Deploy NGC on On-Premises Servers

Documentation Introduction
Table of Contents

Chapter 1. NGC Software is Certified on On-Premises Systems and the Cloud.................1
Chapter 2. NVIDIA-Certified Systems.................................................................................. 2
Chapter 3. NGC-Ready Servers............................................................................................3
Chapter 1. NGC Software is Certified on On-Premises Systems and the Cloud

NVIDIA certification programs validate the performance of AI, ML and DL workloads using NVIDIA GPUs on data center and edge servers and public clouds. It allows server manufacturers and public clouds to qualify their NVIDIA GPU equipped systems on a wide variety of AI workloads ranging from training to inference on on-premise servers and public cloud infrastructures.
Chapter 2. NVIDIA-Certified Systems

NVIDIA is working with its partners to certify the industry’s most advanced AI systems - EGX servers running NVIDIA A100 and HGX servers built with HGX A100-4 and HGX A100-8 baseboards - both as single-node servers and in cluster configurations using Mellanox ConnectX-6, ConnectX-6DX, and Bluefield 2 network adapters and Mellanox Quantum HDR InfiniBand and Mellanox Spectrum® Ethernet switches.

The NVIDIA-Certified Systems program is the evolution of the NGC-Ready program that tests single node servers against a diverse range of workloads. The NVIDIA-Certified Systems program introduces new tests of multi-node deep learning training, distributed training and inference workloads, high-speed remote storage access, and cluster health, performance, and security tests.
Chapter 3. NGC-Ready Servers

The NGC-Ready program validates the performance of AI, ML and DL workloads using NVIDIA GPUs on leading on-premises servers. It allows server manufacturers to qualify their NVIDIA GPU equipped systems on a wide variety of AI workloads ranging from training to inference on on-premise servers, cloud infrastructure and edge devices.

Servers certified as NGC-Ready for Edge undergo additional testing for the security and remote system management capabilities required for remote deployment at the edge.

Performance-validated NGC-Ready systems accelerate time to solution, increase business agility, and are backed by enterprise-grade support, reducing the risk of deploying complex AI solutions.

- For the current list of NGC-Ready systems, visit https://docs.nvidia.com/ngc/ngc-ready-systems/index.html
- For detailed guidelines for configuring NGC-Ready systems, visit https://docs.nvidia.com/ngc/ngc-ready-recommended-configs/
- For a detailed specification for configuring a NGC-Ready server using the NVIDIA T4 GPU, visit https://docs.nvidia.com/ngc/ngc-t4-platform-design-guide/index.html

DGX Systems

All NGC content is tested on NVIDIA DGX systems, NVIDIA’s best-in-class portfolio of purpose-built AI systems. To learn more about NVIDIA DGX systems visit: https://www.nvidia.com/en-us/data-center/dgx-systems/.

Deploying NGC on a Public Cloud

For documentation on deploying NGC on a public cloud, visit Deploying NGC on a Public Cloud
Notice

THE INFORMATION IN THIS GUIDE AND ALL OTHER INFORMATION CONTAINED IN NVIDIA DOCUMENTATION REFERENCED IN THIS GUIDE IS PROVIDED "AS IS." NVIDIA MAKES NO WARRANTIES, EXPRESSED, IMPLIED, STATUTORY, OR OTHERWISE WITH RESPECT TO THE INFORMATION FOR THE PRODUCT, AND EXPRESSLY DISCLAIMS ALL IMPLIED WARRANTIES OF NONINFRINGEMENT, MERCHANTABILITY, AND FITNESS FOR A PARTICULAR PURPOSE. Notwithstanding any damages that customer might incur for any reason whatsoever, NVIDIA’s aggregate and cumulative liability towards customer for the product described in this guide shall be limited in accordance with the NVIDIA terms and conditions of sale for the product.

THE NVIDIA PRODUCT DESCRIBED IN THIS GUIDE IS NOT FAULT TOLERANT AND IS NOT DESIGNED, MANUFACTURED OR INTENDED FOR USE IN CONNECTION WITH THE DESIGN, CONSTRUCTION, MAINTENANCE, AND/OR OPERATION OF ANY SYSTEM WHERE THE USE OR A FAILURE OF SUCH SYSTEM COULD RESULT IN A SITUATION THAT THREATENS THE SAFETY OF HUMAN LIFE OR SEVERE PHYSICAL HARM OR PROPERTY DAMAGE (INCLUDING, FOR EXAMPLE, USE IN CONNECTION WITH ANY NUCLEAR, AVIONICS, LIFE SUPPORT OR OTHER LIFE CRITICAL APPLICATION). NVIDIA EXPRESSLY DISCLAIMS ANY EXPRESS OR IMPLIED WARRANTY OF FITNESS FOR SUCH HIGH RISK USES. NVIDIA SHALL NOT BE LIABLE TO CUSTOMER OR ANY THIRD PARTY, IN WHOLE OR IN PART, FOR ANY CLAIMS OR DAMAGES ARISING FROM SUCH HIGH RISK USES.

NVIDIA makes no representation or warranty that the product described in this guide will be suitable for any specified use without further testing or modification. Testing of all parameters of each product is not necessarily performed by NVIDIA. It is customer’s sole responsibility to ensure the product is suitable and fit for the application planned by customer and to do the necessary testing for the application in order to avoid a default of the application or the product. Weaknesses in customer’s product designs may affect the quality and reliability of the NVIDIA product and may result in additional or different conditions and/or requirements beyond those contained in this guide. NVIDIA does not accept any liability related to any default, damage, costs or problem which may be based on or attributable to: [i] the use of the NVIDIA product in any manner that is contrary to this guide, or [ii] customer product designs.

Other than the right for customer to use the information in this guide with the product, no other license, either expressed or implied, is hereby granted by NVIDIA under this guide. Reproduction of information in this guide is permissible only if reproduction is approved by NVIDIA in writing, is reproduced without alteration, and is accompanied by all associated conditions, limitations, and notices.