
GOTC 2023
全球开源技术峰会
THE GLOBAL OPENSOURCE TECHNOLOGY CONFERENCE
OPEN SOURCE, INTO THE FUTURE
「AI is Everywhere」专场
本期议题: DeepRec: 面向推荐场景的高性能深度学习框架
丁 辰 2022年5月28日

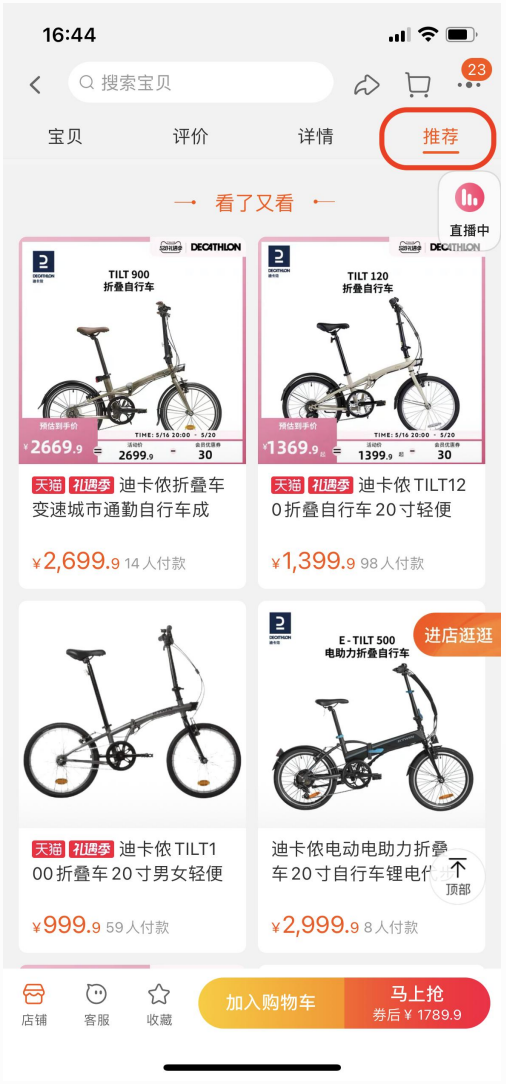
大纲

DeepRec GOTC

DeepRec 背景及开源历程

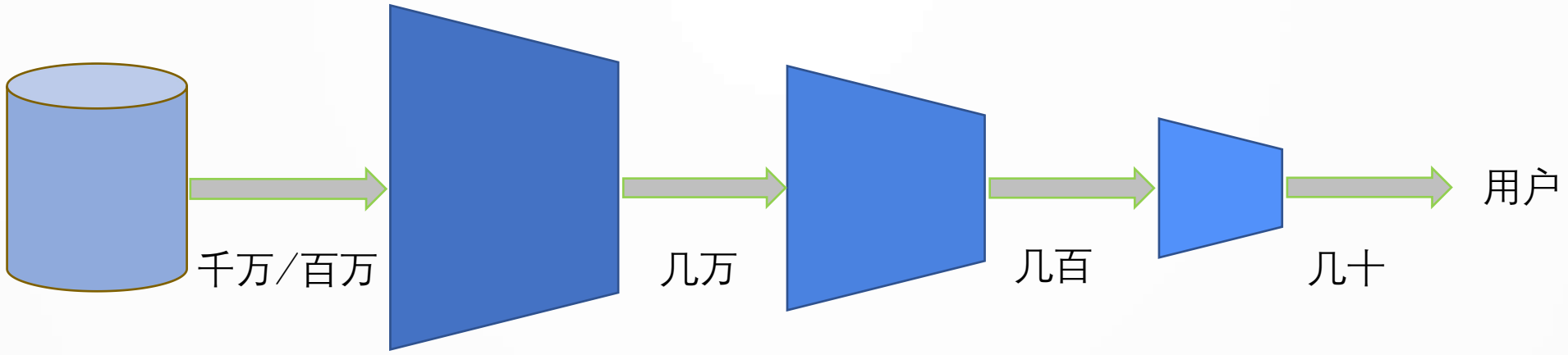
业务场景

