# GTC 2026 最新情報 for JAWS-UG HPC

2026/04/17

佐々木 邦暢 (@_ksasaki)
エヌビディア合同会社


NVIDIA
GTC
It All Starts Here
SAN JOSE McENERY CONVENTION CENTER
GTC26

# NVIDIA は AI インフラ企業

AI は多層インフラとして、世界的な雇用創出と経済成長を促進

アプリケーション

モデル

インフラ

チップ

エネルギー

CHATBOTS

DIGITAL BIOLOGY

ROBOTAXI

ENTERPRISE AI AGENTS

SCIENCE

ROBOTICS

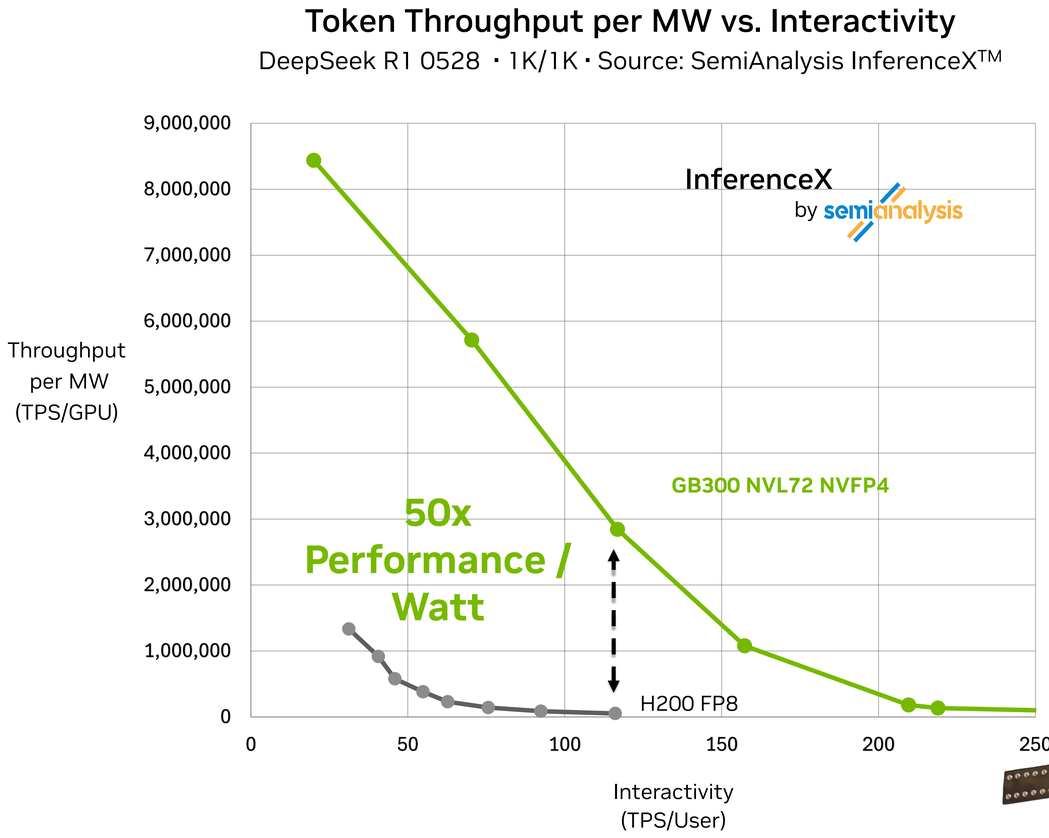
MANUFACTURING

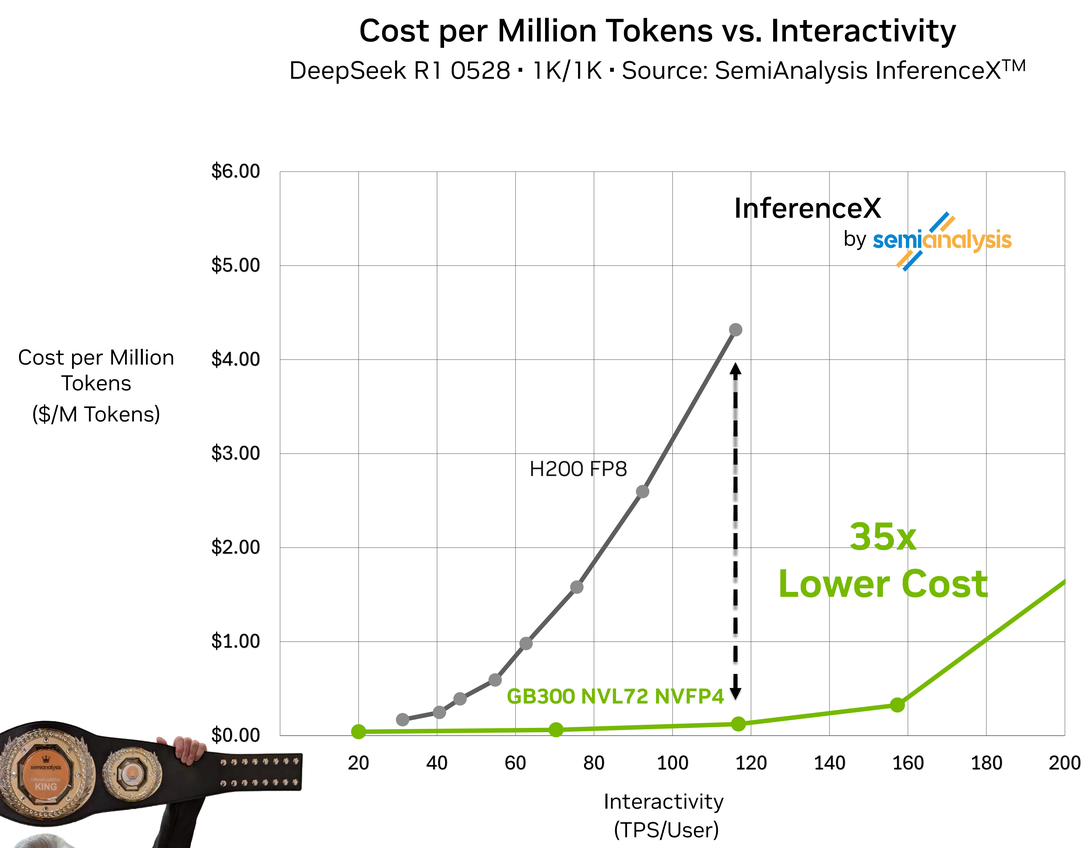
AI CODER

| | | | | |
|---|---|---|---|---|
| LLM | VLM | VLA | MMLLM | GPT |
| DM | GNN | MOE | SSM | LBM |

# GB300 NVL72：スループット/MWが50倍、トークンコストが35分の1に低下

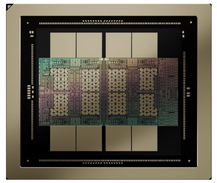
大幅なパフォーマンス向上により、大きな収益機会が生まれる



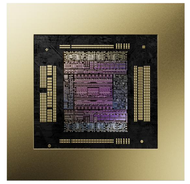



# Vera Rubin がエージェント AI の最前線を開拓
# 7 つの新チップ — 大きな飛躍

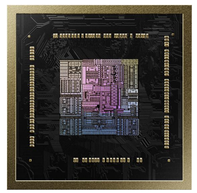
Vera



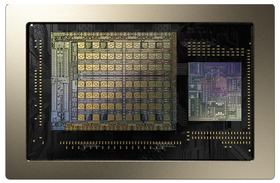
Rubin



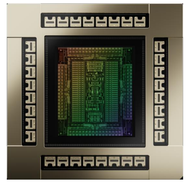
NVLink6 Switch



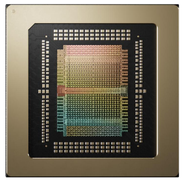
CX9



BF4



Spectrum-X
102.4T CPO



NVIDIA
Groq 3 LPU

Vera Rubin SuperPOD

STX
Storage

LPX

Vera Rubin NVL72

Vera
CPU

Spectrum-6
SPX (E-W traffic)

Vera
CPU

Vera Rubin NVL72

LPX

STX
Storage

Spectrum-6 SPX
(N-S traffic)

Spectrum-6 SPX
(N-S traffic)

NVIDIA

# NVIDIA Vera Rubin NVL72 : AI ファクトリーの性能を 10 倍に

| | | |
|---|---|---|
| NVFP4 Inference | 3.6 EFLOPS | **5X Blackwell** |
| NVFP4 Training | 2.5 EFLOPS | **3.5X** |
| Fast Memory | 75 TB | **2.5X** |
| HBM4 Bandwidth | 1.6 PB/s | **2.8X** |
| Scale-Up Bandwidth | 260 TB/s | **2X** |
| In-Network Compute | 130 TFLOPS | **2X** |

