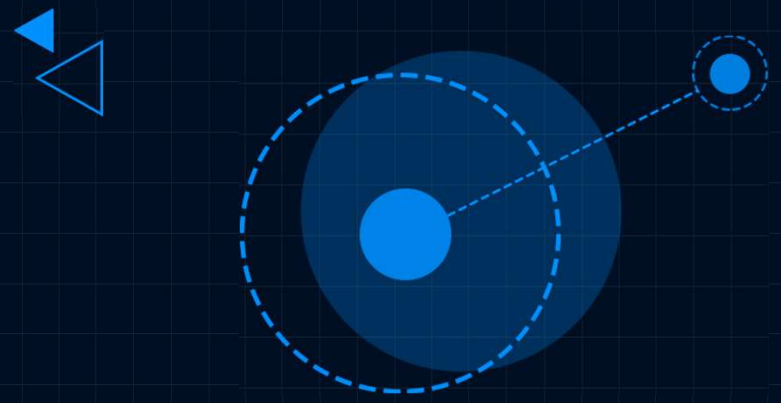
AI加速引擎 – TACO优化实践

腾讯云面向IaaS层的AI训练和推理套件

冯克环 张锐

腾讯云虚拟化研发工程师

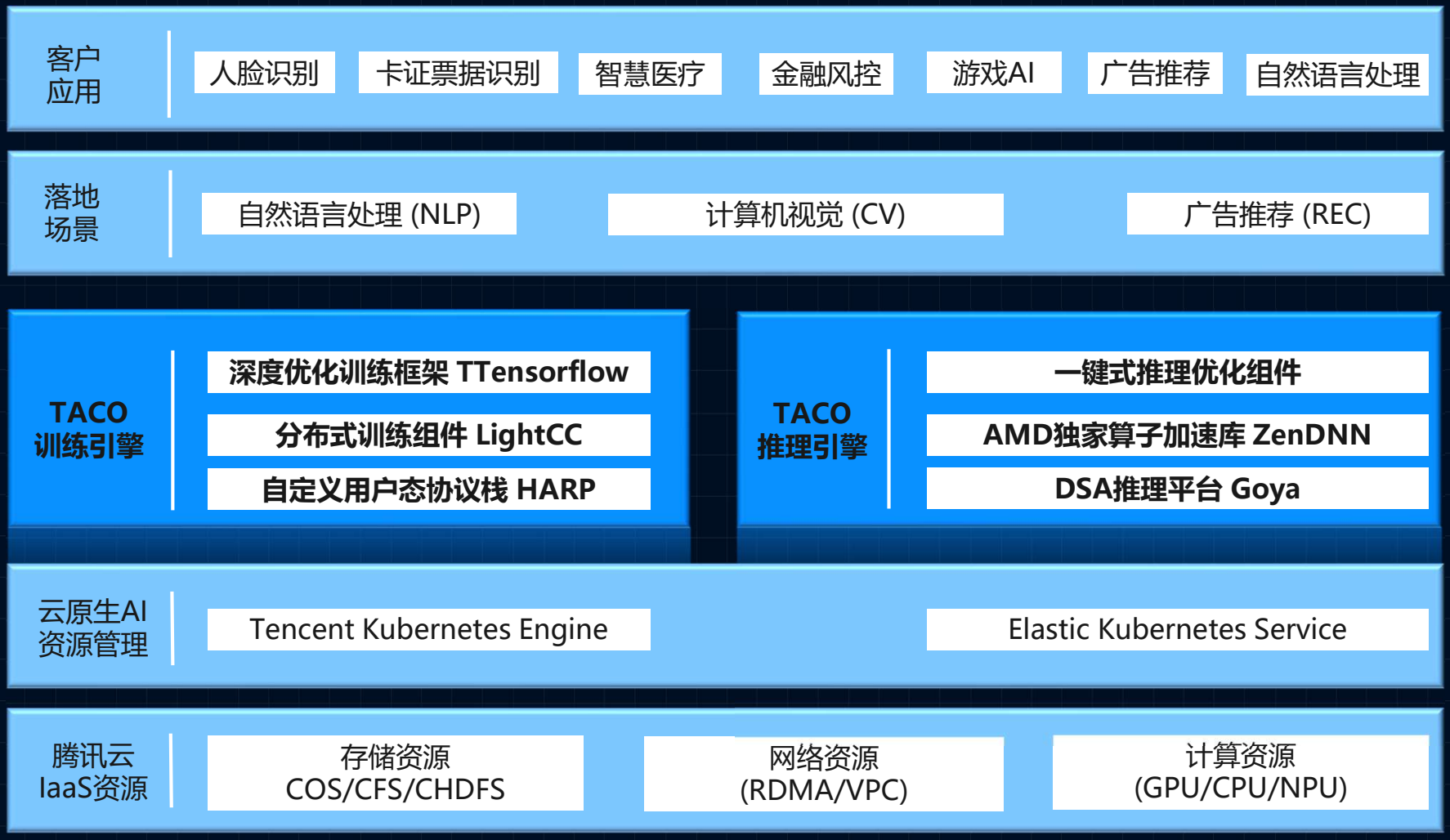


腾讯云异构AI加速方案 ——TACO

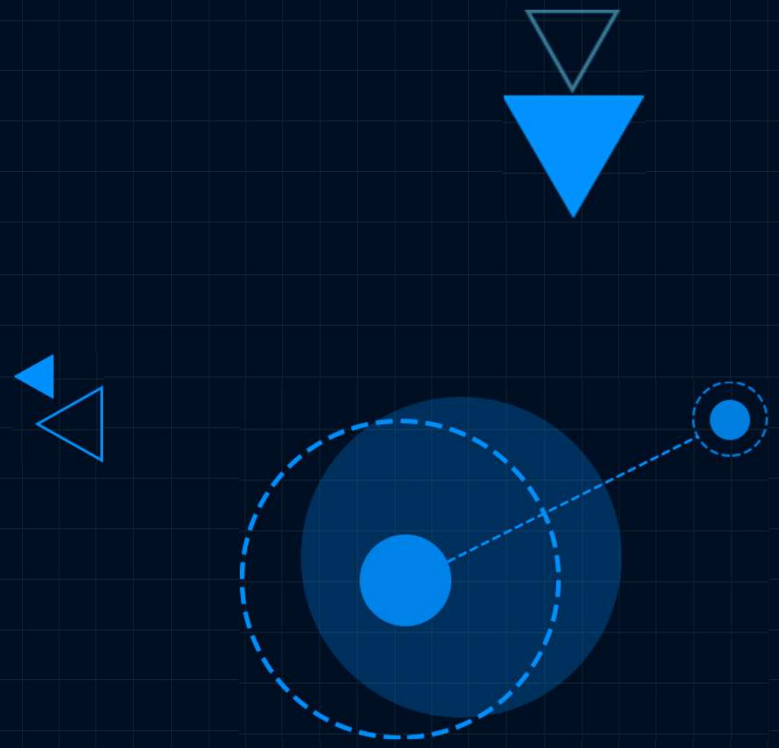
背景介绍

TACO是腾讯云虚拟化团队，依托云帆团队（来自18个部门的AI专家组成的虚拟团队），立足于腾讯内部丰富的AI业务场景，深耕训练框架优化，分布式框架优化，网络通信优化，推理性能优化等关键技术，携手打造的一整套AI加速方案。为了更好的服务内外部客户，腾讯云决定将内部深度优化的加速方案免费提供给公有云客户，助力广大用户提高AI产品迭代效率。

