White Paper

Training, Simulation, Testing for Autonomous Driving Infrastructure

Sizing and Software for Autonomous Driving Infrastructure

2018-08-29

Contents

Abstract ii

1 Data Factory 1

2 AI Model Training 2

3 Simulation 3

4 Testing 4

5 DRIVE Constellation POD 5

6 Summary 7

Abstract

The NVIDIA® Training Simulation Testing for Autonomous Driving Infrastructure (TSTADI) reference architecture describes the infrastructure hardware and software needed for building autonomous vehicle (AV) software. TSTADI is based on the NVIDIA DGX™ POD reference architecture used in the NVIDIA DGX SATURNV supercomputer and defines a new NVIDIA DRIVE Constellation™ POD for hardware-in-the-loop (HIL) simulation. The reference architecture also includes a data factory to pre-process and label the hundreds of petabytes of data collected and simulated during the AV development process. Sizing guidelines for a customer implementation of TSTADI are given based on the experience of several thousand NVIDIA engineers using the platform.

Figure 1 shows the main TSTADI components and workflow described in detail in later section of this paper. A fleet of data collection vehicles generate raw data which is fed into a data factory where it is pre-processed and labeled. AI model training then executes on DGX PODs using a petabyte-sized data cache. Additional data is generated by HIL simulation running on DRIVE Constellation PODs. Finally, completed models need to be regularly tested, both during the AV development processes and on an ongoing basis to provide updated software via over-the-air (OTA) updates to AV fleets. This is an iterative process that needs to be repeated many times.

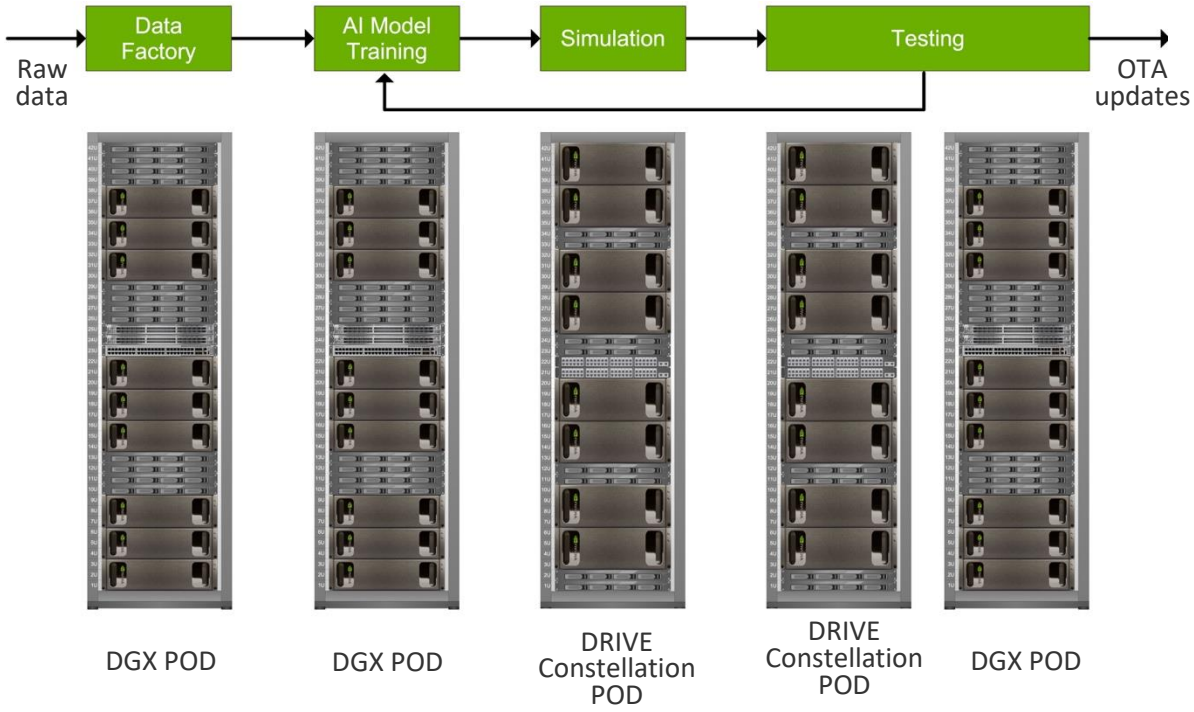

Figure 1. TSTADI workflow

Sizing demands will increase over time. An AV development process can begin with a few DGX PODs and DRIVE Constellation PODs and grow over time as additional data is collected.

1 Data Factory

The data factory consists of infrastructure for data ingestion from vehicles, computing resources for transcoding and compression, labeling services, and a data lake — a centralized repository for storage and retrieval of all AV data including sensor inputs, labeled data, simulation data, performance metrics, and trained models.

