NVIDIA A30 GPU Accelerator

Product Brief

PB-10418-001_v01 | March 2021

Document History

PB-10418-001_v01

Version	Date	Authors	Description of Change
01	March 23, 2021	AV, AS, SM	Initial Release

Table of Contents

List of Figures

List of Tables

Overview

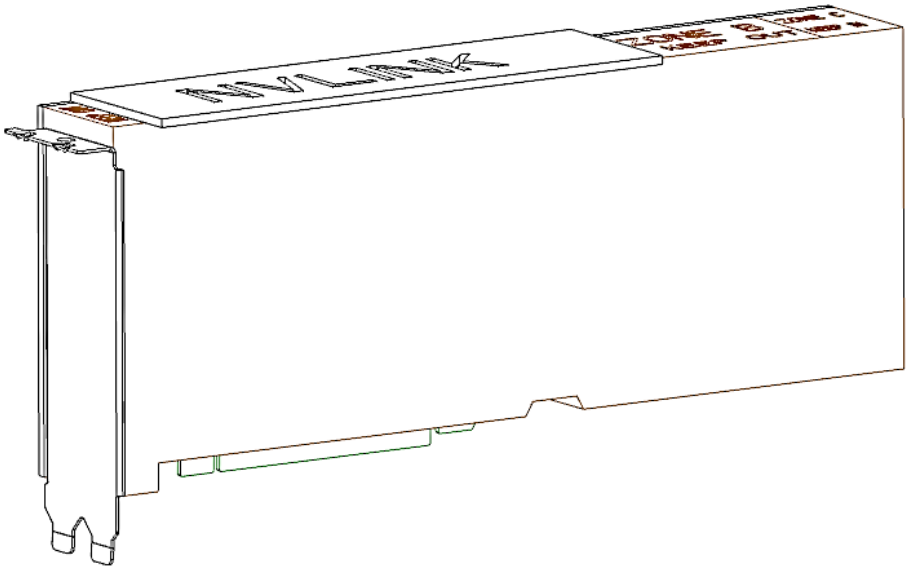
The NVIDIA® A30 Tensor Core graphics processing unit (GPU) delivers a versatile platform for mainstream enterprise workloads like AI inference, training, and high-performance computing (HPC). It combines 3rd generation tensor cores with 24 GB of HBM2 memory in a dual-slot 10.5-inch PCI Express Gen4 form factor, with 165 W maximum board power. The card is passively cooled that requires system airflow to operate within its thermal envelope.

Built on the latest NVIDIA Ampere architecture, the NVIDIA A30 brings innovations like Tensor Float 32 (TF32) and Tensor Core FP64, as well as end-to-end software stack solutions, including the NVIDIA AI Enterprise suite to ensure that mainstream AI and HPC jobs can be rapidly solved. In addition to these features, the A30 supports double precision (FP64), single precision (FP32), half precision (FP16), Brain Float 16 (BF16) and Integer (INT8) computations, unified virtual memory, and page migration engine capability. The Multi-Instance GPU (MIG) feature ensures quality of service (QoS) with secure, hardware-partitioned, right-sized GPUs across all compute workloads for a diverse set of users and maximizes the utilization of GPU resources.

The NVIDIA A30 ships with ECC enabled to protect the GPU's memory interface and the on-board memories from detectable errors. A30's HBM2 memory has native support for ECC with no ECC overhead, both in memory capacity and bandwidth.

The card is designed to meet the requirements of NEBS Level 3 compliant servers and supports security features like secure boot and hardware root-of-trust.

Figure 1. NVIDIA A30 PCIe Card



Specifications

Product Specifications

Table 1 through Table 3 provide the product, memory, and software specifications for the NVIDIA A30 GPU card.

