NVIDIA SETS EIGHT RECORDS IN AI PERFORMANCE

Nowhere is relentless innovation more apparent than in the field of AI. It's driven by the researchers and developers discovering new network architectures, algorithms, and optimizations—doing much of this leading-edge work on NVIDIA data center platforms.

In MLPerf 0.6, NVIDIA has set eight new AI performance records—three at scale and five in per-accelerator comparisons. These breakthrough results come from a combination of unprecedented scale and a nearly 40% improvement in software optimizations across MLPerf's six workloads.

GETTING TO KNOW MLPERF 0.6

MLPerf is a set of benchmarks that enable the machine learning (ML) field to measure training performance across a diverse set of usages.

IMAGE CLASSIFICATION
Assigns a label from a fixed set of categories to an input image, i.e., applies to computer vision problems such as autonomous vehicles.

OBJECT DETECTION (LIGHTWEIGHT)
Finds instances of real-world objects such as faces, bicycles, and buildings in images or videos and specifies a bounding box around each.

OBJECT DETECTION (HEAVYWEIGHT)
Detects distinct objects of interest appearing in an image and identifies a pixel mask for each.

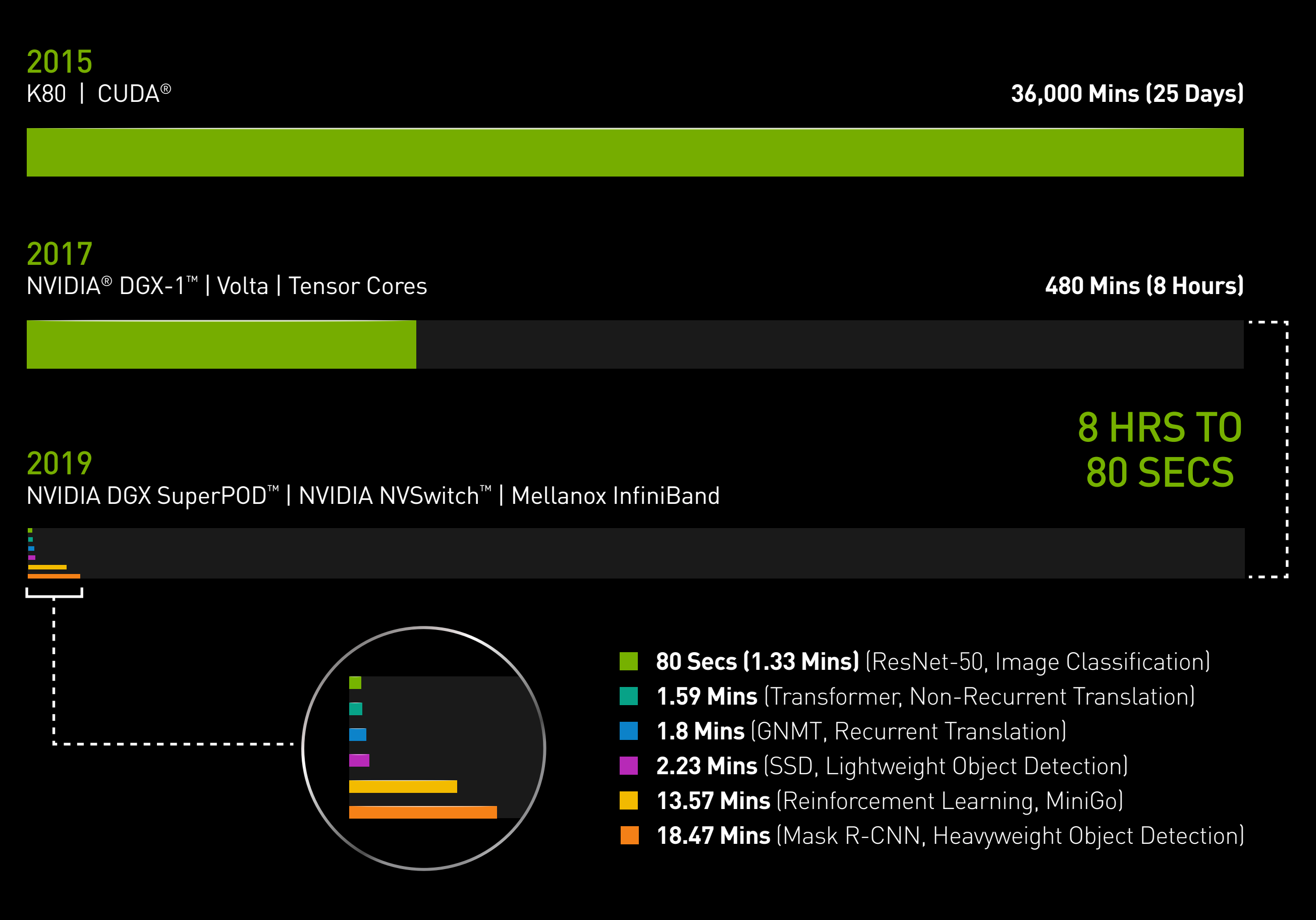
TRANSLATION (RECURRENT)
Translates text from one language to another using a recurrent neural network (RNN).

