



# Inference at Scale

Andrea Pilzer, Solutions Architect | NVAITC @ Helsinki  
University/16.05.025



# Agenda

- Introduction & Motivation
- NVIDIA Inference Microservices (NIM)
- NVIDIA Dynamo

# The Generative AI Journey



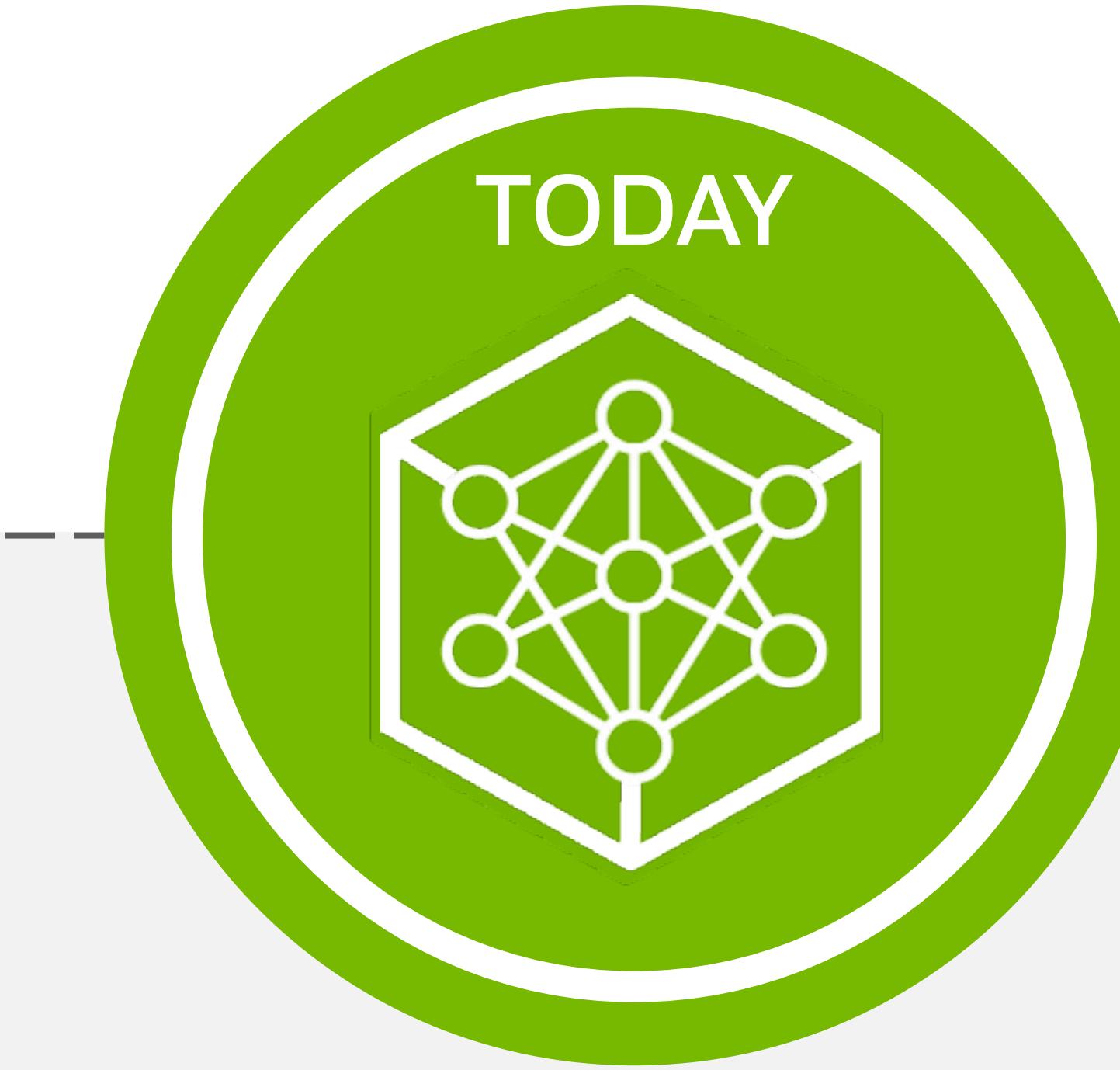
## Explosion

ChatGPT gets announced late in 2022, gaining over 100 million users in just two months. Users of all levels can experience AI and feel the benefits firsthand.



## Experimentation

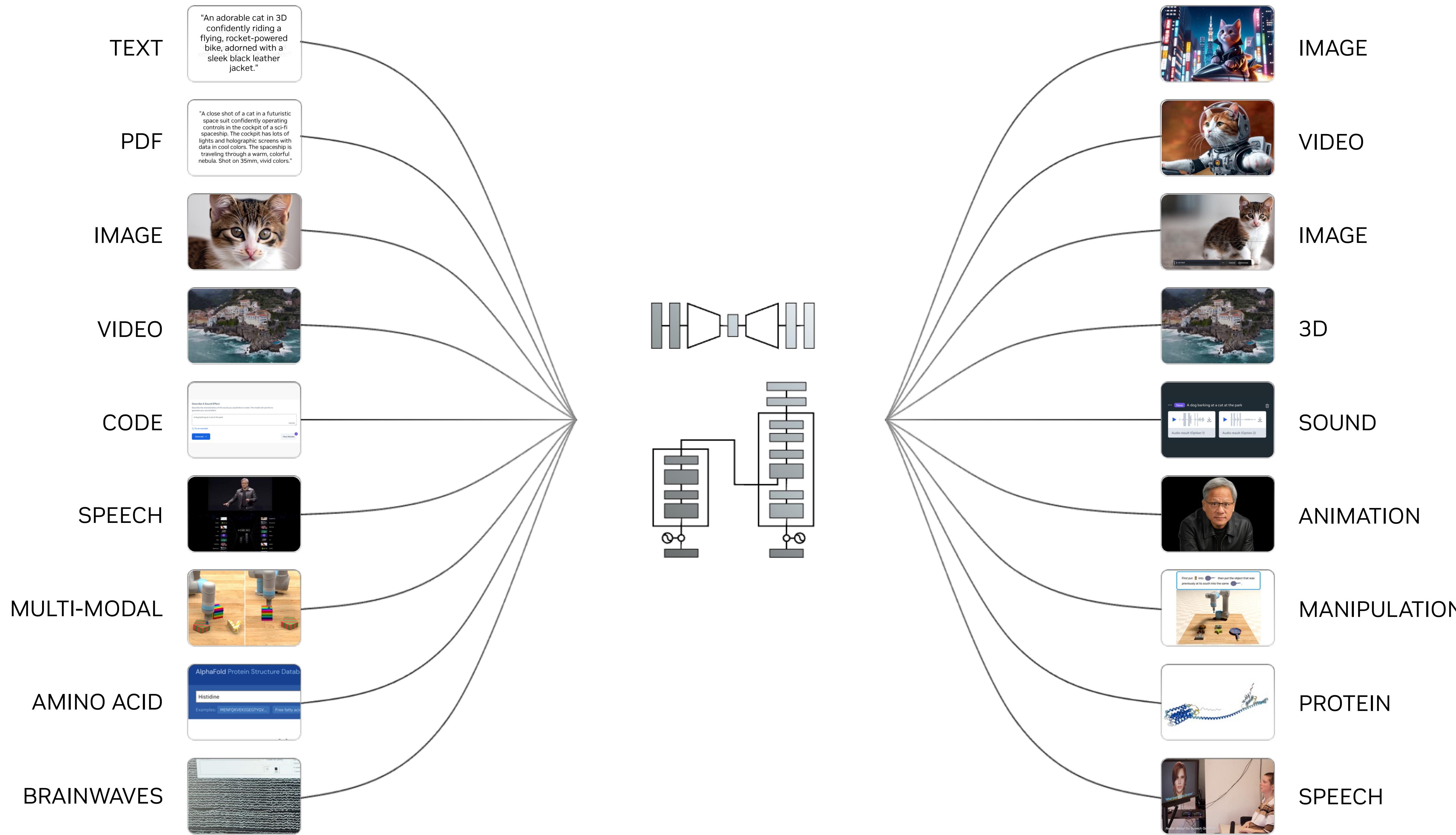
Enterprise application developers kick off POCs for generative AI applications with API services and open models including Llama 2, Mistral, NVIDIA, and others.



## Production

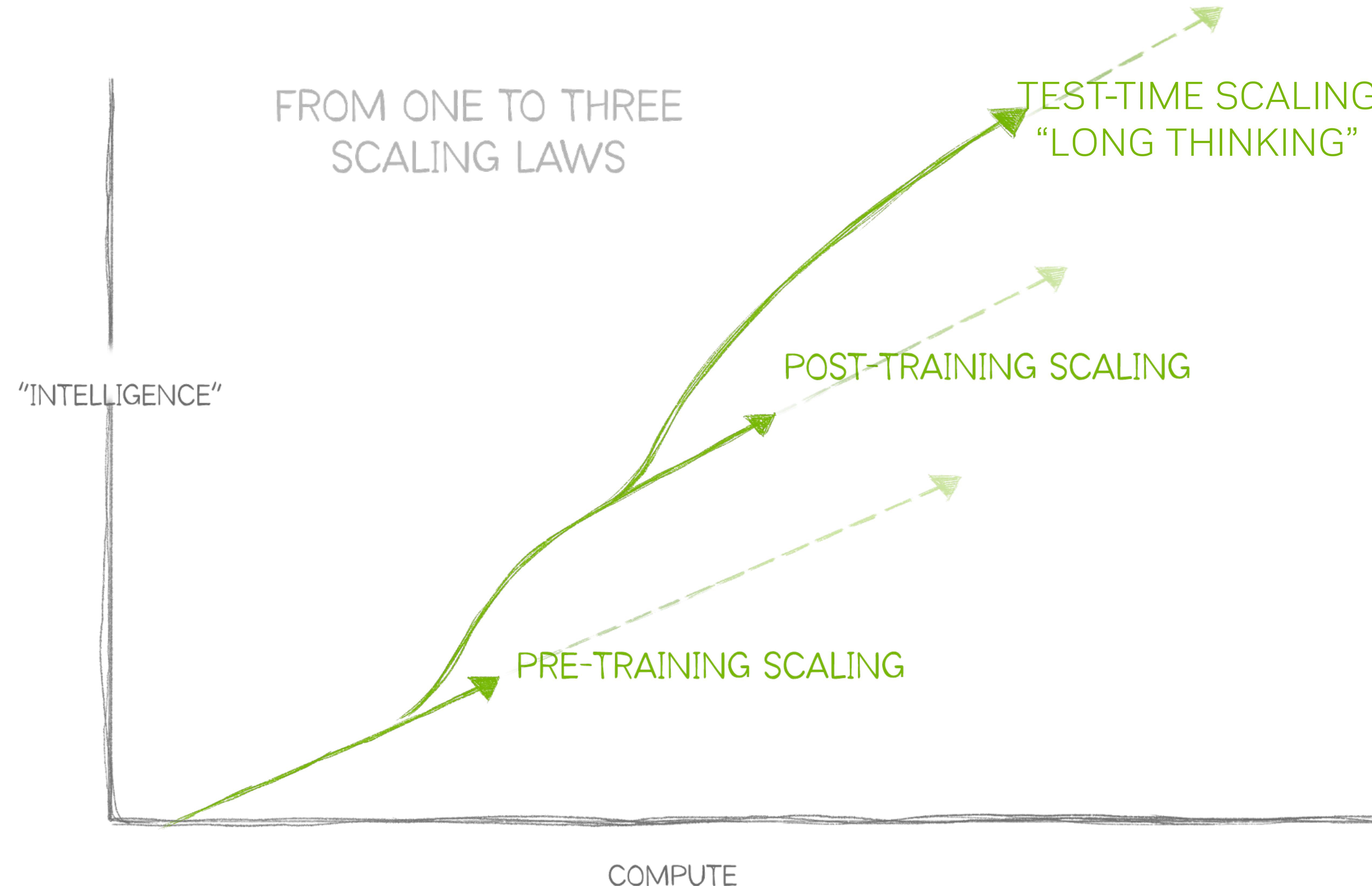
Organizations have set aside budget and are ramping up efforts to build accelerated infrastructure to support generative AI in production.

# Generative AI Can Learn and Understand Everything



# AI Scaling Laws Drive Exponential Demand for Compute

New “long thinking” supercharges inference scaling



# Inference Compute Requirements Scaling Exponentially

Fueled by reasoning models and AI agents



Hundreds of billions of  
parameters



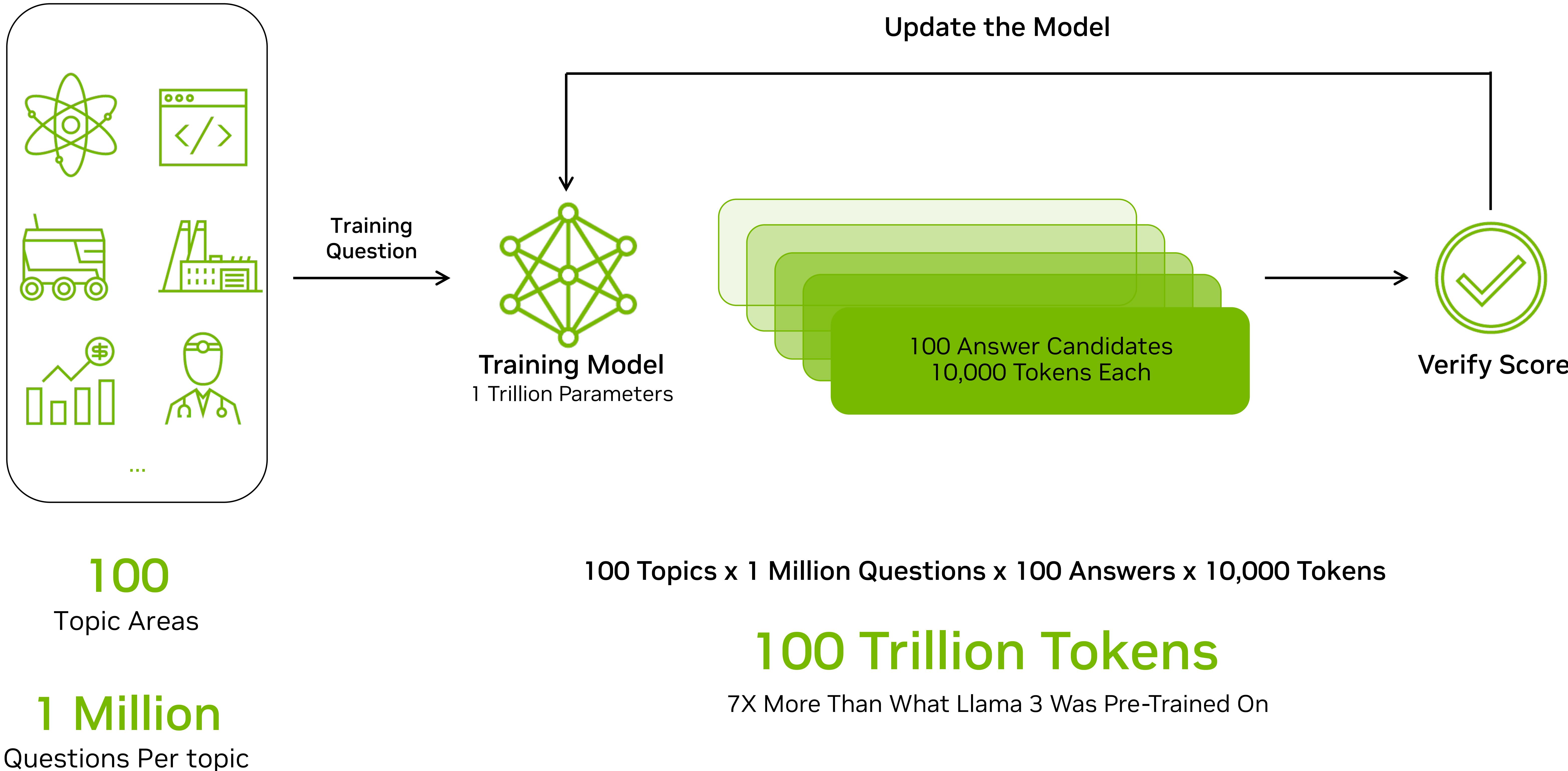
100x more thinking  
tokens



Millions of input  
tokens

# Post-Training Scaling in an Era of Reasoning

Teaching models to think requires a lot of compute



# Reasoning Drives 100X More Tokens

Cost increasing with compute – requires full stack innovation to lower cost/token

## KNOWLEDGE

Pre-trained models simulate internet knowledge  
100 tokens | \$0.50 / M Token

### What is NVIDIA?

NVIDIA is a multinational technology company that specializes in designing and manufacturing graphics processing units (GPUs), high-performance computing hardware, and artificial intelligence (AI) technologies. The company was founded in 1993 by Jensen Huang, Chris Malachowsky, and Curtis Priem.

