NVIDIA AI Enterprise Test Drive

Application Note

SWE-NVAIE-001-APNT | July 2021

Document History

SWE-NVAIE-001-APNT

Version	Date	Authors	Description of Change
01	30 July 2021	VNK, JC, JL, CW	Initial release

Table of Contents

List of Figures

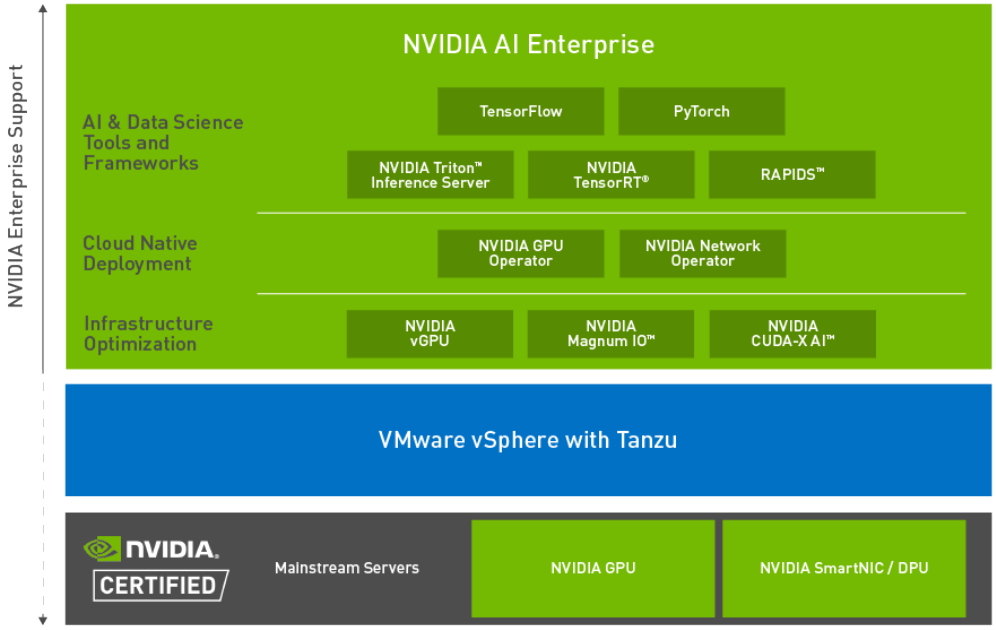
Chapter 1. Introduction

Artificial intelligence (AI) is transforming every industry, whether it is by improving customer relationships in financial services, streamlining manufacturer supply chains, or helping doctors deliver better outcomes for patients. While most organizations know they need to invest in AI to secure their future, they struggle with finding the strategy and platform that can enable success.

Unlike traditional enterprise applications, AI applications are a relatively recent development for many IT departments. They are anchored in rapidly evolving, open-source, bleeding-edge code and lack proven approaches that meet the rigors of scaled production settings in enterprises.

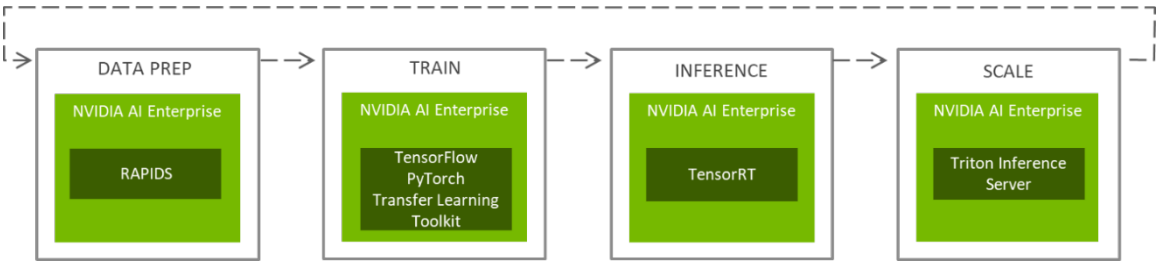
VMware and NVIDIA have partnered to unlock the power of AI for every business by delivering an end-to-end enterprise platform optimized for AI workloads. This integrated platform delivers best-in-class AI software, the NVIDIA AI Enterprise Suite, optimized and exclusively certified for the industry's leading virtualization platform, VMware vSphere®. Running on NVIDIA-Certified Systems™, industry-leading accelerated servers, this platform accelerates the speed at which developers can build AI and high-performance data analytics, enables organizations to scale modern workloads on the same VMware vSphere infrastructure they have already invested in, and delivers enterprise-class manageability, security, and availability.