Long-term storage required for ongoing upgrade and maintenance of AV software will range from tens to hundreds of petabytes depending on the number of data collection vehicles used. Data storage can be either cloud-based or on-premises. TSTADI allows the training, simulation, and testing workflows to operate physically separate from the data factory by using a petabyte-sized local cache in each DGX and DRIVE Constellation POD. This allows customers to use the best combination of cloud-based or on-premise as required by their individual requirements.

Sizing begins with the number and types of sensors used in data collection vehicles. A typical data collection vehicle for autonomous driving uses multiple sensors which may include cameras, radar, and lidar. Since a significant amount of redundant information is collected by cars in the AV fleet, only about one out of every 1,000 images need to be labeled. DGX-1 servers are used to inference all data and select the images for labeling using pre-trained AI models.

In the example shown in Table 1, 30 DGX-1 servers are required in a data factory for a single data collection car driving 2,000 hours over one year (eight hours a day and 250 days per year).

Sensor type	Resolution	Frames per second	Sensors per car	Images per year	Storage needed	Labeled images	DGX-1 servers
Camera	2 MPixel	30	5	1B	2 PB	1M	25
Lidar	0.13 MPixel	10	3	200M	400 TB	200K	5

Table 1. Data factory sizing

As AI model training progresses and as collection vehicles are added, turn-around time (TAT) can be lengthened to support additional collection vehicles without adding additional DGX-1 servers. Storage, however, continues to grow linearly as all data from all vehicles needs to be retained for future AI model maintenance.

2 AI Model Training

The AI model training workflow is as follows:

1. Data is selected and labeled.
2. The labeled data is used to train the model. This process is highly iterative.
3. The trained model is used to inform the next selection of data to be labeled. This continues until the desired model performance is achieved.
4. The model is validated against a large and growing collection of test data.

Since AI training algorithms iterate many times on the same sets of training data, a local multi-tier caching system provides the training data to the GPUs at high speeds. Each DGX POD includes 1 PB of data cache.

A single DGX-1 server can complete an AI model training experiment on 300,000 images using ResNet-18 in about one day. For the shortest development times, ten experiments are run in parallel on the AI model. An AV development program training ten AI models thus requires 100 DGX-1 servers.

3 Simulation

Given the high-risk and costly nature of testing AVs in the real-world, it is necessary to leverage simulation for development, regression testing, and validation of AV systems. Newly developed or improved AI models need to be tested for performance and regressions against real images from millions of driven miles or against synthetically generated 3D GPU rendered images.

Both SIL (software-in-the-loop) and HIL (hardware-in-the-loop) systems can be used when testing AV systems. Most simulation, however, will be done with HIL as it provides the closest environment to real-world driving using DRIVE Constellation systems. SIL will still occasionally be used for R&D of new simulation scenarios as playback speeds can exceed real-time. Since this testing can be done using spare capacity of the DGX PODs used for AI Model training, additional DGX PODs for SIL simulation are not covered in this reference architecture.

Billions of simulated driving miles are needed to ensure safety requirements are met. However, the miles themselves are not the most important factor – what matters is the diversity in the miles driven and capturing all use cases and scenarios.

A trial is a single test of a scenarios under one set of randomized parameters. Simulation testing is based around each scenario having multiple trials to test different directed random parameters.

One example of a base scenario is a new feature that an engineering team is developing. This feature would need to be tested using multiple trials in a simulated environment.

The sizing model in this reference architecture is:

1. 100 environmental conditions – such as rain, snow, sunlight, and clouds
2. 100 transients – such as an adversarial car moving to close off space during a lane change
3. 100 fault injections – introduced via test code paths to trigger error handling routines

This implies every base scenario has one million trials (100x100x100).

Table 2 shows the number of DRIVE Constellation systems needed for HIL testing these parameters. Results are rounded, and the system uptime is estimated to be 20 hours per day.

Base scenarios	Trials	Running time per trial (seconds)	Total time of testing (hours)	TAT (hours)	DRIVE Constellation Systems (1x speed)
100	1M	15	400k	1000	400

Table 2. Simulation sizing

Simulation testing will have a minimum of ten base scenarios to start with, which will result in the need for 40 DRIVE Constellation systems. Complete testing will need a minimum of 100 base scenarios, which will result in 400 DRIVE Constellation systems over a period of time.

4 Testing

Regression testing on DRIVE Constellation PODs plays back previously recorded sensor data to the AV software. It is used to test against labeled ground-truth, and to test against recorded and expected vehicle behavior. For example, automatic emergency breaking (AEB) is checked for zero false-positives using regression testing.

