Relevant Technologies

Table of contents

Holoscan accelerates streaming AI applications by leveraging both hardware and software. The Holoscan SDK relies on multiple core technologies to achieve low latency and high throughput:

- Rivermax and GPUDirect RDMA
- Graph Execution Framework
- TensorRT Optimized Inference
- Interoperability between CUDA and rendering frameworks
- Accelerated Image Transformations
- Unified Communications X

Rivermax and GPUDirect RDMA

The NVIDIA Developer Kits equipped with a ConnectX network adapter can be used along with the NVIDIA Rivermax SDK to provide an extremely efficient network connection that is further optimized for GPU workloads by using GPUDirect for RDMA. This technology avoids unnecessary memory copies and CPU overhead by copying data directly to or from pinned GPU memory, and supports both the integrated GPU or the discrete GPU.

> **Note**
>
> NVIDIA is also committed to supporting hardware vendors enable RDMA within their own drivers, an example of which is provided by the AJA Video Systems as part of a partnership with NVIDIA for the Holoscan SDK. The `AJASource` operator is an example of how the SDK can leverage RDMA.

For more information about GPUDirect RDMA, see the following:

- GPUDirect RDMA Documentation

- Minimal GPUDirect RDMA Demonstration source code, which provides a real hardware example of using RDMA and includes both kernel drivers and userspace applications for the RHS Research PicoEVB and HiTech Global HTG-K800 FPGA boards.

Graph Execution Framework

The Graph Execution Framework (GXF) is a core component of the Holoscan SDK that provides features to execute pipelines of various independent tasks with high performance by minimizing or removing the need to copy data across each block of work, and providing ways to optimize memory allocation.

GXF will be mentioned in many places across this user guide, including a dedicated section which provides more details.

TensorRT Optimized Inference

NVIDIA TensorRT is a deep learning inference framework based on CUDA that provided the highest optimizations to run on NVIDIA GPUs, including the NVIDIA Developer Kits.

The inference module leverages TensorRT among other backends, and provides the ability to execute multiple inferences in parallel.

Interoperability between CUDA and rendering frameworks

Vulkan is commonly used for realtime visualization and, like CUDA, is executed on the GPU. This provides an opportunity for efficient sharing of resources between CUDA and this rendering framework.

The Holoviz module uses the external resource interoperability functions of the low-level CUDA driver application programming interface, the Vulkan external memory and external semaphore extensions.

Accelerated Image Transformations

Streaming image processing often requires common 2D operations like resizing, converting bit widths, and changing color formats. NVIDIA has built the CUDA accelerated NVIDIA Performance Primitive Library (NPP) that can help with many of these common

transformations. NPP is extensively showcased in the Format Converter operator of the Holoscan SDK.

Unified Communications X

The Unified Communications X (UCX) framework is an open-source communication framework developed as a collaboration between industry and academia. It provides high performance point-to-point communication for data-centric applications. Holoscan SDK uses UCX to send data between fragments in distributed applications. UCX's high level protocols attempt to automatically select an optimal transport layer depending on the hardware available. For example technologies such as TCP, CUDA memory copy, CUDA IPC and GPUDirect RDMA are supported.