Figure 1-1. NVIDIA AI Enterprise Software Suite



NVIDIA AI Enterprise is an end-to-end AI software suite that includes AI frameworks and tools that provide performance-optimized deep learning, machine learning, and data science tools that simplify building, sharing, and deploying AI software, so enterprises can gather insights faster and deliver business value sooner.

Figure 1-2. Example of Data Science Workflow



Data Prep – Data science software such as the RAPIDS suite of software libraries, built on CUDA-X AI, enable data scientists to execute end-to-end data science and analytics pipelines entirely on GPUs.

AI Training – Deep learning frameworks for training and machine learning, such as PyTorch and TensorFlow, are optimized for GPU acceleration. With the NVIDIA Transfer Learning Toolkit, enterprises can take these pre-trained models and refine them for their own usage

domain, greatly reducing development time. Leveraging these tools and pre-trained models, can reduce development time by up to 10x, from 80 weeks to 8 weeks, for example.

AI Inference – The newly trained models can be optimized using deep learning inference SDKs and libraries such as NVIDIA TensorRT. This is done by fusing layers and eliminating unneeded steps.

Scale – Triton Inference Server simplifies and optimizes the deployment of AI models at scale in production. It integrates with Kubernetes for orchestration and auto-scaling and allows front-end client applications to submit inference requests from an AI inference cluster and can service models from an AI model repository.

The NVIDIA AI Enterprise Test Drive is not a commercially available product. It is a technology demonstration developed by NVIDIA and VMware to showcase the benefits of adding NVIDIA AI Enterprise software suite to boost AI performance, deploy with confidence, and scale without compromise.

Take 48 hours to test drive a high-performance enterprise-ready AI experience.

Chapter 2. Getting Started with NVIDIA AI Enterprise Test Drive

Your personal virtual machine (VM) has been preconfigured with demos that showcase the benefits of the NVIDIA AI Enterprise Software Suite. These applications are categorized by type and found on the **Home Tabs** of the **Google Chrome Browser**.

To run any of these demos, simply click on any of the **Tabs** after launching Google Chrome.

The demos are divided into the following categories:

- Inference
- Training

The NVIDIA AI Enterprise Software Suite includes a curated set of AI and data science frameworks and tools, NVIDIA operators for cloud native deployment, and infrastructure optimization software.

Chapter 3. Testing the NVIDIA and VMware Test Drive

3.1 Data Prep (RAPIDS)

RAPIDS is a suite of GPU-accelerated data science libraries with APIs that should be familiar to users of Pandas, Dask, and Scikitlearn. This demo focuses on showing how to use cuDF with Dask and XGBoost to scale GPU DataFrame ETL-style operations. Anaconda has graciously made some of the NYC Taxi dataset available in a public Google Cloud Storage bucket. We use our vGPU-enabled VM to process it and train a model that predicts the fare amount.

3.2 Training (TensorFlow)

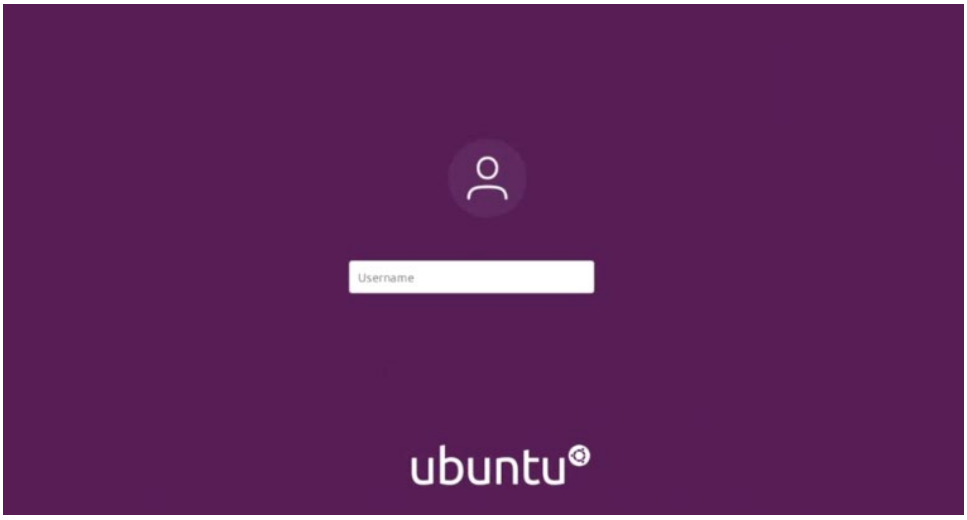
Bidirectional Embedding Representations from Transformers (BERT) is a method of pre-training language representations that obtains state-of-the-art results on a wide array of natural language processing (NLP) tasks. The original paper can be found here: https://arxiv.org/abs/1810.04805.

