Cloud connect the world as a Glue

AWS Dev Day 2017 Track 2

Masahiro Nagano @kazeburo

Me

- Masahiro Nagano / 長野雅広
- @kazeburo
- 

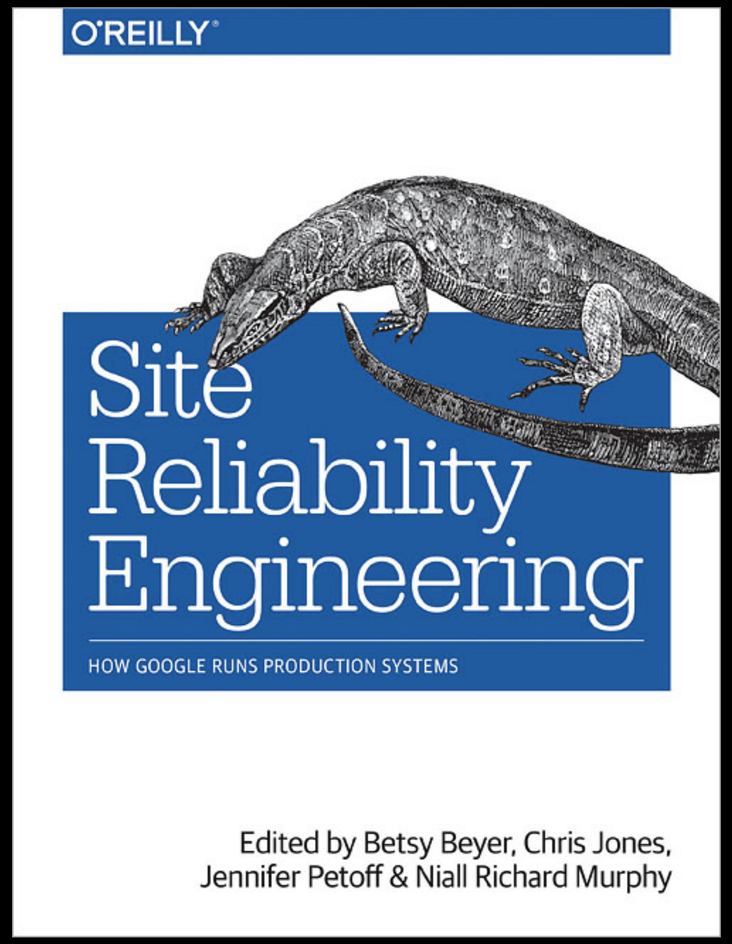
  Mercari, Inc
  Principal Engineer
  Site Reliability Engineering (SRE) Team
- BASE, Inc Technical Advisor

SRE Team の紹介

SRE

- Site Reliability Engineering の略
- Google の運用チームを率いる Ben Treynor が提唱
- Google の様々なプロダクト・サービスを横断して、ソフトウェアエンジニアリングよりサイト/サービスの信頼性を向上させる Software Engineering/Teamとその実践 = Google SRE

Google SRE



- ソフトウェアエンジニア(SWE)として採用
- 運用の業務を50%以下に抑える
  - 50%はオペレーションの自動化、ソフトウェアの信頼性向上にあてる
- エラーバジェットという考え方
  - SREとSWEのSLAを取り決め、利害を一致させる

http://landing.google.com/sre/book.html


Site Reliability Engineering
mercari
SRE
Our mission: To boldly go where no one has gone before.

Mercari SRE

- いつでも快適かつ安全に利用できる「信頼性の高い」サービスの実現
  - 「新規サービスの開発以外のエンジニアリングは全部やる」
- 2015/11 「インフラチーム」からSREへ
  - 「インフラ」よりもサービス指向
- 現在メンバーは「6人」絶賛募集中

Mercari SRE の業務範囲

Operations

OnCall (障害対応)

Automation

Software Eng.

スケーラビリティ・可用性改善

ミドルウェア構築

アプリケーションの設計レビュー

ログ収集・分析基盤の構築、運用

サーバプロビジョニング・デプロイの整備

セキュリティ／不正利用検出

基盤構築

Agenda

- メルカリとは / 世界3拠点での開発運用体制について
- メルカリのアーキテクチャ / クラウドの利用
- 距離を超える、世界を繋ぐシステム

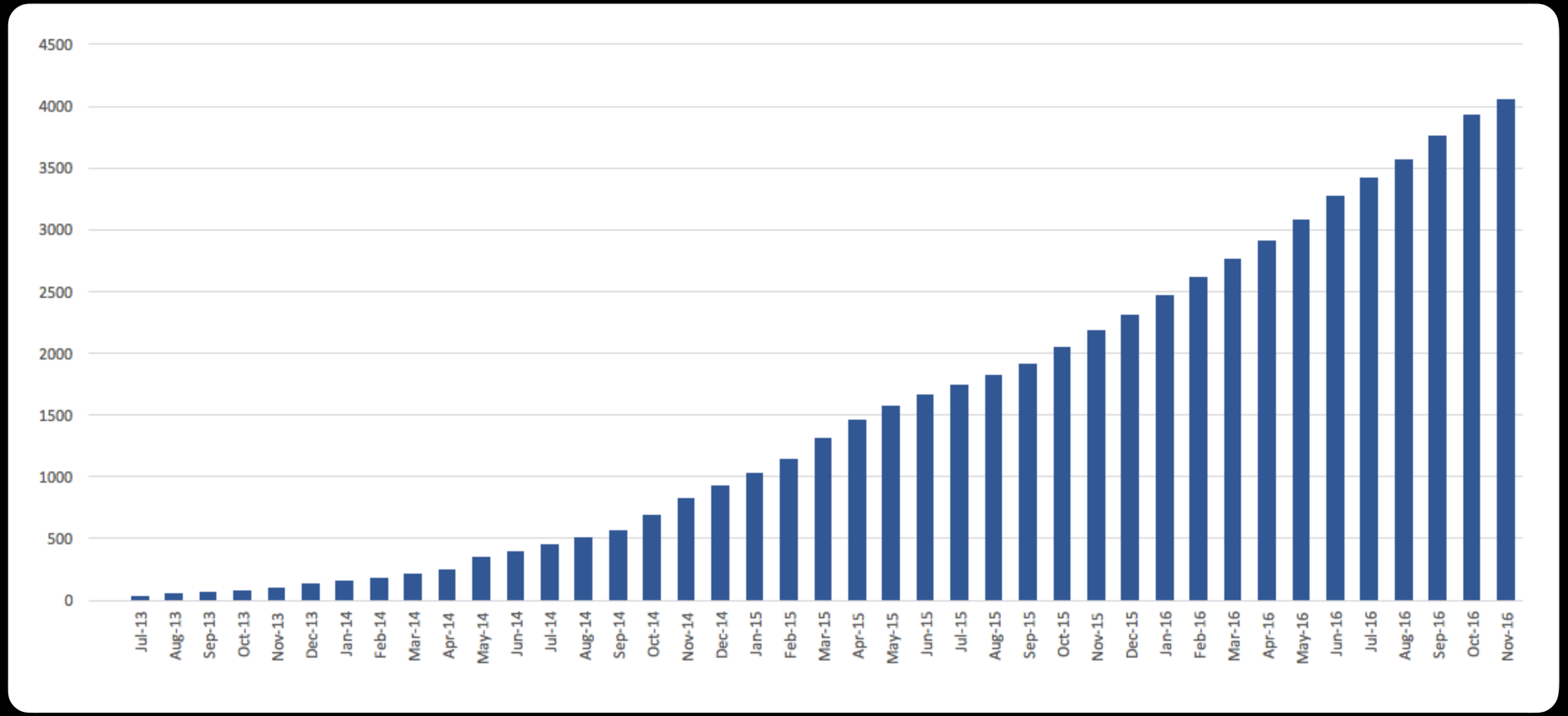
Mercari

- 国内最大級のフリマアプリ
- 3分で簡単に出品
- 安心安全な決済

Mercari KPI

ダウンロード数	6500万DL(JP+US)
出品数	1日100万品以上
GMV(総取引額)	月間100億円以上

ダウンロード数推移(JP)