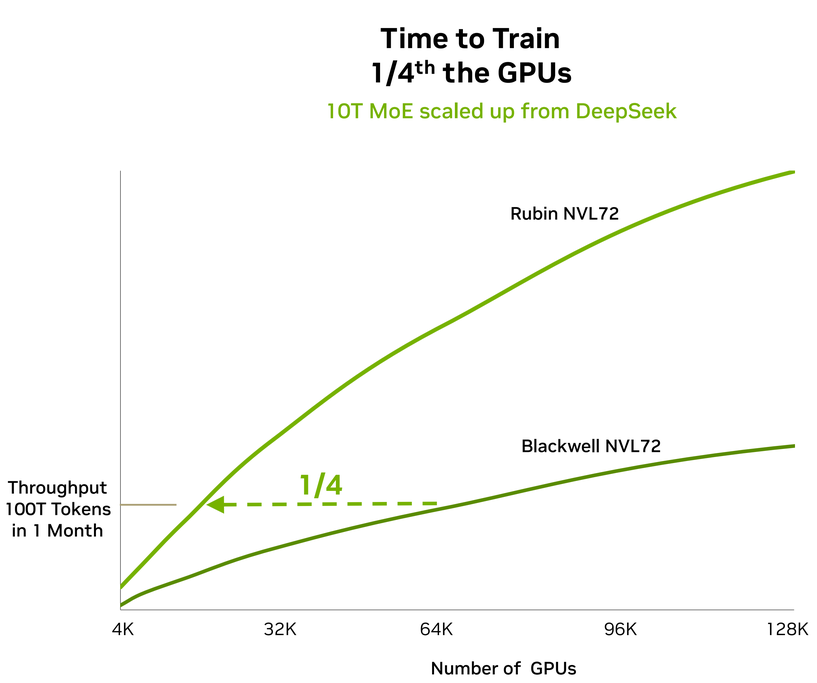
**10X**
Perf-per-Watt

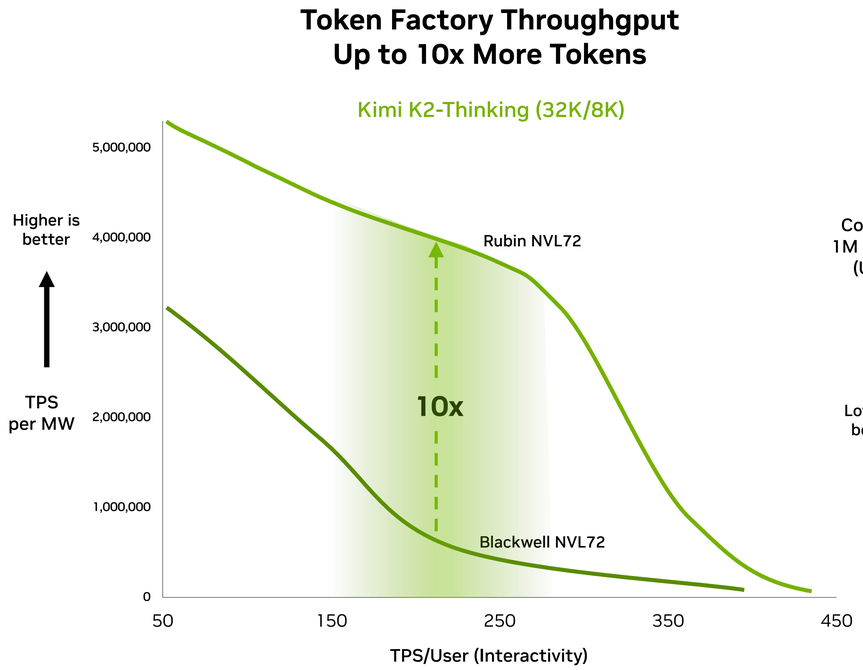
**1/10th**
Lower Token Cost

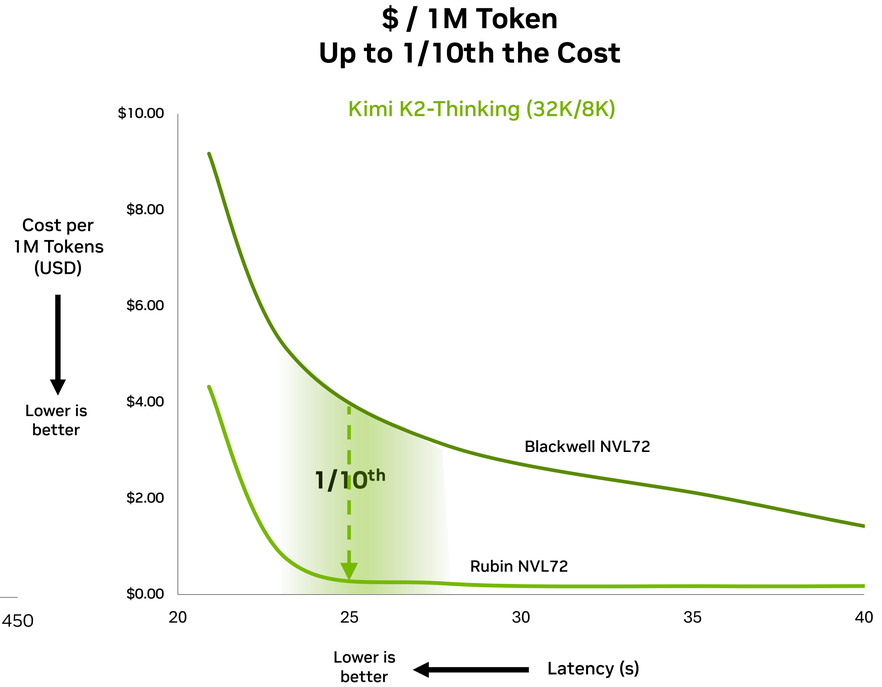
**1/4th**
Fewer GPUs to Train

*Results from SemiAnalysis InferenceX 2.0 comparing NVIDIA GB300 NVL72 to Hopper platform*

# Vera Rubin NVL72 MoE トレーニングおよび推論性能



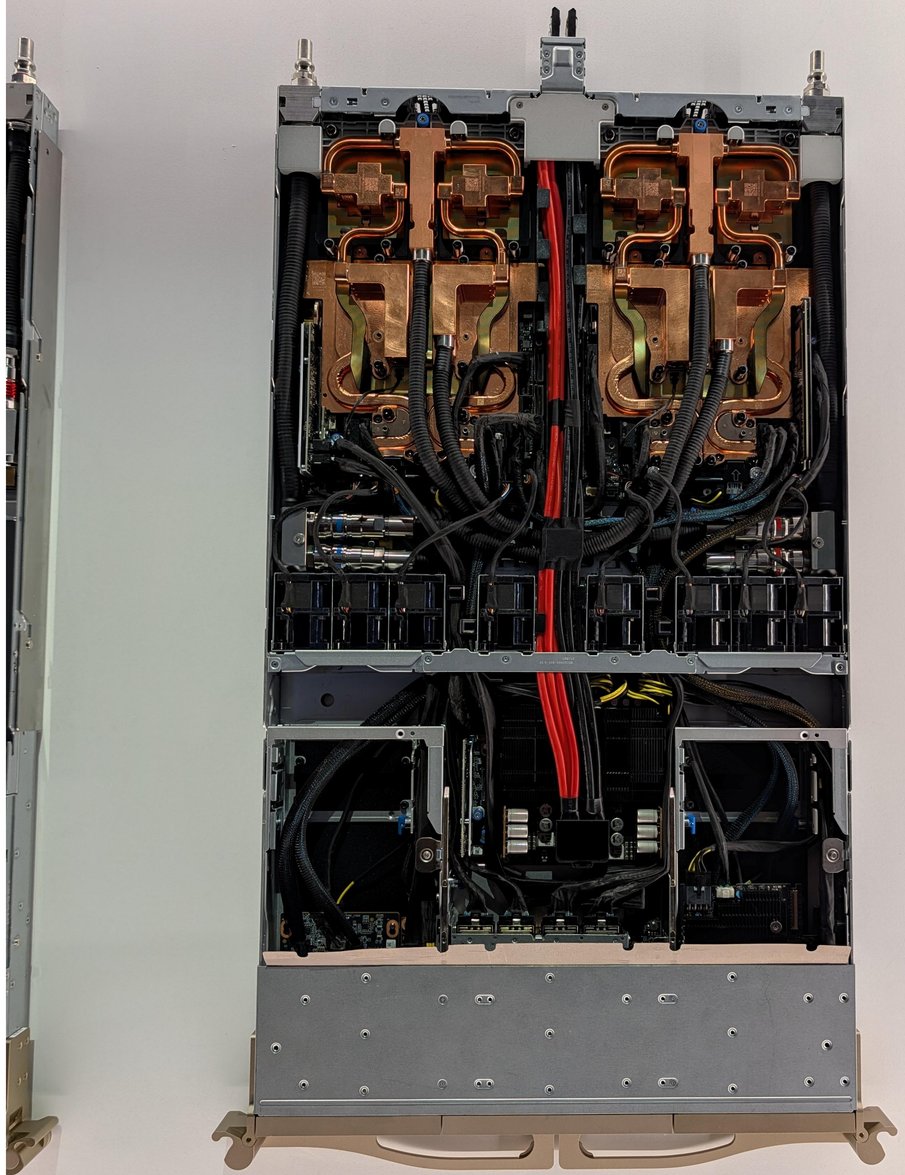



# Vera Rubin NVL72 Compute & Switch Trays

Cable-free, Hose-free, Fanless

NVLink Switch Tray

Compute Tray



NVIDIA GB300 NVL72
Compute Tray

NVIDIA Vera Rubin NVL72

# HGX Rubin NVL8

## Scaling AI Factories for Enterprise

### Supercharging AI and HPC for Every Data Center

400 PFLOPS NVFP4 Inference – 5.5x of Blackwell

280 PFLOPS NVFP4 Training

### Next-Generation Scale-Up and Scale Out

28.8TBps NVLink Switch Bandwidth

1.6 TBps Networking Bandwidth

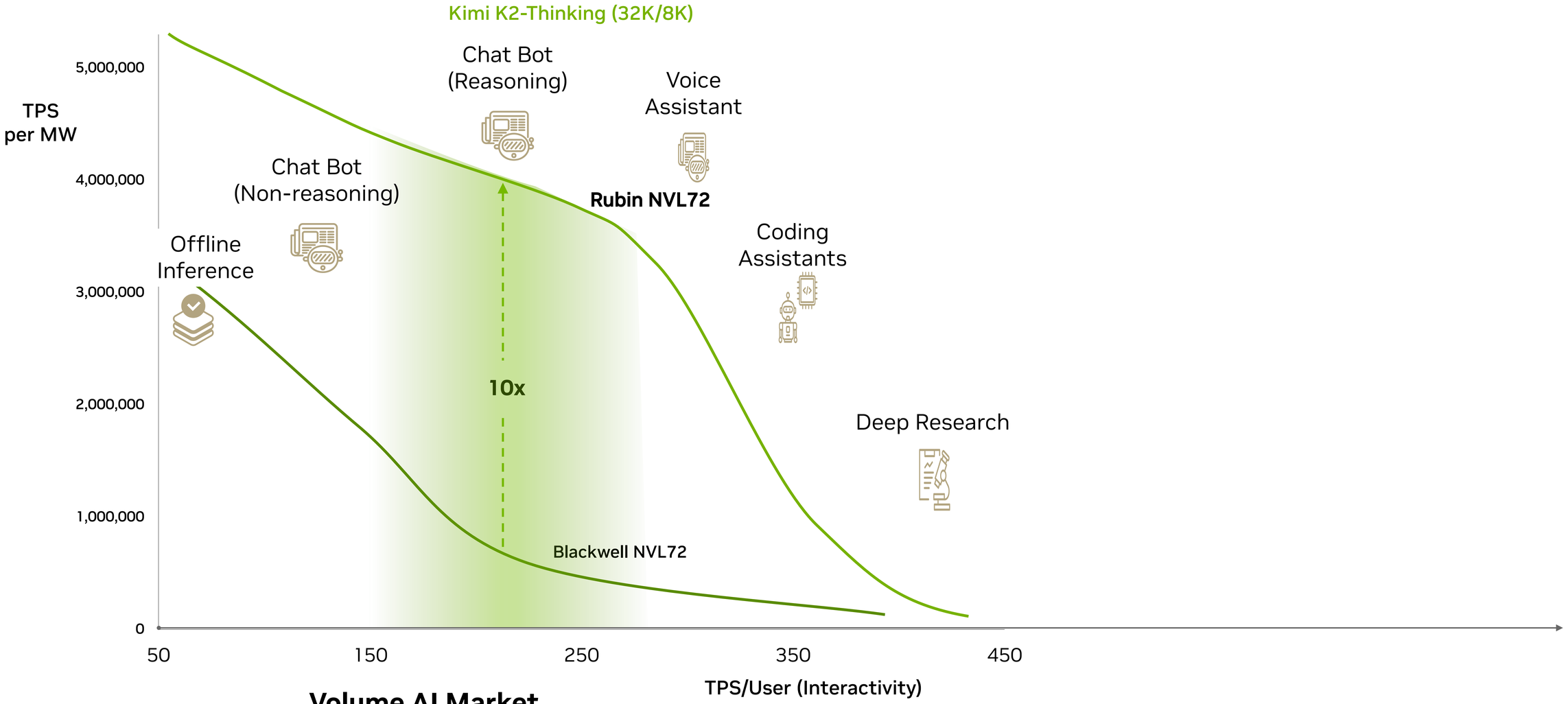
### Efficient Cooling for Maximum Performance