NVIDIA's BERT is an optimized version of Google's official implementation, leveraging mixed precision arithmetic target accuracy. This demo will demonstrate inference on question/answering tasks and the use of mixed precision models.

3.2.1 Launching NVIDIA AI Enterprise Test Drive

After login, click **NVIDIA AI Enterprise Test Drive** and launch the Ubuntu Desktop.

Figure 3-1. NVIDIA AI Enterprise Test Drive – Ubuntu Desktop



Login credentials are as follows:

Username: nvidia

Password: nvidiaAI!

Once the Ubuntu desktop is launched, open **Google Chrome** from the favorites **left side bar** for access to the demos.

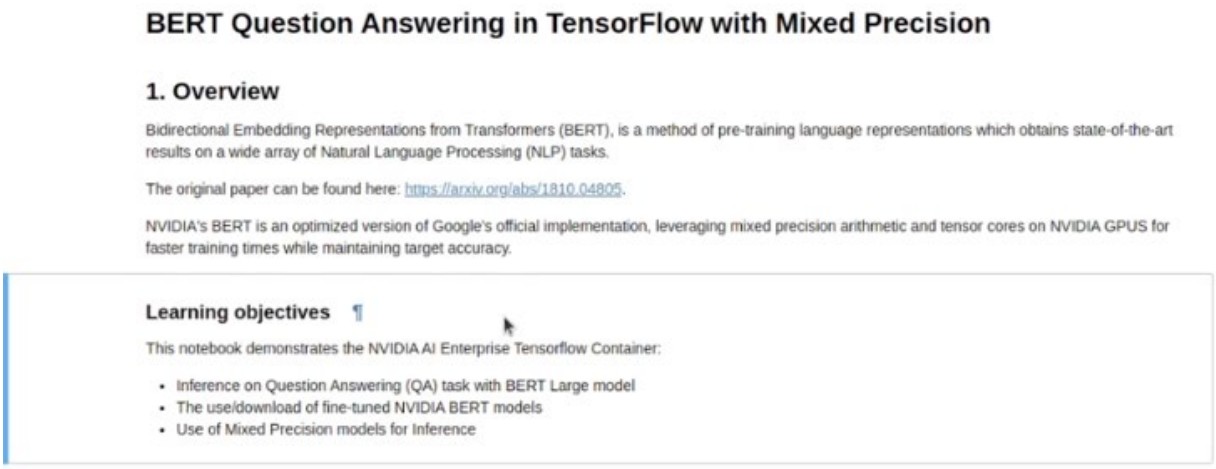
Chapter 4. Running the NVIDIA AI Enterprise Test Drive Demos

This section describes how to execute the pre-installed demos in the NVIDIA AI Enterprise Test Drive.

- Training (TensorFlow)
- Bert Q&A System Data Prep (RAPIDS) – The section goes through the listed rich web experience demos.
- Data Prep (RAPIDS)
  - NYC Taxi Fair

4.1 Training (TensorFlow)

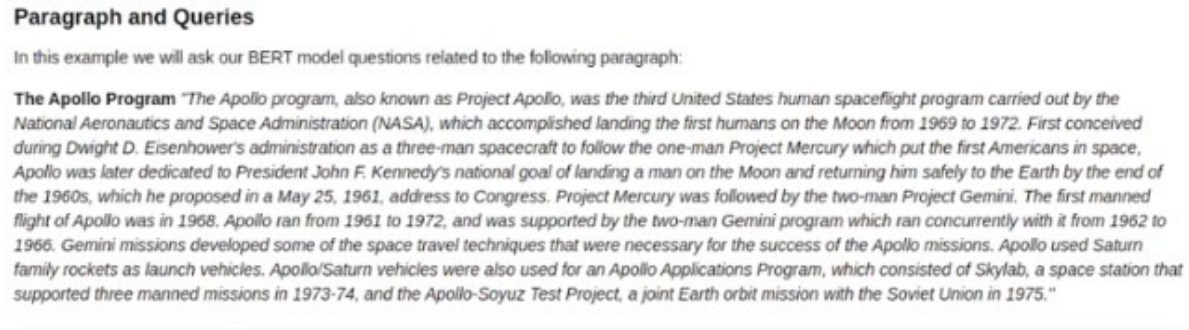
Figure 4-1. BERT Question/Answer



To interact with this demo, click on each step then press Shift + Enter to execute each Jupyter notebook text or code block.

