DLRover

DLRover Open Day

大模型时代的 AI 基建

2023 年 7 月 8 日

北京市朝阳区东三环环球金融中心（WFC）东塔 9F 大厅

OSCHINA segmentfault

Alibaba Group
阿里巴巴集团

DeepRec: 面向推荐场景的高性能深度学习框架

彭陶

2023年07月08日

Alibaba Group
阿里巴巴集团

Contents
目录

业务场景

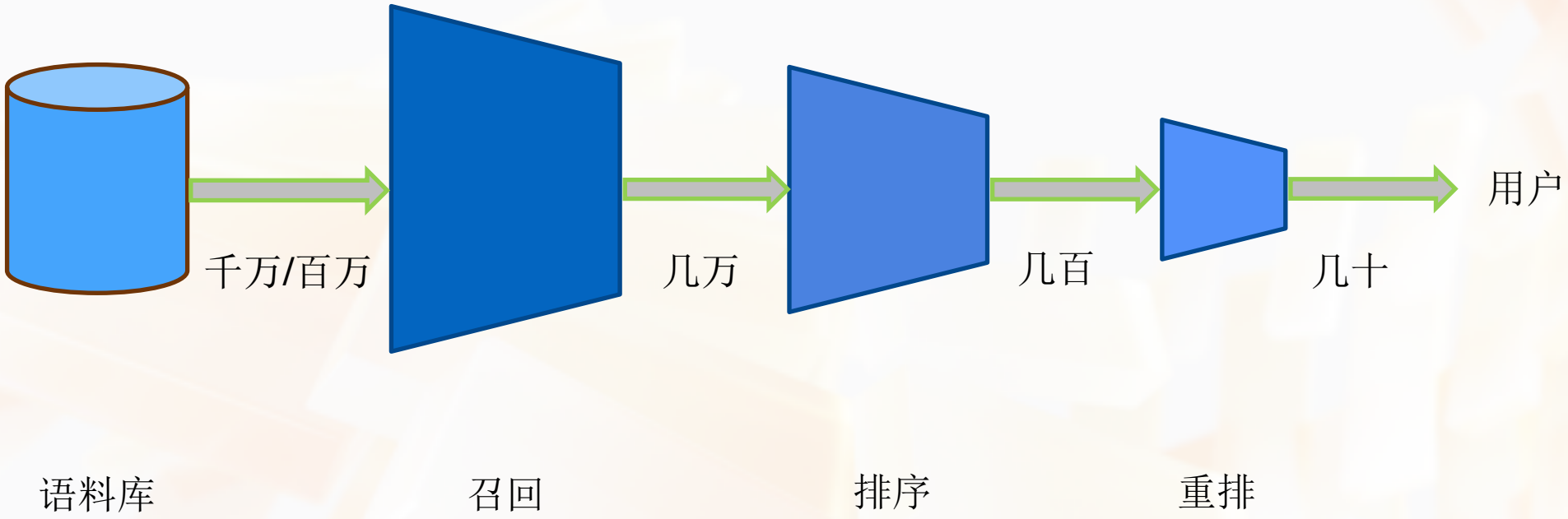
- 搜索
- 推荐
- 广告


16:50
折叠自行车
综合
销量
筛选
迪卡侬官方旗舰店20/24寸6-12儿童自行车变速山地车学
¥789.9
dahon大行20寸6速通勤折叠自行车
¥1248
官方旗舰店永久牌男女折叠自行车
¥999
迪卡侬TILT 120折叠自行车
¥1389.9
折叠自行车 11年老店 折叠自行车男女狼途铝合金架16寸女款成


16:44
宝贝
评价
详情
推荐
看了又看
迪卡侬折叠车变速城市通勤自行车成
¥2,699.9
迪卡侬TILT120折叠自行车20寸轻便
¥1,399.9
迪卡侬TILT100折叠车20寸男女轻便
¥999.9
迪卡侬电动电助力折叠车20寸自行车锂电
¥2,999.9
加入购物车
马上抢

