



# 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

蚂蚁北京 | 京程说

稀土掘金

COSCHINA



示说网



segmentfault 思否

# 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

# Contents

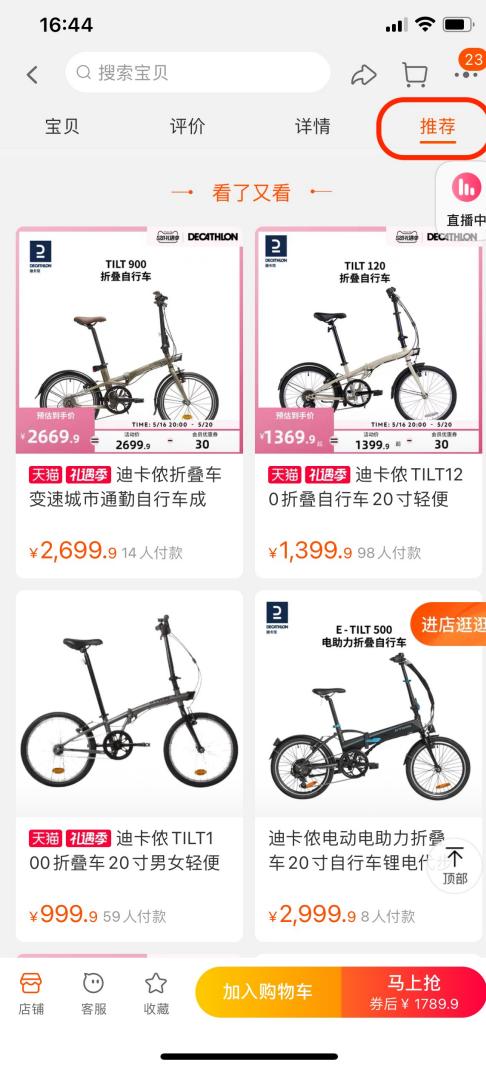
## 目录

- 01 DeepRec 背景及开源历程
- 02 DeepRec 概述及关键技术
- 03 DeepRec 展望

# 业务场景

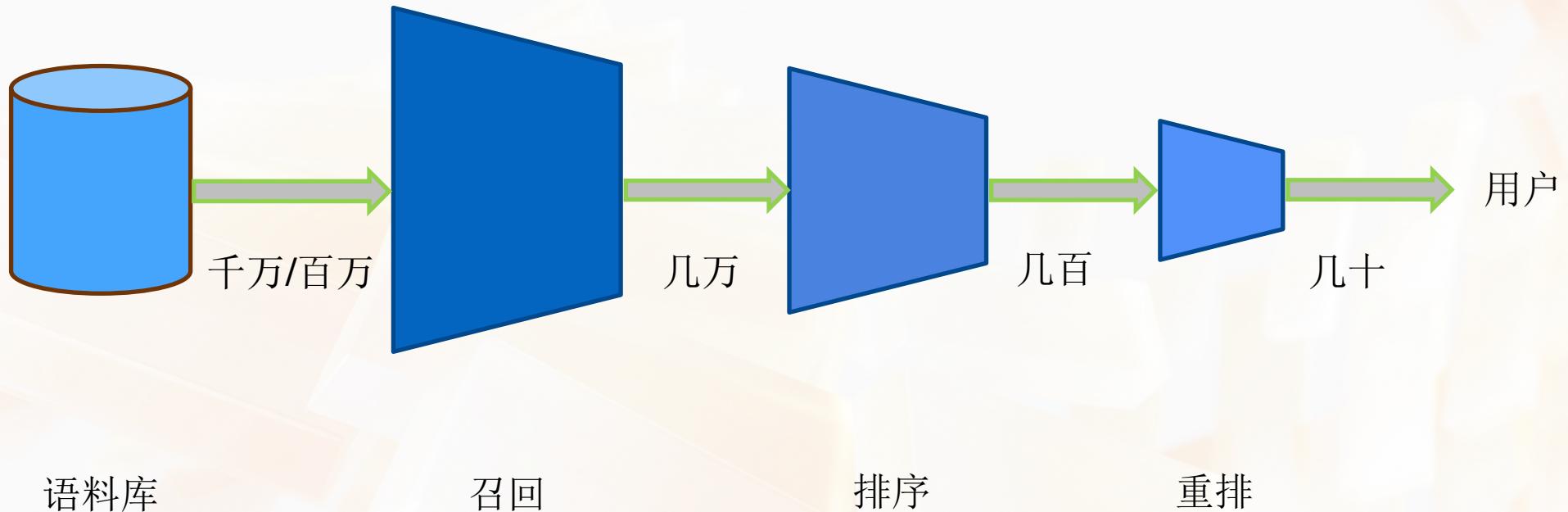
- 搜索
- 推荐
- 广告



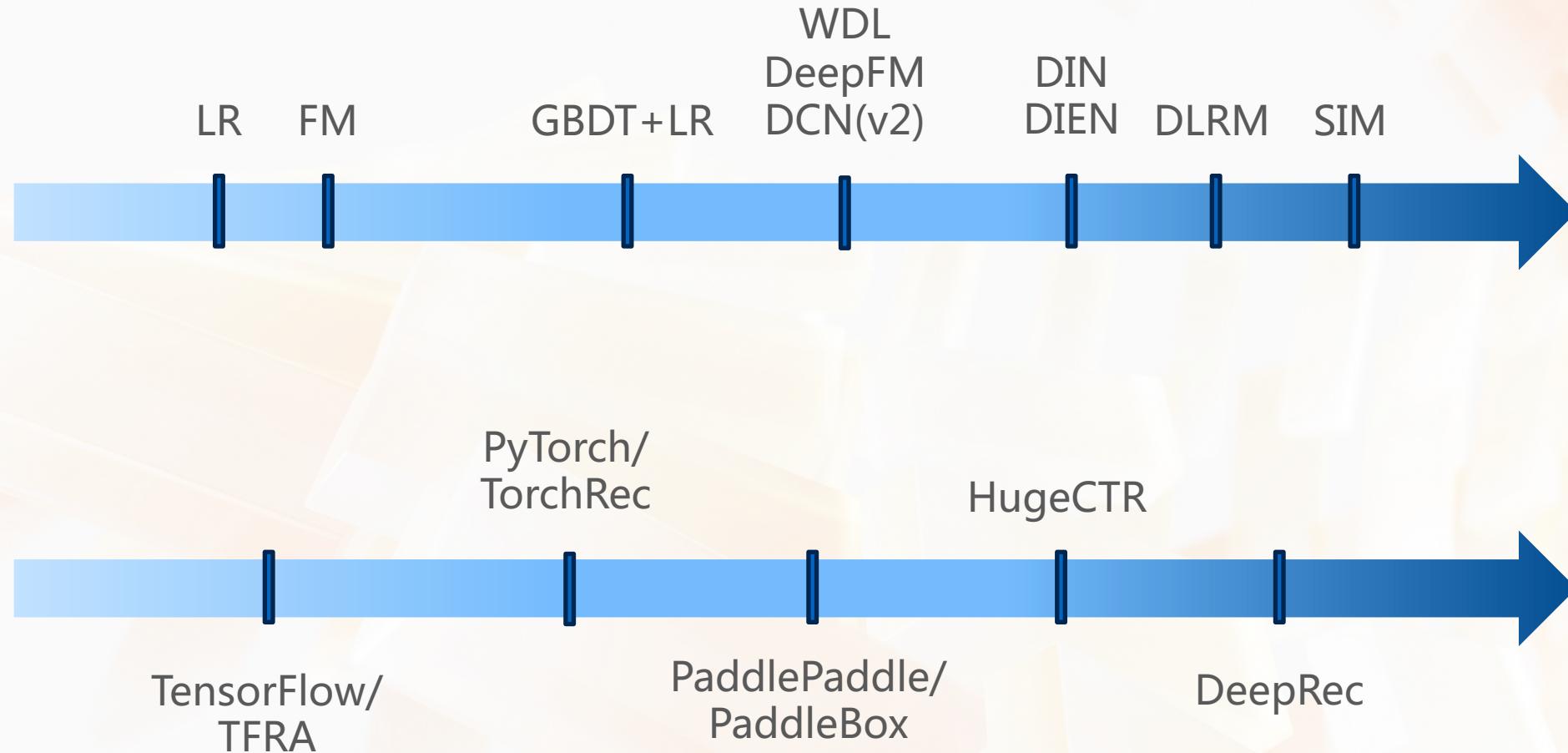




## 一个典型的系统架构



## 稀疏模型和计算框架的发展



# 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

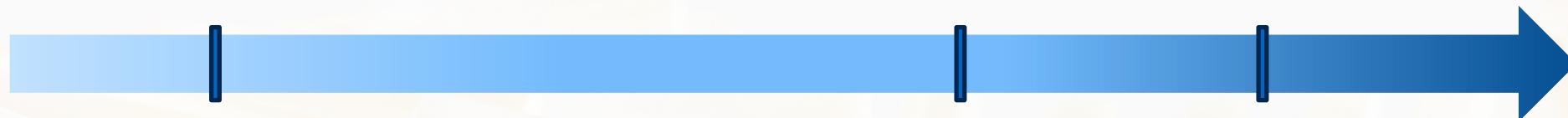
## - 框架挑战

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

## 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://deeprrec.readthedocs.io/en/latest/>

# Contents 目录

- 01 DeepRec 背景及开源历程
- 02 DeepRec 概述及关键技术
- 03 DeepRec 展望

# DeepRec 功能 概览

## Embedding & Optimizer

- Embedding Variable
- Feature Eviction and Filter
- Dynamic dimension EV
- Adaptive Embedding Variable
- Multi-Hash Embedding
- AdamAsync Optimizer
- AdagraDecay Optimizer