100% Liquid-Cooled

# NVIDIA Groq 3 LPU
# NVIDIA Groq 3 LPX

# Vera Rubin NVL72 による AI ファクトリー全体のワークロードの高速化

From Chatbots to Deep Thinking

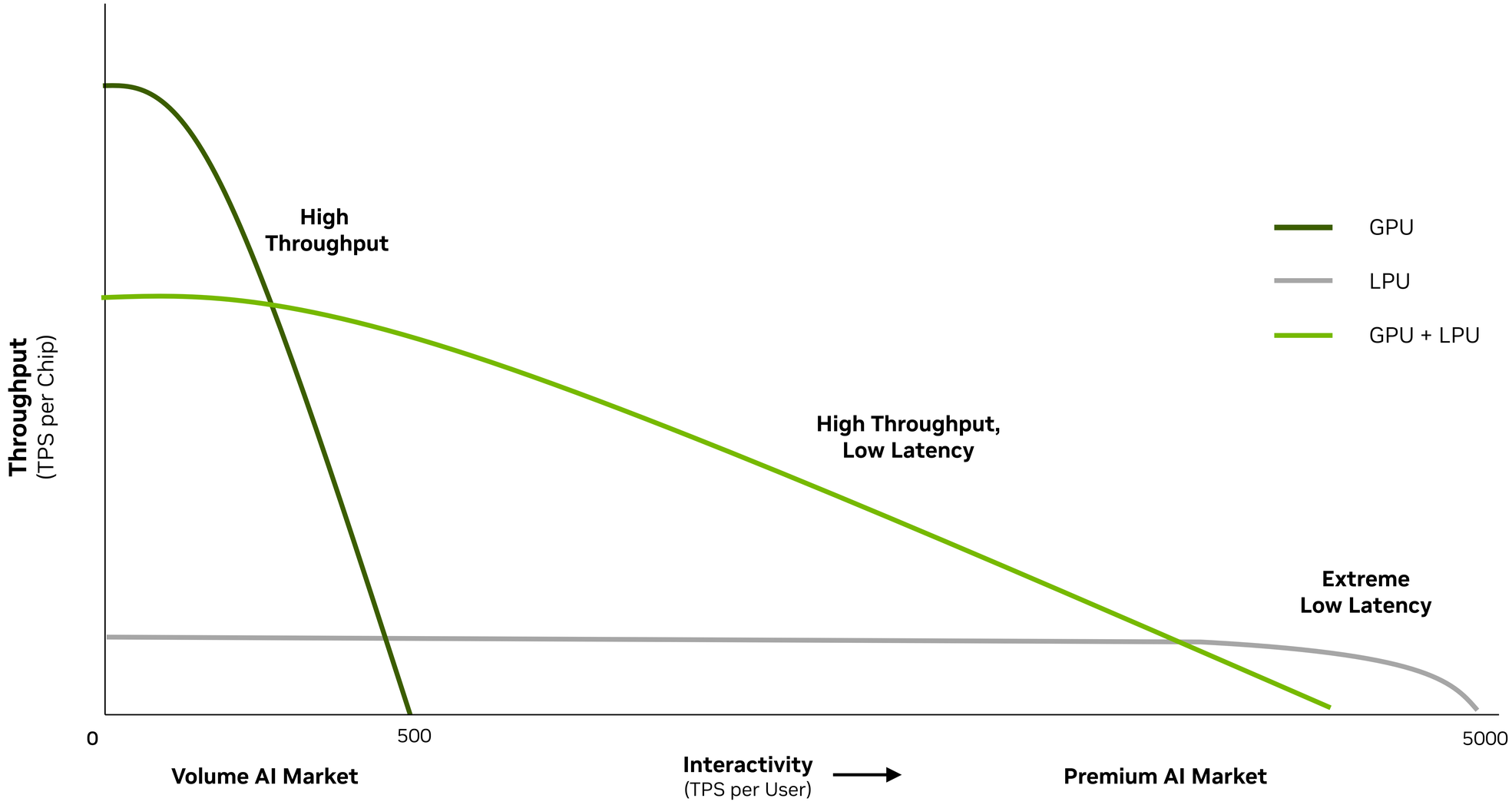



**Volume AI Market**

100B Params | 8K Context | 100 Tok/s

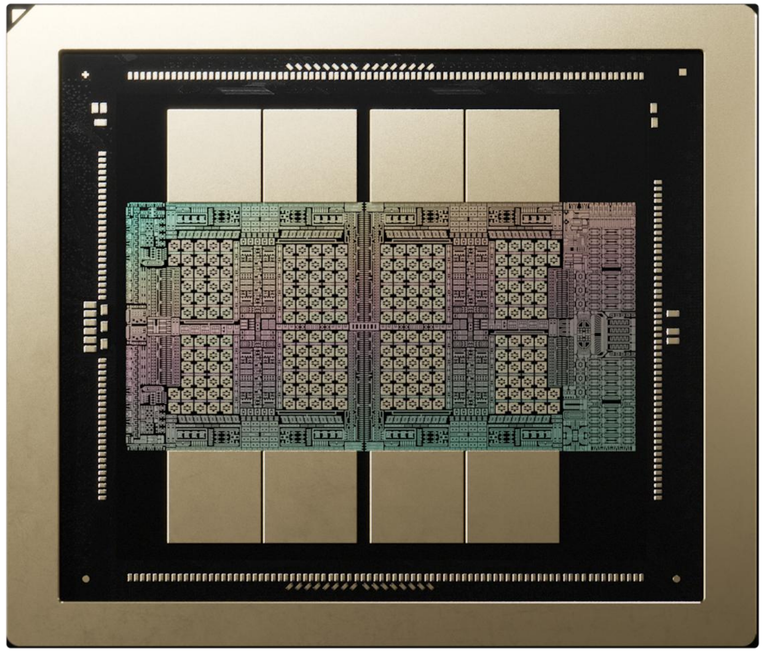
# GPU と LPU の統合が AI 性能を飛躍させる

低遅延と高スループットの両立



# NVIDIA Groq 3 LPU

極めて高い FLOPS と帯域幅を持つプロセッサの統合



| | |
|---|---|
| 288 GB HBM4 | 500 MB SRAM |
| 22 TB/s | 150 TB/s SRAM Bandwidth |
| 50 PFLOPs (NVFP4) | 1.2 PFLOPs (FP8) |
| 336B Transistors | 98B Transistors |
| Scale Up NVLink | Custom Chip to Chip |

# NVIDIA Groq 3 LPX

| | |
|---|---|
| NVIDIA Groq 3 LPUs | **256** |
| AI Inference Compute | **315 PFLOPS** |
| SRAM Capacity | **128 GB** |
| Memory Bandwidth | **40 PB/s** |
| Scale-Up Density | **256 Chips** |
| Scale-Up Bandwidth | **640 TB/s** |
| Scale-Out Density | **1000+ LPUs** |

8x NVIDIA Groq 3 LPU

1.2 PB/s SRAM Bandwidth

BlueField-4 AI Runtime Security

100% Liquid-Cooled

# NVIDIA Vera Rubin NVL72

| | | |
|---|---|---|
| NVFP4 Inference | 3.6 EFLOPS | **5X Blackwell** |
| NVFP4 Training | 2.5 EFLOPS | **3.5X** |
| Fast Memory | 75 TB | **2.5X** |
| HBM4 Bandwidth | 1.6 PB/s | **2.8X** |
| Scale-Up Bandwidth | 260 TB/s | **2X** |
| In-Network Compute | 130 TFLOPS | **2X** |

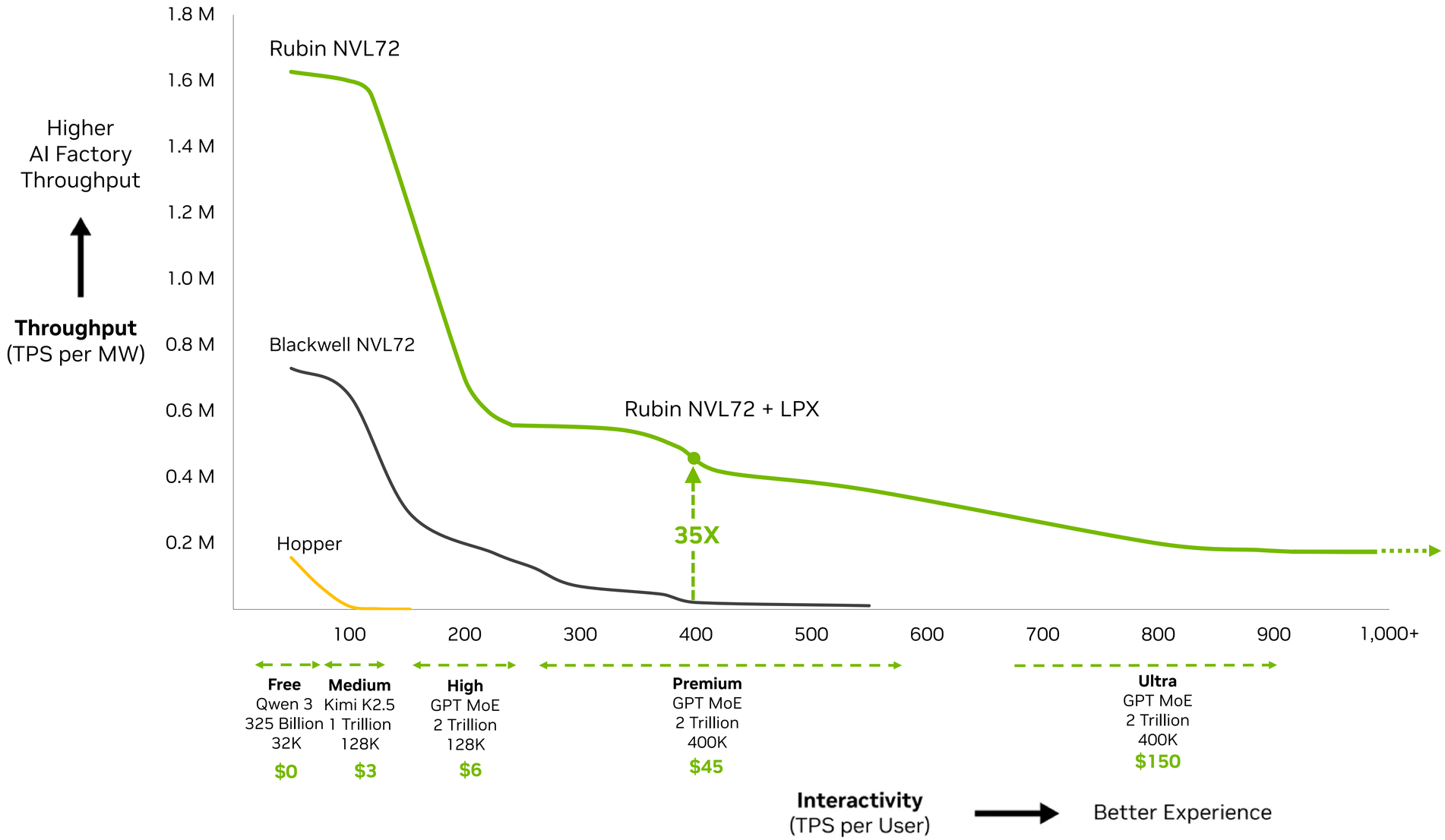
**10x**
**Perf-per-Watt**

**1/10th**
**Lower Token Cost**

**1/4th**
**Fewer GPUs to Train**

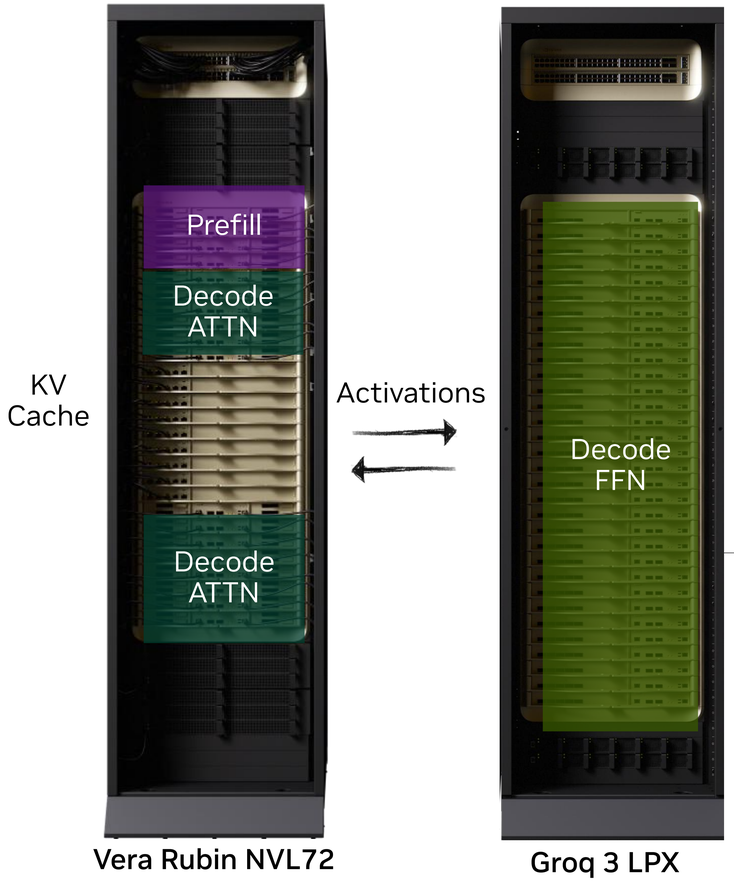
# AI 体験の新たなカテゴリーを切り拓く

Vera Rubin NVL72 + Groq 3 LPX により、高スループットと低遅延が実現



# Extreme Co-Design of Hardware and Software Drives NVIDIA Inference

## NVIDIA Dynamo 1.0 Production Ready – Available Today







### Broad Adoption

amazon · baseten · ByteDance · CoreWeave · Meituan · NEBIUS · PayPal · SoftBank · together.ai

