

# ACTAPIO



LINEヤフー米国データセンタ技術の最前線  
LLM(大規模言語モデル)と水冷技術への挑戦

Norifumi Matsuya



# 自己紹介



## 自己紹介 in JAPAN

- 松谷 憲文(まつや のりふみ)
- 2000年から2014年まで  
ヤフーにてインフラの設計、構築、運用を担当
- 主な実績
  - プロダクションネットワークの BGP化
  - JPNIC、APNIC、ARINでAS取得  
(AS23816、AS24572、AS18140)
  - バックボーンネットワークを仮想化、VPLS化
  - 国内最大規模の Openstack クラスタ作成
  - 国内データセンターの建設  
(東京、大阪、中部地区、九州地区、東北地区)



米国データセンター見学を経て、米国進出を決意



米国データセンターの見学 '09

## 自己紹介 in U.S.



Actapio データセンター

- 2014年に渡米し、Actapioに出向
- ActapioはLINEヤフー株式会社の米国法人でAI・ビッグデータ用のデータセンター建設、サーバ運用を行う会社
- 現在は会社を経営しながら、OCP(Open Compute Project)サーバの運用やAIビジネス開発などを担当

# AGENDA

- **本日の発表ダイジェスト**
- **第一部：水冷技術で解決できる空調機の風量不足**
  - LLMの出現とサーバリソースの変化
  - データセンタの現状と予測される空調機の風量問題
  - 空調機の風量不足への解決策検討
- **第二部：水冷技術で解決できるサーバ近接性制約**
  - GB200 NVL72 (GPUサーバ)の登場とその背景
  - 電源の480V化(調査中)
- **まとめ**

# 本日の発表ダイジェスト



## 第一部

## 第二部

GTCでGB200 NVL72の発表  
GB200 NVL72の電源が480Vと判明

'22  
Nov

'23  
Mar

'24  
Feb

'24  
Mar

'24  
May

ChatGPT 3.5がリリースされる

ヤフーのサービスでもLLMの本格利用開始

データセンタの風量不足を調査

- ChatGPTは対話型の**LLM**(大規模言語モデル)
- LLMは**トランスフォーマモデル**を使用し、**大量のテキストデータ**で訓練されている
- トランスフォーマは**自然言語処理で高い性能を発揮**するニューラルネットワークアーキテクチャ

## 第一部

## 第二部

GTCでGB200 NVL72の発表  
GB200 NVL72の電源が480Vと判明

'22  
Nov

'23  
Mar

'24  
Feb

'24  
Mar

'24  
May

ChatGPT 3.5がリリースされる

ヤフーのサービスでもLLMの本格利用開始

データセンタの風量不足を調査

- LLMは**大量のデータ**と**膨大なコンピューティングリソース**が必要
- その結果、サーバの**ワークロードが大幅に増加**



- 何か問題がないか、データセンタの**各リソース**をチェックするも**問題ない**ことを確認
- 安心していたところに、データセンタ**空調機の風量が不足する**情報を米国企業から入手



## 第一部

## 第二部

GTCでGB200 NVL72の発表  
GB200 NVL72の電源が480Vと判明

'22  
Nov

'23  
Mar

'24  
Feb

'24  
Mar

'24  
May

ChatGPT 3.5がリリースされる

ヤフーのサービスでもLLMの本格利用開始

データセンターの風量不足を調査

- 空調機の風量不足問題を**水冷技術**で解決できるか検証
- 水冷技術の利用によって、
  - **冷却効率が上昇**
  - 空調機の**風量不足が解決**
  - **より多くのサーバを設置可能**

## 第一部

## 第二部

GTCでGB200 NVL72の発表  
GB200 NVL72の電源が480Vと判明

'22  
Nov

'23  
Mar

'24  
Feb

'24  
Mar

'24  
May

ChatGPT 3.5がリリースされる

ヤフーのサービスでもLLMの本格利用開始

データセンタの風量不足を調査

- 2024年3月、GTC(GPU Technology Conference)でNVIDIAが新しいGPUサーバである**GB200 NVL72**を発表
  - GB200 NVL72には**水冷技術**が必須となっている
- 水冷技術が必要な理由を**米国企業からヒアリング**
  - 理由はモデル学習時間の短縮を図るためにGPU間で**高帯域の通信が必要**

## 第一部

## 第二部

GTCでGB200 NVL72の発表  
GB200 NVL72の電源が480Vと判明

'22  
Nov

'23  
Mar

'24  
Feb

'24  
Mar

'24  
May

ChatGPT 3.5がリリースされる

ヤフーのサービスでもLLMの本格利用開始

データセンターの風量不足を調査

- ・GB200 NVL72の電源が**480V入力(高電圧)**
- ・現在、わかっている状況までを本編で説明



## LLMの出現とサーバリソースの変化

イマココ！



第一部

第二部

GTCでGB200 NVL72の発表  
GB200 NVL72の電源が480Vと判明

'22  
Nov

'23  
Mar

'24  
Feb

'24  
Mar

'24  
May

ChatGPT 3.5がリリースされる

ヤフーのサービスでもLLMの本格利用開始

データセンタの風量不足を調査



2022年11月 ChatGPT 3.5リリース



2022年11月 ChatGPT 3.5リリース

今までのAIと何が違うの？



## 従来型「タブラー形式データ学習」と 「LLMのトランスフォーマー学習」の違い

特徴	タブラー形式データ	LLM(大規模言語モデル)
データの種類	表形式データ(例: Excel, CSV)	テキストデータ
得意なこと	数値予測、分類、データ分析	文章生成、翻訳、チャットボット
トレーニングデータ	構造化データ(数値、カテゴリ)	大量の文書データ
モデルの規模	比較的小規模 (特徴量やデータ数に依存)	大規模(数十億のパラメータ)
柔軟性	特定のデータセットに特化	多様なタスクに対応可能

## 従来型「タブラー形式データ学習」と 「LLMのトランスフォーマー学習」の違い

特徴	タブラー形式データ	LLM(大規模言語モデル)
	(Excel, CSV)	テキストデータ
	分析	文章生成、翻訳、チャットボット
		<b>大量の文書データ</b>
モデルの規模	比較的小規模 (特徴量やデータ数に依存)	<b>大規模(数十億のパラメータ)</b>
柔軟性	特定のデータセットに特化	多様なタスクに対応可能

今までのAIと比べて桁違いのデータ  
が必要😱



## ヤフーのサービスでもLLMを利用



- ヤフーではユーザの投稿など「言語を理解」するためにLLMを使用
- LLMの利用によって大きな精度向上を達成



ただし、大きな課題も

## ヤフーのサービスでもLLMを利用



- ヤフーではユーザの投稿など「言語を理解」するためにLLMを使用
- LLMの利用によって大きな精度向上を達成

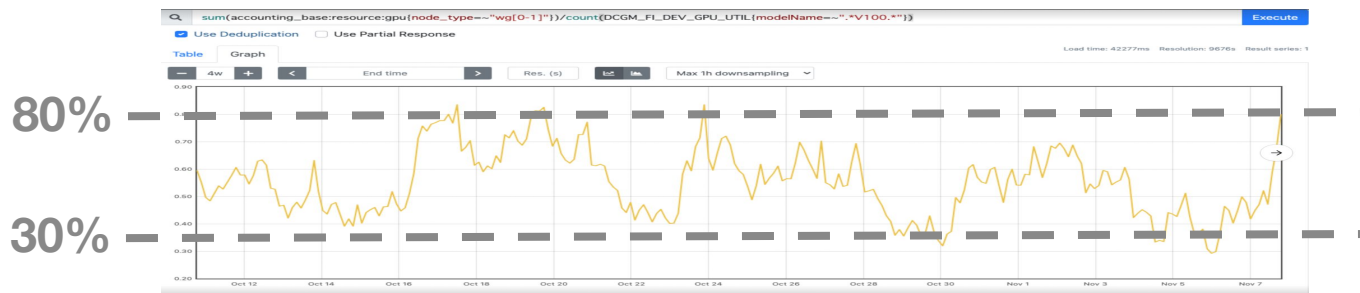


**ただし、大きな課題も**



## サーバワークロードの変化 (with A100 GPU)

タブラー形式データ ワークロード 30%-80%が1週間ほど



LLM ワークロード **80%-100%**が1、2ヶ月継続





## データセンタの現状と予測される空調機の風量問題

イマココ！



第一部

第二部

GTCでGB200 NVL72の発表  
GB200 NVL72の電源が480Vと判明

'22  
Nov

'23  
Mar

'24  
Feb

'24  
Mar

'24  
May

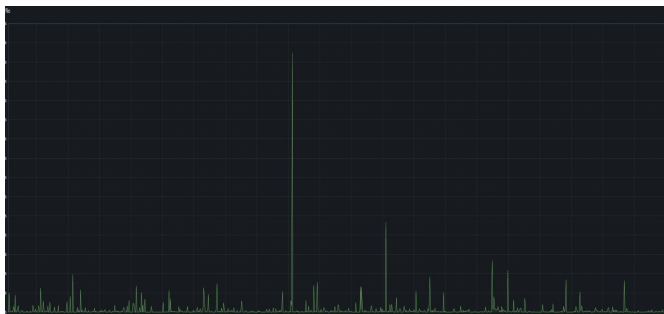
ChatGPT 3.5がリリースされる

ヤフーのサービスでもLLMの本格利用開始

データセンタの風量不足を調査

## ネットワーク: LLMと別システム(Hadoop)のサーバ帯域比較

Hadoopのサーバインターフェースのグラフ



LLMのサーバインターフェースのグラフ



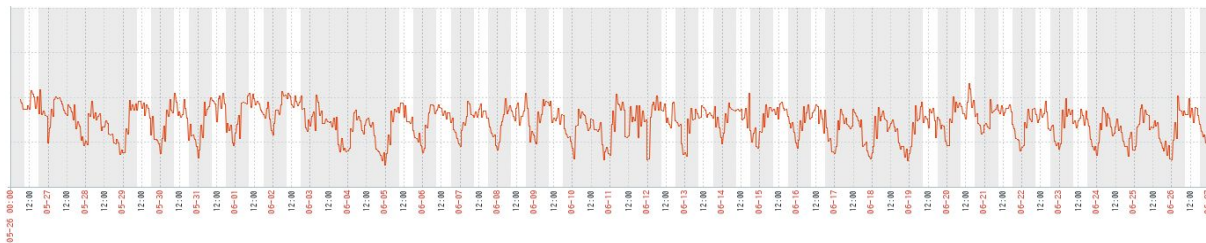
	サーバポートの帯域キャパシティ	利用状況
<u>Hadoopのサーバインターフェース</u>	25 Gbps	13.50 Gbps (54%)
<u>LLMのサーバインターフェース</u>	100 Gbps	23.5 Gbps (24%)

- 帯域も足りており、エラーやドロップも見当たらない
- **ネットワークにおける課題は発生していない**
- ※モデル作成時に少し工夫していることも影響

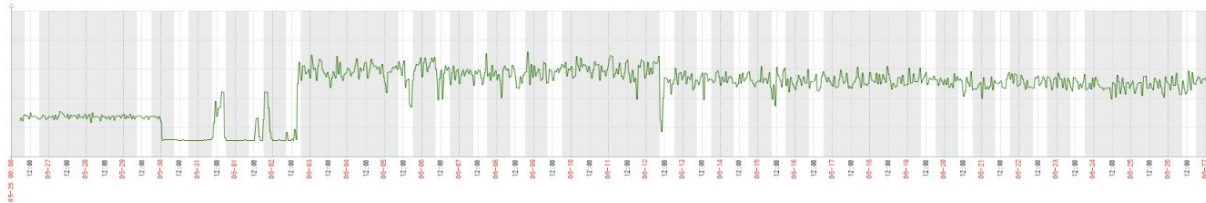


## 電力：LLMと別システム（Hadoop）のサーバ電力比較

### Hadoopのサーバ (最大稼働率：76.12%)



### LLMのサーバ (最大稼働率：95.44%)



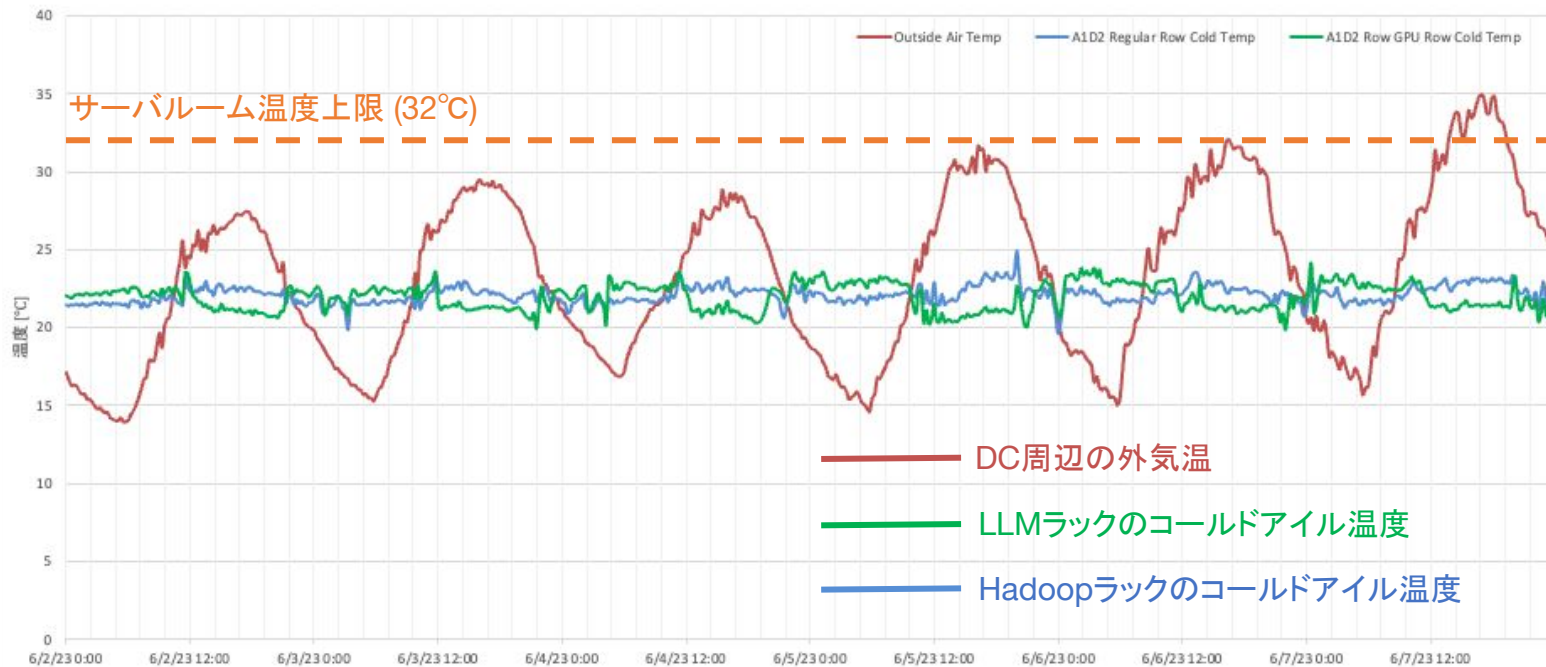
Hadoop Rack  
(30 nodes / rack)

LLM GPU Rack  
(2 nodes / rack)



サーバワークロードが長期間にわたって高負荷状態にあるため、消費電力が上昇

## 冷却: LLMと別システム (Hadoop) の温度比較



温度の観点ではLLMとHadoopのラック列とで大きな差異はなく、  
温度上の問題は発生していない

現時点ではネットワーク、電力、冷却に関して概ね問題なし



米国企業と水冷技術について対話

- GPUのTDP (Thermal Design Power)が増加傾向
- TDPの増加によりサーバ風量(ファンスピード)が最大に
- サーバ風量増加により空調機の風量が増加する懸念が明らかになった



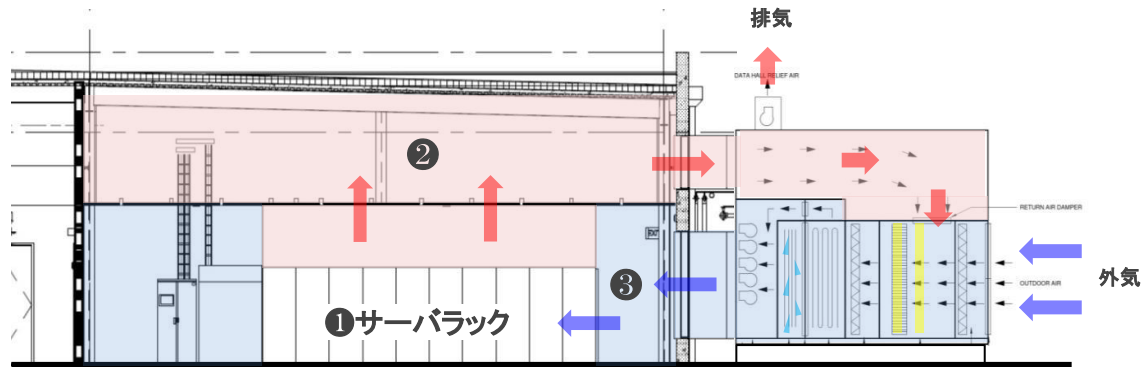
現時点ではネットワーク、電力、冷却に関して概ね問題なし？



米国企業と水冷技術について対話

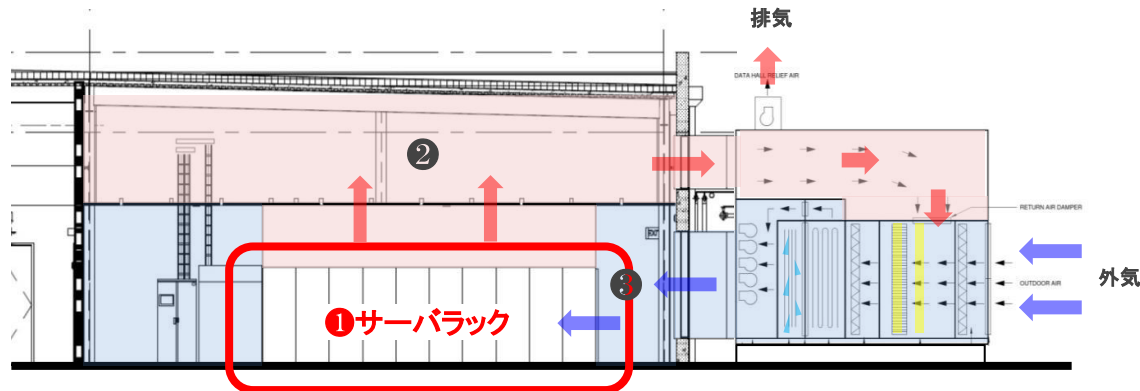
- GPUの**TDP** (Thermal Design Power)が**増加傾向**
- TDPの増加により**サーバ風量** (ファンスピード)が**最大に**
- サーバ風量増加により**空調機の風量が不足する懸念** が明らかになった

## Actapioデータセンターの空調方式と断面図



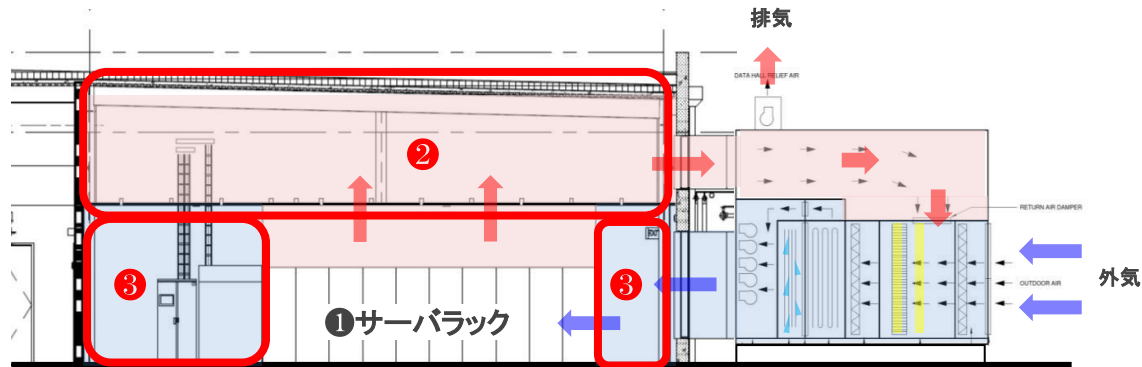
- ①のサーバラックの風量が増加
- ②(ホットアイル)の圧力増加と③(コールドアイル)の圧力減少が発生
- ②、③部分に対して排気、排出できる、十分な空調機の風量が必要となる
- 空調機の風量不足でデータセンター内部の圧力バランスが保てなくなると温度制御が出来なくなる