腾讯云异构AI解决方案 ——TACO



TACO训练引擎



TACO训练引擎

TTensorflow

LightCC

HARP

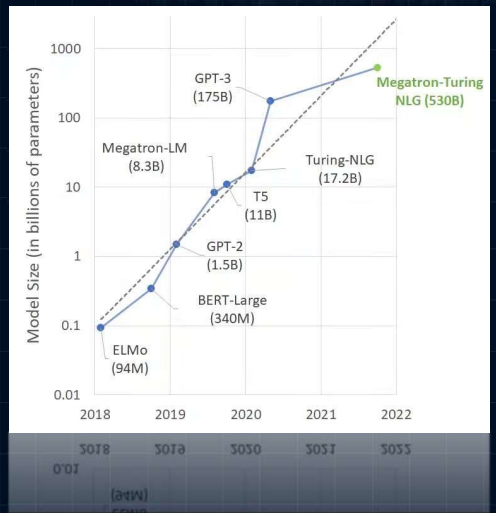
TTensorflow

基于 HashTable 的 TensorFlow 大规模分布式 Embedding 方案

推荐模型业务痛点

- 参数变化快 → 样本持续增长；用户兴趣变化快

- 模型规模超大 → GB到TB；特征稀疏

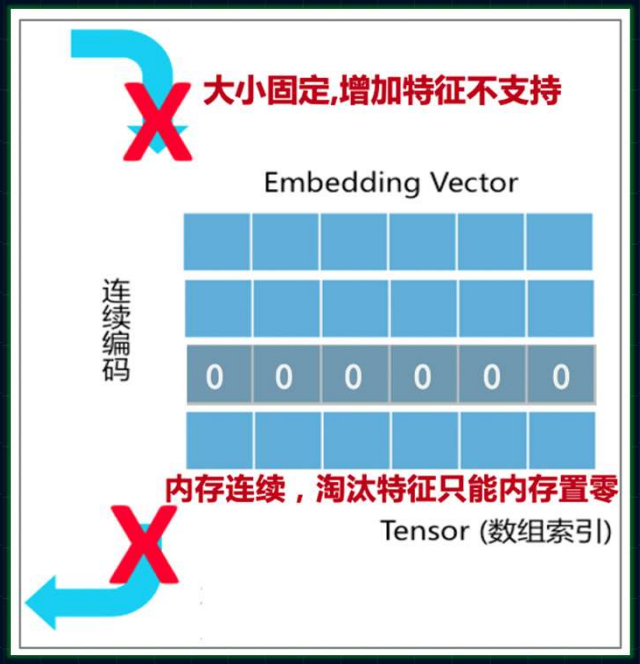


TTensorflow

基于 HashTable 的 TensorFlow 大规模分布式 Embedding 方案

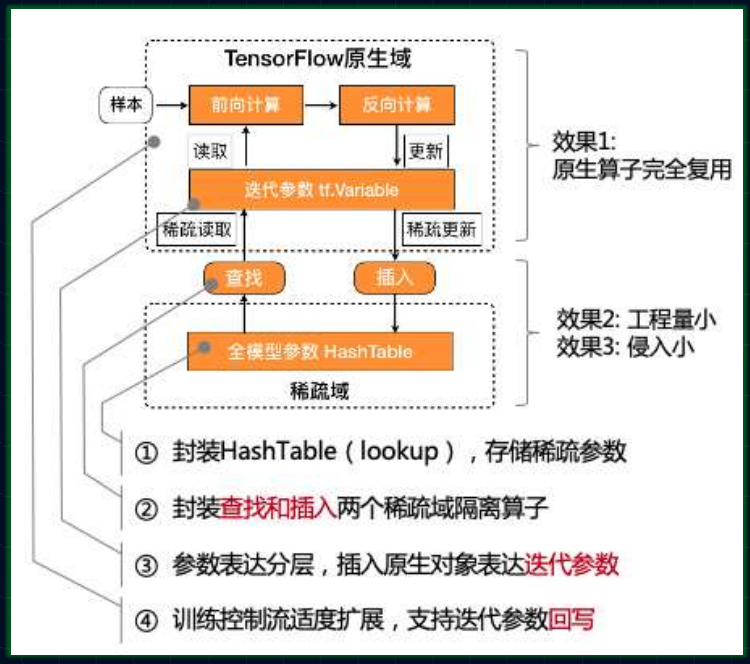
开源Tensorflow的不足

- 内存连续，不支持参数动态增删
- 超大规模样本存在Hash冲突



TTensorflow的改进

- 支持动态增删
- 按需使用内存，避免Hash冲突
- 保留原始API设计风格



TTensorflow

基于 HashTable 的 TensorFlow 大规模分布式 Embedding 方案

TTensorflow其他改进及业务效果

- ◆ 自适应动态编译框架，解决冗余重编译问题，提升特定业务场景的性能
- ◆ 混合精度编译，自动切换全精度和半精度，避免精度损失
- ◆ TF 1.5版本支持Ampere GPU和CUDA 11+