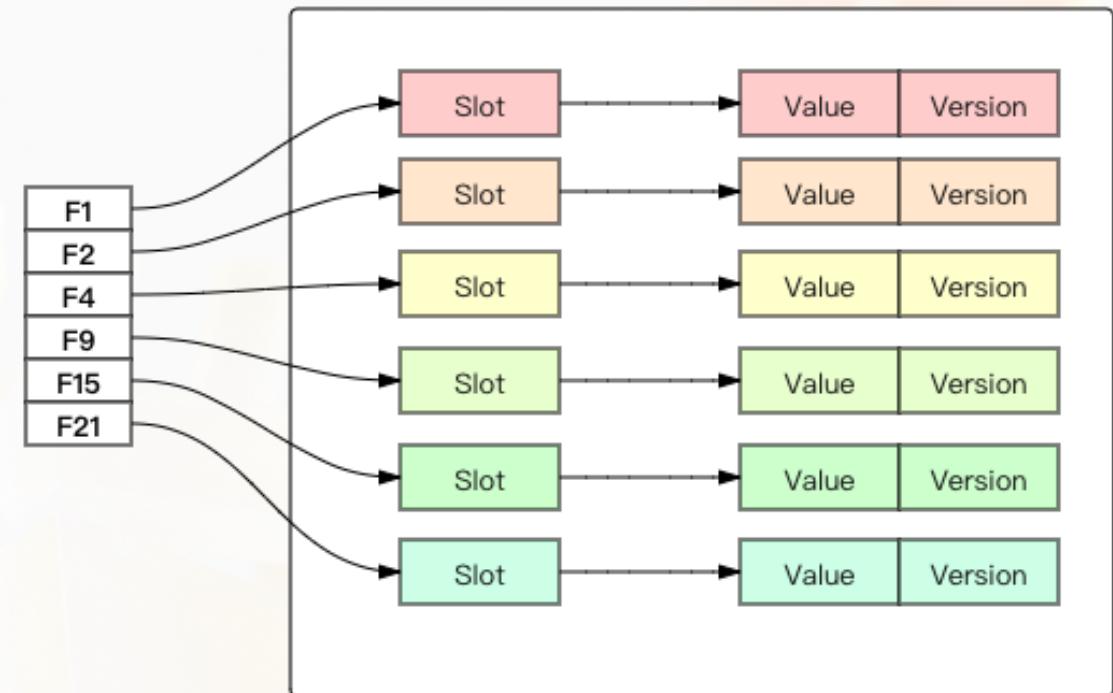
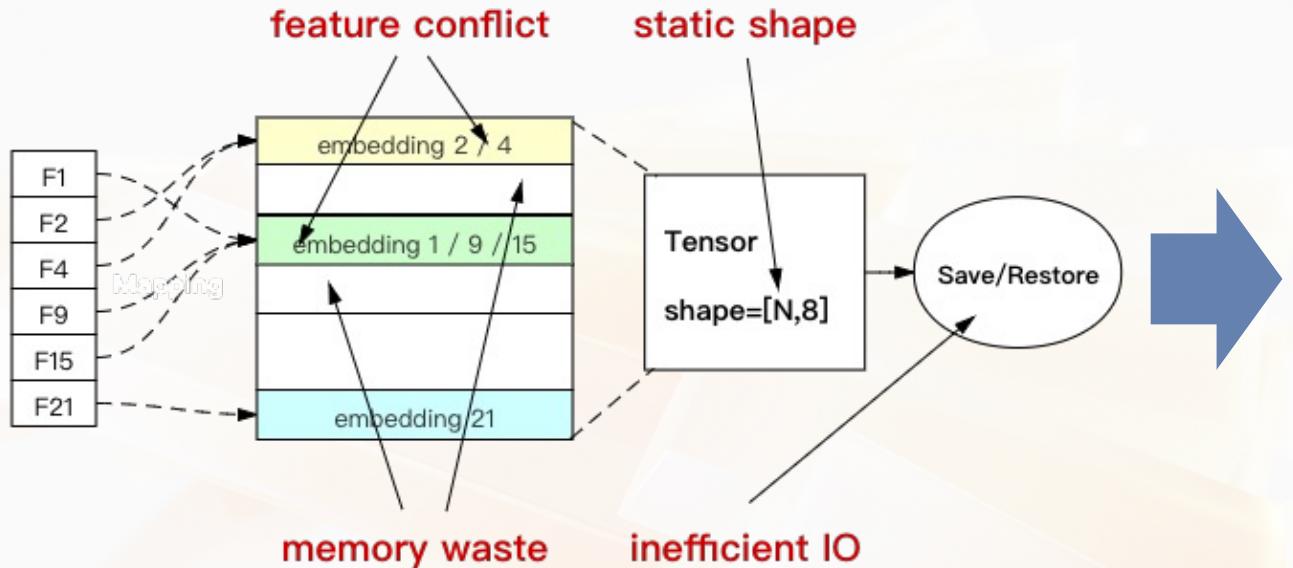
## Training

- Async/Sync Distributed Training
- Distributed Training based on GBA
- Graph Aware Memory Allocator
- Automatic Pipeline
- Graph Template Engine
- Critical-path based Executor
- GPU Multi-Stream Engine
- Multi-tier Embedding

## Serving & Deployment

- Share-nothing SessionGroup
- Multi-tier Embedding
- GPU Multi-stream In SessionGroup
- Dynamic Shape Compiler (BladeDISC)
- CUDA Graph Execution Engine
- Delta checkpoint
- Online Deep Learning
- Model Quantization

