2023.08.30

Private Cloud Meetup #4

サイバーエージェントの機械学習基盤 ML Platform の紹介 & 開発運用の苦労話

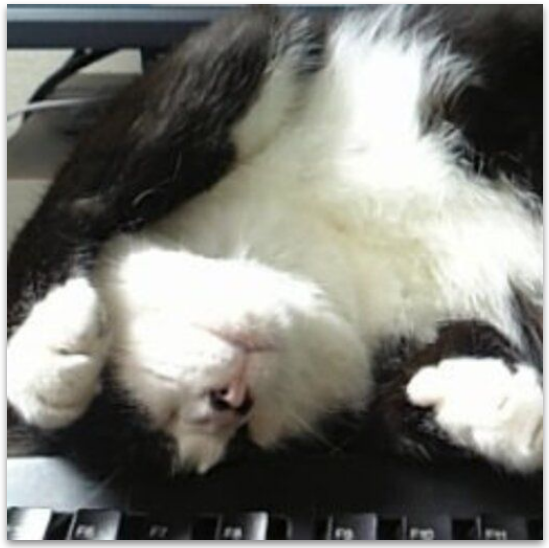
CyberAgent group Infrastructure Unit

青山 健人 (Kento Aoyama)

目次

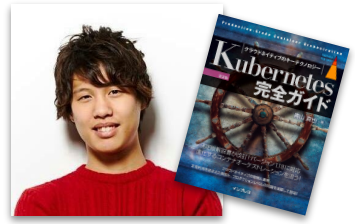
自己紹介

青山 健人 – Kento Aoyama



- 株式会社サイバーエージェント
  グループIT推進本部 CIU所属
  2022年7月中途入社
- 前職は東京工業大学の研究員
- 機械学習基盤 ML Platform の開発・運用を担当
- 趣味はビデオゲーム全般

≠



Masaya Aoyama

別人です

@ meta1127

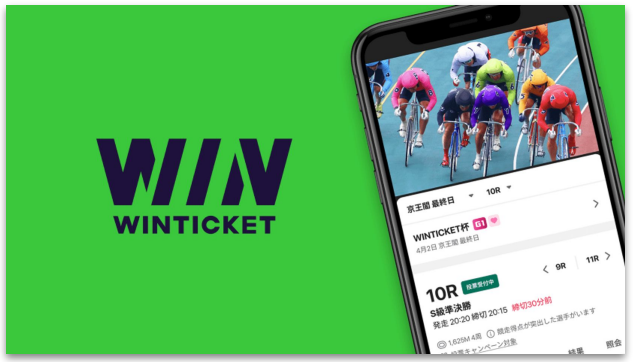
CyberAgent group
Infrastructure Unit

サイバーエージェントの主要事業と関連会社

主な事業

- メディア事業（e.g. ABEMA）
- 広告事業
- ゲーム事業（e.g. Cygames）
- その他

ABEMA



tapple

Cygames

CIU CyberAgent group Infrastructure Unit

CIU（CyberAgent group Infrastructure Unit）は、
CyberAgentグループ全体のインフラを支える組織です

Cycloud というブランドでプライベートクラウドを展開しており、
IaaSとしてのOpenStackやKaaSであるAKEなど様々なサービスを提供しています

サイバーエージェントのプライベートクラウドサービス

Cycloud

このサービスの話をします

Compute

AKE

GKE相当のKubernetes as a Service

IaaS

VM、ブロックストレージ、ロードバランサー、ファイアウォールなどのIaaS機能を提供

ML Platform

GPUが利用可能なコンテナプラットフォーム、および機械学習基盤

Storage

CDB

マネージド MySQL サービス

S4

[BETA] AWS S3に互換性をもったオブジェクトストレージ

CI / CD

hosted runner

安価かつ強力なスペックを持つGitHub Actions Runner

CCR

コンテナレジストリ

Platform Management

IAM

Cycloud Platformに統合されたアクセス制御を提供するサービス

Tools

Cycloud CLI

Cycloud Platform のためのコマンドラインインターフェイス

Terraform Provider

Cycloud Platform のための Terraform Provider

action-configure-credentials

クレデンシャルレスにCycloud Platformを操作するための GitHub Action

action-setup-cycloud

Cycloud CLI をインストールする GitHub Action

最近の関連プレスリリース

https://www.cyberagent.co.jp/news/detail/id=28484

https://www.cyberagent.co.jp/news/detail/id=28817

サイバーエージェントの社内向け機械学習基盤

機械学習基盤 ML Platform

- CIU が開発する機械学習基盤のサービス総称
  - GPU as a Service (GPUaaS)
  - Cycloud AI Platform Training
  - Cycloud AI Platform Prediction
  - Distributed

