平衡 AI 推理成本与性能的艺术

电子书

本电子书是您应对当今快速发展的人工智能技术格局中基础设施复杂性的必备指南。本书专为负责人工智能的 IT 领导者量身定制，旨在揭示 AI 用例如何直接影响性能评估与基础设施优化。

AI 工作负载需要提供高性能、可靠性和效率的基础设施。但您如何判断现有基础设施是否能满足特定 AI 用例的需求？

本电子书将带您探索:

用例在性能评估中的作用：了解不同 AI 应用如何驱动差异化的基础设施需求。
AI 推理关键指标：掌握需要衡量的核心指标 —— 延迟、吞吐量、能效等，以确保成功。
基础设施优化最佳实践：获取可落地的策略，将技术栈与业务目标对齐。

通过洞见、框架与案例，本《IT 领导者指南：AI 推理与性能》电子书为您提供必要的知识，帮助您高效评估、部署和扩展 AI 解决方案 —— 是希望在 AI 时代自信决策的管理者的必读之作。

目录

推理性能的定义与重要性

AI 推理是将人工智能技术落地的关键环节，它通过解决复杂的部署挑战，释放实际应用价值 —— 从代码生成到长文档摘要，甚至生成图像和短视频。推理技术在自动化工作流程、创新商业模式和推动突破性产品方面的潜力具有变革意义 —— 尤其对负责在企业内部推动 AI 创新的首席信息官（CIO）和 IT 领导者而言。

然而，这一机遇也带来一个核心挑战：在扩展推理工作负载时，需在平衡延迟与性能的同时，控制持续性成本。

应对每 Token 成本挑战

在维护本地 AI 基础设施时，每次推理请求的收入必须足以覆盖硬件加速器、推理软件、持续优化与管理等成本。对于采用基础设施即服务（IaaS）的企业，财务成本则转移到按月/年计费的加速计算实例和推理软件即服务（SaaS）费用上。

无论采用何种部署模式，确保 AI 推理的 ROI（投资回报率）合理化这些成本，是 IT 领导者的核心关切。

我们的使命是通过优化 AI 性能并降低推理成本，帮助企业应对这些挑战。通过提升 AI 推理系统的效率，NVIDIA 使企业能够在满足不断变化的 AI 部署需求的同时，降低成本并保持可扩展性、性能与用户体验。

推理成本与两个核心因素紧密相关。第一是 AI 平台本身的性能（吞吐量）。系统处理请求的效率越高，组织越能通过分摊成本降低单次请求成本。然而，由于 AI 模型生成的 token 复杂度和类型差异显著，每 token 成本成为衡量系统性能与整体成本效率的关键指标。

然后，降低每 token 成本需与保持高质量的用户体验取得平衡。以牺牲用户体验为代价提升请求量，可能导致用户放弃使用 AI 应用或服务。糟糕的用户体验会引发客户流失，进而侵蚀收入。因此，必须同时衡量每 token 成本与系统延迟。

延迟通常通过两个维度衡量 —— 首 token 延迟（TTFT）和除首 token 外的每输出 token 延迟（TPOT）。正如后文将详细探讨的，优化这些维度往往存在权衡。

提升一个指标可能导致另一个指标恶化，因此企业必须找到正确的平衡点。

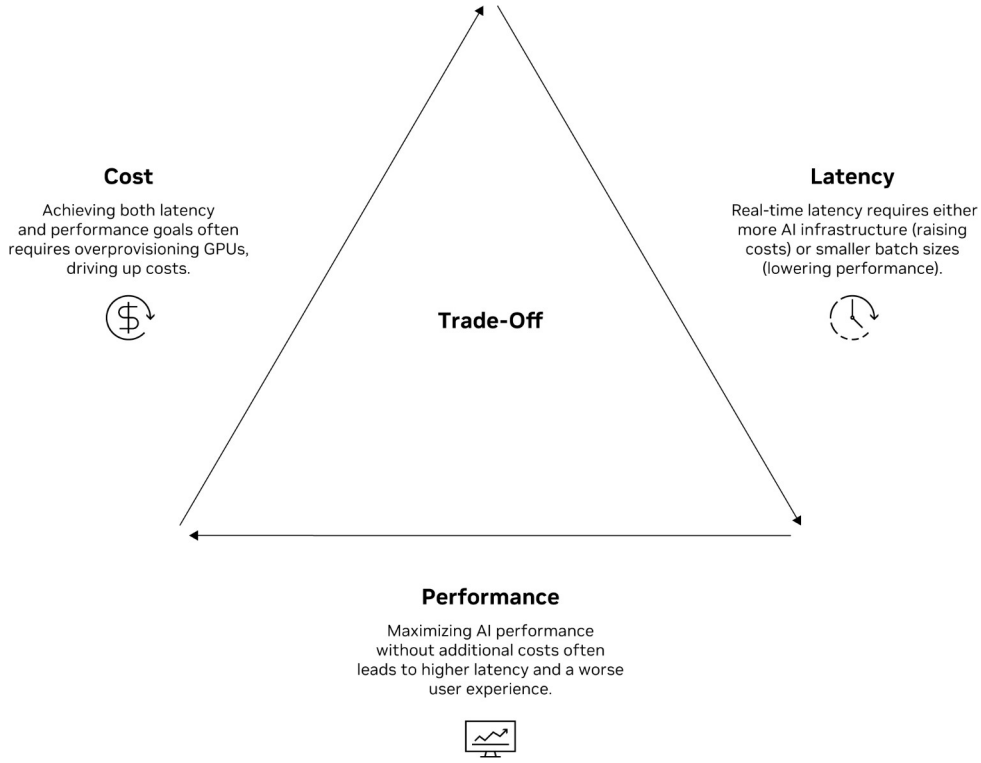


图1. 首 token 延迟与 token 间延迟的优化若未正确平衡，通常会产生三重权衡

为应对这些相互依赖关系，IT 领导者开始采用“有效吞吐量”（Goodput）作为衡量指标 —— 即系统在满足目标 TTFT 和 TPOT 时实现的吞吐量。这一指标使组织能够以更全面的方式评估性能，确保吞吐量、延迟与成本协同优化，同时支持运营效率与卓越的用户体验。

驱动推理成本的第二个关键因素是上市时间（time to market）。人工智能领域发展迅速，率先推出产品可带来显著竞争优势。

然而，若推理软件栈不可靠、测试不充分，或缺乏完善的文档与强大的社区支持，可能导致高昂的延迟成本、集成难题以及内部开发团队陡峭的学习曲线。

这些综合障碍会阻碍组织把握新兴机遇的能力。尽管量化这些问题对推理成本的确切影响具有挑战性，但其影响是可感知的，IT 领导者应将其纳入整体推理成本管理策略中。

在后续内容中，我们将深入探讨如何应对这些因素，以优化 AI 推理性能、降低成本，并最大化您的 AI 投资长期价值。

客户案例: Wealthsimple

图2. Wealthsimple 利用 NVIDIA AI 推理平台缩短模型交付时间并提升整体客户体验

Wealthsimple 是加拿大领先的投资管理公司，致力于优化机器学习（ML）模型部署以提升客户体验。该公司管理超过 500 亿美元资产，因缺乏标准化的 AI 推理平台，面临新模型部署的延迟问题。为解决这一挑战，Wealthsimple 采用 NVIDIA 的 AI 推理平台，取得了显著成果。

过去 12 个月中，Wealthsimple 部署了超过 30 个 AI 模型，生成 1.45 亿次预测。通过 NVIDIA 平台，该公司将模型交付时间从数月缩短至 15 分钟以内，显著提升了速度与效率。平台的可靠性也得到验证：过去一年半内未收到任何与 AI 推理相关的 IT 报修工单。

NVIDIA Triton 推理服务器增强了 Wealthsimple 提供机器学习驱动的个性化服务的能力，例如欺诈检测与交易优化，确保更快速且精准的金融流程。该平台还将正常运行时间从 95% 提升至 99.999%，基本上减少了客户的交易延迟并优化电子转账流程。

Wealthsimple 在云上利用 NVIDIA 加速计算高效运行模型，并依托 NVIDIA AI Enterprise 确保安全性与稳定性。这一转型使 Wealthsimple 能够提供领先的服务能力，同时最大限度减少停机时间，显著推动其基于机器学习的金融服务发展。

关键成果：

- 将模型部署时间从数月缩短至 15 分钟
- 推理正常运行时间从 95% 提升至 99.999%
- 推理服务延迟降低 20%
- 生成 1.45 亿次预测，且无任何 IT 报修工单

用例对推理的影响

在评估不同 AI 加速器性能时，仅依赖单一基准来制定硬件基础设施决策可能颇具吸引力。然而，这种过度简化的方法往往会导致系统成本随时间上升。

初始配置的系统可能无法满足特定用例的性能需求，从而迫使未来追加资源投资。这是因为不同用例的推理负载对加速计算资源和软件栈优化的需求各不相同 —— 这些因素对实现理想的总拥有成本（TCO）和确保最佳用户体验至关重要。

