

# Express5800/R120k-2M NVIDIA RTX PRO 6000 Blackwell Server Edition NVIDIA Run:ai MCP Server 動作検証レポート

日本電気株式会社



## 商標

Linux は Linus Torvalds 氏の日本およびその他の国における商標または登録商標です。

Red Hat、Red Hat Enterprise Linux は、米国およびその他の国における Red Hat, Inc. の商標または登録商標です。

NVIDIA は、米国およびその他の国における NVIDIA Corporation の商標または登録商標です。

Kubernetes は、The Linux Foundation の米国およびその他の国における登録商標または商標です。

その他、記載されている会社名、製品名は、各社の登録商標または商標です。

## 免責条項

本書または本書に記述されている製品や技術に関して、日本電気株式会社またはその関連会社が行う保証は、製品または技術の提供に適用されるライセンス契約で明示的に規定されている保証に限ります。このような契約で明示的に規定された保証を除き、日本電気株式会社およびその関連会社は、製品、技術、または本書に関して、明示または黙示を問わず、いかなる種類の保証も行いません。

# 目次

|                                                             |    |
|-------------------------------------------------------------|----|
| 動作検証について .....                                              | 3  |
| 1    ご利用にあたっての注意事項 .....                                    | 3  |
| 2    NVIDIA RTX PRO 6000 Blackwell Server Edition の概要 ..... | 3  |
| 3    NVIDIA Run:ai MCP Server の 概要 .....                    | 3  |
| 4    検証目的 .....                                             | 3  |
| 5    動作検証の準備 .....                                          | 4  |
| 5.1 動作検証システム構成 .....                                        | 4  |
| 5.2 動作検証済のサーバ構成 .....                                       | 5  |
| 5.2.1 Express5800/R120k-2M サーバ手配構成 .....                    | 5  |
| 5.2.2 NVIDIA GPU 搭載時に搭載可能な電源ユニット .....                      | 6  |
| 5.3 Kubernetes 構築手順 .....                                   | 7  |
| 5.4 NVIDIA Run:ai 構築手順 .....                                | 7  |
| 5.5 NVIDIA Run:ai MCP Server 構築手順 .....                     | 7  |
| 6    検証結果 .....                                             | 8  |
| 6.1 クラスタと GPU リソースの状況確認 .....                               | 8  |
| 6.2 ワークロードの問題調査 .....                                       | 9  |
| 6.3 性能問題の調査と改善 .....                                        | 10 |
| 7    関連リンク .....                                            | 12 |
| 8    お問い合わせ先 .....                                          | 12 |
| 9    改版履歴 .....                                             | 12 |

# 動作検証について

## 1 ご利用にあたっての注意事項

本レポートは動作検証レポートであり、弊社が動作保証するものではありません。  
動作確認情報は、各ページに掲載されている評価環境での検証結果に基づいたものです。  
導入に際しては個々の環境で十分な確認を実施してください。

## 2 NVIDIA RTX PRO 6000 Blackwell Server Edition の概要

革新的な NVIDIA Blackwell アーキテクチャを搭載した NVIDIA RTX PRO™ 6000 Blackwell Server Edition は、最先端の AI 処理能力と卓越したビジュアル・コンピューティング性能を統合し、企業データセンターの複雑なワークロードを劇的に加速します。超高速 GDDR7 メモリを 96GB 搭載した NVIDIA RTX PRO 6000 Blackwell は、エージェント型 AI、フィジカル AI、科学計算からレンダリング、3D グラフィックス、ビデオ処理まで幅広いユースケースに対応する比類なきパフォーマンスと柔軟性を提供します。

## 3 NVIDIA Run:ai MCP Server の 概要

NVIDIA Run:ai は、Kubernetes を基盤として、複数のユーザーや AI ワークロードに対する GPU リソースの割り当てやスケジューリングを管理し、GPU の利用効率を向上させるためのオーケストレーションプラットフォームです。