- 社内のML/DSエンジニア向けに開発
  - プロダクト開発・運用コストの削減
  - GPU利用ハードルの低減と導入の推進

ML PLATFORM

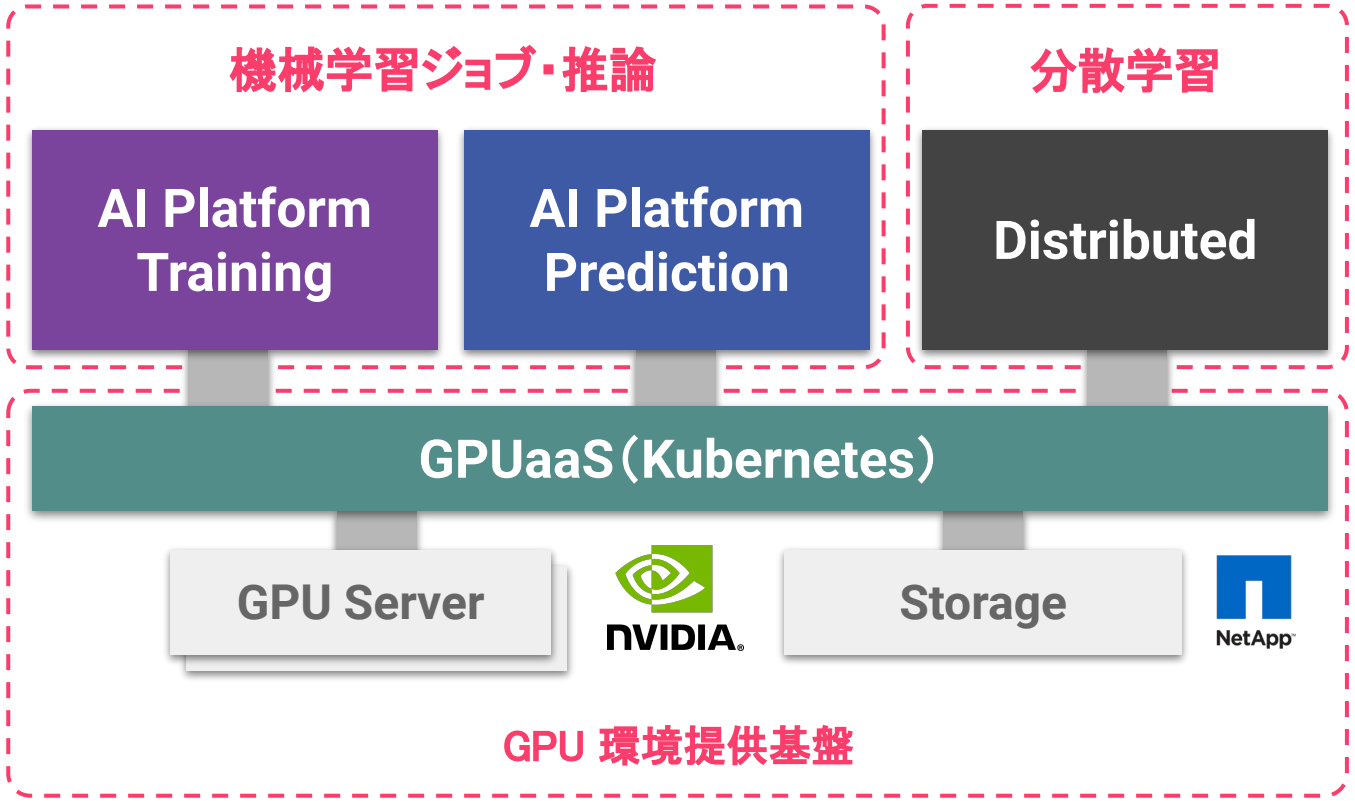
ML Platform のサービス開発経緯

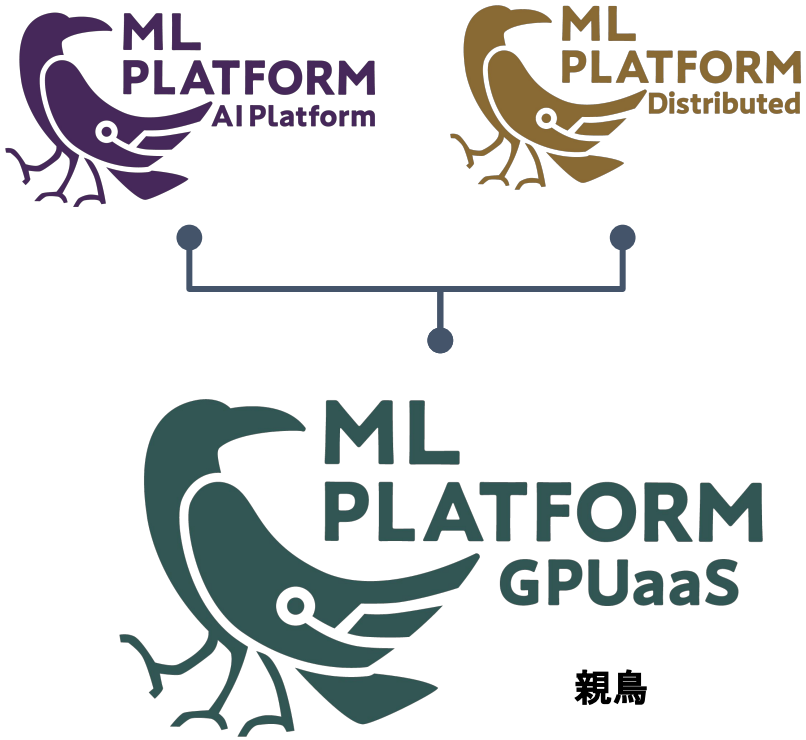
- 2019年に AI 事業本部が新設
  - 各プロダクトでのAI利用が推進
- 社内のGPUの需要が急激に増加
  - オンプレのメリットが多かった
    - 長時間利用時のコスト優位性
    - 最新GPUが利用可能になるまでの時間、パブリッククラウドのGPUの奪い合いの回避
    - 既存のDC運用とプライベートクラウドのノウハウを活かせる
- 横断インフラ組織（CIU）としてGPU基盤提供を開始
  - パブリッククラウドのGPU利用費用と比較して 約 46 % のコスト減
  - 双方のメリットを考慮して弊社内では両方を活用

ML Platform のシステム基盤と提供サービス

- GPU 搭載ノード上に構成された Kubernetes クラスタ

  GPU as a Service (GPUaaS) の基盤上で提供



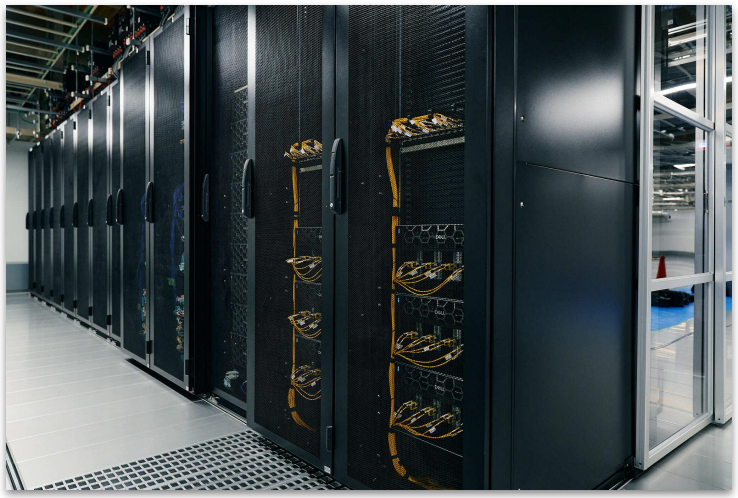


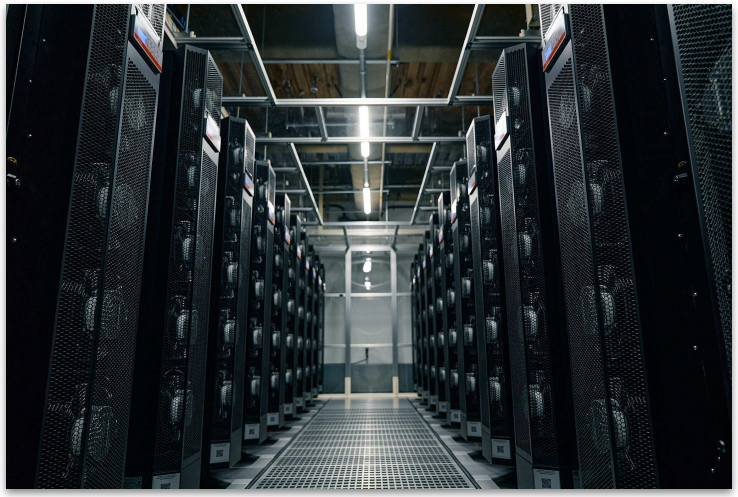
ML Platform (GPUaaS) のシステム情報

ベアメタルなGPU搭載ノードによるKubernetesクラスタ

- ベアメタルKubernetesな理由
  - VMによるパフォーマンスの低下の懸念がない
  - OpenStackベースのIaaSやKaaSサービス（AKE）で VM関連コンポーネント起因の障害が色々と見えていた
  - サイバーエージェントの事業のスピード感

vCPU	約 5,500 Cores
NFS ストレージ	約 250 TB (NetApp AFF A800)
GPU	• NVIDIA H100 80GB x80 • NVIDIA A100 80GB x48 • NVIDIA A100 40GB x16 • NVIDIA A2 x16 • NVIDIA T4 x11
外部ネットワーク	50 Gbps (Bonding 25GbE)
インターコネクト	400 Gbps (ConnectX-7) x 8 port　※ H100 only





（余談）GPUaaS のサービス提供形態の変遷

- 元々GPUaaSは GPUマシンを一元管理するサービス として誕生してきた
- 徐々に規模を拡大しつつ、その上で様々な改善とサービス開発を続けて今に至る
  - GPUaaSv1　　GPUコンテナ（ワークステーション, 占有ホスト）
  - GPUaaSv2　　GPUコンテナ + Notebook（マルチテナント化, 共有ホスト）
  - GPUaaSv3　　GPUコンテナ + Notebook + AI Platform（DC移設）
  - （～現在）　GPUコンテナ + Notebook + AI Platform + Distributed（さらにDC移設）

