

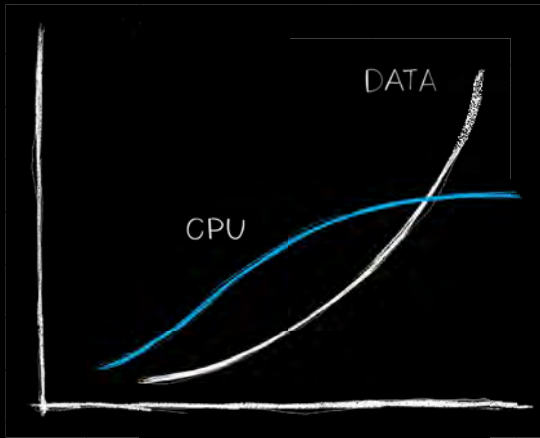


# AI Factories and Community Models to Transform Financial Services

**Dr. Jochen Papenbrock**

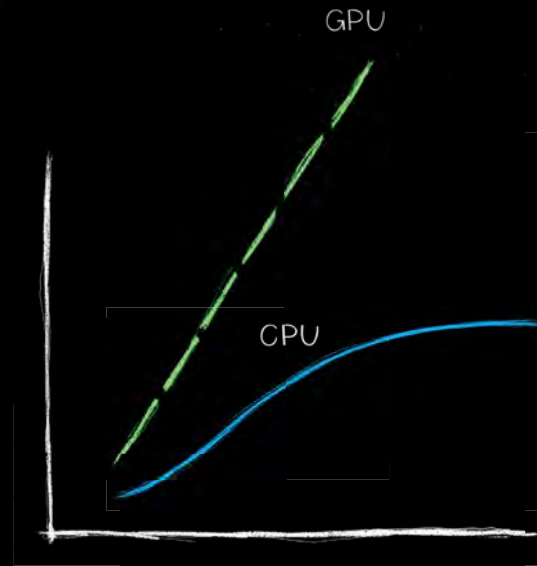
Head of Financial Technology EMEA / Lead Developer Relations Manager Banking Global

[jpapenbrock@nvidia.com](mailto:jpapenbrock@nvidia.com)



CPU SCALING SLOWS

... AND  
COMPUTE DEMAND GROWS  
EXPONENTIALLY



GPU-ACCELERATED COMPUTING

2006  
CUDA

ACCELERATE EVERY APPLICATION

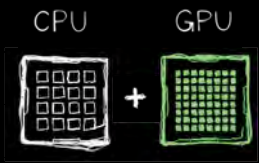
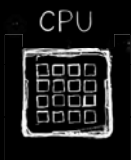
CPU



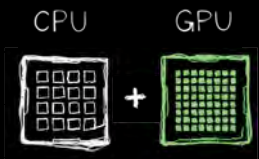
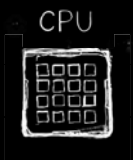
1t