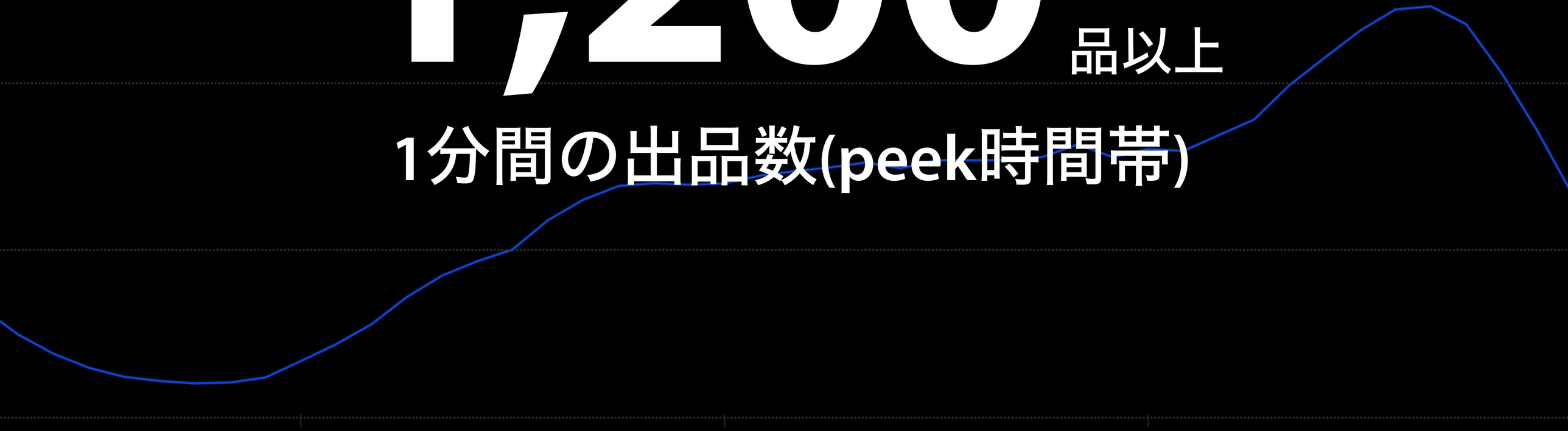


JP ダウンロード数 4000万 (2016/11)

日本最大のフリマアプリ

1,200 品以上

1分間の出品数(peek時間帯)



出品からすぐに売れる

24

時間以内

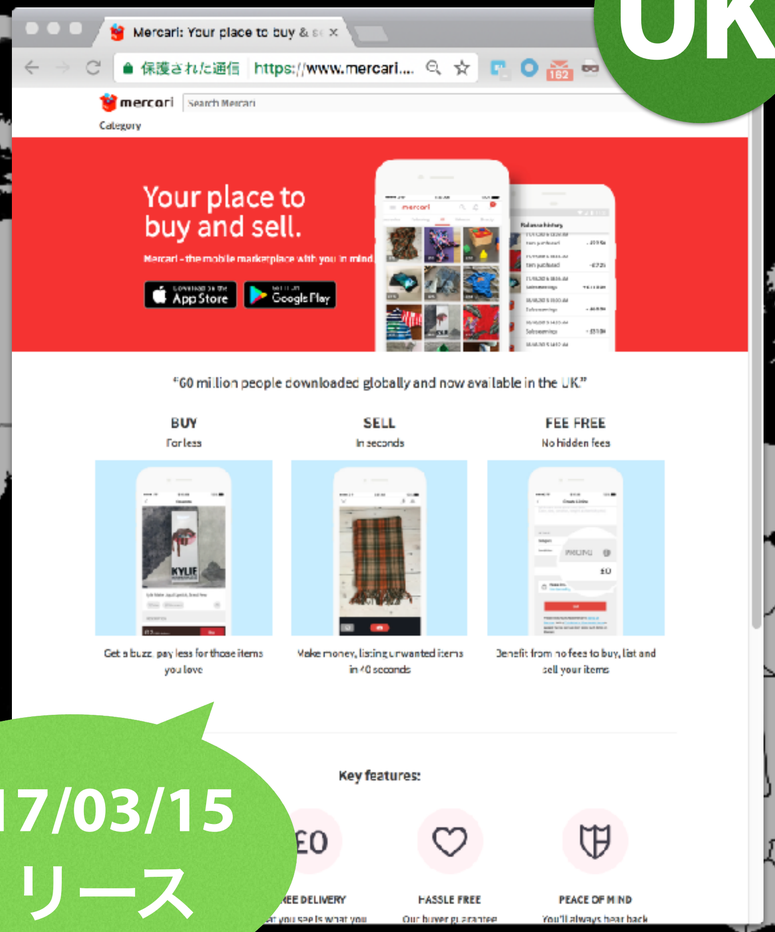
売れた商品の約50%が
24時間以内に取引成立

Global Service

JP

US

UK



2016/08
US AppStore
3位

2017/03/15
リリース

Global Development Team

San Francisco/London にオフィス

現地採用、出向者、長期出張合わせて

エンジニアが数名から数十名

Global Development Team

- Tokyo
  - 開発の中心。JPに加えて全てのregionの開発
- San Fransisco
  - サービスのローカライズ
  - 言語だけではなく、文化や習慣のローカライズ
- London
  - サービス立ち上げフェーズ / 現地の法令などに合わせたローカライズ

Global Development の難しさ

Global Development の進め方(1)

- クラウドを活用してコミュニケーションを図る
  - 太平洋・大西洋をまたいだPull Reqeuestレビュー
  - Slack
  - Video Conference
  - リモートペアプロ(スクリーン共有)

Global Development の進め方(2)

- 自立したチームとして課題解決する
  - プロジェクトマネージャと、クライアントからサーバサイドエンジニアまでフルスタックのプロダクトチームを現地で結成
  - チーム丸ごと出張
  - iOS/AndroidはRegionによってforkあるいは、branchを分け、互いの影響を減らす

SREのケース

- 6人のうち、1人が長期US出張中
  - 現地開発のサービスのオペレーションの把握
- 週1でUSとのSync MTG
  - 朝9時(PDT 17:00) に自宅にて Video Conference
  - UKとは案件ベースで夕方にMTG
- OnCall 当番は朝9時から自宅待機。USからの作業依頼にあたる