Table 1. Product Specifications

Specification	NVIDIA A30
Product SKU	P1001 SKU 205 NVPN: 699-21001-0205-xxx
Total board power	165 W
Thermal solution	Passive
Mechanical Form Factor	Full-height, full-length (FHFL) 10.5", dual-slot
PCI Device IDs	Device ID: 0x20B7 Vendor ID: 0x10DE Sub-Vendor ID: 0x10DE Sub-System ID: 0x1532
GPU clocks	Base: 930 MHz Boost: 1440 MHz
Performance States	P0
VBIOS	EEPROM size: 8 Mbit UEFI: Supported
PCI Express interface	PCI Express 4.0 ×16 Lane and polarity reversal supported
Multi-Instance GPU (MiG)	Supported (up to 4 instances)
Secure Boot	Supported
Zero Power	Not supported
NEBS readiness	Supported
Power connectors and headers	One CPU 8-pin auxiliary power connector

Specification	NVIDIA A30
Weight	Board: 1240 grams (excluding bracket, extenders, and bridge) NVLink Bridge: 20.5 grams Bracket with screws: 20 grams Long offset extender: 64 grams Straight extender: 39 grams

Table 2. Memory Specifications

Specification	Description
Memory clock	1215 MHz
Memory type	HBM2
Memory size	24 GB
Memory bus width	3072 bits
Peak memory bandwidth	Up to 933 GB/s

Table 3. Software Specifications

Specification	Description[1]
SR-IOV support	Supported: 8 VF (virtual functions)
BAR address (physical function)	BAR0: 16 MiB[1] BAR1: 32 GiB[1] BAR3: 32 MiB[1]
BAR address (virtual function)	BAR0: 2 MiB, (256 KiB per VF)[1] BAR1: 32 GiB, 64-bit (4 GiB per VF)[1] BAR3: 256 MiB, 64-bit (32 MiB per VF)[1]
Message signaled interrupts	MSI-X: Supported MSI: Not supported
Multi-Instance GPU (MIG)	Supported
ARI Forwarding	Supported
Driver Support	Linux: R460.65 or later Windows: R461.98 or later
CEC Firmware	v6.01 or later
NVIDIA® CUDA® Support	CUDA 11.2.1 (or later)
Virtual GPU Software Support	Supports vGPU 13.0 (or later): NVIDIA Virtual Compute Server Edition
NVIDIA AI Enterprise	Supported with VMware
NVIDIA® NGC-Ready™ Test Suite	NGC-Next Certification 2.2 (or later)
PCI class code	0x03 – Display Controller

Specification	Description[1]
PCI sub-class code	0x02 – 3D Controller
Primary Boot Device Capability	Not supported
ECC support	Enabled (by default). Can be disabled via software
SMBus (8-bit address)	0x9E (write), 0x9F (read)
SMBus direct access	Supported
Reserved I2C addresses[2]	0xAA, 0xAC
SMBus Post-Box Interface (SMBPBI)	Supported

Note:

[1]The KiB, MiB and GiB notation emphasizes the "power of two" nature of the values. Thus,

- 256 KiB = 256 x 1024
- 16 MiB = 16 x 1024^2
- 64 GiB = 64 x 1024^3

[2]See "CEC Hardware Root of Trust" section of this product brief.

The operator is given the option to configure this power setting to be persistent across driver reloads or to revert to default power settings upon driver unload.

Environmental and Reliability Specifications

Table 4 provides the environment conditions specifications for the NVIDIA A30 card.

Table 4. Board Environmental and Reliability Specifications

Specification	Description
Ambient operating temperature	0 °C to 55 °C
Storage temperature	-40 °C to 75 °C
Operating humidity	5% to 95% relative humidity
Storage humidity	5% to 95% relative humidity
Mean time between failures (MTBF)	Uncontrolled environment:[1] TBD hours at 35 °C Controlled environment:[2] TBD hours at 35 °C

Notes:

