

Cloud Operator Days Tokyo 2026

日本のAI社会実装を支える 次世代GPUクラウドサービス開発への挑戦

KDDI株式会社 コア技術統括本部 ネットワーク開発本部 クラウドサービス開発部

谷岡 功基 | 杉田 修斗

2026-07

登壇者紹介



谷岡 功基

タニオカ アツキ

KDDI株式会社
クラウドサービス開発部
コアスタッフ

経 歴

KDDI入社以来、KDDIクラウドプラットフォームサービスにおけるサーバ開発業務に従事。その後、KDDI Video Management Service および KDDI GPU Cloud のサービス立ち上げに参画し、クラウドサービスの開発を担当。HaaSからSaaSまで多様な提供形態におけるクラウドサービス開発を経験し、基盤設計から実装、運用を見据えた技術検討までを一貫して推進。

資 格 / 得 意

PMP(Project Management Professional)
VMware VCP-DCV

趣 味

旅行・キャンプ
ポーカー



杉田 修斗

スギタ シュウト

KDDI株式会社
クラウドサービス開発部
コアスタッフ

経 歴

2020年KDDI入社後、5年間にわたり社内・モバイル系向けIaaS基盤の開発を経験。開発・運用業務の自動化を主として推進しつつ、ハイパーバイザーの大規模アップグレードや基盤移行プロジェクトに参画。現在はKDDI GPU Cloudのお客さま向けポータル・オーケストレーターの開発チームにて開発者兼プロダクトオーナーを務める。

資 格 / 得 意

Scrum Inc. 認定スクラムマスター
VMware VCP-DCV

趣 味

ギター・DTM
ハロプロ(BEYOOOOOONDS)

本発表で伝えたいこと

#サービス観点での優先順位の決め方

#元工場をデータセンターに転用

#未知の機器・液冷ファシリティへの挑戦

前例の少ない大規模な最先端GPU基盤を

超短期間で法人提供

/ 運用開始したプロジェクトの裏側と実践知

#AIエージェントで過去資産を最大限活用

#企画・開発・運用・ファシリティの四位一体

注意点

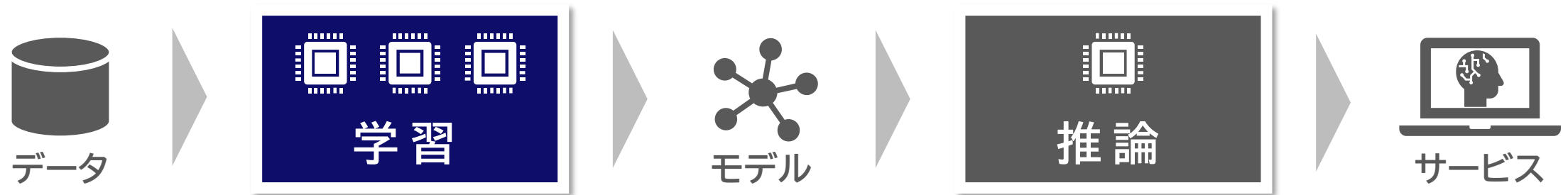
- 本資料は説明用に作成した資料です。
- 内容は作成時点の情報であり、最新の情報と異なる場合があります。
- 本資料に記載された内容は発表者個人の見解に基づくものであり、所属する組織の公式見解を示すものではありません。

本発表の流れ

- 導入・課題の共有
- 前例のない基盤構築に挑むための4つの挑戦
 1. 使えるものは使え！ データセンター編
 2. 使えるものは使え！ ポータル編
 3. 未知を歓迎せよ！ GPU・液冷編
 4. 未知を歓迎せよ！ SLA編
- まとめ

