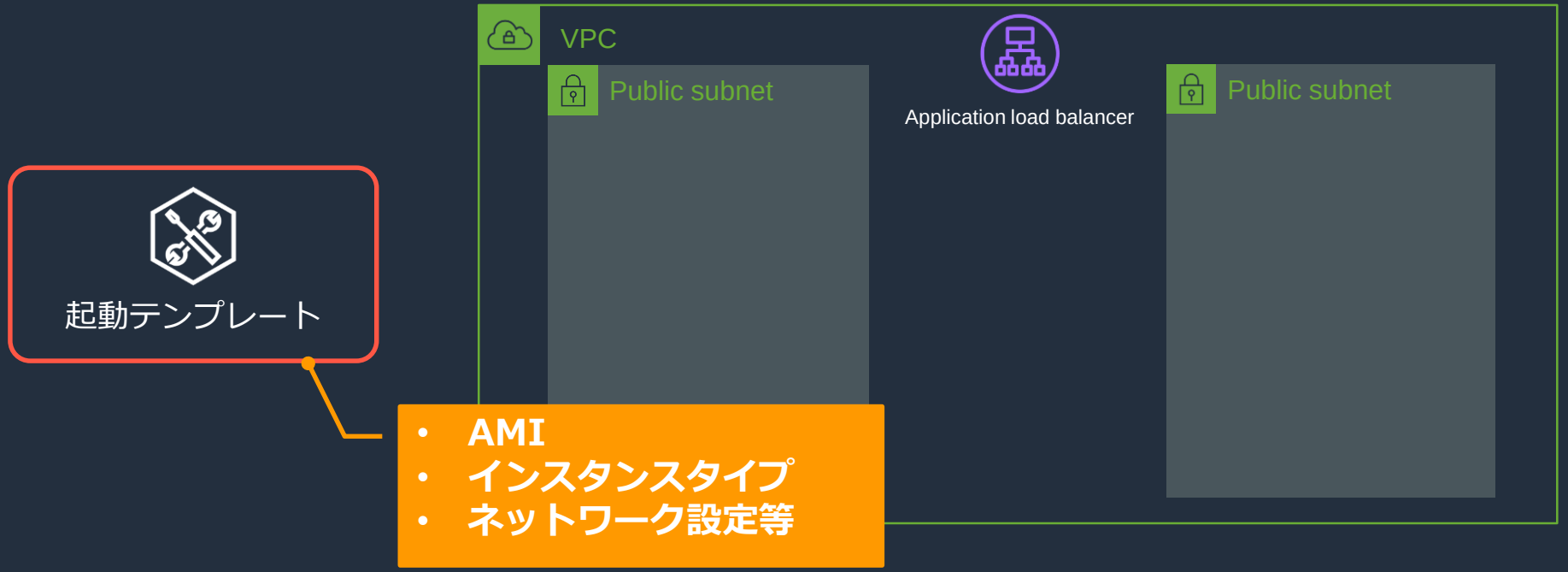
起動テンプレート
の作成

Auto Scaling
グループの作成

スケジュール
の設定

aws

起動テンプレートの作成



スケジュールスケーリング
ハンズオンの流れ

起動テンプレート
の作成

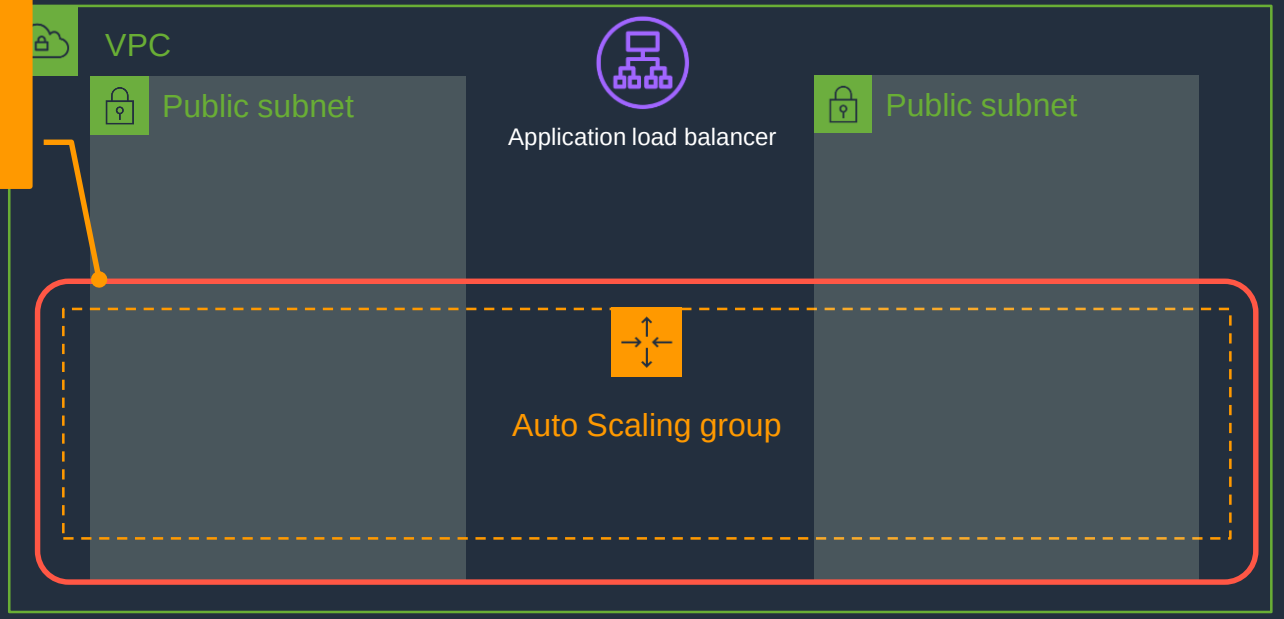
Auto Scaling
グループの作成

スケジュール
の設定

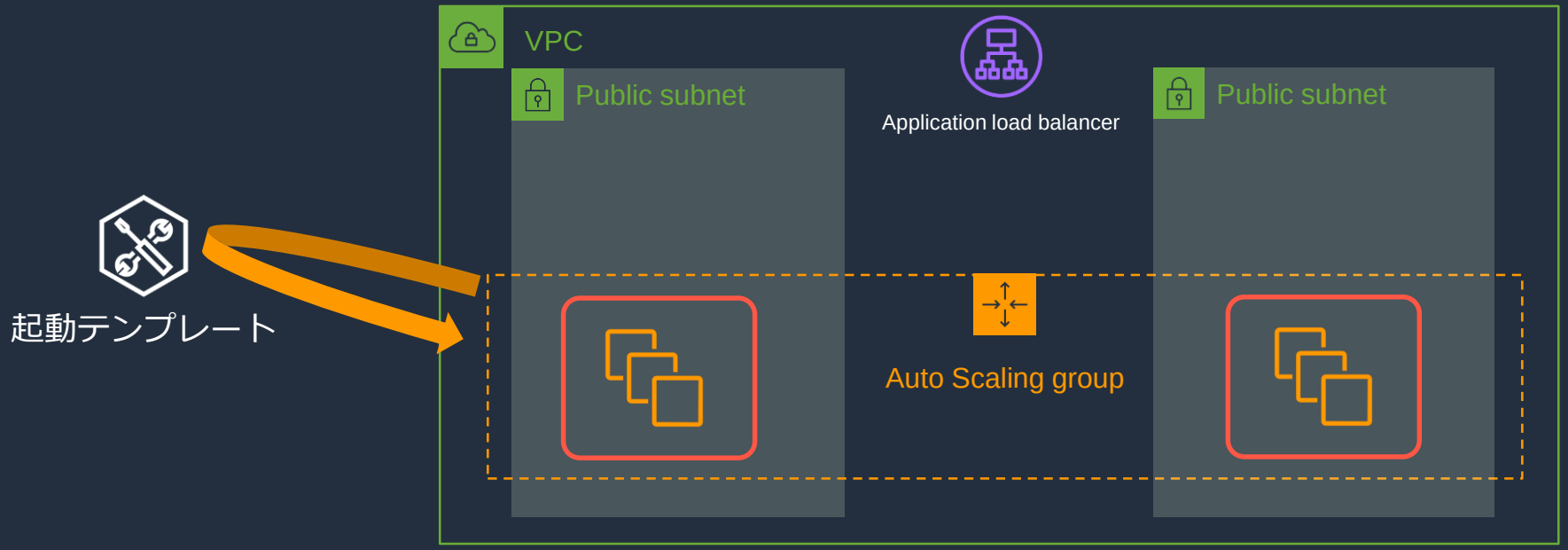
Auto Scalingグループの作成

- 起動テンプレート
- 何台構成として起動するか
- スケールアウトの最大値
- スケールインの最小値

起動テンプレート



希望する容量に応じたEC2インスタンスの起動確認



スケジュールスケーリング
ハンズオンの流れ

起動テンプレート
の作成

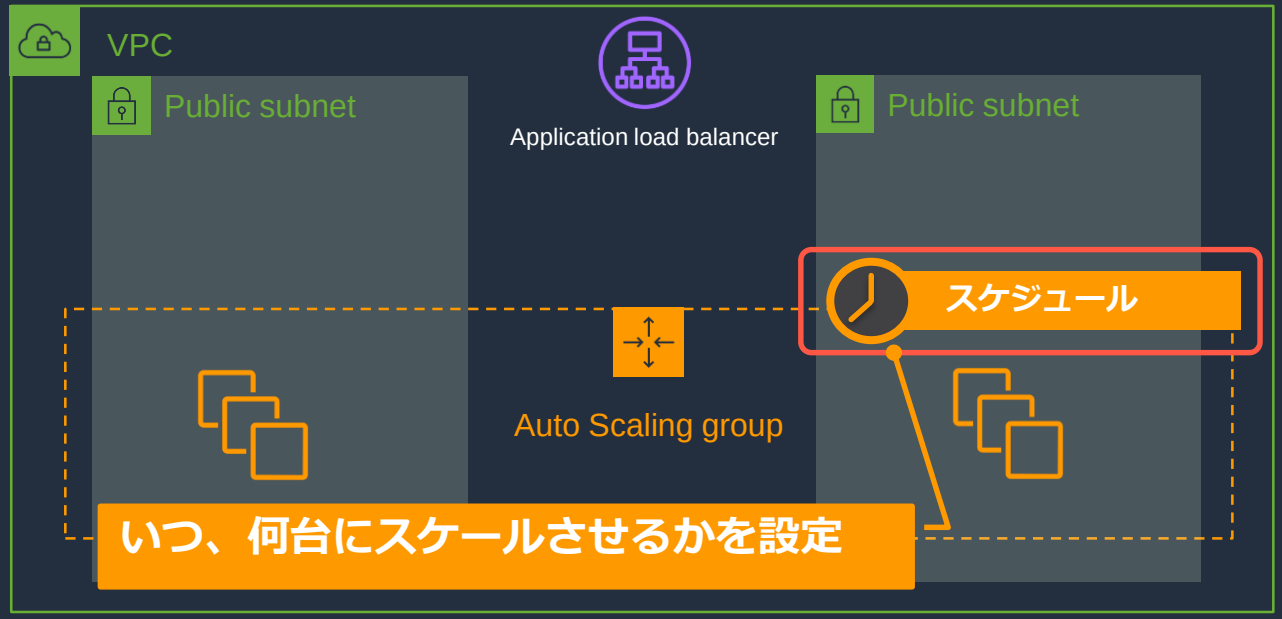
