NVIDIA AI for Live Media

Thomas True, Senior Applied Engineer for Professional Video
SMPTE Montreal AI Workshop

SMPTE_Montreal_AI_Workshop.pptxSMPTE_Montreal_AI_Workshop.pptx

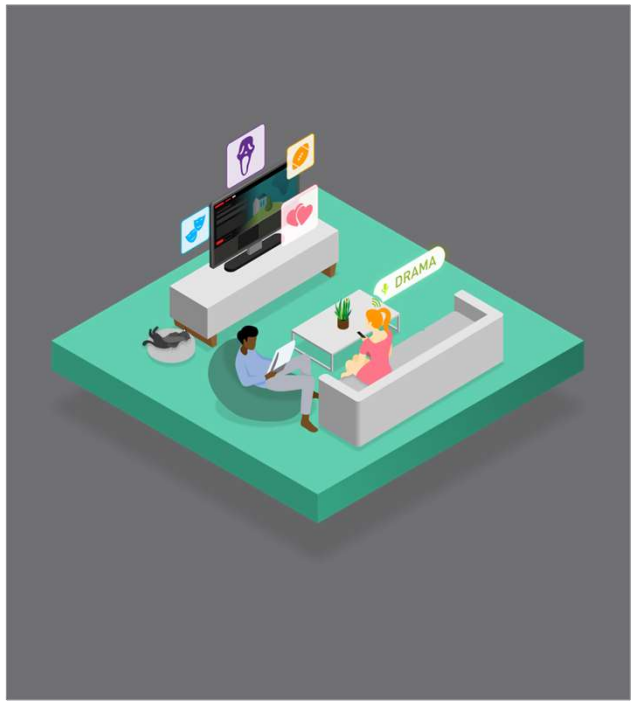
Agenda

- AI Opportunities in Live Media
- Generative AI
- Predictive AI
- Video AI
- Holoscan for Media