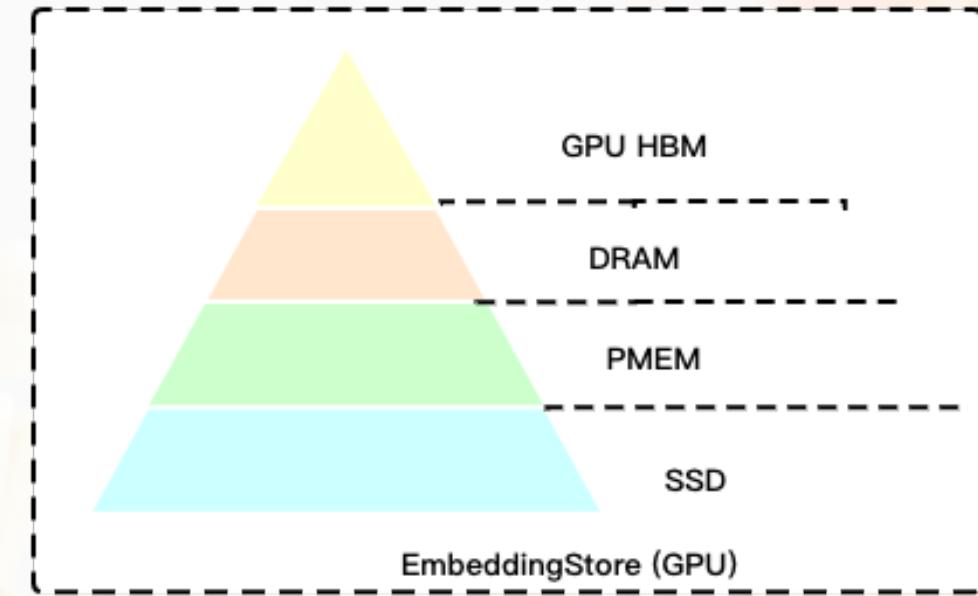
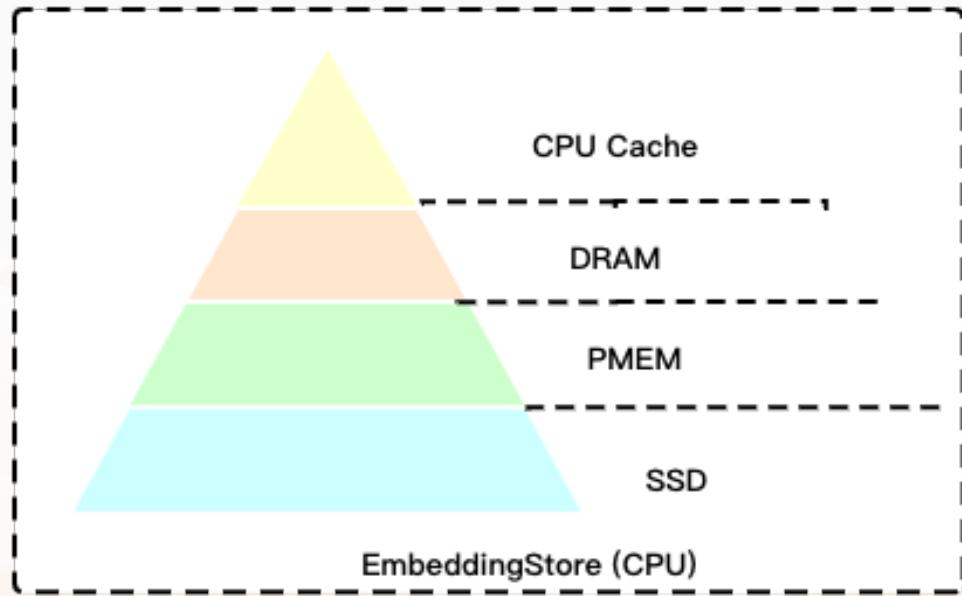
# Embedding Variable 功能



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

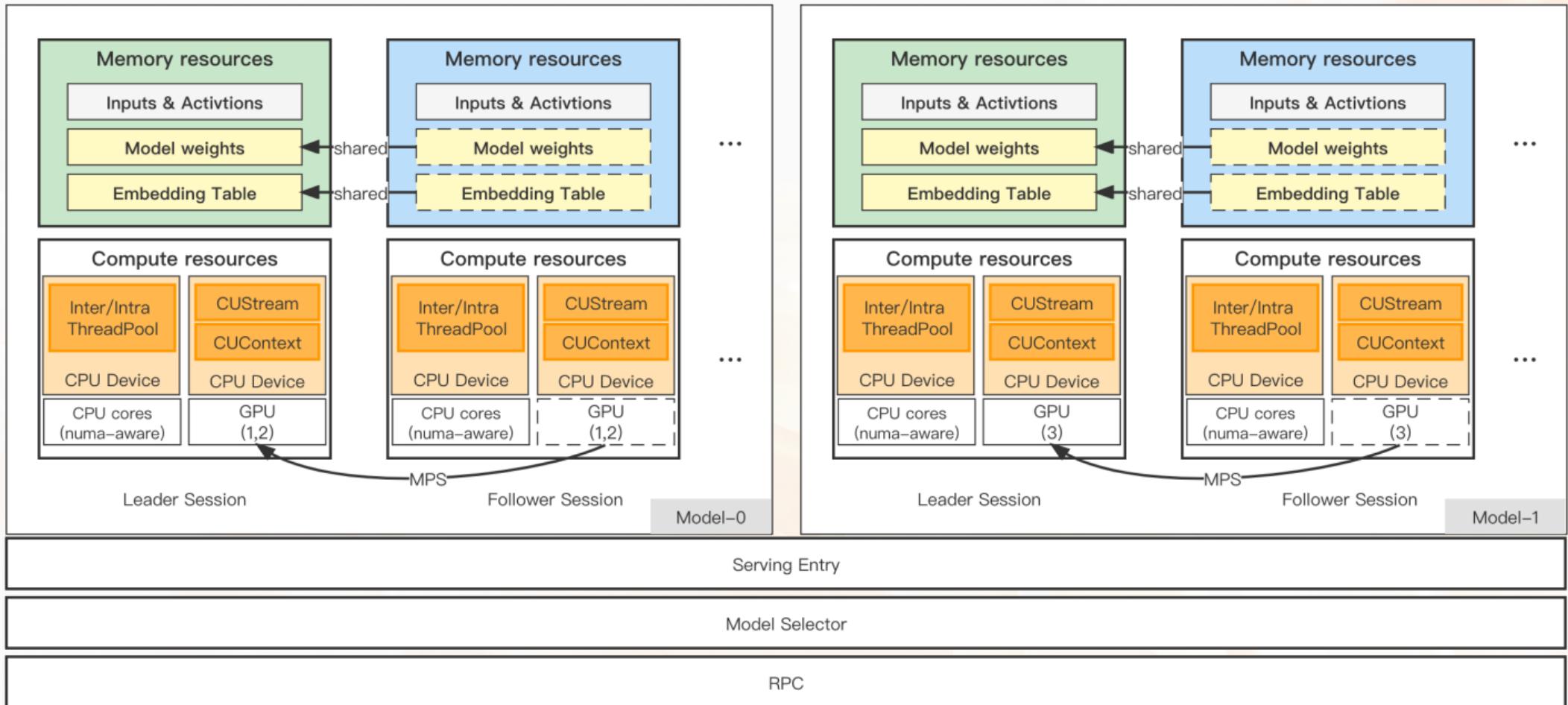
Feature Eviction:  
 1. Based on Global Step  
 2. Based on l2weights