Auto Scaling
グループの作成

スケジュール
の設定

aws

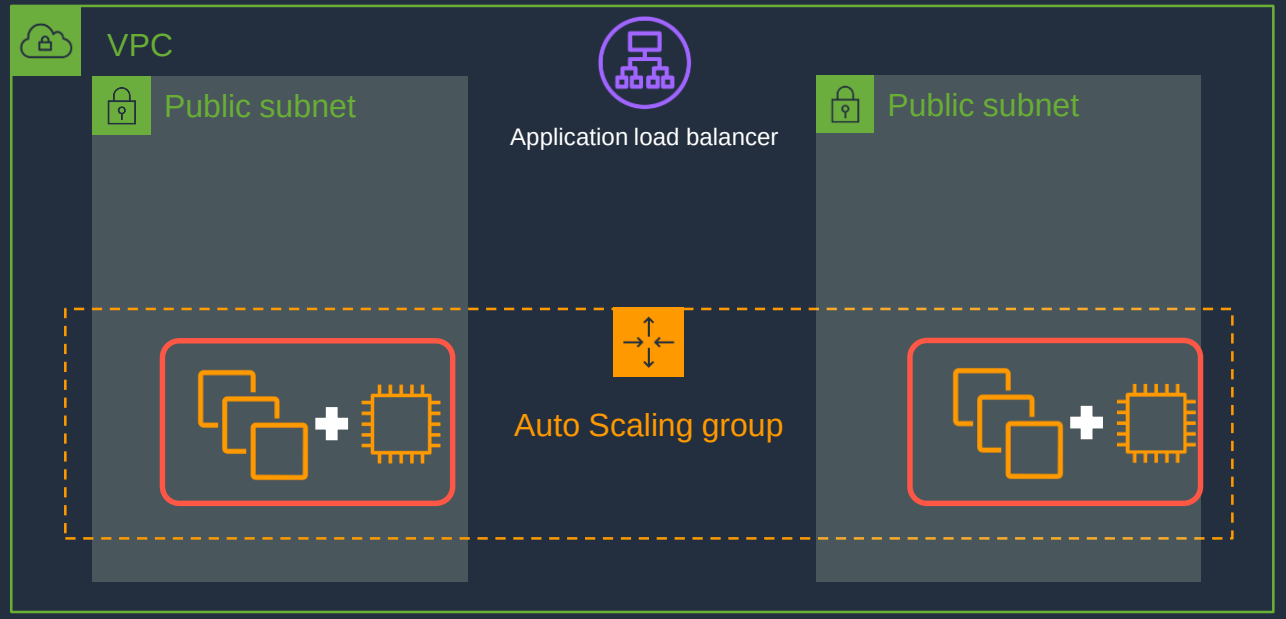
スケジュールスケーリングの設定

起動テンプレート



スケジュールスケーリングの確認

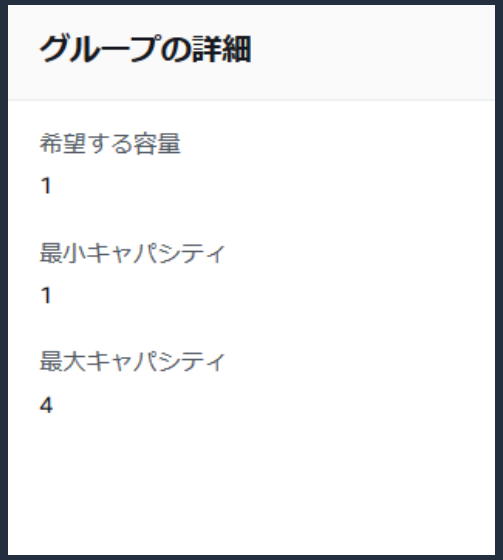
起動テンプレート



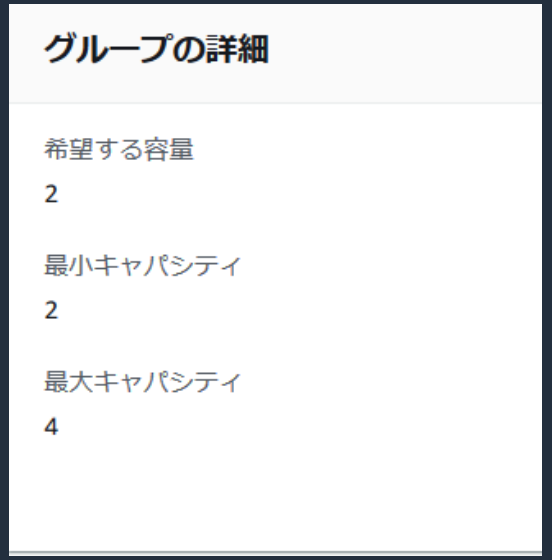
補足：スケジュールスケーリングによるグループサイズ変更

グループサイズの変更は手動で行えるが、スケジュールをトリガーに自動で
変更を行うのがスケジュールスケーリング

スケジュールスケーリング実行前



スケジュールスケーリング実行後



アジェンダ

- はじめに
    - スケーリングにまつわる従来の課題
    - Amazon EC2 Auto Scalingとは？
- Hands-on
    - 事前準備
    - スケジュールスケーリング
