各位早上好。构建基础设施是一项充满挑战且需要勇气的事业，而构建一套不仅能支持单个机柜、还能支撑整个数据中心的基础设施，更是一项挑战 —— 毕竟要提供足够多的计算资源，才能让所有研究人员高效工作、持续推进项目。如今，**我们正在打造的正是这类数据中心 —— 下一代数据中心**。每年我们都会推出新的平台，这正是人工智能发展的速度。

因此，在本次 AI 推理峰会上，我想和大家分享英伟达对这个市场的一些看法，尤其是围绕“推理”这一领域 —— 它是如今我们许多人关注的重点和核心。AI 推理的领域格局相当复杂。通常人们认为模型训练难度更大，毕竟这需要用到我们刚才提到的十万块 GPU。但推理本身同样涉及诸多需要权衡的优化维度。

首先是模型智能度：我想部署的模型规模有多大？这直接决定了模型的价值，但模型规模越大，运行成本就越高；其次是模型响应速度：为某个特定模型配置多少计算资源，才能让它交互性更强、速度更快 —— 是提升单用户每秒令牌数，还是提升整个数据中心的每秒令牌数？这些都需要我们在基础设施投入规模、预期营收与用户体验之间做出权衡。

成本也是关键因素。显然，硬件平台的配置选项繁多，从 H100、B200、GB200 到即将推出的 GB300，选择适合自身需求的基础设施成本方案，来构建和运行这些模型，是必须考虑的问题。此外，吞吐量与单用户每秒令牌数之间也存在权衡：如果为单个查询投入大量 GPU，模型响应速度确实会极快，但数据中心的营收来自所有查询的总和 —— 在单个用户查询上投入越多，能处理的总查询量就越少。我们必须在这种帕累托最优关系中权衡取舍。

最后是能效：衡量数据中心性能的指标，往往不是占地面积，而是兆瓦或吉瓦。因此，了解单个模型运行时的能效水平至关重要。以上所有权衡，都是我们在规划基础设施、决定技术路线图以及推动创新时必须面对的问题。

而且，硬件研发周期很长，芯片研发周期更长。我们必须提前 1 到 2 年预判这些权衡因素，但 AI 领域的变化速度几乎让这种预判变得不可能。而我的工作中最具挑战性的部分，就是思考这些权衡并做出正确的决策 —— 这也是英伟达的工作方式之一。

另一种思考方式是参考 “TRICKLE SMART” 这一缩写所代表的维度：**首先要考虑扩展复杂度** —— 你正在构建的基础设施，其扩展复杂度如何？我们该如何实现更高效的扩展？其次要关注多维度性能 —— 每一个决策都需要考量智能度、性能、吞吐量、成本、能效等不同维度，以及这些指标如何共同决定最终的端到端解决方案。

最终，这需要一套全栈解决方案：既包括芯片架构、节点架构、机柜架构，也包括多层软件及软件优化。所有这些要素结合在一起，才能形成完整的解决方案。因此，是 “硬件 + 软件” 共同决定了整体吞吐量性能。

最终，这套方案要能带来投资回报率（ROI）—— 尤其是**在推理领域，性能就等于营收**。我之后会举例子说明其中的逻辑，但毋庸置疑，性能是生成令牌、处理查询的核心。可以说，推理领域是 “性能决定营收”，而训练领域更偏向 “能力决定成本”。

此外，所有这些进展都不是在真空环境中发生的，而是**依赖整个社区的力量**。在座各位都在

为构建未来的 AI 推理生态贡献力量：你们中许多人代表的公司，要么在建设数据中心、研发硬件或机柜，要么在整合这些资源并提供服务；同时，开放软件社区也功不可没。 比如 OCP（Open Compute Project）、开源软件比如 OpenAI Triton、PyTorch 等软件栈，这些创新都来自社区。

与过去许多次计算革命不同，AI 领域的开放性尤为突出：研究人员和企业会公开分享想法、发布研究成果、探讨如何推动未来发展。这是因为 “水涨船高” —— 当我们所有人都投身其中，AI 的发展机会就会增多，更多模型、更多功能会形成良性循环。