NVIDIA Run:ai MCP Server は、Model Context Protocol (MCP) を介して、MCP に対応した AI アシスタントと NVIDIA Run:ai クラスターを連携させるためのコンポーネントです。AI アシスタントから自然言語で指示することにより、MCP Server が提供するツールを通じて、NVIDIA Run:ai の UI や API を直接操作することなく、ワークロードやクラスターの状態、GPU リソースの利用状況などの確認、問題の調査、構成情報の確認や設定変更を行うことができます。

これらにより、GPU リソースを効率的に活用した AI ワークロードの運用と、AI アシスタントを介した NVIDIA Run:ai クラスターの管理が可能になります。

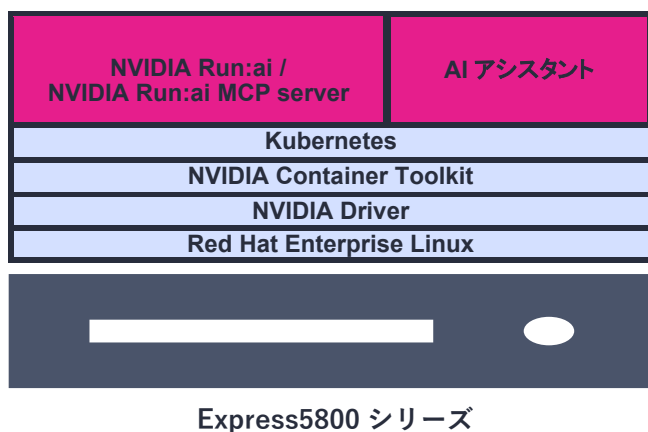
## 4 検証目的

今回の検証では、Express5800 シリーズに NVIDIA RTX PRO 6000 Blackwell Server Edition を搭載し、後述の環境下にて、NVIDIA Run:ai MCP Server の基本動作検証を実施しました。その結果を記載します。

# 5 動作検証の準備

## 5.1 動作検証システム構成

弊社において動作検証を実施した構成を掲載いたします。なお、下記は一例ですので、お客様の環境や用途にあわせてシステムを構成してください。



検証作業者

## 5.2 動作検証済のサーバ構成

本章では、動作検証を実施した Express5800 R120k-2M についてのサーバ手配構成 / 構成に応じた電源ユニットの選択方法 / 動作検証条件について説明します。

### 5.2.1 Express5800/R120k-2M サーバ手配構成

| 製品名                                                       | 対象型名          | 数量 | 補足事項               |
|-----------------------------------------------------------|---------------|----|--------------------|
| Express5800/R120k-2M 8x2.5型ドライブモデル (U.3 NVMe x1/SAS/SATA) | N8100-3034Y   | 1  | 8x2.5 型<br>ドライブモデル |
| CPUボード (16C/3.20GHz/6517P)                                | N8101-1913    | 2  |                    |
| 64GB増設メモリボード (1x64GB/R/DR)                                | N8102-773     | 16 |                    |
| メモリダミーキット                                                 | N8102-746     | 1  |                    |
| 製造指示 (温度条件なし)                                             | NESV16-074    | 1  |                    |
| 増設用2.5型3.84TB SATA R1 SSD                                 | N8150-1829    | 4  |                    |
| RAIDコントローラ (MR, 4GB, RAID 0/1/5/6, OCP)                   | N8103-249     | 1  |                    |
| フラッシュバックアップユニット                                           | N8103-218     | 1  |                    |
| 増設バッテリー用ケーブル                                              | K410-513(00)  | 1  |                    |
| 内蔵NVMe/SAS/SATA OCP型RAIDコントローラ接続ケーブル                      | K410-573(00)  | 1  |                    |
| 2U内蔵DVDドライブ増設キット                                          | N8154-195     | 1  |                    |
| 内蔵DVD-ROM ドライブ                                            | N8151-137     | 1  |                    |
| 内蔵 DVD ドライブ接続ケーブル                                         | K410-569(00)  | 1  |                    |
| 1stライザカード (3xPCI + 1xGPU搭載キット)                            | N8116-112     | 1  |                    |
| 2ndライザカード (3xPCI + 1xGPU搭載キット)                            | N8116-113     | 1  |                    |
| 3rdライザカード (2xPCI)                                         | N8116-119     | 1  |                    |
| 製造指示 (x16カード接続対応)                                         | NESV16-063    | 1  |                    |
| ライザカード接続ケーブル                                              | K410-583(00)  | 2  |                    |
| 1000BASE-T 接続LOMカード (4ch)                                 | N8104-222     | 1  |                    |
| OCPカード接続ケーブル (1st CPU側)                                   | K410-570(00)  | 1  |                    |
| 10GBASE-T 接続ボード (2ch)                                     | N8104-219     | 1  |                    |
| 電源ユニット (1800W)                                            | N8181-F210    | 2  |                    |
| ACケーブル (3m)                                               | K410-E162(03) | 2  |                    |
| 2U 高性能ヒートシンク                                              | N8101-1926    | 2  |                    |
| 2U 高性能ファン                                                 | N8181-209     | 1  |                    |

