Solution Overview

NVIDIA NIM: The Fastest Path to Enterprise Generative AI

The Challenges on the Path to Production

In the rapidly evolving landscape of generative AI, enterprises are looking to leverage this cutting-edge technology to gain a competitive advantage and fast-track innovation.

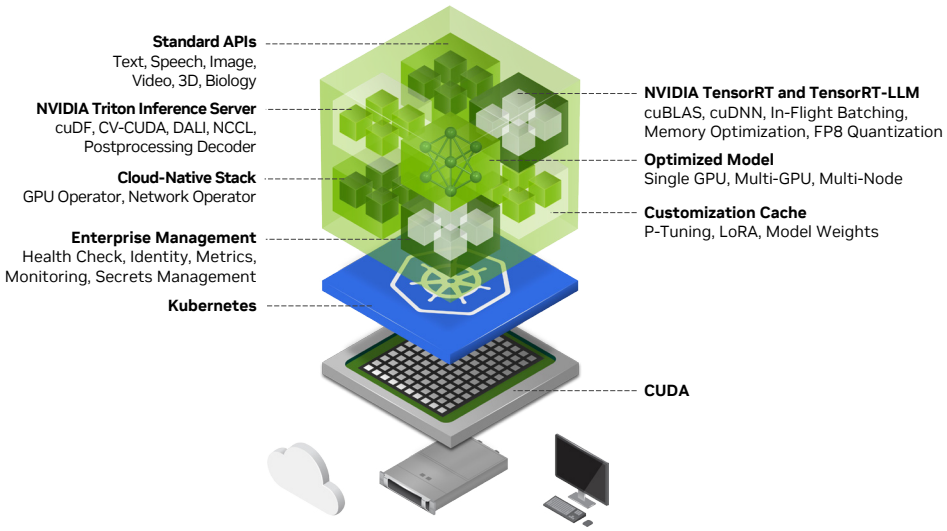
However, there are some significant challenges to integrating generative AI into existing business processes. Enterprises are concerned about protecting their intellectual property (IP), maintaining brand integrity, ensuring client confidentiality, and meeting regulatory standards.

Five Minutes to Inference

NVIDIA NIM helps overcome these challenges, making it easy for IT and DevOps teams to self-host AI models in their own managed environments, while providing developers with industry-standard APIs for building powerful copilots, chatbots, and AI assistants that can transform their business.

Part of NVIDIA AI Enterprise, NVIDIA NIM is a set of easy-to-use microservices designed to accelerate deployment of generative AI. These prebuilt microservices support a broad spectrum of AI models—from open-source community models to NVIDIA AI Foundation and custom models. NIM microservices can be deployed with a single command and quickly integrated into applications with just a few lines of code.

NVIDIA NIM



The Timeline of Generative AI

2022: Explosion

ChatGPT is announced in late 2022 and gains over 100 million users in just two months. Users of all levels experienced AI and its benefits firsthand.

2023: Experimentation

Enterprise application developers kick off proofs of concept (POCs) for generative AI applications with API services and open models, including Llama 2, Mistral, NVIDIA, and others.

Today: Production

Organizations have set aside budget and are ramping up efforts to build accelerated infrastructure to support generative AI in production.

Benefits

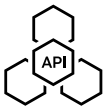
- Deploy anywhere with security and control.
- Empower developers with industry-standard APIs and tools.
- Lower costs and scale performance on accelerated infrastructure.

Built on robust foundations, including inference engines like NVIDIA Triton™ Inference Server, TensorRT™, TensorRT-LLM, and PyTorch, NIM is engineered to facilitate seamless AI inferencing at scale, ensuring that AI applications can be deployed anywhere with confidence. Whether on premises or in the cloud, NIM is the fastest way to achieve accelerated generative AI inference at scale.

The NVIDIA API Catalog

The latest community-built AI models—optimized and accelerated by NVIDIA—are available at **ai.nvidia.com**. With API access to these models, developers can experiment, prototype, and ultimately deploy anywhere, whether in the cloud or on premises, with NVIDIA NIM.

Experience Models



Prototype Wth APIs

Deploy With NIMs

Ready to Get Started?

To learn more about NVIDIA NIM, visit: **ai.nvidia.com**

3159981. MAY24