参考文献（過去のGPUaaS関連の発表資料）

- KubernetesベースのGPU as a Service Platform ～サイバーエージェントにおけるGPU活用の取り組み～ [INSIGHT Japan 2022 Digital]
- AI事業本部におけるGPU活用の取り組みとKubernetes - CloudNative Days Spring 2021 Online / cndo2021-ca-ml-gpuaas-aiplatform

（宣伝）DCネットワーク周辺 や Distributed の詳細

DCネットワーク関連の発表 @ JANOG52

https://www.janog.gr.jp/meeting/janog52/aiml400/

ML Platform Distributed 関連の発表 @ CADC2023

https://cadc.cyberagent.co.jp/2023/sessions/distributed-ml-with-kubernetes/

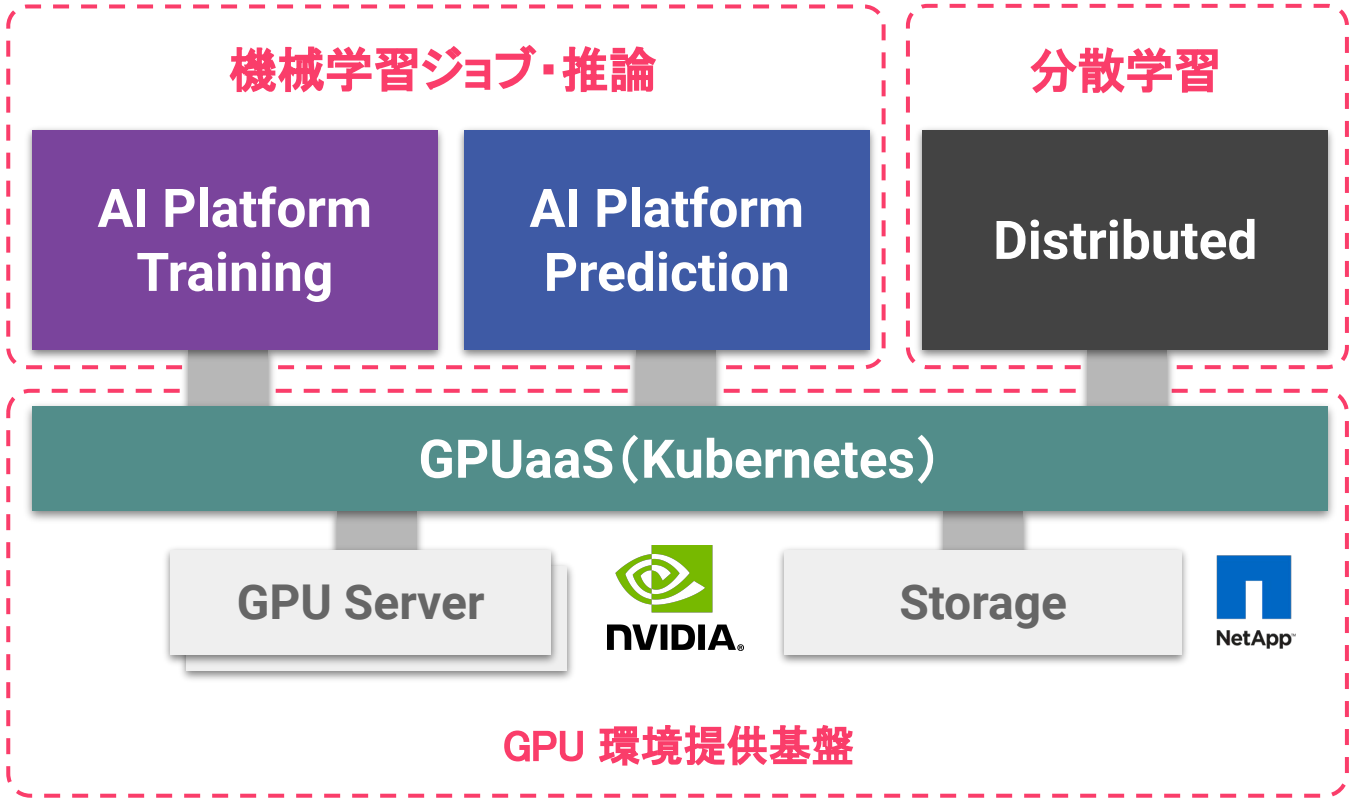
2 | ML Platform の機械学習サービス概要

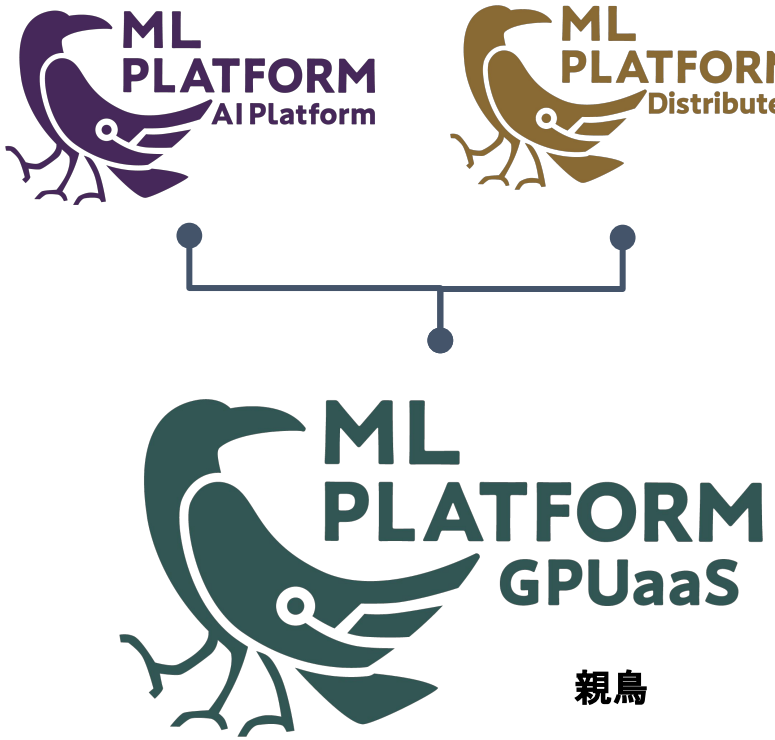
- GPU as a Service (GPUaaS)
- Cycloud AI Platform Training
- Cycloud AI Platform Prediction
- Distributed

（再掲）ML Platform の計算基盤と提供サービス

- GPU 搭載ノード上に構成された Kubernetes クラスタ

  GPU as a Service (GPUaaS) の基盤上で提供





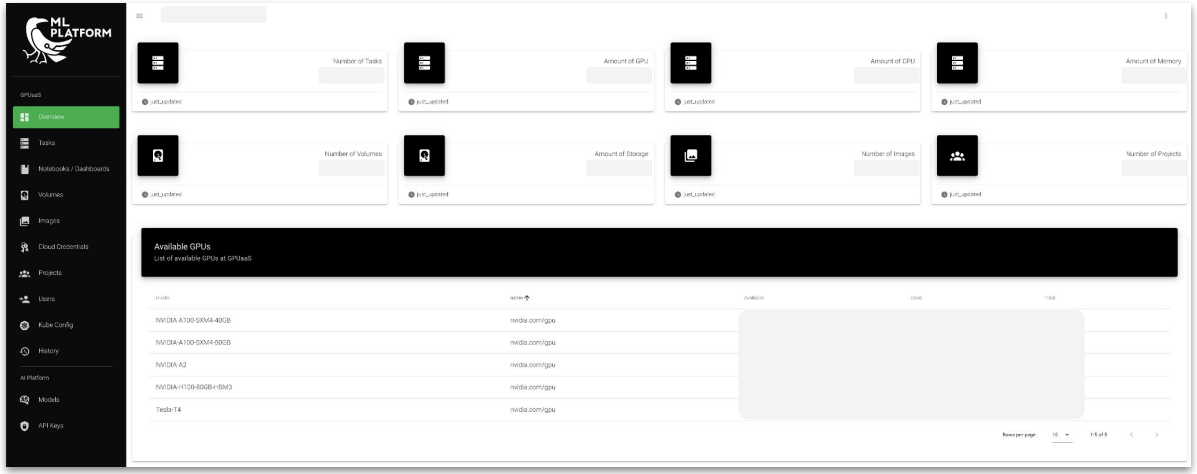
GPU as a Service (GPUaaS)

