NVIDIA

NVIDIA Holoscan for Media

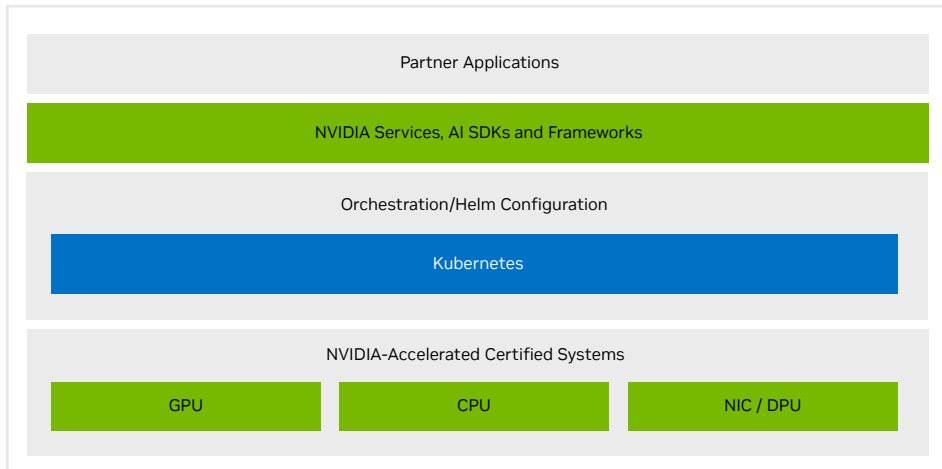A software-defined, AI-enabled platform for live media.

Companies in broadcast, streaming, and live sports are transforming content production and delivery as well as their operations by transitioning from appliance-based to software-defined infrastructure. Using a software-defined approach lets you use resources in a way that's easy to deploy and grow. It also lets you use the latest technologies for video, predictive, and generative AI.

The Platform for the Future of Live Media

NVIDIA Holoscan for Media is an AI-enabled, software-defined platform for live media. Core elements of this platform are that:

> It enables live video pipelines to run on the same infrastructure as AI.

> Software applications in its expanding catalog are built by established and emerging vendors within the live media ecosystem.

> Applications run on NVIDIA accelerated commercial-off-the-shelf hardware on premises—with cloud and edge coming soon.

> It's an IP-based solution built on industry standards and APIs such as ST 2110, and inclusive of essential services like PTP for timing and NMOS for interoperability and manageability.

Partner Applications
NVIDIA Services, AI SDKs and Frameworks
Orchestration/Helm Configuration
Kubernetes
NVIDIA-Accelerated Certified Systems

GPU	CPU	NIC / DPU

Holoscan for Media Platform

Benefits of Holoscan for Media

> **AI-Enabled:** Live video pipelines run on the same software and hardware infrastructure that's used to deploy AI at scale.

> **Repurposable:** Applications from a variety of vendors can be deployed on the same hardware, meaning the hardware can be repurposed for multiple use cases, including as backup. This reduces the infrastructure footprint and its associated costs.

> **Flexible:** Applications can be dynamically connected to media streams and to each other to create any desired workflow. They can also be turned on and off as needed. This delivers flexibility.

> **Agile:** Infrastructure resources can be allocated where and when they are needed and applied to any use case required through GPU partitioning. Scaling out resources is as easy as adding new server nodes.

> **Resilient:** The platform allows users to recover automatically through network redundancy, failover, and high availability (HA) cluster support.

> **Upgradeable:** Software and hardware upgrades are independent of each other. This makes the platform and its applications easily upgradeable.

Platform Components

Holoscan for Media is a platform comprising key technologies from NVIDIA and our ecosystem of partners.

Technologies	Description
NVIDIA OVX with L40S GPUs	**NVIDIA OVX™ with L40S GPUs** is a transformative data center–scale computing system reference architecture. It's optimized to deliver breakthrough performance for building and deploying generative AI with NVIDIA AI Enterprise. NVIDIA OVX with L40S GPUs server and cluster configurations deliver powerful graphics and network performance at scale and a universal data center platform to accelerate a broad range of use cases. These include generative AI use cases like RAG, large-language model (LLM) inference and fine-tuning, to visual computing use cases like graphics, rendering, and uncompressed media streaming. The NVIDIA OVX with L40S GPUs foundation is a performance-optimized server configuration that combines the NVIDIA RTX™-accelerated AI and graphics performance of L40S GPUs with ultra-high-bandwidth, low-latency ConnectX-7 and BlueField-3 networking technologies. Together, they deliver performance at scale for computationally intensive AI applications and demanding visual computing workloads.
NVIDIA AI Enterprise	**NVIDIA AI Enterprise** is the NVIDIA AI platform's operating system. It's important for making and supporting applications that use the NVIDIA library of frameworks and models that are trained before being used. These include NVIDIA® Riva for speech AI; NVIDIA Merlin™ for recommendation engines; NVIDIA Maxine™ for audio, video, and augmented reality AI effects; and more. NVIDIA AI Enterprise is certified to deploy on broadly adopted enterprise platforms—including multi-cloud instances as well as as NVIDIA-Certified Systems™ from leading workstation and server vendors.
NVIDIA Rivermax	Key to SMPTE ST 2110 adoption, **NVIDIA® Rivermax® SDK** increases GPU processing capacity with direct data transfers to and from the GPU. It reduces CPU load by 65%, works on any CPU architecture, and is OS agnostic. It's a fully virtualized streaming solution with natively integrated PTP time service and NMOS Node. It delivers hardware time scheduling and accurate bitrate for any data stream and hardware-enabled SMPTE ST 2110-21 specification compliance.

> **Efficient:** Moving to IT-oriented software-defined infrastructure means that users can benefit from the cyclical cost reductions that IT brings. Over the lifecycle of the infrastructure, this will lower their total cost of ownership.

> **Legacy support:** The platform is built on standards like ST 2110 and includes PTP as a service. This means it can work with SDI gateways, allowing for gradual migration to IP.

Technologies	Description
NVIDIA-Certified Systems	**NVIDIA-Certified servers** create the essential platform for the evolution of enterprise data centers, delivering infrastructure that can handle a diverse range of accelerated workloads. The certification process employs the performance and functionality of a wide range of real-world applications in AI, data science, visualization, and high-performance computing (HPC).
Red Hat OpenShift	**Red Hat® OpenShift®** is the industry-leading Kubernetes-powered hybrid cloud platform that accelerates secure application development and delivery—in the public cloud, on premise, or at the edge. OpenShift provides high availability through automated failover and replication, ensuring that broadcasting services remain online—even in the event of hardware failures or other issues. It combines built-in security features with dedicated Enterprise support, a trusted software supply chain, and self-service provisioning, allowing teams to work together more efficiently. Broadcasting companies often deal with varying workloads, especially during live events or peak viewing times. OpenShift's ability to dynamically scale applications ensures that the infrastructure can handle sudden spikes in demand without compromising performance. OpenShift offers a complete set of integrated tools and services for cloud-native, AI, and traditional workloads alike.
Partner Applications	Applications from established and emerging vendors run on top of the platform and facilitate live video workflows. See our app catalog for a full list of available partner applications.

Ready to Get Started?

Take advantage of flexible deployment, resource scalability, and the latest video, predictive, and generative AI technologies by transitioning true software-defined infrastructure with Holoscan for Media. To learn more, visit: nvidia.com/holoscan-for-media

Partner Logo