## Actapioデータセンタの空調方式と断面図



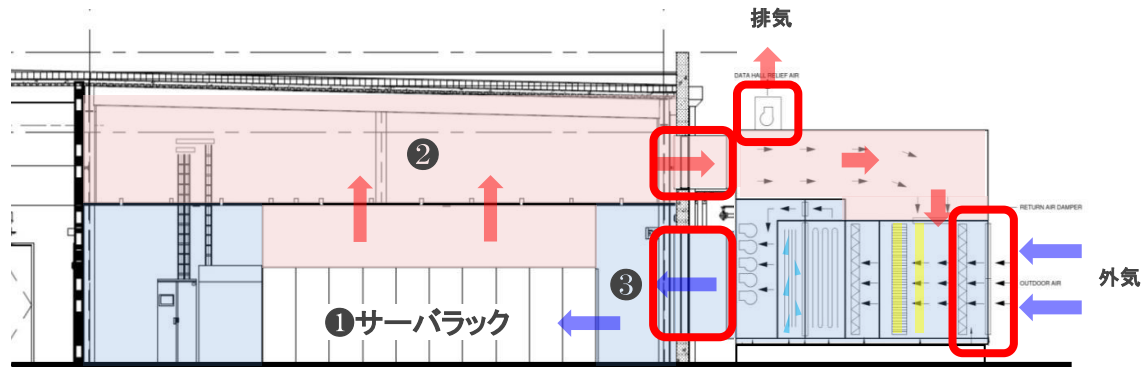
- **①のサーバラックの風量が増加**
- **②**(ホットアイル)の圧力増加と**③**(コールドアイル)の圧力減少が発生
- **②**、**③**部分に対して排気、排出できる、十分な空調機の風量が必要となる
- 空調機の風量不足でデータセンタ内部の圧力バランスが保てなくなると温度制御が出来なくなる

## Actapioデータセンタの空調方式と断面図



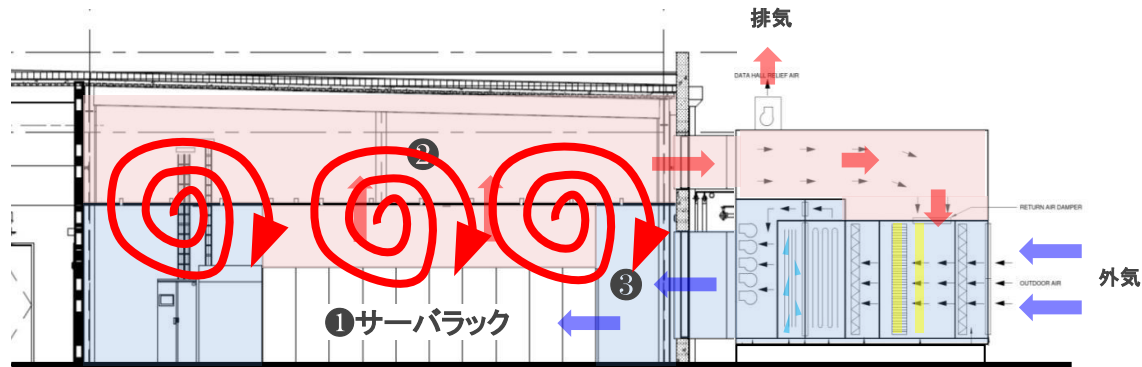
- ①のサーバラックの風量が増加
- ②ホットアイルの圧力増加 と ③コールドアイルの圧力減少 が発生
- ②、③部分に対して排気、排出できる、十分な空調機の風量が必要となる
- 空調機の風量不足でデータセンタ内部の圧力バランスが保てなくなると温度制御が出来なくなる

## Actapioデータセンターの空調方式と断面図



- ①のサーバラックの風量が増加
- ②(ホットアイル)の圧力増加と③(コールドアイル)の圧力減少が発生
- ②、③部分に対して排気、排出できる、十分な**空調機の風量が必要となる**
- 空調機の風量不足でデータセンター内部の圧力バランスが保てなくなると温度制御が出来なくなる

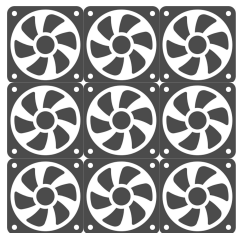
## Actapioデータセンタの空調方式と断面図



- ①のサーバラックの風量が増加
- ②(ホットアイル)の圧力増加と③(コールドアイル)の圧力減少が発生
- ②、③部分に対して排気、排出できる、十分な空調機の風量が必要となる
- 空調機の風量不足でデータセンタ内部の圧力バランスが保てなくなると**温度制御が出来なくなる**

データセンタの空調機が問題なく稼働するためには、  
**空調機の風量**が**サーバの風量**を上回る必要があります

サーバの風量



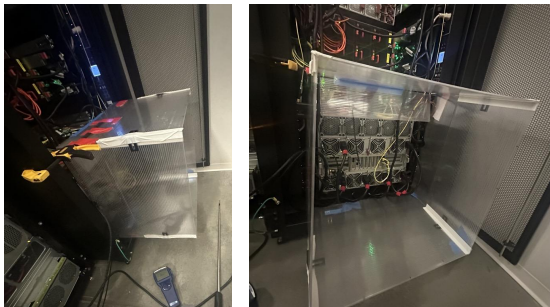
空調機の風量



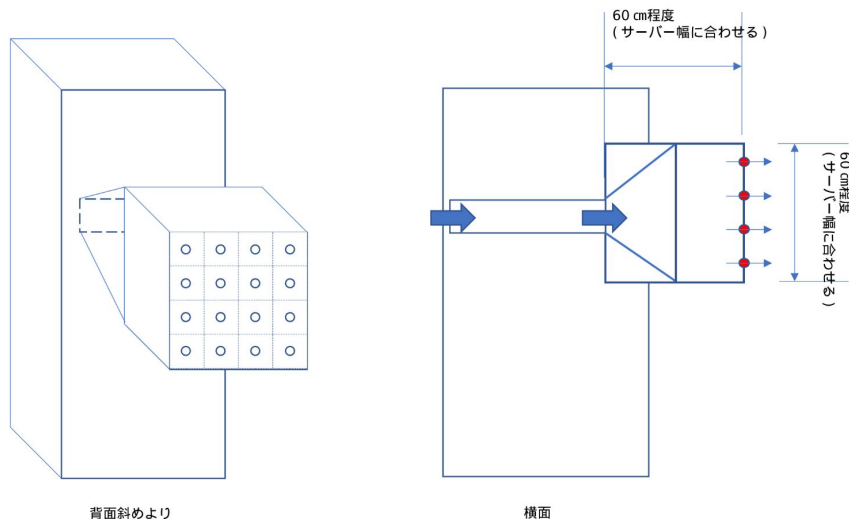
## 空冷GPUサーバの風量を測定

- サーバは**フル負荷**で測定
- サーバ背面をプラスチック製の風防で囲う
- 断面を4x4の区間に糸で区切り、それぞれの区間を**風速計**で測定

## 実際の様子



## 測定方法



## 風速計

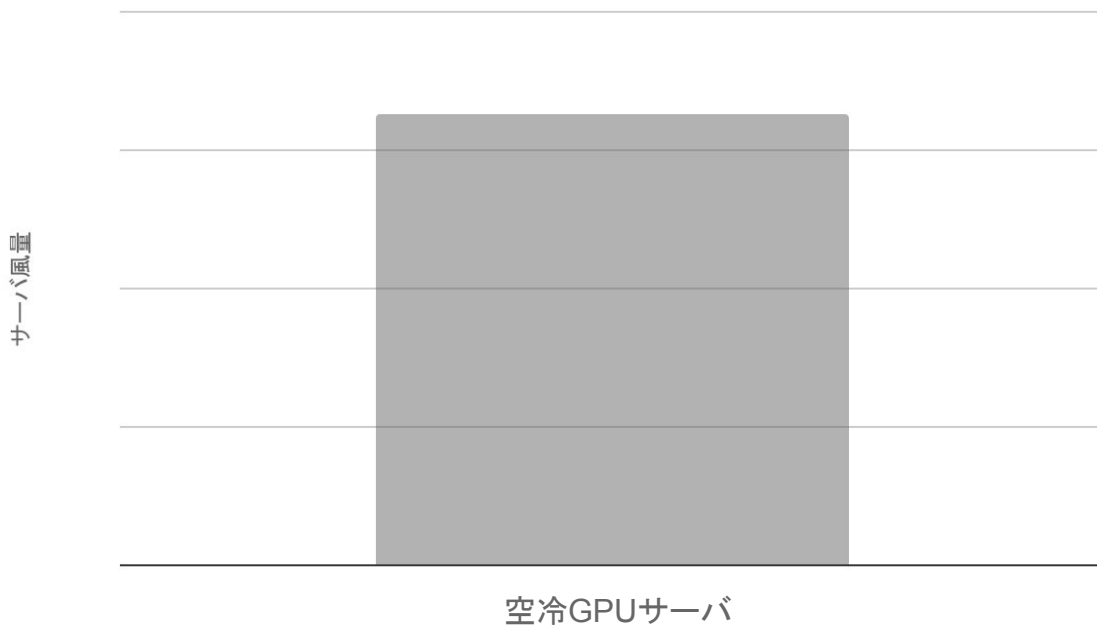






## Actapioの2MWのサーバールームに LLM用の空冷GPUサーバをフル設置した場合のシナリオ

空冷GPUサーバの2MW分のトータル風量

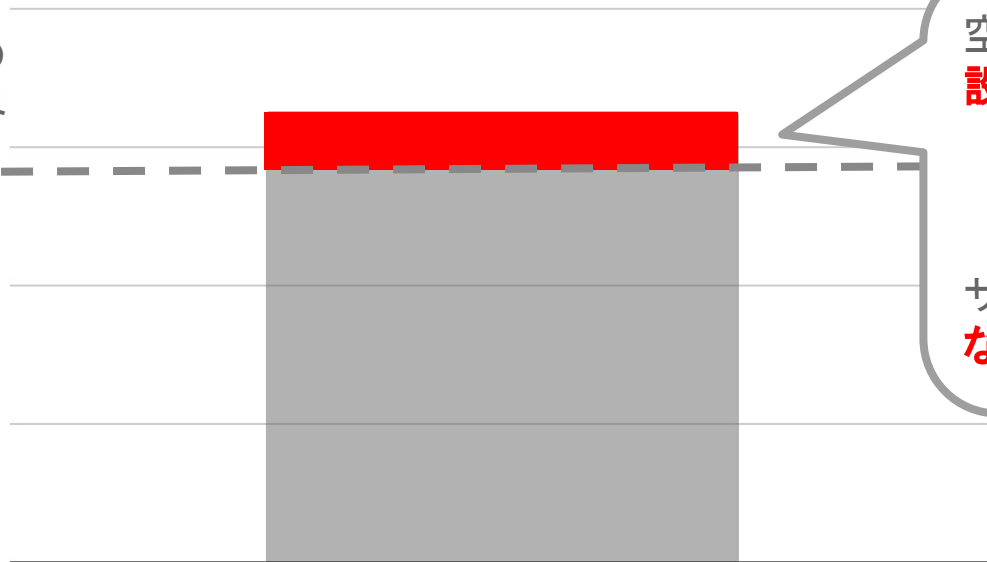


## Actapioの2MWのサーバールームに LLM用の空冷GPUサーバをフル設置した場合のシナリオ

空冷GPUサーバの2MW分のトータル風量

空調機の  
風量限界

サーバ風量



空冷GPUサーバ風量が空調機の  
設備限界を上回る



サーバールームの温度制御が出来  
なくなる

空冷GPUサーバ



## 空調機の風量不足への解決策検討

イマココ！



第一部

第二部

GTCでGB200 NVL72の発表  
GB200 NVL72の電源が480Vと判明

'22  
Nov

'23  
Mar

'24  
Feb

'24  
Mar

'24  
May

ChatGPT 3.5がリリースされる

ヤフーのサービスでもLLMの本格利用開始

データセンターの風量不足を調査

## 空調機の風量不足への解決案

案① 設置サーバ数の抑制

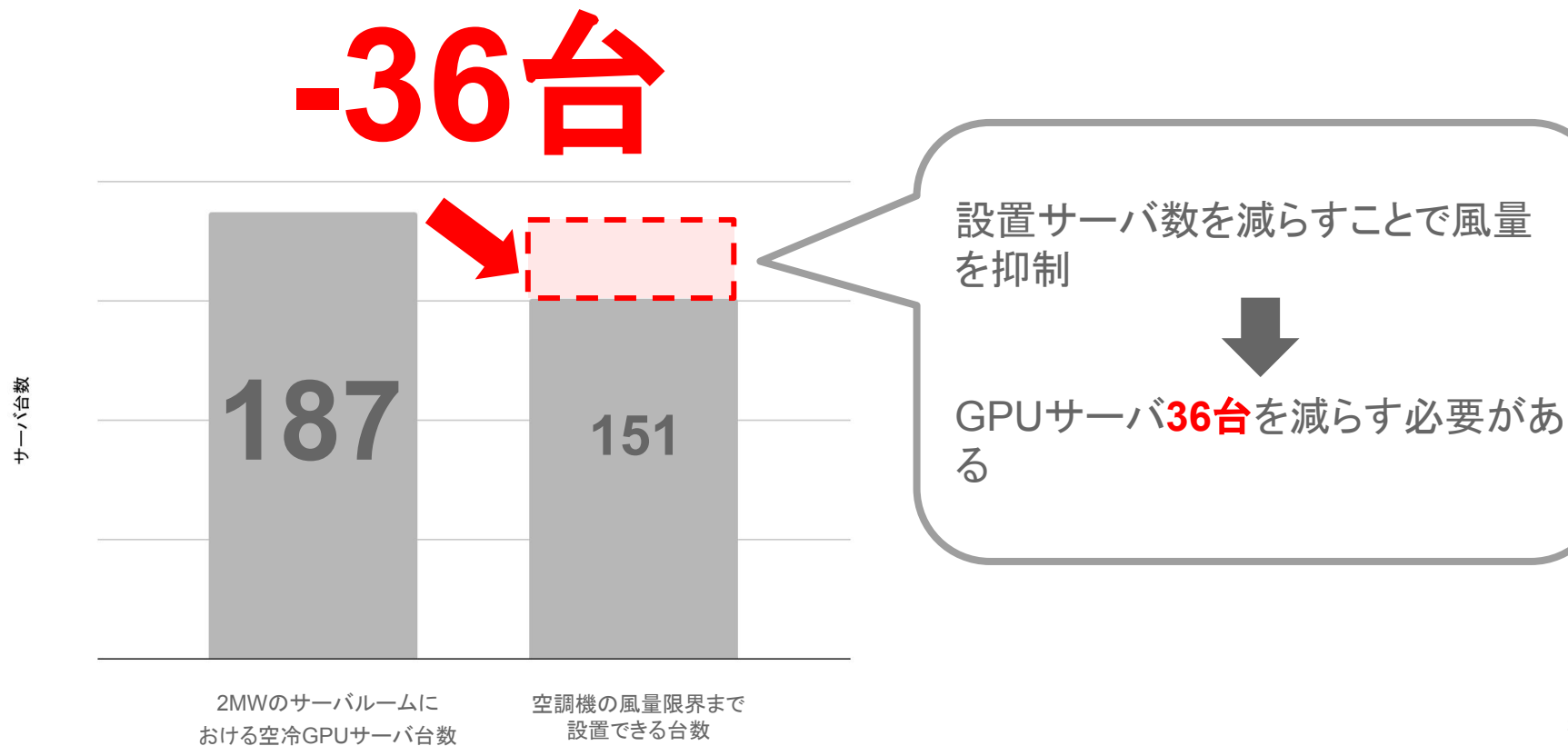
案② 空調機の増強

- a. 空調機の増設
- b. 空調機のスペック増強

案③ 水冷方式の検討: DLC (Direct Liquid Cooling)

- a. L to L: Liquid to Liquid
- b. AALC: Air Assisted Liquid Cooling

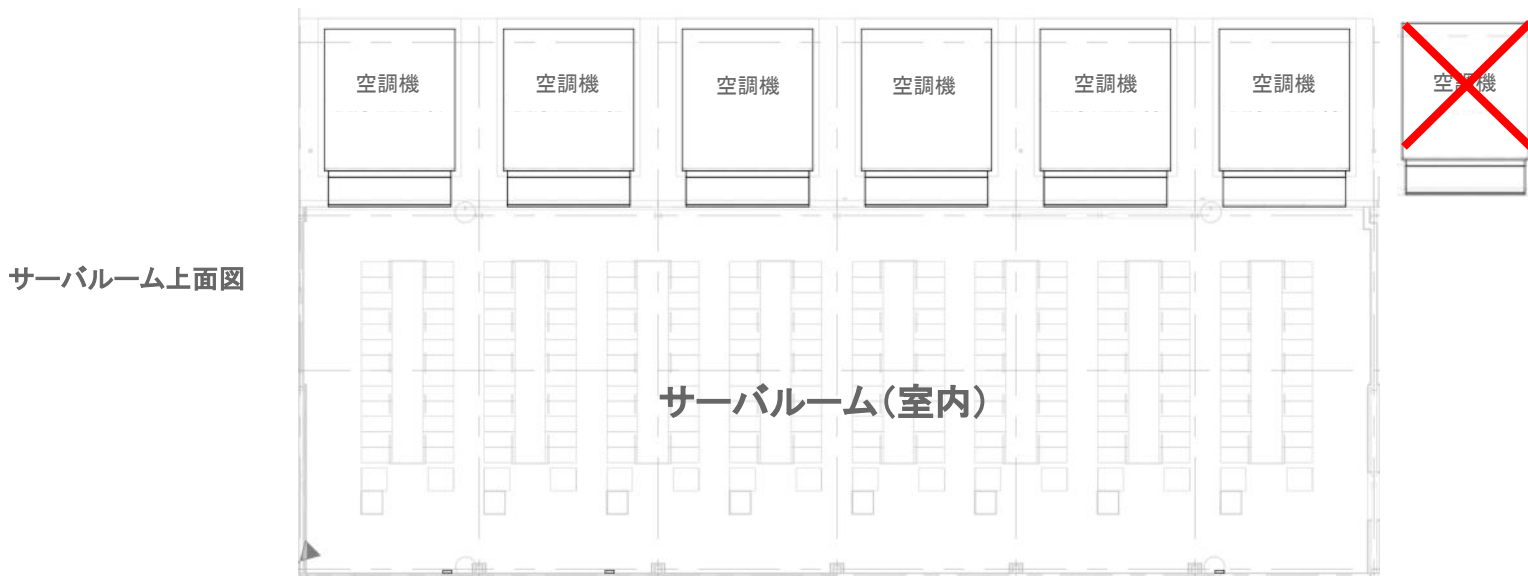
## 案① 設置サーバ数の抑制



## 案② 空調機の増強

### a. 空調機の増設

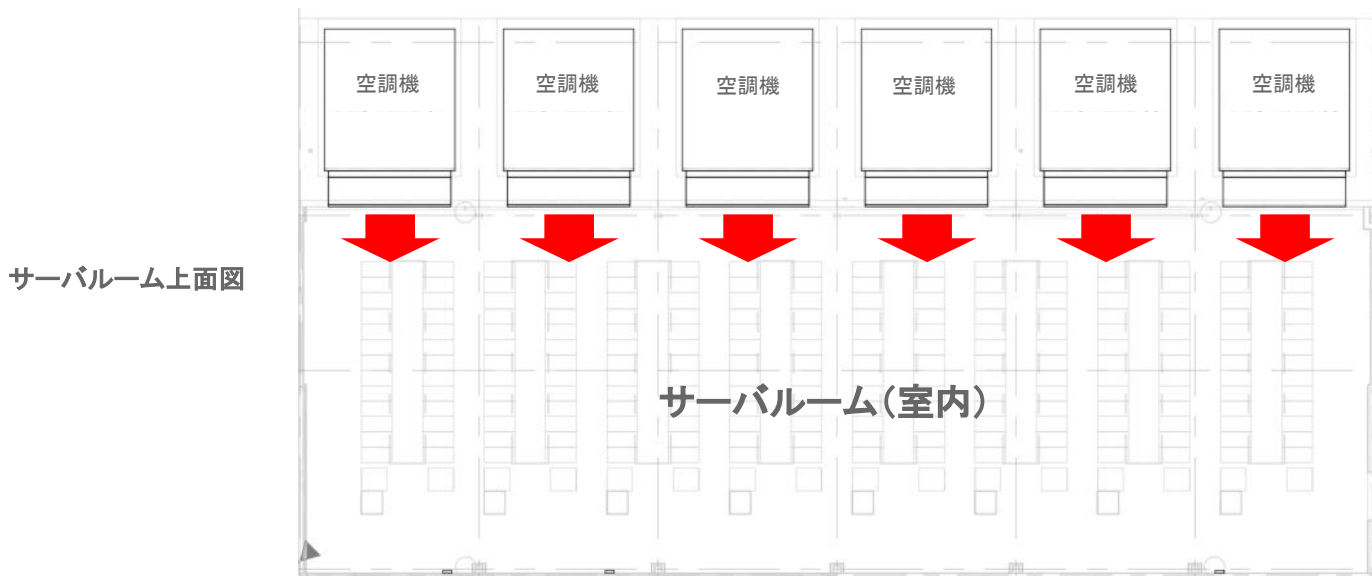
データホールのレイアウト(下図)のため物理的に**増設負荷**



## 案② 空調機の増強

### b. 空調機のスペック増強

既存空調機のスペックアップはコスト、工事レベル含め**増強は困難**



### 案③ 水冷方式の検討

水冷方式(DLC)は2種類に分類できます

DLC(Direct Liquid Cooling)