|                                            |              |   |
|--------------------------------------------|--------------|---|
| リモートマネジメント拡張ライセンス (Advanced)               | N8115-33     | 1 |
| 2 U ケーブルアーム                                | N8143-154    | 1 |
| グラフィックスカード電源ケーブル (12 + 4 pin)              | K410-527(00) | 1 |
| 1 U 高性能ヒートシンク                              | N8101-1924   | 2 |
| GPU 冷却キット                                  | N8116-124    | 1 |
| GPU コンピューティングカード (NVIDIA RTX PRO 6000 BSE) | N8105-75     | 1 |

その他増設オプションについては、Express5800/R120k-2 システム構成ガイドを参照の上、手配ください。  
(<https://jpn.nec.com/pcserver/systemguide/100guide.html>)

## 5.2.2 NVIDIA GPU 搭載時に搭載可能な電源ユニット

搭載するオプション(メモリ、ディスク等)により、システムに要求される電力量が異なります。導入に際しては個々の環境で十分な確認を実施してください。

### 動作検証条件および搭載制限オプション

| 分類          | GPU 搭載枚数 : 1 枚                                                                                                | GPU 搭載枚数 : 2 枚                                                                                                | GPU 搭載枚数 : 3 枚       |
|-------------|---------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------|----------------------|
| CPU         | CPU TDP: 300W まで搭載可能                                                                                          | CPU TDP: 270W まで搭載可能                                                                                          | CPU TDP: 225W まで搭載可能 |
| 内蔵ドライブ      | 搭載可能台数 : 8 台以下                                                                                                |                                                                                                               |                      |
| メモリ         | RDIMM: 16 枚以下                                                                                                 |                                                                                                               |                      |
| 増設ドライブケース   | 搭載不可                                                                                                          |                                                                                                               |                      |
| PCI カード     | 4 枚まで搭載可能                                                                                                     |                                                                                                               | 2 枚まで搭載可能            |
| 防塵フィルタ      | 搭載不可                                                                                                          |                                                                                                               |                      |
| RAID コントローラ | 制限なし                                                                                                          |                                                                                                               |                      |
| 動作環境温度      | N8100-3034Y 8x2.5 型ドライブモデル(U.3 NVMe x1/SAS/SATA) : 25 度以下<br>N8100-3035Y 8x2.5 型ドライブモデル(U.3 NVMe x4) : 25 度以下 | N8100-3034Y 8x2.5 型ドライブモデル(U.3 NVMe x1/SAS/SATA) : 20 度以下<br>N8100-3035Y 8x2.5 型ドライブモデル(U.3 NVMe x4) : 20 度以下 |                      |

#### 補足事項:

- PCI カードの枚数に N8105-75 GPU コンピューティングカード(NVIDIA RTX PRO 6000 BSE)、RAID コントローラ(専用スロット型)、LOM カードは含みません。
- K410-527(00)グラフィックスカード電源ケーブル(12+4Pin)は 1 セットで 3 本の補助電源ケーブルが含まれます。
- 本 GPU 搭載時、電源を必ず 2 台搭載してください。また、電源は非冗長構成となります。
- ファンによる騒音が大きいいため、専用のサーバ室での運用をお願いいたします。

