蚂蚁大模型存储加速
PCache

蚂蚁大模型存储加速团队

ANT GROUP

InfoQ 极客传媒

极客邦科技 2024 年会议规划

促进软件开发及相关领域知识与创新的传播

访问大会官网

参会咨询

ANT GROUP

1. 大模型存储的问题和挑战
2. 蚂蚁 AI 存储加速方案

   （整体架构 + 各场景方案）
3. 未来计划

ANT GROUP

1. 大模型存储的问题和挑战

训练任务各 IO 阶段的影响

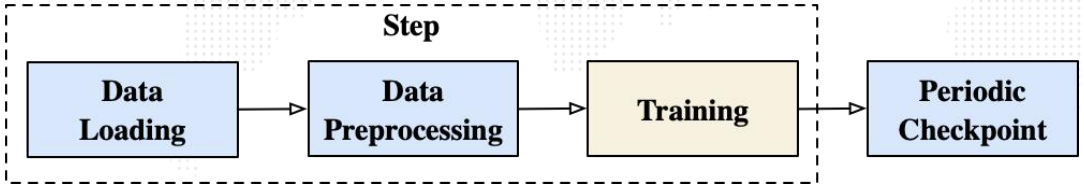
数据加载阶段（数据读取 + 预处理）

- 数据读取：IO wait 会导致 GPU 资源浪费。
- 预处理：计算性能不足会导致 GPU 资源闲置。

Checkpoint 阶段

- 写 chkpt：IO wait 会导致 GPU 资源浪费。
- 降低写入频率同样会导致 GPU 资源浪费。

  e.g., chkpt / 3h，故障时浪费 3h GPU 资源。



AI 数据读取的挑战

数据规模大

- 多模态任务的训练集达到百亿，PB级数据。

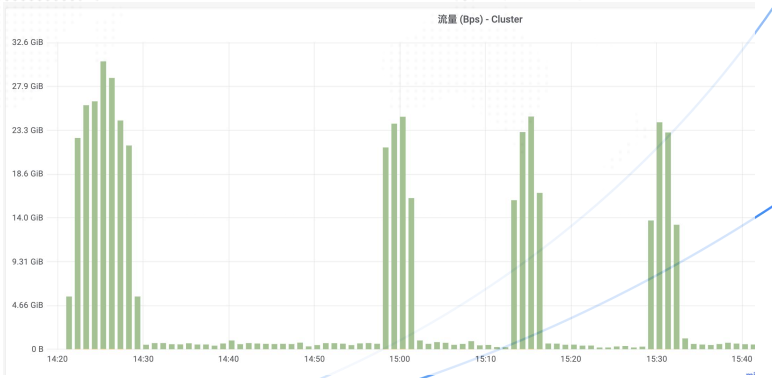
数据 & 读写操作类型多

- 图片、视频、文本、checkpoint，以及结构化数据等多种类型。
- 涵盖顺序读和随机读，甚至在一次数据加载中。

流量特性复杂

- 各类大模型训练任务数据读取时流量特性多样。

文件类型	文件大小	数量	读写操作
图片	1KB～100KB	百亿级	顺序读
视频	10MB～1GB	千万级	随机读
Checkpoint	1GB～10GB	百万级	顺序读、写
NLP 文本	10MB～10GB	千万级	顺序读，随机读
列存结构化数据	100MB～1GB	百万级	顺序读，随机读



ANT GROUP

Checkpoint 写入的挑战

Checkpoint size 不断增大，对写入性能要求越来越高（可靠性 + 吞吐）

- 千亿参数 checkpoint TB级
- 万亿参数 checkpoint 10TB级

为了减少 GPU 故障对训练的影响，checkpoint 频率越来越快

- 从天级 -> 小时级 -> 分钟级 -> 每个 step
- 虽然 FSDP 等并行模式可以减少每卡的写入量，但是 per step 的写入频率对存储高并发下写入性能的要求仍然非常高。