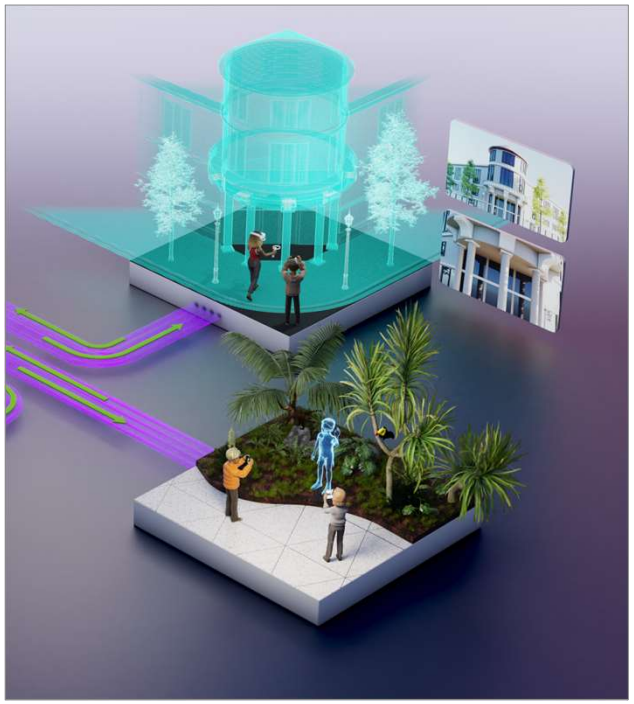
AI Opportunities

Live Media Opportunities

Engaging and retaining audiences with new viewer experiences



Personalized

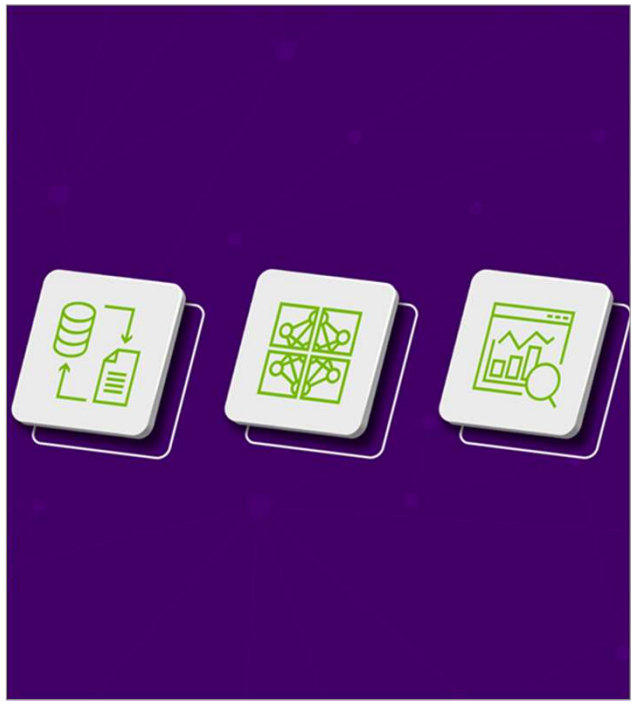


Immersive

Interactive

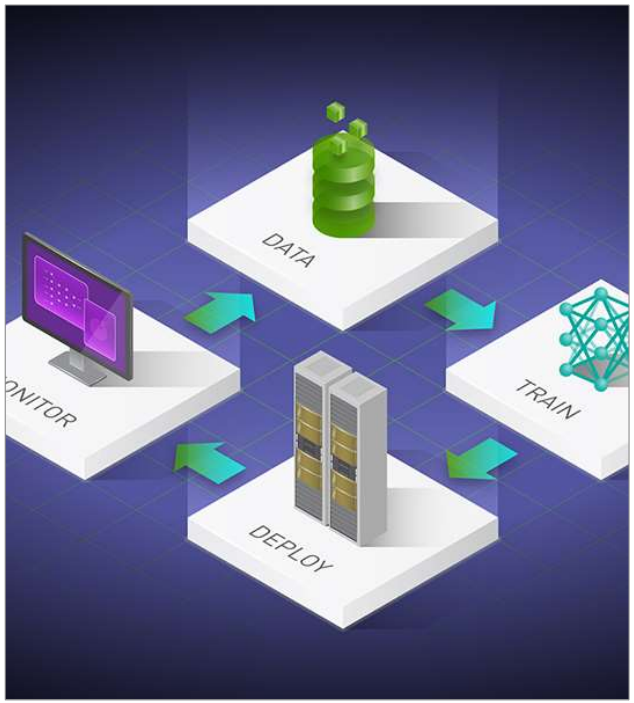
NVIDIA AI for Live Media

Models, foundries, and infrastructure enabling core AI competencies



Generative AI

Enhanced productivity through GenAI tools



Predictive AI

Analytics and recommenders

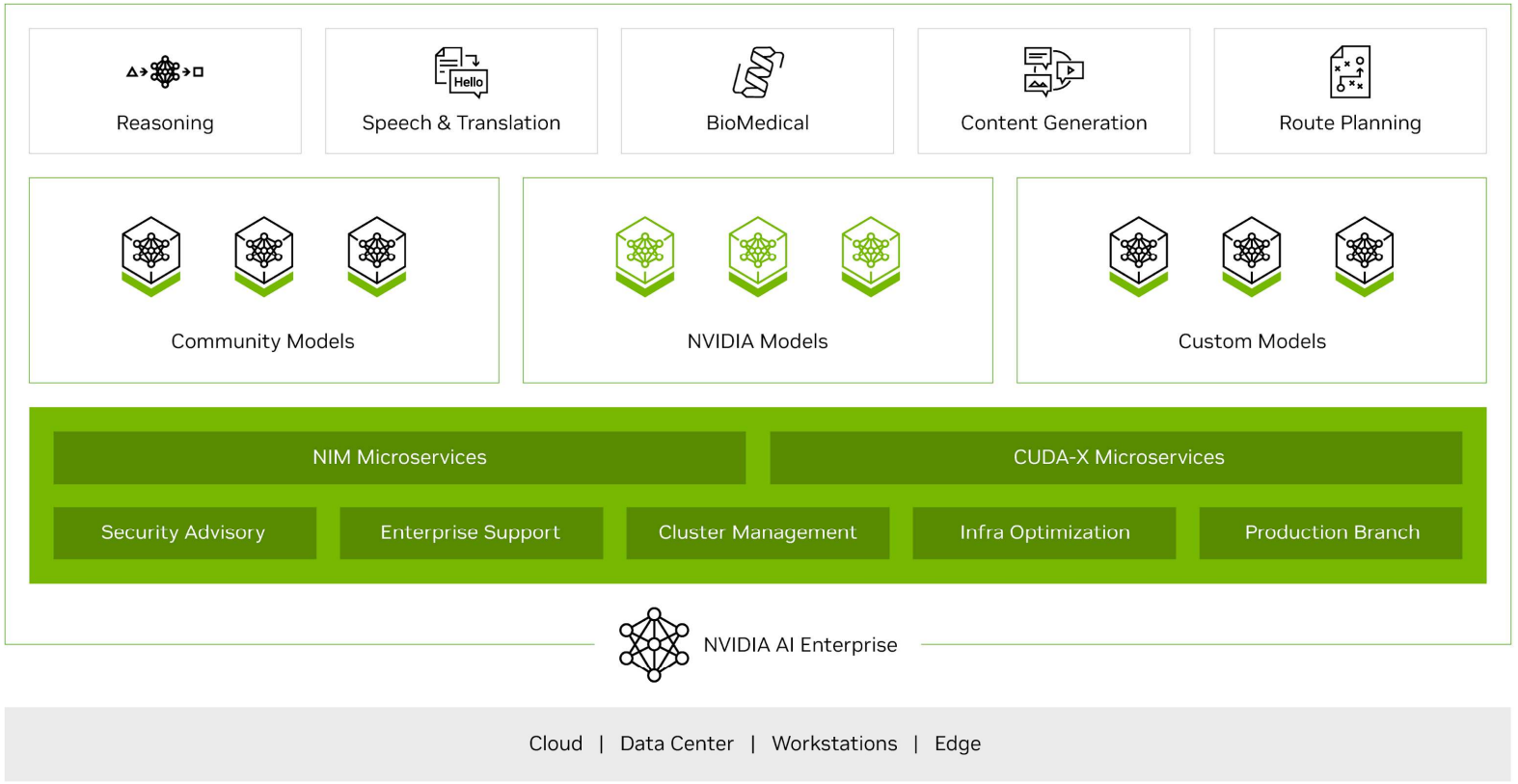
Video AI

Upscaling, image recognition, subtitling and translation

Generative AI Solutions

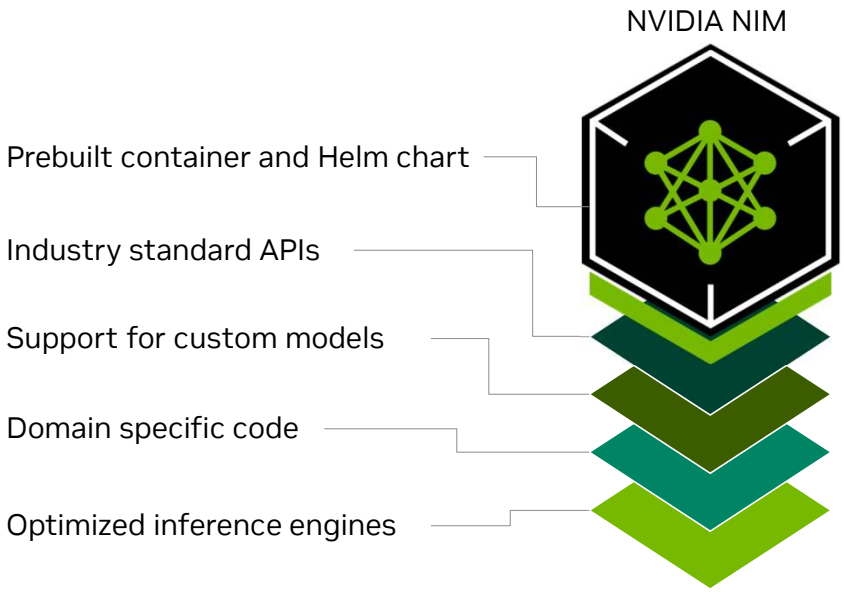
NVIDIA AI Enterprise

High performance and efficient runtime for generative AI



Inference Microservices for Generative AI

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

Microsoft Azure · aws · Google Cloud · ORACLE · DGX & DGX Cloud · DELL Technologies · Hewlett Packard Enterprise · Lenovo · SUPERMICRO

