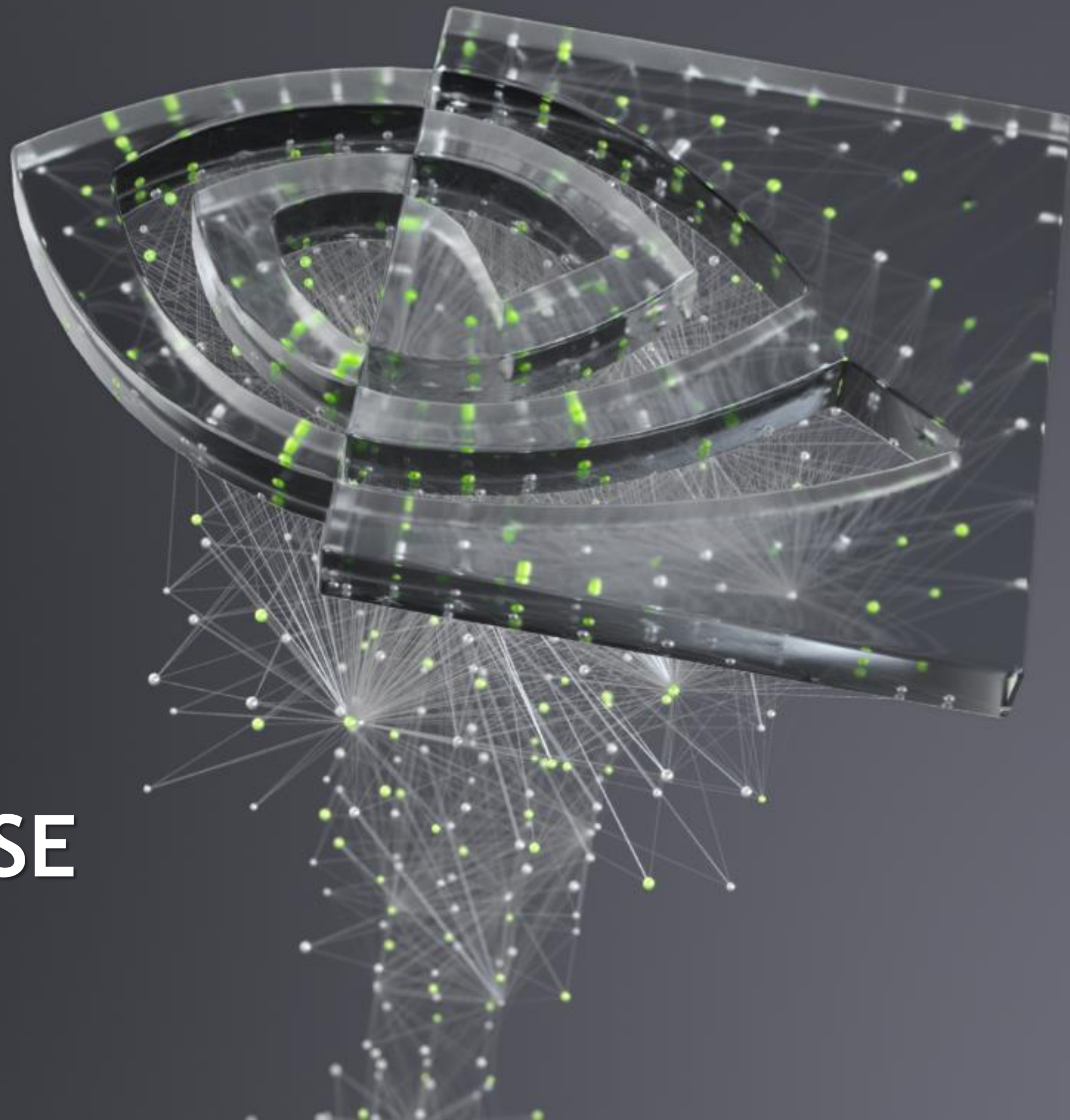




NVIDIA AI ENTERPRISE

March , 2022





目录

- AI在企业应用的机遇和挑战
 - NVIDIA AI Enterprise解决方案和适用场景
 - NVIDIA AI Enterprise技术亮点
 - NVIDIA AI Enterprise性能
 - 成功案例
-