During this demo, we demonstrate inference on the paragraph below:

Figure 4-2. BERT Model Example Paragraph



Paragraph and Queries

In this example we will ask our BERT model questions related to the following paragraph:

The Apollo Program *"The Apollo program, also known as Project Apollo, was the third United States human spaceflight program carried out by the National Aeronautics and Space Administration (NASA), which accomplished landing the first humans on the Moon from 1969 to 1972. First conceived during Dwight D. Eisenhower's administration as a three-man spacecraft to follow the one-man Project Mercury which put the first Americans in space, Apollo was later dedicated to President John F. Kennedy's national goal of landing a man on the Moon and returning him safely to the Earth by the end of the 1960s, which he proposed in a May 25, 1961, address to Congress. Project Mercury was followed by the two-man Project Gemini. The first manned flight of Apollo was in 1968. Apollo ran from 1961 to 1972, and was supported by the two-man Gemini program which ran concurrently with it from 1962 to 1966. Gemini missions developed some of the space travel techniques that were necessary for the success of the Apollo missions. Apollo used Saturn family rockets as launch vehicles. Apollo/Saturn vehicles were also used for an Apollo Applications Program, which consisted of Skylab, a space station that supported three manned missions in 1973-74, and the Apollo-Soyuz Test Project, a joint Earth orbit mission with the Soviet Union in 1975."*

The demo answers the questions provided.

Figure 4-3. BERT Demo Provided Questions

In [4]:
```
# Create BERT input file with (1) context and (2) questions to be answered based on that context
predict_file = '/data/workspace/qa/bert/config.qa/input.json'
```

In [ ]:
```
%%writefile $predict_file
{"data":
 [
     {"title": "Project Apollo",
      "paragraphs": [
          {"context":"The Apollo program, also known as Project Apollo, was the third United States human spaceflig
           "qas": [
               { "question": "What project put the first Americans into space?",
                 "id": "Q1"
               },
               { "question": "What program was created to carry out these projects and missions?",
                 "id": "Q2"
               },
               { "question": "What year did the first manned Apollo flight occur?",
                 "id": "Q3"
               },
               { "question": "What President is credited with the notion of putting Americans on the moon?",
                 "id": "Q4"
               },
               { "question": "Who did the U.S. collaborate with on an Earth orbit mission in 1975?",
                 "id": "Q5"
               },
               { "question": "How long did Project Apollo run?",
                 "id": "Q6"
               },
               { "question": "What program helped develop space travel techniques that Project Apollo used?",
                 "id": "Q7"
               },
               {"question": "What space station supported three manned missions in 1973-1974?",
                 "id": "Q8"
               }
]}]}]}
```

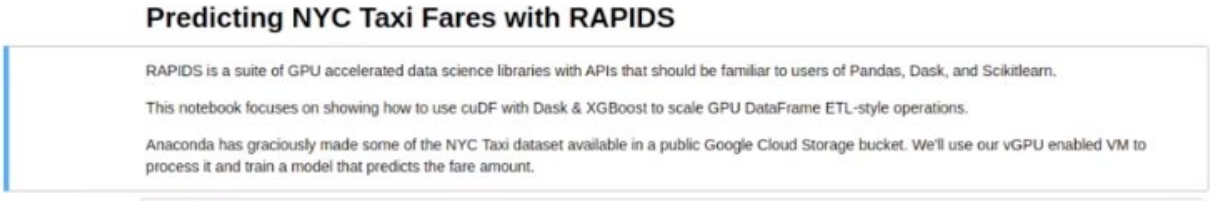
After running inference, you are able to display and compare the Question/Answer results.

You can also customize the number of questions, content, and questions. At the end, the answers are displayed based on the content and questions provided.