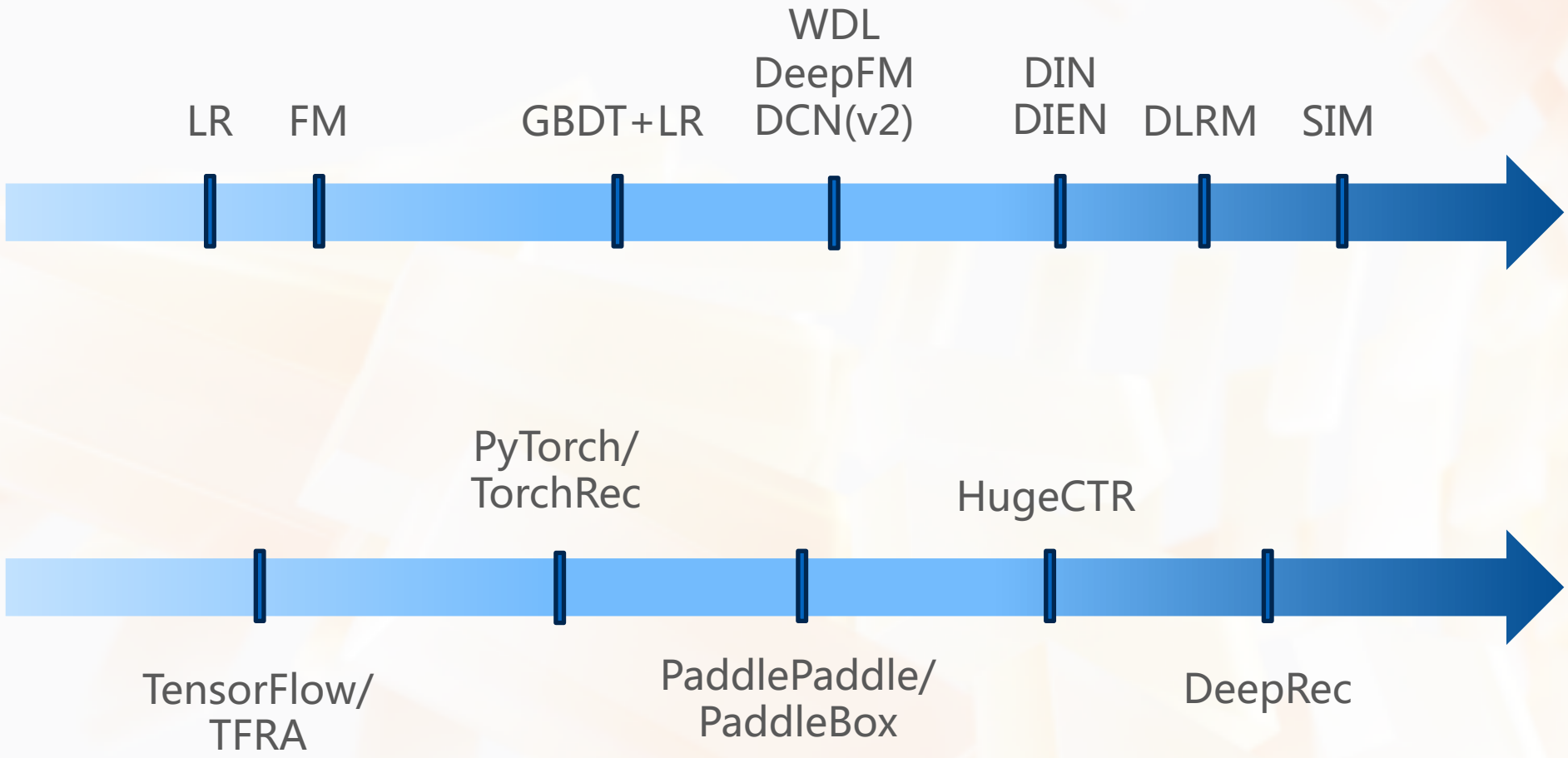

16:51
折叠自行车
综合
销量
筛选
迪卡侬TILT 120折叠自行车
¥1389.9
折叠自行车 11年老店 折叠自行车男女狼途铝合金架16寸女款成
¥1168
喜德盛铝合金折叠车架
¥1249
java折叠自行车铝合金碟刹变速
¥2198
莫曼顿iThink Expressway折叠车

一个典型的系统架构



Alibaba Group 阿里巴巴集团

稀疏模型和计算框架的发展



Alibaba Group 阿里巴巴集团

稀疏模型特点以及对框架的挑战

- **模型特点**

1. 样本量大，PB级样本，上千worker分布式训练
2. 参数量大，模型尺寸百GB-TB
3. 特征高维稀疏，千亿规模
4. 模型复杂，计算量大，异构计算
5. 在线学习趋势显著

- **框架挑战**

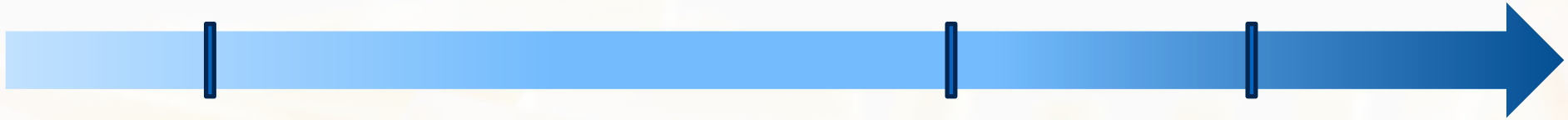
1. 易用性 (部署/调试)
2. 功能完善
3. 高性能
4. 框架灵活，高可扩展性
5. 完善的社区支持

