The Exponential Growth of Computing

Accelerated computing is being rapidly adopted across industries and large-scale production deployments. Because new compute demands are outstripping the capabilities of traditional CPU-only servers, enterprises need to optimize their data centers—making this acceleration a must-have. The NVIDIA data center platform is the world’s leading accelerated computing 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 AI training and inference to supercomputing and virtual desktop infrastructure (VDI) applications, with a diverse range of GPUs. For optimal performance, it’s essential to identify the ideal GPU for a specific workload. A guide to those workloads and the corresponding NVIDIA GPUs that deliver the best results is provided on the next page.
# Choose the Right NVIDIA Data Center GPU for You

<table>
<thead>
<tr>
<th>WORKLOAD</th>
<th>DESCRIPTION</th>
<th>NVIDIA V100 SXM2 with NVIDIA® NVLink™</th>
<th>NVIDIA V100S with PCIe</th>
<th>NVIDIA T4</th>
<th>NVIDIA Quadro RTX™ 6000/8000</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deep Learning Training</td>
<td>For the absolute fastest model training time</td>
<td>8–16 GPUs</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>High-Performance Computing (HPC)</td>
<td>For scientific computing centers and higher education and research institutions running HPC and AI (training/inference) workloads</td>
<td>4 GPUs for supercomputing centers</td>
<td>Up to 4 GPUs for higher education and research use cases</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Render Farms</td>
<td>For batch and real-time rendering</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Performance Graphics</td>
<td>For the best graphics performance on professional virtual workstations</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mainstream Graphics</td>
<td>For cost-effectively running virtual desktops for knowledge workers</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Enterprise Acceleration</td>
<td>For enterprises running mixed workloads (e.g., graphics, machine learning, deep learning, data science, and analytics)</td>
<td>4 GPUs for compute-intensive, multi-GPU workloads</td>
<td>Up to 4 GPUs for compute-intensive, single-GPU workloads</td>
<td>4–8 GPUs</td>
<td></td>
</tr>
<tr>
<td>Edge Acceleration</td>
<td>For deploying AI to the edge with multiple use cases and locations</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

## Recommended number of GPUs per workload

### Deep Learning Training
- Recommended: 8–16 GPUs

### High-Performance Computing (HPC)
- Recommended: 4 GPUs for supercomputing centers
- Higher education and research use cases: Up to 4 GPUs

### Render Farms
- Recommended: 4–8 GPUs

### Performance Graphics
- Recommended: 2–8 GPUs for midrange virtual workstations (e.g., ProE or Autodesk)
- Running graphics and simulation applications (e.g., CATIA and live rendering): 4–8 GPUs

### Mainstream Graphics
- Recommended: 2–8 GPUs for graphics use cases only (e.g., running CATIA software)
- Lowest total cost of ownership per user: 2–4 M10 GPUs

### Enterprise Acceleration
- Recommended: 4–8 GPUs for balanced workloads
- Graphics-intensive workloads: 2–4 GPUs

### Edge Acceleration
- Inference and video-code-intensive workloads (e.g., intelligent video analytics, industrial inspection): 1–8 GPUs
- Graphics-intensive workloads (e.g., augmented reality, virtual reality): 2–4 GPUs

## Key Features

**NVIDIA V100 SXM2 with NVIDIA® NVLink™**
- 32 GB HBM2 memory
- 125 teraFLOPS of tensor operations for deep learning
- 15.7 teraFLOPS of single-precision performance
- 7.8 teraFLOPS of double-precision performance
- 300 GB/s NVIDIA NVLink interconnect bandwidth
- 900 GB/s memory bandwidth

**NVIDIA V100S with PCIe**
- 32 GB HBM2 memory
- 130 teraFLOPS of INT8 inference performance
- 16.4 teraFLOPS of single-precision performance
- 8.1 teraFLOPS of double-precision performance
- 32 GB/s PCIe interconnect bandwidth
- 1,134 GB/s memory bandwidth

**NVIDIA T4**
- 16 GB memory
- 130 teraFLOPS of INT8 inference performance
- 8.1 teraFLOPS of single-precision performance
- Dedicated video decode and encode engines
- 70 W power
- Low-profile form factor

**NVIDIA Quadro RTX™ 6000/8000**
- 24/48 GB memory
- 130 teraFLOPS of tensor operations for deep learning
- 261 teraFLOPS of INT8 inference performance
- 16 teraFLOPS of single-precision performance
- RT Cores for high-performance rendering
- Dedicated video decode and encode engines
- 100 GB/s NVIDIA NVLink

© 2019 NVIDIA Corporation. All rights reserved. NVIDIA, the NVIDIA logo, NVLink, and Quadro RTX are trademarks and/or registered trademarks of NVIDIA Corporation in the U.S. and other countries. All other trademarks and copyrights are the property of their respective owners. DEC19