Express5800/R120j-2M(2nd-Gen)
NVIDIA H100 NVL
NVIDIA NIM 動作検証レポート

日本電気株式会社

目次

NVIDIA NIM 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 H100 NVL の概要 3
3 NVIDIA NIM 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/R120j-2M(2nd-Gen)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 NVIDIA NIM 構築手順 7
6 検証結果 8
7 関連リンク 9
8 お問い合わせ先 9
9 改版履歴 9

NVIDIA NIM 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 H100 NVL の概要

NVIDIA H100 NVL は、NVIDIA Hopper™ アーキテクチャを基盤とする次世代 GPU で、業界最高水準の対話型 AI を提供し、大規模言語モデル（LLM）を高速化します。また、卓越したパフォーマンス、拡張性、セキュリティをあらゆるワークロードに提供します。
NVIDIA H100 NVL を搭載したサーバであれば、電力に制限のあるデータセンター環境で低遅延性を維持しながら、パフォーマンスを向上することが可能になります。

3　NVIDIA NIM の概要

NVIDIA NIM は、NVIDIA AI Enterprise に含まれる、生成 AI モデルの展開を加速するために設計されたマイクロサービスのセットです。
NVIDIA NIM には業界標準 API、最適化された推論エンジンなどが含まれています。NVIDIA GPU コンピューティングカード(以降、GPU と表記します)を搭載したオンプレミス環境のコンテナ基盤にデプロイするだけで、NVIDIA NIM コンテナが起動し、標準の REST API を使ったリクエストをすることが可能となります。

NVIDIA NIM for VLMs は、NVIDIA NIM のうち、VLM(Vision Language Models)に対応するもので、
NVIDIA NIM for LLMs は、NVIDIA NIM のうち、LLM(大規模言語モデル)に対応するものです。

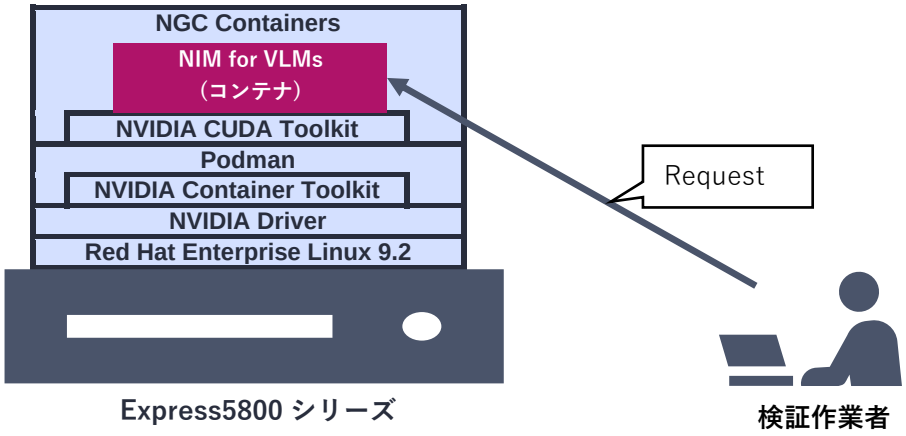
4　検証目的

今回の検証では、Express5800 シリーズに NVIDIA H100 NVL を搭載し、後述の環境下にて、NVIDIA NIM for VLMs および NVIDIA NIM for LLMs 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

弊社において検証済みの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。



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 R120j-2M (2nd-Gen)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/R120j-2M(2nd-Gen)サーバ手配構成