AI在企业应用的机遇和挑战

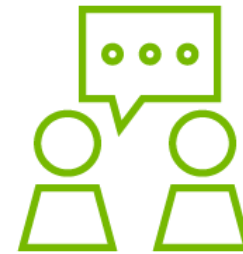
AI加速各行各业

到 2024 年，60% 的 G2000 将在所有关键业务中使用 AI/ML 技术

人力资源
人才招聘



市场营销
个性化推荐



IT系统和安全
网络安全/欺诈检测



企业运营
预测性维护和供应链



客户服务
虚拟助手和聊天机器人

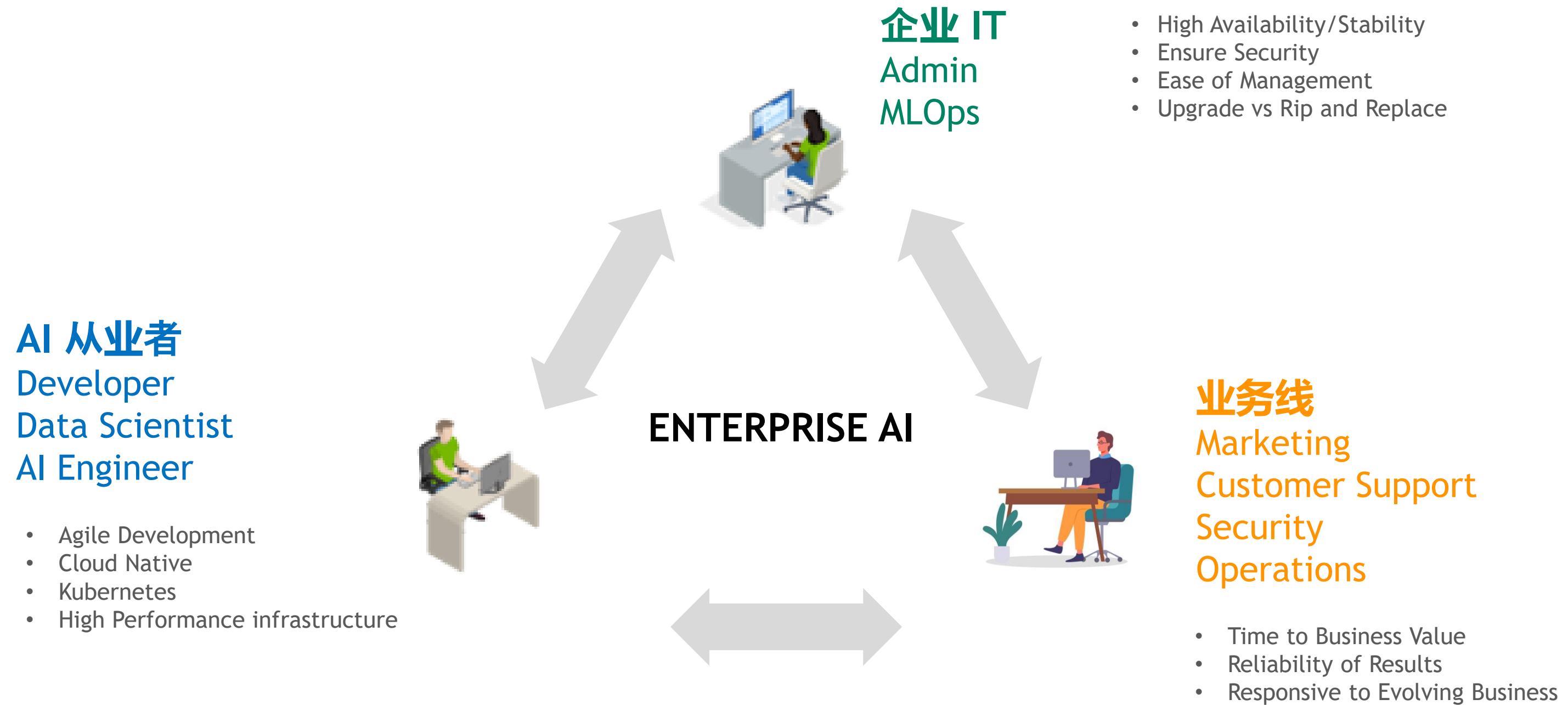


产品开发
数字原型和过程自动化



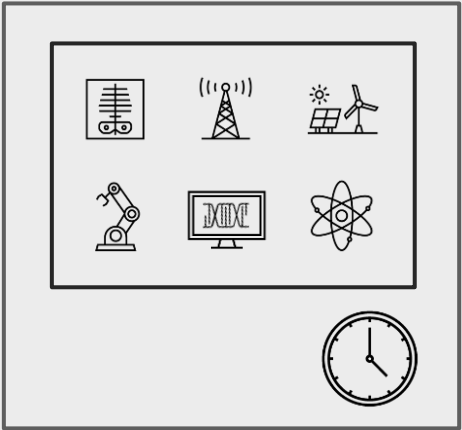
企业 AI 是一个团队项目

每个团队成员都有不同的优先级和要求



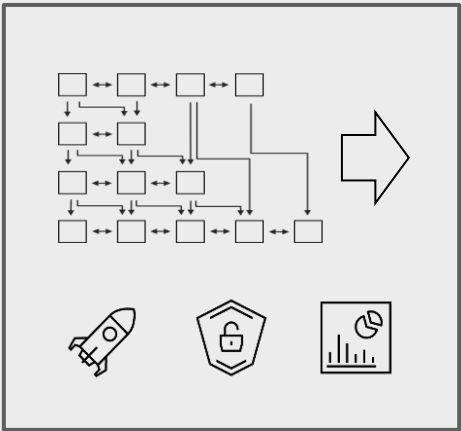
企业落地AI应用的新途径

几乎一半的 AI 项目从未投入生产



不算增加的复杂性

32% 的 IT 专业人员将不断增加的数据复杂性和数据孤岛视为人工智能采用的最大障碍²



缺乏AI工具和平台

28% 的 IT 专业人员认为缺乏基础设施是人工智能采用的最大障碍²

用于简化开发和部署的集成解决方案



Process Automation



Conversational AI



Image Analysis

...



