Google Cloud

Google Cloud サーバーレス最新情報

頼兼 孝幸

Google Cloud

アプリケーション モダナイゼーション スペシャリスト

スピーカー自己紹介

頼兼 孝幸

Google Cloud
アプリケーション モダナイゼーション スペシャリスト

担当製品エリア:

- Anthos
- GKE
- **サーバーレス（App Engine、Cloud Run、Cloud Functions など）**
- CI / CD

本セッションでは、サーバーレスの話をします！

本セッションの内容

Google Cloud が提供するサーバーレスでアプリケーション
を実行する環境として、以下の 3 つのプロダクトが存在します。

- **Cloud Run**
- **App Engine（GAE）**
- **Cloud Functions**

この 1 年で多くのアップデートがありました。
本セッションでは、主に直近のアップデートを振り返って、
どのようにプロダクトのユースケースが広がったかを説明します。

各プロダクトの概要

各プロダクトの特徴と、主なユースケース

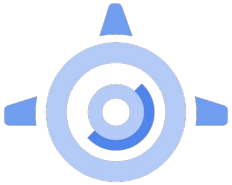
Cloud Run

コンテナ実行環境

マイクロ サービス向き
（REST API、gRPC）

イベント駆動も可

App Engine



アプリケーション実行環境

静的、動的コンテンツを
利用した **WebApp**

Cloud Functions

関数の実行環境

イベント駆動向き

Firebase との連携

サーバーレス アプリと連携する周辺プロダクト

Eventarc

Cloud Storage、Pub/Sub、
Cloud Audit Logs などの
イベントをトリガー

cloudevents.io の仕様に
準拠したデータ形式

Cloud Scheduler

フルマネージド
cron ジョブ スケジューラ

HTTP、Pub/Sub、
App Engine から
ターゲットを選択

Pub/Sub

NoOps でスケーラブルな
**メッセージング、または
キューシステム**

pull / push モードが選択可

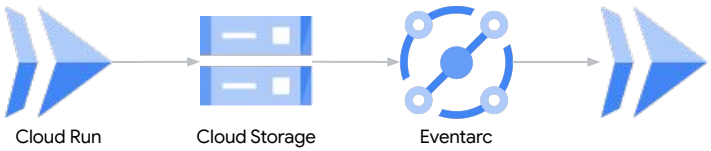
少なくとも 1 回の
メッセージ配信保証
（At-least-once delivery）

Workflows

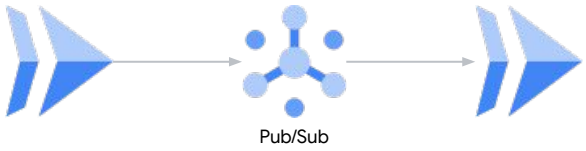
定義した順序で
**サービスを実行する
オーケストレーション**
プラットフォーム

マイクロ サービスの統合やビ
ジネスプロセスの自動化など

サーバーレス アプリと連携するユースケース



イベント駆動



非同期データ連携（Pull / Push どちらでも可）

定期的な呼び出し

複数サービスのオーケストレーション

サーバーレスの最新情報
（2021 年下半期以降を対象）

App Engine のランタイムについて

App Engine のランタイム

環境	スタンダード 第 1 世代	スタンダード 第 2 世代	フレキシブル
開発言語	Python 2.7 Java 8 PHP 5.5 Go 1.11	Python 3.7,3.8,3.9 Java 11 Node.js 10,12,14,16 PHP 7.2,7.3,7.4 Ruby 2.5,2.6,2.7 Go 1.12+	Node.js, Ruby, Java, Python, Go, PHP, .NET カスタムコンテナ イメージ
実行環境	サンドボックス		仮想マシン上の Docker コンテナ
Google Cloud 機能の利用	GAE 専用 の API Google Cloud の API	Google Cloud の API を利用 ※	
サードパーティ バイナリのインストール	不可	可	

※ Python 3、Java 11、Go 1.12+ では、第 1 世代からのマイグレーション手段の提供を目的とし、一部 GAE 専用の API が利用できるようになっています