    - 動的スケーリング - ターゲット追跡スケーリング
    - 異常なインスタンスの置き換え
- 本ハンズオンのまとめ、リソースの削除

スケーリングポリシー

スケーリングポリシー

- スケーリングポリシーには、以下の3種類が設定できる
    - ターゲット追跡スケーリング
    - ステップスケーリング
    - シンプルなスケーリング

- 要件に応じてスケーリングポリシーを設定することになりますが、
本ハンズオンでは**ターゲット追跡スケーリング**を利用します

- その他スケーリングポリシーの詳細はドキュメントをご確認ください

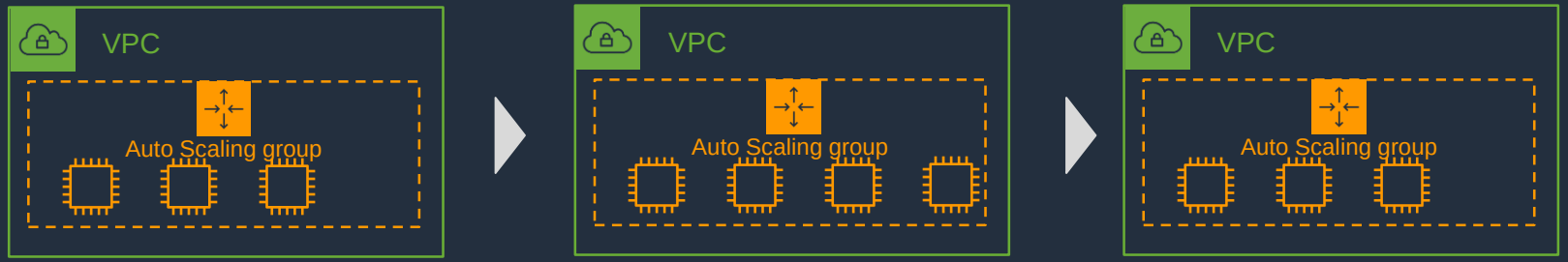
  スケーリングポリシータイプ
  https://docs.aws.amazon.com/ja_jp/autoscaling/ec2/userguide/as-scale-based-on-demand.html#as-scaling-types