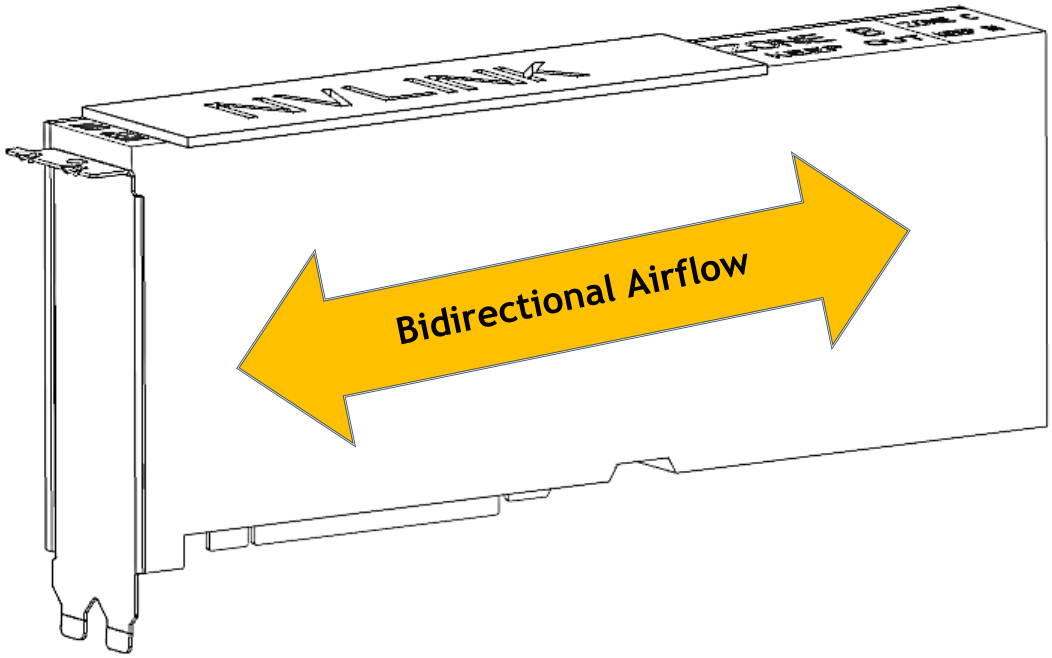
[1]Some environmental stress with limited maintenance (GF35).

[2]No environmental stress with optimum operation and maintenance (GB35).

Airflow Direction Support

The NVIDIA A30 PCIe card employs a bidirectional heat sink, which accepts airflow either left-to-right or right-to-left directions.

Figure 2. NVIDIA A30 Airflow Directions



Product Features

PCI Express Interface Specifications

The following sub-sections describe the PCIe interface specifications for the NVIDIA A30 PCIe card.

PCIe Speed Support

The A30 card supports PCIe Gen4.

Polarity Inversion and Lane Reversal Support

Lane Polarity Inversion, as defined in the PCIe specification, is supported on the A30 PCIe card.

Lane Reversal, as defined in the PCIe specification, is supported on the A30 PCIe card. When reversing the order of the PCIe lanes, the order of both the Rx lanes and the Tx lanes must be reversed.

CEC Hardware Root of Trust

The NVIDIA A30 provides secure boot capability via CEC. Implementing code authentication, rollback protection and key revocation, the CEC device authenticates the contents of the GPU firmware ROM before permitting the GPU to boot from its ROM.

It also provides out-of-band (OOB) secure firmware update, secure application processor recovery, and remote attestation.

The Hardware Root of Trust feature occupies up to two I2C addresses (in addition to the SMBus addresses). I2C addresses 0xAA and 0xAC should therefore be avoided for system use.

Multi-Instance GPU Support

The A30 PCIe card supports Multi-Instance GPU (MIG) capability by providing up to 4 GPU instances per NVIDIA A30 GPU. MIG technology can partition the A30 GPU into individual instances, each fully isolated with its own high-bandwidth memory, cache, and compute cores, enabling optimized computational resource provisioning and quality of service (QoS).

For detailed information on MIG provisioning and use, consult the *Multi-Instance GPU User Guide*: https://docs.nvidia.com/datacenter/tesla/mig-user-guide/index.html

Programmable Power

The Programmable Power feature provides partners the general ability to configure the power cap of the card for system power/thermal budget or performance-per-watt reasons.