ANT GROUP

多云数据互通问题

算力资源紧张，多算力中心（私有云 + 公有云）

- 数据分布在多中心，跨云访问效率低。
- 缺少数据同步工具，导致训练效率低。
- 数据管理混乱，多云存在重复数据，导致存储空间浪费。

2. 蚂蚁 AI 存储加速方案

PCache 整体架构

用户接入

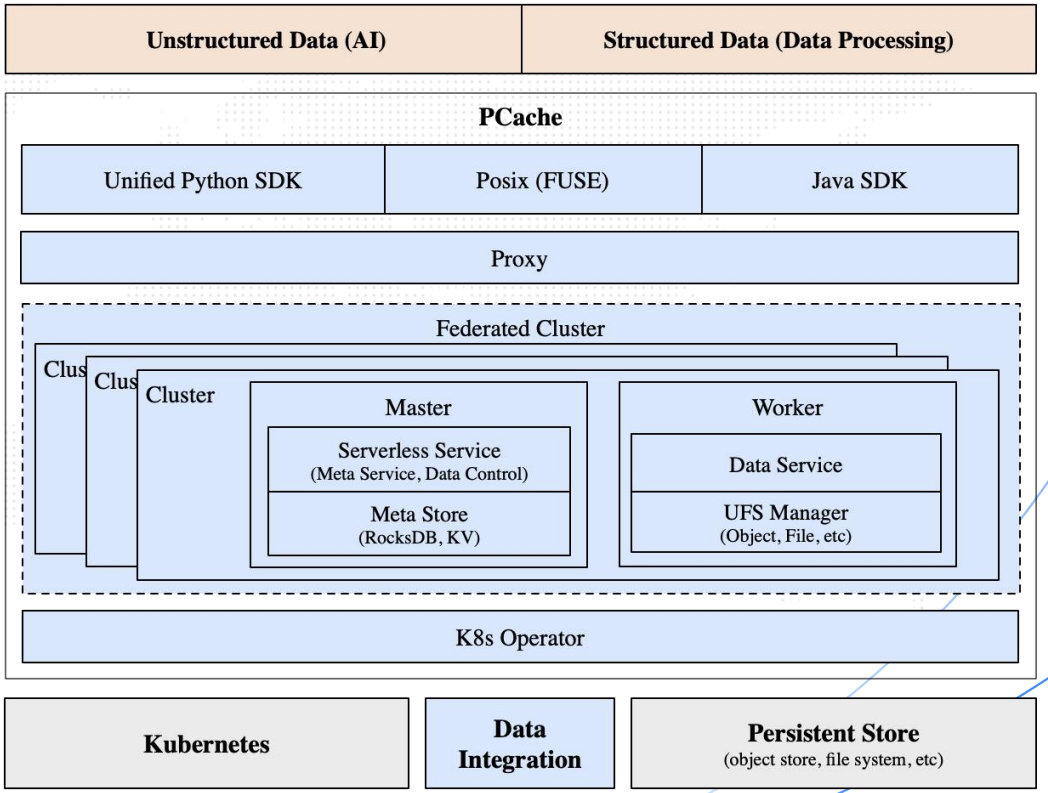
- 多类型 + 多语言API， 支持结构化和非结构化多计算场景的缓存加速需求。

PCache Runtime

- 基于联邦集群的横向扩展，Proxy 统一数据操作入口屏蔽用户对联邦集群的感知。
- Master 负责元数据服务，支持内置存储和分离 KV 两种模式。
- Worker 负责数据块的读写、副本、生命周期、存储分层管理，以及 UFS 的管理。

