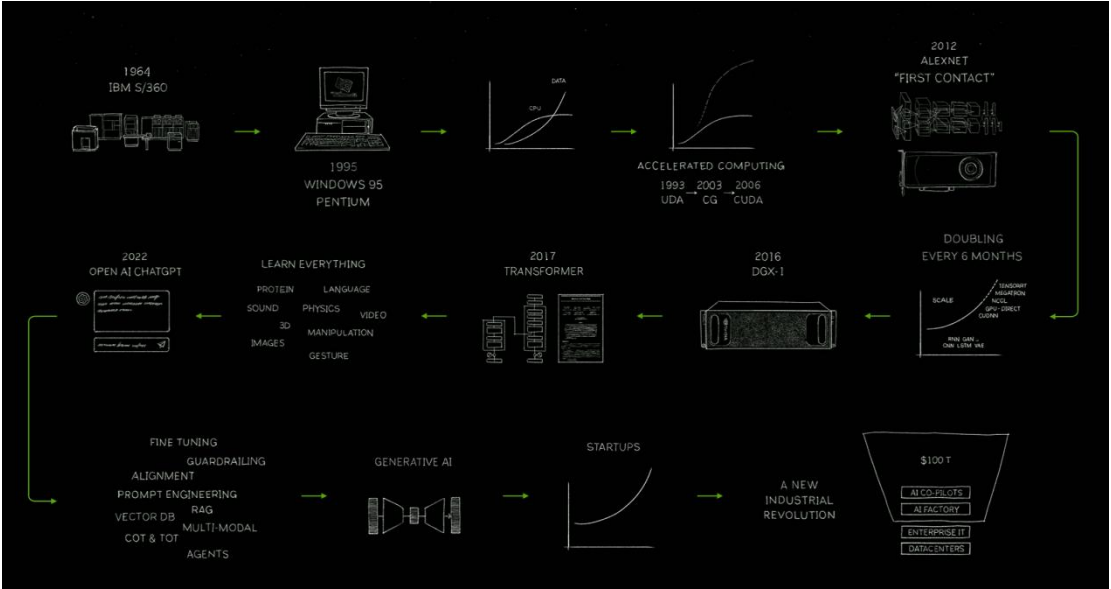


会议时间：2024.3.19（北京时间）
回放链接：NVIDIA CEO 黄仁勋主题演讲 | GTC 2024 | NVIDIA

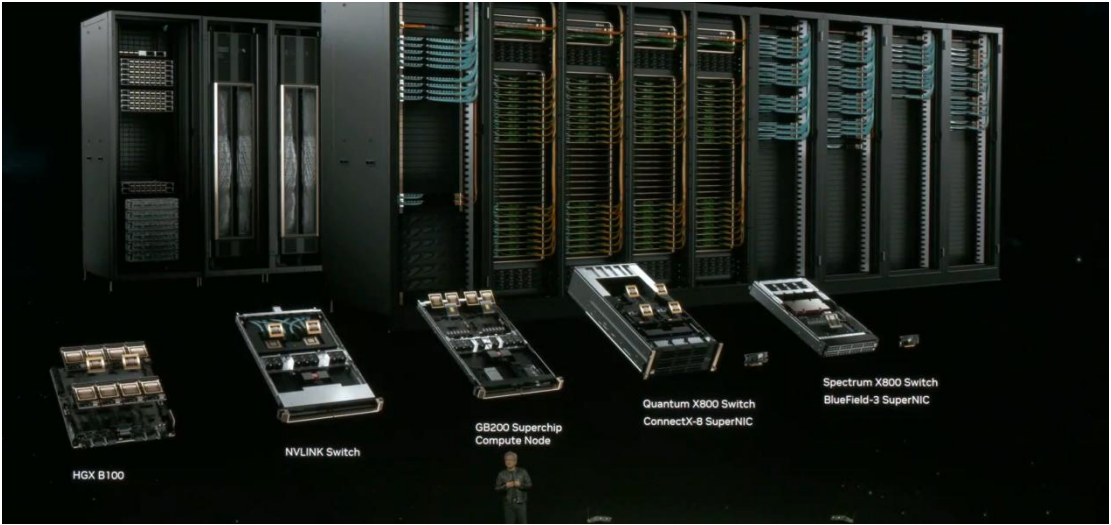
注：以下材料仅为公开资料整理，不涉及分析师的研究观点和投资建议，记录和翻译可能产生误差，仅供参考，如有异议，请联系删除。

Nvidia、CUDA 发展历程



Blackwell 平台（详见云笔记最下方的附图）

- Blackwell 小芯片：1040 亿个晶体管，TSM N4P
- B100 GPU：两个 Blackwell 小芯片组成一个 B100
- GB200 超级芯片：两个 B100 GPU+一个 Grace CPU
- GB200 计算单元：两个 GB200 超级芯片
- GB200 NVL72:36 个 Grace CPU+72 个 B100 GPU
- GB200 NVL 72 机架：8 个 GB200 NVL72



提到的合作伙伴

Synopsys: 台积电使用的计算光刻项目（GTC 2023 曾提过），台积电宣布英伟达 CULITHO 投入使用

Cadence: 芯片 EDA、Cadence copilot

AWS: 正在构建 222ExaFLOPS 的系统，并且正在合作通过 CUDA 加速 SageMaker AI、Bedrock AI，双方在 Omniverse、Isaac Sim 上展开合作，AWS Health 已经集成了 Nvidia Health;

Google: GCP 已经拥有 A100、H100、T4、L4 等一系列 CUDA GPU，并在上面部署 Gemma;

甲骨文: 加速 Oracle 数据库

微软: 打造最大的英伟达 infiniband 超级计算机

纬创: 打造 Omniverse 方案

其他: servicenow、snowflake、西门子、dell、比亚迪、

AI 场景:

Earth-2: 天气模型

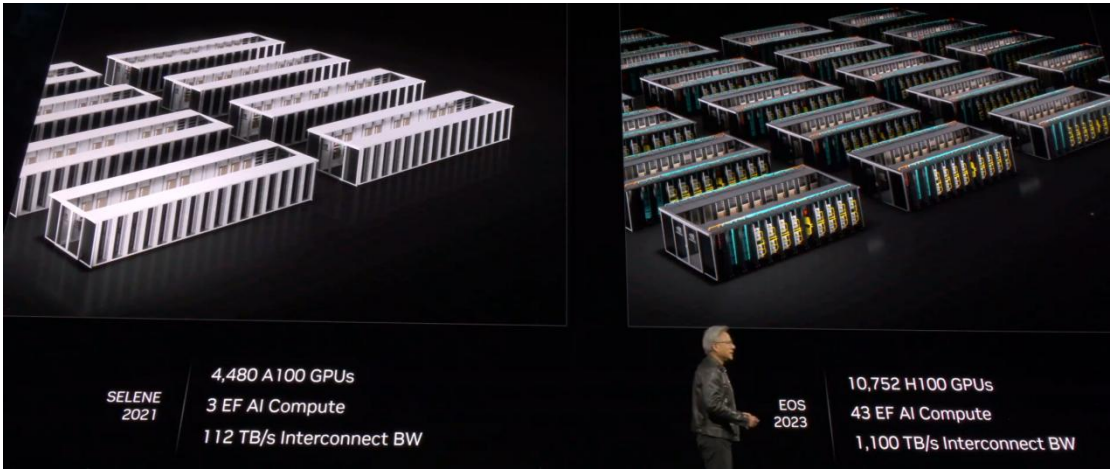
ChipNeMo: 基于 Llama2 的芯片设计聊天机器人

VR: Vision Pro

机器人

AI 基建

HGX 集群发展



B100

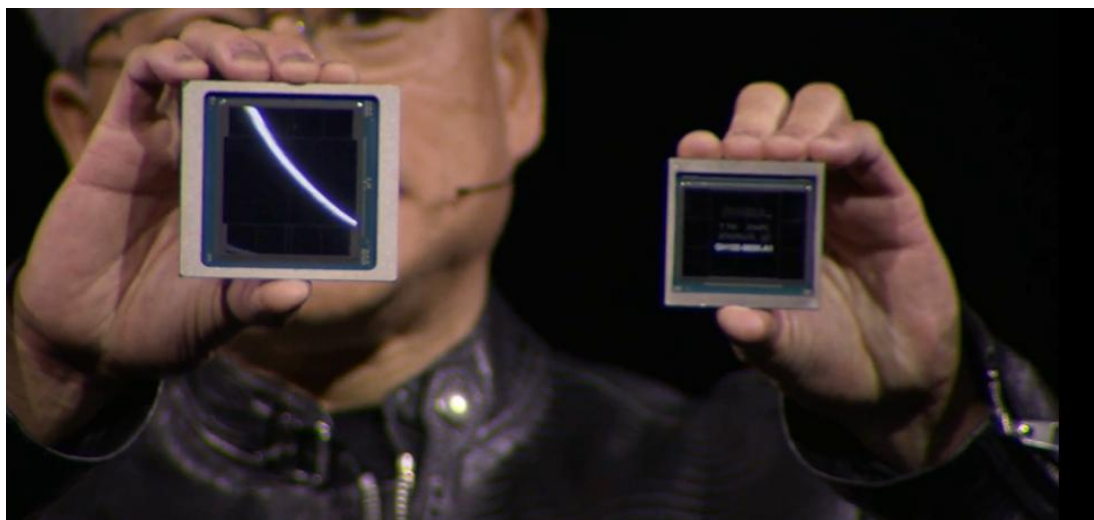
2080 个晶体管

采用 TSMC N4P 制程

194

通过两个芯片集成，C2C 传输速度达到 10TB/S

Hopper（右边）、Blackwell（左边）

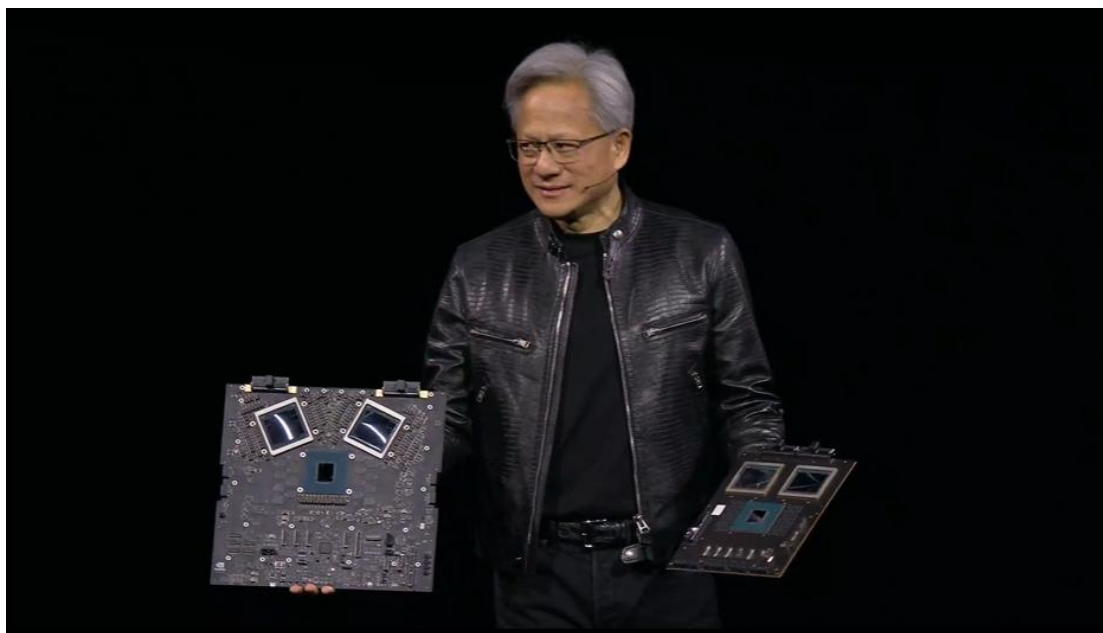


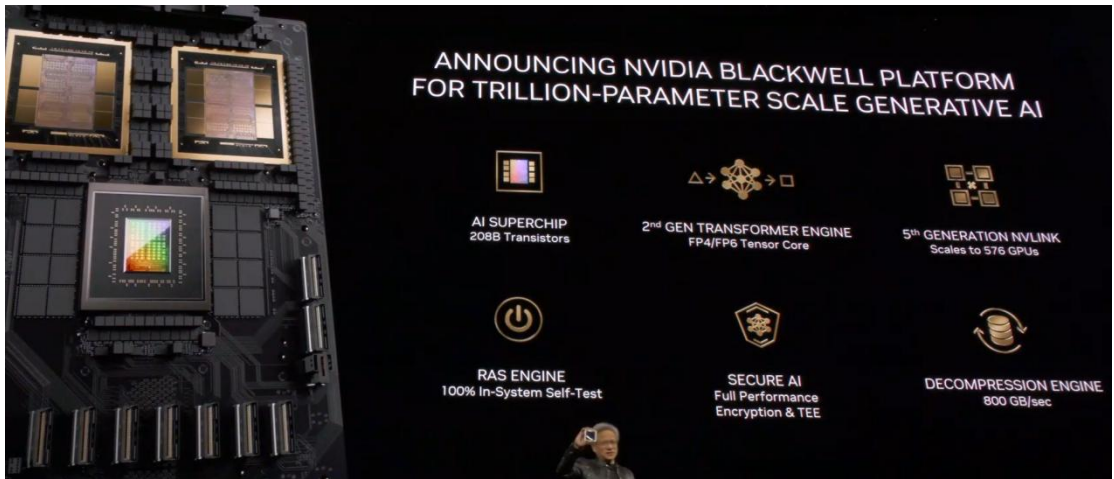
GB 系统及升级点

GB200: 通过 900GB/s 的 NVLink C2C 互联，将两个 B100 连接到 Grace CPU

升级点: B100 C2C 方案、二代 Transformer、五代 NVLINK、RAS 引擎、Secure AI、解压缩引擎

测试样品（左边）、最终成品（右边）





第二代 transformer 引擎