NVIDIA

Inference Microservices for Generative AI

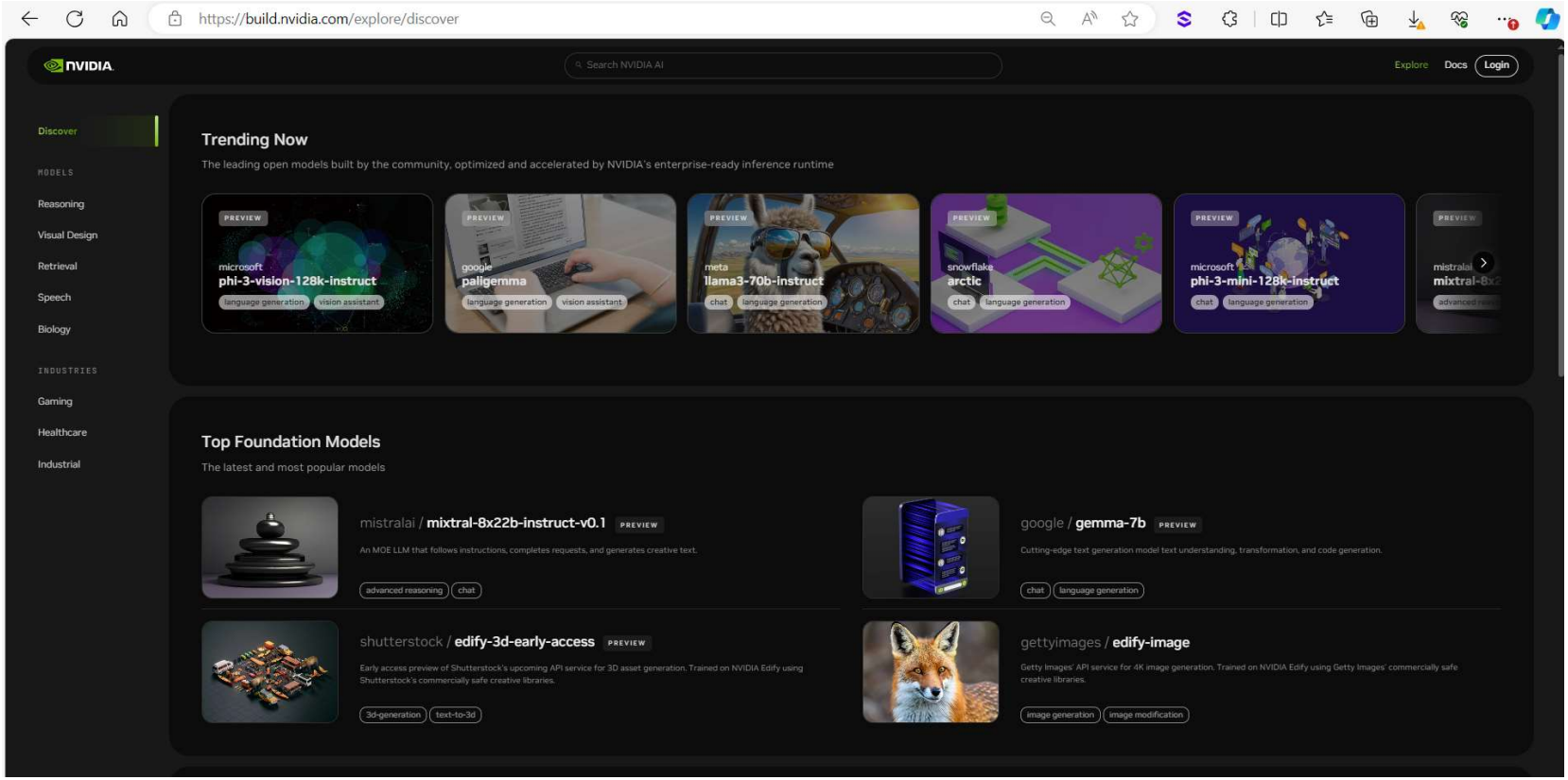
NVIDIA NIM is the fastest way to deploy AI models on accelerated infrastructure across cloud, data center, and PC

NVIDIA API Catalog

Microsoft Azure · aws · Google Cloud · ORACLE · DELL Technologies · Hewlett Packard Enterprise · Lenovo · SUPERMICRO