某推荐业务使用
动态Embedding AUC提升

某游戏AI业务训练
使用自适应动态编译框架
吞吐提升

某业务使用混合精度训练
性能提升1倍，精度和全精度持平

SIG Recommenders Addons
TensorFlow

TACO训练引擎

TTensorflow

LightCC

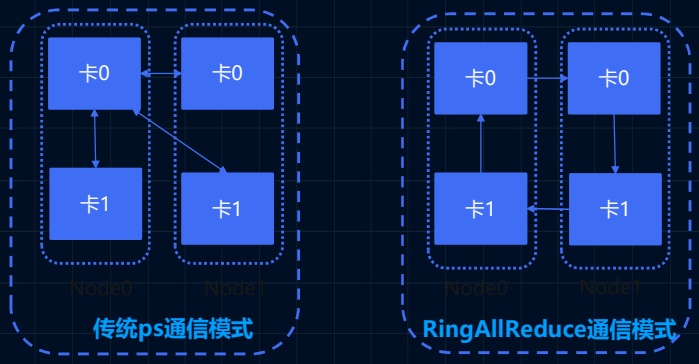
HARP

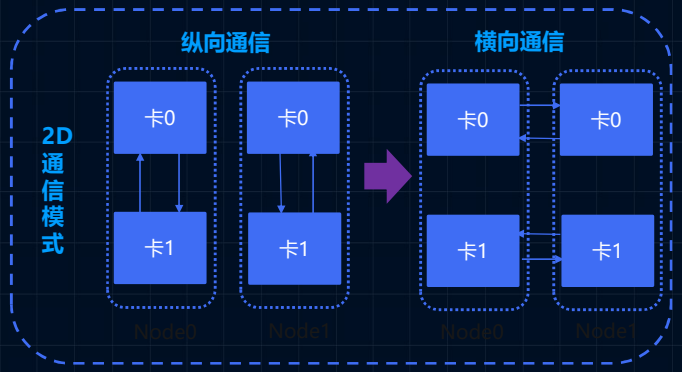
LightCC

基于Horovod深度优化的分布式训练框架

LightCC相比开源Horovod改进

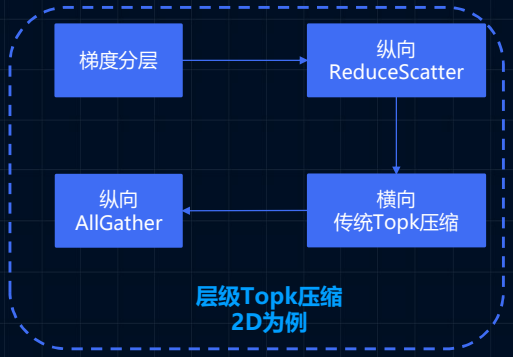
◆ 2D AllReduce充分利用网络带宽



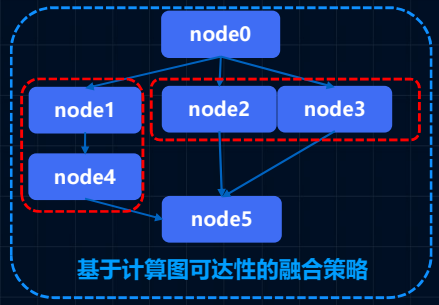


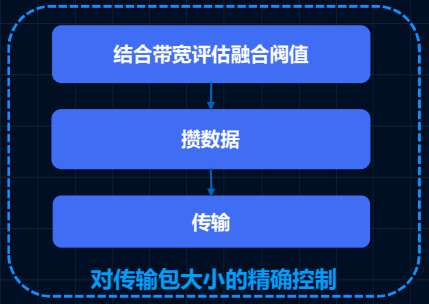
根据场景不同，支持多种2D/3D的通信模式

◆ TOPK压缩通信，降低通信量，提高传输效率



◆ 自适应梯度融合方式

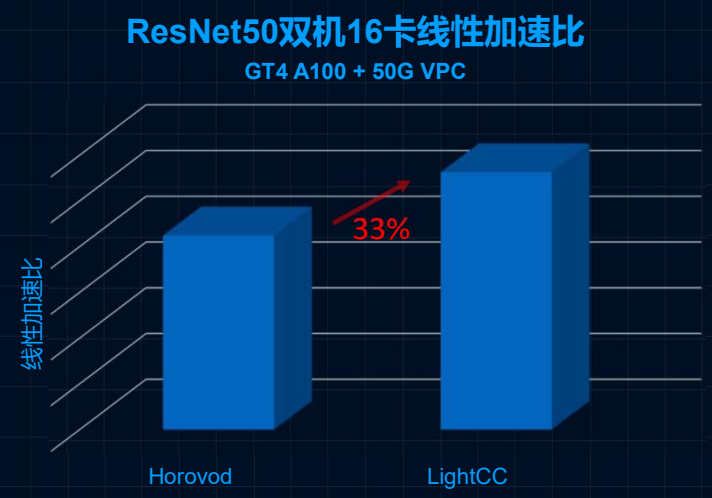


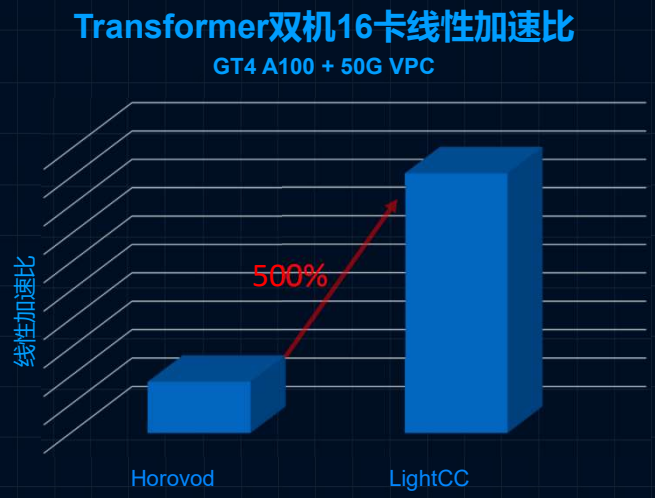


LightCC