而英伟达的贡献，始终是探索提升 “每瓦性能”“每美元性能”的方法，最终帮助 AI 数据中心实现盈利。我们最近的创新重点之一是 NVL 72：我们将原本集成 GPU 的服务器单元进行拆解，把 GPU 从服务器箱体内移出，实现了机柜级扩展 —— 这是一项极具挑战性的工作，但最终打造出了如今已投入部署和使用的出色基础设施。

我们还重新思考了以太网的扩展方式：传统上，我们通过 InfiniBand 连接 1 万、2 万、5 万块 GPU，而现在我们希望**通过以太网基础设施，将 GPU 连接规模提升到 10 万甚至 20 万块。**这需要一种全新的以太网 —— 一种能让每块 GPU、每个节点都能以满性能与其他任何 GPU、节点通信的以太网。这与以太网的原始设计用途截然不同：传统以太网用于连接客户端与服务器，其使用场景并非 “所有设备时刻相互通信”，因此我们需要全新的网络和全新的交换机。

这正是我们推出 Spectrum-X 的基础 —— 即便在数值方法领域，AI 本质上也是一个统

计问题：无非是输入数据、生成预测。而结合统计技术与数值计算，我们可以拓展数值表示的边界。

一切始于 FP32（32 位浮点精度），这是大家熟知的格式。10 年前，谷歌推出了 FP16（16 位浮点精度）；如今，IEEE 又新增了 FP8（8 位浮点精度），我们也支持这一格式，现在甚至能支持 4 位浮点精度。事实上，4 位浮点精度还有多种格式，最近我们在 Blackwell 平台上推出了 NVFP4 格式（英伟达 4 位浮点格式）—— 该格式采用微张量缩放技术，确保所有 AI 计算都能在 4 位精度下运行，同时保证数值在统计范围内，避免溢出。我们不断提升硬件能力，通过持续的偏差调整和计算修正，确保计算始终处于合理范围。

NVFP4 目前已用于推理场景，而最近我们还在研究数值算法技术，让 NVFP4 也能支持训练—— 这是我们在近期活动中宣布的进展。计算本身并不难，难的是数值计算的复杂性。

软件也是关键一环，而英伟达只是软件生态的贡献者之一。我们在 GTC（英伟达 GPU 技术大会）上发布了 NVIDIA Dynamo，之后会详细介绍。要实现推理任务在多台服务器间的拆解、完成概念处理与生成，并扩大单个模型可分配的 GPU 数量，就需要一套全新的模型服务软件栈。

最终，这套方案能显著提升 ROI：我们的 Blackwell 平台能带来约 10 倍的投资回报率 —— 也就是说，对比你在基础设施上的投入成本，AI 数据中心能产生的营收规模是前者的 10 倍。

而这一切都离不开合作。我所管理的团队合作推进 PyTorch 项目，同时也参与 JAX 及其他软件栈（如 OpenAI Titan）的开发；在推理领域，我们有 SGY VLM（英伟达生成式视觉语言模型），而其他机构也有自己的推理库，比如 TensorRT-LLM（张量 RT 大语言模型推理库）。所有这些进展都发生在开放社区中 —— 通过贡献、支持并提供这些技术的软件基础，我们让研究人员、开发者、企业和云服务商能够使用这些技术，从而加速这些技术的市场化进程，这个速度非常惊人。

在推理领域，衡量性能并非易事。声称 “性能出色” 很简单，但模型的准确性如何？我可以运行一个模型并降低精度，但如果降低精度后，模型速度变快了，准确性却下降了，那还不如直接使用一个更简单的模型。

MLPerf（机器学习性能基准测试）是一个已存在多年的基准测试项目，2019 年他们推出了 MLPerf Inference。谷歌、英伟达、Meta 等业内众多企业都会参与其中，共同制定评估标准—— 我们会相互评审结果，通过基准测试证明：在特定用户场景下，我们的模型能达到特定性能水平和令牌生成速率。