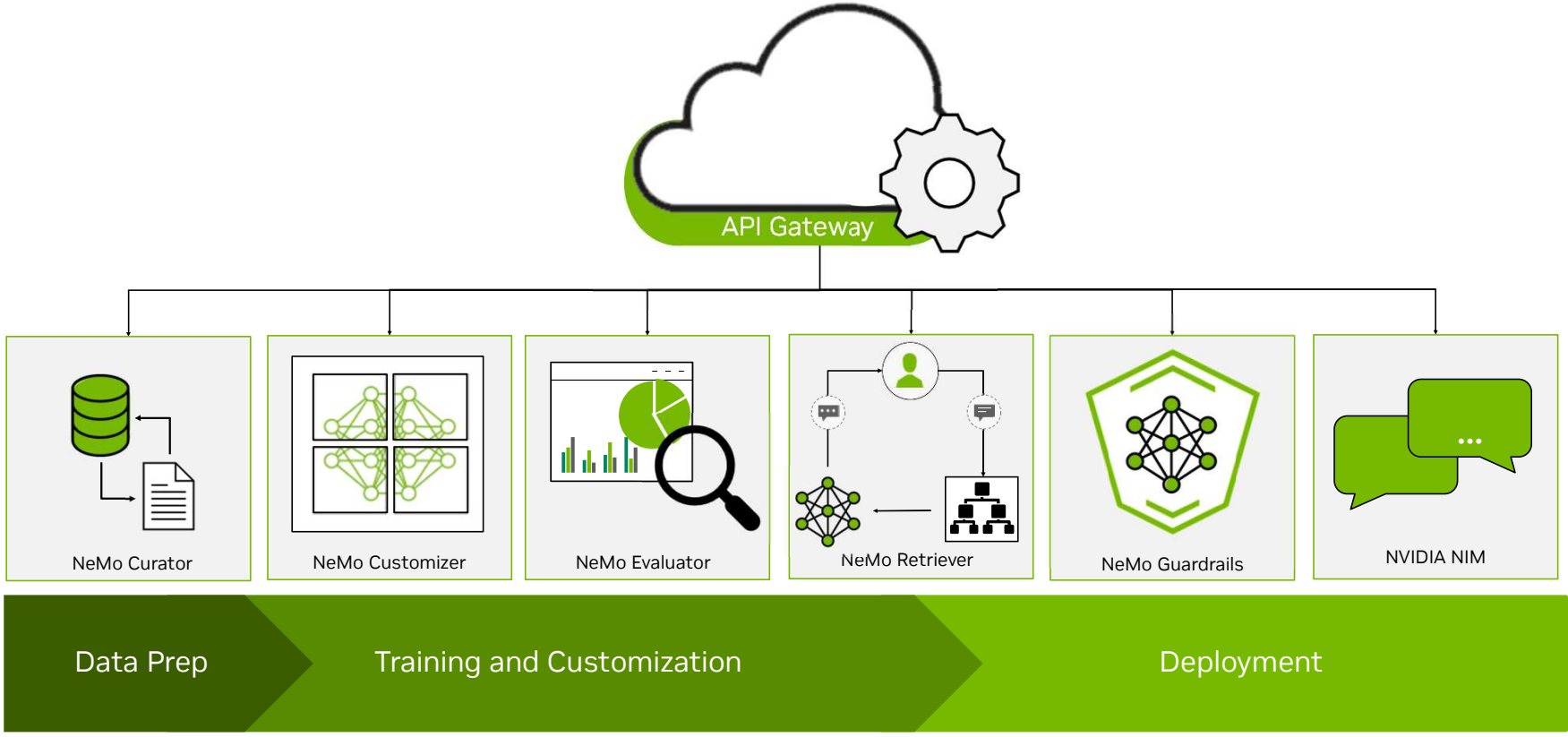
How to Get Started

Tuned Foundation Models



Building Generative AI Applications

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

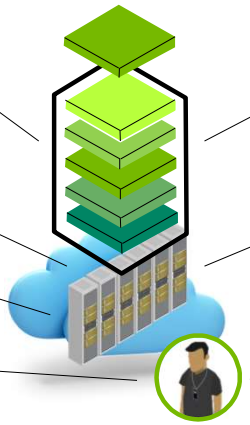


Microsoft Azure · aws · Google Cloud · ORACLE · NVIDIA DGX Cloud · DELL Technologies · Hewlett Packard Enterprise · Lenovo · SUPERMICRO · NVIDIA

NVIDIA DGX Cloud

AI Training-as-a-Service Platform for the era of generative AI

NVIDIA DGX Cloud



NVIDIA AI Enterprise
Base Command Platform

DGX Cloud Instance
(8) A100 80GB GPUs
with multi-node scale

10TB storage / instance
10TB egress

NVIDIA AI Expertise

One price
with no surprises

Organizations doubling-down on generative AI need:

Software that unleashes developer productivity

Multi-node training performance + easy scale

Access AI practitioners who help maximize performance

Predictable pricing that includes everything

Microsoft Azure · Oracle Cloud Infrastructure · Google Cloud

Hosted in **leading clouds**

NVIDIA

NVIDIA-Certified Systems

Simplifies deployment of generative AI at scale

SYSTEM CERTIFICATION CATEGORIES

Data Center

Reference Architectures
NVIDIA OVX
Spectrum-X Networking

Certified Servers
Multi-Node Compute & Graphics
High-Density VDI

DESKTOP & MOBILE WORKSTATIONS

INDUSTRIAL & ENTERPRISE EDGE

NVIDIA Certified

Validates the Best System Configuration for

- PERFORMANCE
- MANAGEABILITY
- SECURITY
- SCALABILITY

NVIDIA-Certified

Example data center systems

DELL Technologies

Dell PowerEdge R760xa

Hewlett Packard Enterprise

HPE ProLiant DL380a Gen 11

Lenovo

Lenovo ThinkSystem SR675 V3

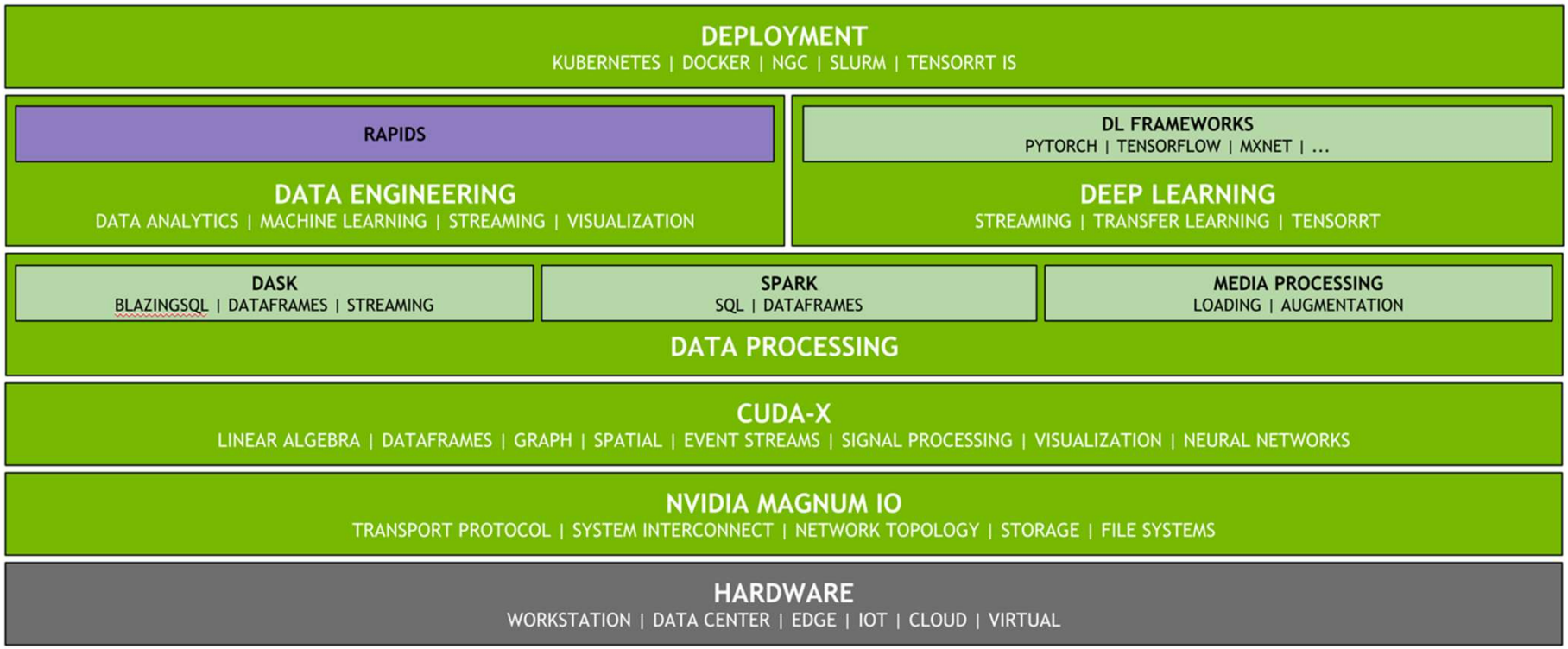
NVIDIA Certified

Systems Catalog

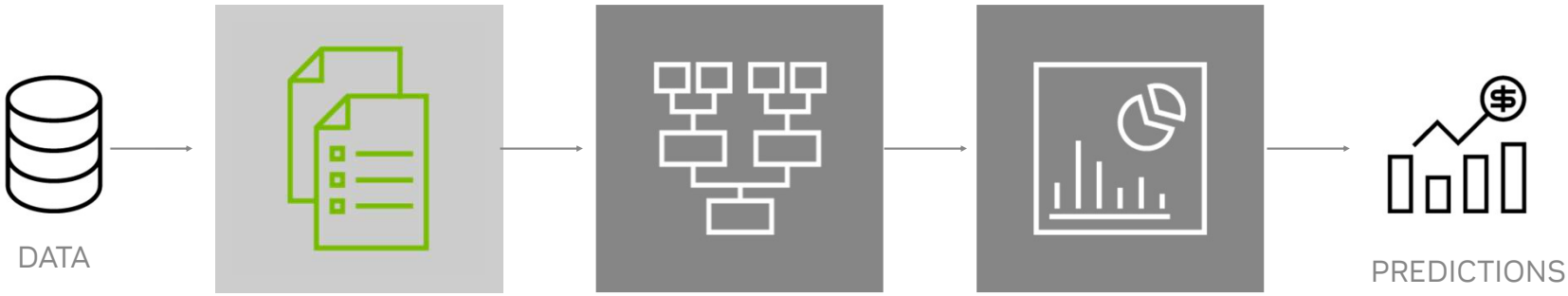
Predictive AI Solutions

NVIDIA RAPIDS Transforms Data Science

From analytics to NVIDIA **accelerated** data science



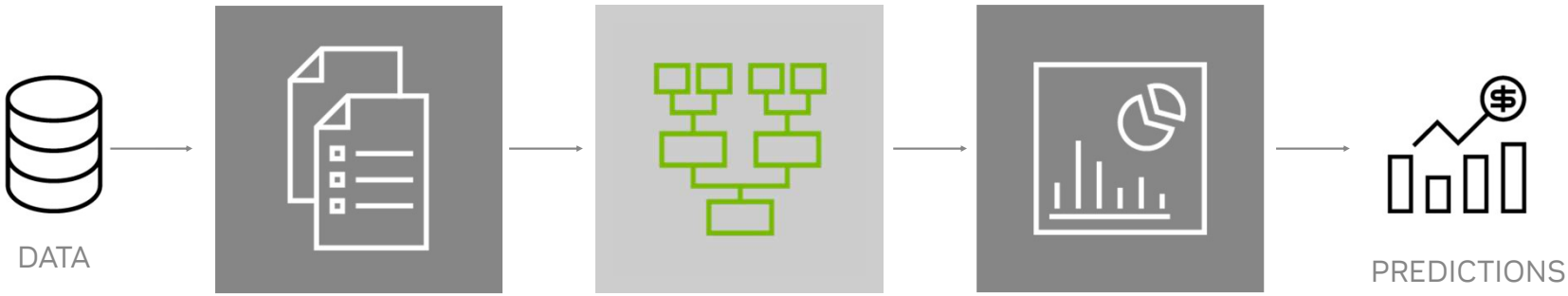
GPU-Accelerated Data Science Workflow With RAPIDS



Data Preparation

- GPUs accelerated compute for in-memory data preparation
- Simplified implementation using familiar data science tools
- RAPIDS is a Pythonic drop-in replacement for pandas built on CUDA C++
- GPU-accelerated Spark

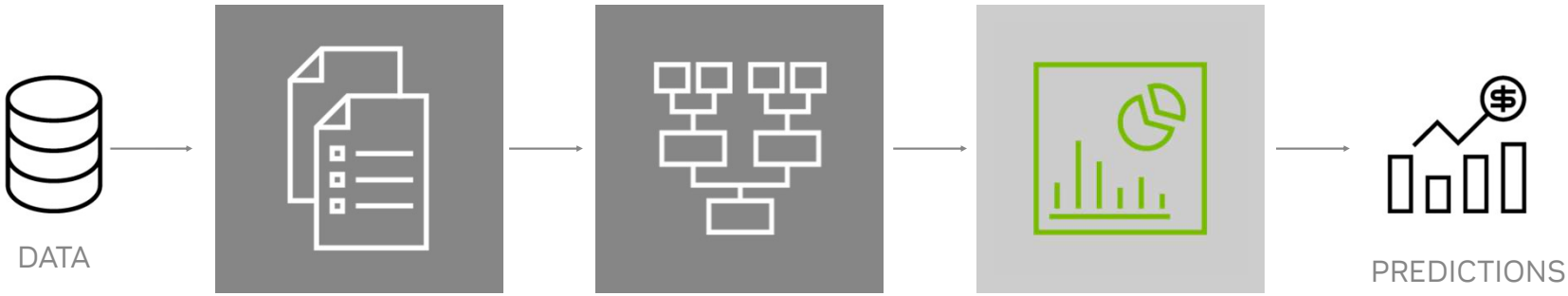
GPU-Accelerated Data Science Workflow With RAPIDS