開発・運用しているサービスのご紹介 - 1/2

AIサービス提供の流れ



2026年4月リリース

KDDI GPU Cloud

開発・運用しているサービスのご紹介 - 2/2

KDDI GPU Cloud

最新のGPUサーバーをHaaSとしてクラウド提供する大規模AI学習基盤

スモールスタート&スケール可能



管理コスト不要



電力



冷却



セキュリティ

キャリアグレードの選べるネットワーク



* KDDI Wide Area Virtual Switch
セキュアな通信を提供する閉域型ネットワークサービス

国内完結運用によるソブリン性担保

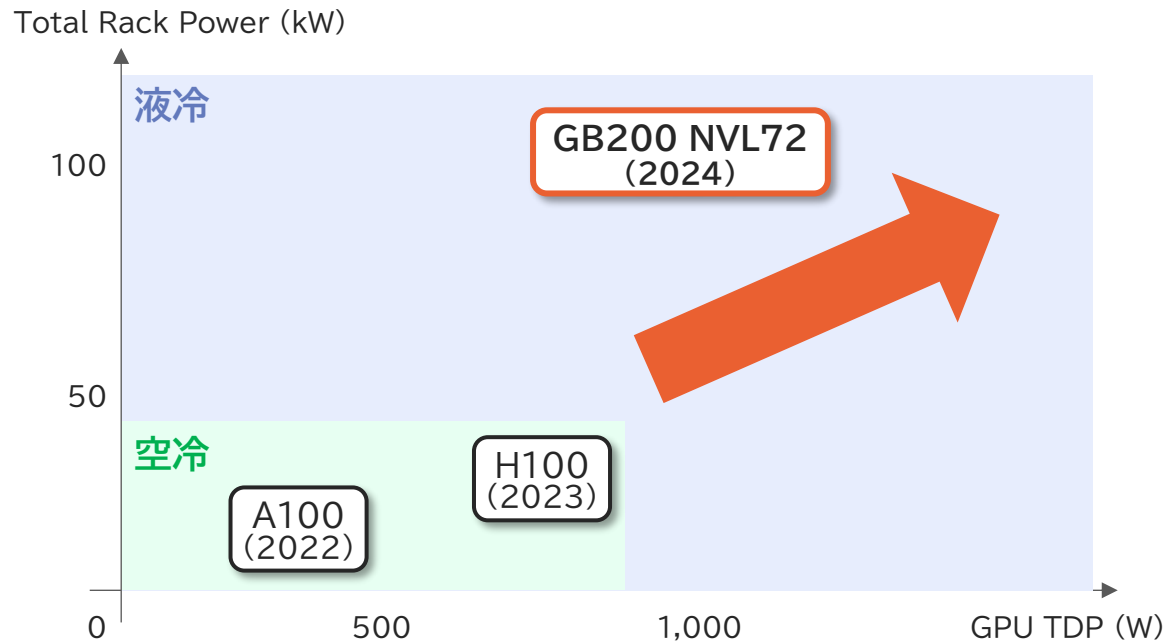


- ✓ データ主権
- ✓ システム主権
- ✓ 運用主権

AI-GPU基盤の特徴

電力と発熱

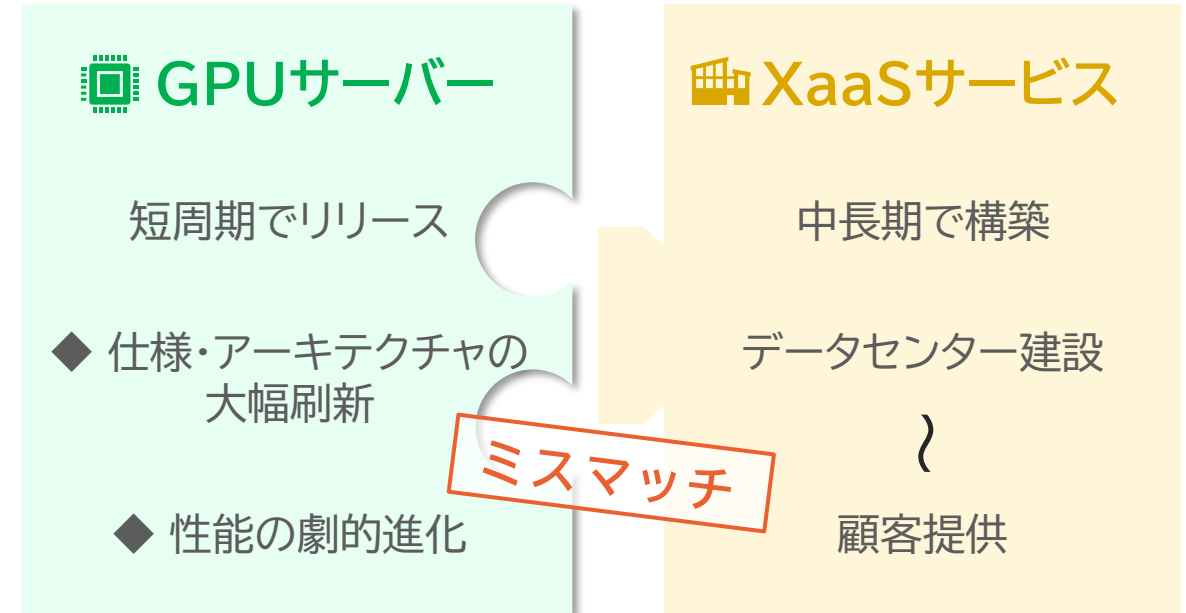
高密度化で消費電力・発熱量が急増



空冷から液冷の時代へ

