Solving the Generative AI Infrastructure Challenge in 2024 [S62227]

Charlie Boyle, VP, DGX Platforms, NVIDIA
William Mayo, Senior Vice President for Research IT, Bristol Myers Squibb

Agenda

- GTC Key Announcements
- DGX Platform Update
- Bristol Myers Squibb Overview with William Mayo
- Wrap-up and Q&A

State of Generative AI Adoption

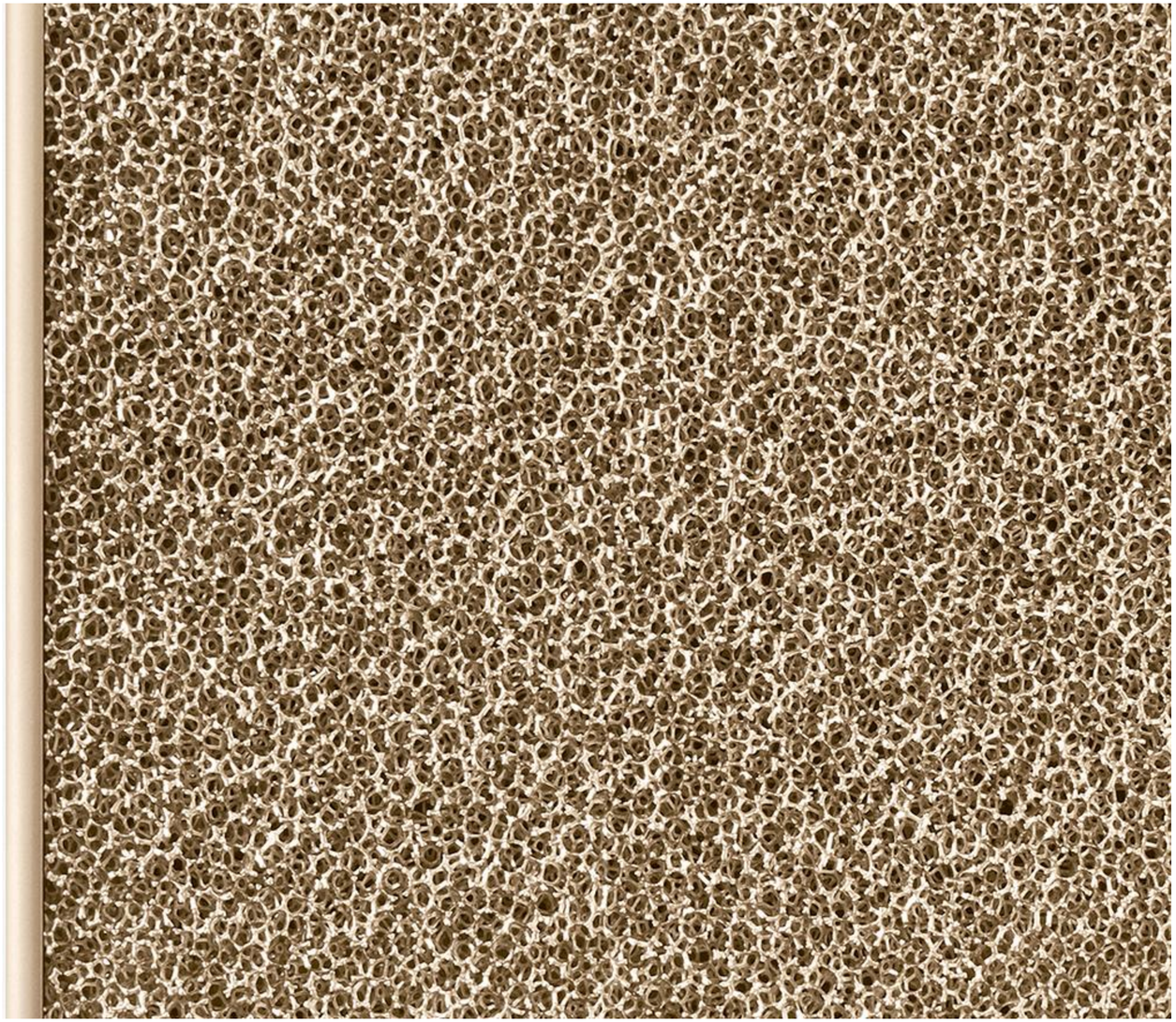
- Organizations are reporting an average of **3.5X ROI** for every $spent and 5% report and average of 8X ROI[1]
- By 2024, 33% of G2000 companies will exploit innovative business models to **double their monetization potential** of generative AI[2]
- By 2028, GenAI-based tools will be capable of **writing 80% of software tests** decreasing the need for manual testing, resulting in improvements to test coverage, software usability and code quality[3]

