



## IT MAKING AI WORK

AI is transforming every industry, but many companies are starting their journey without an IT-led strategy. The result is silos of innovation that can't scale efficiently. The right AI infrastructure strategy attracts talent, consolidates resources, and drives innovation.

AI at scale puts unprecedented demands on your data center. Powering your company's business transformation with machine and deep learning requires infrastructure that's optimized for the unique demands of AI.



We've already built the world's most advanced AI infrastructure. We call it SATURNV and it represents all that we've learned designing and deploying scalable AI in the world's largest data centers.

**NVIDIA DGX™ A100**  
The Universal Building Block for the AI Data Center

4.6 ExaFLOPS Total Capacity and Growing



### NVIDIA SELENE RESEARCH CLUSTER WITHIN SATURNV

No. 2 on the Green500 list for supercomputers at 20.5 gigaflops per watt  
No. 7 in the TOP500 of supercomputers at 27.5 petaFLOPS on the LINPACK Benchmark



## SCALABLE AI AT WORK

At NVIDIA, IT infrastructure built on NVIDIA DGX, powers the most important NVIDIA work; NVIDIA DGX infrastructure and revolutionary performance drives the AI behind rapid advancements across industries from self-driving cars, to maximizing product quality and increasing customer satisfaction and driving the next breakthroughs in AI research. Every day, NVIDIA's developers, operations, and customers benefit from its revolutionary performance, effortless productivity, and pervasive reach across our business.



## RECORD-SETTING PERFORMANCE

In MLPerf v0.7 training, the leading benchmark suite for AI performance, NVIDIA DGX SuperPOD and DGX systems set world records in all 8 of the at scale benchmarks for commercially available systems. This winning infrastructure solution was built following the DGX SuperPOD reference architecture and was assembled in under 1 month.

<b>RECOMMENDATION (DLRM)</b>	---->	Delivers personalized results in user-facing services such as social media or e-commerce websites by understanding interactions between users and service items, like products or ads.
<b>3.33 mins</b>		
<b>NATURAL LANGUAGE PROCESSING (BERT)</b>	---->	Understands text by using the relationship between different words in a block of text. Allows for question answering, sentence paraphrasing, and many other language-related use cases.
<b>0.81 mins</b>		
<b>REINFORCEMENT LEARNING (MINIGO)</b>	---->	Evaluates different possible actions to maximize reward using the strategy game Go played on a 19x19 grid.
<b>17.07 mins</b>		
<b>TRANSLATION (NON-RECURRENT) TRANSFORMER</b>	---->	Translates text from one language to another using a feed-forward neural network.
<b>0.62 min</b>		
<b>TRANSLATION RECURRENT (GNMT)</b>	---->	Translates text from one language to another using a recurrent neural network (RNN).
<b>0.71 min</b>		
<b>OBJECT DETECTION</b>	---->	Finds instances of real-world objects such as faces, bicycles, and buildings in images or videos and specifies bounding box around each.
<b>10.46 mins</b>		
(heavy weight) Mask R-CNN		
<b>0.82 min</b>		
(light weight) SSD		
<b>IMAGE CLASSIFICATION (RESNET-50 V1.5)</b>	---->	Assigns a label from a fixed set of categories to an input image, i.e. applies to computer vision problems such as autonomous vehicles.
<b>0.76 min</b>		

Test Platform: NVIDIA DGX SuperPOD, multi-node cluster built with NVIDIA DGX A100. DGX A100 configuration: Dual socket Epyc 7742, 1TB System memory, Mellanox ConnectX-6 VPI HDR InfiniBand/Ethernet, and 8x NVIDIA A100 Tensor Core GPU

Max Scale: All results from MLPerf v0.7 using NVIDIA DGX A100 (8x A100s). MLPerf ID Max Scale: ResNet50 v1.5: 0.7-37; Mask R-CNN: 0.7-28; SSD: 0.7-33; GNMT: 0.7-34; Transformer: 0.7-30; Minigo: 0.7-36; BERT: 0.7-38; DLRM: 0.7-17

MLPerf name and logo are trademarks. See [www.mlperf.org](http://www.mlperf.org) for more information.

## BUILD YOUR OWN WORLD-CLASS AI INFRASTRUCTURE

Leverage NVIDIA's modular reference architecture, NVIDIA DGX SuperPOD™, based on insights from SATURNV and powered by DGX A100 systems and our ecosystem of trusted IT solutions providers.