MLPerf Inference 已成为衡量推理性能的公认标准、。自 2019 年以来，英伟达一直在提交测试结果，覆盖从 Ampere 架构、Hopper 架构到 Blackwell 架构的所有平台。我们每年都会提交新的基准测试结果，不断提升 AI 模型在推理领域的性能，并且自 MLPerf 数据中心基准测试启动以来，我们保持着每块 GPU 的性能纪录。

事实上，就在今天晚些时候，我们将公布最新一轮测试结果：我们新增了 DeepSeek-1、

Llama 3 145B 的测试结果，而 Llama 2 的测试结果则是基于全新的 Blackwell Ultra 或 GB300 平台。测试中，我们使用了 NVFP4 格式、Dynamo 软件和全新的 TensorRT，并在整个 GB300 机柜的 GPU 间实现了推理任务的完全分布式处理。

要实现这一切，背后的软件非常复杂。事实上，**从去年 Blackwell 平台发布到现在，我们仅通过软件优化，就将 Blackwell 的性能提升了一倍** —— 这涉及通信分布式处理、数值技术应用等复杂工作，最终实现性能提升、成本降低和吞吐量优化。在相同的 Blackwell 硬件上，仅通过软件优化，性能就提升了两倍，而且这是 “免费” 的。

Hopper 架构也有类似经历：在其生命周期内，我们通过软件优化将其性能提升了四倍 —— 同样的 Hopper 硬件，速度快了四倍，这完全得益于软件优化和众多工程师的努力，更离不开开源社区和研究人员的贡献 —— 他们探索出了在保证模型精度的前提下，更快、更低成本运行模型的新方法，最终帮助提升营收。

这就是当下的实际情况：一套价值 300 万美元的 GB200 和 NVL72 机柜，实际上能产生约 3000 万美元的令牌营收 —— 以至于 “免费的 GPU” 都显得不够划算。如果对比性能仅为其 1/4 的上一代或其他平台，你会发现：即便把 GPU 成本和服务器及机柜的其他成本都算进去，多年下来产生的营收依然非常可观。

这就是我们对推理领域的看法：平台性能就是 AI 工厂（AI factory）的营收。而 Blackwell 平台的投入产出比，确实能达到 10 倍。

接下来，我们再深入聊聊推理的工作原理、当前的创新方向，以及我们的关注重点。推理领域主要涉及两类工作负载：一类是传统的推理服务 —— 当用户输入查询时，AI 模型首先会进行上下文处理。

上下文处理处理的是什么？其实就是你向 ChatGPT 或聊天机器人提出的问题，同时也包括所有与你相关的独特令牌或系统提示(prompt) —— 比如你之前提出的问题，或是与你的查询相关、能帮助 AI 回答问题的信息。也就是说，AI 不仅会分析你的当前问题，还会处理所有输入令牌 —— 这一阶段被称为 “预填充阶段”（prefill)。

在处理完所有查询、相关数据和输入令牌后，AI 才会开始输出你看到的令牌 —— 这一阶段被称为 “生成阶段”（generation)，也叫 “解码阶段”（decode)。

通常，我们会在 GPU 集群上运行这一过程：根据模型规模和性能需求，可能需要 4 块、8 块甚至更多 GPU，但一般是在一组 GPU 上运行一个模型。

有趣的是，**上下文处理和生成这两个阶段其实存在差异**：虽然它们运行的是同一个模型，但上下文处理可以通过大规模并行的方式进行 —— 我们可以同时处理所有输入令牌；而 AI 生成阶段则往往是自回归式的：每输出一个令牌，都需要重新运行模型来计算下一个令牌，循环往复。

我们可以并行处理 1.6 万、3.2 万甚至 10 万个输入令牌，因此这两个阶段的性能存在差异：上下文处理因大规模并行而速度极快，而生成 / 解码阶段因自回归特性，需要结合内