## 5.3 Kubernetes 構築手順

『Kubernetes Documentation』を参照し、Kubernetes クラスターを構築してください。  
また、containerd を導入し、GPU Operator をインストールしてください。

<https://kubernetes.io/docs/setup/>

<https://containerd.io/docs/2.2/>

<https://docs.nvidia.com/datacenter/cloud-native/gpu-operator/26.3/>

## 5.4 NVIDIA Run:ai 構築手順

『NVIDIA Run:ai DOCS』を参照し、NVIDIA Run:ai を構築してください。  
また、NVIDIA Run:ai の構築にあたり、Ingress Controller、MetalLB、Local Path Provisioner、Knative Serving、および Prometheus を導入してください。

<https://run-ai-docs.nvidia.com/self-hosted/2.26/getting-started/installation>

## 5.5 NVIDIA Run:ai MCP Server 構築手順

『NVIDIA Run:ai DOCS』を参照し、NVIDIA Run:ai MCP Server を構築してください。  
本動作検証では、NVIDIA Run:ai MCP Server を Local Mode (STDIO) で構築し、AI アシスタントとして GitHub Copilot CLI を使用しました。検証時の AI モデルには Claude Sonnet 5 を使用しています。  
AI アシスタントとの接続設定例については、以下の NVIDIA Run:ai DOCS を参照してください。

<https://run-ai-docs.nvidia.com/self-hosted/2.26/getting-started/mcp-server#configure-your-ai-tool>

## 6 検証結果

NVIDIA RTX PRO 6000 Blackwell Server Edition を搭載した Express5800 R120k-2M において、NVIDIA Run:ai の動作検証を行った結果、本レポートに記載した検証範囲において問題なく動作することを確認しました。なお、本検証では MIG を有効化し、1 枚の物理 GPU を 2 つの GPU インスタンスとして利用する構成で評価を実施しています。

動作検証時の主要ソフトウェアのバージョン・内容は下記になります。

- OS Red Hat Enterprise Linux 9.6
- GPU ドライバー NVIDIA Driver 590.48.01
- NVIDIA GPU Operator v26.3.3
- NVIDIA Run:ai v2.26.23

### 6.1 クラスターと GPU リソースの状況確認

NVIDIA Run:ai MCP Server を使用して、クラスターの稼働状況、ワークロードの状態、および GPU リソースの割り当て・利用状況を確認しました。AI アシスタントへ自然言語で指示することにより、これらの情報を確認できるかを検証しました。

```
Copilot v1.0.83 uses AI.
Check for mistakes.

Tip: /clear
  Abandon this session and start fresh

> クラスタ全体は正常? GPUの空き状況とワークロードの状態を教えてください

Thought for 1s
I need to find the cluster ID first. There's no explicit list_clusters tool, so I should check whoami or try listing projects/departmen

whoami (MCP: runai-stdio) {"email":"sa-mcp-server","subject":"sa-mcp-server","token_type":"Bearer","use...

list_node_pools (MCP: runai-stdio) {"hasMore":false,"nodePools":[{"clusterId":"66df36d7-7bf4-43a6-8702-b66c8a289...

クラスタIDを取得できました。健全性・GPU状況・ワークロード状況をまとめて確認します。

get_cluster_infrastructure_health (MCP: runai-stdio) · clusterId: "66df36d7-7bf4-43a6-8702-b66c8a2892ef" {"asOf":"2026-09-15T10:35:28Z"

get_cluster_physical_inventory (MCP: runai-stdio) · clusterId: "66df36d7-7bf4-43a6-8702-b66c8a2892ef" {"asOf":"2026-09-15T10:35:29Z","b

get_workloads_summary (MCP: runai-stdio) · clusterId: "66df36d7-7bf4-43a6-8702-b66c8a2892ef" {"asOf":"2026-09-15T10:35:28Z","byPhase":{

クラスタ状況 (2026-09-15 10:35 UTC時点)

インフラ健全性: ☒ 正常 - ノード1台、すべてHealthy、GPU使用不可(unschedulable)な状態なし

GPU空き状況:

・ GPUモデル: NVIDIA RTX PRO 6000 Blackwell (MIG 2g.48gb)
・ 合計2 GPU、空き0 (全て割当済み・使用率100%)
・ CPU: 21.4/64コア使用、メモリ: 30.6/1007 GiB使用 (余裕あり)

ワークロード状態:

・ 合計2件、すべて Running
・ Pending/Failed/Stopped は0件 (バックログなし)

結論: クラスタ自体は健全ですが、GPUは完全に使い切っている状態です。新規GPUワークロードを投入する場合は待機が発生する可能性があります。
```