Mercari Architecture

Infrastructure

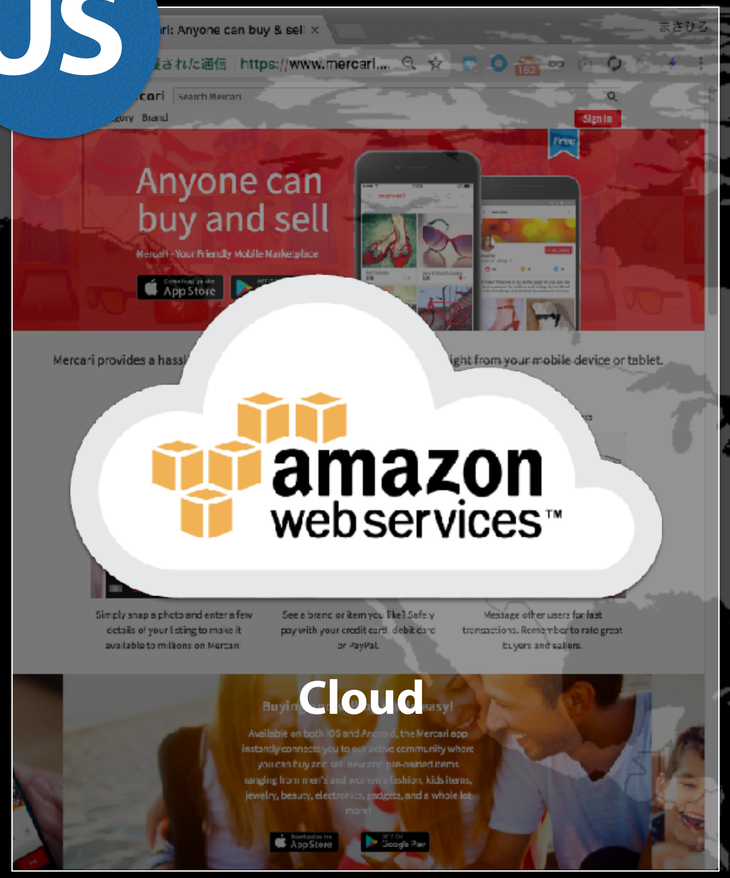
JP

SAKURA internet

石狩DC

専用サーバ

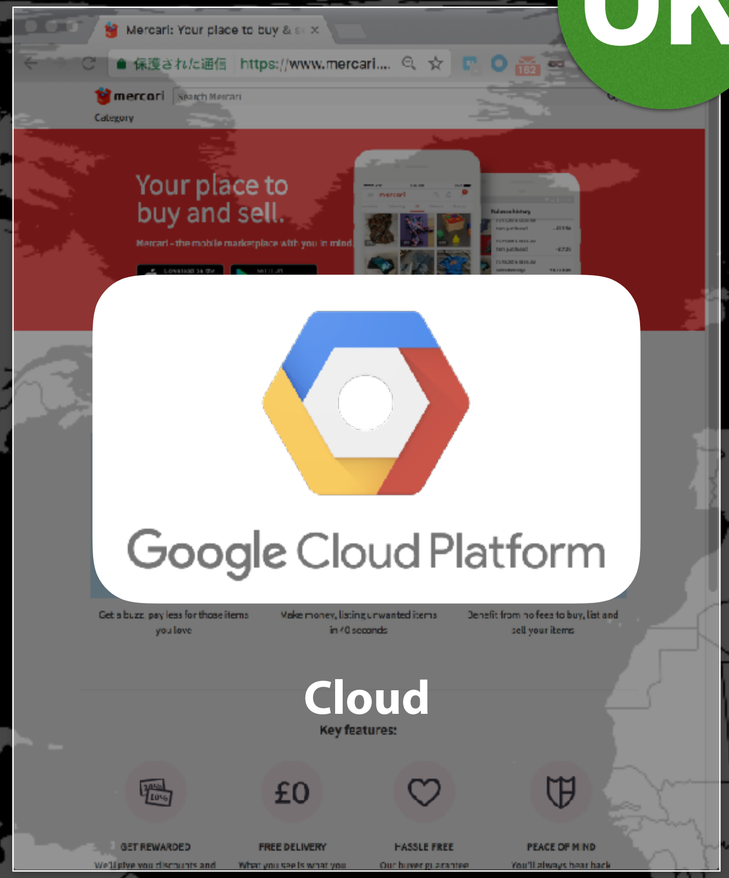
US



amazon web services™

Cloud

UK



Google Cloud Platform

Cloud

Hybrid & Multi Cloud

Infrastructure history (1)

- 2013/07 JP リリース
  - さくらインターネットのVPS 1台にWebもDBもすべて載せた
    - インフラストラクチャ専任者いない中で、身近な技術を選択
  - リリース後2ヶ月でさくらクラウド、専用サーバに移行してきた

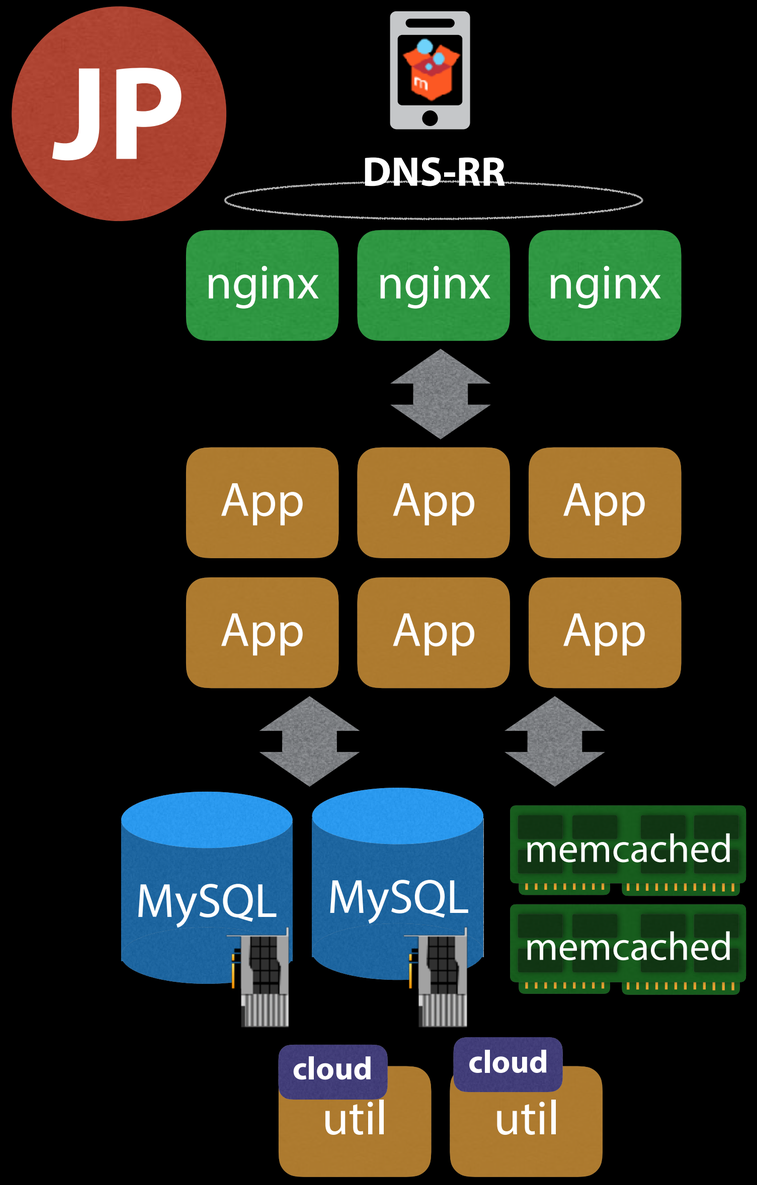
Infrastructure history (2)

- 2014/09 US リリース
  - AWS (Oregon) にてサービス構築
    - JPリリース当初に比べてエンジニアが増え、AWS経験者も多くなった
    - それでもまだインフラストラクチャ専任者は少なく、AWSのマネージドサービスを多く利用してサービスを構築
  - US国内の専用サーバ利用も検討したが、USのスケールは予想しづらく、クラウドの柔軟さを日本よりも重要視した

Infrastructure history (3)

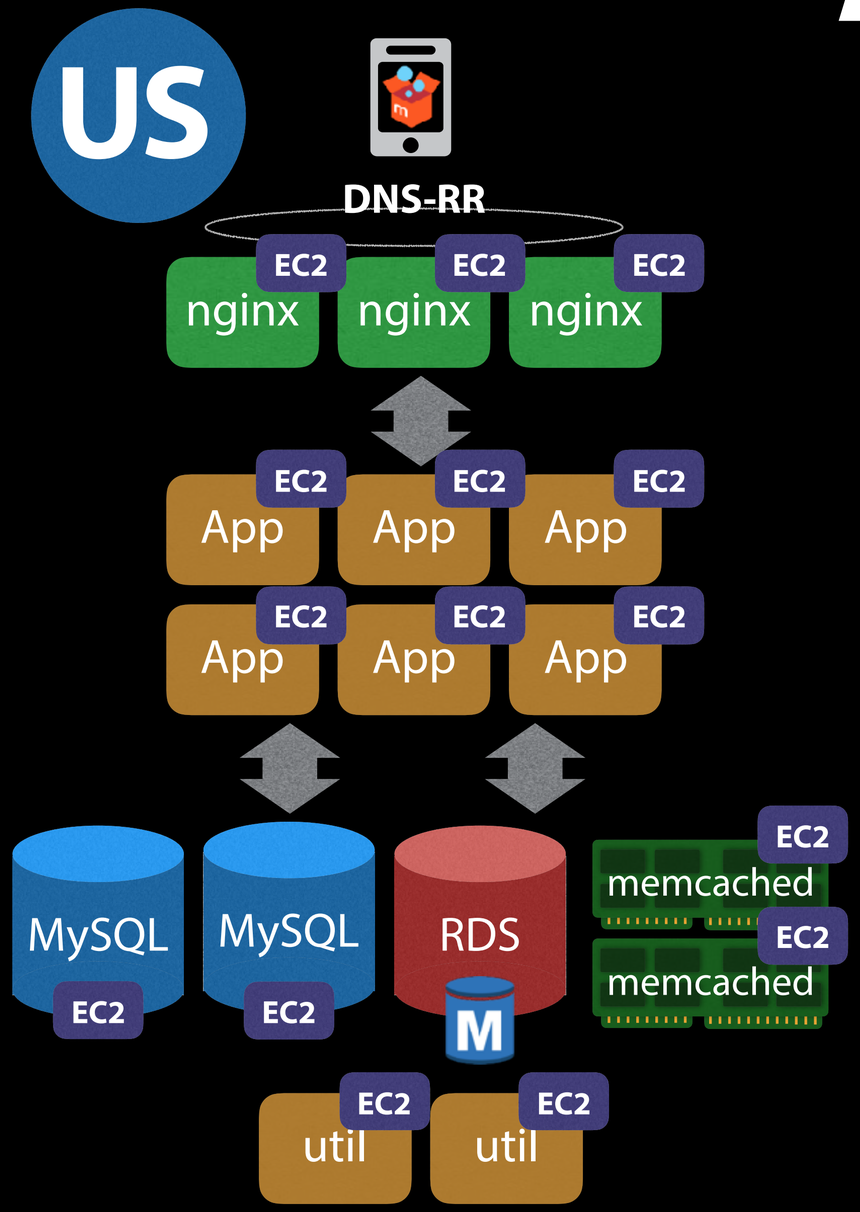
- (2015/02 kazeburo 入社)
- 2015/11 SREチーム発足
  - さくらインターネットとAWSのハイブリッドなインフラストラクチャの上のアーキテクチャを進化させ、信頼性とスケーラビリティの向上
- 2017/03 UK リリース
  - 新しい技術的チャレンジとしてGCPを選択