[1] Source: IDC Infographic, sponsored by Microsoft, The Business Opportunity of AI, doc #US51315823, November 2023
[2] Source: Worldwide AI and Automation 2024 Predictions; IDC #AP50341323, Oct 2023
[3] Source: Worldwide Generative AI 2024 Predictions; IDC #US51291623, Oct 2023

GTC Key Announcements for DGX Customers

- NVIDIA DGX B200 provides enterprises with a unified AI platform optimized for every stage of the AI pipeline, from training to fine-tuning to inference.

- DGX SuperPOD with DGX GB200 systems is liquid-cooled, rack-scale AI infrastructure for training & inferencing multi-trillion parameter gen AI models.



The NVIDIA DGX Platform

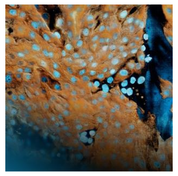
Benefits

Best-of-Breed for AI Development

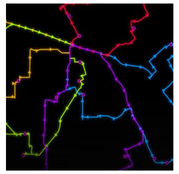
Direct Access to NVIDIA AI Experts

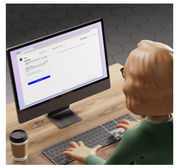
Best Performance, Predictable Cost

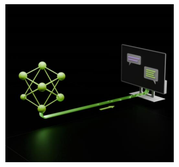
Modern AI Use Cases


Biomedical


Speech & Translation


Route Planning


Content Generation


Reasoning

More....

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

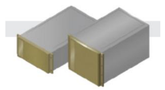
NVIDIA AI Enterprise

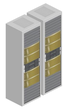
Developer Workflow, Cluster Management, & Orchestration

Base Command

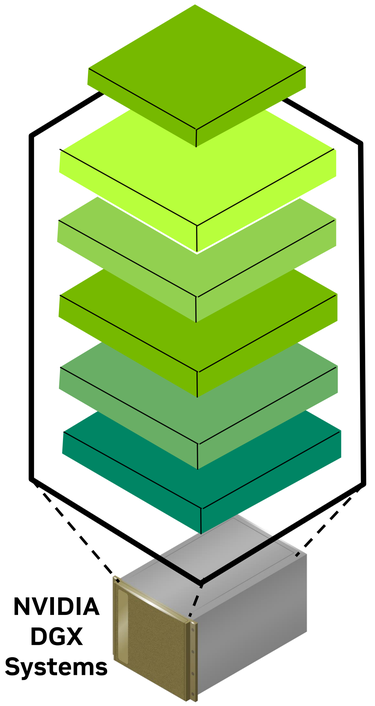
DGX Infrastructure Options


DGX Cloud


DGX B200 and DGX H100


DGX BasePOD


DGX SuperPOD



The Power of DGX SuperPOD in Action

Delivering generative AI solutions – accelerating drug discovery, enhancing productivity, and delighting customers

Bristol Myers Squibb®

Developing an AI center of excellence for drug discovery

servicenow

Customizing LLMs to bring intelligent workflow automation to enterprises

Opera

Empowering browser users with custom capabilities and a cutting-edge browser AI

NYU Langone Health

Optimizing Patient Outcomes: Leveraging LLMs for Personalized Care

SoftBank

Developing custom LLMs for Japanese language, enabling tailored AI solutions

DeepL

Building LLMs to empower businesses and individuals to break down language barriers