聊天机器人

以电商和客户服务流程中常见的大语言模型（LLM）驱动聊天机器人为例。这些聊天机器人需快速响应用户的简短咨询与问题，响应速度对防止用户流失至关重要。

在此场景中，TTFT —— 衡量聊天机器人开始输出响应的速度 —— 是提供积极用户体验的关键，而 TPOT 只需达到或略微超过人类阅读速度即可接受。典型的生产级聊天机器人服务超大用户群，采用大批次处理模式，将多个用户请求分组后通过单次请求发送至 AI 系统进行推理。

摘要生成

另一极端是文档摘要生成用例，该场景在多个行业迅速普及，例如医疗领域总结医学研究论文、新闻媒体提炼关键事件与进展，以及视频会议服务商从会议中生成行动项。

在这些场景中，输入序列通常较长，从数页到数百页不等，系统需优先快速生成完整摘要而非即时响应。因此，系统必须优化出更快的 TPOT，使其远超人类阅读速度，而用户对 TTFT 的容忍度更高 —— 尤其是在长输入序列场景下。为实现快速 TPOT，此类用例的批次规模通常较小，需通过高级软件栈优化在低批次规模下最大化吞吐量。

问答系统

问答机器人与聊天机器人有很多相似之处，尤其是在快速响应和处理简短用户查询。但其核心差异在于，问答机器人需从预定义知识库或数据集中生成答案。尽管用户查询本身可能简短，但发送给大语言模型（LLM）的实际推理请求更长，因其包含以嵌入数据形式从知识库中提取的相关信息。这一工作流被称为检索增强生成（RAG）。

在此场景中，选择能在长输入序列与中短输出序列下最优运行的 AI 系统与加速器，对实现最佳总拥有成本（TCO）和用户体验至关重要。此外，由于知识库数据通常存储于 GPU 内存之外的更大容量的 CPU 内存中，部署具备超越传统 PCIe 互联速度的 CPU-GPU 高速互联的系统，可进一步优化用户体验并降低 TCO。

AI 智能体

当前的 AI 聊天机器人利用生成式 AI 基于单次交互提供响应，通过自然语言处理回复用户查询。AI 的下一阶段进化是代理式 AI（Agentic AI），其超越了简单响应模式，通过高级推理与迭代规划解决复杂的多步骤问题。此类 AI 从多个来源吸收大量数据，使其能够独立分析挑战、制定策略并执行任务。通过持续的反馈学习与改进，代理式 AI 系统可增强决策能力与运营效率。

代理式 AI 需要调用函数并协调多个模型，这在推理过程中会产生额外 token。对于在组织内部部署 AI 智能体的领导者而言，选择合适的推理平台栈 —— 提供必要的抽象工具与蓝图以高效编排这些交互，对确保高性能推理操作至关重要。

如上所述，不同用例对输入序列长度、批次规模、TTFT、TPOT 以及 CPU 与 GPU 互联提出了差异化需求。在选择合适实例时，聚焦与目标生产用例高度匹配的基准测试，对维持一致的用户服务协议（SLA）和防止因硬件资源不足导致的基础设施成本上升至关重要。

客户案例: Perplexity

Perplexity AI 是一家基于人工智能的搜索引擎，每月处理超过 4.35 亿次查询，每项查询均涉及多个 AI 推理请求。为在成本与用户体验之间取得平衡，Perplexity 的推理团队采用了 NVIDIA 端到的推理部署解决方案，包括 Triton 推理服务器 和 TensorRT-LLM。该团队同时部署了 20 余个 AI 模型，包括多种 Llama 3.1 系列模型，并使用小型分类模型将用户请求匹配至合适的模型。其部署通过 Kubernetes 集群 管理，确保在流量波动时仍能满足 服务等级协议（SLA）。

Perplexity 团队专注于通过高效的调度与批处理策略优化性能与成本。他们利用 Triton 推理服务器 对传入请求进行批处理以优化 GPU 利用率，并根据需求动态扩展部署规模。团队持续评估和调整调度算法，以最小化 token 间延迟（Inter-Token Latency），提升对延迟敏感任务的性能。此外，通过 A/B 测试 精调配置，确保模型在满足 SLA 的同时最大化 GPU 资源利用率。

通过在内部部署 NVIDIA GPU，Perplexity 实现了显著的成本节约。其部署策略包括跨多块 GPU 并行化模型以及使用 TensorRT-LLM，使模型在严苛 SLA 要求下以最低成本运行。

影响推理性能的部署因素

在上一节中，我们探讨了推理解决方案中可测量和优化的不同性能指标，以及这些指标如何受特定用例的影响。在本节中，我们将聚焦部署因素与架构设计对性能的影响，包括以下主题：

- 硬件架构考量
- 模型规模
- 通过模型并发最大化利用率
- 模型中的多模态
- 通过高级批处理技术优化性能

本节的目标并非提供特定解决方案的部署或架构设计指南，而是为您提供设计高性能系统的必要知识。

硬件架构考量

选择合适的 GPU 是显著影响推理性能的关键因素。然而，仅依赖芯片或实例的峰值指标（如标称 FLOPS 或内存规格）可能无法全面反映实际表现，因为实际工作负载性能取决于多种因素，尤其是软件栈的效率。同样，GPU 的绝对价格或其在云端的每小时成本并非有意义的指标。真正重要的是 每瓦特每美元所交付的性能（Performance per Watt per Dollar）。这意味着，某块 GPU 虽然每小时成本高于另一块，但其 每 token 的整体成本可能反而更低。

数据中心和 AI 工厂通常受电力限制，因此需在固定功耗预算内最大化推理吞吐量。单位能耗下的推理吞吐量（即能效） 对最大化数据中心推理吞吐能力及最终收益潜力至关重要。

推理优化型 GPU 架构

自 2012 年在数据中心首次亮相以来，NVIDIA GPU 通过支持并行计算，显著缩短了资源密集型任务的耗时，彻底改变了行业格局。这一变革带来了显著提升：相比传统 CPU 系统，NVIDIA GPU 提供了 高达 30 倍的每瓦特性能 和 60 倍的每美元性能。

NVIDIA 持续推动创新边界，每代架构均实现性能突破。最新推出的 NVIDIA Blackwell 架构 相比前代，在万亿参数模型的推理速度上 提升达 30 倍。这一突破得益于 第二代 Transformer 引擎 的集成、更高速且更宽的 NVLink 互联技术 等革新，使得其性能相比上一代 提升达数量级。

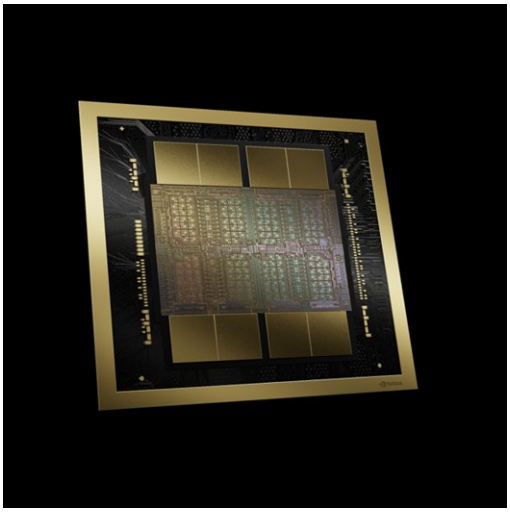


图3：NVIDIA Blackwell GPU 采用 2080 亿个晶体管，打造无与伦比的推理性能

平衡 AI 推理成本与性能的艺术

Blackwell GPU 采用 2080 亿个晶体管，是前代产品晶体管数量的 2.5 倍以上，成为目前体积最大的 GPU。该架构引入 第二代 Transformer 引擎，结合定制化的 Blackwell 张量核心 与 TensorRT-LLM，加速大语言模型（LLM）和专家混合模型（MoE）的推理任务。

Blackwell 张量核心支持全新的精度格式和微缩放格式，在提升计算吞吐量的同时优化了模型精度。该 Transformer 引擎通过微张量缩放技术实现 FP4 AI 精度，相比 FP8 精度，性能、HBM 带宽及每块 GPU 支持的模型规模平均提升一倍。

GPU 与能效型 CPU 的协同设计

NVIDIA Grace CPU：72 核高性能能效架构与突破性内存带宽。

NVIDIA Grace CPU 集成 72 个高性能且能效优化的 Arm Neoverse V2 核心，并通过 NVIDIA 可扩展一致性结构（SCF） 互联。该 SCF 是一种高带宽芯片内互连架构，提供 3.2 TB/s 的双工带宽，是传统 CPU 的 2 倍。