基于Horovod深度优化的分布式训练框架

LightCC相比开源Horovod性能提升





TACO训练引擎

TTensorflow

LightCC

HARP

HARP

高性能用户态协议栈

NCCL存在的问题

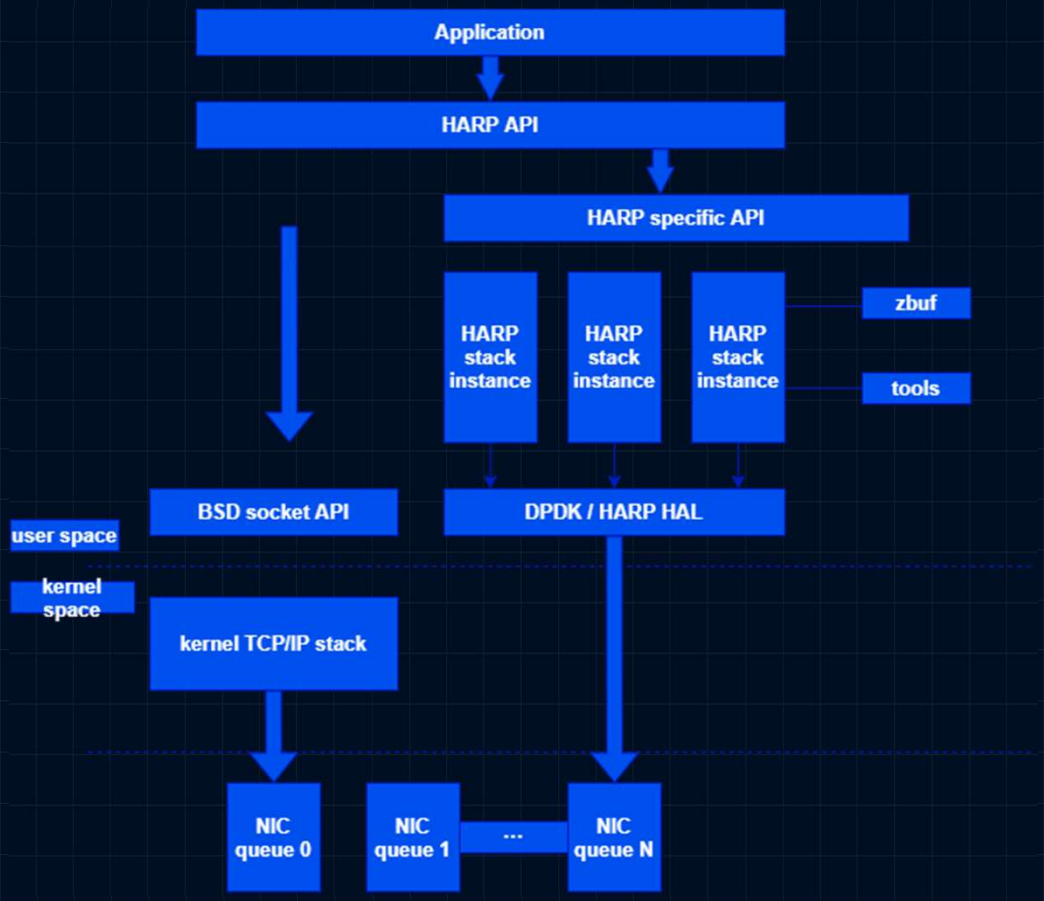
- 训练规模及数据集越来越大，多机分布式训练中网络通信占比越来越重
- 云上没有RDMA的环境下，内核Socket通信效率低

- Bypass Kernel，全路径内存零拷贝
- 多核性能线性提升
- 无锁设计，低cache miss，高速处理IO等
- Plugin的方式集成，无需任何业务改动

未来：

- 支持多路径传输
- 小数据量延迟更低
- IO长尾延迟低

HARP架构图

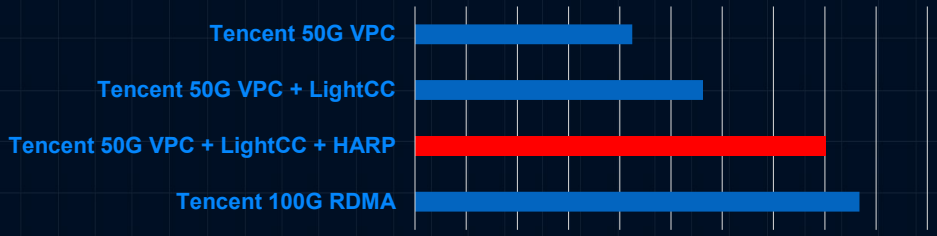


HARP

高性能用户态协议栈

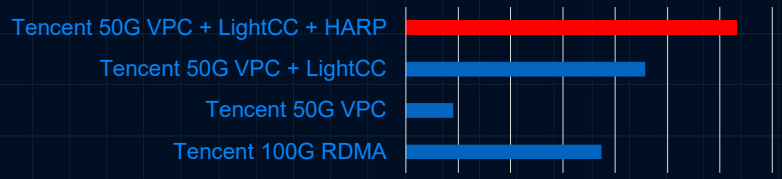
Resnet50 双机16卡AI训练线性加速比

(Synthetic | batch=256)

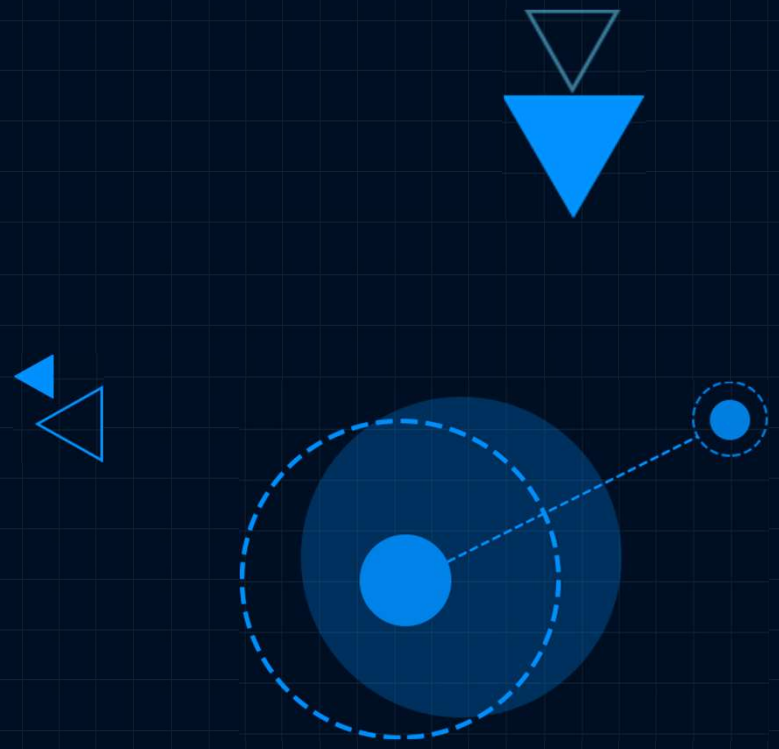


Transformer-XL 双机16卡AI训练线性加速比

(wikitext-103 | batch=32)