### CSP Integration

Alibaba · aws · Google Cloud · Microsoft · ORACLE CLOUD Infrastructure

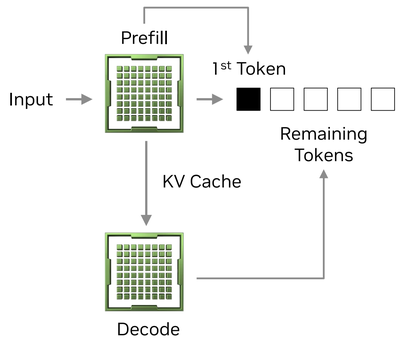
### OSS Integration

llm-d · LMCache · SGL · vLLM

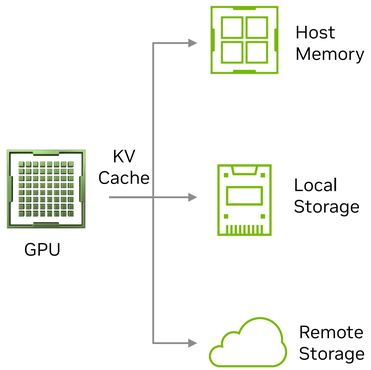
### Performance Optimizations

50x more performance when paired with Blackwell Ultra vs. Hopper

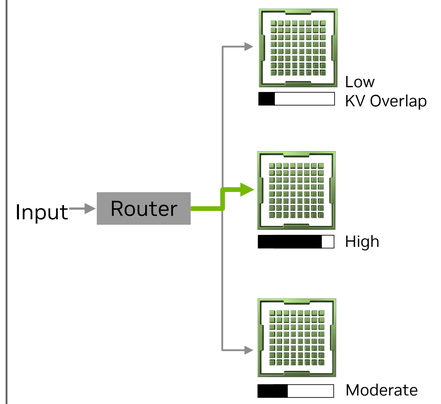
#### Disaggregated Serving




#### KV Offloading




#### KV-aware Routing



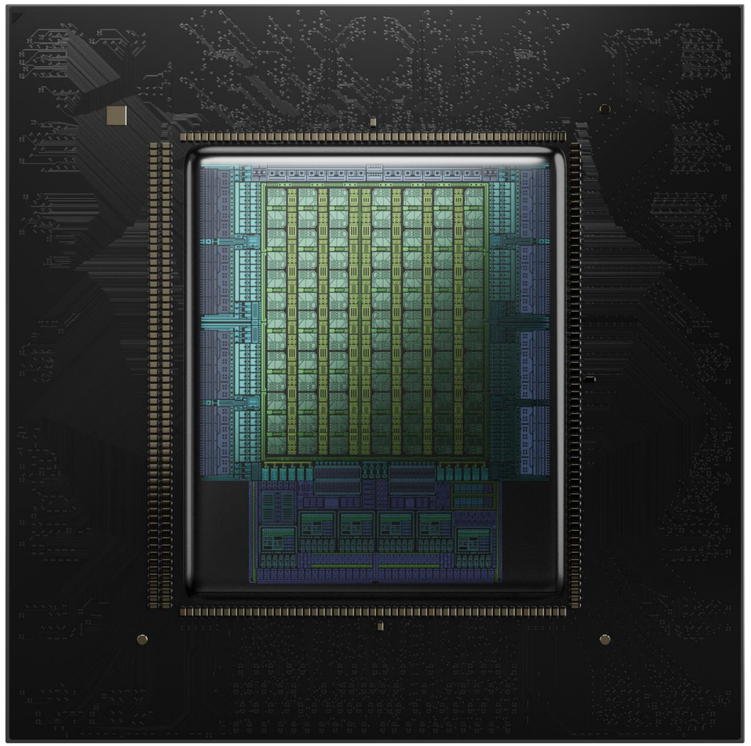

### Production Ready Features

Offline Configurator · Online Optimizer · Topology-Aware Scaling · Fault Tolerance

# NVIDIA Vera CPU

# NVIDIA Vera – AI エージェント時代の CPU

AI 実行向けに設計されたオリンパス コア — GPU ホスティングからエージェントワークロードまで



88 NVIDIA Custom Olympus Core– Fastest Single-threaded Performance

Neural Branch Predictor. 1.5x IPC vs Grace

Custom Graph Data Analytics Prefetch Engine

Spatial Multithreading – 176 Threads Full Core Utilization

2nd Gen Scalable Coherency Fabric – Uniform Data Access at Scale

SOCAMM LPDDR5X - Half the memory power, twice the bandwidth

Full Confidential Computing Support

**1.5x**
Perf-per-Sandbox
Compared to x86

**3x**
Memory BW
per Core

**2x**
More Power
Efficient

# NVIDIA Vera – AI エージェント時代の CPU

大規模 AI の継続的な実行に向けて

**3x**
perf for SQL (vs x86)

Starburst

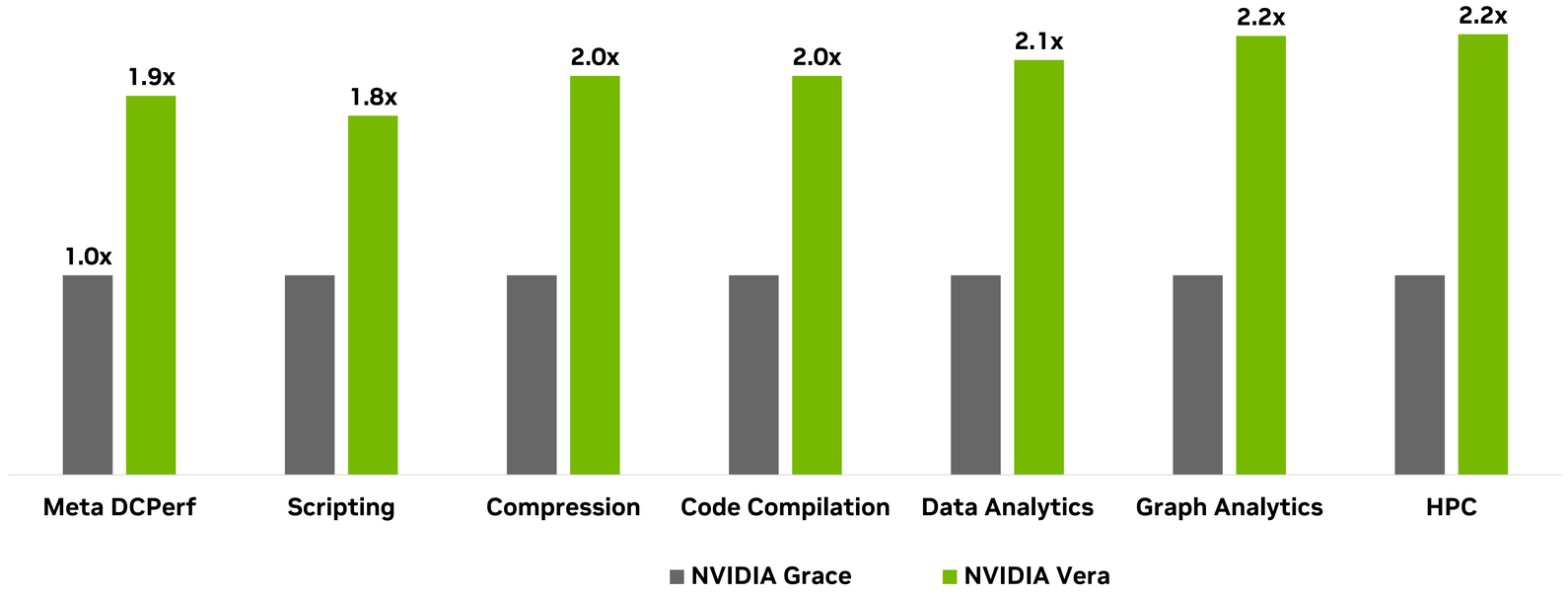
**2x**
perf for hyperscale cloud (vs Grace)

Meta DCPerf

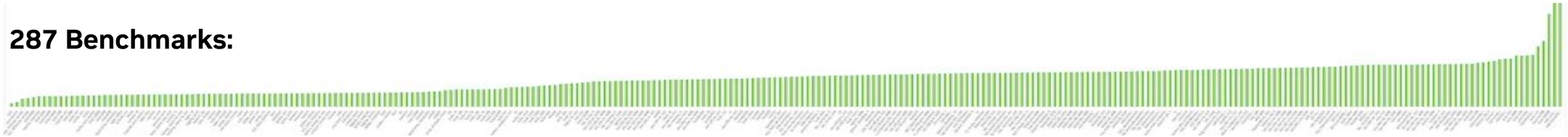
**2x**
perf HPC Golden Suite (vs Grace)

TACC

**5.5x**
lower latency Apache Kafka (vs x86)
Redpanda




**287 Benchmarks:**



# NVIDIA Vera CPU ラック

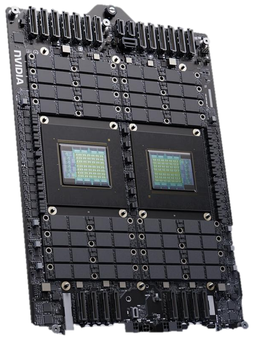
エージェント型AIと大規模な強化学習

| | |
|---|---|
| Vera CPUs | 256 |
| Cores \| Threads | 22,528 \| 45,056 |
| Memory | Up to 400 TB \| 300 TB/s LPDDR5X |
| BlueField-4 DPUs | 64 |
| Rack | MGX Compatible |
| Cooling | Liquid Cooled |
| Networking | Spectrum-X Ethernet |
| Energy Efficiency | 2x |

# Vera エコシステムパートナー

データ駆動型コンピューティングおよび AI コンピューティング向けに設計

Vera CPU Rack



2S and 1S
Vera CPU Servers

Hyperscalers & Clouds

Meta
ORACLE CLOUD Infrastructure
CoreWeave
Crusoe
Lambda
NEBIUS
NSCALE
together.ai