Grace 成为 首款采用高带宽 LPDDR5X 内存 的数据中心 CPU，通过纠错码（ECC）等机制实现服务器级可靠性。其 内存带宽高达 500 GB/s，而功耗仅为传统 DDR 内存的 1/5（成本相近）。

这些创新使 NVIDIA Grace 在性能、内存带宽及数据传输能力上表现卓越，并实现 突破性的每瓦特性能（Performance per Watt）。

推理阶段多 GPU 协同：以 “单一大 GPU” 模式运行

尽管选择合适的 GPU 对 AI 部署至关重要，但 百亿亿次计算（Exascale Computing）和 万亿参数 AI 模型的需求，要求 GPU 之间实现无缝通信，使多个 GPU 能在推理阶段协同工作，如同单一大型 GPU。

NVIDIA NVLink Switch 支持 130TB/s GPU 带宽，实现超大规模模型并行。结合 NVLink 与 NVLink Switch，多服务器集群可实现 1.8 TB/s 的互连带宽，显著扩展 GPU 通信与计算能力。

模型规模

在为工作负载部署选择大语言模型（LLM）时，一个重要的考量因素是模型的参数数量，即模型大小。随着时间推移，前沿 LLM 的参数数量持续增长，从而提升了模型的能力和响应质量。

平衡AI推理成本与性能的艺术

为满足广泛的用例、部署场景和预算需求，模型开发者通常会提供多种尺寸的模型变体。例如，Llama 3.1 开源模型家族提供了 8B、70B 和 405B 参数三种尺寸。通常，在同一模型家族中，更大的模型能提供更精准的响应。但需注意，较大模型生成 token 所需的计算资源也显著高于较小模型。

这意味着模型大小的选择需同时权衡目标用例和可用计算资源。对于需要复杂任务最高响应精度的场景，较大模型可能是首选；而在对输出速度要求严苛或计算资源有限的场景中，选择较小模型可能更为合适。

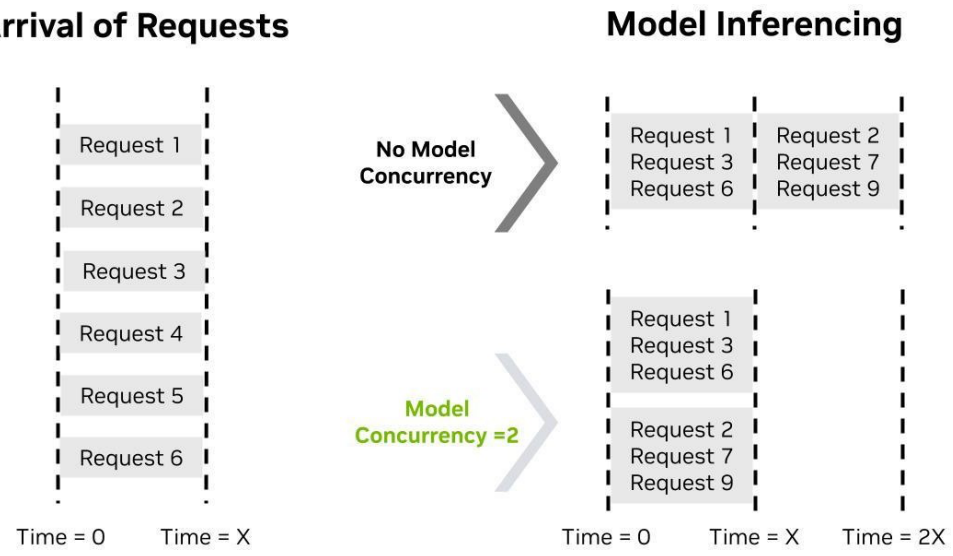


图 4：模型并发度提升对吞吐量与延迟的影响

NVIDIA Dynamo-Triton 简化了启用并发的模型部署流程。开发者只需在模型配置文件中开启一个标志位（flag），即可轻松实现模型并发。随后，他们可指定需部署的并发模型实例数量及目标 GPU。这种简化设计使开发者无需复杂配置或手动扩展流程，即可充分利用 GPU 基础设施的完整性能。

在需同时服务多个独立模型但无法全部同时容纳于同一实例或节点的场景中，NVIDIA Dynamo-Triton 可作为多路复用器运行。它会根据实时用户请求动态加载和卸载模型，确保所需模型在需要时即时可用。这使组织无需额外基础设施投资，即可维持高服务可用性与性能。

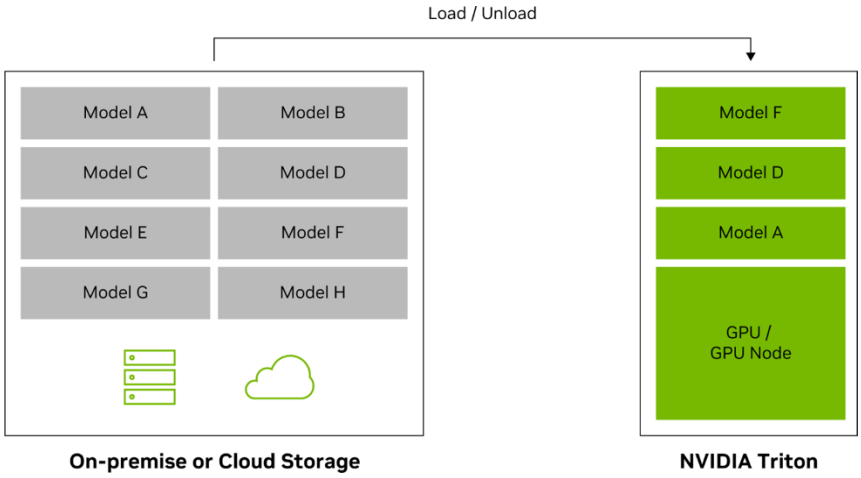


图 5：NVIDIA Dynamo-Triton 如何在 GPU 上动态加载与卸载模型，提升服务可用性与性能

除了提升吞吐量并降低成本外，模型并发还能应对 用户需求不可预测 的场景。IT 团队可从单个模型实例起步，根据需求增长逐步扩展为多个并发实例。通过调整并发模型数量，既能满足波动需求，又能最小化未使用计算资源的成本。

模型并发是一个强大的推理性能工具，尤其是在数据中心 GPU 中部署中、小规模 AI 模型的时候。采用此方法后，组织将获得以下显著优势：

- 提升吞吐量：并行运行多个模型或模型实例可增强系统整体吞吐量，使其能够同时处理更多推理请求。
- 降低每 token 成本：通过充分利用现有资源，模型并发减少了每个服务 token 的成本，确保基础设施投资的最大化。
- 降低延迟：并发运行模型或模型实例可减少等待时间，提升服务响应速度并改善用户体验。

模型中的多模态

能够同时处理和整合来自多种数据模态（如文本、图像、音频和视频）信息的 AI 模型被称为多模态模型。与仅处理单一输入模态的传统单模态 AI 模型不同，多模态模型具有复杂的推理服务需求，必须能够跨多种模态处理输入并生成输出。

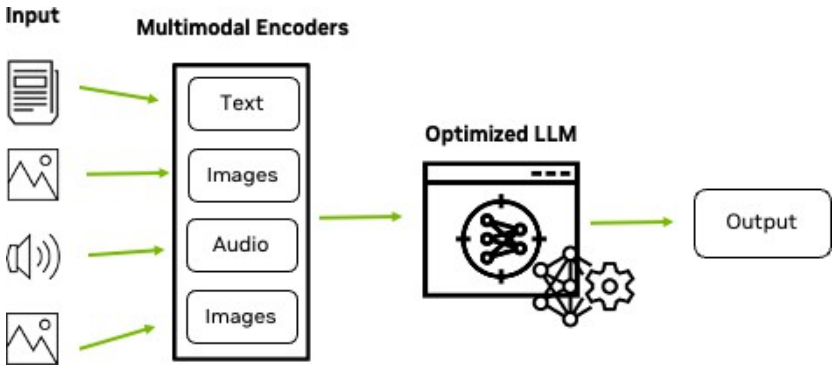


图6：多模态 LLM 融合多种模态信息并优化高吞吐量输出

NVIDIA AI 推理平台为多模态模型提供优化支持，通过 NVIDIA TensorRT 推理加速库实现高性能的音频和图像输入编码器，并通过 TensorRT-LLM 推理加速库优化文本解码器。这些模型可借助 NVIDIA Dynamo-Triton 部署，其支持多项高级功能，例如：

- 多图像推理（multi-image inference）：单个请求可包含多张图像；
- 高效 KV 缓存复用（KV cache reuse）：图像可在多个文本输入 token 间共享，从而减少延迟，并根据 KV 缓存复用的长度比例提升性能。