Existing Applications

AI Tools and Frameworks

Management

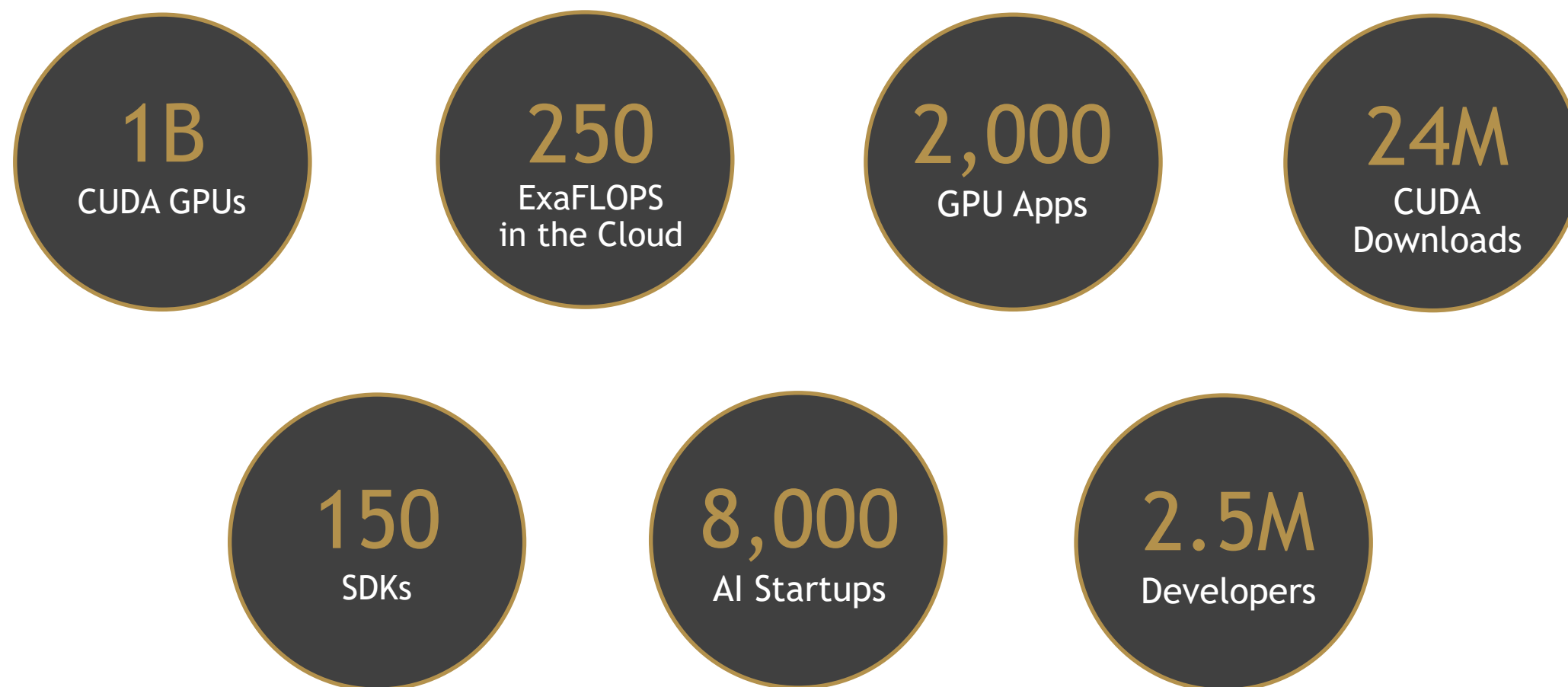
Accelerated Infrastructure



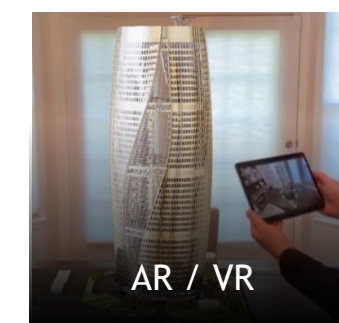
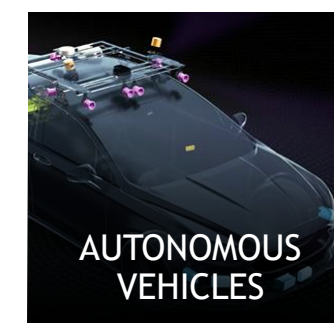
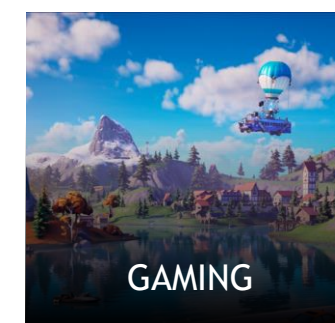
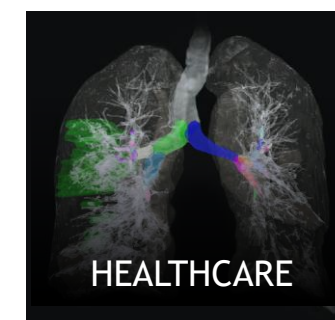
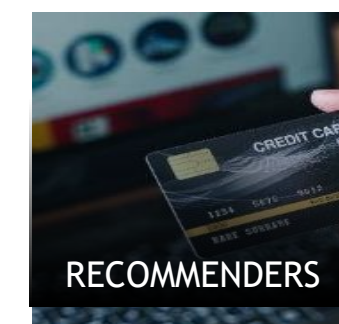
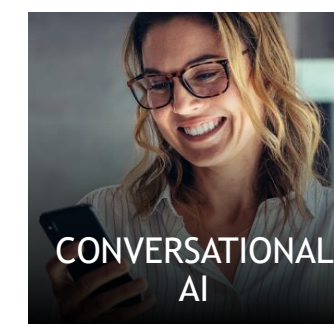
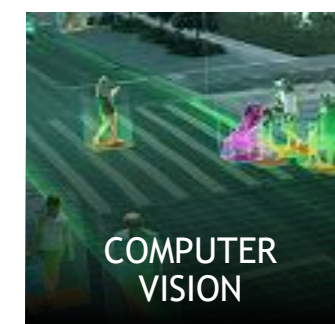
NVIDIA AI ENTERPRISE解决方案和适用场景

