NVIDIA H100 Tensor Core GPU 架构白皮书

为数据中心提供卓越的性能、可扩展性和安全性

V1.02

目录

插图目录

表格列表

简介

NVIDIA® 加速计算技术可以应对远超普通计算机能力的计算挑战。加速计算需要的不止是强大的 GPU。NVIDIA® CUDA® 通用可编程 GPU 与众多 GPU 加速的 SDK、API 和算法相结合，可提供全栈计算解决方案，为多个领域的应用带来惊人的加速效果。分布式 GPU 计算系统和软件已将处理范围扩展到整个数据中心。全球有越来越多的云数据中心使用 NVIDIA GPU 加速的系统和架构进行纵向扩展和横向扩展，运行各种 AI、高性能计算 (HPC) 和数据分析应用。

15 多年前，NVIDIA 推出了搭载 G80 GPU 的 CUDA 并行计算平台。从那时起，CUDA 工具和库的下载量已超过 3000 万次，服务了近 300 万名开发者。NVIDIA 一直在不断改进、优化和扩展 CUDA 平台，为之配备支持 CUDA 的更强大 GPU、各式新型 GPU 加速库、工作站、服务器和应用，用以扩大 NVIDIA 加速计算的覆盖范围。

NVIDIA 现在可以为不同的行业、科学领域和应用提供全栈解决方案。有 450 余个 NVIDIA SDK、工具包、库和模型为各种行业和应用提供服务，包括游戏与设计、生命与地球科学、机器人、自动驾驶汽车、量子计算、供应链物流、网络安全、5G、气候科学、数字生物学等。目前有超过 25000 家公司在使用 NVIDIA AI 技术。

NVIDIA CUDA 平台的编程简便性和丰富性使设计师、研究人员和工程师能够快速进行创新。随着平台软件的持续优化，用户通常会在 NVIDIA 产品的整个生命周期中体验到数倍的提速。

全球许多大型数据中心使用 NVIDIA GPU，为 AI、HPC 和数据分析系统及应用提供大幅提速。云数据中心正在借助 NVIDIA GPU 快速纵向扩展 AI 训练并横向扩展推理应用。许多不同类型的 AI 模型现已成熟并已工业化，可供企业广泛使用，并通过使用 NVIDIA GPU 进行训练和不断改进。成熟 AI 模型示例包括：计算机视觉模型、语音识别、推荐系统、图和树、时间序列模型、生成模型、可变编码器和大型语言模型。事实上，为新的语言和领域定制大型语言模型很可能是有史以来最大规模的超算应用之一。

NVIDIA 的新 Omniverse™ 平台将为众多元宇宙环境提供助力，这需要强大的 GPU 计算能力。除了将在许多基于 Omniverse 打造的元宇宙中支持实时渲染和模拟的 NVIDIA RTX GPU 之外，我们预计支持 H100 的系统能够为复杂的数字孪生挑战增加额外的 AI 和仿真能力。NVIDIA 自身的 Earth-2 超级计算机项目将是一个规模较大的超算项目，该项目将持续不断地将大量数据传输到在 Omniverse 中运行物理模拟的地球数字孪生中，以预测全球未来的天气模式。

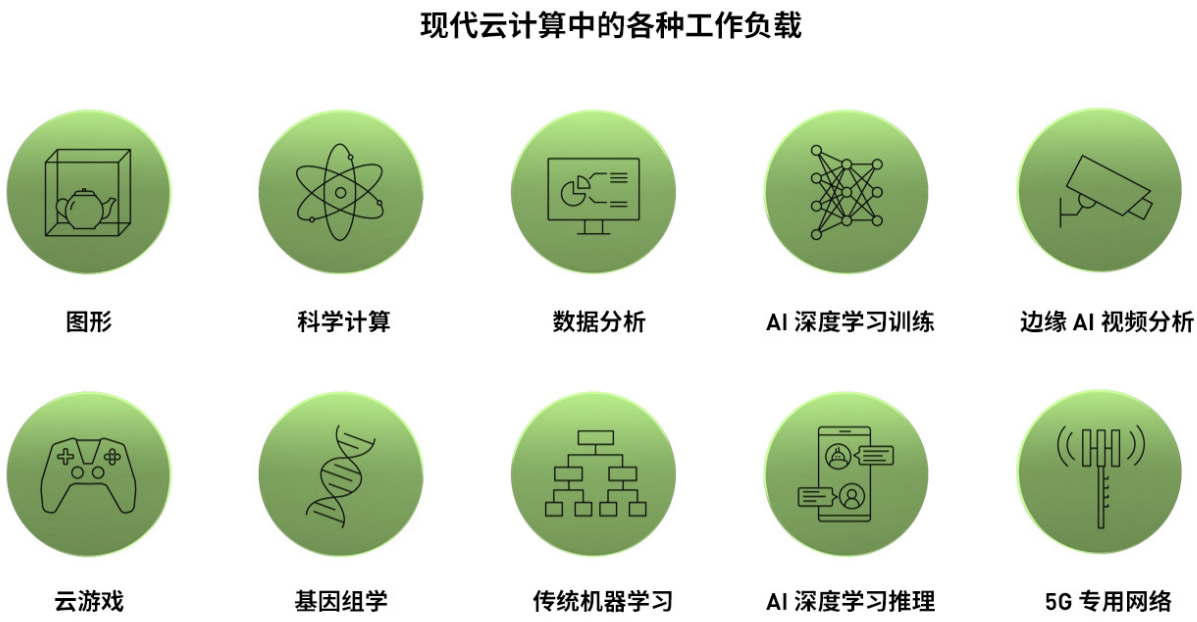


图 1. 现代云数据中心工作负载需要 NVIDIA GPU 加速

在本白皮书中，我们将介绍全新 NVIDIA H100 Tensor Core GPU，这是我们的新一代超高性能数据中心 GPU。H100 基于 NVIDIA Hopper GPU 架构构建，将加速云数据中心、服务器、边缘系统和工作站中的 AI 训练和推理、HPC 以及数据分析应用。

本文将简要介绍 H100、基于 H100 的新 DGX、DGX SuperPOD 和 HGX 系统以及基于 H100 的新融合加速器，然后深入探讨 H100 硬件架构、效率提升和新的编程功能。

NVIDIA H100 Tensor Core GPU 概述

人工智能 (AI)、高性能计算 (HPC) 和数据分析的复杂程度呈指数级上升，这要求科学家和工程师使用非常先进的计算平台。NVIDIA Hopper GPU 架构能够安全地提供低延迟的超高性能计算，并集成数据中心级计算的各种功能。

NVIDIA® H100 Tensor Core GPU 采用 NVIDIA Hopper GPU 架构，使 NVIDIA 数据中心平台的加速计算性能再次实现了重大飞跃。H100 可安全地加速各种工作负载，包括小型企业工作负载、百亿亿次级 (Exascale) HPC、万亿级参数 AI 模型等。

H100 是一款超先进的芯片，采用专为 NVIDIA 定制的 TSMC 4N 工艺制造，拥有 800 亿个晶体管，并包含多项架构改进。

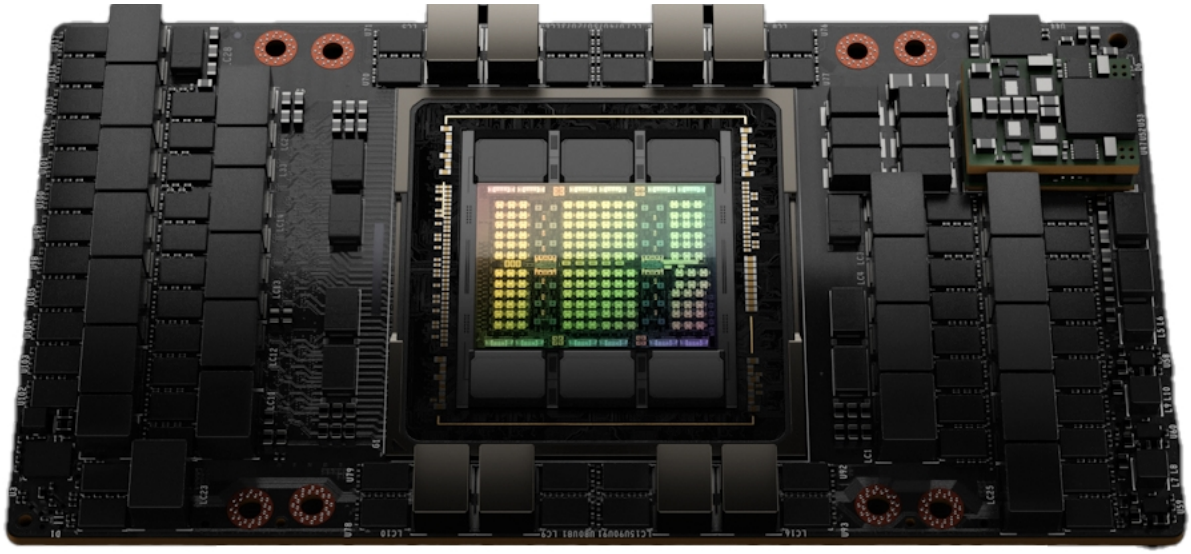
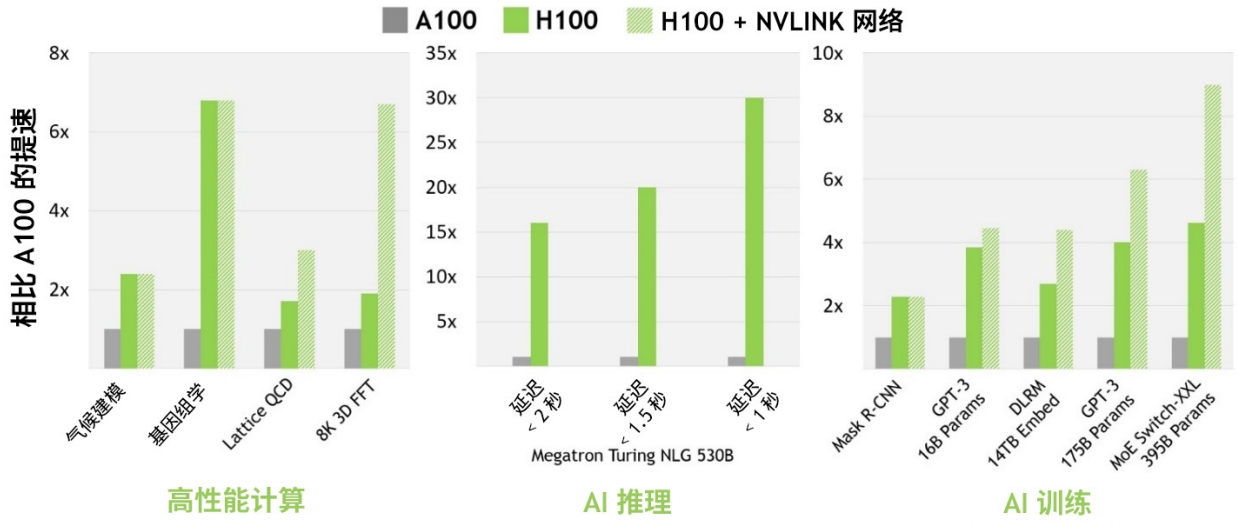


图 2　新 SXM5 模组上的 NVIDIA H100 GPU

H100 是 NVIDIA 的第 9 代数据中心 GPU，旨在为大规模 AI 和 HPC 实现相比于上一代 NVIDIA A100 Tensor Core GPU 数量级的性能飞跃。H100 延续了 A100 的主要设计重点，可提升 AI 和 HPC 工作负载的强大扩展能力，并显著提升架构效率。

对于当今主流的 AI 和 HPC 模型，采用 InfiniBand 互连技术的 H100 所提供的性能最高可达 A100 的 30 倍（见图 3）。

新的 NVLink Switch 系统互连技术面向的是一些具有超高挑战性的大型计算工作负载。这些工作负载需要跨多个 GPU 加速节点实现模型并行，从而又一次获得代际性能飞跃。在某些情况下，工作负载的性能将提升至使用 InfiniBand 技术的 H100 的 3 倍。



所有性能数据均为基于当前预期的初步数据，在交付产品中可能会有变化。A100 集群：HDR IB 网络。H100 集群：带 NVLink Switch 系统的 NDR IB 网络（若注明）。

GPU 数量：气候建模 1K、LQCD 1K、基因组学 8、3D-FFT 256、MT-NLG 32（批量大小：1 秒时 A100 为 4、H100 为 60，1.5 秒和 2 秒时 A100 为 8、H100 为 64），MRCNN 8（批量 32），GPT-3 16B 512（批量 256），DLRM 128（批量 64K），GPT-3 16K（批量 512），MoE 8K（批量 512，每个 GPU 一位专家）

图 3. H100 助力新一代 AI 和 HPC 实现突破

在 2022 年春季 GTC 大会上，新款 NVIDIA Grace Hopper 超级芯片产品发布。Hopper H100 Tensor Core GPU 将为 NVIDIA Grace Hopper 超级芯片 CPU+GPU 架构提供支持，该架构专为 TB 级加速计算而构建，并在大模型 AI 和 HPC 上提供 10 倍的性能。

NVIDIA Grace CPU 利用 Arm® 架构的灵活性来从头开始设计 CPU 和服务器架构，用于加速计算。H100 与 Grace 搭配，使用 NVIDIA 超快速的芯片间互连技术，可提供 900GB/s 的带宽，比 PCIe 5.0 快 7 倍。与当今运行超快的服务器相比，这种创新设计将总带宽提升 30 倍，并为运行 TB 级数据的应用程序提供高达 10 倍的性能。

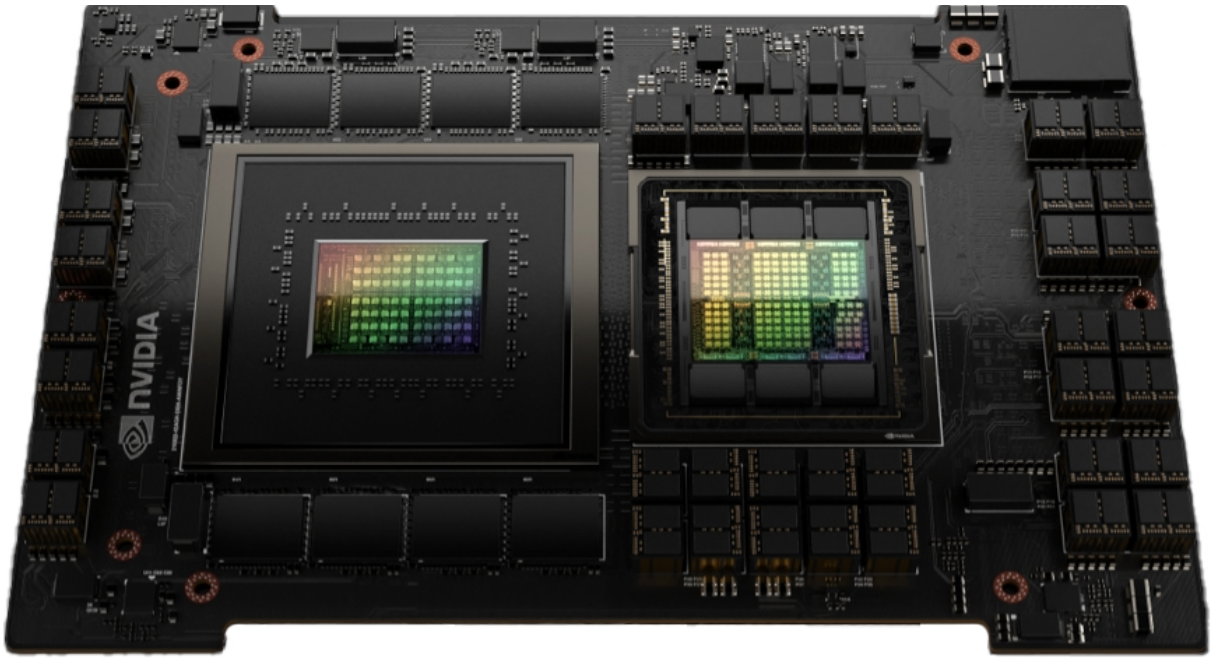


图 4. Grace Hopper 超级芯片

NVIDIA H100 GPU 关键特性摘要

- **新型流式多处理器 (SM)** 在性能和效率方面有许多改进。新的关键特性包括：
  - 与 A100 相比，新的**第四代 Tensor Core** 的芯片间速度最高可提升 6 倍，包括每 SM 提速、额外的 SM 数量以及更高的 H100 时钟频率。基于单个 SM ，与上一代 16 位浮点选项相比，Tensor Core 在同等数据类型上计算 MMA（矩阵乘积累加）速度是 A100 SM 的 2 倍，而在使用新的 FP8 数据类型时，计算速度是 A100 的 4 倍。稀疏功能利用深度学习网络中的细粒度结构化稀疏，使标准 Tensor Core 运算的性能提高了一倍。
  - 与 A100 GPU 相比，新的 **DPX 指令**最高可将动态编程算法的速度提升 7 倍。其中的两个示例包括用于基因组学处理的 Smith-Waterman 算法，以及在动态仓储环境中用于为机器人寻找最优路线的 Floyd-Warshall 算法。
  - 与 A100 相比，**IEEE FP64 和 FP32** 的芯片间处理速度可**提升 3 倍**，这是因为每个 SM 的时钟频率提升了 2 倍，此外还有额外的 SM 数量以及更高的 H100 时钟频率。
  - **新的线程块簇功能**允许以比单个 SM 上的单个线程块更大的粒度对局部性进行编程控制。这通过在编程层次结构中添加了另一个层级扩展了 CUDA 编程模型，现在其中包括线程、线程块、线程块簇和网格。簇支持多个线程块在多个 SM 上并发运行，以进行同步并以协作方式获取和交换数据。
  - 新的**异步执行**功能包括新的 **Tensor Memory Accelerator (TMA)** 单元，此单元可以在全局显存和共享内存之间非常高效地传输大数据块。TMA 还支持集群中线程块之间的异步拷贝。此外，还新增了**异步事务屏障**功能，用于执行原子数据移动和同步。
- 新的 **Transformer 引擎**结合了软件和定制的 Hopper Tensor Core 技术，专门用于加速 Transformer 模型的训练和推理。Transformer 引擎能够智能管理并动态选择 FP8 和 FP16 计算，自动处理每层中 FP8 和 FP16 之间的重铸和缩放，与上一代 A100 相比，可令大型语言模型的 AI 训练速度最高提升 9 倍、AI 推理速度最高提升 30 倍。
- 与上一代产品相比，**HBM3 显存子系统**的带宽提升了近 2 倍。H100 SXM5 GPU 率先采用 HBM3 显存，可提供 3TB/s 的超高显存带宽。
- **50 MB 二级缓存架构**可缓存大量模型和数据集以便于重复访问，从而减少对 HBM3 的访问。
- 与 A100 相比，**第二代多实例 GPU (MIG) 技术**提供的计算容量大约增加了 3 倍，每个 GPU 实例的显存带宽提升了近 2 倍。现在首次提供具有 MIG 级别可信执行环境 (TEE) 的机密计算能力。支持多达七个单独的 GPU 实例，每个实例均配备专门的 NVDEC 和 NVJPG 单元。每个实例现在都包含一套性能监控器，可与 NVIDIA 开发工具配合使用。

- **新的机密计算支持**可保护用户数据，抵御硬件和软件攻击，并能更好地隔离和保护虚拟化及 MIG 环境中的虚拟机 (VM)。H100 是全球首款支持原生机密计算的 GPU，并能够以 PCIe 全线速搭配 CPU 扩展可信执行环境。
- 与上一代 NVLink 相比，**第四代 NVIDIA NVLink®** 可将全局归约操作的带宽提升 3 倍，通用带宽提升 50%，同时多 GPU IO 的总带宽为 900GB/s，是 PCIe 5.0 的 7 倍。
- **第三代 NVSwitch** 技术包括位于节点内部和外部的交换机，用于连接服务器、集群和数据中心环境中的多个 GPU。节点内的每个 NVSwitch 具有 64 个第四代 NVLink 链路端口，可加速多 GPU 连接。交换机总吞吐量从上一代的 7.2Tb/s 提升到 13.6Tb/s。新的第三代 NVSwitch 技术还通过组播和 NVIDIA SHARP 在网计算，为集合运算提供硬件加速。
- 新的 **NVLink Switch 系统**互连技术和基于第三代 NVSwitch 技术的新的二级 **NVLink 交换机**引入了地址空间隔离和保护，使多达 32 个节点或 256 个 GPU 能够以 2:1 收敛比在胖树拓扑架构中通过 NVLink 进行连接。这些连接的节点能够提供 57.6TB/s 的多对多带宽，并可以提供惊人的 1 exaFLOP FP8 稀疏 AI 计算性能。
- **PCIe 5.0** 的总带宽为 128GB/s（每个方向 64GB/s），而 PCIe 4.0 的总带宽为 64GB/s（每个方向 32GB/s）。PCIe 5.0 支持 H100 与超高性能的 x86 CPU 和智能网卡 / DPU（数据处理器）交互。