マルチテナント Kubernetes 基盤 + JupyterLab (Notebook) 環境

- 主な特徴や機能
  - WEB UIから Notebook / GPUコンテナ をデプロイ可能
  - プロジェクト管理 / ユーザ管理
  - ボリューム管理 (NFS)
  - クラウド認証情報のコンテナ埋め込み
  - GCS / S3 / Cycloud S4 (S3互換) のデータロード機能

ML PLATFORM GPUaaS

- Notebook 利用ユーザは WEB UIのみで完結可能
- ユーザが任意イメージを 登録してデプロイ可能
- KubeConfig を使って 直接Pod へのアクセスも可能



GPUaaS (GPU as a Service) の補足情報

- 単一のKubernetesクラスタですべてのGPUノードを管理
- スケジューラ は Bin-Packing (GPU) + Gang-Scheduling でカスタムしている
- Pod に割り当てられたGPUは基本的に占有する
  - Pod の Preemption は導入していない（検討中）
- ノード間 Interconnect ありの H100 GPU搭載ノードは Distributed 専用
  - 現在は一部のユーザのみの限定提供 (beta)
  - kubernetes-sigs/kueue の ClusterQueue によって資源管理
    - 将来的にはGPUaaS全体で Kueue による管理を採用

Cycloud AI Platform Training

機械学習ジョブの実行基盤

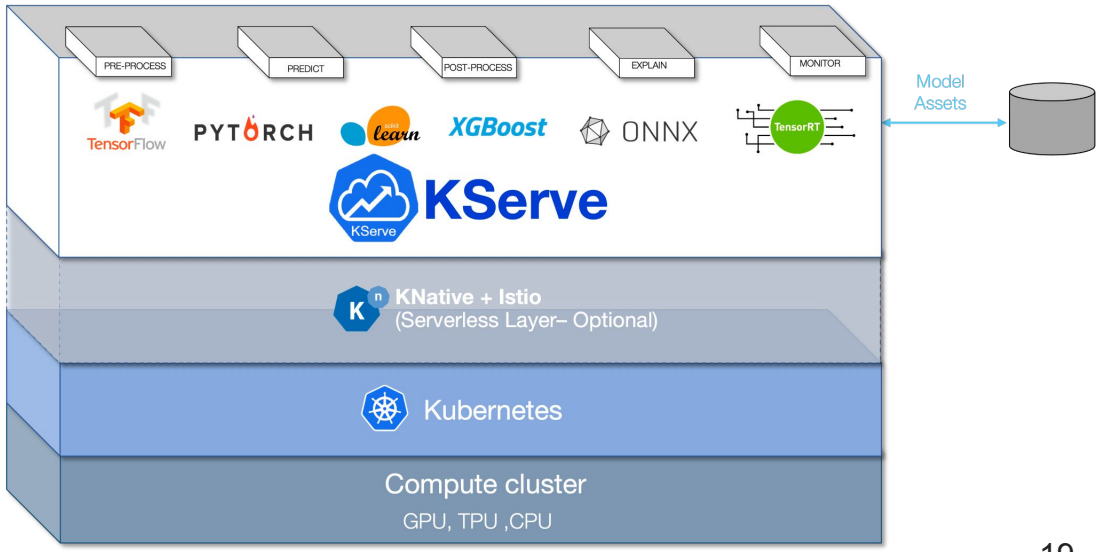
- 機械学習タスクをジョブとして実行可能
- ジョブは kubeflow/training-operator を採用

- 特徴
  - GCP AI Platform に準拠したコマンド体系
    - 現在はKubectl Plugin による提供
  - Tensorflow/PyTorch のフレームワークに対応
  - 独自定義の機械学習カスタムコンテナを実行可能
  - モデルの出力先として GCS/S3 に対応
  - kubeflow/katib によるハイパーパラメータ最適化（HPO）

サーバーレス推論基盤

- 学習済みモデルをホスティングするサービス
- kserve/kserve を採用

- 特徴
  - GCP AI Platform に準拠したコマンド体系
    - Kubectl Plugin → Cycloud CLI による提供
  - オンライン推論・バッチ推論
  - パブリッククラウド（GCS/S3）対応
  - GPU 推論
  - グローバルエンドポイント



https://github.com/kserve/kserve

大規模モデル向け分散学習基盤 (2023/06~, beta)

ML PLATFORM Distributed

Kubeflow Kueue

特徴

- ノード間並列分散学習ジョブの実行環境
  - MPIJob 対応（今後 TFJob, PyTorchJob 対応予定）
- MPIJob実行のバックエンドは kubeflow/mpi-operator を採用
- ジョブのキューイングは kubernetes-sigs/kueue を採用
- Podネットワークは SR-IOV Network Device Plugin + SR-IOV CNI + Multus CNI
  - RoCEv2を利用したPod間のMPI/NCCL通信が可能
- WEB UI と Cycloud CLI による提供
- 現在は インターコネクトありのH100 GPU のみ利用可能