FP4/FP6 tensor core（Hopper、Ada Lovelace 为 FP8）

通过新的 FP4 引擎，以及动态管理算法，实现训练、推理性能的提升（吞吐量提升）

网络：

第五代 NVLink

为每个 GPU 提供 1.8TB/s 的双向吞吐量

RAS 引擎

用于检测故障芯片，及时在集群中找到效率低的 GPU

最大限度延长系统正常运行时间，并提高 AI 系统的弹性，使系统可以连续运行数月，减少运维成本

Secure AI

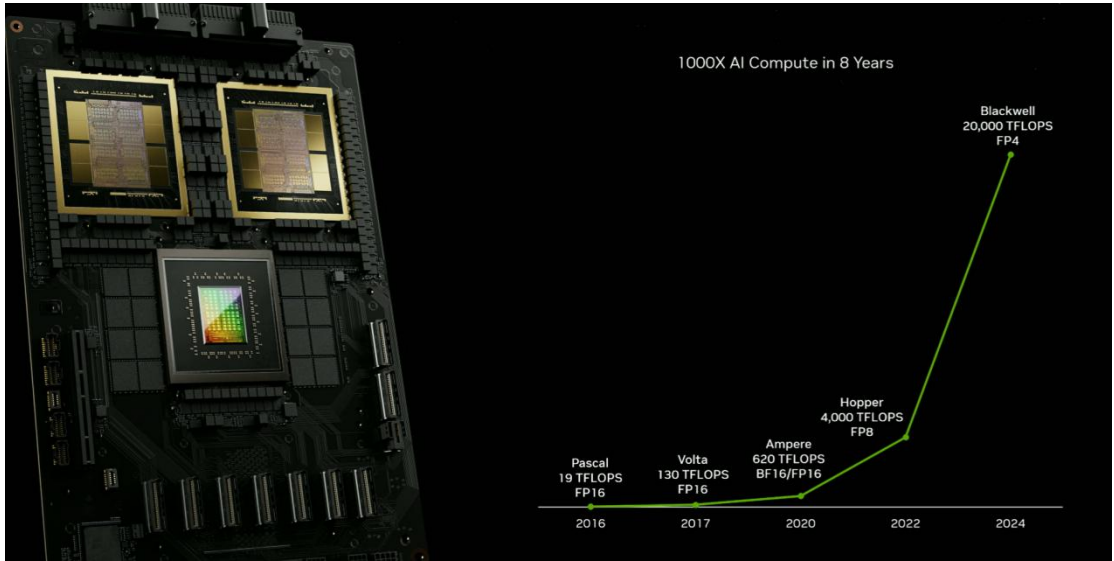
保障模型安全性、保护客户数据

Decompression Engine（解压缩引擎）

通过专用解压缩引擎支持最新的数据库格式，并推动更多的数据通过 GPU 加速计算

Blackwell GPU		
FP8	20 PFLOPS	2.5X Hopper
NEW FP6	20 PFLOPS	2.5X
NEW FP4	40 PFLOPS	5X
HBM Model Size	740B param	6X
HBM Bandwidth	34T param/sec	5X
NVLINK All-Reduce with SHARP	7.2 TB/s	4X