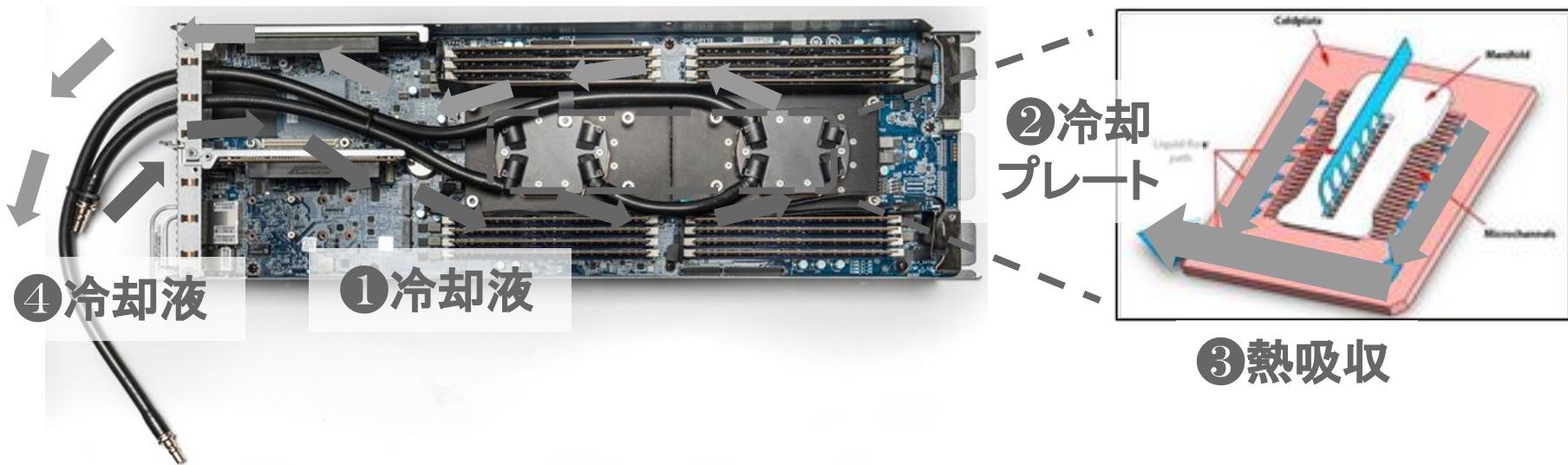
1. L to L : Liquid to Liquid

2. **AALC : Air Assisted Liquid Cooling**

水冷方式の中でも**AALC**を検討しています、  
その選択理由を説明します

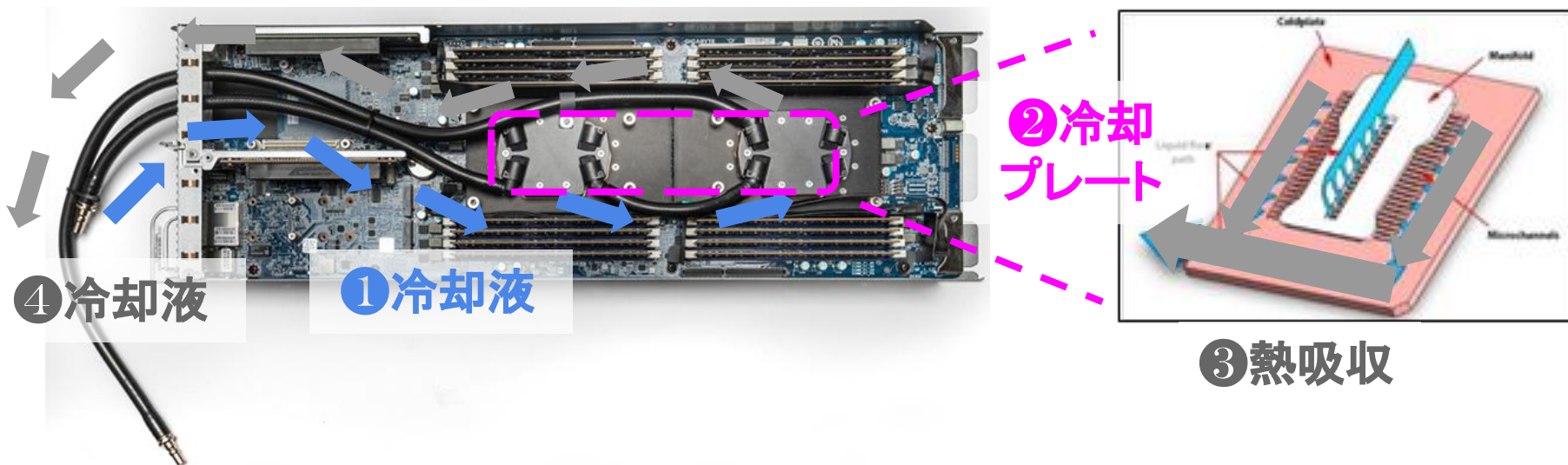


## DLC(Direct Liquid Cooling)について



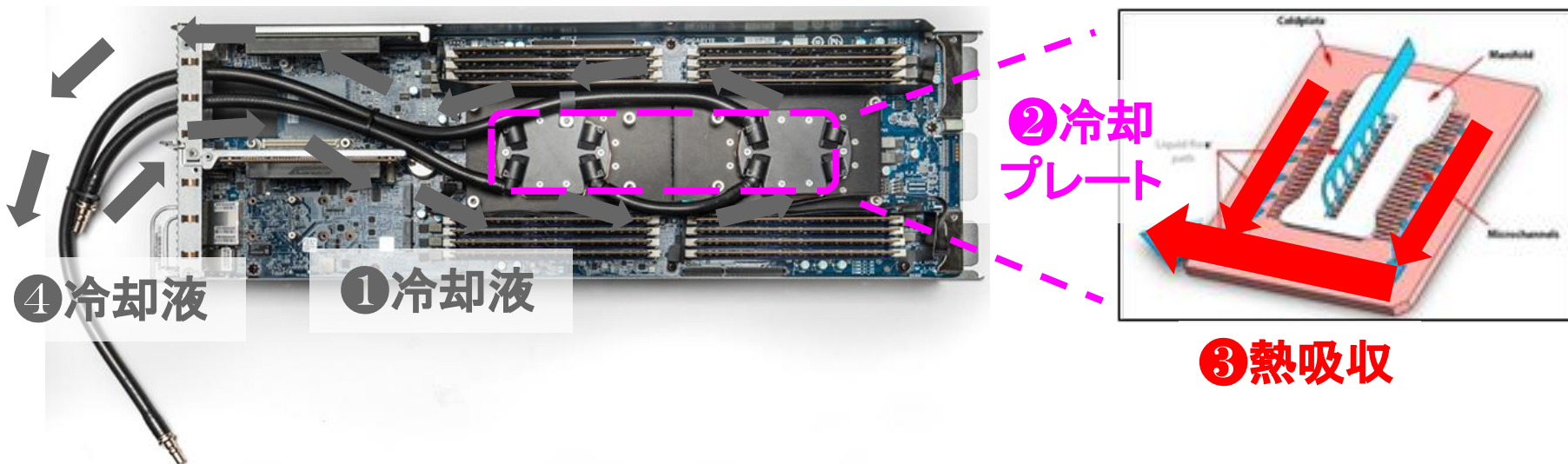
- ① 冷たい冷却液が②冷却プレートに送られる
- ②冷却プレートがプロセッサから③熱を吸収
- ④温かい冷却液は熱交換器に送られる

## DLC (Direct Liquid Cooling) について



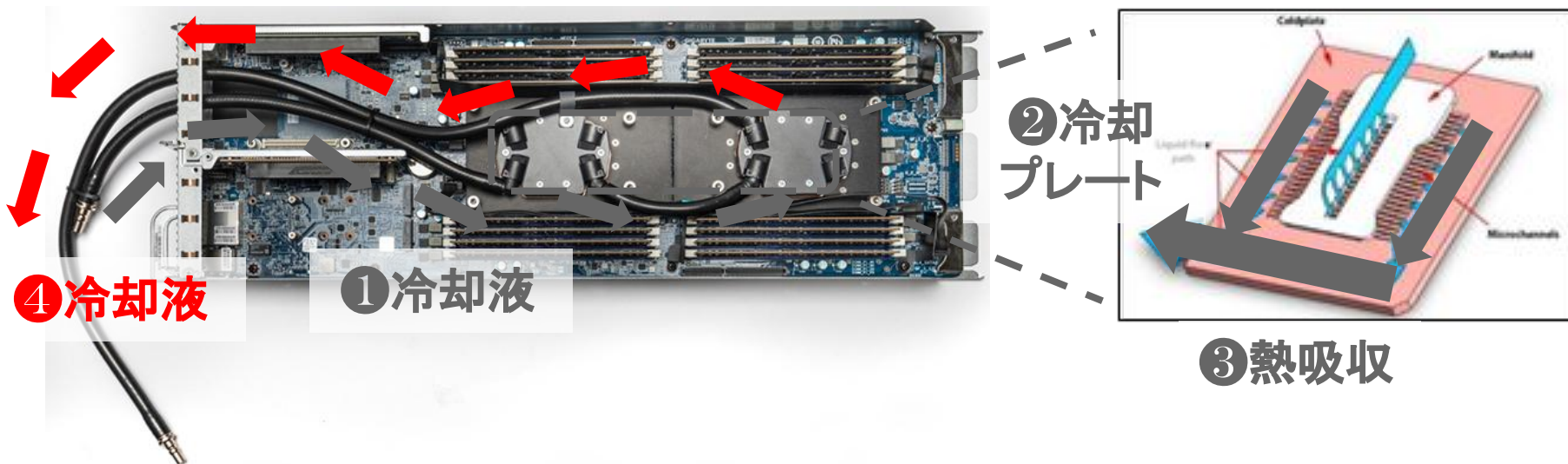
- ① 冷たい冷却液 が ② 冷却プレート に送られる
- ② 冷却プレートがプロセッサから ③ 熱を吸収
- ④ 温かい冷却液 は熱交換器に送られる

## DLC (Direct Liquid Cooling) について



- ①冷たい冷却液が②冷却プレートに送られる
- ②冷却プレートがプロセッサから③熱を吸収
- ④温かい冷却液は熱交換器に送られる

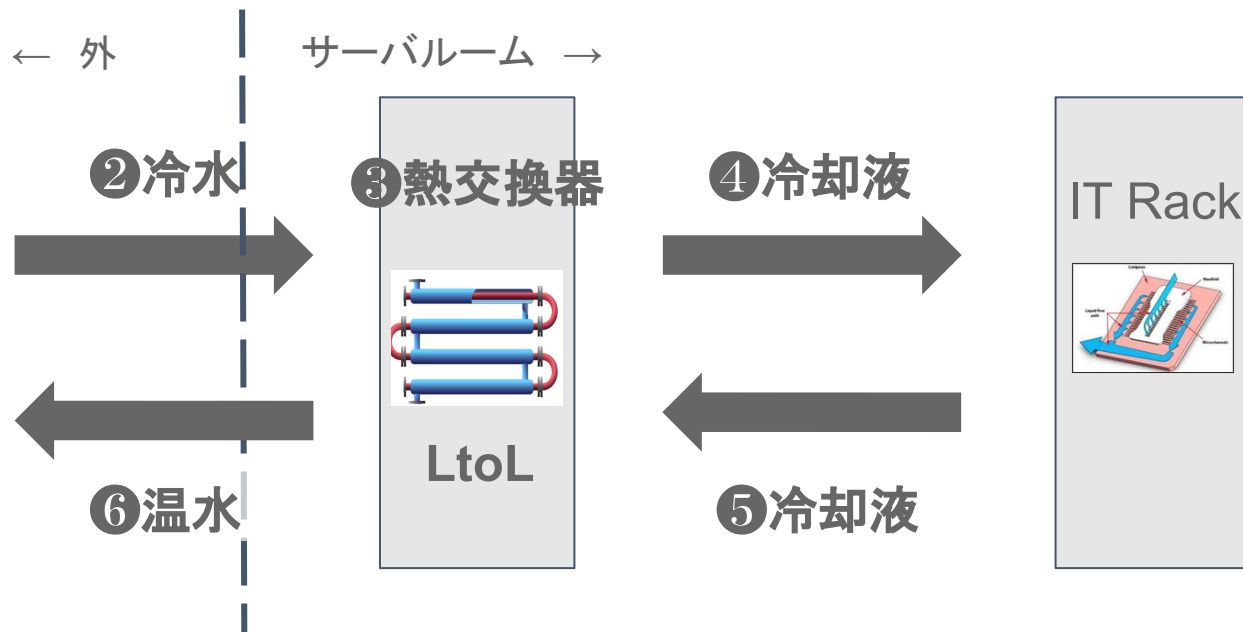
## DLC (Direct Liquid Cooling) について



- ① 冷たい冷却液 が ② 冷却プレート に送られる
- ② 冷却プレート が プロセッサ から ③ 熱を吸収
- ④ 温かい冷却液 は 熱交換器 に送られる

## LtoLによる冷却

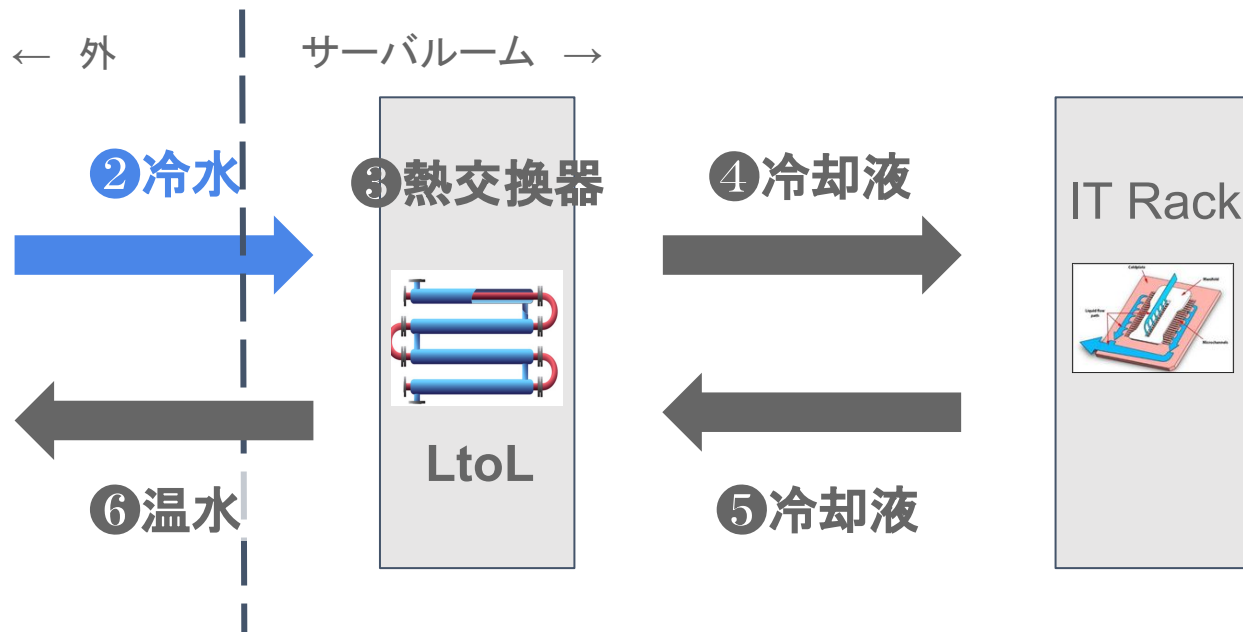
## ①チラー/冷却塔



- ①チラーや冷却塔で②冷水を作る
- ②冷水はサーバールーム内の③LtoLで冷却液と熱交換を行う
- ④冷たい冷却液はサーバの冷却プレートに送られる
- ⑤温かい冷却液は④CDUに戻され、冷水と熱交換を行う
- ⑥温水は①チラーや冷却塔に戻され、再び②冷水を作る

## LtoLによる冷却

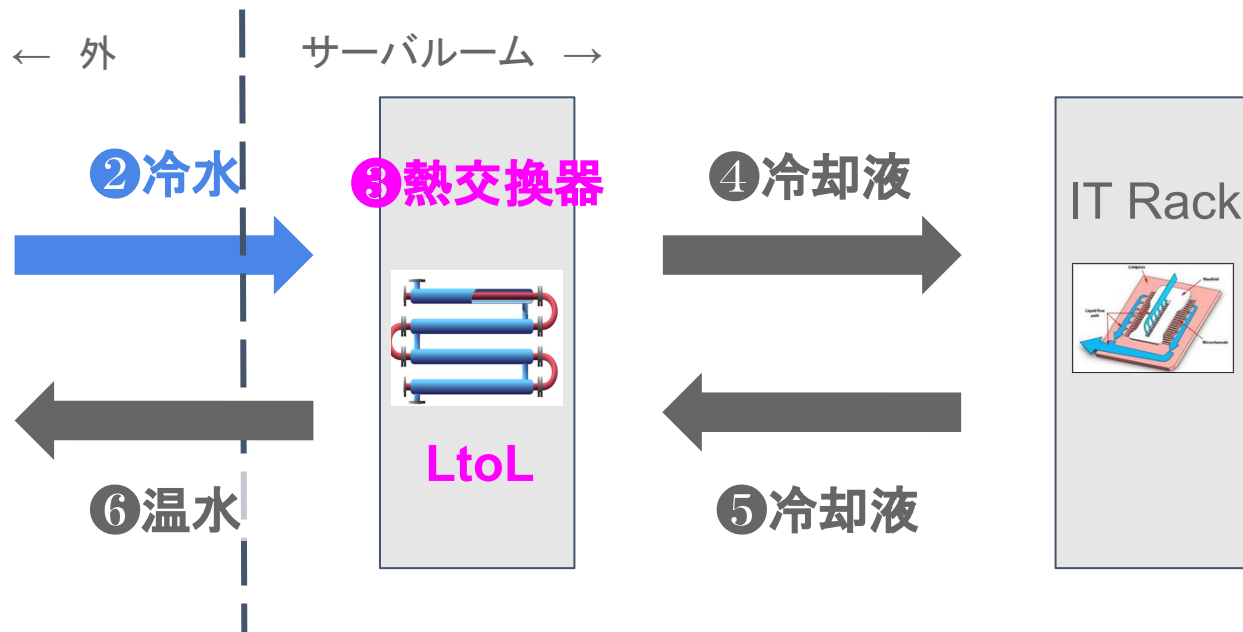
## ①チラー/冷却塔



- ①チラーや冷却塔で②冷水を作る
- ②冷水はサーバールーム内の③LtoLで冷却液と熱交換を行う
- ④冷たい冷却液はサーバの冷却プレートに送られる
- ⑤温かい冷却液は④CDUに戻され、冷水と熱交換を行う
- ⑥温水は①チラーや冷却塔に戻され、再び②冷水を作る