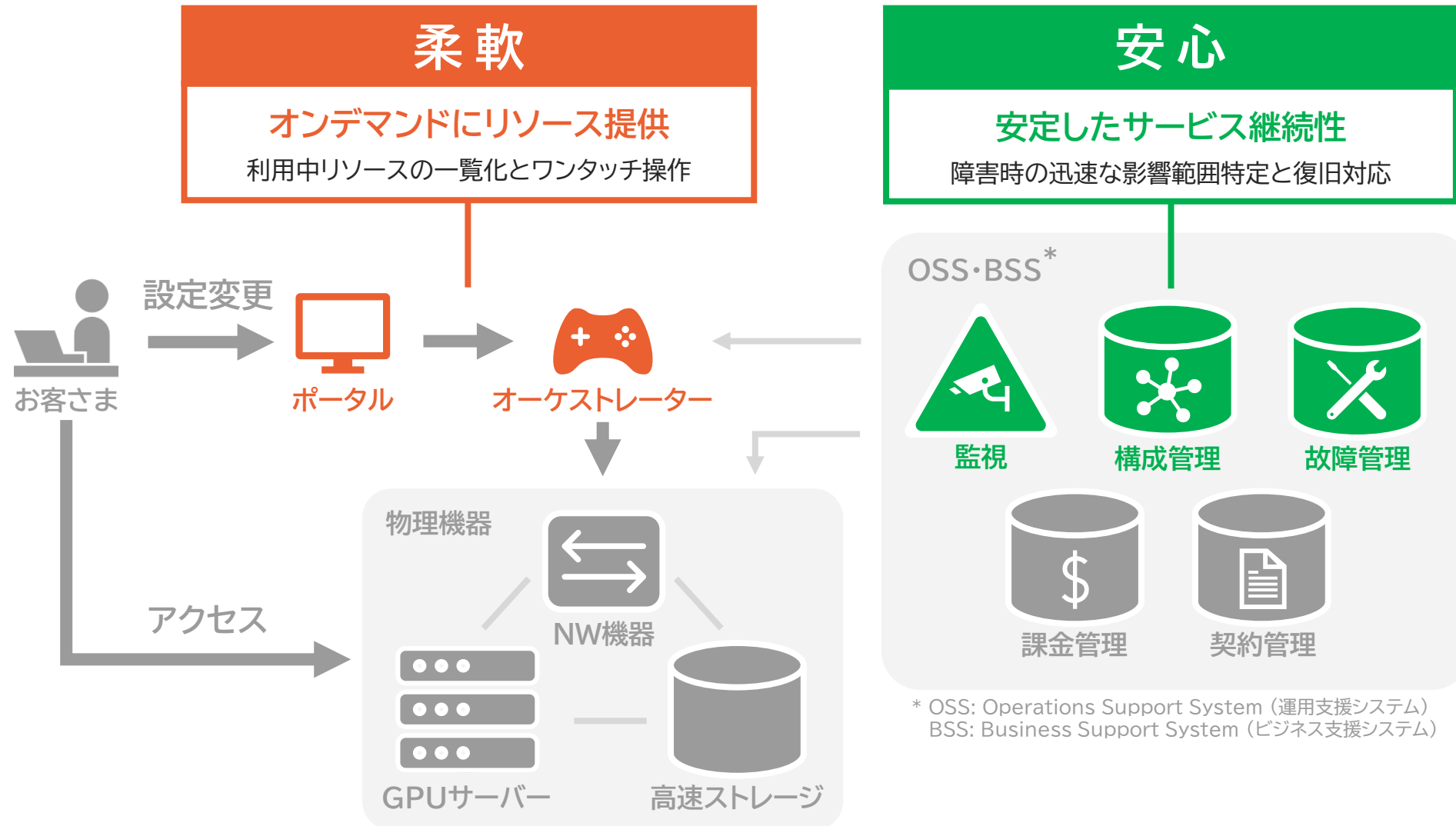
スピード感

AI向けGPUサーバーは急速に進化

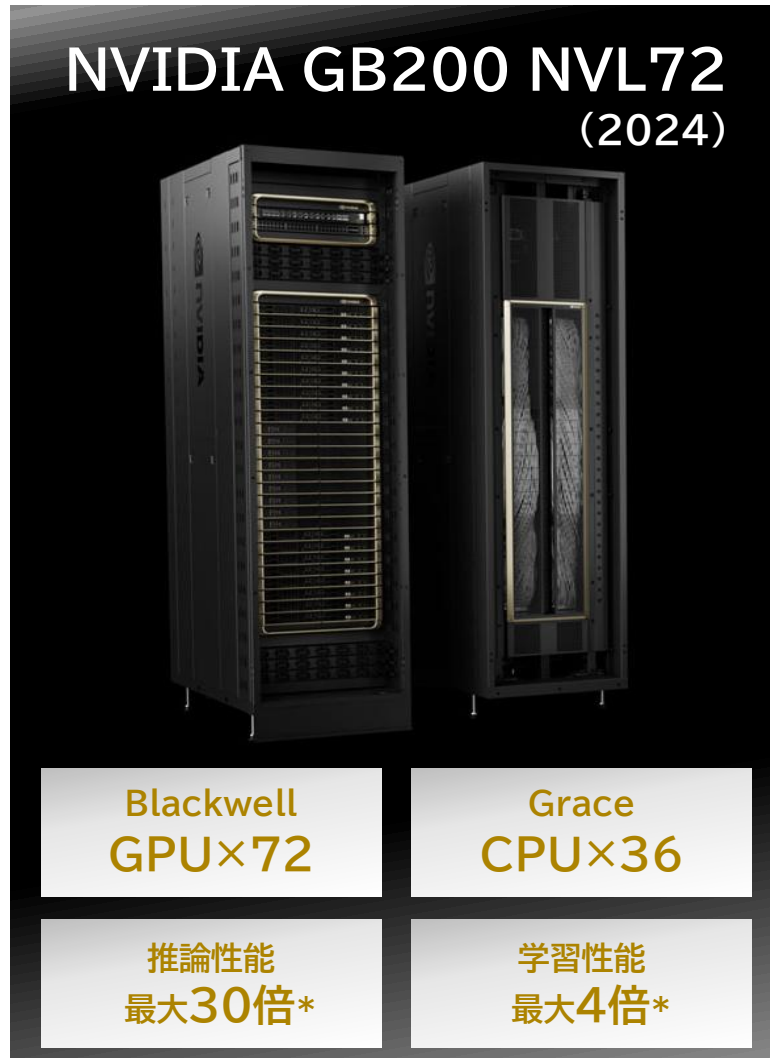


早く顧客提供する必要がある

法人クラウドサービスとして必要なこと



採用した最新のGPUサーバー



*NVIDIA H100 (2022) 比較

<https://www.nvidia.com/en-us/data-center/gb200-nvl72/>

理想の進め方

● 機器仕様の理解

過去事例がない

仕様が特殊

● サービス仕様の理解

仕様確定には個別確認が必要

● システム設計

ラック利用前提のアーキテクチャを分割して提供することでスモールスタート可能に！

● 運用設計

液冷などファシリティーの考慮も必要

● サービスリリース

実機搬入が直前に

前例のない基盤構築に挑むための4つの挑戦

サービスとしては

 早く届けなければ！

 法人サービスとして
仕上げなければ！

課題