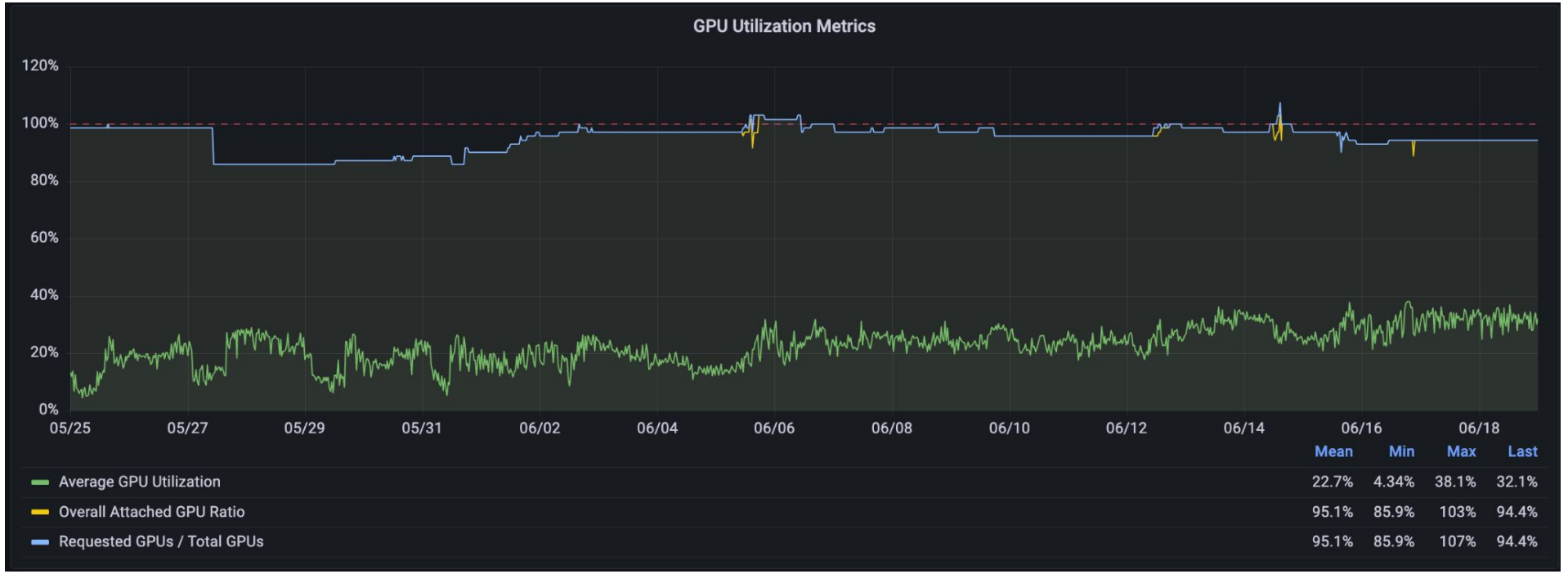
詳細はこちら↓↓

3 | 開発運用の課題と取り組み

ML Platformの開発運用の苦労話と現在の取り組み

平均GPU利用率（Utilization）が低い問題

- サービスの元々の提供形態の経緯もあり、対話的なNotebook の利用ユーザが主流
- GPU占有率は高い が、GPU使用率が低い 状態



（2023/05/25 ～ 06/18）

・全体GPU要求率
約 80% ～ 100%超

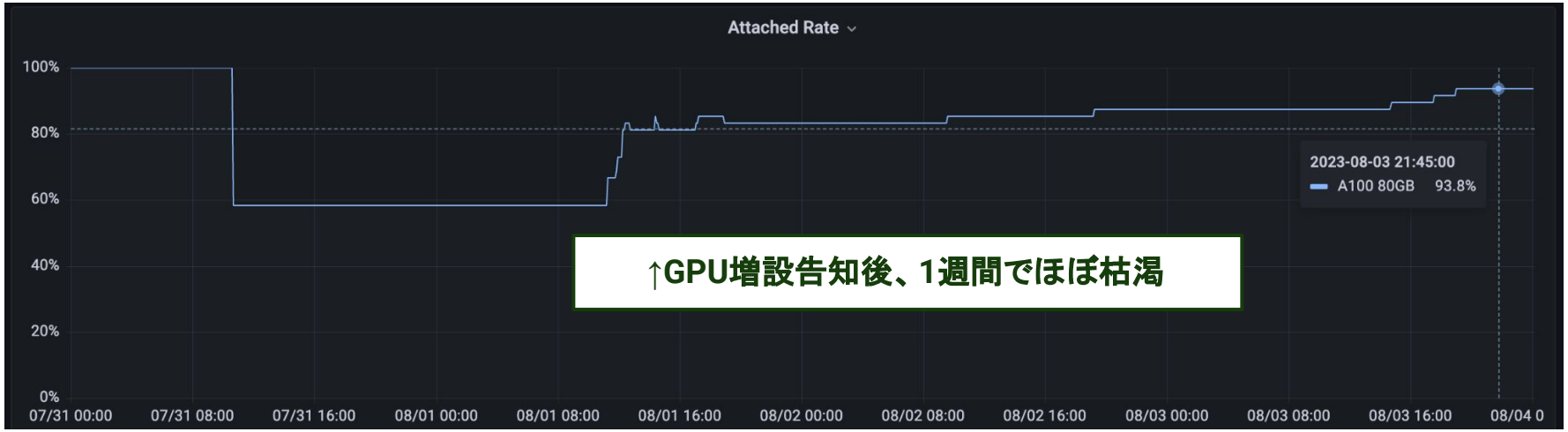
・全体GPU占有率
約 80% ～ 100%

・平均GPU使用率
約 20 ～ 40%

- GPU 要求率（Requested Ratio）　= 全体のGPU枚数に対して、要求されたGPU枚数の割合
- GPU 占有率（Attached Ratio）　= 全体のGPU枚数に対して、Podと紐づいているGPU枚数の割合
- GPU 利用率（Utilization）　= 各GPUの使用率の平均

問題の背景1: GPU需要の増加に追いつけていない

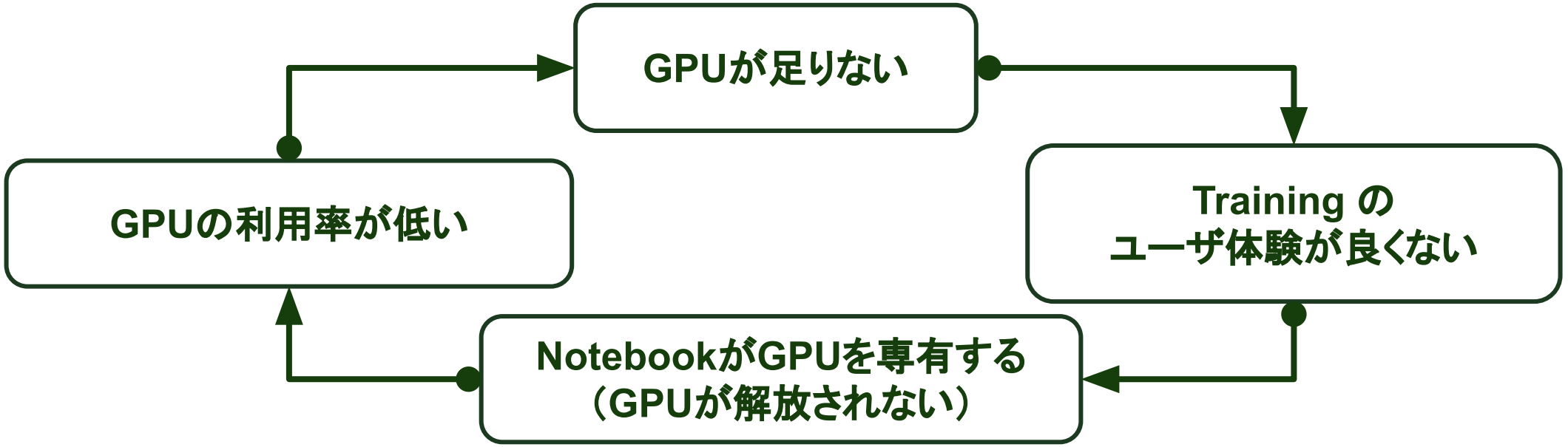
- 近年の社内のGPU需要増大に計算資源の拡充が追いついていない
  - 「ユーザ数」も「ユーザあたりの要求数」もずっと増加傾向にある
    - パブリッククラウドのGPUが確保できなくて相談に来るユーザが多い
  - GPUを増設しても、1週間も経過すると空きがなくなるのが現状... 😇（図）
  - 確保しているNotebook (GPU) を解放すると次に確保できる保証がないので、ユーザはずっと専有したい（治安問題）



