NVIDIA GPU BOOST FOR TESLA

DA-06767-001_v03 | November 2014

Application Note

DOCUMENT CHANGE HISTORY

DA-06767-001_v03

Version	Date	Authors	Description of Change
01	March 28, 2013	GG, SM	Initial Release
02	January 20, 2014	GG, SM	•Updated product name •Added figures and a table •Added new sections •General updates throughout application note
03	November 11, 2014	GG, SM	Updated to include Tesla K80

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

INTRODUCTION

NVIDIA GPU Boost™ is a feature available on NVIDIA® GeForce,® NVIDIA® Quadro® and NVIDIA® Tesla® graphics processing units (GPUs) that boosts application performance by increasing GPU core and memory clock rates when sufficient power and thermal headroom are available. This application note is useful for anyone who wants to use NVIDIA GPU Boost on a Tesla K40 and Tesla K80 to maximize application performance.

> **Note:** Tesla K40 is available both as a workstation and server module. Within this application note, Tesla K40 refers to both of them.
>
> Tesla K80 is only available as a server module.

NVIDIA GPU BOOST FOR TESLA

The Tesla boards are designed for a specific power budget (for example, 235 W in the case of the Tesla K20X) assuming a highly optimized compute workload. However, HPC workloads vary in the power consumption and profile. The following chart in Figure 1 shows the average power consumption from various workloads measured on the Tesla K20X. This shows that several workloads are not using the full 235 W and hence have power headroom. NVIDIA GPU Boost for Tesla allows customers to use available power headroom to select higher graphics clocks using NVML or `nvidia-smi`.