Cloud Run の新しい機能について

App Engine と Cloud Run の比較

App Engine 第 2 世代を利用すること自体に問題はなし

一方、Cloud Run の方が縛りも少なく、コンテナのスケール性能は高い

- 言語やライブラリの制約なし
- WebSocket や gRPC サポート
- App Engine フレキシブル環境よりもスケール性能が高い
- マルチリージョンのサービス展開
- イベント駆動
- コンテナ化や CI / CD サポート（Buildpacks、ソースコードからのデプロイ）

新しい機能も頻繁にリリースされている

Preview

第 2 世代の実行環境

性能の高速化

- CPU パフォーマンスの高速化
- ネットワーク パフォーマンスを高速化

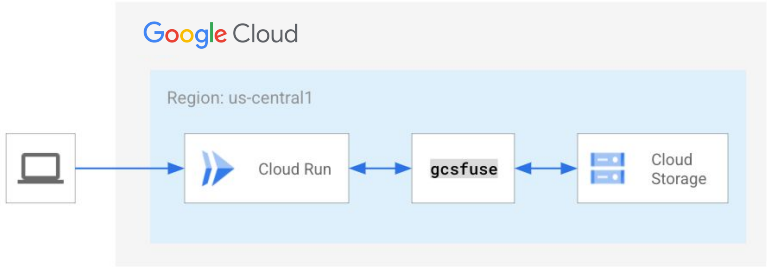
ユースケースの拡大

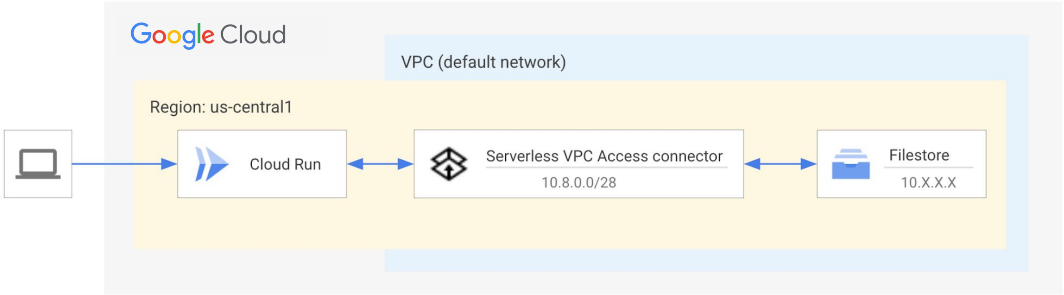
- すべてのシステムコール、名前空間、cgroup のサポートを含む、Linux との完全な互換性
- ネットワーク ファイル システムのサポート

※ Preview 段階では、第 1 世代よりもコールド スタート時間が少し長くなる点に注意

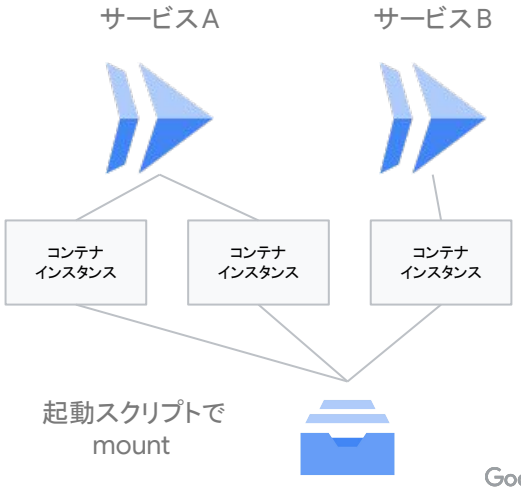
第 2 世代の実行環境
- NFS や Cloud Storage FUSE の利用