通过高级批处理技术优化性能

数据中心 AI 加速器（如 NVIDIA GPU）利用多个核心（Core）并行处理请求。为最大化这些核心的潜力，请求通常会被分组为批次并发送至推理。然而，批大小（batch size）、吞吐量（throughput）和延迟（latency）之间存在关键权衡关系。调整其中任意两个因素可能对第三个产生负面影响，这对寻求最佳性能的企业构成挑战。

例如，较小的批大小可降低延迟，但会导致 GPU 资源未充分利用，从而产生效率低下和更高的运营成本。相反，较大的批大小可最大化 GPU 资源利用率，但会增加延迟，可能损害用户体验。

IT 团队通常通过 A/B 测试确定其用例的理想平衡点，但这可能是一个复杂且耗时的过程。借助 NVIDIA Dynamo-Triton 和 NVIDIA TensorRT-LLM 等高级推理软件平台，

可缓解这一权衡问题。NVIDIA AI 推理平台采用复杂的批处理技术，帮助组织在吞吐量与延迟之间实现更优平衡，从而提升整体系统性能。

动态批处理（Dynamic Batching）

动态批处理根据特定标准（如最大或首选批大小、最大等待时间）创建批次。如果能以首选大小形成批次，则按该大小处理；否则，系统将按配置的最大批大小生成尽可能大的批次。通过允许请求在队列中短暂停留，动态批处理可等待更多请求到达以形成更大批次，从而提升吞吐量和资源利用率，最终降低每 token 成本。

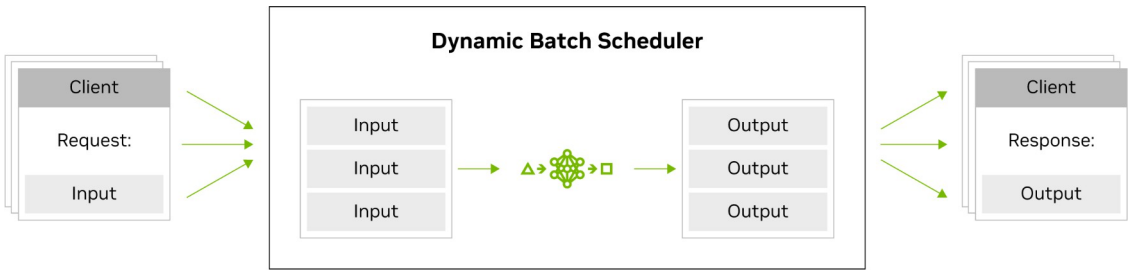


图 7：动态批处理如何创建多个请求的批次

序列批处理（Sequence Batching）

对于需要特定请求顺序的用例（例如视频流），序列批处理确保相关请求（例如视频中的帧）以有意义的顺序进行处理。当请求存在关联性时（例如视频流中模型需要在帧之间保持状态），序列批处理会确保这些请求在相同实例上按顺序处理。这既保证了输出的正确性，又优化了性能。

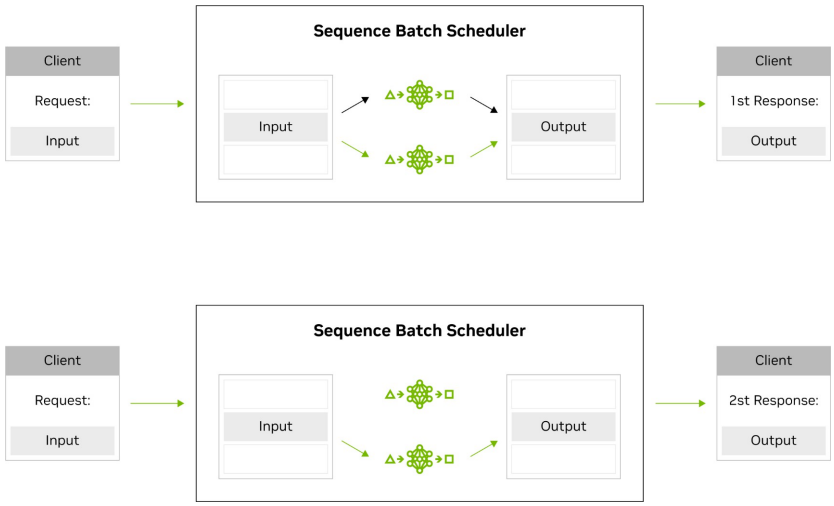


图 8：序列批处理如何在保持请求关系与顺序的同时创建不同请求的批次

实时批处理（Inflight Batching）

实时批处理通过实现连续的请求处理增强了传统批处理技术。借助实时批处理，TensorRT-LLM 运行时（NVIDIA 用于优化大语言模型推理的 SDK）会在请求就绪后立即处理，而无需等待整个批次完成。这通过在其他请求仍在处理时立即启动新请求，实现了更快的处理速度和更高的吞吐量。

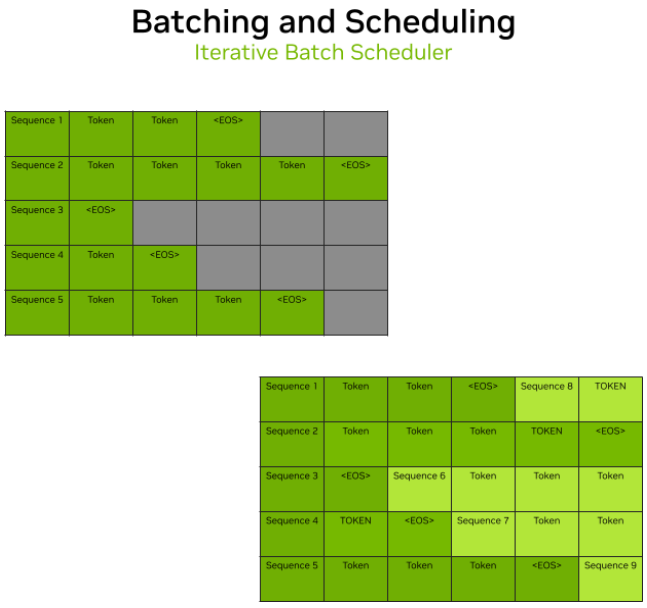


图9：实时批处理如何在批次中完成请求后立即替换为新请求

高效利用计算资源需要对批大小、吞吐量和延迟进行精细管理。NVIDIA AI 推理平台提供的高级批处理技术（如动态批处理、序列批处理和实时批处理）为组织提供了灵活性，可在优化性能、降低成本和提升用户体验之间取得平衡。这些技术帮助在吞吐量与延迟之间实现更优权衡，使企业能够更高效地扩展 AI 驱动的解决方案。

大模型推理的并行化：关键方法与权衡

随着大语言模型（LLMs）的规模和复杂性不断增加，确保这些模型能够高效地跨多个 GPU 部署变得至关重要。当模型的内存需求超过单个 GPU 的容量时，需采用并行化技术对工作负载进行拆分并优化性能。对于企业高管和 IT 领导者而言，理解主要的并行化方法及其对吞吐量与用户交互性的影响，是做出基础设施决策的关键。以下介绍了大模型推理的主要并行化方法，并结合领先案例进行说明。

数据并行（Data Parallelism, DP）
数据并行通过在多个 GPU 或 GPU 集群上复制模型实现。每个 GPU 独立处理用户请求组，确保这些请求组之间无需通信。该方法的扩展性与 GPU 数量呈线性关系，即分配的 GPU 资源越多，可处理的请求量直接增加。

然而需注意，单独的数据并行通常无法满足最新大规模 LLM 的需求，因为其模型权重往往无法适配单个 GPU。因此，DP 通常需与其他并行化技术结合使用。

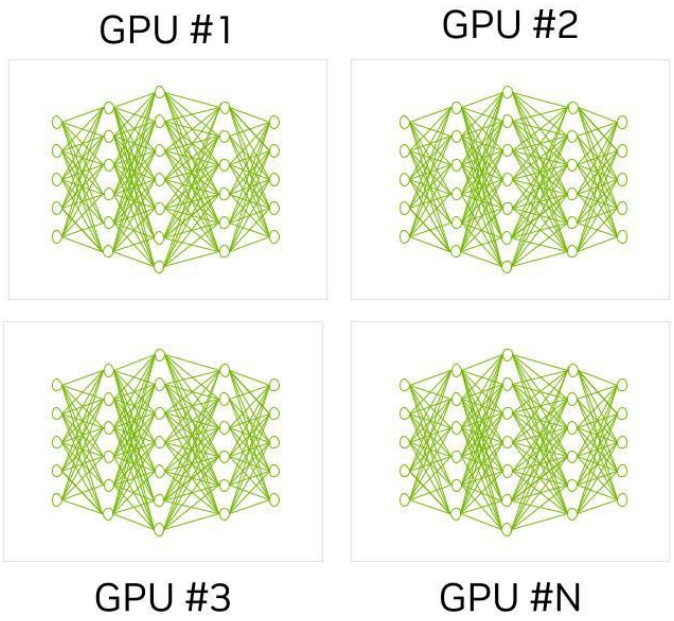


图10：在多个 GPU 上应用数据并行

性能影响：

- GPU 吞吐量（GPU Throughput）：不受数据并行单独影响，因为模型被复制了。
- 用户交互性（User Interactivity）：无显著影响，因为请求独立处理。