基础设施层

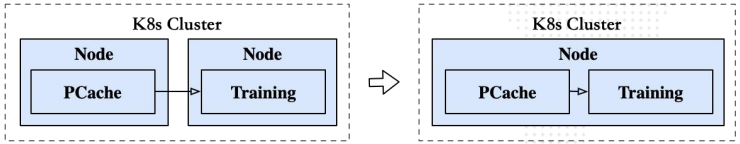
- 云原生存储
- 支持多类型持久化存储
- 分布式数据集成系统

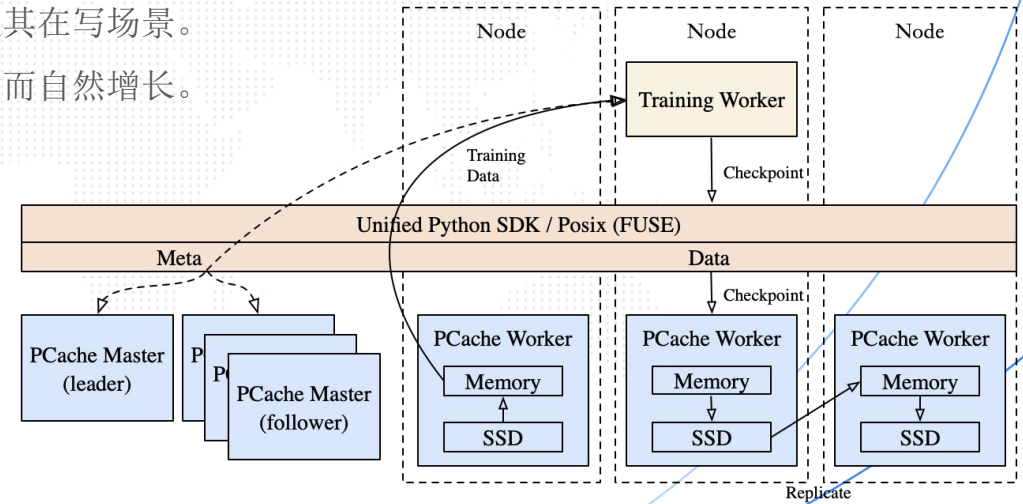


AI 数据链路 & 部署架构

部署形态的变化，存储分离 -> 混部

- 低成本：充分利用 GPU 机器上的存储和计算资源。
- 高性能：Co-locate 带来的局部性能提升，尤其在写场景。
- 扩展性：存储能力能够随着训练集群规模扩大而自然增长。





