

How to Deploy NVIDIA Inference Microservices (NIMs) on Vultr

Learn how to deploy NVIDIA Inference Microservices (NIMs) on Vultr cloud platform with our step-by-step guide for efficient AI model deployment and inference.

Contents

01	Introduction	3
02	Prerequisites	3
03	Setting up NVIDIA NIM	3
04	More Applications of NIMS	6
05	Conclusion	7

Introduction

[NVIDIA NIM \(NVIDIA Inference Microservices\)](#) is a set of containerized services designed to streamline the deployment of generative AI models across various computing environments. It offers easy-to-use APIs for integrating large language models, image generation, and other AI capabilities into enterprise applications. NIM enables organizations to self-host powerful AI models securely, leveraging NVIDIA's GPU acceleration for high-performance inference.

In this article, you will set up NVIDIA NIM on an A100 Server. You'll learn how to deploy a containerized AI model, specifically the Llama 3 8B Instruct model, and interact with it using simple API calls. These steps will demonstrate leveraging NVIDIA's powerful GPU acceleration for AI inference in a secure, self-hosted environment.

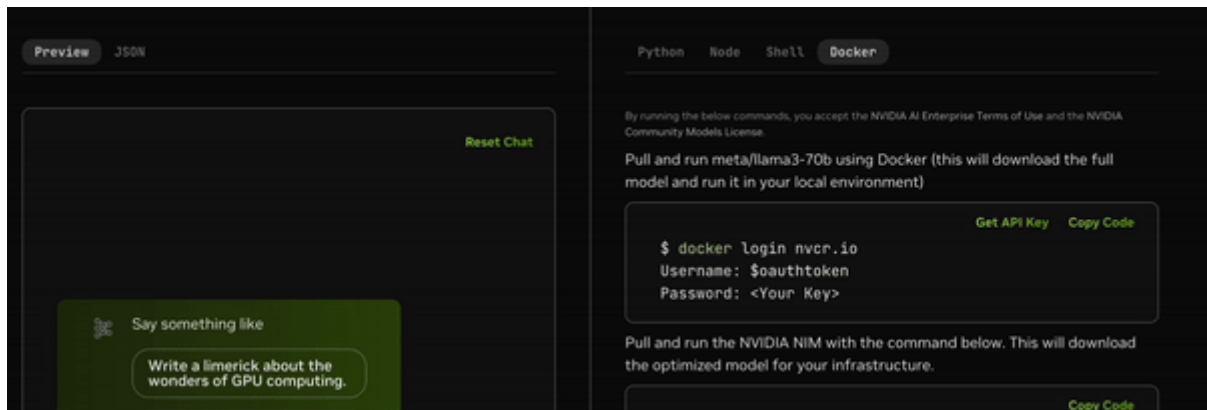
Prerequisites

Before you begin:

- Deploy a fresh [Ubuntu 22.04 NVIDIA A100 Server](#) on Vultr.
- Access the server [using SSH](#) as a non-root user with sudo privileges.

Setting up NVIDIA NIM

1. Explore models in the [Model Catalogue](#).
2. Choose a model.
3. Generate an API key to access the model.



4. Export the API key.

CONSOLE

```
$ export NGC_API_KEY=<value>
```

5. Login to NVIDIA Container Registry to pull the NIM container image.

CONSOLE

```
$ echo "$NGC_API_KEY" | docker login nvcr.io --username '$oauthtoken' --password-stdin
```

You should be able to see the login succeeded output.

6. Export the container image name.

CONSOLE

```
$ export CONTAINER_NAME=llama3-8b-instruct
```

Try [more models](#).

7. Export the NGC image name.

CONSOLE

```
$ export IMG_NAME="nvcr.io/nim/meta/${CONTAINER_NAME}:1.0.0"
```

8. Declare a path to cache downloaded models.

CONSOLE

```
$ export LOCAL_NIM_CACHE=~/.cache/nim && mkdir -p "$LOCAL_NIM_CACHE"
```

9. Allow incoming connections to port `8000`.

CONSOLE

```
$ ufw allow 8000
```

10. Start the Docker container.

CONSOLE

```
$ docker run -it --rm --name=$CONTAINER_NAME \
  --runtime=nvidia \
  --gpus all \
  --shm-size=16GB \
  -e NGC_API_KEY \
  -v "$LOCAL_NIM_CACHE:/opt/nim/.cache" \
  -u $(id -u) \
  -p 8000:8000 \
  $IMG_NAME
```

Initially, the container will be downloaded and then the container will run. The download procedure can take up to 10 minutes.

11. Once the application is running, list the models available for inference.

CONSOLE

```
$ curl -X GET 'http://<server-ip>:8000/v1/models'
```

The above command will list the model you mentioned in the container image, the mentioned model will be available for inference.

12. Infer the model, and send a chat completion request.

CONSOLE

```
$ curl -X POST "http://<server-ip>:8000/v1/chat/completions" \
\
-H "accept: application/json" \
-H "Content-Type: application/json" \
-d "{
    "model": "meta/llama3-8b-instruct",
    "messages": [
      {
        "role": "user",
        "content": "Hello! How are you?"
      },
      {
        "role": "assistant",
        "content": "Hi! I am quite well, how can I
help you today?"
      },
      {
        "role": "user",
        "content": "Can you write me a song?"
      }
    ],
    "max_tokens": 32
  }"
```

More Applications of NIMS

- **Enterprise Chatbot Development:** NVIDIA NIM can be used to deploy Large Language Models for creating advanced customer service chatbots. These chatbots could handle complex queries, provide detailed product information, and offer personalized recommendations, all while running on the company's own secure infrastructure. This would allow businesses to maintain control over their data while leveraging powerful AI capabilities.
- **Financial Fraud Detection:** Banks or financial institutions could implement NVIDIA NIM to host machine learning models that analyze transaction patterns in real time to detect potential fraud. This would allow them to process large volumes of data quickly while keeping sensitive financial information within their own infrastructure.

- **Automated Code Review and Assistance:** Software development firms implement NVIDIA NIM to host coding assistant models. The AI models could be integrated into the development workflow to provide real-time code suggestions, identify potential bugs, and even generate unit tests. By self-hosting the models, companies can ensure that their proprietary code never leaves their secure environment while still benefiting from AI-powered development tools.

Conclusion

By following this article, you've successfully set up NVIDIA NIM on your A100 Server and deployed a large language model. You've learned how to start the Docker service, authenticate with NGC, run the NIM container, and make inference requests to the model. This setup provides a foundation for integrating advanced AI capabilities into your applications while maintaining control over your infrastructure and data.



VULTR

