NVIDIA NIM Offers Optimized Inference Microservices for Deploying AI Models at Scale

Mar 18, 2024 +59 Like Discuss (0)

By Amanda Saunders, Neal Vaidya, Adam Tetelman and Nik Spirin

The rise in **generative AI** adoption has been remarkable. Catalyzed by the launch of OpenAI's ChatGPT in 2022, the new technology amassed over 100M users within months and drove a surge of development activities across almost every industry.

By 2023, developers began POCs using APIs and open-source community models from Meta, Mistral, Stability, and more.

Entering 2024, organizations are shifting their focus to full-scale production deployments, which involve connecting AI models to existing enterprise infrastructure, optimizing system latency and throughput, logging, monitoring, and security, among others. This path to production is complex and time-consuming—it requires specialized skills, platforms, and processes, especially at scale.

NVIDIA NIM, part of **NVIDIA AI Enterprise**, provides a streamlined path for developing AI-powered enterprise applications and deploying AI models in production.

NIM is a set of optimized cloud-native microservices designed to shorten time-to-market and simplify deployment of generative AI models anywhere, across cloud, data center, and GPU-accelerated workstations. It expands the developer pool by abstracting away the complexities of AI model development and packaging for production using industry-standard APIs.

NVIDIA NIM for optimized AI inference

NVIDIA NIM is designed to bridge the gap between the complex world of AI development and the operational needs of enterprise environments, enabling 10-100X more enterprise application developers to contribute to AI transformations of their companies.

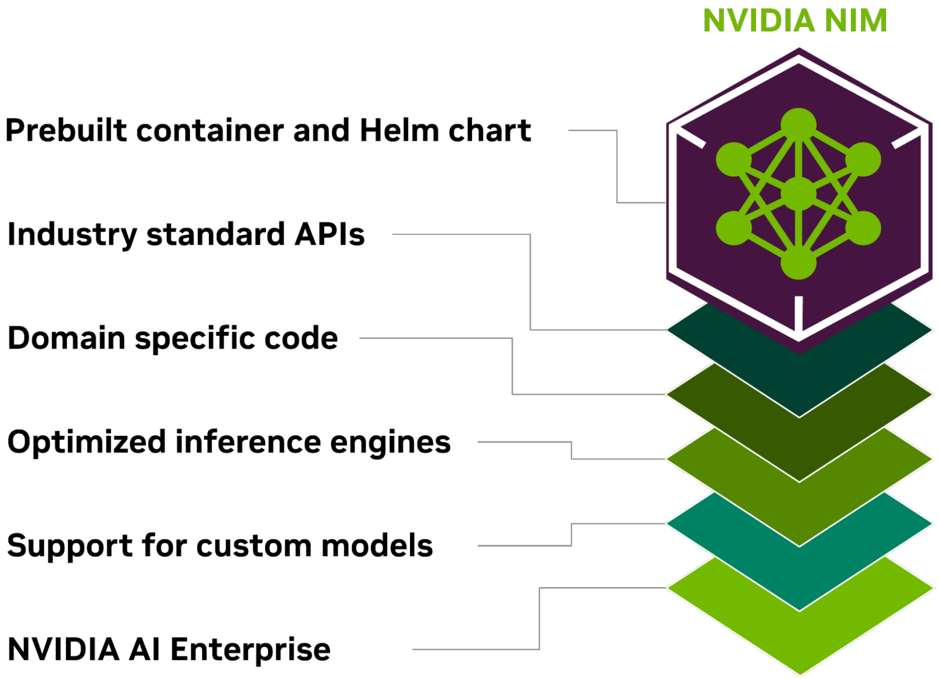
NVIDIA NIM
Prebuilt container and Helm chart
Industry standard APIs
Domain specific code
Optimized inference engines
Support for custom models
NVIDIA AI Enterprise

Figure 1. NVIDIA NIM is a containerized inference microservice including industry-standard APIs, domain-specific code, optimized inference engines, and enterprise runtime

Some of the core benefits of NIM include the following.

Deploy anywhere

NIM is built for portability and control, enabling model deployment across various infrastructures, from local workstations to cloud to on-premises data centers. This includes NVIDIA DGX, NVIDIA DGX Cloud, NVIDIA Certified Systems, NVIDIA RTX workstations and PCs.

Prebuilt containers and Helm charts packaged with optimized models, are rigorously validated and benchmarked across different NVIDIA hardware platforms, cloud service providers, and Kubernetes distributions. This enables support across all NVIDIA-powered environments and ensures that organizations can deploy their generative AI applications anywhere, maintaining full control over their applications and the data they process.

Develop with industry-standard APIs

Developers can access AI models through APIs that adhere to the industry standards for each domain, simplifying the development of AI applications. These APIs are compatible with standard deployment processes within the ecosystem, enabling developers to update their AI applications swiftly—often with as little as three lines of code. This seamless integration and ease of use facilitate rapid deployment and scaling of AI solutions within enterprise environments.

