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

日本電気株式会社

目次

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 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 を導入することで、GPU リソースの共有、負荷に応じた推論ワークロードの水平スケーリング、NVIDIA NIM を含む推論ワークロードのデプロイおよび稼働状況の監視が可能になります。これにより、限られた GPU リソースを効率的に活用しながら、AI ワークロードを運用できます。

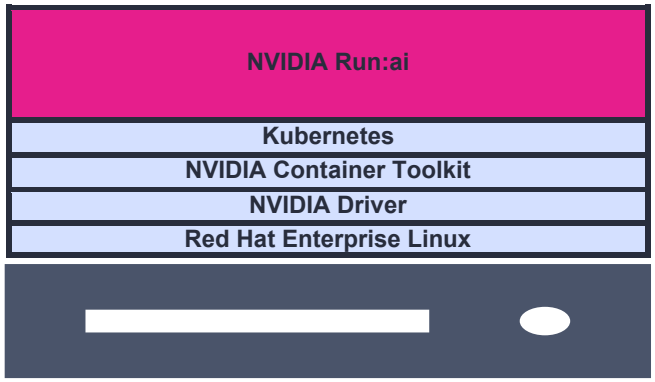
4　検証目的

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 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。



Express5800 シリーズ

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 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　ＲＩ　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
２Ｕ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＋４ｐｉｎ）	K410-527(00)	1
１Ｕ高性能ヒートシンク	N8101-1924	2
GPU ｻﾎﾟｰﾄｷｯﾄ	N8116-124	1
GPU ｺﾝﾋﾟｭｰﾃｨﾝｸﾞ ｶｰﾄﾞ (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		
増設ドライブケージ	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 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 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

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。

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。

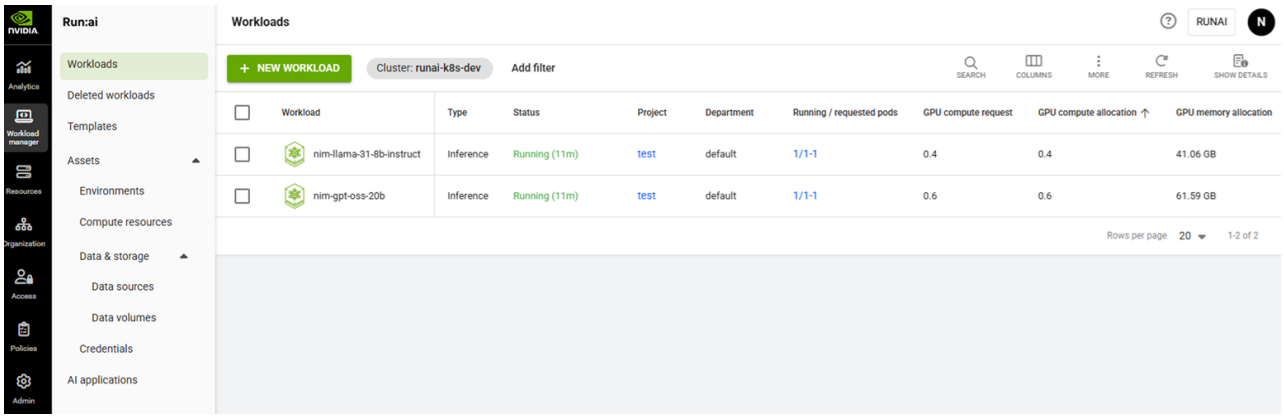
- OS　Red Hat Enterprise Linux 9.6
- GPU ドライバー　NVIDIA Driver 580.82.07
- NVIDIA GPU Operator　v26.3.3
- NVIDIA Run:ai　v2.26.23

6.1 GPU フラクション

GPU フラクションは、1 枚の GPU を複数のワークロードで分割し、それぞれに決まった割合の GPU メモリを割り当てて同時に共有利用できるようにする機能です。

NVIDIA NIM の meta/llama-3.1-8b-instruct と openai/gpt-oss-20b の 2 モデルを、1 枚の GPU 上にそれぞれ 40%および 60%の GPU フラクションを固定割り当てし、同時に稼働させました。
両モデルに対して同時に推論リクエストを送信した結果、両モデルとも正常に応答を返し、同時稼働できることを確認しました。