Cloud Filestore や、Cloud Storage FUSE を利用し、複数のコンテナやサービス間のデータを共有





VPC Access Connector 経由で
VPC 内の Filestore へアクセス



第 2 世代の実行環境のよくある質問

第 1 世代は今後使わない方が良いのか

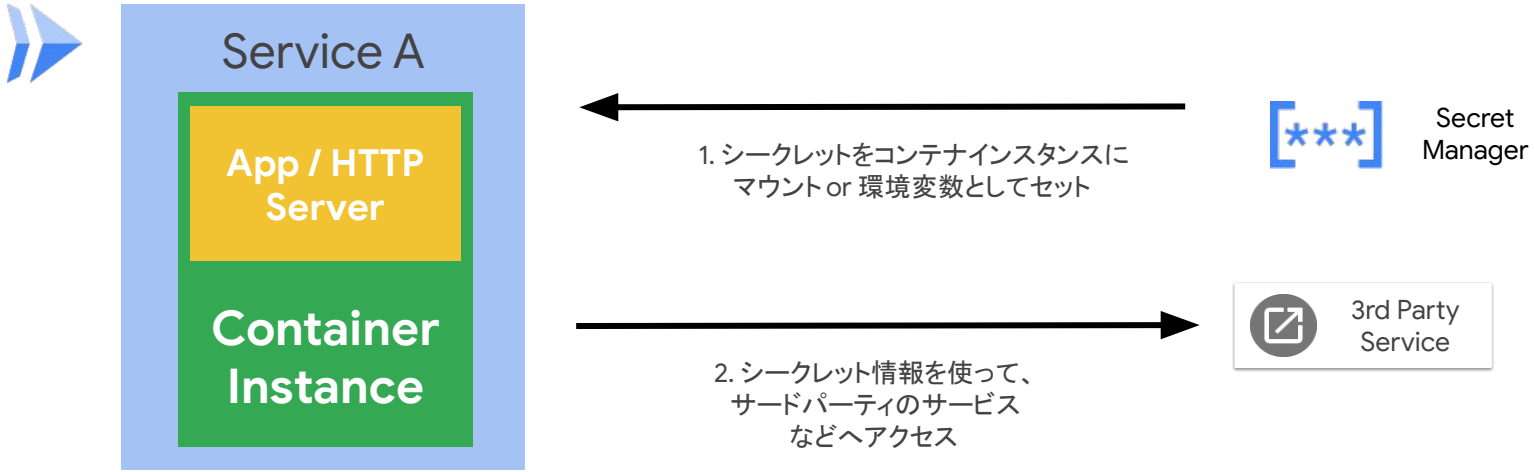
- スパイク アクセスに応じた高速なスケールアウトや、0 から 1 にスケールする頻度が多い場合などは、第 1 世代の方が向いている
- 第 2 世代が GA になるまでに、このギャップは小さくなる
- GA 後も、しばらくはユースケースに応じて使い分けることになる

実行環境についての詳細はこちらを参照
https://cloud.google.com/run/docs/about-execution-environments

GA

Secret Manager の統合

API キーやパスワードなどの機密情報を、安全に管理することが可能

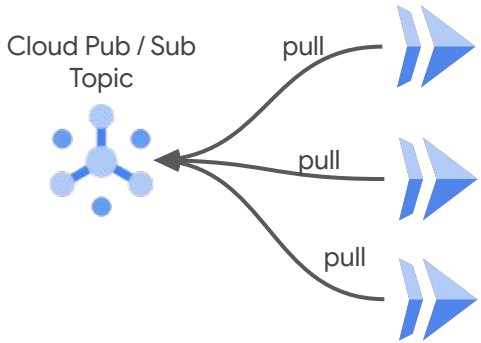


GA

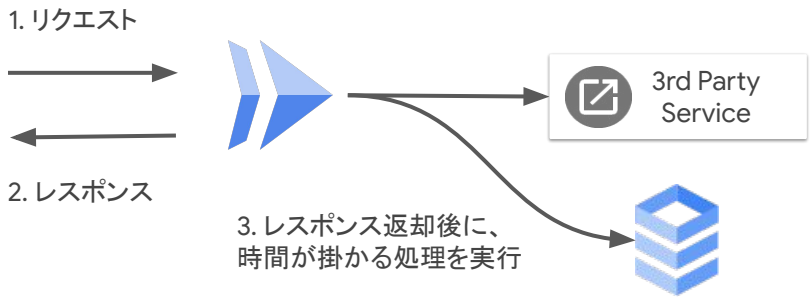
インスタンス時間に応じた CPU Allocation と課金（Always on CPU）

HTTP リクエストの有無に関わらず、常に CPU が割り当てられ、コンテナ インスタンスが存在している時間に対して課金が行われる（バックグラウンド タスクや非同期処理などに最適）