NVIDIA 是企业人工智能领域的领导者

为开发和部署提供框架和平台

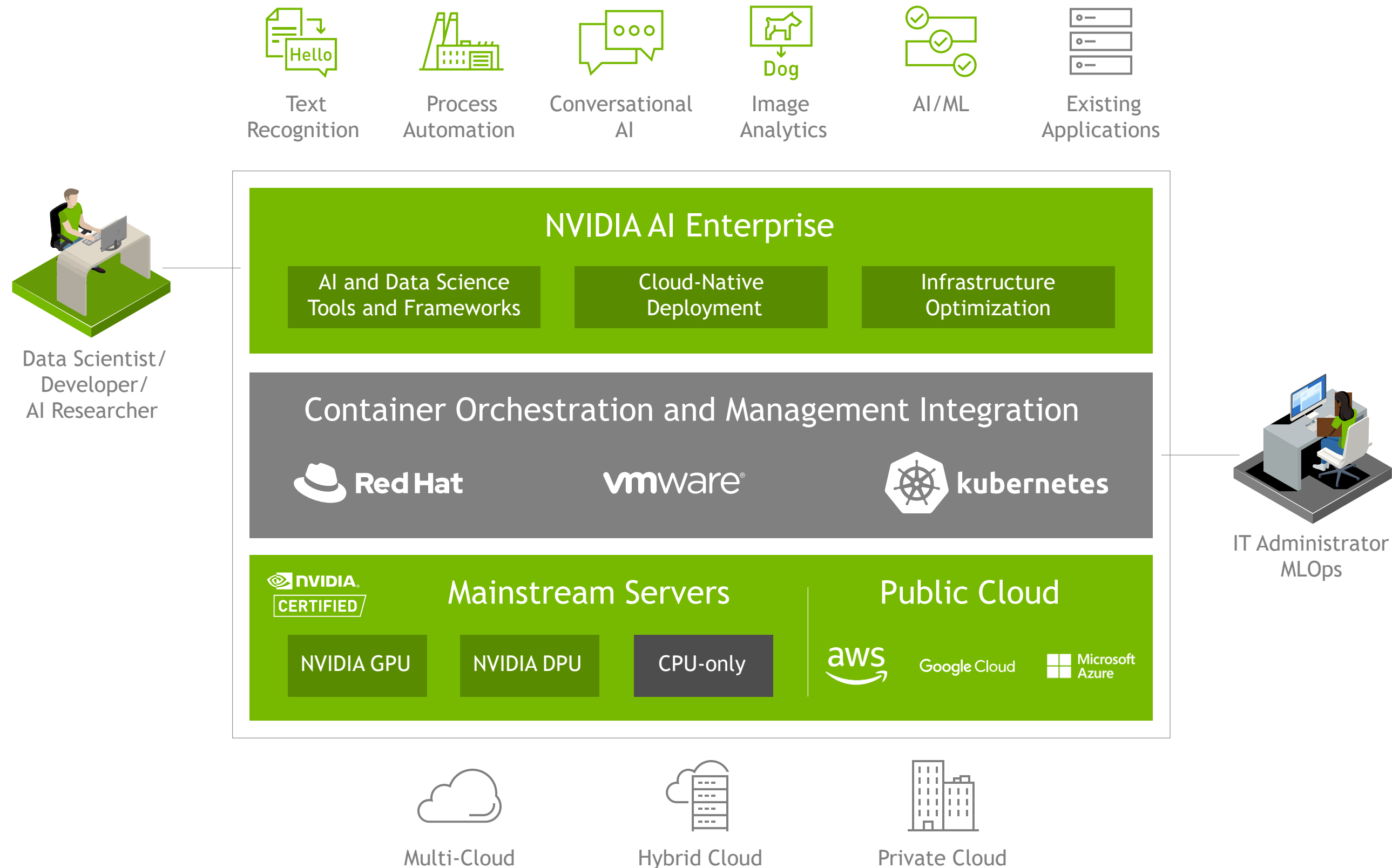


TRANSFORMING INDUSTRIES



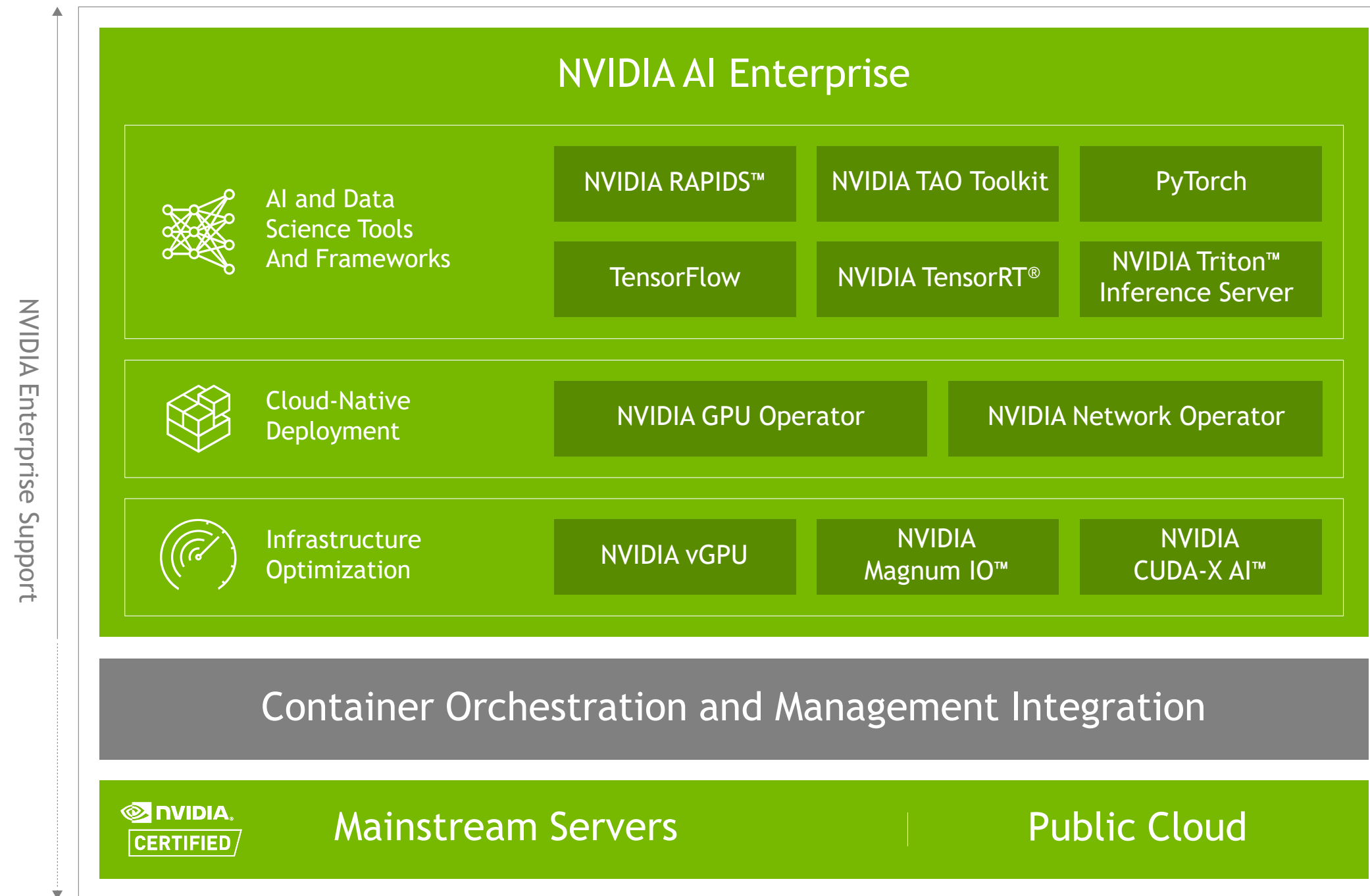
人工智能就绪的企业平台

面向每个人、任何地方、每个平台的企业 AI



NVIDIA AI ENTERPRISE 软件套件

面向每个人、任何地方、每个平台的企业 AI



NVIDIA AI ENTERPRISE

为企业释放AI的能量



针对AI部署的优化

包括了性能优化的 DL/ML 工具的 AI 框架和容器，以简化 AI 的构建和部署。



企业认证

在主流企业平台上部署，降低部署风险 - 支持裸机、虚拟化、公共云，由 GPU 加速或仅 CPU。



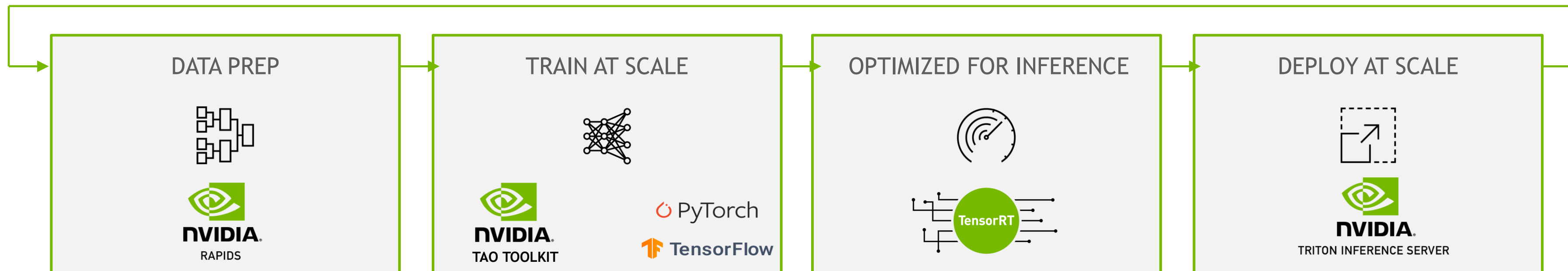
NVIDIA企业级技术支持

借助 NVIDIA 专家、长期支持、最新产品更新和企业培训，确保关键 AI 项目按计划进行。



NVIDIA 端到端 AI 软件套件

让深度学习应用从探索到大规模生产



- 将数据科学流程从几小时缩短到几秒
- 性能比类似 CPU 配置快 70 倍
- 比同类 CPU 配置更具成本效益 20 倍

- 在数小时而不是数月内训练、调整和优化模型
- 针对 GPU 优化的开源 ML 框架
- 与 NVIDIA RAPIDS 集成以简化开发

- 最大化吞吐量
- 优化每个网络（CNN、RNN 和 Transformer）
- 优化 GPU 内存带宽的使用

- 应用程序的快速且可扩展的 AI
- 多种查询类型——实时、离线批处理、集成
- 与仅 CPU 相比，性能提升高达 266 倍
- 带有 FIL 后端的 Triton 为 GPU 上的基于树的模型提供最佳推理性能

NVIDIA 端到端 AI 软件套件

让深度学习应用从探索到大规模生产

 NVIDIA NGC | CATALOG

CATALOG

Explore Catalog

Enterprise Catalog

Collections

Containers

Helm Charts

Models

Resources

Catalog > Collections > NVIDIA AI Enterprise 2.0

NVIDIA AI Enterprise 2.0



Description

NVIDIA AI Enterprise is an end-to-end, cloud-native, suite of AI and data analytics software, optimized and certified by NVIDIA to run on VMware vSphere with NVIDIA-Certified Systems.