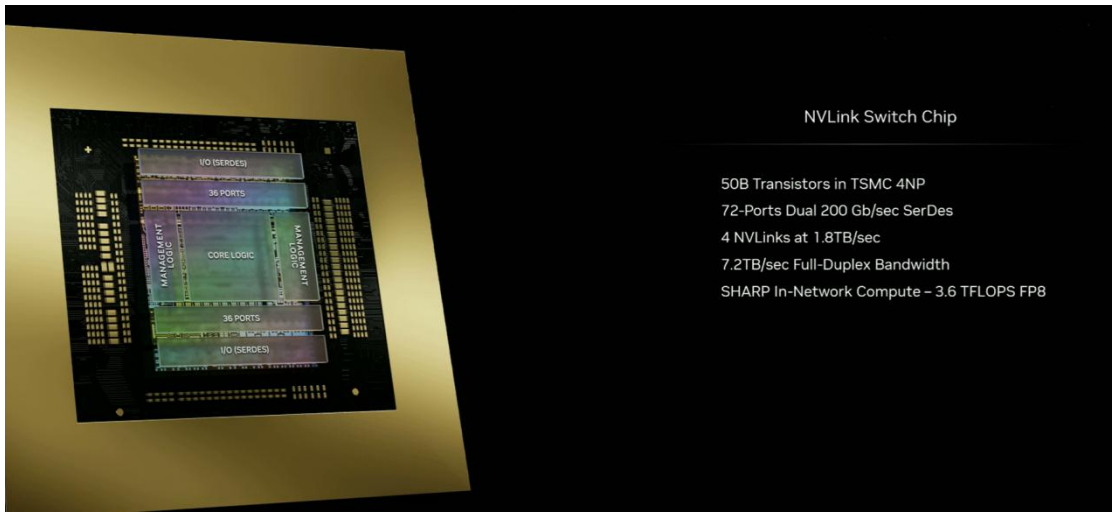
AI 芯片发展路径、支持精度（FP 规格）



网络产品

NVLink Switch Chip

500 亿晶体管，采用台积电 N4P 制程
有 4 个 NVLink，分别提供 1.8TB/s 的速度
可以实现 GPU 同时与其他 GPU 实现全速通信



Nvidia Quantum-X800 Infiniband、Spectrum-X800 以太网网络平台

可为每个 GPU 提供高达 1800GB 的带宽
Quantum-X800 包括 Quantum Q3400 交换机和 ConnectX-8 SuperNIC，与上一代相比，带宽容量提高 5 倍，网络内计算速度提高 9 倍，采用英伟达 SHARPV4 协议，可达到 14.4TFlops；
Spectrum X800 包括 Spectrum SN5600 800GB/s 交换机和 BlueField 3 SuperNIC

服务器产品

DGX GB200 NVL72

结合 36 个 Grace Blackwell 超级芯片，对应 72 个 GPU、36 个 CPU，通过 5 代 NVLink 互联

使用 BlueField-3 DPU；

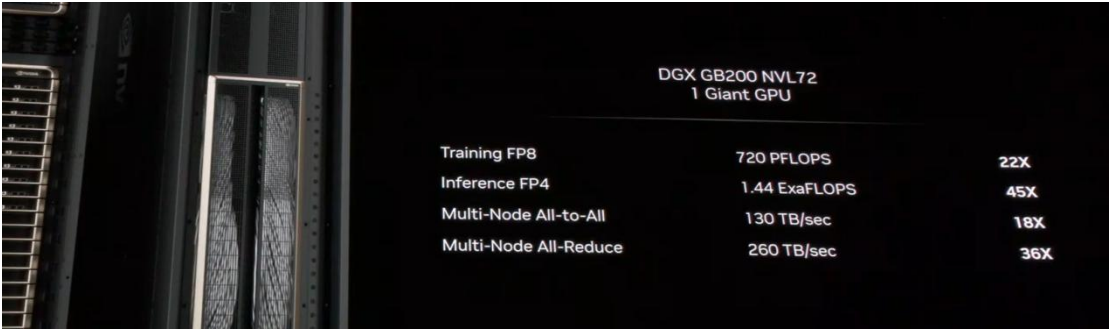
使用液冷；

与 H100 方案相比，GB200 NVL72 对 LLM 推理工作负载的性能提高 30 倍，成本和能耗降低幅度达 25 倍

FP8 训练算力 720PFLOPS，提升 22 倍；

FP4 推理算力 1.44 ExaFLOPS，提升 45 倍；

通过 NVLink Switch 实现计算能耗的降低：训练一个 1.8 万亿的 GPT 模型，使用 Hopper，需要 8000 个 GPU，消耗 15 兆瓦，耗时 3 个月，使用 Blackwell，同样 3 个月，只需要 2000 个 GPU，消耗 4 兆瓦；



DGX B200 系统：

包括 8 个 Blackwell GPU 和 2 个五代 Intel XEON 处理器，

提供 FP8 精度，AI 性能可达 144 PetaFLOPS，实现 1.4TBGPU 内存、64TB/s 内存带宽，与上一代相比，万亿参数模型推理速度提高 15 倍

网络方面，配置 8 个英伟达 connectX-7 NIC 和 2 个 BlueField-3 DPU，每个连接点可提供 400GB/s 的带宽，并通过 Quantum-2 infiniband 或 Spectrum-X 加速。