Cloud Pub / Sub への Pull Subscribe



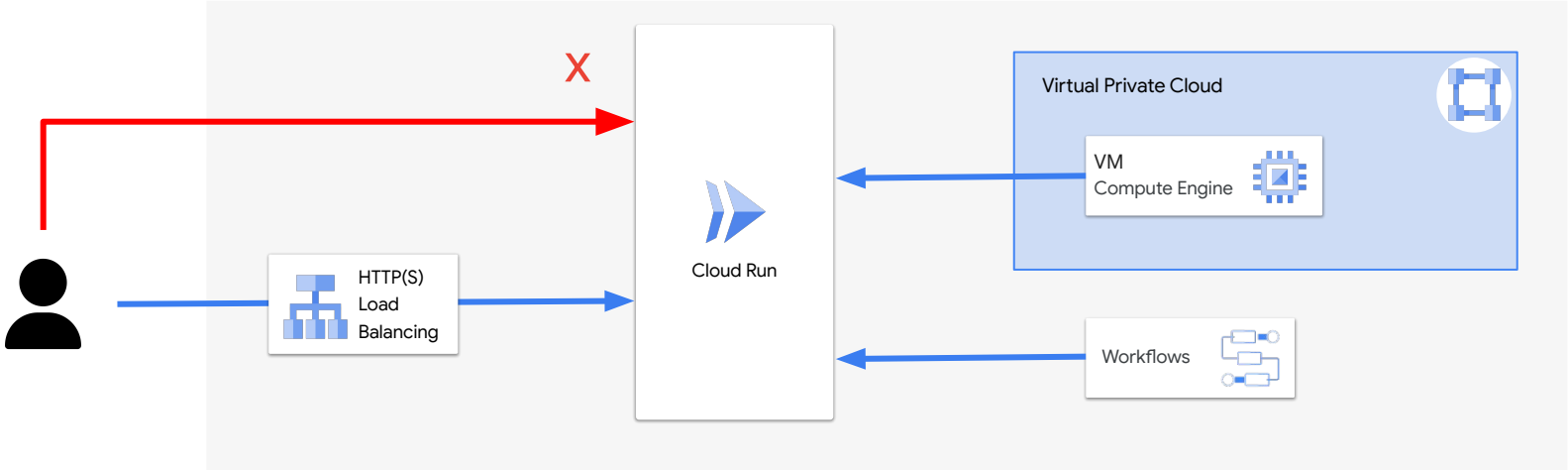
レスポンス後にタスクを実行する



GA

内部 Ingress に限定されたサービスを Workflows から呼び出す

内部トラフィックのみ、または、内部トラフィックと Cloud Load Balancing からのトラフィックを許可した Ingress 設定の場合に、Workflows から Cloud Run の URL（run.app）を呼び出せる



Preview

CPU とメモリの適用範囲拡張

CPU

- 1 CPU 未満の割り当てが可能に。適用範囲は 0.08 から 1 まで 0.01 単位で指定が可能
  - CPU 1 未満を割り当てた場合、コンテナあたりの 最大リクエスト数は 1 固定
  - 開発用途や、単発起動などの軽量な処理でコストを最適化

メモリ

- 16 GiB のメモリが割り当て可能に
- 割り当てメモリ相当の CPU は指定が必要（逆も然り）
  - 16 GiB メモリ割り当ての場合、最低 4 つの vCPU が必要、など

新しい Cloud Functions

Preview

Cloud Functions 第 2 世代

Cloud Functions 第 2 世代の環境が選択可能に

Cloud Run の実行基盤で Cloud Functions が動く

東京リージョンも利用可能


```
gcloud beta functions deploy [FUNCTION_NAME] \
--gen2 ...
```

https://cloud.google.com/functions/docs/2nd-gen/overview

今までの Cloud Functions（第 1 世代）の課題

- 1 実行環境に、同時実行性がない
  - 呼び出し毎に、別の実行環境が起動される
  - 呼び出し毎の料金設定（最初の 200 万回無料。以降 100 万回あたり $0.40）
- 長時間の処理を実行する際の制限
  - 第 1 世代では、最長 10 分でタイムアウト
- 柔軟なトラフィック管理は不可
  - 現在のバージョンと、次のバージョンでトラフィック分割する、といった柔軟なトラフィック制御などは行えない

