ECS on EC2 環境における
キャパシティープロバイダーとスポットインスタンスの活用事例

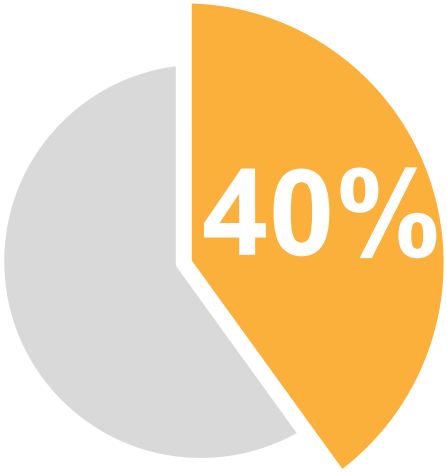
2021年8月

シルバーエッグ・テクノロジー株式会社
エンジニアリング部
白井 慎太郎

お問い合わせ先： シルバーエッグ・テクノロジー株式会社　Tel: 06-6386-1931(Osaka) / 03-5357-1255(Tokyo) https://www.silveregg.co.jp / E-mail: inquiry@silveregg.co.jp

弊社のご紹介

シルバーエッグ・テクノロジーは、1998年創業以来一貫してレコメンド技術の開発・提供を行う、シェアNo.1のレコメンド企業です。



SaaS型レコメンドツール市場占有率

2020年9月 富士キメラ総研調べ「ソフトウェアビジネス新市場 2020年版」
SaaS型レコメンドツールの市場占有率推移調査

国内No.1のレコメンドベンダー

最高精度のレコメンド
最先端テクノロジーとリアルタイム処理による業界最高精度のレコメンド

コンサルティングサポート
コンサルタントによる成果に繋げるためのカスタマーサービス

各種ツールとの連携実績
WEB接客ツール、検索ツール、メール配信ツール、MAとの連携

お問い合わせ先： シルバーエッグ・テクノロジー株式会社　Tel: 06-6386-1931(Osaka) / 03-5357-1255(Tokyo) https://www.silveregg.co.jp / E-mail: inquiry@silveregg.co.jp

SILVER EGG TECHNOLOGY

aigent recommender

リアルタイム・レコメンドサービス『アイジェント・レコメンダー』

AIエンジンが実現する、高精度のレコメンデーション

『アイジェント・レコメンダー』は、サイトやMoile Appを利用しているユーザーの過去と現在の行動情報から、ユーザーが求めているものを即座に推測し、提示します。

✔ 高精度のレコメンドを実現：ユーザー導線分析（パス・ディペンデンシー）

ユーザーがどのようなものを、どの順番で見てきたのか、買ってきたのか？ 過去の閲覧・行動の“経路”を分析することで、ユーザーの嗜好をより高い精度で理解します。

✔ 商機を逃さないリアルタイム相関生成

リアルタイムでユーザーデータを解析し、自動学習することで、最新の在庫の中から顧客ニーズに合ったレコメンドを表示します。

✔ 無償コンサルサービスによる、成果の最大化

価格体系はレコメンドによる売り上げに対してのみ費用が発生する成果報酬型。また、運用中はレコメンドのエキスパートが、設定変更や改善提案など、継続的なサポートを行います。

独自の機械学習技術＋インフラ技術による高速エンジン

追加ラインナップ

aigent recogazo

『レコガゾウ』

アイジェント・レコメンダーのオプション機能で、メールコンテンツを1通1通パーソナライズ

aigent X

『アイジェント・エックス』

企業の高度なマーケティングニーズに対応する、パーソナライゼーション・プラットフォーム

お問い合わせ先： シルバーエッグ・テクノロジー株式会社　Tel: 06-6386-1931(Osaka) / 03-5357-1255(Tokyo) https://www.silveregg.co.jp / E-mail: inquiry@silveregg.co.jp