Alibaba Group 阿里巴巴集团

开源历程

2022 DeepRec 开源

- 数十家公司在使用，包括微博，得物，喜马拉雅，VIVO等



2018 PAI-TF

- 阿里巴巴内部统一的深度学习训练框架
- 基于TensorFlow，性能优化，功能增强

2023 DeepRec 捐献

LF AI & Data Foundation

Github https://github.com/DeepRec-AI/DeepRec/

Docs https://deeprec.readthedocs.io/en/latest/

Alibaba Group
阿里巴巴集团

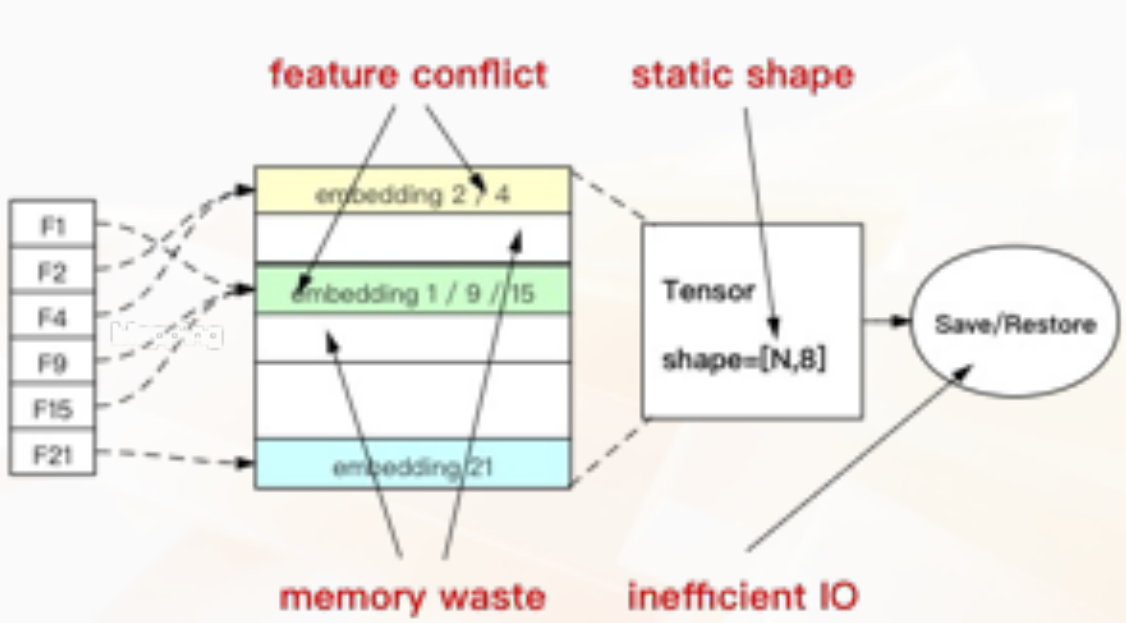
Contents
目录

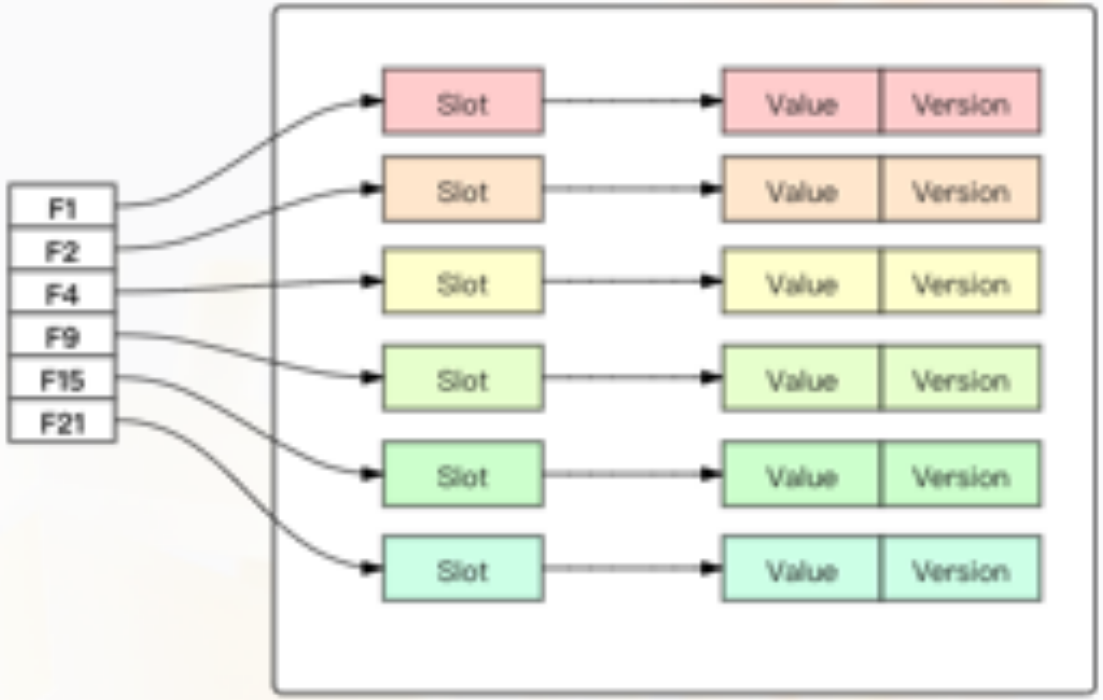
Alibaba Group 阿里巴巴集团

DeepRec 功能 概览

Embedding & Optimizer	Training	Serving & Deployment
Embedding Variable	Async/Sync Distributed Training	Share-nothing SessionGroup
Feature Eviction and Filter	Distributed Training based on GBA	Multi-tier Embedding
Dynamic dimension EV	Graph Aware Memory Allocator	GPU Multi-stream In SessionGroup
Adaptive Embedding Variable	Automatic Pipeline	Dynamic Shape Compiler (BladeDISC)
Multi-Hash Embedding	Graph Template Engine	CUDA Graph Execution Engine
AdamAsync Optimizer	Critical-path based Executor	Delta checkpoint
AdagraDecay Optimizer	GPU Multi-Stream Engine	Online Deep Learning
	Multi-tier Embedding	Model Quantization

