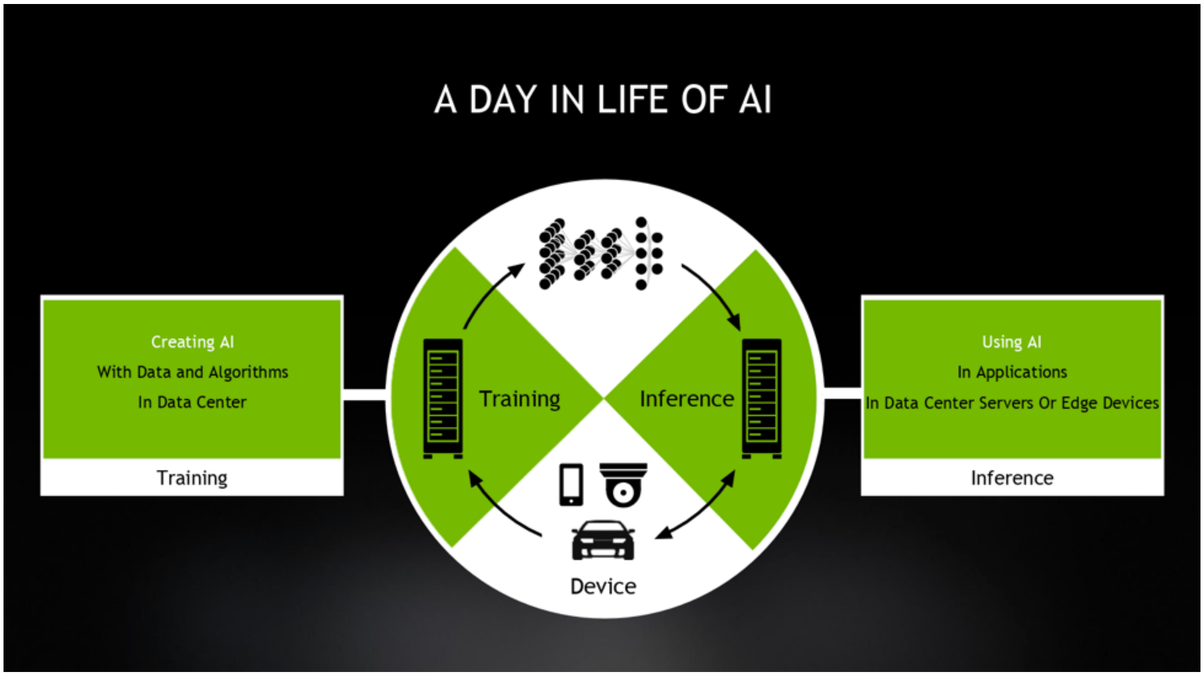
NVIDIA Xavier Wins Critical AI Performance Benchmarks

THE BRAIN OF THE DRIVE AGX PLATFORM LEADS MLPERF INFERENCE TESTS.

Cars and trucks of the future will be driven by an AI supercomputer running a wide range of diverse deep neural networks. This is a massive AI workload, and the new MLPerf Inference 0.5 benchmark suite gives great insight into the true performance of solutions being used for inferencing, including autonomous vehicles.

That's why it is so significant that the NVIDIA Xavier system-on-a-chip and our Turing GPUs — the engines behind the NVIDIA DRIVE AGX platform — just achieved top results in MLPerf Inference, the first industry-standard, independent AI benchmarks. This important accomplishment shows NVIDIA's leadership in AI inferencing for a variety of different AI tasks and scenarios, a critical capability for safe operation of a self-driving vehicle.

Inference is the process of running AI models in real-time to extract insights from enormous amounts of data. NVIDIA DRIVE runs numerous deep neural networks simultaneously to perceive a vehicle's environment via various sensors generating terabytes of data. These DNNs must be able to analyze key data in real time to perform redundant and diverse functions, such as identifying intersections and classifying drivable paths.



AI encompasses both training and inference to power intelligent applications.

THE CAR OF THE FUTURE

The car of the future will be an amazing edge computing device. We will converse with the car and it will respond, answering questions, directing us to destinations, warning us of road conditions. It will be our co-pilot and guardian, able to take over and drive autonomously, monitor our alertness and safeguard us. To accomplish this, it will combine the most diverse AI algorithms together and run them on what will certainly be a high performance computer capable of operating many diverse deep neural networks, simultaneously.

The MLPerf 0.5 inference tests represent a suite of benchmarks to assess this type of complex workload. Many different benchmark tests, across multiple scenarios, including edge computing, verify whether a solution can perform exceptionally at not just one task, but many, as would be required in a modern car.