The power cap can be modified using either of these two NVIDIA tools:

- nvidia-smi (power cap adjustment must be re-established after each new driver load)
- SMBPBI (power cap adjustment remains in force across driver loads and system boots)

Power limit specifications for the NVIDIA A30 are presented in Table 1.

nvidia-smi

nvidia-smi is an in-band monitoring tool provided with the NVIDIA driver and can be used to set the maximum power consumption with driver running in persistence mode. An example command to reduce the power cap to 100 W is shown:

```
nvidia-smi -pm 1
nvidia-smi -pl 100
```

To restore the A30 back to its default TDP power consumption, either the driver module can be unloaded and reloaded, or the following command can be issued:

```
nvidia-smi -pl 165
```

SMBPBI

An out-of-band channel exists through the SMBus Post-Box Interface (SMBPBI) protocol to set the power limit of the GPU, but this also requires that the NVIDIA driver be loaded for full functionality. The power cap can be adjusted through the following asynchronous command:

Table 5. SMBPBI Commands

Specification	Value
Opcode	10h – Submit/poll asynchronous request
Arg1	0x01 – Set total GPU power limit
Arg2	0x00

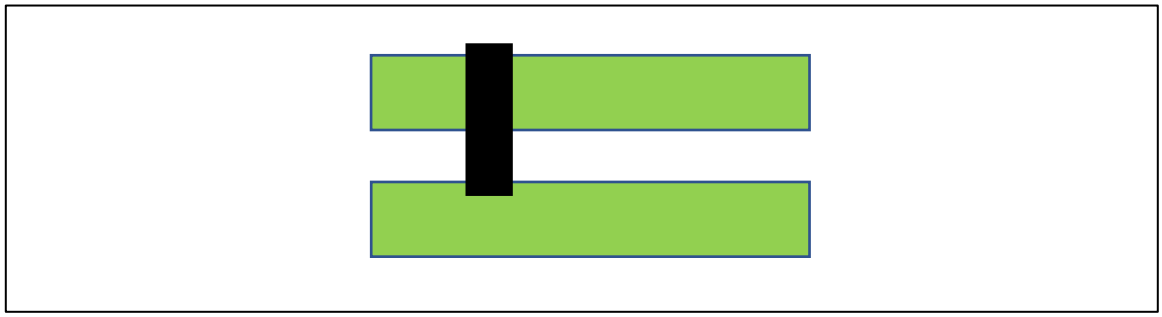
Using SMBPBI, the configured power limit setting can be made persistent across driver reloads. Refer to the *SMBus Post-Box Interface (SMBPBI) Design Guide* (DG-06034-002) for full implementation details.

NVLink Bridge Support

NVIDIA® NVLink® is a high-speed point-to-point peer transfer connection, where one GPU can transfer data to and receive data from one other GPU. The NVIDIA A30 card supports NVLink bridge connection with a single adjacent A30 card.

The attached bridge spans two PCIe slots. Wherever an adjacent pair of A30 cards exists in the server, for best bridging performance and balanced bridge topology, the A30 pair should be bridged.

Figure 3. A30 NVLink Connection – Top View



For systems that feature multiple CPUs, both A30 cards of a bridged card pair should be within the same CPU domain—that is, under the same CPU's topology. Ensuring this benefits workload application performance. There are exceptions, for example in a system with dual CPUs wherein each CPU has a single A30 PCIe card under it; in that case, the two A30 PCIe cards in the system may be bridged together.

NVLink Connector Placement

Figure 4 shows the connector keep-out area for the NVLink bridge support of the A30.

Figure 4. NVLink Connector Placement – Top View

NVIDIA A30 NVLink speed and bandwidth are given in the following table.