## LtoLによる冷却

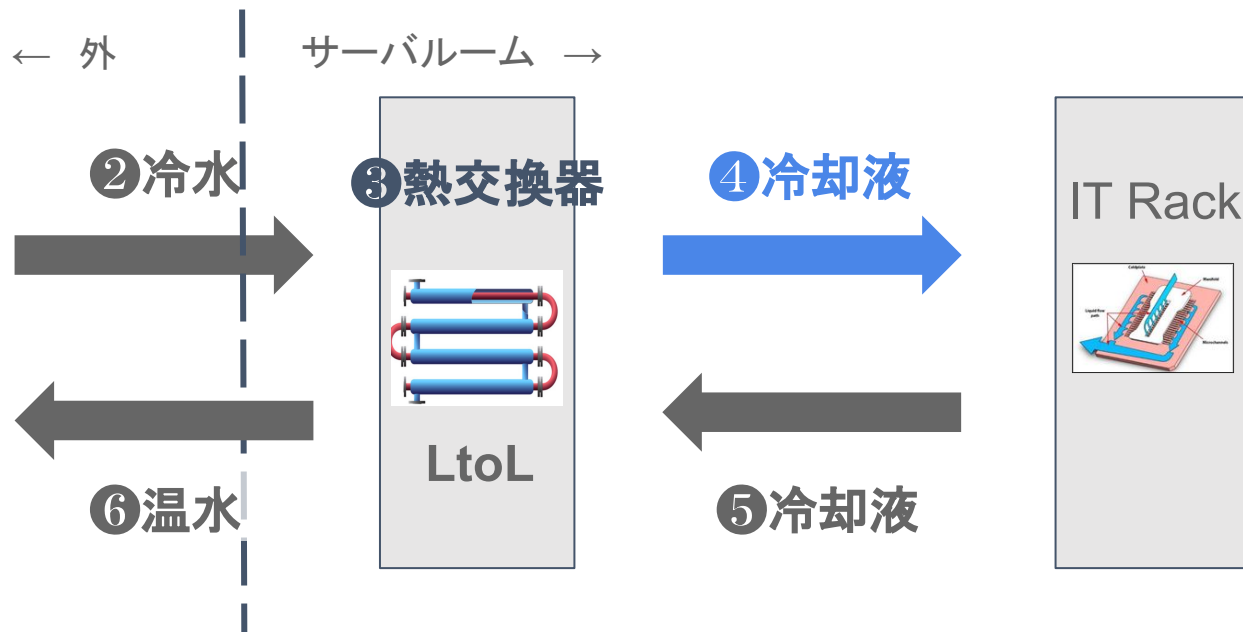
## ①チラー/冷却塔



- ①チラーや冷却塔で②冷水を作る
- ②冷水はサーバールーム内の③LtoLで冷却液と熱交換を行う
- ④冷たい冷却液はサーバの冷却プレートに送られる
- ⑤温かい冷却液は④CDUに戻され、冷水と熱交換を行う
- ⑥温水は①チラーや冷却塔に戻され、再び②冷水を作る

## LtoLによる冷却

## ①チラー/冷却塔

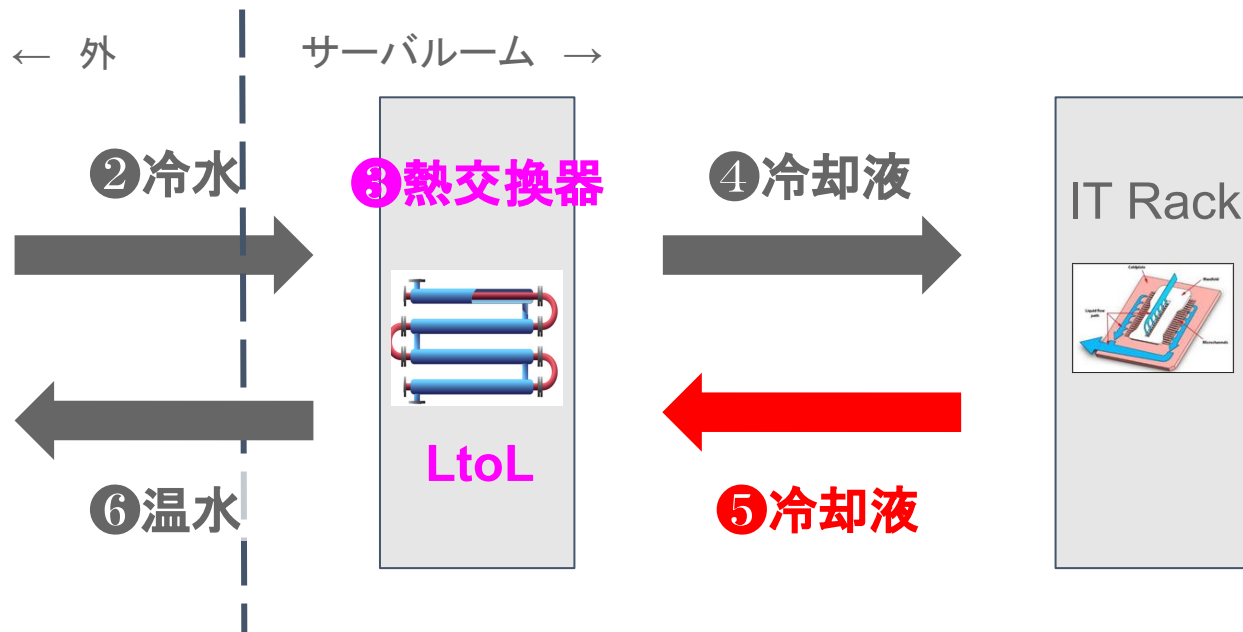


- ①チラーや冷却塔で②冷水を作る
- ②冷水はサーバールーム内の③LtoLで冷却液と熱交換を行う
- ④冷たい冷却液はサーバの冷却プレートに送られる
- ⑤温かい冷却液は③LtoLに戻され、冷水と熱交換を行う
- ⑥温水は①チラーや冷却塔に戻され、再び②冷水を作る



## LtoLによる冷却

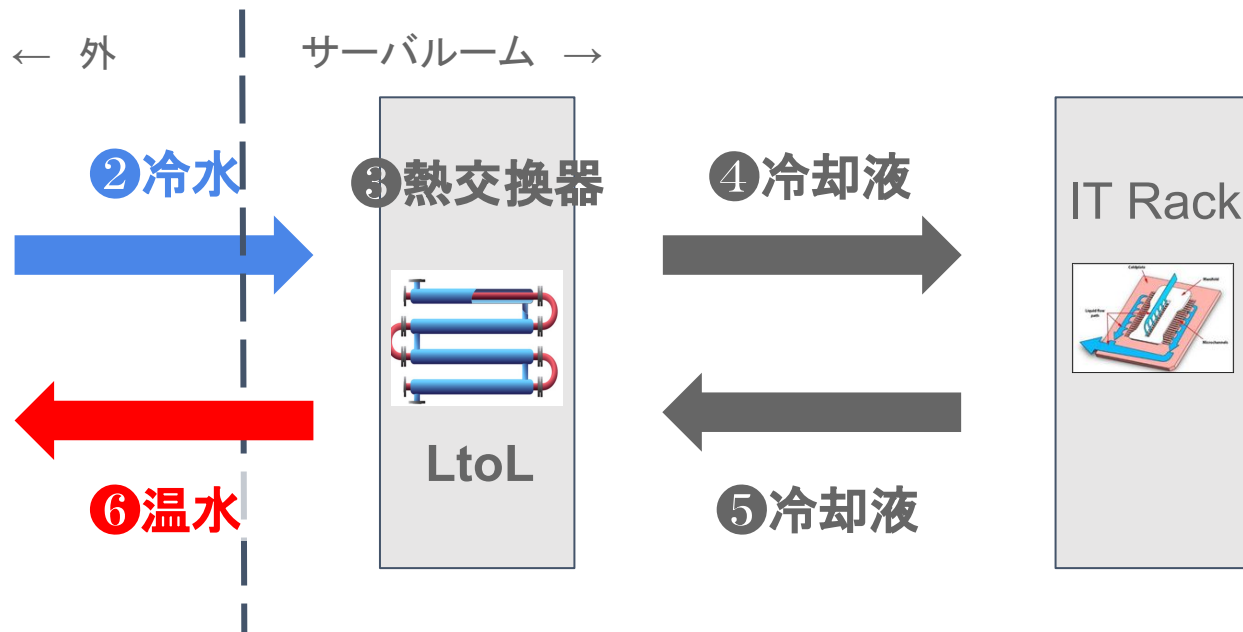
## ①チラー/冷却塔



- ①チラーや冷却塔で②冷水を作る
- ②冷水はサーバールーム内の③LtoLで冷却液と熱交換を行う
- ④冷たい冷却液はサーバの冷却プレートに送られる
- ⑤温かい冷却液は③LtoLに戻され、冷水と熱交換を行う
- ⑥温水は①チラーや冷却塔に戻され、再び②冷水を作る

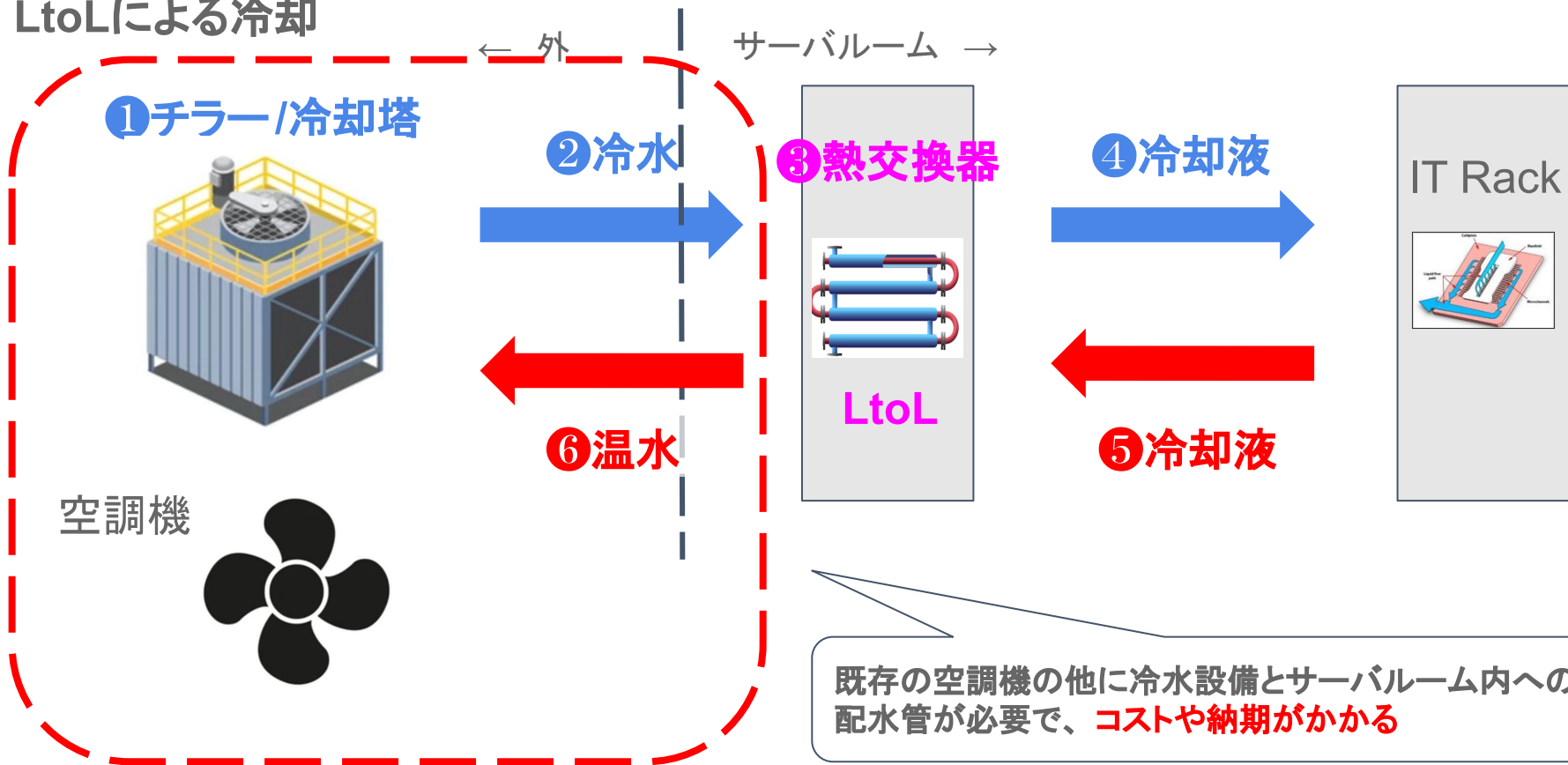
## LtoLによる冷却

## ①チラー/冷却塔

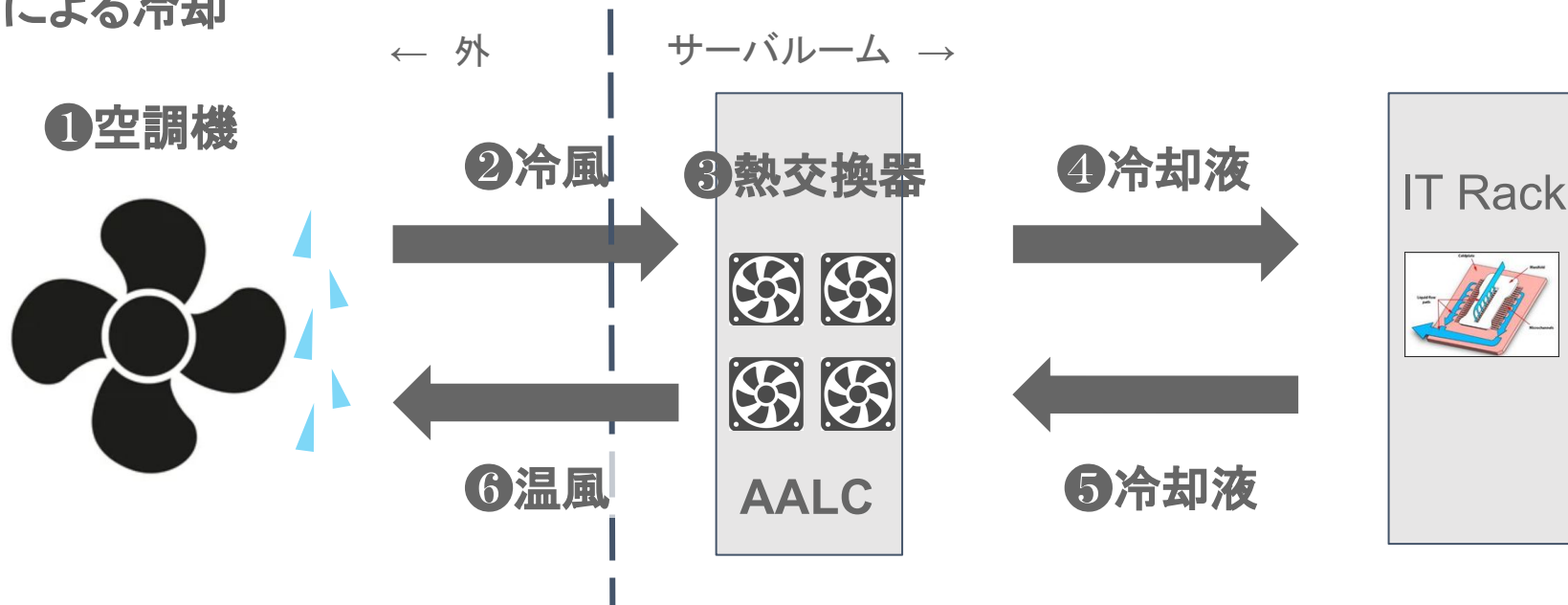


- ①チラーや冷却塔で②冷水を作る
- ②冷水はサーバールーム内の③LtoLで冷却液と熱交換を行う
- ④冷たい冷却液はサーバの冷却プレートに送られる
- ⑤温かい冷却液は③LtoLに戻され、冷水と熱交換を行う
- ⑥温水は①チラーや冷却塔に戻され、再び②冷水を作る

## LtoLによる冷却

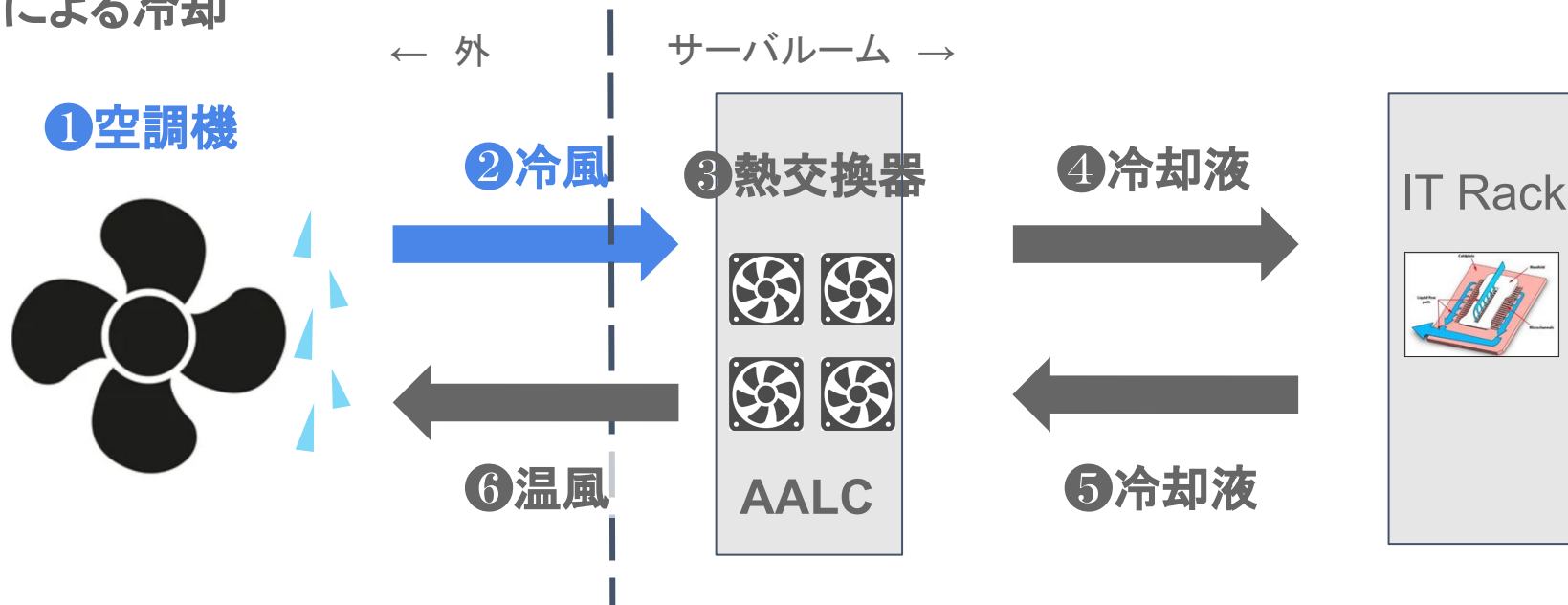


## AALCによる冷却



- ①空調機で②冷風を作る
- ②冷風はサーバールーム内の③AALCで冷却液と熱交換を行う
- ④冷たい冷却液はサーバの冷却プレートに送られる
- ⑤温かい冷却液は③AALCに戻され、冷風と熱交換を行う
- ⑥温風は①空調機に戻され、再び②冷風を作る