TRANSLATION (NON-RECURRENT)
Translates text from one language to another using a feed-forward neural network.

REINFORCEMENT LEARNING
Evaluates different possible actions to maximize reward using the strategy game *Go* played on a 9x9 grid.

THE AI TIME MACHINE

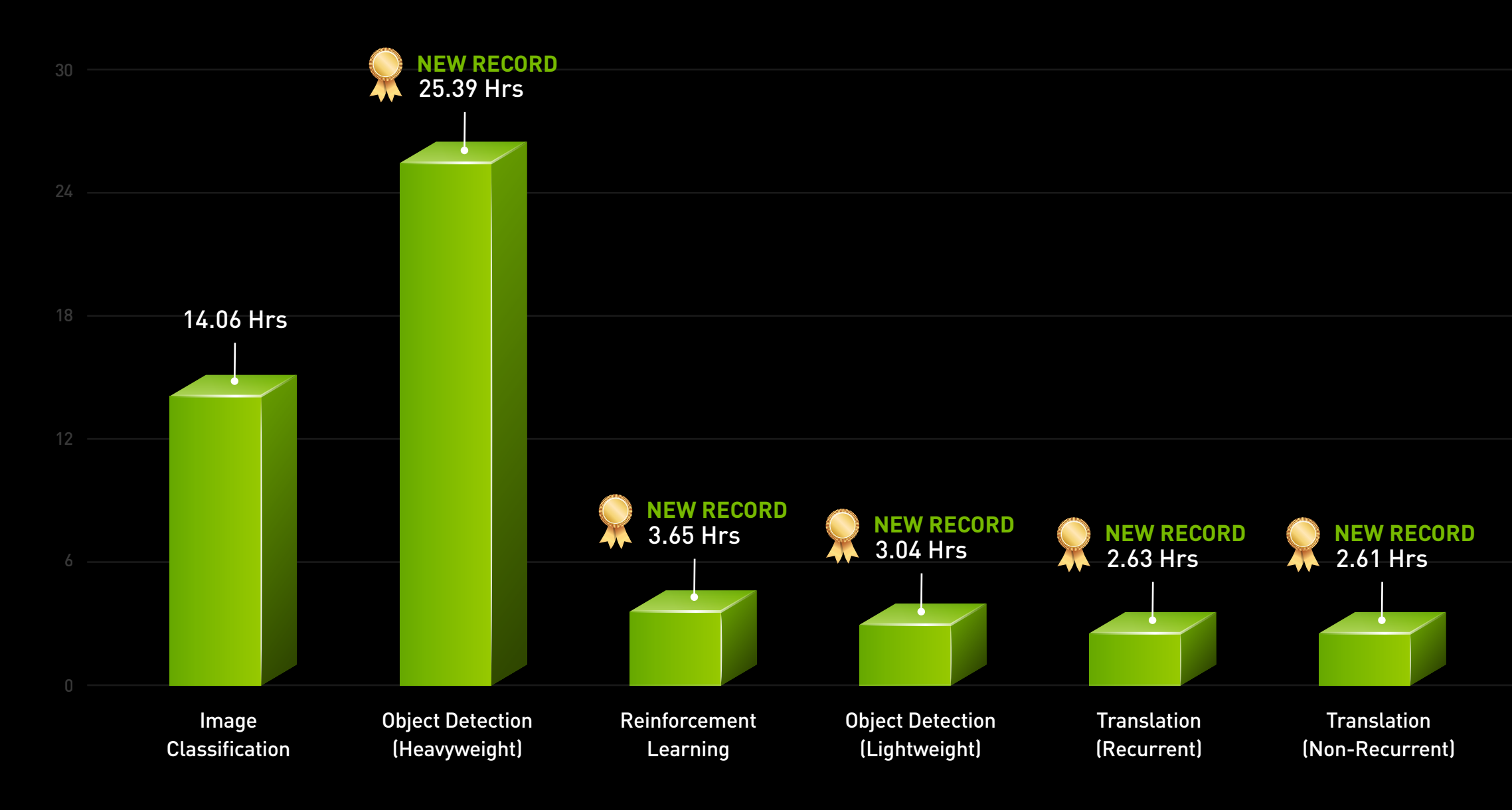
The NVIDIA AI platform has delivered market-leading performance through constant hardware and software innovation, slashing training time from 8 hours to 80 seconds in just two years.