TACO推理引擎



TACO推理引擎

推理优化组件

ZenDNN

Goya

推理优化组件

一键式推理优化组件，性能最优的同时不影响业务部署

训练框架众多

TensorFlow

PyTorch

mxnet

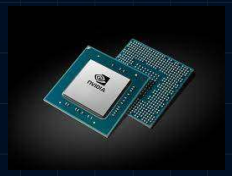
Caffe2

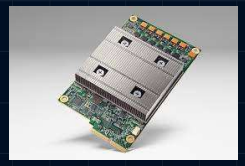
Microsoft Cognitive Toolkit

飞桨

...

硬件平台众多



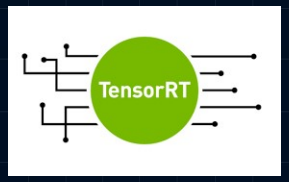


...

推理优化组件

一键式推理优化组件，性能最优的同时不影响业务部署

如何推出一款适合云端的推理框架？



推理优化组件

一键式推理优化组件，性能最优的同时不影响业务部署

TACO Inference 推理组件的优势

- ◆ 全面支持主流模型格式 TF, pyTorch
- ◆ 优化对用户透明（不改变模型格式），不影响模型部署
- ◆ 自研编译优化
- ◆ 充分集成社区优秀推理优化技术：TensorRT, TVM
- ◆ 独家高性能算子库集成 AMD ZenDNN

TACO推理引擎

推理优化组件

ZenDNN

Goya

ZenDNN

针对AMD CPU定制优化算子库

由于推荐场景模型尺寸非常大，GPU显存容量有限，CPU进行模型训练和推理占主流。

问题

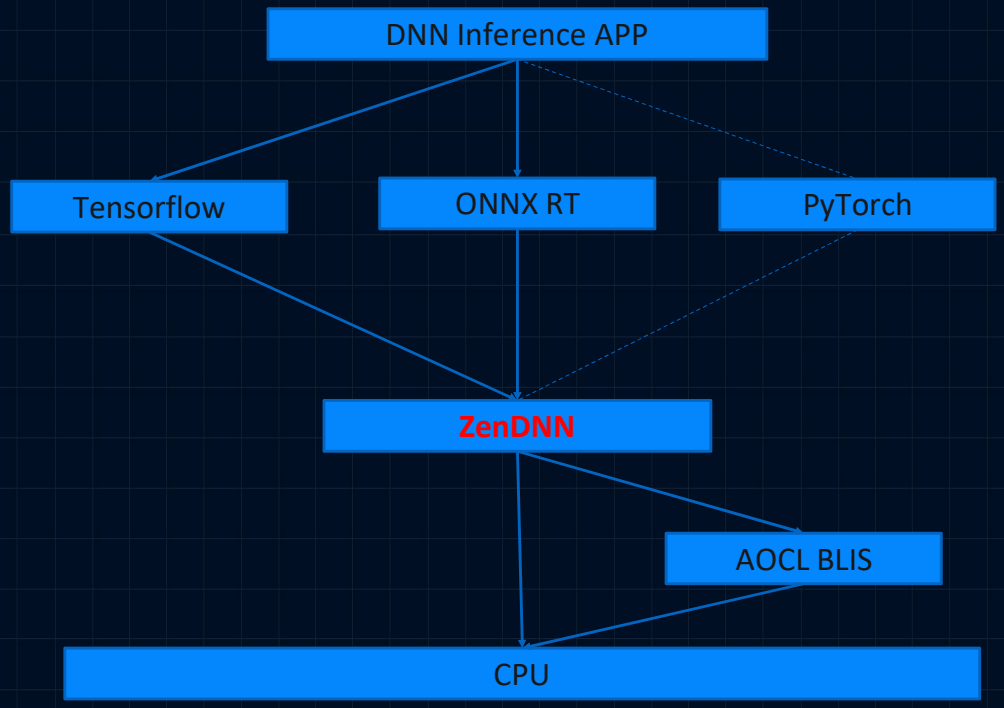
- AMD CPU价格低
- 主流深度学习库MKL-DNN针对AMD CPU没有优化