NVIDIA is headquartered in Santa Clara, California, and has become one of the leading companies in the field of computer graphics, gaming, and AI.

## REASONING

Reasoning models think through problems  
~10K tokens → 100X cost

Okay, so I need to figure out seating for a family dinner with 8 people. Let's list out the requirements first to make sure I don't miss anything. Firstly, the family has 8 members. Let me confirm who they are. The user mentions: in-laws (which I assume are the wife's parents), parents (user's own parents), sister, little brother, wife, and the user themselves. That adds up to 8: user, wife, user's mom, user's dad, wife's mom, wife's dad, sister, and little brother. Wait, that's 9 people, maybe I missed someone? Let's look again. User says my family of 8 includes user, the spouse (user's parents, mom and dad), their in-laws (wife's parents, mom and dad), sister, and little brother. That's 8: user, wife, user's mom, user's dad, wife's mom, wife's dad, sister, little brother. Yes, that's 8 people. Now, let's list out the constraints:

I need to figure out where to seat my family of 8 for dinner at a round table. My in-laws don't get along with my parents. My sister needs to sit next to my little brother to help him eat. My wife really doesn't want to sit next to my mom. Where should each person be seated?

Additional considerations: Typically, seating arrangements at a rectangular table have two ends and sides. But sometimes it's a round table. The problem doesn't specify, so maybe I should assume a round table because it's common for family dinners and allows for a circular arrangement. If it's a round table, seating is in a circle, so everyone has two neighbors. If it's rectangular, the ends have only one neighbor, but the problem might be easier with a round table. Let me proceed with a round table unless otherwise specified.

So, the goal is to seat 8 people harmoniously at a round table while respecting all constraints. Here's the optimal arrangement:

You Wife Dad Mom Sister Little Brother  
Mother-in-law Father-in-law

### Key Constraints Met:

- . In-laws vs. Parents: Separated by your sister, brother, and you.
- . Sister & Brother: Adjacent for assistance.
- . Wife & Mom: Not next to each other (separated by you and dad).

## ACCELERATING AI FACTORY VALUE

>\$1/M Token | >1T Param | >300 TPS

Need for increasing tokens

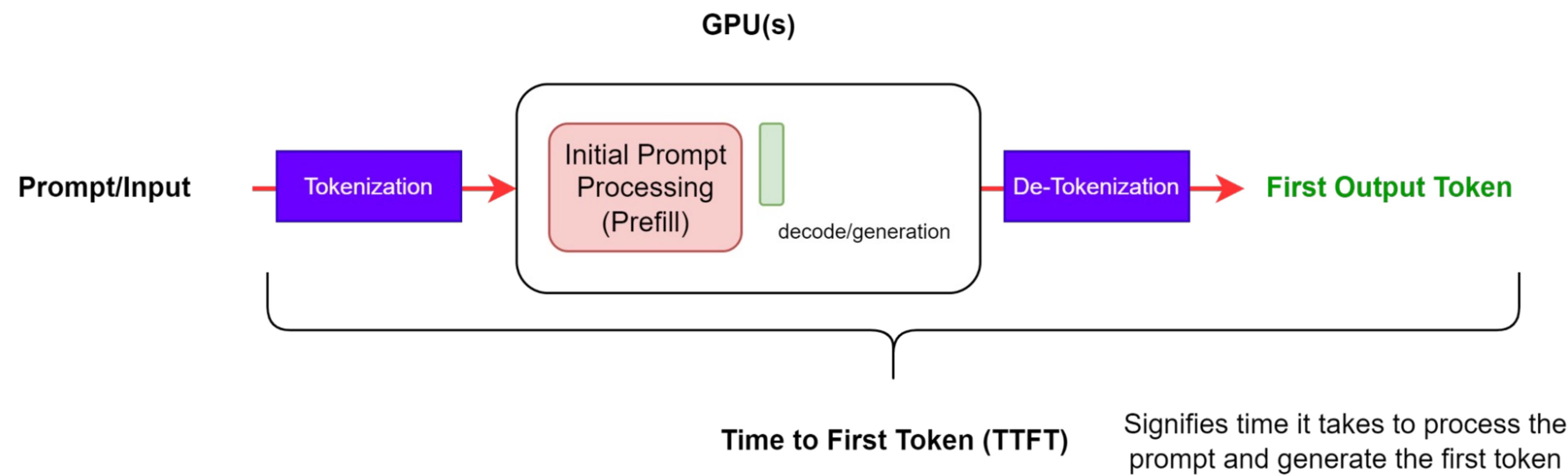
Increasing revenue customer experience

Decrease cost

# LLM Inference Terminology

<https://docs.nvidia.com/nim/benchmarking/llm/latest/overview.html>

- **Prompt:** User provides a query
- **Queuing:** Query joins the queue for processing
- **Prefill:** The LLM model processes the prompt
- **Generation:** The LLM model outputs a response, one token at a time

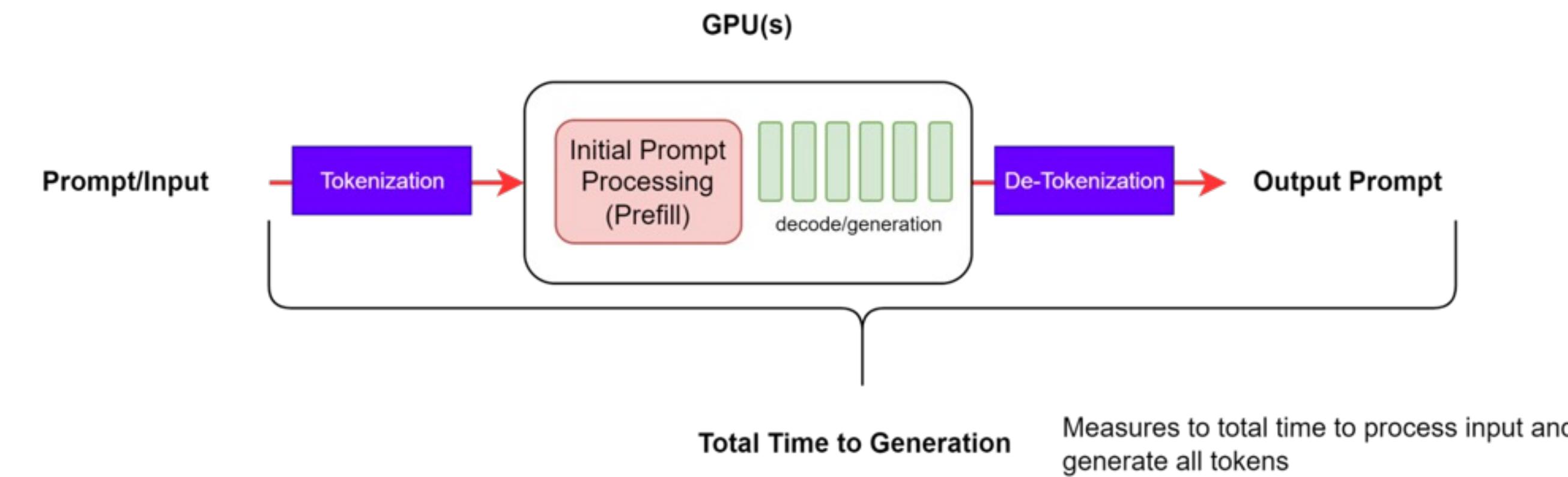


- **Time to First Token (TTFT):** Includes request queuing, prefill time, network latency.
- **Token:** Smallest lingual entity that an LLM uses to represent natural language. A collection of them is the Vocabulary
- **Sequence Length:** Length of sequence of data, it includes prompts, chat history, documents and more. We can define Input Seq. Length (ISL), and Output Seq. Length (OSL).
- **Context Length:** Maximum number of tokens used by the model at inference

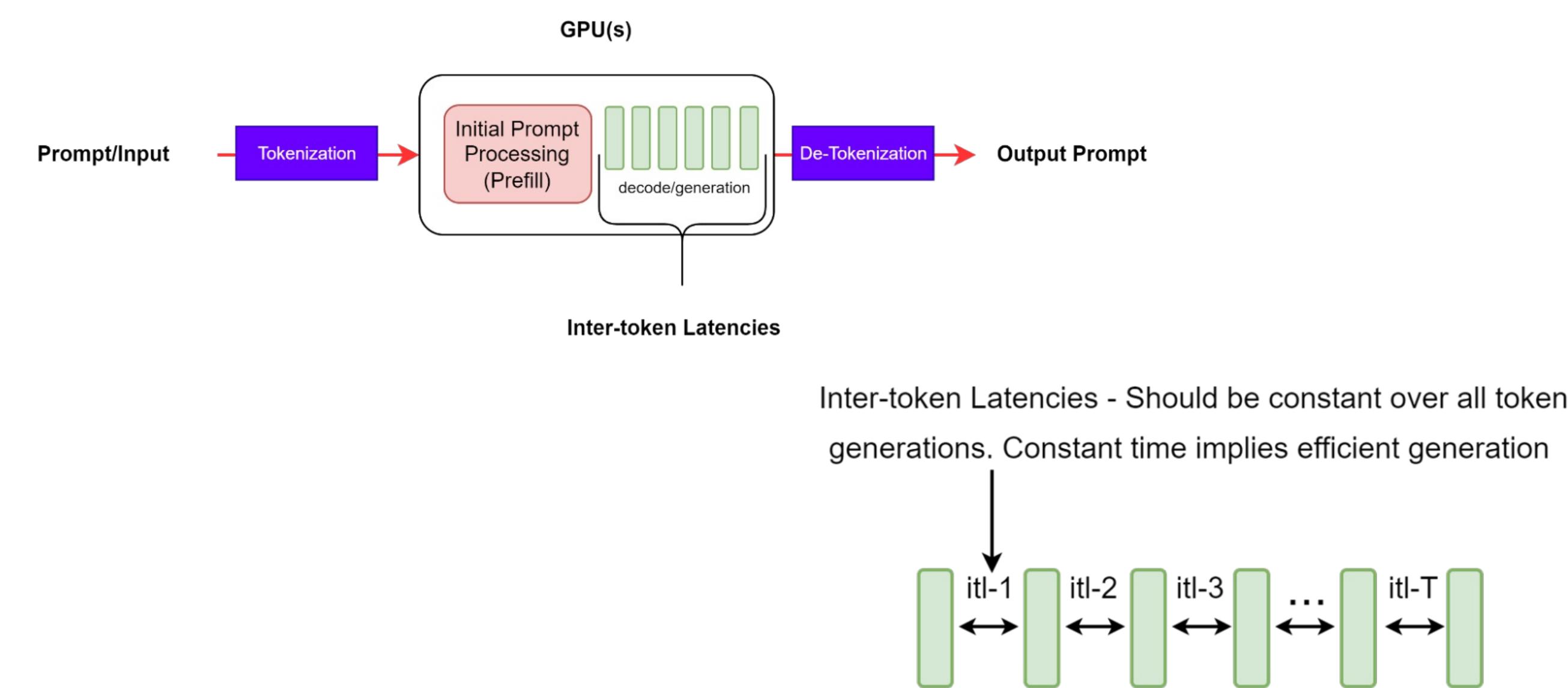
# Latency

## End-2-End, Intertoken

- **End-to-End Request latency:** Time it takes from submitting a query to receiving the full response



- **Streaming Mode:** Optionally we can stream to the users chunks of tokens decoded incrementally (e.g. chatbots)
- **Intertoken Latency:** average time between the generation of consecutive tokens in a sequence



- **Tokens per second:** Total throughput of a system in output tokens per second
- **Requests per second:** Average number of requests that can be completed by the system per second

# Use Cases

## And their Impact on Performance

- **Translation:** Includes translation between languages and code and is characterized by having similar ISL and OSL of roughly 500~2000 tokens each.
- **Generation:** Includes generation of code, story, and email content and generic content through search. This is characterized by having an OSL of  $O(1,000)$  tokens, much longer than an ISL of  $O(100)$  tokens.
- **Summarization:** Includes retrieval, chain-of-thought prompting, and multturn conversations. This is characterized by having an ISL of  $O(1000)$  tokens, much longer than an OSL of  $O(100)$  tokens.
- **Reasoning:** Recent reasoning models generate a large number of output tokens in an explicit chain-of-thought, self-reflection-and-verification reasoning approach to solve complex problems, like coding, maths or puzzles. This is characterized by short ISL of  $O(100)$  tokens and a large OSL of  $O(1000-10000)$  tokens.