100t





= ~100X SPEED-UP  
~3X POWER  
~1.5X COST



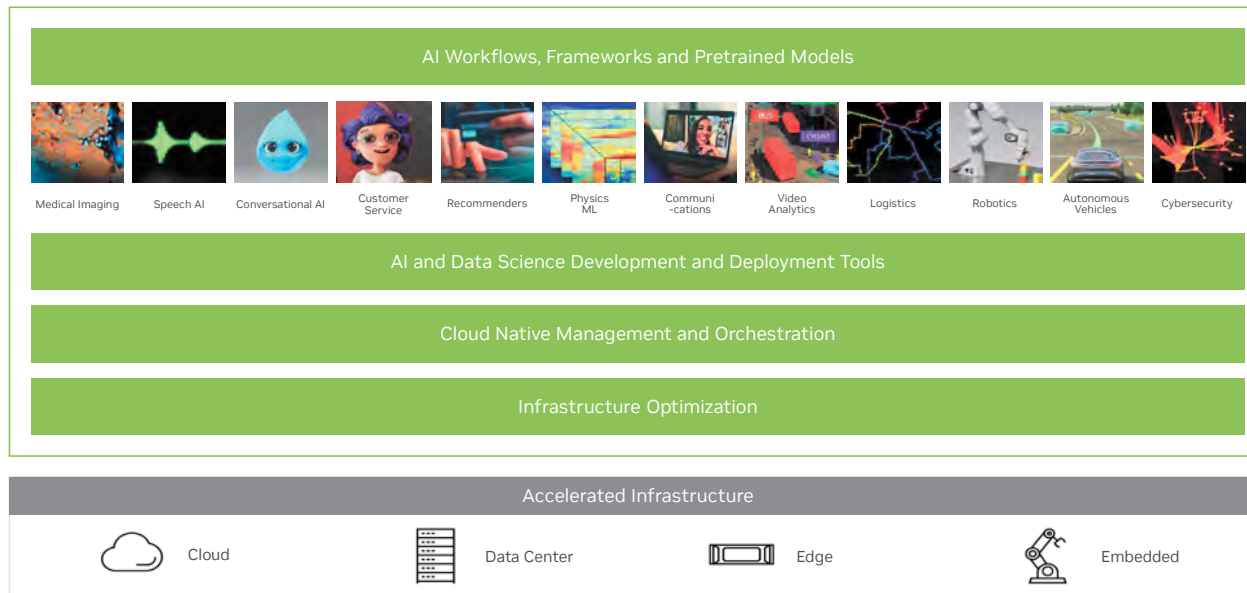
= ~100X SPEED-UP  
~3X POWER  
~1.5X COST

60X PERF / \$ OR 98% SAVINGS

30X PERF / W OR 97% SAVINGS

# AI is a full stack problem

NVIDIA AI Enterprise – the 'Operating System' for AI



- ✓ Cloud Native, Hybrid Optimized
- ✓ Deploy anywhere: on-prem and in the cloud
- ✓ Reduce OSS development complexity
- ✓ Secure and Scalable
- ✓ Certifications with broad partner ecosystem
- ✓ Improved AI model accuracy
- ✓ Standard Support 9 x 5, Premium 24x7





## References in for building a hybrid AI platform/factory



**BNY Mellon**  
646,702 followers  
2W • Edited • 🌐

ANNOUNCEMENT **BNY Mellon** becomes the first global bank to deploy an AI supercomputer powered by **NVIDIA**.

With more than 600 opportunities in #AI identified and dozens already in development, this collaboration will streamline and accelerate innovation within our business and across the global financial system.

"Key to our technology strategy is empowering our clients through scalable, trusted platforms and solutions," said #BNYMellon Chief Information Officer **Bridget Engle**. "By deploying NVIDIA's AI supercomputer, we can accelerate our processing capacity to innovate and launch AI enabled capabilities that help us manage, move and keep our clients' assets safe."