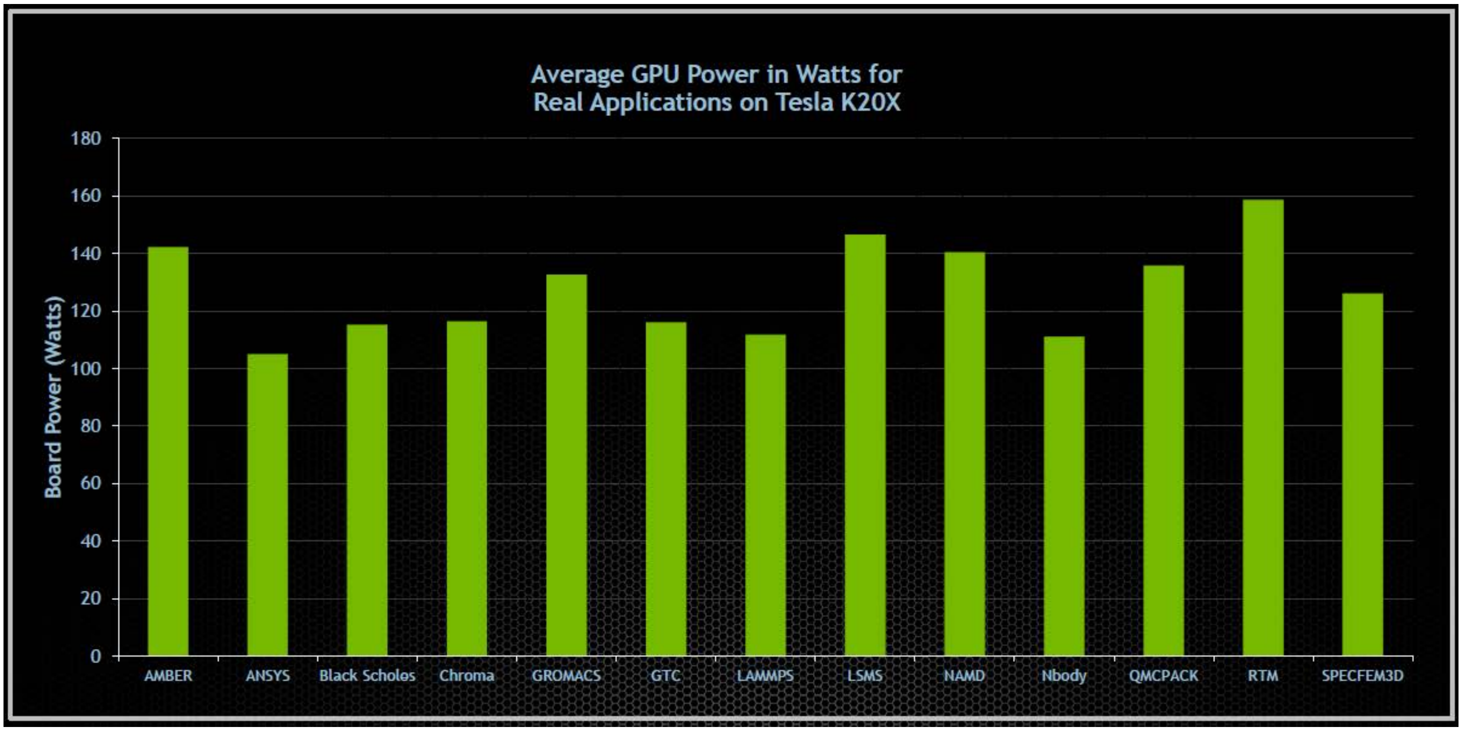

Figure 1. Average Power Consumption

When the GPU is in a lower performance (idle) state, the GPU clock is fixed. However, when the GPU is operating in a high performance state (P0), the highest GPU performance is typically desired. NVIDIA GPU Boost maximizes the GPU performance by automatically raising the GPU clock when there is thermal and power headroom available. Likewise, if the power or thermal limit is reached, the GPU clock scales down to the next available clock setting so that the board remains below the power and thermal limit.

NVIDIA products that support NVIDIA GPU Boost have multiple high-performance GPU clocks defined. That is, when the GPU is operating in its high performance mode (P0 state; determined automatically by the driver software), it has an array of GPU clocks available.

The GPU clocks available under NVIDIA GPU Boost are referred to as follows:

- **Base Clock:** Base clock is a clock defined to run the thermal design power (TDP) application under TDP test conditions (worst-case board under worst-case test conditions). For Tesla products, the TDP application is typically specified to be a variation of DGEMM.
- **Boost Clock(s):** These are the clocks above the base clock and they are available to the GPU when there is power headroom. The number of boost clocks supported, vary from product to product.

Note: In some cases, products may have clocks defined below the base clocks as a protection against extreme power spikes situations.

NVIDIA GPU BOOST FOR HPC WORKLOADS

NVIDIA GPU Boost for Tesla K40 and Tesla K80 is optimized to deliver a robust and deterministic boost behavior for a wide range of HPC workloads.

For Tesla K40 and Tesla K80 the end users have an option to select the NVIDIA GPU Boost behavior and the GPU clock frequency based on the workload characteristics. The workload may have one or more of the following characteristics.

- Problem set is spread across multiple GPUs and requires periodic synchronization.
- Problem set spread across multiple GPUs and runs independent of each other.
- Workload has “compute spikes.” For example, some portions of the workload are extremely compute intensive pushing the power higher and some portions are moderate.
- Workload is compute intensive throughout without any spikes.

- Workload requires fixed clocks and is sensitive to clocks fluctuating during the execution.
- Workload runs in a cluster where all GPUs need to start, finish, and run at the same clocks.
- Workload or end user requires predictable performance and repeatable results.
- Datacenter is used to run different types of workload at different hours in a day to better manage the power consumption.
- Some boards in a cluster have access to better cooling than others.

NVIDIA GPU BOOST IN TESLA K40

The Tesla K40 ships with the GPU clock set to the base clock. To enable the NVIDIA GPU Boost, the end user can use the NVML or `nvidia-smi` to select one of the available GPU clocks or boost levels.

In the case of Tesla K40 the clocks available to the end user are:

- Base Clock: 745 MHz
- Boost Clock 1: 810 MHz
- Boost Clock 2: 875 MHz

NVIDIA GPU BOOST IN TESLA K80

The Tesla K80 ships with Autoboost. This means that NVIDIA GPU Boost is enabled by default. This ensures that out of the box, the Tesla K80 will always try to achieve the best possible GPU clock and maximize performance for a given workload, power and thermal condition. At any point in time, an end user can disable this behavior via NVML or `nvidia-smi`.

Figure 2 and Table 1 summarizes the NVIDIA GPU Boost behavior and features for Tesla K40 and Tesla K80.