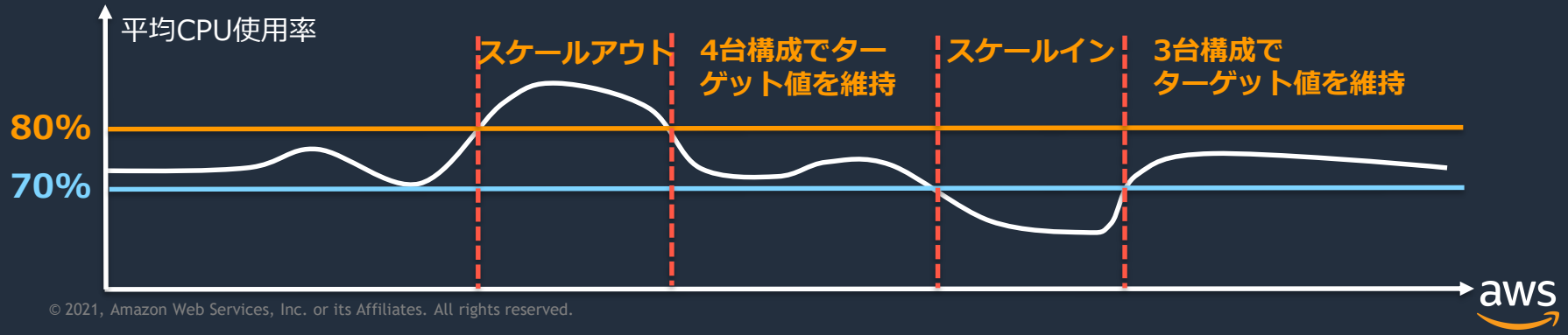
ターゲット追跡スケーリング

ターゲット追跡スケーリング

ターゲット追跡スケーリングは、設定した**ターゲット値を維持**するように
スケールアウト・スケールインを行います

ex. 平均CPU使用率を70%～80%で維持する場合





ターゲット追跡スケーリング
ハンズオンの流れ

ターゲット追跡
スケーリング設定

負荷をかけて
スケーリング確認

負荷を止めて
スケールイン確認

ターゲット追跡スケーリング
ハンズオンの流れ

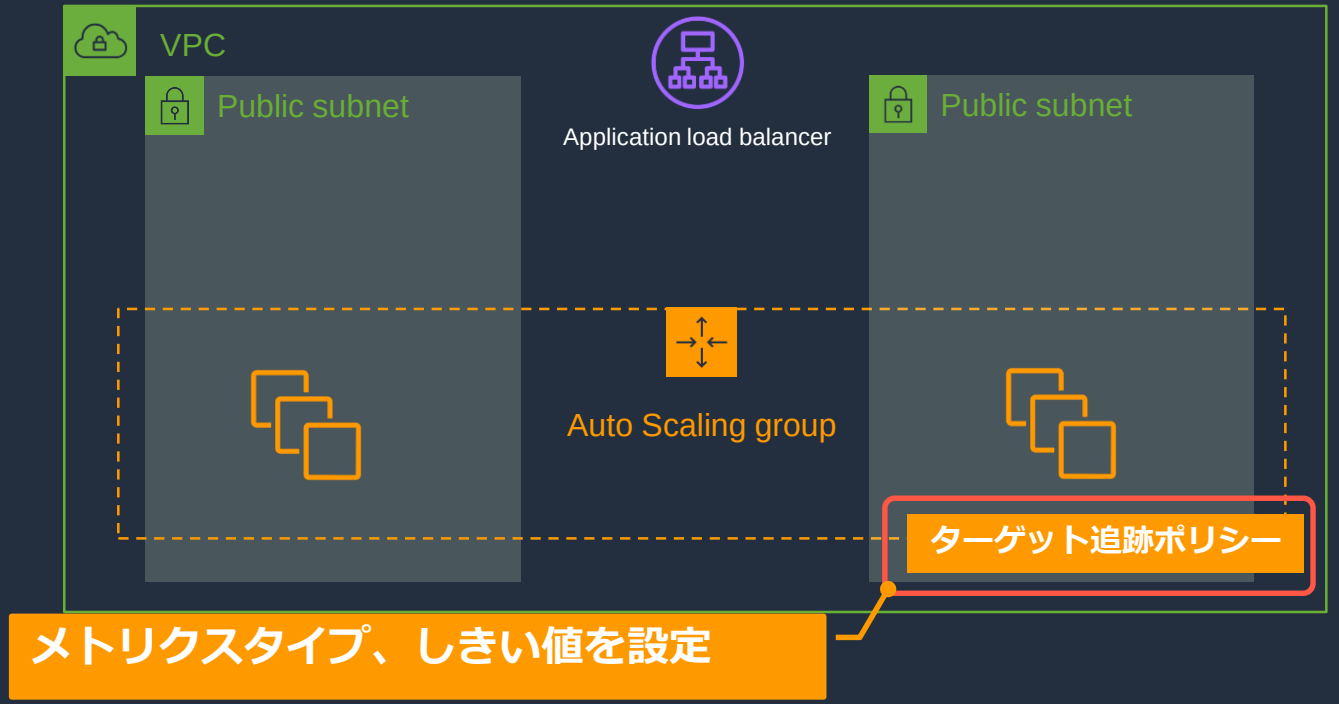
ターゲット追跡
スケーリング設定

負荷をかけて
スケーリング確認

負荷を止めて
スケールイン確認

ターゲット追跡スケーリングの設定

起動テンプレート