#Nvidia #supercomputing #artificialIntelligence #cio

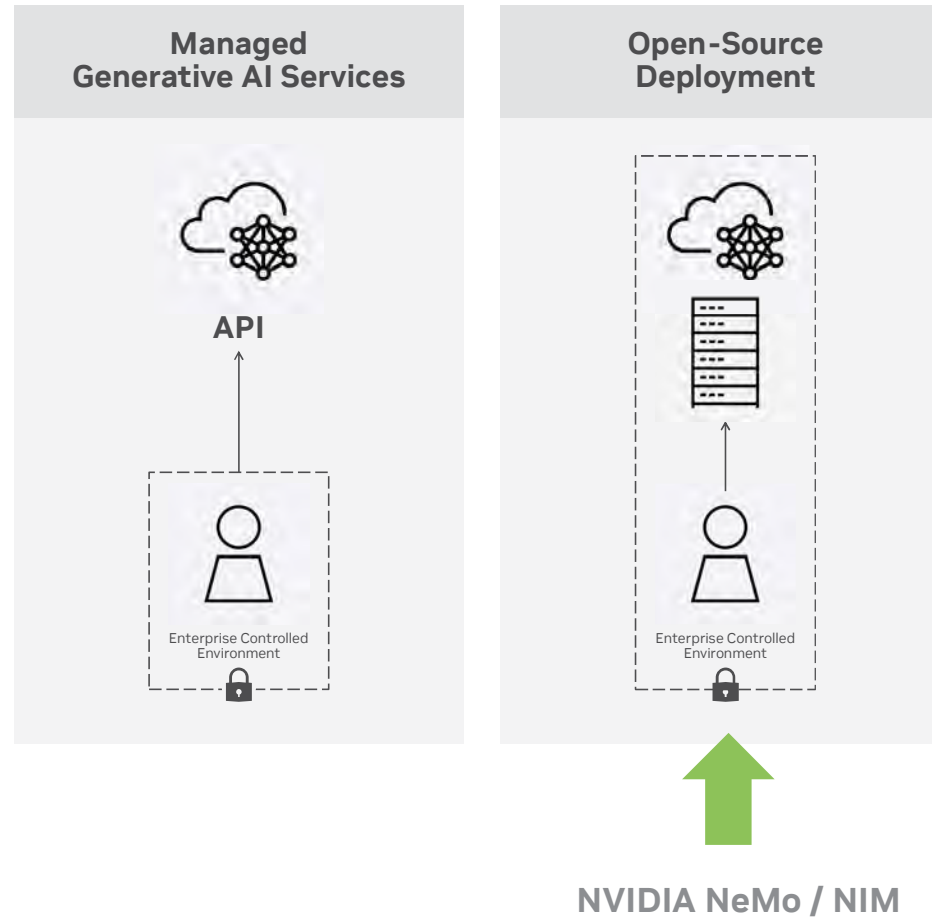


The graphic features the NVIDIA logo on the left and the text "BNY MELLON DEPLOYS NVIDIA AI SUPERCOMPUTER" in large, bold, white and blue letters on the right. A small BNY Mellon logo is in the bottom right corner of the graphic.

BNY Mellon, First Global Bank to Deploy AI Supercomputer Powered by NVIDIA DGX SuperPOD With DGX H100

# Enterprises Face Challenges Experimenting With Generative AI

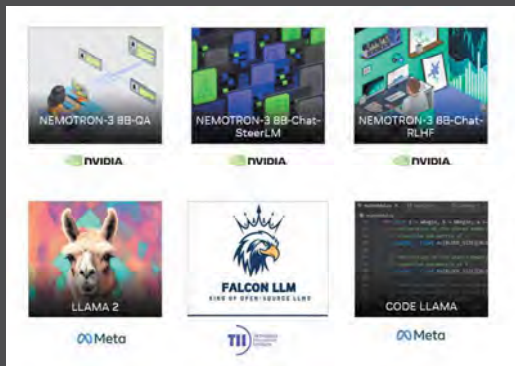
Organizations must choose between ease of use and control



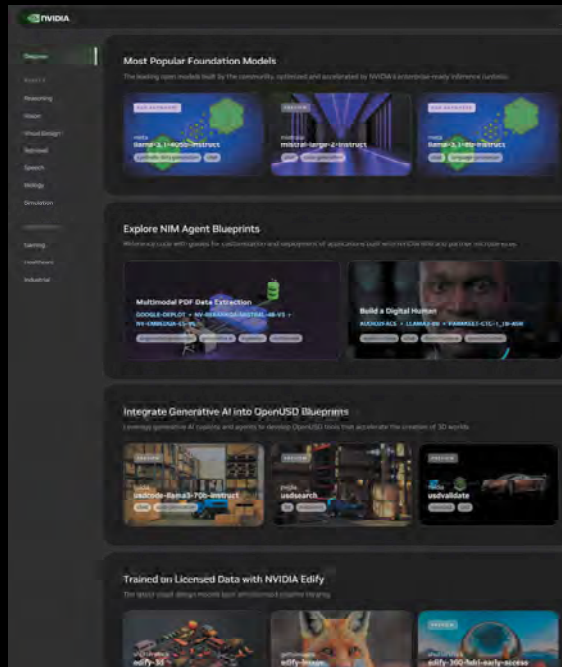


# NIM: NVIDIA Inference Microservice

Experience, prototype, and deploy the latest AI models at [ai.nvidia.com](https://ai.nvidia.com)



- State-of-the-art community, commercial and NVIDIA-built models
- Performance-optimized for GPU-accelerated stack



