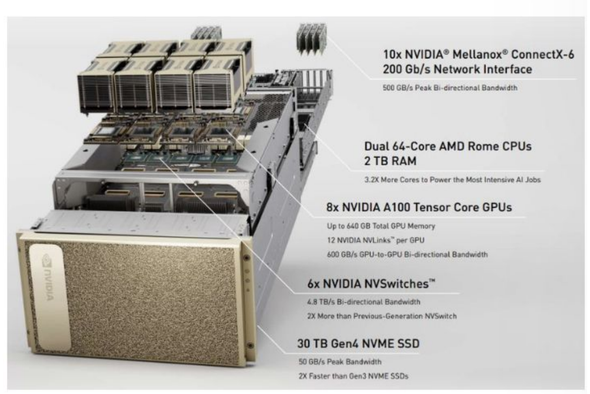
About SandBox/SuperPOD：sandweb.beyondai.jp

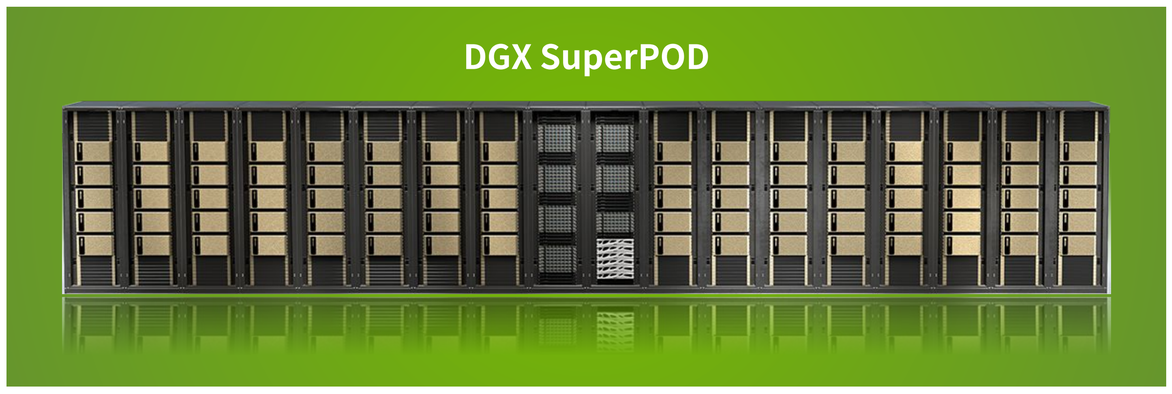
鶴長 鎮一、Khan Md Anwarus Salam
ソフトバンク株式会社

In 2021, SoftBank introduced SANDBOX: 40 NVIDIA DGX A100, for Beyond AI



Added New SSD (30.72TB x 11) Effective capacity of about 240TB

in 2023, Softbank launched AI SuperPOD equipped wilh 2,000 NVIDIA Tensor Core GPUs



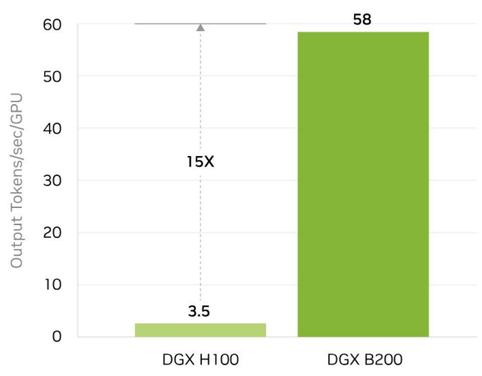
In 2024, Added 4,000 NVIDIA Hopper GPU (NVIDIA DGX B200): 4.7ExaFLOPS (10^{18} FLOPS) on the way of making it one of the largest computing infrastructures in the world!