张量并行（Tensor Parallelism, TP）
在张量并行中，模型参数被拆分至多个 GPU，用户请求在 GPU 或 GPU 集群间共享。不同 GPU 上的计算结果通过 GPU 间网络（GPU-to-GPU network）进行整合。该方法可通过为每个请求分配更多 GPU 资源（从而加快处理速度）提升用户交互性，尤其适用于基于 Transformer 的模型（如 Llama 405B 和 GPT4 1.8T MoE（1.8 万亿参数专家混合模型））。

性能影响：

- GPU 吞吐量（GPU Throughput）：在无高速互联（如 NVIDIA NVLink）的情况下，将张量并行扩展至大规模 GPU 数量可能会因通信开销增加而降低吞吐量。
- 用户交互性（User Interactivity）：增强，因为每个请求分配更多 GPU 资源，从而加快处理速度。

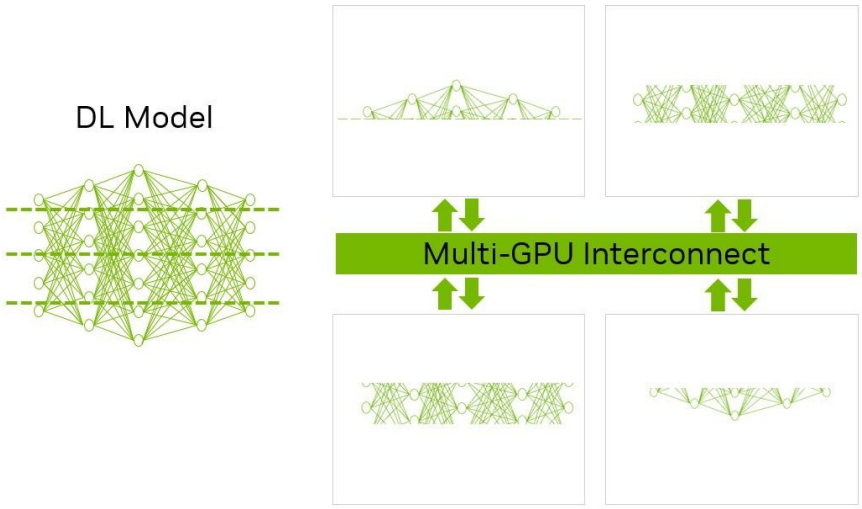


图11：在深度神经网络上应用流水线并行

流水线并行（Pipeline Parallelism, PP）

流水线并行将模型层划分为多个组，每组分配至不同的 GPU。模型按顺序跨 GPU 处理请求，每个 GPU 对其分配的模型部分执行计算。该方法适用于单个 GPU 无法容纳的大型模型权重，但在效率方面存在一定局限性。

性能影响：

- GPU 吞吐量（GPU Throughput）：因模型权重分布可能导致效率降低。
- 用户交互性（User Interactivity）：无法显著优化用户交互性，因为处理必须跨 GPU 按顺序进行。

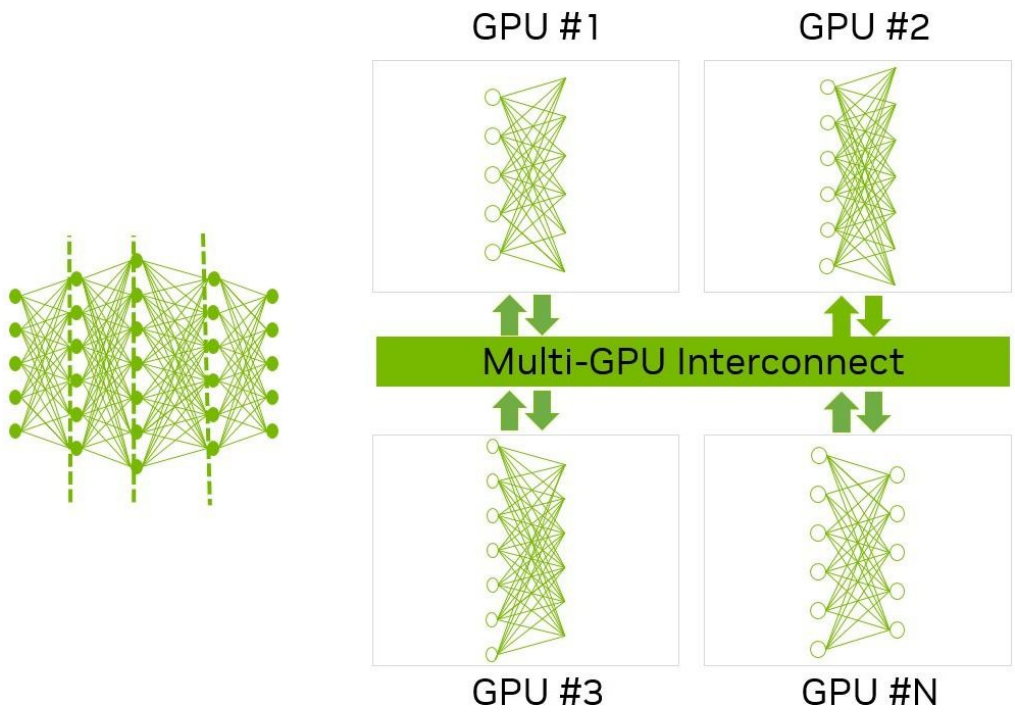


图 12：在深度神经网络上应用流水线并行

专家并行（Expert Parallelism, EP）

专家并行（expert parallelism）通过将用户请求路由至模型内的专业化“专家”（expert）实现处理。通过限制每个请求仅涉及较少的模型参数（即特定专家），系统可降低计算开销并优化处理效率。请求由独立专家处理后，在最终输出阶段重新组合，这需要高带宽的 GPU-to-GPU 通信。

性能影响：

- GPU 吞吐量（GPU Throughput）：专家并行可在与其他技术结合时显著提升吞吐量。
- 用户交互性（User Interactivity）：EP 配置可在避免其他方法中常见权衡问题的前提下，维持或提升用户交互性。

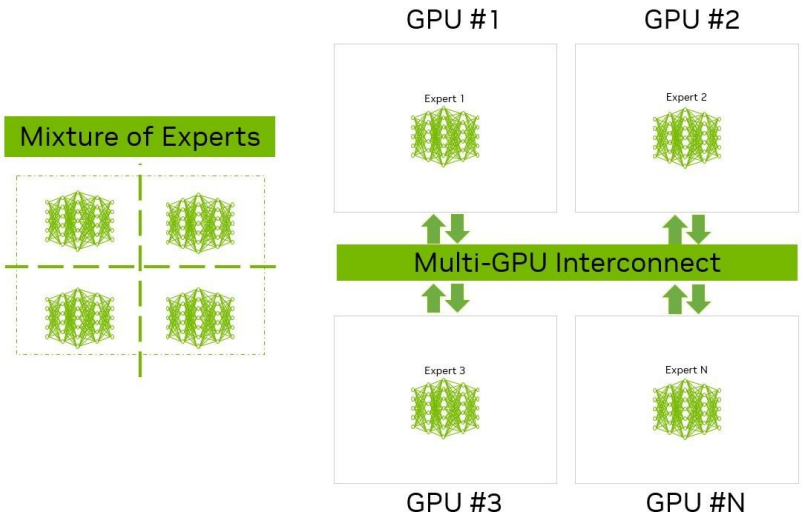


图 13：在由四个专家组成的深度神经网络上应用专家并行

结合并行方法以实现最佳性能

在多个 GPU 上优化 LLM 性能的最有效方式是结合多种并行化技术。通过这种方式，吞吐量与用户交互性之间的权衡可被最小化，从而确保高性能与响应式用户体验并存。

例如，在部署具有 16 个专家的 GPT 1.8T MoE 模型时，使用 64 块 GPU（每块配备 192 GB 显存），我们观察到：与仅使用专家并行相比，专家 + 流水线并行（EP16PP4）使用户交互性提升 2 倍，且 GPU 吞吐量的降低可忽略不计；而张量 + 专家 + 流水线并行（TP=4，EP=4，PP=4）在保持用户交互性的同时，GPU 吞吐量比仅使用张量并行提升了 3 倍。

