QCon
全球软件开发大会

InfoQ 极客传媒

释放生成式AI推理潜力：
分布式LLM基础设施与llm-d实践

张家驹

红帽首席架构师兼大中华区CTO

目录

InfoQ 极客传媒

北京

QCon 全球软件开发大会

会议时间：4月10-12日

- 大模型赋能 AIOps
- 越挫越勇的大前端
- 云上业务架构演进
- 多模态大模型及应用
- 海外 AI 应用创新实践

北京

AiCon 全球人工智能开发与应用大会

会议时间：6月27-28日

- 端侧智能
- AI Agent
- 多模态大模型
- 金融+大模型

上海

QCon 全球软件开发大会

会议时间：10月23-25日

- AI Agent
- 大模型训练推理
- 端侧 AI
- 搜推深度融合
- 数智金融

4月 — 5月 — 6月 — 8月 — 10月 — 12月

上海

AiCon 全球人工智能开发与应用大会

会议时间：5月23-24日

- 通用大模型
- AI Agent
- 垂直领域应用
- 模型可解释性

深圳

AiCon 全球人工智能开发与应用大会

会议时间：8月22-23日

- 模型效率与部署
- 多模态大模型
- 大模型安全
- 智能硬件

北京

AiCon 全球人工智能开发与应用大会

会议时间：12月19-20日

- 通用大模型
- 智能硬件
- LMOps