Architecture



- 三層+αなシンプルなアーキテクチャ
  - Reverse Proxy = nginx
    Application = Apache+mod_php
    Database = MySQL
    Cache = memcached
    Search = Solr
- 多くを物理サーバにて構成
  - スケールアップもスケールアウトも行うDiagonal Scale指向
  - DatabaseにはioMemoryやNVMeを搭載したサーバを採用

Architecture



JP のアーキテクチャを基本踏襲

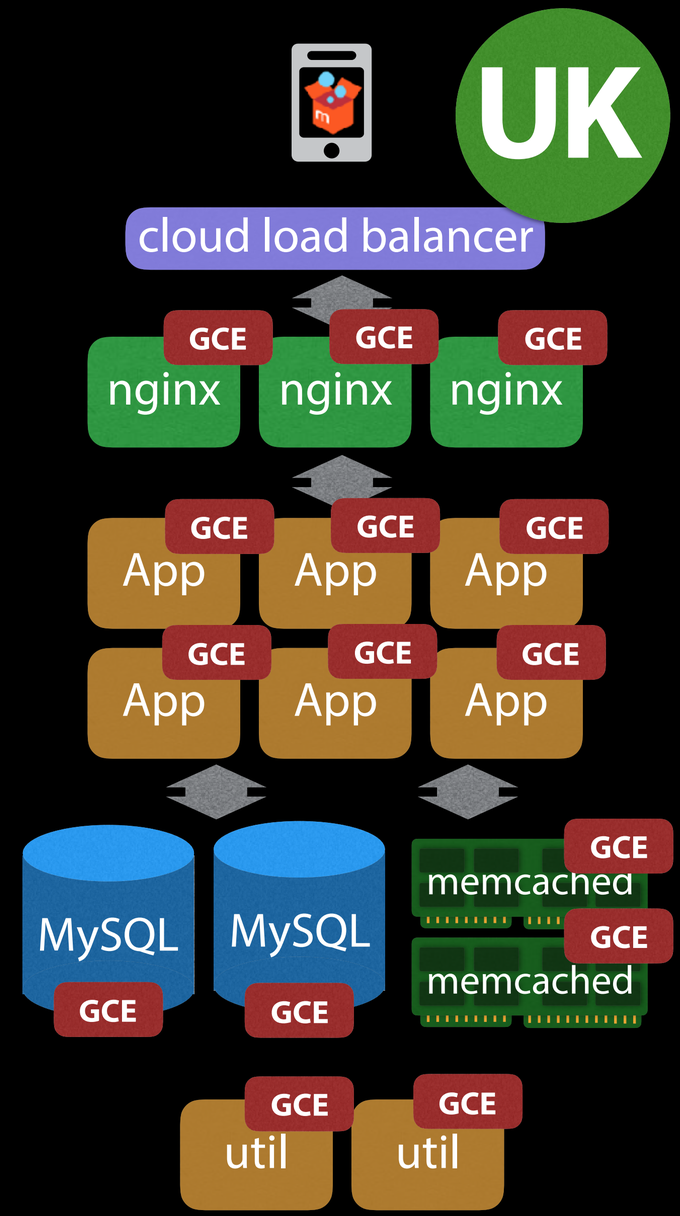
EC2/GCE (サーバ) を中心した構成

・

US独自のサービスや

小規模〜中規模DBには RDS

UKではCloud Load Balancerを利用



サーバ中心の Architecture

- メンテンスビリティ・スケーラビリティ戦略の共通化
  - 少人数での運用
  - Ansible Playbook 再利用
  - スケールが先行しているJPで実績ある構成
    - US での App Store ランキング3位のトラフィックも問題なく運用
- EC2のIaaSとしてのパフォーマンス、信頼性はかなり向上している

Mercari Architecture まとめ

- 3つのRegionで採用するインフラストラクチャが異なる
  - JP/US/UK はサーバを中心としたArchitectureを採用
  - AWSでもクラウドらしい設計はせず、規模で先行するJPに合わせることで、運用の共通化と省力化
- メルカリではクラウドを積極的に使っていない？
  - JP/US/UK 共通のインフラストラクチャで利用しています

Mercari Global Infrastructure

Global Infrastructure

- Mercari JP/US/UK のインフラストラクチャは独立している
  - データをサービスを行う域内に留める必要性
- いくつかのクラウドサービスを共通して利用
  - 海外でのアクセス改善
  - クラウドの高いスケーラビリティ・信頼性によりサービスの可用性を保つ

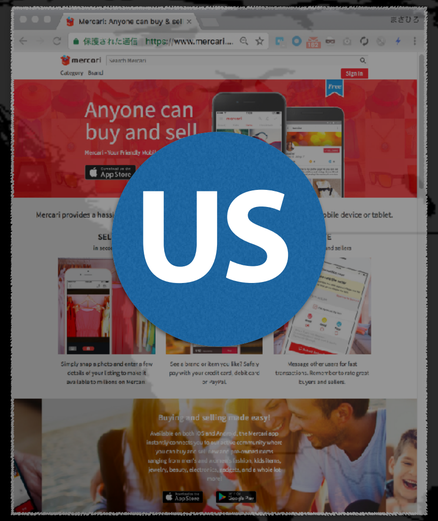
Global Infrastructure

DNS: Amazon Route53

CDN: Akamai, CloudFront

各Region

サーバが中心



共通アーキテクチャ

クラウドが中心
信頼性の高いAWSの
サービスが挟み込む

決済/物流/Domestic Service

決済/物流/Domestic Service

決済/物流/Domestic Service

Common Micro Services

Analysis: Google BigQuery

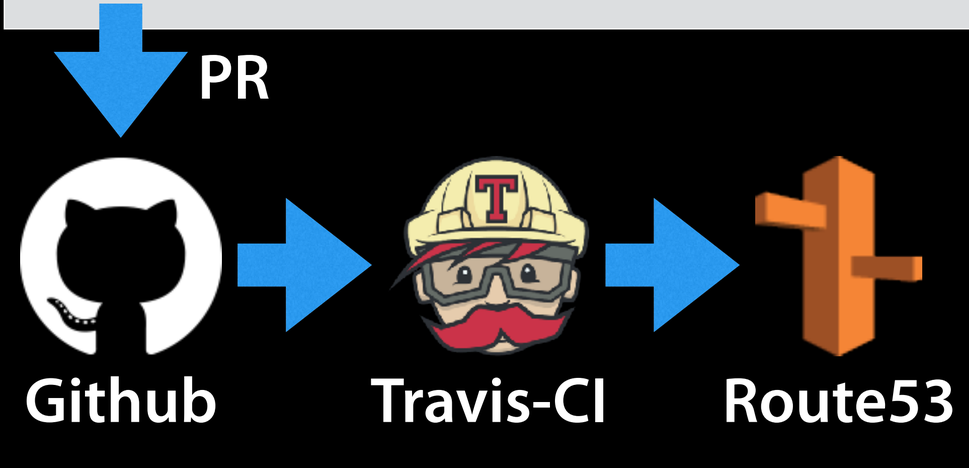
Storage: Amazon S3

Amazon Route53

- 高い可用性と信頼性のDNS
- Roadworker を利用
  - github.com/codenize-tools/roadworker
  - Routefile をGithubで管理
  - Pull Requestのmerge後、CIを経由して自動反映