## AALCによる冷却

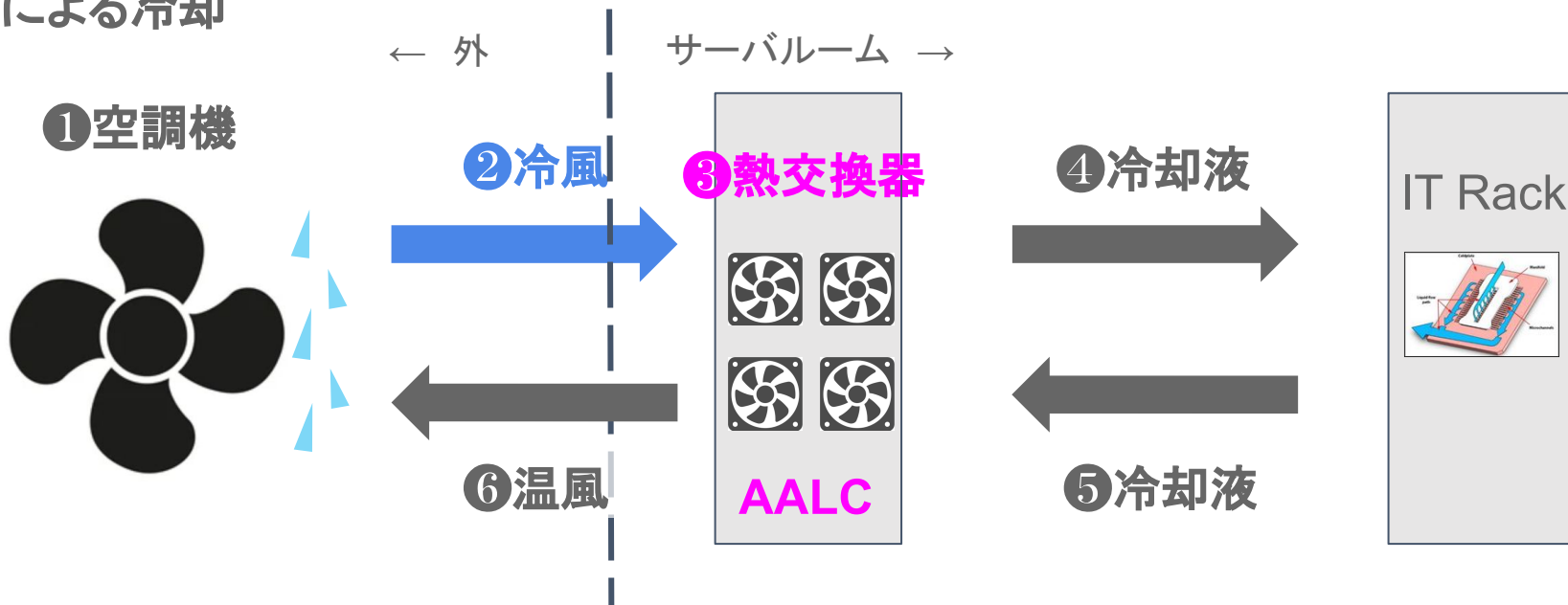


- ①空調機で②冷風を作る
- ②冷風はサーバールーム内の③AALCで冷却液と熱交換を行う
- ④冷たい冷却液はサーバの冷却プレートに送られる
- ⑤温かい冷却液は③AALCに戻され、冷風と熱交換を行う
- ⑥温風は①空調機に戻され、再び②冷風を作る



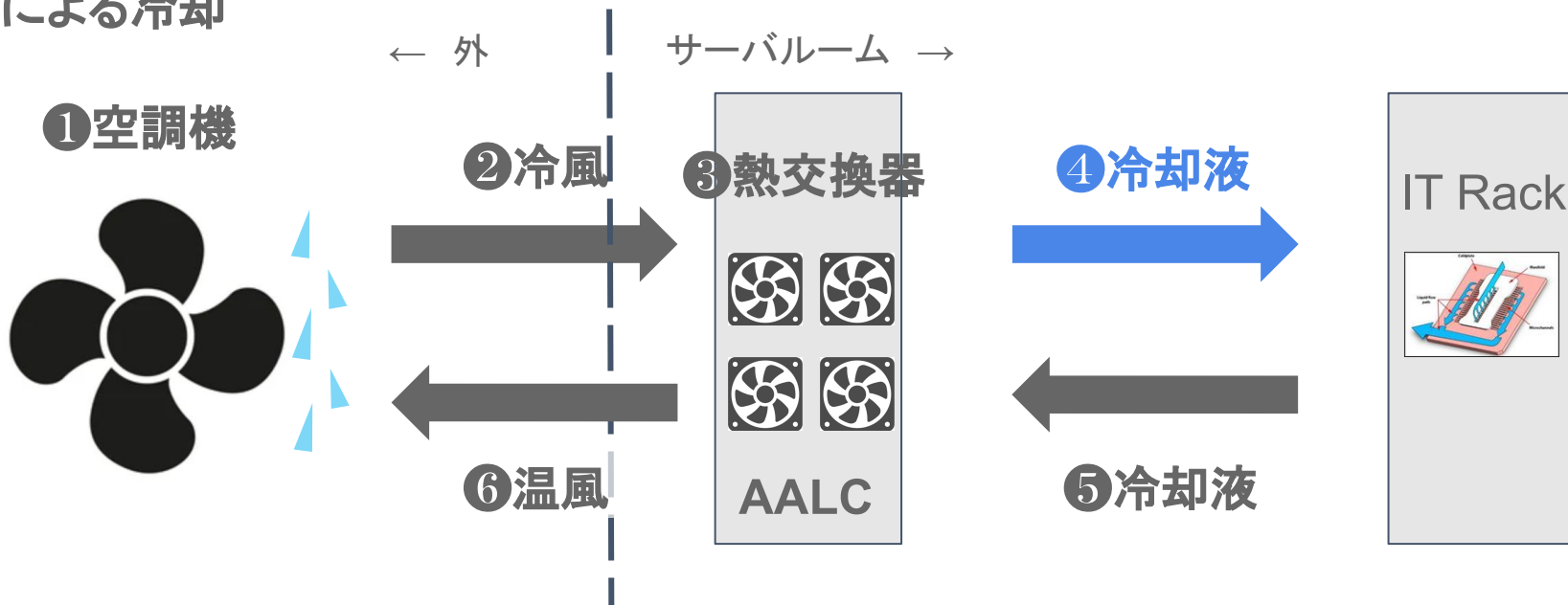
# 第一部：水冷技術で解決できる空調機の風量不足

## AALCによる冷却



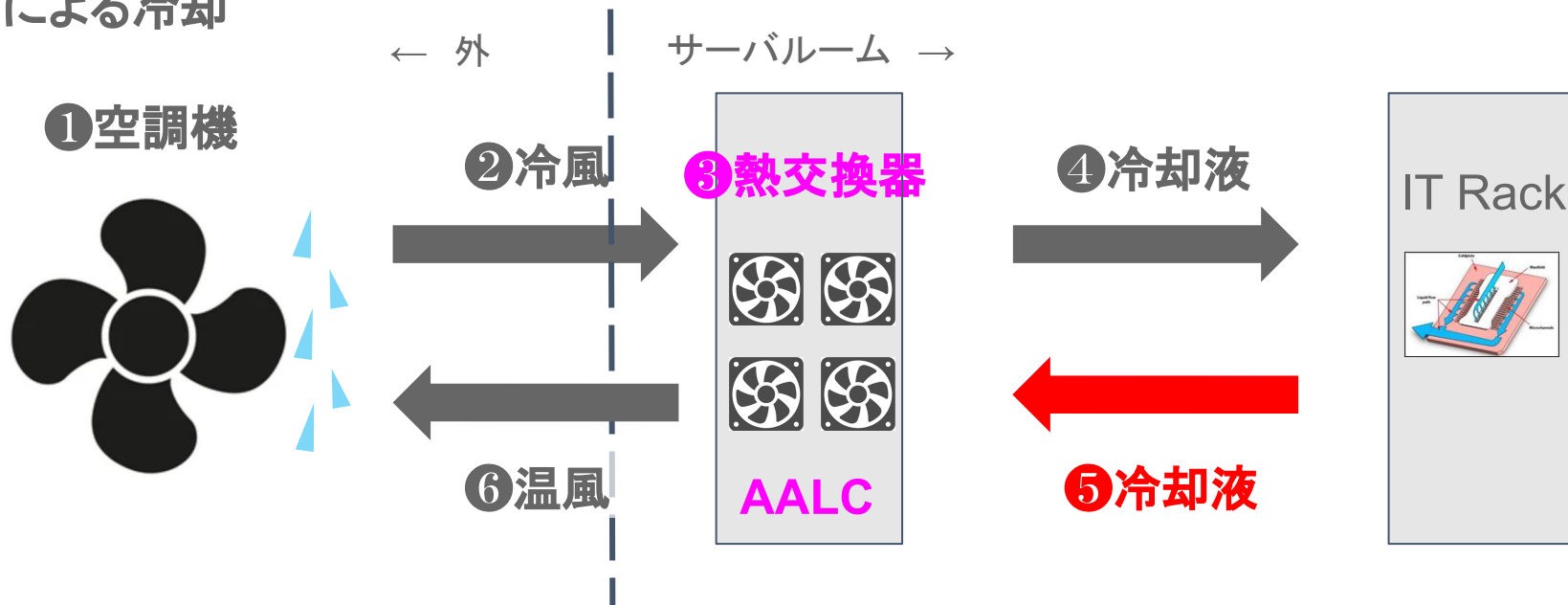
- ①空調機で②冷風を作る
- ②冷風はサーバールーム内の③AALCで冷却液と熱交換を行う
- ④冷たい冷却液はサーバの冷却プレートに送られる
- ⑤温かい冷却液は④AALCに戻され、冷風と熱交換を行う
- ⑥温風は①空調機に戻され、再び②冷風を作る

## AALCによる冷却



- ①空調機で②冷風を作る
- ②冷風はサーバールーム内の③AALCで冷却液と熱交換を行う
- ④冷たい冷却液はサーバの冷却プレートに送られる
- ⑤温かい冷却液は④AALCに戻され、冷風と熱交換を行う
- ⑥温風は①空調機に戻され、再び②冷風を作る

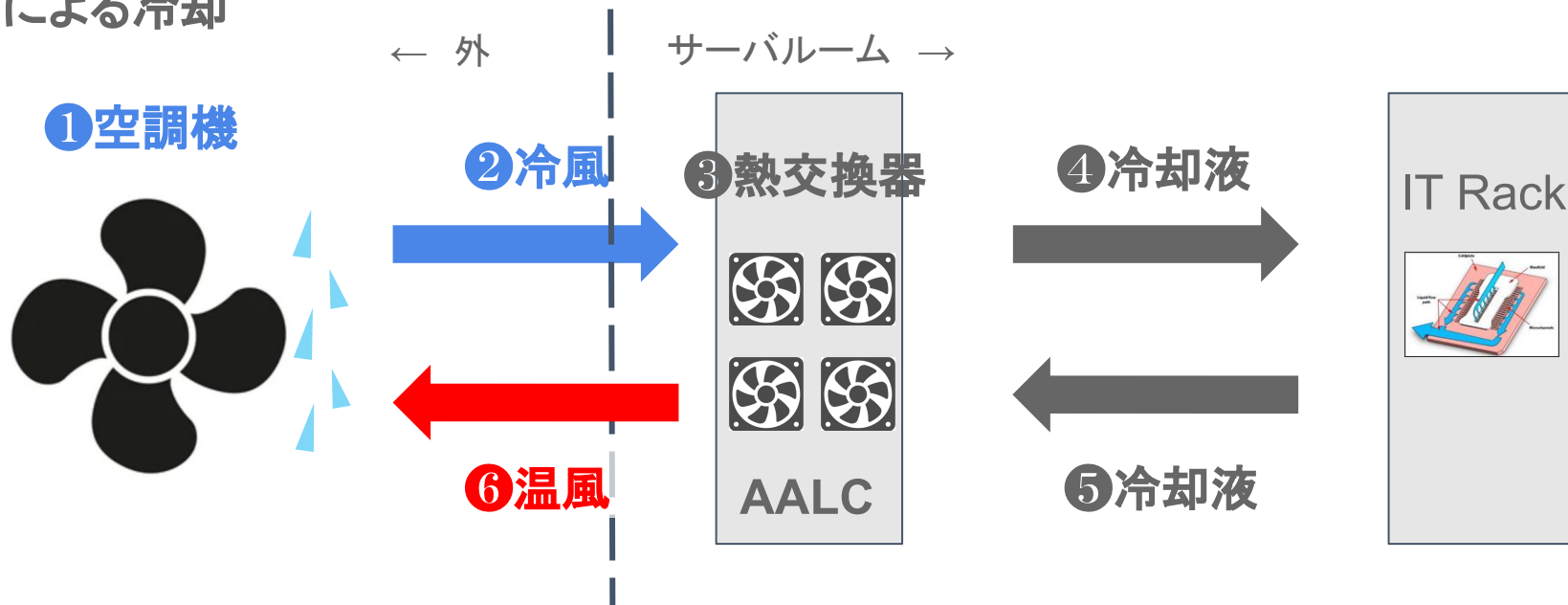
## AALCによる冷却



- ①空調機で②冷風を作る
- ②冷風はサーバールーム内の③AALCで冷却液と熱交換を行う
- ④冷たい冷却液はサーバの冷却プレートに送られる
- ⑤温かい冷却液は④AALCに戻され、冷風と熱交換を行う
- ⑥温風は①空調機に戻され、再び②冷風を作る



## AALCによる冷却



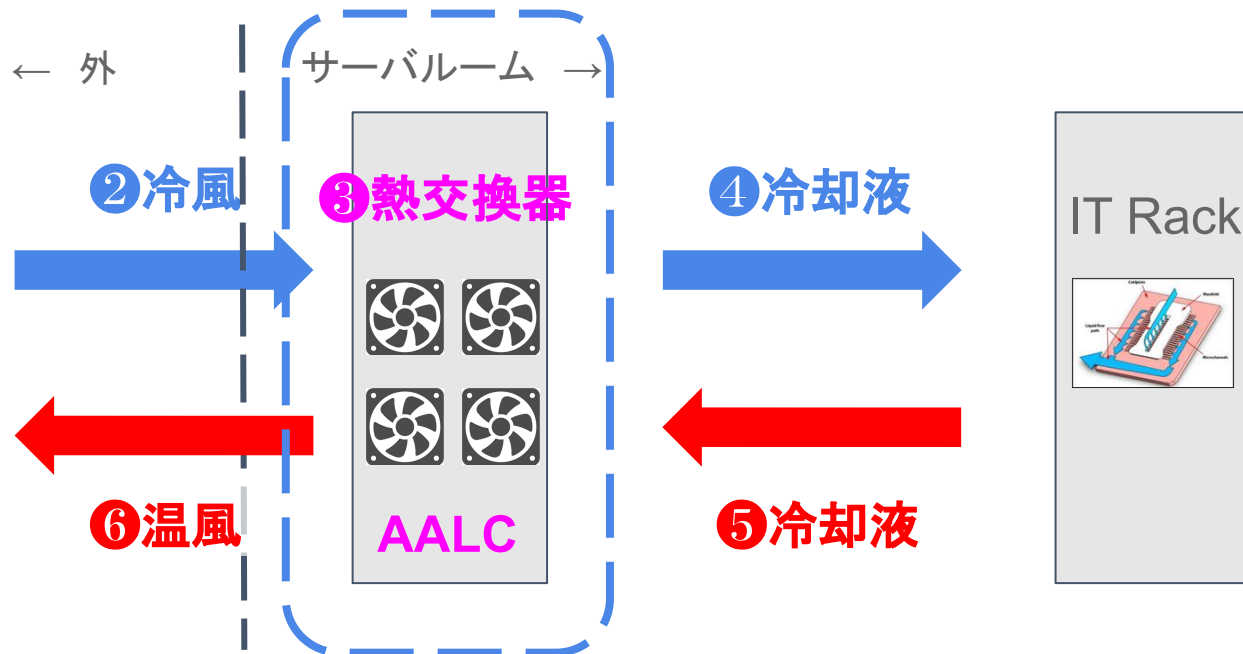
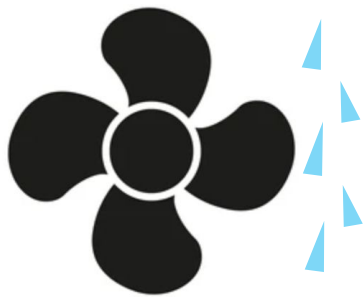
- ①空調機で②冷風を作る
- ②冷風はサーバルーム内の③AALCで冷却液と熱交換を行う
- ④冷たい冷却液はサーバの冷却プレートに送られる
- ⑤温かい冷却液は④AALCに戻され、冷風と熱交換を行う
- ⑥温風は①空調機に戻され、再び②冷風を作る



