

Ragate LT会

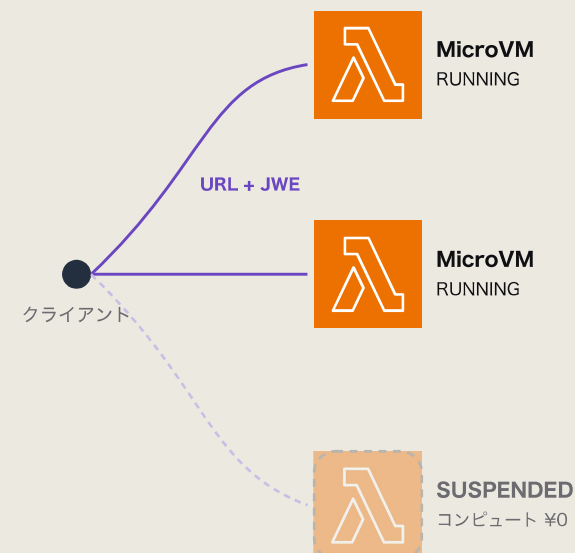
サーバーレスが “記憶”を持った日

Lambda MicroVMs — 状態を持つ新しいサーバーレス。
聞いたあと、Lambda と MicroVMs の境目を判断できる。

久保 / Ragate ・ 2026年6月

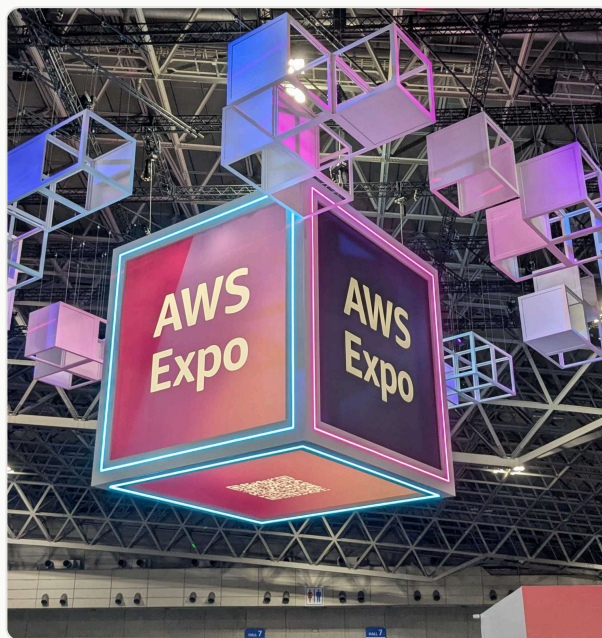
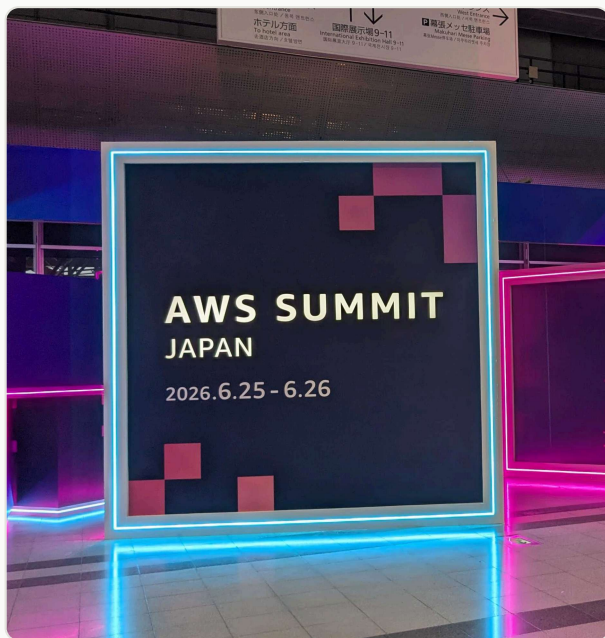
Lambda MicroVMs

Firecracker ・ suspend ⇄ resume



昨日までのサーバーレスは、
呼ばれるたびに“記憶喪失”
していた。

この話、AWS Summit 2026 で見つけた



会場は AIエージェント・Claude・Bedrock 一色。その“実行基盤”が Lambda MicroVMs — 現地の一次情報です。

今日の3つ

1 サーバーレスに「記憶」が付いた

2 呼び出しが逆転する（フリースは自前）

3 コストは「使い方」で激変する

なぜ“記憶”を持てなかったのか

これまで = 使い捨て

呼ばれた時だけ起動 → 終わると破棄

毎回 別の箱に当たる = ステートレス

記憶は外部DBに毎回 出し入れ

MicroVMs = 鍵付きの自分の部屋

VMを 起動したまま生かす (最大8h)

同じVMを名指し = 箱が固定

“記憶”がVMに残る

たとえ：使う時だけ借りる会議室 → 荷物を置いて出られる自分の部屋へ。

“ブート”ではなく“復元”