#Routefile

```
hosted_zone "mercari.jp." do
  rrset "api.mercari.jp.", "CNAME" do
    ttl 30
    resource_records(
      "endpoint-api.mercari.jp"
    )
  end
end
```



Amazon Route53 + HealthCheck

- DNS-RR 運用時の問題点
  - サーバ障害時にDNSの書き換えに時間がかかる
  - ブラウザなどの一部クライアントはDNS-RRの場合、一部のサーバに接続ができない場合、他のサーバへ接続し直すので障害による影響は大きくなりにくい。
  - マイクロサービス化が進むと様々なブラウザ以外のクライアントが接続する。多くはDNS-RRの障害時の再接続は実装されてない
- Route53 の Health Checkを使い解決(を検証中)

Route53 + Health Check with Roadworker

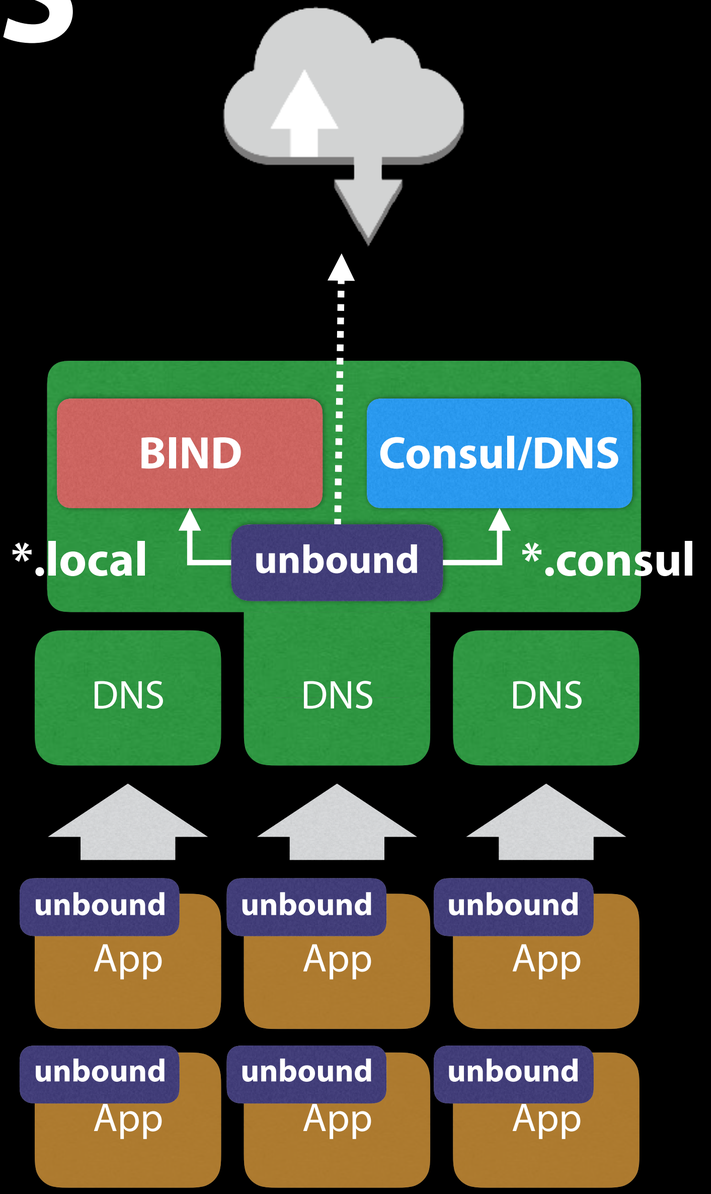
#Routefile

```
[“153.x.y.150”, "153.x.y.151"].each do |ip|
  rrset "endpoint-ha.mercari.jp.", "A" do
    ttl 30
    weight 1
    set_identifier “endpoint-ha-“ + ip.gsub(/\./,'-')
    health_check "http://#{ip}/hc", :request_interval => 30, :failure_threshold => 3
    resource_records(
      "#{ip}"
    )
  end
end
```

Health Checkにより DNS-RR でも可用性を高められる

(話はそれますが) 内部 DNS

- 全てのサーバにunboundを導入
  - ローカルキャッシュによるパフォーマンス向上
  - resolv.confより可用性が上がる
- DNSサーバのunboundでリクエストを振り分け
  - *.local はBINDが権威サーバ
  - *.consul はconsul DNS interface



(話はそれますが) 内部DNSでCNAME

- 内部DNSでマネージドサービスのエンドポイントのCNAMEを設定
  - アプリケーションから接続はCNAME経由

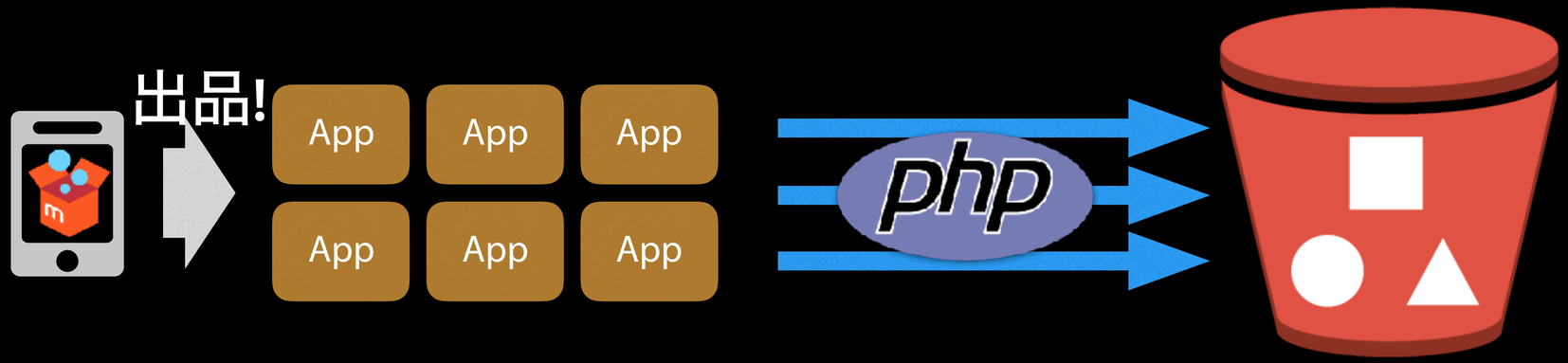
```
db-cstool-master    IN    CNAME    cstool-db.XXXXX.us-west-2.rds.amazonaws.com.
```

  - マネージドサービスからマネージドサービスへの移行、マネージドサービスからEC2への移行、またその逆の移行がやりやすい

Amazon S3

- 高い可用性と信頼性のストレージ
  - 商品画像、ログ、データベースのバックアップなどあらゆるデータを格納
- IAMを利用した高度なアクセス管理と疎結合の実現
  - サブシステムからのデータインポート・エクスポート
  - 外部サービス・パートナーとのデータ受け渡し手段