# 第一部：水冷技術で解決できる空調機の風量不足

## AALCによる冷却

① 空調機



サーバールームに設置するだけなので **低コスト、短納期**

**AALCによるサーバールームの風量減少を検証した**



AALCを利用することでサーバールーム風量は減少した

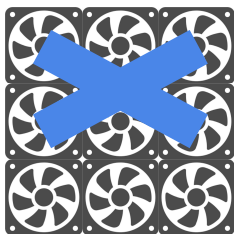
AALCによるサーバールームの風量減少を検証した



AALCを利用することでサーバールーム風量は減少した

サーバ風量とAALCの風量が空調機の風量を下回れば OK

サーバ風量



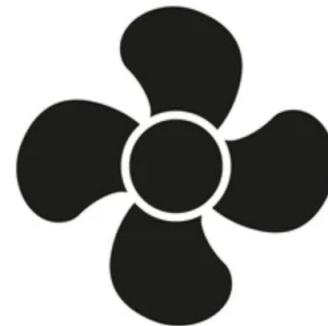
AALC風量



ラック型AALC

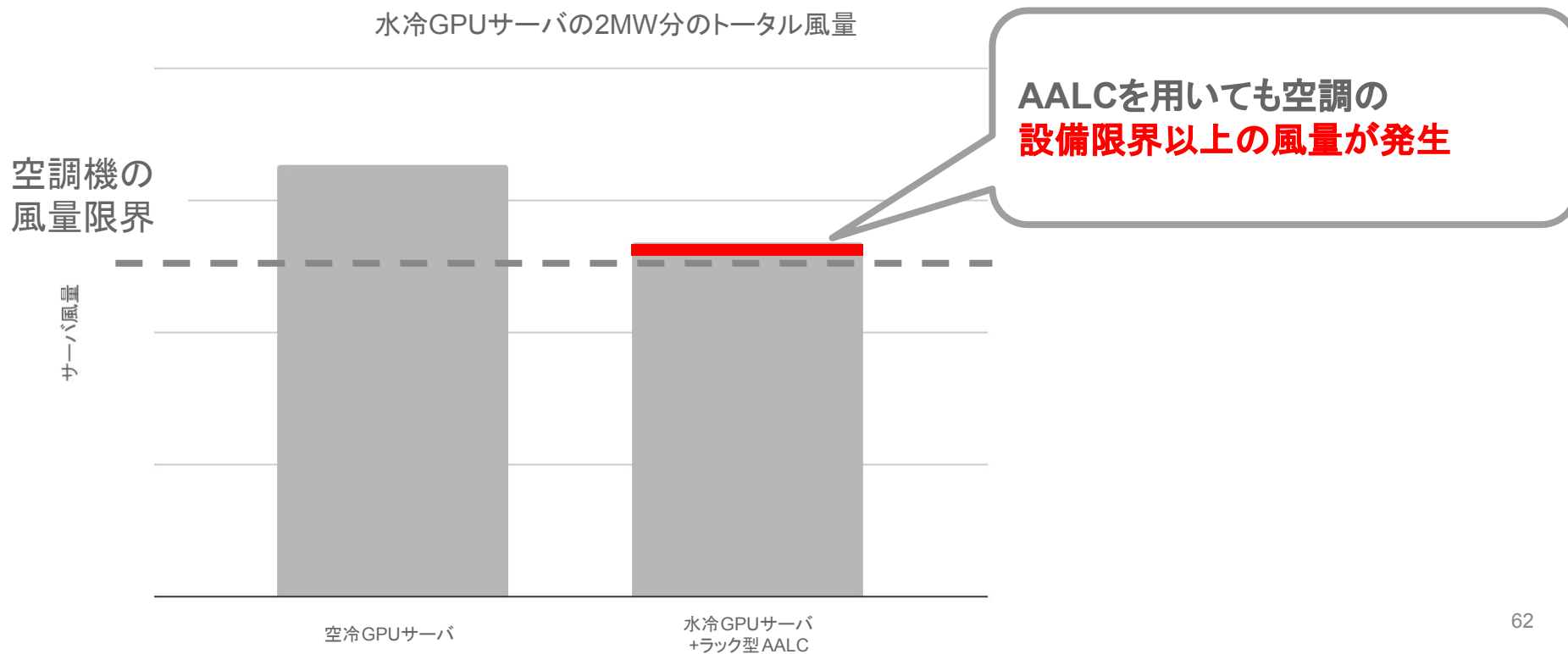


空調機の風量

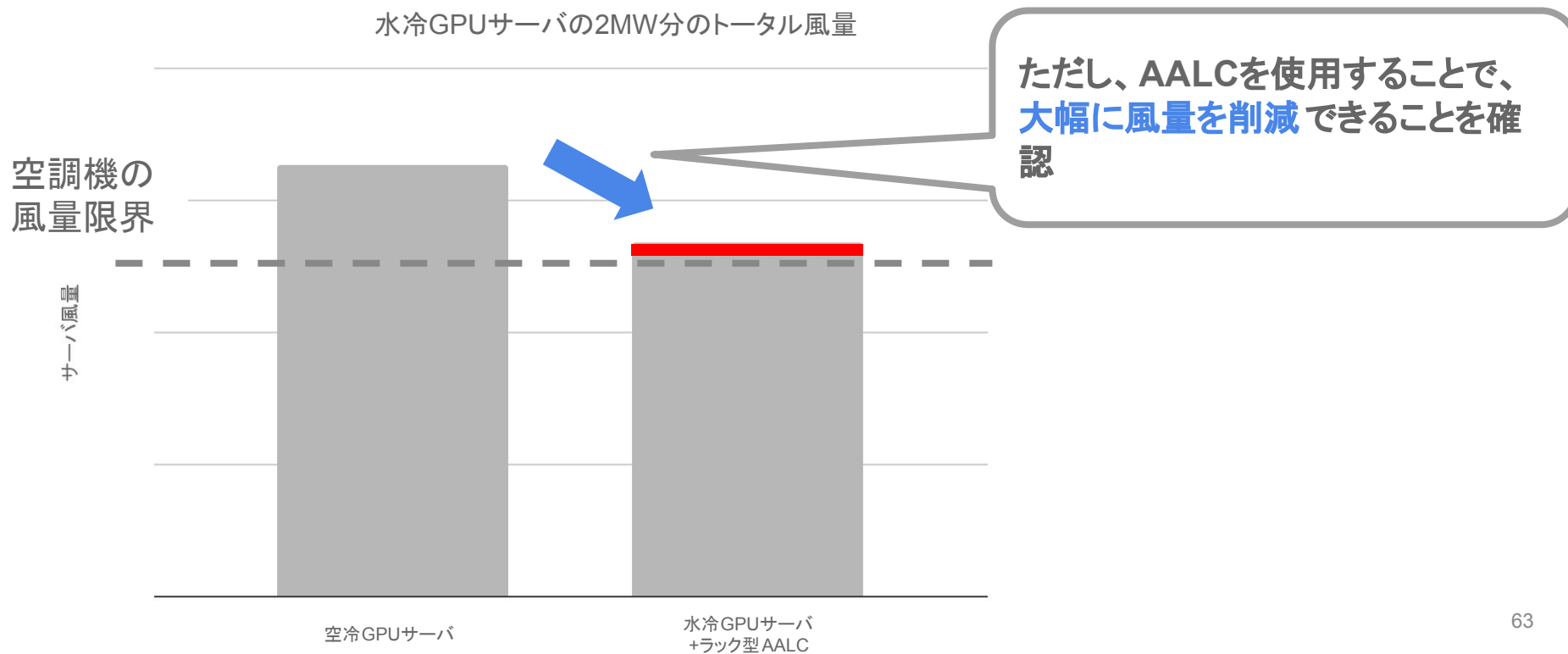


ポイント：GPUとCPUの風量が**減少した**  
分、AALCの風量は**増加**

## Actapioの2MWのサーバールームにLLM用の水冷GPUサーバとAALCをフル設置した場合のシナリオ



## Actapioの2MWのサーバールームにLLM用の水冷GPUサーバとAALCをフル設置した場合のシナリオ



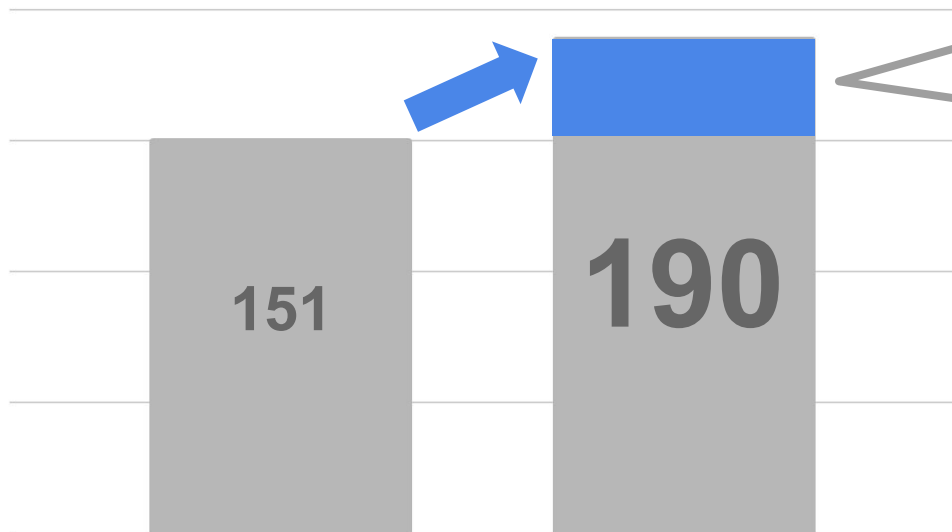


# 第一部:水冷技術で解決できる空調機の風量不足

Actapioの2MWのサーバールームにLLM用の水冷GPUサーバとAALCをフル設置した場合のシナリオ

+ 39台

サーバ台数



AALCを利用することで、  
最大で**39台**多く設置可能

空冷GPUサーバ

水冷GPUサーバ  
+ラック型AALC



## 第一部サマリ

第一部では、

- LLMは**大量のデータ**と**膨大なコンピューティングリソース**を必要とする
- その結果、**サーバワークロードが増加**し、**データセンタリソースを逼迫**させる
- 米国企業から空調機の**風量不足の可能性**を教えてもらった
- 空調システムの風量不足を**AALCでの解決**

第一部

第二部

GTCでGB200 NVL72の発表  
GB200 NVL72の電源が480Vと判明

'22  
Nov



ChatGPT 3.5がリリースされる

'23  
Mar



ヤフーのサービスでもLLMの本格利用開始

'24  
Feb



データセンタの風量不足を調査

'24  
Mar



'24  
May





## GB200 NVL72の登場とその背景

イマココ！



第一部

第二部

GTCでGB200 NVL72の発表  
GB200 NVL72の電源が480Vと判明

'22  
Nov

'23  
Mar

'24  
Feb

'24  
Mar

'24  
May

ChatGPT 3.5がリリースされる

ヤフーのサービスでもLLMの本格利用開始

データセンタの風量不足を調査

## GB200 NVL72とは？

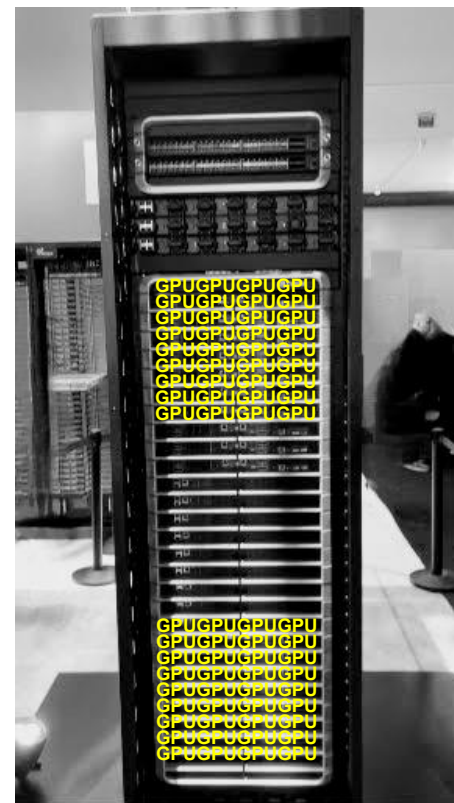
- 最新のGPU72基が1つのラックに搭載されたLLM向けのGPUシステム
- サーバ単体ではなく、ラック一式で提供
- ラック内のGPU全てがNVLinkにより相互接続され、高速なGPU間通信を実現
- 前モデルのチップであるH100と比較して、トレーニング速度が4倍、電力効率率は1/25（GPT-MoE-1.8Tモデルの場合）



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

## GB200 NVL72とは？

- 最新の**GPU72基**が1つのラックに搭載された**LLM向け**のGPUシステム
- サーバ単体ではなく、ラック一式で提供
- ラック内のGPU全てがNVLinkにより相互接続され、高速なGPU間通信を実現
- 前モデルのチップであるH100と比較して、トレーニング速度が4倍、電力効率率は1/25（GPT-MoE-1.8Tモデルの場合）



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

## GB200 NVL72とは？

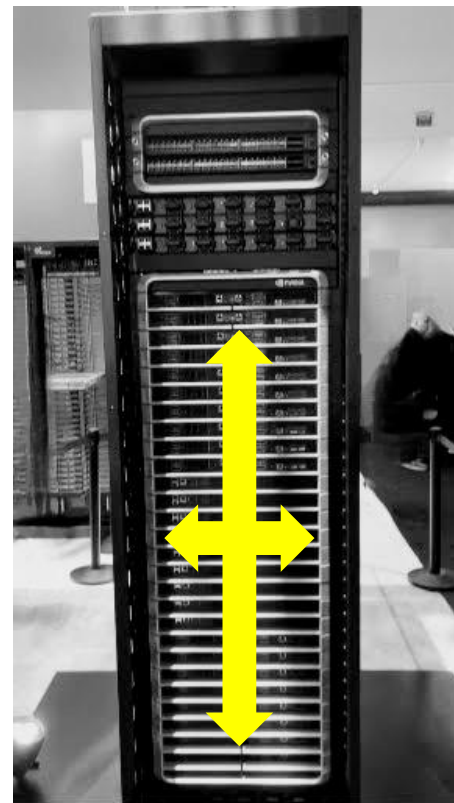
- 最新のGPU72基が1つのラックに搭載されたLLM向けのGPUシステム
- サーバ単体ではなく、**ラック一式**で提供
- ラック内のGPU全てがNVLinkにより相互接続され、高速なGPU間通信を実現
- 前モデルのチップであるH100と比較して、トレーニング速度が4倍、電力効率率は1/25（GPT-MoE-1.8Tモデルの場合）



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

## GB200 NVL72とは？

- 最新のGPU72基が1つのラックに搭載されたLLM向けのGPUシステム
- サーバ単体ではなく、ラック一式で提供
- ラック内のGPU全てが**NVLink**により相互接続され、**高速なGPU間通信を実現**
- 前モデルのチップであるH100と比較して、トレーニング速度が4倍、電力効率率は1/25（GPT-MoE-1.8Tモデルの場合）



## GB200 NVL72とは？

- 最新のGPU72基が1つのラックに搭載されたLLM向けのGPUシステム
- サーバ単体ではなく、ラック一式で提供
- ラック内のGPU全てがNVLinkにより相互接続され、高速なGPU間通信を実現
- 前モデルのチップであるH100と比較して、**トレーニング速度が4倍**、電力効率率は1/25（GPT-MoE-1.8Tモデルの場合）



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

## GB200 NVL72とは？

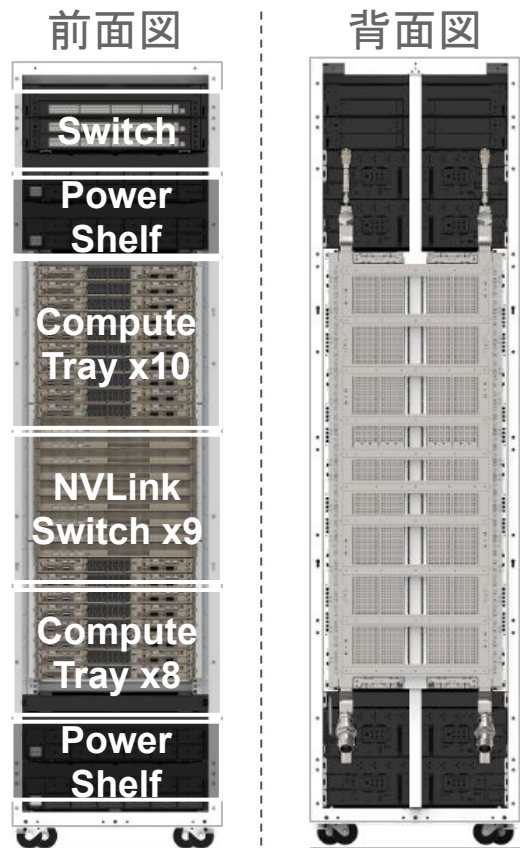
- 最新のGPU72基が1つのラックに搭載されたLLM向けのGPUシステム
- サーバ単体ではなく、ラック一式で提供
- ラック内のGPU全てがNVLinkにより相互接続され、高速なGPU間通信を実現
- 前モデルのチップであるH100と比較にて、トレーニング速度が4倍、**電力効率は1/25**（GPT-MoE-1.8Tモデルの場合）



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



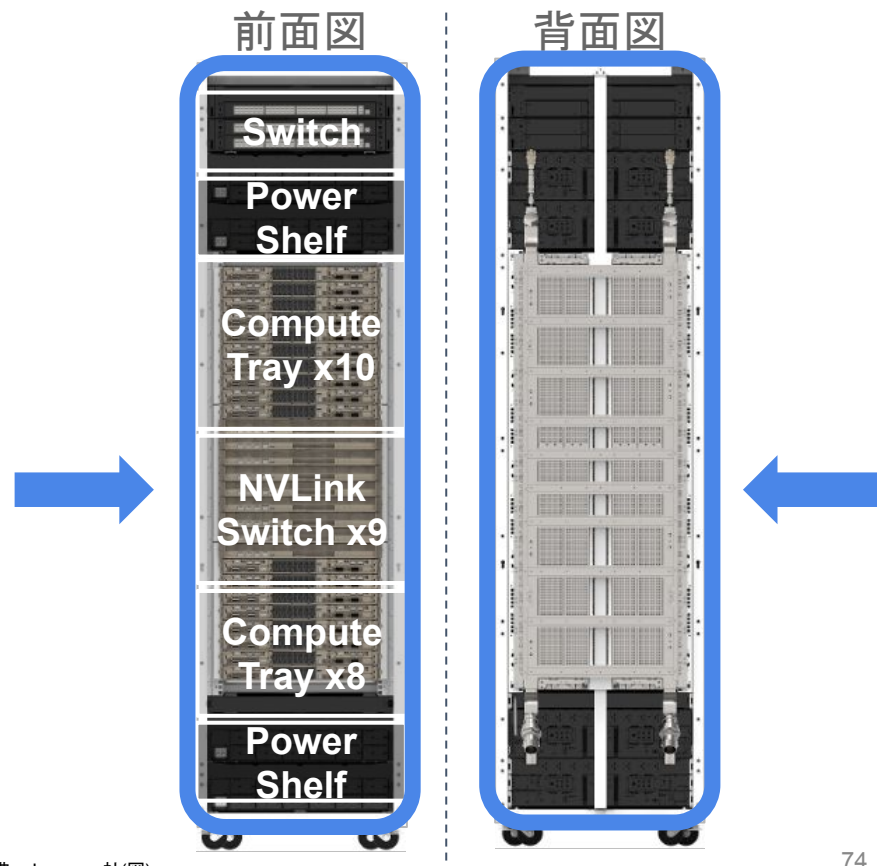
## GB200 NVL72の特徴



## GB200 NVL72の特徴

## ラック：OCP仕様のORv3ラック

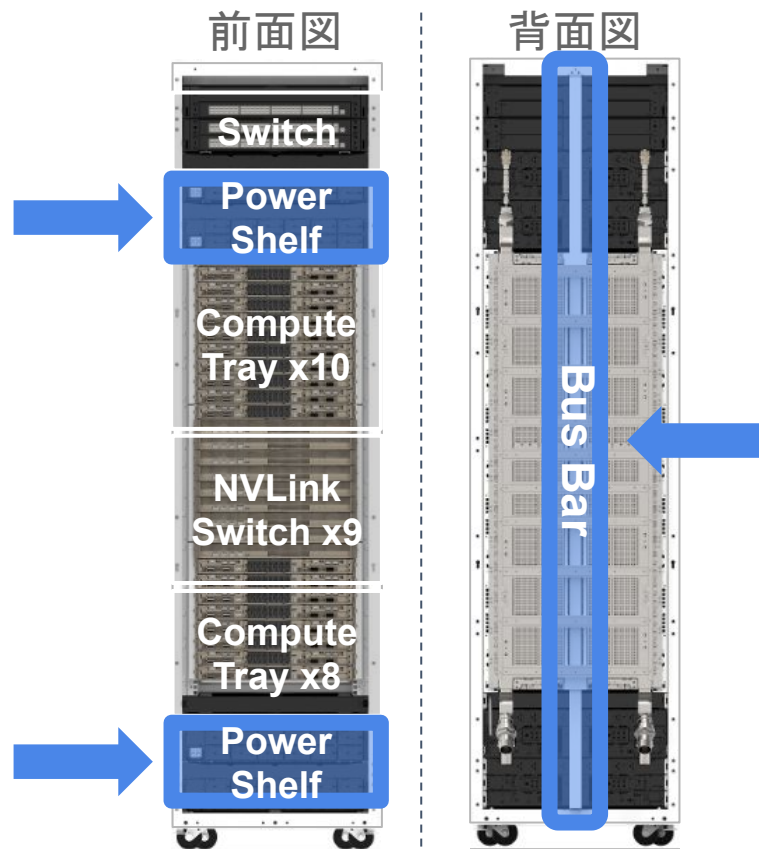
- ORv3ラックはラック幅が**21インチ**
- **19インチ、21インチ両方**の機器を設置可能
- GB200 NVL72ではPower Shelf以外は19インチ



## GB200 NVL72の特徴

## 電源：OCP仕様のBus BarとPower Shelf

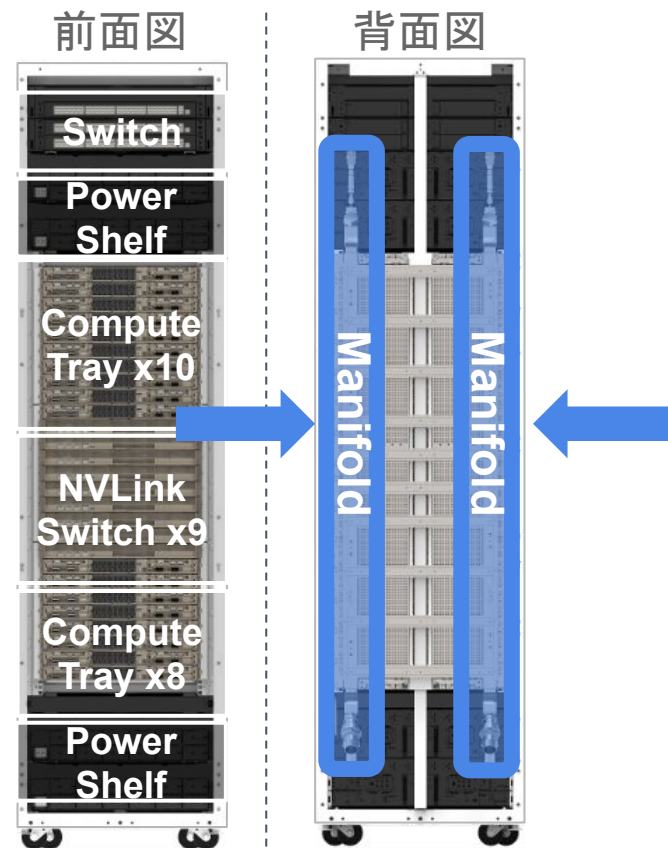
- Bus Barは48V直流電圧
- Power Shelfは必要に応じて電力やバックアップシステムを追加可能



## GB200 NVL72の特徴

Manifold: **OCP仕様**の**水冷用の集合配管**

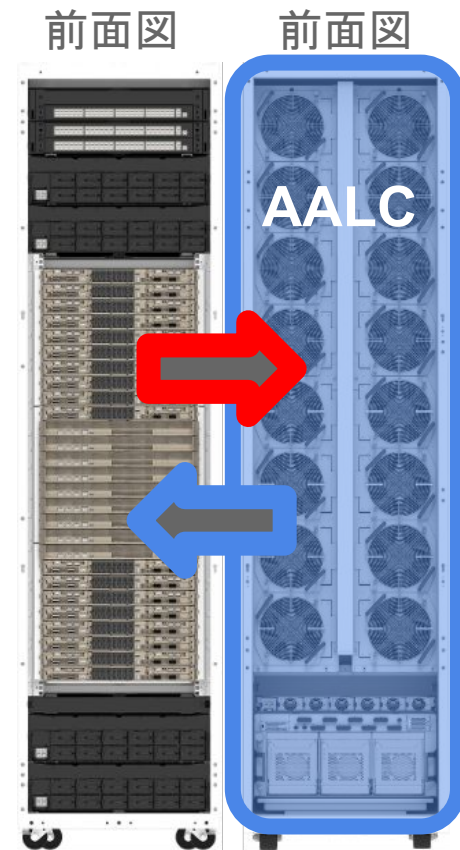
- ラック内の各IT機器に**冷却液を分配**
- ユニバーサルクイックディスコネクトを採用し、**迅速で安全な接続・切断**が可能
- OCPで標準化することにより、**互換性が確保**