此外，H100 还包括许多其他的新功能，以提升强大的扩展性、减少延迟和开销，并从总体上简化 GPU 编程。

本白皮书中的 **NVIDIA 加速数据中心**部分讨论了基于 H100 的新 DGX、HGX、融合加速器以及 AI 超算系统。

NVIDIA H100 GPU 架构深度解析包含 H100 GPU 架构特点、新的编程功能和性能提升的详细介绍。

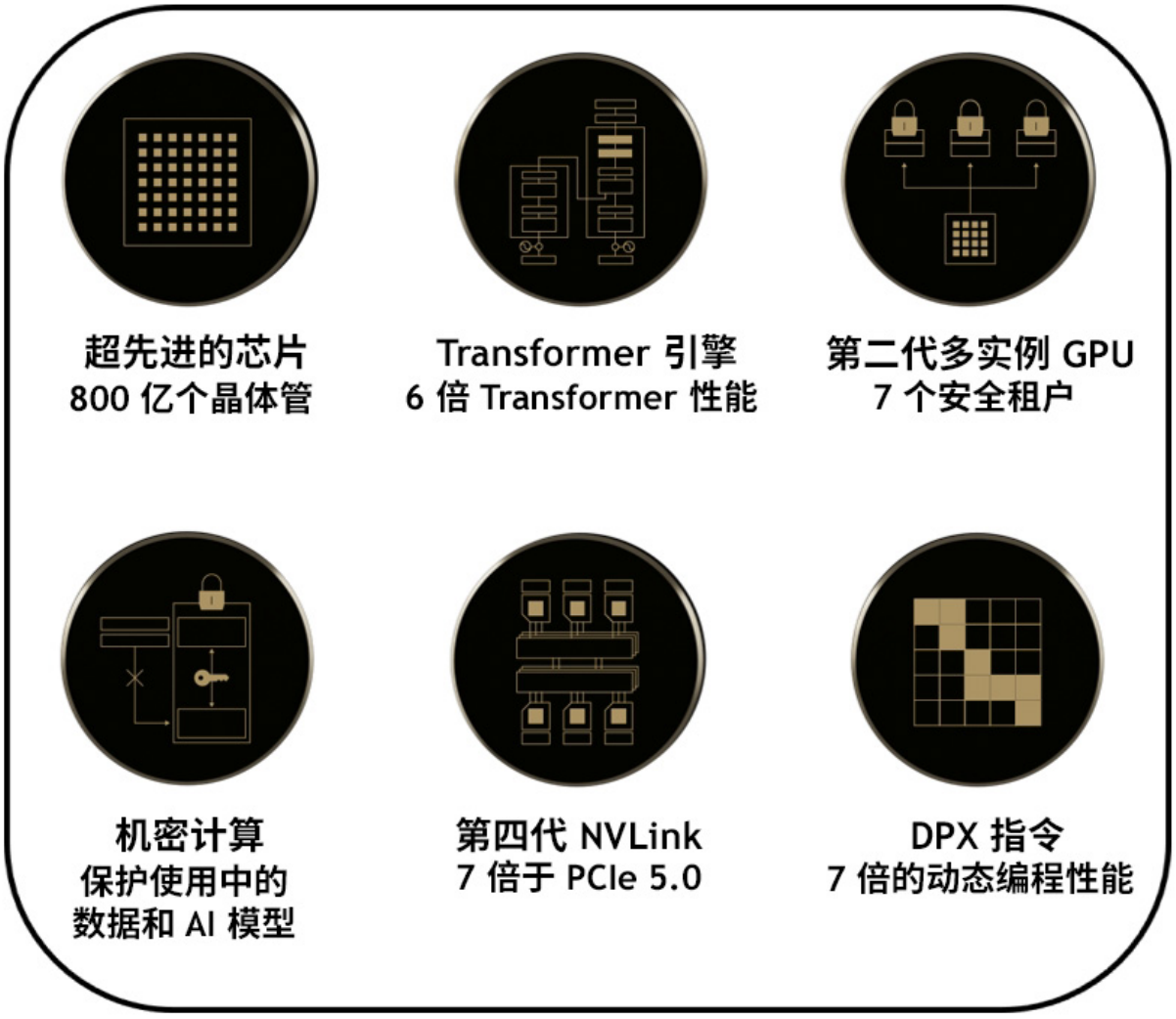


图 5. Hopper H100 中采用的新技术

基于 NVIDIA GPU 加速的数据中心

从 AI 和数据分析到高性能计算 (HPC)，数据中心是攻克某些重要挑战的关键。端到端的 NVIDIA 加速计算平台跨硬件和软件进行了集成，可为企业构建出鲁棒而安全的基础架构蓝图，支持在所有现代化工作负载中实施开发到部署的操作。

深度学习数据集正变得愈发庞大而复杂，诸如对话式 AI、推荐系统和计算机视觉之类的工作负载已在整个行业中变得越来越常见。NVIDIA 数据中心平台（包括硬件和软件）可显著加快 AI 训练速度，从而打造高效的数据科学团队、节省大量成本，同时缩短投资回报时间。

如要加速数据中心中的推理工作负载，则需要使用灵活且有弹性的基础架构，这样的基础架构可以横向扩展并能充分利用所有可用的计算资源。采用多实例 GPU (MIG) 等新技术，NVIDIA 解决方案的独特之处就是加速推理工作负载（例如图像识别、推荐系统和自然语言处理），提供应用 AI 所需的超高吞吐量和实时响应能力。

HPC 是加速数据中心中相关科学发展的重要工具之一。NVIDIA GPU 是现代 HPC 数据中心的引擎。NVIDIA 数据中心平台以更少的服务器提供突破性性能，可帮助用户更快地分析并显著降低成本，从而为科学发现铺平了道路。

企业正在生成和收集海量数据。当越多的数据用于分析，越多的信息就可以被学习到。借助 NVIDIA 数据中心平台和分析解决方案，企业能够比以往更快地从其数据中获得可行见解。

用于数据中心的 NVIDIA GPU 加速功能可通过 NVIDIA 庞大的服务器合作伙伴生态中的各种服务器来实现。H100 GPU 提供不同的配置，以满足不同的服务器设计需求。

以下部分简要介绍了适用于数据中心且基于 H100 的 NVIDIA 系统和主板，包括 SMX5 和 PCIe 5.0 外形规格的 H100 GPU、DGX H100 和 DGX SuperPOD 系统、HGX H100，以及结合了 NVIDIA H100 GPU 的强大功能和 NVIDIA® ConnectX-7 智能网卡的高级网络功能的 H100 CNX 融合加速器。请参阅**附录 A - NVIDIA DGX - 数据中心 AI 的基础构件**，详细了解 DGX H100 系统。

H100 SXM5 GPU

H100 SXM5 配置采用 NVIDIA 定制的 SXM5 主板，该主板包含 H100 GPU 和 HBM3 显存堆栈，并提供第四代 NVLink 和 PCIe 5.0 连接，可提供极高的应用性能。这个配置是客户应用扩展到单机多卡和多机应用的理想配置。HGX H100 服务器主板可提供 4 GPU 和 8 GPU 配置。4 GPU 配置包含 GPU 之间的 P2P NVLink 连接，并能提高服务器中的 CPU 与 GPU 的比率，而 8 GPU 配置包含 NVSwitch，可在任意一对 GPU 之间提供 SHARP 在网计算和 900GB/s 的完整 NVLink 带宽。H100 SXM5 GPU 还可用于功能强大的新型 DGX H100 服务器和 DGX SuperPOD 系统。

H100 PCIe 5.0 GPU

H100 PCIe 5.0 配置可在仅为 350 瓦的热设计功耗 (TDP) 下提供 H100 SXM5 GPU 的所有功能。此配置可以选择使用 NVLink 桥接器以 600GB/s 的带宽连接最多两个 GPU，该带宽几乎是 PCIe 5.0 的 5 倍。H100 PCIe 非常适合放入到标准机架的主流加速服务器（降低了每台服务器的功耗），它为一次扩展到 1 或 2 个 GPU 的应用（包括 AI 推理和一些 HPC 应用）提供了出色的性能。在排名前 10 的数据分析、AI 和 HPC 应用中，单个 H100 PCIe GPU 可高效提供达到 H100 SXM5 GPU 65% 的性能，同时功耗仅为 50%。

DGX H100 和 DGX SuperPOD

NVIDIA DGX H100 是用于训练、推理和分析的通用高性能 AI 系统。DGX H100 配备 BlueField-3、NDR InfiniBand 和第二代 MIG 技术。单个 DGX H100 系统可提供独一无二的 16 petaFLOPS FP16 稀疏 AI 计算性能。通过将多个 DGX H100 连接到集群中，如 DGX POD 或 DGX SuperPOD ，可以轻松扩展这种性能。DGX SuperPOD 至少可支持 32 个 DGX H100 系统（称为“可扩展单元”），集成了 256 个 H100 GPU，这些 GPU 通过基于第三代 NVSwitch 技术的新的二级 NVLink 交换机连接，可提供出色的 1 exaFLOP FP8 稀疏 AI 计算性能。DGX H100 SuperPOD 可同时支持 InfiniBand 和 NVLINK 交换机网络选项。

请参阅**附录 A - NVIDIA DGX - 数据中心 AI 的基础构件**，了解更多详情。

HGX H100

随着工作负载的复杂度持续呈爆炸式增长，人们需要多个 GPU 协同工作，且需要它们彼此之间超高速通讯。NVIDIA HGX H100™ 利用 NVLink 和 NVSwitch 驱动的高速互联技术将多个 H100 GPU 结合起来，可打造出功能超强的垂直扩展式服务器。

HGX H100 可作为服务器构件，以集成主板的形式提供 4 个或 8 个 H100 GPU 配置。4 GPU 配置的 HGX H100 在 GPU 之间提供完全互连的 P2P NVLink 连接，而 8 GPU 配置通过 NVSwitch 提供完整的 GPU 到 GPU 带宽。8 路 HGX H100 利用 H100 多精度 Tensor Core 的强大功能，通过稀疏 FP8 运算可实现超过 32 petaFLOP 的深度学习计算性能。HGX H100 支持标准化的高性能服务器，基于各种应用工作负载提供可预测的性能，同时还能缩短 NVIDIA 生态系统中服务器制造商合作伙伴的产品上市时间。

H100 CNX 融合加速器

NVIDIA H100 CNX 结合了 NVIDIA H100 GPU 的强大功能和 NVIDIA® ConnectX-7 智能网卡的高级网络功能，最高可提供 400Gb/s 的带宽，并包含 NVIDIA ASAP2（加速交换和数据包处理）以及用于 TLS/IPsec/MACsec 加密/解密的在线硬件加速等创新功能。这种独特架构可为 GPU 驱动的 I/O 密集型工作负载提供出色的性能，例如企业数据中心中的分布式 AI 训练或边缘侧的 5G 信号处理。

NVIDIA H100 GPU 架构深度解析

基于新 Hopper GPU 架构的 NVIDIA H100 GPU 具有多项创新：

- 在应对更广泛的 AI 和 HPC 任务时，新的第四代 Tensor Core 能够执行速度更胜以往的矩阵计算。
- 与上一代 A100 相比，采用新 Transformer 引擎的 H100 可令大型语言模型的 AI 训练速度提升 9 倍，AI 推理速度提升 30 倍。
- 新的 NVLink 网络互连可在跨多个计算节点的多达 256 个 GPU 之间，支持 GPU 到 GPU 通信。
- 安全 MIG 可将 GPU 划分为大小合适的独立实例，以更大限度地提升较小工作负载的服务质量 (QoS)。

NVIDIA H100 是第一款完全异步 GPU。H100 扩展了 A100 跨所有地址空间的全局到共享异步传输，并增加了对张量内存存取模式的支持。它使应用能够构建端到端的异步流水线，将数据移入和移出芯片，在完成计算同时完全隐藏数据搬运。

现在，只有少量 CUDA 线程需要使用新的 Tensor 内存加速器来管理 H100 的全部显存带宽，而大多数其他 CUDA 线程可以专注于通用计算，例如为新一代 Tensor Core 预处理和后处理数据。

H100 在 CUDA 线程组层次结构中增加了一个名为“线程块簇”的新层级。簇是一组保证可以并发调度的线程块，支持跨多个 SM 的线程进行高效协作和数据共享。簇还可以更高效地协同驱动 Tensor 内存加速器和 Tensor Core 等异步单元。

编排不断增多的片内加速器和各种通用线程组需要同步。例如，消费输出的线程和加速器必须等待生产输出的线程和加速器产生输出。

NVIDIA 的异步事务屏障使集群内的通用 CUDA 线程和片内加速器能够高效地同步，即使它们位于不同的 SM 上。所有这些新功能使每个用户和应用能够始终充分利用其 H100 GPU 的所有单元，这使 H100 成为功能强大、可编程性强的节能高效 GPU。

为 H100 GPU 提供支持的完整 GH100 GPU 核心采用专为 NVIDIA 定制的 TSMC 4N 工艺制造，拥有 800 亿个晶体管，芯片大小为 814 mm^2，并采用较高频率的设计。

NVIDIA GH100 GPU 由多个 GPU 处理集群 (GPC)、纹理处理集群 (TPC)、流多处理器 (SM)、二级缓存和 HBM3 内存控制器组成。

完整的 GH100 GPU 架构包括以下单元：

- 8 个 GPC、72 个 TPC（9 个 TPC/GPC）、2 个 SM/TPC、每个完整 GPU 内含 144 个 SM
- 每个 SM 内含 128 个 FP32 CUDA Core 核心、每个完整 GPU 内含 18432 个 FP32 CUDA Core 核心
- 每个 SM 内含 4 个第四代 Tensor Core 核心、每个完整 GPU 内含 576 个第四代 Tensor Core 核心
- 6 个 HBM3 或 HBM2e 堆栈、12 个 512 位内存控制器
- 60MB 二级缓存
- 第四代 NVLink 和 PCIe 5.0

采用 SXM5 主板封装的 NVIDIA H100 GPU 包括以下单元：

- 8 个 GPC、66 个 TPC、2 个 SM/TPC、每个 GPU 内含 132 个 SM
- 每个 SM 内含 128 个 FP32 CUDA Core 核心、每个 GPU 内含 16896 个 FP32 CUDA Core 核心
- 每个 SM 内含 4 个第四代 Tensor Core 核心、每个 GPU 内含 528 个第四代 Tensor Core 核心
- 80GB HBM3、5 个 HBM3 堆栈、10 个 512 位内存控制器
- 50MB 二级缓存
- 第四代 NVLink 和 PCIe 5.0

采用 PCIe 5.0 主板封装的 NVIDIA H100 GPU 包括以下单元：

- 7 或 8 个 GPC、57 个 TPC、2 个 SM/TPC、每个 GPU 内含 114 个 SM
- 每个 SM 内含 128 个 FP32 CUDA Core 核心、每个 GPU 内含 14592 个 FP32 CUDA Core 核心
- 每个 SM 内含 4 个第四代 Tensor Core 核心、每个 GPU 内含 456 个第四代 Tensor Core 核心
- 80GB HBM2e、5 个 HBM2e 堆栈、10 个 512 位内存控制器
- 50MB 二级缓存
- 第四代 NVLink 和 PCIe 5.0

与采用 TSMC 7nm N7 制造工艺的上一代 GA100 GPU 相比，采用 TSMC 4N 制造工艺使 H100 能够增加 GPU 核心频率、提高性能功耗比，并整合更多的 GPC、TPC 和 SM。

图 6 显示了配备 144 个 SM 的**完整 GH100 GPU** 核心。H100 SXM5 GPU 配备 132 个 SM，PCIe 版本配备 114 个 SM。请注意，H100 GPU 的主要用途为执行 AI、HPC 和数据分析的数据中心及边缘计算工作负载，而非图形处理。SXM5 和 PCIe H100 GPU 中只有两个 TPC 具备图形处理能力（也就是说，它们可以运行顶点、几何图形和像素着色器）。

图 6. 配备 144 个 SM 的完整 GH100 GPU 核心

H100 SM 架构

H100 SM 基于 NVIDIA A100 Tensor Core GPU SM 架构而构建。由于引入了 FP8，与 A100 相比，H100 SM 将每 SM 浮点计算能力峰值提升了 4 倍，并且对于之前所有的 Tensor Core 和 FP32 / FP64 数据类型，将各个时钟频率下的原始 SM 计算能力增加了一倍。

与上一代 A100 相比，采用 Hopper 的 FP8 Tensor Core 的新 Transformer 引擎使大型语言模型的 AI 训练速度提升 9 倍，AI 推理速度提升 30 倍。针对用于基因组学和蛋白质测序的 Smith-Waterman 算法，Hopper 的新 DPX 指令可将其处理速度提升 7 倍。

Hopper 新的第四代 Tensor Core、Tensor 内存加速器以及许多其他新 SM 和 H100 架构的总体改进，在许多其他情况下可令 HPC 和 AI 性能获得最高 3 倍的提升。

表 1. NVIDIA H100 Tensor Core GPU 初步性能规格

	NVIDIA H100 SXM5[1]	**NVIDIA H100 PCIe**[1]
FP64 峰值性能 [1]	30 TFLOPS	24 TFLOPS
FP64 Tensor Core 峰值性能 [1]	60 TFLOPS	48 TFLOPS
FP32 峰值性能 [1]	60 TFLOPS	48 TFLOPS
FP16 峰值性能 [1]	120 TFLOPS	96 TFLOPS
BF16 峰值性能 [1]	120 TFLOPS	96 TFLOPS
TF32 Tensor Core 峰值性能 [1]	500 TFLOPS | 1000 TFLOPS[2]	400 TFLOPS | 800 TFLOPS[2]
FP16 Tensor Core 峰值性能 [1]	1000 TFLOPS | 2000 TFLOPS[2]	800 TFLOPS | 1600 TFLOPS[2]
BF16 Tensor Core 峰值性能 [1]	1000 TFLOPS | 2000 TFLOPS[2]	800 TFLOPS | 1600 TFLOPS[2]
FP8 Tensor Core 峰值性能 [1]	2000 TFLOPS | 4000 TFLOPS[2]	1600 TFLOPS | 3200 TFLOPS[2]
INT8 Tensor Core 峰值性能 [1]	2000 TOPS | 4000 TOPS[2]	1600 TOPS | 3200 TOPS[2]

1. **基于当前预期对 H100 进行初步性能评估，交付产品可能会有变化**
2. **使用稀疏特性实现有效 TFLOPS / TOPS**

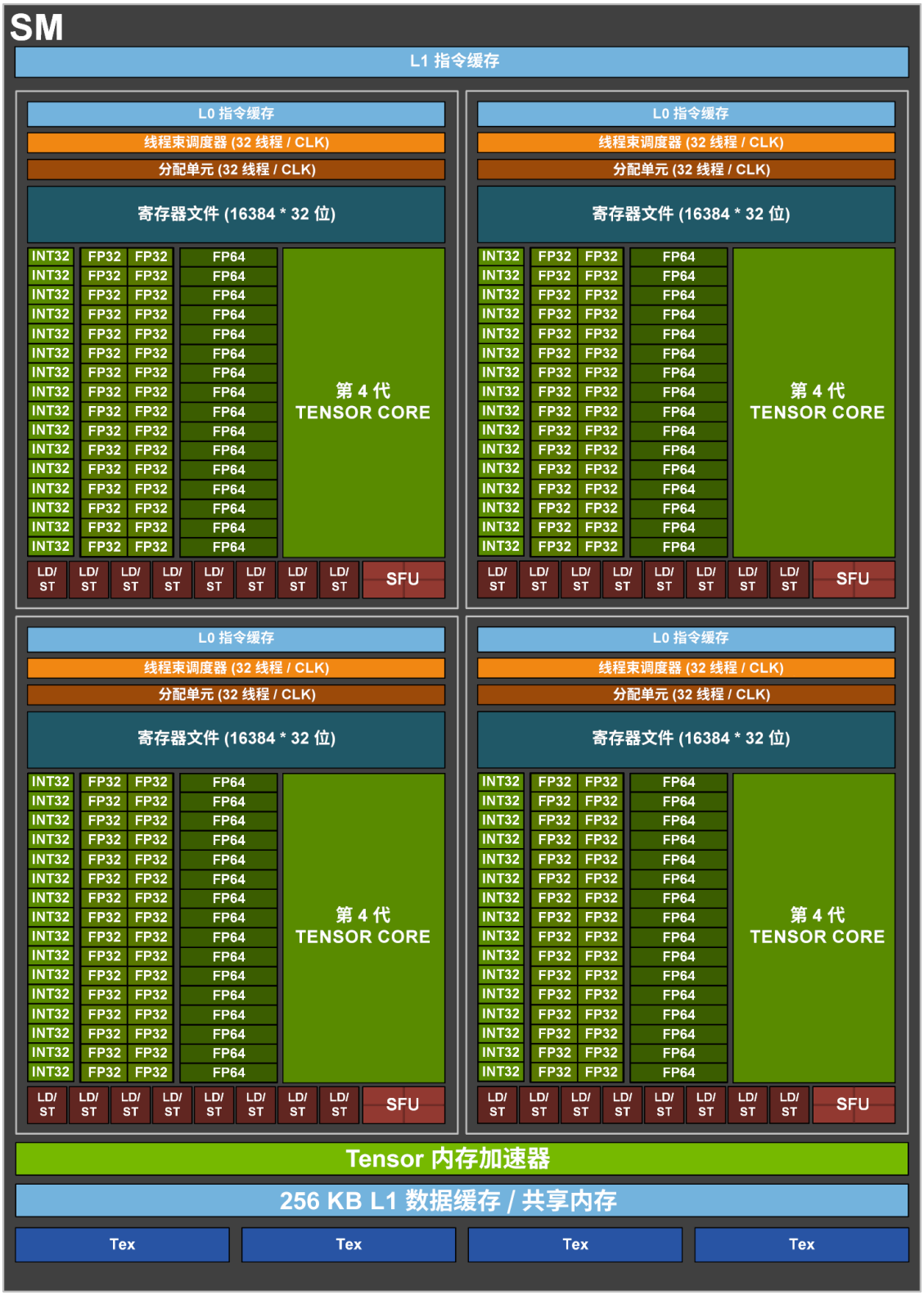


图 7. GH100 流式多处理器 (SM)

H100 SM 关键特性摘要

- 第四代 Tensor Core：
  - 与 A100 相比，芯片间速度最高可提升 6 倍，包括每 SM 提速、额外的 SM 数量以及更高的 H100 时钟频率。
  - 在每个 SM 的基础上，与上一代 16 位浮点计算相比，Tensor Core 在同等数据类型上计算速度是 A100 SM 的 MMA（矩阵乘积累加）2 倍，而在使用新的 FP8 数据类型时，计算速度是 A100 的 4 倍。
  - 稀疏特性利用深度学习网络中的细粒度结构化稀疏，使标准 Tensor Core 运算的性能提高了一倍。
- 与 A100 GPU 相比，新的 **DPX 指令集**最高可将动态编程算法的速度提升 7 倍。其中的两个示例包括用于基因组学处理的 Smith-Waterman 算法，以及在动态仓储环境中用于为机器人寻找最优路线的 Floyd-Warshall 算法。
- 与 A100 相比，**IEEE FP64 和 FP32** 的芯片间处理速度**可提升 3 倍**，这是因为每个 SM 的时钟频率提升了 2 倍，此外还有额外的 SM 数量以及更高的 H100 时钟频率。
- 256KB 的组合共享内存和 L1 数据缓存，比 A100 大 1.33 倍。
- 新的**异步执行**功能包括新的 **Tensor Memory Accelerator (TMA)** 单元，其可以在全局显存和共享显存之间高效地传输大量数据块。TMA 还支持簇中线程块之间的异步拷贝。此外，还新增了**异步事务屏障**功能，用于执行原子数据移动和同步。
- 新的**线程块簇**功能支持跨多个 SM 局部性控制。
- **分布式共享内存**允许跨多个 SM 共享内存块进行加载、存储和原子操作的 SM 到 SM 直接通信。