あらゆるデータのストレージ: 商品画像

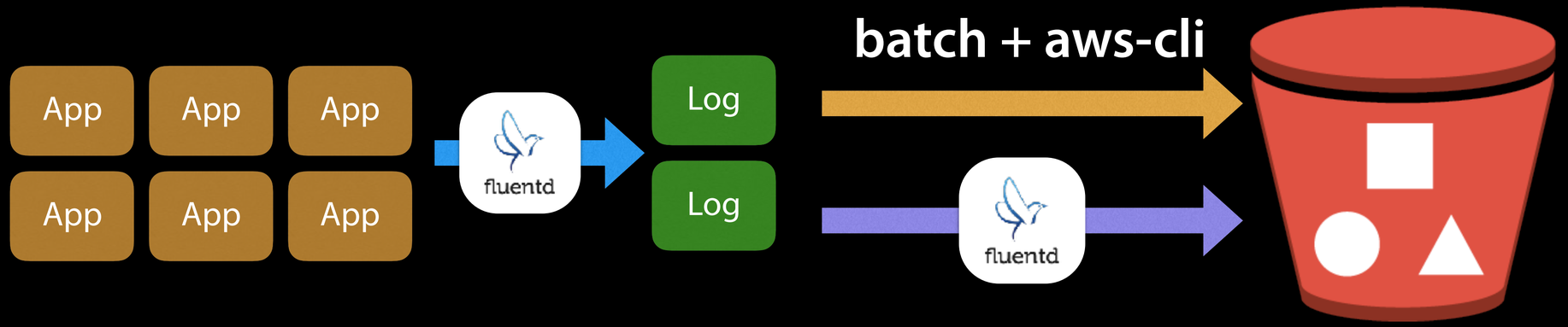


数百万枚/day

商品画像データは同期的に縮小/アップロード

AWS SDK for PHPを利用。複数の画像を並行してPUTして速度向上

あらゆるデータのストレージ: ログ

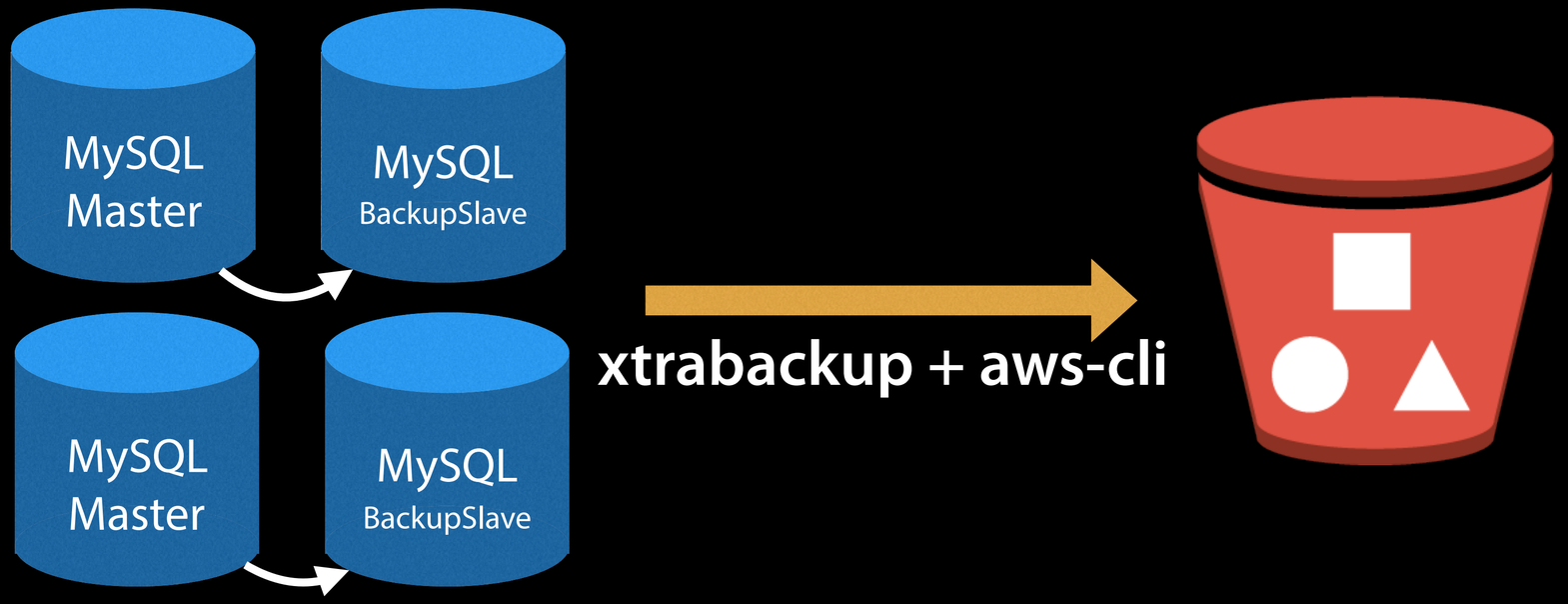


> 1TB/day

アクセスログ／エラーログなど各種ログはfluent経由で集約してS3に格納

aws-cli または fluent-plugin-s3

あらゆるデータのストレージ: バックアップ

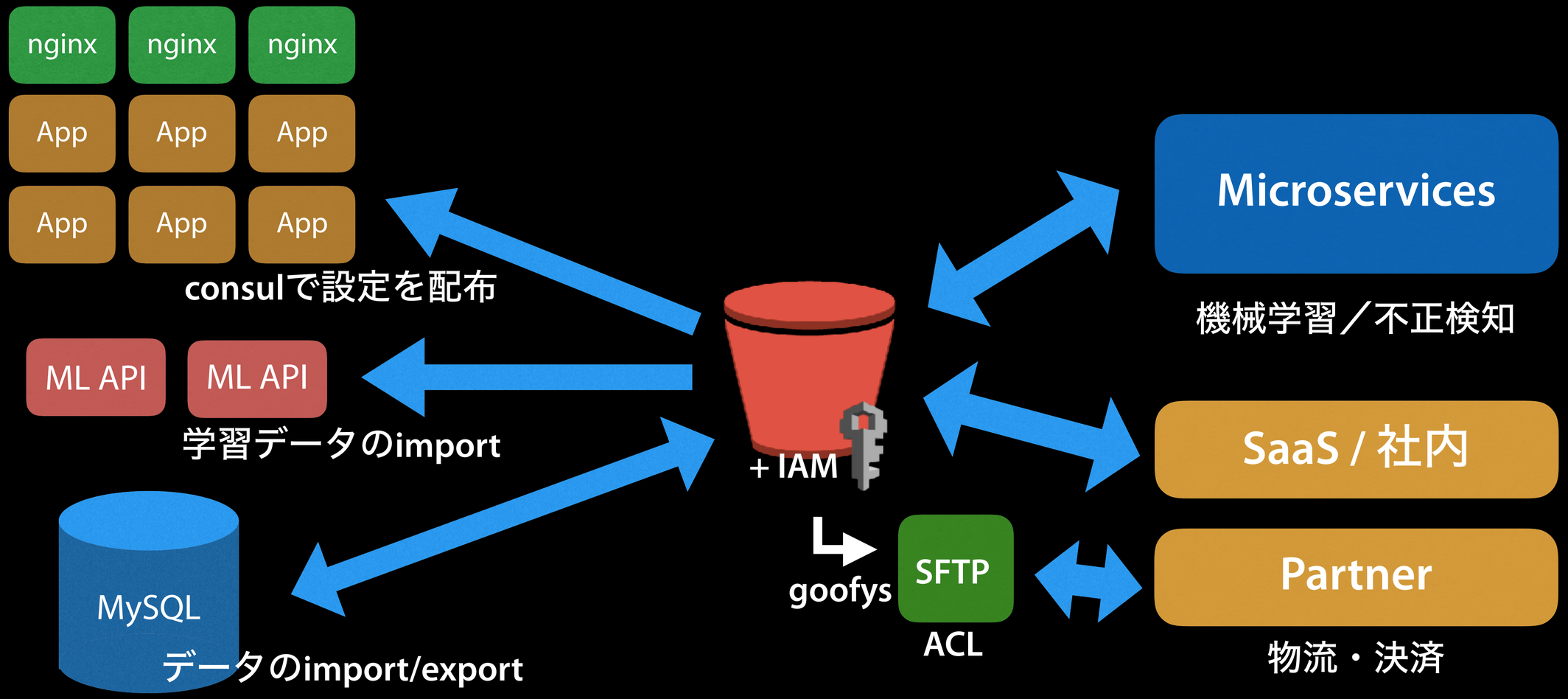


> 1.2TB(圧縮済)/day

MySQLは毎日xtrabackup(週1でmysqldump)