Embedding Variable 功能



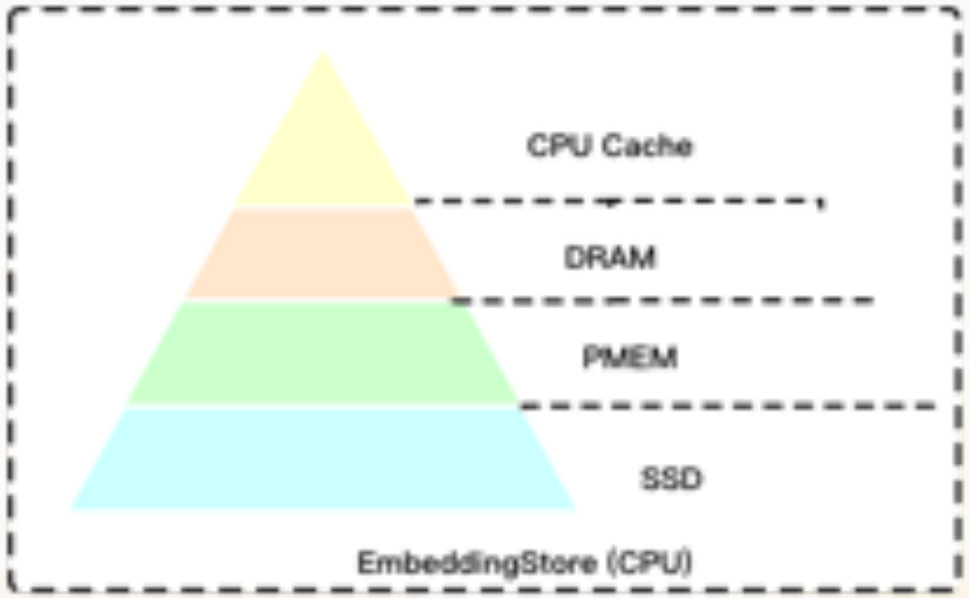


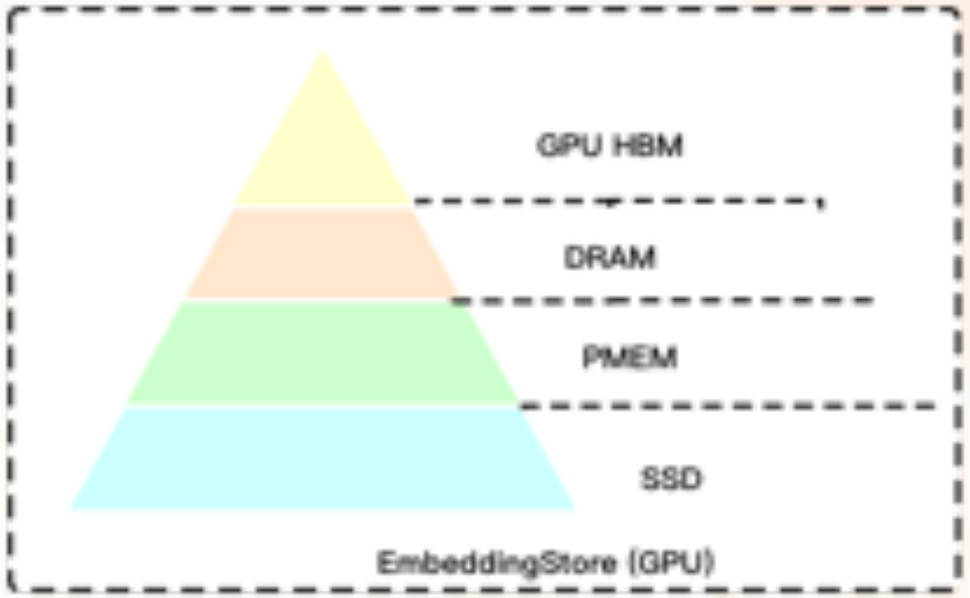
Feature Filter
1. Based on Counter
2. Based on Bloom Filter