HGX B200 服务器主板：

通过 NVLink 连接 8 个 B200 GPU，以支持基于 X86 的 Gen AI 平台
可选择 Quantum-2 infiniband 或 Spectrum X 以太网平台，达到 400GB/s 的网络速度

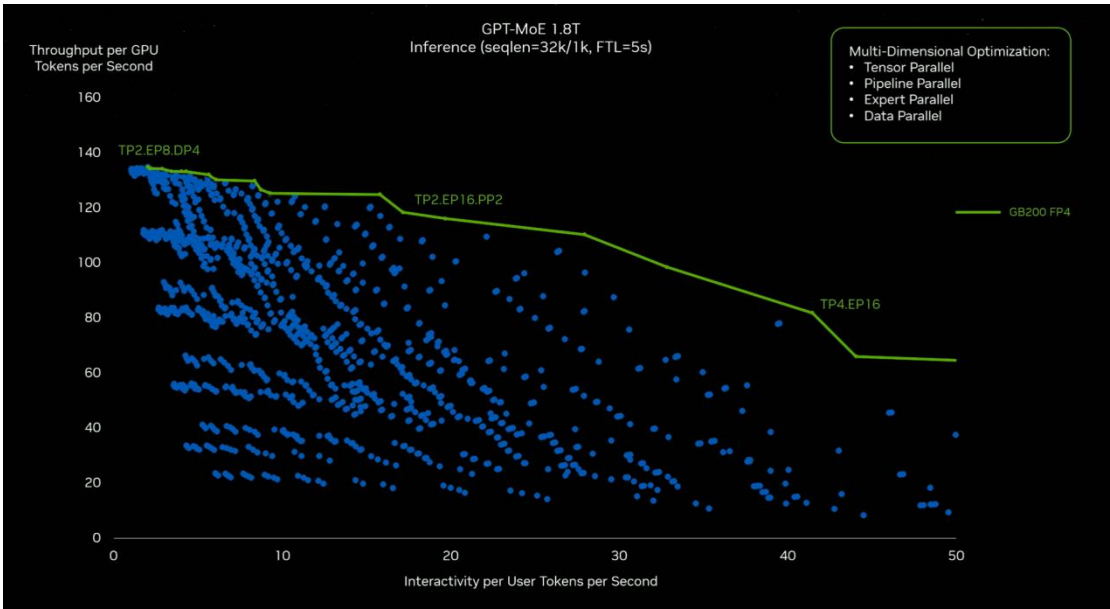
Blackwell 平台推理性能

大语言模型的推理复杂度：（1）规模大，不适合单个 GPU；（2）每个聊天机器人数量亿万个参数、token，以及大量用户访问，需要 GPU 集群才能实现目标；（3）需要高吞吐量来降低成本、提高服务质量；

下图：

y 轴代表每秒的 token 数，x 轴代表每秒用户的 token 交互数
越往右上角越好

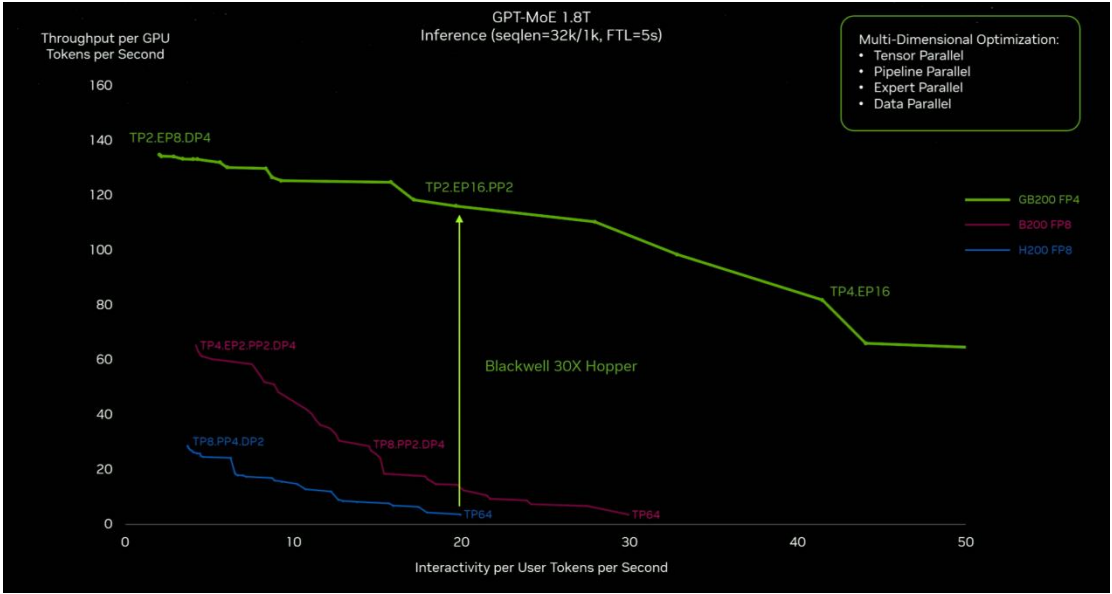
英伟达通过不断优化找到每个 xy 对应的最优方案，这离不开英伟达 GPU 可编程性和 CUDA 的生态系统，绿色代表最优曲线（TP 代表 tensor 并行、EP 代表 Expert 并行、DP 代表 Data 并行）



下图

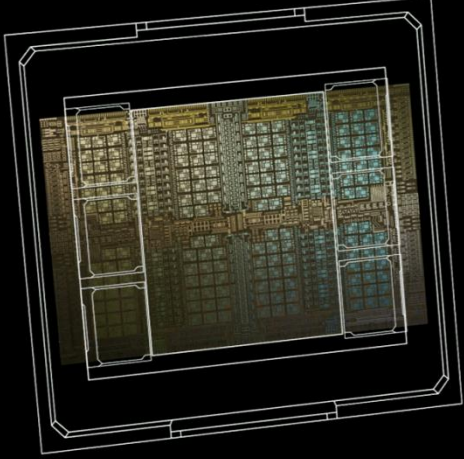
Blackwell 的推理性能是 Hopper 的 30 倍（蓝色是 H100）