# Where to Optimize?

LLM inference can be done in a smart way

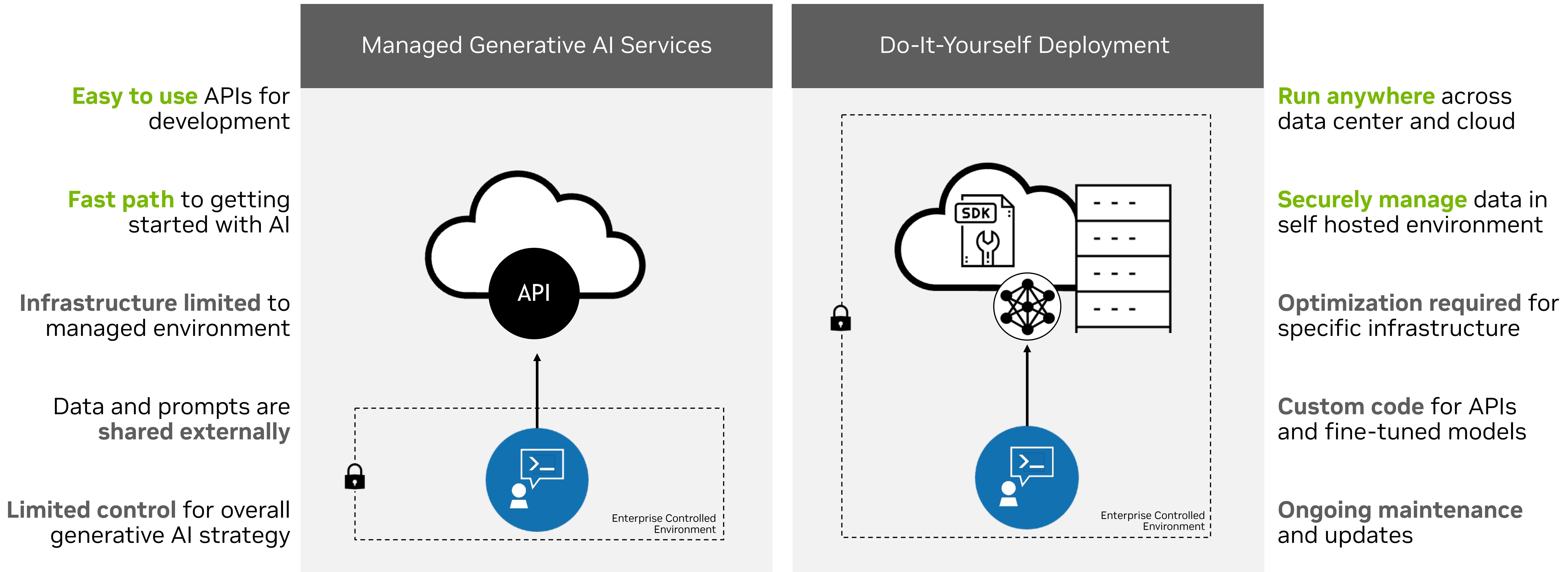
We defined lots of concepts that can help us through the rest of the presentation.

Now we can identify some improvements (assuming we have a model, and we don't change it):

- Optimize the SW stack, use performant engines to power the LLM, API usage -> **NIMs**
- Work around bottlenecks of the LLM inference pipeline, prefill vs decoding, network, KV cache -> **Dynamo**

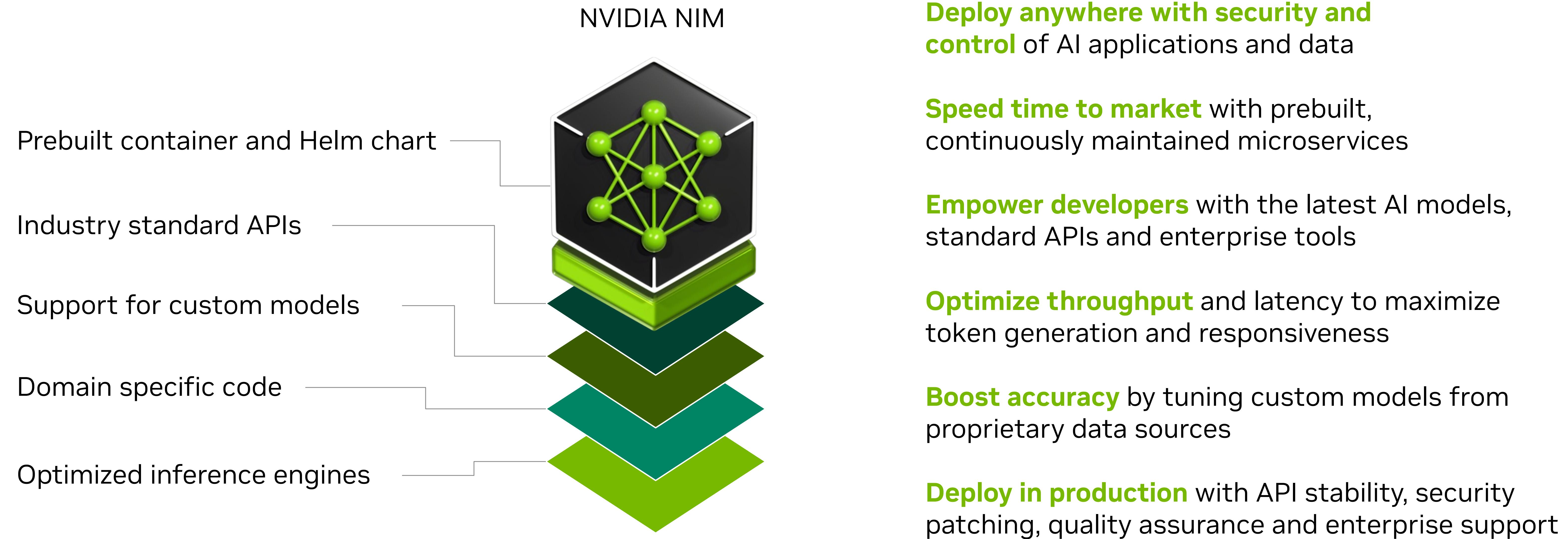
# **NVIDIA Inference Microservices (NIM)**

# Enterprises Face Challenges Experimenting with Generative AI



# NVIDIA NIM Optimized Inference Microservices

Accelerated runtime for generative AI

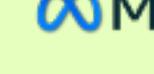
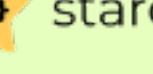


DGX &  
DGX Cloud



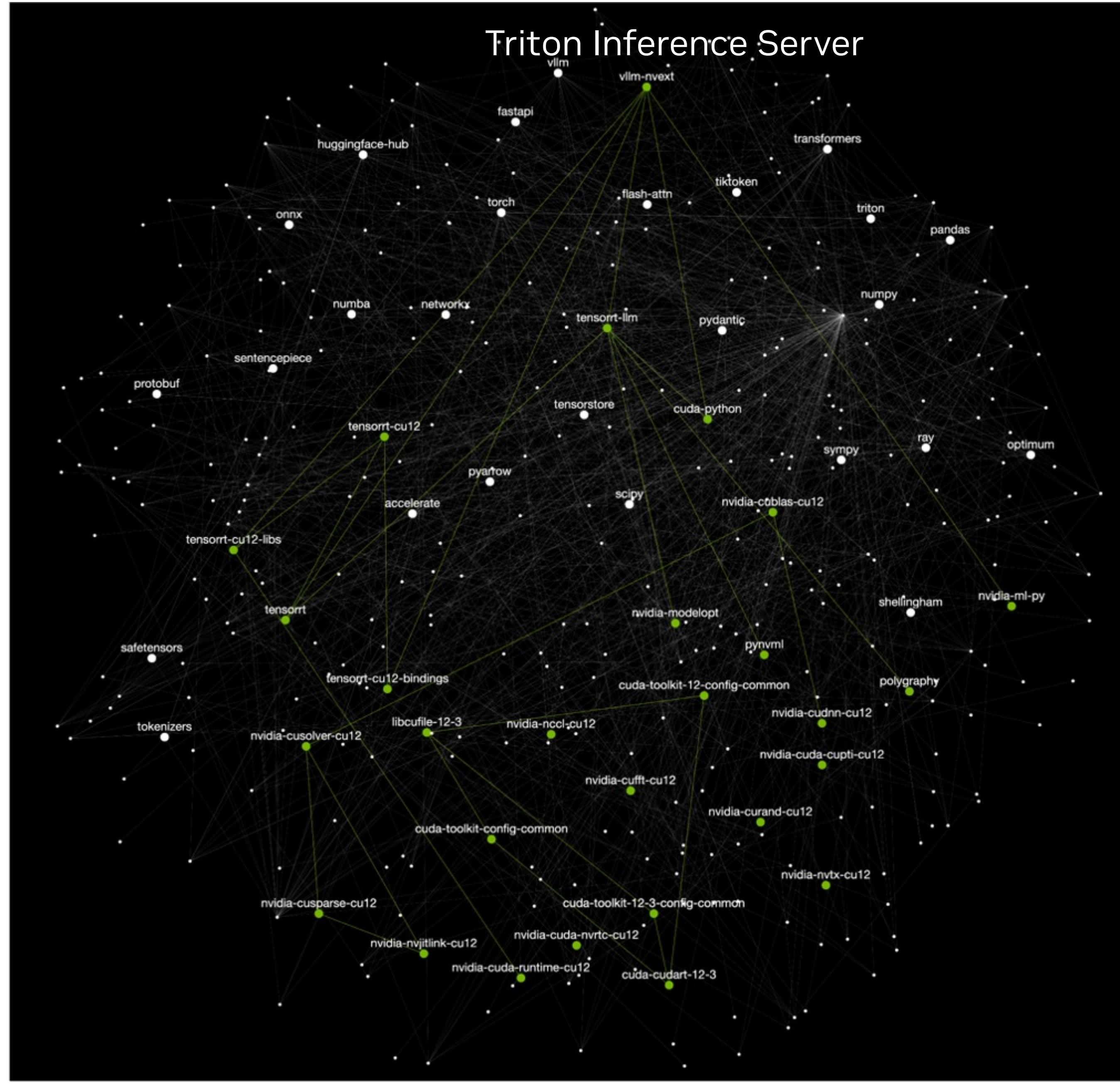
# NVIDIA NIM is the Fastest Path to AI Inference

Reduces engineering resources required to deploy optimized, accelerated models

	NVIDIA NIM	Do It Yourself
Deployment Time	5 minutes	1 week +
API Standardization	Industry standard protocol OpenAI for LLMs, Google Translate for Speech	Implement the API layer for each domain and model family according to industry standard specifications
Optimized Engines	Pre-built engines for NVIDIA and community models  Mistral AI  Meta  starcoder  Nemotron	Build your own engine and manually customize for workload and hardware specific requirements
Pre and Post Processing Pipelines	Pre-built with optimized pipeline engines to handle pre/post processing (tokenization)	Implement custom logic
Model Server Deployment	Automated	Manual setup and configuration
Customization	LoRA is supported, more planned	Create custom logic
Container Validation	Extensive workload specific QA support matrix validation	No validation
Enterprise Support	Delivered with NVIDIA AI Enterprise Security and CVE scanning/patching and tech support	Self supported

# AI Software Security and Stability

# Growing Complexity Increases Challenges of Maintenance



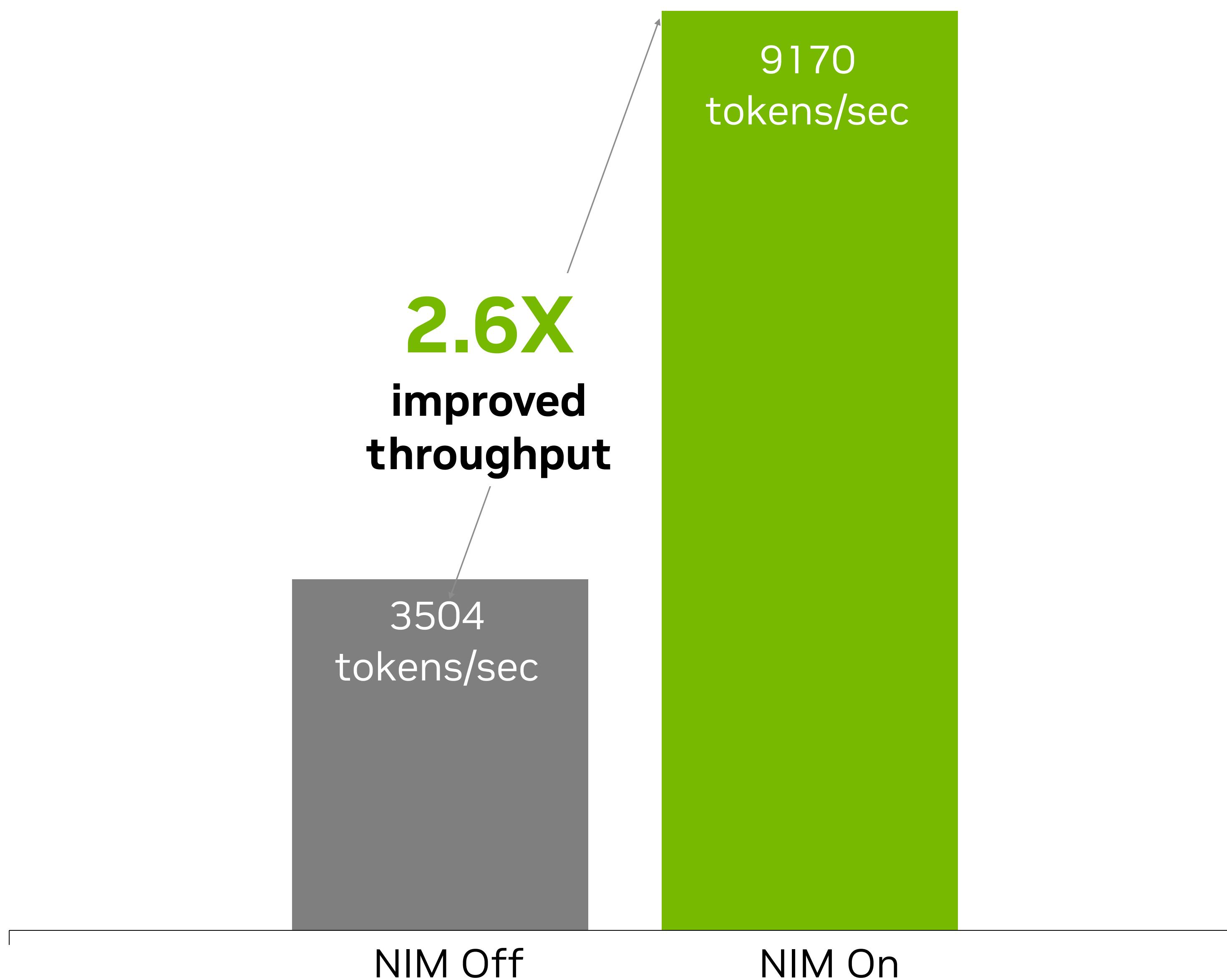
- . AI software stack is complex
    - . Consists of hundreds of dependencies
    - . Heavily leverages third-party OSS libraries
  - . Software complexity leads to
    - . Increased chance of vulnerability
    - . Challenges to maintaining compatibility
  - . Maintaining an AI software stack is resource-intensive:
    - . Assessment and exploration are needed after a CVE scan for further actions and measures
    - . A software package update requires a subsequent change in a dependent API, which can break the stack