- 搜索
- 推荐
- 广告



DeepRec 背景及开源历程

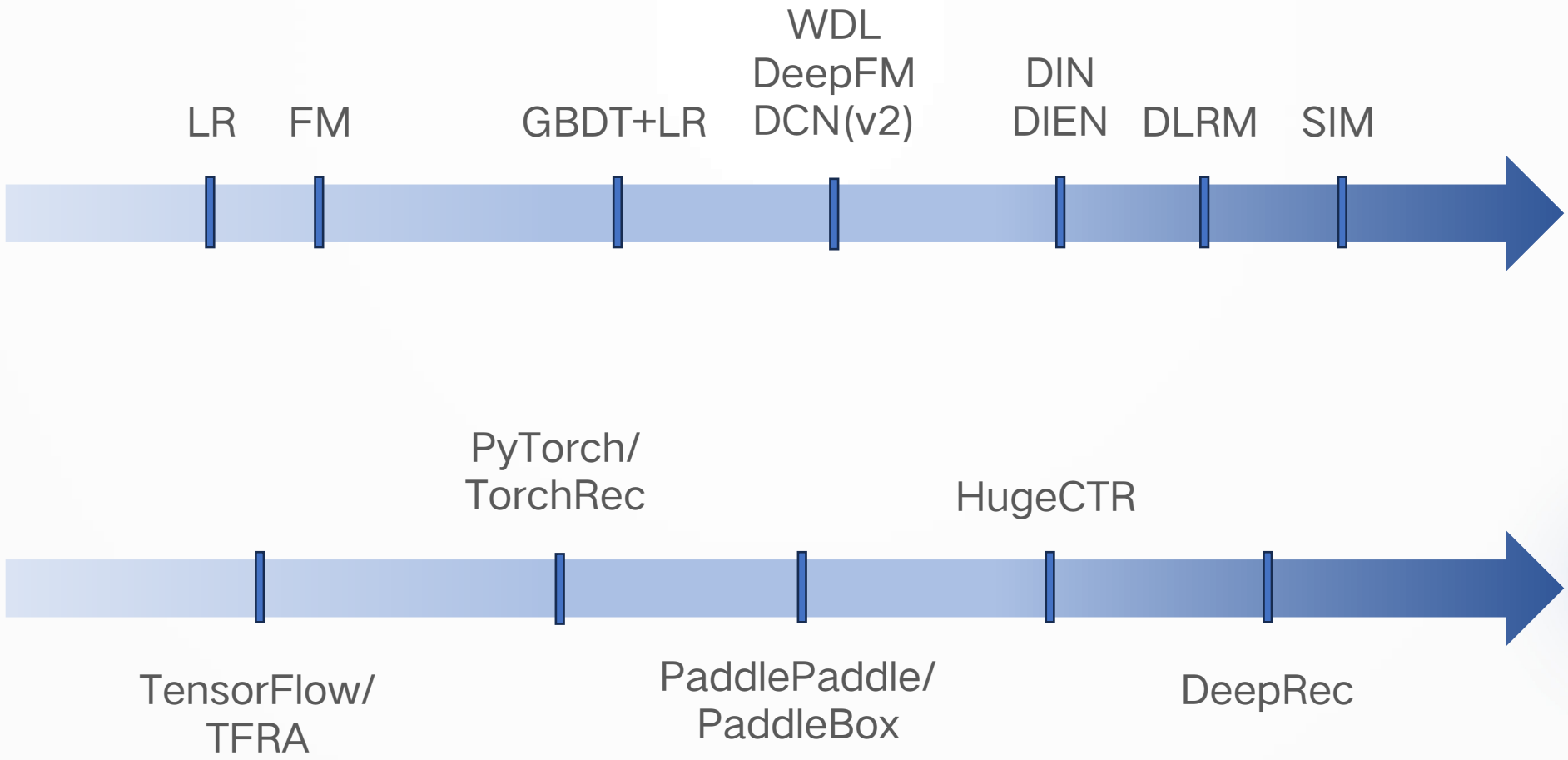
一个典型的系统架构



语料库　　召回　　排序　　重排

DeepRec 背景及开源历程

稀疏模型和计算框架的发展



DeepRec 背景及开源历程

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

DeepRec 背景及开源历程

开源历程

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

