如何理解 ZAI 中的多 AI 引擎

ZAI 的多 AI 引擎是基于 DNN-Thread 的多 GPU 引擎.

每个 GPU 引擎都需要有自己的管理体系,状态呈现,api 驱动,使用这些东西既要从 app 层面入手也要从底层建立的各种库入手.换个角度来说:曾经的 app 架构无法直接扩展使用新技术.

Z.AI.pas 使用 TAI_DNN_Thread_Pool 的 GPU 引擎,支持的 AI 模型 11 种

Z.AI.Tech2022.pas 使用 TAI_TECH_2022_DNN_Thread_Pool 引擎,支持 AI 模型 3 种

把引擎分离出来是因为编译太过耗时(例如 build 一个 cuda12 的平台是 30-90 分钟起步,并且需要人值守编译,因为这需要很多操作),另一方面,AI 库代码已经达到 2.3W 行,往里面无限堆大会非常逆天,难以维护.

未来会不会开辟新库 Z.AI.Tech2024/ Z.AI.Tech2025,这是未知的,目前 tech2022 只有 6k 行,大概率会直接往 tech2022 里面继续堆,大约到 2025 会考虑开新库吧.

未来非常确定会接入 pybind11,并且引入 openai,Stable Diffusion,segment-anything 这类大模型技术.

对于编程的影响

Demo 的 API caller 方法,不会是 AI 应用,使用 AI 必然会开多 DNN-Thread 来充分利用 GPU,这对编程会出现更高的架构要求.

早期使用 TAI_DNN_Thread_Pool 的架构将会很难兼容 TAI_TECH_2022_DNN_Thread_Pool.因为这种架构从设计之初就是单引擎架构,未来 AI 的程序架构会走多引擎路线.

未来多引擎路线需要从 app 架构+库架构同时入手才能享受到新技术的价值.

本文撰写与 2024-1 月,在未来设计架构时一定要考虑多引擎架构的方式.

By.qq600585