对于负责 LLM 部署的首席信息官（CIO）和 IT 领导者而言，理解各种并行化策略的权衡与优势至关重要。结合数据、张量、流水线和专家并行等方法，组织可优化 GPU 资源利用率，同时提升吞吐量与用户交互性。实践中，精心设计的配置能在跨多 GPU 的高吞吐量与响应式用户体验之间实现平衡。随着对更大、更复杂模型的需求增长，这些并行化技术将在确保 AI 模型可扩展性与高效部署中发挥关键作用。

在云中解锁 AI 推理性能与成本效率

云计算长期以来一直是企业的变革性力量，提供可扩展、高效且全球可访问的 IT 基础设施。随着企业将存储、数据库和企业资源规划等工作负载迁移至云端以实现灵活性和成本节约，AI 推理也正沿着这一路径发展。根据 IDC 最近的一份报告，企业中 AI 推理增长最快的领域将是云中的加速计算实例，其采用速度预计将以三倍于其他部署模式的速度增长。

对于 IT 领导者而言，这一转变带来了在不增加成本的前提下提升性能的重大机遇。通过将基于云的加速计算与优化的 AI 软件相结合，组织可实现更高的速度、可扩展性和成本效率。为优化云中的 AI 推理性能，IT 领导者必须确保团队具备灵活性，能够从广泛的 GPU 类型中选择，并为每种特定工作负载匹配最合适的 GPU。这一战略选择是避免加速计算资源过度配置或不足配置、充分发挥云中 AI 推理潜力的关键。

同时，选择一个与 IT 团队已接受培训并认证的云工具深度集成的 AI 推理堆栈同样重要。这种集成可最小化培训成本，使团队能够专注于推理性能优化，而非处理复杂的软件集成任务。为进一步提升灵活性并避免供应商锁定（vendor lock-in），IT 领导者应优先选择在所有主要云服务提供商（CSPs）上均可提供的 GPU 和推理堆栈。这不仅能够将 GPU 部署在更接近终端用户的位置，还赋予 AI 推理团队自由选择最适合其特定应用场景的云平台的权利。

在云中使用 NVIDIA 加速计算解决方案的一个显著优势在于其提供的灵活性和选择权。所有主要云服务提供商的云计算实例均搭载了不同代的 NVIDIA GPU，为企业和开发者提供了广泛的选择，帮助其选择最符合需求的加速计算解决方案。

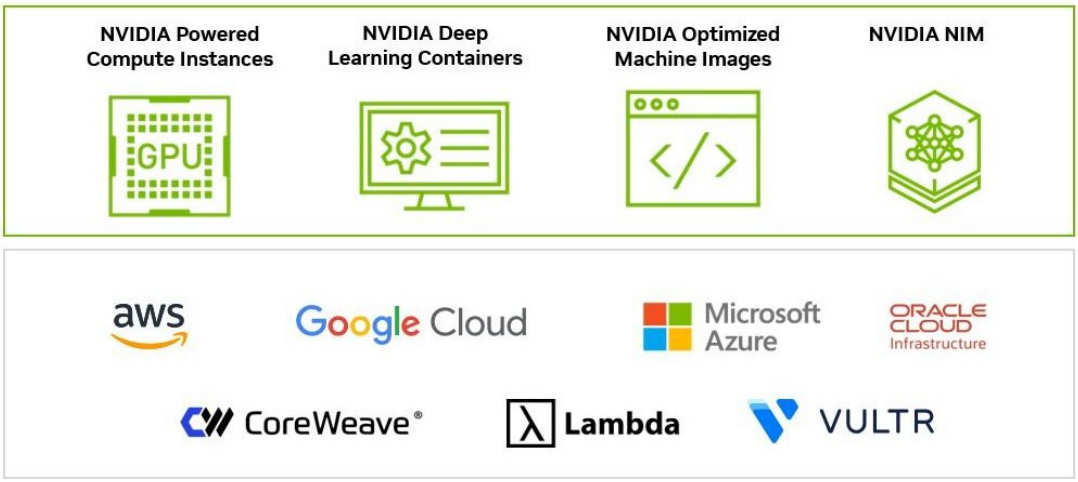


图 14：NVIDIA AI 推理平台在云中的广泛支持

此外，NVIDIA AI 推理堆栈已与所有主流云 MLOps 工具完全集成，简化了 AI 工作负载的部署与扩展。例如，NVIDIA Dynamo-Triton 和 NVIDIA NIM 已与所有主流云 AI 工具集成，允许 IT 团队以最低的代码量或无需代码的方式快速部署模型。无论是使用 AWS SageMaker、Google Vertex AI、Azure ML，还是 OCI 的 Data Science Cloud，企业均可一键启动 NVIDIA Dynamo-Triton 和 NVIDIA NIM，节省时间和 IT 资源。

对于寻求更简化部署方式的组织，NVIDIA 还提供预配置的云镜像，其中所有必要的驱动程序和依赖项均已预装。这使得 AI 工作负载可快速部署而无需手动配置，帮助企业节省部署时间和成本。

通过利用 NVIDIA 的加速计算与集成式 AI 推理解决方案，IT 领导者可确保其 AI 推理性能在云中高效扩展，兼顾灵活性、成本效率与最小化运维开销。

客户案例研究：Amdocs
Amdocs 是通信和媒体行业软件与服务的领先提供商，其开发的 amAIz 是一款专为电信行业定制的生成式 AI 平台。为提升性能并加速部署，Amdocs 在 NVIDIA DGX Cloud 上采用 NVIDIA NIM 推理微服务，实现对实时客户服务与销售咨询场景的微调大语言模型（LLMs）的无缝部署。

Amdocs 从匿名通话记录、客户交互记录和账单中整理了人工标注数据集，并采用参数高效微调技术（如 Meta 的 Llama 系列模型 和 Mistral AI 的 Mixtral 8x7B 的低秩适应（LoRA））。团队通过 NIM 自动化微调模型部署，实现延迟降低 80%，支持近实时响应，同时准确性提升 30%。通过领域定制化微调，Token 消耗量减少 40%，从而降低基础设施成本。

借助 DGX Cloud 上的 NVIDIA NIM，Amdocs 显著提升了 LLM 的准确性与响应速度并降低成本，使 amAIz 平台能够满足现代电信应用的需求。

扩展 AI 推理以应对波动需求
随着 IT 领导者推进 AI 应用落地，最棘手的决策之一是预测用户需求及其波动趋势。这些预测直接影响基础设施决策，尤其是 GPU 等预置资源（provisioned resources）的分配，进而影响成本与性能。平衡这些要素是确保 AI 系统动态扩展同时控制运维开销的关键。

AI 应用（尤其是基于优化后 LLMs 的应用）需要强大的计算能力以支持实时推理。常见的扩展方式是根据预估的峰值需求配置 GPU 资源，但这种方法在峰值需求短暂或不规律时可能效率低下。为应对这一挑战，组织需要灵活调整基础设施规模以响应需求波动。Kubernetes 提供了理想的解决方案，使 IT 团队能够动态扩展 LLM 的部署规模。无论是从单个 GPU 扩展至多个 GPU，还是在低峰期减少资源，Kubernetes 均能以成本效益和性能优化的方式管理基础设施。

Kubernetes 与 NVIDIA AI 推理平台的灵活性
对于电商或客服等行业的企业而言，推理请求量波动频繁。无论是促销期间处理大量

客户咨询，还是深夜时段处理较少请求，高效扩展能力至关重要。通过 Kubernetes 扩展，组织可动态调整所需 GPU 数量，确保硬件资源与实时需求精准匹配。相比为峰值负载过度配置硬件资源，这种方式可节省显著成本。

NVIDIA AI 推理平台（包括 NVIDIA Dynamo-Triton 和 NVIDIA NIM）支持 Kubernetes，以简化扩展流程。当推理请求数量增加时，Prometheus 会从 NVIDIA Dynamo-Triton 抓取指标，并将其发送至 Kubernetes 水平 Pod 自动扩缩器（HPA），后者会向部署中添加更多 Pod，每个 Pod 配备一个或多个 GPU。

一个可从 NVIDIA Dynamo-Triton 抓取并通过 Prometheus 发送至 Kubernetes HPA 以指导扩缩决策的自定义指标是 队列计算比（queue-to-compute ratio）。该比率反映了推理请求的响应时间，其定义为：队列时间 ÷ 推理请求的计算时间。