本质上 hopper 架构没有改变，公司只是将其变成更大的芯片，并将两个芯片集成在一起，叠加 NVLink 交换机对网络的提升，Blackwell 成为了非常高效的 Gen AI 系统



其他图

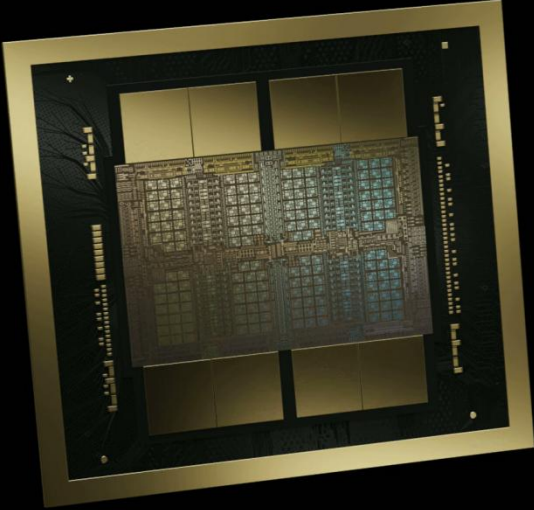




BLACKWELL VERSUS HOPPER

TWICE THE SIZE, A MASSIVE LEAP IN COMPUTE

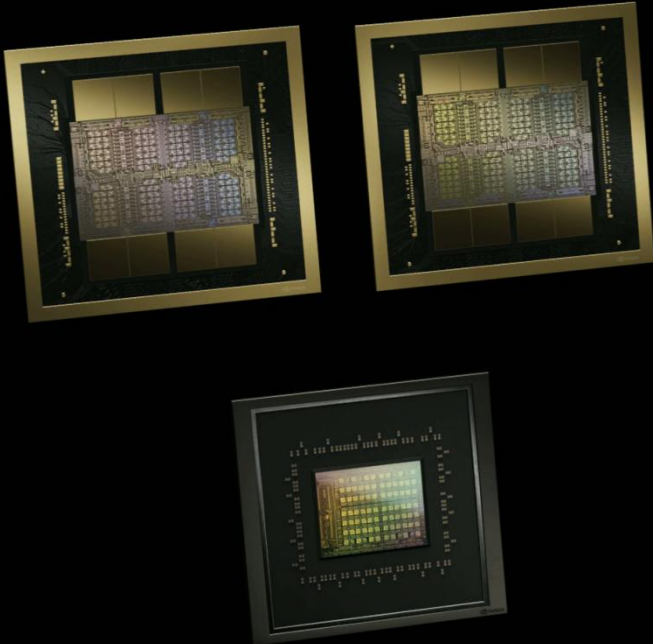
- 128 billion more transistors
- 5X the AI performance
- 4X the on-die memory



BLACKWELL

THE ENGINE OF THE NEW INDUSTRIAL REVOLUTION

- 20 petaFLOPS of AI performance
- 192GB of HBM3e
- 8TB/s of memory bandwidth
- Full stack, CUDA enabled

Three semiconductor dies are shown against a dark background. Two larger dies at the top are Blackwell GPUs, featuring a complex grid of circuitry and a central square area. A smaller die at the bottom is a Grace CPU, showing a different layout with a central square and surrounding circuitry.

TWO BLACKWELL GPUs AND ONE GRACE CPU

BUILDING BLOCKS OF THE GB200 SUPERCHIP

384GB of HBM3e

72 Arm Neoverse V2 CPU cores

900GB/s of NVLink-C2C bandwidth

A perspective view of the GB200 Grace Blackwell Superchip. It is a large, rectangular integrated circuit with a complex internal structure, including several square dies and a dense network of interconnects. The chip is mounted on a dark substrate.

GB200 GRACE BLACKWELL SUPERCHIP

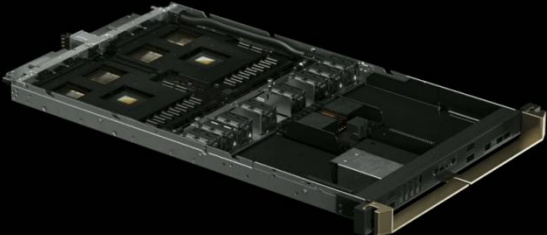
PROCESSOR FOR TRILLION-PARAMETER-SCALE GENERATIVE AI

40 petaFLOPS of AI performance

864GB of fast memory

16TB/s of HBM

3.6TB/s of NVLink bandwidth

A perspective view of the Blackwell Compute Node. It is a large, rectangular circuit board with a complex internal structure, including several square dies and a dense network of interconnects. The board is mounted on a dark substrate.

