In the scenarios below, a private/on-prem AI system is necessary

1. **You cannot upload any data to the cloud** – in most cases, for legal or regulatory restrictions.

2. **Nvidia DGX cloud service or Google TPU cloud service are not available in your region**.

3. **Sensitive output of AIGC deduction cannot be exposed, even training can be done in public clouds.**

4. **Applications or calculations can only be executed practically by AI cards/CUDA, like**

    a) **Computer Vision:** While TPUs can handle some computer vision tasks, complex image recognition and video analysis usually require CUDA support.
    b) **High-Performance Computing (HPC):** Many HPC applications, such as climate or other complex systems modeling, rely on CUDA for acceleration.
    c) **Quantum Computing:** Quantum computing simulations and research typically require CUDA acceleration.
    d) **Drug Development:** The versatility of GPUs makes them more advantageous for handling diverse tasks, while TPUs perform better in training large-scale deep learning models.
    e) **Genomics Research:** Genomic data analysis usually involves processing large amounts of data, and the parallel computing capabilities of GPUs are well-suited for this type of task.
    f) **Antibody/Antigen Development:** These tasks typically involve complex molecular simulations and large-scale data analysis, where the versatility and powerful computing capabilities of GPUs excel.

Below are the typical libraries and elements only supported by CUDA with GPU/AI cards.

Mathematical Libraries

- cuBLAS: GPU-accelerated Basic Linear Algebra Subprograms (BLAS) library for matrix and vector operations.
- cuFFT: GPU-accelerated Fast Fourier Transform library for signal and image processing.
- cuRAND: GPU-accelerated random number generation library.
- cuSOLVER: GPU-accelerated dense and sparse direct solver library.
- cuSPARSE: GPU-accelerated sparse matrix BLAS library.
- cuTENSOR: GPU-accelerated tensor linear algebra library.

Operators

- cuDNN: Deep neural network library providing efficient operations for convolution, pooling, and activation functions.
- TensorRT: High-performance library for deep learning inference, supporting model optimization and deployment.

Computer Graphics

- OptiX: GPU-accelerated ray tracing engine for high-performance graphics rendering.
- CUDA Graphics API: Provides direct access to the GPU for efficient graphics rendering and computation.

Other Support

- RAPIDS: GPU-accelerated data science library supporting data processing and analysis. Source
- Thrust: GPU-accelerated C++ parallel algorithms and data structures library.

In comparison, TPUs primarily focus on accelerating deep learning tasks, especially large-scale matrix operations and neural network computations. While TPUs excel in these specific tasks, their support for low-level mathematical libraries and computer graphics is not as extensive and mature as that of GPUs.