OEM and ODM Partners

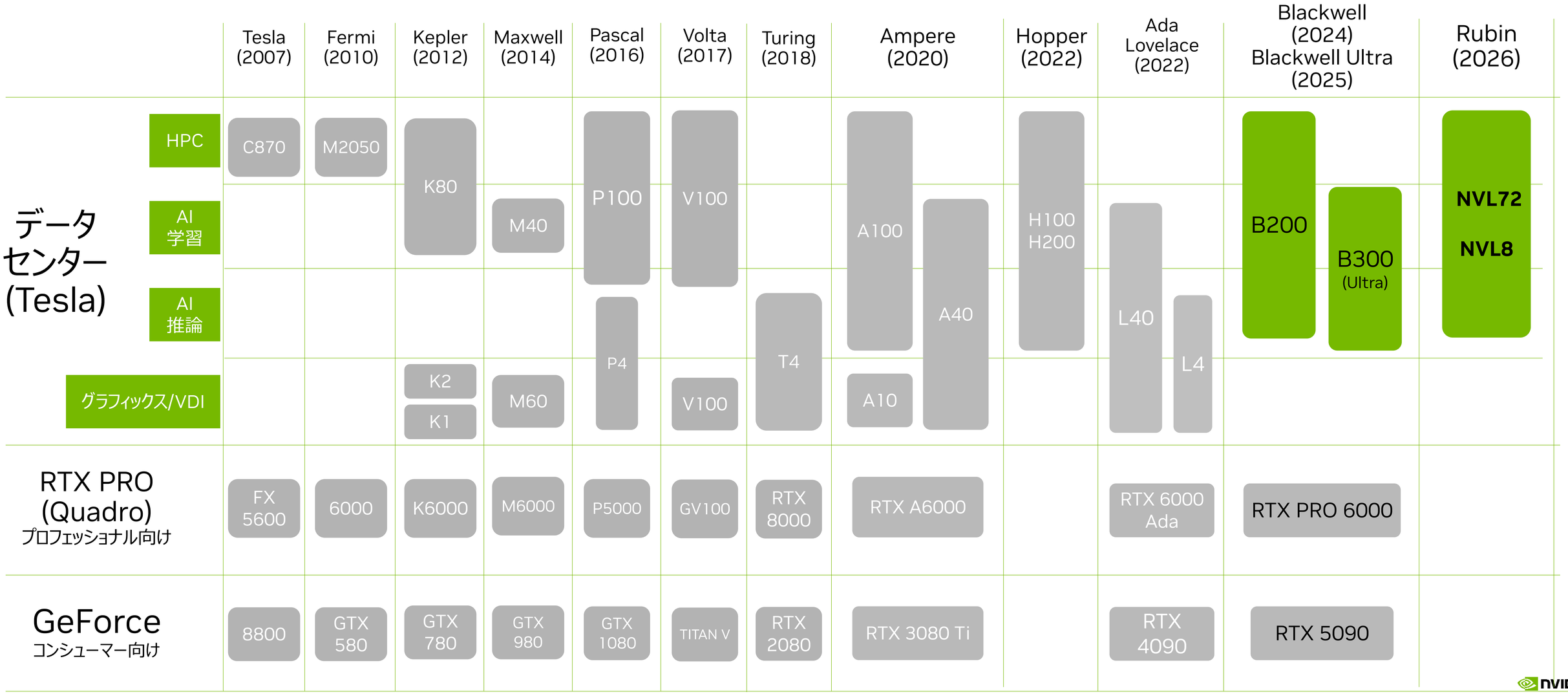
CISCO
DELL Technologies
HPE
Lenovo
SUPERMICRO
AIVRES
ASRock Rack
ASUS
COMPAL
FOXCONN
GIGABYTE
hyve solutions
Inventec
MiTAC
msi
PEGATRON
QCT
wistron
wiwynn

NVIDIA

# 最新 GPU 製品ラインナップ

# NVIDIA GPU 製品のおおまかな一覧

PC / ワークステーション / データセンター向け



# NVIDIA Vera Rubin NVL72

AI Supercomputer for the next generation of AI - 36 Vera CPUs and 72 Rubin GPUs

| | | |
|---|---|---|
| NVFP4 Inference | 3.6 EFLOPS | 5X Blackwell |
| NVFP8 Training | 1.2 EFLOPS | 3.5X |
| LPDDR5X Capacity | 54 TB | 2.5X |
| HBM Capacity | 20.7 TB | 1X |
| HBM4 Bandwidth | 1.6 PB/s | 2.5X |
| Scale-Up Bandwidth | 260 TB/s | 2X |

Also offering NVIDIA HGX Rubin NVL8

# Blackwell Ultra | Blackwell for Every AI Use Case

Delivering the Highest Performance for AI Factories

## Blackwell Ultra

**P6e-GB300**

GB300 NVL72

Maximum Performance & Lowest TCO for AI Reasoning Liquid-Cooled

**P6-B300**

HGX B300

Best Performance & TCO for Air-cooled HGX

## Blackwell

**P6e-GB200**

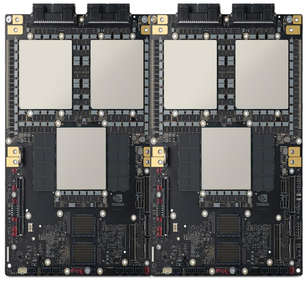
GB200 NVL72

Compute for at Scale AI Training & Inference Liquid-Cooled

**P6-B200**

HGX B200

Optimized Performance & TCO for Air-cooled HGX



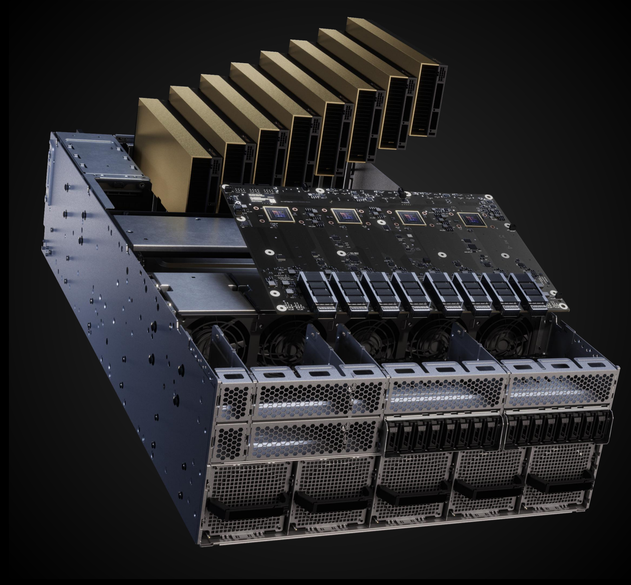
GB200 NVL4

Best HPC + AI Performance in a single server

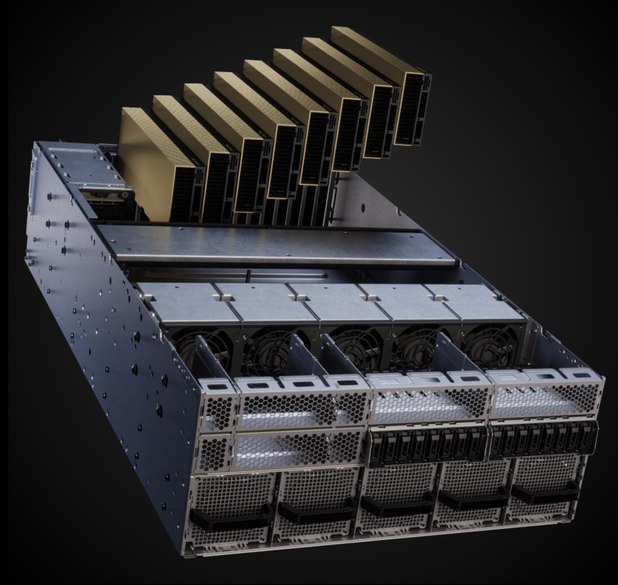
# RTX PRO Server

G7e

NVIDIA Blackwell for every enterprise data center



**MGX 4U Configuration**
8X RTX PRO 6000 Blackwell Server Edition GPUs
ConnectX-8 SuperNIC with PCIe Gen 6 Switch



**4U Configuration**
8X RTX PRO 6000 Blackwell Server Edition GPUs

**2U Configuration**
2X RTX PRO 6000 Blackwell Server Edition GPUs

cisco DELL Technologies HPE Lenovo SUPERMICRO

ADVANTECH aetina ASRock Rack ASUS COMPAL GIGABYTE ingrasys

Inventec MiTAC MiTAC Computing msi PEGATRON QCT wistron wiwynn

NVIDIA

# NVIDIA HGX H200

**P5e**

The world's leading AI computing platform

## Highest Performance for AI and HPC

8-way or 4-way H200 GPUs

Up to 32 PetaFLOPs FP8

Up to 1.1TB High Bandwidth Memory

## Fastest, Scalable Interconnect

4th Gen NVLINK with 2X faster All-Reduce communications

3.6 TB/s bisection bandwidth

## Fully Compatible with Partner H100 Systems

Supported by Leading Major OEMs and CSPs


Coming to Leading OEM and CSP Partners

Starting Q2 2024

ASRock Rack · ASUS · aws · CoreWeave · DELL Technologies · EVIDEN an atos business

GIGABYTE · Google Cloud · Hewlett Packard Enterprise · Lambda · Lenovo · Microsoft Azure

ORACLE CLOUD Infrastructure · QCT · SUPERMICRO · VULTR · wistron · wiwynn

NVIDIA

# NVIDIA HGX B300

**P6-B300**

## Purpose Built for AI Reasoning

### Increased Performance for AI Compute

7X increase in AI compute[1]
2TB of HBM3E memory
Integrated NVIDIA ConnectX-8 SuperNICs

### Boost Revenue

30x overall increase in AI factory output performance

### Next-Level AI Training Opportunities

2.6x higher LLM training performance
14.4 TB/s of NVLink Switch bandwidth

1. 7x increase over NVIDIA Hopper platform

# HGX Rubin NVL8

???