- 具身智能

极客邦科技 2025 年会议规划

促进软件开发及相关领域知识与创新的传播

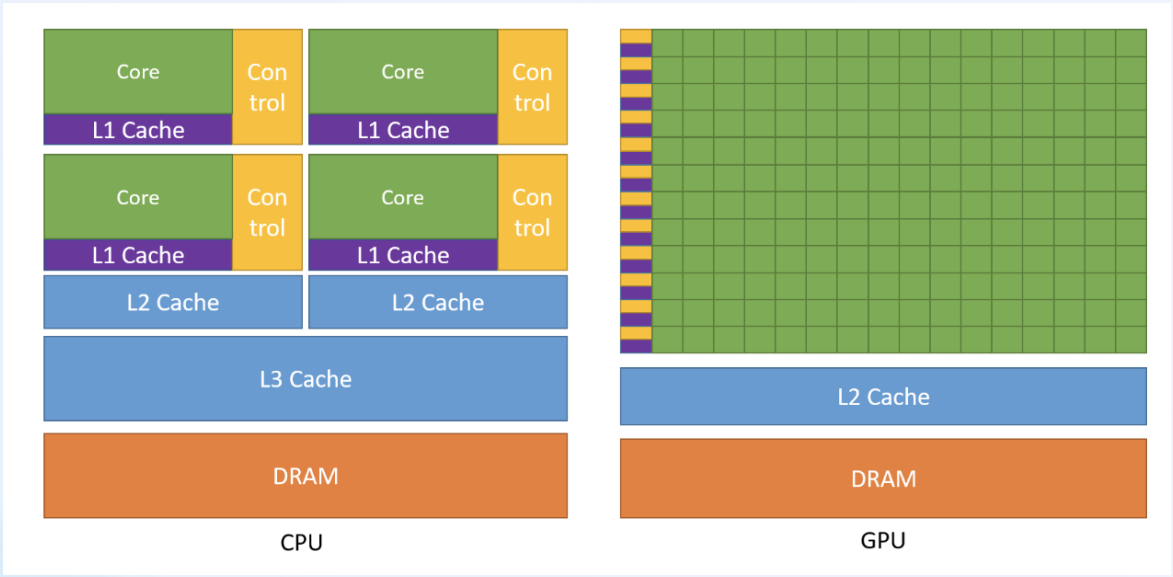
参会咨询

查看会议

01

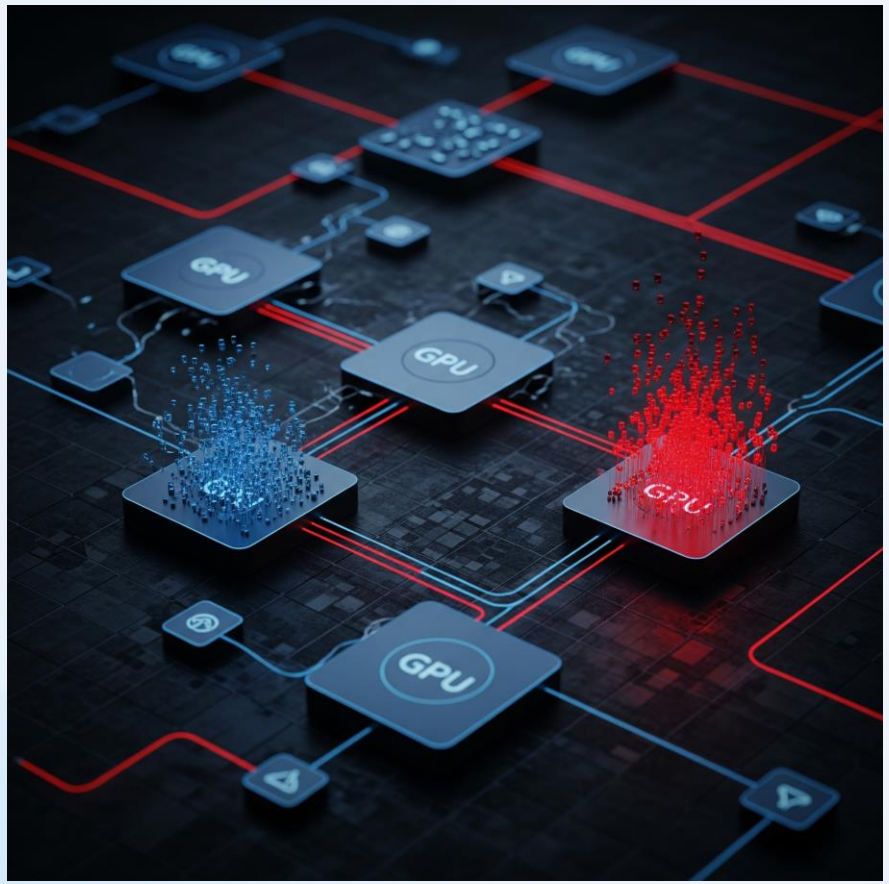
背景与问题

CPU与GPU不同的编程模型



Source:
https://docs.nvidia.com/cuda/cuda-c-programming-guide/index.html

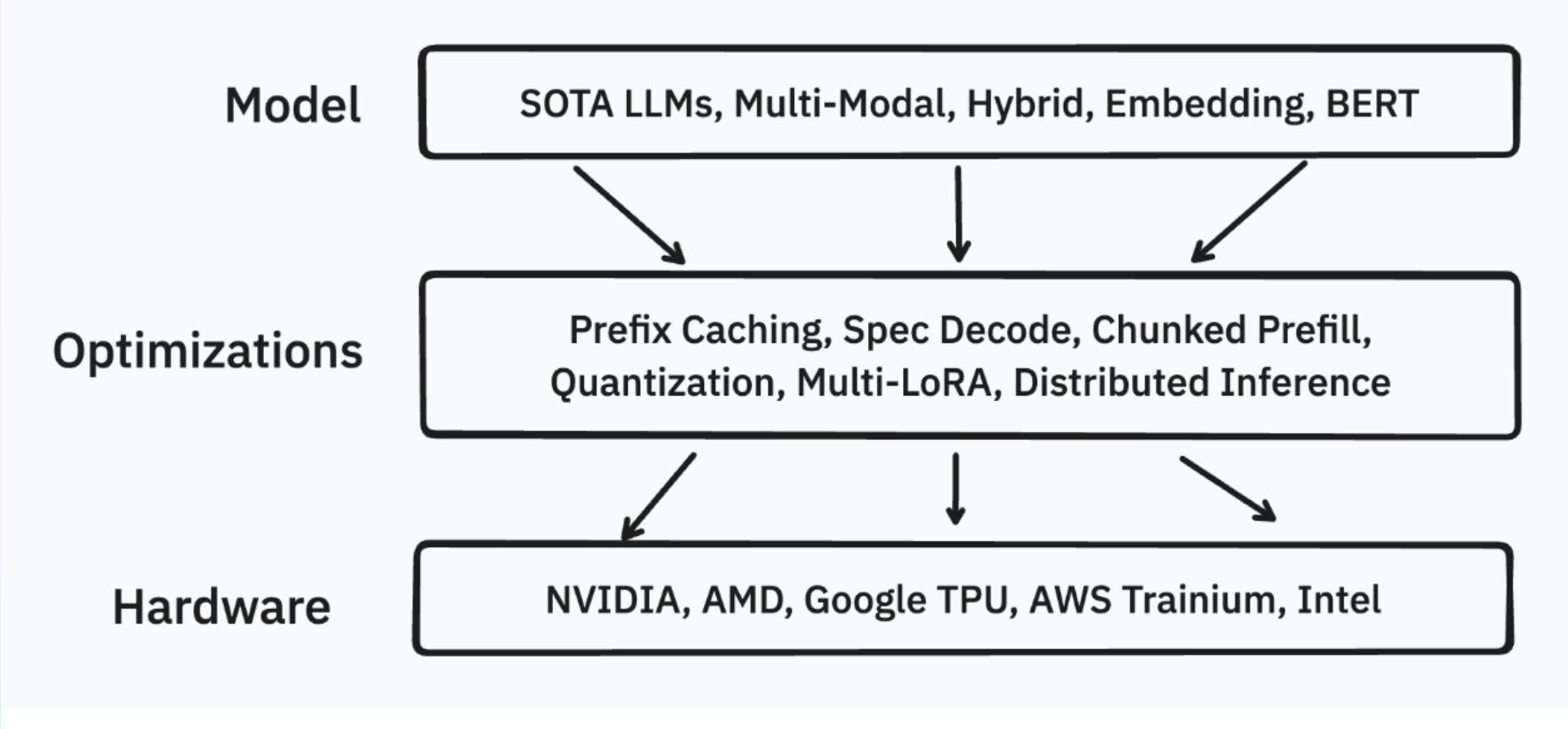
LLM推理面临的挑战



在生产环境中运行大语言模型，会面临一系列工程和运维层面的挑战，这些问题直接影响延迟、可扩展性和资源利用效率。随着模型规模增大、并发量提升以及实时响应需求增强，这些挑战会愈发突出。

- 高延迟：自回归式解码在 70 亿参数以上的模型中，每生成一个 token 可能会带来 200–600 毫秒的延迟。
- 资源利用低效：在突发流量下，批处理效率下降，GPU 空闲时间增加，导致整体算力浪费。
- 巨大的 KV 缓存需求：在长上下文窗口（4k–16k tokens）下，单个请求的 KV 缓存占用可超过 10–12GB。
- 复杂的分布式调度：跨节点的推理复制需要复杂的编排、动态的自动扩缩容机制以及资源调优策略。