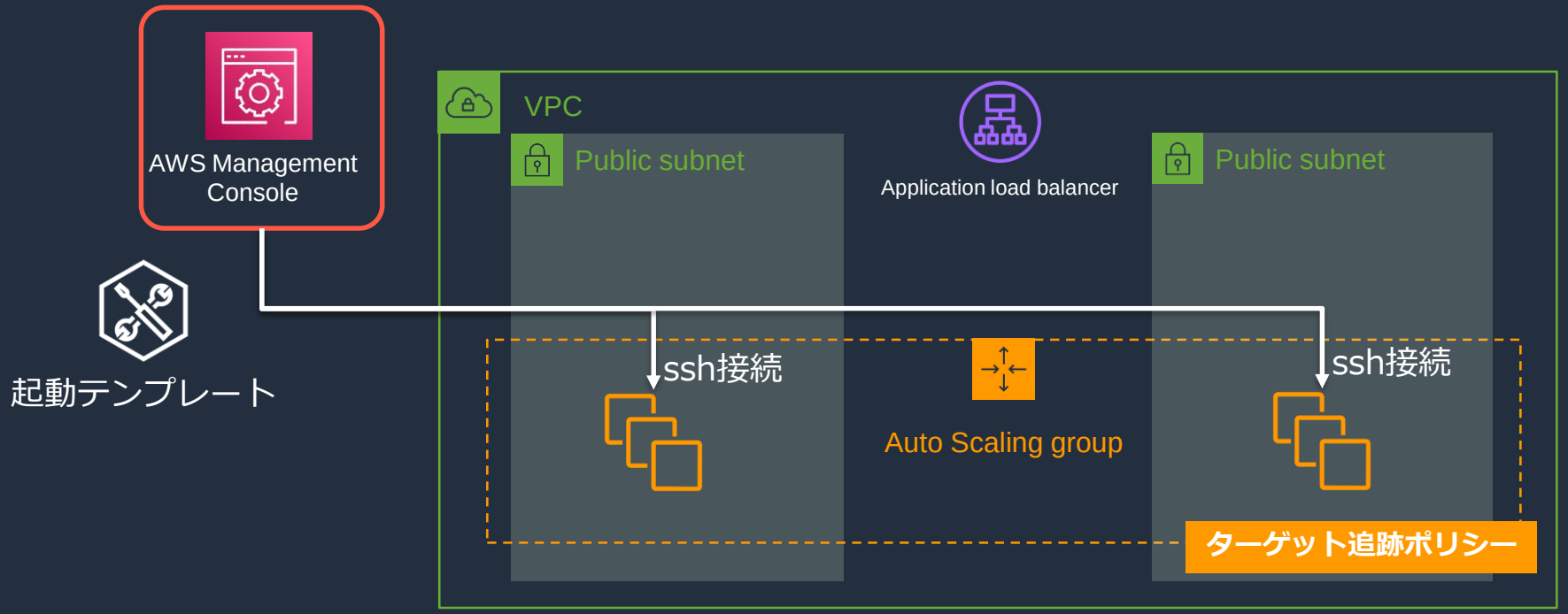
ターゲット追跡スケーリング
ハンズオンの流れ

ターゲット追跡
スケーリング設定

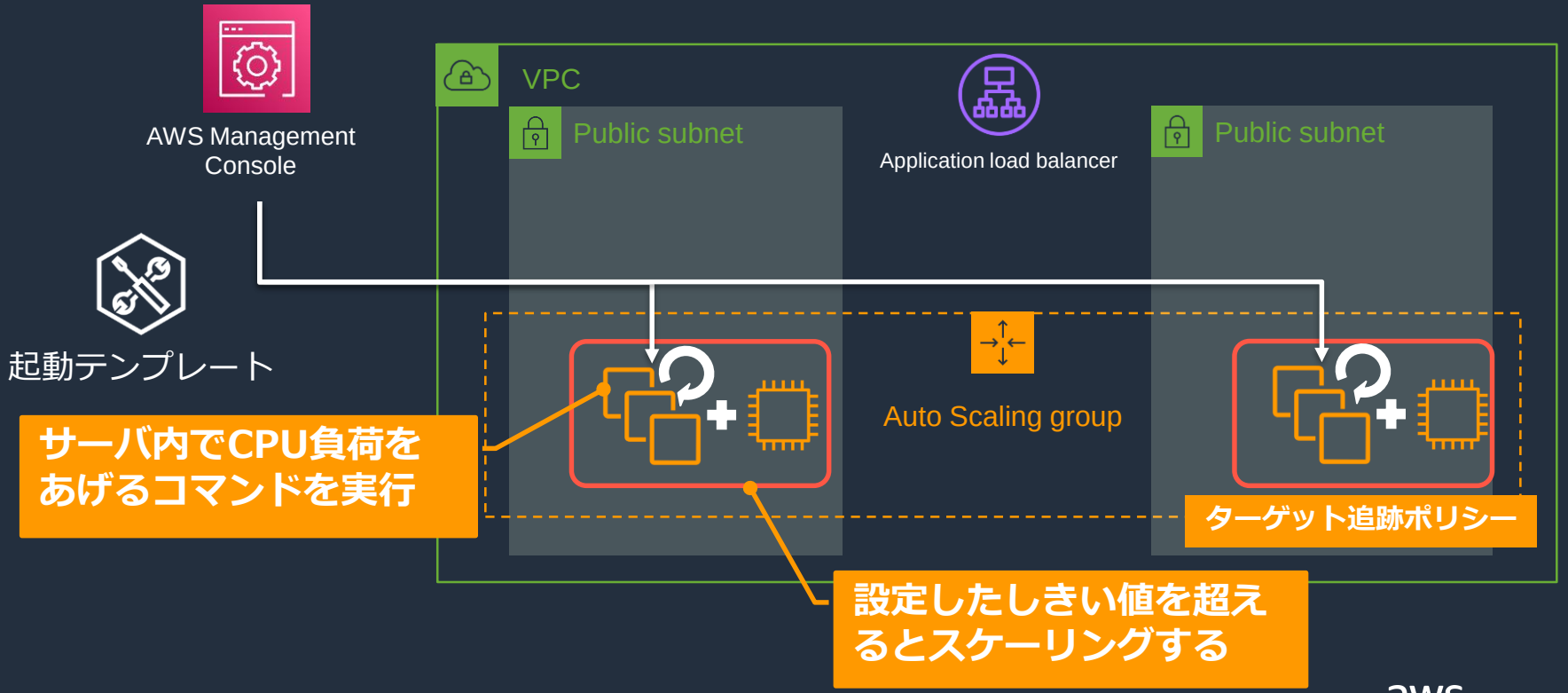
負荷をかけて
スケーリング確認

負荷を止めて
スケールイン確認

EC2インスタンスへのssh接続



負荷をかけてスケーリングを確認



ターゲット追跡スケーリング ハンズオンの流れ

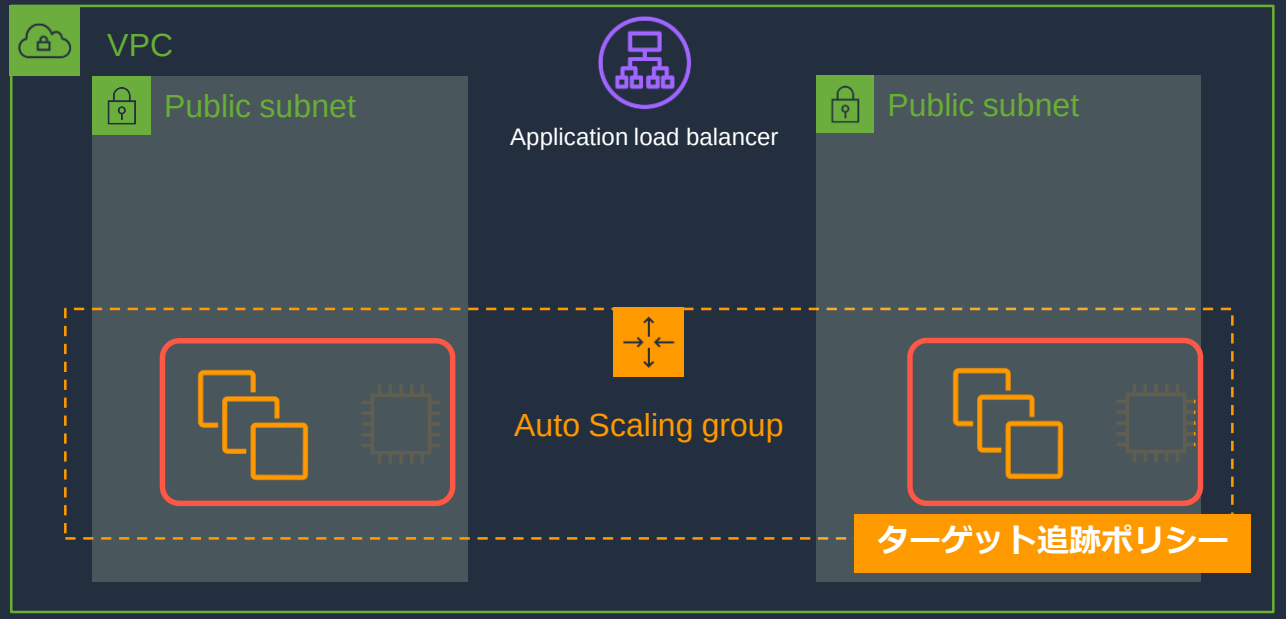
ターゲット追跡
スケーリング設定

負荷をかけて
スケーリング確認

負荷を止めて
スケールイン確認

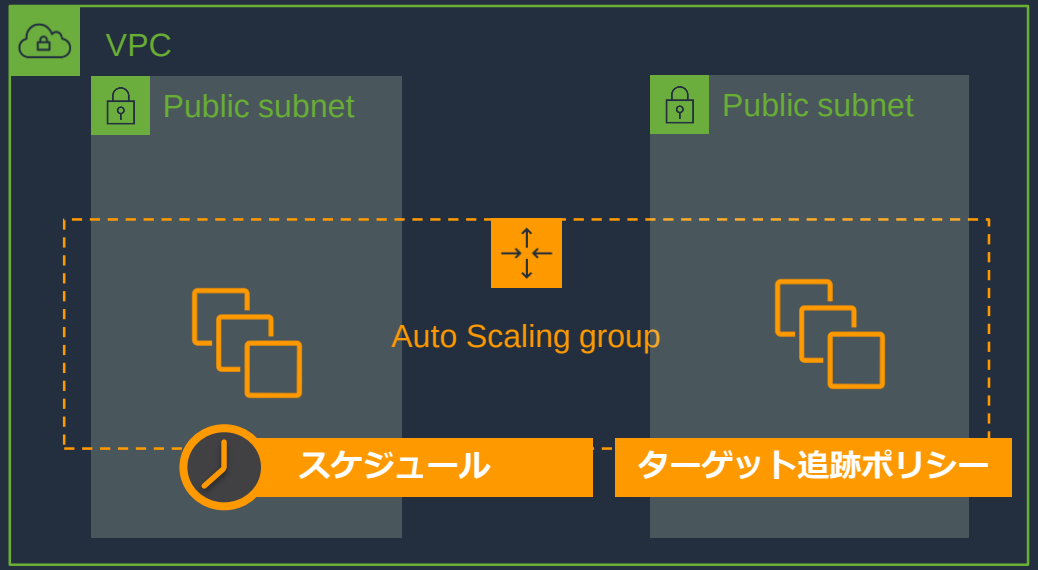
負荷を止めてスケールインを確認

