PRODUCTS		NVIDIA H100 PCIe	NVIDIA A100 80GB PCIe	NVIDIA A30	NVIDIA H100 CNX PCIe	NVIDIA A100X	NVIDIA A30X
PNY PART NUMBERS		NVH100XTCGPUCA-KIT	NVA100TCGPU80NC-KIT / NVA100TCGPU80-KIT	NVA30TCGPUNC-KIT / NVA30TCGPU-KIT	NVH100XTCGPUCNX-KIT	NVA100XTCGPUCA-KIT	NVA30XTCGPUCA-KIT
NVIDIA PART NUMBERS		900-21010-0000-000	900-21001-0020-100 / 900-21001-0020-000	900-21001-0040-100 / 900-21001-0040-000	900-21011-0000-000	900-21004-0030-000	900-21004-0010-000
WORKLOAD	DESCRIPTION	Ultimate Performance Compute	High Performance Compute	Mainstream Compute	Ultimate Performance Converged Accelerator	High Performance Converged Accelerator	Mainstream Converged Accelerator
		Recommended Number of GPUs or Converged Boards per Server					
Deep Learning (DL) Training and Data Analytics	For the absolute fastest model training and analytics	**4-8 GPUs** 80GB: 8n+ parameter models (DLRM, GPT-3)	**4-8 GPUs** 80GB: 8n+ parameter models (DLRM, GPT-3)		**1-2 cards** for multi-node training	**1-2 cards** for multi-node training	
DL Inference	For batch and real-time inference	**1-2 GPUs** w/ multi-instance GPU (MIG) 80GB: large batch size constrained models (RNN-T)	**1-2 GPUs** w/ multi-instance GPU (MIG) 80GB: large batch size constrained models (RNN-T)	**2–4 GPUs** with MIG			
High Performance Computing HPC / AI	For Higher Education Research and scientific computing centers	**1-4 GPUs** with MIG	**1-4 GPUs** with MIG	**2–4 GPUs** with MIG	**1-2 cards** for multi-node workloads	**1-2 cards** for multi-node workloads	
Render Farms	For batch and real-time photorealistic rendering						
Graphics	For the highest user density and graphics performance on professional VDI						
Cloud Gaming	For 4K resolution on Android devices						
Enterprise Acceleration	For mixed workloads, including graphics, DL, ML, analytics, training, and inference	**1-2 GPUs** with MIG for compute workloads	**1-2 GPUs** with MIG for compute workloads	**1-2 GPUs** with MIG for compute workloads			**1 card** for compute acceleration with software-defined infrastructure
Edge Acceleration	For differing use cases and deployment locations	**1-2 GPUs** with MIG	**1-2 GPUs** with MIG	**1-2 GPUs** with MIG	**1 card** for AI-on-5G with heavy workloads	**1 card** for AI-on-5G with heavy workloads	**1 card** for AI-on-5G with average workloads
5G vRAN	For low-latency wireless GPU to network communication						**1-2 cards**
AI-Based Security	For AI-based and GPU accelerated network security processing						**1 card**

* NVIDIA RTX Virtual Workstation (vWS) software license required for virtual workstation workloads.
** NVIDIA Virtual PC (vPC) software license required for VDI workloads.

PNY 3 THREE YEAR WARRANTY

PNY | NVIDIA.

REV. 08.12.22

PRODUCTS		NVIDIA T4	NVIDIA A2	NVIDIA A40	NVIDIA A10	NVIDIA A16
PNY PART NUMBERS		TCSCT4-KIT	NVA2TCGPUNC-KIT NVA2TCGPU-KIT	NVA40TCGPUNC-KIT NVA40TCGPU-KIT	NVA10TCGPUNC-KIT NVA10TCGPU-KIT	NVA16TCGPUNC-KIT NVA16TCGPU-KIT
NVIDIA PART NUMBERS		900-2G183-0000-001	900-2G179-0020-101 900-2G179-0020-001	900-2G133-0000-100 900-2G133-0000-000	900-2G133-0020-000 900-2G133-0020-100	900-2G171-0000-100 900-2G171-0000-000
WORKLOAD	DESCRIPTION	Small Footprint Low Power	Entry-Level Compact AI	Highest Performance Graphics	Mainstream Graphics	Optimized for VDI
		Recommended Number of GPUs or Converged Boards per Server				
Deep Learning (DL) Training and Data Analytics	For the absolute fastest model training and analytics					
DL Inference	For batch and real-time inference	**1–4 GPUs**	**1–4 GPUs**			
High-Performance Computing (HPC) / AI	For Higher Education Research and scientific computing centers					
Render Farms	For batch and real-time rendering			**4–8 GPUs**	**4–8 GPUs**	
Graphics	For the best graphics performance on professional VDI	**1–4 GPUs** for entry-level virtual workstations*	**1–4 GPUs** for entry-level virtual workstations*	**2–4 GPUs** for midrange to high-end virtual workstations*	**2-4 GPUs** for midrange virtual workstations*	**2-4 GPUs** for highest virtual desktop and workstation user density**
Cloud Gaming	For 4K resolution / Android	**1–4 GPUs** for mobile android	**1–4 GPUs** for mobile android	**4–8 GPUs** (4K resolution)	**4–8 GPUs** (4K resolution)	
Enterprise Acceleration	For mixed workloads, including graphics, ML, DL, analytics, training, and inference	**1–4 GPUs** for balanced workloads*	**1–4 GPU** for balanced workloads*	**1–2 GPUs** for graphics-intensive workloads*	**1-2 GPUs** for graphics workloads*	
Edge Acceleration	For differing use cases and deployment locations	**1–4 GPUs** for inference and video workloads	**1-4 GPUs** for inference and video workloads	**1-4 GPUs** for graphics-intensive workloads & AR / VR*	**1-4 GPUs** for graphics workloads & AR / VR*	
5G vRAN	For low-latency GPU-network communication					
AI-Based Security	For GPU-powered network processing					

* NVIDIA RTX Virtual Workstation (vWS) software license required for virtual workstation workloads.
** NVIDIA Virtual PC (vPC) software license required for VDI workloads.

PNY 3 THREE YEAR WARRANTY

PNY | NVIDIA.

REV. 08.12.22