LLM基础设施功能概览



02

单节点LLM推理优化

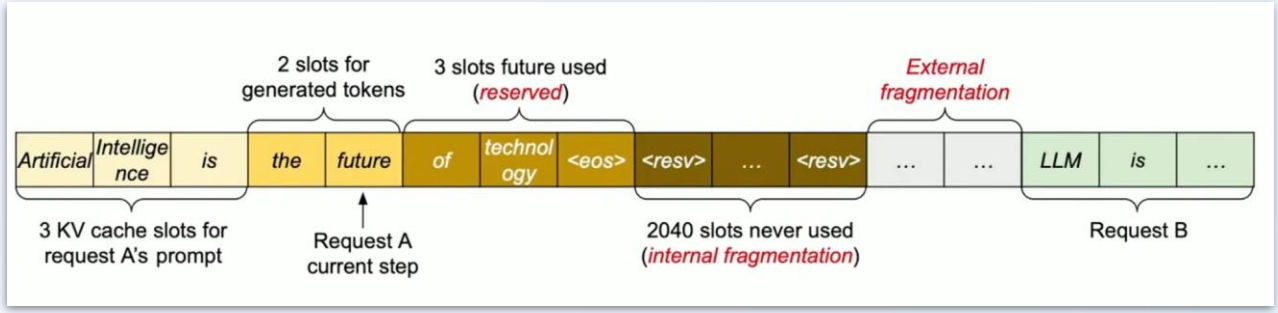
PagedAttention & KV Cache

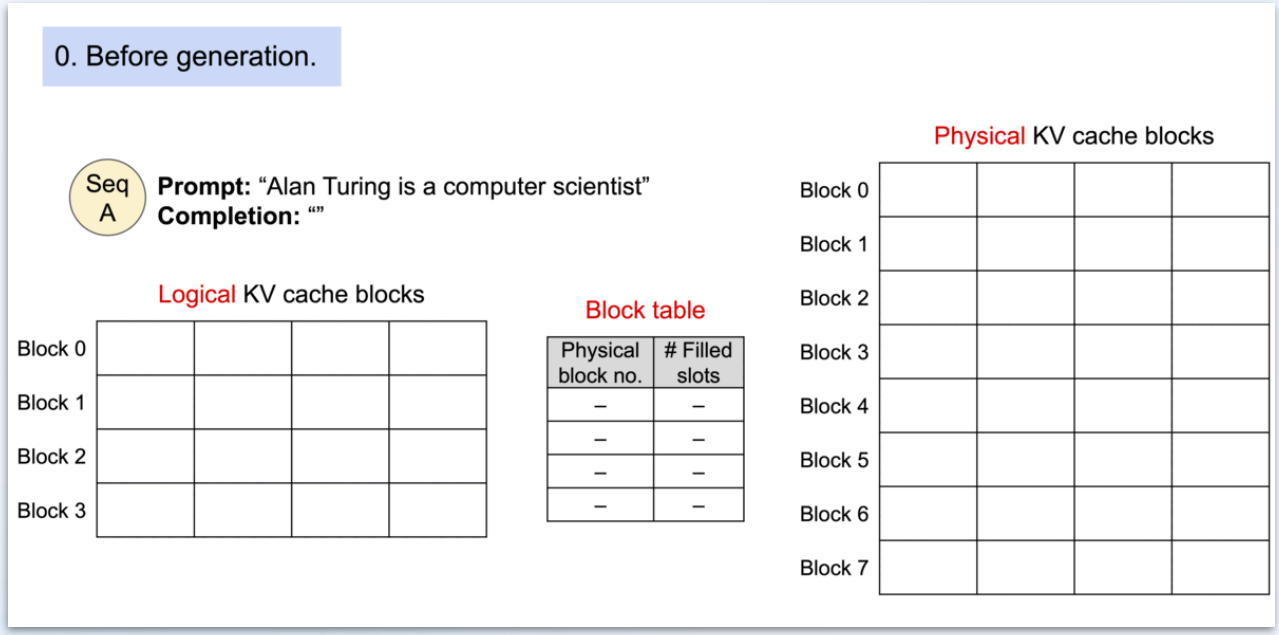
问题：KV 缓存挑战

- 在推理过程中，键值缓存（Key-Value Cache）相当于大语言模型的“记忆”，其大小会随着上下文长度线性增长。
- 传统系统中常出现严重的内存碎片化问题，甚至触发 GPU 显存溢出（OOM）错误。
- 这导致显存利用率极低——实际有效使用率往往只有分配空间的 20% ~ 40%，严重限制批处理规模并推高成本。

解决方案：PagedAttention

- 借鉴虚拟内存的设计理念，PagedAttention 将 KV 缓存拆分为多个固定大小的小块进行管理。
- 它将逻辑序列与 GPU 物理内存解耦，消除了内存碎片问题，并支持灵活的显存分配。
- PagedAttention 还能在不同请求之间动态共享缓存块，这对于多轮对话和复杂的解码策略至关重要。





连续批处理

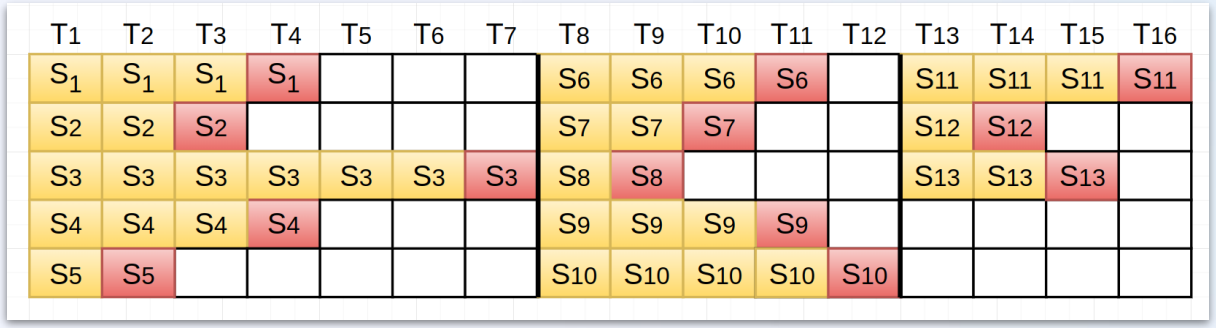
问题：静态批处理（Static Batching）

- 传统的批处理机制会等待固定数量的请求到达后，才开始执行前向计算。
- 在真实业务场景中，这种方式会导致 GPU 大量空闲、算力浪费严重。
- 同时，请求需要排队等待下一个批次填满，进一步增加了响应延迟。

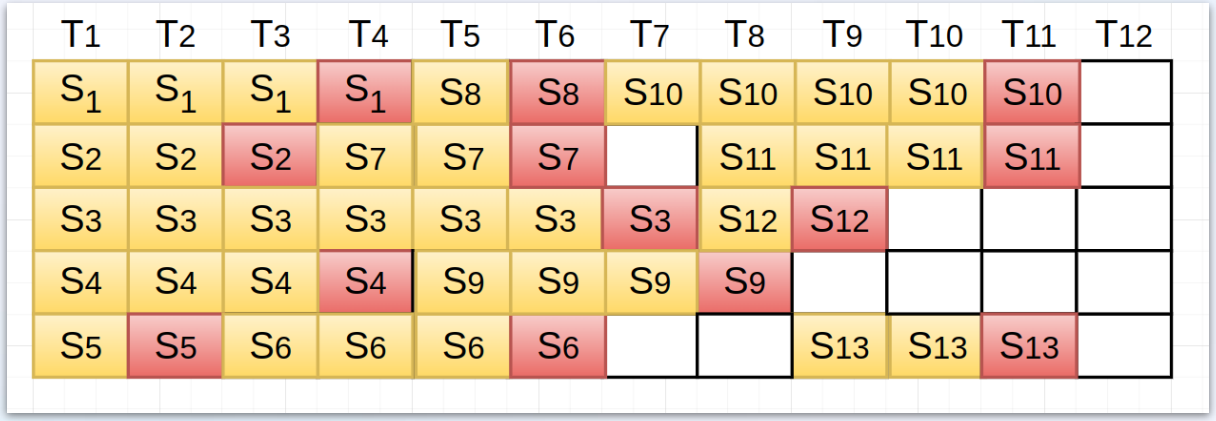
解决方案：连续批处理（Continuous Batching）

- 动态地将新到达的请求即时加入当前批次进行处理。
- 通过让请求随时进入 GPU 管线，保持 GPU 长时间“满负荷”工作。
- 这种机制有效消除了排队延迟，大幅提升了 GPU 利用率和整体吞吐性能。