Like simulation, regression testing can be done with either SIL or HIL – with HIL providing the most realistic and accurate testing. Sizing HIL hardware requirements is simply a matter of determining how many hours of testing are needed and the desired TAT. At the start of an AV development program, testing might begin with 2,000 hours of raw data (one car driving for one year) and a TAT of 100 hours, which would require 20 DRIVE Constellation systems. A complete AV development program with ten cars will end up testing on 20,000 hours of raw data with a 100 hour TAT, thus requiring 200 DRIVE Constellation systems.

SIL sizing requirements are very similar to the data factory process used to select images for labeling except that the actual production AI models would be used instead of using pre-trained AI models. Thus 30 additional DGX-1 servers will be required for SIL testing.

5 DRIVE Constellation POD

The DRIVE Constellation POD is a rack that has four DRIVE Constellation systems and eight storage nodes. A DRIVE Constellation system mimics vehicle sensors and controls with eight camera channels over GMSL2 and 1 GbE for radar and lidar systems. One chassis runs NVIDIA DRIVE Sim™ software to simulate sensors on a self-driving car such as cameras, lidar, and radar. In this chassis, NVIDIA GPUs are used to generate photoreal data streams that create a wide range of testing environments and scenarios. The other chassis contains a DRIVE Pegasus AI car computer that runs the complete, binary compatible software stack which operates inside an autonomous vehicle. The Simulation servers use a MaxQ performance setting to reduce the power used, while optimizing performance per watt.

The in-band switch provides 25 GbE to the Simulation nodes and 50 GbE to the storage nodes. An out-of-band switch with 1 GbE ports is used to manage IPMI, OS installation, etc. on the nodes. The Constellation server does not have its own networking.

NOTE: The DGX POD is a single-rack design that includes nine DGX-1 servers for high performance GPU compute and twelve storage nodes. Details about it are in the *DGX Data Center Reference Design*.

Preparing the data center to support DRIVE Constellation PODs is straightforward. As with all IT equipment installation, it is important to work with the data center facilities team to ensure the environmental requirements can be met.

These requirements are further detailed in the *NVIDIA DGX Site Preparation Guide* but important items to consider include:

Area	Design Guidelines
Rack	Supports 2000 lbs of static load Dimensions of 1200 mm depth x 700 mm width Structured cabling pathways per TIA 942 standard
Cooling	Removal of 68,240 BTU/hr ASHRAE TC 9.9 2015 Thermal Guidelines “Allowable Range”
Power	North America: A/B power feeds, each three-phase 400V/60A/20kW International: A/B power feeds, each 380/400/415V, 32A, three-phase – 21-23kW each.

Table 3. Rack, cooling, and power considerations for a DRIVE Constellation POD

Figure 2 shows the placement and requirements for a DRIVE Constellation POD.

Four sets of DRIVE Constellation systems
(4 x 8 = 32 RU)

Eight storage servers
(8 x 1 RU = 8 RU)

In-band switch (data)
(1 RU)

Out-of-band switch (management)
(1 RU)

Figure 2. DRIVE Constellation POD

6 Summary

To develop autonomous vehicles, the transportation industry will need to use new end-to-end systems for software development. This begins with a data factory with petabytes of storage for the billions of sensor images generated from data collection fleets and DGX PODs to pre-process the raw data and select images for labeling. An even larger number of DGX PODs are required to train the AI Models at the heart of all AV software.

Safety is critical to the acceptance of autonomous vehicles and this requires billions of miles of driving under a diverse set of conditions. Even with large test fleets, HIL simulation will be required to achieve acceptable safety levels. Regular testing of new AI models using the same HIL technique will also be required, both during the initial AV development phase and for AV software maintenance over the life of the vehicle. The computational scale challenges are significant and ideally suited for GPU-accelerated computing using DRIVE Constellation PODs.

NVIDIA brings a unique perspective to AV development with several thousand engineers working on AV software using our SATURNV deep learning supercomputer. The DGX and DRIVE Constellation PODs are based on the same designs used in the SATURNV Supercomputer. This has resulted in the TSTADI reference architecture developed to help all companies developing AV software to accelerate their schedules and bring safer vehicles to market.

Legal Notices and Trademarks

Legal Notices

ALL INFORMATION PROVIDED IN THIS WHITE PAPER, INCLUDING COMMENTARY, OPINION, 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 MATERIALS, AND EXPRESSLY DISCLAIMS ALL IMPLIED WARRANTIES OF NONINFRINGEMENT, MERCHANTABILITY, AND FITNESS FOR A PARTICULAR PURPOSE.

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.

Trademarks

NVIDIA, the NVIDIA logo, CUDA, Tesla, NVLink, DGX, and DGX-1 are trademarks or registered trademarks of NVIDIA Corporation in the United States and other countries. Other company and product names may be trademarks of the respective companies with which they are associated.