前例のない機器 + 前例のないスピード感
= 通常フローで進められない

挑戦 1

使えるものは使え！
データセンター編

挑戦 2

使えるものは使え！
ポータル編

挑戦 3

未知を歓迎せよ！
GPU・液冷編

挑戦 4

未知を歓迎せよ！
SLA編

挑戦1 使えるものは使え！ データセンター編

通常3～4年*程度かかってしまうデータセンター構築

立地

電力

冷却

耐震

建設

セキュリティ

日本初*

元工場をデータセンターとして転用

大型液晶パネルの生産をしていた工場を活用

◎電力・冷却設備などをそのまま再利用

◎新設を最小化し、設計・施工・設備調達を大幅圧縮

約半年で構築、1年未満で顧客提供開始



* 当社調べ

挑戦2 使えるものは使え！ ポータル編（1/2）

HaaSの柔軟性を担うポータル/オーケストレーターを
10年近く開発してきたチーム
その資産を最大限活用することで開発を高速化したい

“すぐ”使える状態にはない宝の山

コード

レビューの記録

議事録

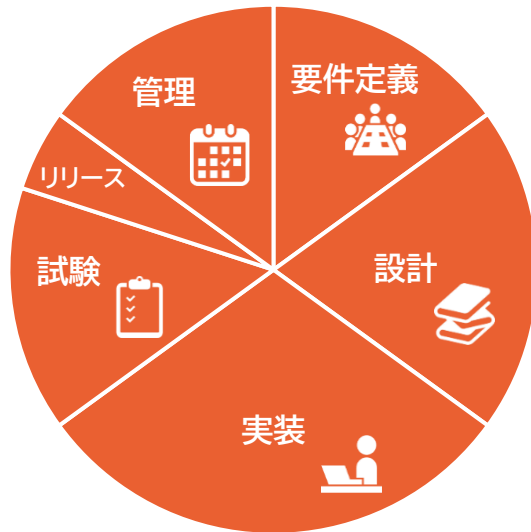
設計資料

今回の仕様

メンバーの変遷

特有の文脈

AI駆動開発によって過去資産をフル活用



直前にリリースされたGitLab Duo Agent Platformを即導入

◎GitLabに蓄積されていた各種データを
標準提供された専門エージェントが自律的に統合

◎実装以外の工程も大幅圧縮：工程別で最大60%の工数削減

機器納入後、わずか3か月でリリース

挑戦2 使えるものは使え！ ポータル編（2/2）

要件定義フェーズでの活用例

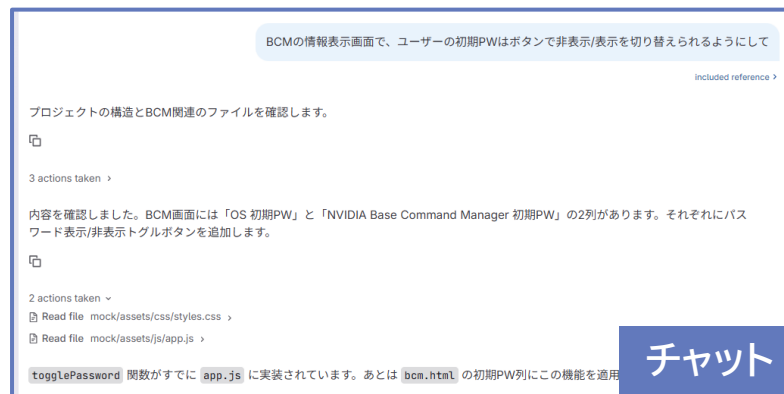
 サービス仕様が曖昧な状態で進めざるを得ない

▶ AI駆動なモックで企画・運用との認識合わせをリード

過去プロダクトの
機能・UIをベースに生成

AIチャットベースで
打ち合わせ中に修正

デプロイまで自動化し、
迅速に共有



設計

検討・ドキュメンテーション

実装

過去実装からブレないコード生成

試験

自動試験生成

オンボーディング