Curator

NVIDIA

Modified

March 26, 2022

Ai

Deep-Learning

Containers

11

Helm Charts

2

Models

0

Resources

3

Overview

Entities

What is NVIDIA AI Enterprise?

NVIDIA AI Enterprise is an end-to-end, cloud-native suite of AI and data analytics software, optimized for enterprise data centers, public cloud—and includes global enterprise support and training. It includes key enabling technologies from NVIDIA for rapid deployment, management, and scaling of AI workloads in the modern cloud.

NVIDIA AI Enterprise is licensed and supported by NVIDIA. For more information on getting started, see [documentation](#).

NVIDIA AI Enterprise

March 26, 2022

Ai

Deep-Learning

Containers

NAME	REPOSITORY	PUBLISHER	SIZE
 NVIDIA GPU Operator v1.10.0	nvaie/gpu-operator-2-0	NVIDIA	171.14 ...
 NVIDIA Network Operator v1.1.0	nvaie/network-operator-2-0	NVIDIA	18.55 MB
 NVIDIA RAPIDS 22.02-cuda1...	nvaie/nvidia-rapids-2-0	NVIDIA	5.33 GB
 PyTorch 22.02-nvaie-...	nvaie/pytorch-2-0	Facebook	6.51 GB
 TAO Toolkit for Langua... v3.21.08-py3	nvaie/tao-toolkit-lm-2-0	NVIDIA	5.57 GB
 TAO Toolkit for Conv AI v3.22.02-py3	nvaie/tao-toolkit-pyt-2-0	NVIDIA	7.91 GB
 TAO Toolkit for CV v3.21.11-tf1....	nvaie/tao-toolkit-tf-2-0	NVIDIA	7.73 GB

13





NVIDIA AI ENTERPRISE技术亮点

灵活多样的部署方式

部署在经过认证的任意位置



AI and Data
Science Tools
and Frameworks

NVIDIA RAPIDS™

TensorFlow

PyTorch

NVIDIA TAO Toolkit

NVIDIA TensorRT®

NVIDIA Triton™
Inference Server

	 VMware vSphere® with Tanzu™	 Red Hat OpenShift ¹	 kubernetes ²	Deep Learning Frameworks/Data Science Containers ³
GPU	Yes	Yes	Yes	Yes
CPU	Yes	Yes	Yes	Yes
Bare-Metal	No	Yes	Yes	Yes
Virtualized ⁴	Yes	Yes	Yes	Yes
Public Cloud	No	No	No	Yes ⁵

¹With Red Hat CoreOS

²Upstream Kubernetes only

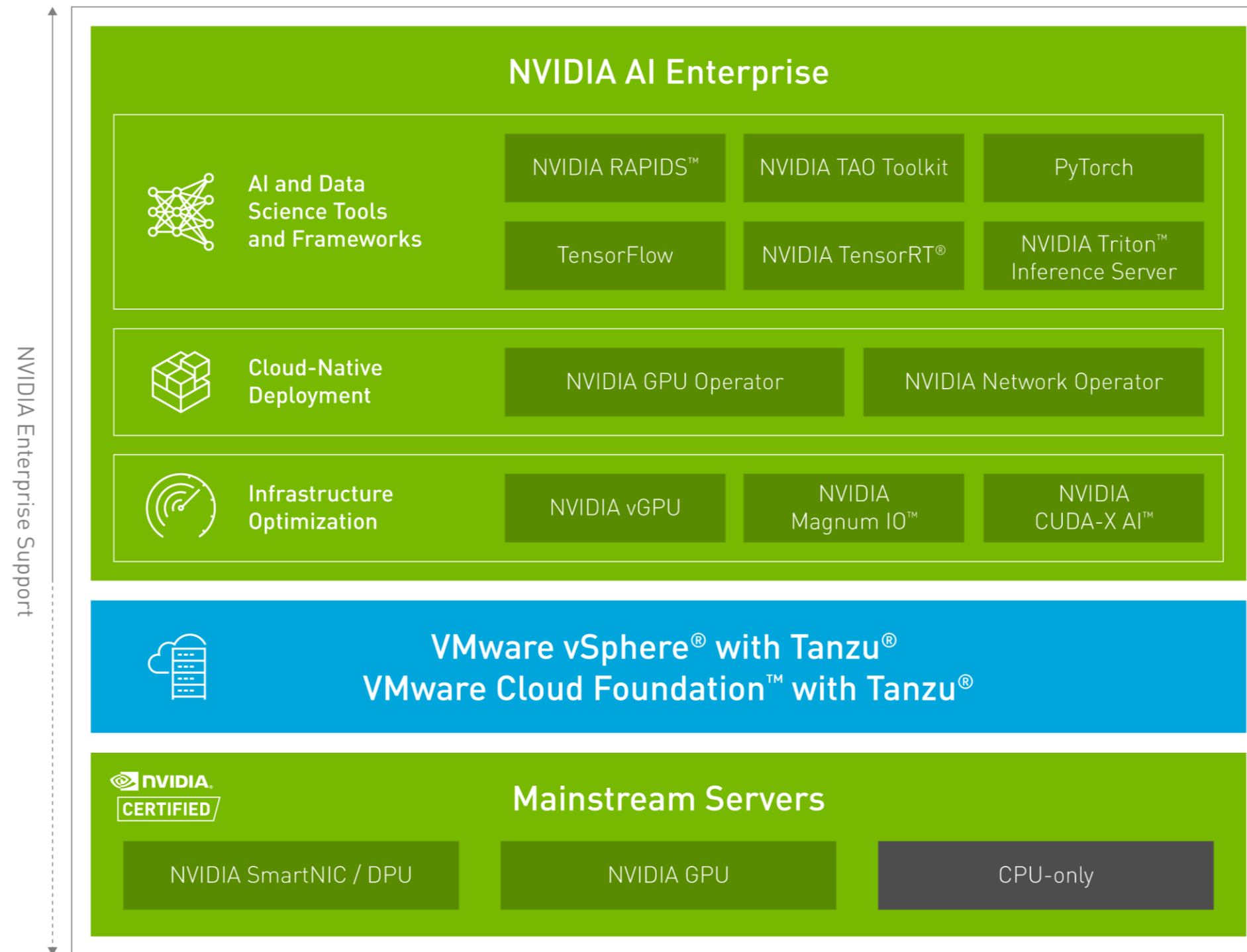
³For RHEL 8.4 and Ubuntu 20.04, no orchestration support needed