INTEL® MKL-DNN

ZenDNN

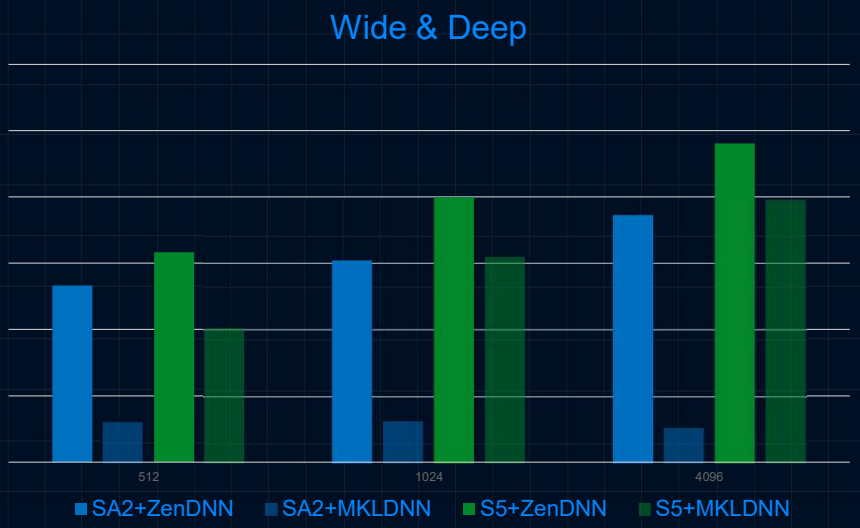
针对AMD CPU定制优化算子库

腾讯云与AMD签署战略合作协议，将ZenDNN集成到TACO推理引擎当中，致力于提升公有云内外部客户的业务性能



ZenDNN

针对AMD CPU定制优化算子库



- 腾讯云SA2机型使用AMD EPYC Rome(2.6GHz)
- 腾讯云S5机型使用Intel Xeon Cascade Lake 8255C(2.5 GHz)