エンゲージメント向上

計画・管理

タスク記述の明確化

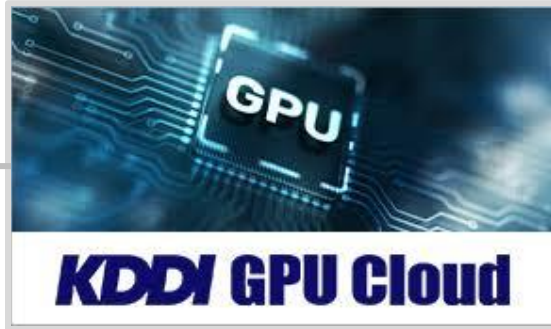
詳しくはウェビナーで



**KDDI 社が明かす
AI駆動開発の実践知**

ソフトウェア開発工程の70%を占める
「見えないボトルネック」を制す

挑戦3 未知を歓迎せよ！ GPU・液冷編



サービス

論理

物理

ファシリ
ティー

① Purpose

情報の一元管理で安心したサービス提供を可能とする
ex.故障の即時検知と影響範囲の即時特定

② Problem

未知の設備”液冷”の登場
運用実績も少ない中で、どのようにすれば
適切な体制、適切な管理を行うかが最大の課題

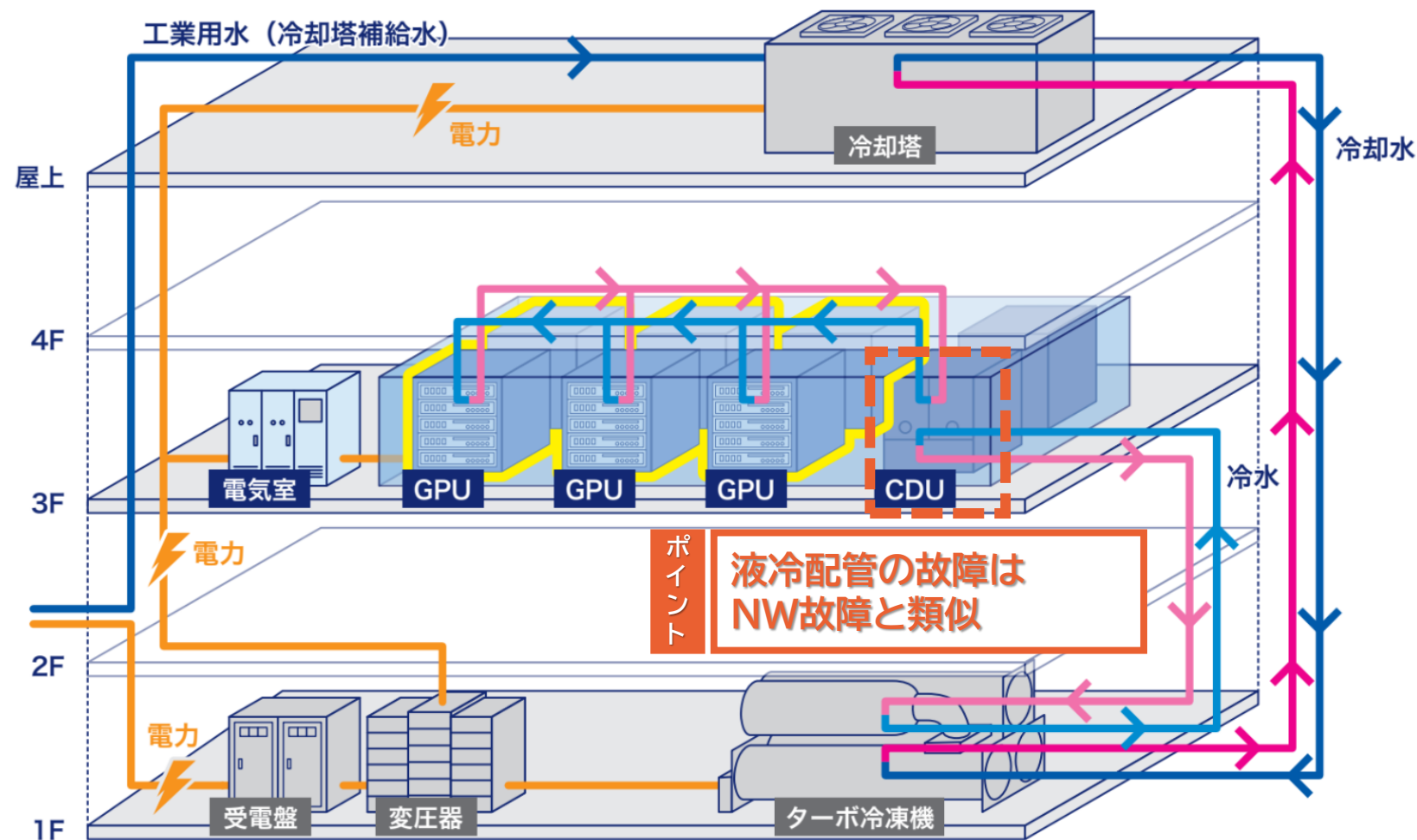
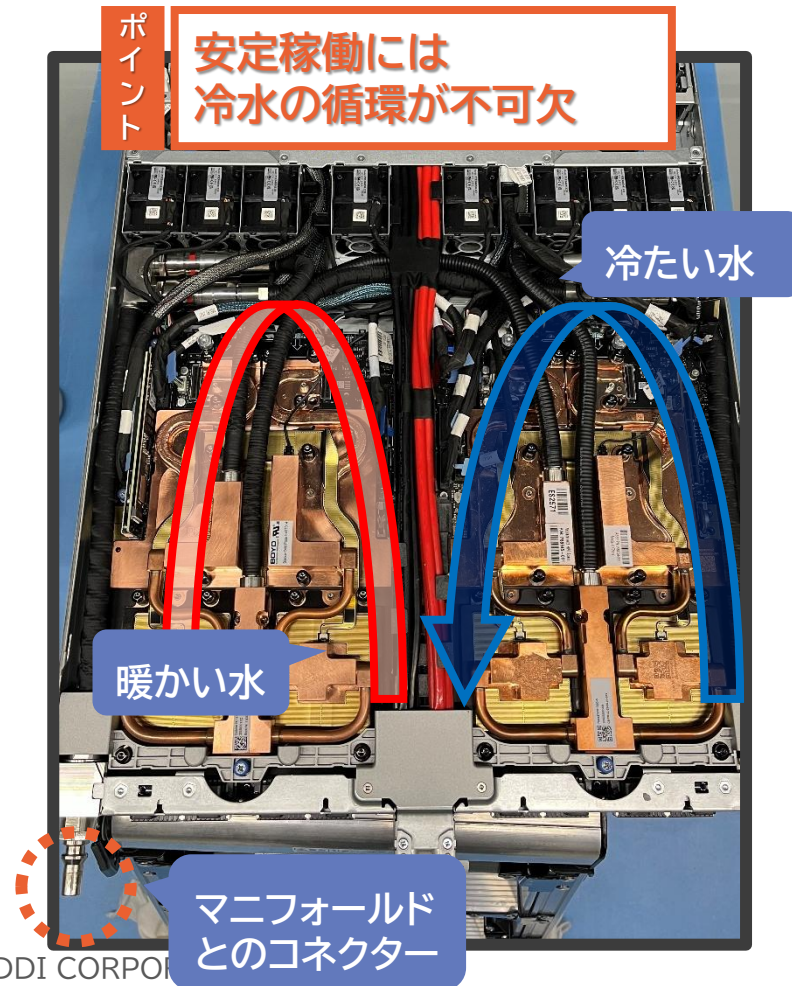
	DCファシリティ保守	ICT機器保守
冷却設備	○	
ラック		○
サーバ		○
液冷	境界があいまいな中で適切な管理方法の確立が必要	