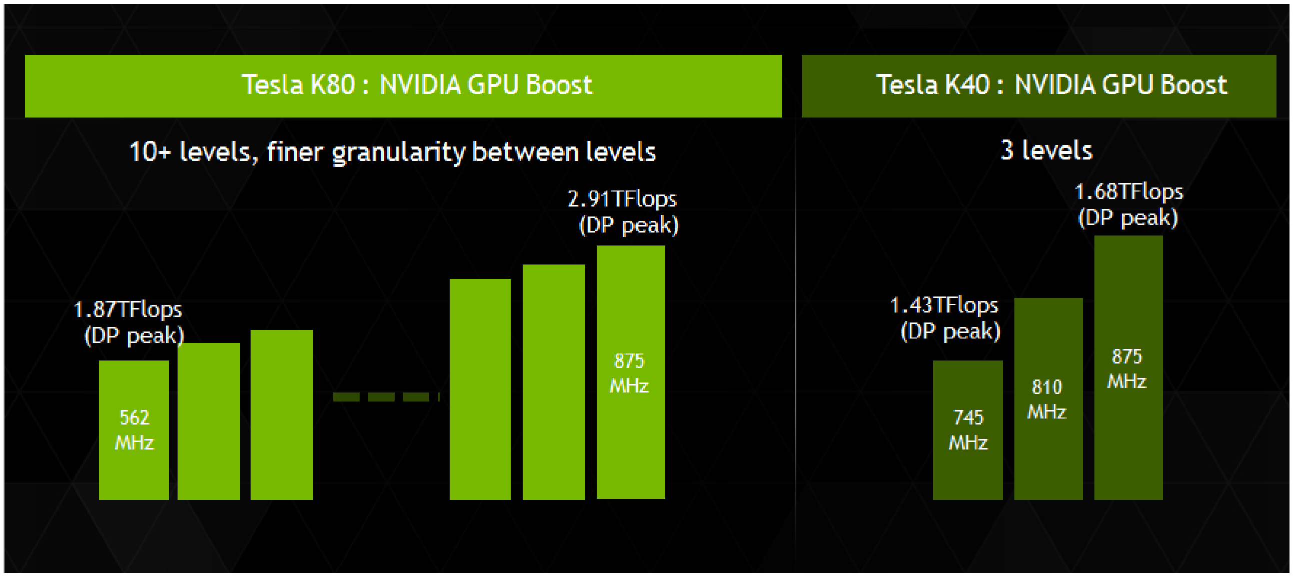

Figure 2. NVIDIA GPU Boost Levels in Tesla K80 and Tesla K40

Table 1. Summary of NVIDIA GPU Boost Features in Tesla K40 and Tesla K80

Feature	**Tesla K40**	**Tesla K80**
GPU clocks	•745 MHz •810 MHz •875 MHz	562 MHz to 875 MHz at 13 MHz increments
Base clock	735 MHz	560 MHz
Autoboost: NVIDIA GPU Boost enabled by default	No. End user has to explicitly select using nvidia-smi/NVML	Yes. Enabled by default to boost the clock based on power headroom
Ability to select clocks via nvidia-smi/NVML	Yes	Yes
Ability to disable NVIDIA GPU Boost	Yes. Via nvidia-smi/NVML	Yes. Via nvidia-smi/NVML

API FOR NVIDIA GPU BOOST ON TESLA

The Tesla K40 and Tesla K80 give full control to end-users to select the core clock frequency via NVML or `nvidia-smi`. NVML is a C-based API for monitoring and managing the various states of Tesla products. It provides a direct access to submit queries and commands via `nvidia-smi`. NVML documentation is available at https://developer.nvidia.com/nvidia-management-library-nvml

Table 2 gives a summary of the `nvidia-smi` commands for using NVIDIA GPU Boost on Tesla.

Table 2. nvidia-smi Commands

Usage	Command
View the clocks the Tesla board supports	nvidia-smi –q –d SUPPORTED_CLOCKS
Set one of the supported clocks	nvidia-smi -ac <MEM clock, Graphics clock>
Make the clock settings persistent across driver unload	nvidia-smi -pm 1
Make the clock settings revert to base clocks after driver unloads (or turn off the persistent mode)	nvidia-smi -pm 0
To view the clock in use, use the command	nvidia-smi -q –d CLOCK
To reset clocks back to the base clock (as specified in the board specification)	nvidia-smi –rac
To allow “non-root” access to change graphics clock	nvidia-smi -acp 0
Enable auto boosting the GPU clocks	nvidia-smi --auto-boost-default=ENABLED -i 1
Disable auto boosting the GPU clocks	nvidia-smi --auto-boost-default=ENABLED -i 0
To allow “non-root” access to set autoboost	nvidia-smi --auto-boost-permission=UNRESTRICTED -i 0

When using non-default clocks, driver persistence mode should be enabled. Persistence mode ensures that the driver stays loaded even when no NVIDIA® CUDA® or X applications are running on the GPU. This maintains current state, including requested boost clocks. If persistent mode is not enabled, and no applications are using the GPU, the driver will unload and any current user settings will revert back to default for the next application. To ensure that the next application also runs at boost clocks, select persistent mode using `"nvidia-smi -pm 1"` As with application clocks, this setting requires administrative privileges, and the GPU should have persistent mode enabled. Autoboost permissions can be relaxed similarly to application clock permissions.

```
sudo nvidia-smi --auto-boost-permission=UNRESTRICTED -i 0
```

The driver will attempt to maintain requested applications clocks whenever a CUDA context is running on the GPU. However, if no contexts are running the GPU will revert back to idle clocks to save power and will stay there until the next context is created. Thus, if the GPU is not busy, you may see idle current clocks even though requested applications clocks are much higher.