Which is why the NVIDIA Xavier benchmark results are so important to automakers. The Xavier processor ranked as the highest performer under both edge-focused scenarios (single- and multi-stream) among commercially available edge and mobile SoCs*. Today, the Xavier processor powers the brain of the NVIDIA DRIVE AGX computer for both self-driving and cockpit applications — it's an AI supercomputer running up to 20 DNNs simultaneously. These include models to understand the environment, such as LaneNet for lane markings, PathNet which detects driveable edges, PilotNet which determines centerlines, SignNet for street signs, LightNet for traffic lights, WaitNet for intersections, DriveNet for object detection, OpenRoadNet for free space detection, ParkNet for locating parking spots, and more.

In addition, inside the vehicle, there are networks for driver monitoring, including those that can determine head pose, detect blinking eyes, and read lips. The future car will be one capable of carrying on a conversation and will require advanced speech recognition, natural language processing, and text-to-speech, all with incredibly low latency.

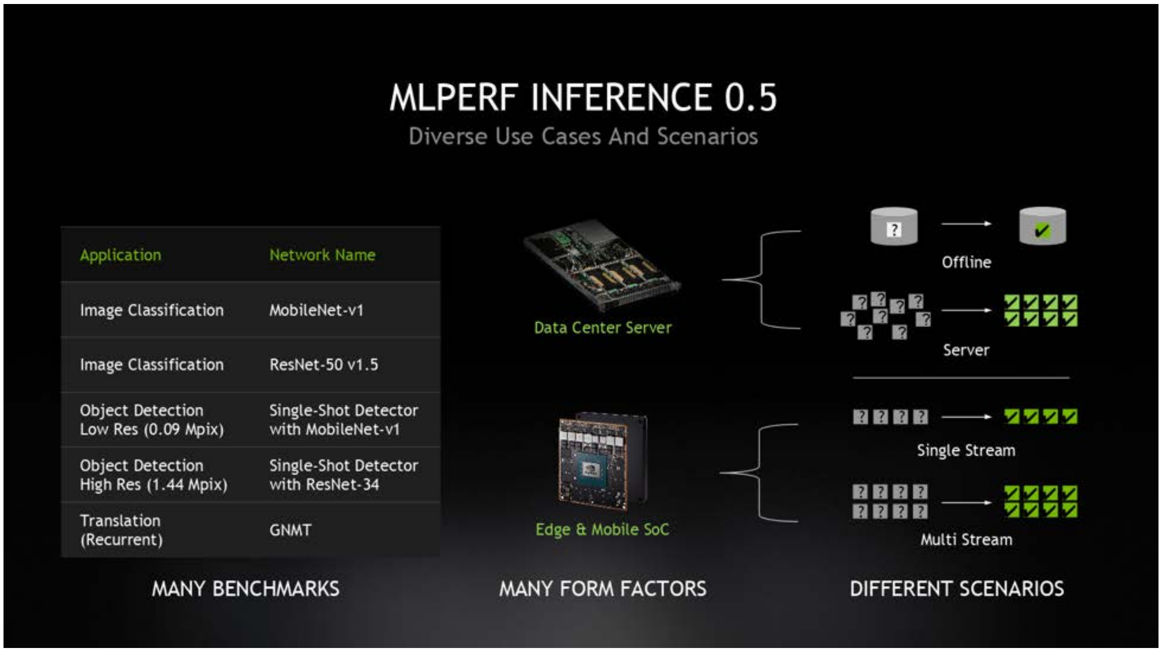
These diverse neural networks all process different types of data, through a wide range of different types of neural nets. The net net is that a tremendous amount of processing performance is required, to deliver a safe AV with an intelligent voice activated user interface.

But how can you tell that a processor is capable of delivering on this promise when all manufacturers seem to be quoting higher and higher TOPS for their future processors? What is clear is that it is not just about the raw peak performance numbers, but rather how these processors can handle actual AI workloads. And the MLPerf benchmark is the industry's way to measuring true AI inference performance.

TOP OF THE CLASS

NVIDIA was the only AI platform among 14 organizations to submit results across all five MLPerf inference tests. These benchmarks follow the tests released earlier this year to measure the industry's capabilities in AI training.

MLPerf measures performance in three AI applications — image classification, object detection and translation — using five distinct benchmarks. The server and offline scenarios defined in the benchmark are most relevant for data center uses cases, while single- and multi-stream scenarios are used for edge devices, like autonomous vehicles.