問題の背景2: ジョブ形式の利用が少ない

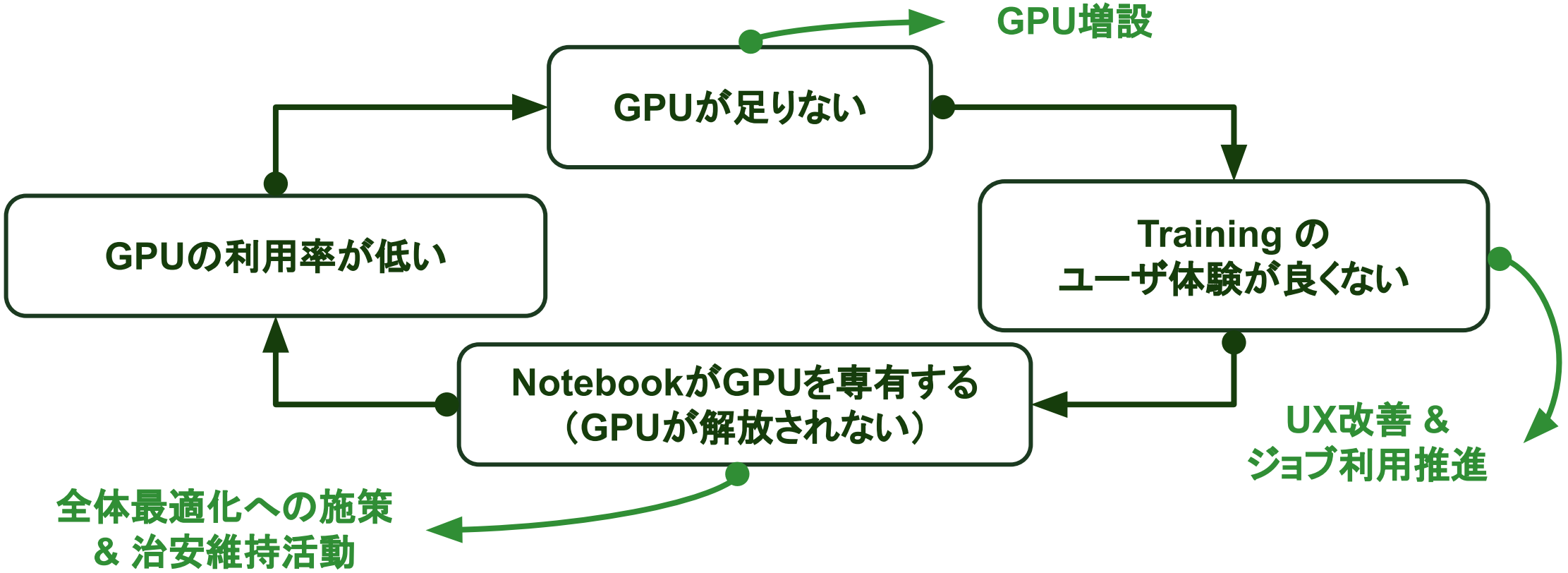
- Cycloud AI Platform Training が一部のユーザにしか使われていない
  - ユーザがジョブの形態に慣れていない と有効利用やデバッグがそもそも難しい
    - キューイングされたらジョブの書き方の問題で即エラー終了していたり...
    - そもそもGPUが空いていないと使い始められない
      - GPUを専有できて準備と実験ができるなら、Notebookの方が楽
  - サービスの使い勝手（ユーザ体験）が悪い
    - WEB UIなし、SDKなし、Kubectl Pluginのコマンドのみの提供
    - Kubectl Plugin 提供はユーザがK8s慣れしていないと心理的距離がある
  - ML Platform (GPUaas) は 元々NotebookとGPUコンテナを払い出すサービスだった
    - ユーザの意識や利用形態を上手く変化させるのは簡単ではない

運用上の課題: GPU不足とGPU利用率低迷の悪循環



- 全体最適に至るための道筋（ジョブ形式の利用）が十分に整備されていない
- さらに慢性的なGPU不足により、ユーザの行動がそれぞれで局所最適に陥っている
  - 「組織全体の効率と成長のためのインフラ」として全体最適化を促進する施策が必要

運用上の課題: GPU不足の悪循環に対する取り組み



- 慢性的なGPU不足 → GPUの継続的な増設
- Trainingが利用されない → ジョブのユーザ体験改善と利用推進の施策
- NotebookによるGPU専有 → 全体最適化への施策や治安維持活動

平均GPU利用率を
改善する

課題に対する現在と今後の取り組み

GPUの継続的な増設

- GPU総台数は直近1年で2倍以上、かつ今後も見据えてDCを移設済み
- 一時的なパブリッククラウドのリソースの組み込み等もPoC検証中

ジョブのユーザ体験改善と利用推進の施策

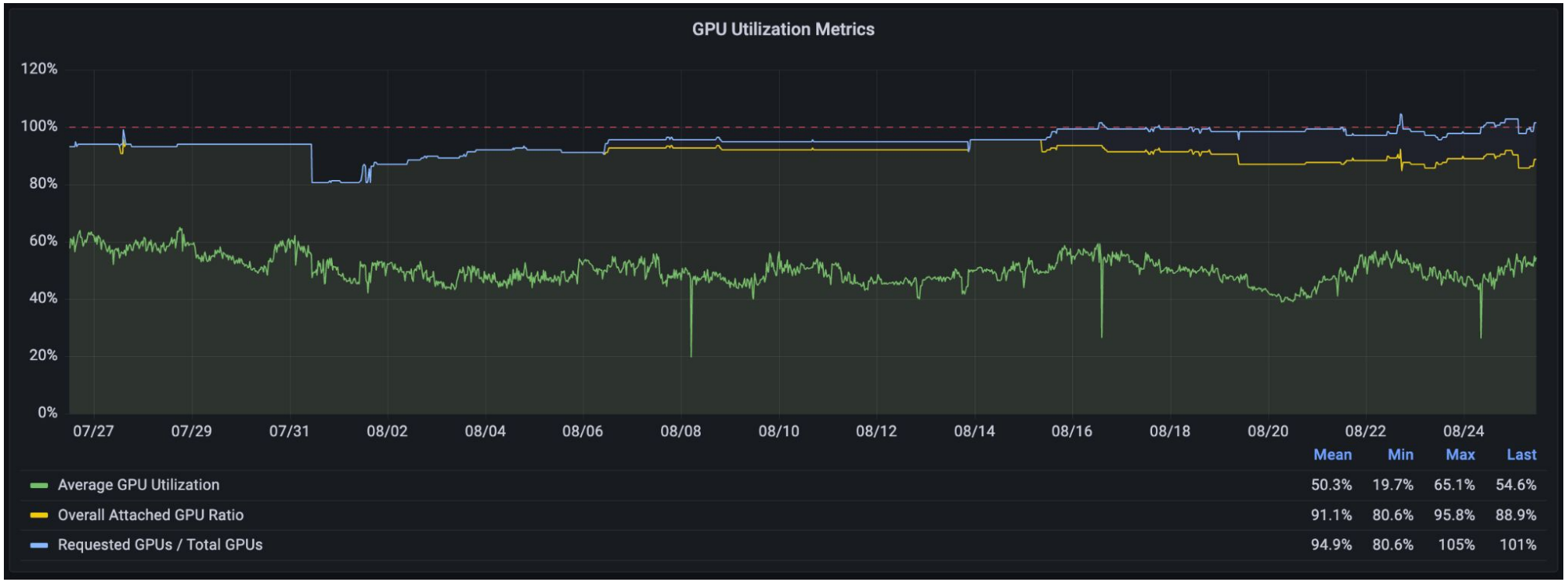
- 直近リリースの Distributed では Kueue によるジョブのキューイングを前提とした設計
- Trainingサービスとコマンドの抜本的な改修（ongoing）
  - WEB UIの提供、Cycloud CLI 移行、サービスのgRPC化と諸々改修
- ジョブのキューイング専用のノードの導入の検討中

全体最適化への施策・治安維持活動

- 暫定で人間が月毎の集計を見て利用率の低いPodの掃除を交渉している（人力パトロール）
- Notebookの時間制限や、PodのPreemption等の施策の準備中

現在の状況（7/25 ～ 8/25）

- 平均GPU使用率は多少改善している（約+20%程度）
  - Distributed による分散学習実行（H100 GPU）が大きく寄与
  - 現在の Distributed は一部ユーザのみ提供で限定的だが、
    今後は Training 含めたジョブの利用推進 & 全体最適化の施策を進めていく