Feature Eviction:
1. Based on Global Step
2. Based on l2weights

Alibaba Group 阿里巴巴集团

Multi-Tired Embedding Variable

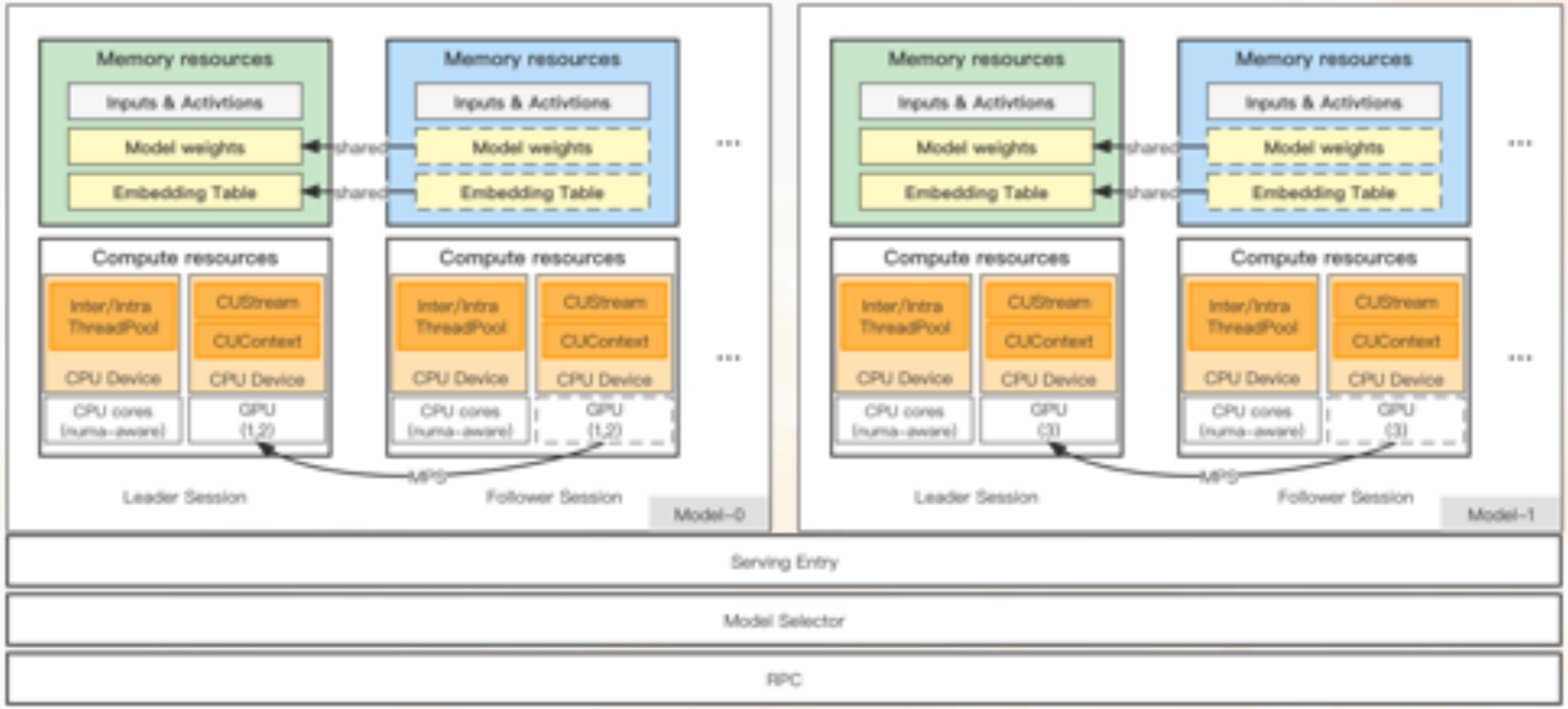




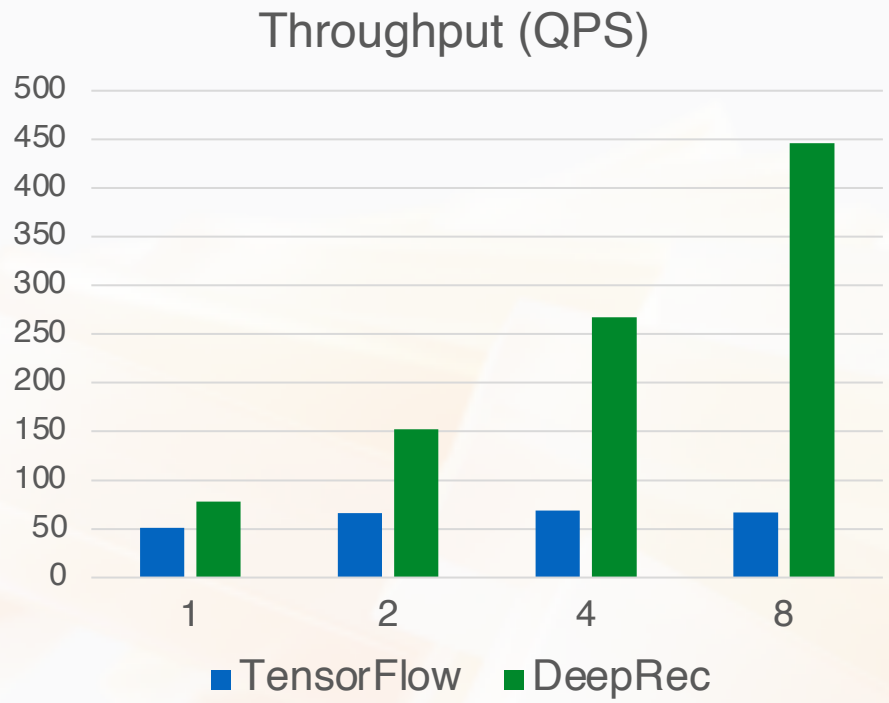
- (1-10ns) -> 100ns -> 1us ->100ns
- Support Model with 10TB+ (Training and Inference)
- Less Memory or GPU Memory Usage
- Hot/Cold Features for Embedding (Pareto Principle, hot features 20%)
- Compare with Distributed Serving, (latency TP99 80ms->25ms)