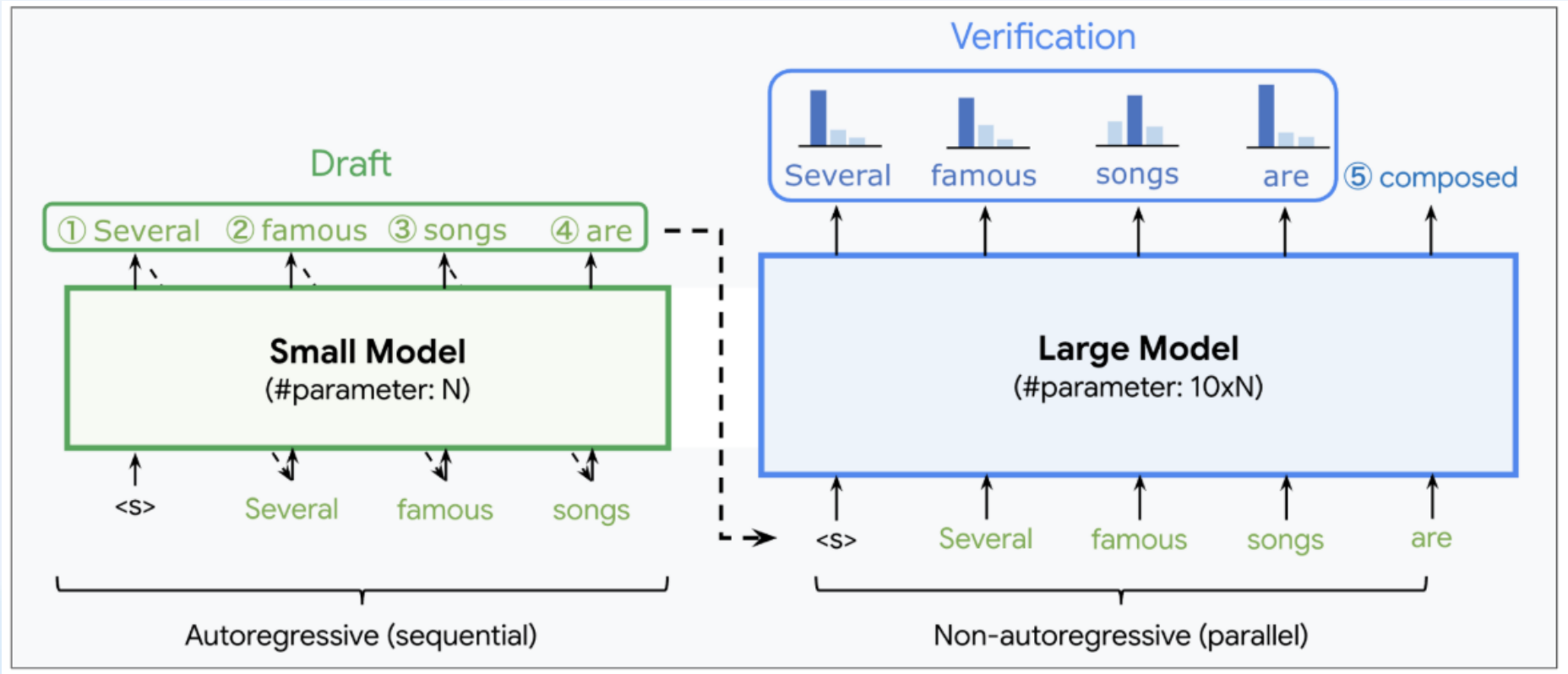
Naive Batching



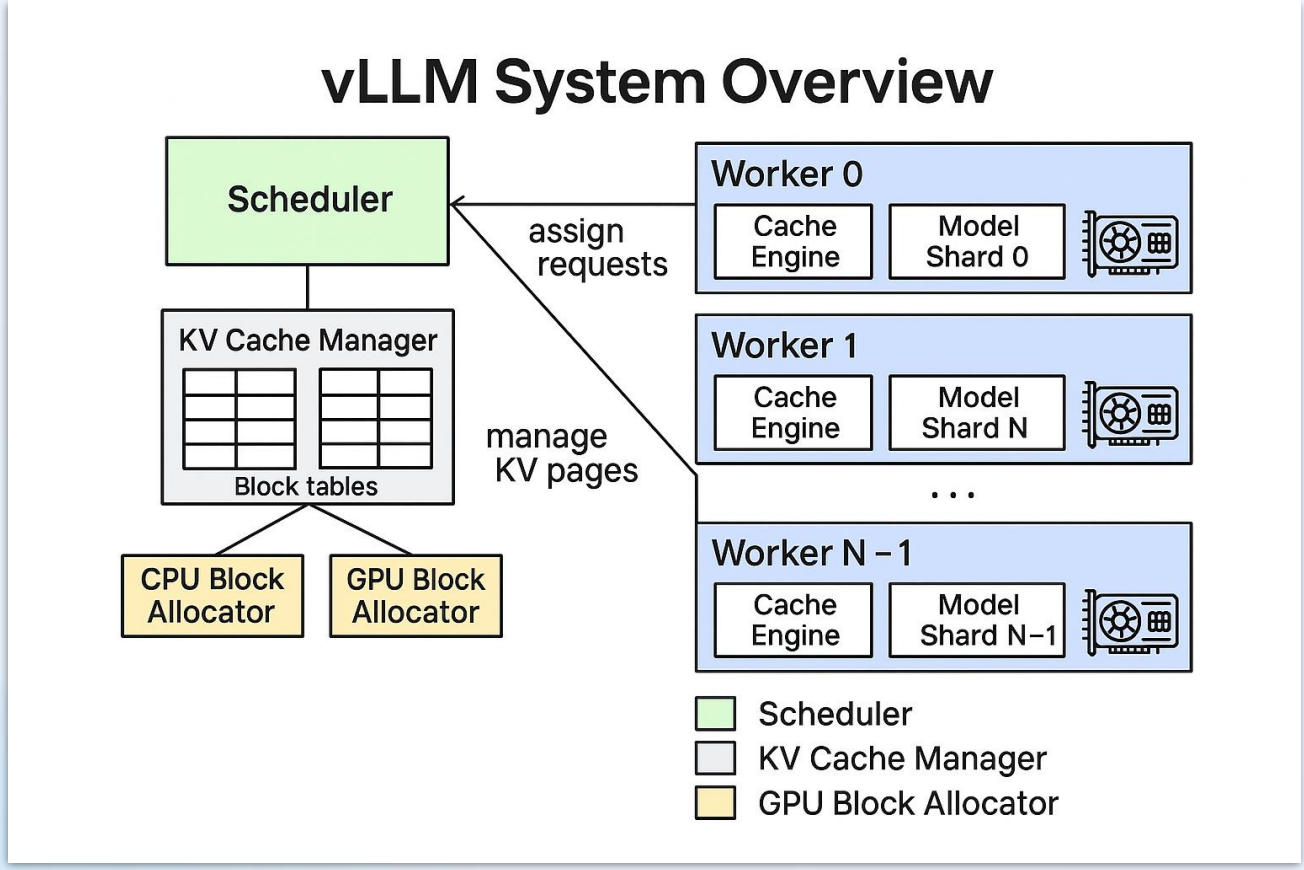
Continuous Batching



推测解码


Verification
Several
famous
songs
are
⑤ composed
Draft
① Several ② famous ③ songs ④ are
Small Model
(#parameter: N)
Large Model
(#parameter: 10xN)
<s>
Several
famous
songs
<s>
Several
famous
songs
are
Autoregressive (sequential)
Non-autoregressive (parallel)

其他特性



模型优化（Model Optimization）：采用 FP8 量化等技术，在保持精度的同时减少模型规模和延迟，更高效地利用 GPU 资源。

张量并行（Tensor Parallelism）：将超大模型拆分到单节点的多张 GPU 上协同运行，实现大模型的高效推理。

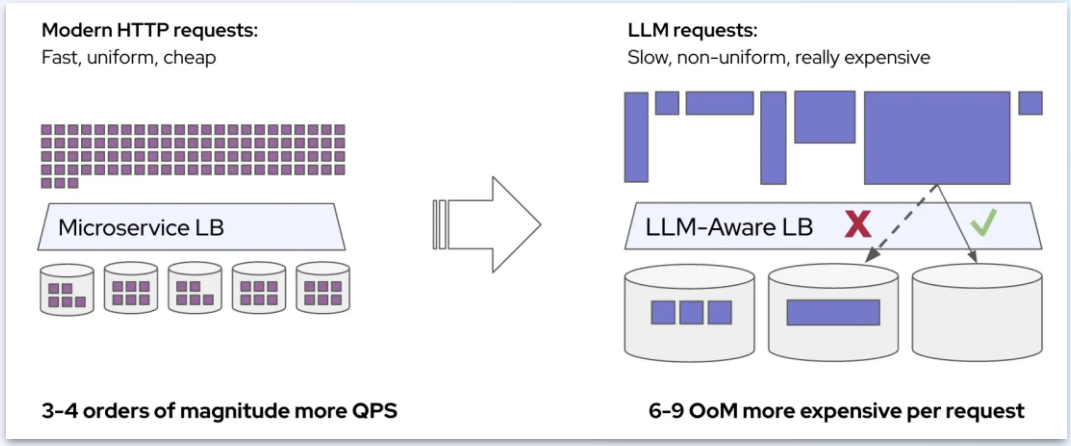
兼容 OpenAI API（OpenAI API Compatibility）：提供标准化接口，方便开发者无缝集成到现有工具和应用中。

