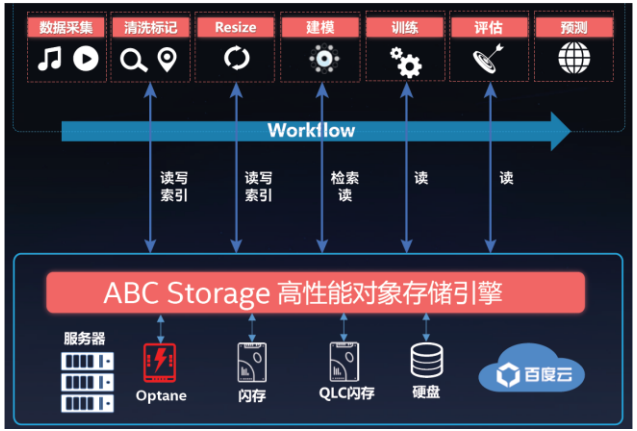
为AI训练和超算定制高性能存储

百度智能云* 全闪对象存储方案导入QLC固态盘+傲腾™ 固态盘组合

高性能存储

AI, 大数据，高性能计算



由 ABC Storage 高性能存储解决方案支持的AI训练流程

5%以内
文件数据增加10倍时，QPS和延迟波动保持在

60%
TCO 降低

1-3倍
用户业务效率提高

可部署于私有云，专攻 AI 训练、大数据和高性能计算

将久经考验的高性能对象存储引擎引入方案

“全闪” = 傲腾™ 固态盘 + QLC 3D NAND固态盘

intel XEON GOLD inside

intel OPTANE DC SOLID STATE DRIVE

intel SSD inside

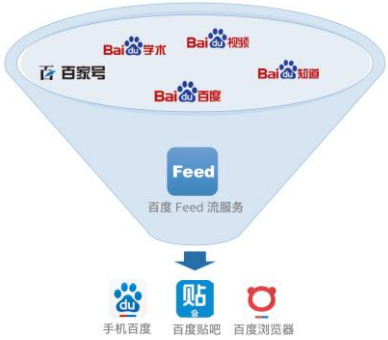
来自英特尔的全闪产品组合，及其与英特尔® 至强® 可扩展处理器的配合，帮助我们的方案在稳定性、IOPS等方面实现了更优表现，成为应对海量非结构化小文件的得力手段。

百度智能云私有云存储团队

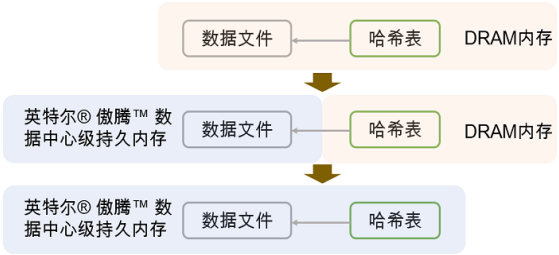
借助傲腾技术重构内存数据库

百度* Feed 流服务采用傲腾™ 数据中心级持久内存承载海量数据

内存数据库　搜索引擎



百度核心应用中的Feed流服务



百度Feed-Cube内存硬件迁移路径

千万级
每秒查询量
考验顺利过关

TCO
大大降低，相比
采用DRAM

PB级
未来将承载的数
据量高达

在更多核心业务场景中考验傲腾数据中心级持久内存

将承载 Feed-Cube* 全部数据的内存全部
换为傲腾数据中心级持久内存

将Feed流内存数据库Feed-Cube* 从纯DRAM部署方式
转向DRAM+持久内存混合部署方式

intel XEON PLATINUM inside

intel OPTANE DC PERSISTENT MEMORY

来自英特尔的傲腾™ 数据中心级持久内存，可以帮助Feed流服务的核心模块Feed-Cube在保证高并发、大容量和高性能的同时，大大降低TCO。

汪瑁
推荐技术架构部主任架构师
百度

为飞桨* 增添“加力推进器”

百度* 飞桨INT8方案借英特尔® 深度学习加速技术提升推理效率

深度学习推理　智能图像分析

模型	数据集	FP32 准确率	INT8 准确率	准确率差值
ResNet-50	Full ImageNet Val	76.63%	76.23%	0.40%
MobileNet-V1	Full ImageNet Val	70.78%	70.47%	0.31%

FP32 和 INT8 推理准确度结果比较

模型	数据集	FP32 准确率	INT8 准确率	INT8/FP32 吞吐量比比率
ResNet-50	Full ImageNet Val	11.54 images/s	32.2 images/s	2.79
MobileNet-V1	Full ImageNet Val	49.21 images/s	108.37% images/s	2.2

FP32 和 INT8 推理吞吐量结果比较

2-3倍
使用INT8时的推理速度是FP32时的

1%以内
INT8与FP32的深度学习模型推理准确度差值

能效更优
提升推理效率、降低功耗和部署复杂度

在图像识别与分类等场景的深度学习中发挥INT8的优势

基于MKL/MKL-DNN对不同深度学习模型进行特定优化

利用第二代至强可扩展处理器集成的英特尔深度学习加速技术对INT8更优的支持

英特尔® 深度学习加速技术

英特尔® MKL　英特尔® MKL-DNN

第二代英特尔® 至强® 可扩展处理器的强劲算力及英特尔® 深度学习加速，让飞桨 INT8方案在不影响推理准确度的情况下，推理速度得以显著提升。

高铁柱
高级经理
百度深度学习平台部