H100 Tensor Core 架构

Tensor Core 是专门用于矩阵乘积累加 (MMA) 数学运算的高性能计算核心，可大大提升 AI 和 HPC 应用的性能。与标准浮点 (FP) 运算、整数 (INT) 运算和融合乘加 (FMA) 运算相比，在一个 NVIDIA GPU 内跨 SM 并行运行的 Tensor Core 可大幅提高吞吐量和效率。我们在 NVIDIA Tesla® V100 GPU 中首次引入了 Tensor Core，并会在每一代新的 NVIDIA GPU 架构中不断增强这一核心。

与 A100 相比，H100 中新的第四代 Tensor Core 架构可使每时钟每个 SM 的原始密集计算和稀疏矩阵运算吞吐量提升一倍，考虑到 H100 比 A100 拥有更高的 GPU 加速频率，其甚至会达到更高的吞吐量。其支持 FP8、FP16、BF16、TF32、FP64 和 INT8 MMA 数据类型。新的 Tensor Core 还能够实现更高效的数据管理，最高可节省 30% 的操作数传输功耗。

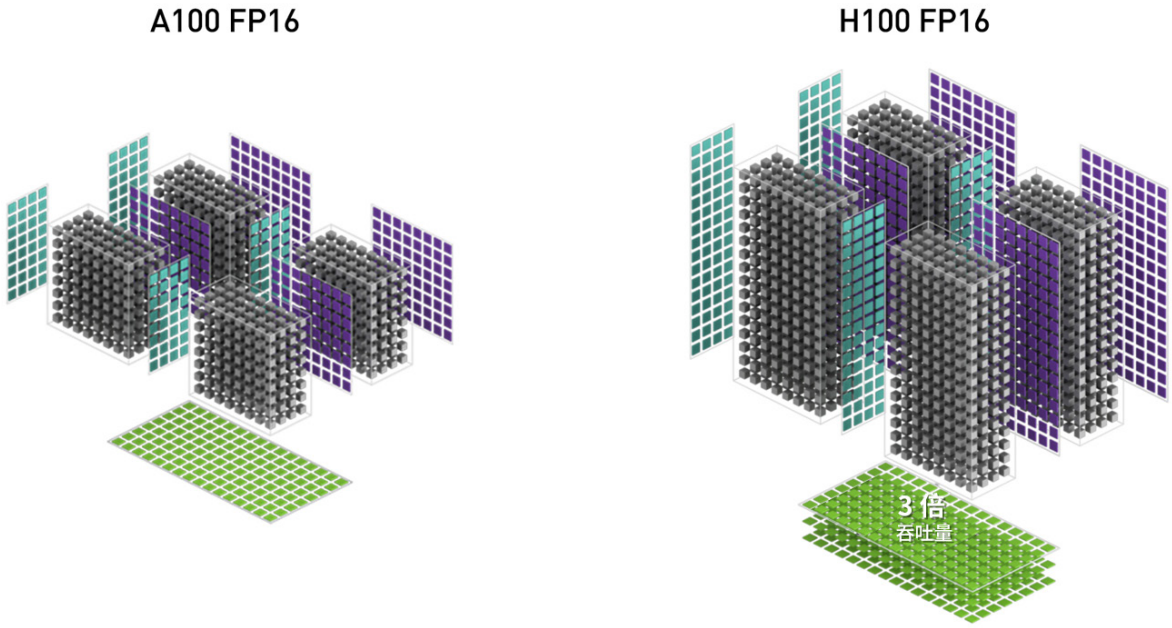


图 8. H100 FP16 Tensor Core 的吞吐量是 A100 FP16 Tensor Core 的 3 倍

Hopper FP8 数据格式

H100 GPU 增加了 FP8 Tensor Core，可加速 AI 训练和推理。如图 9 所示，FP8 Tensor Core 支持 FP32 和 FP16 累加器，以及两种新的 FP8 输入类型：

- E4M3：具有 4 个指数位、3 个尾数位和 1 个符号位
- E5M2：具有 5 个指数位、2 个尾数位和 1 个符号位。

E4M3 支持动态范围更小、精度更高的计算，而 E5M2 可提供更宽广的动态范围和更低的精度。与 FP16 或 BF16 相比，FP8 可将所需要的数据存储空间减半，并将吞吐量提升一倍。

新的 Transformer 引擎（如下文所述）可结合使用 FP8 和 FP16 精度，减少内存使用并提高性能，同时仍能保持大型语言模型和其他模型的准确性。

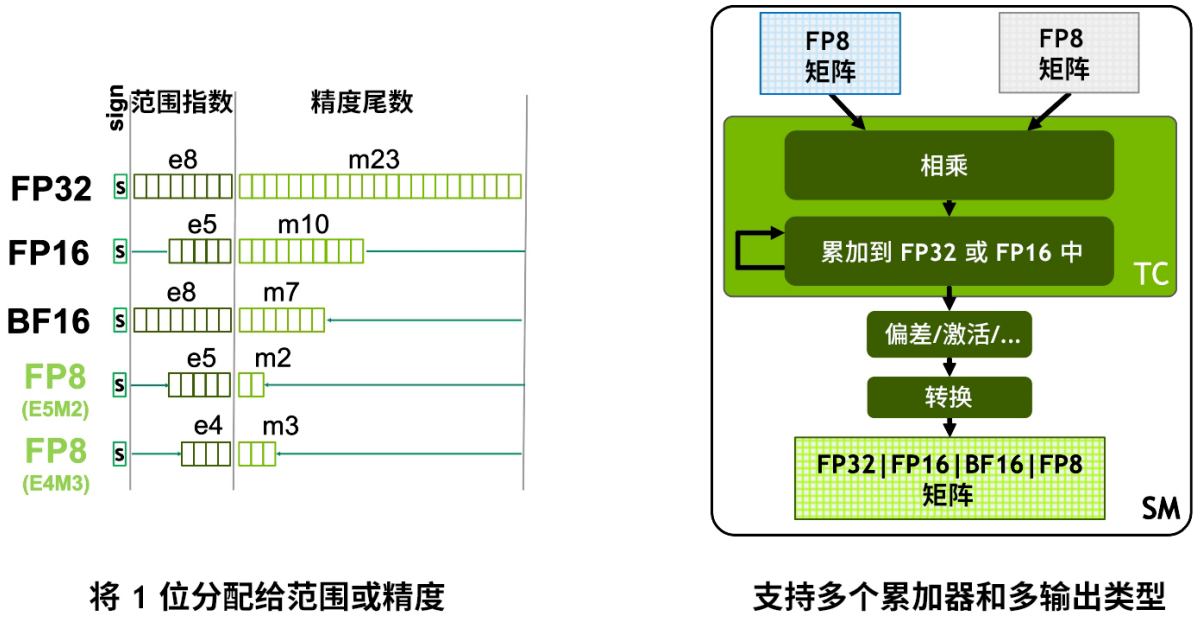


图 9. 新 Hopper FP8 精度 - 相比于 H100 FP16 / BF16，吞吐量提升一倍、占用空间减半

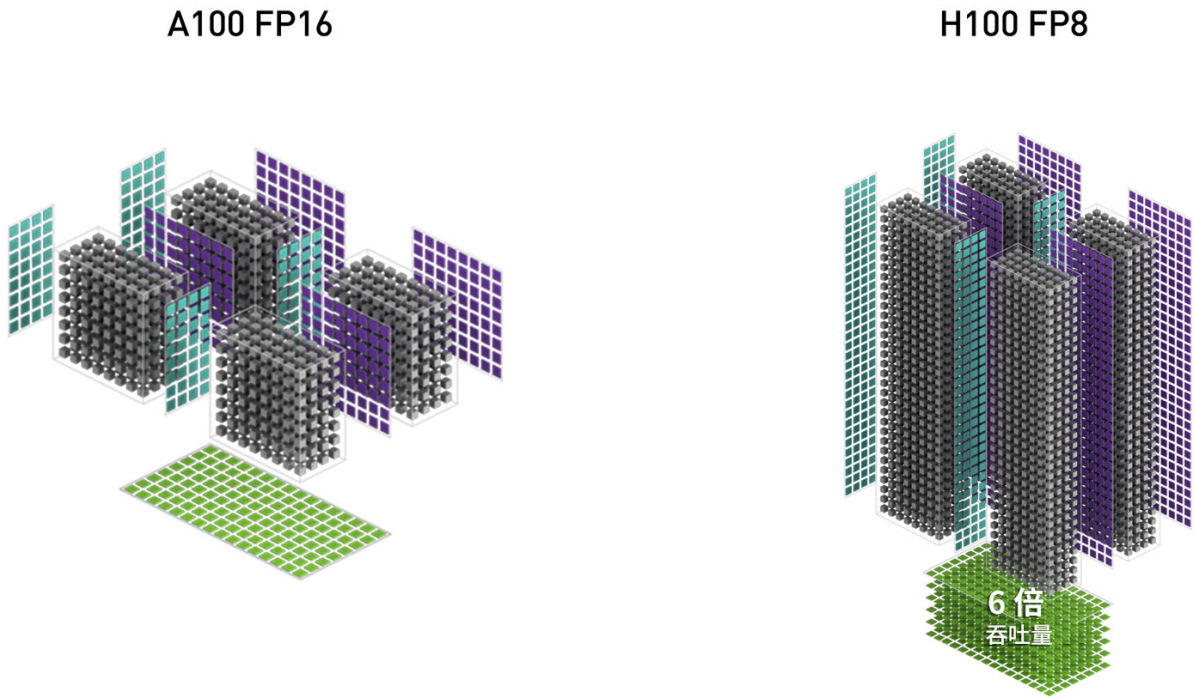


图 10. H100 FP8 Tensor Core 的吞吐量是 A100 FP16 Tensor Core 的 6 倍

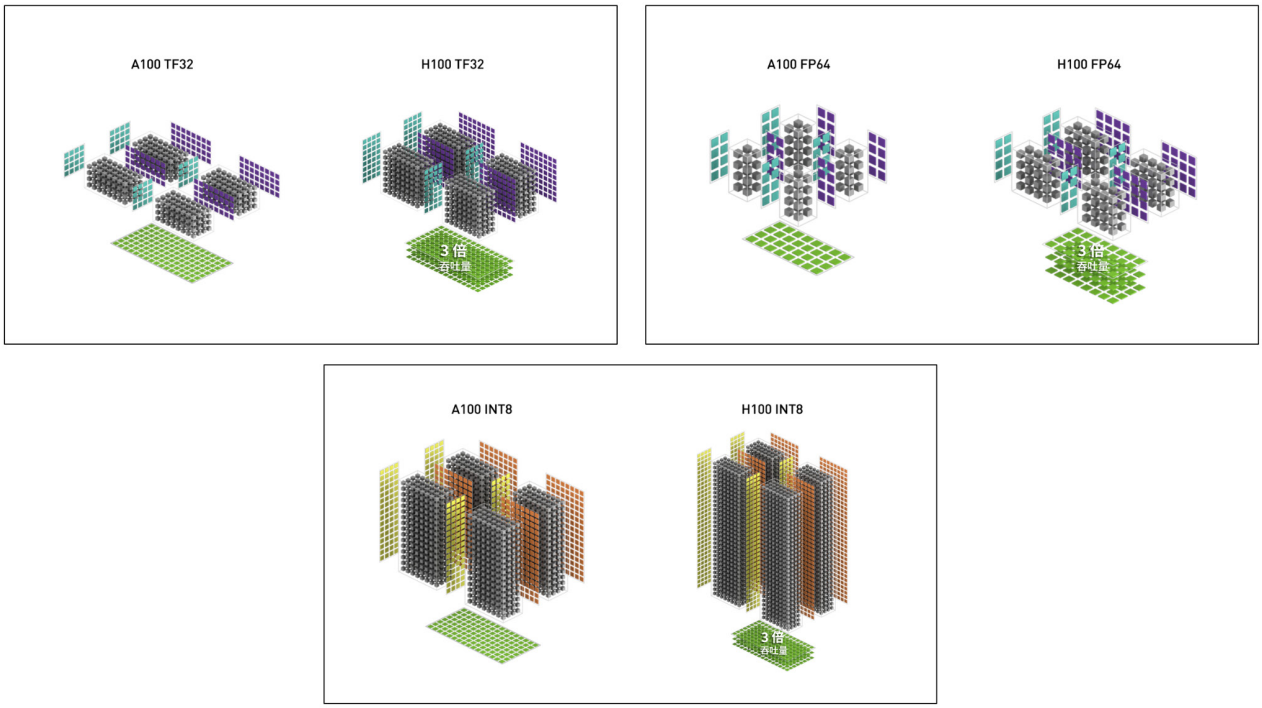


图 11. H100 TF32、FP64 和 INT8 Tensor Core 的吞吐量均为 A100 的 3 倍

对于多种数据类型，H100 的数学运算速度相比 A100 的提升如下表 2 所示。

表 2. H100 相比 A100 的提速（初步 H100 性能，TC = Tensor Core）

	A100	A100 稀疏	H100 SXM5[1]	H100 SXM5[1] 稀疏	H100 SXM5[1] 相比 A100 的提速
FP8 Tensor Core	NA	NA	2000 TFLOPS	4000 TFLOPS	6.4 倍 （相比于 A100 FP16）
FP16	78 TFLOPS	NA	120 TFLOPS	NA	1.5 倍
FP16 Tensor Core	312 TFLOPS	624 TFLOPS	1000 TFLOPS	2000 TFLOPS	3.2 倍
BF16 Tensor Core	312 TFLOPS	624 TFLOPS	1000 TFLOPS	2000 TFLOPS	3.2 倍
FP32	19.5 TFLOPS	NA	60 TFLOPS	NA	3.1 倍
TF32 Tensor Core	156 TFLOPS	312 TFLOPS	500 TFLOPS	1000 TFLOPS	3.2 倍
FP64	9.7 TFLOPS	NA	30 TFLOPS	NA	3.1 倍
FP64 Tensor Core	19.5 TFLOPS	NA	60 TFLOPS	NA	3.1 倍
INT8 Tensor Core	624 TOPS	1248 TOPS	2000 TFLOPS	4000 TFLOPS	3.2 倍

1 - 基于当前预期对 H100 进行初步性能评估，交付产品可能会有变化

用于加速动态规划的新 DPX 指令

许多“暴力”优化算法具有这样的特性：在解决更大的问题时，会多次重复使用子问题的解决方案。作为一种算法技术，动态规划的解决方式是将复杂递归问题分解为更简单的子问题。动态规划算法可以存储子问题的结果，而无需在以后需要时重新加以计算，从而将指数级问题集的计算复杂性降低到了线性规模。

动态规划通常广泛应用于优化、数据处理和基因组学算法中。在快速发展的基因组测序领域，Smith-Waterman 动态规划算法是目前使用的最重要的方法之一。在机器人开发领域，Floyd-Warshall 是一种用于在动态仓储环境中为机器人实时寻找最优路线的关键算法。

H100 引入了 DPX 指令，使动态规划算法的性能提高到 Ampere GPU 的 7 倍。这些新指令为许多动态规划 (DP) 算法的内循环提供了对高级融合操作数的支持。对于疾病诊断、物流路线优化以及图形分析领域，这将大幅缩短解决问题的时间。

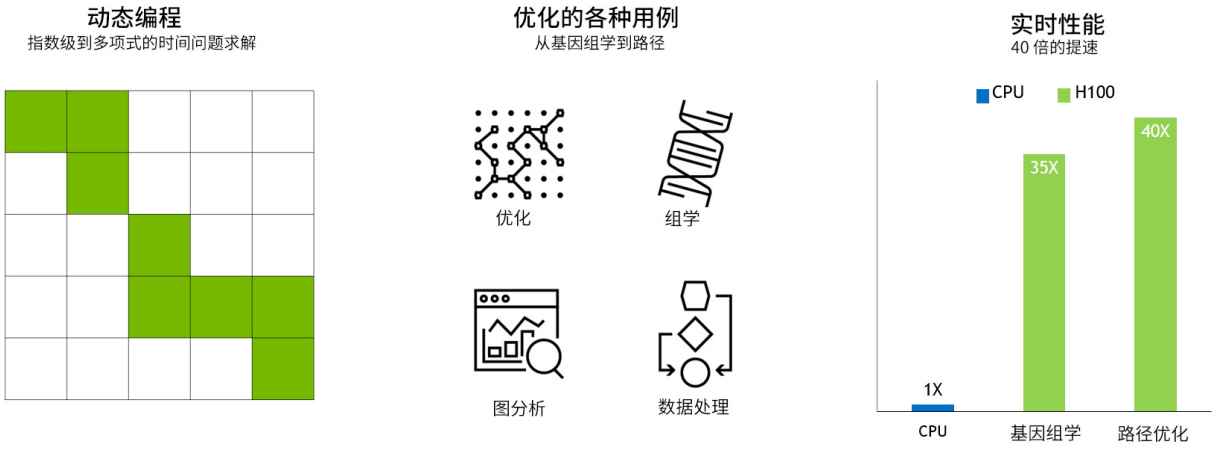


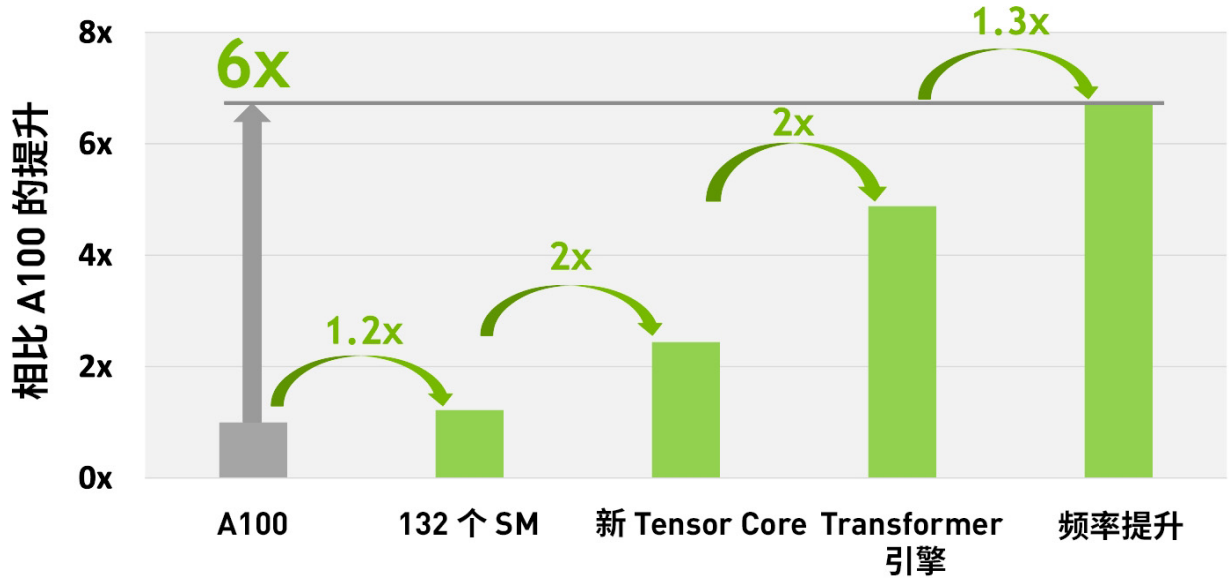
图 12. DPX 指令加速动态规划

L1 数据缓存和共享内存合并

在 Volta V100 中首次引入的 NVIDIA 合并 L1 数据缓存和共享内存子系统架构可大幅提升性能，同时还简化编程并减少实现应用性能峰值或接近峰值所需的调优。将数据缓存和共享内存功能合并到同一内存块中，可为两种类型的内存访问都提供出色的整体性能。H100 中 L1 数据缓存和共享内存的合并容量为 256KB/SM，而 A100 中的容量为 192KB/SM。在 H100 中，SM 共享内存容量可以配置，最高支持 228KB 。

H100 计算性能总结

总体而言，综合 H100 中所有新的计算技术进步的因素，H100 的计算性能比 A100 提高了约 6 倍。图 13 总结了 H100 的提升，首先是 H100 配备 132 个 SM，比 A100 的 108 个 SM 增加了 22%。由于采用新的第四代 Tensor Core，每个 H100 SM 的速度都提升了 2 倍。在每个 Tensor Core 中，新的 FP8 格式和相应的 Transformer 引擎又将性能提升了 2 倍。最后，H100 中更高的时钟频率将性能再提升了约 1.3 倍。通过这些改进，总体而言，H100 的峰值计算吞吐量大约为 A100 的 6 倍，对于计算密集型工作负载而言，这是一次重大飞跃。



H100 可为计算密集型工作负载提供 6 倍的吞吐量

图 13. H100 计算提升总结

H100 GPU 层次结构和异步改进

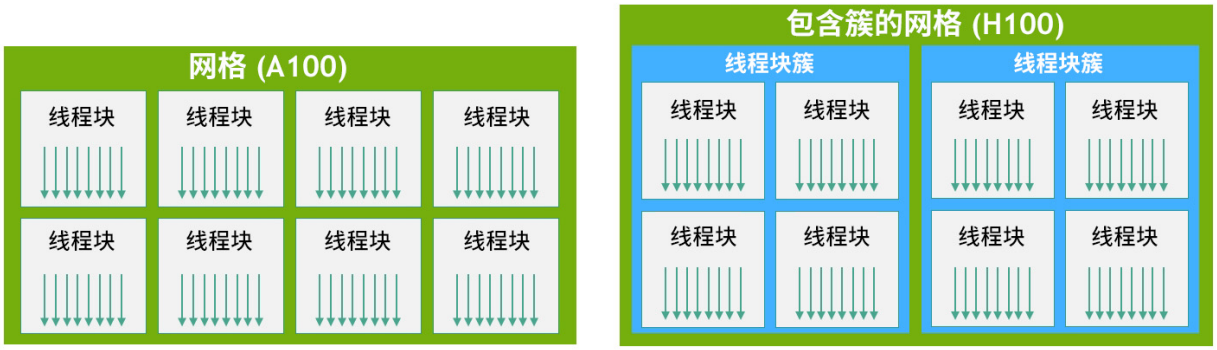
要保障并行程序的高性能，两个关键要素是数据本地性和异步执行。通过将程序数据尽可能靠近执行单元，程序员可利用本地数据的低延迟和高访问带宽所带来的性能。异步执行包括寻找独立任务来掩藏内存传输和其他处理。这样做旨在充分利用 GPU 中的所有单元。我们将探索在 Hopper 的 GPU 编程层次结构中添加一个重要的新层级，该层级支持比单个 SM 上的单个线程块更大规模的局部性。我们还将介绍新的异步执行功能，这些功能可以提高性能并减少同步开销。