⁴ With VMware vSphere

⁵Designated GPU instances on AWS, Google Cloud, Azure

NVIDIA AI ENTERPRISE 软件套件

在 VMware vSphere 和 VMware Cloud Foundation 上启用 AI 和数据分析




Optimized for
Performance

为大型、复杂的训练和机器学习工作负载提供近似裸机的性能和多节点支持


Certified for
VMware vSphere

使用经过 VMware 数据中心认证的一整套 NVIDIA AI 软件降低部署风险


NVIDIA Enterprise
Support

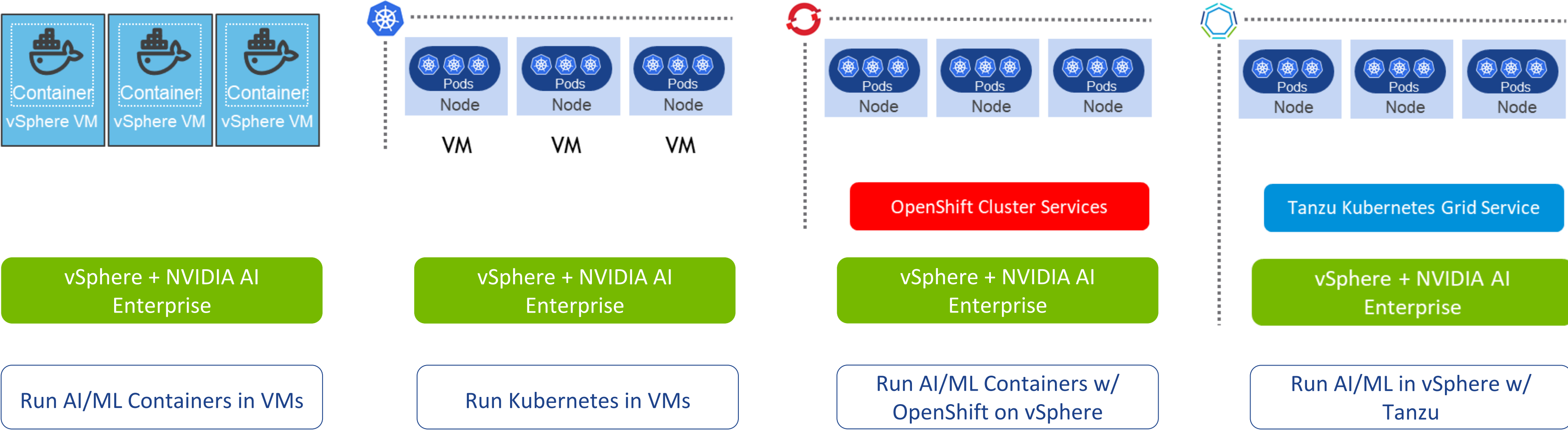
借助 NVIDIA 专家，确保 AI 关键项目按计划进行

*TensorFlow and PyTorch are integrated in the NVIDIA RAPIDS containers in NVIDIA AI Enterprise 1.1 and later

<https://www.nvidia.com/ai-enterprise-suite/>

使用 NVIDIA AI ENTERPRISE 交付 AI 工作负载

使用 VMware vSphere 进行编排的方法



NVIDIA认证系统

简化大规模加速计算的部署

SYSTEM DESIGN OPTIONS



NVIDIA DATA CENTER GPUs

A100 and A30 recommended
for training and inferencing workloads
with NVIDIA AI Enterprise



NVIDIA SMARTNICs AND DPUs



MAINSTREAM ACCELERATED SERVERS



Validates the Best Baseline
Configuration for



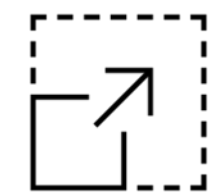
PERFORMANCE



MANAGEABILITY



SECURITY



SCALABILITY

List of [NVIDIA-Certified Systems that are NVIDIA AI Enterprise Compatible](#)

NVIDIA AI ENTERPRISE 支持和培训

确保企业级支持的开源透明度

Extend Your Team



Access to NVIDIA AI Experts

8-5 local business hours
Guidance on configuration + performance
Access to engineering

Stay Up to Date



Priority Notifications

Latest security fixes, maintenance
releases, coordinated support across
partners

Control Upgrade and Maintenance Schedule



Long Term Support

Up to 3 years for designated SW
branches

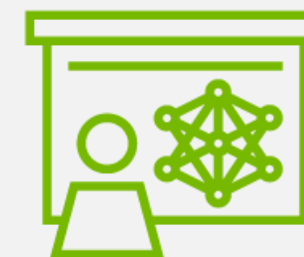
Customized Support



Mission Critical*

Designated Technical Account Manager
Business critical 24/7 live agent access

Upskill Your Workforce



Enterprise Training Services

Instructor-led workshops and self-paced
trainings

* Available as upgrade options.