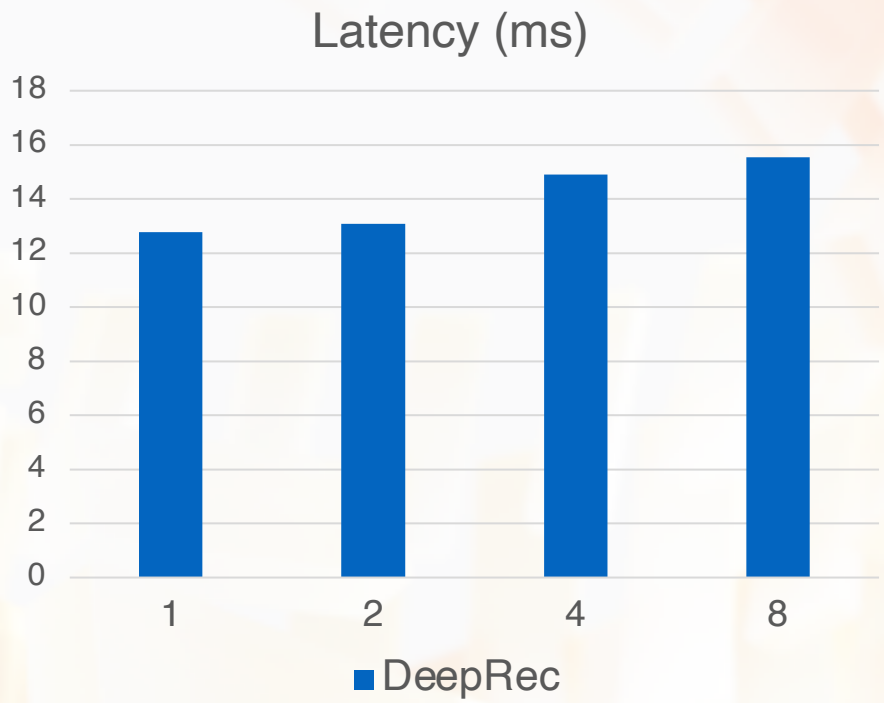
Alibaba Group
阿里巴巴集团

Share-nothing SessionGroup



Share-nothing SessionGroup

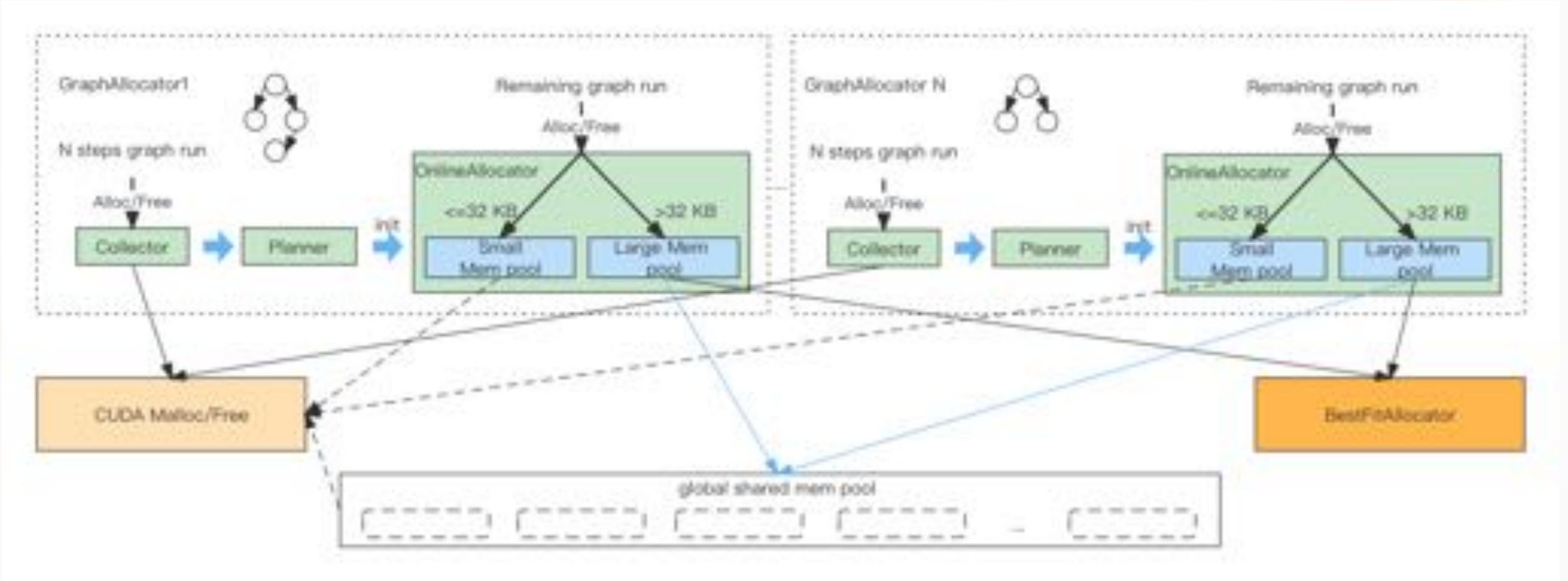




- 8 CUDA stream and share-nothing architecture could brings 6X QPS than TensorFlow