存带宽、链路带宽和计算资源，才能提升令牌输出速度。

如果我们在同一平台或同一组 GPU 上运行这两个阶段，就只能在两者之间寻求平衡，但无法实现各自的最优性能。

如今，大多数现代化数据中心都会采用 “推理拆解”（disaggregate inference）的方式：他们会先接收输入查询，在单独的 GPU 上运行上下文处理，生成所谓的 “缓存（KV Cache）”—— 本质上就是第一个令牌；然后将 KV Cache 传递给另一组 GPU，由这组 GPU 专门负责生成阶段。

这种方式能让我们灵活分配用于上下文处理和生成的 GPU 数量，从而大幅提升整体性能。英伟达 Dynamo 软件就是为此设计的，而且它是开源的 —— 大家可以去 GitHub 上查看并获取，我们所有的开发工作都在 GitHub 上进行。

通过这种优化，我们可以配置 GPU，为上下文处理阶段选择适合并行计算的 AI 核和数学方法，为自回归生成阶段选择不同的核和并行化技术，最终在 GPU 数量不变的情况下提升总吞吐量。事实上，仅针对大模型，这种方式就能带来约 6 倍的性能提升；对于其他模型，提升幅度也能达到 2-4 倍 —— 同样的 GPU 数量，仅通过 “拆解” 就能实现更快的速度。

当然，这种方式难度更高：系统需要同时运行两组工作负载，还要在两组平台间传递 KV Cache，并确保所有资源都处于繁忙状态。不过目前这一技术已投入生产。

Base Ten 就是一个例子 —— 这家公司是推理聚合（inference aggregate）领域的模型服务提供商，他们在谷歌云等多个云上部署了超过 8000 块 Hopper 和 Blackwell GPU。在 GPT-4 刚推出时，Base Ten 的推理性能是所有云服务商中最快的 —— 原因就是他们使用英伟达 Dynamo 进行了深度优化，实现了上下文处理与生成阶段的拆分。

这个例子充分说明软件的重要性 —— 尤其是当软件与 NVL72（或 GB200）这类机柜基础设施结合时，效果更为显著。总体而言，**“拆解” 能让首令牌速度提升约 6 倍，让 DeepSeek 等模型的令牌输出速度提升 3 倍**，最终将推理问题转化为数据中心级、基础设施级的挑战。

此外，我们还发现，上下文处理的重要性和价值正不断提升。如今大多数模型最多可接收约 25.6 万个输入令牌，而每个单词约对应 2-3 个令牌 —— 大家可以据此估算，向普通聊天机器人提问时，输入令牌的规模大概是多少。

不过有一类用户对 “超长输入令牌” 需求强烈 —— 先进编码（advanced coding）就是典型场景。我们都听说过编码聊天机器人（coding chat bots），它们能帮助编写代码；而先进编码聊天机器人则能接收整个程序代码，利用 AI 添加新功能 —— 不再是帮你写一个小循环或修复小 bug，而是能接收 10 万行代码或 100 万个编码输入令牌，输出新功能、完整代码块甚至应用的部分模块，将 AI 真正转变为能与软件开发者深度协作的 “软件代理(Agent)”。

要实现这一点，就必须能处理数百万个令牌，但其价值也极高：有了这样的 AI 工具，软件开发者的工作效率能提升 10 倍 —— 因为 AI 会生成初始代码，开发者在此基础上进行优化即可。

另一类当前热门的应用场景是视频处理与生成 —— 比如处理 1 小时的高清（HD）视频并生成新的视频内容。生成式视频涉及大量数据，对应数百万个令牌。**目前，AI 视频生成市场规模约为 40 亿美元，而到下一个十年初，这一市场规模预计将超过 400 亿美元。**这一领域不仅涵盖娱乐行业，还包括媒体、营销和广告行业。

可以这样理解：过去我们回家看电视，只能看电视台播放的内容；进入数字时代后，我们有了 “点播” 功能，可以看自己想看的内容；而到本十年末，我们将进入 “交互式媒体” 时代 —— 不再是 “点播想看的内容”，而是所有娱乐互动都能通过视频实现交互。因此，“超长上下文处理能力” 的价值不言而喻。