Leverage domain-specific models

NIM also addresses the need for domain-specific solutions and optimized performance through several key features. It packages domain-specific NVIDIA CUDA libraries and specialized code tailored to various domains such as language, speech, video processing, healthcare, and more. This approach makes sure that applications are accurate and relevant to their specific use case.

Run on optimized inference engines

NIM leverages optimized inference engines for each model and hardware setup, providing the best possible latency and throughput on accelerated infrastructure. This reduces the cost of running inference workloads as they scale and improves the end-user experience. In addition to supporting optimized community models, developers can achieve even more accuracy and performance by aligning and fine-tuning models with proprietary data sources that never leave the boundaries of their data center.

Support for enterprise-grade AI

Part of NVIDIA AI Enterprise, NIM is built with an enterprise-grade base container providing a solid foundation for enterprise AI software through feature branches, rigorous validation, enterprise support with service-level agreements, and regular security updates for CVE. The comprehensive support structure and optimization capabilities underscore the role of NIM as a pivotal tool in deploying efficient, scalable, and customized AI applications in production.

Accelerated AI models ready for deployment

With support for many AI models, such as community models, **NVIDIA AI Foundation models**, and custom AI models provided by NVIDIA partners, NIM supports AI use cases across multiple domains. This includes **large language models** (LLMs), vision language models (VLMs), and models for speech, images, video, 3D, drug discovery, medical imaging, and more.

Developers can test the latest generative AI models using NVIDIA-managed cloud APIs from the **NVIDIA API catalog**. Or they can self-host the models by downloading NIM and rapidly deploy with Kubernetes on major cloud providers or on-premises for production, cutting development time, complexity, and cost.

NIM microservices simplify the AI model deployment process by packaging algorithmic, system, and runtime optimizations and adding industry-standard APIs. This enables developers to integrate NIM into their existing applications and infrastructure without extensive customization or specialized expertise.

Using NIM, businesses can optimize their AI infrastructure for maximum efficiency and cost-effectiveness without worrying about AI model development complexities and containerization. On top of accelerated AI infrastructure, NIM helps with performance and scalability, while reducing hardware and operational costs.

For businesses looking to tailor models for enterprise applications, NVIDIA provides microservices for model customization across different domains. **NVIDIA NeMo** offers fine-tuning capabilities using proprietary data for LLMs, speech AI, and multimodal models. **NVIDIA BioNeMo** accelerates drug discovery with a growing collection of models for generative biology chemistry, and molecular prediction. NVIDIA Picasso enables faster creative workflows with Edify models. These models are trained on licensed libraries from visual content providers, enabling the deployment of customized generative AI models for visual content creation.

Getting started with NVIDIA NIM

Getting started with NVIDIA NIM is easy and straightforward. Within the **NVIDIA API catalog**, developers have access to a wide range of AI models that can be used to build and deploy their own AI applications.

Begin prototyping directly in the catalog using the graphical user interface or interact directly with the API for free. To deploy the microservice on your infrastructure, simply sign up for the **NVIDIA AI Enterprise 90-day evaluation license** and follow these steps.

1. Download the model you want to deploy from NVIDIA NGC. In this example, we'll download a version of the Llama-2 7B model built for a single A100 GPU.

```
ngc registry model download-version "ohlfw0olaadg/ea-participants/llama-2-
7b:LLAMA-2-7B-4K-FP16-1-A100.24.01"
```

If you have a different GPU, you can list the available versions of the model with ngc registry model list "ohlfw0olaadg/ea-participants/llama-2-7b:*"

2. Unpack the downloaded artifact into a model repository:

```
tar -xzf llama-2-7b_vLLAMA-2-7B-4K-FP16-1-A100.24.01/LLAMA-2-7B-4K-FP16-1-
A100.24.01.tar.gz
```

3. Launch the NIM container with your desired model:

```
docker run --gpus all --shm-size 1G -v $(pwd)/model-store:/model-store --
net=host nvcr.io/ohlfw0olaadg/ea-participants/nemollm-inference-ms:24.01
nemollm_inference_ms --model llama-2-7b --num_gpus=1
```

4. Once NIM is deployed, you can start making requests using a standard REST API:

```
import requests

endpoint = 'http://localhost:9999/v1/completions'

headers = {
    'accept': 'application/json',
    'Content-Type': 'application/json'
}

data = {
    'model': 'llama-2-7b',
    'prompt': "The capital of France is called",
    'max_tokens': 100,
    'temperature': 0.7,
    'n': 1,
    'stream': False,
    'stop': 'string',
    'frequency_penalty': 0.0
}

response = requests.post(endpoint, headers=headers, json=data)
print(response.json())
```

NVIDIA NIM is a powerful tool to help organizations accelerate their journey to production AI. Get started on your AI journey today.

Related resources

- **DLI course:** Deploying a Model for Inference at Production Scale
- **GTC session:** Accelerating Enterprise: Tools and Techniques for Next-Generation AI Deployment
- **GTC session:** LLM Inference Sizing: Benchmarking End-to-End Inference Systems
- **SDK:** Triton Management Service
- **SDK:** Triton Inference Server
- **SDK:** NeMo Inferencing Microservice