以下は GPU フラクションを適用したワークロードの実行状況を示しています。単一 GPU 上で 2 つの推論ワークロードが稼働しており、それぞれに GPU Memory Allocation を約 41GB および約 62GB で割り当てていることを確認できます。

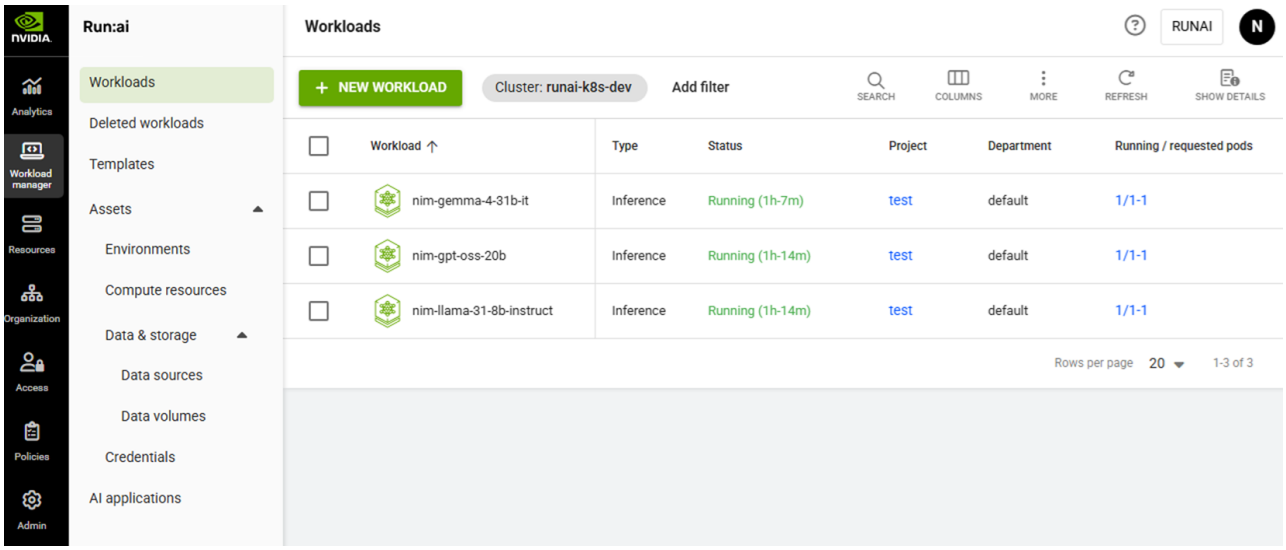


6.2 GPU メモリスワップ

GPU メモリが不足した際に、ワークロードの GPU メモリ上のデータを一時的にホストの CPU メモリへ退避し、必要に応じて GPU メモリへ復元する機能です。これにより、GPU メモリを効率的に活用し、GPU の物理メモリ容量を超える数のワークロードを 1 枚の GPU 上で運用することができます。

NVIDIA NIM の google/gemma-4-31B-it、openai/gpt-oss-20b、meta/llama-3.1-8b-instruct の 3 つのモデルを 1 枚の GPU 上で運用し、それぞれのモデルに対して継続的に推論リクエストを送信する評価を実施しました。GPU メモリスワップを有効化した構成において、各モデルが正常に応答できることを確認しました。

以下は NVIDIA Run:ai 上にデプロイした 3 つの NVIDIA NIM（google/gemma-4-31B-it、openai/gpt-oss-20b、meta/llama-3.1-8b-instruct）の実行状況を示しています。