Table 6. NVLink Speed and Bandwidth

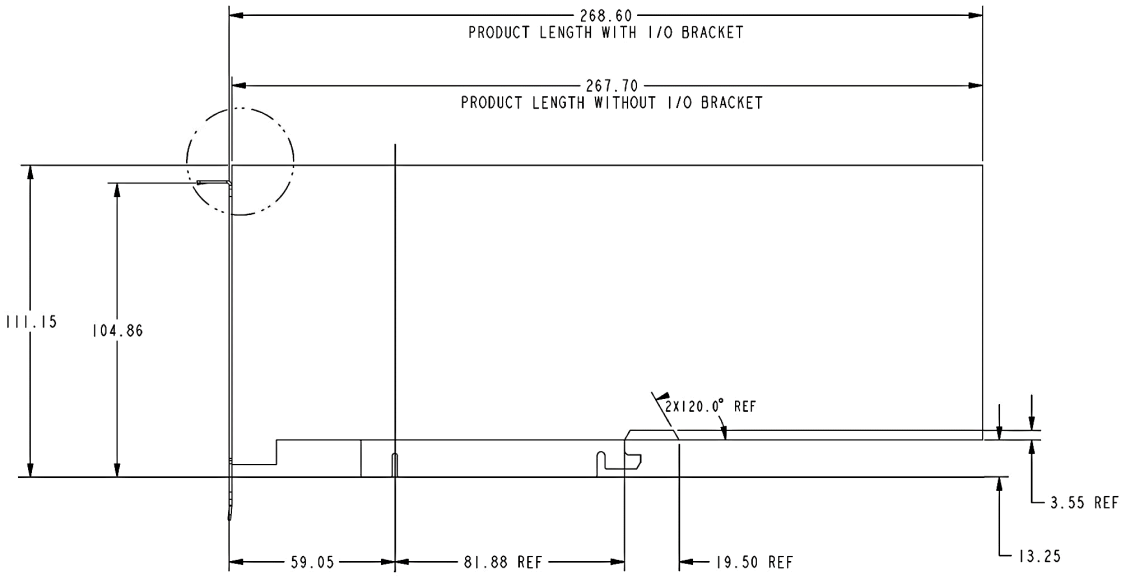
Parameter	Value
Total NVLink Bridges Supported by NVIDIA A30	1
Total NVLINK Rx and Tx lanes supported	32
Data rate per NVIDIA A30 NVLink lane (each direction)	50 Gbps
Total Maximum NVLink Bandwidth	200 GB/s

Sufficient clearance must be provided both above the card's north edge and behind the backside of the card's PCB to accommodate NVIDIA A30 NVLink bridge. The clearance above the north edge should meet or exceed 2.5 mm. The backside clearance (from the rear card's rear PCB surface) should meet or exceed 2.67 mm.

Form Factor

In this product brief, nominal dimensions are shown in Figure 5.

Figure 5. NVIDIA A30 PCIe Card Dimensions



Power Connector Placement

The board provides a CPU 8-pin power connector on the east edge of the board.

Figure 6. CPU 8-Pin Power Connector

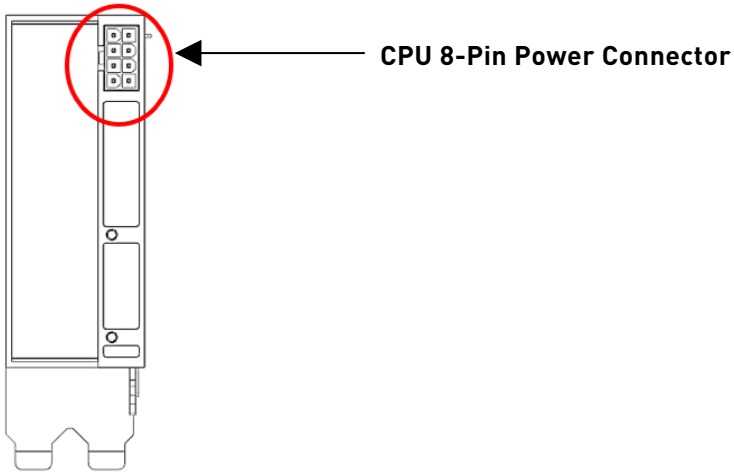


Table 7 lists supported auxiliary power connections for the NVIDIA A30 GPU card.

