

AZURE CONFIDENTIAL COMPUTING

Confidential GPU VM

NDA Private Preview | Frequently Asked Questions

Last Updated: May 2022

This article provides answers to common questions about Confidential GPU VMs.

1. What is confidential computing?

Confidential computing is a privacy preserving technology that allows you to isolate your sensitive code and data while it's being processed. Many industries use confidential computing for scenarios such as securing financial information, protecting patient records, running machine learning algorithms on sensitive data, and processing sensitive datasets from multiple sources.

2. What are Confidential GPU VMs?

Confidential GPU VMs are currently IaaS VMs for tenants looking to protect sensitive data and proprietary AI models from unauthorized access using strong hardware-based security while utilizing the computing capabilities of GPUs. These VMs are available as NCC A100 v4 series VMs in Azure, and are powered by NVIDIA A100 PCIe GPU and 3rd-generation AMD EPYC™ 7V13 (Milan) processors. Workloads appropriate to preview on these VMs include:

- AI /ML training and inferencing workloads in industries like finance, healthcare, retail, advertising, and public sector where data used for both AI model training and inference may contain personally identifiable information (PII).
- Independent Software Vendors (ISVs) who want to protect Intellectual Property (IP) of AI models while distributing and deploying AI models on shared or remote infrastructures for consumption by their customers.
- Multi-Party data analysis in use cases like fraud detection, medical imaging, and drug development without compromising the confidentiality and integrity of data.

3. How can I deploy Confidential GPU VMs?

{Placeholder: Will update after finalizing onboarding document.}

4. Which Guest OS images are supported?

We support Ubuntu 20.04 during private preview. The images we support may be subject to change during public preview and GA.

5. What are the supported VM Sizes in Private preview?

We support one (1) GPU instance with 24 CPU cores and 192GB RAM in private preview.

VM	CPU cores	Network bandwidth (Gigabit)	RAM (GB)	Temporary Disk (TB)	A100 GPU instances
NCC24_ads_A100_v4	24	20	192	1	1

1 GPU instance = 1 A100 card

We will support for large number of cores, more RAM and multiple GPUs in later phases.

6. In which regions are Confidential GPU VMs available?

During private preview, Confidential GPU VMs are available only in the East US2 region.

7. What is the cost for Confidential GPU VMs?

The cost of a Confidential GPU VM depends on your usage and storage needs. Currently the pricing is the same as NC A100 v4 series VM SKU. Pricing information can be found at

[Pricing - Linux Virtual Machines | Microsoft Azure](#)

8. How do I attest Confidential GPU VMs?

During private preview, attestation of the host side components (boot loader, UEFI, guest OS) and GPU are done separately.

Confidential GPU VMs are based on trusted launch VMs. Azure automatically attests these VMs during boot to ensure that the VMs are running expected boot components. Additionally, customers can use the virtual TPM in their VMs to attest VM firmware and OS as described [here](#).

After secure boot, when the guest loads the NVIDIA GPU driver that supports APM, the driver attests that the GPUs are running expected firmware and are in good security state as part of the SPDM based key exchange protocol. This attestation is transparent to guest applications.

After the driver has loaded, guest applications must invoke GPU attestation using the cc-admin.py script and verifier tool provided in the GitHub repository. Running of CUDA programs is blocked until the guest runs GPU attestation.

In public preview, we will expand attestation capabilities e.g., enable VM attestation rooted in SEV-SNP hardware and support for attestation using Microsoft Azure Attestation service.

9. What do Confidential GPU VMs protect against and not protect against?

Where's the trust boundary?

In private preview, Confidential GPU VMs based on A100 GPUs offer safeguards to protect user data confidentiality. For example, Confidential GPU VMs protect against certain kinds of privileged attacks on confidentiality e.g., from compromised hypervisors to man-in-the-middle connections between the VM and the device. Since all user data transfers are encrypted, such attackers will not be able to compromise confidentiality.





However, Confidential GPU VMs based on A100 do not offer protection for code or data integrity. A malicious actor could attack the user data confidentiality protections, so these VMs should not be used for production use cases that require full confidential computing protections. Please refer below for a full list of attacks that are in and out of scope.