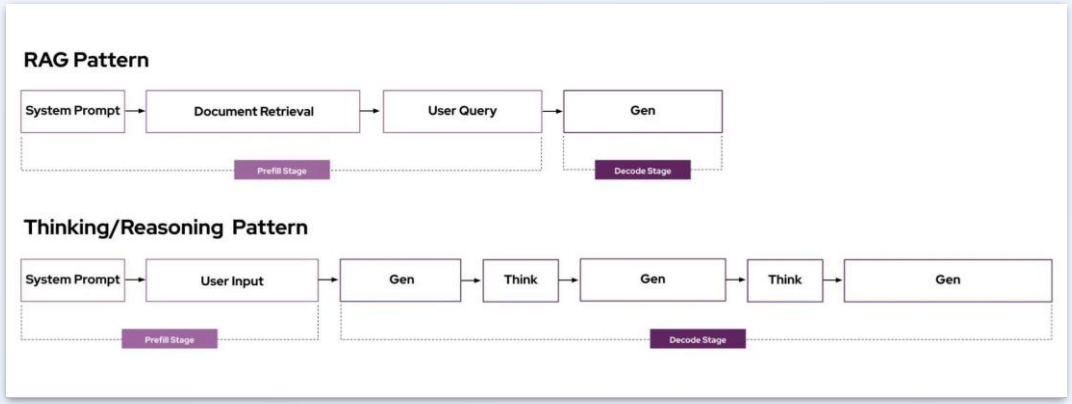
03

分布式推理

vLLM推理的局限

- **高成本、非均匀且有状态的请求**
  LLM 推理每个请求的计算成本比传统应用高 6–9 个数量级，输入输出长度差异巨大且存在关键状态依赖。
- **资源盲目负载均衡**
  传统 Kubernetes 的轮询或随机路由无法感知 AI 特性，忽略了 GPU 负载、队列深度、提示复杂度和 KV 缓存状态，导致 GPU 利用率低下和延迟峰值。
- **资源利用低效**
  在同一实例上同时处理计算密集的 prefill 阶段和内存密集的 decode 阶段效率低下，尤其在长序列推理中，浪费了宝贵的 GPU 计算周期。
- **优化机会错失**
  缺乏 KV 缓存和前缀感知路由，使得重复提示被当作冷启动处理，无法利用缓存计算来降低延迟。
- **可扩展性瓶颈**
  单体部署限制了推理组件的独立扩缩容，在动态、突发流量下难以满足服务等级目标（SLO）。

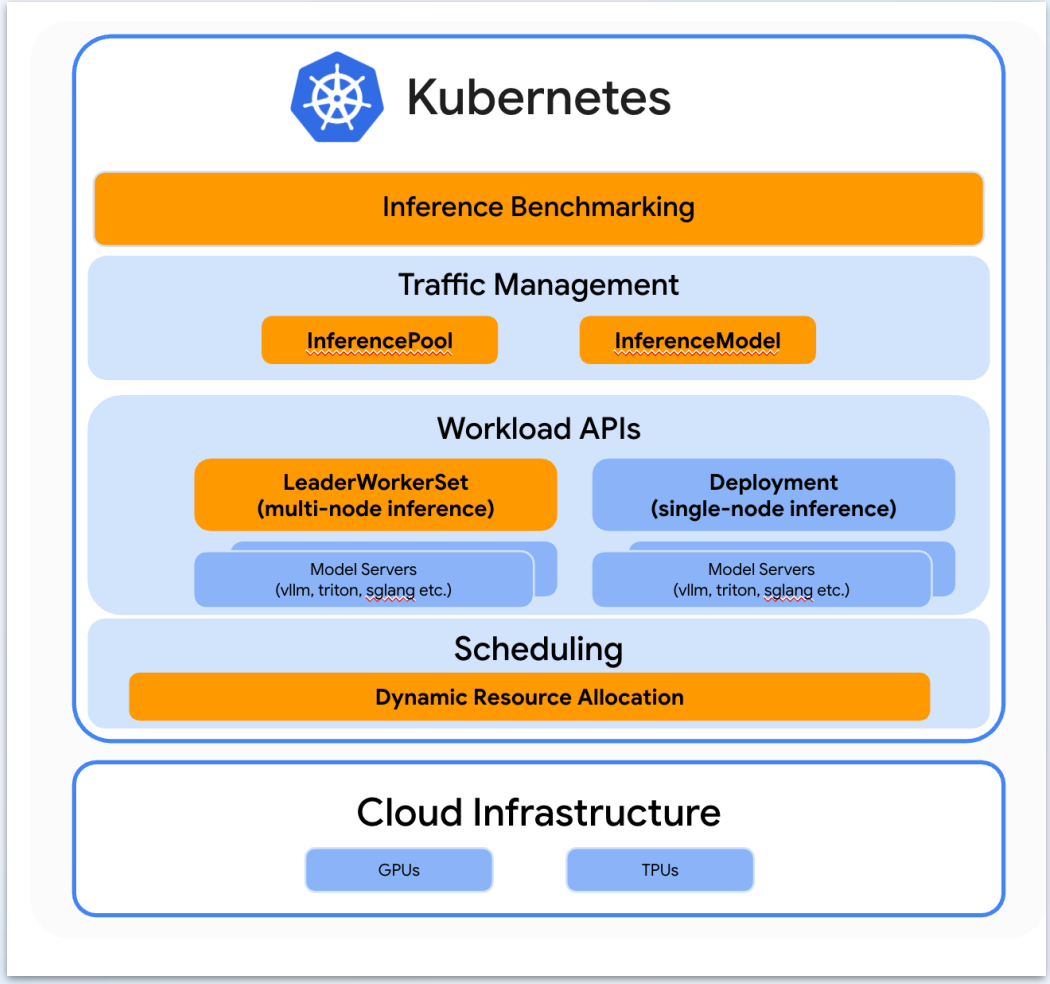




Kubernetes生态下GenAI工作负载的使能

在过去的十年里，Kubernetes 已成为云原生应用的事实标准，但生成式 AI（GenAI）工作负载带来了独特的挑战。社区正在致力于让 Kubernetes 在核心层面变得“AI 感知”。