BNY MELLON

Leveraging AI to help power the future of financial markets

SONY

Providing computing resources for R&D projects across all Sony Group companies

ebay

Building e-commerce foundation models from the ground up

Introducing DGX SuperPOD With DGX GB200 Systems

Turnkey supercomputing for trillion-parameter AI

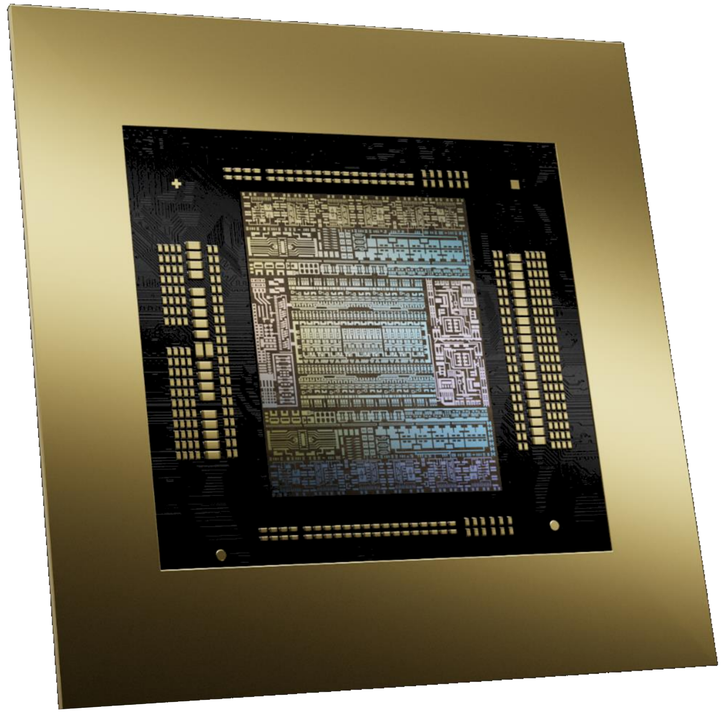
- Highly efficient, liquid-cooled, rack-scale design built with NVIDIA GB200 Grace Blackwell Superchips
- **36** NVIDIA Grace CPUs and **72** NVIDIA Blackwell GPUs per rack, connected via fifth-generation NVLink
- Scale to tens of thousands of GB200 Superchips with Quantum-2 InfiniBand
- Intelligent, full-stack resilience for constant uptime
- Integrated hardware and NVIDIA AI software
- Built, cabled, and factory tested before delivery and installation
- Optional **576** NVLink configuration for memory-limited workloads

The World's Most Efficient AI Infrastructure

288 Grace CPUs | 576 Blackwell GPUs
240TB Fast Memory | 11.5 ExaFLOPS FP4
30X Inference | 4X Training | 25X Energy Savings

Announcing Fifth Generation NVLink and NVLink Switch Chip

Efficient Scaling for Trillion Parameter Models



7.2 TB/s Full all-to-all Bidirectional Bandwidth

Sharp v4 plus FP8

3.6 TF In-Network Compute

Expanding NVLink up to 576 GPU NVLink Domain

18X Faster than Today’s Multi-Node Interconnect

DGX SuperPOD with DGX GB200 Systems Delivers Constant Uptime

Full-stack resiliency and predictive maintenance

Application Resiliency

Auto Checkpoint and Restart

System Software Resiliency

Unified Error Collection, Reporting, and Prediction

Rack Resiliency

Auto Node Lifecyle Management

Warm Spares

Data Center Resiliency

Preventative Maintenance

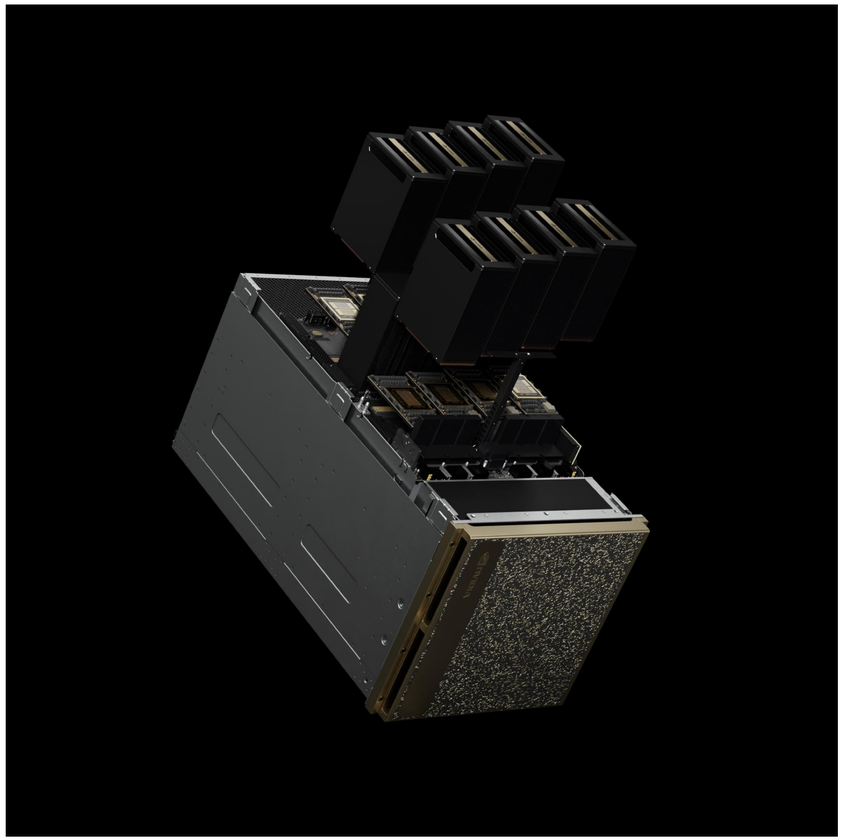
Cluster Health Checks ⟷ Predictor Engine