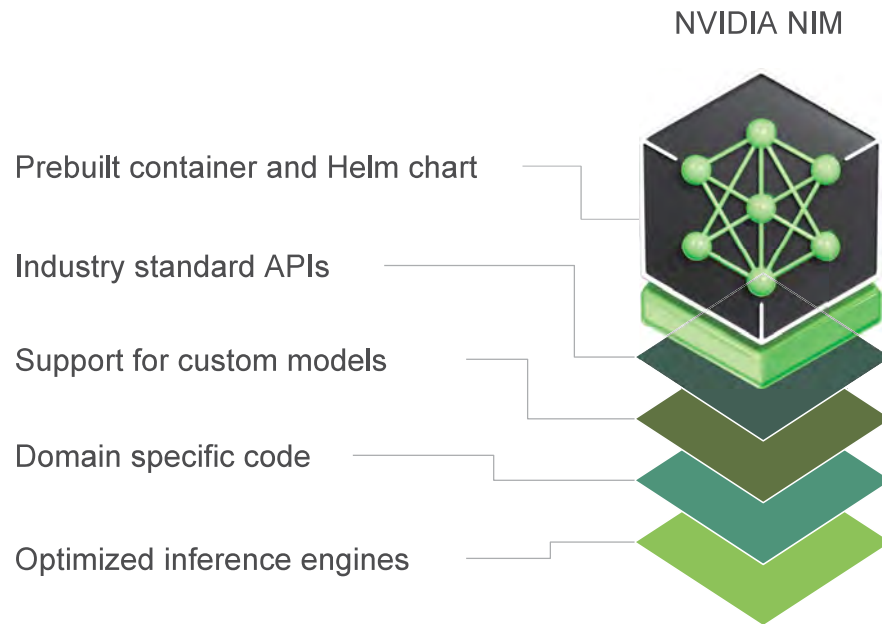
Prototype with NVIDIA-hosted API endpoints on [ai.nvidia.com](https://ai.nvidia.com)



