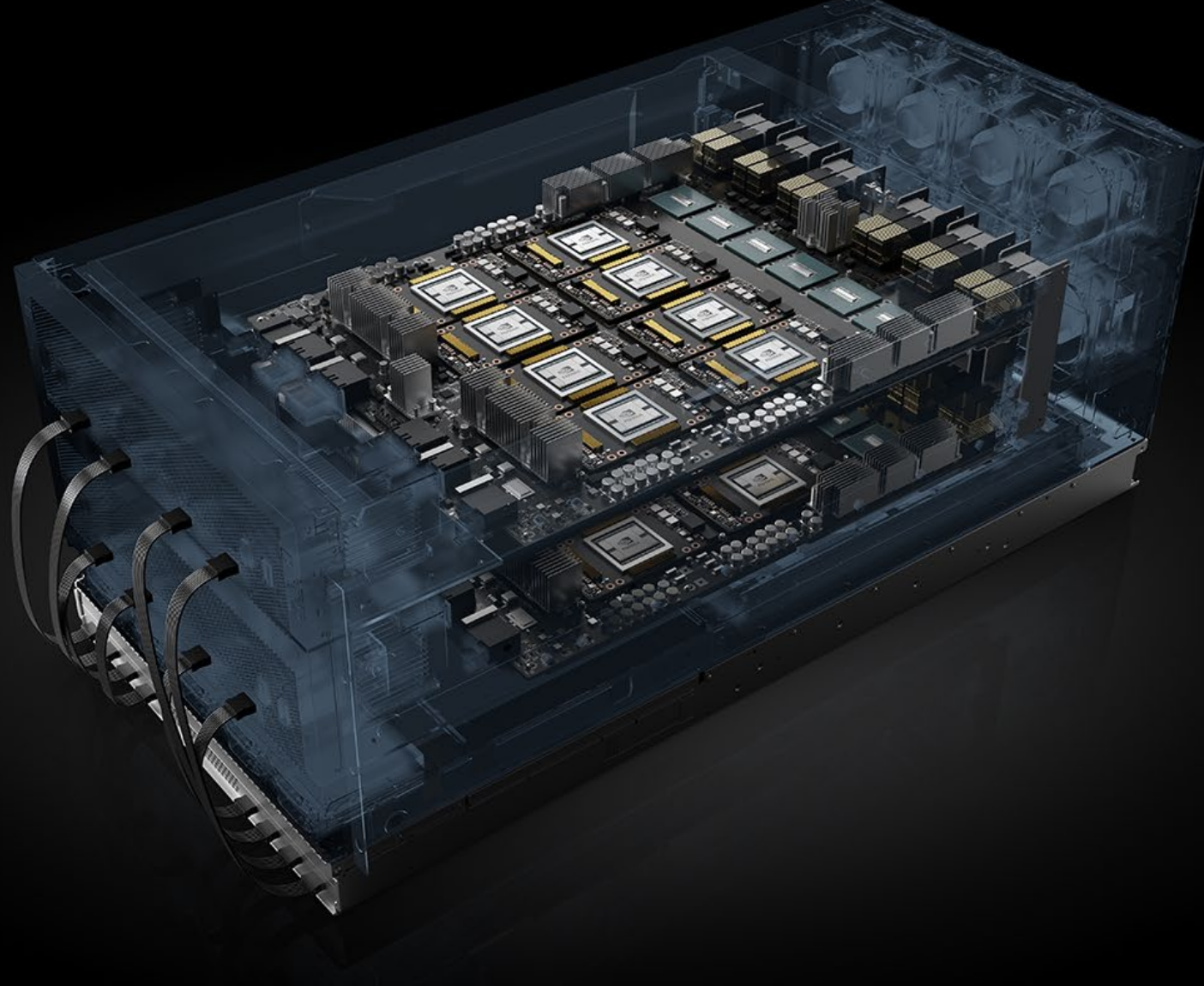


NVIDIA HGX-2

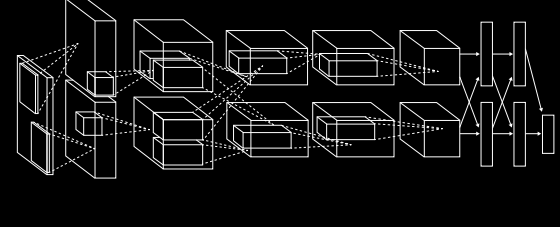
FUSING HPC AND AI COMPUTING INTO A UNIFIED ARCHITECTURE



EXPLOSION OF NETWORK COMPLEXITY

AI models are becoming increasingly complex and diverse, from translating languages to autonomous driving. Solving these models requires massive compute capability, large memory, and extremely fast connections between the GPUs.