概要

AIエンジンで高精度なパーソナライゼーションをWEBサービスで提供する弊社では、安定性と費用対効果の追求は全社的な課題です。

弊社はEC2 on ECSを基盤としており、EC2 Auto Scaling, LambdaおよびReserved Instanceの組み合わせでは、「安定性の追求に工数が必要」かつ
「スケーリングさせても月額費用が硬直的」という問題を抱えていました。

試行錯誤の結果、Capacity ProviderとSpot Instanceを組み合わせることでそれらの問題を解決しつつ、安定化、省力化、効率化を実現しました。

お話しすること

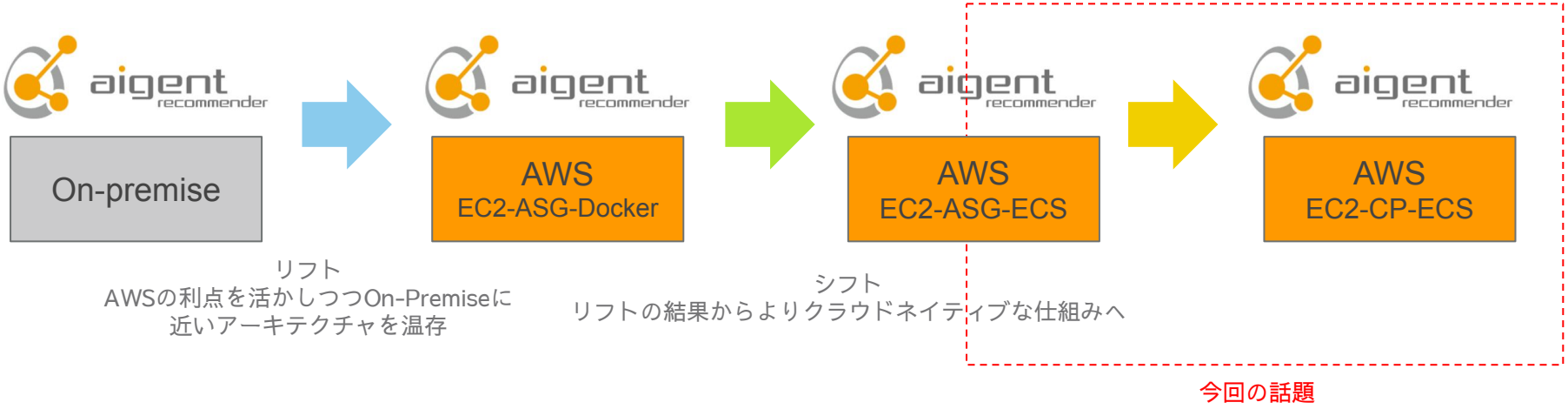
- 実装の勘所
- 背景
- 非Capacity Provider(以下CP)環境の模式図
- ECS内のアプリケーション配置
- 非CP環境の課題
- CPとSpot 活用への道中
- CP環境の模式図
- ある本番環境の動き(24時間)
- 削減効果