# ○ **378** 3rd party software

## ● 31 NVIDIA software

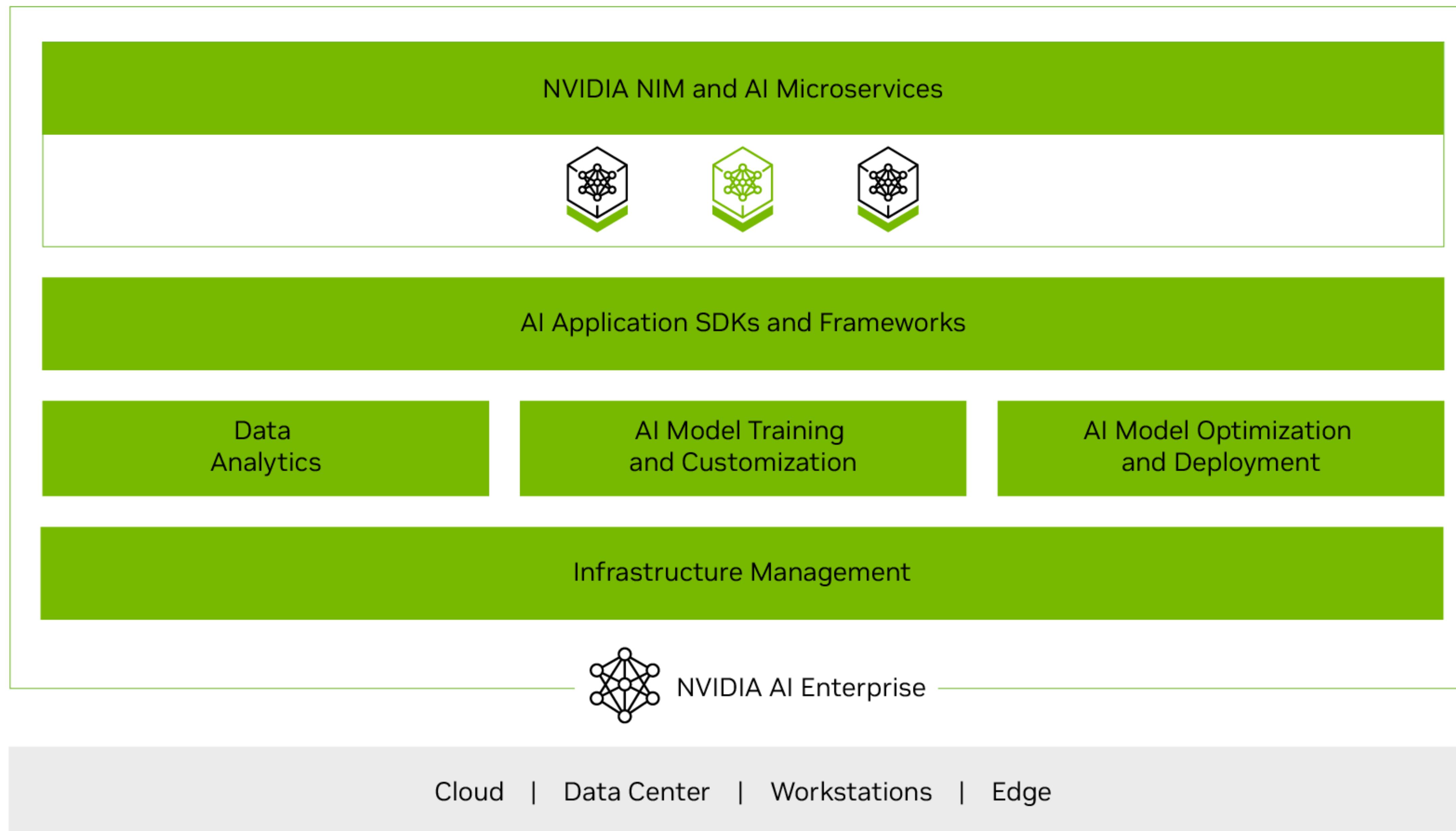
# — Dependency

# Improved Efficiency Out of the Box



# NVIDIA AI Enterprise

Cloud-native software platform for developing and deploying production-grade AI solutions



- Optimized NIM inference microservices enhance model performance and speed time to deployment
- SDKs and frameworks support AI applications across many domains
- Tools and libraries accelerate data analytics, AI model training and customization, and AI model optimization and deployment
- Infrastructure software helps manage AI clusters at scale, across the edge and data center, both bare-metal and virtualized
- Support across on-prem, edge, and cloud provides maximum flexibility

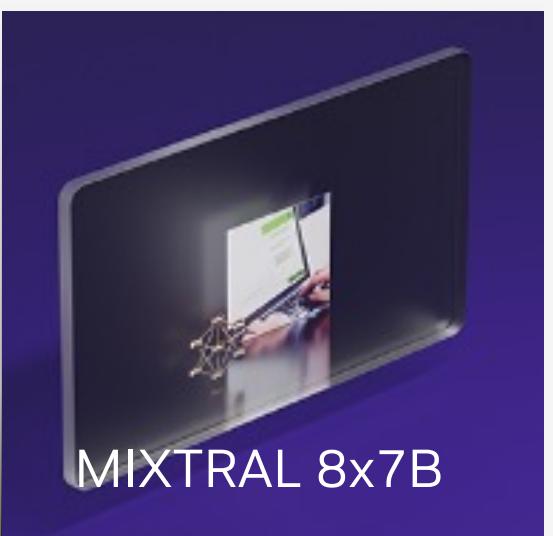
# Inference Microservices for Generative AI

NVIDIA NIM is the fastest way to deploy AI models on accelerated infrastructure across cloud, data center, and PC

NVIDIA API Catalog



Meta



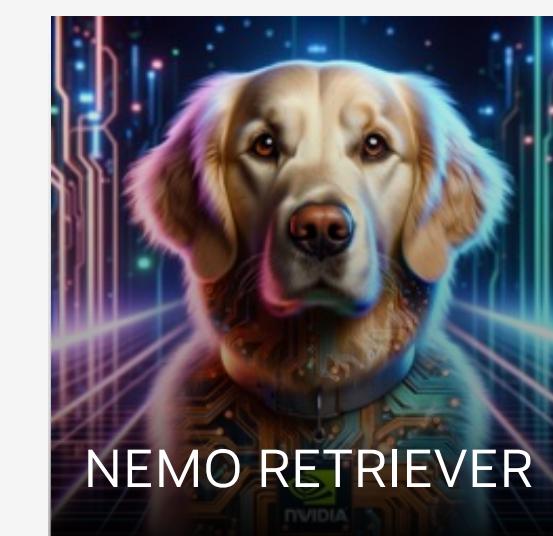
MISTRAL  
AI\_



Google



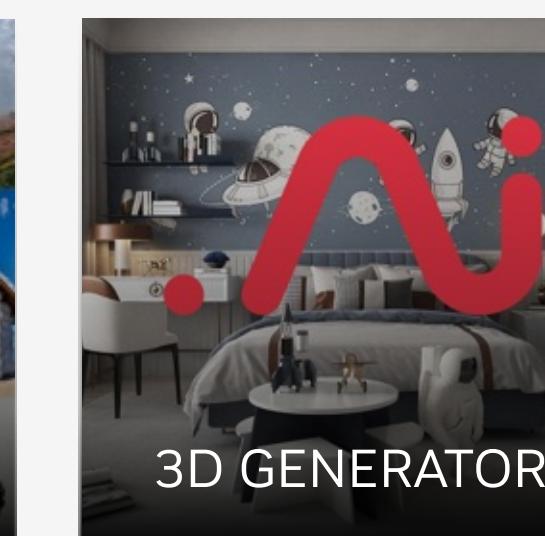
A D E P T



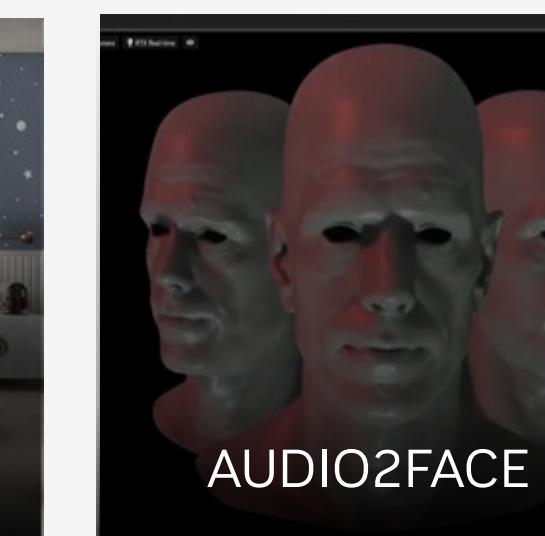
NVIDIA.  
FUYU



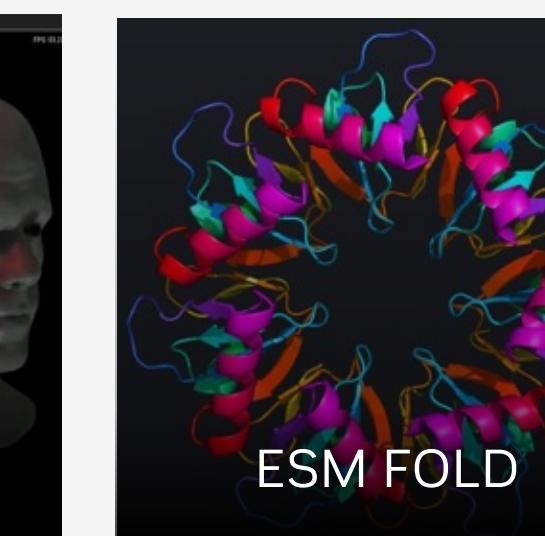
gettyimages



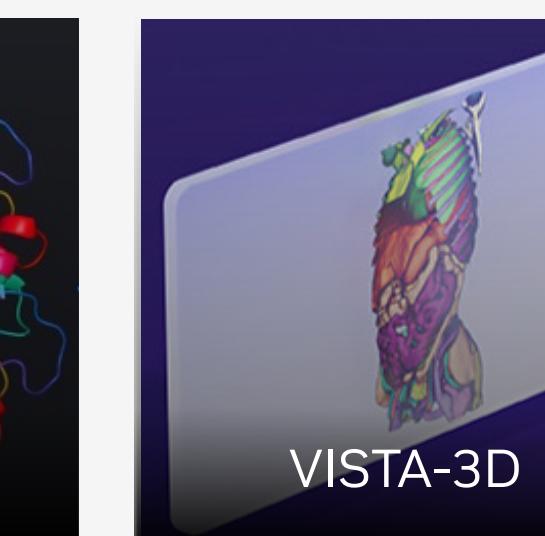
shutterstock



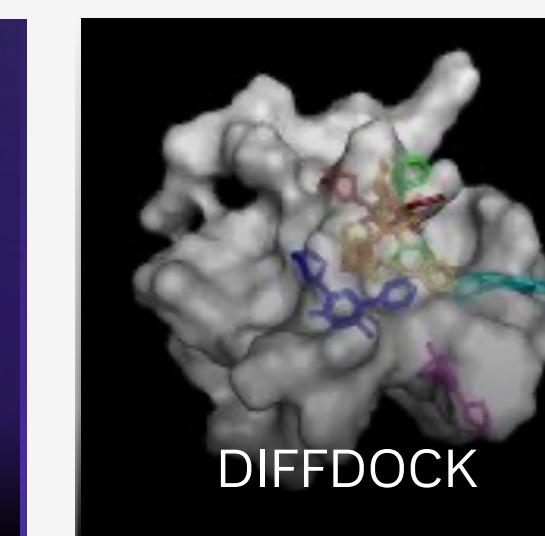
NVIDIA.  
AUDIO2FACE



Meta



NVIDIA.