Note: By default changing the application clocks requires root access. If the user does not have root access, the user can request his or her cluster manager to allow non-root control over application clocks. Once changed, this setting will persist for the life of the driver before reverting back to root-only defaults. Persistence mode should always be enabled whenever changing application clocks, or enabling non-root permissions to do so.

APPLICATION BEHAVIOR WITH NVIDIA GPU BOOST FOR TESLA

In the previous sections we learned about various types of application characteristics and the APIs to use. Let's take a look at a few scenarios to explain what an end user might see when one of the boost clocks are selected on the Tesla K40 and Tesla K80.

It's highly likely that the application exhibits a combination of those during its entire execution period. The following scenarios can serve as a reference to understand the application behavior with NVIDIA GPU Boost on either the Tesla K40 or Tesla K80.

An important point to remember is that no matter which clocks the end user selects, if at any time the power monitoring algorithm detects that the application may exceed the board power or thermal limit, the GPU comes down to a lower clock level as a precaution.

SCENARIO 1: USER SELECTS BASE CLOCK

The GPU will run at the base clock for the entire duration. After completion, the next job will also run at the same base clock. As long as the board does not exceed power and thermal limit, the GPU will run at the base clock. Even if there is power headroom, the GPU will not select the boost clock automatically. This is by design and well suited for workloads that may be running on multiple GPUs and require all GPUs to run in lock-step. If during the run the board starts exceeding power/thermal limit, the power monitoring algorithm may lower the GPU clock for a brief period as a precaution and bring it back up to the base clock once the power spike comes down.

SCENARIO 2: USER SELECTS BOOST CLOCK WITHOUT SELECTING PRESISTENT MODE

The GPU will run at the selected boost clock for the entire duration of the workload. As long as the board does not exceed the power/thermal limit, the GPU will run at selected clock.

If during the run the board starts exceeding power/thermal limit, the power monitoring algorithm may lower the GPU clock for a brief period as a precaution and bring it back up to the selected clock once the power spike comes down. After completion when the driver unloads, the GPU will revert back to the base clock.

Note: Using persistent mode when setting boost clocks is highly recommended.

SCENARIO 3: USER SELECTS BOOST CLOCK AND SPECIFIES PERSISTENT MODE

The GPU will run at selected clock for the entire duration of the workload. After completion when the driver unloads, the GPU will start the next job also at a selected clock. In this scenario the GPU behaves as if it has to always run at a selected clock unless the end user selects a different clock or removes the persistent option. Under this scenario the clocks behavior will be as shown in Figure 3.

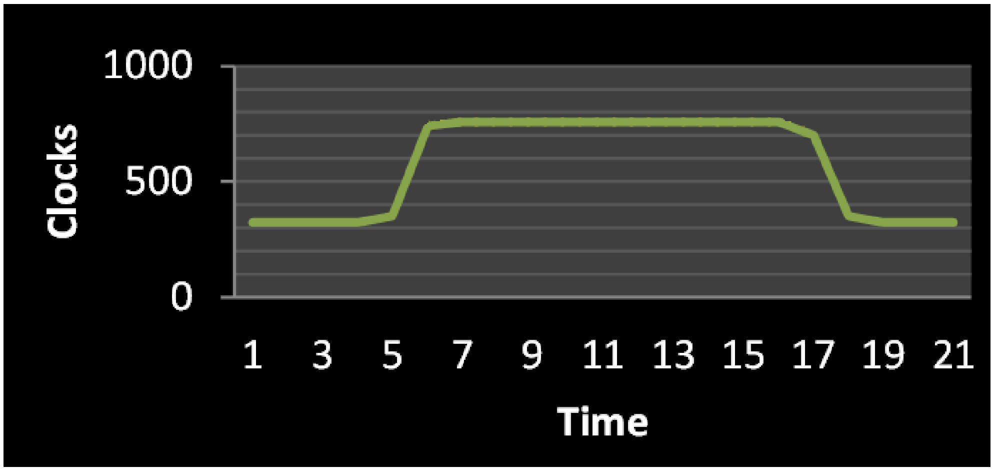

Figure 3. Tesla K40 and Tesla K80 with a Specific Boost Clock

SCENARIO 4: AUTOBOOST ON TESLA K80

Only Tesla K80 supports Autoboost, where the GPU clocks boost dynamically as they detect power headroom. As the application runs, and depending on the power and thermal profile, the GPU clocks will dynamically adjust as shown in the Figure 4 while staying within the power limit of 300 W.