2022 DeepRec 开源

- 数十家公司在使用，包括微博，得物，喜马拉雅，VIVO等

2023 DeepRec 捐献

LF AI & Data Foundation

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

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

大纲

DeepRec GOTC

01 DeepRec 背景及开源历程

02 DeepRec 概述及关键技术

03 DeepRec 展望

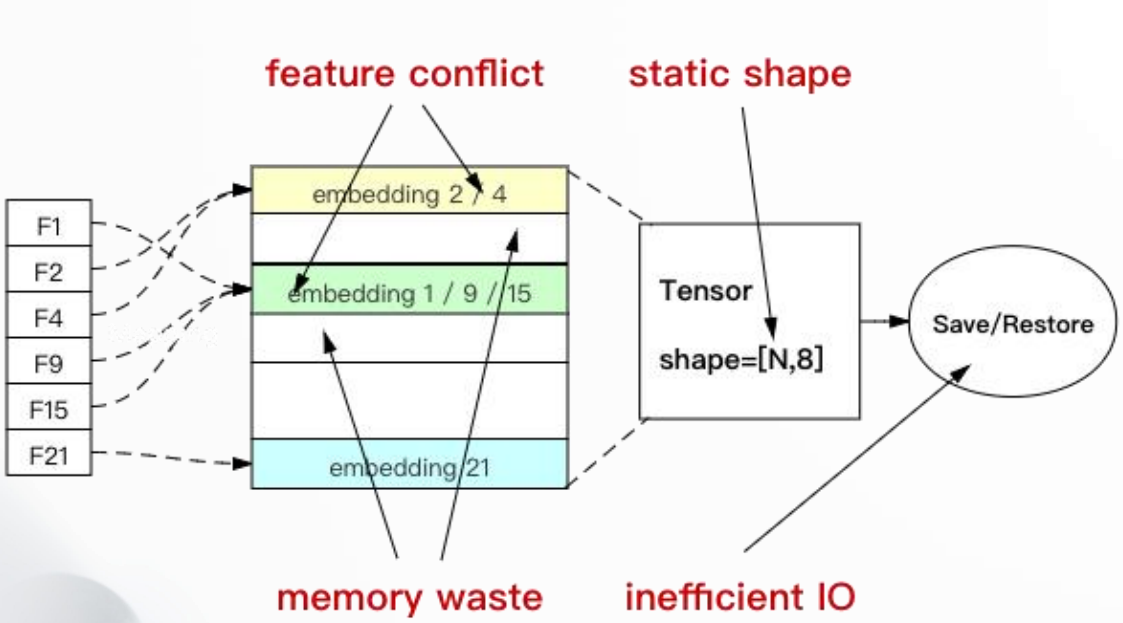
DeepRec 概述及关键技术

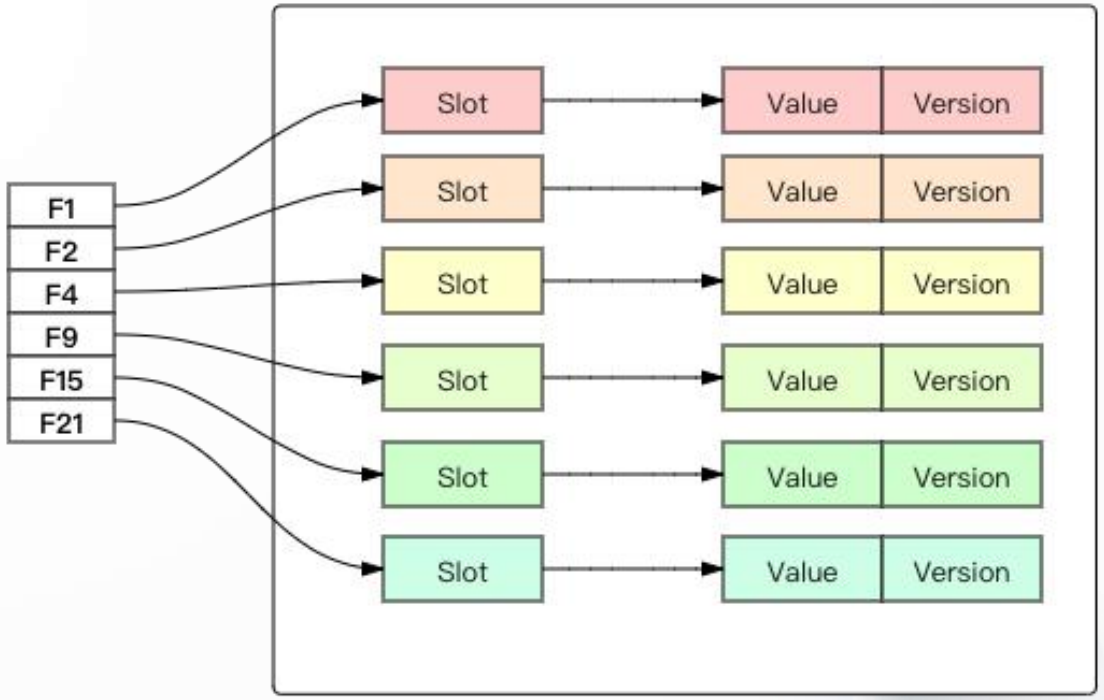
DeepRec 功能 概览

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

DeepRec 概述及关键技术

Embedding Variable 功能



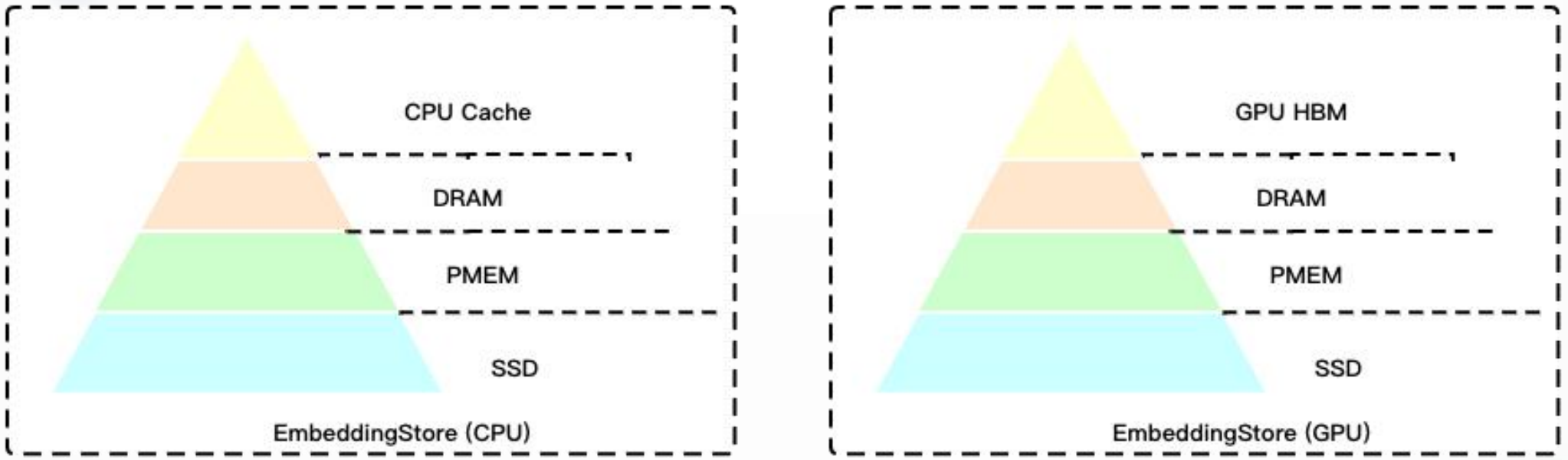


Feature Filter
1. Based on Counter