実装の勘所

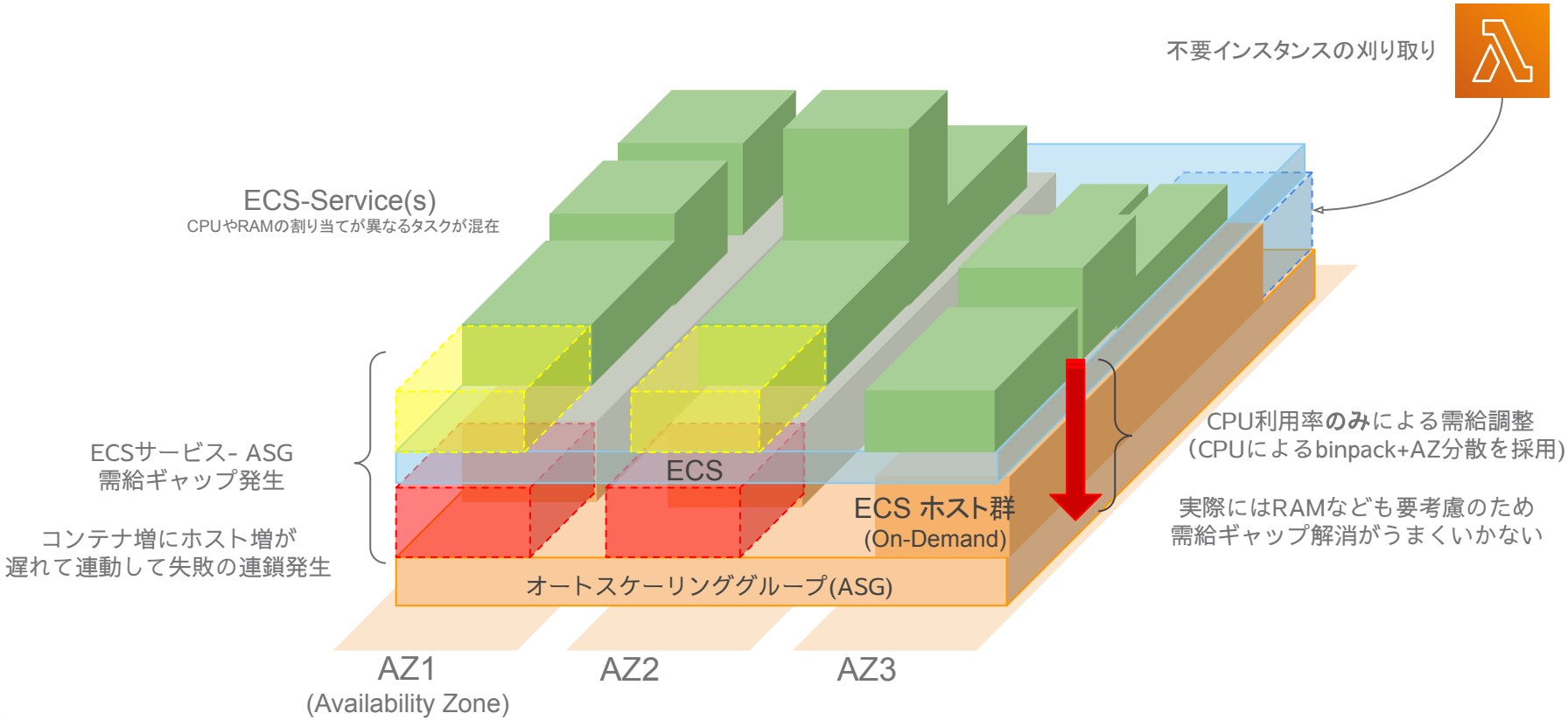
- ECS インスタンスとオートスケーリンググループをSpot対応へ
  - ECS_ENABLE_SPOT_INSTANCE_DRAININGをインスタンスの userdataに設定
  - 配分戦略として「最適化されたキャパシティー」を選択
  - (容認できるなら) CPのターゲットキャパシティー % を 100未満 (例: 97)に設定する
- ECSインスタンスとECS Serviceの設定を整合
  - インスタンスがDrain状態に入りコンテナがSIGTERMを受け取って終了することを確認
  - ECSサービスのサービスデプロイオプションで最大率を 200%にする
- 中断確率の低いSpot インスタンスタイプの使用
  - Spot instance advisor の定期チェックして5種ほどインスタンスタイプを登録し都度入れ替え

