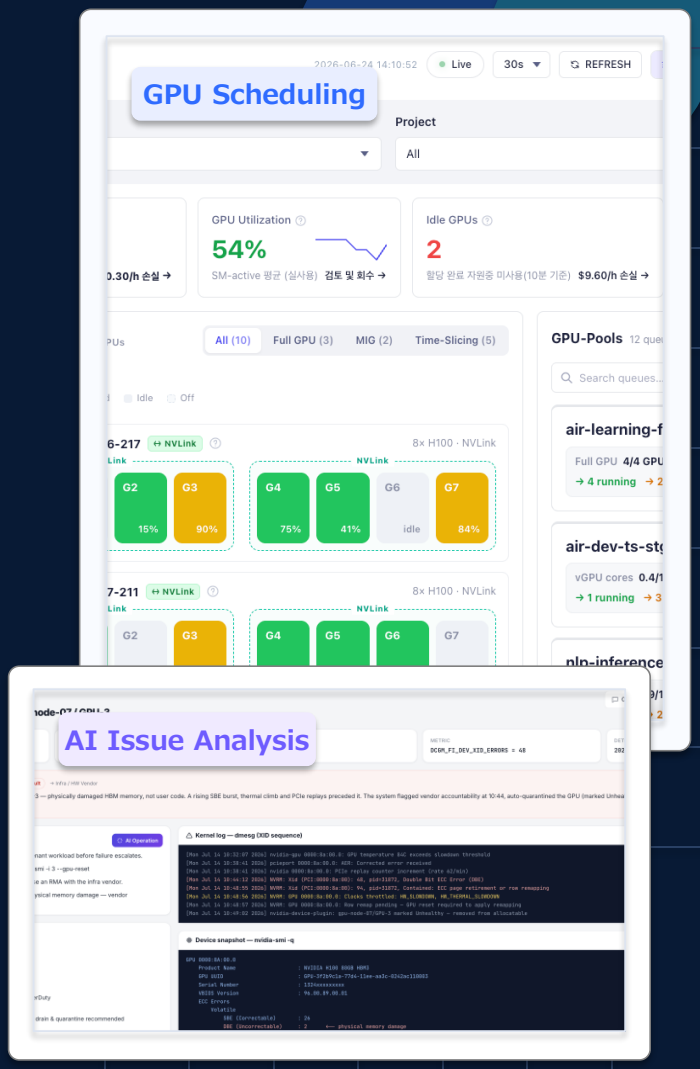
GPUaaS

GPUをインフラから“運用サービス”へ

動的割当 × AI監視・診断 × 高可用性 ×
モデル運用でGPUの生産性と運用効率を高める

GPU利用率 **最大化**	開発待ち時間 **最小化**
AI運用コスト **削減**	AIインフラ **安定運用**



ビジネス要件に応じてGPUリソースを動的に割り当て、
推論エンジンを最適化
ワークロード変動に柔軟に対応しながら、
GPU利用率と運用効率を向上

Monitoring / Diagnosis / Notification / HA / MIG・vGPU / Scheduling / Model Manager

KEY FEATURES

GPUaaSの中核機能

MON

可視化・統合管理

Cluster／Project／GPU Type別の利用状況、コスト・性能、GPU Pool／Quotaを一元管理

AIOps

AI監視・診断・通知

DCGM／XID／ECC等を監視し、原因をSW・Infra・HWへ分類。対応ガイドを自動通知

MaaS

モデル運用まで一体化

SOTAモデル登録・サービング、Playground、vLLM Benchmarkで本番最適化を支援

GPUaaS

GPU運用を検知から復旧・最適化まで一気通貫に

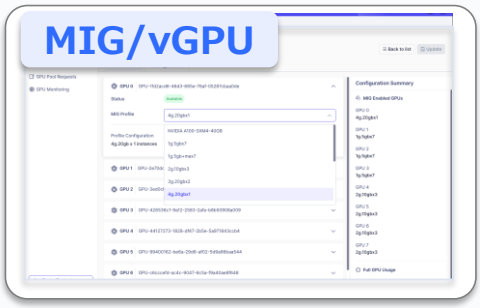
可視化→異常検知→Deep Diagnosis→AI分析・通知→GPU再配置→モデル運用

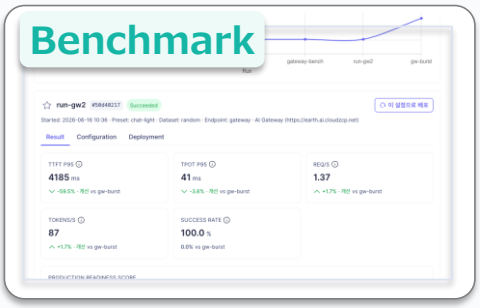
ISSUE & SOLUTION

課題と解決策

01	GPUの遊休・偏在	GPU Pool／Quota、MIG・vGPU、優先度ベースの動的割当により共有効率を高める
02	障害原因の切り分けに時間	DCGM／XID／定期診断とAI分析で、SW・Infra・HWの責任区分と対応ガイドを提示
03	通知・初動対応が遅い	Teams／Slack／LINE／Webhook／Email等へ重要度・担当別に自動配信し、Runbookまで連携
04	障害GPUによるサービス影響	Unhealthy GPUを割当対象から除外し、Workloadを利用可能なGPUへ再スケジューリング







USE CASE

活用シーン

01 共用GPU基盤

複数Project／ClusterでGPU PoolとQuotaを分離・統合管理し、Self-Serviceで必要なGPUを即時割当

02 生成AI／LLMサービス

OSS／Fine-Tunedモデルのサービング、vLLM／TensorRT-LLM最適化、Playground・Benchmarkに対応

03 大規模GPUフリート運用

数十〜数百GPUを一画面で監視し、異常兆候・劣化を早期検知 RMA／SLA向け証跡も保管

FUNCTION / SPEC

主要機能・対応範囲

Monitoring	GPU使用量・コスト・性能、ノード／GPU／MIG、Pool／Quota
Diagnosis	DCGM 7段階診断、夜間Idle診断、履歴・PDF Report
Notification	HW特化ルール、AI原因分析、監査証跡・多チャネル通知
HA / Scheduling	Failover、Preemption、Gang Scheduling、TimeTable
Model Manager	SOTA Model、Playground、vLLM Benchmark、AI推奨

AIインフラ運用の潮流

- 専有 → 共有・統合管理
- 事後対応 → 予兆検知・証跡
- 手動運用 → AIOps
- モデル配備 → MaaS化

お問い合わせ・サンプル相談
株式会社CHO&Company
〒108-0074
東京都港区高輪4-10-63
エミナンス高輪4F
TEL：03-6721-7387
MAIL：sales@cho-co.jp
HP：www.cho-co.jp