## Scaling AI Factories for Enterprise

### Supercharging AI and HPC for Every Data Center

400 PFLOPS NVFP4 Inference – 5.5x of Blackwell

280 PFLOPS NVFP4 Training

### Next-Generation Scale-Up and Scale Out

28.8TBps NVLink Switch Bandwidth

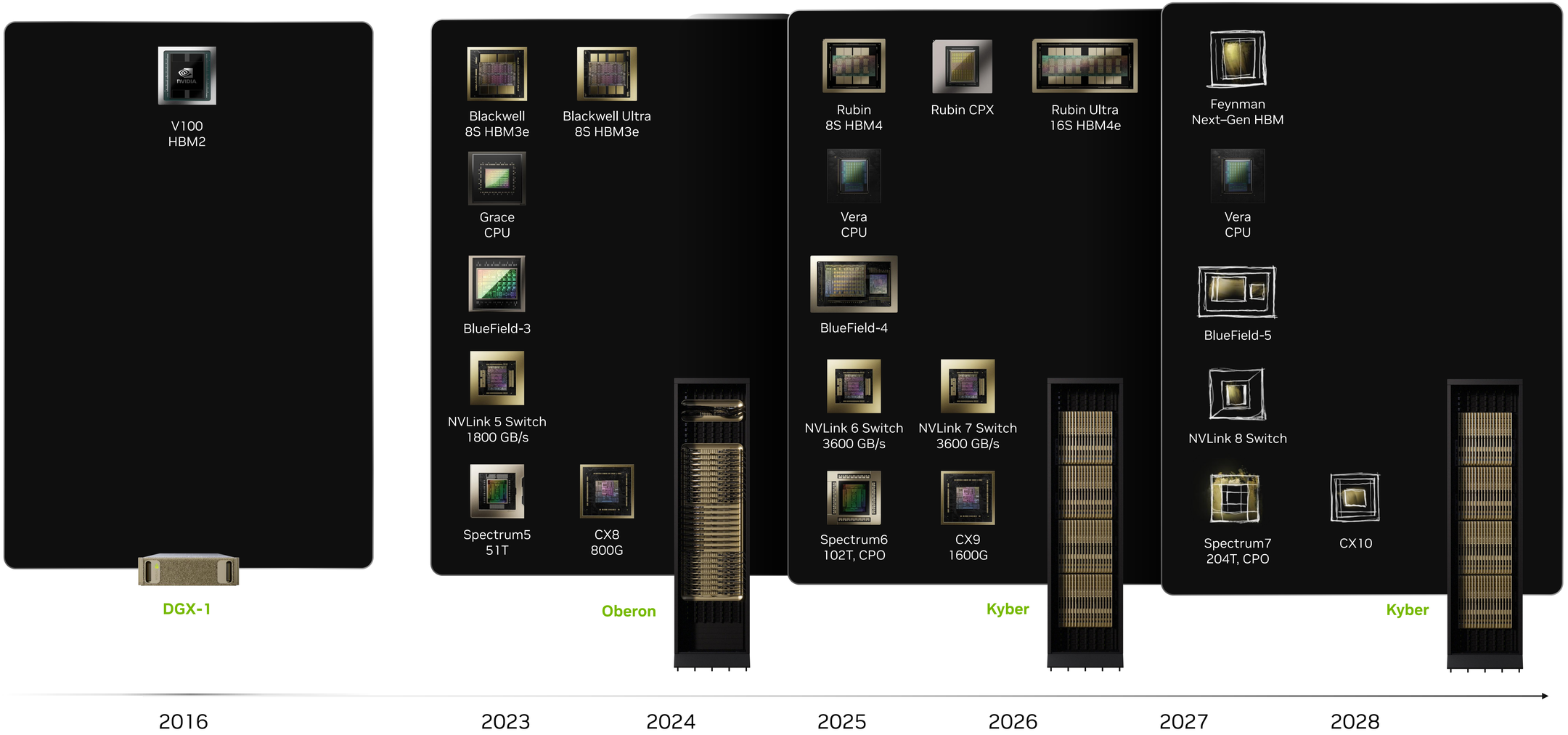
1.6 TBps Networking Bandwidth

### Efficient Cooling for Maximum Performance

100% Liquid-Cooled

# NVIDIA Platform Delivers Continuous Innovation With One-Year Rhythm

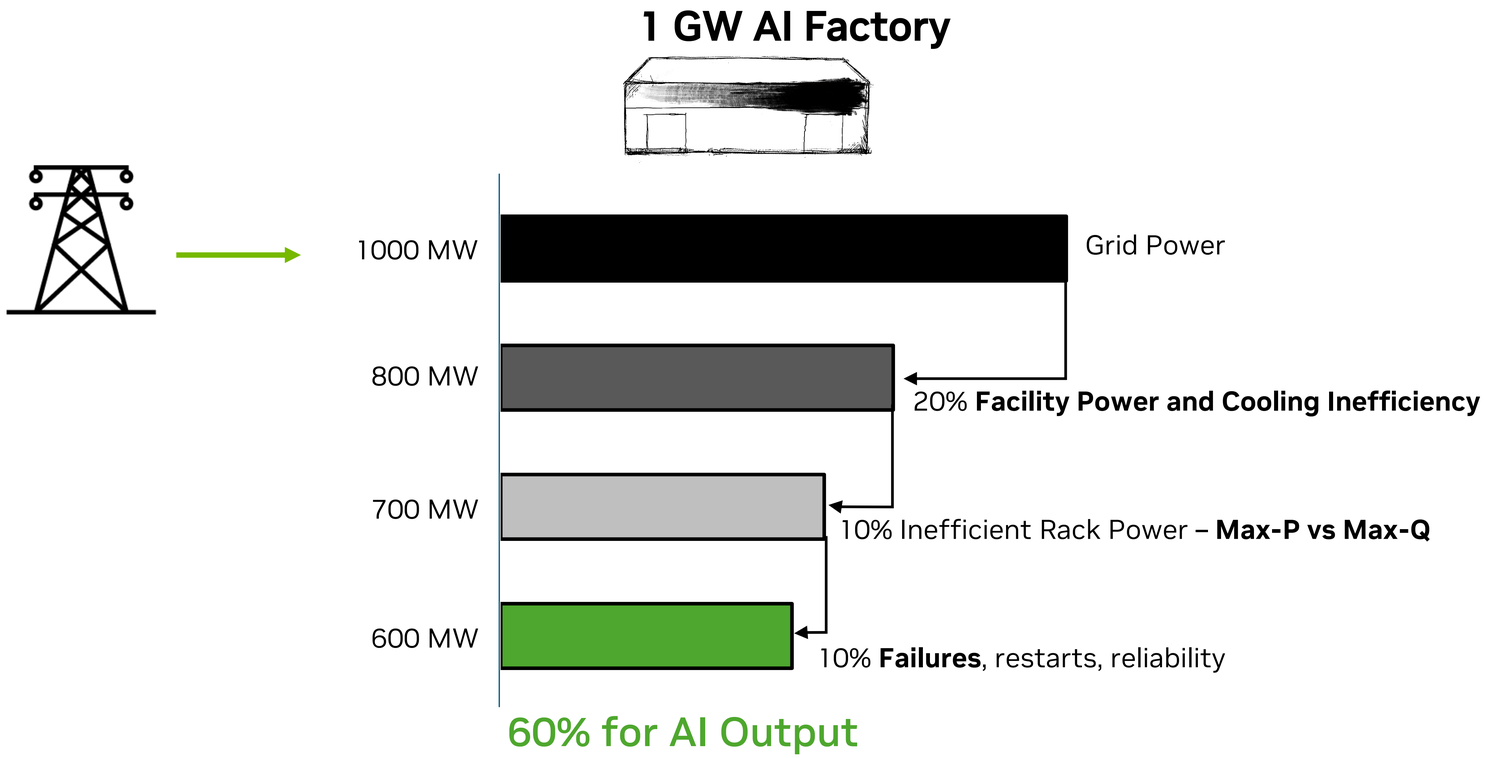
## Full-stack | One architecture | CUDA everywhere



# Vera Rubin DSX Reference Design & Omniverse DSX Blueprint

# AI インフラの電力効率における課題

設備、ラック、ダウンタイムの非効率性により、グリッド電力の一部しかコンピューティングに届かない




- Stranded Power
- Fixed and Static, Conservative Power / Cooling Configurations
- Overprovisioned for Failures and Extremes

# NVIDIA Vera Rubin DSX with Omniverse DSX Blueprint

パートナー各社との共同設計による耐障害性と最大トークンスループット | AI コンピューティング能力が30%向上

Vera Rubin DSX
Hardware and Software
Design Guides

Developer Libraries,
APIs, Software

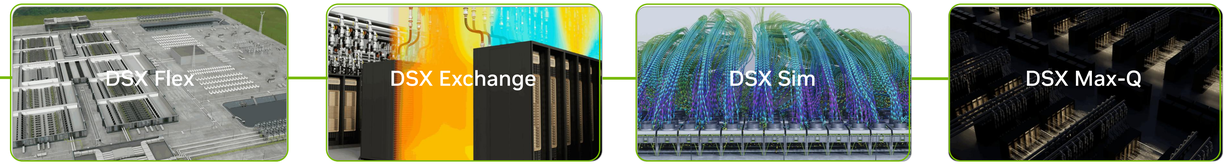
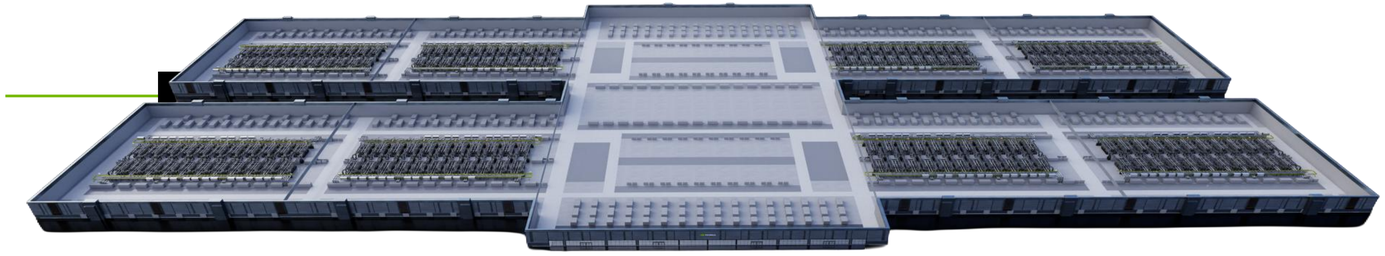



ABB BECHTEL cadence CATERPILLAR DASSAULT SYSTEMES EATON emeraldai etap GE VERNOVA HITACHI Jacobs

ptc paloalto phaidra Schneider Electric SIEMENS SIEMENS energy TRANE TECHNOLOGIES VAST VERTIV WEKA

Omniverse DSX
Digital Twin Blueprint



NVIDIA

# NVIDIA DSX

AI ファクトリーリファレンス設計 | 全 5 層にわたる最適化

| **5 Layer Cake of AI Factories** | **DSX Data Center Reference Design:** Modular codesign from rack to factory scale Facilities, Hardware and Software for fastest time to production token, maximizing token per megawatt |
|---|---|
| Applications | Optimized by DSX for all Third-Party Applications |
| AI Models | Optimized by DSX for all Third-Party Models |
| Infrastructure | Software libraries, Ecosystem integration, Power, Cooling, Controls systems |
| Chips & Compute | Compute, Network, Storage - Vera-Rubin and SPX6 |
| Energy | Connectivity to the Grid and onsite Power generation |