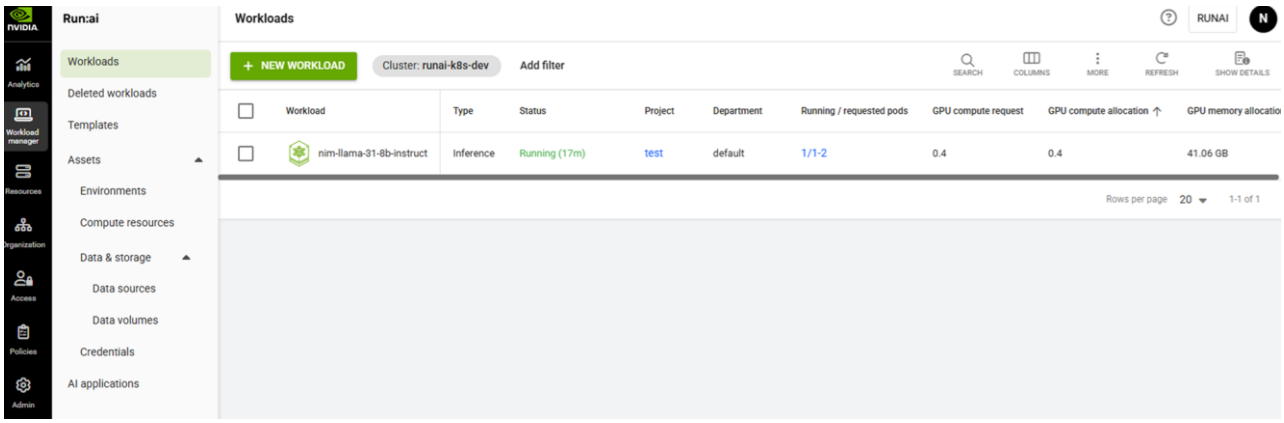


6.3 水平スケーリング

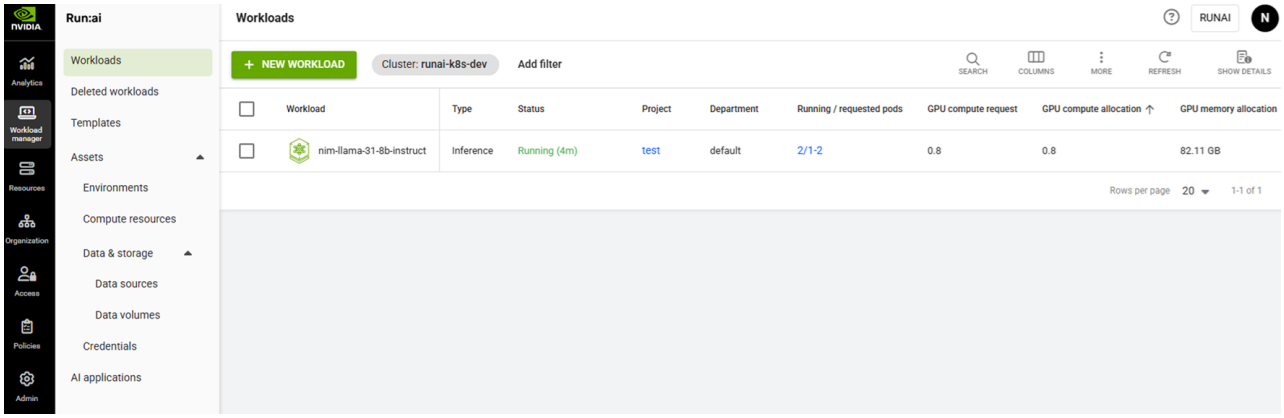
ワークロードの負荷状況に応じて Pod 数を自動的に増減させる機能です。

NVIDIA NIM の meta/llama-3.1-8b-instruct を対象に、リクエスト数（req/sec）を段階的に変化させる評価を実施しました。設定した閾値を超える高負荷時には Pod 数が自動的に増加し、負荷を低下させた際には Pod 数が自動的に減少することを確認しました。

以下は低負荷時の NVIDIA Run:ai の実行状況を示しています。ワークロード nim-llama-31-8b-instruct が Pod 数 1 で稼働しており、Running pods が 1 であることを確認できます。