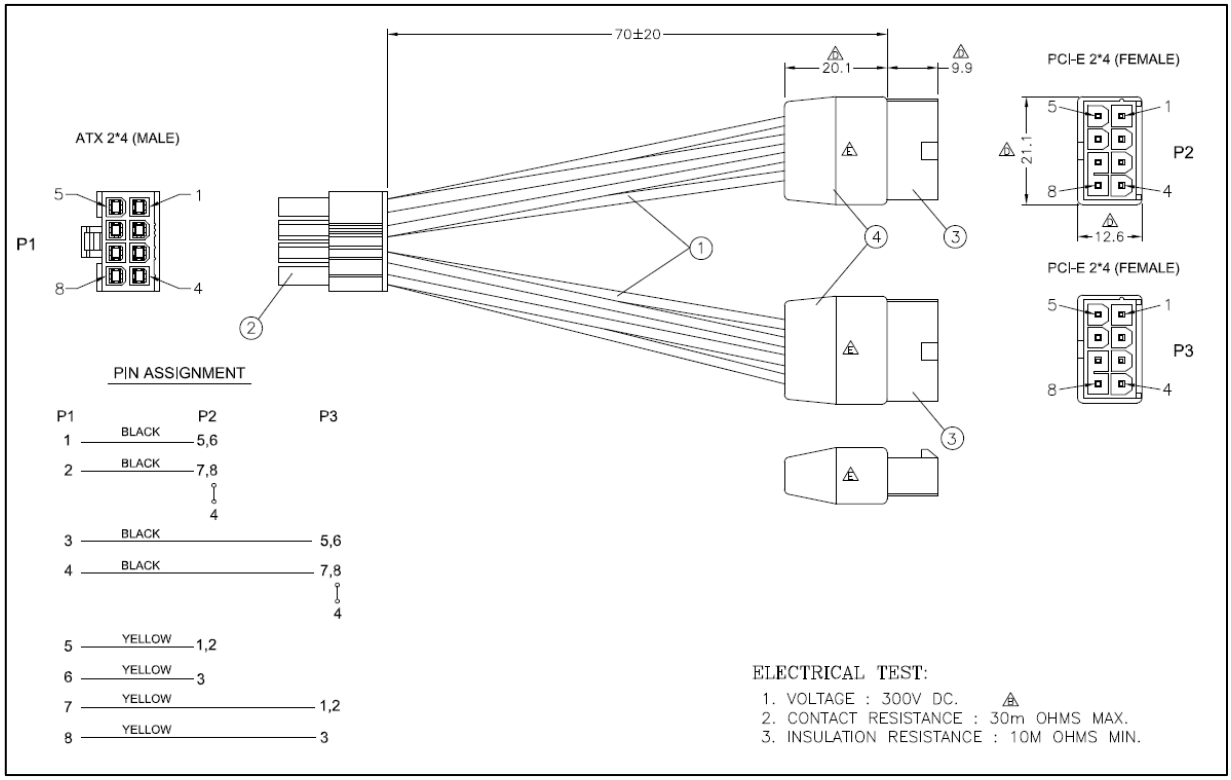
Table 7. Supported Auxiliary Power Connections

Board Connector	PSU Cable
CPU 8-pin	1× CPU 8-pin cable
CPU 8-pin	CPU 8-pin to PCIe 8-pin cable adapter

CPU 8-Pin to PCIe 8-Pin Power Adapter

Figure 7 lists the pin assignments of the power adapter.