## GB200 NVL72の特徴

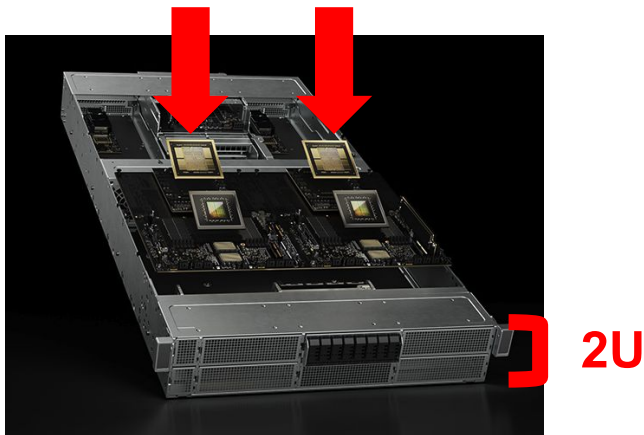
冷却システム：**水冷方式（DLC）**が必須

- 各社**ラック型AALC**で繋ぐモデルで考えている模様



## なぜ？ 水冷方式が必須

- NVIDIA Blackwellの **GPUチップが2つ**配置されているGB200 NVL2は、**空冷ヒートシンクが2U**サイズです

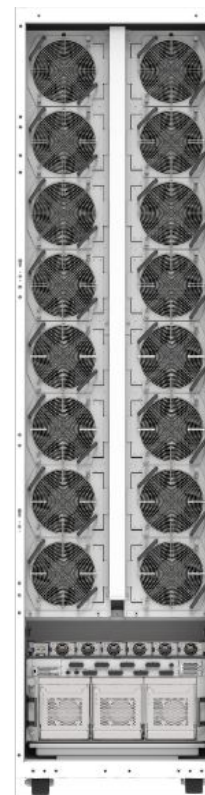


GB200 NVL2

前面図



前面図

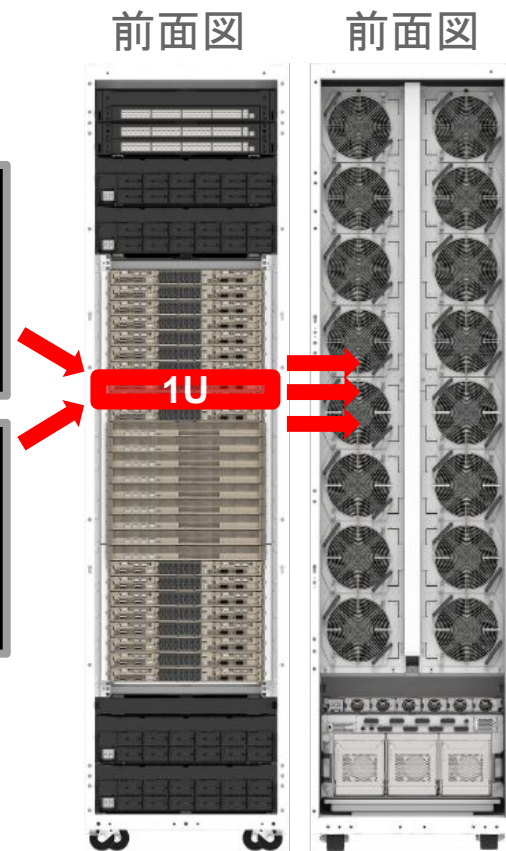


## なぜ？ 水冷方式が必須

- GB200 NVL72は、NVIDIA Blackwell **GPUチップが4つ、1Uに配置**
- 1Uに空冷ヒートシンクを配置するのは**スペース的に難しい**
- そのため、**水冷技術**を使用して熱を運ぶ必要があった



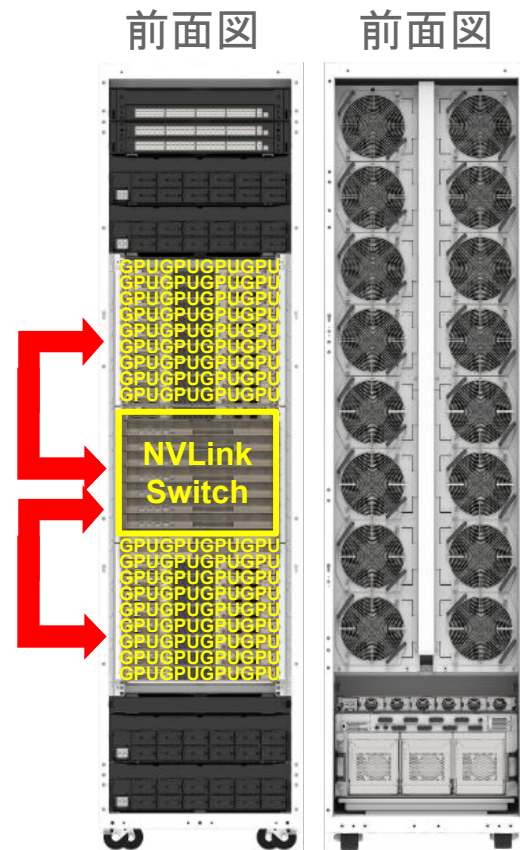
GB200 Superchip



なぜ？ 水冷方式が必須

なぜ？ 1Uにしている？

- 72個のGPUを**NVLink**で接続するには**Copper(銅線)**で接続する必要がある
- そのCopperの**長さ**に**制限**がある
- そのためGPU(サーバ)間を**近くに配置**する必要がある





なぜ？ 水冷方式が必須

なぜ？ 1Uにしている？

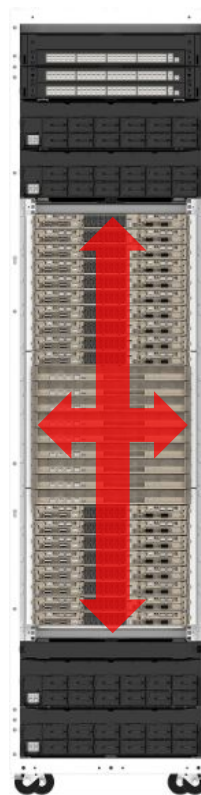
なぜ？ NVLinkで接続しないといけない？

- LLMの課題は**モデル学習時間の短縮**
- そのモデル学習時間の短縮には**GPU間的高速通信が必須**となっている
- **NVLink**でGPU間を接続することにより高速通信が可能になる

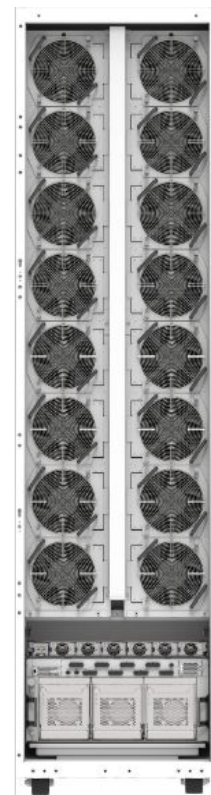
速度比較 (bi-directionally)

	GB/s	Gbps
<b>NVLink 5th</b>	<b>1,800</b>	<b>14,400</b>
PCIe-7 x16	512	4,096
800G Ether	200	1,600
400G Ether	100	800

前面図



前面図



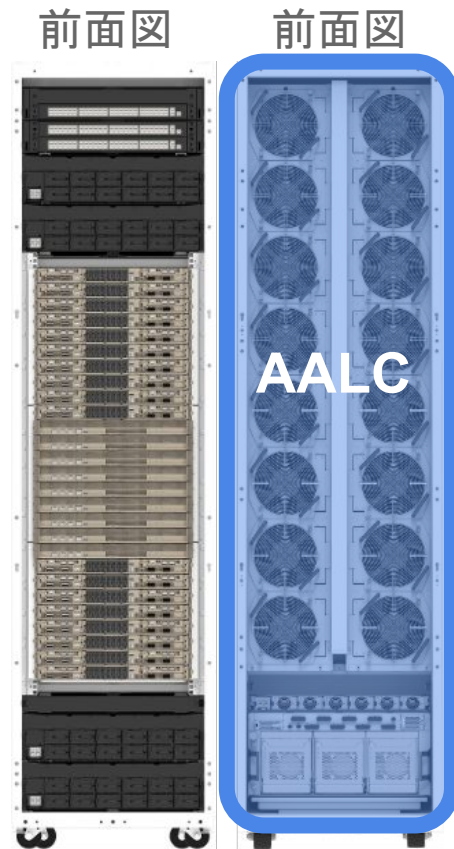
なぜ？「水冷方式が必須」

なぜ？ 1Uにしている？

なぜ？ NVLinkで接続しないとイケない？



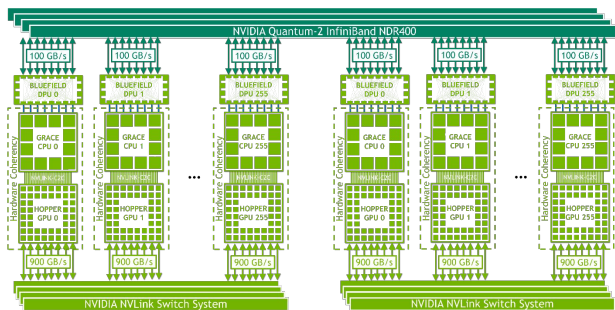
なぜならば、  
「LLMの課題であるモデル学習時間の短縮のため」



## 少し脱線

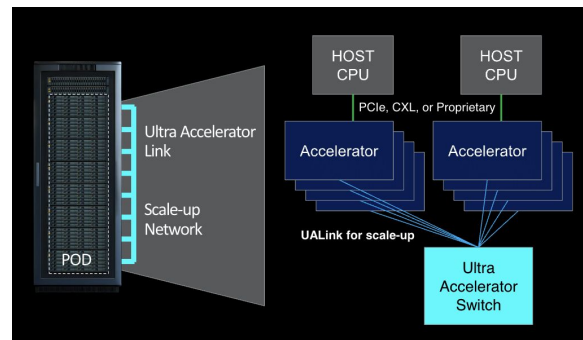
## Processor to Processor networks: Future of Connectivity?

## NVlink



VS

## UALink(Ultra Accelerator Link)



開発者：NVIDIA

主な用途：プロセッサ間のデータ転送

特徴：高帯域、低レイテンシー、576のGPUを接続

拡張性：最大576のGPUを接続(図はGH200)

開発者：Intel, AMD, Broadcom, Cisco, HP, Google, Microsoft, Meta

主な用途：プロセッサ間のデータ転送

特徴：高帯域、低レイテンシー、オープンスタンダード

拡張性：最大1,024個のAIアクセラレータを接続



## 電源の480V化(新しい課題)

イマココ！



第一部

第二部

GTCでGB200 NVL72の発表  
GB200 NVL72の電源が480Vと判明

'22  
Nov

'23  
Mar

'24  
Feb

'24  
Mar

'24  
May

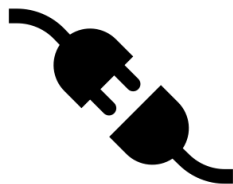
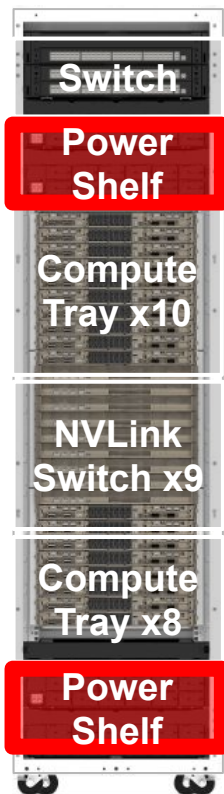
ChatGPT 3.5がリリースされる