MIT



NVIDIA.  
MolMIM



Microsoft  
Azure

aws

Google Cloud

ORACLE®



DELL Technologies

Hewlett Packard  
Enterprise

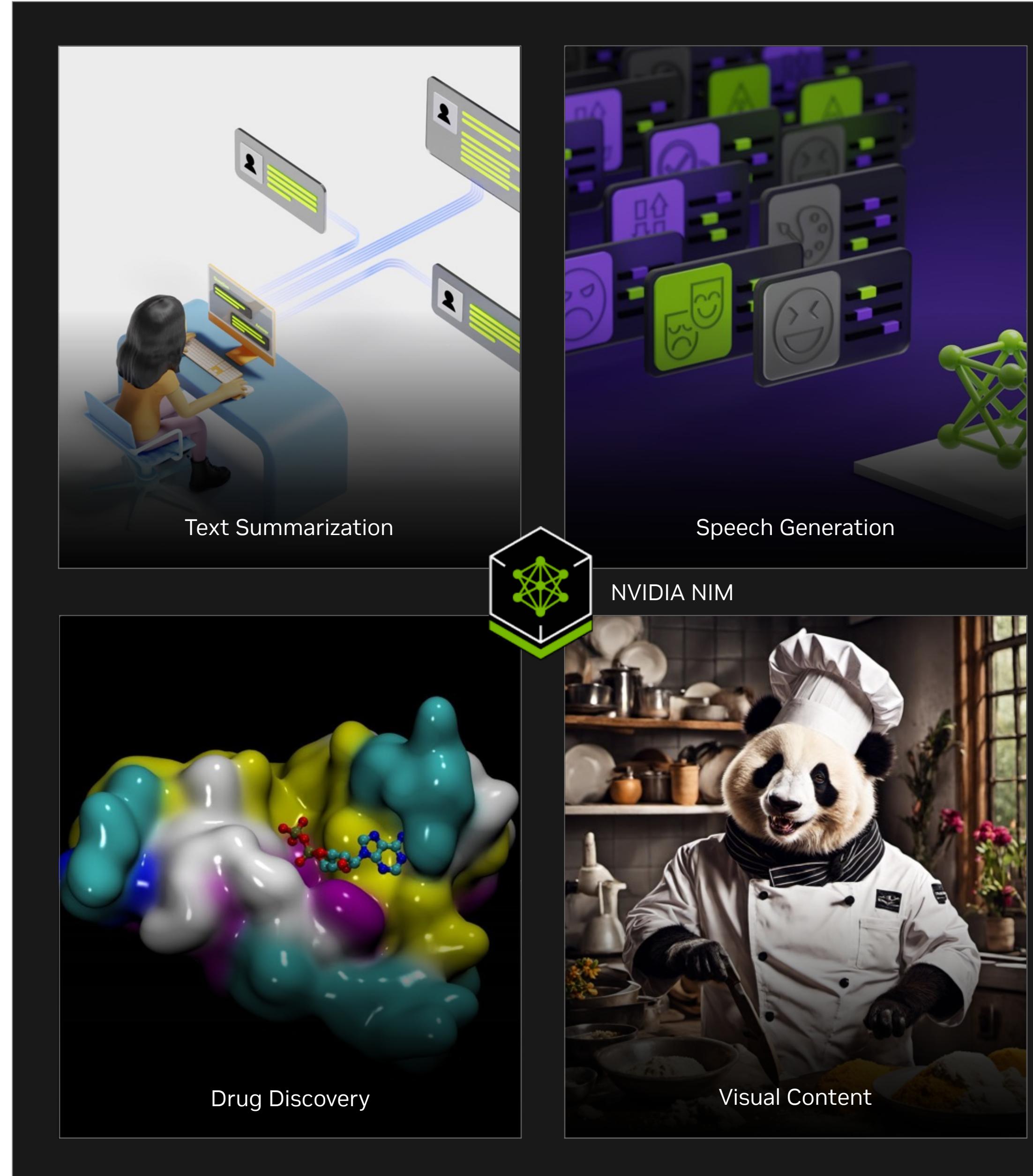
Lenovo

SUPERMICRO®

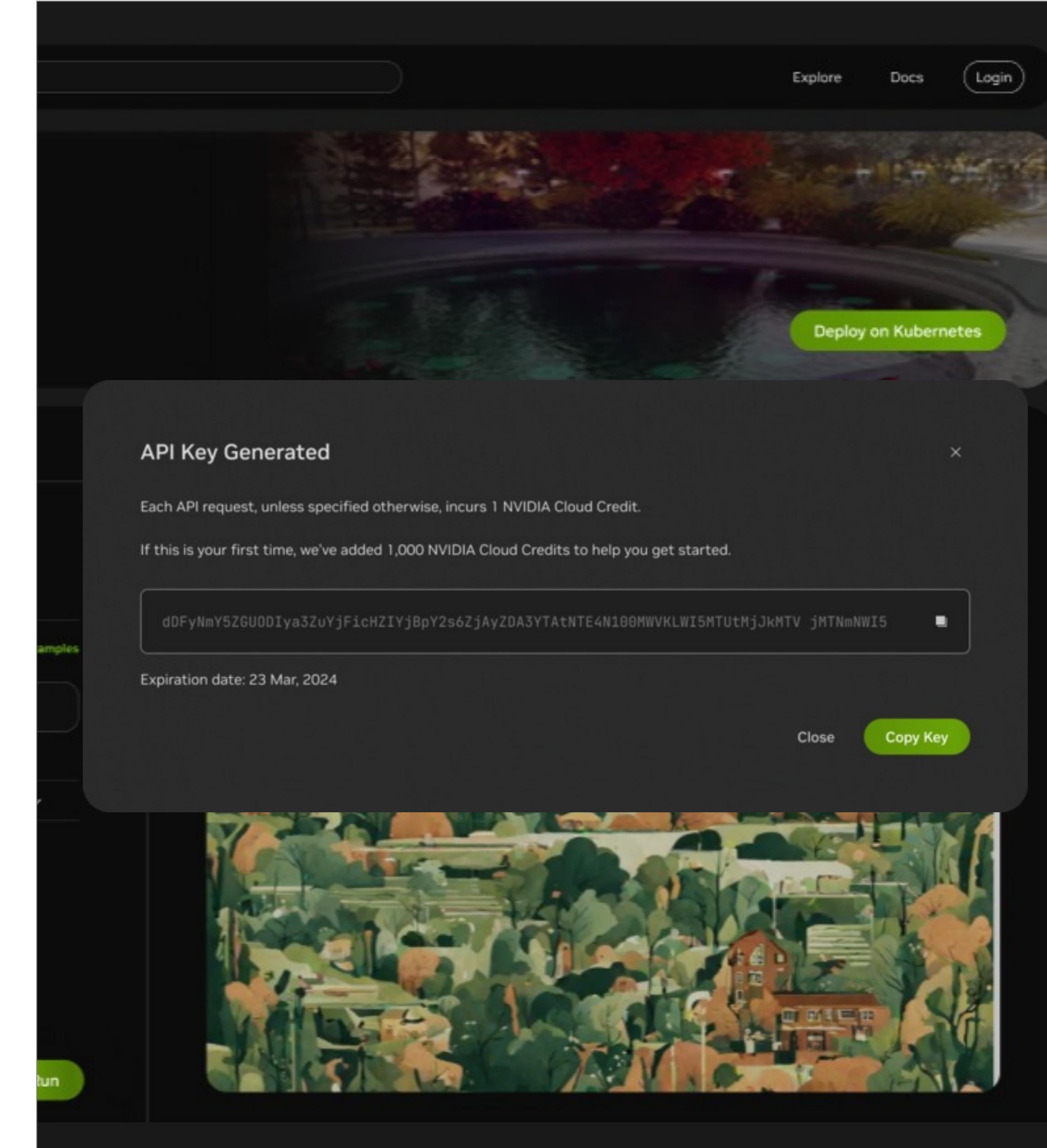
NVIDIA.

# Experience and Run Enterprise Generative AI Models Anywhere

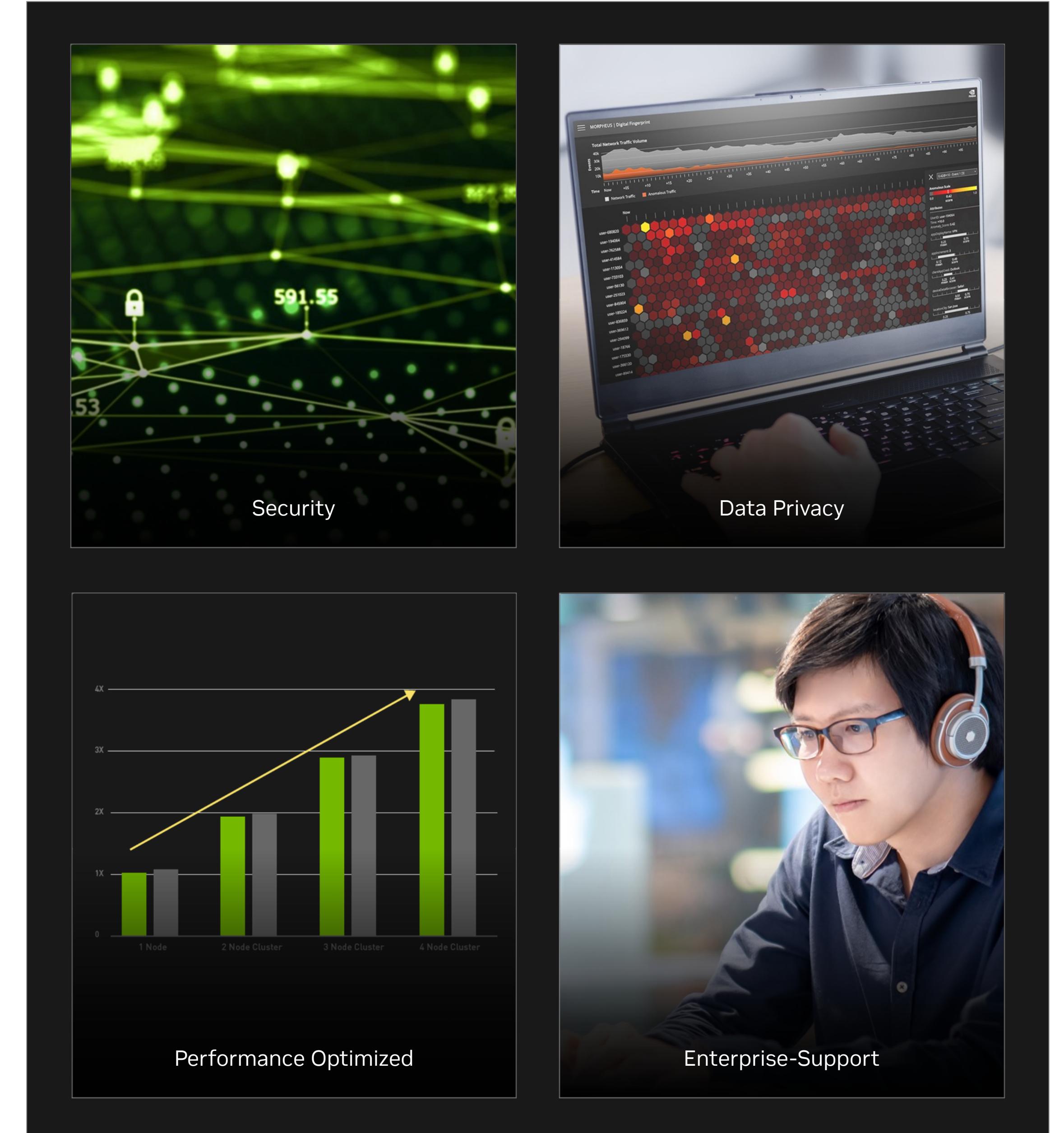
Seamlessly integrate AI in business applications with NVIDIA AI APIs