Figure 7. CPU 8-Pin to PCIe 8-Pin Power Adapter



Extenders

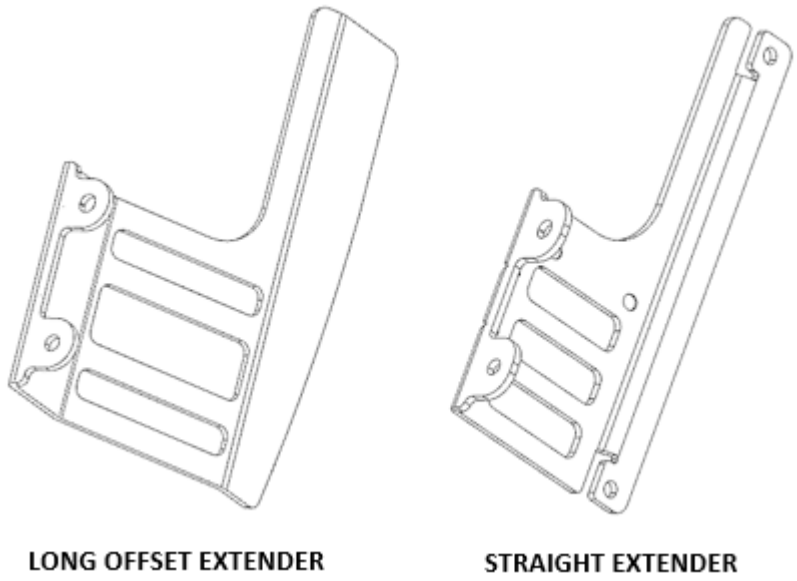
The A30 PCIe card provides two extender options, shown in Figure 8.

- NVPN: 682-00007-5555-000 – Long offset extender
  - Card + extender = 339 mm
- NVPN: 682-00007-5555-001 – Straight extender
  - Card + extender = 312 mm

Using the standard NVIDIA extender ensures greatest forward compatibility with future NVIDIA product offerings.

If the standard extender will not work, OEMs may design a custom attach method using the extender mounting holes on the east edge of the PCIe card.

Figure 8. Long Offset and Straight Extenders



Support Information

Certifications

- Windows Hardware Quality Lab (WHQL):
  - Certified Windows 7, Windows 8.1, Windows 10
  - Certified Windows Server 2008 R2, Windows Server 2012 R2
- Ergonomic requirements for office work W/VDTs (ISO 9241)
- EU Reduction of Hazardous Substances (EU RoHS)
- Joint Industry guide (J-STD) / Registration, Evaluation, Authorization, and Restriction of Chemical Substance (EU) – (JIG / REACH)
- Halogen Free (HF)
- EU Waste Electrical and Electronic Equipment (WEEE)

Agencies

- Australian Communications and Media Authority and New Zealand Radio Spectrum Management (RCM)
- Bureau of Standards, Metrology, and Inspection (BSMI)
- Conformité Européenne (CE)
- Federal Communications Commission (FCC)
- Industry Canada - Interference-Causing Equipment Standard (ICES)
- Korean Communications Commission (KCC)
- Underwriters Laboratories (cUL, UL)
- Voluntary Control Council for Interference (VCCI)

Languages

Table 8. Languages Supported

Languages	Windows[1]	Linux
English (US)	Yes	Yes
English (UK)	Yes	Yes
Arabic	Yes	
Chinese, Simplified	Yes	
Chinese, Traditional	Yes	
Czech	Yes	
Danish	Yes	
Dutch	Yes	
Finnish	Yes	
French (European)	Yes	
German	Yes	
Greek	Yes	
Hebrew	Yes	
Hungarian	Yes	
Italian	Yes	
Japanese	Yes	
Korean	Yes	
Norwegian	Yes	
Polish	Yes	
Portuguese (Brazil)	Yes	
Portuguese (European/Iberian)	Yes	
Russian	Yes	
Slovak	Yes	
Slovenian	Yes	
Spanish (European)	Yes	
Spanish (Latin America)	Yes	
Swedish	Yes	
Thai	Yes	
Turkish	Yes	

Note:

[1]Microsoft Windows 7, Windows 8, Windows 8.1, Windows 10, Windows Server 2008 R2, Windows Server 2012 R2, and Windows 2016 are supported.

Notice

This document is provided for information purposes only and shall not be regarded as a warranty of a certain functionality, condition, or quality of a product. NVIDIA Corporation ("NVIDIA") makes no representations or warranties, expressed or implied, as to the accuracy or completeness of the information contained in this document and assumes no responsibility for any errors contained herein. NVIDIA shall have no liability for the consequences or use of such information or for any infringement of patents or other rights of third parties that may result from its use. This document is not a commitment to develop, release, or deliver any Material (defined below), code, or functionality.