## Multi-Tired Embedding Variable

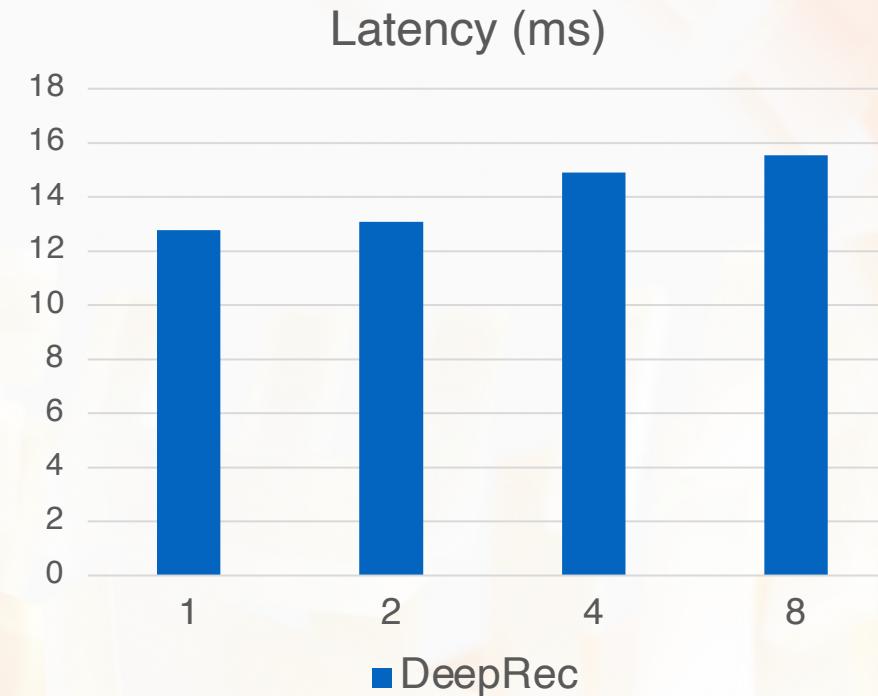
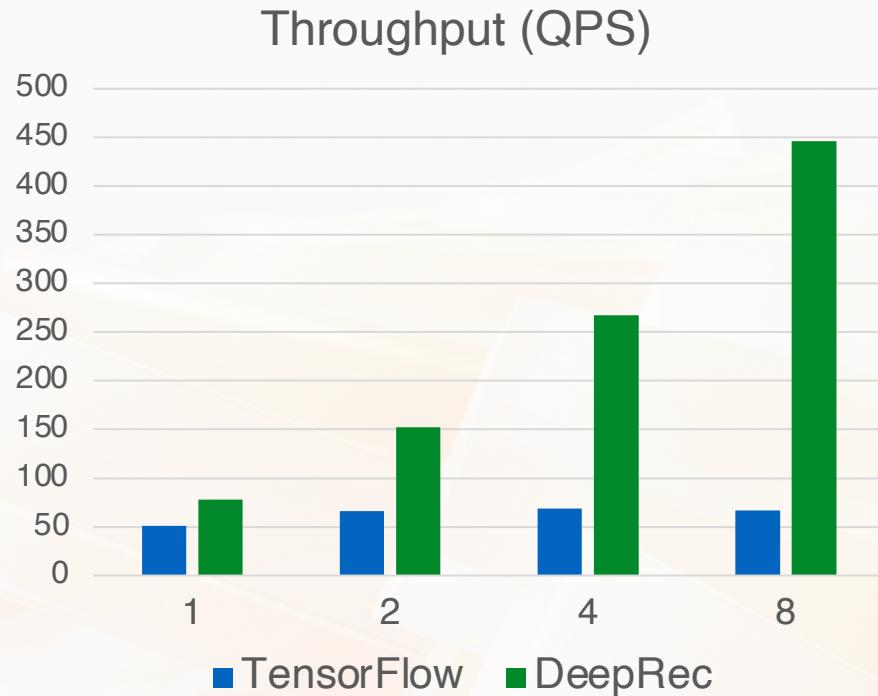