製品名	対象型名	補足事項
Express5800/R120j-2M(2nd Gen)	N8100-3008Y	8x 2.5 型ドライブモデル
1st ライザカード接続ケーブル	K410-509(00)	
2U 用 OS ブートデバイス接続ケーブル	K410-511(00)	
増設バッテリ用ケーブル	K410-513(00)	
OCP カード接続ケーブル（1st CPU 側）	K410-525(00)	
AC ケーブル（3m）	K410-E162(03)	
2U 高性能ヒートシンク	N8101-1857	
CPU ボード（16C/2.80GHz/Gold 6526Y）	N8101-1888	
メモリダミーキット	N8102-746	
64GB 増設メモリボード（1x64GB/R/DR）	N8102-768	
フラッシュバックアップユニット	N8103-218	
RAID コントローラ（SR, 2GB, RAID 0/1/5/6, OCP）	N8103-243	
480GB OS ブート専用 SSD ボード（RAID 1, HS）	N8103-247	
1000BASE-T 接続 LOM カード（4ch）	N8104-206	
10GBASE-T 接続ボード（2ch）	N8104-219	
リモートマネジメント拡張ライセンス（Advanced）	N8115-33	
2nd ライザカード（3xPCI + 1xGPU 搭載キット）	N8116-113	
3rd ライザカード（2xPCI）	N8116-115	
増設用 2.5 型 3.84TB SATA RI SSD	N8150-1829	
内蔵 DVD-ROM ドライブ	N8151-137	
2U 内蔵 DVD ドライブ増設キット	N8154-181	
2U 高性能ファン	N8181-209	
電源ユニット（1800W）	N8181-210	
ケーブルアーム	N8143-150	
GPU コンピューティングカード(NVIDIA H100 NVL)	N8105-71	

その他増設オプションについては、Express5800/R120j-2M(2nd Gen)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**
CPU	CPU TDP: 300W 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: 制限なし
増設ドライブケージ	搭載不可
PCI カード	4 枚まで
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	N8100-3009Y 8x2.5 型ドライブモデル(U.3 NVMe x4/SAS/SATA) ： 25 度以下) N8100-3008Y 8×2.5 型ドライブモデル（U.3 NVMe x1/SAS/SATA） ： 25 度以下）

補足事項:

- PCI カードの枚数に N8105-71 GPU コンピューティングカード(NVIDIA H100 NVL)、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。

5.3 NVIDIA NIM 構築手順

NVIDIA Documentation を参照し、NVIDIA NIM for VLMs および NVIDIA NIM for LLMs を構築してください。

- NVIDIA Documentation『NVIDIA NIM for Vision Language Models(VLMs)』
  https://docs.nvidia.com/nim/vision-language-models/latest/index.html

- NVIDIA Documentation『NVIDIA NIM for Large Language Models(LLMs)』
  https://docs.nvidia.com/nim/large-language-models/latest/introduction.html

6 検証結果

NVIDIA H100 NVL を搭載した Express5800 R120j-2M (2nd-Gen)において、NVIDIA NIM for VLMs および NVIDIA NIM for LLMs の動作検証を行なった結果、問題が発生しない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 9.2
- GPU ドライバー　NVIDIA Driver 550.90.07
- podman　4.9.4
- ContainerToolkit　1.17.4
- NVIDIA NIM　meta/llama-3.2-11b-vision-instruct:latest (検証時点の latest は 1.1.1) (VLM)
  openai/gpt-oss-120b:latest (検証時点の latest は 1.12.1) (LLM)

7 関連リンク

- NVIDIA Documentation『NVIDIA NIM for Vision Language Models(VLMs)』
  https://docs.nvidia.com/nim/vision-language-models/latest/index.html

- NVIDIA Documentation『NVIDIA NIM for Large Language Models(LLMs)』
  https://docs.nvidia.com/nim/large-language-models/latest/introduction.html

- NVIDIA API カタログ『llama-3.2-11b-vision-instruct』
  https://build.nvidia.com/meta/llama-3.2-11b-vision-instruct

- NVIDIA API カタログ『gpt-oss-120b』
  https://build.nvidia.com/openai/gpt-oss-120b

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	2025 年 7 月	第 1 版リリース
第 2 版	2025 年 9 月	openai/gpt-oss-120b の検証結果を追記