Convolutional Networks



Encoder/Decoder

ReLU

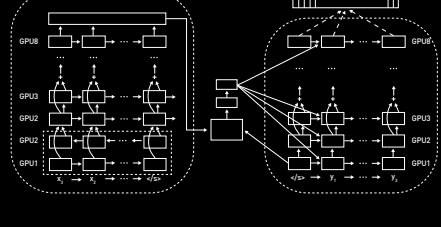
BatchNorm

Concat

Dropout

Pooling

Sequence & Attention Networks



LSTM

GRU

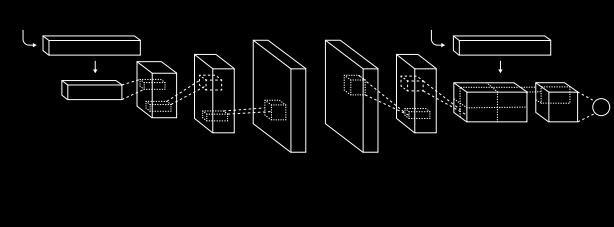
Beam Search

WaveNet

CTC

Attention

Generative Adversarial Networks



3D-GAN

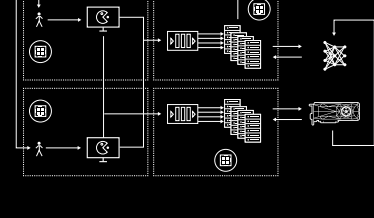
MedGAN

Conditional GAN

Attention

Speech Enhancement

Reinforcement Learning

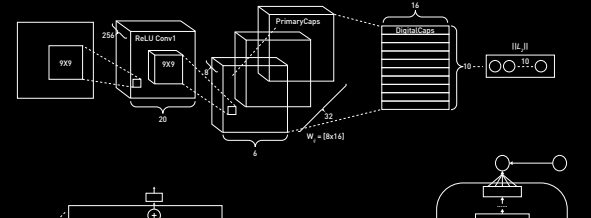


DQN

Simulation

DDPG

New Species



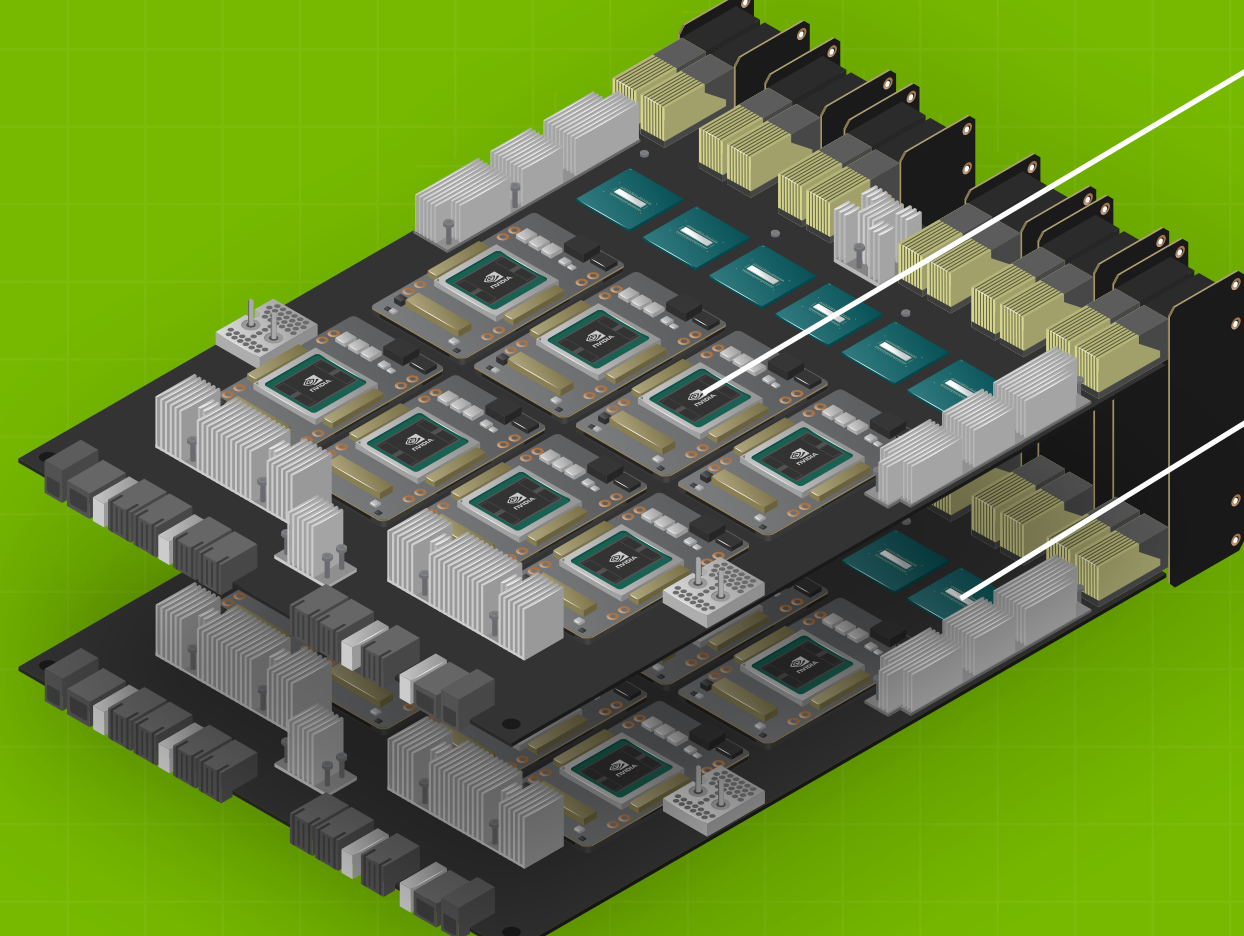
Mixture of Experts

Neural Collaborative Filtering

Block Sparse LSTM

REDEFINING THE FUTURE OF COMPUTING