- 标准化基准测试：社区项目正在开发评估模型和加速器配置的标准，将部署决策从经验猜测转向数据驱动。
- AI 感知负载均衡：Kubernetes Gateway API 推理扩展正在开发中，支持 AI 原生路由，根据 LLM 工作负载特性和 KV 缓存使用情况智能分配请求。
- 硬件可移植性：生态系统正在向 GPU、TPU 等多种加速器的无缝互操作和可替换性发展，提供更多选择与灵活性。



分布式推理的实践

P/D分离（Disaggregated Inference）

当计算密集型的 prefill 阶段和内存密集型的 decode 阶段在同一个副本上运行时，资源利用效率低下。

基于 KV 缓存和负载感知的智能调度

传统负载均衡忽略了 LLM 特有特性，导致 GPU 利用率和整体性能不佳。

混合专家模型（MoE）推理的扩展

大型 MoE 模型占用显存巨大，需要复杂的并行策略才能实现多节点的高效执行。

分布式推理：llm-d

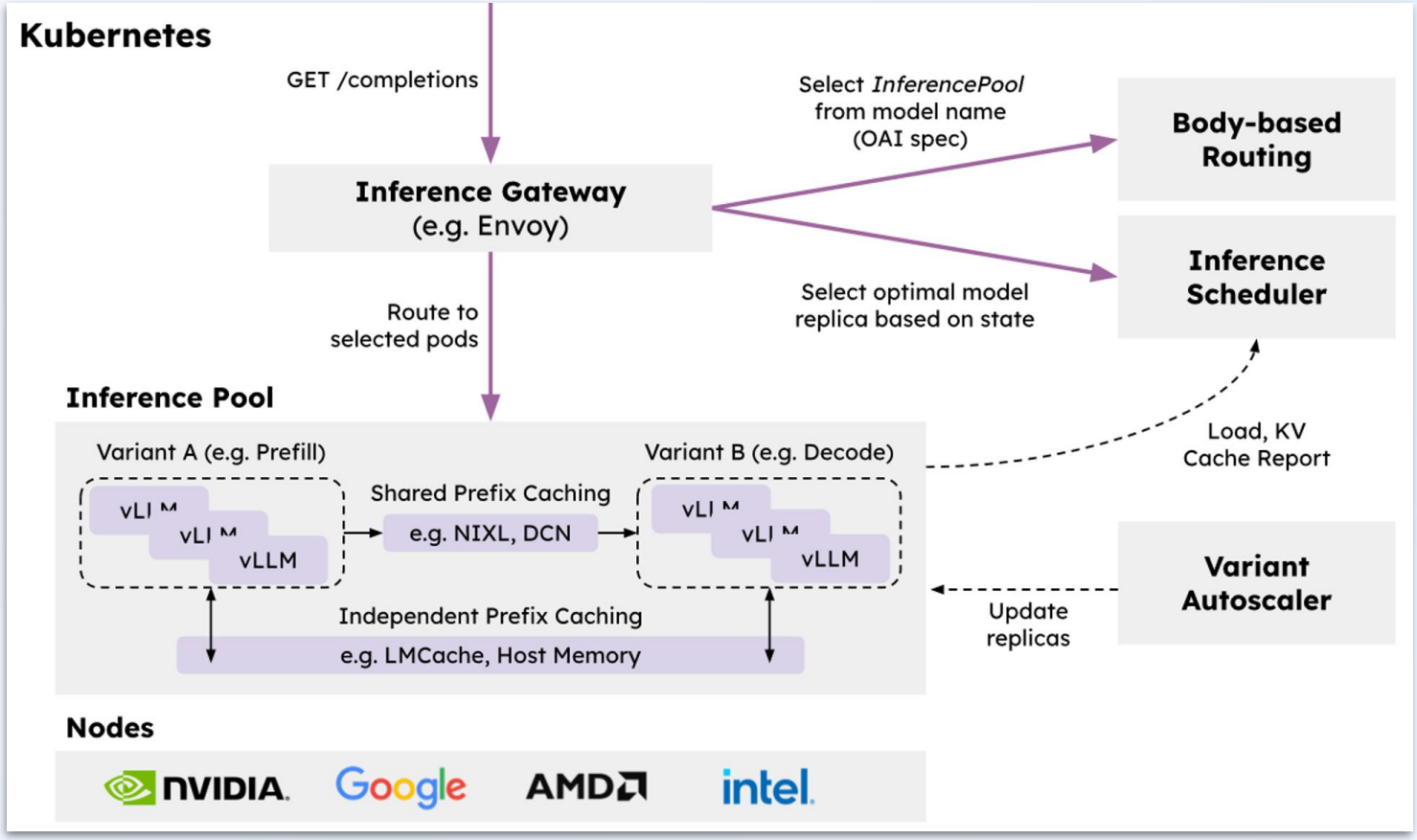
用于分布式大语言模型（LLM）推理，能够原生运行在 Kubernetes 上的开源框架

- **原生 Kubernetes 架构**：一个开源的 Kubernetes 原生平台，用于高性能分布式 LLM 推理，通过 AI 感知路由和高级调度扩展 Kubernetes 能力。
- **由 vLLM 驱动**：利用 vLLM 作为优化执行引擎，将其高吞吐、低延迟能力集成到分布式框架中。
- **跨平台 & 灵活**：支持多种硬件（NVIDIA、AMD、Google TPU、Intel）以及广泛的 LLM 模型，为企业部署提供高度适配性和选择自由。
- **产业协作**：由 Red Hat、Google、IBM Research、NVIDIA、AMD 和 CoreWeave 等领先组织支持，推动社区驱动的 GenAI 推理普及。
- **性能与成本优化：**通过智能推理调度、分离式服务以及先进的分布式 KV 缓存管理，降低推理成本并提升效率。

llm-d的架构

设计原则

- **可落地性（Operationalizability）：**模块化且高可用的架构，通过 Inference Gateway API 与 Kubernetes 原生集成。
- **灵活性（Flexibility）：**跨平台支持（正在积极支持 NVIDIA、Google TPU、AMD 和 Intel），关键可组合层可扩展实现。
- **性能（Performance）：**利用分布式优化技术，如分离式推理和前缀感知路由，在满足 SLO 的同时，实现最高的 token/$ 性能比。



通过 Inference Gateway 实现智能推理调度

Inference Gateway 位于 LLM 推理的前端，作为 Kubernetes 原生的控制平面，用于管理流量并执行策略。

智能流量调度

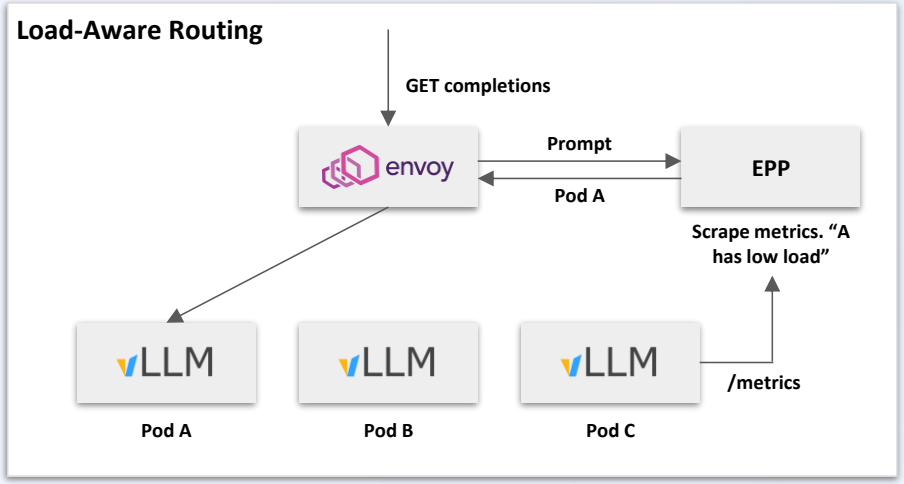
利用推理调度器（Inference Scheduler）将请求动态路由到最优的 vLLM 实例。