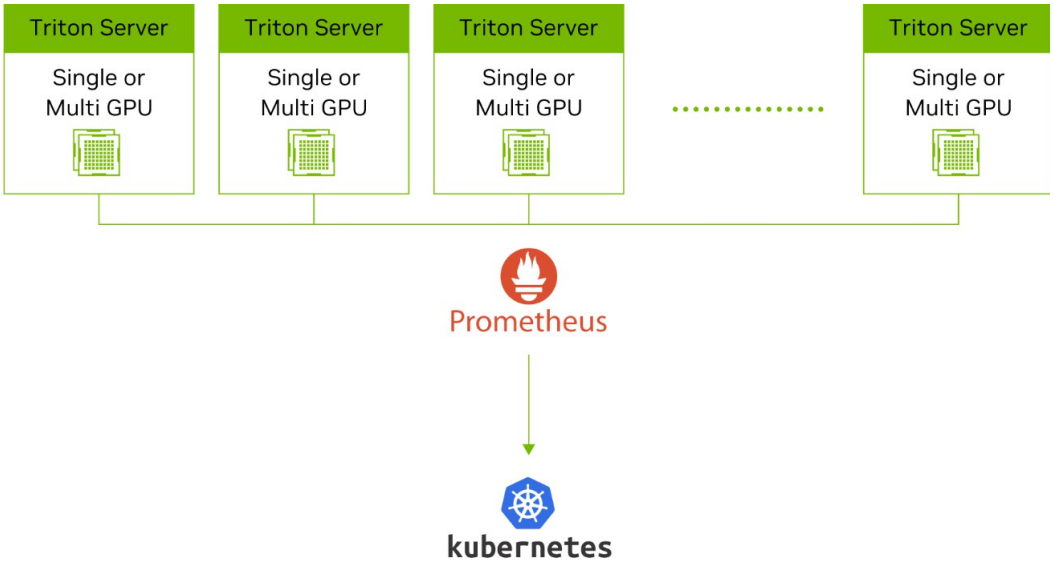


图 15：基于 Kubernetes 架构的 NVIDIA Triton 推理服务器扩展

可扩展 AI 部署的关键优势：

- 成本效率：根据需求扩展可帮助企业优化基础设施成本，确保仅按需配置处理实时请求所需的硬件资源。这种动态扩展能力减少了过度配置（over-provisioning）的需求，避免高昂的硬件开销。

- 性能优化：Kubernetes 与 NVIDIA Triton 推理服务器通过动态平衡 GPU 资源，确保即使在高负载（heavy load）期间仍能保持高性能。这使得 AI 应用能够满足严格的延迟和准确性要求，尤其在面向客户的场景中至关重要。

- 跨行业灵活性：无论是应对在线购物活动期间的流量激增，还是管理呼叫中心中波动的请求量，可扩展的 AI 部署提供了广泛的灵活性，可满足 多行业（如电商、客服等）的多样化需求。

如需详细了解如何使用 Kubernetes 优化 NVIDIA Dynamo-Triton 和 AI 部署，请访问技术博客《利用 NVIDIA Triton 和 NVIDIA TensorRT-LLM 及 Kubernetes 实现 LLM 扩展》。

如需详细了解如何使用 Kubernetes 优化 NVIDIA NIM 和 AI 部署，请访问技术博客《在 Kubernetes 上实现 NVIDIA NIM 微服务的横向自动扩展》和《使用 NVIDIA NIM Operator 在 Kubernetes 上管理 AI 推理流程》。

通过模型集成简化 AI 管道

AI 和 ML 系统很少以单一模型独立部署，通常作为更复杂管道的一部分，整合多个模型、预处理步骤和后处理任务。这种 模型集成（Model Ensembles）方法已成为现代 AI 工作流的核心，尤其在企业级应用中。

模型集成指将多个机器学习模型、预处理和后处理步骤整合为统一的管道。与独立运行模型不同，集成技术使这些组件协同工作，从而简化流程并提升整体效率。每个模型通常负责任务的不同环节，例如数据处理、生成预测或优化输出。

在实际场景中（如企业应用），一个 AI 管道可能包含多个计算机视觉模型、图像编辑工具及其他专用模型。例如，使用生成式 AI 模型 SDXL 创建营销活动时，流程可能包括：

- 初始预处理：放大图像以聚焦特定区域；
- 模型生成：通过 SDXL 生成场景；
- 后处理：提升图像分辨率以适配高清显示。

通过将这些步骤无缝串联为统一的管道，企业可确保数据在模型间高效流转，同时降低延迟并优化资源使用。

NVIDIA Dynamo-Triton 推理服务器在模型集成中的作用

构建和管理这类复杂模型管道颇具挑战，但 NVIDIA AI 推理平台工具（如 NVIDIA Dynamo-Triton）提供了强大的自动化解决方案，可高效协调模型集成。

Dynamo-Triton 的模型集成功能 无需手动编写代码管理各步骤，显著降低了复杂性并减少了出错风险。此外，它支持在 CPU 上运行预处理和后处理，同时将核心 AI 模型部署在 GPU 上，灵活平衡计算资源。

更进一步，Dynamo-Triton 允许在集成中添加高级特性，例如 条件判定逻辑（conditional logic）和 循环（loops），使开发者能够构建更复杂、灵活的 AI 工作流。例如，根据输入数据动态选择模型分支，或通过循环优化迭代任务，从而实现更智能的端到端解决方案。

使用 Dynamo-Triton 实现模型集成的优势
将模型集成与 Dynamo-Triton 结合使用可带来以下关键优势：

- **降低延迟：**通过将预处理、推理和后处理整合为单一管道，企业可显著减少模型间的数据传输时间。这对生成式 AI 和计算机视觉等实时应用至关重要。
- **优化资源利用：**模型集成允许团队优化资源部署。例如，将轻量级的预处理和后处理任务运行在 CPU 上，同时将 GPU 资源保留给密集的模型任务，从而在不牺牲性能的前提下实现成本效益的扩展。
- **减少网络开销：**传统机器学习管道中，每个模型需通过网络通信传输中间数据。而模型集成可最小化这类中间数据传输，减少网络调用次数并优化带宽使用。
- **无代码集成：**Dynamo-Triton 的模型集成功能提供低代码或无代码方式连接不同 AI 模型。这使 IT 领导者和推理团队能够轻松将预处理和后处理工作流（通常基于 Python 构建）集成到统一的管道中。通过单个推理请求即可触发整个流程，进一步提升效率。

实际应用示例：AI 驱动的生成式应用
以文本生成图像的生成式 AI 应用为例：输入文本被转换为合成图像。典型管道包含两个核心组件：

- LLM（大语言模型）：对输入文本进行编码；
- 扩散模型（Diffusion Model）：生成图像。

流程步骤：

- 预处理：对输入文本进行清洗、分词或格式化以适配 LLM；

- 文本编码：LLM 将文本转换为潜在表示；
- 图像生成：扩散模型基于编码生成图像；
- 后处理：调整图像分辨率或添加特效以适配最终应用需求。

通过模型集成，Dynamo-Triton 可无缝串联从文本预处理到图像生成的全流程，甚至自动处理后处理步骤，形成统一的高效管道。例如，用户输入“一只在森林中的发光狐狸”后，系统可一键生成高清图像并优化显示效果，全程无需人工干预。

优势体现：

- 延迟优化：数据在管道内直接流转，避免多次网络传输；
- 资源灵活分配：CPU 处理文本预处理，GPU 专注图像生成；
- 开发效率提升：通过 Dynamo-Triton 无代码配置可集成全流程，缩短开发周期。

模型集成在 AI 优化的威力

通过将多个模型及预处理、后处理步骤整合为单一管道，IT 领导者可显著提升 AI 应用效率，降低延迟并优化资源使用。

NVIDIA AI 推理平台为此类集成提供了强大支持，其灵活的资源分配能力简化了组件整合流程。凭借低代码开发模式与自动化优化，团队可更轻松地构建和扩展 AI 系统，使其成为企业提升模型性能的关键工具。

案例研究：Let’s Enhance 将产品照片转化为营销资产

AI 初创公司 Let’s Enhance 是成功应用 NVIDIA AI 推理平台部署 SDXL 模型的典范。该公司已使用 Triton 推理服务器在 NVIDIA Tensor Core GPU 上部署超过 30 个 AI 模型，持续运行三年以上。

近期，Let’s Enhance 推出新产品 AI Photoshoot，利用 SDXL 模型将普通产品照片转化为适用于电商网站和营销活动的视觉资产。

通过 Triton 推理服务器对多框架和后端的强大支持，以及动态批处理（dynamic batching）等核心功能，公司创始人兼首席技术官 Vlad Pranskevichus 仅需少量机器学习工程师团队参与，便成功将 SDXL 模型无缝集成至现有 AI 管道，释放团队精力用于研发创新。

在成功完成概念验证后，该公司通过将 SDXL 模型迁移至 Google Cloud G2 实例的 NVIDIA GPU，实现成本降低 30%。

提升推理性能的高级技术

在上一章节中，我们探讨了基础的推理性能优化策略。本章将深入更高级的技术，针对那些将 AI 推理视为企业营收战略核心的 IT 领导者，或已完成系统基础优化、希望从AI投资中挖掘下一阶段性能提升的技术负责人。