Model Training

- GPU-acceleration of today's most popular ML algorithms
- **XGBoost and Random Forest**
- RAPIDS accelerated ML algorithms like UMAP, PCA, K-means, k-NN, DBScan, tSVD, and more
- Easy-to-adopt, scikit-learn like interface

GPU-Accelerated Data Science Workflow With RAPIDS



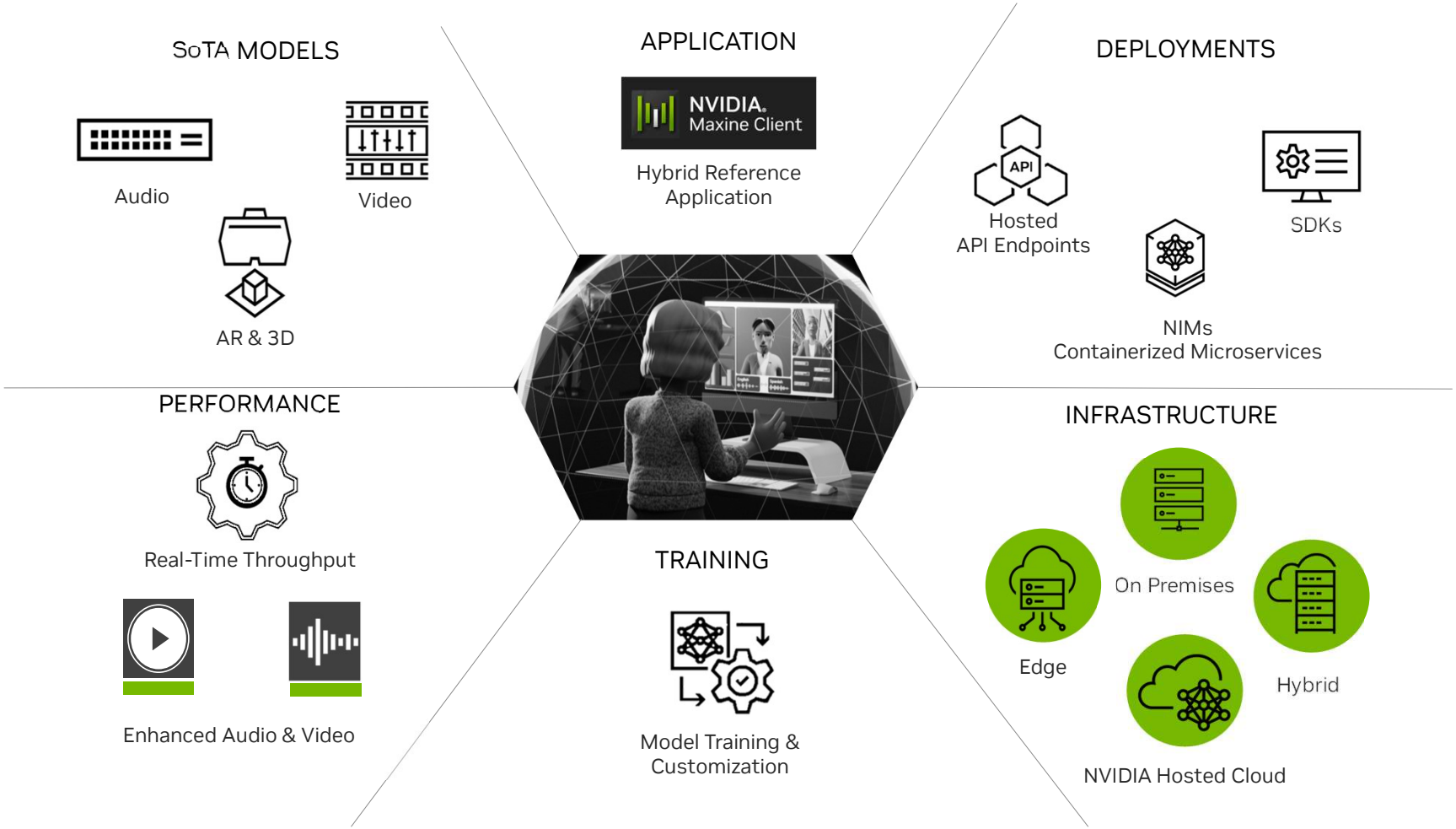
Visualization

- Effortless exploration of datasets, billions of records in milliseconds
- Dynamic interaction with data = faster ML model development
- Integration with data visualization ecosystem leaders like Bokeh, Datashader, Plotly, and more

Video AI Solutions

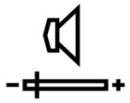
Maxine AI Developer Platform

Customized real-time models integrated in your application



Maxine Features

NVIDIA Maxine AI Models Deliver Breakthrough Audio and Video Quality, and Unique AR Effects



Audio Effects

- Room Echo Cancellation
- Acoustic Echo Cancellation
- Audio Super Resolution
- Speaker Focus
- Background Noise Removal
- *Voice Font
- Studio Voice

Video Effects

- Artifact Reduction
- Video Noise Removal
- Video Upscaler
- Video Super Resolution
- Virtual Background
- Background Blur
- Video Relighting

Augmented Reality & 3D

- Eye Contact
- Face Mesh
- Face Tracking
- Face Landmark Tracking
- Face Expression Estimation
- 3D Body Pose Estimation
- Video Live Portrait
- Speech Live Portrait
- Maxine 3D

Green indicates Early Access
Audio and Video effects are HELM-deployable.