HGX-2 multi-precision computing platform allows high-precision calculations using FP64 and FP32 for scientific computing and simulations, while also enabling FP16 and Int8 for AI training and inference. This unprecedented versatility provides unique flexibility to support the future of computing.



16

NVIDIA® Tesla® V100 GPUs
0.5TB Memory

12

NVIDIA NVSwitches
Direct GPU-to-GPU Connection
Between All 16 GPUs

24X

Higher GPU-to-GPU
Bandwidth*

0.5TB

Aggregate High-Bandwidth
GPU Memory

2 PFLOPS

Total Compute

* Compared to two HGX-1-based servers connected with 4x 1B ports



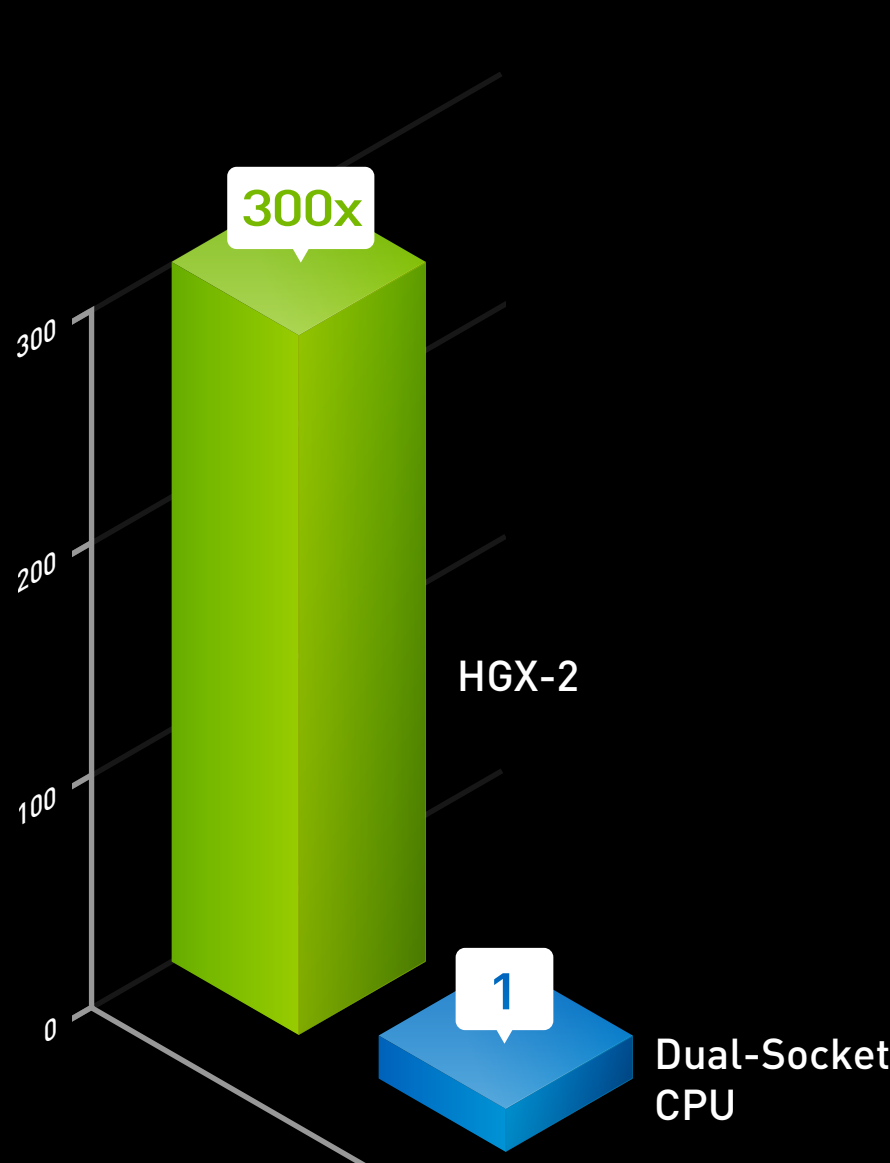
RECORD PERFORMANCE

The HGX-2 platform is powered by NVIDIA NVSwitch™ which enables every GPU to communicate with every other GPU at full bandwidth of 2.4TB/sec to solve the largest of AI and HPC problems.