## 6.2 ワークロードの問題調査

起動に失敗するワークロードを対象として、NVIDIA Run:ai MCP Server を使用して問題調査を実施しました。AI アシスタントへ自然言語で指示することにより、ワークロードの状態確認、関連情報の取得、および原因調査ができるかを検証しました。

2つのワークロードが起動に失敗している状態になりました。

| Workload ↑                                                                                    | Type      | Status              | Project                         | Department | Preemptibility  | Running / requested pods | GPU compute request | GPU compute allocation |
|-----------------------------------------------------------------------------------------------|-----------|---------------------|---------------------------------|------------|-----------------|--------------------------|---------------------|------------------------|
|  chat-api-v1 | Inference | Initializing (6m) ⓘ | <a href="#">pj-chat-service</a> | default    | Non-preemptible | 0/1-1                    | 1                   | 1                      |
|  chat-api-v2 | Inference | Pending (23m) ⓘ     | <a href="#">pj-chat-service</a> | default    | Non-preemptible | 0/1-1                    | 3                   | -                      |

上記の状態を対象に AI アシスタントへ調査を依頼し、原因を特定しました。

```
Copilot v1.0.83 uses AI.
Check for mistakes.

Tip: /diff
  Review the changes made in the current directory

> chat-api-v1 が起動しない、原因を教えてください

Thought for 1s
I need to locate the workload named chat-api-v1, so I'll search for it using list_workloads.

list_workloads (MCP: runai-stdio) · name: "chat-api-v1" {"hasMore":false,"workloads":[{"allocatedGpu":1,"clusterId":"66df36d7-7bf4-43...

ワークロードが見つかりました。状態を確認します。

get_workload_status (MCP: runai-stdio) · workloadId: "7bb1cf26-185d-4a8f-86f7-b7d8a99dccc4" {"clusterId":"66df36d7-7bf4-43a6-8702-b66c8a2892ef","conditions":[{"creat...

get_workload_spec (MCP: runai-stdio) · workloadId: "7bb1cf26-185d-4a8f-86f7-b7d8a99dccc4" {"allocatedResources":{"cpu":0.13,"cpuMemory":"104857600","gpu":1,"gpuMemory...

原因: コンテナイメージが存在しない (ImagePullエラー)

指定されたイメージタグが誤っています。

nvcr.io/nim/openai/gpt-oss-20b:this-tag-does-not-exist-demo

このタグ this-tag-does-not-exist-demo は NGC レジストリに存在せず、ErrImagePull / BackOff を繰り返して起動できていません (Readiness probe も当然タイムアウト)。

対処法: runai inference submit の -i に指定する正しいイメージタグ (例:
nvcr.io/nim/openai/gpt-oss-20b:<正しいバージョンタグ>) に修正し、再作成してください。NGC上で利用可能なタグを確認の上、再デプロイをおすすめします。

> chat-api-v2 が Pendingなのはなぜ?

list_workloads (MCP: runai-stdio) · name: "chat-api-v2" {"hasMore":false,"workloads":[{"allocatedGpu":null,"clusterId":"66df36d7-7bf4...

get_workload_status (MCP: runai-stdio) · workloadId: "322b4b46-02d7-485a-85d5-f61bc0cccd797" {"clusterId":"66df36d7-7bf4-43a6-8702-b66c8a2892ef","conditions":[{"creat...

原因: GPUクォータ超過 + ノード単体GPU上限超過 (2つの問題)

1. クォータ超過: pj-chat-service プロジェクトのGPUクォータは 1。既に非プリエンプティブルで1GPU使用中のため、chat-api-v2 が要求する 3 GPU は割り当て不可 (NonPreemptibleOverQuota)。
2. ノード上限超過: そもそも default ノードプールの1ノードあたり最大GPU数は 2 で、3 GPU要求は単一ノードに収まらずスケジューリング不可。

対処法:

・ GPU要求数を減らす (最大2以下、かつクォータ内に収める)
・ プロジェクトのGPUクォータを増やす (list_project_resources / manage_org_unit で確認・変更)
・ または preemptible ワークロードとして提出しオーバークォータを許可する
```