（2023/07/25 ～ 08/25）

- 全体GPU要求率
  約 80% ～ 100%超
- 全体GPU占有率
  約 80% ～ 100%
- 平均GPU使用率
  約 40 ～ 60%
  （ 約 +20% 改善）

これまでの取り組み、今後の方針のまとめ

- 対話的なNotebook利用による GPUの占有率が高く、GPU利用率が低迷する問題
- 以下の取り組みにより、ひと月の平均GPU利用率は 約20%程度 の改善に成功
  - 分散学習基盤 Distributed でジョブ形式の推奨
  - GPU利用率の監視による人力パトロール
- 今後の方針
  - GPUの増設、規模拡大の継続
  - ジョブ形式の利用にユーザを誘導して、全体最適化を促進
    - Notebookの時間制限の導入、PodのPreemptionの導入
    - Training の改修によるジョブ形式の利用のUX改善

4 | おわりに

まとめ

- サイバーエージェントの機械学習基盤 ML Platform のサービス概要を紹介
  - GPUaaS / Training / Prediction / Distributed
- 運用上の課題として全体のGPU利用率の問題について、
  問題原因の考察と改善のための取り組みの一部を紹介
- 今後も ML Platformの規模拡大と計算基盤としての最適化を継続していきます

We're Hiring!

- 一緒に機械学習基盤を開発する仲間を絶賛募集しています！
  - 新卒採用 / キャリア採用
  - Twitter(X) @meta1127 でもお気軽に！

CyberAgent group Infrastructure Unit

Appendix | 補足情報

Cycloud AI Platform Training のコマンド例

- **Training は Kubectl Plugin として提供 → 将来的にCycloud CLI に統合予定**

【 Job の発行】

```
kubectl ai-platform jobs submit training sample
\
  --config ./config.yaml \
  --package-path ./trainer
```

【 Job の発行用スクリプト】

```
trainer/
├── __init__.py
├── experiment.py
├── model.py
└── task.py
```

【 Job の設定ファイル（config.yaml）】

```
trainingInput:
    scaleTier: CUSTOM
    jobDir: "gs://test-bucket/test-job"
    masterType: n1-standard-16
    masterConfig:
        acceleratorConfig:
            count: 1
            type: NVIDIA_A100_40GB
    args:
        - "--test-arg1=test1"
        - "--test-arg2=test2"
    pythonModule: trainer.task
    pythonVersion: 3.8
    runtimeVersion: 2.5
```

基盤側で用意しているコンテナイメージ
で指定したスクリプトが実行されます

Cycloud AI Platform Prediction のコマンド例

- 利用感は GCP 準拠、コマンドは Cycloud CLI (+ Kubectl Extension) で提供

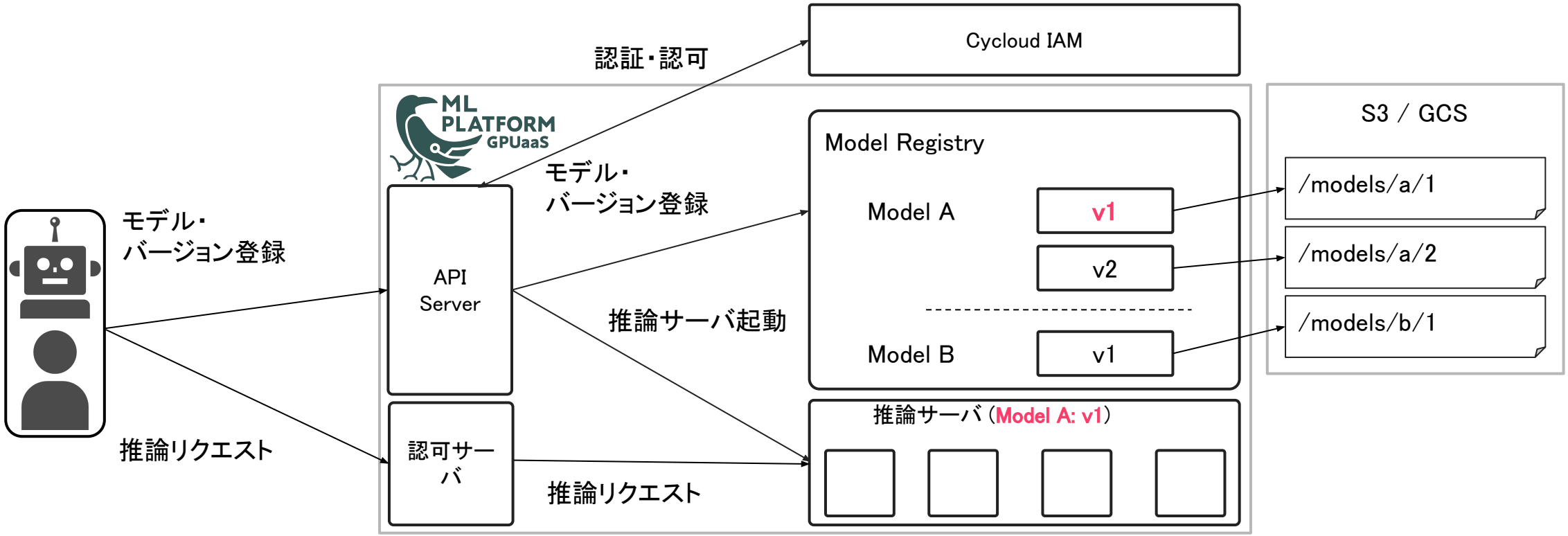
```
# 推論サーバの作成
$ cycloud ai-platform models create mnist

$ cycloud ai-platform versions create v1 \
    --model mnist --framework tensorflow \
    --origin gs://your-bucket/path/to/model
```

```
# 推論
$ cycloud ai-platform predict \
    --json-instances prediction_input.json \
    --model mnist --version v1
# {
#     "predictions": [
#         {
#             "scores": [0.999114931, 9.20989623e-05, 0.000136786344,
0.000337257865, 0.000300533458, 1.84813962e-05],
#             "prediction": 0,
#             "key": "   1"
#         }
#     ]
# }
```