线程块簇

长期以来，CUDA 编程模型一直依赖于 GPU 计算架构，该架构使用包含多个线程块的网格来利用程序中的本地性。线程块包含在单个 SM 上并发运行的多个线程，这些线程可以与快速屏障同步，并使用 SM 的共享内存交换数据。但是，随着 GPU 的 SM 数量超过 100，且计算程序变得更加复杂，线程块作为编程模型中表示的唯一本地性单元，不足以更大限度地提高执行效率。

H100 引入了一种新的线程块簇架构，支持以比单个 SM 上的单个线程块更大粒度的局部控制。线程块簇扩展了 CUDA 编程模型，并在 GPU 的物理编程层次结构中添加了另一个层级，现在包括线程、线程块、线程块簇和网格。线程块簇是一组可以确保并发调度到一组 SM 的线程块，旨在支持跨多个 SM 的线程进行高效协作。

H100 中的线程块簇在 GPC 内跨 SM 并发运行。GPC 是硬件层次结构中的一组 SM，它们在物理上始终保持紧密相邻。线程块簇具有如下所述的硬件加速屏障和全新的内存访问协作功能。GPC 中用于 SM 的专用 SM 到 SM 网络可在簇中的线程之间实现快速的数据共享。如图 14 所示，在 CUDA 中可以选择在内核启动时将网格中的线程块分组为簇，并且可以通过 CUDA cooperative_groups API 使用簇功能。



在传统 CUDA 编程模型中，网格由线程块组成，如上图左半部分的 A100 所示。Hopper 架构添加了一个可选的簇层次结构，如上图右半部分所示。

图 14. 线程块簇和包含簇的网格

分布式共享内存

在使用簇的情况下，所有线程都可以通过加载、存储和原子操作直接访问其他 SM 的共享内存。由于共享内存的虚拟地址空间在逻辑上分布在簇中的所有块上，所以此功能又名为“分布式共享内存”(DSMEM)。DSMEM 支持在 SM 之间进行更高效的数据交换，在这种方式下，不再需要通过写入和读取显存来传递数据。用于簇之间的专用 SM 到 SM 网络可确保快速、低延迟地访问远程 DSMEM。与使用显存相比，DSMEM 可将线程块之间的数据交换速度提升约 7 倍。

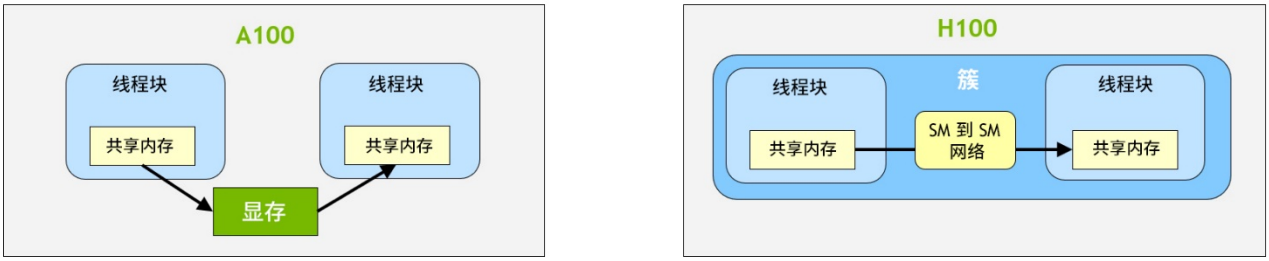


图 15. 线程块到线程块的数据交换（A100 与包含簇的 H100 的对比）

在 CUDA 层级，系统会将簇中所有线程块的所有 DSMEM 段映射到每个线程的通用地址空间，因此用户可以直接使用简单的指针引用所有 DSMEM。CUDA 用户可以利用 cooperative_groups API 构建指向簇中任何线程块的通用指针。DSMEM 传输也可以表示为异步拷贝操作，这些异步拷贝操作可以被基于共享内存的屏障来同步，以追踪其完成情况。

下方的图 16 显示了在不同算法上使用簇的性能优势。程序员可以使用簇来直接控制 GPU 的较大部分而不是单个 SM，从而提高性能。簇允许系统协同执行更多线程，与仅使用单个线程块相比，可以访问更大的共享内存池。

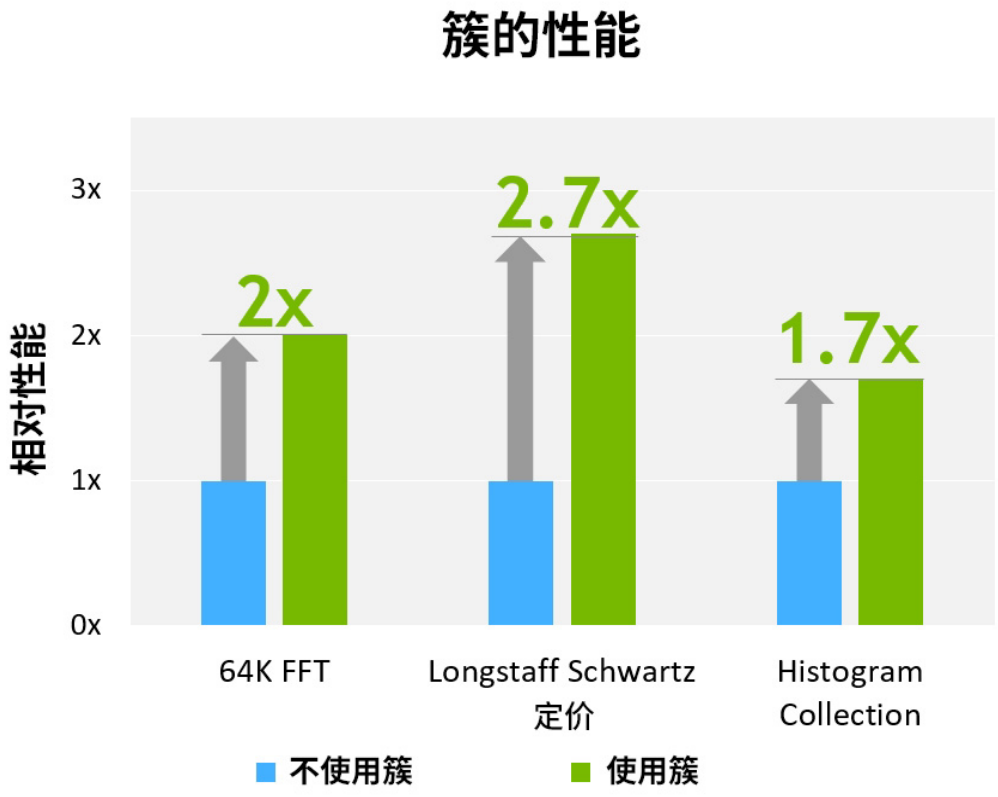


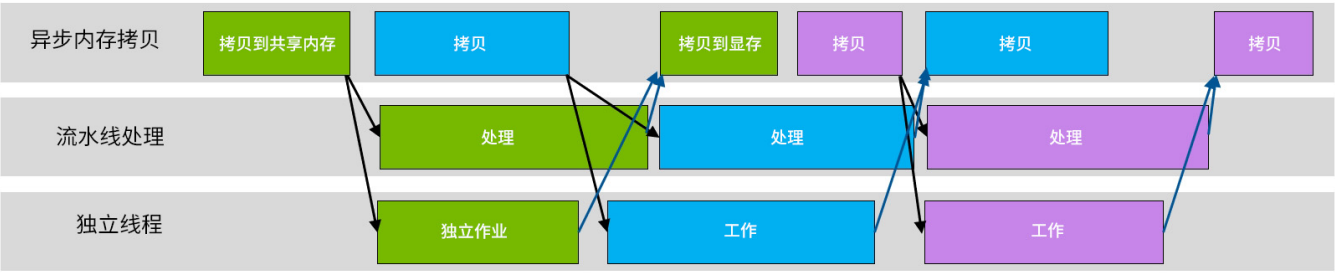
图 16. 使用簇与不使用簇的性能比较

H100 的初步性能评估基于当前预期，交付产品可能会有变化

异步执行

每一代 NVIDIA GPU 都包括多项架构改进，这些改进可提高性能、可编程性、能效、GPU 利用率和许多其他因素。最新几代的 NVIDIA GPU 内置异步执行功能，可支持数据移动、计算和同步之间的更多重叠。Hopper 架构提供了提升异步执行性能的新功能，这允许进一步掩藏内存拷贝与计算以及其他的独立作业，同时还能尽量减少同步点。

下文介绍了称为“Tensor 内存加速器 (TMA)”的新异步存储复制单元和新的异步事务屏障功能。



CUDA 编程模型接口	A100	H100 的新功能
Barrier.arrive(), Barrier.wait()	异步屏障	**异步事务屏障**
	共享内存中的等待程序空转	**等待程序休眠，直到所有线程到达**
Memcpy_async()	直接拷贝到共享内存	**异步内存拷贝单元(称为 TMA)**

数据移动、计算和同步的编程重叠。异步并发和尽量减少同步点是保证性能的关键。

图 17. Hopper 中的异步执行并发和改进

Tensor 内存加速器 (TMA)

为了向功能强大的新 H100 Tensor Core 传输数据，新的 Tensor 内存加速器 (TMA) 提高了数据预取效率，可以在显存和共享显存之间双向传输大量数据块和多维张量。

TMA 操作需通过复制描述符启动，该描述符使用张量维度和块坐标来指定数据传输，而非为每个元素寻址（请参阅下方的图 18）。用户可以指定大量数据块（最高为共享显存容量），并将这些数据块从显存加载到共享内存中，或者从共享内存回存到显存。TMA 支持不同的张量布局（一维到五维张量）、不同的数据访问模式、归约和其他功能，可显著减少寻址开销并提高效率。

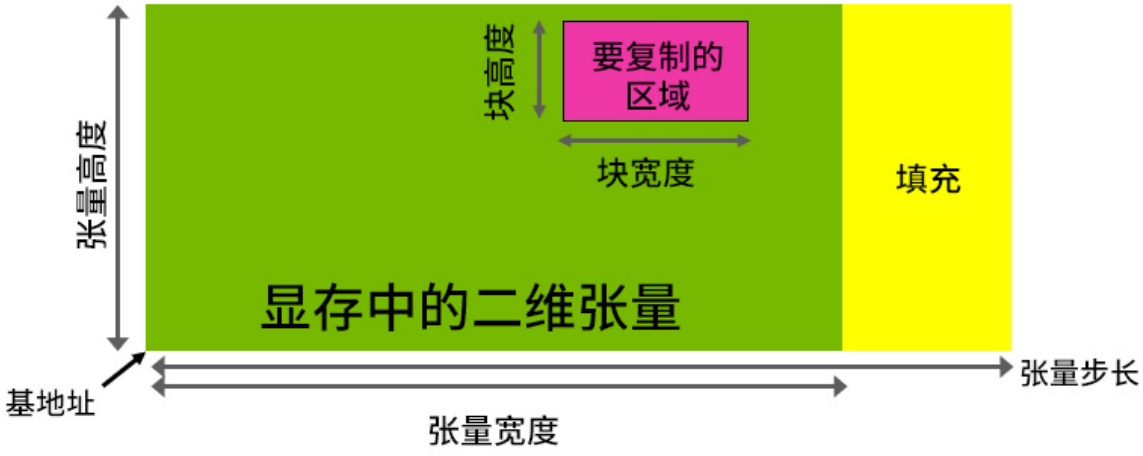


图 18. 通过复制描述符生成 TMA 地址

TMA 操作是异步操作，并且利用了 A100 中引入的基于共享内存的异步屏障。此外，TMA 编程模型是单线程的，系统会在线程束中选择单个线程发出异步 TMA 操作 (cuda::memcpy_async) 以复制张量，随后多个线程可以在 cuda::barrier 上等待数据传输完成。为了进一步提高性能，H100 SM 增加了硬件来加速这些异步屏障等待操作。

TMA 的一个关键优势在于它可以释放线程来执行其他独立的作业。如图 19 的左侧部分所示，在 A100 上，异步显存复制经由特殊的 LoadGlobalStoreShared 指令执行，因此线程负责生成所有地址，并在整个复制区域内循环。

在 Hopper 上，TMA 负责处理一切。在启动 TMA 之前，单个线程会创建一个复制描述符，这之后，地址生成和数据移动均由硬件负责处理。TMA 提供了一个更为简单的编程模型，因为它能够在复制张量段时接管计算步长、偏移和边界计算的任务。

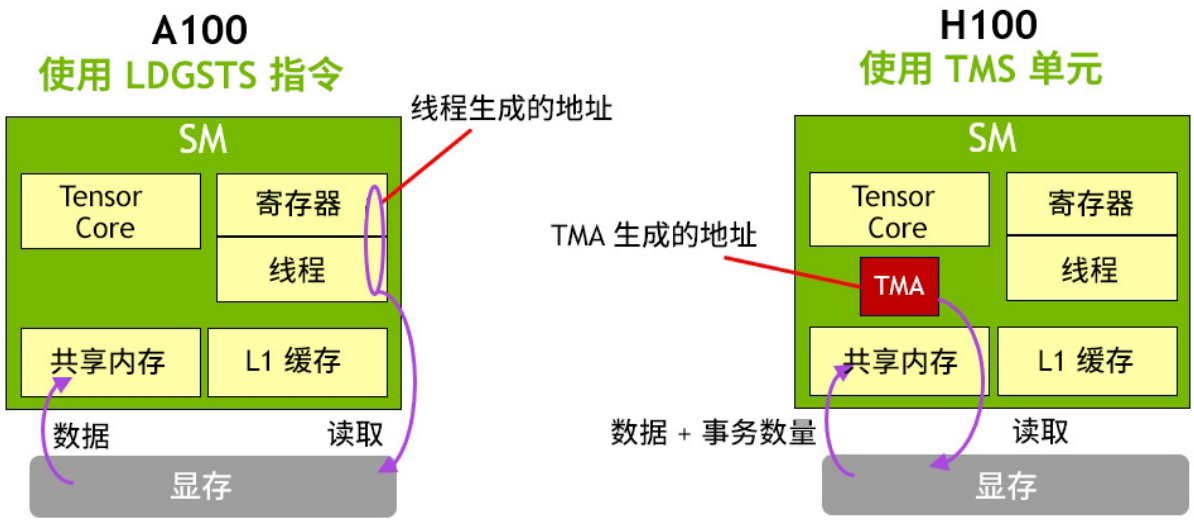


图 19. 在 H100 上使用 TMA 与在 A100 上使用 LDGSTS 进行异步内存复制的对比情况

异步事务屏障

异步屏障最初在 Ampere GPU 架构中引入。请参阅图 20 的左侧部分。举个例子，一组线程正在生成数据，并会在一个屏障之后使用这些数据。异步屏障将同步过程分为两步。首先，线程在生成自己的那部分共享数据后，会发出 “Arrive” 信号。这一信号是非阻塞的，因此线程处于空闲状态，可以执行其他独立的作业。最后，线程需要所有其他线程生成的数据。此时，线程会发出 “Wait” 指令以阻塞线程执行其他作业，直到每个线程都发出 “Arrive” 信号。

异步屏障的优势在于它允许提前到达的线程在等待时执行独立的作业。这种重叠正是额外性能收益的来源。如果所有线程都有足够的独立作业，则屏障实际上变成了“无成本的”，因为所有线程均已到达，所以就无需再使用 “Wait” 指令。

Hopper 的一个全新功能是在所有其他线程到达之前，让“等待”线程休眠。在之前的芯片上，等待线程会在共享内存中的屏障对象上空转。

虽然异步屏障仍然是 Hopper 编程模型的一部分，但 Hopper 增加了称为“异步事务屏障”的新屏障形式。异步事务屏障与异步屏障非常相似。请参阅图 20 的右侧部分。异步事务屏障也是一种拆分屏障，但它不只计算到达的线程数，还计算事务数。Hopper 包含一个用于写入共享内存的新命令，该命令同时传递要写入的数据和事务计数。事务计数本质上是字节计数。异步事务屏障在收到 Wait 命令时会阻塞线程，直到所有生成者线程都执行了 Arrive，并且所有事务计数的总和达到预期值。

异步事务屏障是一种强大的新基元，用于异步内存复制或数据交换。如前所述，簇可以通过隐式同步进行线程块到线程块的数据交换通信，并且簇功能建立在异步事务屏障之上。

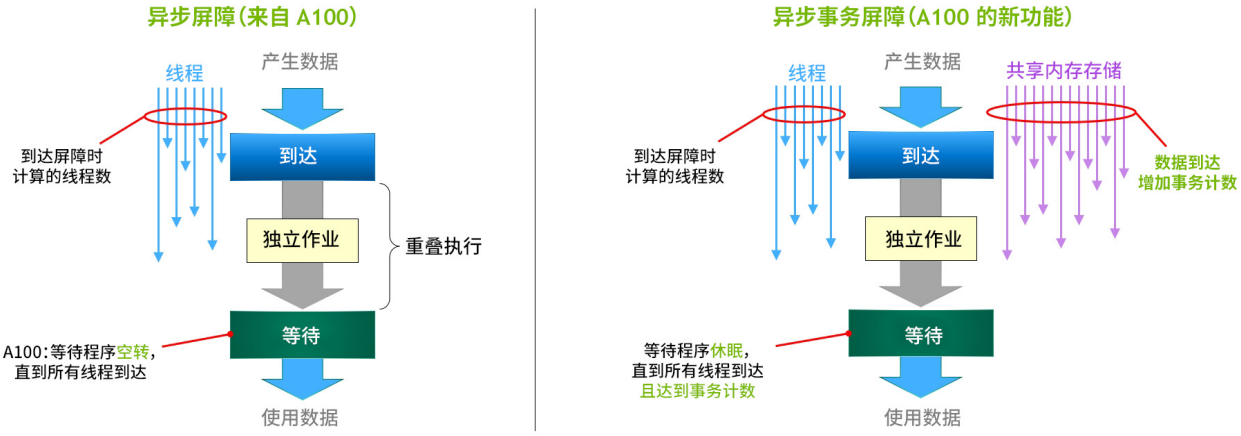


图 20. A100 中的异步屏障与 H100 中的异步事务屏障对比

H100 HBM 和二级缓存存储架构

GPU 存储架构和层次结构的设计对应用性能至关重要，且会影响 GPU 大小、成本、功耗和可编程性。GPU 中存在许多不同的存储子系统，包括大量芯片外 DRAM（帧缓存）设备显存、不同级别和类型的片内内存，以及 SM 计算中使用的寄存器堆。

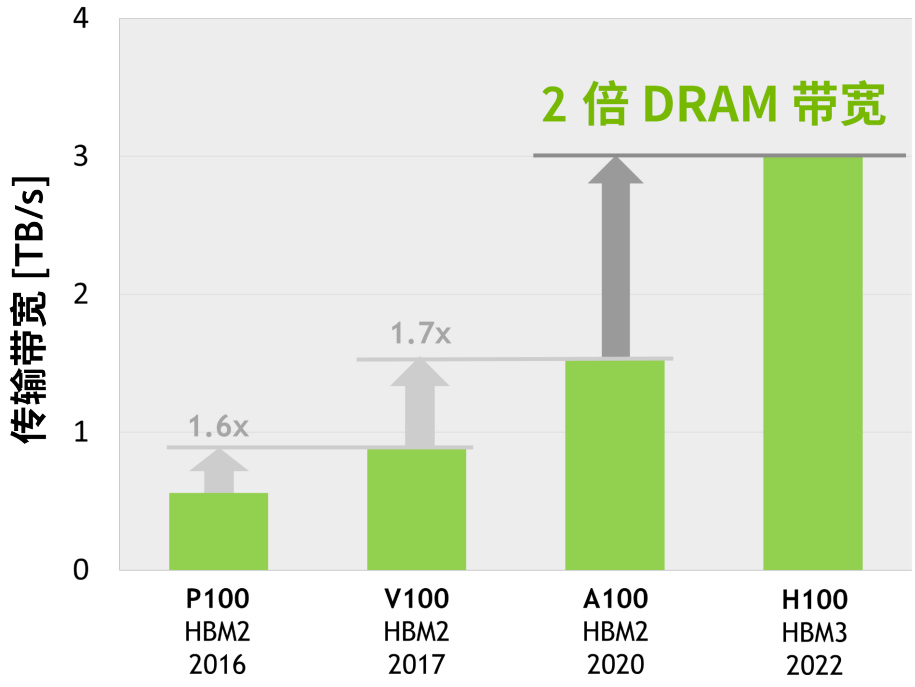
高性能 HBM3 和 HBM2e 分别是 H100 SXM5 和 PCIe H100 GPU 中所用的 DRAM 技术。HBM 显存由显存堆栈组成，与 GPU 位于同一物理封装内，相较于传统的 GDDR5/6 显存设计，可显著节省能耗和占用空间，便于在系统中安装更多 GPU。

CUDA 程序访问的全局显存（Global Memory）和局部显存（Local Memory）位于 HBM 显存空间中，这类显存在 CUDA 中被称为“设备显存”（Device Memory）。常量显存 (Constant Memory) 空间位于设备显存 (Device Memory) 中，并在常量缓存上进行缓存。纹理和表面显存空间位于设备显存中，并在纹理缓存上进行缓存。二级缓存用于缓存从 HBM（设备）显存读取或写入其中的内容，并为从 GPU 内不同子系统中发出的显存请求提供服务。所有 SM 以及 GPU 上运行的所有应用均可访问 HBM 和二级缓存空间。

H100 HBM3 和 HBM2e DRAM 子系统

随着 HPC、AI 和数据分析数据集规模的不断增加，计算问题的复杂程度也日益加深，而这必然也需要更高的 GPU 显存容量和带宽。NVIDIA P100 是全球首款支持高带宽 HBM2 显存技术的 GPU 架构，而 NVIDIA V100 则提供了更快速、更高效且更高容量的 HBM2 实现。NVIDIA A100 GPU 进一步提升了 HBM2 的性能和容量。

H100 SXM5 GPU 支持 80 GB（5 个堆栈）的快速 HBM3 显存，可提供超过 3 TB/s 的显存带宽，相较于两年前推出的 A100 显存带宽，实际上提高了 2 倍，因此大大抬高了标准。PCIe H100 具备 80 GB 的快速 HBM2e，且内存带宽也在 2 TB/s 以上。