ANT GROUP

多模态场景碰到的问题

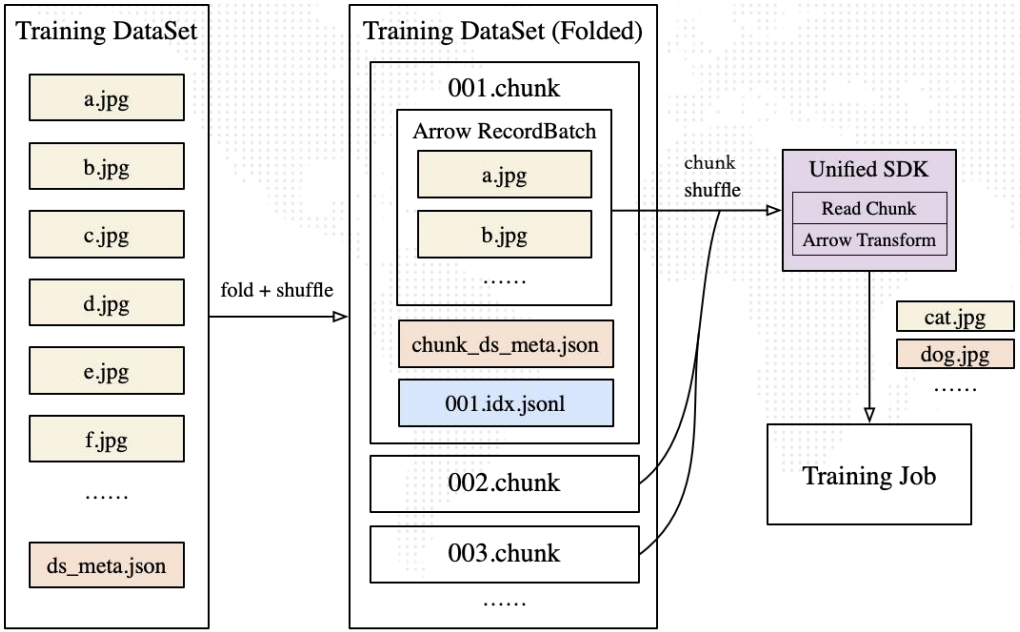
挑战1：支持海量图片的训练数据

- 如何支持亿级甚至百亿级的元数据管理。
- 如何保障百亿规模下的元数据读写性能。

挑战2：多模态场景下数据读取性能

- 图片、视频、音频、文本等不同模态数据读取时如何保障顺序 + 随机混合读取的性能。

文件折叠 - 减少元数据规模



性能提升

- 大幅减少元数据数量和读取请求。
- 线上的多模态任务的数据读取性能提高 2～4 倍。

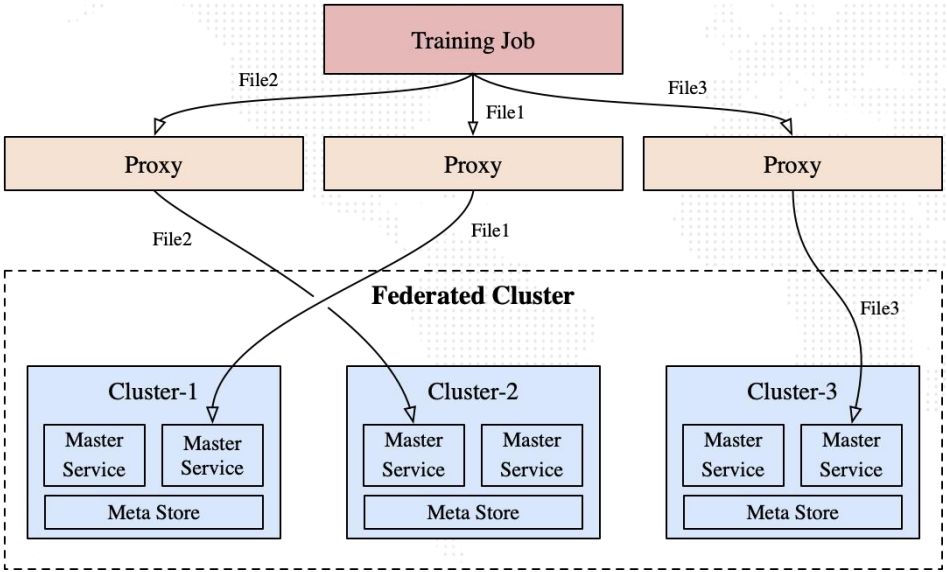
训练效果

- 从 training loss 等指标来看，从文件变为 chunk 级别的 shuffle，对训练效果没有影响。

多维度折叠

- 除了数量单一维度的折叠外，现在也出现了越来越多的多维折叠需求，e.g.，卫星图片场景下的时空维度。

元数据管理优化



联邦集群

- 提供集群级别的元数据横向扩展能力。
- 通过 Proxy 屏蔽用户对联邦集群的感知。

元数据存储 & 服务分离

- Serverless master 提供横向扩展能力。
- 支持内置和外置两种 meta store 模式。

预取优化

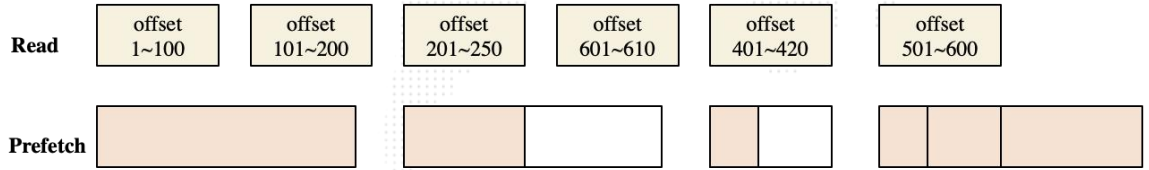
预取的问题

- 对顺序读友好，随机读时有读放大问题。
- 在混合读取时，开启预取有明显的抖动。

启发式的预取

- 根据历史的读取操作，动态的调整预取窗口大小。