AI PERFORMANCE LEADERSHIP ACROSS DIVERSE USAGES

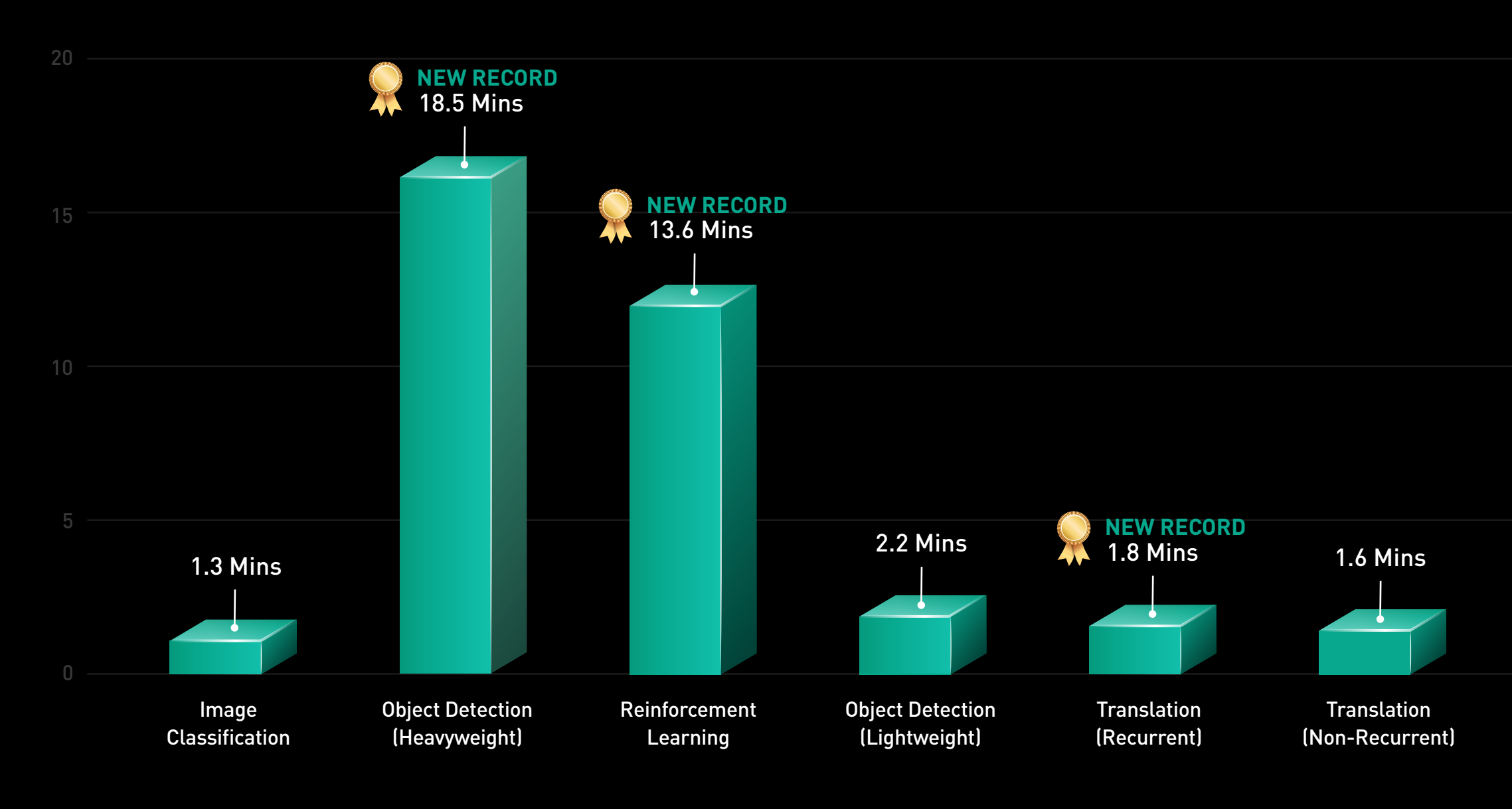
Performance leadership in AI means accelerating every workload and accelerating every framework. NVIDIA delivered its results using three different frameworks—MXNet, PyTorch, and TensorFlow—showing not only the platform's great performance but also its great versatility.