ヤフーのサービスでもLLMの本格利用開始

データセンタの風量不足を調査

## GB200 NVL72は1ラック120kwのお化けマシン

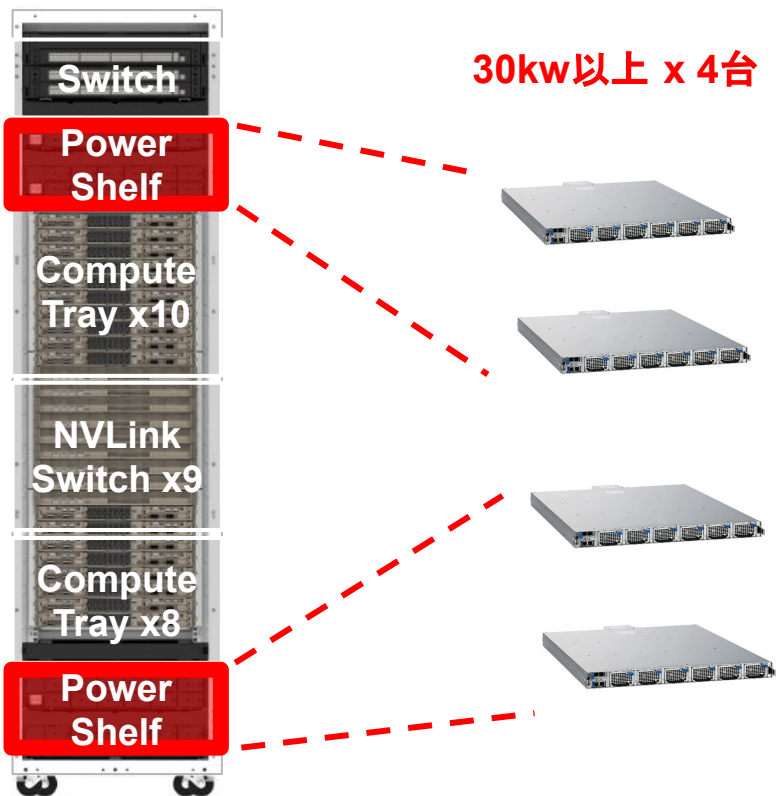
前面図



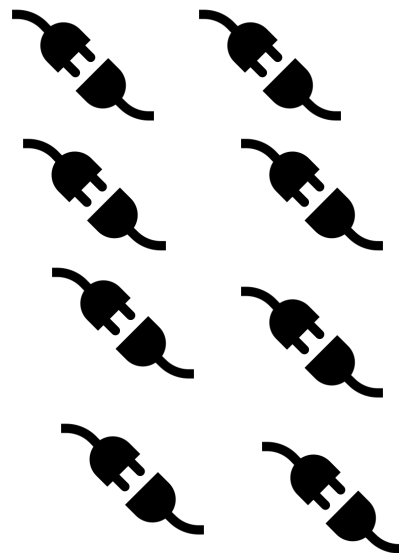
1Rackに詰め込んでいるため  
**120kw分**の高電力が必要

## GB200 NVL72は1ラック120kwのお化けマシン

前面図

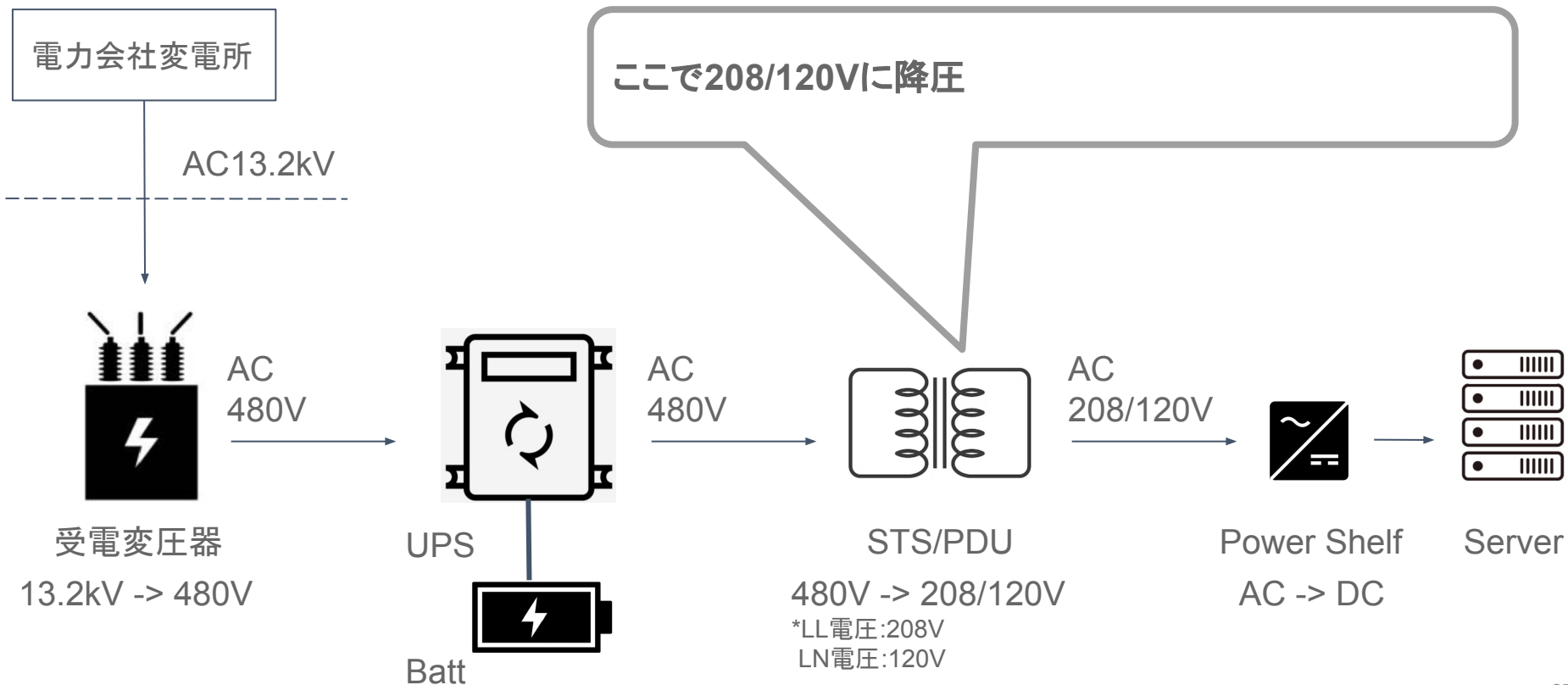


480V x 8本  
19.0A

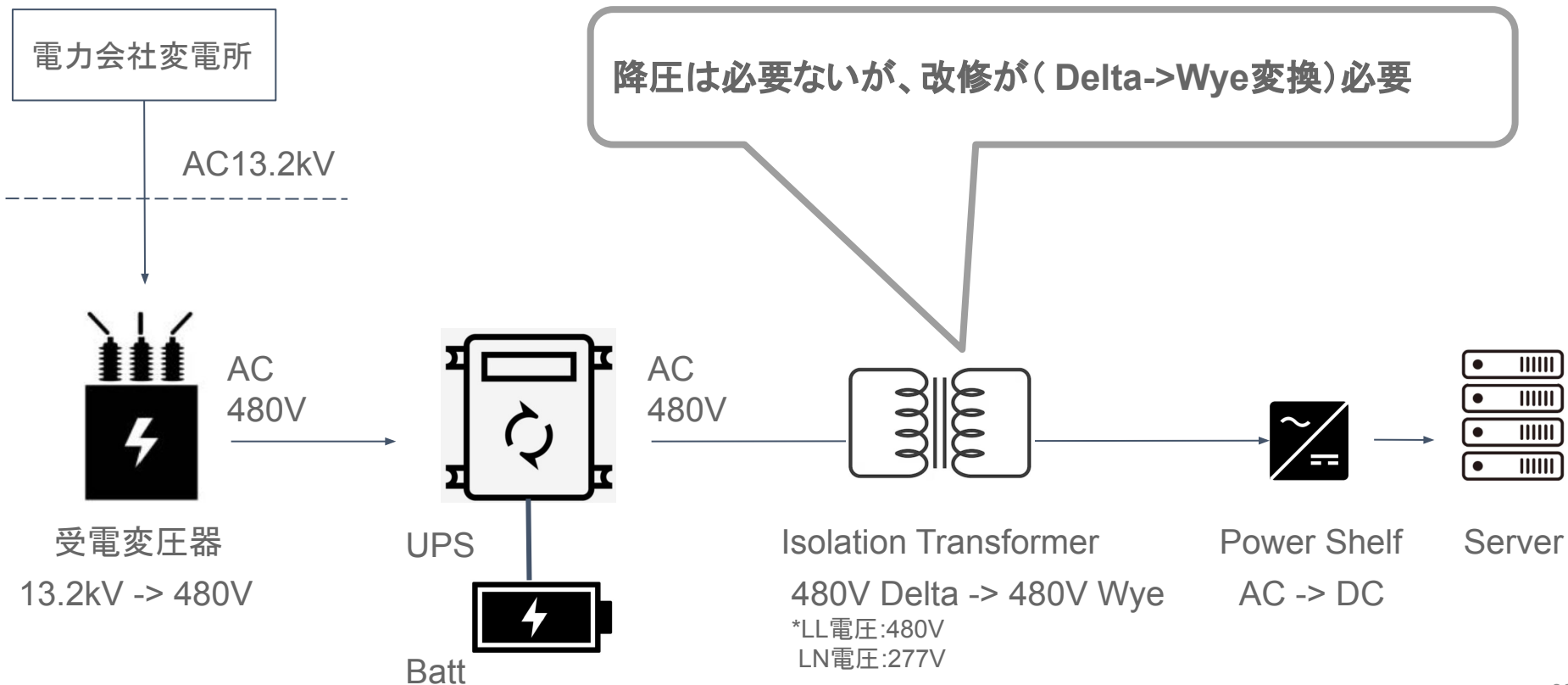


サーバラックまでの  
接続に高電圧が使用  
されるため、  
安全性に関する懸念

## 電力供給スキーム Actapioデータセンターの場合



## 電力供給スキーム 480Vの場合

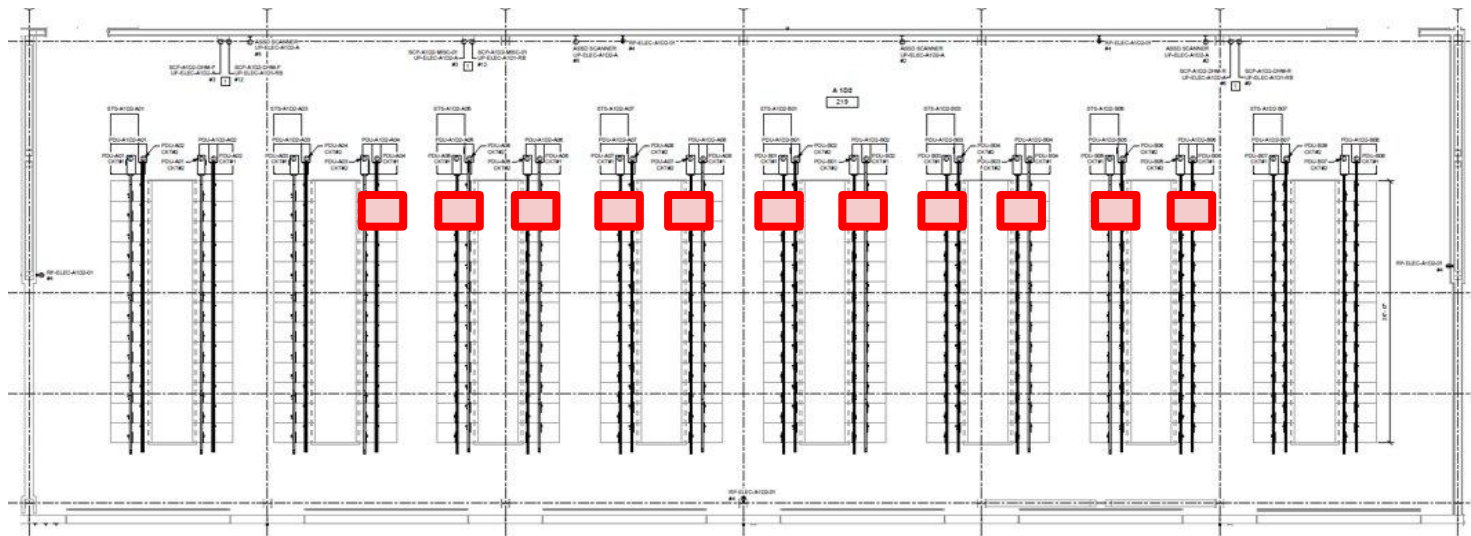






To be continued 🥲

## 2MWのサーバルームに10セットくらいしか置けない



スペース効率は悪いが、データセンターの**土地、建物のOpex**は**とても小さい**ので気にしない👩🏻

## 第二部サマリ

第二部では、

- LLMは**モデル学習時間の短縮が大きな課題** になっている
- GTCで発表されたGB200 NVL72は、**この問題を解決できる技術**
- 近接性のためにはヒートシンクでの対応が難しく、効率的な冷却のために**水冷技術**が必要
- 1ラックに集積されたサーバは、新たな課題として**電力供給の問題を生み出しました**

第一部

第二部

GTCでGB200 NVL72の発表  
GB200 NVL72の電源が480Vと判明

'22  
Nov

'23  
Mar

'24  
Feb

'24  
Mar

'24  
May

ChatGPT 3.5がリ  
リースされる

ヤフーのサービスでも  
LLMの本格利用開始

データセンタの風  
量不足を調査

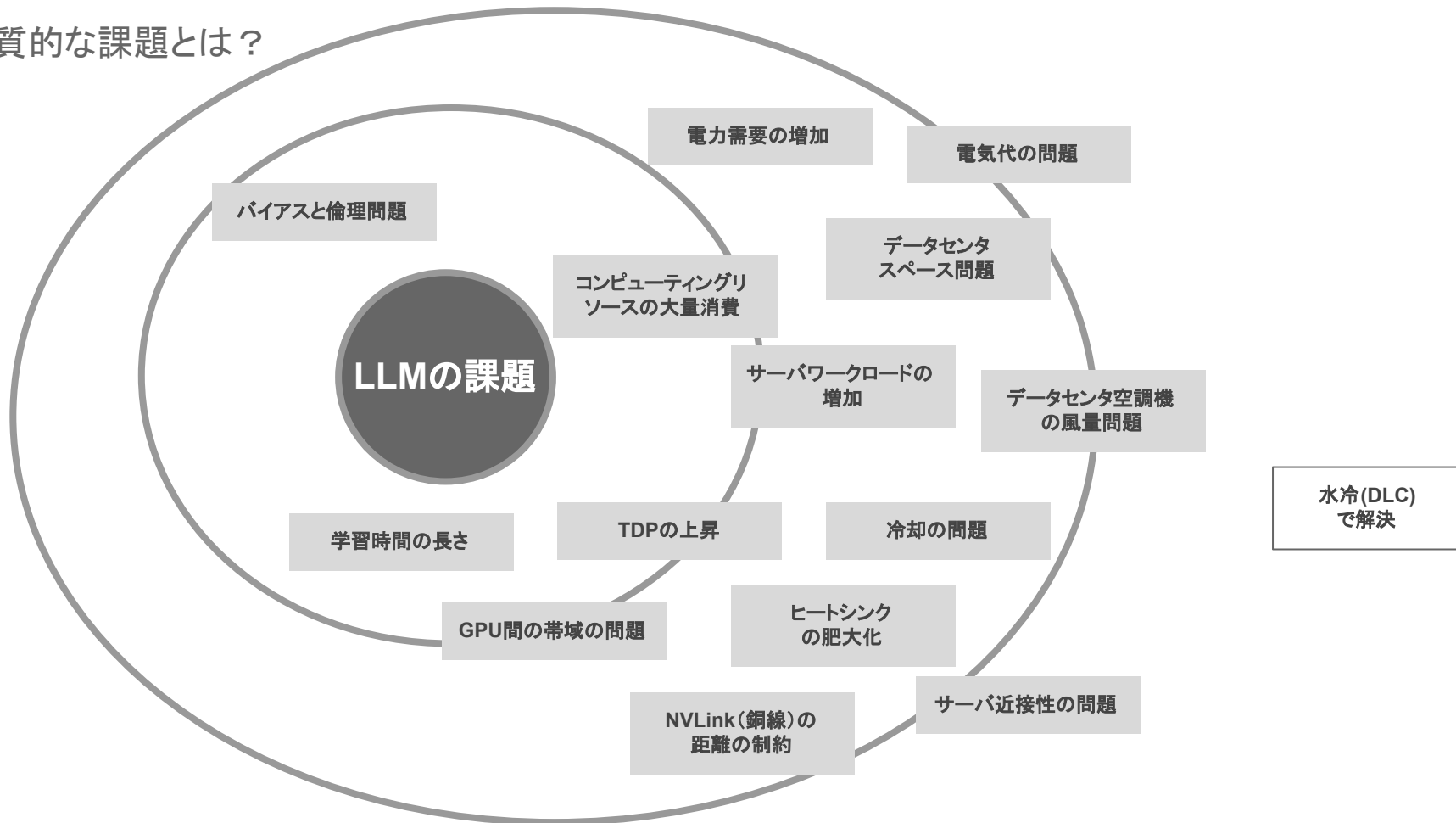
## まとめ

- **LLMの普及**は、私たちのデータセンタ運用に新たな挑戦をもたらした
- LLMによってサーバワークロードの増加とそれに伴う冷却の課題が発生
- 具体的には**空調システムの風量不足** や**サーバ近接性制約** という問題として表面化

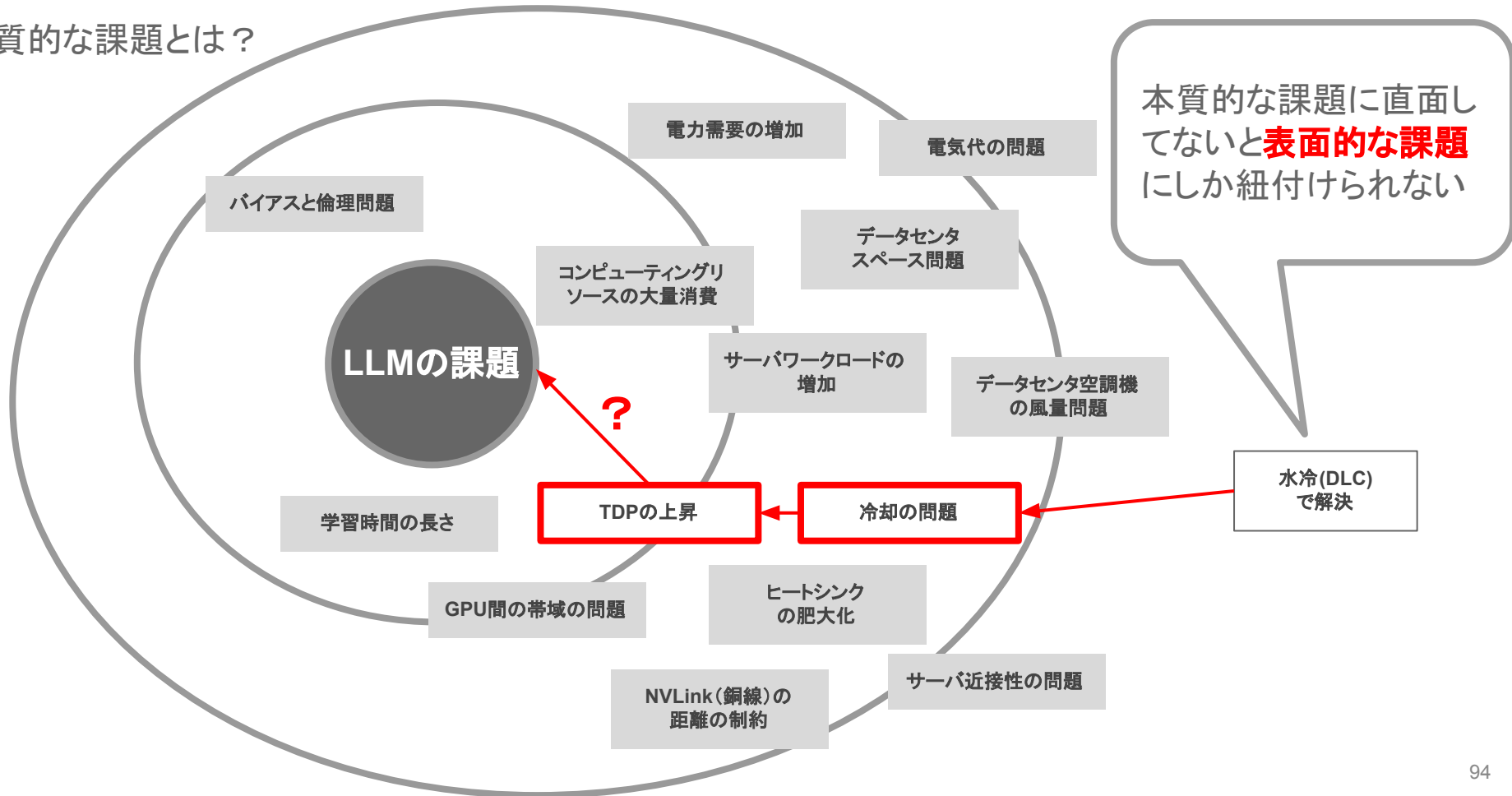


- 私たちは、GTCやOCPサミットなどで米国企業やハイパースケールと対話し、**水冷技術**がこれらの「**本質的な課題**」を解決する鍵であることを慎重に検討しました

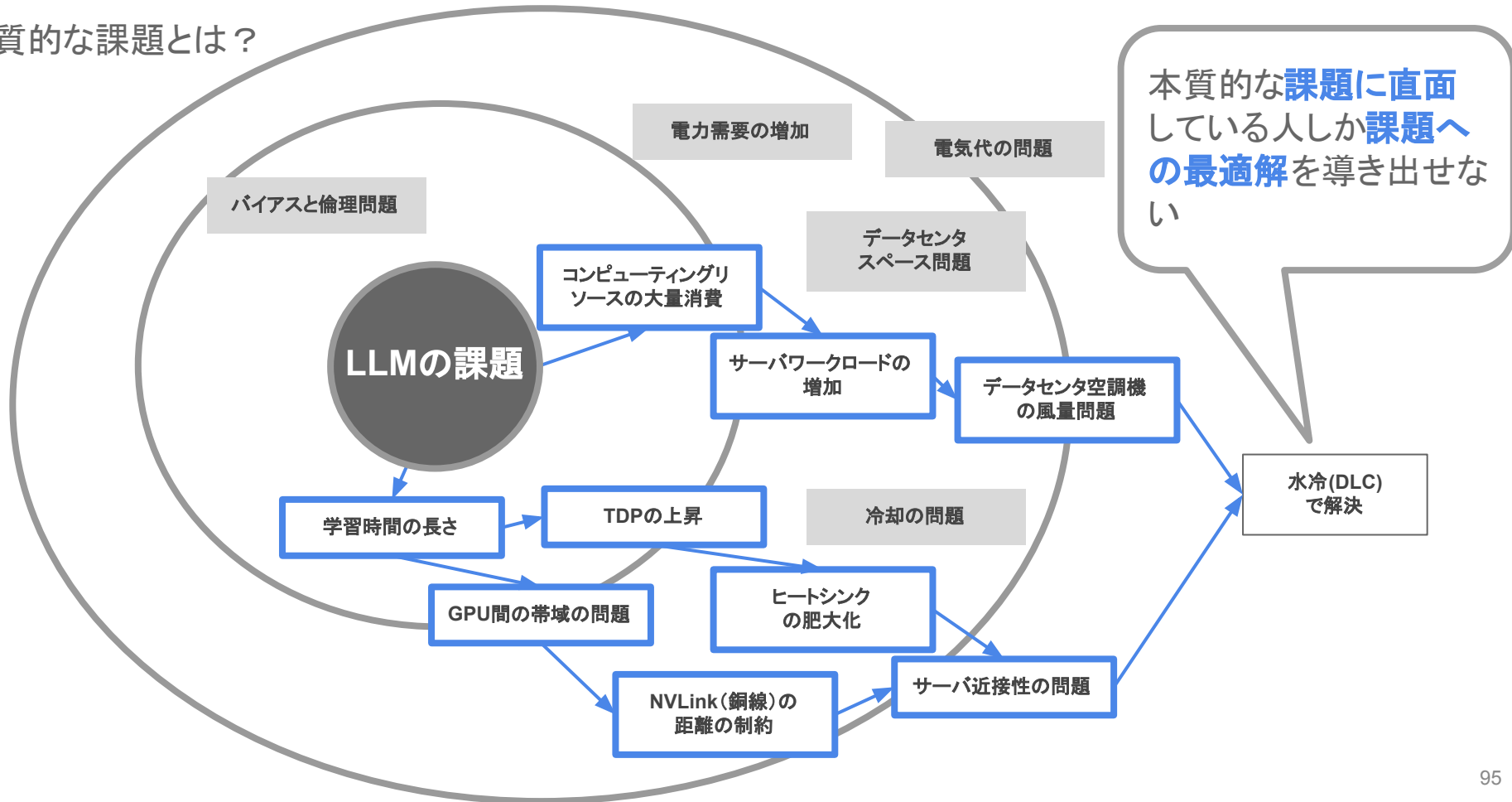
## 本質的な課題とは？



## 本質的な課題とは？



本質的な課題とは？



本質的な課題に直面している人しか課題への最適解を導き出せない

## 第一部

## 第二部

GTCでGB200 NVL72の発表  
GB200 NVL72の電源が480Vと判明

'22  
Nov

'23  
Mar

'24  
Feb

'24  
Mar

'24  
May

ChatGPT 3.5がリリースされる

ヤフーのサービスでもLLMの本格利用開始

データセンタの風量不足を調査

米国企業(ハイパースケラ)などはこの時点で**直面した課題**の解決を進めている



## 第一部

## 第二部

GTCでGB200 NVL72の発表  
GB200 NVL72の電源が480Vと判明

'22  
Nov

'23  
Mar

'24  
Feb

'24  
Mar

'24  
May

ChatGPT 3.5がリリースされる

ヤフーのサービスでもLLMの本格利用開始

データセンタの風量不足を調査



- ・我々はやっこの時点でいろいろな課題に**遭遇し始める**
- ・そもそも課題に遭遇するのが**米国企業より遅い**

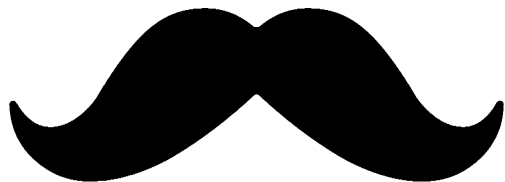
課題に直面してない人が、  
これから直面するであろう課題の解決方法を見誤らない方法



- **課題に直面**した人に、**本質的な課題**は何かを情報収集
- LLMで言うと**米国企業などハイパースケーラ**
- 集めた情報を**近く(米国)ですぐに実践**することが大事

- Actapioの課題解決アプローチ
  - 課題に直面した人(米国企業)から情報を収集・検証・実施しているので**最善の技術を選択**できている
- 未来へのビジョン
  - **「利用する側」は後手**にまわってしまう
  - より業界をリードするには**「課題解決の先駆者」**にならないと
  - 私たちが未来に向けて目指していくのは、迅速に技術を取り入れる「利用する側」に徹するのか、「課題解決の先駆者」を目指し業界のリーダを目指すのか。どちらなのでしょう？
  - どちらの道を選ぶにしても、**今日から**必要な取り組みを共に考え、未来を切り拓きませんか？





# THANKS

## Our Team

Aaron Kirkpatrick  
Andrew Arnold  
Anthony Skwiat  
Atsuko Ishigaki  
Eiji Kawauchi  
Jason Van Winkle  
JD Salling  
Kai Fukazawa  
Ken Tairabune  
Koyo Uemizu  
Kyoya Torikai  
Hideki Mikami  
Hisatomo Tanaka  
Marie Rudolph

Masahiko Matsui  
Masayuki Ashida  
Osamu Kurino  
Shinichiro Okamoto  
Susan McKee  
Takashi Watanabe  
Takuya Kitano  
Takumi Uematsu  
Tatsumi Yuusuke  
Travis Mather  
Tsubasa Yoshimura  
Yuji Kohata  
Yukihito Imai

## Our Datacenter