每当英伟达发现这类 “高价值市场 + 技术突破点” 的机会 —— 比如如何提升输入上下文规模，我们就会进一步优化技术。或许我们可以为这些高价值、超长上下文场景专门设计解决方案，而不是让上下文处理和生成共用同一组 GPU—— 正是基于这一思路，我们在今天的 AI 大会上宣布了**一款全新的 Rubin 处理器，专门用于超长上下文处理。**

这款处理器名为 Rubin CPX GPU，是专门为 “百万级令牌处理” 这类高价值场景打造的 GPU，其核心优化方向就是上下文处理，同时也具备其他能力。这是一款全新的 Rubin GPU，基于与现有 Rubin 架构相同的技术，但属于全新产品形态：它的 NVFP4 精度算力超过

30 拍次 / 秒 (over 30 petaflops of NVFP4)，同时我们还大幅强化了注意力处理能力 —— 注意力机制是如今许多 AI 模型的核心构建模块。

我们为这款芯片新增了注意力加速核心，**其速度是当前 GB300 GPU 的 3 倍**；在内存方面也进行了优化：上下文处理的计算密集度高，对 HBM（高带宽内存）带宽、内存带宽以及跨节点扩展性的依赖较低，因此可以使用目前市场上大多数 GPU 都采用的标准 GDDR7 内存。

此外，我们还强化了视频处理能力：新增了 4 个视频编码器和 4 个视频解码器，用于处理和生成 AI 视频内容。这款处理器将于 2026 年底推出，紧随英伟达 Rubin 系列首款产品之后。

如何将这款单芯片 Rubin 集成到 Rubin 机柜中？今年 GTC 大会上我们发布了 Vera Rubin 机柜，单个机柜的 AI 算力超过 3.6 艾次 / 秒 (3.6 exaflops of AI performance)，将于 2026 年下半年上市。大家可以看到，机柜中的每个托盘都包含 4 块 Rubin GPU、多块 CPU 以及用于扩展互联的 ConnectX-9 —— 这是一个性能非常出色的平台：**单个机柜的算力是目前正在部署的 GB300 机柜的 3.3 倍，拥有 7075 太字节（terabytes）的高速内存和 1.4 拍字节（petabytes）的 HBM4 内存（HBM4）**，本身就是一款极具竞争力的机柜产品。

更重要的是，它采用与 GB300 相同的机柜架构 —— 这能帮助合作伙伴更便捷地部署，因为它在机械结构和空间占用上与 GB300 完全兼容。大家可以看到，单个机柜中集成了

72 块封装 GPU —— 由于这些 GPU 是双芯片设计（dual die），实际等效于 144 块 GPU，我们将其称为 NVL 144。

回到 CPX 处理器：我们可以直接将 CPX 集成到 Vera Rubin 平台中。事实上，在机柜底部有专门的区域，可以插入额外的上下文处理器，从而将机柜的百万级令牌处理能力大幅提升 —— 这就是 Vera Rubin NVL 144 CPX 机柜。我们在原有架构和托盘的基础上，在 Vera Rubin 机柜的 ConnectX-9 网卡后方插入了 8 块 Rubin CPX 处理器，整个机柜都能调用这些处理器进行上下文处理，从而显著提升机柜性能。

优化后，机柜的算力达到了 8 艾次 / 秒（eight exaflops），是当前 GB300 机柜的 7.5 倍；内存容量也再次提升，高速内存达到 1700 太字节（1.7 petabytes）—— 所有这些都能无缝适配现有机柜基础设施，让那些希望优先支持 “百万级令牌输入上下文” 的客户，能轻松升级或集成到现有数据中心中。