# NVIDIA DSX Air

AI ファクトリー設計・検証プラットフォーム

## Deploying AI Factories in Days Instead of Weeks

*from*

**6 Months to 1 Week**

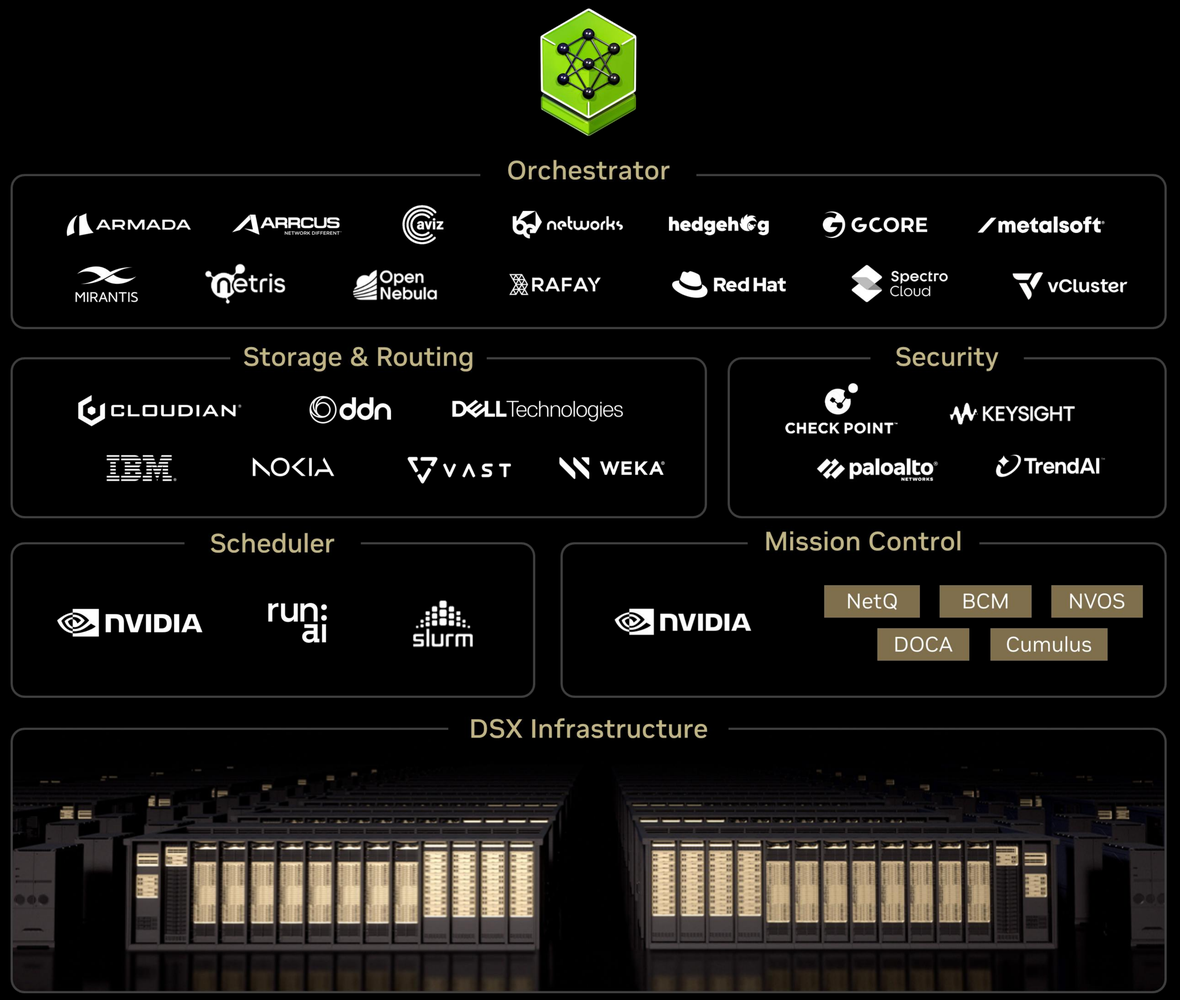
**Infrastructure Validation**

**3 Months to 1 Week**

**Software Bring-Up**

**2 Weeks to 1 Day**

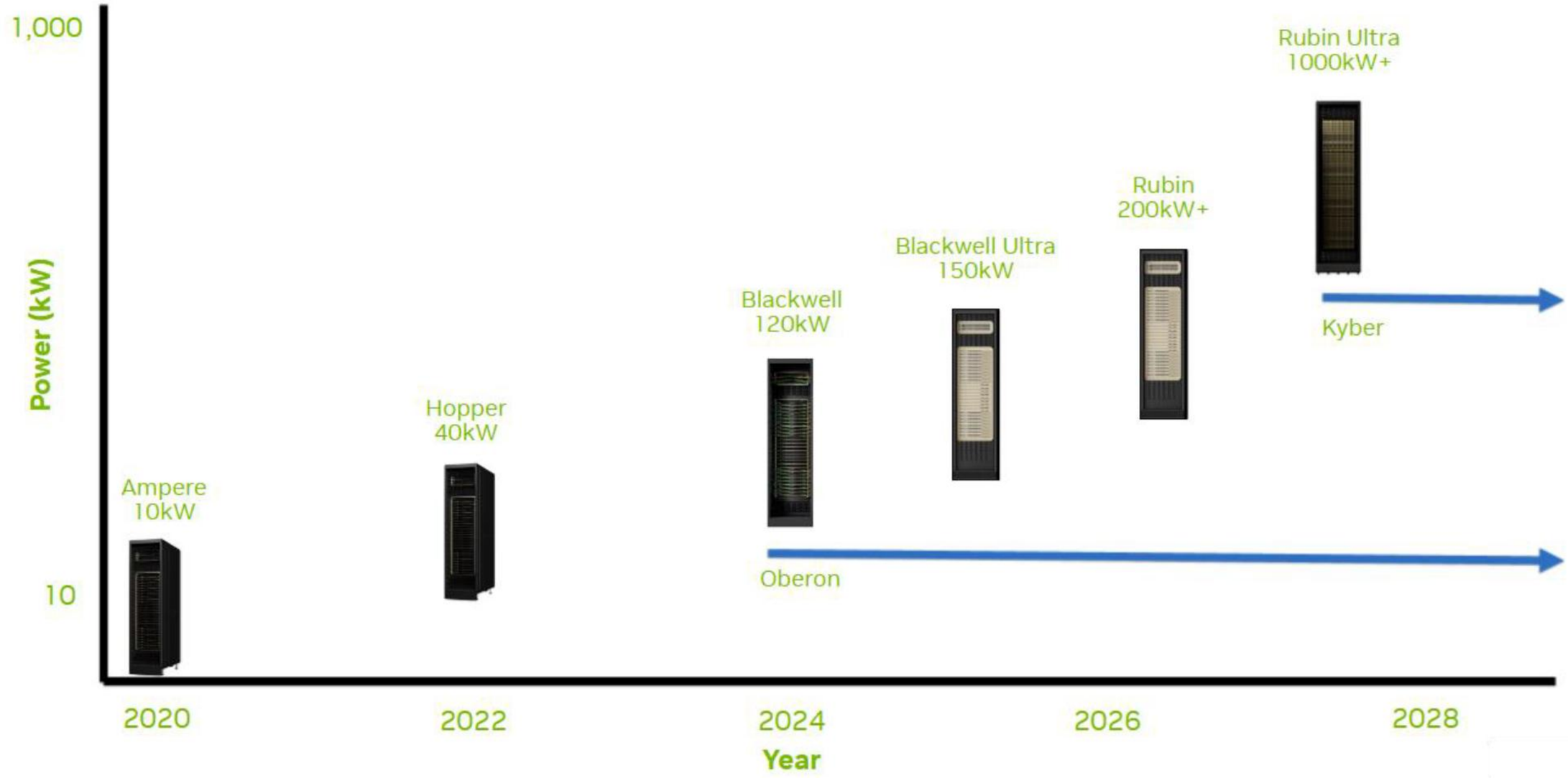
**Deployment Time**



# 800 VDC Architecture for Next-Generation AI Infrastructure

# 世代ごとに増大する消費電力

新しい GPU アーキテクチャではワットあたりの性能が常に向上するが、絶対的な消費電力も増大



# 800 VDC Architecture for Next-Generation AI Infrastructure

Jared Huntington & Mike Tu

**The Architectural Imperative of 800 VDC and Integrated Energy Storage**

## 800 VDC MGX Architecture

The above requirements for exponentially increasing power, pushing power out of the NVLink domain, and energy storage for mitigating load swings all come together to define the requirements for the 800 VDC power architecture.

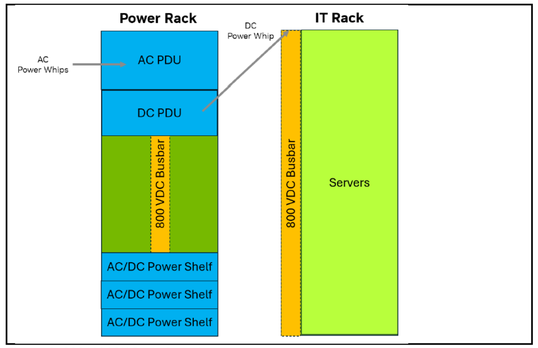



*Figure 8: 800 V MGX Rack Overview*

The overall rack architecture takes 800 VDC directly into the rack and distributes 800 VDC direct to the compute node.

Safety is a critical design feature for the 800V system. Because of this touch-safe connectors are used in all locations that are user accessible. To ensure safety while connecting and disconnecting loads mechanical interlocks are also used in all connectors to ensure that there is no disconnect occurring under load. This helps to minimize the risk of arcing as well as act as a secondary protection ensuring that there are no live accessible parts. These techniques are commonly used in EV chargers which are widely available and used by untrained consumers. The presently planned connectors are leveraging this existing technology and optimizing for a datacenter environment.

To minimize connector sizes and volume of copper, the DC/DC conversion is pushed as close as possible to the GPU. This takes advantage of the 800 VDC to minimize power overhead in the system. The conversion is then done from 800 VDC to 12 VDC and routed a short distance to the voltage regulators.



Each compute rack is supported by local or aggregated energy storage to provide slew-rate control, power smoothing, voltage regulation, and optional backup capability. To maintain concurrent serviceability and uptime, racks are configured with 1+1 redundant feeds sourced from independent rectifiers.

In the event of a single rectifier failure (e.g., loss of Source #2), the system automatically transfers load to an alternate source, ensuring uninterrupted operation. Even under worst-case loading, a single rectifier (e.g., Source #1) can sustain up to 3.3 MW of compute load. This demonstrates full-load survivability and readiness for high-density, mission-critical AI workloads.

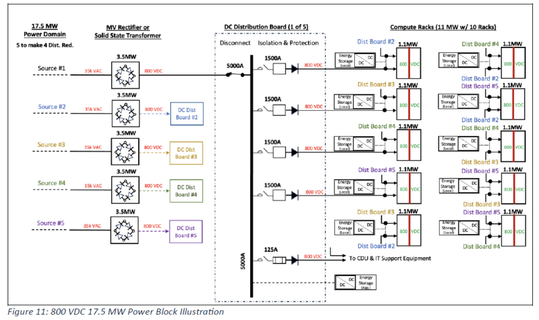



*Figure 11: 800 VDC 17.5 MW Power Block Illustration*

## Industry Collaboration and Path Forward

To realize this architecture requires multiple points of collaboration with the industry moving forward.

*Align on common voltage ranges, connectors, and current levels to maintain modularity*

Establishing standardized voltage levels, current ratings, and connector interfaces is essential to enabling modular and interoperable solutions across vendors and system integrators. For AI Factories adopting 800 VDC architectures, industry-wide alignment will minimize customization, accelerate time-to-deploy, and support scalability. A consistent framework allows for flexible deployment strategies while reducing complexity in design, installation, and maintenance.