③ Try

ファシリティを含めた構成管理の挑戦

挑戦3 未知を歓迎せよ！ GPU・液冷編

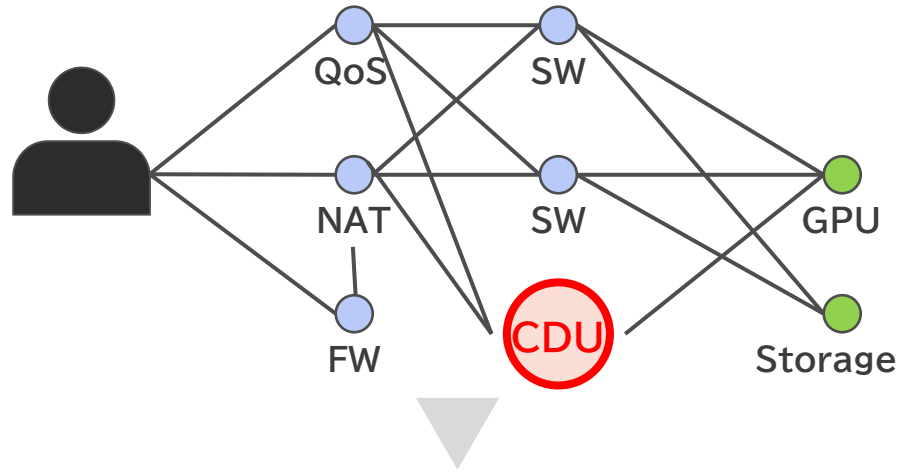
クラウドサービスにおいて
なぜ液冷設備などのファシリティ管理も必要だったか



挑戦4 未知を歓迎せよ！ GPU・液冷編

ポイント

CDUをNW機器に、液冷配管をNWケーブルに見立てて構成管理



活用例：影響ユーザ特定

故障CDU名称で検索

KDDI GPU Cloud 故障管理システム Ver. 1.0

グループ検索 収容/構成検索 管理メニュー

データ取得日時 構成情報: 2026/06/16 14:03 収容情報: 2026/06/16 14:18 ログアウト [at-tanotok@kddi.co.jp]

検索条件

契約管理(M)番号 ※構成検索では無効

GPU物理サーバ機器名

NW機器ホスト名

ストレージホスト名

ファイアウォール名

ゾーン

GPUサーバホスト名

BCMサーバ名

ストレージクラス名

CDU名 moz-ski-xxxx-xxxx-xxx

すべてを選択 すべての選択を解除 障害記録

契約者	サイト	機器種別	機器名
顧客A	gpu-001		
顧客B	gpu-002		
顧客C	gpu-003		
顧客D	gpu-004		

冷却水の供給が止まるGPUと利用ユーザー一覧表示

顧客A gpu-001
顧客B gpu-002
顧客C gpu-003
顧客D gpu-004

影響規模の特定と顧客通知

影響数

影響数:
(MOZU影響数)
お客様契約: 50契約

影響数 : GPUサーバ 80
BCMサーバ 20
ストレージ 30
FW 10



即時顧客通知

ファシリティーからIT機器に至るまでの情報を一元管理することで
法人サービス化するための情報管理を実現！

挑戦4 未知を歓迎せよ！ SLA編

プログレッシブな運用設計サイクル



1 Purpose

安心安全の要となるSLAを達成する



サービスネットワーク、ストレージ
稼働率99.95%



GPU交換作業
作業開始から4時間以内

2 Problem

DC建設からスタートしたプロジェクト
多種多様な”未知”がある複雑な要件を整理し、超短期間で
安心安全なSLAを達成するには、部門の垣根を超えて
一体となってプロジェクトを推進しなければならなかった