背景

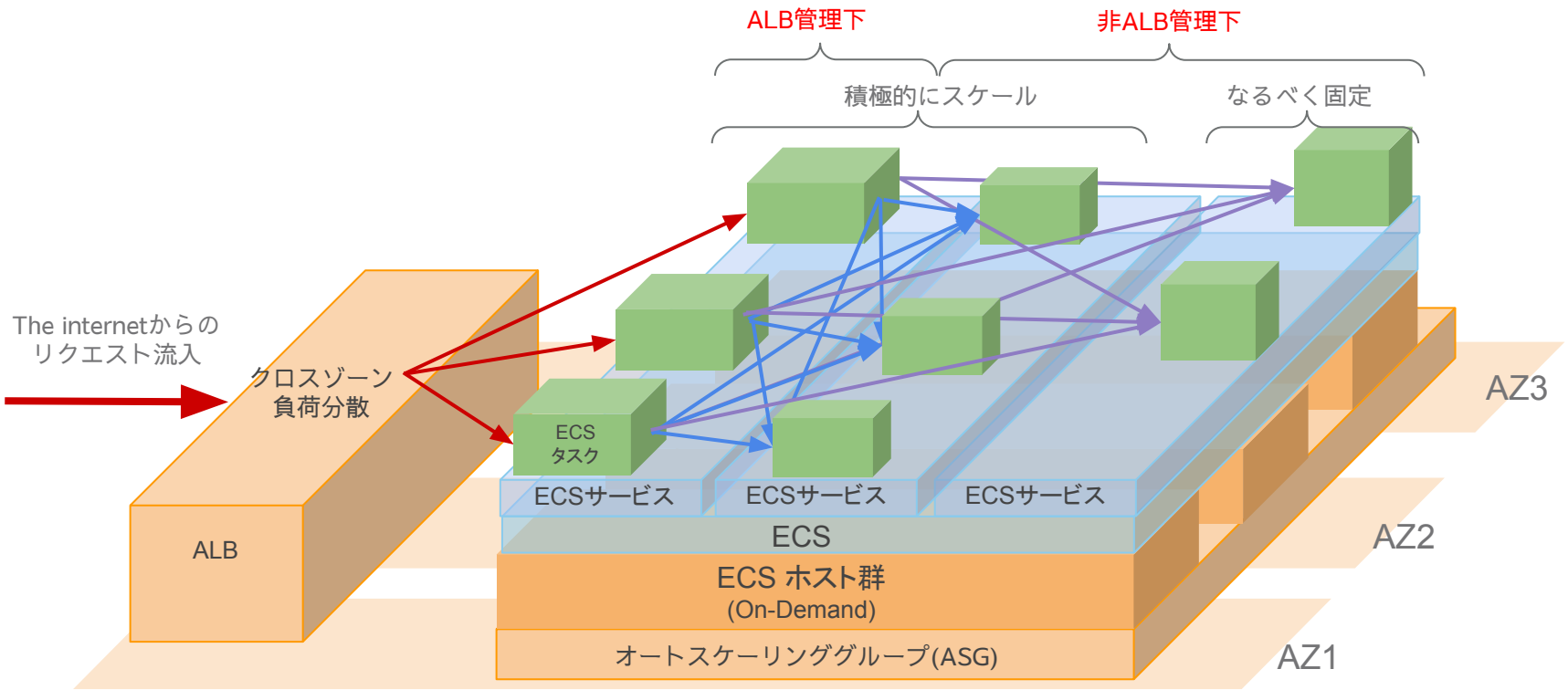
今回の話題はリフト&シフト的な移行を行った後の運用改善のお話です。



非CP環境の模式図



ECS内のアプリケーション配置



非CP環境の課題

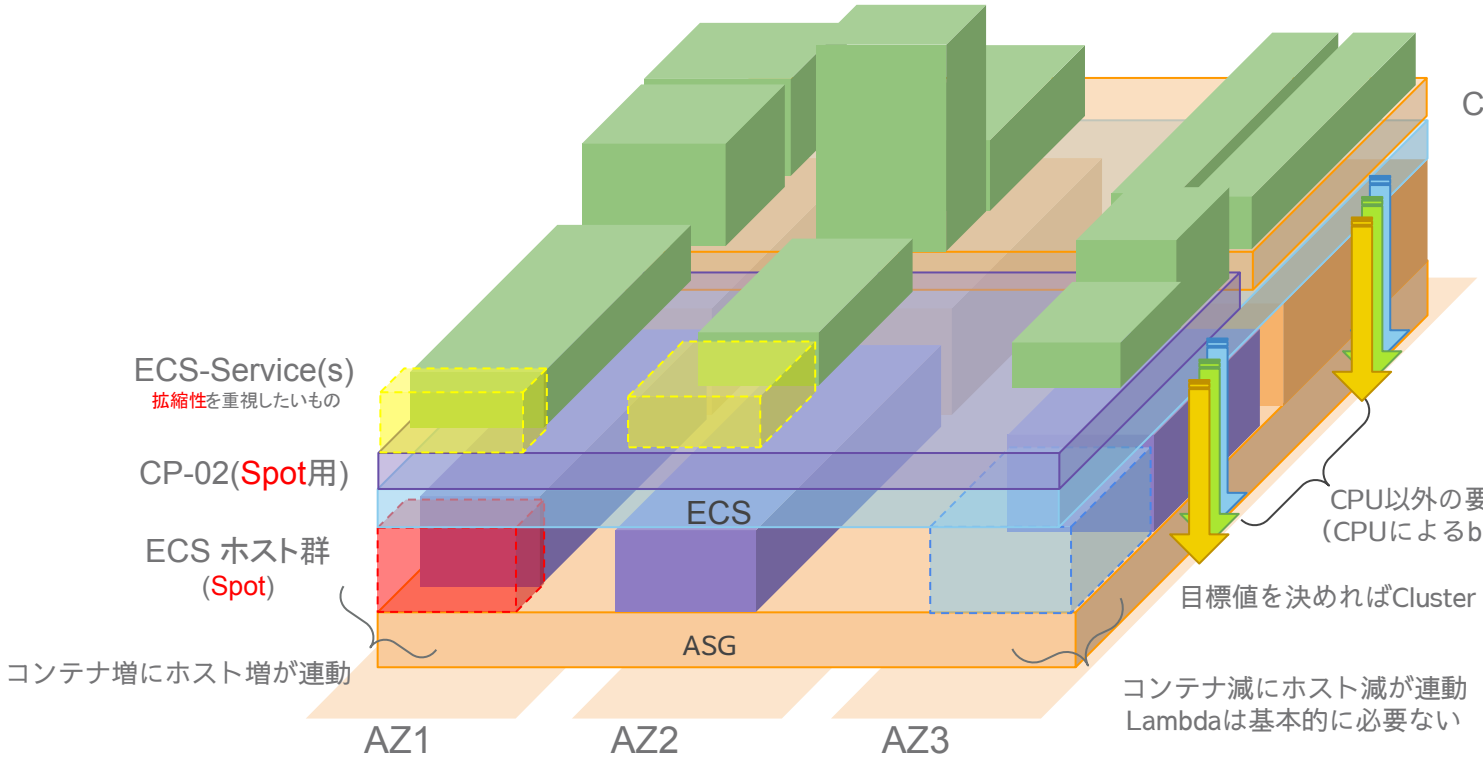
- EC2とECSの2系統のオートスケーリングによる需給ギャップ
  - 規模・性格の異なる多数のクラスタを ECSで運用していた
    - LambdaやASG閾値のカスタムによる実装・維持工数が高かった
- 運用費の硬直性
  - オートスケーリングの課題が未解決のままだった
    - 多数のRIと高余裕率の共依存を続けて障害を減らす他なかった