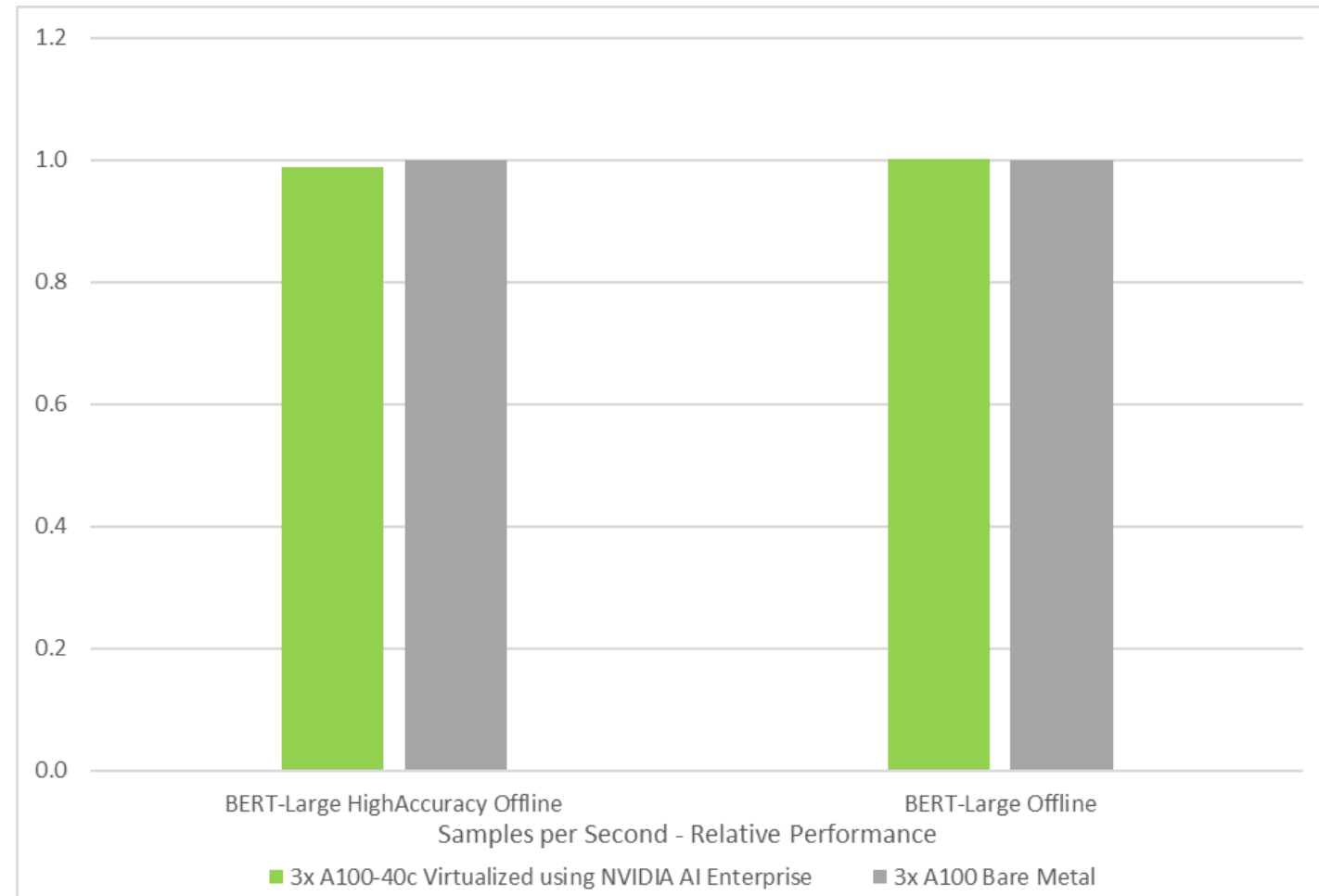


NVIDIA AI ENTERPRISE性能

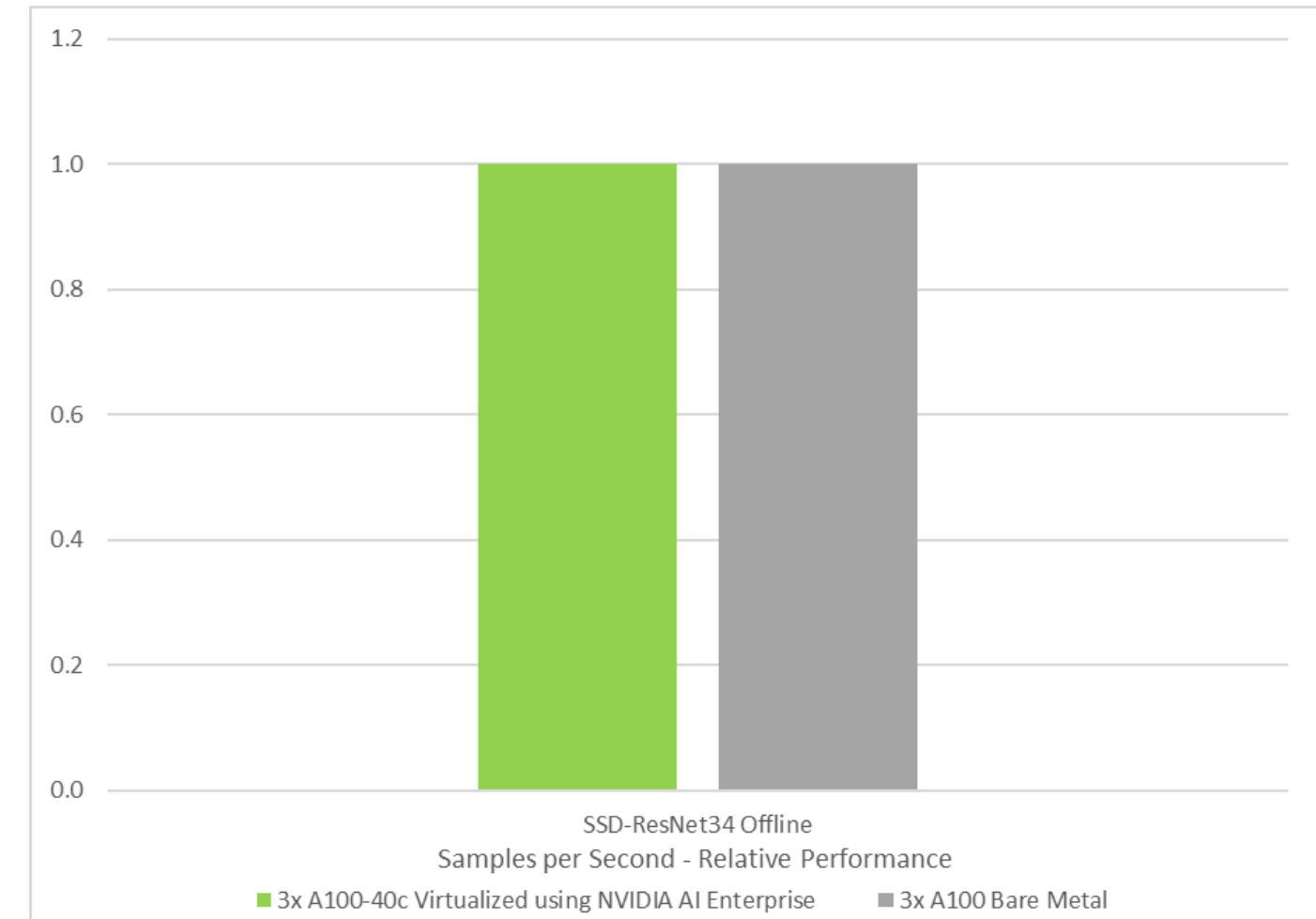
基于 VMWARE VSPHERE 的 NVIDIA AI ENTERPRISE与裸机性能对比