Production Confidential GPU VMs based on Hopper GPUs will offer full confidential computing protections, adding code confidentiality and data and code Integrity. Private Preview Confidential GPU VMs provide early adopter benefits to learn CC workflows on GPUs to ease the transition to Production on Hopper.

The following table summarizes the threat model and responsibilities for Confidential GPUs based on A100 and contrasts with what will come in Hopper:

THREATS & RESPONSIBILITY MODEL

✓=NVIDIA GPU ✗= Cloud CSP

Category	Threat	APM	Hopper-CC
 Confidentiality	Use PCIe/NVLink to read tenant data (e.g. Hypervisor, another VM, PCIe interposer)	✓	✓
	Use PCIe/NVLink to read tenant code (e.g. Hypervisor, another VM, PCIe interposer)	✗	✓
	Use Out-of-band management/debug channels to read tenant data (e.g. SMBus, JTAG)	✗	✓
	Use memory remapping to read tenant data	✓	✓
	Use GPU Cache/Memory based side channels to read tenant data	✓	✓
	Use GPU covert channel attacks to read tenant data	✗	✓
	Use GPU Performance Counters to read tenant data or fingerprint tenant	✓	✓
	Read tenant data via hypothetical physical attacks	✗	✗
 Integrity	Use PCIe/NVLink to modify tenant data (e.g. Hypervisor, another VM, PCIe interposer)	✗	✓
	Use PCIe/NVLink to modify tenant code (e.g. Hypervisor, another VM, PCIe interposer)	✗	✓
	Use Out-of-band management/debug channels to modify tenant data (e.g. SMBus, JTAG)	✗	✓
	Corrupt tenant data by replaying previous data or MMIO transactions (replay attacks)	✗	✓
	Corrupt tenant data via hypothetical physical attacks (fault injection, HBM interposer)	✗	✗
 Availability	Denial of Service to hypervisor by tenant	✓	✓
	Denial of Service to tenant by another tenant	NA	✓
	Denial of Service to tenant by hypervisor	✗	✗
 General	Use a spoofed, non-genuine, or known vulnerable TCB component	✗	✓
	Use hardware side channels to extract persistent device keys	✗	✓
	Use hardware side channels to extract tenant ephemeral session key	✗	✗

10. What happens if there is an attempt to violate confidentiality or integrity?

Some attacks on data confidentiality could be attempted e.g., by attaching a GPU without confidential computing capabilities or a GPU running microcode with known vulnerabilities. In confidential GPUs, such attacks will result in attestation failures and will be reported back to the application. Other integrity attacks, such as tampering with kernel code cannot be detected.

11. Can I convert a Trusted launch VM or any A100 GPU VM to confidential GPU VMs?

No. For security reasons, you must create a Confidential GPU from scratch.

12. Can I enable Confidential Computing features on any A100 GPU VM?

No. Confidential Computing features are enabled only on NCC A100 v4 series VMs and not available in other A100 GPU SKUs.

13. What changes are required for my ML applications to run on Confidential GPU VM?

Most ML applications will not require any changes to their CUDA code to run in Confidential GPU VMs. There are some exceptions. In private preview, multiple GPUs is not supported – an application can only use one GPU in confidential mode. Also, specific CUDA APIs (e.g., `cudaHostRegister` and CUDA peer-to-peer communication APIs) are not supported; please refer to NVIDIA's user guide and driver release notes in the GitHub repo for more details. We do expect changes to the scaffolding in your application e.g., to incorporate remote attestation.

14. How do I run ML applications inside Confidential GPU VM?

The simplest way to run ML applications in Confidential GPU VMs is to use containers from NVIDIA GPU Cloud (NGC). Please see NVIDIA's user guide for specific versions of containers that are supported with confidential GPU VMs.

15. What performance can I get for my ML workloads within confidential GPU VMs?

We will publish a thorough evaluation of performance overheads of using confidential GPU VMs with APM shortly.

16. What happens if I need Microsoft to help me with service or access data on my confidential GPU VM?

Azure doesn't have operating procedures for granting Confidential GPU VM access to its employees even if a customer authorizes the access. As a result, various recovery and support scenarios aren't available for confidential GPU VMs.