backup用slaveからbackupを取得。aws-cliで転送

Amazon S3 as a Hub



信頼性の高いS3をHubとして、疎結合を実現

機械学習への取り組み

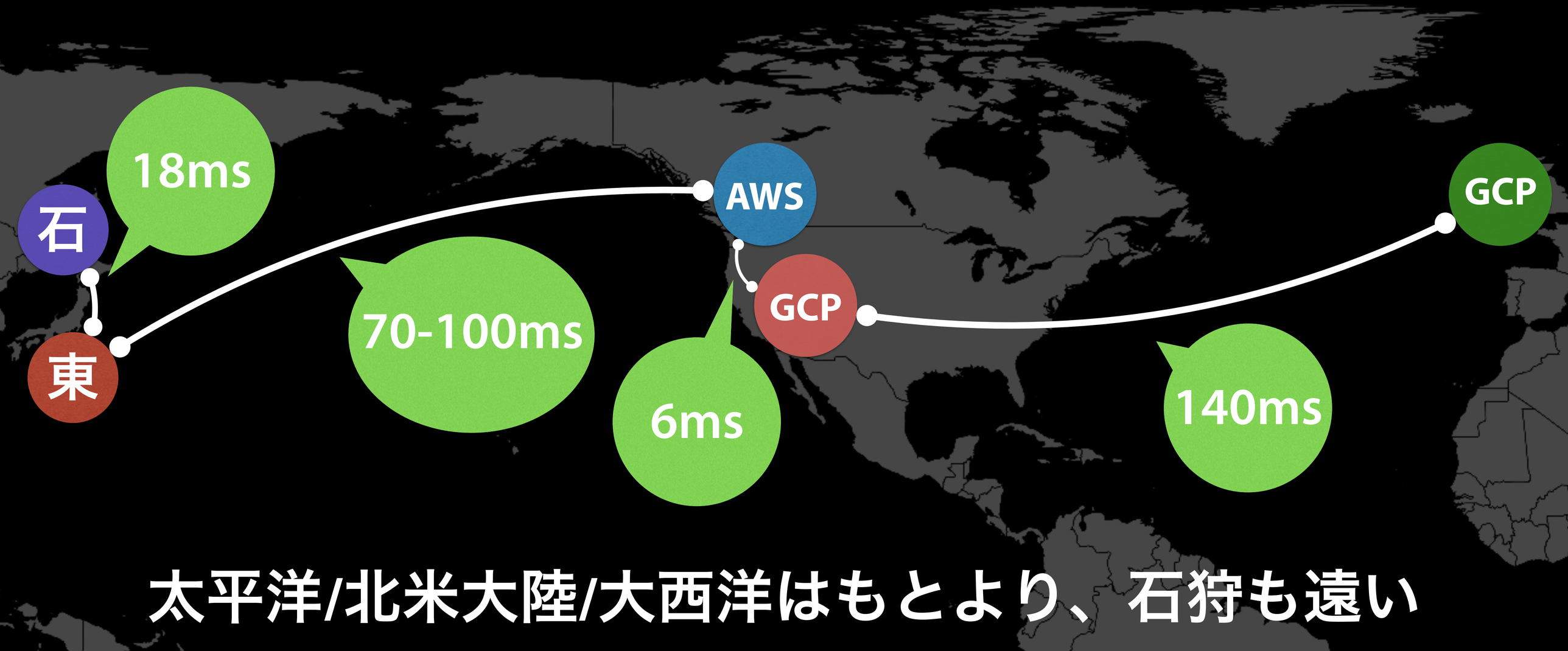
- サービスで利用中・検証中
  - 検索結果の改善。行動解析により、商品の検索インデックスにキーワードを追加し、より見つけやすく
  - 出品時の価格サジェスト
- 機械学習をだれでも試すことができる環境を
  - Amazon MLも検討

距離を超えて世界を繋ぐ

距離とレイテンシ

- 光は50msecに地球半周もできない。遠距離との通信はコストが高い
- データセンター間、クラウド間の距離がある場合には、それを克服し、効率の良い通信を行う必要
- (石狩遠い問題)

国内と国外のレイテンシ



太平洋/北米大陸/大西洋はもとより、石狩も遠い

高レイテンシ環境でのHTTPS通信

- 通常のTCP Handshakingに加え数回のやりとりが必要
  - RTT 26msecでHTTPSの通信を行なった場合、200msec以上かかる
  - RTT 100msec超えると、600msec以上
  - 参考) mercari APIのレスポンスタイム(90percentile)は 100msec

遠距離接続するユースケース

- クライアントが遠いところからサービスにアクセスする
  - 海外, US東海岸/西海岸
- アプリケーションのコードから他のクラウド(データセンター)にアクセスする
  - SaaS、マイクロサービス

クライアントからの接続改善

- CDNを利用する
  - Cloudfront, Akamai, Fastly
  - クライアントは近くにあるCDNのエッジサーバとTLS Handshaking
  - CDN と Origin 間はコネクション集約や専用ネットワークを利用することで高速化
- www.mercari.com はCDNを利用