DGX B200 System Specifications	
Estimated System Power	**14.3 kW Max**
Estimated System Weight	>287.6 lbs. (130 kgs)*
System Dimensions	10 Rack Units (RU) Height: 17.5 in (444 mm) Width: 19.0 in (482.2 mm) Length: 35.3 in (897.1 mm)
Operating Temperature	5ºC to 30ºC (41ºF to 86ºF)
Cooling	Air
Operating Altitude	3048 meters (10,000 feet) maximum 5° to 35°C: altitude 0 – 1000 ft 5° to 30°C: altitude 1000 – 5000 ft 5° to 25°C: altitude 5000 – 10000 ft
Estimated Acoustic Noise Operating	Acoustic Power (LWA,m)* 25 °C /77 °F = 96 dB 30 °C / 86 °F = 101 dB

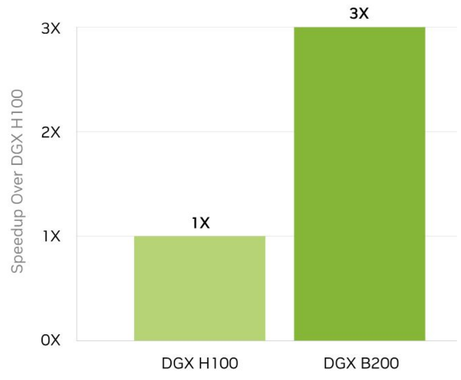
*Estimated. Actual specification to be determined

NVIDIA

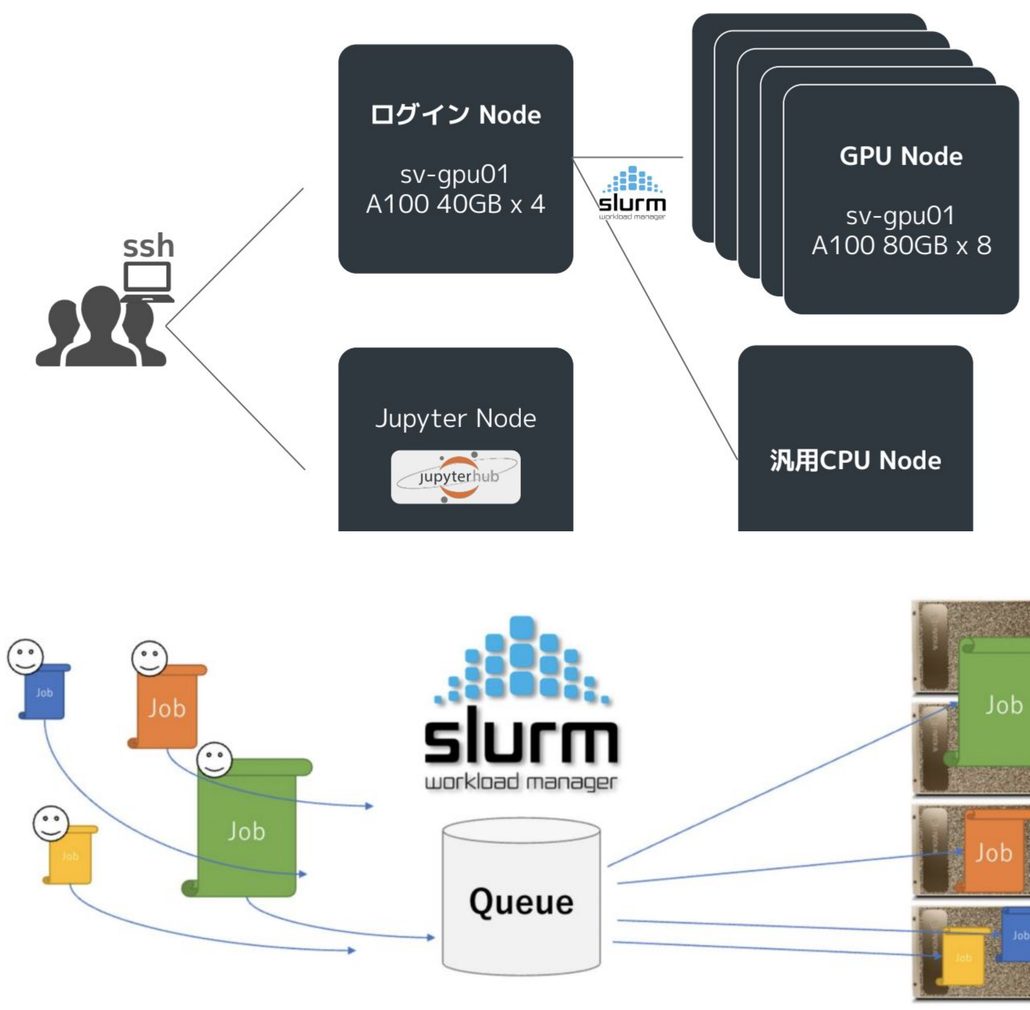
Real Time Large Language Model Inference



