Сервер инференса с ускорителями Atlas 300I Duo

Сервер инференса на китайских ускорителях Ascend — альтернатива NVIDIA по цене за инференс.

Описание

Платформа инференса на архитектуре Huawei Ascend: ускорительные карты Atlas 300I Duo (56 TOPS INT8 каждая). Подходит для видеоаналитики, распознавания речи и текстовых сервисов, где важна цена обработки, а не обучение.Карта: Atlas 300I Duo, 56 TOPS INT8 / 28 TFLOPS FP16Конфигурация: до 8 карт на серверСофт: CANN, MindStudio, поддержка PyTorch, TensorFlow, MindSporeПомогаем с портированием моделей с CUDA на CANN — это основной барьер входа, и он решается инженерно.

Технические характеристики

Карта	Atlas 300I Duo, 56 TOPS INT8 / 28 TFLOPS FP16
Конфигурация	до 8 карт на сервер
Софт	CANN, MindStudio, поддержка PyTorch, TensorFlow, MindSpore

Характеристики могут обновляться производителем. Финальная конфигурация фиксируется в спецификации счёта.