Models “to-go”:  
deploy in production with NIMs

# NVIDIA NIM Optimized Inference Microservices

Accelerated runtime for generative AI



**Deploy anywhere and maintain control** of generative AI applications and data

**Simplified development** of AI application that can run in enterprise environments

**Day 0 support** for all generative AI models providing choice across the ecosystem

**Improved TCO** with best latency and throughput running on accelerated infrastructure

**Best accuracy** for enterprise by enabling tuning with proprietary data sources

**Enterprise software** with feature branches, validation and support



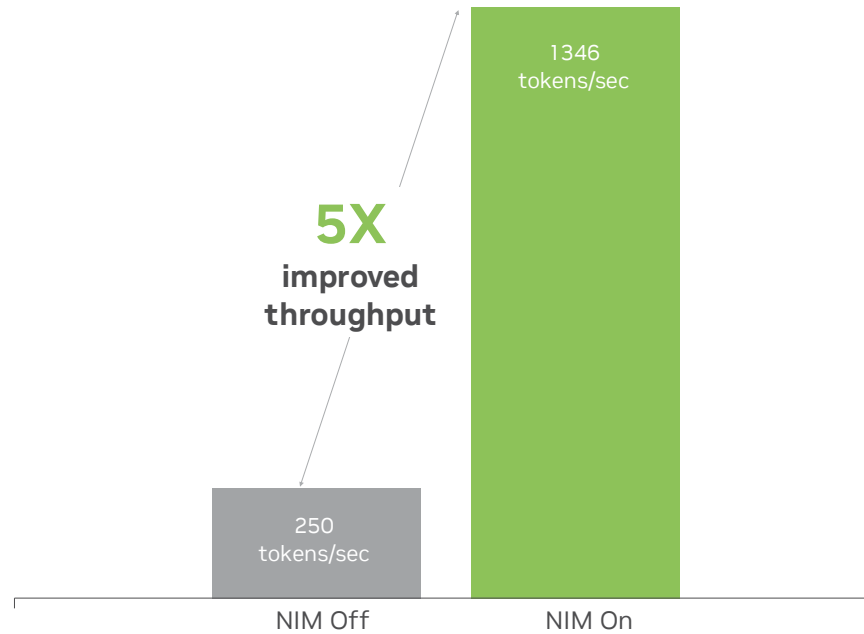
DGX &  
DGX Cloud



# Improved Efficiency Out of the Box

State-of-the-art Throughput Reduces Overall Cost of Solution

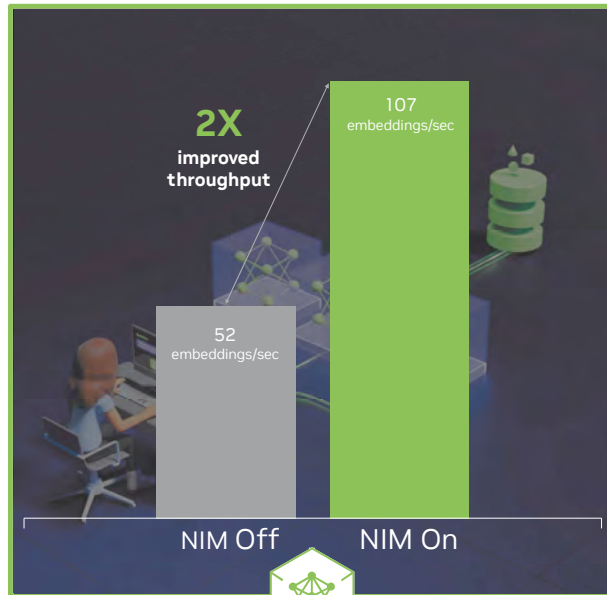
## Llama 3 70B NIM Delivers 5X Higher Throughput



Llama 3-70b-instruct, input token length: 7,000, output token length: 1,000. Concurrent client requests: 100, 4xH100 SXM NVLink. NIM Off: FP16, TTFT: ~120s, ITL: ~180ms. NIM On: FP8, TTFT: ~4.5s, ITL: ~70ms.

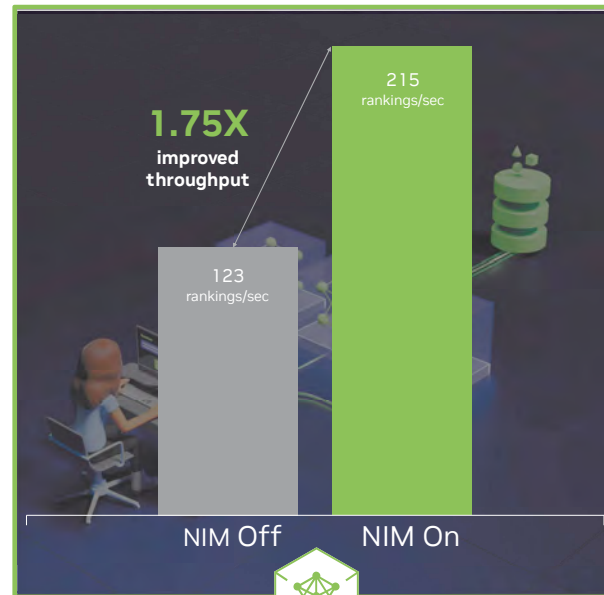
# NVIDIA NeMo Retriever NIMs

Available for download today at [ai.nvidia.com](https://ai.nvidia.com)



**NV-EmbedQA-Mistral7B-v2**  
Multilingual text embedding model

NV-EmbedQA-Mistral7B-v2, 1xH100 SXM; passage token length: 512, batch size: 64, concurrent client requests: 3; NIM Off: FP16, P90 latency: ~3.8s; NIM On: FP8, P90 latency: ~1.8s.



**NV-RerankQA-Mistral4B-v3**  
Text reranking for high accuracy question answering

NV-RerankQA-Mistral4B-v3, 1xH100 SXM; query token length: 20, passage token length: 512, batch size: 40, concurrent client requests: 3; NIM Off: FP16, P90 latency: ~1s; NIM On: FP8, P90 latency: ~0.56s

# Building Generative AI Applications for the Enterprise

Build, customize, and deploy generative AI models with NVIDIA NeMo.