Experience Models

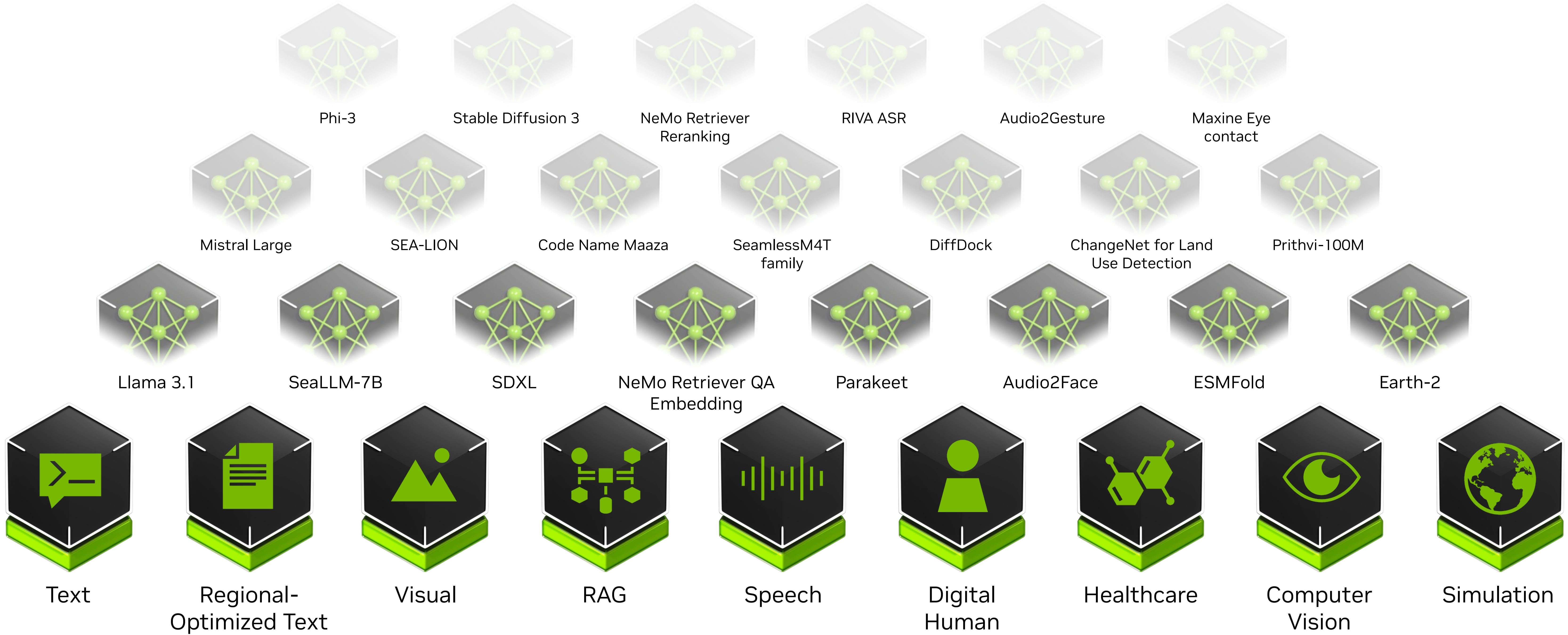


Prototype with APIs



Deploy with NIMs

# NVIDIA NIM For Every Domain



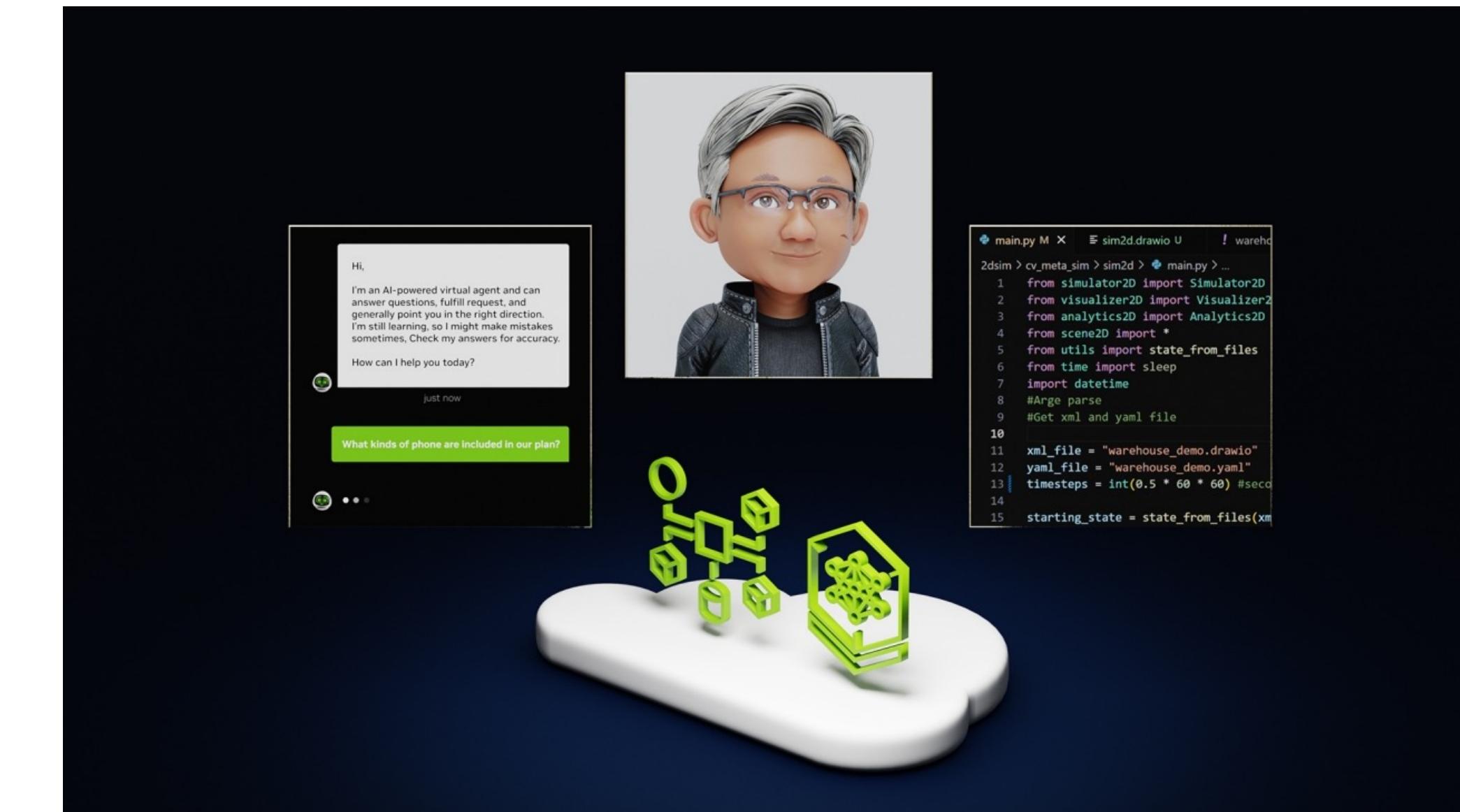
# NVIDIA Dynamo

# NVIDIA Dynamo Use Cases

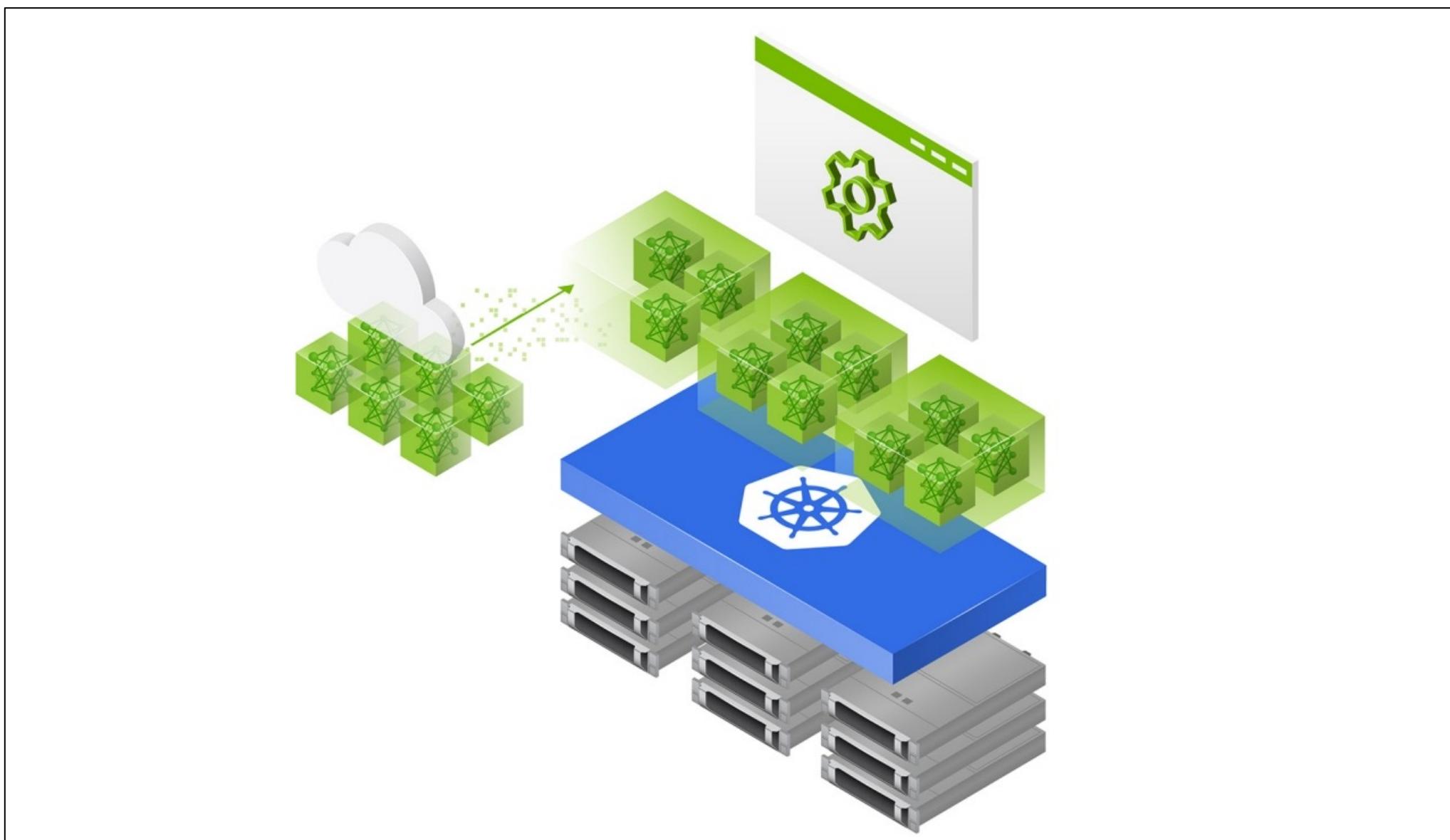
Unlock the full potential of reasoning models and AI agents



Serving Reasoning Models



Scalable AI Agents



Distributed Inference

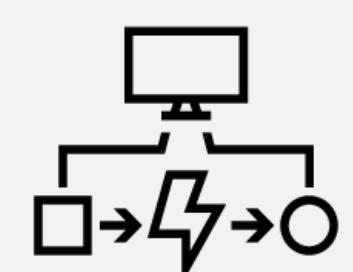


Code Generation

# NVIDIA Dynamo Breakthrough Features

A modular generative AI inference server designed for distributed and disaggregated serving