我们的目标是概述 NVIDIA AI 推理平台中的前沿优化技术，同时指导领导者如何引导团队进一步学习。我们将探索 NVIDIA Dynamo 与 TensorRT-LLM 提供的高级策略，包括分块预填充（Chunked Prefill）、多块注意力（Multiblock Attention）、KV 缓存预复用（KV Cache Early Reuse）、P/D 分离服务（Disaggregated Serving）和投机采样（Speculative Decoding），这些技术均旨在突破生产环境中的 AI 推理边界。

分块预填充：最大化 GPU 利用率并降低延迟

大语言模型（LLM）推理通常包含两个关键阶段：预填充（prefill）和解码（decode）。在预填充阶段，系统针对用户的输入来计算上下文理解（KV 缓存），这一阶段计算密集；随后进入解码阶段，按顺序生成 token，首个 token 基于 KV 缓存生成。然而，传统方法在预填充阶段的高计算需求与解码阶段的较低负载之间难以平衡。

分块预填充（Chunked Prefill）是一种优化技术，通过将预填充阶段拆分为更小的分块（chunk），提升与解码阶段的并行性，减少瓶颈。这种方法有助于处理更长的上下文和更高并发请求，同时充分利用 GPU 的内存和计算资源。

它还通过将内存使用与输入序列长度分离，提供了灵活性，使处理大规模请求时不易受内存容量限制。TensorRT-LLM 的动态分块大小调整功能可根据 GPU 利用率智能调节分块大小，简化部署并消除手动配置需求。

多块注意力：通过长上下文提升吞吐量

随着 AI 模型的发展，上下文窗口（允许更全面的认知理解）的规模呈指数级增长。Llama 2 的上下文窗口为 4K token，而 Llama 3.1 已扩展至惊人的 128K token。在实时推理场景中处理这些长序列时，尤其是在 GPU 资源分配方面，存在独特的挑战。

TensorRT-LLM 中的多块注意力（Multiblock Attention）技术通过将解码过程高效分配到 GPU 的所有流式多处理器（SM）上，解决了这些挑战。

这一方法显著提升了吞吐量，并减少了因大型 KV 缓存规模导致的瓶颈。即使在实时应用中常见的较小批量（batch size）场景下，该技术也能充分利用 GPU 资源，从而在 NVIDIA HGX GPU 平台上实现每秒生成 token 数量提升最高达 3.5 倍。这种性能提升并未以牺牲 TTFT 为代价，确保了复杂查询仍能快速响应。

KV 缓存预复用：缩短 TTFT

KV 缓存的生成是一个计算密集型过程。因此，通过避免重复计算，KV 缓存复用在加速 token 生成中起到关键作用。然而，传统方法通常需要等待整个 KV 缓存生成完成后才能复用，这在高负载场景中导致效率低下。

KV 缓存预复用（KV Cache Early Reuse）通过允许在缓存生成过程中分段复用，而非等待完整生成，优化了这一流程。该技术显著加速推理过程，尤其适用于需要系统提示或预定义指令的场景。

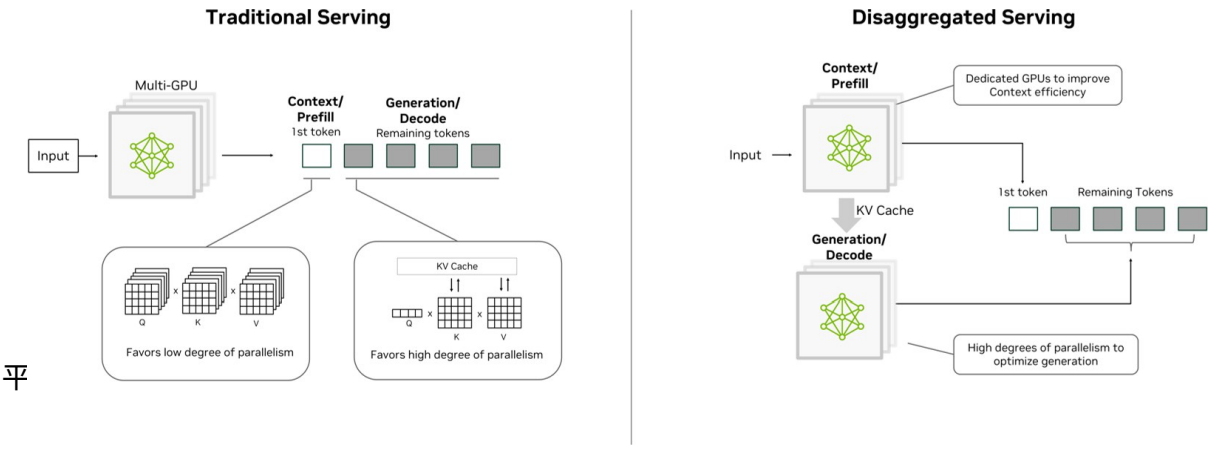
例如，在企业聊天机器人中，用户请求通常共享相同的系统提示，该功能可在高负载时段将 TTFT 缩短最高达 5 倍，提供更快速且响应更灵敏的用户体验。

P/D 分离服务：独立资源分配与分离式扩展

传统推理架构通常将预填充（prefill）与解码（decode）阶段共存于同一 GPU，导致资源分配低效和吞吐量欠佳。

NVIDIA Dynamo 提供的 P/D 分离服务（Disaggregated Serving）将这两个阶段解耦，使 AI 推理团队能根据各阶段特定需求独立分配资源。这实现了独立资源分配与解耦扩展：更多 GPU 可被分配至预填充阶段以优化 TTFT，同时更多 GPU 可专用于解码阶段以降低后续 token 间的生成延迟（ITL，Inter-token latency）。

这种分离还引入了阶段专用并行性 —— 预填充阶段的计算密集型任务可受益于管道并行（Pipeline Parallelism），而解码阶段可利用张量并行（Tensor Parallelism）和专家并行（Expert Parallelism）。



平

投机采样：加速 token 生成

通过自回归方式（逐个生成）生成 token 的过程可能缓慢且效率低下。投机采样通过并行生成多个候选 token 序列，优化这一流程，减少 token 生成所需时间。TensorRT-LLM 集成了多种投机采样方法，例如 Draft Target 和 Eagle Decoding ，允许系统预测多个 token 并选择最合适的输出。

对于 Llama 3.3 70B 等模型，推测解码使每秒生成 token 数量显著提升 3.5 倍，在不牺牲输出质量的前提下提高吞吐量和用户体验。该技术在低延迟、高吞吐量场景中尤为有益，可最大化每个计算步骤的效率。

上下文感知路由：避免 KV 缓存重新计算

复用 KV 缓存可避免从头开始重新计算，从而降低推理时间和计算资源消耗。这一特性在相同请求频繁执行的场景中尤为关键，例如系统提示、单用户多轮对话的聊天机器人交互以及代理工作流。NVIDIA Dynamo 在多节点和 P/D 分离部署场景中，跨大规模 GPU 集群跟踪 KV 缓存，并高效路由新请求，最大限度减少昂贵的重复计算。

当新推理请求到达时，NVIDIA Dynamo 计算传入请求与分布式集群中所有 GPU 内存中活跃的 KV 缓存块之间的重叠得分。通过结合重叠得分及 GPU 集群的工作负载分布，系统智能地将请求路由至最合适的计算节点，最小化 KV 缓存重新计算的同时确保集群负载均衡。

通过减少不必要的 KV 缓存重新计算，NVIDIA Dynamo 释放了 GPU 资源。这使得 AI 服务提供商能够响应更多用户请求，从而最大化他们在加速计算投资中的回报。

解锁 AI 推理的未来

NVIDIA AI 推理平台旨在支持处于 AI 推理旅程任何阶段的组织。对于仍处于实验阶段或专注于加速上市的企业，本章《影响推理部署的部署因素》中概述的策略提供了良好的起点。

然而，对于将 AI 推理视为影响毛利率的主要成本驱动因素的组织，本章《在云端解锁 AI 推理性能与成本效益》中涵盖的高级性能与成本节省技术，将帮助最大化系统效率、性能并降低成本。

入门 NVIDIA AI 推理平台

NVIDIA AI 推理平台为推理任务提供了灵活的部署选项，以满足广泛的业务和 IT 需求。

希望最快实现价值的企业可选择 NVIDIA NIM，该方案提供预封装、优化的推理微服务，可在任何 NVIDIA 加速基础设施上运行最新的 AI 基础模型。

若需要最大程度的灵活性、可配置性及可扩展性以满足独特的 AI 推理需求，NVIDIA 提供 NVIDIA Triton 推理服务器和 NVIDIA TensorRT，支持根据特定需求定制和优化推理服务平台。

访问 “NVIDIA AI 推理解决方案” 以了解更多信息并开始部署。