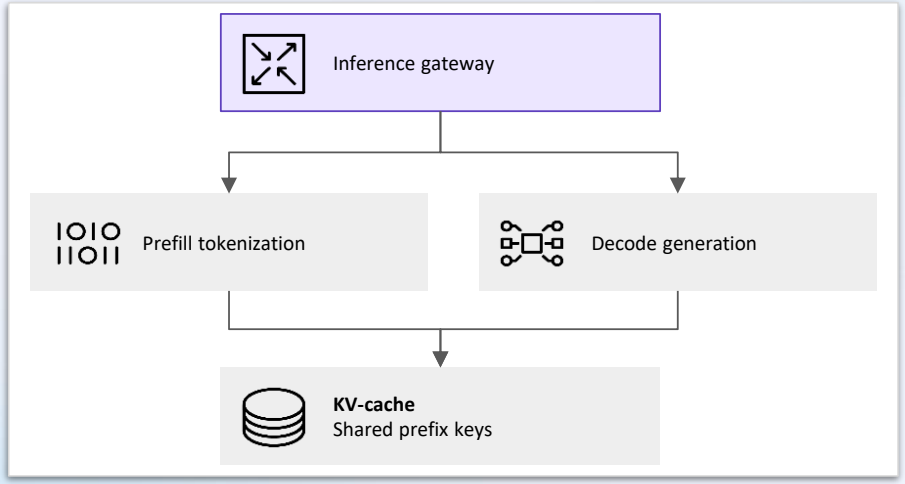
性能优化（通过 Prefill/Decode 分离）

支持高级路由决策，高效处理复杂工作负载。

安全与合规

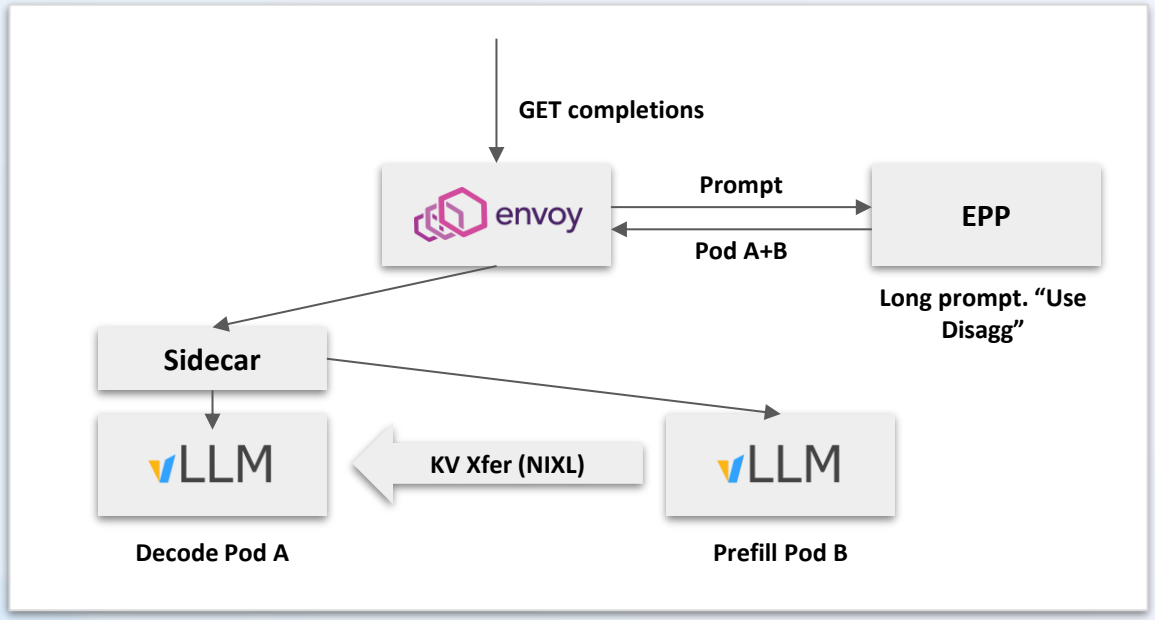
执行严格的安全策略和请求保护机制，同时提供提示日志记录和审计功能。





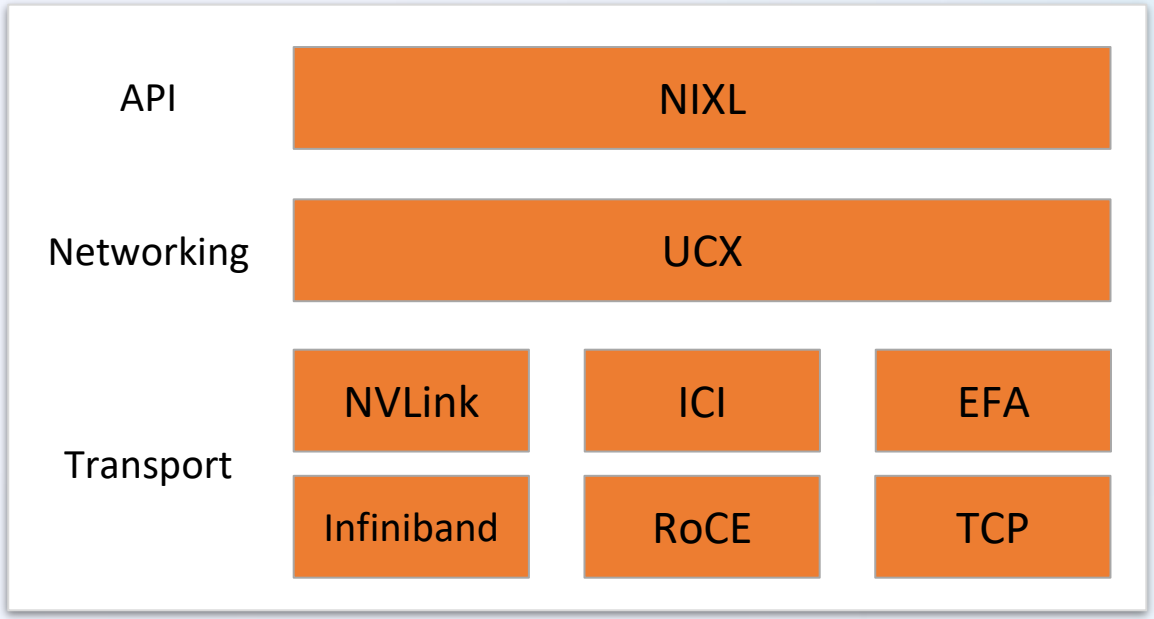
P/D 分离

Disaggregated Serving



Specialization of pods for prefill and decode phase

Heterogenous Transfer Protocols

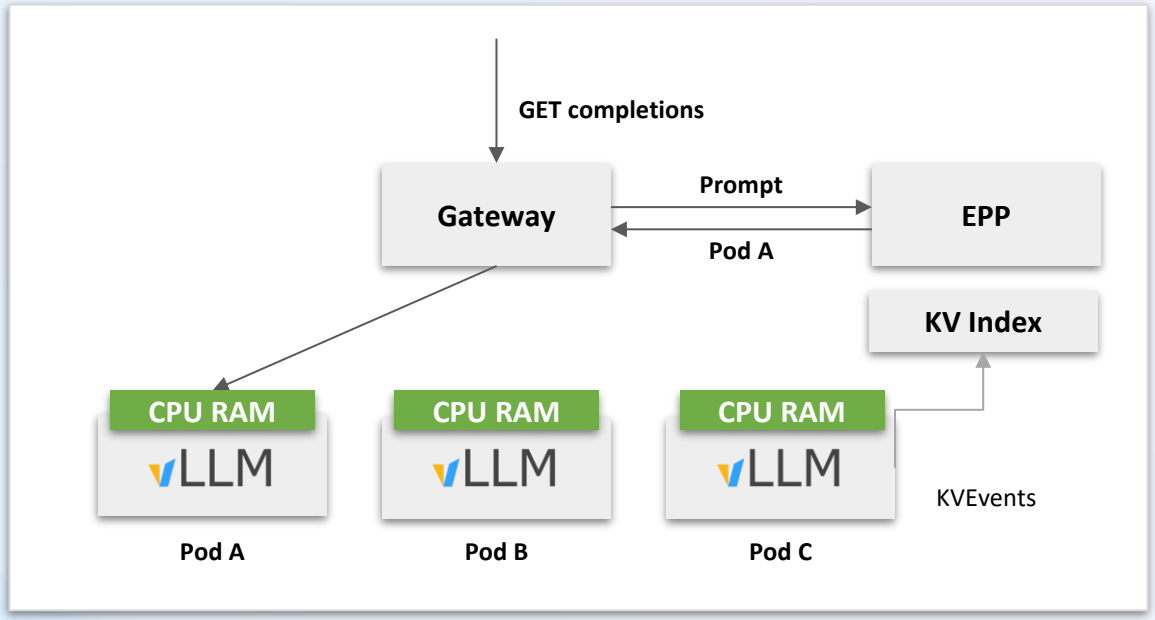


Support variety of transports via UCX

P/D disaggregation is a key optimization for demanding inference workloads

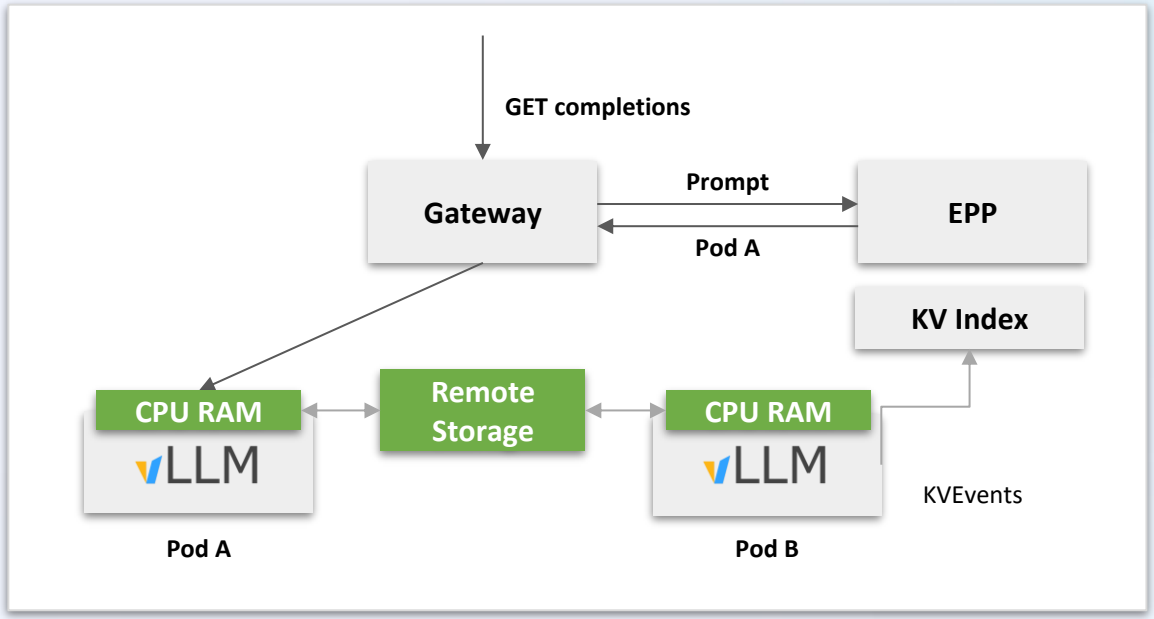
多层 KV 缓存管理

North/South KV Management



Offload KV caches to local memory with global index

East/West KV Management



Offload KV caches to global shared remote storage

Leverage all system resources to maximize prefix cache hit rate within the cluster