## NVIDIA Dynamo



### Distributed Inference Serving

Seamlessly scale LLMs from a single GPU to thousands of GPUs



### GPU Planning & Scheduling

Meet changing demand patterns w/o over or under provisioning of resources



### Smart Request Router

Free up GPU resources by reducing re-computations for similar requests



### Low-latency Inference Data Transfer Library

Accelerate GPU-to-GPU communication to enhance user experience



### KV Cache Manager

Preserve GPU memory by offloading context (KV\$) to cheaper storage

# NVIDIA Dynamo

AI Inference Software for Reasoning Inference at Scale

**30X**

AI Factory Throughput  
& Revenue  
DeepSeek models  
on Blackwell

**1000+**

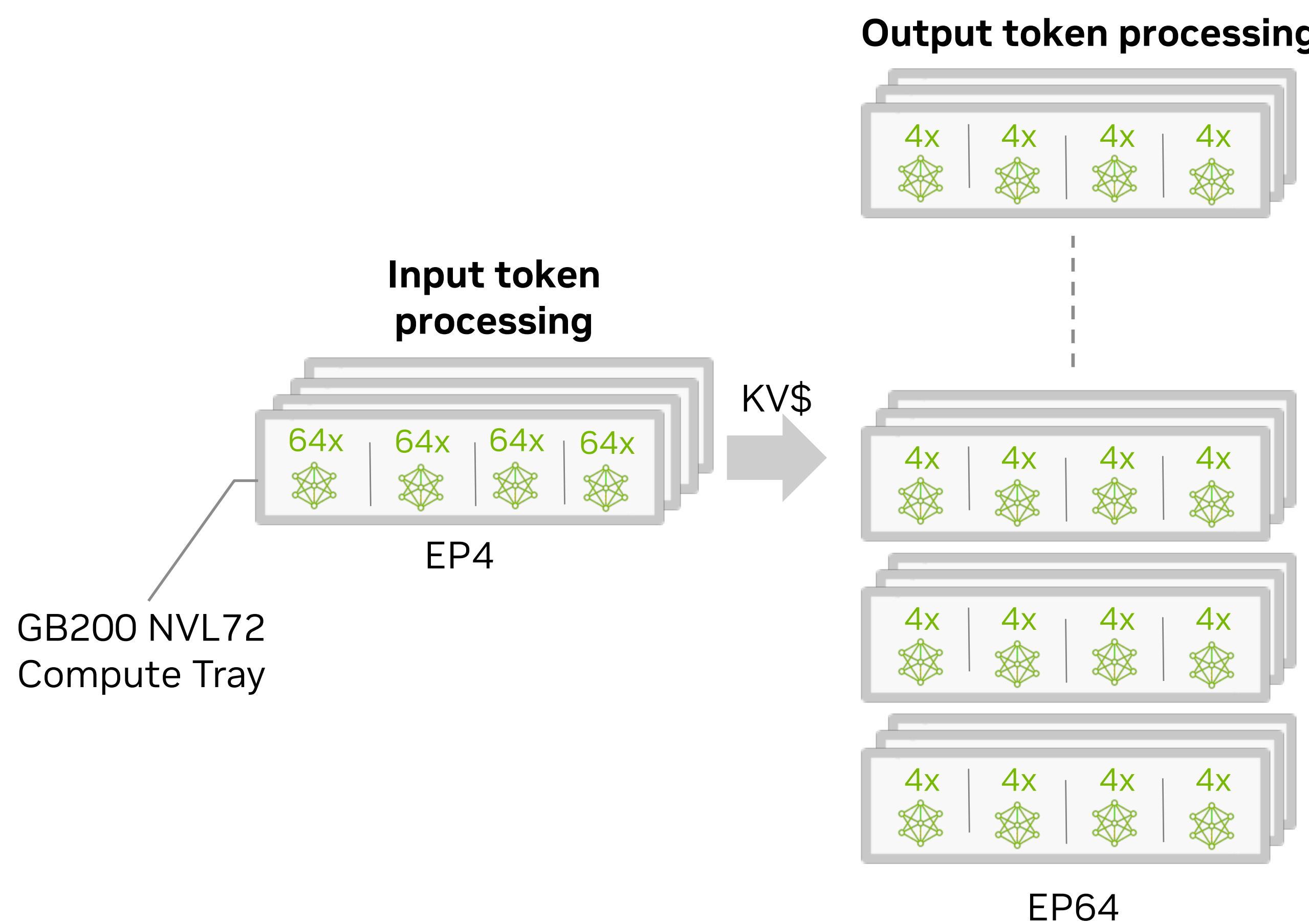
GPU Scale for  
a single query

**2X**

Throughput &  
Revenue  
Llama Models  
On Hopper

## Distributed and Disaggregated Serving

DeepSeek R1



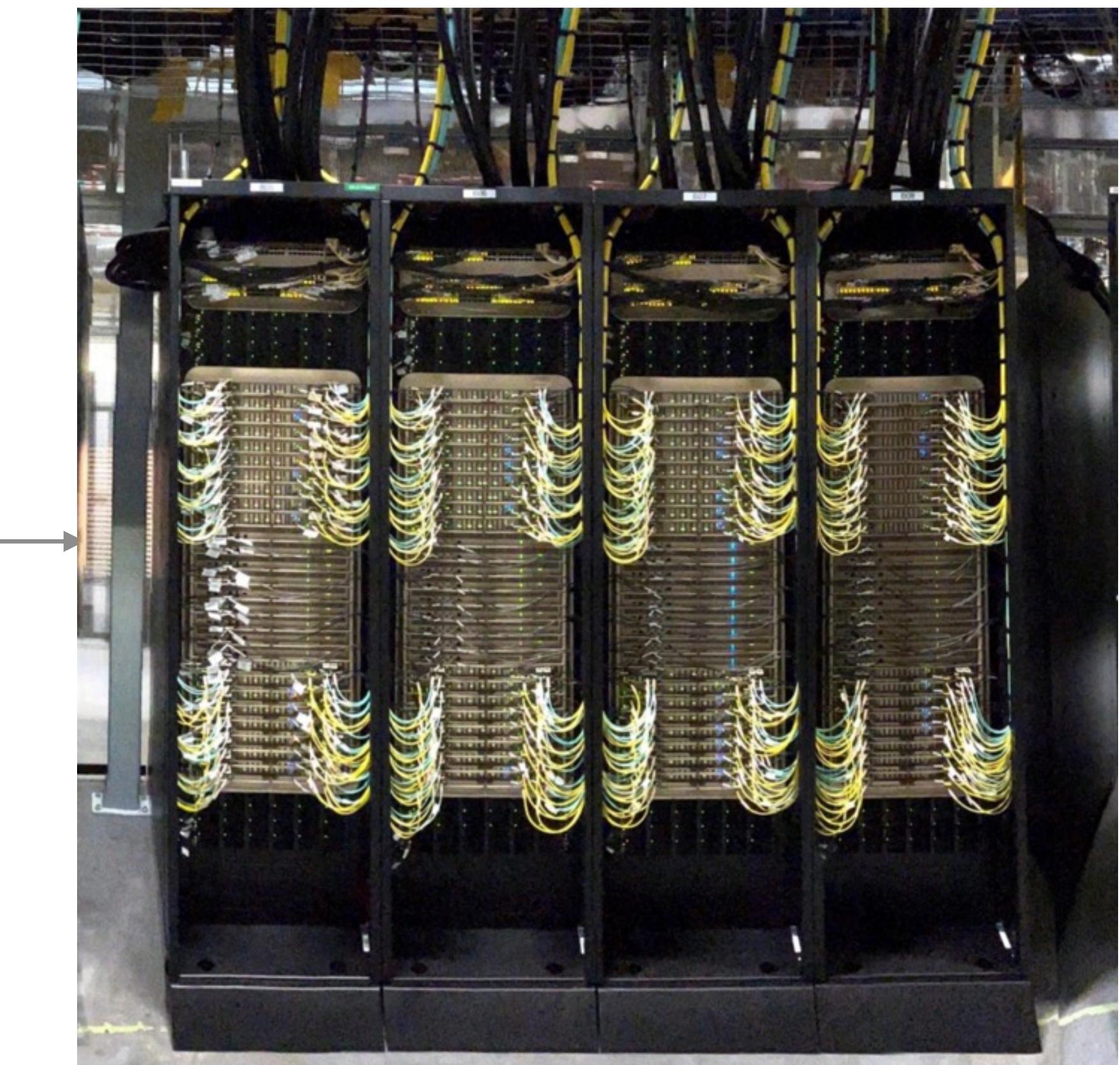
## NVIDIA Dynamo

Smart Router

GPU Planner

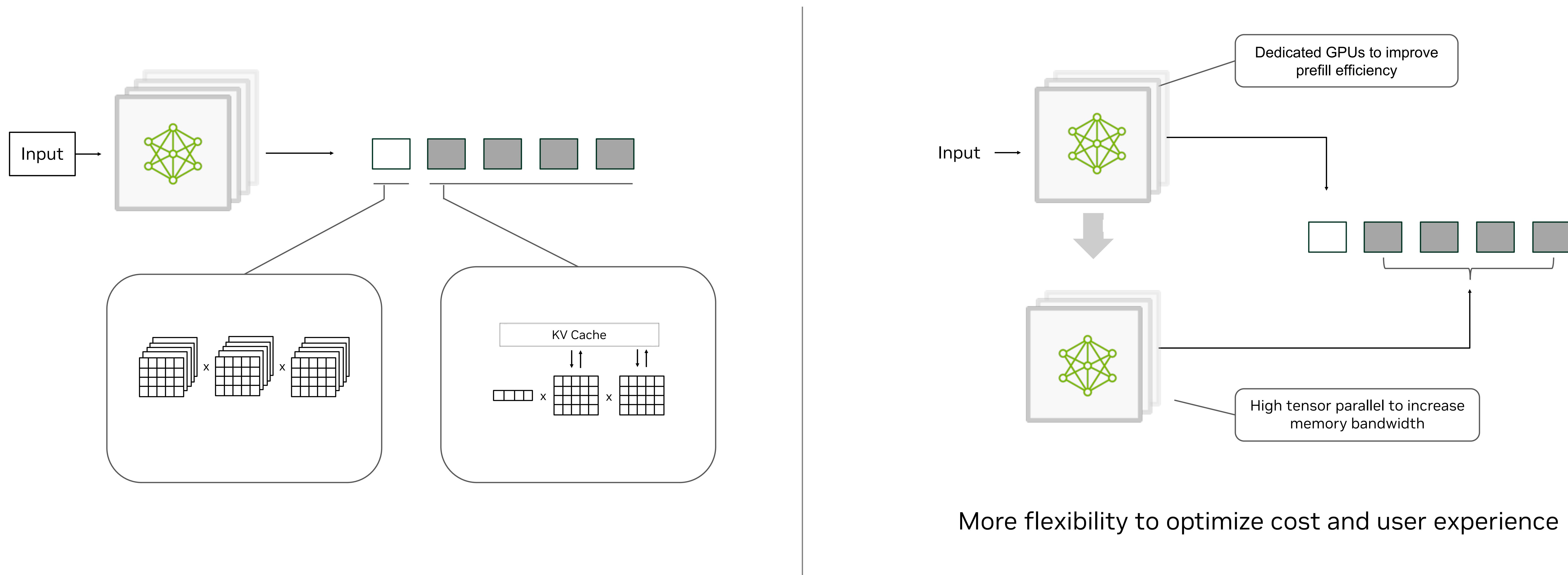
Low-latency  
Communication  
Library

KV-Cache  
Offload  
Manager

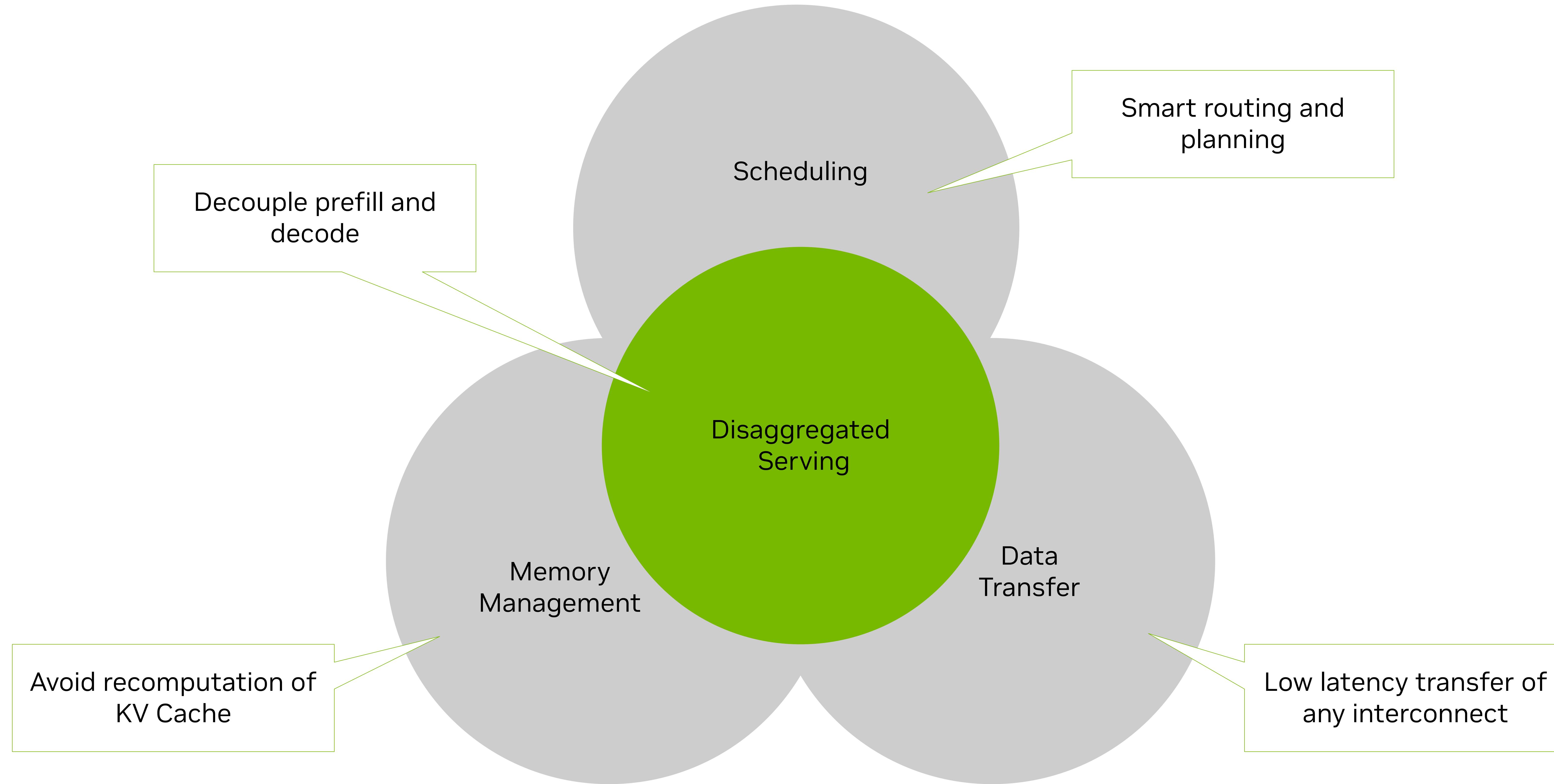


# New Inference Optimization Techniques to Boost Inference

Disaggregated serving separates prefill and decode allowing each to be optimized independently

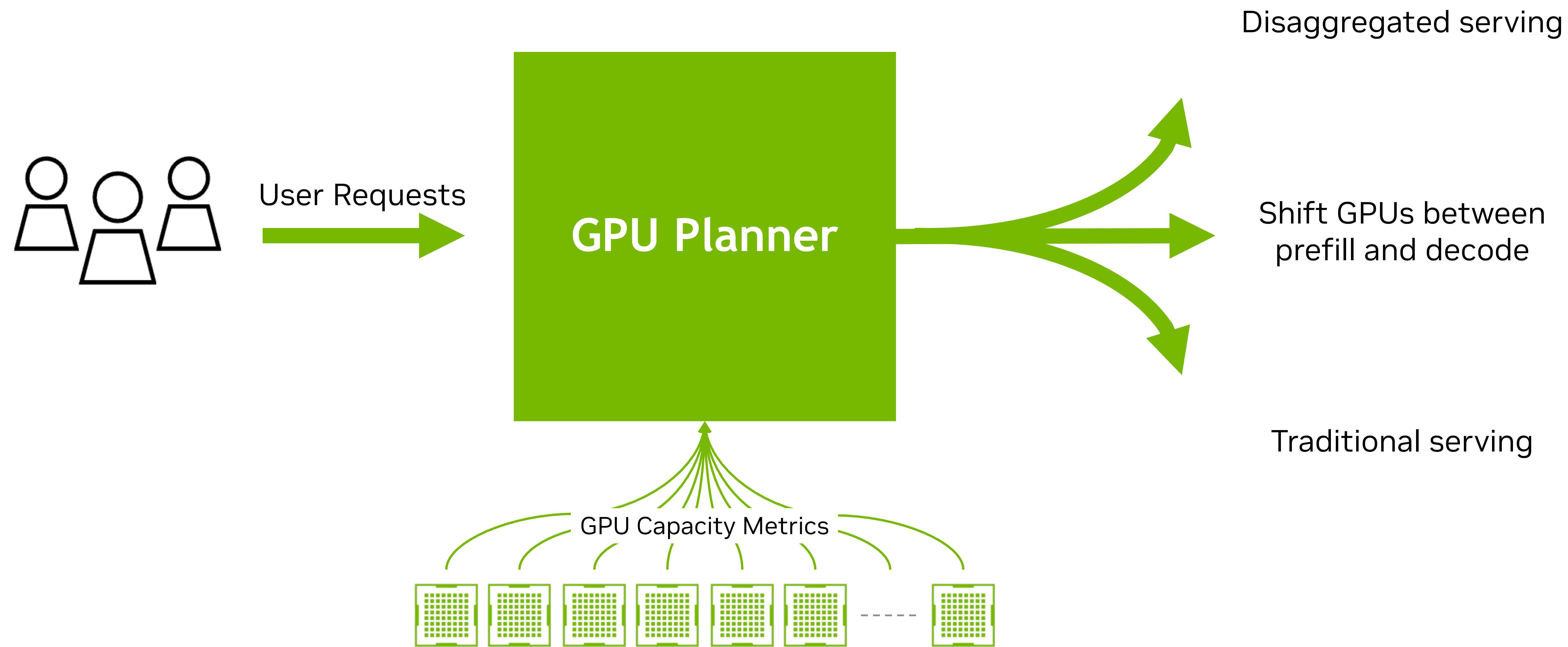


# Disaggregated Serving Requires Complex Resource Orchestration



# NVIDIA Dynamo: GPU Planner

Optimizing GPU resources for distributed inference



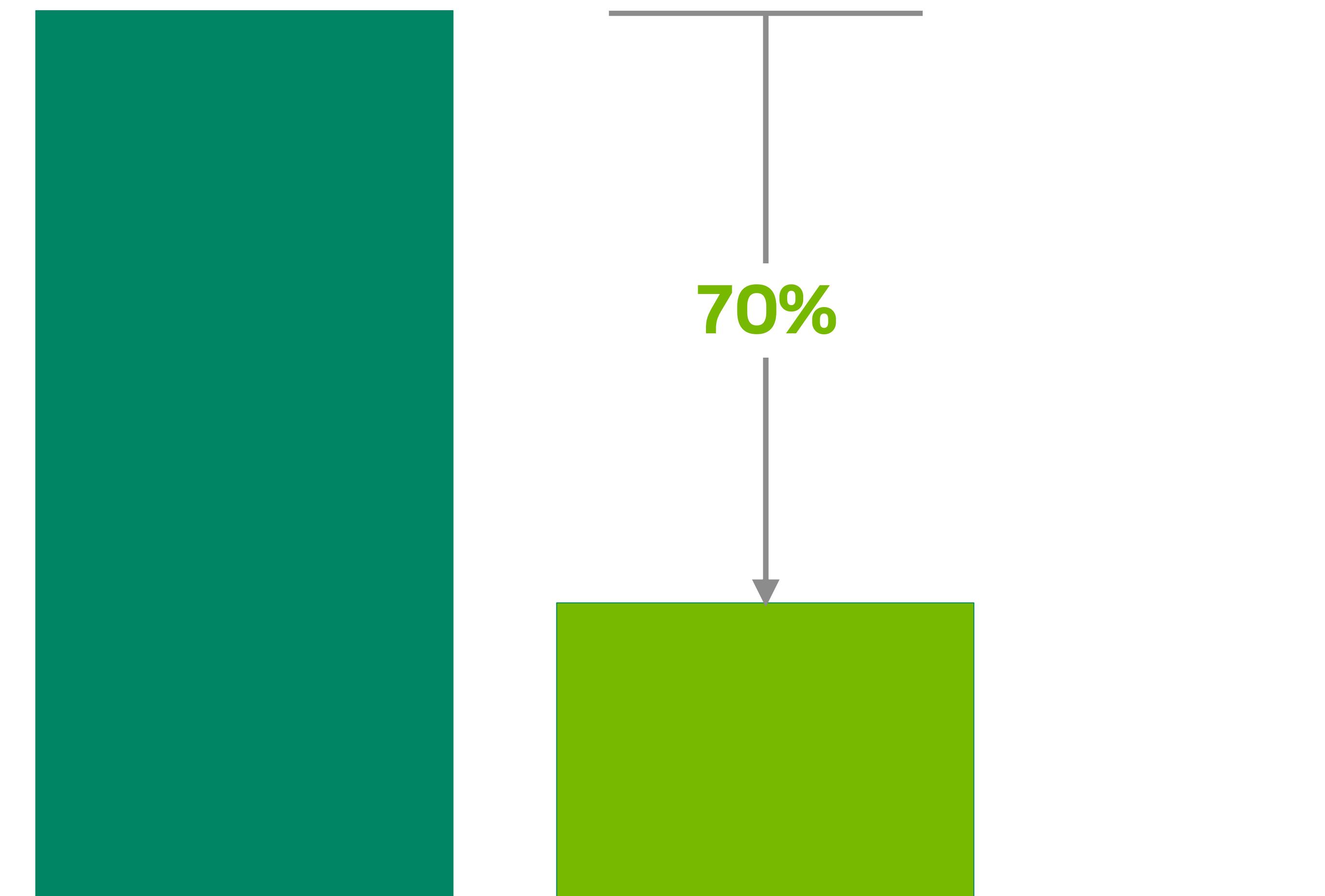
**Efficient Resource Allocation | Adjust to Fluctuating Demand | Lower Inference Costs**