- 在混合读取场景下，能够有效减少抖动，提高整体吞吐。

  Note: 窗口策略可调整，从简单的2分到基于历史的moving window，甚至是预测。



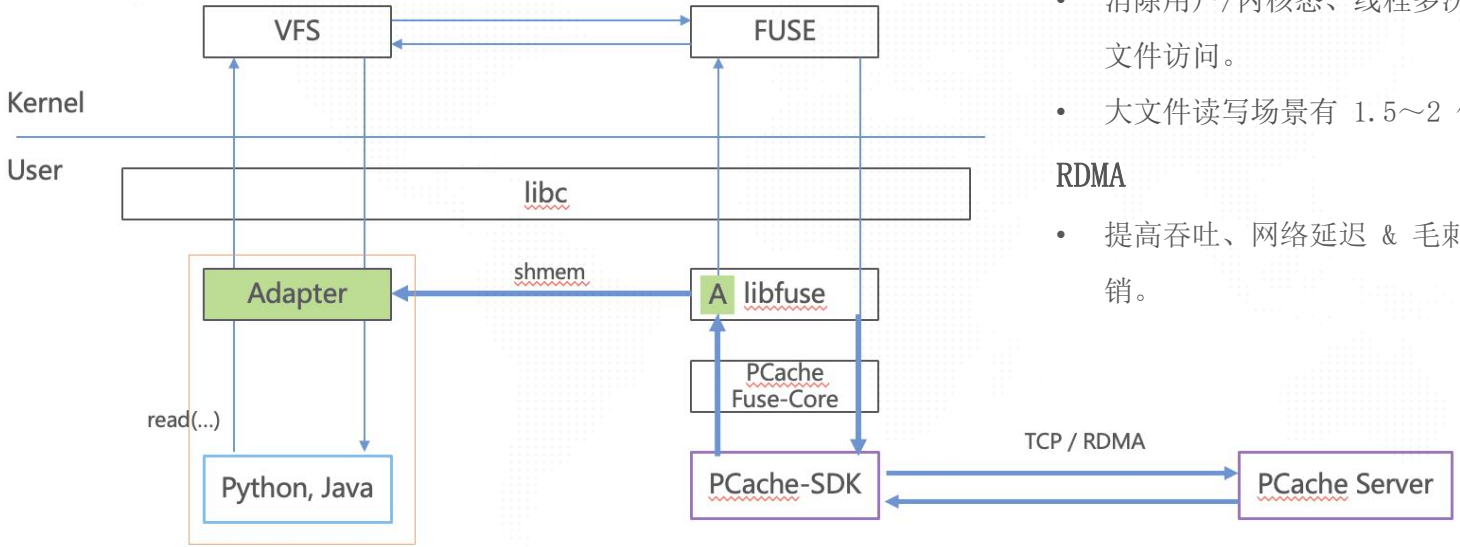
NLP 场景碰到的问题

挑战：支持大规模训练任务 checkpoint 的高频写入

- 文件大：千亿参数 checkpoint TB 级，万亿参数 checkpoint 10TB 级。
- 写入频率高：为了减少故障时的 GPU 资源浪费，需要提高 checkpoint 写入频率，甚至到每个 step。

以一个 1K 卡以上集群平均每天发生一次 failover，如果 3 小时做一次 checkpoint，那对千亿参数的训练任务来说平均每天就会浪费 3 小时的 GPU 资源。

用户态 FUSE + RDMA



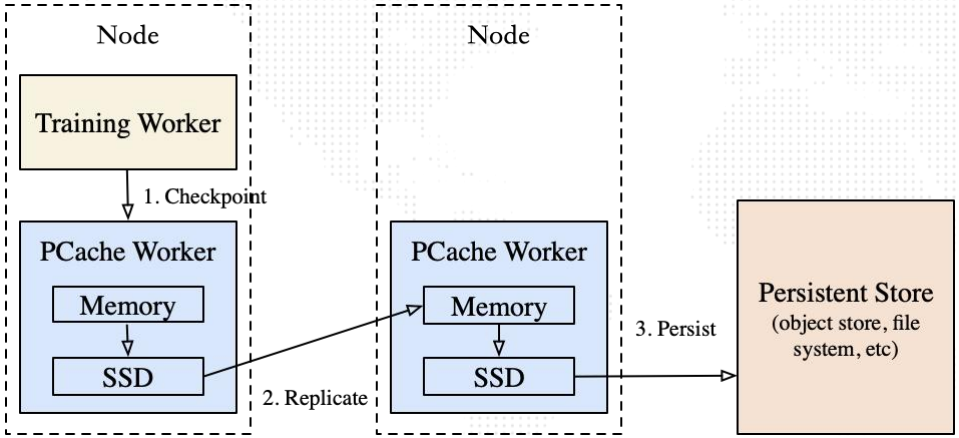
用户态FUSE

- 消除用户/内核态、线程多次切换拷贝，加速中大文件访问。
- 大文件读写场景有 1.5～2 倍左右的性能提升。

RDMA

- 提高吞吐、网络延迟 & 毛刺、客户端 CPU 开销。

Checkpoint 写方案

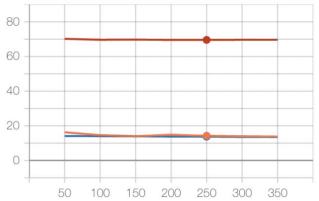


本地加速 + 写入流程异步化

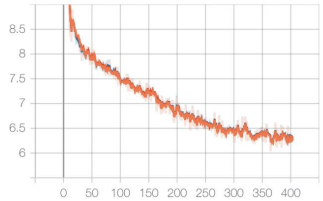
- 优先写入本地 worker，加速写入性能。
- 让副本同步和持久化异步化，不会阻塞 chkpt 过程。

