Microsoft Azure NVv4 GPU Virtual Machines

Microsoft Azure’s NVv4 GPU-enabled instances are built on 2nd Gen AMD EPYC™ 7002 processors and AMD Radeon™ Instinct MI25 GPU. The new NVv4 instances offers customers unprecedented GPU resourcing flexibility, giving customers more choices than before. The NVv4 GPU instance provides a more flexible option for customer use cases where the availability of a partial GPU is a cost-effective solution. A perfect example of this would be applications in industries such as AEC (e.g. AutoCAD) where graphics needs are light and a whole GPU is cost-prohibitive.

The NVv4 instance type now offers a customer the option to select a VM with a whole GPU all the way down to 1/8th of a GPU. This becomes a great cost-effective option for lightweight 3D applications and use cases that are less demanding but still require GPU.

NVv4 instance types support up to 32 vCPUs, 112GB of RAM, and 16GB of GPU memory. The instances below can allocate 1/8, 1/4, 1/2 or all of the 4096 stream processors available on the MI25 GPU:

<table>
<thead>
<tr>
<th>Size</th>
<th>vCPU</th>
<th>vCPU Memory</th>
<th>GPU partition</th>
<th>GPU Memory (VRAM)</th>
<th>Azure network</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard_NV4as_v4</td>
<td>4</td>
<td>14 GB</td>
<td>1/8</td>
<td>2 GB</td>
<td>50 Gbps</td>
</tr>
<tr>
<td>Standard_NV8as_v4</td>
<td>8</td>
<td>28 GB</td>
<td>1/4</td>
<td>4 GB</td>
<td>50 Gbps</td>
</tr>
<tr>
<td>Standard_NV16as_v4</td>
<td>16</td>
<td>56 GB</td>
<td>1/2</td>
<td>8 GB</td>
<td>50 Gbps</td>
</tr>
<tr>
<td>Standard_NV32as_v4</td>
<td>32</td>
<td>112 GB</td>
<td>1 (unpartitioned)</td>
<td>16 GB</td>
<td>50 Gbps</td>
</tr>
</tbody>
</table>

With this new GPU virtualization option, customers now have the option of running workloads on virtual GPUs with dedicated GPU frame buffers.

To understand performance considerations (network bandwidth, frame rates, etc.) and configuration requirements for different remote session use cases, review the Session Planning Guide and Workload Analysis section.

Teradici Cloud Access Software

Teradici Cloud Access Software is the Teradici end-to-end solution using PCoIP technology, that allows customers to deliver Windows and Linux desktops and applications from both an on-prem and cloud infrastructure such as Azure. The PCoIP technology that drives the Cloud Access Software solution helps with desktop and application remoting use cases in providing the best secure high-performance user experience on the market today.
An important part of the Cloud Access Software solution is the PCoIP Graphics Agent. The Graphics Agent is the component that gets installed on a GPU enabled desktop. The PCoIP Graphics Agent is built on and is driven by Teradici PCoIP technology, a display protocol with advanced display compression that encodes a complete desktop, which then in turn is displayed through a PCoIP client device over a standard IP network.

PCoIP also features a 'build-to-lossless' capability which ensures lossless reproduction of the original display image at the PCoIP client endpoint. Lossless reproduction is critical particularly in instances such as medical diagnostics, geospatial analysis, and media production, where the image itself contains important visual information. PCoIP protocol uses the User Datagram Protocol (UDP) which is much better suited for streaming media and real time display situations than TCP-based alternatives, especially over high latency networks.

Basic Teradici Cloud Access Software integration with Azure NVv4 instances

This architecture demonstrates the capability for both corporate and remote users to access NVv4 virtual machines running in Azure.

- Users connect to NVv4 virtual machines using a PCoIP client (Windows, macOS, or Linux)
- Cloud Access Connector is deployed in public facing subnet to provide secure connectivity between PCoIP clients and the Graphics Agent Host.
- Azure network security groups will apply the correct network ACLs.
- Active Directory is deployed to join and manage Azure NVv4 (PCoIP Graphics Agent Host).
- User domain accounts are used for authentication and authorization to virtual desktops.
- Radius server is used for secondary authentication method (Optional).
For this architecture the following components are deployed.

**Azure**

This architecture demonstrates the capability for both corporate and remote users to access NVv4 virtual machines running in Azure.

- **Azure Virtual Networks** - allows resources such as VMs to securely communicate with each other, the internet, and on-premises networks. Virtual networks provide isolation and segmentation, filter and route traffic, and allow connection between locations. One virtual network will be used for all resources in this scenario.

- **Network Security Groups** - contain a list of security rules that allow or deny inbound or outbound network traffic based on source or destination IP address, port, and protocol. The virtual networks in this scenario are secured with network security group rules that restrict the flow of traffic between the application components.

- **Azure NVv4 Virtual Machines** – Windows AMD Radeon™ Instinct MI25 GPU enabled VM for hosting visualization and simulation applications in industries such as Media and Entertainment, Manufacturing, AEC, and Finance.

- **Azure D2av4 Virtual Machines** – one instance to host Active Directory and second instance to host Radius Server.

**Teradici**

- **PCoIP Client** – installed and launched from an end-user’s workstation to initiate a PCoIP connection to the remote workstations.

- **Cloud Access Connector** - provides connectivity between the PCoIP clients and the remote workstations. The Cloud Access Connector provides a single point of entry into a deployment of remote workstations. It includes a NAT function for access through a single IP address.

- **Cloud Access Software Graphics Agent** - The Graphics Agent for Windows and Linux are part of Teradici Cloud Access Software. It enables the delivery of GPU-powered virtual workstations to end users via the PCoIP client.

**Solution Use Cases**

- Providing secure remote access to mission-critical, specialized virtualized desktops from any region.

- Provide a cost-effective alternative to virtual machines with dedicated GPUs. Use cases that have applications only requiring a fractional GPU would benefit from the NVv4 series.

- Entry-level virtual workstations for lightweight graphics applications.

- Organizations that will be deploying Windows Virtual Desktop.
Solution Considerations

- Check the Teradici Cloud Access Software solution page.
- Check the Graphics Agent for Windows system requirements for Azure NVv4 page.
- Latency and network conditions can have an impact on the overall solution. Refer to the Workload Analysis section of the Teradici Session Planning Guide.
- A trial or purchased license is required for user connections from a PCoIP client to a Cloud Access Software Graphics Host Agent.

Next Steps

- Review install steps for Cloud Access Software Graphics Host Agent software.
- Review deployment steps for Teradici Cloud Access Connector.
- Review NVv4 documentation and pricing.