- (1-10ns) -> 100ns -> 1us ->100ns
- Support Model with 10TB+ (Training and Inference)
- Less Memory or GPU Memory Usage
- Hot/Cold Features for Embedding (Pareto Principle, hot features 20%)
- Compare with Distributed Serving, (latency TP99 80ms->25ms)

# Share-nothing SessionGroup

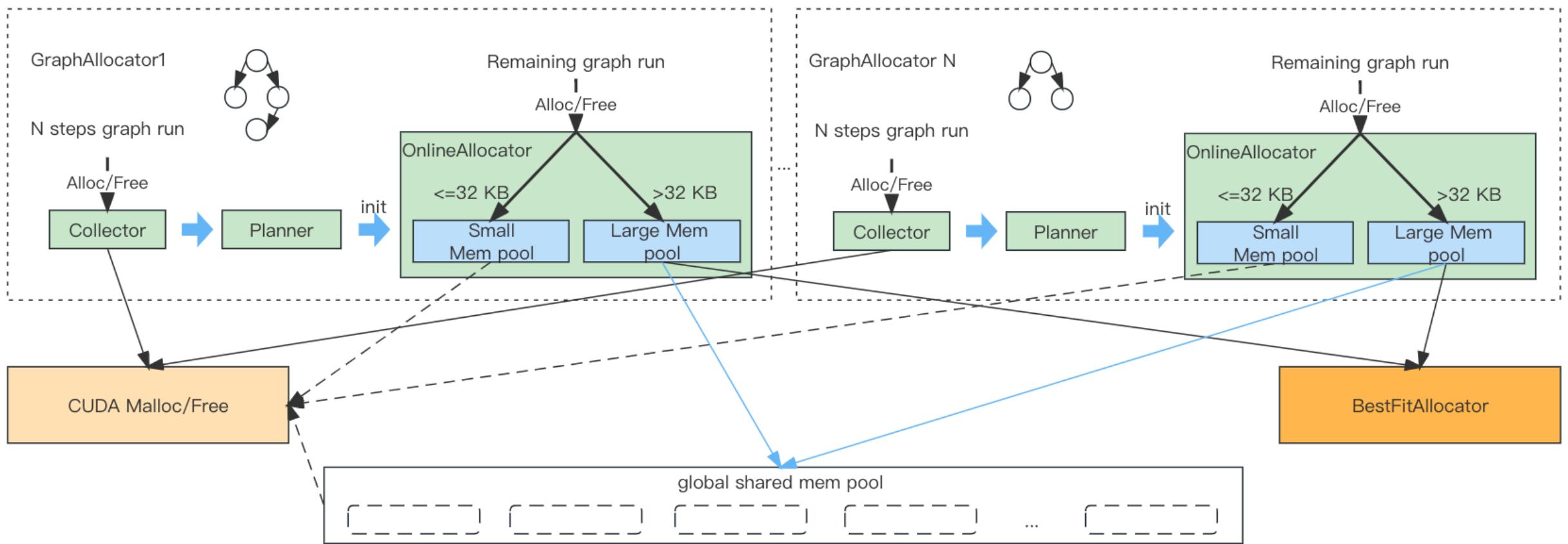


## Share-nothing SessionGroup

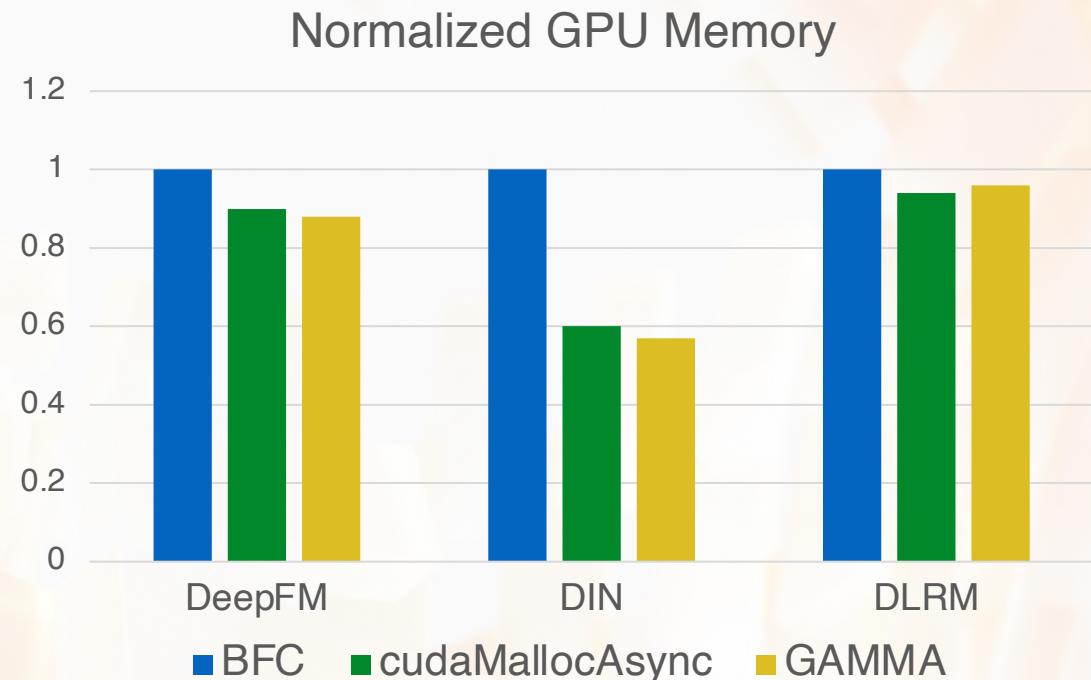
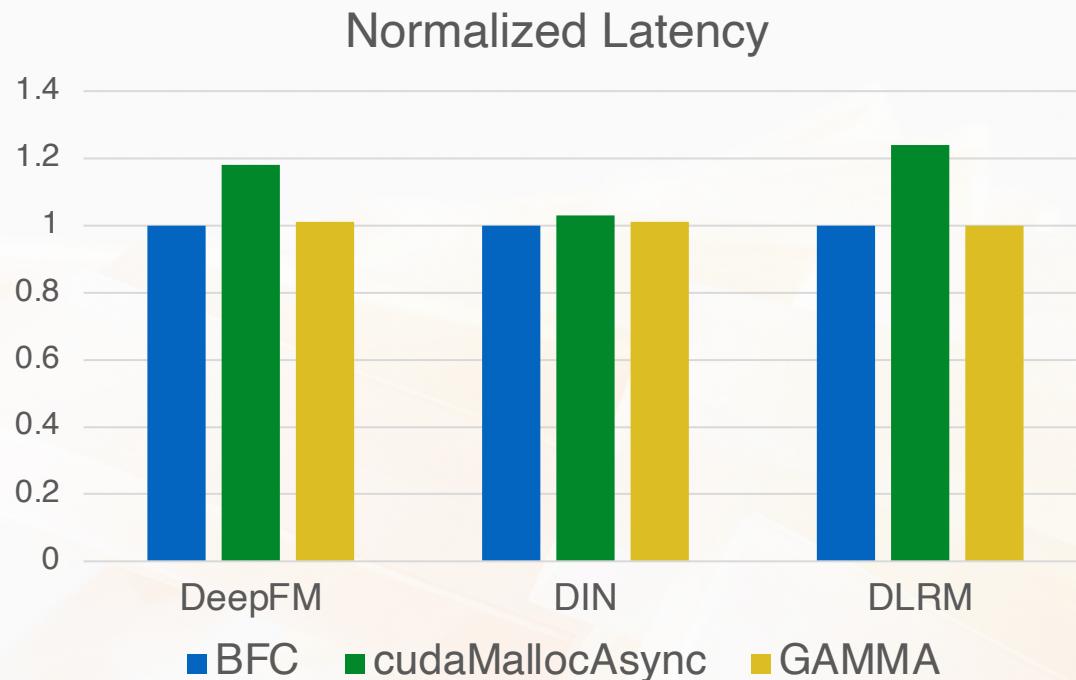


- 8 CUDA stream and share-nothing architecture could brings **6X** QPS than TensorFlow