Cycloud AI Platform Prediction の対応フレームワーク

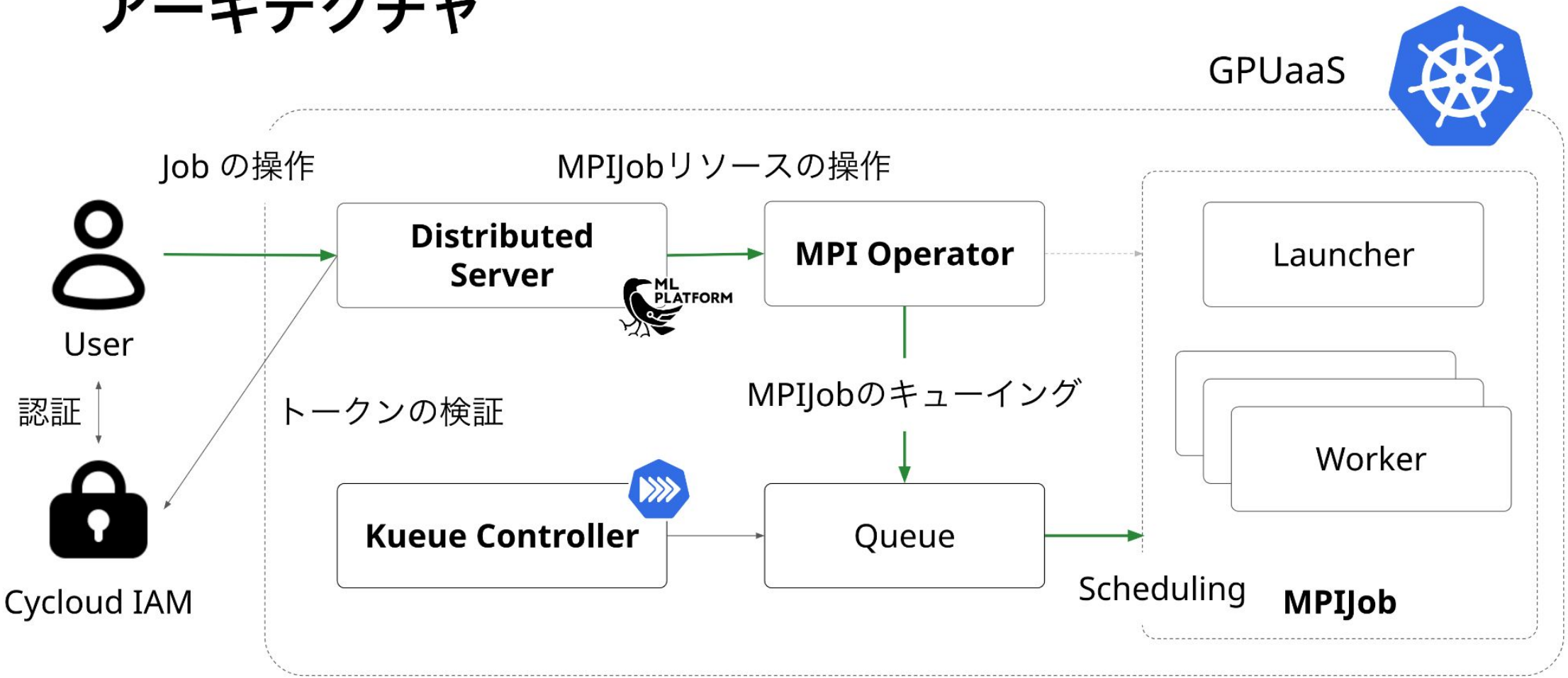
- Tensorflow
- PyTorch
- scikit-learn
- XGBoost
- LightGBM
- ONNX
- Triton Inference Server
- カスタム予測コンテナ

Cycloud AI Platform Prediction のアーキテクチャ図



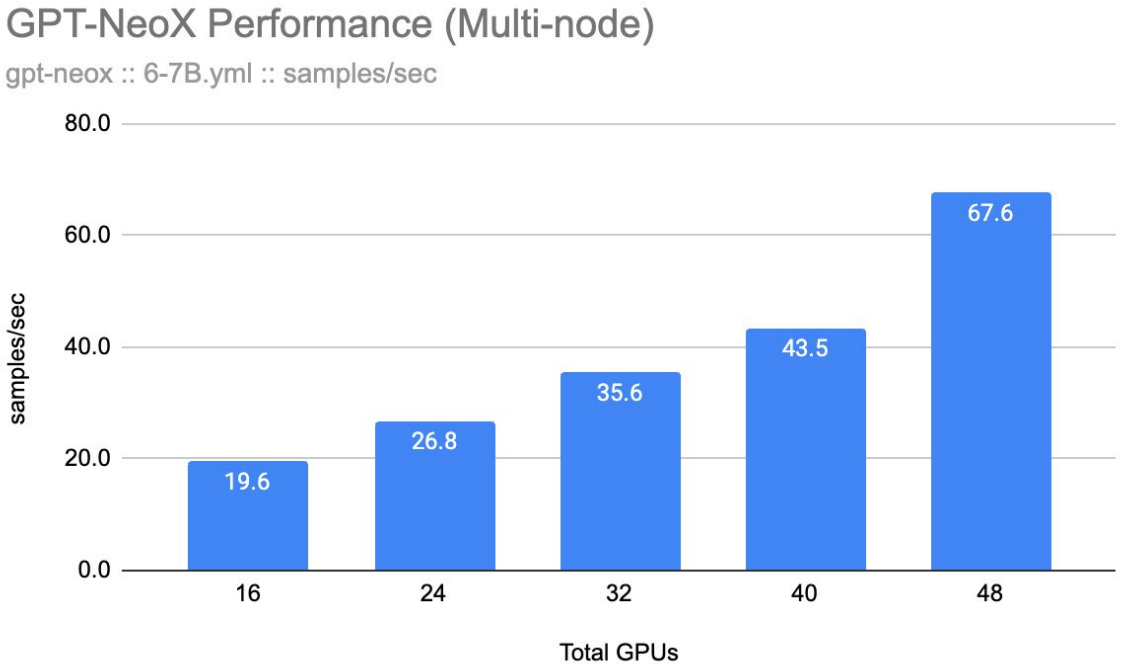
ML Platform Distributed のアーキテクチャ図

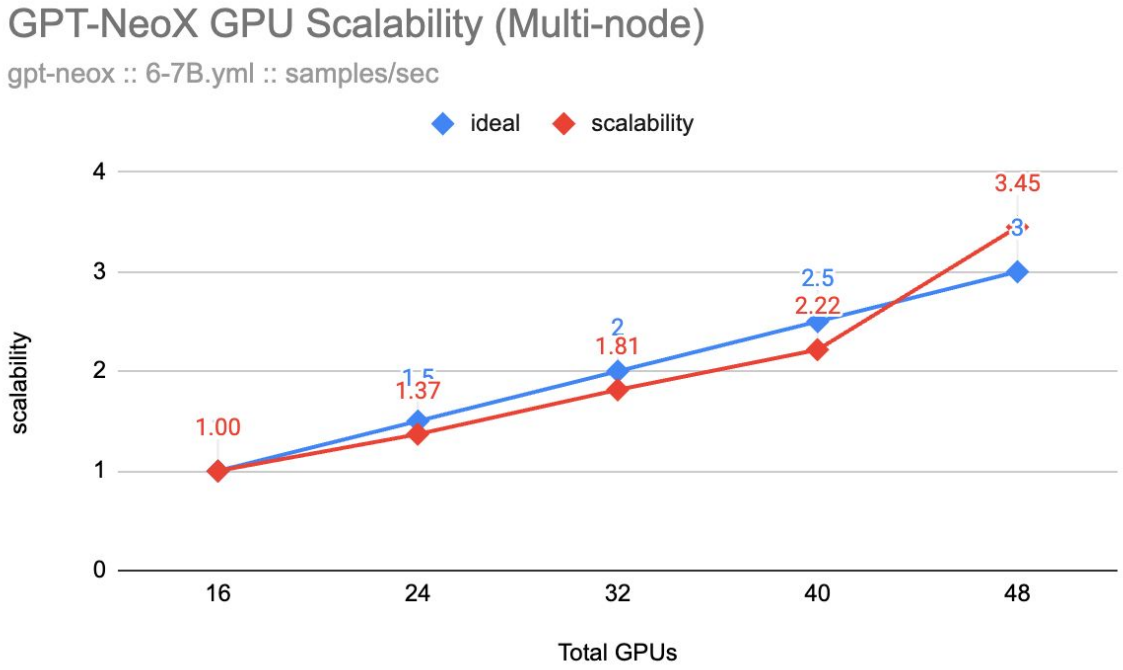
アーキテクチャ



Distributed によるLLM学習の並列性能検証

- LLM学習に用いるフレームワーク (GPT-NeoX) のサンプルデータで動作&並列性能検証





- サンプルデータの60-70億パラメータの学習速度とスケーラビリティは良好
- より大規模なモデルやノード数での性能計測は準備中