第 2 世代で、より高機能に

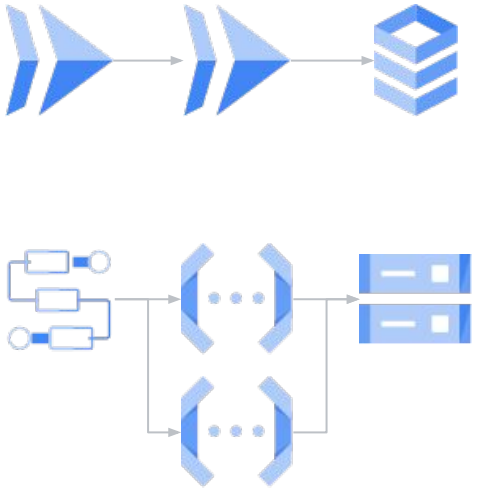
- Cloud Run と Eventarc を基盤としたアーキテクチャ
  - Cloud Run の特性を利用し、**同時実行性を実現し、コストを最適化**
  - Eventarc を統合したイベント駆動により、CloudEvents 標準の一貫したデータ形式
- 長時間の処理にも対応
  - **最長 60 分**のタイムアウト設定
- Cloud Run と同様、リビジョン毎のトラフィック管理を提供
  - 複数の関数のリビジョンを利用可能
  - **リビジョン間のトラフィック分割** や、**以前のバージョンへのロールバック** など可

Cloud Run と Cloud Functions の使い分け

Cloud Functions の単位が「関数」なのは変わらない

複数サービスの連携 など、**一定規模のシステム** を
サーバーレスで構築するなら Cloud Run が良い

Workflows を併用した、一連の **ワークフロー作成** や、
3rd パーティとの連携、Google Cloud **プロダクト間をつなぐような処理** は、関数単位で閉じやすく、
Cloud Functions だと管理単位も小さくなる



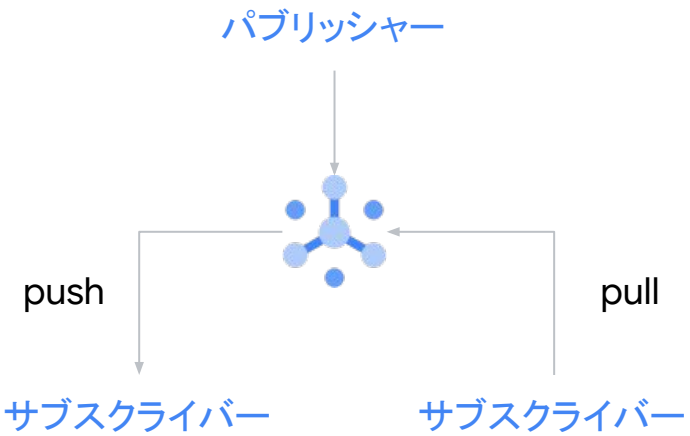
Pub/Sub のメッセージ配信保証

Pub/Sub の配信保証と順序指定

Pub/Sub は少なくとも 1 回はメッセージ配信する（At least once delivery）ことを保証している

2020 年には、順序指定（Ordering delivery）を保証する機能が追加

メッセージの重複については、必要に応じて
後続タスクでチェックする必要あり



Preview

Pub/Sub の 1 回限りの配信保証（Exactly once delivery）

1 回限りの配信が有効になっている、サブスクリプションに対して、
配信が重複しないことを保証

順序指定配信のサポートはない（ 2022 年 3 月時点）

通常のサブスクリプションよりも、パブリッシュとサブスクライブ間の
レイテンシが大幅に高くなる点に注意が必要

まとめ

Cloud Run を中心に、幅広いユースケースに対応

Cloud Run 第 2 世代のように、**Cloud Run 自身のユースケースが広がるだけでなく、その機能を Cloud Functions にも活用**し、関数としてのサーバーレス製品のユースケースもより広がった

Workflows や Eventarc、Pub/Sub などの製品を組み合わせ、**ワークフローやイベント駆動、非同期処理の連携**など、サーバーレスだけで実現できることが非常に増えた

Thank you.

Google Cloud