# AZURE CONFIDENTIAL COMPUTING
## Confidential GPUs
## NDA Private Preview | Participant On-boarding Guide & Preview Scope

Last Updated: May 2022

## Contents

## Disclaimer

**The information shared in this document is covered under the NDA agreement between the two parties and is not to be shared publicly.**

This document contains preliminary information that may be changed substantially prior to final commercial release of the software described herein.

The information contained represents the current view of Microsoft Corporation on the issues discussed as of the date of the document. Because Microsoft must respond to changing market conditions, it should not be interpreted to be a commitment on the part of Microsoft, and Microsoft cannot guarantee the accuracy of any information presented after the date of the document.

This document is for informational purposes only. Microsoft makes no warranties, express, implied, or statutory, as to information regarding product roadmap or future capabilities.

Microsoft may have patents, patent applications, trademarks, copyrights, or other intellectual property rights covering subject matter in this document. Except as expressly provided in any written license agreement from Microsoft, the furnishing of this information does not give you any license to these patents, trademarks, copyrights, or other intellectual property.

# Foreword

Thank you for your participation in this preview of **Azure Confidential GPUs**!

Many industries such as healthcare, finance, transport, and retail are going through a major AI-led disruption. However, a broader democratization of AI is limited by concerns regarding sharing and use of personal data. In this context, Confidential Computing becomes an important tool to help organizations meet their privacy and security needs surrounding business and consumer data.

During this preview you should expect a gradual rollout of new security enhancements and controls. While our aim is to bring these capabilities across a wide set of OSes and VM sizes, some options may be limited as we work to address all compatibility needs. We will point these out in the applicable sectors.

## Disclaimer on private preview limitations

The Confidential GPU private preview is a cutting-edge exclusive Azure offer that makes use of new hardware capabilities. By participating in the preview, we hope you can assess the benefits of this new technology and start prototyping and evaluating new privacy-preserving machine learning applications; however, we do not recommend the use of sensitive data/production workloads during the private preview phase due to the following limitations:

- During the private preview phase, isolation of VMs relies on Azure Trusted VM (TVM), making use of virtual TPMs and virtualization-based security to protect and measure the boot, instead of AMD SEV-SNP hardware attestation capabilities used in Azure Confidential VMs (CVM). This means that during the private preview phase, the hypervisor and host operating system are trusted, and the VM memory is not hardware encrypted. We expect to replace TVM with CVM in a future preview phase; however, given that CVM attestation is also exposed through TPM (with the additional hardware attestation of the vTPM root key), the transition from TVM to CVM has limited impact on the design of confidential applications.
- **Known hardware limitations on A100:**
    - A100 only supports a limited encrypted data transfer bandwidth to and from the GPU memory. In the preview phase, you can expect CUDA memory copy operations to be limited to around 250 MB/s. This bottleneck can significantly impact the performance of applications that are IO limited (massively parallel inference, pictures/video streams) rather than compute limited (DNN training, NLP models, etc.).
    - On A100, only data confidentiality is protected, while CUDA kernels are loaded unencrypted. This means that an attacker may be able to launch a malicious CUDA kernel that has access to the Ampere protected memory region (however, if the kernel tries to write back protected data to the host, it will be encrypted by the GPU root of trust and only the guest who initiated the SPDM session will be able to decrypt it). Even though data transfers are integrity protected, a malicious kernel can tamper with confidential data in the protected memory region, so it is not safe to assume that the result of computation on A100 in protected memory mode can be trusted.
    - While our goal is enabling the execution of any unmodified GPU application, there are a limited number of unsupported CUDA APIs when the GPU is in protected memory mode.
- **Software limitations on attestation**
    - During the private preview phase, some parts of the security sensitive state of the GPU may not be initially measured and reported in attestation reports. The documentation of the Nvidia reference attestation verifier utility may include more details on the state captured by attestation; this is subject in future updates to the GPU drivers.
    - Validation of Confidential GPU VMs platform attestation on the Microsoft Azure Attestation (MAA) service will not be offered during the private preview phase.
    - Users will be able to request the GPU attestation report by using the GPU attestation tools supplied in the GitHub repository.

# What are Confidential GPU VMs?

Confidential GPU VMs are designed as an IaaS Virtual Machine for tenants with stringent security and confidentiality requirements. The threat model for confidential GPU VMs includes the host operating system and hypervisor, which includes Azure administrators and other Microsoft employees who can access Azure servers as well as malicious tenants who manage to escape VM isolation or gain unauthorized access to Azure management systems. It also includes some forms of hardware-level attacks, such as PCIe interposers between the server motherboard and the accelerator cards.

In combination with encryption of data at rest and in transit, confidential GPUs enable protection of data throughout its lifecycle by protecting data in use, both in the CPU and on the GPU. Data remains encrypted even when it is transferred from the CPU to a GPU over PCIe with keys that are securely exchanged between the CPU and the GPU. The only place where data is decrypted is within a hardware-protected, isolated environment within the GPU package where it can be processed to generate models or inference results.

Confidential GPU VMs bring together the security of trusted VMs with secure boot coupled with NVIDIA Ampere A100 GPUs with 80GB high-bandwidth memory. With Confidential GPU VMs, you can setup a secure isolated environment in the Azure public cloud and run your machine learning workloads utilizing any of the existing ML frameworks such as TensorFlow or PyTorch.

Confidential VMs require specialized hardware and software. As such, their initial availability is limited to certain regions, clusters and VM sizes.

To join this preview, you will need to sign up aka.ms/accgpusignup (if you haven't already).

- After signing up, please allow a few days for your subscription to be set up for this preview.

## Prerequisites

To participate in the private preview, you will need to have the following resources:

- A **fast** Internet connection (or a good deal of patience) since you will need to download/upload a combined 100GB-200GB. It is advised to do this from your fast corporate network as opposed to a home network. To be sure, you can do this from the safety of your home by remotely accessing a corporate workstation which will perform these bandwidth intensive operations over your corporate network.

## Guest OS

The following OS images are supported during private preview:

- Ubuntu 20.04 LTS

## VM Size

The new NCCadsA100v4-Series will be supported for these VMs. The VM sizes supported are:

1. Standard_NC24ads_A100_v4

### Regions

- East US 2

### Disclaimer

During this preview, hardware configuration and various firmware components remain in preproduction mode. Accordingly, Microsoft does not recommend using this preview for running production workloads or processing highly sensitive information.

## Creating a C-GPU VM

Please refer the GitHub location Azure-Confidential-Computing/PrivatePreview (github.com) for detailed steps.

Note: Please work with Confidential GPU team to get access to this location by sharing your GitHub account.

## Frequently Asked Questions

Please refer to the FAQ document (shared separately) for details.

## Feedback

Once you are done testing out the preview (successfully or not), please submit your feedback to cgpupreview@microsoft.com . Your comments and suggestions will be very carefully reviewed and considered by the Confidential GPU team and will likely have a direct impact on its future releases.

## Contact Us

Contact cgpupreview@microsoft.com for questions and support.