# NVIDIA Dynamo: Smart Router

Reducing costly re-computation of KV cache

**DeepSeek-R1 Distill Llama 70B | NVIDIA HGX-H100**  
(Lower is Better)

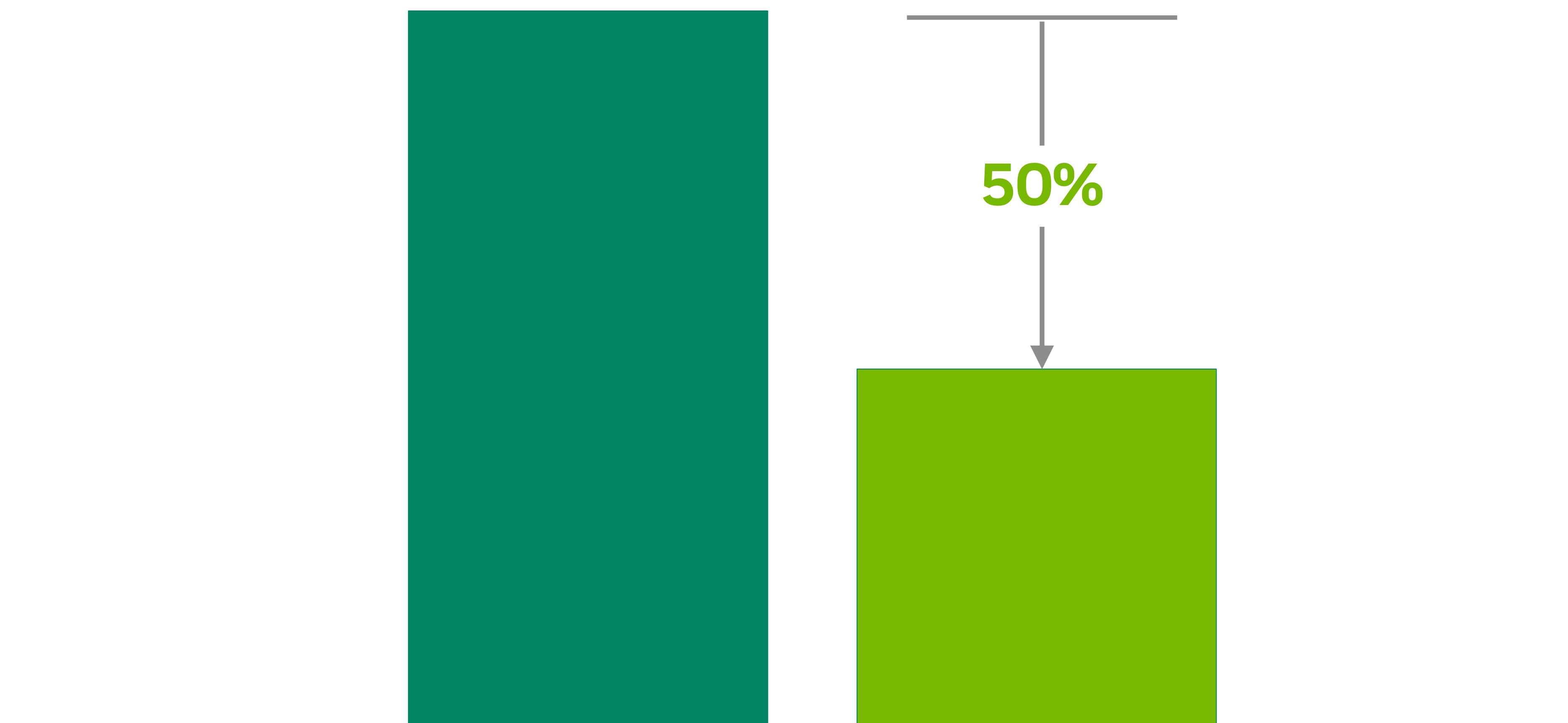
**Time to First Token**



■ NVIDIA Dynamo w/ Random Routing

■ NVIDIA Dynamo w/ Smart Router

**Avg. Request Latency**



■ NVIDIA Dynamo w/ Random Routing

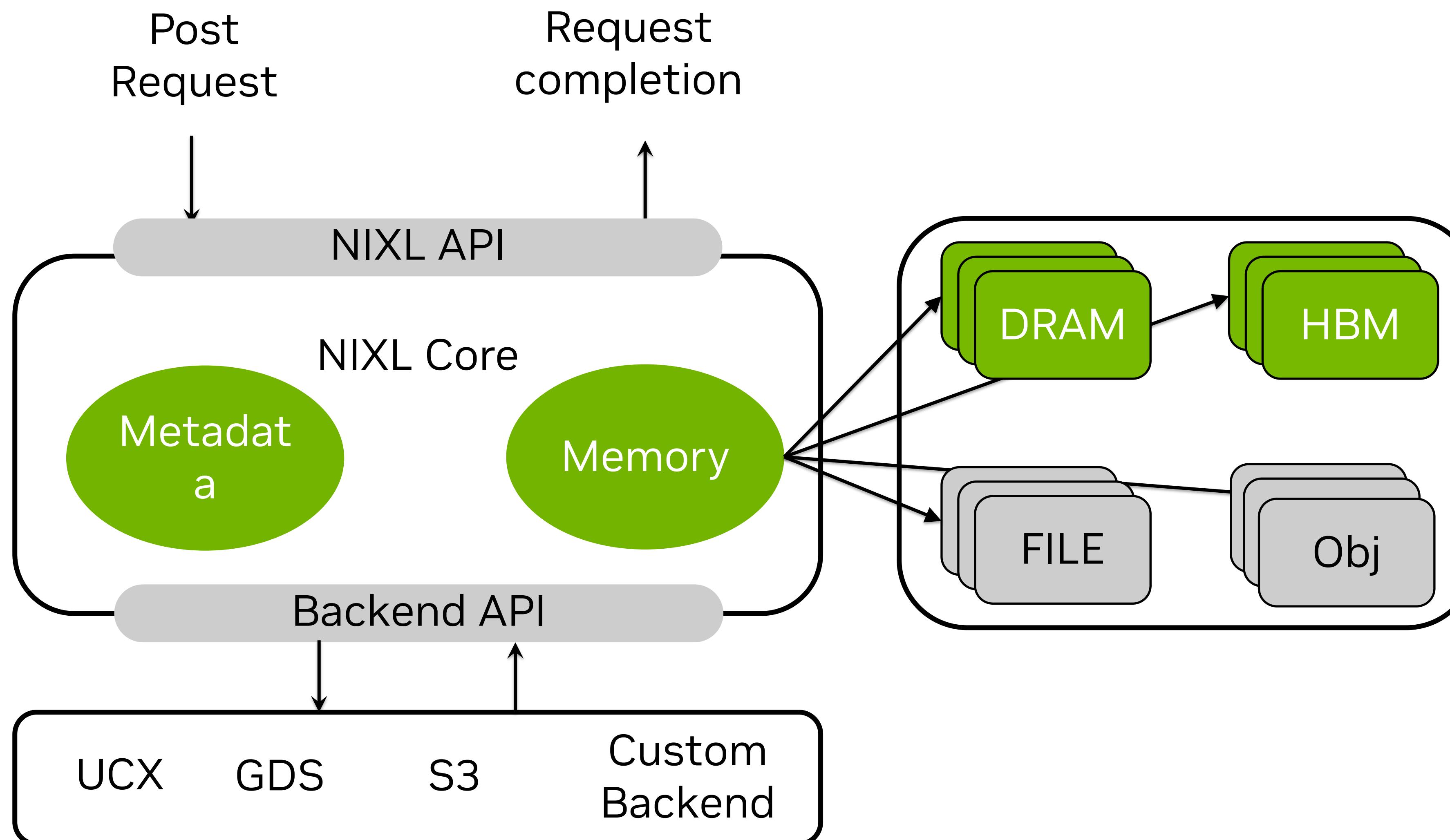
■ NVIDIA Dynamo w/ Smart Router

2x HGX-H100 nodes  
8x DeepSeek-R1- Distill-Llama-70B. vLLM, FP8, Tensor Parallel: 2  
Data Source: 100K real R1 requests, Avg ISL/OSL: 4k/800



# NVIDIA Inference Transfer Library (NIXL)

Low-latency, hardware-agnostic communication



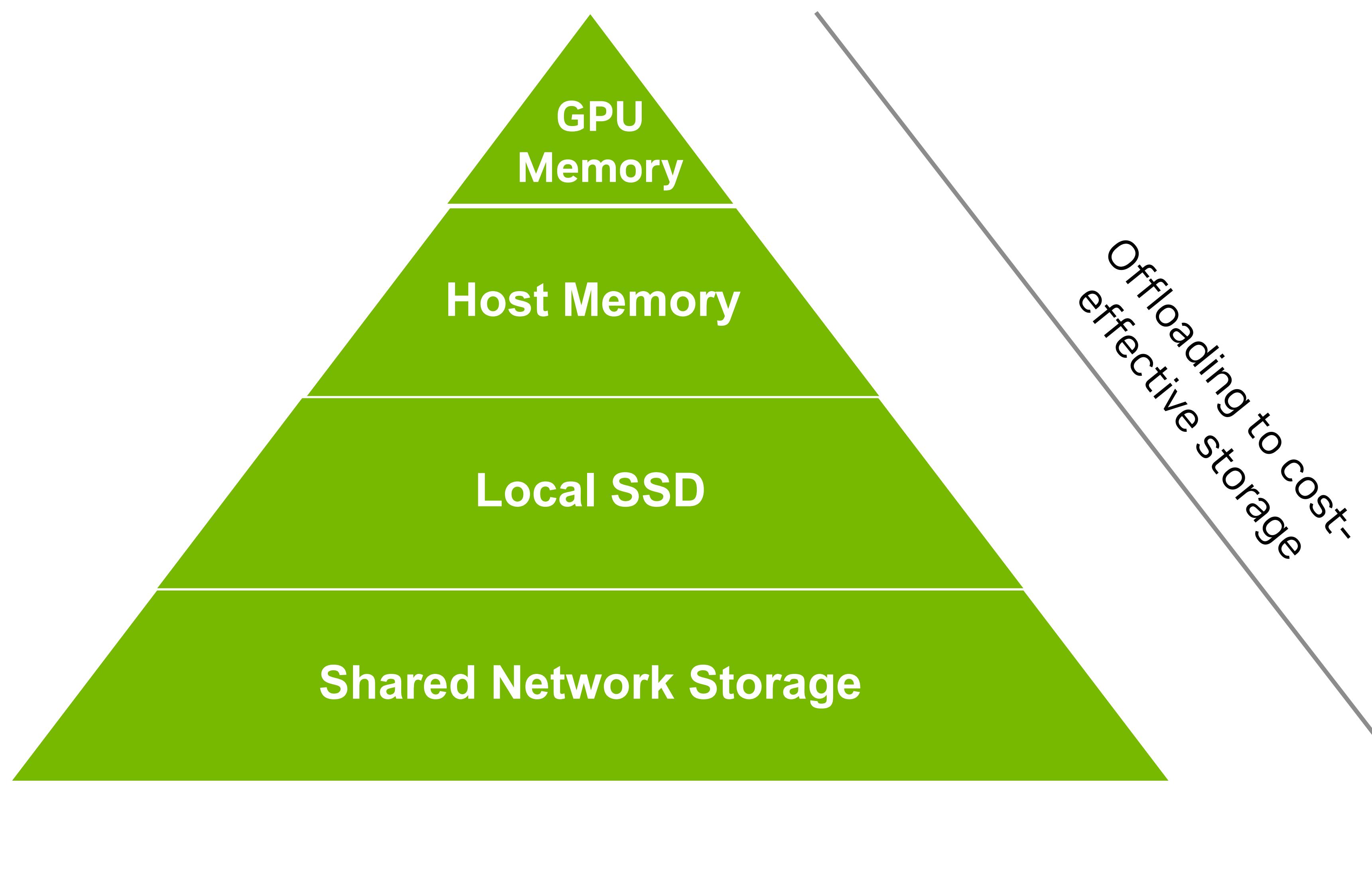
Optimized for inference data movement

Consistent API across heterogenous data paths

Supports different types of memory, SSDs, and networked storage

# NVIDIA Dynamo: KV Cache Manager

Offloading KV cache to cost-effective storage

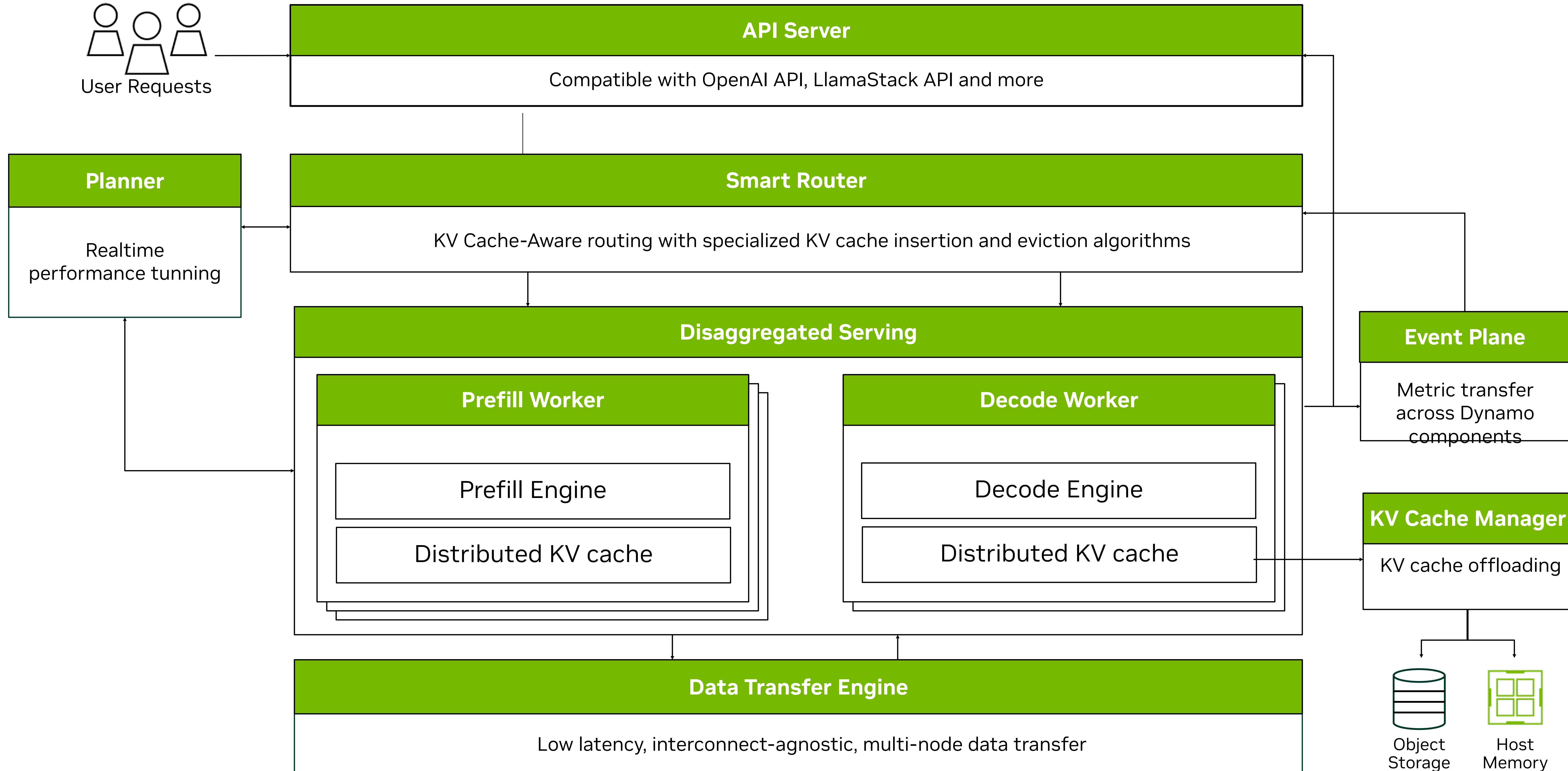


Store petabytes of KV\$ data at low cost

Offload older KV\$ to cost-effective storage

Strike a balance between over and under caching

# Architecture and Components



# Solutions Tailored to Customer Inference Needs

## NVIDIA Dynamo Platform

### NVIDIA Dynamo (announced at GTC25)

Distributed and Disaggregated  
Generative AI Serving

### NVIDIA Dynamo Triton (formerly Triton Inference Server)

Standardized Model Deployment  
Across Every AI Workload

## NVIDIA NIM

Fastest and Easiest Way to Deploy NVIDIA Dynamo Platform in Production

# Choose the Optimal Solution for Your Needs

	NVIDIA Dynamo Platform		NIM w/ Dynamo Platform
	Dynamo Triton (formerly Triton Inference Server)	Dynamo (launched at GTC25)	
<b>Key Benefits</b>	Standardizes AI model deployment and execution across every workload	Serves Gen AI models in large scale distributed environments	Fastest path for deployment
<b>Reasons to select</b>	Consolidate inference serving on a single server that supports all models, different backends, and runs on any H/W platform (GPUs, CPUs)	Advanced System Level Inference Optimizations: <ul style="list-style-type: none"><li>• Disaggregated Serving</li><li>• KV Cache aware routing</li><li>• KV Cache Offloading</li></ul>	Fastest path to deployment with a Turnkey solution for enterprises
<b>Supported via NVAIE</b>	Yes	Planned	Yes
<b>Support Disagg. Serving</b>	No	Yes	Planned
<b>Support VLMs</b>	Yes	Planned	Yes
<b>Support K8s</b>	Yes	Yes	Yes
<b>Supports for any LLM</b>	Yes	Yes	coming soon
<b>Target Audience</b>	Enterprises, ISVs and GSIs	CSPs, CRISP, LLM Providers	Enterprises, ISVs and GSIs, GenAI Ops, OEMs/ODMs

# Get started with NVIDIA Dynamo



[Deploy NVIDIA Dynamo on GitHub](#)

[Watch NVIDIA Dynamo Demos](#)

[Read NVIDIA Dynamo Technical Blog](#)

NVIDIA Dynamo will be made available in NVIDIA NIM microservices for simplified deployment and supported by NVIDIA AI Enterprise in a future release with production-grade security, support and stability.



# Customer Testimonial Highlights



## Cohere

"Scaling advanced AI models requires sophisticated multi-GPU scheduling, seamless coordination and low-latency communication libraries that transfer reasoning contexts seamlessly across memory and storage. We expect NVIDIA Dynamo will help us deliver a premier user experience to our enterprise customers." **Saurabh Baji, Senior Vice President of Engineering at Cohere**



## Perplexity AI

"Handling hundreds of millions of requests monthly, we rely on NVIDIA's GPUs and inference software to deliver the performance, reliability, and scale our business and users demand, "We'll look forward to leveraging NVIDIA Dynamo with its enhanced distributed serving capabilities to drive even more inference serving efficiencies and meet the compute demands of new AI reasoning models." **Denis Yarats, CTO of Perplexity AI.**



## Together AI

"Scaling reasoning models cost-effectively requires new advanced inference techniques, including disaggregated serving and context-aware routing. Together AI provides industry leading performance using our proprietary inference engine. The openness and modularity of NVIDIA Dynamo will allow us to seamlessly plug its components into our engine to serve more requests while optimizing resource utilization—maximizing our accelerated computing investment." **Ce Zhang, CTO of Together AI.**

# Accelerating AI Ecosystem

Fully open source and supports all major AI frameworks



PyTorch SGL together.ai VAST vLLM