NVIDIA reserves the right to make corrections, modifications, enhancements, improvements, and any other changes to this document, at any time without notice.

Customer should obtain the latest relevant information before placing orders and should verify that such information is current and complete.

NVIDIA products are sold subject to the NVIDIA standard terms and conditions of sale supplied at the time of order acknowledgement, unless otherwise agreed in an individual sales agreement signed by authorized representatives of NVIDIA and customer ("Terms of Sale"). NVIDIA hereby expressly objects to applying any customer general terms and conditions with regards to the purchase of the NVIDIA product referenced in this document. No contractual obligations are formed either directly or indirectly by this document.

NVIDIA products are not designed, authorized, or warranted to be suitable for use in medical, military, aircraft, space, or life support equipment, nor in applications where failure or malfunction of the NVIDIA product can reasonably be expected to result in personal injury, death, or property or environmental damage. NVIDIA accepts no liability for inclusion and/or use of NVIDIA products in such equipment or applications and therefore such inclusion and/or use is at customer's own risk.

NVIDIA makes no representation or warranty that products based on this document will be suitable for any specified use. Testing of all parameters of each product is not necessarily performed by NVIDIA. It is customer's sole responsibility to evaluate and determine the applicability of any information contained in this document, ensure the product is suitable and fit for the application planned by customer, and perform 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 document. NVIDIA accepts no 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 document or (ii) customer product designs.

No license, either expressed or implied, is granted under any NVIDIA patent right, copyright, or other NVIDIA intellectual property right under this document. Information published by NVIDIA regarding third-party products or services does not constitute a license from NVIDIA to use such products or services or a warranty or endorsement thereof. Use of such information may require a license from a third party under the patents or other intellectual property rights of the third party, or a license from NVIDIA under the patents or other intellectual property rights of NVIDIA.

Reproduction of information in this document is permissible only if approved in advance by NVIDIA in writing, reproduced without alteration and in full compliance with all applicable export laws and regulations, and accompanied by all associated conditions, limitations, and notices.

THIS DOCUMENT AND ALL NVIDIA DESIGN SPECIFICATIONS, REFERENCE BOARDS, FILES, DRAWINGS, DIAGNOSTICS, LISTS, AND OTHER DOCUMENTS (TOGETHER AND SEPARATELY, "MATERIALS") ARE BEING PROVIDED "AS IS." NVIDIA MAKES NO WARRANTIES, EXPRESSED, IMPLIED, STATUTORY, OR OTHERWISE WITH RESPECT TO THE MATERIALS, AND EXPRESSLY DISCLAIMS ALL IMPLIED WARRANTIES OF NONINFRINGEMENT, MERCHANTABILITY, AND FITNESS FOR A PARTICULAR PURPOSE. TO THE EXTENT NOT PROHIBITED BY LAW, IN NO EVENT WILL NVIDIA BE LIABLE FOR ANY DAMAGES, INCLUDING WITHOUT LIMITATION ANY DIRECT, INDIRECT, SPECIAL, INCIDENTAL, PUNITIVE, OR CONSEQUENTIAL DAMAGES, HOWEVER CAUSED AND REGARDLESS OF THE THEORY OF LIABILITY, ARISING OUT OF ANY USE OF THIS DOCUMENT, EVEN IF NVIDIA HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES. Notwithstanding any damages that customer might incur for any reason whatsoever, NVIDIA's aggregate and cumulative liability towards customer for the products described herein shall be limited in accordance with the Terms of Sale for the product.

Trademarks

NVIDIA, the NVIDIA logo, CUDA, NGC-Ready, and NVLink are trademarks and/or registered trademarks of NVIDIA Corporation in the U.S. and other countries. Other company and product names may be trademarks of the respective companies with which they are associated.

Copyright

NVIDIA Corporation | 2788 San Tomas Expressway, Santa Clara, CA 95051
http://www.nvidia.com