04

结论与行动呼吁

结论与行动呼吁

- 大模型正在重新定义软件
- 从优化到架构的范式转变
- 开放协作是前进的关键
- 行动呼吁：加入我们的旅程，共建可扩展、可移植、高性能、易运维的GenAI基础设施

InfoQ 极客传媒

北京

QCon 全球软件开发大会

会议时间：4月10-12日

- 大模型赋能 AIOps
- 越挫越勇的大前端
- 云上业务架构演进
- 多模态大模型及应用
- 海外 AI 应用创新实践

北京

AiCon 全球人工智能开发与应用大会

会议时间：6月27-28日

- 端侧智能
- AI Agent
- 多模态大模型
- 金融+大模型

上海

QCon 全球软件开发大会

会议时间：10月23-25日

- AI Agent
- 大模型训练推理
- 端侧 AI
- 搜推深度融合
- 数智金融

4月 — 5月 — 6月 — 8月 — 10月 — 12月

上海

AiCon 全球人工智能开发与应用大会

会议时间：5月23-24日

- 通用大模型
- AI Agent
- 垂直领域应用
- 模型可解释性

深圳

AiCon 全球人工智能开发与应用大会

会议时间：8月22-23日

- 模型效率与部署
- 多模态大模型
- 大模型安全
- 智能硬件

北京

AiCon 全球人工智能开发与应用大会

会议时间：12月19-20日

- 通用大模型
- 智能硬件
- LMOps
- 具身智能

极客邦科技 2025 年会议规划

促进软件开发及相关领域知识与创新的传播

参会咨询

查看会议

QCon
全球软件开发大会

InfoQ 极客传媒

THANKS

大模型正在重新定义软件

Large Language Model Is Redefining The Software