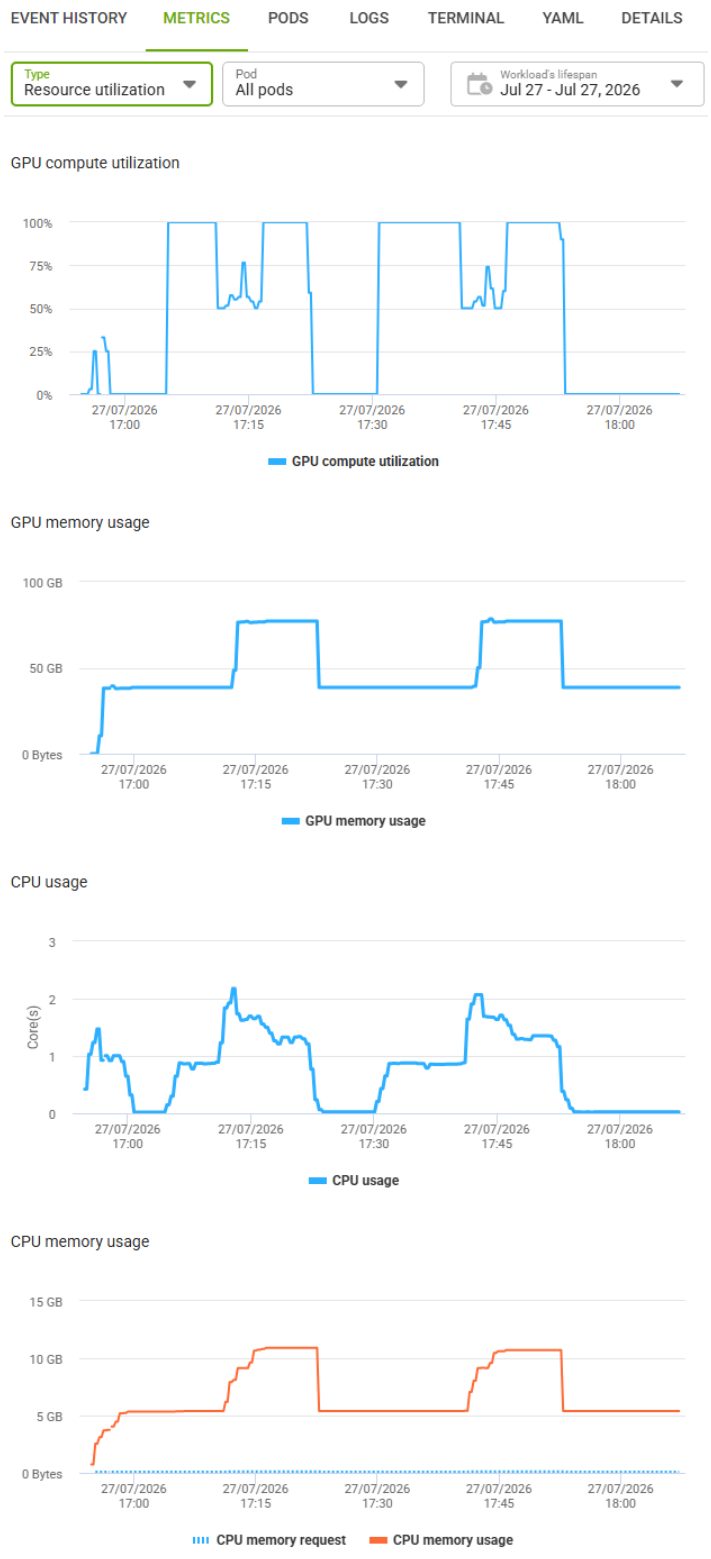
高負荷時の実行状況を示しています。負荷の増加に伴いワークロードの Pod 数が増加し、水平スケーリングによって Pod 数が 2 に増加していることを確認できます。



6.4 リソース監視

推論ワークロードの稼働中に GPU 使用率、GPU メモリ使用量、CPU 使用率および CPU メモリ使用量が取得できることを確認しました。

以下は NVIDIA Run:ai のメトリクス画面を示しています。



7 関連リンク

- NVIDIA Run:ai DOCS
  https://run-ai-docs.nvidia.com/self-hosted/2.26/
- NVIDIA NIM『google/gemma-4-31b-it』
  https://build.nvidia.com/google/gemma-4-31b-it/modelcard
- NVIDIA NIM『meta/llama-3.1-8b-instruct』
  https://build.nvidia.com/meta/llama-3_1-8b-instruct/modelcard
- NVIDIA NIM『openai/gpt-oss-20b』
  https://build.nvidia.com/openai/gpt-oss-20b/modelcard

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 コンピュート統括部
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 7 月	第 1 版リリース