## 6.3 性能問題の調査と改善

負荷の増加に対して推論ワークロードがスケールアウトできない状況を対象として、NVIDIA Run:ai MCP Server を使用して性能問題の調査を実施しました。AI アシスタントへ自然言語で指示することにより、問題の原因調査、改善策の提示、設定変更の実施、および改善後の状態確認を行えるかを検証しました。

負荷増加により推論ワークロードがスケールアウトできず、Pending 状態となる状況が発生させました。

| Workload ↑                                                                                      | Type      | Status       | Project         | Department | Preemptibility  | Running / requested pods | GPU compute request | GPU compute allocation |
|-------------------------------------------------------------------------------------------------|-----------|--------------|-----------------|------------|-----------------|--------------------------|---------------------|------------------------|
|  chat-api-prod | Inference | Pending (1m) | pj-chat-service | default    | Non-preemptible | 1/1-2                    | 2                   | 1                      |

上記の Pending 状態を対象に AI アシスタントへ調査を依頼し、原因と改善策の提案を得ました。

```
Copilot v1.0.83 uses AI.
Check for mistakes.

Tip: /tasks
  View and manage tasks (subagents and shell commands)

> chat-api-prod の応答が遅いと相談がきている。原因を調べて、改善策があれば教えて

Thought for 1s
I should look into "chat-api-prod" —it's probably a Run:ai workload, so I'll search for it using list_workloads.

list_workloads (MCP: runai-stdio) · name: "chat-api-prod" {"hasMore":false,"workloads":[{"allocatedGpu":1,"clusterId":"66df36d7-7bf4-4

Thought for 1s
The pod's phase is Pending, which would explain why it's not actually running yet rather than just being slow. I should check it's stuck there.

chat-api-prodは現在Pending状態。原因を詳しく調べます。

get_workload_status (MCP: runai-stdio) · workloadId: "0c721b4e-216b-41a0-a2bb-938e49bff323" {"clusterId":"66df36d7-7bf4-4

get_workload_spec (MCP: runai-stdio) · workloadId: "0c721b4e-216b-41a0-a2bb-938e49bff323" {"allocatedResources":{"cpu":0.

Thought for 1s
Found it —scaling to 2 replicas requires a second non-preemptible GPU, but the project quota only allows 1 GPU which is why while requests queue on the single existing replica past its concurrency threshold. I should check the GPU quota for pj-chat-service.

list_project_resources (MCP: runai-stdio) · projectName: "pj-chat-service" {"hasMore":false,"resources":[{"cpuCores":{"a

原因

chat-api-prod は現在 Pending（新レプリカが起動できずスケールアウト失敗）状態です。これが応答遅延の直接原因です。

・ 設定: min-replicas 1, max-replicas 2, concurrency閾値2 でオートスケーリングが有効
・ 同時実行数が閾値を超え、2台目のレプリカ（GPU 1基要求・non-preemptible）を起動しようとした
・ しかし pj-chat-service プロジェクトのGPUクォータは1基で、既に1基が非preemptibleで使用中
・ Run:aiスケジューラが NonPreemptibleOverQuota で2台目をスケジュール不可 →Pendingのまま
・ 結果、既存の1台だけで捌ききれずリクエストが詰まり、応答が遅くなっている

改善策（優先度順）

1. プロジェクトのGPUクォータを増やす（推奨・恒久対策）
   pj-chat-service の default ノードプールでのGPU deserved quotaを1-2以上に引き上げる
2. ワークロードをpreemptibleに変更
   over-quota分をpreemptibleにすれば、クォータ超過でもスケジュール可能に（ただし他ジョブに割り込まれるリスクあり）
3. 暫定対応: concurrency閾値を上げてスケールアウトを抑制、または他プロジェクトから使っていないGPUを回収してクォータ移管

対応を進めますか？（例：クォータを1-2に増やす、など）具体的な変更値を教えてください。実行します。
```