PER ACCELERATOR
Derived from a Single DGX Server



Per-accelerator comparison using reported performance for MLPerf 0.6 NVIDIA DGX-2H (16x V100s) compared to other submissions at same scale except for MiniGo, which used NVIDIA DGX-1 (8x V100s).

AT SCALE
DGX SuperPOD



Test platform: NVIDIA DGX SuperPOD in different configurations. Each DGX-2H, Dual-Socket Xeon Platinum 8189, 384 GB system RAM, 16x 32 GB Tesla V100 GPUs connected via NVSwitch. All results from MLPerf 0.6 training closed; retrieved from **www.mlperf.org** 10 July 2019. MLPerf name and logo are trademarks. See **www.mlperf.org** for more information.

THE NVIDIA PLATFORM APPROACH

A great AI platform should deliver great performance both at scale and in a single-server node. From Tensor Cores to NVSwitch, our continuous innovation in architecture, systems, and ecosystem support make NVIDIA the platform of choice.

MOST ACCESSIBLE PLATFORM
It's available from every cloud and every server maker, from edge to desktop, and includes free integrated software stacks from NGC.

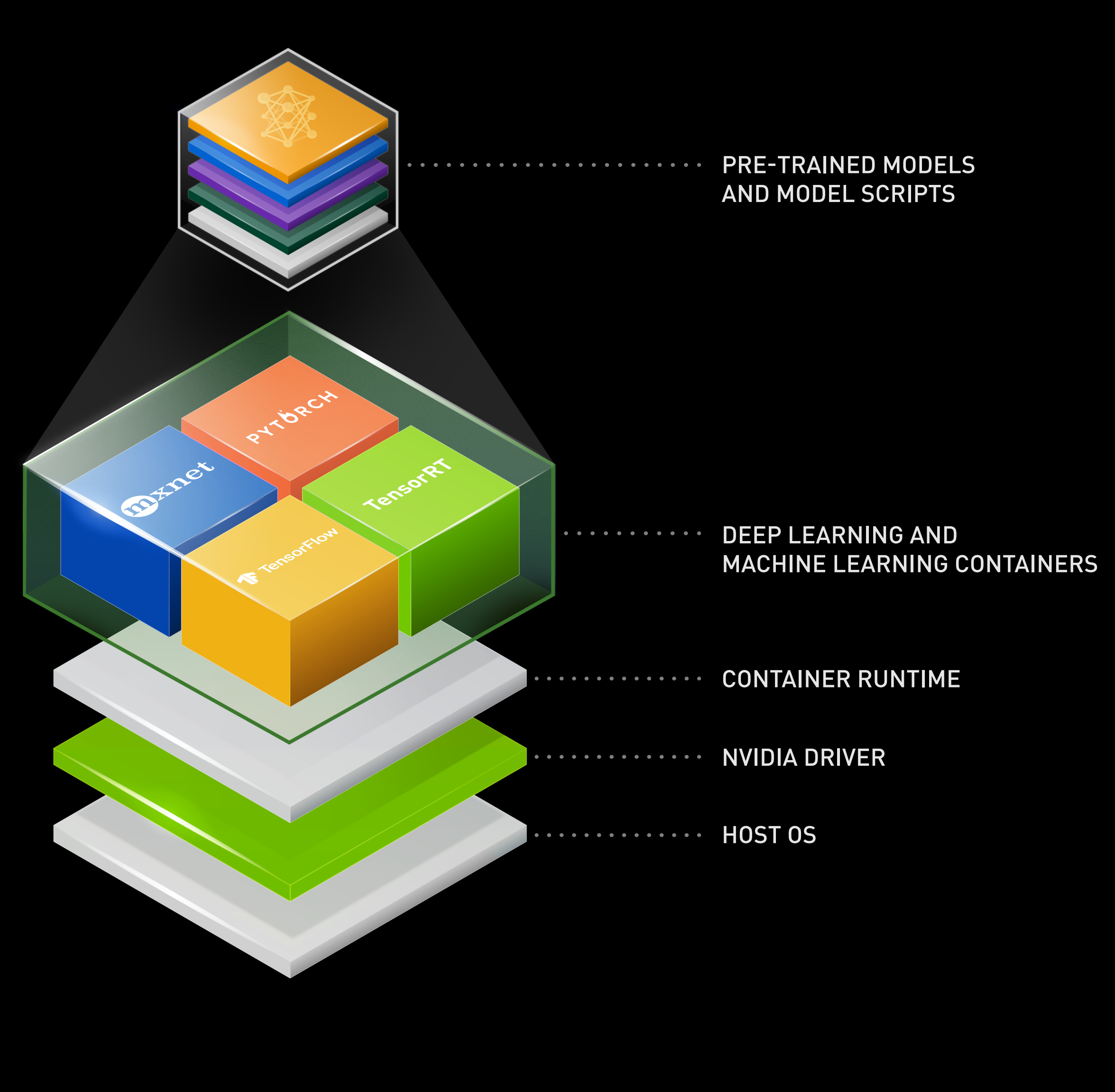
BROADEST DEVELOPER ECOSYSTEM
The platform is supported by 1.3 million developers, supports all frameworks and development environments, and delivers continuous software optimizations in the latest NGC containers.

LEADERSHIP-CLASS AI INFRASTRUCTURE
NVIDIA DGX SuperPOD is used for record-setting performance, delivering massive computing power to enterprises and a modular and scalable approach for AI.

READY-TO-RUN AI SOFTWARE