效果

- 配合 FSDP 等并行模式，千亿参数的 chkpt 在训练每个 step 的开销占比可以降低到 < 0.1%，实现 chkpt per step



save_chkpt time



train_loss

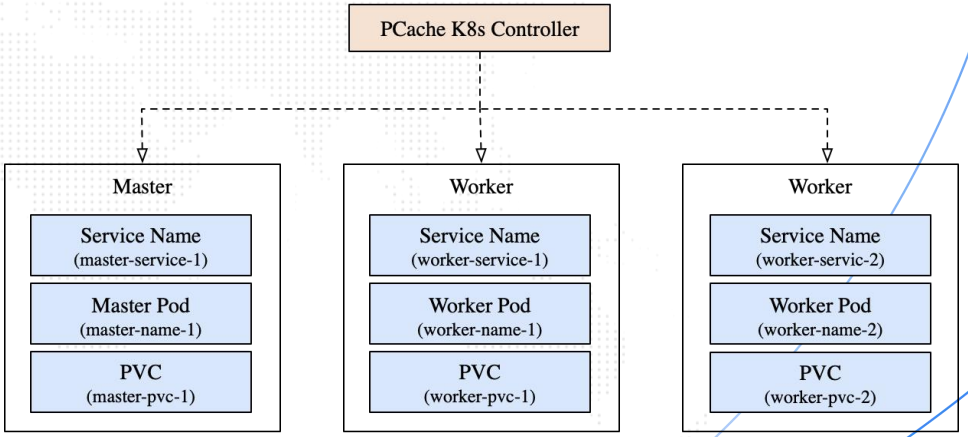
稳定性优化 - 云原生存储

基于云原生的方式部署服务

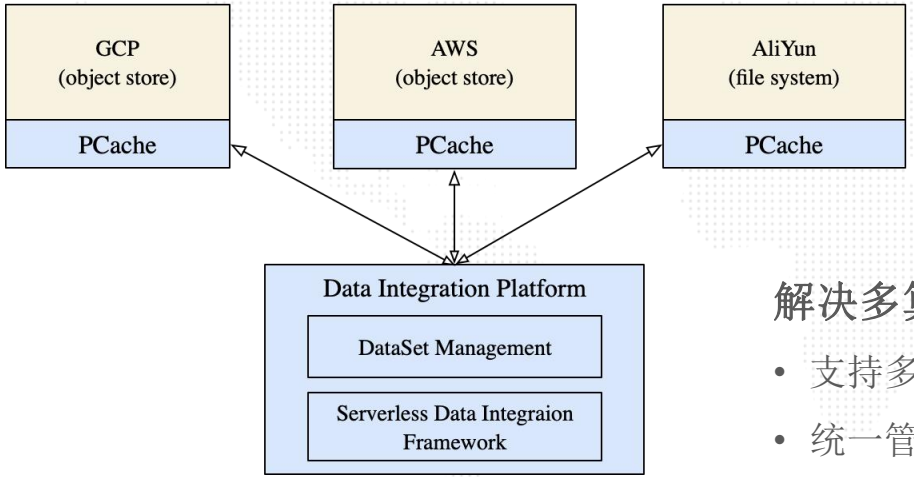
- POD 管理计算资源，PVC 管理存储资源

故障时的自动回复

- 通过 POD name、PVC name 的管理，保障容器重启后对外服务地址不变，数据不丢。
- 物理机故障下线时，通过 K8s 编排能力，在容器恢复后，自动做数据预热。



多云数据同步



解决多算力中心下的数据问题

- 支持多种持久化存储，能够在多云环境提供加速。
- 统一管理多云环境的数据集，避免大量重复数据。
- 集成高性能分布式数据集成工具，提高数据迁移效率。

3. 未来计划

ANT GROUP

面向 AI 数据特性的缓存策略

AI 数据特性

- 训练样本数据有可替代、可预测、重要性等特性。

基于 AI 数据特性的缓存策略

- 性能：基于有效的数据特性，可以让缓存更加高效，保证数据始终在缓存里。
- 缓存效率：加速数据淘汰，可以支持较低的缓存 & 持久化存储比。
- 稳定性：在缓存节点故障时，通过数据可替代性可以减少穿透读带来的性能下降。

ANT GROUP

计算、存储统一调度

训练过程中计算 & 资源问题

- 计算资源：当前 GPU 机器上会空闲大量的计算资源，e.g. CPU、MEM。
- 数据预处理问题：在多模态场景中，有不少数据预处理的计算，传统的串行 pipeline 会导致训练长时间等待，浪费 GPU 资源。

计算、存储统一调度/编排

- 可以在数据传输前，根据数据分布统一调度预处理计算、数据读取和训练任务，提高训练的效率，以及 GPU 机器的资源利用率。

InfoQ 极客传媒

极客邦科技 2024 年会议规划

促进软件开发及相关领域知识与创新的传播

访问大会官网

参会咨询

THE END

THANK YOU!