Figure 4-4. BERT Demo Custom Inputs

4.2 Data Prep (RAPIDS)

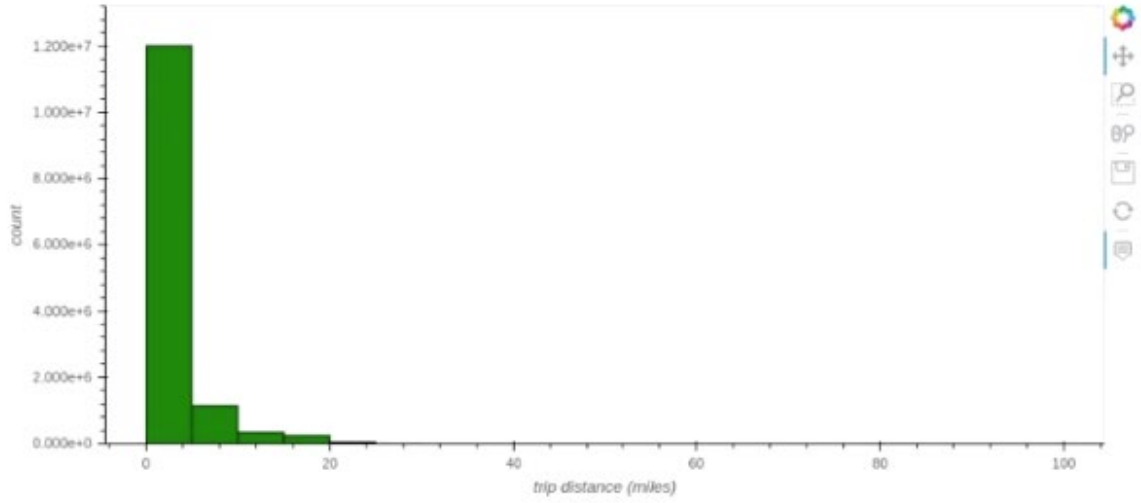
Figure 4-5. Taxi Fare Demo Overview



To interact with this demo, click on each step then press Shift + Enter to execute each Jupyter notebook text or code block.

First, you install tools needed for the demo. This should take about 1-2 minutes. This demo will go through the steps to inspect, clean up, and analyze the taxi fare data. You can also display the data based on distance, fare, and passenger count.

Figure 4-6. Example of Taxi Fare Data



You can perform training based on the specified training set. Notice that the wall time is under 12 seconds, significantly faster than a CPU-only VM.

Figure 4-7. Training and Wall Time

Train the XGBoost Regression Model

The wall time output below indicates how long it took your GPU cluster to train an XGBoost model over the training set.

In [23]:
```
dtrain = xgb.dask.DaskDMatrix(client, X_train, Y_train)
```

In [24]:
```
%%time

trained_model = xgb.dask.train(client,
                               {
                                'learning_rate': 0.3,
                                 'max_depth': 8,
                                 'objective': 'reg:squarederror',
                                 'subsample': 0.6,
                                 'gamma': 1,
                                 'silent': True,
                                 'verbose_eval': True,
                                 'tree_method':'gpu_hist'
                               },
                               dtrain,
                               num_boost_round=100, evals=[(dtrain, 'train')])
```

```
CPU times: user 248 ms, sys: 46.3 ms, total: 294 ms
Wall time: 11.5 s
```

You can then compare the prediction to the actual fare.

Figure 4-8. Comparison of Taxi Fare Prediction and Actual Data

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.

VESA DisplayPort

DisplayPort and DisplayPort Compliance Logo, DisplayPort Compliance Logo for Dual-mode Sources, and DisplayPort Compliance Logo for Active Cables are trademarks owned by the Video Electronics Standards Association in the United States and other countries.

HDMI

HDMI, the HDMI logo, and High-Definition Multimedia Interface are trademarks or registered trademarks of HDMI Licensing LLC.

Arm

Arm, AMBA and Arm Powered are registered trademarks of Arm Limited. Cortex, MPCore and Mali are trademarks of Arm Limited. All other brands or product names are the property of their respective holders. "Arm" is used to represent Arm Holdings plc; its operating company Arm Limited; and the regional subsidiaries Arm Inc.; Arm KK; Arm Korea Limited.; Arm Taiwan Limited; Arm France SAS; Arm Consulting (Shanghai) Co. Ltd.; Arm Germany GmbH; Arm Embedded Technologies Pvt. Ltd.; Arm Norway, AS and Arm Sweden AB.

OpenCL

OpenCL is a trademark of Apple Inc. used under license to the Khronos Group Inc.

Trademarks

NVIDIA, the NVIDIA logo, NVIDIA AI Enterprise 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