NGC containers, including TensorFlow, PyTorch, MXNet, NVIDIA TensorRT™, and more, help data scientists and researchers rapidly build, train, and deploy AI models. Each is optimized, tested, and ready to run on supported NVIDIA GPUs in workstations, on-prem, and in the cloud.

NGC also offers pre-trained models and model scripts for natural language processing, text-to-speech, classification, and many other popular use cases. Data scientists and developers can speed up time to solution by utilizing the models and scripts as is or modifying them to fit their needs.



ENTERPRISE AI INFRASTRUCTURE

NVIDIA DGX SuperPOD revolutionizes supercomputing, delivering a new, enterprise-grade infrastructure solution that any large organization can use to access massive computing power and propel business innovation. Built on NVIDIA DGX-2 with Mellanox networking, DGX SuperPOD provides a modular and cost-effective approach for AI at scale.

Explore the NVIDIA AI platform, how it achieved these great MLPerf results, and how it can accelerate your projects, products, and services.

LEARN MORE

PER-ACCELERATOR MLPERF RESULTS
MLPerf 0.6 submission information: Per-accelerator comparison using reported performance for MLPerf 0.6 NVIDIA DGX-2H (16x V100s) compared to other submissions at same scale except for MiniGo where NVIDIA DGX-1 (8x V100s) submission was used. See **mlperf.org** for more information.

AT-SCALE MLPERF RESULTS
MLPerf ID max scale: Mask R-CNN: 0.6-23, GNMT: 0.6-26, MiniGo: 0.6-11 | MLPerf ID per accelerator: Mask R-CNN, SSD, GNMT, transformer: all use 0.6-20, MiniGo: 0.6-10. See **mlperf.org** for more information.