MLPERF INFERENCE 0.5

Diverse Use Cases And Scenarios

Application	Network Name
Image Classification	MobileNet-v1
Image Classification	ResNet-50 v1.5
Object Detection Low Res (0.09 Mpix)	Single-Shot Detector with MobileNet-v1
Object Detection High Res (1.44 Mpix)	Single-Shot Detector with ResNet-34
Translation (Recurrent)	GNMT

MANY BENCHMARKS

Data Center Server

Edge & Mobile SoC

MANY FORM FACTORS

Offline

Server

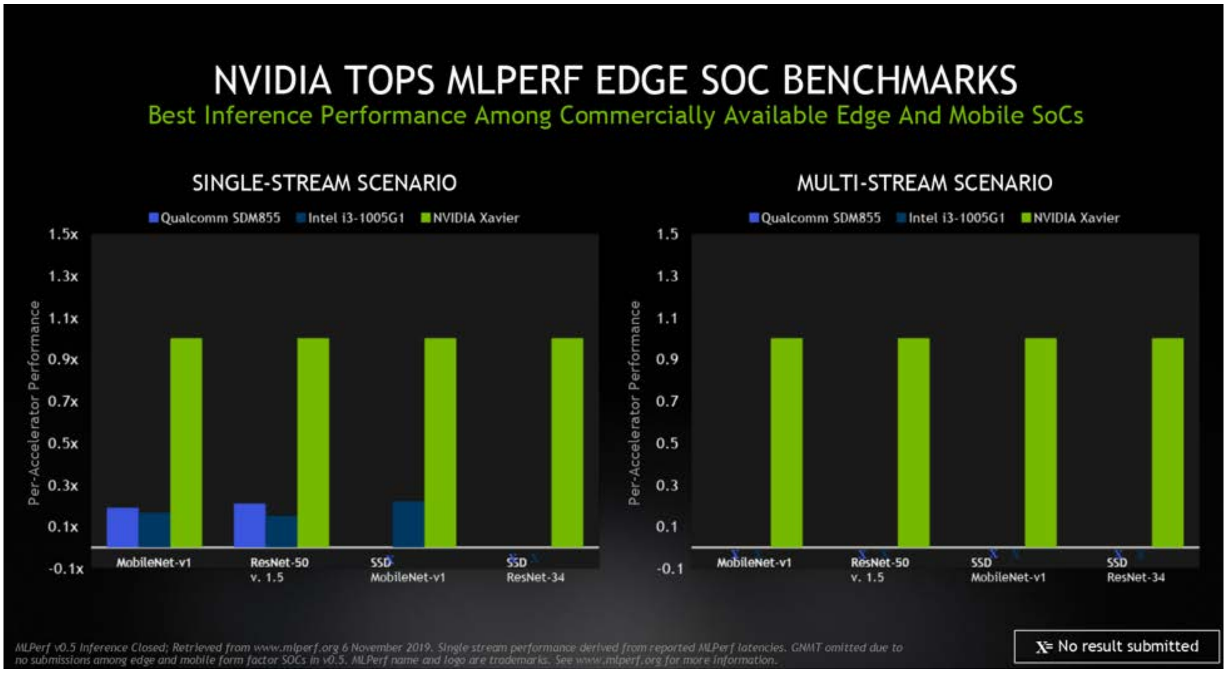
Single Stream

Multi Stream

DIFFERENT SCENARIOS

MLPerf Inference 0.5 covers Data Center Server as well as Edge and Mobile SoC scenarios.

NVIDIA topped the benchmarks for data center scenarios, with Turing GPUs providing the highest performance per processor among commercially available products*. Xavier ranked as the highest performer under edge-focused scenarios among commercially available edge and mobile SoCs*.



NVIDIA TOPS MLPERF EDGE SOC BENCHMARKS

Best Inference Performance Among Commercially Available Edge And Mobile SoCs

MLPerf v0.5 Inference Closed; Retrieved from www.mlperf.org 6 November 2019. Single stream performance derived from reported MLPerf latencies. GNMT omitted due to no submissions among edge and mobile form factor SOCs in v0.5. MLPerf name and logo are trademarks. See www.mlperf.org for more information.

The Xavier SoC topped tests for edge computing.