显存数据速率尚未最终确定，可能会在最终产品中发生变化。

图 21. 带宽提升 2 倍的全球首款 HBM3 GPU 显存架构

H100 二级缓存

H100 中的 50 MB 二级缓存比 A100 的 40 MB 二级缓存大 1.25 倍。它能够缓存更多部分的模型和数据集以供重复访问，从而减少 HBM3 或 HBM2e DRAM 的访问次数并提高性能。通过使用分区交叉结构，二级缓存可定位和缓存数据，以便在直接连接到分区的 GPC 的 SM 中访问显存。二级缓存驻留控制可优化容量利用率，支持编程者有选择地管理应保留在缓存中和应被逐出的数据。

HBM3 或 HBM2e DRAM 和二级缓存子系统均支持数据压缩和解压缩技术，能够优化显存和缓存的使用情况和性能。

显存子系统 RAS 功能

已针对 H100 中的 HBM3 和 HBM2e 显存子系统实现以下两个主要 RAS（可靠性、可用性和可服务性）功能。

ECC 显存抗误码技术

H100 HBM3/2e 显存子系统支持通过单错纠正/双错检测 (SECDED) 纠错码 (ECC) 来保护数据。纠错码可为易受数据损坏影响的计算应用程序提供更高可靠性。这在大型集群计算环境中尤为重要，因为其中的 GPU 需处理非常大的数据集亦或长时间运行应用程序。H100 支持 HBM3/2e 显存的 “Sideband ECC”（边带纠错码），其中一个与主 HBM 显存分开的小型显存区域用于存储纠错码位元，这与 “Inline ECC”（内联纠错码）相反，后者会在主显存中开拓一部分空间来存储纠错码位元。H100 中的其他关键显存结构也受 SECDED 纠错码的保护，包括二级缓存和一级缓存，以及所有 SM 中的寄存器文件。

显存行重映射

H100 HBM3/HBM2e 子系统可以让显存单元生成纠错码错误的显存行失效，并在引导时通过行重映射逻辑将这些行替换为预留的正确数据行。每个 HBM3/HBM2e 内存库 (Memory Bank) 中存在预留为备用行的多个显存行，可在需要时激活以替换确定为损坏的行。

表 3. NVIDIA A100 和 H100[1] 数据中心 GPU 对比

GPU 特征	NVIDIA A100	NVIDIA H100 SXM5[1]	NVIDIA H100 PCIe[1]
GPU 架构	NVIDIA Ampere	NVIDIA Hopper	NVIDIA Hopper
GPU 主板外形规格	SXM4	SXM5	PCIe 5.0
SM 数量	108	132	114
TPC 数量	54	66	57
FP32 Core 核心数 / SM	64	128	128
FP32 Core 核心数 / GPU	6912	16896	14592
FP64 Core 核心数 / SM（不包括 Tensor）	32	64	64
FP64 Core 核心数 / GPU（不包括 Tensor）	3456	8448	7296
INT32 Core 核心数 / SM	64	64	64
INT32 Core 核心数 / GPU	6912	8448	7296
Tensor Core 核心数 / SM	4	4	4
Tensor Core 核心数 / GPU	432	528	456
GPU 加速频率（H100 未最终确定）[3]	1410 MHz	尚未最终确定	尚未最终确定
使用 FP16 累加的 FP8 Tensor TFLOPS 峰值 [1]	NA	2000/4000[2]	1600/3200[2]
使用 FP32 累加的 FP8 Tensor TFLOPS 峰值 [1]	NA	2000/4000[2]	1600/3200[2]
使用 FP16 累加的 FP16 Tensor TFLOPS 峰值 [1]	312/624[2]	1000/2000[2]	800/1600[2]
使用 FP32 累加的 FP16 Tensor TFLOPS 峰值 [1]	312/624[2]	1000/2000[2]	800/1600[2]
使用 FP32 累加的 BF16 Tensor TFLOPS 峰值 [1]	312/624[2]	1000/2000[2]	800/1600[2]
TF32 Tensor TFLOPS 峰值 [1]	156/312[2]	500/1000[2]	400/800[2]

FP64 Tensor TFLOPS 峰值 [1]	19.5	60	48
INT8 Tensor TOPS 峰值 [1]	624/1248[2]	2000/4000[2]	1600/3200[2]
FP16 TFLOPS 峰值（非 Tensor）[1]	78	120	96
BF16 TFLOPS 峰值（非 Tensor）[1]	39	120	96
FP32 TFLOPS 峰值（非 Tensor）[1]	19.5	60	48
FP64 TFLOPS 峰值（非 Tensor）[1]	9.7	30	24
INT32 TOPS 峰值 [1]	19.5	30	24
纹理单元数	432	528	456
显存位宽	5120 位 HBM2	5120 位 HBM3	5120 位 HBM2e
显存容量	40 GB	80 GB	80 GB
显存数据速率 [1]	1215 MHz DDR	尚未最终确定	尚未最终确定
显存带宽（H100 未最终确定）[1]	1555 GB/s	3000 GB/s	2000 GB/s
二级缓存大小	40 MB	50 MB	50 MB
共享显存容量 / SM	最大可配置为 164 KB	最大可配置为 228 KB	最大可配置为 228 KB
寄存器堆大小 / SM	256 KB	256 KB	256 KB
寄存器堆大小 / GPU	27648 KB	33792 KB	29184 KB
TDP[1]	400 瓦	700 瓦	350 瓦
晶体管数量	542 亿	800 亿	800 亿
GPU 芯片尺寸	826 mm2	814 mm2	814 mm2
TSMC 制程	7nm N7	专为 NVIDIA 定制的 4N	专为 NVIDIA 定制的 4N

1. **H100 的初步规格均为预期结果，实际交付的产品可能会有所不同**
2. **使用稀疏功能实现有效的 TOPS / TFLOPS**
3. **GPU 峰值频率和 GPU 加速频率对于 NVIDIA 数据中心 GPU 来说含义相同**

注意：由于 H100 和 A100 Tensor Core GPU 被设计为安装在高性能服务器和数据中心机架中使用，以助力 AI 和高性能计算工作，因此不包括显示器接口、用于光线追踪加速的 NVIDIA RT Core 或 NVENC 编码器。

计算能力

H100 GPU 支持全新的计算能力 9.0。表 4 比较了 不同 NVIDIA GPU 架构之间的计算能力参数。

表 4. 计算能力：V100、A100 与 H100

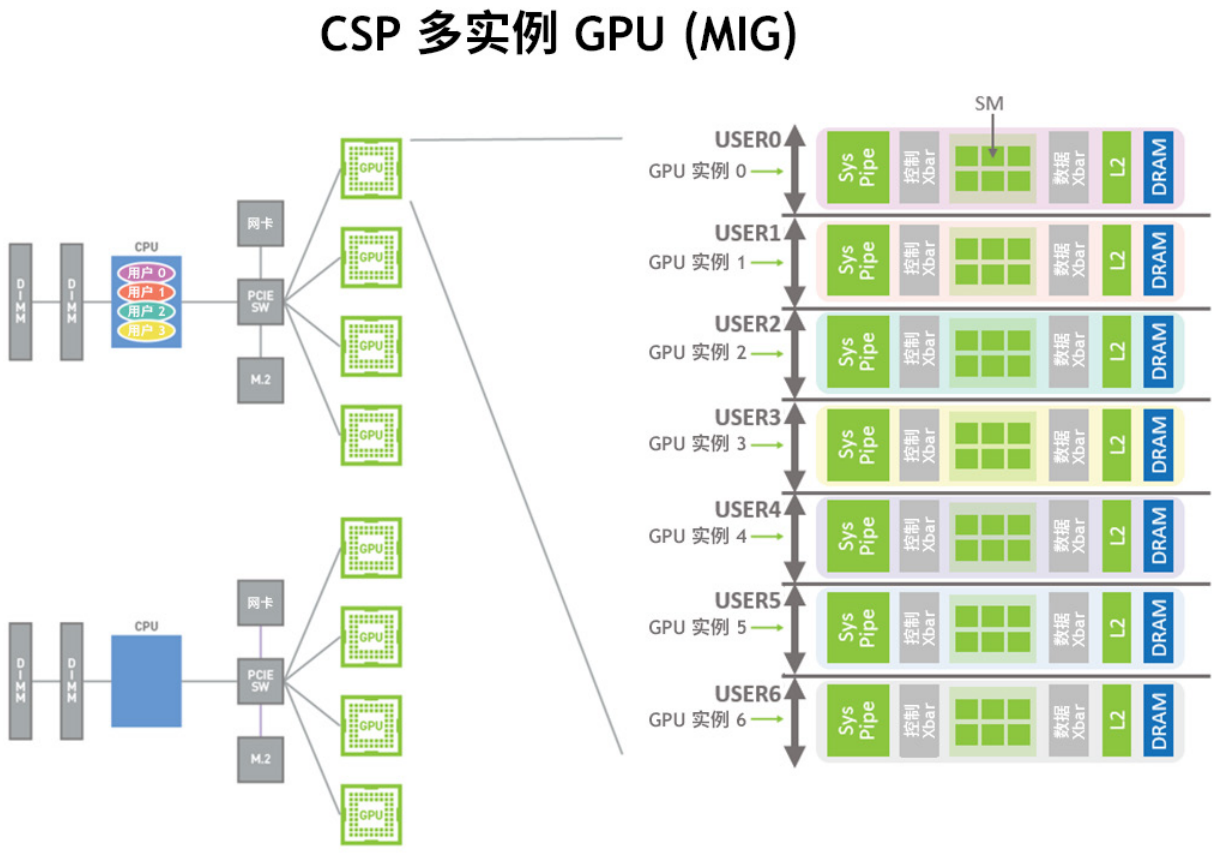
数据中心 GPU	NVIDIA Tesla V100	NVIDIA A100	NVIDIA H100
GPU 架构	NVIDIA Volta	NVIDIA Ampere	NVIDIA Hopper
计算能力	7.0	8.0	9.0
线程 / 线程束	32	32	32
最大线程束数 / SM	64	64	64
最大线程数 / SM	2048	2048	2048
最大线程块 (CTA) / SM	32	32	32
最大线程块 / 线程块簇	NA	NA	16
最大 32 位寄存器数 / SM	65536	65536	65536
最大寄存器数 / 线程块 (CTA)	65536	65536	65536
最大寄存器数 / 线程	255	255	255
最大线程块大小（线程数）	1024	1024	1024
FP32 Core 核心数 / SM	64	64	128
SM 中寄存器数与 FP32 核心数的比率	1024	1024	512
共享显存大小 / SM	最大可配置为 96 KB	最大可配置为 164 KB	最大可配置为 228 KB

第二代可靠 MIG

在基于 NVIDIA Ampere 架构的 A100 GPU 中，NVIDIA 引入了 GPU 多实例技术 (MIG)。MIG 技术为共享同一 GPU 硬件的多个用户，提供相互独立、完全隔离且安全的 GPU 实例，这已成为扩展云服务提供商 (CSP) 数据中心的一项极其重要的功能。

MIG 技术回顾

MIG 技术支持将每个 A100 或 H100 GPU（H100 SXM5 和 H100 PCIe 版本）划分为最多 7 个 GPU 实例，以优化 GPU 利用率，并且 MIG 技术为不同的客户端（例如虚拟机、容器和进程）之间提供定义明确的 QoS 和隔离机制。MIG 对拥有多租户用例的云服务提供商尤其有价值，该技术可确保一个客户端不会影响其他客户端的工作或调度，以此为客户提供更强的安全性并让 GPU 利用率得到保证。



此 CSP MIG 示意图展示了如何在单个物理 GPU 中为来自相同或不同组织的多个独立用户分配专用的、受保护且相互隔离的 GPU 实例。

图 22. CSP MIG 配置示例

MIG 技术的一个特性是能对 vGPU（虚拟 GPU）虚拟机 (VM) 配置进行管理、调整、维护和负载均衡。这一重要特性，使得 vGPU 可以在单个 GPU 上的不同 GPU 实例之间迁移，以及在集群中不同 GPU 之间迁移，后者在实际情况中更频繁发生。

每个 GPU 实例都有跨越整个内存系统的独立且隔离路径：片内交叉开关端口、二级缓存库、显存控制器和 DRAM 地址总线，上述所有资源都会单独地分配给每个实例。这确保单个用户的计算任务能在可预测的吞吐量和延迟下运行，即使其他任务造成其自身缓存抖动或其 DRAM 接口饱和情况下，也能实现相同的二级缓存分配和 DRAM 带宽分配。

（有关基本 MIG 技术的更多详细信息，请参阅 NVIDIA A100 Tensor Core GPU 白皮书。）

H100 MIG 增强功能

与 A100 相比，H100 中的全新第二代 MIG 技术为每个 GPU 实例提供了约 4 倍的计算能力和近 3 倍的显存带宽。NVIDIA Hopper 架构提供了充分安全、支持云原生多租户、多用户的 MIG 配置，这些特性又一步增强了 MIG 技术。通过硬件层和虚拟化层的加密计算功能，MIG 技术可以将最多 7 个 GPU 实例彼此安全隔离（请参阅**后文安全增强和机密计算**部分，了解有关机密计算的更多详细信息）。

图 23 展示了 CPU 和 GPU 如何协同为共享单个 GPU 的多名用户提供多个可信执行环境 (TEE) 的**系统配置示例**。CPU 提供了多个带有 NVIDIA 安全驱动的可信安全 VM。此示例中的 H100 GPU 分为 4 个安全 MIG 实例。CPU 和 GPU 之间会进行加密传输。GPU 硬件虚拟化由 PCIe SR-IOV 负责实现，其中每个 MIG 实例拥有一个虚拟函数 (VF)。机密性和数据完整性基于多个硬件的安全功能实现，显存的隔离通过硬件的防火墙机制实现。

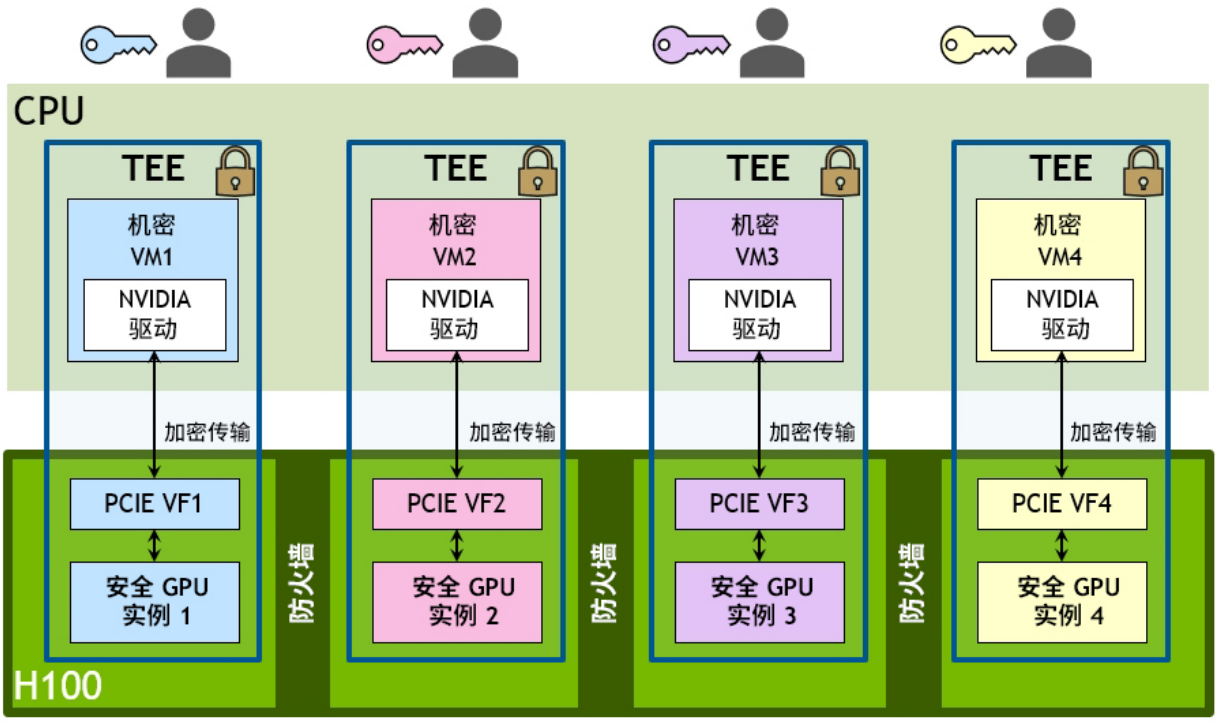


图 23. 多租户单 GPU 配置中的安全 MIG 示例

Hopper 架构现在还允许为每个 GPU 实例提供专用的图像和视频解码器，以在共享基础架构上实现安全、高吞吐量的智能视频分析 (IVA) 系统。每个 MIG 实例至少可接收一个 NVDEC 和 NVJPG 单元。

此外，H100 MIG 实例现在还拥有各自的性能监控功能，这些性能监控功能可与 NVIDIA 开发者工具配合使用。借助 Hopper 架构的并发分析，管理员可以监控合理分配后 GPU 的加速情况，同时可以无缝地优化多个用户之间的资源分配方式。

Transformer 引擎

Transformer 模型是今天被广泛运用于语言模型的骨架网络，例如 BERT 和 GPT-3，同时 Transformer 模型需要庞大的计算资源。Transformer 最初是针对自然语言处理 (NLP) 而开发的模型，但是如今已经被越来越多地应用于计算机视觉、药物研发等多种领域。这一类的模型，现在其规模呈指数级增长，已包含数万亿个参数，导致训练时间长达数月，由于需要的计算量过于庞大，在实际场景中是不切实际的。例如，Megatron Turing NLG (MT-NLG) 模型需要 2048 个 NVIDIA A100 GPU 运行 8 周才能完成训练。整体看来，在过去五年中，Transformer 模型的计算量增长比大多数其他 AI 模型要快得多，每两年即要增长 275 倍（参见图 24）。

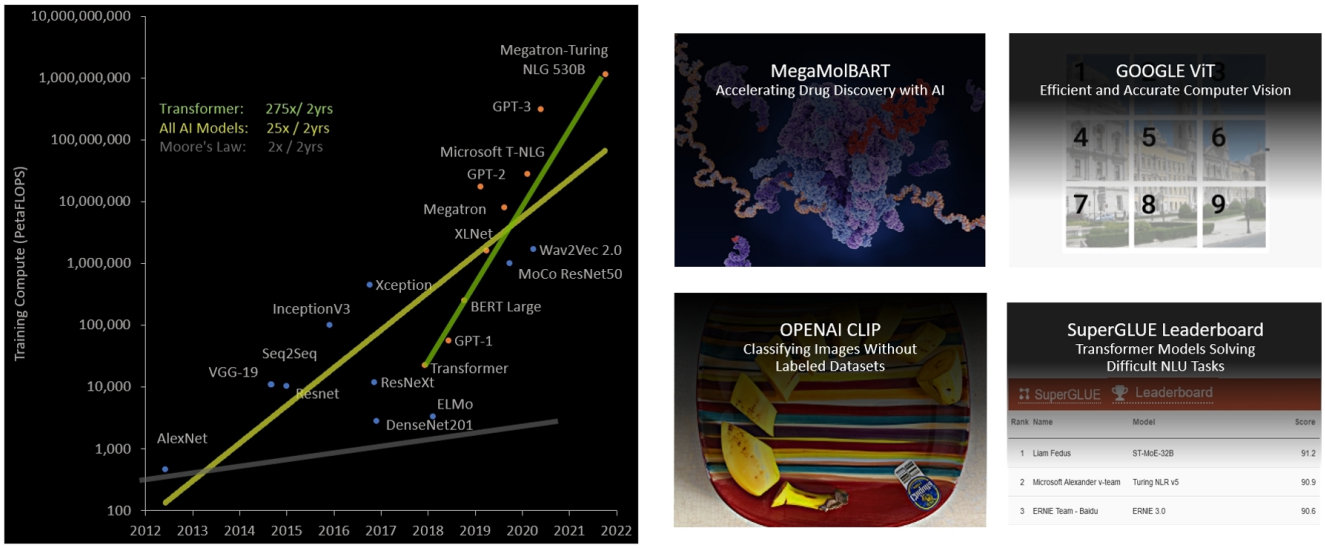


图 24. Transformer 模型大小随不同用例呈指数级增长

H100 包含一个新的 **Transformer 引擎**，这是一种定制 Hopper Tensor Core 技术，可显著加速 Transformer 模型的计算。

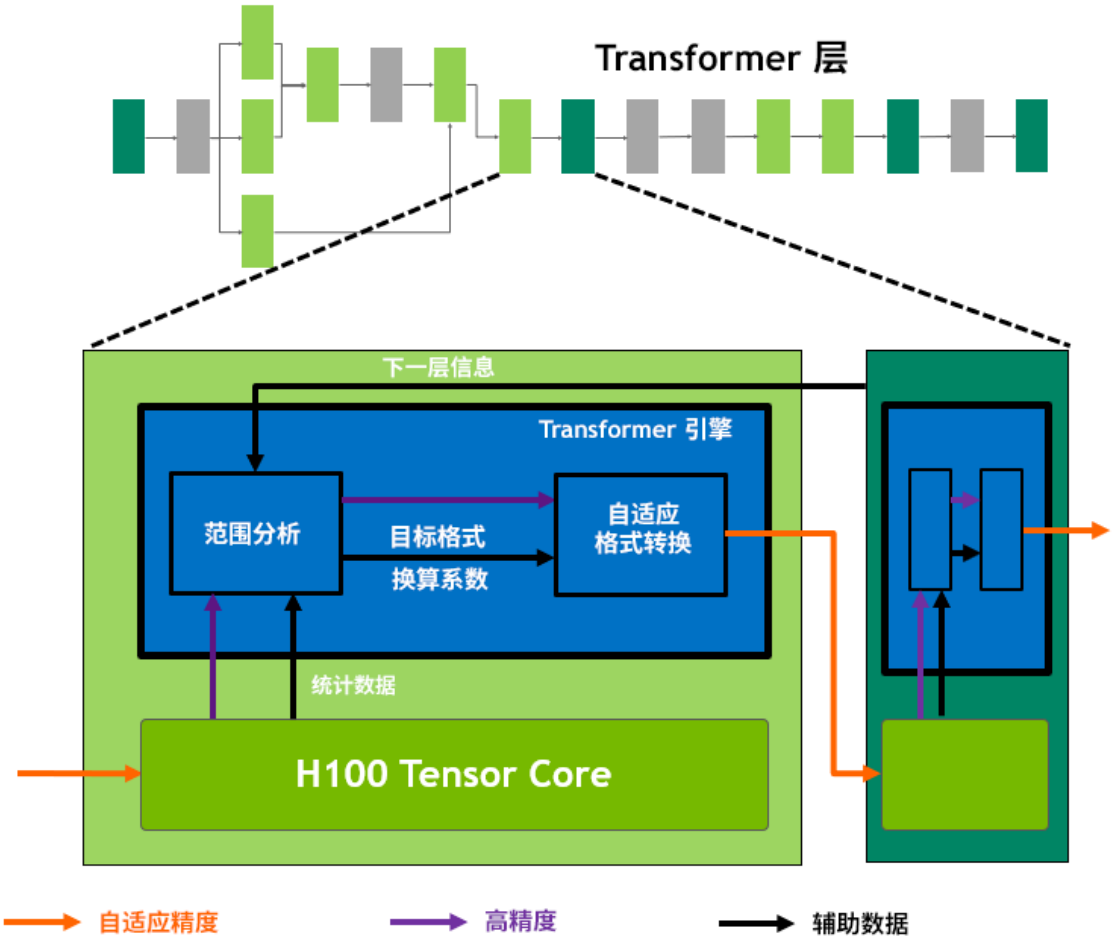


图 25. Transformer 引擎的运行概念

混合精度的目标是智能地管理精度以保持计算准确性，同时从更小、更快的数值精度中提升计算性能。针对 Transformer 模型的每一层参数，Transformer 引擎都会分析由 Tensor Core 所生成的输出值的统计分布。通过知道下一层神经网络的类型和精度，Transformer 引擎决定在张量存储到显存之前将其转换为哪种数值格式。与其他数值格式相比，FP8 的数值范围是相对有限的。为了最优地使用可用范围，Transformer 引擎还使用从张量统计数据中计算出的缩放因子，将张量的数据动态地扩展到可表示的范围。这样一来，每个层将以其需要的精度进行计算，并且以一个最优的策略进行加速。