TACO推理引擎

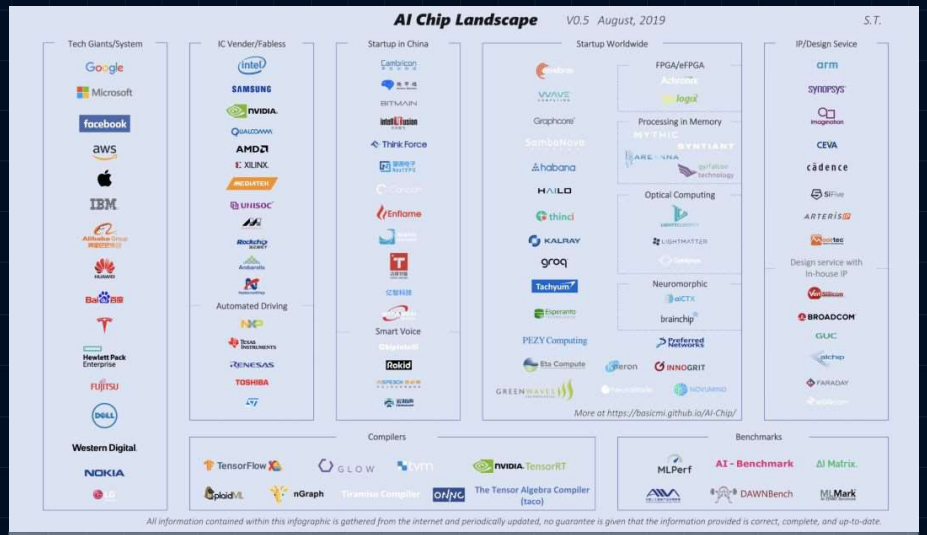
推理优化组件

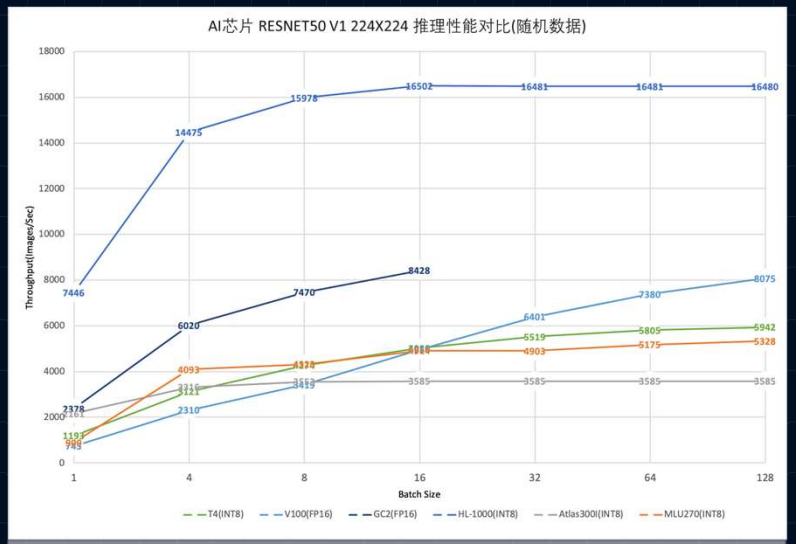
ZenDNN

Goya

Goya

高性价比NPU硬件

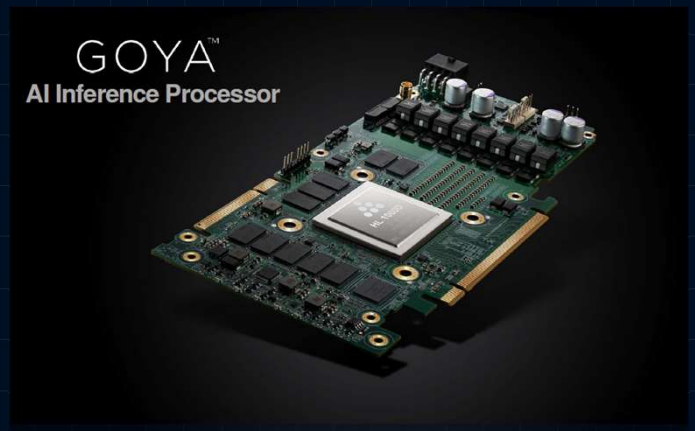




Goya

高性价比NPU硬件

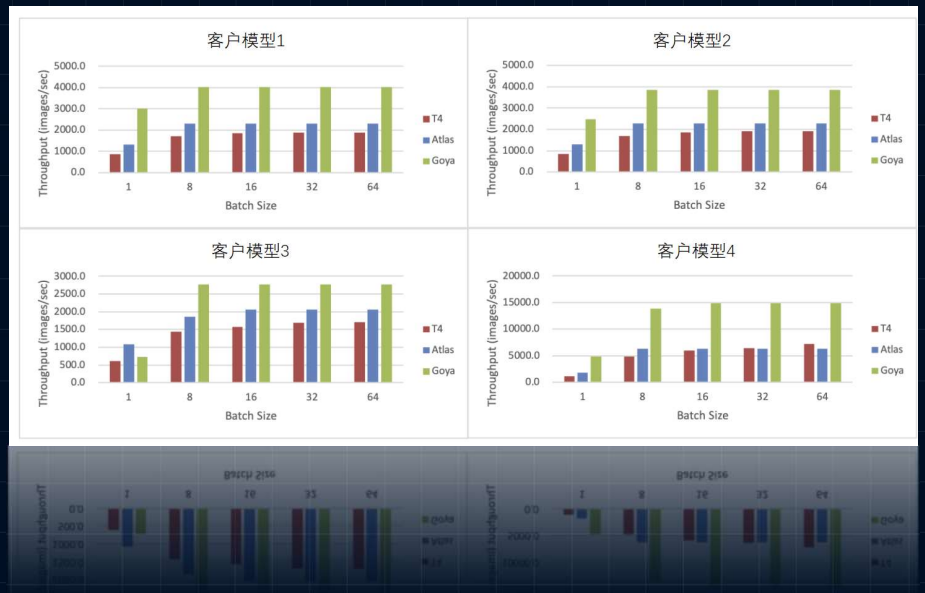
Habana Labs 成立于2016年，目前为 Intel 的全资子公司。