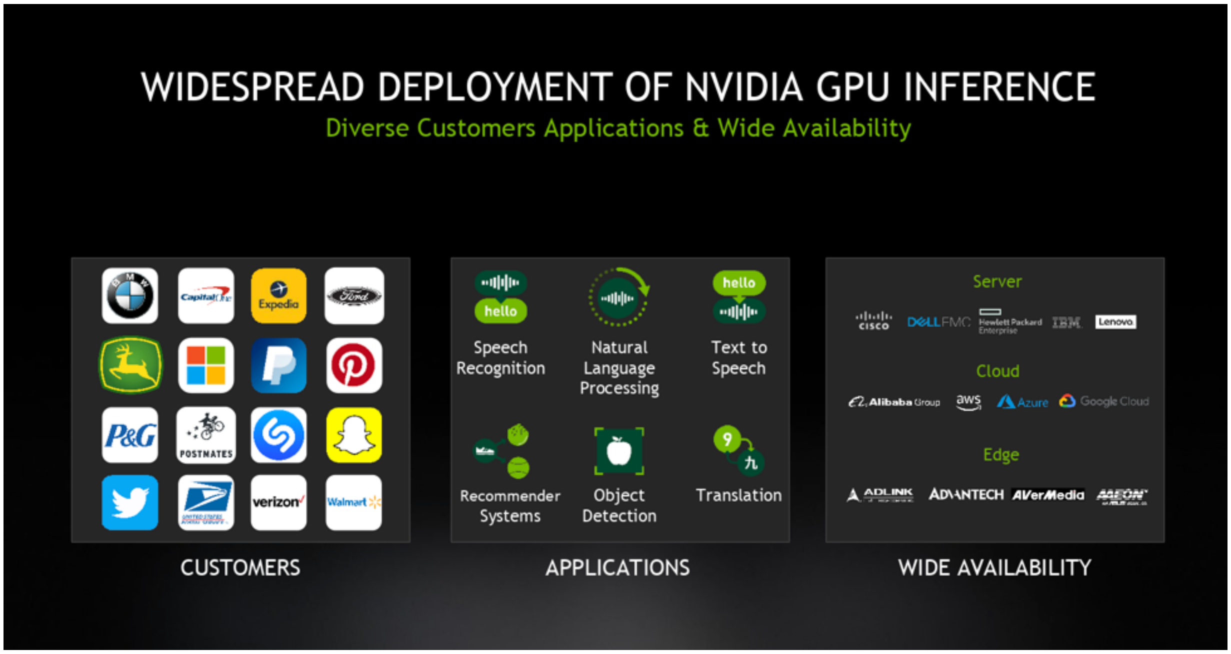
The multistream scenario tests how many feeds a chip can handle. Equipped with a halo of various types of sensors, self-driving cars are a prime example of edge devices that must be able to handle these diverse sources of data in real time.

THE INDUSTRY AGREES

The results reveal the power of our CUDA and TensorRT software running on our processors. They provide a common platform that delivers leadership results across multiple products and use cases, a capability unique to NVIDIA.

MLPerf has broad backing from industry and academia. Its members include Arm, Facebook, Futurewei, General Motors, Google, Harvard University, Intel, MediaTek, Microsoft, NVIDIA and Xilinx.

When it comes to actual workloads, inference can be incredibly demanding, requiring significant pre- and post-processing steps. That's why a wide range of industries are increasingly adopting high-performance NVIDIA platforms to handle inference jobs. They include a who's who of forward-thinking transportation and technology companies such as BMW, Cisco, Ford Motor Co., John Deere, Microsoft, PayPal, Pinterest, Postmates, Shazam, Snap, Toyota, Twitter, Verizon, Volvo and Walmart.



WIDESPREAD DEPLOYMENT OF NVIDIA GPU INFERENCE

Diverse Customers Applications & Wide Availability

A wide range of companies around the world have turned to NVIDIA GPU technology for top inference performance.

THE CONVERSATION CONTINUES

Looking ahead, conversational AI represents a giant set of opportunities and technical challenges on the horizon — and NVIDIA is a clear leader here, too.

NVIDIA DRIVE IX is an intelligent experience platform that provides AI-powered interactions between the vehicle and its occupants. Using embedded speech software, it enables natural conversation with the car.

In addition to the AI cockpit, NVIDIA offers optimized reference designs for conversational AI services such as automatic speech recognition, text-to-speech and natural-language understanding. Our open-source optimizations for AI models such as BERT, GNMT and Jasper give developers a leg up in reaching world-class inference performance.

We are excited to share with the transportation industry that the NVIDIA Turing GPU and NVIDIA Xavier SoC have achieved the fastest results in their segments on the MLPerf benchmarks. In a world where daily announcements professing AI excellence are made by numerous companies causing confusion, the MLPerf 0.5 inference results are intended to separate noise from what matters—real performance and excellence across multiple different tests. The results from those tests show Xavier as the best performer for building the car of the future.