AI アシスタントが提示した改善策に基づき、設定変更の実施を指示しました。

```
> 1を実施してください

Thought for 1s
I need to update the GPU deserved quota from 1 to 2 on the default node pool for pj-chat-service, using the org unit update action target. I will dry run first to verify before applying the actual change.

• manage_org_unit (MCP: runai-stdio) · action: "update", orgType: "project", name: "pj-chat-s- {"action":"update","after":{"clusterId":"66df36d7-7bf4-43a6-8702-b66c8801-216b-41a0-a2bb-938e49bff323"}

• プレビュー確認: GPUクォータ 1→2 (default プール)。適用します。

• manage_org_unit (MCP: runai-stdio) · action: "update", dryRun: false, confirm: true, name: "... {"action":"update","after":{"clusterId":"66df36d7-7bf4-43a6-8702-b66c8801-216b-41a0-a2bb-938e49bff323"}

• 適用しました (GPUクォータ1→2)。ワークロードの状態を確認します。

• get_workload_status (MCP: runai-stdio) · workloadId: "0c721b4e-216b-41a0-a2bb-938e49bff323" {"clusterId":"66df36d7-7bf4-43a6-8702-b66c8801-216b-41a0-a2bb-938e49bff323"}

• まだイベント/フェーズが更新前の状態を表示しています (スケジューラの再評価に数十秒～数分かかることがあります)。少し待って再確認します。

$ Shell Wait for scheduler to re-evaluate pending workload 1 line...
sleep 45

• 待機中です。完了後に再確認します。

• Read shell output Waiting up to 20 seconds for command output 1 line...

• get_workload_status (MCP: runai-stdio) · workloadId: "0c721b4e-216b-41a0-a2bb-938e49bff323" {"clusterId":"66df36d7-7bf4-43a6-8702-b66c8801-216b-41a0-a2bb-938e49bff323"}

• ✓ 解消しました。

  • pj-chat-service のGPUクォータを 1→2 に更新
  • 2台目のレプリカがスケジュール可能になり、chat-api-prod は Pending →Running に遷移 (コンテナ起動完了)
  • これで concurrency閾値超過時も2レプリカで負荷分散でき、応答遅延は解消される見込みです

引き続き様子を見て、再度遅延が出るようであれば get_workload_metrics でレイテンシ/スループットを監視することをお勧めします。
```

設定変更後、推論ワークロードがスケールアウトして Running 状態となったことを確認しました。

| Workload ↑                                                                                        | Type      | Status        | Project                         | Department | Preemptibility  | Running / requested pods | GPU compute request | GPU compute allocation |
|---------------------------------------------------------------------------------------------------|-----------|---------------|---------------------------------|------------|-----------------|--------------------------|---------------------|------------------------|
|  chat-api-prod | Inference | Running (10m) | <a href="#">pj-chat-service</a> | default    | Non-preemptible | 2/1-2                    | 2                   | 2                      |

## 7 関連リンク

- NVIDIA Run:ai DOCS  
<https://run-ai-docs.nvidia.com/self-hosted/2.26/>
- NVIDIA Run:ai MCP Server  
<https://run-ai-docs.nvidia.com/self-hosted/2.26/getting-started/mcp-server>
- NVIDIA runai-mcp-server  
<https://catalog.ngc.nvidia.com/orgs/nvidia/runai/containers/runai-mcp-server/2.26.27/>
- GitHub Copilot CLI  
<https://github.com/features/copilot/cli>

## 8 お問い合わせ先

NEC インフラ・テクノロジーサービス事業部門 コンピュータ統括部  
[pp\\_support@hetero.jp.nec.com](mailto:pp_support@hetero.jp.nec.com)

## 9 改版履歴

| 版数    | 公開日時       | 変更内容      |
|-------|------------|-----------|
| 第 1 版 | 2026 年 9 月 | 第 1 版リリース |