BLACKWELL COMPUTE NODE

MOST POWERFUL COMPUTE NODE EVER CREATED

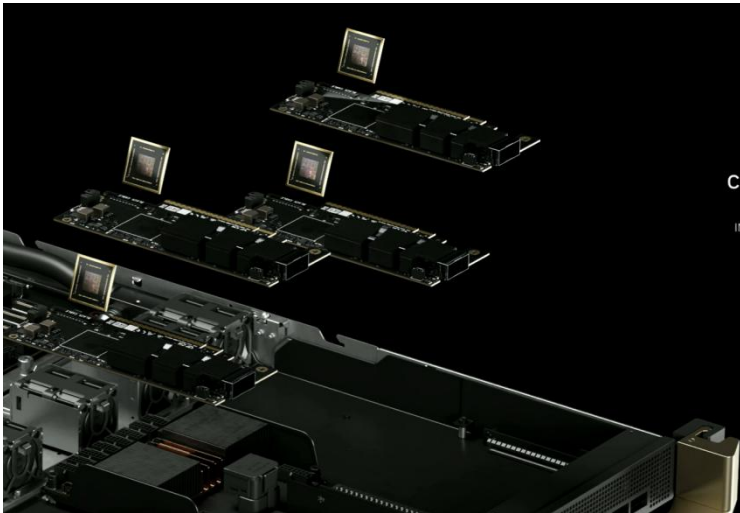
Two Grace CPUs and four Blackwell GPUs

80 petaFLOPS of AI performance

1.7TB of HBM3e

32TB/s of memory bandwidth


Liquid-cooled MGX design



CONNECTX-800G INFINIBAND SUPERNIC

INDUSTRY'S MOST ADVANCED GPU RDMA, ADAPTIVE ROUTING,
AND PROGRAMMABLE CONGESTION CONTROL

Optimized for AI processing



BLUEFIELD-3 DPU

POWERFUL INFRASTRUCTURE PROCESSOR

Line-speed processing of networking,
storage, and cybersecurity

In-Network Computing

80GB/s of memory bandwidth

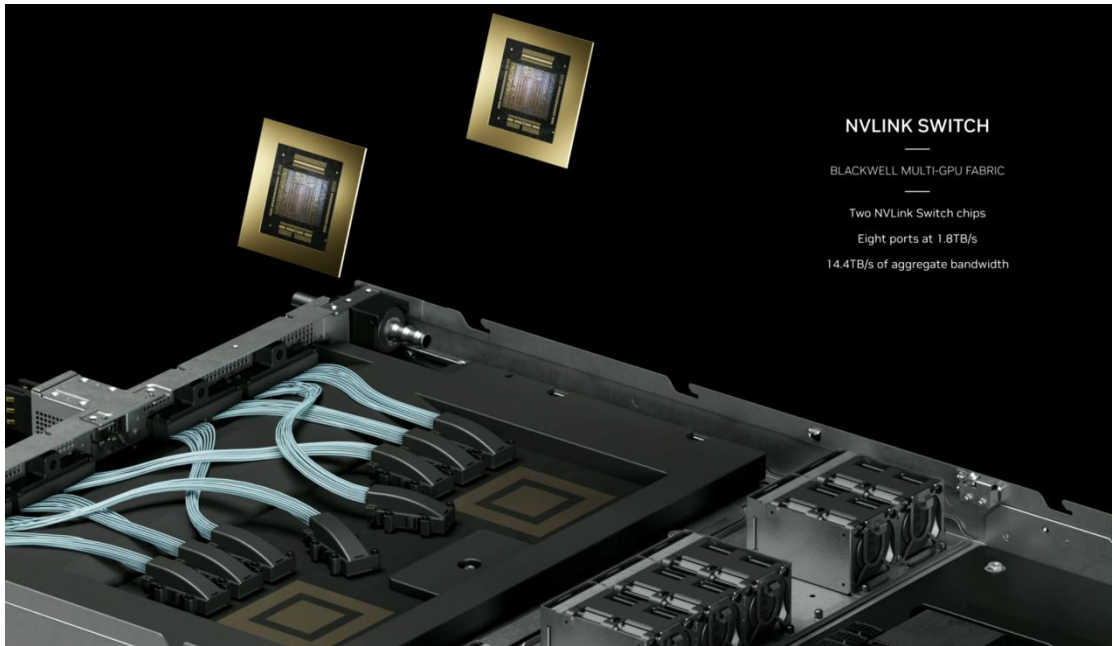


GB200 NVL72

COMPUTE FOR THE MODERN DATA CENTER

18 compute trays in a rack

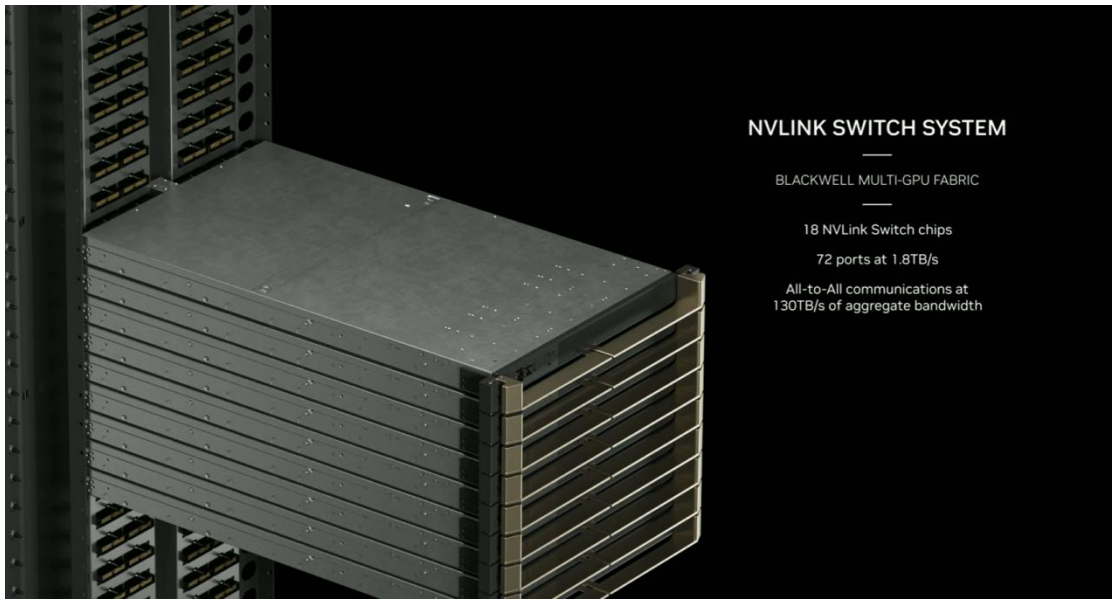
36 Grace CPUs and 72 Blackwell GPUs



NVLINK SWITCH

BLACKWELL MULTI-GPU FABRIC


- Two NVLink Switch chips
- Eight ports at 1.8TB/s
- 14.4TB/s of aggregate bandwidth