MLPerf 1.1 推理 Benchmark: Datacenter - Virtualized

Natural Language Processing (BERT-Large) Inference on NVIDIA A100



Object Detection (SSD-ResNet34) Inference on NVIDIA A100

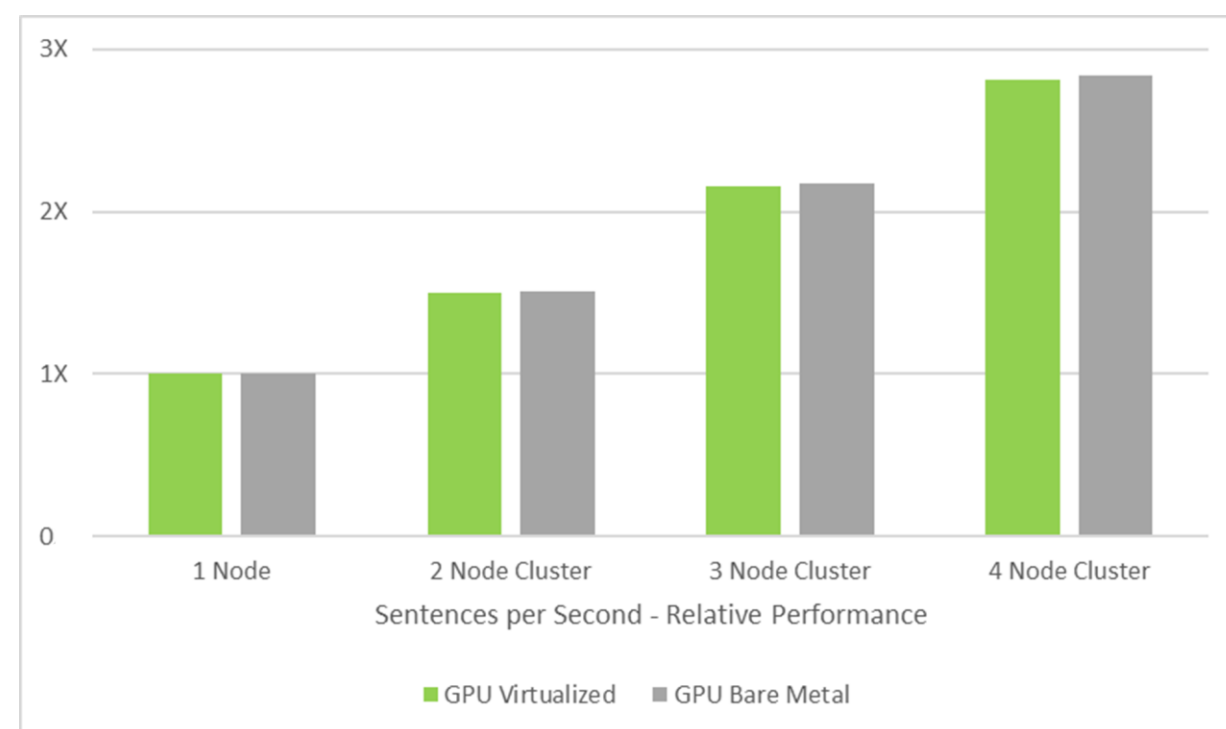


- Near Bare Metal performance
- Consistent performance across multiple MLPerf inference benchmarks
- NVIDIA AI Enterprise certified and optimized on VMware vSphere 7.0.2 with Dell EMC PowerEdge R7525 (NVIDIA-Certified)

对话式 AI 的裸机性能对比

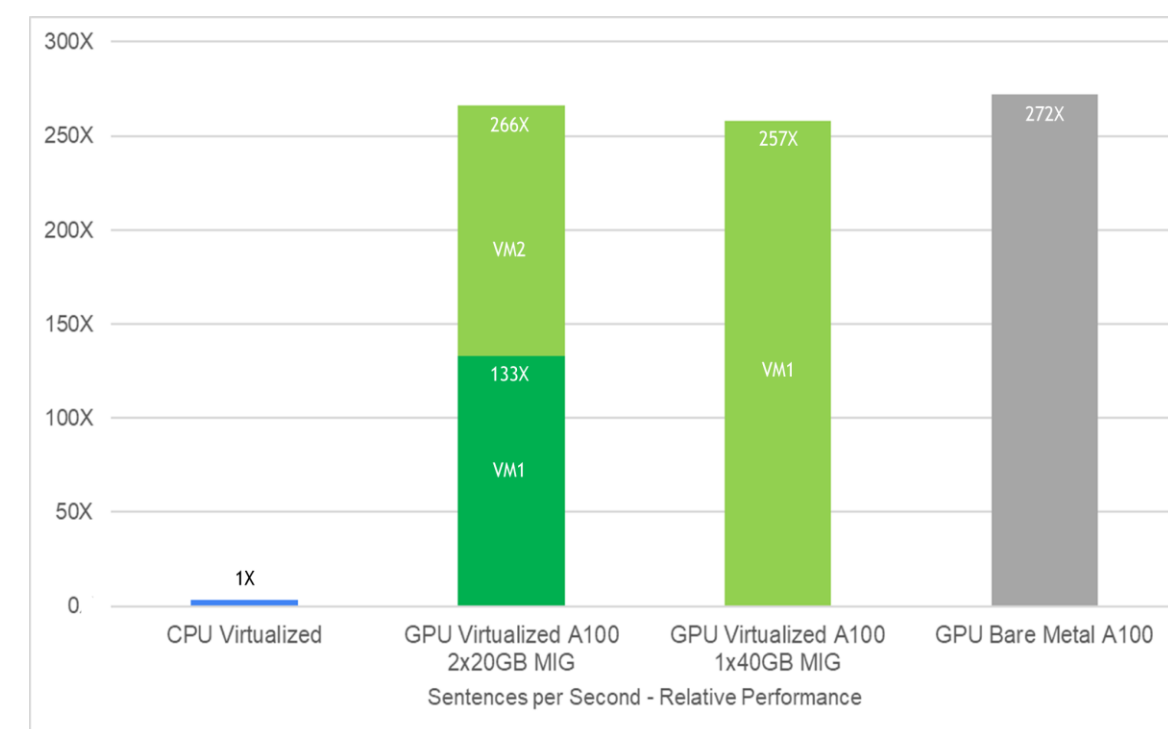
BERT 分布式深度学习训练和推理

NLP (BERT) Large Training on A100



- Multi-node scale out
- Bare Metal performance
- Train Natural Language Processing (NLP) models for Chatbots and Virtual Assistants in VMs

NLP (BERT) Inference on A100 w/ MIG



- Up to 266x Performance increase over CPU-only
- Near Bare Metal performance
- GPU Sharing without performance loss



成功案例

NTT COMMUNICATIONS 采用 NVIDIA AI ENTERPRISE 的可扩展 AI 就绪型基础架构

- Creates an AI-ready environment for research and development teams building NLP and intelligent video analytics applications
- Utilizes In-house infrastructure to support AI and reduce costs
- Large-scale development across multiple management platforms (VMware vSphere, Kubernetes and cloud infrastructure)
- Simplifies IT maintenance and management
- Expands instant access to NVIDIA GPUs for more AI developers



比萨大学借助 NVIDIA AI ENTERPRISE 简化 AI

- Simplified deployment and management of AI
- Increased system utilization and reduced costs
- Built flexibility into IT processes and services
- Extended NVIDIA GPUs to more users
- Hardware: Dell EMC PowerEdge servers, NVIDIA A100, NVIDIA BlueField DPU
- Software: NVIDIA AI Enterprise, VMware vSphere



ALLA GIORNATA

支付服务公司选择 NVIDIA AI ENTERPRISE

- Leveraging AI for Natural Language Processing (NLP) chatbot application
- Required VMware virtualized infrastructure for simplified management
- Achieved significant value from NVIDIA AI frameworks & tools - including 20x performance boost with Triton Inference Server
- Hardware: Dell EMC PowerEdge R740, NVIDIA V100
- Software: NVIDIA AI Enterprise, VMware vSphere





NVIDIA AI ENTERPRISE 助力医疗保健组织轻松采用 AI

- New to AI workloads, leveraging AI for data science with EMR data
- Included AI frameworks and tools simplified AI adoption
- Required an AI platform that integrated with existing VMware infrastructure
- Hardware: HPE ProLiant DL385, NVIDIA A100
- Software: NVIDIA AI Enterprise, VMware vSphere