第四代 NVLink 和 NVLink 网络

新兴的百亿亿次级 HPC 和万亿级参数的 AI 模型，比如超越人类的对话式 AI 模型，需要数月时间才能完成训练，即使采用超级计算机也是如此。为了将训练时间从几个月缩短到几天，从而对企业发挥更大效用，需要在服务器集群中的每个 GPU 之间建立高速、无缝的通信。PCIe 由于其带宽有限，在这方面会产生瓶颈。为了构建功能强大的端到端计算平台，我们需要速度更快、扩展性更强的 NVLink 互连技术。

NVLink 是 NVIDIA 的高带宽、高能效、低延迟、无损的 GPU 到 GPU 互连技术，其中包含诸如链路级错误检测和数据包回放机制等弹性特性，可保证数据的成功传输。H100 GPU 中集成了全新第四代 NVLink 技术 ，它的通信带宽是 NVIDIA A100 GPU 上所用第三代 NVLink 的 1.5 倍。

新 NVLink 技术中多 GPU 的 IO 和共享显存的访问总带宽可达 900 GB/s，是第五代 PCIe 技术的 7 倍。在 A100 GPU 中的第三代 NVLink ，在每个方向上使用四个差分对（4 通道）来创建单条链路，以提供 25 GB/s 的有效带宽，而在第四代 NVLink 中， 在每个方向上仅采用两个高速差分对即可形成单条链路，同样能够在每个方向提供 25 GB/s 的有效带宽。H100 包含 18 条第四代 NVLink 链路，提供 900 GB/s 的总带宽，而 A100 包含 12 条第三代 NVLink 链路，提供 600 GB/s 的总带宽。

除了第四代 NVLink 以外，H100 还引入了新的 NVLink 网络互连技术，这是 NVLink 的一种可扩展版本，可在最多 256 个 GPU 之间实现跨计算节点的 GPU 到 GPU 通信。

常规 NVLink 中的所有 GPU 共享一个通用地址空间，且请求直接使用 GPU 物理地址进行路由，而 NVLink 网络与此不同，它引入了一个由 H100 中的新地址转换硬件提供的新网络地址空间，以将所有 GPU 的地址空间相互隔离并与网络地址空间隔离。这使 NVLink 网络能够安全扩展至更多 GPU。

由于 NVLink 网络端点不共享通用显存地址空间，因此整个系统中不会自动建立 NVLink 网络连接。因此，与其他网络接口（如 InfiniBand）类似，用户软件应根据需要在端点之间显式地建立连接。

第三代 NVSwitch

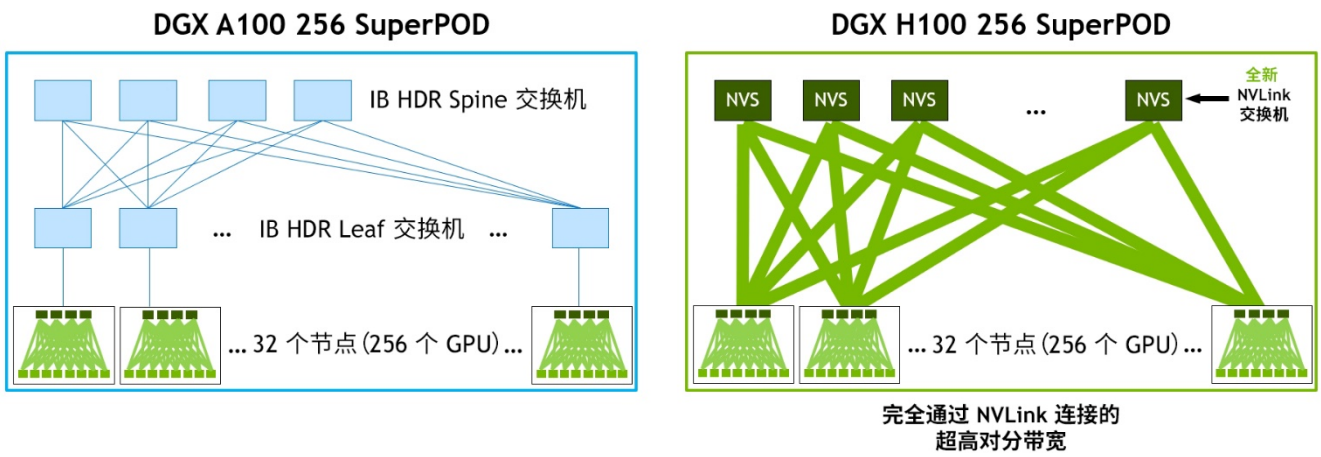
新的第三代 NVSwitch 技术同时包括位于节点内和节点外的交换机，可连接服务器、集群或者数据中心环境中的多个 GPU。节点内的全新第三代 NVSwitch 可支持 64 个第四代 NVLink 连接端口，加速多 GPU 的连接。交换机总吞吐量从上一代的 7.2 Tbits/s 增加到 13.6 Tbits/s。

全新第三代 NVSwitch 还通过组播和 NVIDIA SHARP 在网计算，实现了针对集合运算的硬件加速。加速的集合运算包括写入广播 (all_gather)、reduce_scatter 和广播原子。与在 A100 上使用 NCCL（NVIDIA 集合通信库）相比，网络内组播和归约可令吞吐量获得高达 2 倍的提升，同时显著缩短小型块集合的延迟。NVSwitch 集合加速可显著减少集合通信的 SM 负载。

全新 NVLink Switch 系统

通过结合全新 NVLINK 网络技术和第三代 NVSwitch 技术，使得 NVIDIA 有能力去构建大型 NVLink Switch 系统网络，实现前所未有的通信带宽水平。每个 GPU 节点均显示节点中所有 GPU NVLink 带宽呈 2:1 锥形递减。这些节点通过 NVLink Switch 系统中包含的第二层 NVSwitch 相互连接，这些模块位于计算节点外部，并将多个节点连接在一起。

NVLink Switch 系统最多可支持 256 个 GPU。互连节点能够提供 57.6 TB 的多对多带宽，并且可提供难以置信的、高达 1 exaFLOP 级别的 FP8 稀疏计算算力。图 26 表明了，分别基于 A100 和 H100 的 32 个节点、256 个 GPU 的 DGX SuperPOD 的对比情况。请注意，基于 H100 的 SuperPOD 可选用最新的 NVLink Switch 实现 DGX 节点间的互连。



	A100 SuperPod			H100 SuperPod			加速	
	密集 PFLOP/s	对分 [GB/s]	归约 [GB/s]	密集 PFLOP/s	对分 [GB/s]	归约 [GB/s]	对分	归约
1 个 DGX 包含 8 个 GPU	2.5	2400	150	16	3600	450	**1.5x**	**3x**
32 个 DGX 包含 256 个 GPU	80	6400	100	512	57600	450	**9x**	**4.5x**

DGX H100 SuperPOD 通过基于第三代 NVSwitch 技术，在新的 NVLink Switch 系统上可实现最多覆盖 256 个 GPU 的 NVLink 全连接。2:1 锥形胖树拓扑结构中的 NVLink 网络互连可使对分带宽实现令人惊叹的 9 倍提升，例如，用于多对多交换，同时实现比一代 InfiniBand 系统高 4.5 倍的 allreduce 吞吐量。DGX H100 SuperPOD 可选用 NVLINK Switch 系统。

图 26. 基于 DGX A100 与 DGX H100 的 32 节点、256 GPU NVIDIA SuperPOD 对比

交换机到交换机的最大线缆长度从 5 米增加到 20 米。现支持由 NVIDIA 打造的 OSFP（八通道小卡可插拔）LinkX 线缆。其每个 OSFP 均配备四端口光学收发器，同时拥有 8 个 100G PAM4 信号通道。通过四端口 OSFP 收发器的创新，使得拥有单个 RU、32 端口的 NVLink Switch 总共可支持 128 个 NVLink 端口，每个端口传输数据的速度为 25GB/s。

PCIe 5.0

H100 采用 PCI Express 5.0 x16 的通道接口，可提供 128 GB/s 的总带宽（每个方向 64 GB/s），相较而言，A100 的 PCIe 4.0 的总带宽为 64 GB/s（每个方向 32 GB/s）。

H100 可借助 PCIe 5.0 接口与性能超强的 x86 CPU 和 SmartNIC / DPU（数据处理器）进行交互。H100 可以与 NVIDIA BlueField-3 DPU 实现最优连接，并可以借助 400 Gb/s 以太网或 NDR（下一代数据速率）400Gb/s InfiniBand 网络加速，来提升安全的高性能计算和 AI 计算负载。

H100 新增了对原生 PCIe 原子操作（如为 32 位和 64 位数据类型添加 atomic CAS、原子交换和原子访存）的支持，从而可加速 CPU 和 GPU 之间的同步和原子操作。H100 还支持单根输入/输出虚拟化 (SR-IOV)，支持多个进程或虚拟机 (VM) 共享和虚拟化单个 PCIe 连接的 GPU。H100 还支持单个 SR-IOV PCIe 连接的 GPU 中的虚拟函数 (VF) 或物理函数 (PF) 通过 NVLink 访问同等 GPU。

安全增强和机密计算

NVIDIA 向安全敏感型市场出售的 GPU 正在与日俱增。云服务提供商 (CSP)、汽车制造商、国家实验室、医疗健康、金融以及许多其他行业和组织都需要高度的安全性。每一代新的 NVIDIA GPU 都会持续改进安全性。

每天会生成、存储和处理大量敏感数据，这会增加监管工作量和网络攻击业务风险。虽然有先进的加密技术来保护存储的静态数据以及在网络中传输的数据，但如今在数据的处理或使用过程中，对于数据的保护仍存在较大漏洞。新的机密计算技术可保护使用中的数据和应用，从而为管理敏感和受监管数据的企业组织提供进一步的安全。

NVIDIA H100 包含多项安全功能来限制对 GPU 内容的访问，确保只有授权实体才能访问 GPU 内容，提供安全引导和验证功能，并在系统运行时主动监控攻击。此外，专门的片内安全处理器、多种类型和级别的加密支持、硬件保护的内存区域、授权访问控制寄存器、裸片上传感器和许多其他功能可为我们的客户及其数据实现安全的 GPU 处理。

H100 是全球首款配备机密计算功能的 GPU。用户可以在获取 H100 GPU 出色加速功能的同时，保护其使用中的数据和应用的机密性和完整性。H100 可提供一系列其他安全功能，以做到保护用户数据，抵御软硬件攻击，并在虚拟化和 MIG 环境中更好地隔离 VM，保护 VM 免受攻击。

NVIDIA H100 GPU 全面的安全功能的主要目标包括：

- **数据保护和隔离**：防止未经授权的实体访问其他用户的数据，其中实体可以是用户、操作系统、服务器虚拟化平台或 GPU 固件。
- **内容保护**：防止未经授权的实体访问由 GPU 存储或处理中的受保护内容。
- **物理损坏保护**：防止对 GPU 造成物理损坏，无论是恶意行为者造成，还是偶然造成，均可预防。

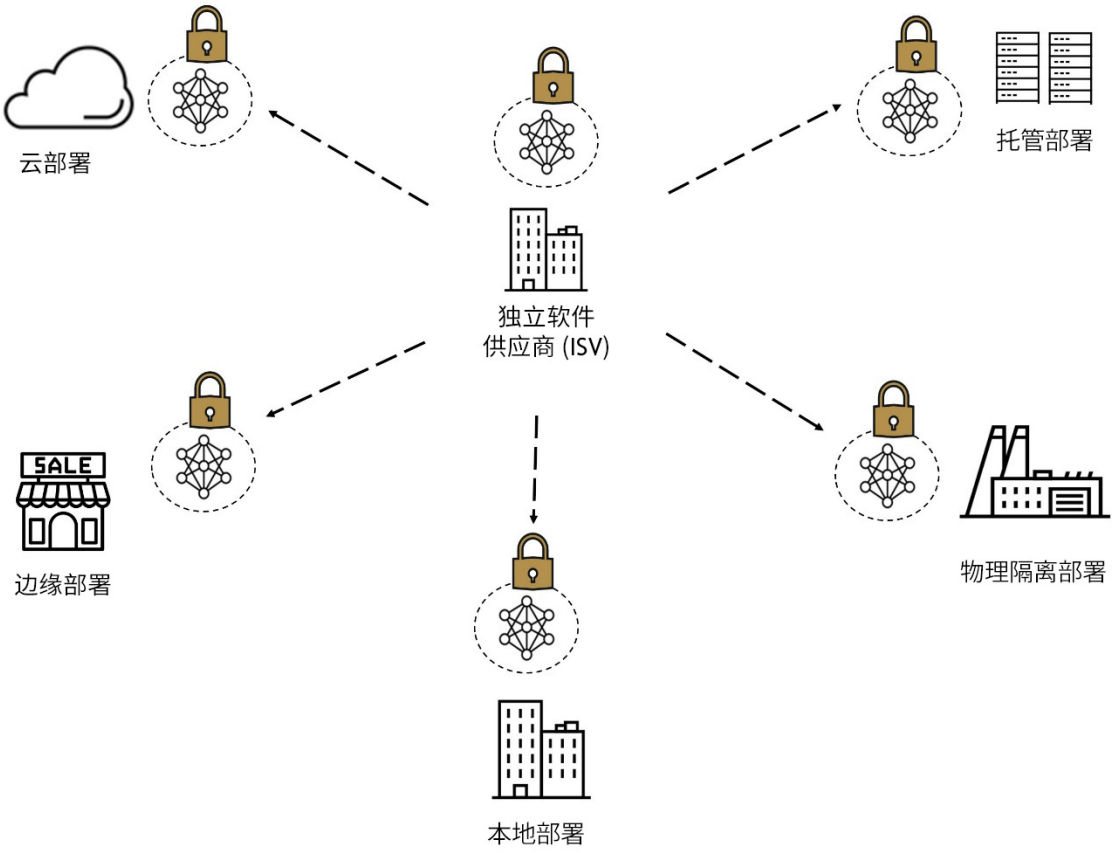
NVIDIA 机密计算

NVIDIA 是机密计算联盟 (C3) 的成员，C3 是由供应商、学术机构、开源项目和软件开发者组成的国际合作联盟，他们共同开发出一系列措施和技术，旨在减少安全威胁，保护在公有云服务、内部数据中心、边缘系统和设备中使用中的敏感数据和应用。

“机密计算”一词的正式定义为“通过在基于硬件的可信执行环境 (TEE) 中执行计算来保护使用中的数据”。该定义与数据的使用位置无关，无论是在云端、最终用户设备或其之间的任一位置使用，均不会影响其定义。此外，这一定义与保护数据所用的处理器或采用的保护技术也无关。

C3 将 TEE 定义为“对三个关键属性（数据保密性、数据完整性和代码完整性）提供一定程度保证的环境”。

如今，数据通常在存储状态和网络传输的过程中受到保护，但在使用中不受操作系统 / 服务器虚拟化平台的保护。这就需要信任操作系统 / 服务器虚拟化平台，而这也因此在保护用户数据和代码方面留下了很大的漏洞。此外，对使用中的数据和代码进行保护的能力在传统计算基础设施中非常有限。对于处理敏感数据——例如个人身份信息（PII）、财务和健康数据——或需要满足数据本地化法规要求的企业组织，他们需要规避针对其应用、模型和数据各个阶段的保密性和完整性方面的威胁。



机密计算可保护在云端、内部部署和边缘上 ISV 客户数据和经过训练的 AI 模型的机密性

图 27. 机密计算可保护多个 ISV 场景

现有的机密计算解决方案均基于 CPU 构建而成，对于 AI 和 HPC 等计算密集型工作负载来说，这些解决方案速度太慢。基于 CPU 的机密计算通常会降低系统性能，这可能会影响工作效率或无法在延迟敏感型数据处理工作负载中加以实现。

借助 NVIDIA Hopper 架构中引入的全新安全功能 NVIDIA 机密计算，H100 成为率先为使用中数据和代码的机密性和完整性提供保护的 GPU。H100 将加速计算引入机密计算领域，并将 CPU 的可信执行环境扩展到 GPU。H100 使得许多过去无法使用共享基础架构（云、托管、边缘）的用例成为可能，这些用例需要保护使用中的数据和代码，以前的机密计算解决方案在性能和灵活性方面都有所欠缺。

NVIDIA 机密计算创建了基于硬件的可信执行环境 (TEE)，用于保护并隔离在单个 H100 GPU、节点内多个 H100 GPU 或单个安全的多实例 GPU (MIG) 实例上运行的整个工作负载。可信执行环境 (TEE) 在 GPU 上的机密 VM 与 CPU 中的机密 VM 之间建立起安全通道。TEE 提供两种操作模式。

1. 将整个 GPU 分配给单个 VM（单个 VM 可能同时获配多个 GPU）。
2. 对 NVIDIA H100 GPU 进行分区，并借助 MIG 技术支持多个 VM，从而实现多租户机密计算。GPU 加速应用可以保持不变运行在 TEE 内，且不必对其进行手动分区。

用户可以将适用于 AI 和 HPC 的 NVIDIA 软件的强大功能与 NVIDIA 机密计算提供的硬件信任根的安全性相结合，从而在最底层的 GPU 架构层级提供安全性和数据保护。用户可以在共享或远程基础设施上运行和验证应用，以确保任何未经授权的实体（包括服务器虚拟化平台、主机操作系统、系统管理员、基础架构所有者或任何具有物理访问权限的人员）在 TEE 内使用应用代码和数据时，均不能对其进行查看或修改。

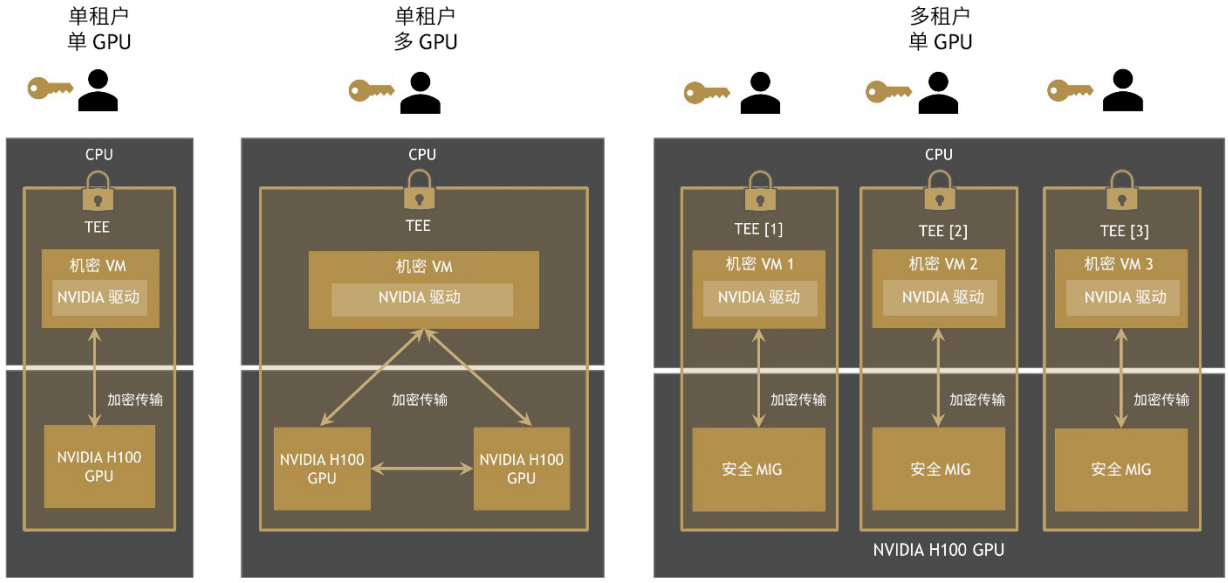


图 28. 面向不同用例的机密计算

Hopper 架构的机密计算功能可进一步提高并加速联合学习等多方协作计算用例的安全性。联合学习支持多家企业组织协同训练或评估 AI 模型，且无需共享每个组的专有数据集。使用 H100 进行的机密联合学习可确保在每个参与站点保护数据和 AI 模型免遭未经授权的访问，避免外部或内部威胁，且每个站点均可了解和验证同伴处运行的软件。这可增强安全协作的信心、推动医学研究的进步、加快药物开发、减少保险和金融欺诈等等，其应用十分广泛，同时保持安全性、隐私性和监管合规性。