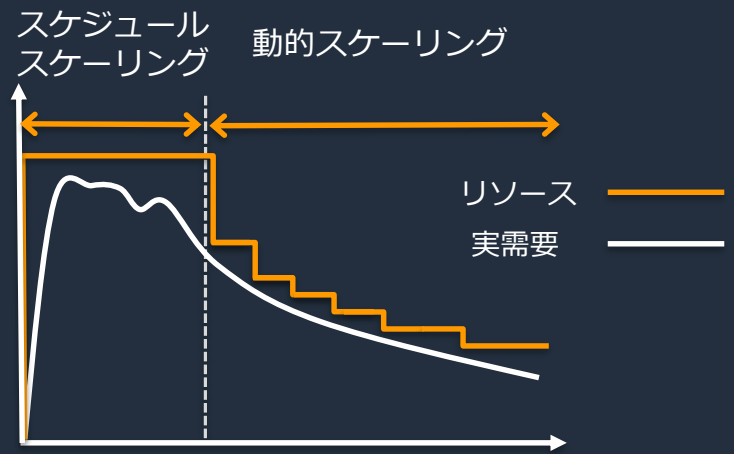
起動テンプレート



補足：動的スケーリングとスケジュールスケーリングの組合せ

動的スケーリングとスケジュールスケーリングを組み合わせることも可能
ex. 急激なアクセス増であるがタイミングが予測でき、その後ゆるやかに減少

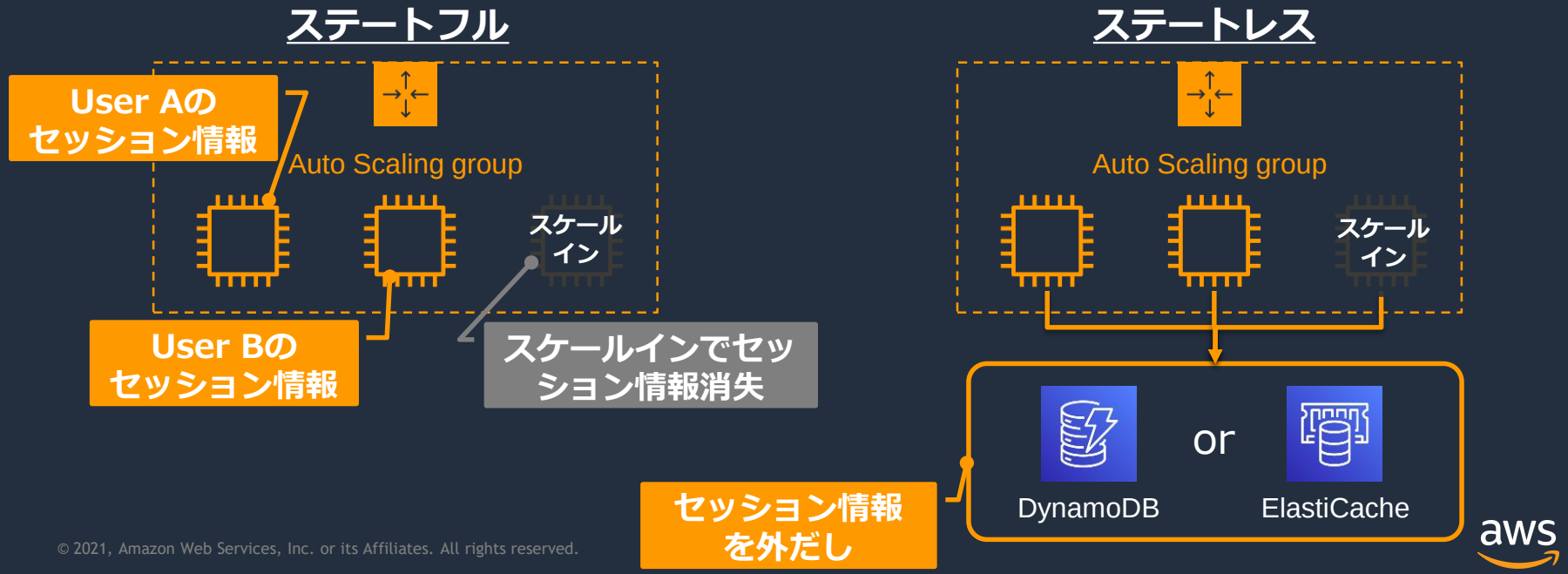




補足：スケーリングを前提としたアプリケーション設計

- スケーリングによってインスタンスが増減するため、特定のサーバに依存するようなステートフルな作りではなく、ステートレスな作りが求められる

ex. セッション管理

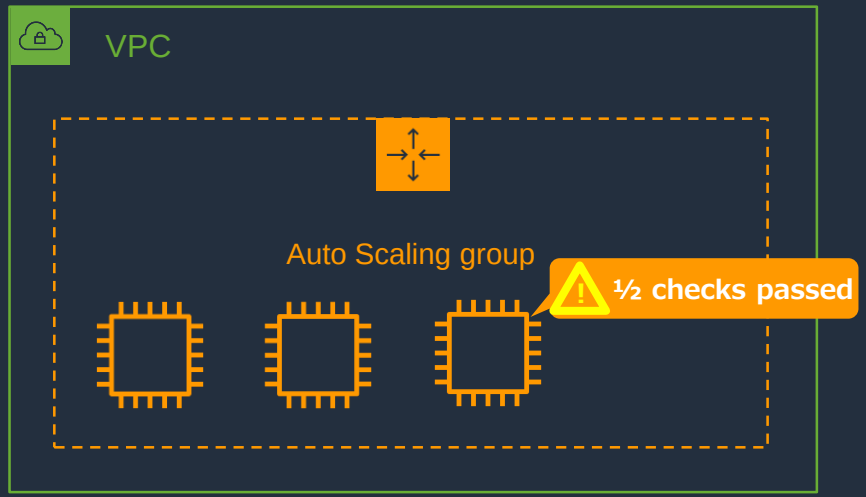


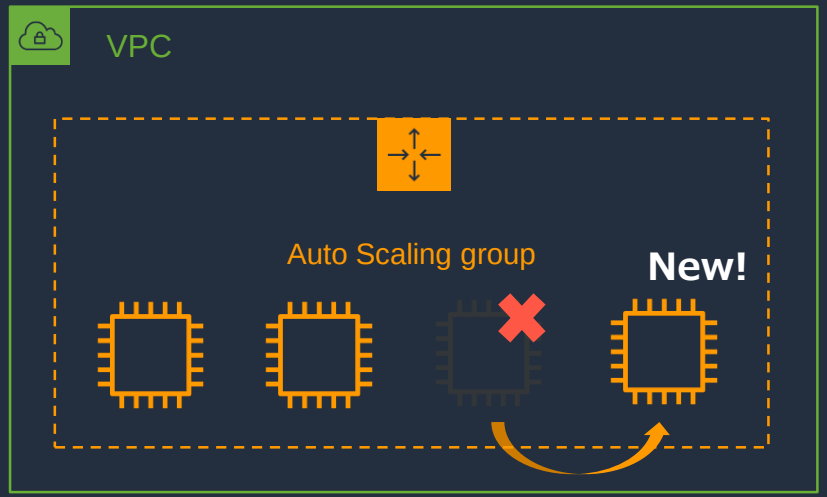
アジェンダ

- はじめに
    - スケーリングにまつわる従来の課題
    - Amazon EC2 Auto Scalingとは？
- Hands-on
    - 事前準備
    - スケジュールスケーリング
    - 動的スケーリング - ターゲット追跡スケーリング
    - 異常なインスタンスの置き換え
- 本ハンズオンのまとめ、リソースの削除

異常なインスタンスの置き換え

- Amazon EC2 Auto ScalingはEC2のステータスチェックまたは、Elastic Load Balancingのヘルスチェックに応じて異常なインスタンスを置き換えます
- アクセス増減に伴うスケーリングの自動化だけでなく、Auto Scalingグループ内の異常なインスタンスを置き換えることによっても**可用性を維持する**