CPによるオートスケーリング改善後にSpot導入を目指すことにした

CPとSpotへの道中課題

- CPに複数AZないし複数インスタンスタイプを束ねるとスケールが遅かった
  - 1CP - 1AZ で運用で回避 (下記の2020/11のupdateで解消し現在は1CP - 3AZ)
    - 3AZで3CPになり、On-DemandとSpotで分けて 6CPになり、運用に苦労があった
      - Amazon ECS Cluster Auto Scaling now offers more responsive scaling (2020/11/24)で解決
- Spotインスタンス大量死とコンテナ道連れ問題
  - Spot instance advisor の定期チェック
    - 中断確率に基づくインスタンスタイプ入れ替え運用(c, m, r等ファミリーを混ぜて5種ほど登録)
  - 配分戦略として「最適化されたキャパシティー」を選択
    - ASG設定内の「インスタンスの分散」より上記のインスタンスファミリー選択と同時に設定
  - コンテナに搭載しているプログラムをチェックして SIGTERMの処理を適正化
    - ELBに属していないサービスをDrainしたときにコンテナが移動しない不具合を解消
  - ECSサービスのサービスデプロイオプションで最大率を 200%にする
    - コンテナ入れ替え時に必ず並行稼働させてリソースを先に確保させる
  - CP の ターゲットキャパシティー % を 100未満に調整
    - インスタンス数に若干の余裕を持たせる(スパイクリクエスト対策の一部で弊社固有の事情もあり)

CP環境の模式図



お問い合わせ先： シルバーエッグ・テクノロジー株式会社　Tel: 06-6386-1931(Osaka) / 03-5357-1255(Tokyo) https://www.silveregg.co.jp / E-mail: inquiry@silveregg.co.jp

ある本番環境の動き(24時間)

削減効果

- コスト効率(EC2の月額コスト)
  - Spot導入開始前月 vs Spot導入完了月 (RIや Savings Plansの量は変更なし)
    - 24%減
      - 導入完了の定義:
        - On-Demand価格で動くインスタンスが0%になった月
      - 完了月の総リクエスト数は開始前月と比して約110%増
        - リクエスト数とインスタンス数は正の相関関係
        - リクエスト増でコスト減 -> コスト効率が改善
- 可用性(spot起因の障害数)
  - 運用調整中 : 月に数件発生
    - 原因:
      - タスクが正常に終了しない
      - Spot在庫払底(インスタンス起動せず)
  - 運用調整完了後 : 0件
    - 先述したspot中断に関連する対策を施した後は、アプリケーションへの影響は皆無

今回の取り組みについて

本件はチーム一丸となって長期に渡って取り組んだ結果です。

貢献・協力してくれた関係者にあらためて感謝します。

お問い合わせ先： シルバーエッグ・テクノロジー株式会社　Tel: 06-6386-1931(Osaka) / 03-5357-1255(Tokyo) https://www.silveregg.co.jp / E-mail: inquiry@silveregg.co.jp

エンジニア募集のお知らせ

リリースしたばかりの次世代プロダクトを共に作り上げていく仲間を募集中です！
最新技術に触れながらスキルアップできる環境で共に働きませんか？

【募集職種】

- **バックエンドエンジニア**
- **AWSクラウドプラットフォームエンジニア**
- **QAエンジニア**

開発拠点は大阪ですが、東京勤務、リモートワークも可能です！

[>>採用ページ/募集詳細はこちら<<]
[Wantedly/シルバーエッグ・テクノロジー会社情報]

問い合わせ先: 採用担当 小嶋(Linkedin kojima.yukari@silveregg.co.jp)

Appendix : CP migrationのやり方

- 以下3ステップを必要CP分繰り返す(弊社ではOn-Demand用とSpot用の2回)
  - 非CPのASGに併置する形でCP用ASGを作る(ASGのCloudwatch metricsも有効化)
  - 先程つくったASGを指定してCPを作成
  - 目的のクラスタにCPを関連づける
- 各新ASGでインスタンスを立ち上げる
  - 放置するとCPによってインスタンスが削除されるので
    希望する容量(Desired capacity)=最小キャパシティ(Minimum capacity)に固定
- 以下2ステップを移行するサービス分繰り返す
  - サービスのCP戦略として On-DemandかSpotかどちらかのCPを紐付ける
  - 新しいデプロイの強制(Force-New-Deployment)で非CPからCPへタスクを移す
- 新ASGのインスタンス数をCPに委ねる
  - 容量(Desired capacity) > 最小キャパシティ(Minimum capacity)にしてCPに委ねる
- 古いASG(非CP)を削除する