NVLINK SWITCH SYSTEM

BLACKWELL MULTI-GPU FABRIC

- 18 NVLink Switch chips
- 72 ports at 1.8TB/s
- All-to-All communications at 130TB/s of aggregate bandwidth



GB200 COMPUTE NODES, NVLINK SWITCH AND SPINE

ONE MASSIVE GPU FOR GENERATIVE AI

72 Blackwell GPUs fully connected by NVLink

Copper cabling for 6X less cost

Blind-mate connectors for
easy installation and serviceability




GB200 NVL72

COMPUTE FOR TRILLION-PARAMETER-SCALE GENERATIVE AI

1.4 exaFLOPS of AI performance

30TB of HBM3e

One giant CUDA GPU

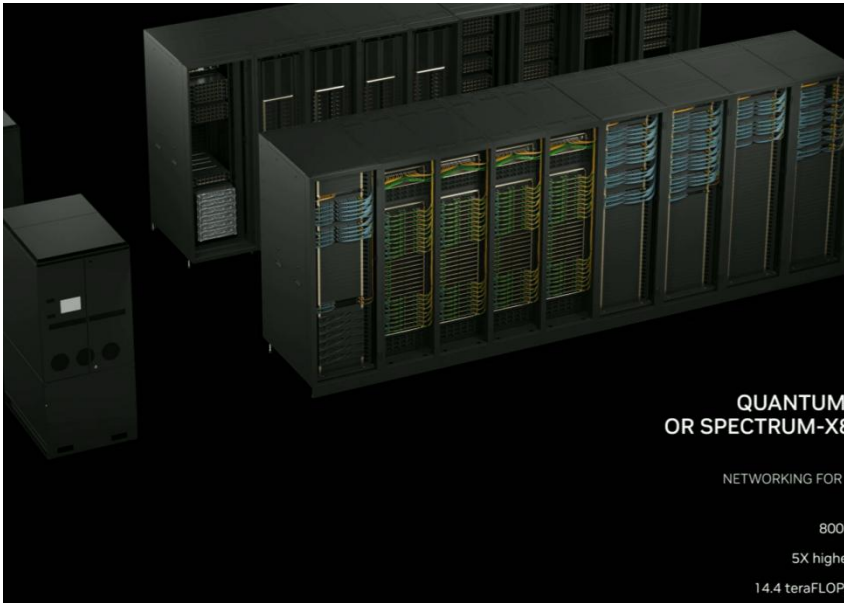


QUANTUM INFINIBAND SWITCH

UNPARALLELED DATA THROUGHPUT AND DENSITY

800Gb/s of throughput per port

230.4Tb/s of aggregated bidirectional throughput



**QUANTUM-X800 INFINIBAND
OR SPECTRUM-X800 ETHERNET SWITCHES**

—

NETWORKING FOR THE HIGHEST AI PERFORMANCE

—

- 800Gb/s connectivity
- 5X higher bandwidth capacity
- 14.4 teraFLOPS of In-Network Computing



GB200 NVL72 COMPUTE RACKS

—

LIQUID COOLED FOR BEST ENERGY EFFICIENCY

—

- Eight GB200 NVL72s
- 288 Grace CPUs and 576 Blackwell GPUs
- 2X reduced rack cooling power

FULL DATA CENTER WITH 32,000 GPUs

—

AI FACTORY FOR THE NEW INDUSTRIAL REVOLUTION

—

- 645 exaFLOPS of AI performance
- 13PB of fast memory
- 58PB/s of aggregate NVLink bandwidth
- 16.4 petaFLOPS of In-Network Computing

免责声明

以上材料仅为公开资料整理，不涉及分析师的研究观点和投资建议，记录和翻译可能产生误差，仅供参考，如有异议，请联系删除。

市场有风险，投资需谨慎。本平台所载内容和意见仅供参考，不构成对任何人的投资建议（专家、嘉宾或其他兴业证券股份有限公司以外的人士的演讲、交流或会议纪要等仅代表其本人或其所在机构之观点），亦不构成任何保证，接收人不应单纯依靠本平台的信息而取代自身的独立判断，应自主做出投资决策并自行承担风险。根据《证券期货投资者适当性管理办法》，本平台内容仅供兴业证券股份有限公司客户中的专业投资者使用，若您并非专业投资者，为保证服务质量、控制投资风险，请勿订阅或转载本平台中的信息，本资料难以设置访问权限，若给您造成不便，还请见谅。在任何情况下，作者及作者所在团队、兴业证券股份有限公司不对任何人因使用本平台中的任何内容所引致的任何损失负任何责任。

本平台旨在沟通研究信息，交流研究经验，不是兴业证券股份有限公司研究报告的发布平台，所发布观点不代表兴业证券股份有限公司观点。任何完整的研究观点应以兴业证券股份有限公司正式发布的报告为准。本平台所载内容仅反映作者于发出完整报告当日或发布本平台内容当日的判断，可随时更改且不予通告。

本平台所载内容不构成对具体证券在具体价位、具体时点、具体市场表现的判断或投资建议，不能够等同于指导具体投资的操作性意见。