CDNの利用: mercari Web

CDN

SAKURA internet
石狩DC

amazon web services™

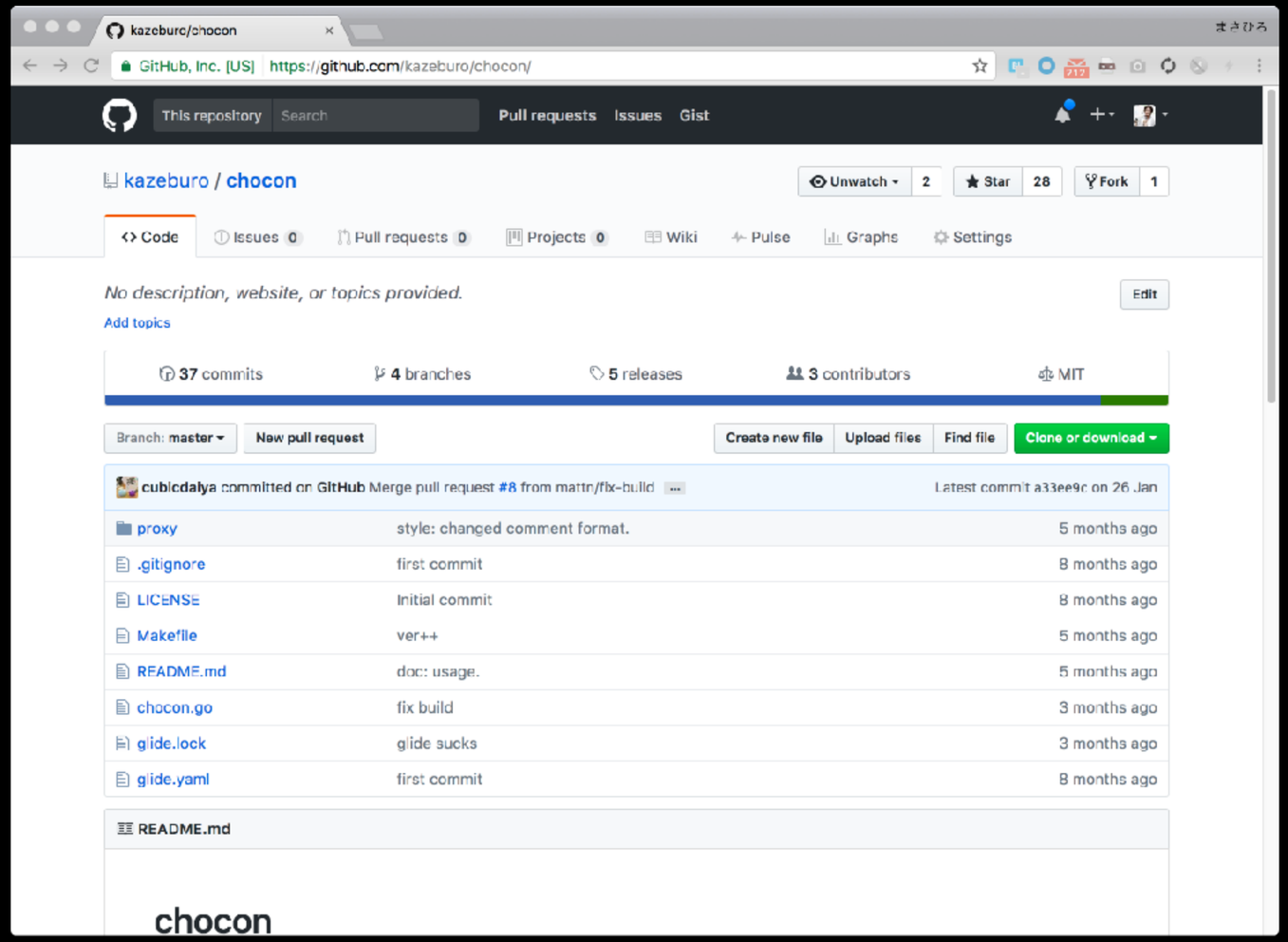
Google Cloud Platform

アプリケーションからクラウドへアクセス

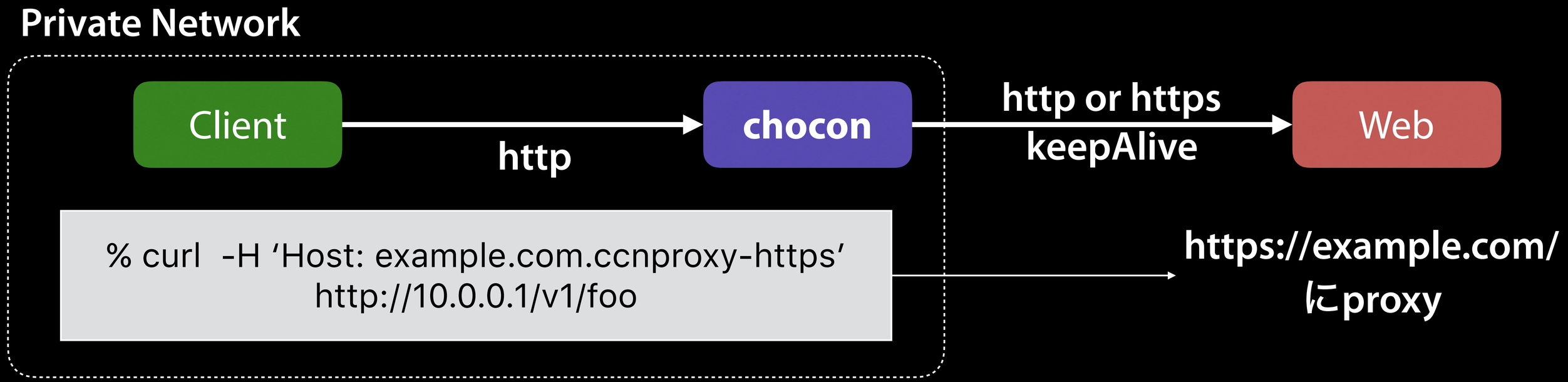
- アプリケーションでHTTPS通信のKeepAliveを行う
- PHP ApplicationでのKeepAliveは難しい
  - リクエスト処理後にメモリがクリアされ、 TCP接続も切れる
  - マルチプロセスであり、KeepAliveしても効率が悪い
- => そこで Connection Poolingを目的とした Proxy Serverを開発

chocon

- Goで実装したシンプルな Proxy Server
- OSSとして公開
  - github.com/kazeburo/chocon
- 半年以上の稼働実績



chocon



内部DNSを活用するとURLのホスト名を変更するだけ

```
*.ccnproxy-https     IN     CNAME     chocon.local.
```

```
% curl  http://example.com.ccnproxy-https/v1/foo
```

Before chocon

```
$ ./httpstat.sh /dev/null https://microservice.example.com/hc
HTTP/1.1 200 OK
Server: nginx/1.11.5
Date: Thu, 01 Jun 2017 00:43:49 GMT
Content-Type: application/json; charset=utf-8
Content-Length: 22
Expires: Thu, 01 Jun 2017 01:43:49 GMT
Cache-Control: max-age=3600,public


Body stored in: /tmp/httpstat-body.2632645114962782 39

  DNS Lookup   TCP Connection   SSL Handshake   Server Processing   Content Transfer
[       2ms   |         24ms    |       197ms   |          25ms     |           0ms     ]
              |                 |               |                   |                   |
    namelookup:2ms              |               |                   |                   |
                        connect:26ms            |                   |                   |
                                      pretransfer:223ms             |                   |
                                                        starttransfer:248ms             |
                                                                                        total:248ms
```

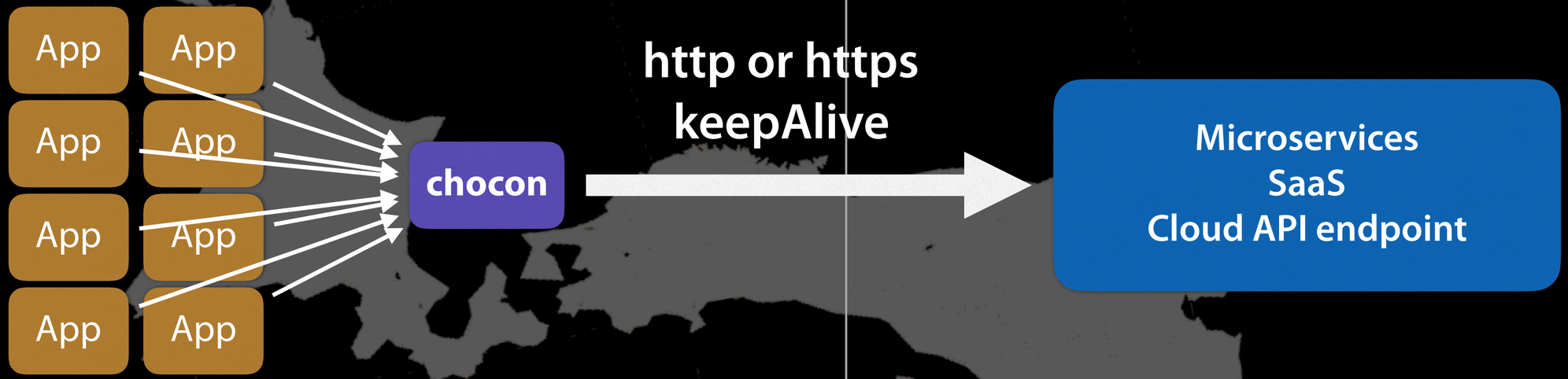
After chocon

```
$ ./httpstat.sh /dev/null https://microservice.example.com.ccnproxy-https/hc
HTTP/1.1 200 OK
Cache-Control: max-age=3600,public
Content-Length: 22
Content-Type: application/json; charset=utf-8
Date: Thu, 01 Jun 2017 00:43:49 GMT
Expires: Thu, 01 Jun 2017 01:43:49 GMT
Server: nginx/1.11.5
X-Chocon-Req: bSCzJrCMZ9wbRN8TYhZ3wV

Body stored in: /tmp/httpstat-body.3901741814962787 75

  DNS Lookup   TCP Connection   Server Processing   Content Transfer
[       1ms   |          1ms    |          19ms      |           0ms     ]
              |                 |                    |                   |
    namelookup:1ms              |                    |                   |
                        connect:2ms                  |                   |
                                      starttransfer:21ms                 |
                                                                 total:21ms
```

Why chocon?

- 似たmiddlewareは見つからない
  - 単純なforward proxyではHTTPS通信の集約はできない
  - HTTPSはend to endで暗号化。MITM Proxyが必要になる
- Go言語標準のHTTP/2により効率の良い集約、高速なアクセスが期待

chocon in JP



90msec が19msec と石狩東京間のRTT実測値まで改善

AWS SDKもendpointを切り替えることで利用可能

chocon & Pacific Ocean

100msec程度まで遅延が抑えられ、他Regionとの連携の実現。

USの先進的なクラウドサービスにアクセスしやすくなる

まとめ

まとめ

- メルカリは JP/US/UK の3拠点でサービス展開、開発も行う
- 各Regionはサーバを中心とした共通したアーキテクチャ
- グローバルではAmazon Route53, Amazon S3の高い信頼性に支えられている
- 世界を結ぶためにクラウドサービスや独自開発のソフトウェアを利用

We're Hiring!

世界に挑む、メルカリ

言い訳ナシのパフォーマンスと信頼性で支えるSRE

www.mercari.com/jp/jobs/