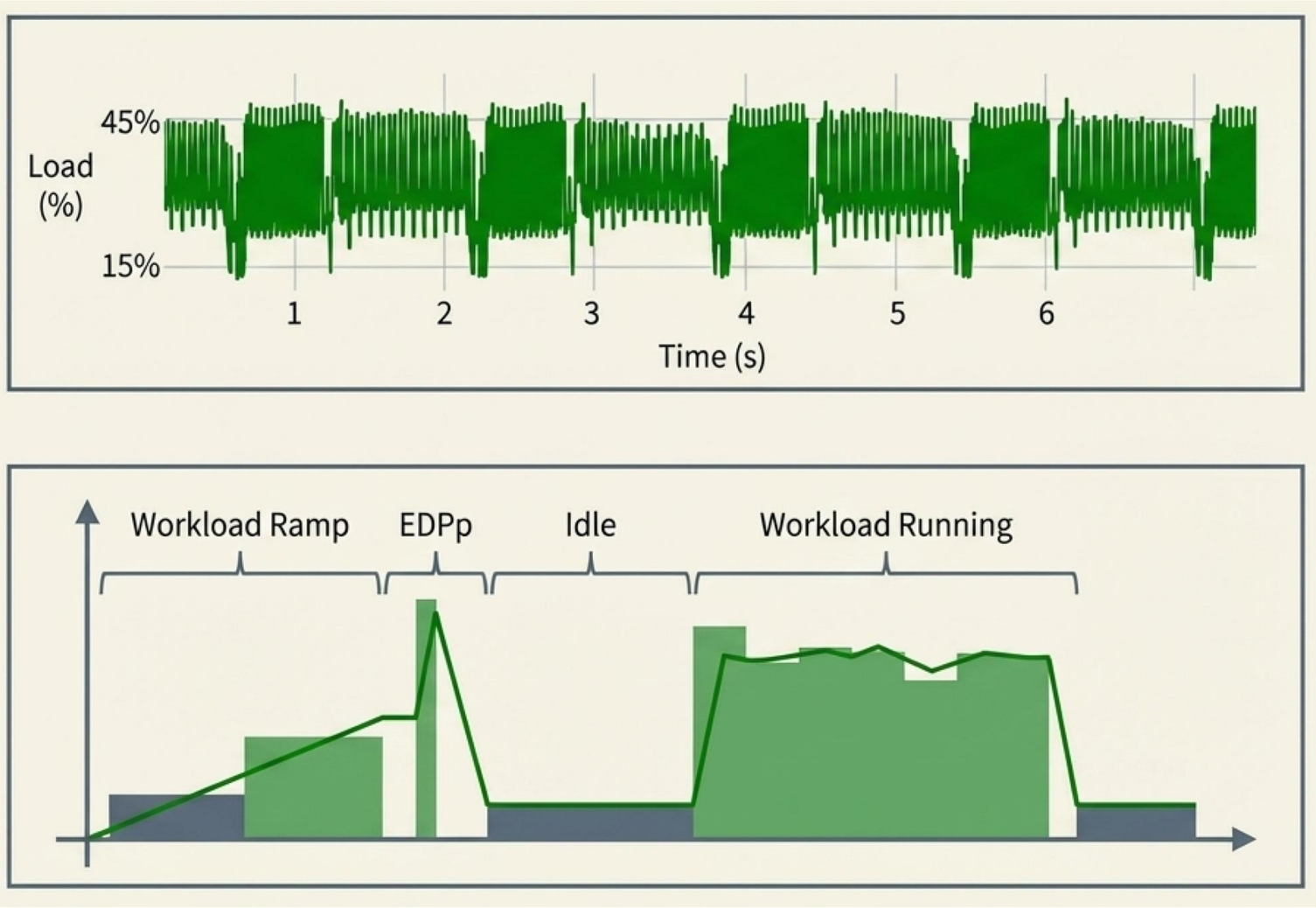
*Develop and certify DC-native equipment*

The transition to facility-level DC distribution requires a new generation of DC-native equipment including rectifier, distribution system, protective devices, cabling, connectors, transfer trip schemes and



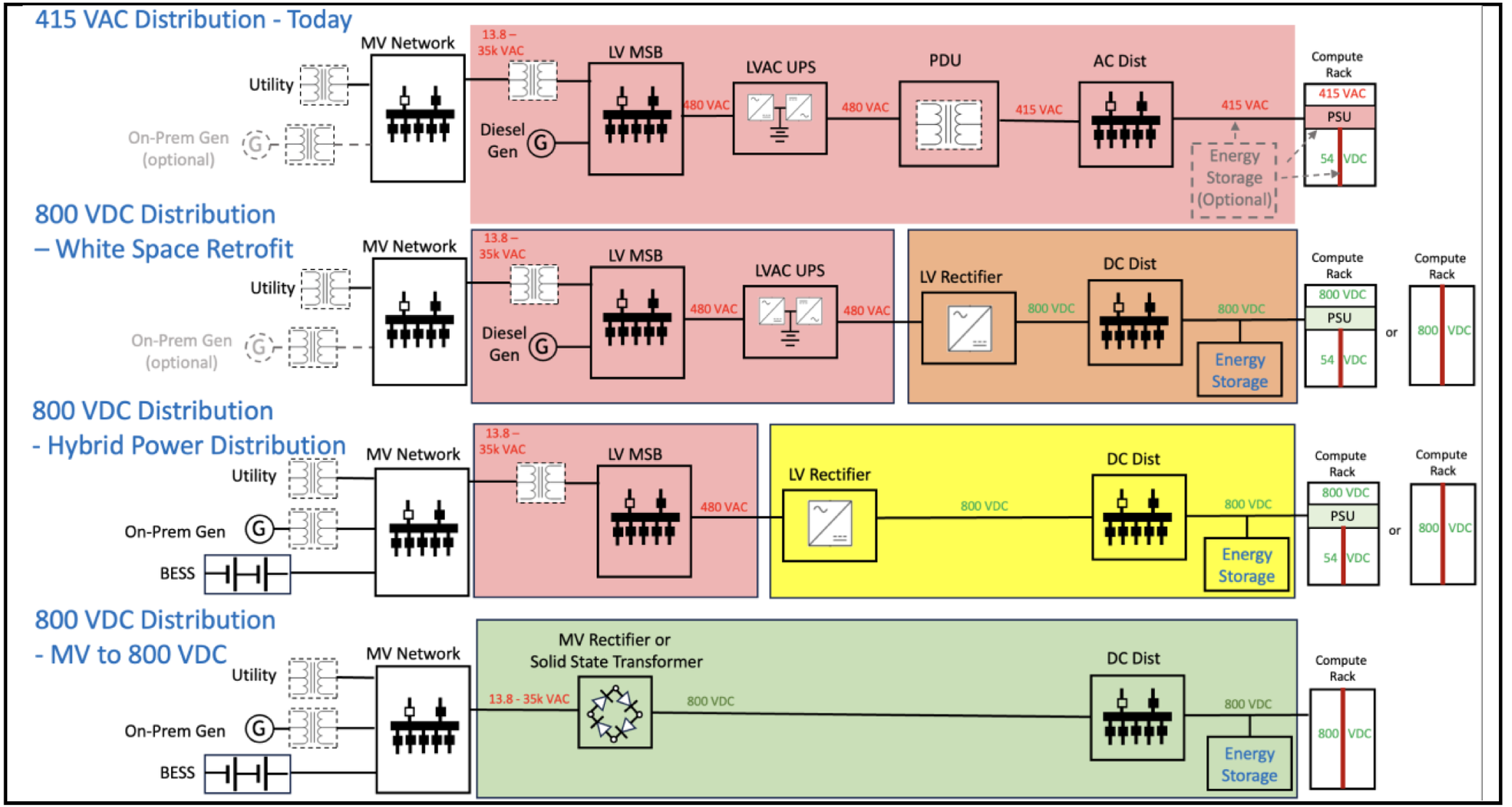
https://www.nvidia.com/en-us/data-center/technologies/800-vdc-architecture/

# 負荷変動への対応



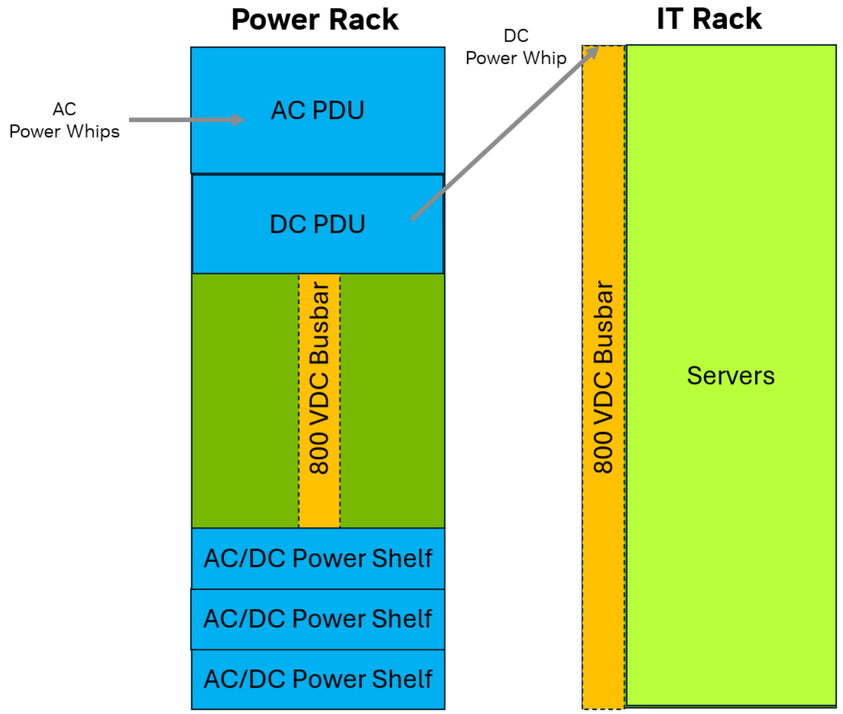

- LLM ワークロードの特性
  - 特に大規模な推論ワークロードでは、ラック電力の短時間での変動が大きい
- 負荷変動への対処
  - ソフトウェアによるアイドル期間の最適化
  - 蓄電の活用
  - 意図的な電力消費 (Power Smoothing 機能)
  - GPU 性能のスロットリング
  - 以上のように複数の手段を総合して対処

# 800VDC 直接給電への段階的な進化




# GTC 2026 での 800VDC 動作デモ

- 左のラックが 800 VDC サイド パワー ラック
  - 負荷変動吸収用に蓄電池内蔵
- 右のラックでは DGX B300 (14.3kW) が4台稼働中
  - DGX B300 自体は既存の通常モデルなので、ラック内のパワーシェルフで 800V -> 54V に降圧して供給




LITEON
In-Row 800 VDC Power
Analog Devices
Delta
Infineon
800V Hot
Monolithic Power Syste
Renesas
STMicroelectronics /
Innoscien


LITEON
Vera Rubin 110kW Powershelf
LITEON
GB300 19in 33kW Powershelf
LITEON
GB200 19in 33kW Powershelf
Vera Rub

NVIDIA

# GTC 2026 お勧めセッション

# GTC’26

Suggested talks

- **GTC 2026 Keynote [S81595]**
  - **Jensen Huang** | Founder and CEO | NVIDIA
- **Inside the NVIDIA Inference Platform and Ecosystem [S81911]**
  - **Ian Buck** | VP Hyperscale and HPC Computing | NVIDIA
- **Foundation Models Revolutionize Molecular Dynamics Simulations for Chemistry and Materials Science [S81802]**
  - **Gabor Csanyi** | Professor | University of Cambridge
- **Cutting-Edge Molecular Dynamics on the Latest Multi-Node NVLink Technology [S81542]**
  - **Alan Gray** | Principal Developer Technology Engineer | NVIDIA
  - **Mahesh Doijade** | Sr. Compute Developer Technology Engineer | NVIDIA
- **Ozaki Scheme: Addressing the Accuracy Challenge in the Era of Low-Precision Accelerators** [S82285]
  - **Katsuhisa Ozaki** | Professor, Shibaura Institute of Technology

# AI スーパーコンピューターを構成する 5 つのラックスケールシステム

## Unified MGX architecture & ecosystem

**NVIDIA MGX NVL**

**NVIDIA MGX ETL**

Fully Configurable up to 256 chips

### NVIDIA Vera Rubin NVL72

NVLink spine

### NVIDIA Groq 3 LPX

Direct Chip-to-Chip spine

### NVIDIA Vera CPU

Spectrum-X Ethernet spine

### NVIDIA BlueField-4 STX Storage

Spectrum-X Ethernet spine

### NVIDIA Spectrum-6 SPX

# NVIDIA コンピューティング プラットフォーム ロードマップ