NVIDIA and Microsoft at Ignite: Vera Rubin on Azure, Foundry post-training and confidential computing

25 November 2026 · 7 min read · blogs.iictp.org/nvidia-microsoft-azure-vera-rubin-foundry-confidential-computing-explained

Short answer: on 19 November 2026, at Microsoft Ignite, NVIDIA and Microsoft said Azure virtual machines powered by NVIDIA Vera Rubin NVL72 systems are available as a preview, that confidential computing is being extended to bare-metal and rack-scale AI infrastructure, and that Microsoft Foundry is adding post-training and custom-weight deployment for NVIDIA's Nemotron models. The release is a joint vendor announcement, with customer results quoted by NVIDIA.

TL;DR The short version

- On 19 November 2026 at Microsoft Ignite, NVIDIA and Microsoft announced Azure VMs on Vera Rubin NVL72 in preview.
- They also extended confidential computing to bare-metal and rack-scale AI infrastructure, and added post-training and custom-weight deployment of Nemotron models in Microsoft Foundry.
- Preview is not general availability and no pricing was given; ask for terms in writing and weigh managed post-training convenience against platform lock-in.

Read online, listen to it, or change the text size: https://blogs.iictp.org/nvidia-microsoft-azure-vera-rubin-foundry-confidential-computing-explained

What was announced?

- **Infrastructure:** Vera Rubin NVL72 systems are running in Azure data centres; NVIDIA's release describes an Azure Vera Rubin preview.
- **Confidential computing:** extending protection to bare-metal and rack-scale infrastructure, with attestation so customers can verify the environment before releasing sensitive models, data or keys.
- **Power efficiency:** collaboration on data-centre power efficiency using NVIDIA MaxLPS.
- **Software and models:** Foundry gets a serverless Programmed Training offering and Managed Compute support for custom model weights, plus an expanded Nemotron 3.5 lineup.

How the pieces fit

Infrastructure supplies the chips; Foundry gives enterprises a managed route to adapt open models; confidential computing addresses regulated data. Azure Local, Foundry Local and Jetson-based edge options extend the same stack to on-premises and industrial sites, according to NVIDIA's post.

Figures in the release

- EY used Programmed Training in Foundry to adapt Nemotron models to classify corporate purchasing records, and NVIDIA reports a 55% accuracy improvement on EY's evaluation set and 15x lower cost compared with proprietary frontier models.
- In MLPerf Training v6.1, Azure scaled to 10,240 NVIDIA GPUs with GB200 NVL72 systems on DeepSeek-V3, the largest scale in the round, per NVIDIA.
- Microsoft is using NVIDIA RTX PRO 6000 Blackwell GPUs for Bing search models, and Dynamo inference software for post-training.

Why it matters commercially

Cloud providers compete on access to the newest accelerators and on how easily customers can customise models without running training clusters. The announcement also shows the confidential-computing angle becoming part of the pitch for regulated industries.

Limitations and open questions

- "Preview" is not general availability; the release text I read gives no pricing, regions or dates for general availability.
- The EY result is a single customer's evaluation, reported through NVIDIA.
- This is a joint marketing release. Sources were read on 25 November 2026. Products, rules and standards change, so check the linked originals before you rely on any detail.

How to evaluate it for your organisation

1. Ask for preview terms, regions, quotas and pricing in writing.
2. List which workloads need confidential computing and what attestation evidence your auditors will accept.
3. Run a cost and accuracy comparison of an adapted open model against your current proprietary model on your own data.
4. Plan for power and networking, not only GPU hours.

What it means for IT leaders and architects

The practical choice is where customisation happens: a managed post-training service saves operational work but ties you to a platform. The EY example is an argument for testing small adapted open models on narrow classification tasks. See Hugging Face on Google Cloud Marketplace for the multi-cloud procurement side.

Frequently asked questions

- **What did NVIDIA and Microsoft announce?** A Vera Rubin NVL72 preview on Azure, confidential computing for rack-scale AI, and Foundry post-training and custom-weight deployment for Nemotron models.
- **What did EY report?** A 55% accuracy gain and 15x lower cost than proprietary frontier models on its purchasing-record classification, per NVIDIA.
- **Is it generally available?** The release describes a preview.

- **Is the confidential computing live?** The companies say they are working to extend it.

Related reading

- NVIDIA Q3 fiscal 2027 results
- Hugging Face on Google Cloud Marketplace
- Cloud Architecture Practitioner certification

> Related credential: Cloud Architecture Practitioner. Explore the exam facts and syllabus.

Sources and further reading

1. NVIDIA. NVIDIA and Microsoft Expand Accelerated Computing From Azure to Windows PCs. 19 November 2026. blogs.nvidia.com/blog/microsoft-ignite-accelerated-computing-azure-windows-pcs

Published by IICTP, an independent technical credential body, on 25 November 2026. The current version of this post is at https://blogs.iictp.org/nvidia-microsoft-azure-vera-rubin-foundry-confidential-computing-explained. More guides: blogs.iictp.org.