AI Training

**HGX-2 Replaces
300 CPU-Only Server Nodes**

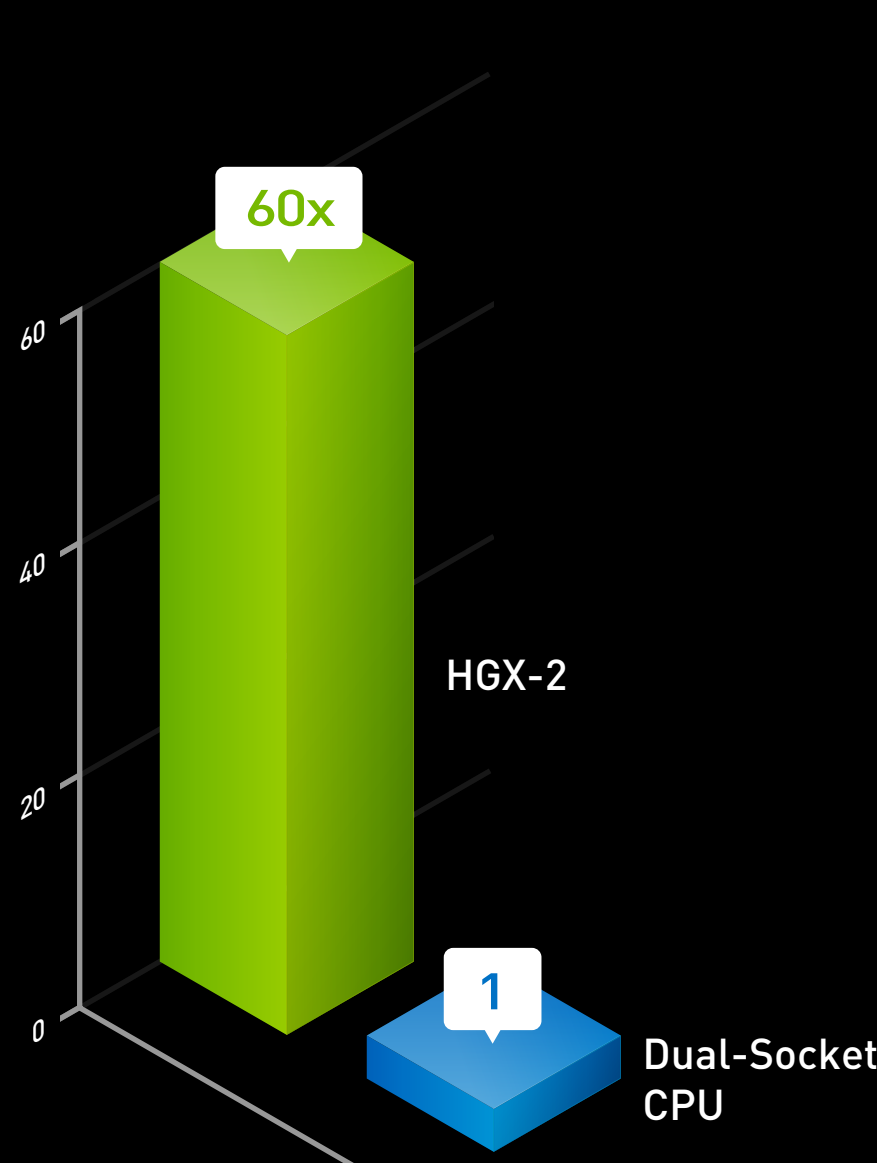
Workload: ResNet50, 90 epochs to solution
CPU server: dual-socket Intel Xeon Gold 6140



HPC

**HGX-2 Replaces
60 CPU-Only Server Nodes**

Workload: MILC (particle physics HPC application)
CPU server: dual-socket Intel Xeon Gold 6140



EMPOWERING THE DATA CENTER ECOSYSTEM

NVIDIA works with a wide range of partners to deliver the ideal AI and HPC solution. With HGX-2, they can now integrate a state-of-the-art platform into their servers to advance the data center ecosystem.

FOXCONN®
Advancing Through Innovation

Inventec

Lenovo®

QCT

Quanta Computer

SUPERMICRO

wlstron

wiwynn®

SEE HOW HGX-2 CAN ACCELERATE
YOUR AI AND HPC WORKLOADS.

www.nvidia.com/hgx