通常は毎回「**ブート＋初期化**」をやり直す。MicroVMs は **初期化済みを“復元”するだけ** —— だから速い。

※ サスペンド中は **コンピュータ課金0**。スナップショット保管のみ微量に課金。

呼び出しモデルが逆転する

① 通常 Lambda



② MicroVMs



API Gateway は不要に。代わりに「どのVMが誰のものか＝**フリート管理を自前**」で持つ。

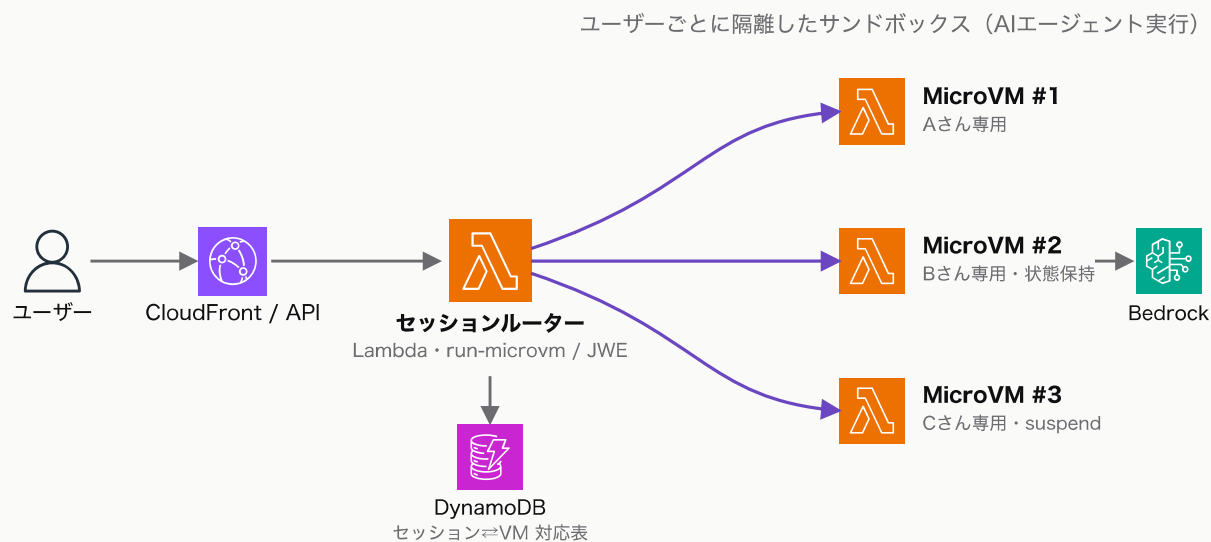
実際に起動すると、URLが返る

```
$ aws lambda-microvms run-microvm --image-identifier my-image
```

```
{  
  "microvmId": "mvm-0123abcd...ef01",  
  "state": "PENDING",  
  "endpoint": "mvm-0123abcd...on.aws"  
}
```

数秒後に RUNNING。あとはこの URL を JWE 付きで“名指し”で叩くだけ。

MicroVM をフル活用した構成



セッション⇔VMを1:1で結び、状態をVMに保持。**terminate** はセッション終了で自動。

費用は「使い方」で激変する

8時間“接続”しても、実稼働が1時間なら

\$0.3

約47円 / 8h — Fargate 常時起動 (\$0.79) の **4割以下**

※ 概算：ARM / us-east-1 / 2vCPU+4GB / \$1≒155円

逆に、割高になる使い方

こういう使い方は	割高・要再考
8時間ずっと高負荷で稼働	同条件 Fargate の 約2.5倍 (8h≈\$2.0)
状態を持たない短時間処理	通常 Lambda で十分・安い
大量の常時稼働ワークロード	ECS / EC2 が安い

「強隔離 × 状態保持 × アイドル0」が要らないなら、**割高なだけ**。安いのは“アイドルが長い”時だけ。

※ 概算：ARM / us-east-1 / 2vCPU+4GB / \$1≒155円

ユースケース別の構成（AWS）

① 生成AIエージェントの PoC



② 対話型 AI アシスタント

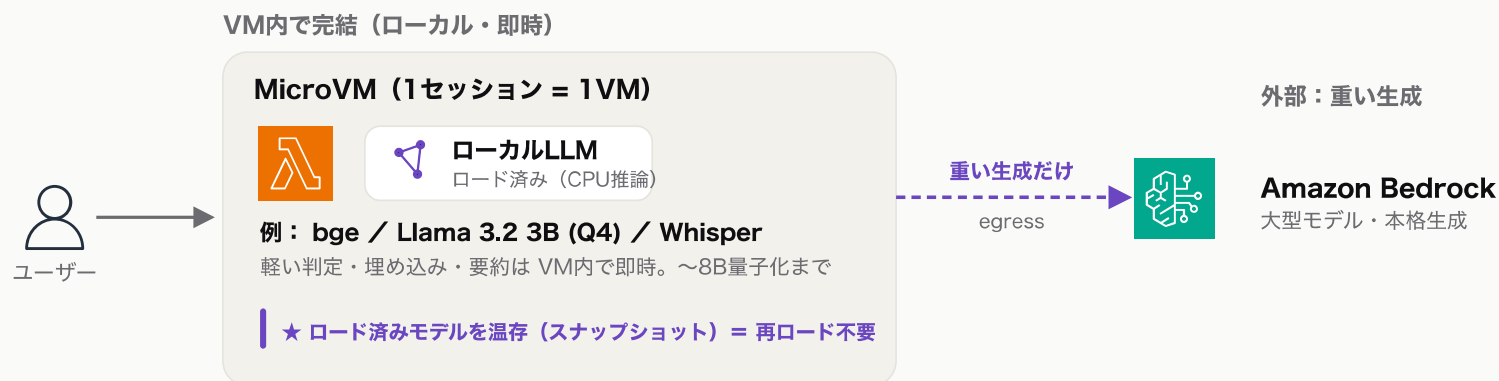


③ 顧客データを扱う分析エージェント



いずれも **1セッション = 1 MicroVM**。状態はVMが持つ。

VM内にローカルLLMも載る（CPU推論）



GPUは無いので**軽量・量子化モデルはVM内（CPU）**、**重い生成だけ Bedrock**。モデルはスナップショットでロード済みのまま起動。