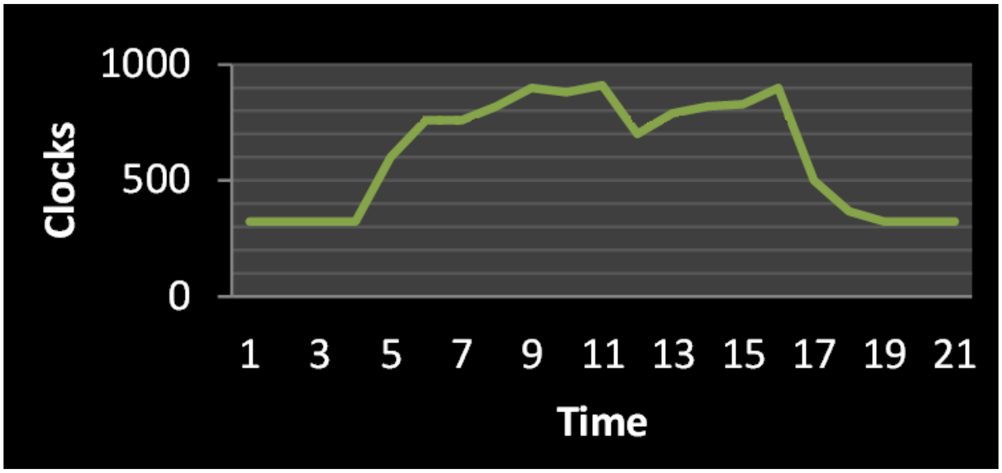

Figure 4. Tesla K80 Clocks with Autoboost Enabled

NVIDIA GPU BOOST AND MEMORY BANDWIDTH

NVIDIA GPU Boost capability allows end users to specific the boost clock which is just the core clock. However, selecting higher boost clocks does improve the effective memory bandwidth utilization for workloads that are sensitive to memory bandwidth. With higher boost clocks some workloads may even see improved PCIe transfer rates. Therefore, NVIDIA GPU Boost on the Tesla K40 and Tesla K80 helps workloads which are sensitive to core clocks, power headroom and also helps workloads that may be more sensitive to memory bandwidth than core clocks.

BEST PRACTICES FOR USING NVIDIA GPU BOOST ON TESLA K40 AND TESLA K80

- Tesla K40: Out of the box, the Tesla K40 will run at a base clock of 745 MHz The end user should try the base clock and check the power draw using the NVML or `nvidia-smi` query. If the power draw is less than 235 W, the end user can select a higher boost clock and re-run the application. This may require a few iterations and experimentation to see what boost clock works the best for a specific workload.
- Tesla K80: The board ships with autoboost. Out of the box, the GPU will start boosting the clock depending on the power headroom.
- If the end user is sharing the Tesla K40 and Tesla K80 with several others in a cluster, the end user may need root access to try and set different clocks. In that case, the end user can request the IT manager to use the following command to grant permission to the end user to set different boost clocks:

  ```
  nvidia-smi -acp 0
  ```
- If the workload runs on multiple GPUs and is sensitive to all GPUs running at the same clock, then the user may need to try out which particular clock works best for all GPUs.
- If the workload is such that each GPU works independently on a problem set and there's little interaction or collaboration between GPUs, then selecting the highest boost clock on a Tesla K40 or running Tesla K80 with autoboost may be the best option.

Notice

The information provided in this specification is believed to be accurate and reliable as of the date provided. However, NVIDIA Corporation (“NVIDIA”) does not give any representations or warranties, expressed or implied, as to the accuracy or completeness of such information. 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 publication supersedes and replaces all other specifications for the product that may have been previously supplied.

NVIDIA reserves the right to make corrections, modifications, enhancements, improvements, and other changes to this specification, at any time and/or to discontinue any product or service without notice. Customer should obtain the latest relevant specification 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. NVIDIA hereby expressly objects to applying any customer general terms and conditions with regard to the purchase of the NVIDIA product referenced in this specification.

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 these specifications 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 specification. 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 specification, 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 specification. 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 specification 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.

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. 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 NVIDIA terms and conditions of sale for the product.

Trademarks

NVIDIA, the NVIDIA logo, CUDA, GeForce, NVIDIA GPU Boost, Quadro, and Tesla 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