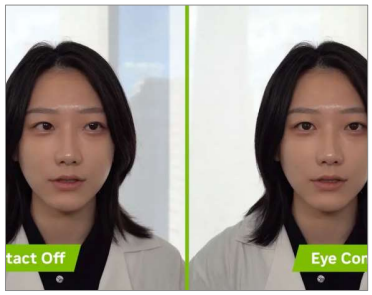
Virtual Fan Engagement

agement Experien

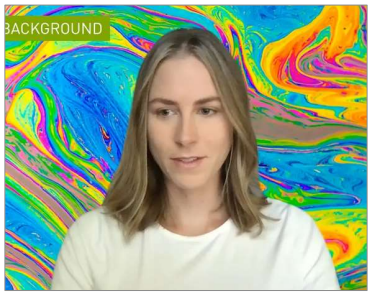
NVIDIA AI NIMs & SDKs for Media

Ready to deploy in your application

Audio Effects SDK
Enhance audio quality

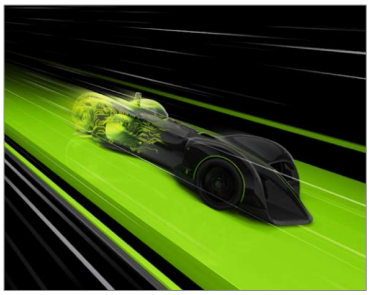


Augmented Reality SDK
Real-time 3D tracking

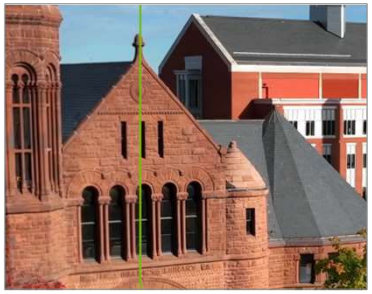


Video Effects SDK
AI-powered visual effects

Audio2Face SDK
Animates 3D character to match voice track

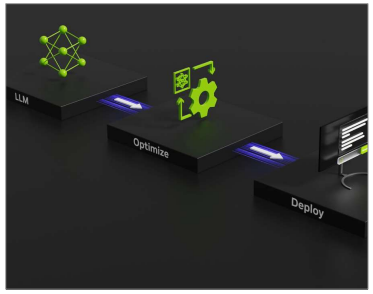


NVIDIA DLSS
Enhance image quality and performance



RTX Video SDK
Upscale and color remap video

NVIDIA ACE
Bring digital avatars to life



TensorRT-LLM API
Optimize LLM performance

Stable Diffusion
Get fastest SD model performance

NVIDIA RTX Video

AI video enhancements for real time playback or offline video processing

RTX Video Super Resolution
Upscale and remove compression artifacts

RTX Video HDR
AI powered real time SDR to HDR tone mapping

Get started on https://developer.nvidia.com/

Slide 25

GD0 This is not an HDR to SDR, it's a super res. [@Brian Choi] do you have a better example?