異常なインスタンス置き換え ハンズオンの流れ

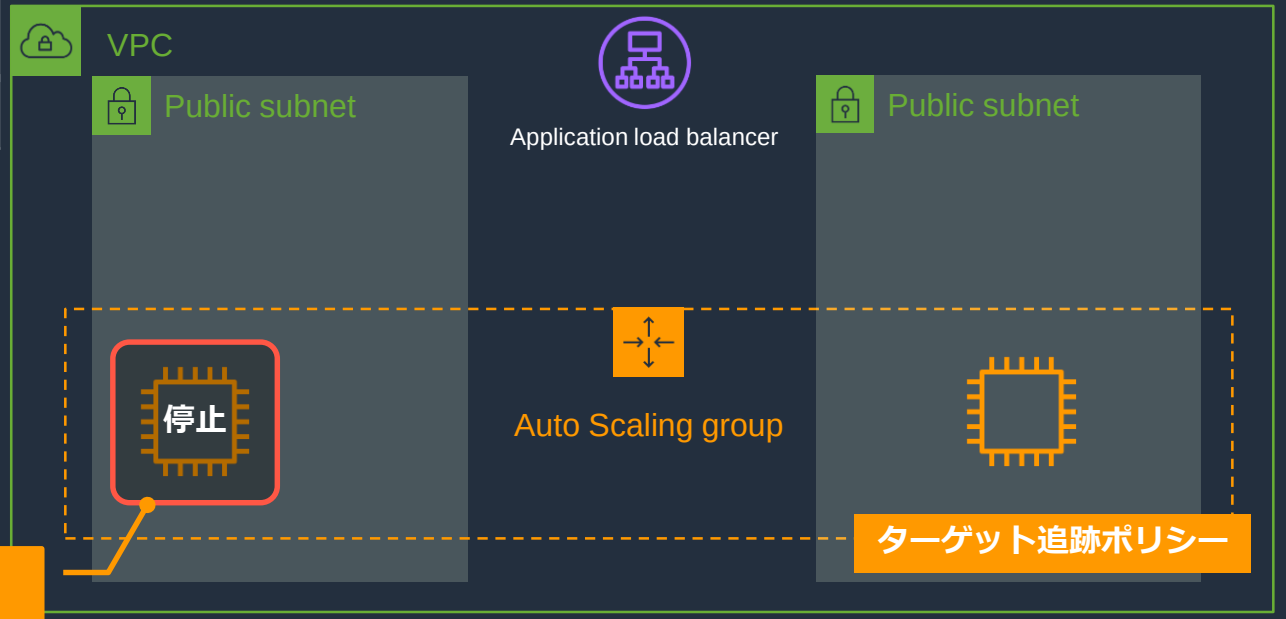
Auto Scalingグループ内の インスタンスを停止する

インスタンスの置き換えを確認

Auto Scalingグループ内のインスタンスを停止する

希望する容量	現在のインスタンス数
2	1

起動テンプレート

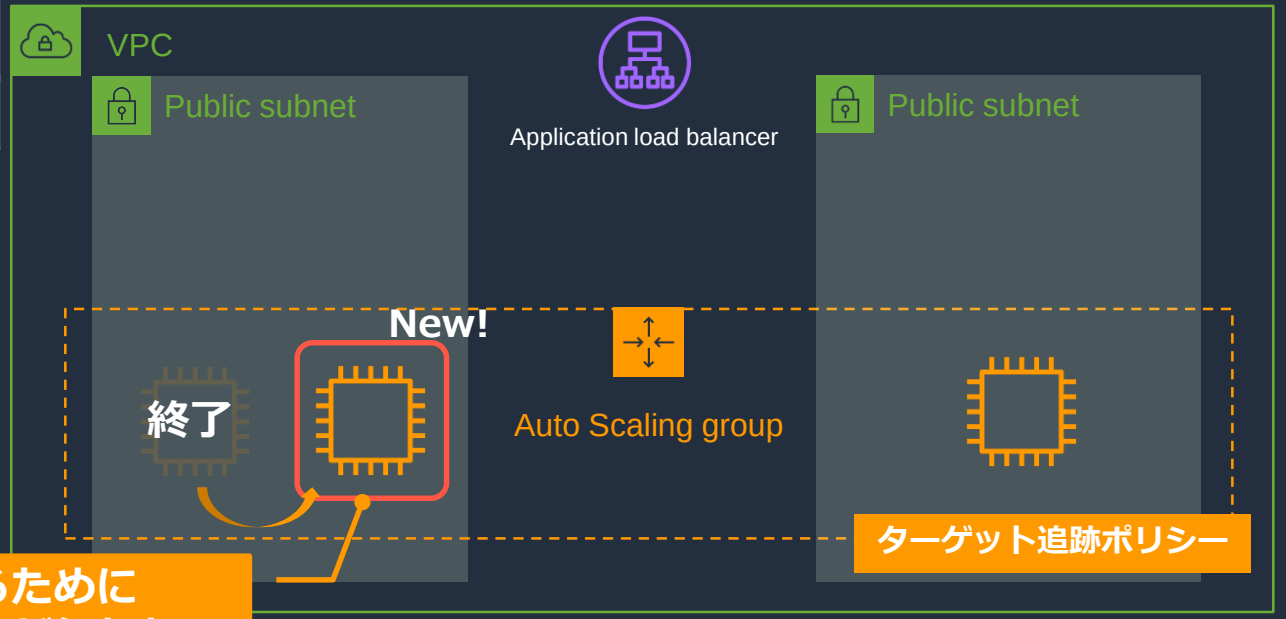


ヘルスチェックで異常なインスタンスとみなされる

インスタンスの置き換えを確認

希望する容量	現在のインスタンス数
2	2

起動テンプレート



希望する容量を維持するために
インスタンスの置き換えが行われる

アジェンダ

- はじめに
    - スケーリングにまつわる従来の課題
    - Amazon EC2 Auto Scalingとは？
- Hands-on
    - 事前準備
    - スケジュールスケーリング
    - 動的スケーリング - ターゲット追跡スケーリング
    - 異常なインスタンスの置き換え
- 本ハンズオンのまとめ、リソースの削除

リソースの削除

1. Auto Scalingグループの削除
    - EC2インスタンスも削除されていること
    - CloudWatch Alarmも削除されていること

2. 起動テンプレートの削除

3. Cloud Formationスタックの削除

本ハンズオンのまとめ

- Amazon EC2 Auto Scalingを使ったスケーリングについて
    - スケジュールスケーリング
    - 動的スケーリング
    - コストの最適化と可用性の維持

- ユースケースに応じたスケーリングの使い分けについて
    - ワークロードの利用のされ方、特性を理解することが重要

- 実際に手を動かし、スケーリングの設定方法を学んでいただきました

Next Step1　サービスを深く知る

- Black Belt：Amazon EC2 Auto Scaling and AWS Auto Scaling

  https://aws.amazon.com/jp/blogs/news/webinar-bb-amazon-ec2-auto-scaling-and-aws-auto-scaling-2019/

- よくある質問：Amazon EC2 Auto Scaling

  https://aws.amazon.com/jp/ec2/autoscaling/faqs/

Next Step2　実運用を想定する

- インスタンスの更新に基づくAuto Scalingインスタンスの置き換え
  https://docs.aws.amazon.com/autoscaling/ec2/userguide/asg-instance-refresh.html

- AWS Well-Architectedフレームワーク　信頼性の柱
  障害を軽減または耐えるために分散システムでの操作を設計する
  https://docs.aws.amazon.com/ja_jp/wellarchitected/latest/reliability-pillar/design-interactions-in-a-distributed-system-to-mitigate-or-withstand-failures.html

アンケートも
よろしくお願いします！