我们也可以不将 CPX 处理器集成到原有托盘，而是推出专门的 “CPX 计算托盘” —— 客户可以将其作为 “侧边设备”（sidecar），与 Vera Rubin 机柜并行部署。左侧的新托盘名为 VRCPX：大家可以看到，每个 VRCPX 托盘包含 2 块 Vera CPU 和 8 块 CPX 处理器，通过后台相同的网络连接。客户可以在数据中心中并行部署 VRCPX 机柜 —— 无论是 1:1、2:1 的比例，还是先部署部分再逐步扩展，都完全可行，而且无需将它们与 Vera Rubin 机柜相邻放置。

上下文处理与生成的工作逻辑是：只要生成第一个令牌，就只需将 KV Cache 发送到数据

中心中任何位置的令牌生成器即可 —— 这无疑是一次大幅升级，能显著提升速度。

目前，我们已与部分对 “超长上下文” 高度感兴趣的标杆客户展开合作，这些客户都是 AI 领域的创新者：比如 Cursor—— 智能代码生成领域的领军企业之一；还有专注于视频领域的公司，以及 Magic—— 我们正与他们合作，探索如何基于 CPX 处理器实现相关功能；此外还有 Runway，以及 Fireworks、Together AI 等领先的推理服务提供商—— 他们拥有最先进的模型服务加速技术，而 CPX 处理器能帮助他们实现 “百万级令牌输入” 的突破。

接下来，我们再梳理一下英伟达的芯片产品矩阵：在 Blackwell 架构下，我们有 Blackwell 和 Blackwell Ultra；此外还有 Gray CPU、Rubin 系列芯片、Spectrum-5 交换机芯片以及 ConnectX-8 网卡—— 所有这些芯片共同支撑 AI 和 AI 推理工作。这从来不是 “单一芯片” 的功劳，而是一个芯片家族的协同。如今，随着 Rubin CPX 处理器的加入 —— 这款专门优化上下文处理的 Rubin GPU，我们能与 Rubin 系列其他产品配合，实现 “百万级上下文处理”。未来，当我们接近 Feynman 架构时，还会分享更多细节。

所有这些要素必须协同工作：AI 服务的提供、数据集的构建，都离不开多类处理器的配合 —— 需要 CPU、GPU，以及不同层级的加速器；网络和基础设施的扩展也必须协同统一，才能支撑这些模型运行，最终实现令牌价值和推理营收的转化。这正是我们的关注重点：英伟达正以最快速度，将全套基础设施和软件栈推向市场。

当前的一个挑战是：如何构建未来的数据中心？我向大家展示了很多集成大量芯片的机柜，

但下一个挑战是 —— 未来的数据中心会是什么样子？英伟达是开放标准的坚定支持者 —— 我们是 OCP 的成员，已将 GB 系列机柜方案贡献给 OCP，未来也会将即将推出的基础设施方案纳入其中。

但现在的问题已上升到 “数据中心规模”：我们如何与社区合作，制定一套 “数据中心路线图” —— 而不仅仅是 “机柜和 GPU 路线图”。这套路线图需要面向未来，支持扩展和演进，同时突破发电、机械管道、汇流排设计、机柜排长度、CDU 等领域的技术限制。所有这些组件如何协同工作，才能让 “数据中心工厂” 高效运行？未来，无论是 Rubin、Rubin Ultra 还是 Feynman 架构，都需要这样的路线图支持扩展。

为此，我们启动了一项新计划 —— **“AI 工厂：数据中心规模参考设计”**。英伟达并非独自推进这项计划，而是与整个社区合作：包括 Cadence、Emerald AI、eTAB、Ge Verona、Schneider Electric、Siemens 和 Verdes—— 我们携手打造能支撑未来数据中心的冷却管道和电气系统，而这类数据中心建设周期长，亟需一套参考架构。

所有这些组件必须在 CDU 层面协同工作，数据中心的电力和运维也必须与 GPU 及计算基础设施无缝衔接，才能保证高可用性和高效率。目前我们正与这些合作伙伴推进项目，预计在下一届 GTC 大会上发布首版参考设计。

以上就是我今天的分享。感谢大家的聆听！