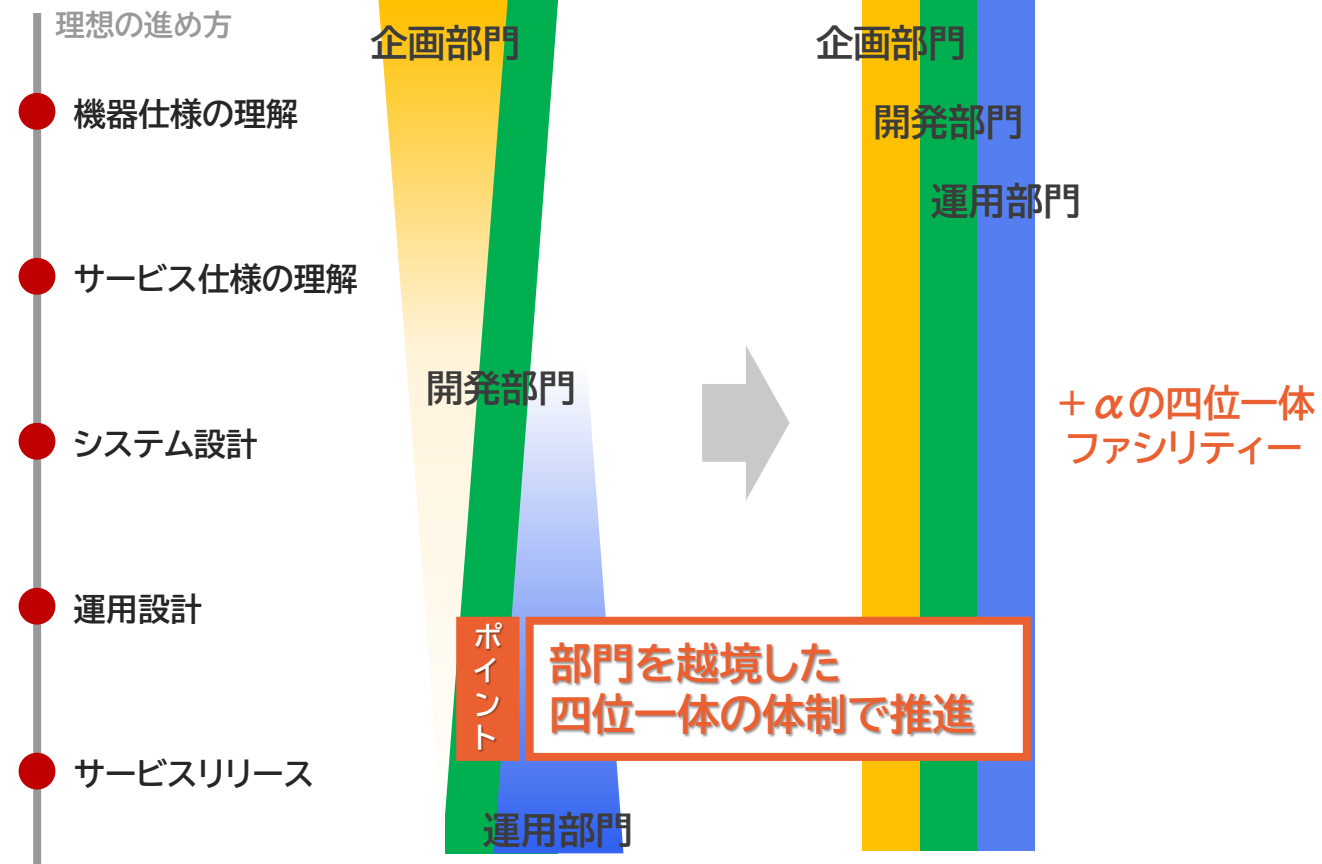
3 Try

企画/開発/運用が四位一体となった仮説駆動の実施
“未知であること”を前提とした攻めの運用設計

挑戦4 未知を歓迎せよ！ SLA編

未知を歓迎するために その①

企画・開発・運用・ファシリティーの四位一体化



企画・開発・運用・ファシリティーが一体となり
意思決定から実行までのリードタイムを大幅に短縮！

挑戦4 未知を歓迎せよ！ SLA編

未知を歓迎するために その② 仮説駆動によるSLA設定と検証



○SLA仮説設定



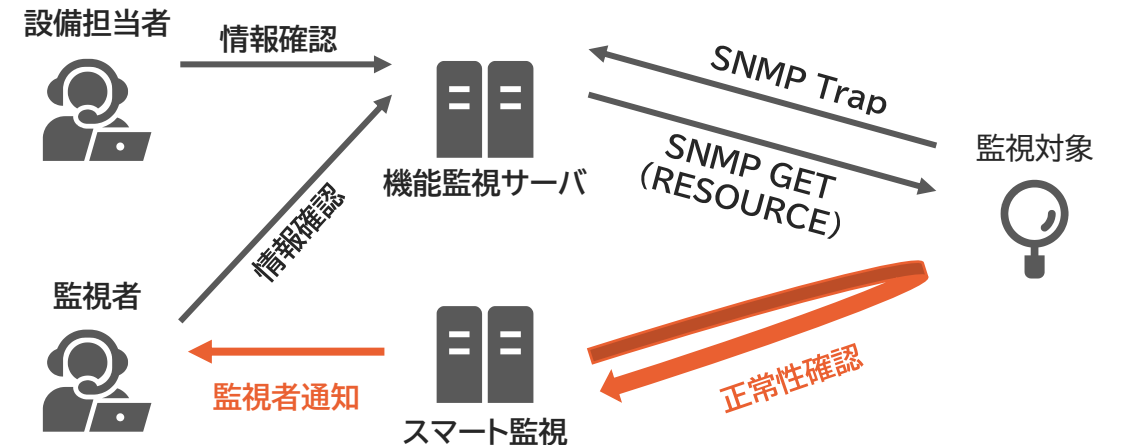
稼働率設定

ポイント

目標に対して仮説立案・検証

- 提供機能不全を検知すれば達成
- 状態確認の稼働を減らせれば達成
- 初動を〇〇分以内にできれば達成

○SLA検証 スマート監視(提供機能監視 + 機器状態の受動監視)



提供価値から仮説を立案、大規模運用実績のノウハウを活かし検証
仮説から確定を最短で達成！

挑戦から得た4つの学び

考えるべきことは山積みだが…

 早く届けなければ！

 法人サービスとして
仕上げなければ！

挑戦 1

まず形にする！

サービスとしての優先順位を決めて過去資産活用

挑戦 2

形が課題を具体化する！

AIの有効利用で形にして議論を加速

挑戦 3

初めての運用設計を喜ぶ！

IT機器の進化で必要な変化する
学び受け入れ常に新しい運用の在り方を考える

挑戦 4

四位一体で仮説駆動する！

決まらないから進められないでは通用しない
企画/開発/運用/ファシリティーが一体となり
仮説検証を繰り返し推進する

「つなぐチカラ」を進化させ、
誰もが思いを実現できる社会をつくる。

KDDI VISION 2030

