## **NVIDIA Data Center Platform**



Accelerate every workload.

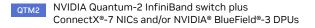
Rapid developments and continuous breakthroughs in AI are fueling transformative change, spanning all industries and revolutionizing the workflows of scientists, engineers, creators, and more. On top of the demand for accelerated computing to power traditional Al applications—such as machine learning, deep learning, natural language processing, and computer vision—a new use case has emerged that's unlocking a frontier of opportunities—generative Al. The NVIDIA data center platform is the world's leading accelerated computing and generative AI solution, deployed by the largest supercomputing centers and enterprises. It enables breakthrough performance with fewer, more powerful servers, driving faster time to insights, while saving money.

The platform accelerates a broad array of workloads, from generative AI training and inference to scientific computing and virtual desktop infrastructure (VDI) applications, with a diverse range of GPUs, from the highest performing to entry level, all powered by a single unified architecture. For optimal performance, it's essential to identify the ideal GPU for a specific workload. Use this as a guide to those workloads and the corresponding NVIDIA GPUs that deliver the best results.

## **GPU Portfolio: NVIDIA Hopper™ and Ada Lovelace Architectures**

| Solution Category                                  | GPU                 | Networking | Training and<br>Data Analytics | Inference | HPC/AI | NVIDIA<br>Omniverse™ /<br>Render Farms | Virtual<br>Workstations | Virtual<br>Desktops<br>(VDI) | Al Video | Far-Edge<br>Acceleration |
|--|---------------------|------------|--------------------------------|-----------|--------|--|-------------------------|------------------------------|----------|--------------------------|
| Compute  | GH200               | QTM2 SPCX  | •                              | •         | •      |  |                         |                              |          |                          |
|  | NVIDIA HGX™<br>H200 | QTM2 SPCX  |                                | •         | •      |  |                         |                              |          |                          |
|  | HGX H100            | QTM2 SPCX  |                                |           |        |  |                         |                              |          |                          |
|  | H100 NVL            | QTM2 SPCX  |                                |           |        |  |                         |                              |          |                          |
| Graphics<br>and Compute                            | L40S                | QTM2 SPCX  |                                | -         |        |  |                         |                              |          |                          |
| Small Form Factor<br>(SFF) Compute<br>and Graphics | L4                  | SPC3       |                                |           |        |  |                         |                              |          |                          |







Spectrum SN5600 Ethernet switch plus BlueField-3 SuperNICs

## **NVIDIA Inference Portfolio**

| GPU                 | Natural Language Processing (NLP) / Large Language Models (LLMs) |                      |                        |                         |                              |                        |                          |                    |          |
|---------------------|--|----------------------|------------------------|-------------------------|------------------------------|------------------------|--------------------------|--------------------|----------|
|                     | Up to 5 Billion (B)<br>Parameters                                | 6B-65B<br>Parameters | 66B-175B<br>Parameters | Over 175B<br>Parameters | Image/Video<br>Generative Al | Recommender<br>Systems | Graph/Vector<br>Database | Computer<br>Vision | Al Video |
| GH200               |  | •                    |                        |                         |                              |                        |                          |                    |          |
| HGX H200<br>(8-way) |  | •                    |                        |                         |                              |                        |                          |                    |          |
| HGX H100<br>(8-way) |  |                      |                        |                         |                              |                        |                          |                    |          |
| H100 NVL            |  |                      |                        |                         |                              |                        |                          |                    |          |
| L40S                |  |                      |                        |                         |                              |                        |                          |                    |          |
| L4                  |  |                      |                        |                         |                              |                        |                          |                    |          |



## **NVIDIA Training Portfolio**

| GPU                      |                        | NLP                  | LLM                    |                         |                              |                        |                          |
|--------------------------|------------------------|----------------------|------------------------|-------------------------|------------------------------|------------------------|--------------------------|
|                          | Up to 5B<br>Parameters | 6B-65B<br>Parameters | 66B-175B<br>Parameters | Over 175B<br>Parameters | Image/Video<br>Generative Al | Recommender<br>Systems | Graph/Vector<br>Database |
| GH200                    | -                      |                      | -                      |                         |                              |                        |                          |
| HGX H200/H100<br>(8-way) |                        |                      |                        |                         |                              |                        |                          |
| H100 NVL                 | •                      |                      |                        |                         |                              |                        |                          |
| L40S                     |                        |                      |                        |                         |                              |                        |                          |



To learn more about NVIDIA data center GPUs, visit www.nvidia.com/data-center-gpus

© 2024 NVIDIA Corporation and affiliates. All rights reserved. NVIDIA, the NVIDIA logo, BlueField, ConnectX, HGX, Hopper, Omniverse, and Spectrum are trademarks and/or registered trademarks of NVIDIA Corporation and affiliates in the U.S. and other countries. Other company and product names may be trademarks of the respective owners with which they are associated. 3267213. MAY24