Supercharged AI Training Performance

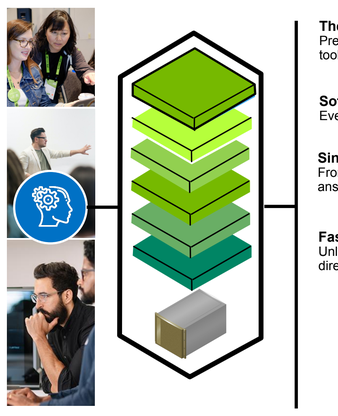


Job Scheduling using Slurm



The DGX Platform: Enterprise-class, Full-stack Validated Solutions

Backed by direct access to full-stack AI expertise found nowhere else



The Best of NVIDIA Software – Included with DGX
Pre-trained models, optimized frameworks, data science tools, infrastructure management tools and more, with full support from NVIDIA

SoftBank and NVIDIA Backed
Even "supported systems" might not be backed by people who know AI intimately – ours do

Single Point of Contact
From framework to libraries to drivers to network, storage and compute – we're one stop for answers and uptime

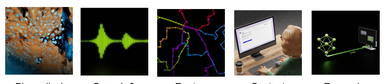
Fastest Path to Resolution
Unlike commodity servers offering commodity support - AI problems resolve quicker with direct access to NVIDIA Enterprise Support

The NVIDIA DGX Platform

Benefits

- Best-of-Breed for AI Development
- Direct Access to NVIDIA AI Experts
- Best Performance, Predictable Cost

Modern AI Use Cases

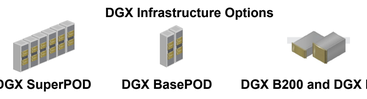


NVIDIA NIM, NVIDIA CUDA-X Microservices, AI Application Frameworks, and End-to-End AI Development Tools

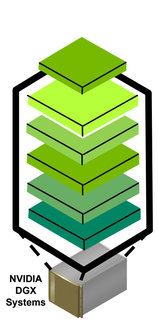
NVIDIA AI Enterprise

Cluster Management, Developer Workflow & Orchestration

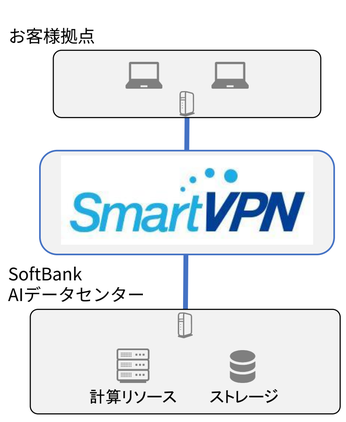
DGX Infrastructure Options

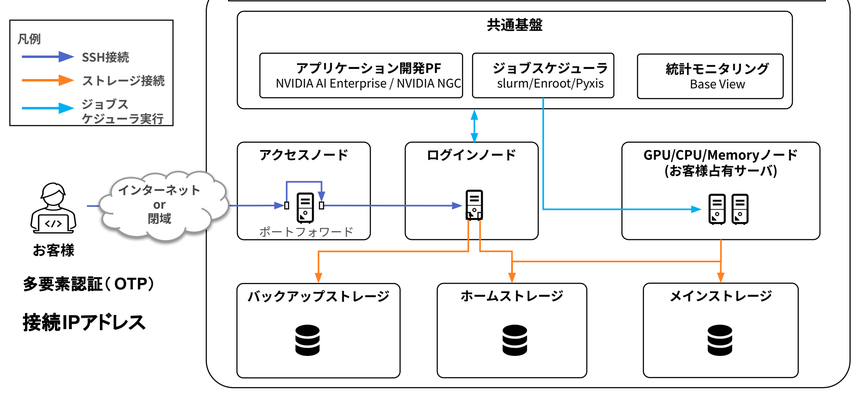


DGX SuperPOD | DGX BasePOD | DGX B200 and DGX H200



Secure connection via SmartVPN





SCALE TO MULTIPLE NODES

Software Stack - Application

Deep Learning Model:

- Hyperparameters tuned for multi-node scaling
- Multi-node launcher scripts

Deep Learning Container:

- Optimized TensorFlow, GPU libraries, and multi-node software

Host:

- Host OS, GPU driver, IB driver, container runtime engine (docker, enroot)