Goya HL-1000 是 Habana Labs 推出的业界首款专为深度学习推理应用而研发的专用 AI 处理器，旨在提供卓越的计算性能，优秀的能耗比和成本节省。
Goya 单芯片 INT8 算力高达 200 TOPS，处于行业领先水平。
Goya 基于 PCIe 4.0 的标准设计，使其可部署在各大主流计算服务器，为云端、数据中心和其他新兴应用提供新的人工智能处理方式。



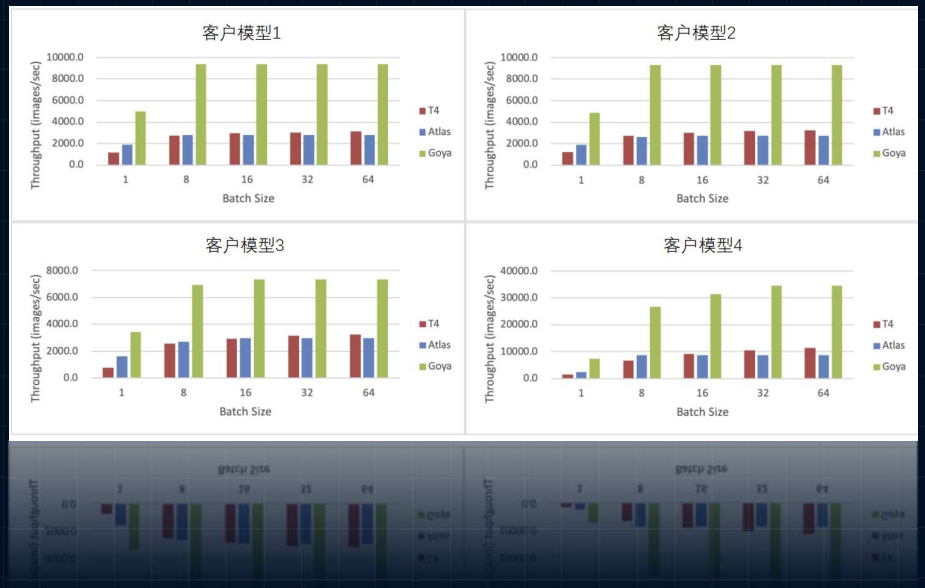
Goya

高性价比NPU硬件

INT16：1.2 ~ 4.0倍性能提升



INT8：2.2~4.8 倍性能提升



谢谢