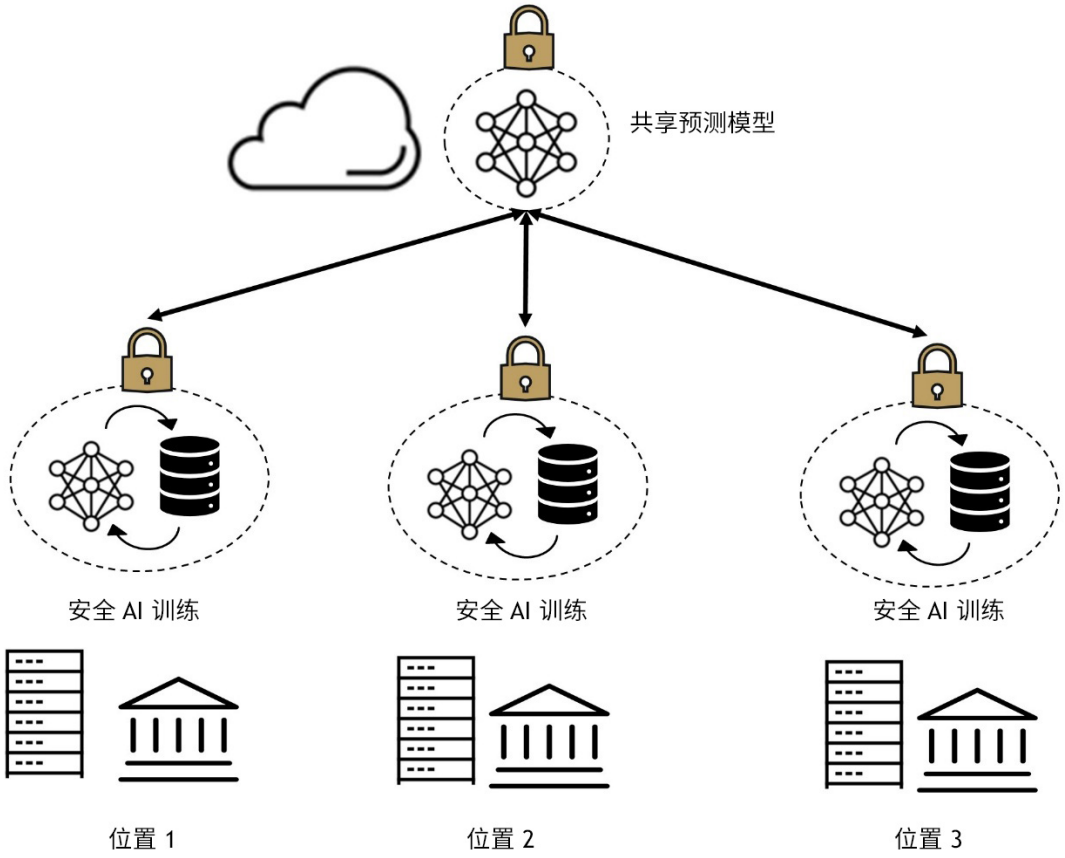


图 29. 机密联合学习

虽然在 GPU 中提供机密计算功能涉及到许多组件，但其中一个更重要的功能是进行安全可靠的引导，具体内容如下所述。

成功衡量标准

NVIDIA Ampere GPU 架构虽已采用安全启动技术，但不支持实现机密计算合规所需要的测量启动。我们简单讨论一下 H100 中实施的安全和测量启动的概念和组成部分。

安全启动是一系列硬件和软件系统技术，可确保 GPU 在已知安全状态下启动，在 GPU 启动时只允许运行经过身份验证的固件和 NVIDIA 所编写与审查过的微码。测量启动过程会收集、安全存储和报告启动过程的特征，这些特征可用于确定 GPU 是否处于安全状态。认证和验证是通过比较测量值和参考值来确保设备处于预期安全状态。NVIDIA 提供证明人、参考值和背书签名。

部署工作流程利用通过测量启动提供的测量值，与 NVIDIA 或服务提供商所提供的参考值进行比较，以确定系统是否处于就绪和安全状态，从而开始处理客户数据。系统验证完毕后，客户可以启动应用，就像在非机密计算环境中运行这些应用一样。

NVIDIA 机密计算实现概述

如图 30 所示，左侧 NVIDIA CC 关闭图代表传统 PC 架构，其中主机操作系统和服务器虚拟化平台具有 GPU 等设备的全部访问权限。右侧 NVIDIA CC 开启图显示 VM 与其他单元完全隔离。

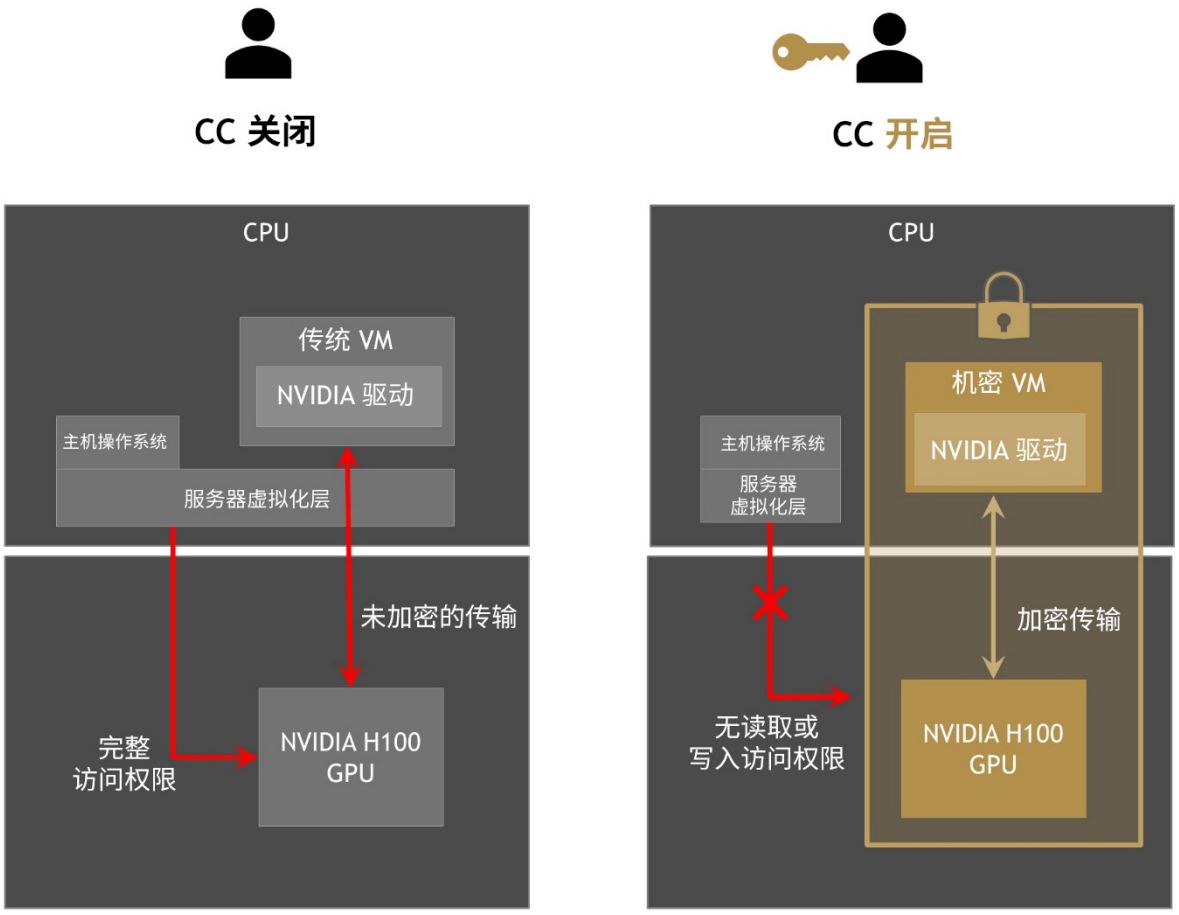


图 30. NVIDIA CC 关闭和 CC 开启时的 VM 隔离情况

VM TEE 和 GPU TEE 完全隔离，形成机密计算环境，是通过基于硬件的强大安全保护功能实现的，其中包括前面已提及部分内容的三个关键要素：

- **裸片上硬件信任根 (RoT)** - 在操作系统与 GPU 通信之前，GPU 使用 RoT 确保设备上运行的固件是可靠的，并且未被设备所有者（CSP 等）篡改。
- **设备认证** - 在启用机密计算时，帮助用户确保与其通信的 NVIDIA GPU 是可靠的，并且 GPU 的安全状态是已知可信的安全状态（包括固件和硬件配置）。
- **AES-GCM 256** – CPU 和 H100 GPU 之间的数据传输通过 AES256-GCM 硬件实现以 PCIe 线速进行加密/解密。此实现通过仅供 CPU 和 GPU TEE 使用的密钥，确保跨总线传输数据的机密性和完整性。此加密实现将通过 FIPS 140-3 2 级认证。

请注意，无需更改 CUDA 应用代码，即可使用 NVIDIA 机密计算技术。

H100 视频及 IO 功能

适用于 DL 的 NVDEC

与 A100 相比，H100 视频解码能力有显著的提升。在 DL 平台中，输入视频可以以任何行业标准进行压缩，例如 H264 / HEVC / VP9 等。在 DL 平台中实现高端到端吞吐量的重大挑战之一是能够平衡视频解码性能与训练和推理性能。否则，GPU 的 DL 性能将无法被充分利用。相较于支持 5 个 NVDEC（NVIDIA 解码器）单元的 A100，H100 支持 8 个 NVDEC 单元，可显著提高解码吞吐量。这也保证了在 MIG 操作中，每个 MIG 分区都可以得到至少 1 个 NVDEC 单元。

表 5. A100 与 H100 视频解码（视频流数量）的对比情况：

#1080p 30 流	**HEVC 解码**	**H264 解码**	**VP9 解码**
H100	340	170	260
A100	157	75	108

表 6. H100 硬件解码支持

	位深度	色度格式
H264	8 位	4:2:0
HEVC	8 / 10 / 12 位	4:2:0 / 4:4:4
VP9	8 / 10 / 12 位	4:2:0

NVJPG (JPEG) 解码

实现 DL 训练和推理高吞吐量的基本瓶颈之一是图像的 JPEG 解码处理（压缩 -> 原始）。由于处理图像位的操作是操作的，所以 CPU 和 GPU 的 JPEG 解码效率都不太高。此外，如果在 CPU 中完成 JPEG 解码，PCIe 将成为另一个瓶颈。

H100 包含八个单核 NVJPG HW 引擎来加速 JPEG 解码，而 A100 中只有一个 5 核引擎。

H100 NVJPG 引擎亮点：

- NVJPG 支持 YUV420、YUV422、YUV444、YUV400 和 RGBA 格式。
- 与 A100 相比，改进了 JPEG 架构：与 A100 的 5 核引擎配置不同，H100 加入了 8 个单核引擎。这显著简化了软件使用模型，因为 JPEG 图像可以独立分配到各个引擎中，而无需按五张图像一个批次收集。此外，它还提高了同一批次中图像分辨率不同情况下的吞吐量。
- 在 MIG 操作中，每个 MIG 分区至少可以得到一个 NVJPG 引擎。
- 与 A100 相比，显著提高了 JPEG 吞吐量。

表 7. NVJPG 解码性能

使用 1080p 分辨率时每秒处理的图像数	JPEG 444 解码	JPEG 420 解码
H100	3310	6350
A100	1490	2950

* 假设上述 JPEG 吞吐量的压缩比为 10:1
** 假设上述吞吐量的分辨率为 1080p。在 224x224 等较小的分辨率下，JPEG 图像吞吐量可能比上面显示的低约 30-40%。

NVIDIA 提供了数据加载库 (DALI)，可自动调用 NVDEC / NVJPG 来管理视频 / 图像工作流的硬件加速。它为 AI 开发人员提供了一种在 DL 工作负载中使用视频 / 图像硬件引擎的简单方法。它还允许使用灵活的图表创建和自定义视频/成像工作流。如需 DALI 的详细说明和用户指南，请访问 https://docs.nvidia.com/deeplearning/dali/user-guide/docs/。DALI 库下载地址为 https://github.com/NVIDIA/DALI。

附录 A - NVIDIA DGX - 数据中心 AI 的基础模块

人工智能 (AI) 现已成为应对艰巨业务挑战的首选解决方案。无论是通过改善客户服务、优化供应链、获取商业智能，还是为几乎所有行业设计尖端产品和服务，人工智能都为企业机构提供了实现创新的机制。作为 AI 基础架构的先驱，NVIDIA DGX 系统提供了最强大、最完善的 AI 平台，可将这些基本想法付诸实践。

NVIDIA DGX H100 - 完善的 AI 平台

NVIDIA DGX H100 助力业务创新和优化。作为 NVIDIA 传奇 DGX 系统的最新版本和 NVIDIA DGX SuperPOD 的基础，DGX H100 搭载了突破性的 NVIDIA H100 Tensor Core GPU。该系统旨在最大限度地提高 AI 吞吐量，为企业提供高度精细化、系统化和可扩展的平台，帮助他们在自然语言处理、推荐系统、数据分析等方面取得突破。DGX H100 可在本地部署，并通过各种访问和部署选项提供企业所需的性能，以解决 AI 面临的最大挑战。

DGX H100 概述

NVIDIA DGX H100 是适用于训练、推理和分析的高性能通用 AI 系统。DGX H100 支持云原生，支持 BlueField-3、NDR InfiniBand 和第二代 MIG 技术。单个 DGX H100 系统即拥有出众性能，运算速度高达 32 Petaflops。通过将多个 DGX H100 系统连接到 DGX POD 甚至 DGX SuperPOD 的集群中，可以轻松扩展性能。

每个 DGX H100 系统由以下组件构成：

- 8 个 H100 Tensor Core GPU
- 第 4 代 Tensor Core
- 第 4 代 NVLink
- 第 3 代 NVSwitch（4 个）
- 8 个 ConnectX-7（400Gb/s InfiniBand 及以太网）
- 2 个 BlueField-3 DPU
- 已启用的 PCIe 5.0

卓越的数据中心可扩展性

NVIDIA DGX H100 是 NVIDIA DGX SuperPOD 等大型 AI 集群的基础模组，为企业打造可扩展的 AI 架构描绘了蓝图。DGX H100 中的 8 个 NVIDIA H100 GPU 使用全新的高性能第四代 NVLink 技术，并通过 4 个第三代 NVSwitch 互连。第四代 NVLink 技术的通信带宽是上一代的 1.5 倍，其速度最多比 PCIe 5.0 快 7 倍。DGX H100 中的 GPU 到 GPU 总吞吐量高达 7.2TB/s，与上一代 DGX A100 相比，提高了近 1.5 倍。DGX H100 系统配备 8 个 NVIDIA ConnectX-7 InfiniBand 及以太网网卡，每个网卡运行速度高达 400Gb/s，可为大规模 AI 工作负载提供强大的高速网络。

每个 DGX H100 还配备两个 NVIDIA BlueField-3 DPU（数据处理器），用于支持智能、硬件加速存储、安全和网络管理功能。BlueField-3 DPU 将传统计算环境转换为安全且经过加速的虚拟私有云，使企业组织能够在安全的多用户环境下运行应用工作负载。BlueField-3 将数据中心基础设施与业务应用程序分离，增强了数据中心的安全性、简化了运营并降低了总成本。BlueField-3 采用 NVIDIA 的网络内计算技术，支持下一代超级计算平台，提供出色的裸机性能并原生支持多节点用户的隔离。

大规模 GPU 加速计算、最先进的网络硬件和软件优化相结合，意味着 NVIDIA DGX H100 可以扩展到数百或数千个节点，以应对下一代 AI 应用程序的最大挑战。

NVIDIA DGX H100 系统规格

表 8. NVIDIA DGX H100 系统规格

规格	DGX A100	DGX H100
GPU	8 个 NVIDIA A100 GPU	8 个 NVIDIA H100 GPU
TFLOPS	5 GPU Tensor PFLOP	32 GPU Tensor PFLOP
GPU 显存	每个 GPU 80GB / 每个 DGX A100 节点 640GB	每个 GPU 80GB / 每个 DGX H100 节点 640GB
系统内存	基础配置为 1TB 3200MHz DDR4，可额外选配 1TB 内存，实现高达 2TB 的总内存	2TB
存储	数据缓存驱动器：15TB（4 块 3.84TB 第四代 NVME 硬盘。可额外选配 15TB，实现高达 30TB 的总存储量） 操作系统驱动器：2 块 1.92TB NVME SSD	数据缓存驱动器：30TB（8 块 3.84TB 硬盘） 操作系统驱动器：2 块 1.92TB NVME SSD
网络	8 个 200Gb/s 的单端口 NVIDIA ConnectX-6 HDR InfiniBand 网卡 2 个 10/25/40/50/100/200Gb/s 的双端口 NVIDIA ConnectX-6 以太网网卡	4 个 OSFP 端口连到 8 块单端口 NV CX7 卡 400Gb/s 的 InfiniBand 及以太网网卡 2 个双端口 NVIDIA BlueField-3 DPU VPI 网卡 1 个 400Gb/s 的 InfiniBand 及以太网网卡 1 个 200Gb/s 的 InfiniBand 及以太网网卡
散热	风冷	风冷

附录 B - NVIDIA CUDA 平台更新

NVIDIA CUDA 是用于加速计算的全面、高效和高性能平台。它利用 GPU、CPU、DPU 和网络内计算加速所有级别的最终用户应用程序，包括系统软件和特定应用程序的库以及框架（参见图 31）。得益于成熟友好的工具链、开发者工具和文档，此平台可提供卓越的开发者体验，助力异构应用加速。

高性能库和框架

CUDA 库可更大限度地提高常用数学运算（CUDA 数学库）、并行算法（CUB 和 Thrust）、线性代数 (cuBLAS)、密集和稀疏线性求解器（cuSOLVER 和 cuSPARSE）、FFT (cuFFT)、随机数生成 (cuRAND)、张量操作 (cuTENSOR)、图像和信号处理 (NPP)、JPEG 解码 (nvJPEG) 和 GPU 管理 (NVML) 的性能。cuNumeric 可通过 Legate 和 Legion 运行时，以透明方式加速 NumPy 程序，并将其分发给各种规模的机器，而无需修改任何代码。libcu++ 可提供异构同步和数据移动基元，帮助构建高度并发、符合 ISO 标准的异构 C++ 应用。

此外，CUDA 平台通信库还支持基于标准的可扩展系统编程。HPC-X 是支持 GPUDirect 的 CUDA 感知型 MPI 库，可直接使用 RDMA 发送和接收 GPU 缓冲区。NVIDIA 集合通信库 (NCCL) 可实现高度优化的多节点集合通信原语。NVSHMEM 基于 OpenSHMEM 构建而成，可为主机和设备线程提供异构多节点通信原语。cuFile 和 MAGNUM IO 可在 GPUDirect Storage 的助力下支持具有高性能文件 I/O 的异构应用。

一整套域特定库和框架可进一步加速各种应用领域的主要算法，例如深度神经网络 (cuDNN)、用于模拟和隐式非结构化方法的线性求解器 (AmgX)、量子计算 (cuQuantum)、数据科学和机器学习 (RAPIDS)、用于机器学习的数据加载和预处理 (DALI) 和实时 3D 仿真及设计协作 (Omniverse) 等。150 余款软件开发套件利用这些库，帮助大量应用领域的开发者提高工作效率，其中包括高性能计算 (NVIDIA HPC SDK)、AI、机器学习、深度学习和数据科学、基因组学 (NVIDIA CLARA)、智慧城市 (NVIDIA Metropolis)、自动驾驶 (NVIDIA Drive SDK)、电信 (NVIDIA Aerial SDK)、机器人开发 (NVIDIA Isaac SDK)、网络安全 (NVIDIA Morpheus SDK)、计算机视觉等。

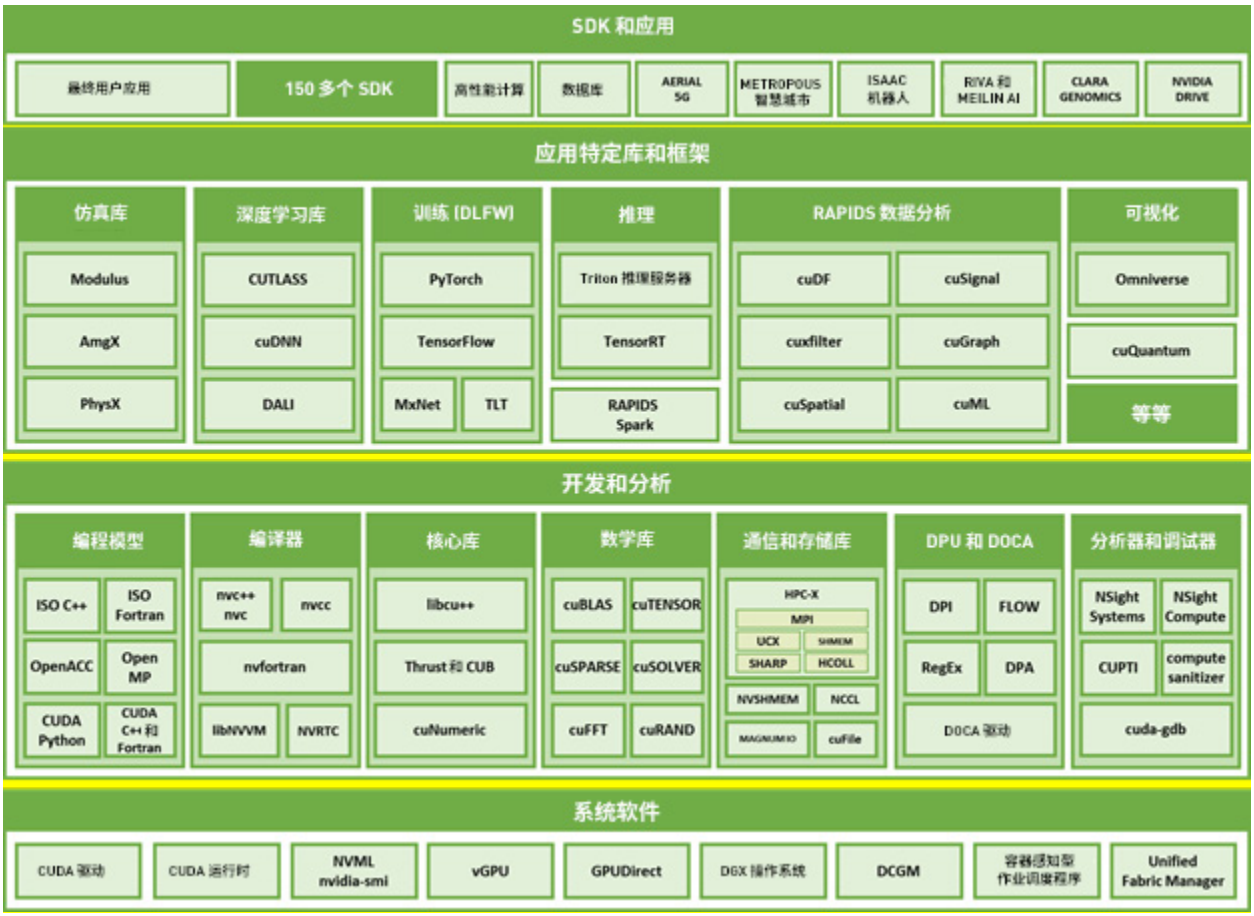


图 31. NVIDIA CUDA 平台及其生态系统

系统软件

NVIDIA CUDA 平台还具备灵活的系统软件组件，可帮助用户高效部署、管理和优化大型异构系统。所提供的软件涵盖设备驱动（CUDA 驱动）、设备管理软件（NVML、NVIDIA-smi、DCGM 和 UFM）、用于异构网络和文件 I/O 的 GPUDirect、容器感知型作业调度系统和操作系统 (DGX OS)。