Data Center Operations Optimizations

Quick Swap Onsite Components & Automated RMA

DGX B200

The foundation of the modern AI data center



DGX B200

- Next generation DGX system with 8X NVIDIA Blackwell GPUs
- 1.4TB of GPU memory, enabling training of large generative AI models
- Purpose-built, unified platform for every workload from training, to fine-tuning, to inference
- Delivers 3X AI training and 15X AI inference performance as previous generation (DGX H100)
- Latest Blackwell architecture in a scalable, air-cooled design

NVIDIA Base Command Powers the DGX Platform

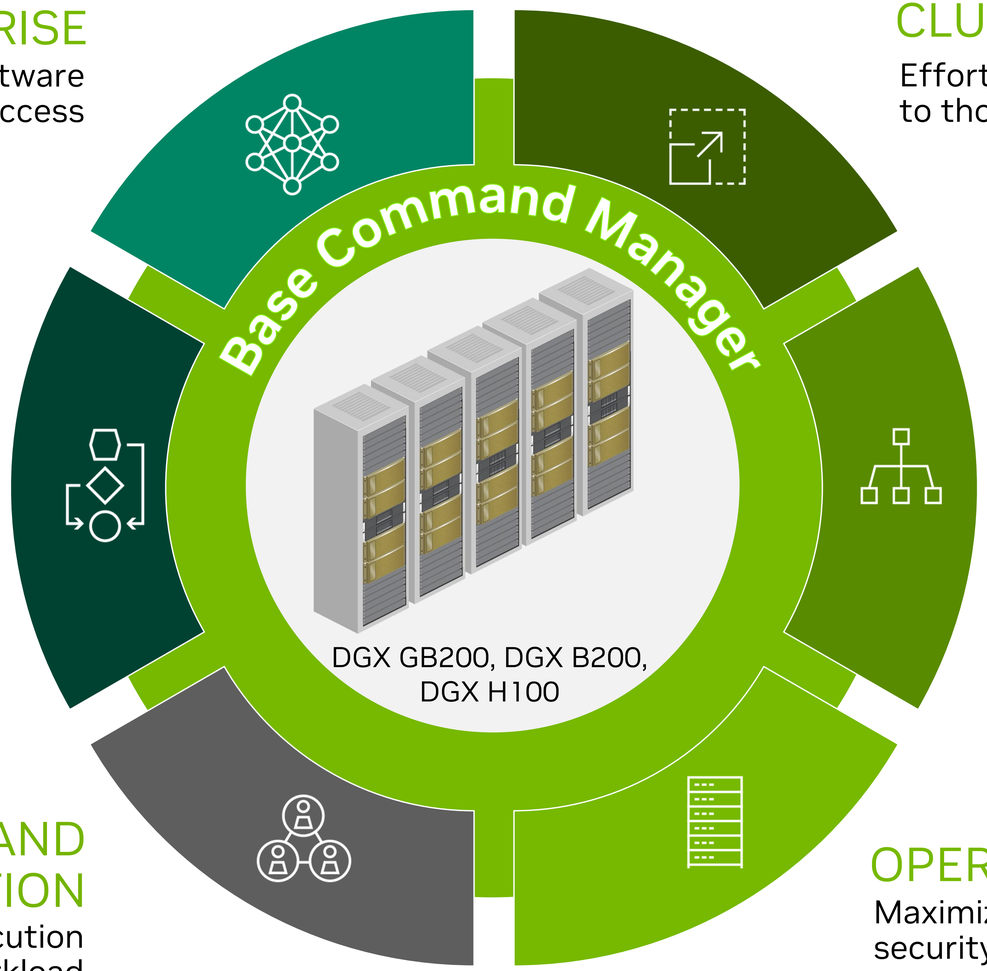
Enterprise software that drives the value of AI investment