# GAMMA - Graph Aware Allocator

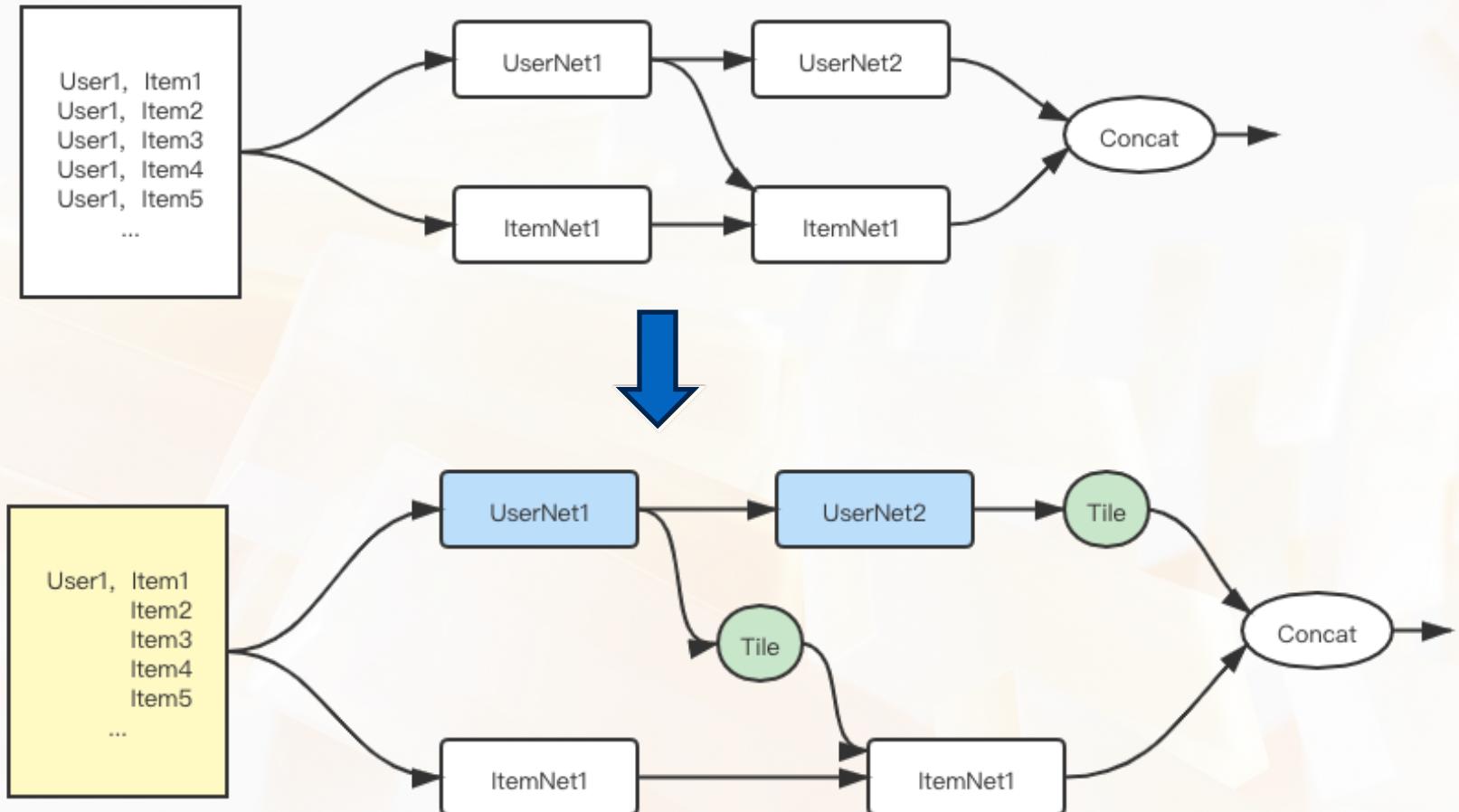


## GAMMA - Graph Aware Allocator

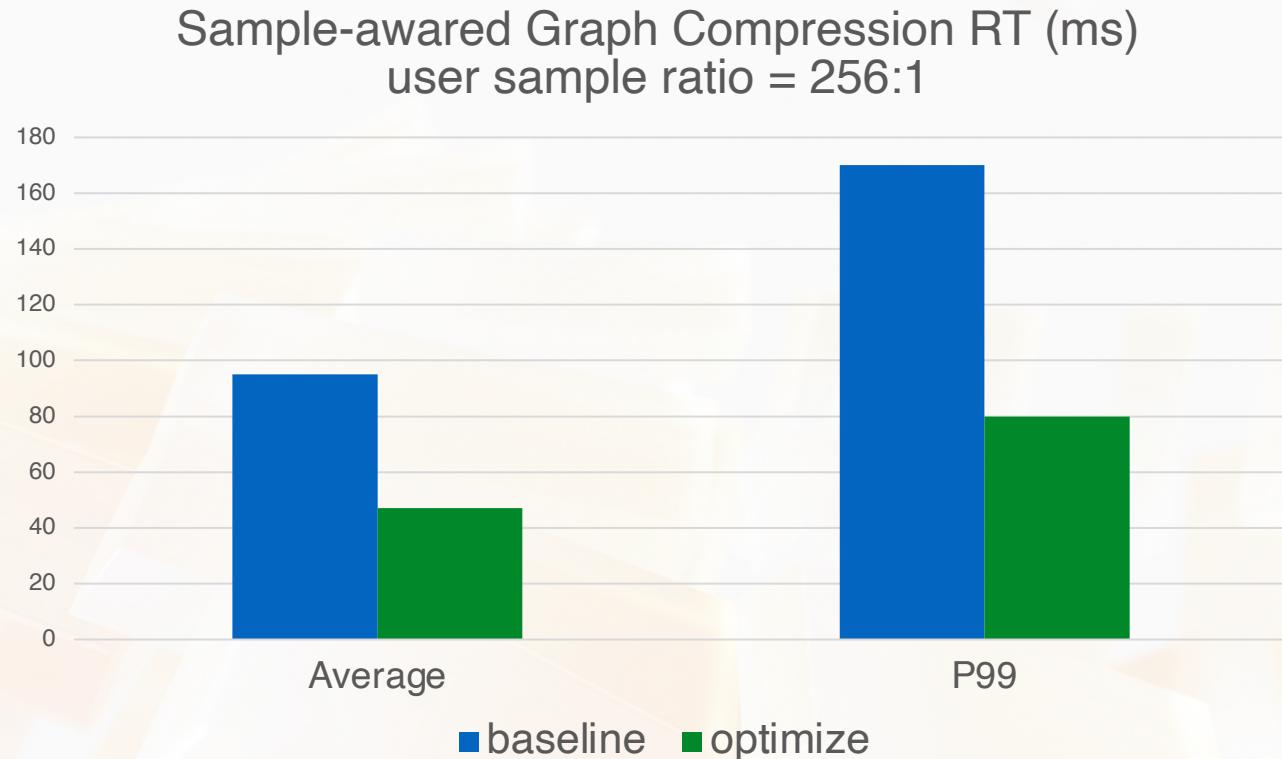


- GAMMA save **12%-43%** GPU memory compare to BFC
- GAMMA improve performance at most **2%-24%** compare to cudaMallocAsync

## Sample-aware Graph Compression



## Sample-aware Graph Compression



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

# Contents 目录

- 01 DeepRec 背景及开源历程
- 02 DeepRec 概述及关键技术
- 03 DeepRec 展望

# 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



# THANKS