文档和培训

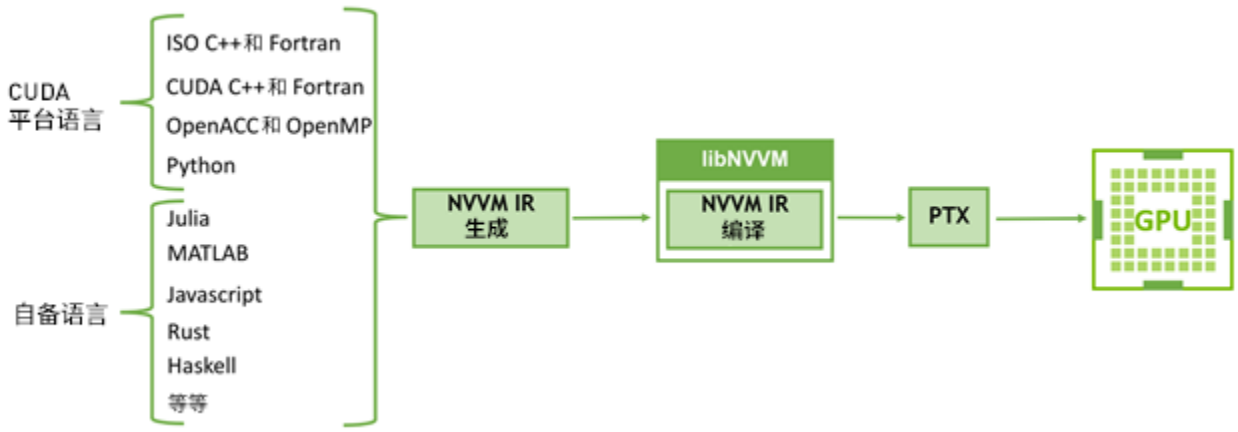
大型 CUDA 软件生态系统与出色的文档相辅相成，这些文档包含编程模型（例如 C++ 并行算法）、库（例如 libcu++）、框架（例如 RAPIDS AI）和 SDK（例如 HPC SDK）等内容。

NVIDIA 深度学习培训中心 (DLI) 针对超级计算机和国际超级计算大会等会议推出了自定进度的实时培训，可帮助个人拓展 AI、加速计算、加速数据科学、图形和仿真等领域的知识。深度学习培训中心 (DLI) 在科研机构和 HPC 中心培训教育工作者，并将合格的教育工作者认证为 DLI 大使，助力他们教授学员，并根据需求定制 DLI 内容。

除官方文档外，NVIDIA 还与不同社区和 HPC 网站合作，推出 GPU 编程黑客松和训练营计划。该计划将专业领域的科学家和研究软件工程师 (RSE) 团队与来自 NVIDIA 和 HPC 社区的 GPU 导师结对，以便传授高效使用现代异构计算系统所需的软件开发、并行计算和优化技能。NVIDIA 每年都会举办 GTC 技术大会，帮助开发者了解最新 NVIDIA 平台和技术。会议内容涵盖 NVIDIA 编程模型、硬件详情以及加速计算在各个领域中的应用。所有会议内容都有相关记录，如需了解详情，请点播观看 GTC 往期内容回放。

语言和编译器

CUDA 平台采用统一的灵活编译器堆栈，可通过 NVIDIA 的 NVVM IR 和 NVIDIA 的 libNVVM 生成高度优化的设备二进制文件。NVVM IR 是基于 LLVM 7 的编译器中间表示 (IR)，提供用于生成 GPU 计算内核的前端编译器目标。 libNVVM 是一个库，用于将 NVVM IR 编译和优化为 PTX，即 NVIDIA GPU 的虚拟 ISA。所有 NVIDIA 计算编译器均使用 libNVVM 确定 NVIDIA GPU 目标（图 32），并帮助用户和框架将其选择的编程语言引入 CUDA 平台，而且代码生成质量和优化效果与本身的 CUDA C++ 语言相同。



前端使用 libNVVM 将 NVVM IR 程序编译为 PTX，并在 GPU 上运行

图 32. 高级语言前端

PTX 是 NVIDIA GPU 的虚拟 ISA，是第三方生产商的目标公共 ISA，可在我们的目标架构上高效运行。PTX 还具有向前兼容的优势，可以离线或在运行时进行组装。

在许多应用中，要生成的 GPU 计算内核取决于程序输入。虽然这些应用可以生成 NVVM IR，但 NVIDIA 运行时编译器可帮助他们改为生成熟悉的 CUDA C++ 语言，从而显著提高这些应用及其用户的工作效率。NVRTC 在运行时使用 libNVVM 将 CUDA C++ 编译为 PTX，或使用嵌入式 PTX 汇编程序编译为原生 GPU 二进制代码。这样，应用（例如 Python 程序）即可为用户输入的程序（例如 C++ 程序）动态生成内核，以便在运行时根据程序输入专门生成计算内核。

NVIDIA HPC SDK 是一套用于异构系统的**工具链**。NVCC 是 CUDA C++ 编译器，可提供用于配对 GPU 编译与 GCC 等外部主机编译器的分割编译模型（图 33：左）。NVIDIA HPC 编译器（NVC、NVC++ 和 NVFortran）可提供统一的异构编译模型（图 33：右）。

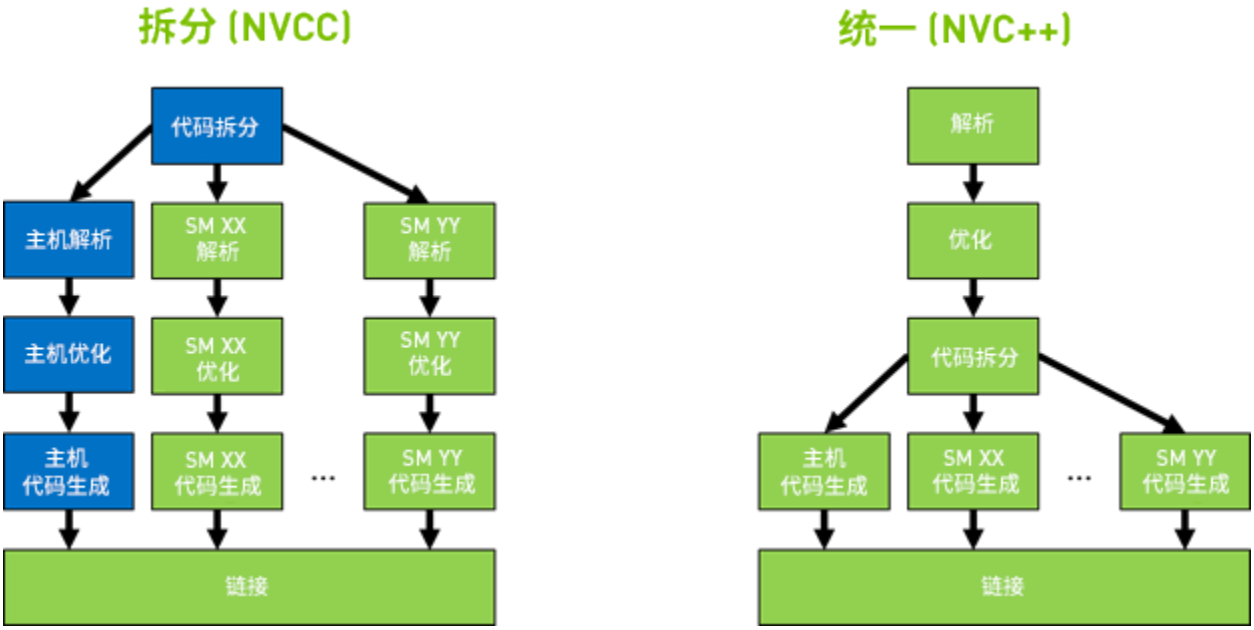


图 33. NVCC 分割编译模型和 NVC++ 统一编译模型

统一编译器在针对不同目标分割编译过程之前，只对程序执行一次解析和优化。此模型启用了 NVCC 中未提供的某些功能。例如，使用 NVCC 时，CUDA C++ 设备代码需要使用 __device__ 注释（图 34：左）。NVC++ 编译器不需要使用这些注释（图 34：右），如果程序使用特定目标的函数，并且可以访问其定义，则此编译器会尝试对其进行编译。

NVCC 需要设备注释

```
__host__ __device__
int square(int x) { return x * x; }

__global__
void square_elements(int* x) {
  int i = threadIdx.x + blockIdx.x * blockDim.x;
  x[i] = square(x[i]);
}
```

NVC++ 推理执行空间

```
int square(int x) { return x * x; }

__global__
void square_elements(int* x) {
  int i = threadIdx.x + blockIdx.x * blockDim.x;
  x[i] = square(x[i]);
}
```

图 34. 统一工具链支持执行空间推理

统一编译简化了开发过程，使初学者能够更轻松地完成 GPU 编程入门，同时可提高经验丰富的开发者的工作效率。此编译方法还增加了主机和设备目标之间的代码重用率，从而简化了 GPU 应用加速过程。

附录 C - 使用 DPX 指令加速基因组学

NVIDIA H100 实现比先前 GPU 和 CPU 性能高 X 倍的多种类型的应用和算法加速。本节会重点介绍 H100 在基因组学领域所带来的显著加速效果。这几年来，随着传染病的肆虐和全球流行病危机的爆发，对于人类而言，基因组和蛋白质分析的重要性远胜以往。

H100 引入了新的 DPX 指令，这一新的专用硬件指令可实现动态编程算法加速，例如用于 DNA 基因测序以及蛋白质分类和折叠的 Smith-Waterman 算法。与 NVIDIA Ampere A100 GPU 相比，H100 最多可将 Smith-Waterman 算法的速度提升 7 倍，从而更快地获得疾病诊断、病毒突变研究和疫苗开发解决方案。下面介绍有关基因组学和基因测序的简短教程。

基因组学领域发展迅猛，引发了医疗健康、农业和生命科学行业的变革，同时还成为我们抗击 SARS-CoV-2 和 COVID-19 的最有力武器之一。对整体或选定部分的人类基因组进行测序，这对于我们了解基因组的工作原理至关重要，这让我们能够识别可引发疾病的基因变异，并提供相应保护和靶向治疗。随着企业组织不断利用基因组来了解疾病、研发新药和加强患者护理，数据分析和管理正成为提取基因组价值的主要工具。

自 2005 年推出新一代测序 (NGS) 以来，行业数据呈爆炸式增长，同时与人类基因组相关的新行业也应运而生，从解决家族病史到临床护理，无所不及。先进的计算系统可加速将原始仪器数据转化为生物学见解所需的计算密集型步骤，从而为基因组学提供支持。单个基因组的原始数据大小约为 100GB。在使用复杂算法和应用（例如深度学习和自然语言处理）进行分析后，这一数据的总量超过了 225GB。使用 GPU 实现数学模型加速可为传统的基因组学分析（例如测序读取处理和变异识别）带来明显优势，不过，这也可能会彻底改变我们对特定基因组变异体如何影响疾病和健康的理解。

NVIDIA Clara™ Parabricks® 是适用于新一代测序数据的加速计算框架，支持 DNA 和 RNA 应用的端到端数据分析工作流程。Clara Parabricks 可在一系列 NVIDIA GPU 平台上运行，能够提供包括 GPU 加速的 Burrows-Wheeler Aligner (BWA-MEM)、Picard 和 Samtools 在内的 50 多种加速工具，以及一套用于标记、筛选和组合多种变体调用格式 (VCF) 的实用程序。在整个工作流程中结合使用加速工具意味着系统无需数小时或数天的检测时间，仅在数分钟内即可生成检测结果。

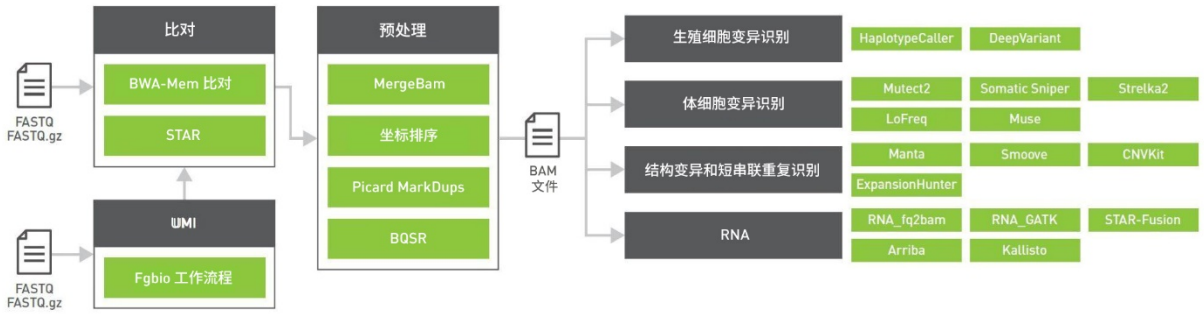


图 35. NVIDIA CLARA Parabricks 加速框架

基因组是生物体的一整套脱氧核糖核酸 (DNA)，此化合物中包含实施和引导每个生物体活动所需的遗传指令。DNA 分子由两条双螺旋链组成。每个链由四个称为核苷酸碱基的化学单元组成。这些碱基分别是腺嘌呤 (A)、胸腺嘧啶 (T)、鸟嘌呤 (G) 和胞嘧啶 (C)。具体来说，两条方向相反的螺旋链上的碱基会进行配对；A 始终与 T 配对，C 始终与 G 配对。人类基因组大约包含 30 亿个这样的碱基对，它们分布在我们所有细胞核中的 23 对染色体中。对基因组测序意味着确定一段 DNA 中碱基对的确切顺序。

个体 DNA 测序的第一个化学过程是：将 DNA 分割成互补碱基对，将 DNA 链切割成特定大小的块（长度为 100 - 2000 个碱基对），然后通过测序机对这些小块测序（称为读取），以生成一连串计算机可读的碱基对代码。然后，通过在参考基因组中搜索序列的位置或使用 De Novo 方法，重新组合这些测序块，De Novo 方法组合测序块的依据是查找碱基的重叠模式，而不是依赖参考基因组序列。

从计算的角度来看，问题在于从长度为数十亿个碱基对的参考基因组中搜索和匹配一组“读数”，或者通过模式匹配算法从头开始组合基因组，此算法需比较数百万个读数以找到重叠项，并按正确的顺序进行对齐。在此过程中，这类算法可能需要插入、编辑或删除序列，以解决不匹配问题，同时还需要指定可能遇到的各类不匹配的成本。因此，用于匹配模式的计算硬件架构需要灵活适应这些要求，同时还需要支持用于基因组学其他问题（如蛋白质测序）的其他类似算法。

用于 DNA 测序的 Smith-Waterman 算法可通过 NVIDIA CLARA Parabricks 加速计算框架的 GPU 加速 BWA-MEM 模组完成。该算法基本上是通过比较两个碱基读数字符串来创建评分矩阵，然后根据矩阵中分数的追踪结果来确定两个字符串较匹配的模式。如需详细了解如何使用此算法进行基因组测序，请点击此处。

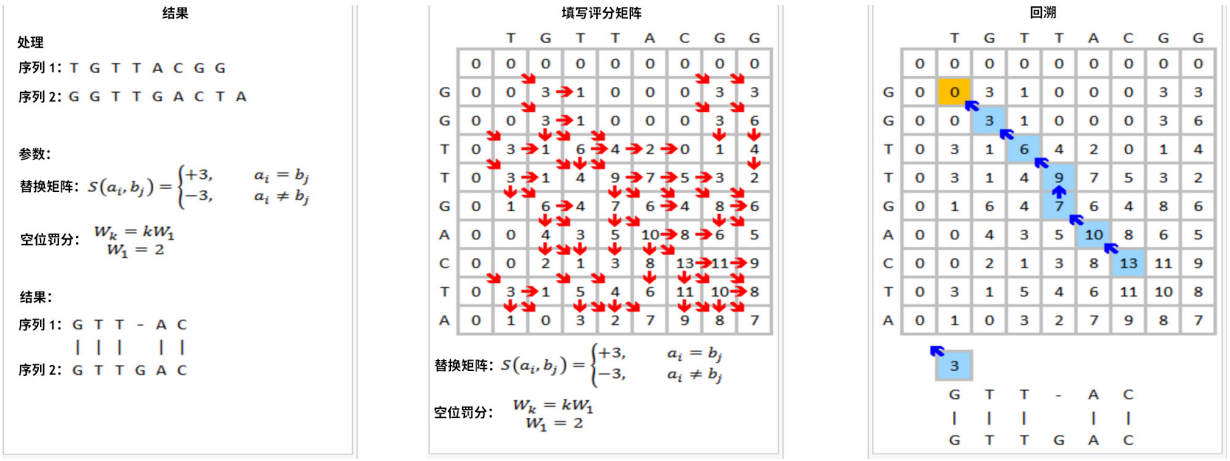


图 36. 用于基因组测序的 Smith-Waterman 算法 [1]

在上面的图示中，矩阵的每个单元更新都需要进行 5 次基本计算。

1. 对角元素匹配时加上 x 值（本图中 x = 3）
2. 对角元素不匹配时减去 x 值
3. 垂直元素不匹配时减去 y 值（本图中 y = 2）
4. 水平元素不匹配时减去 z 值（本图中 z = 2）
5. 找到上述四个运算的最大值（如果计算结果为负值，则将此单元归零）。

H100 中的新 DPX 指令经过优化，可有助于加速上述系列计算及其他类似算法。

[1] 来源：https://en.wikipedia.org/wiki/Smith%E2%80%93Waterman_algorithm

声明

本规范中提供的信息在其发布日期之时是准确可靠的。但是，NVIDIA Corporation（“NVIDIA”）对此类信息的准确性或完整性不作任何明示或暗示的陈述或保证。对使用此类信息的后果或因使用此类信息而造成侵犯第三方专利权或其他权利的后果，NVIDIA 概不负责。本出版物将取代之前可能已提供的所有其他产品规范。

NVIDIA 保留随时对这一规范进行纠正、更改、增强、改进以及其他改动和/或终止任何产品或服务的权利，恕不另行通知。客户在下订单之前应获取最新的相关规范并验证这些信息是否为当前信息以及是否完整。

除非 NVIDIA 授权代表与客户另行签署销售协议，否则 NVIDIA 产品的销售受订单确认时所提供的 NVIDIA 标准销售条款与条件的制约。就购买这一规范中提到的 NVIDIA 产品而言，NVIDIA 在此明确拒绝应用客户的任何一般条款与条件。

NVIDIA 产品并非针对医学、军事、航空、航天或生命保障设备而设计，并未授权用于也不保证适合用于上述设备，亦不得用于 NVIDIA 产品之失效或故障合理预计会造成人身伤亡或财产或环境破坏的应用场合。客户如果在此类设备或应用场合中融入和/或使用 NVIDIA 产品，NVIDIA 不承担任何相关责任，风险由客户自行承担。

在未经进一步测试或改动的情况下，NVIDIA 并不表示，也不担保基于这些规范的产品适合任何具体用途。每款产品所有参数的测试不一定由 NVIDIA 进行。确保产品适合客户所计划的应用场合并针对该应用场合进行必要的测试以避免应用场合出现问题或产品失灵，是客户单方面的责任。客户产品设计中的缺点可能会影响 NVIDIA 产品的质量和可靠性，并且可能会导致超出本规范以外的额外或不同的条件和/或要求。NVIDIA 不承担因下列情况造成失灵、损坏、成本或问题相关的任何责任：(i) 以违反本规范的方式使用 NVIDIA 产品或 (ii) 客户产品设计。

本规范不对 NVIDIA 专利权、版权或其他 NVIDIA 知识产权作任何明示或暗示的许可。NVIDIA 所发布的有关第三方产品或服务的信息并不构成 NVIDIA 对于使用该产品或服务的许可，亦不构成担保或支持。使用此类信息可能需要获得第三方的专利权或其他知识产权的许可，或者需要获得 NVIDIA 的专利权或其他知识产权的许可。只有在获得 NVIDIA 书面批准的情况下才可以复制本规范中的信息，而且必须毫无改动地复制并附带所有相应的条件、限制条款和通知。

所有 NVIDIA 设计规范、参考板、文件、图纸、诊断、列表和其他文档（统称与单称均为“资料”）均“如实”提供。NVIDIA 并未作出与资料相关的明示、暗示、法定或其他形式的保证，并明确否认与非侵权、适销性和特定用途适用性相关的所有暗示保证。尽管客户可能会因任何原因造成损失，但是 NVIDIA 针对本文所述产品向客户承担的全部责任应仅限于该产品的 NVIDIA 销售条款与条件。

特别感谢 NVIDIA GPU 技术专家李蒙、刘洋、史永明、舒引博、王亮、肖骏、张海军、张然对本次白皮书中文版校对的大力支持与帮助。

商标

NVIDIA、NVIDIA 徽标、NVIDIA CUDA、NVIDIA Omniverse、NVIDIA RTX、NVIDIA Tesla、NVIDIA Turing、NVIDIA Volta、NVIDIA Jetson AGX Xavier、NVIDIA DGX、NVIDIA HGX、NVIDIA EGA、NVIDIA CUDA-X、NVIDIA GPU Cloud、GeForce、Quadro、CUDA、GeForce RTX、NVIDIA NVLink、NVIDIA NVSwitch、NVIDIA DGX POD、NVIDIA DGX SuperPOD 和 NVIDIA TensorRT 均为 NVIDIA Corporation 在美国和其他国家/地区的商标或注册商标。其他公司名称和产品名称可能为相应各公司的商标。

版权所有 © 2022 NVIDIA Corporation。保留所有权利。