Gerardo Delgado Cabrera, 2024-05-08T06:54:30.645

BC0 0 Previous image was VSR + HDR. New image is pure HDR

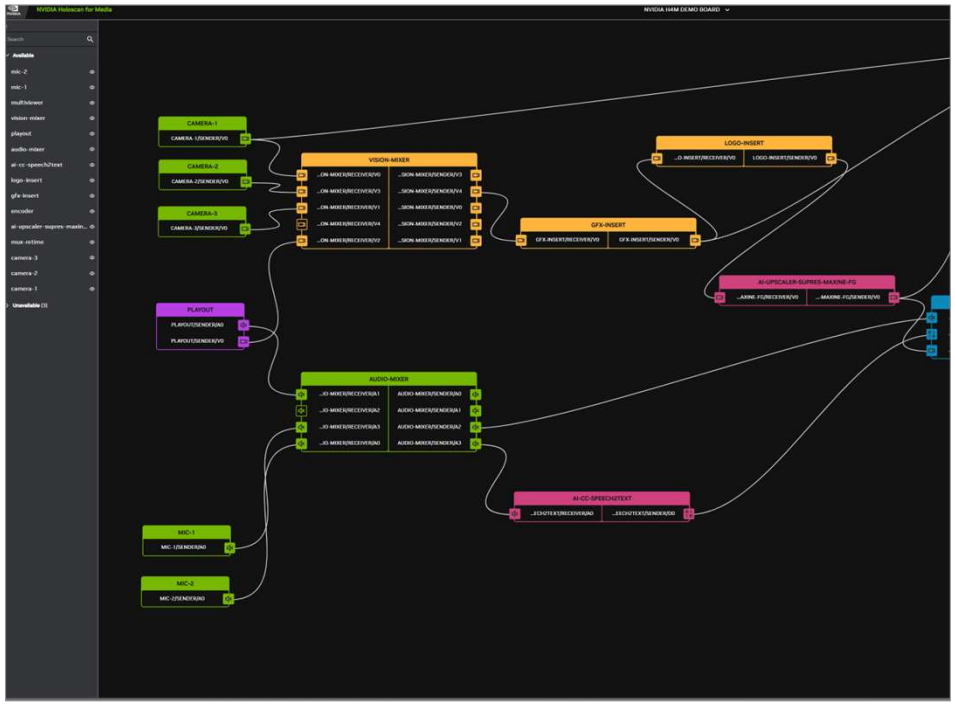
Brian Choi, 2024-05-08T17:36:32.255

Holoscan for Media

Transitioning to Software-Defined

Holoscan for media

Traditional appliances with applications tied to hardware

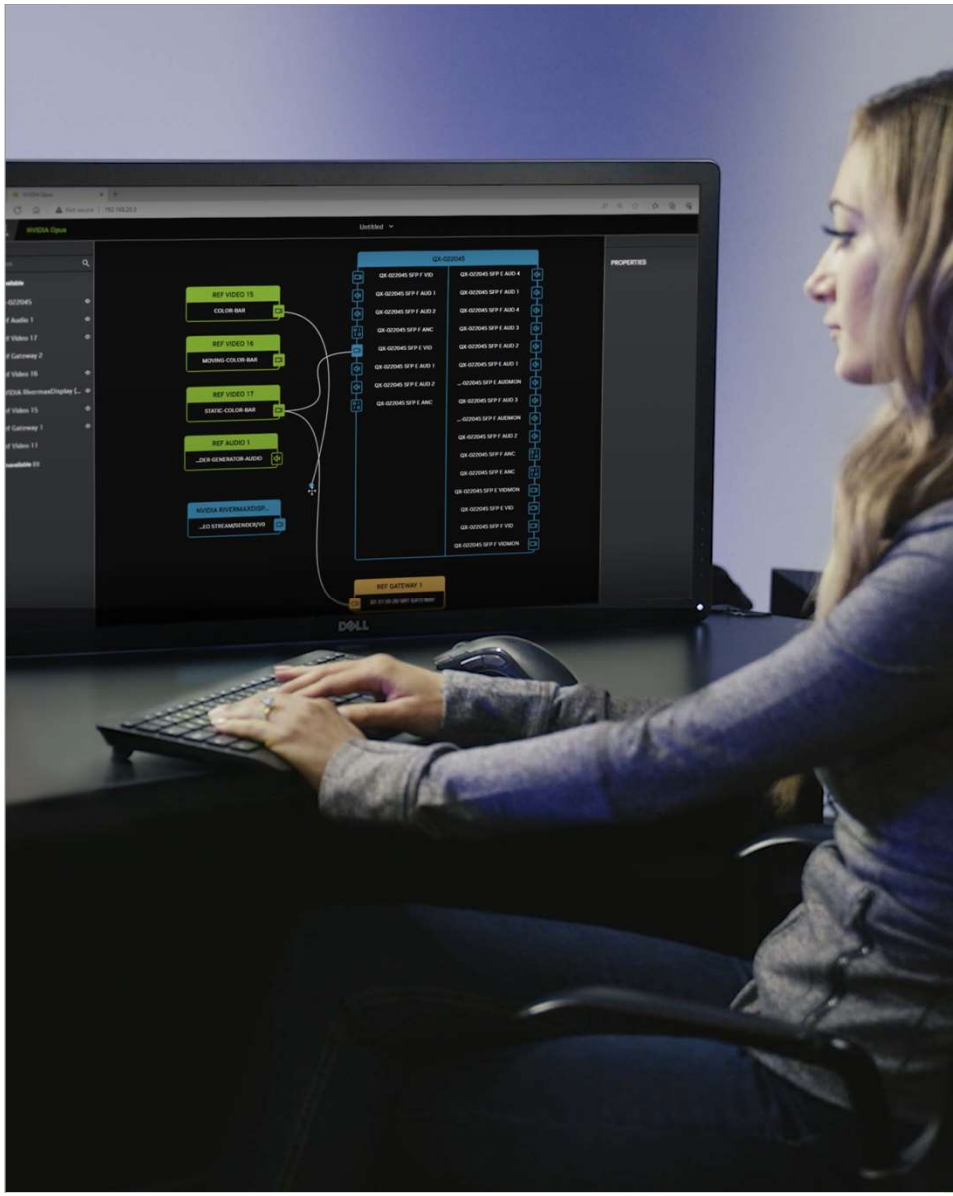


Applications as software running on COTS

What it is

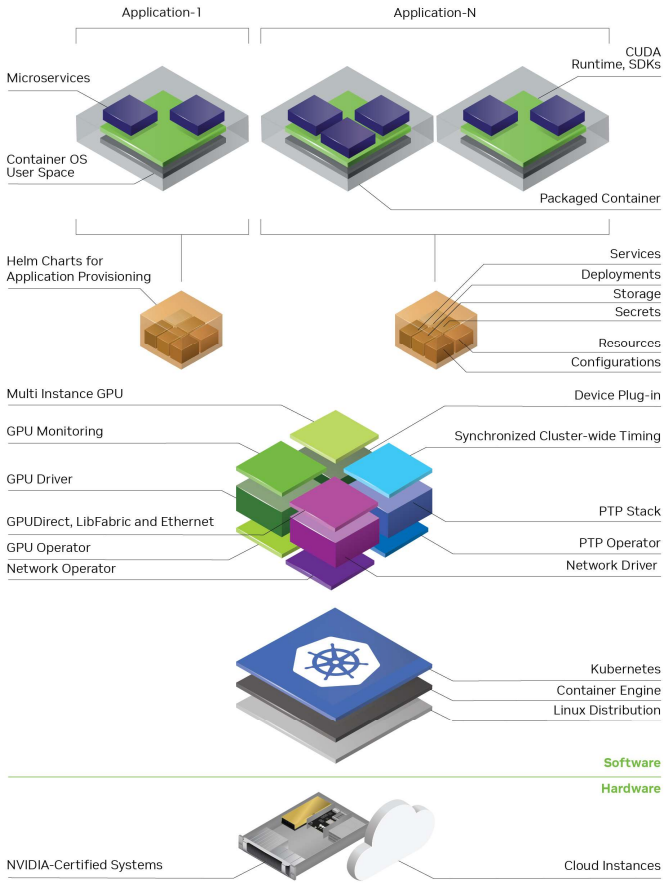
Holoscan for media

- Software-defined platform for building and deploying applications for live media.
- The same architecture can run in the cloud, on premises and at the edge.
- It’s not tied to a specific device, FPGA, or solution.
- It integrates open-source and ubiquitous technologies.
- It is an IP-based solution built on industry standards and APIs.
- Its architecture includes services like authentication, logging, security, PTP timing, NMOS controller and registry.
- It allows for easy integration of NVIDIA AI SDKs for application development.



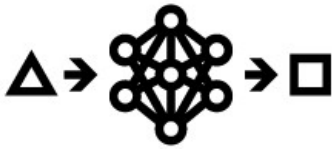
Platform Architecture

Holoscan for media

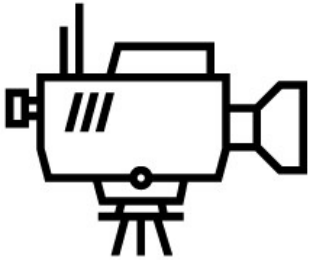


Benefits of Holoscan for Media

Software defined media application running on AI cluster infrastructure



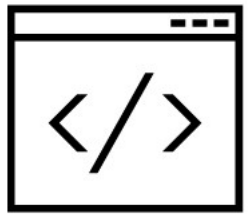
Gen AI Ready
Native AI Cluster Infrastructure



IP-Based Infrastructure
ST 2110, NMOS, SRT, NDI

IT Managed
Kubernetes with centralized PTP, Helm Charts

Develop Once, Deploy Everywhere
Container based software



Open Platform
Based on Open-Source technologies

Partner Ecosystem

Holoscan for Media

Thank You!