NVIDIA AI ENTERPRISE

Ready-to-use, fully supported software that speeds up developer success

CLUSTER MANAGEMENT

Effortlessly scale and manage one node to thousands



AI WORKFLOW MANAGEMENT AND MLOPS

Get more models from prototype to production

NETWORK/STORAGE ACCELERATION LIBRARIES & MANAGEMENT

Accelerate end-to-end infrastructure performance

JOB SCHEDULING AND ORCHESTRATION

Ensure hassle-free execution of every workload

OPERATING SYSTEM

Maximize system uptime, security, and reliability

The DGX Platform: Faster ROI on AI Investments

60%
reduction in system administration time

21%
increase in developer productivity

50%
improvement in IT efficiency

25X
energy savings[1]

30X
speed-up in AI inference performance[1]

4X
speed-up in AI training performance[1]

Achieve lower TCO for development operations, faster ROI on AI projects

AI infrastructure that:

- Enables delivery of more AI prototypes into production
- Orchestrates workloads and optimizes uptime
- Meets growing business demands and user count
- Continuously improves over time

1. X factor compares DGX GB200 to DGX H100

William Mayo
Senior Vice President for Research IT
Bristol Myers Squibb


Bristol Myers Squibb®

Transforming patients' lives through science™

Our mission

To discover, develop and deliver innovative medicines that **help patients prevail** over serious diseases

By the numbers

Science

165 years
of innovation

$9.3B
R&D investment in 2023

45+
unique assets in clinical development*

Patients

85
countries where BMS serves millions of patients*

5
therapeutic areas where BMS helps patients*

People

34K+
employees globally*

13K
employees in eight People Business and Resource Groups*

Society

$663M
in corporate giving to nearly 5,000 organizations in 57 countries over three years (2021 to 2023)

*As of December 2023

Pioneering science for serious diseases with unmet needs

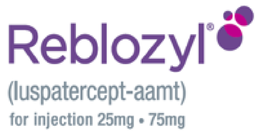
Leading medicines across therapeutic areas

Solid tumor oncology

Cell therapy

Immunology

Hematology



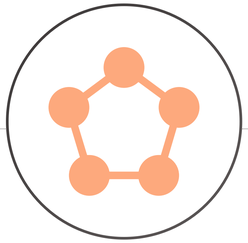
Cardiovascular

Neuroscience

Research is focused on the delivery of high-quality assets that enable clinical success

Strong causal human biology

Use of human data (e.g., genetics, longitudinal profiling) for rigorous target validation in drug discovery



Matching modality to mechanism

Invention of high-quality therapeutic modalities that match a therapeutic modality to a molecular mechanism of action

Path to clinical proof-of-concept

Targeted patient selection (e.g., biomarkers) and clear translational endpoints for greater likelihood of clinical response

Our ambition is to increase probability of success across discovery & development

Fireside chat with William Mayo and Charlie Boyle

Wrap-Up

Some Recommended Upcoming DGX Sessions

S62494: Accelerating the Generative AI Transformation: Expert Insights for Rapid Innovation and Scale

Speakers: Jeremy Barnes (ServiceNow), John Parkhill (Terray), Kristene Aguinaldo (JHU APL), Tony Paikeday (NVIDIA)

Time: Mar 20 | 8:00 AM -8:50 AM

Location: SJCC 210D (L2)

S62421: The Next-Generation DGX Architecture for Generative AI

Speakers: Julie Bernauer (NVIDIA Senior Director, Data Center Systems Engineering), Mike Houston (NVIDIA VP and Chief Architect of AI systems)

Time: Mar 20 | 10:00 AM -10:50 AM

Location: SJCC 230C (L2)

Q&A


Thank you!