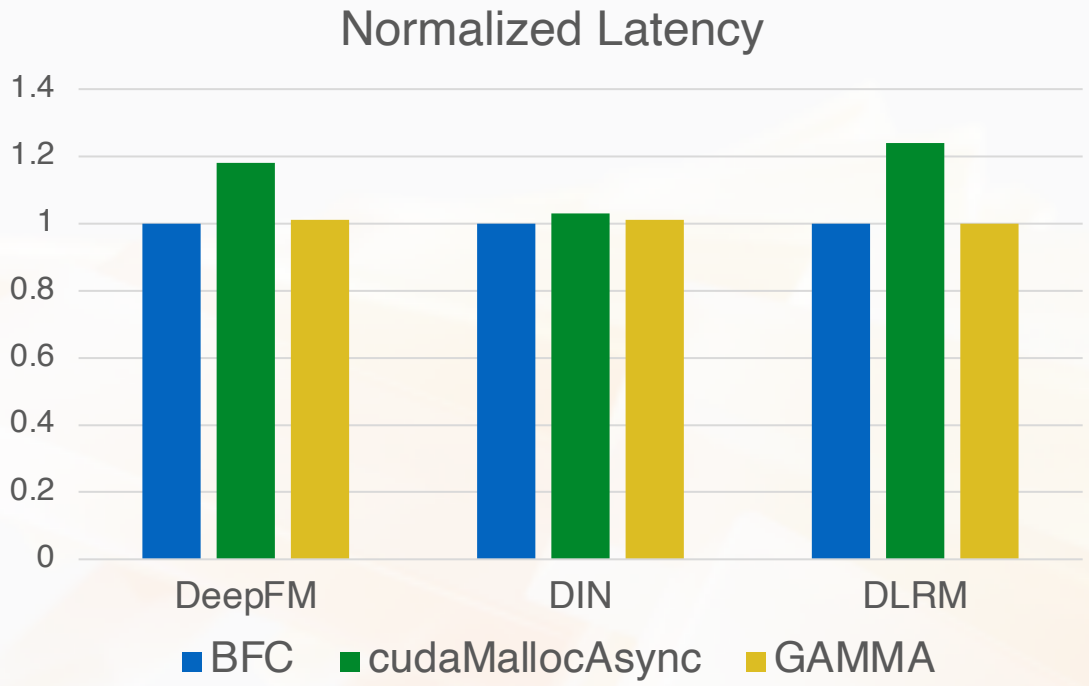
Alibaba Group
阿里巴巴集团

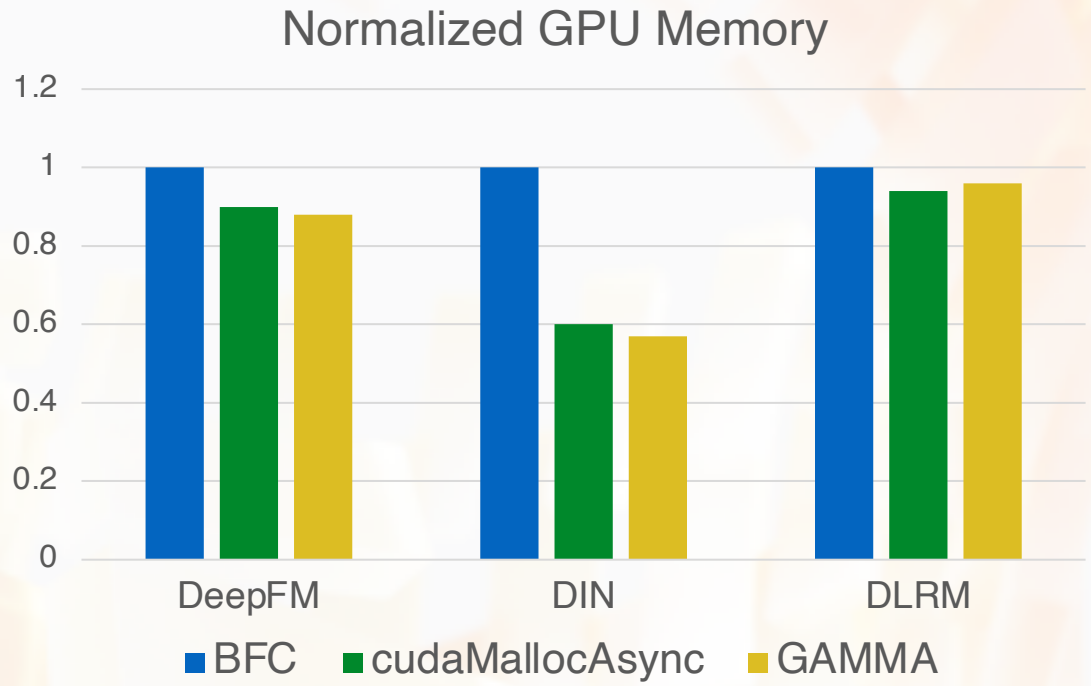
GAMMA - Graph Aware Allocator



Alibaba Group 阿里巴巴集团

GAMMA - Graph Aware Allocator

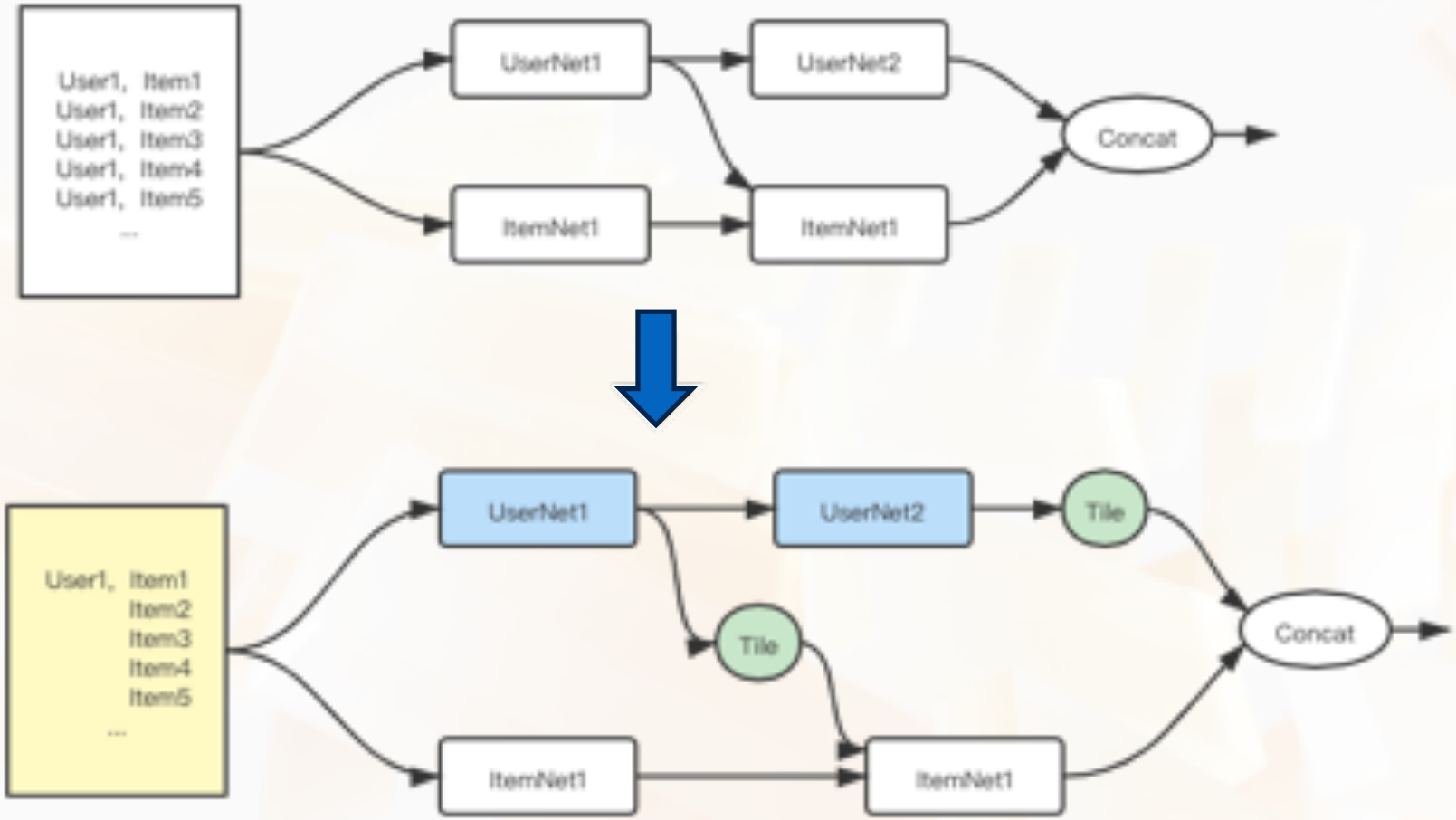




- GAMMA save 12%-43% GPU memory compare to BFC
- GAMMA improve performance at most 2%-24% compare to cudaMallocAsync

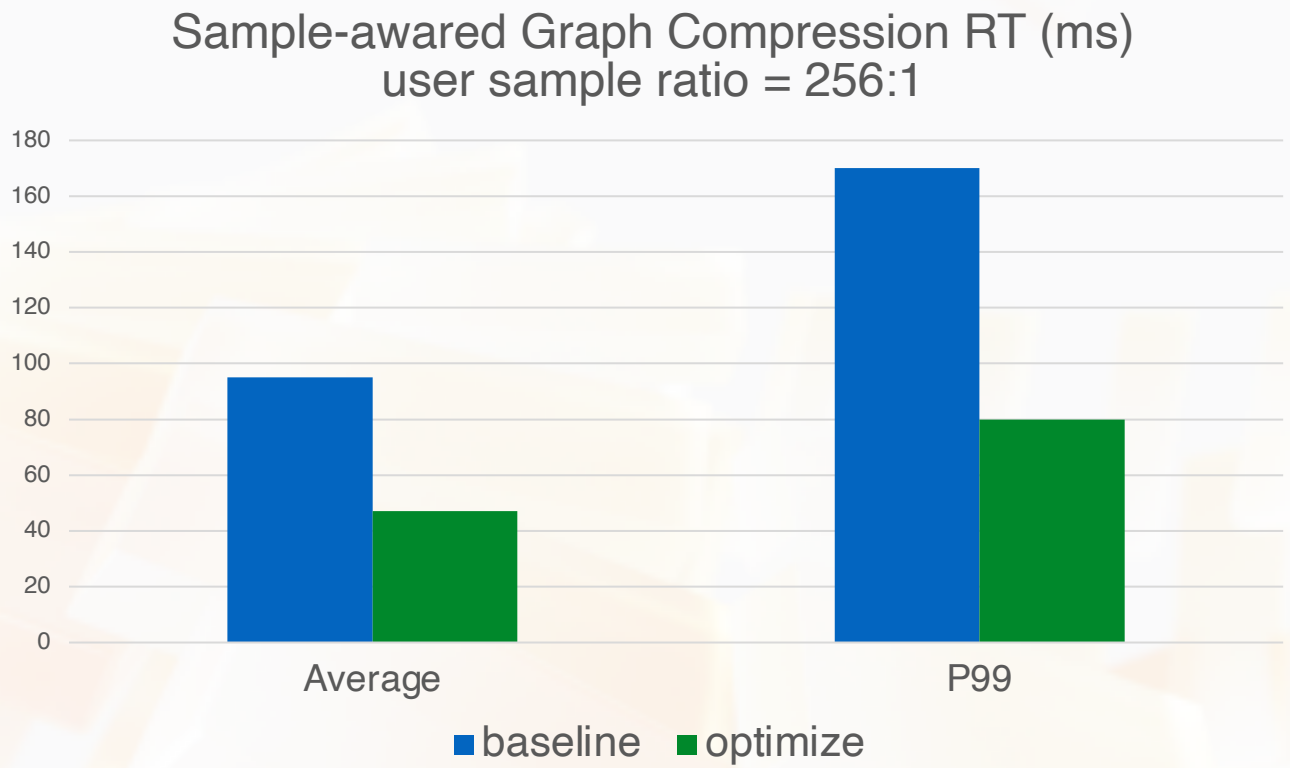
Alibaba Group
阿里巴巴集团

Sample-awared Graph Compression



Alibaba Group 阿里巴巴集团

Sample-awared Graph Compression



- Average RT drops 49%, P99 RT drops 47%

Alibaba Group
阿里巴巴集团

Contents
目录

Alibaba Group 阿里巴巴集团

技术发展的方向

Embedding 功能增强

- Group Embedding
- Embedding 量化

计算优化

- Triton
- OpenXLA

推理优化

- CUDA Graph
- Embedding Service


Alibaba Group
阿里巴巴集团
THANKS