2. Based on Bloom Filter

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

DeepRec 概述及关键技术

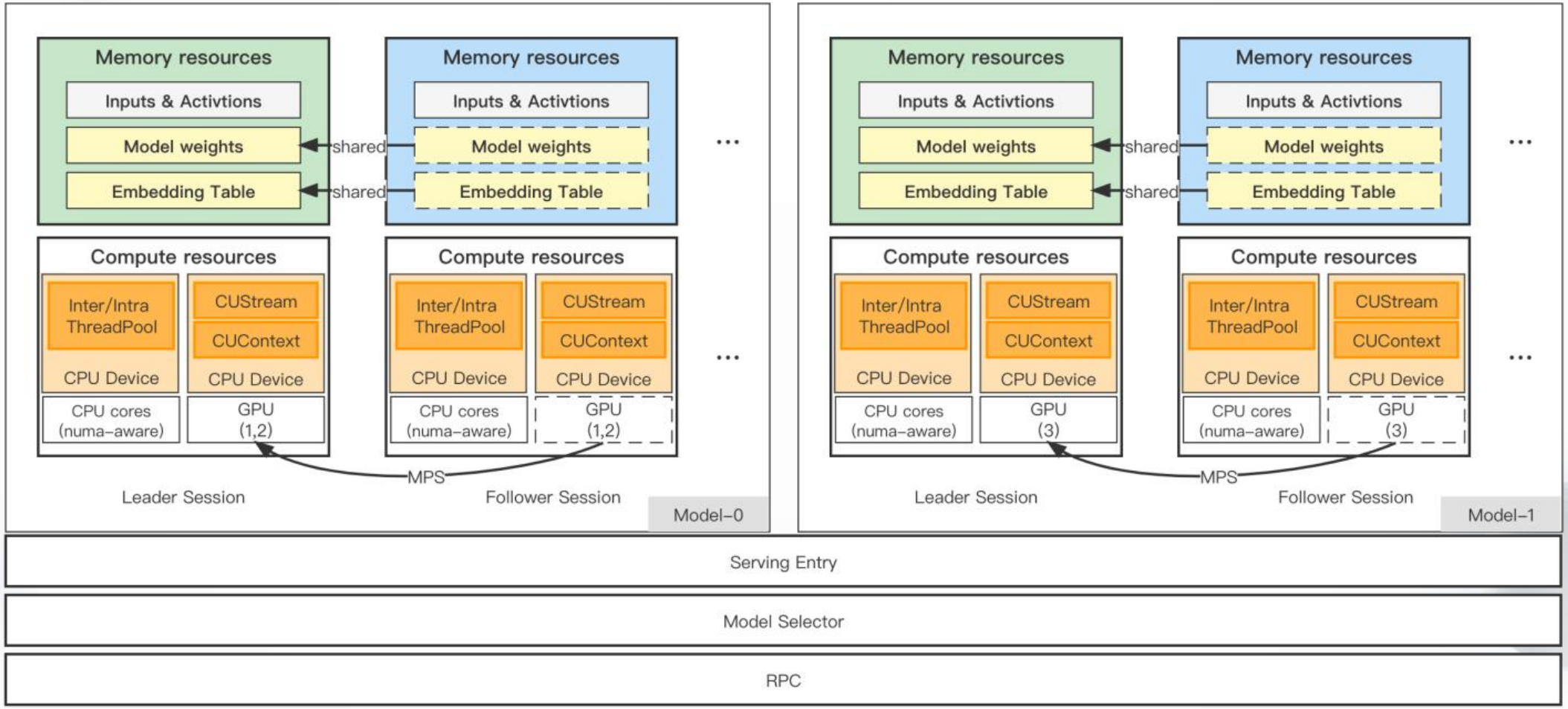
Multi-Tired Embedding Variable



- 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)

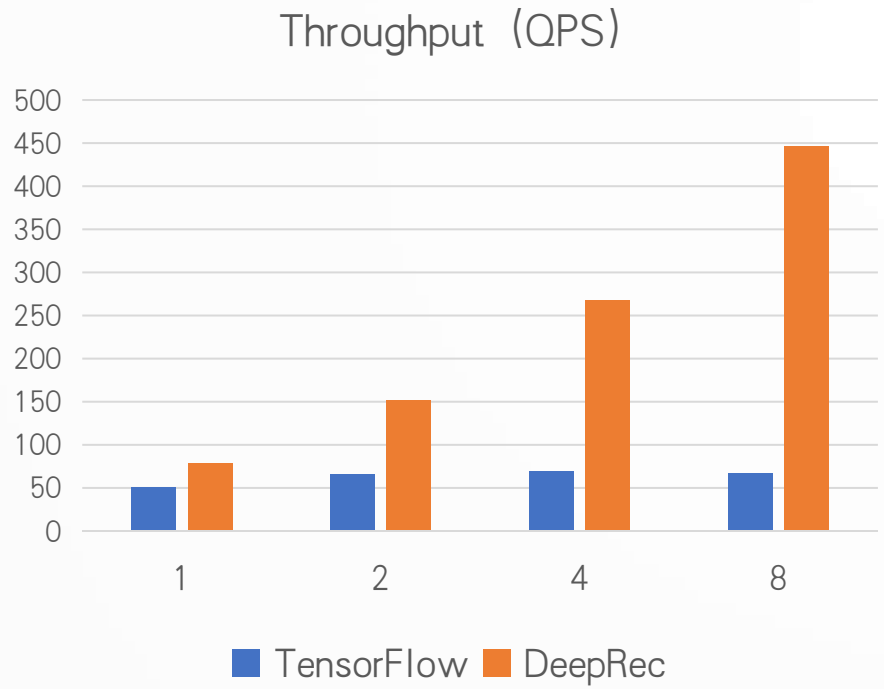
DeepRec 概述及关键技术

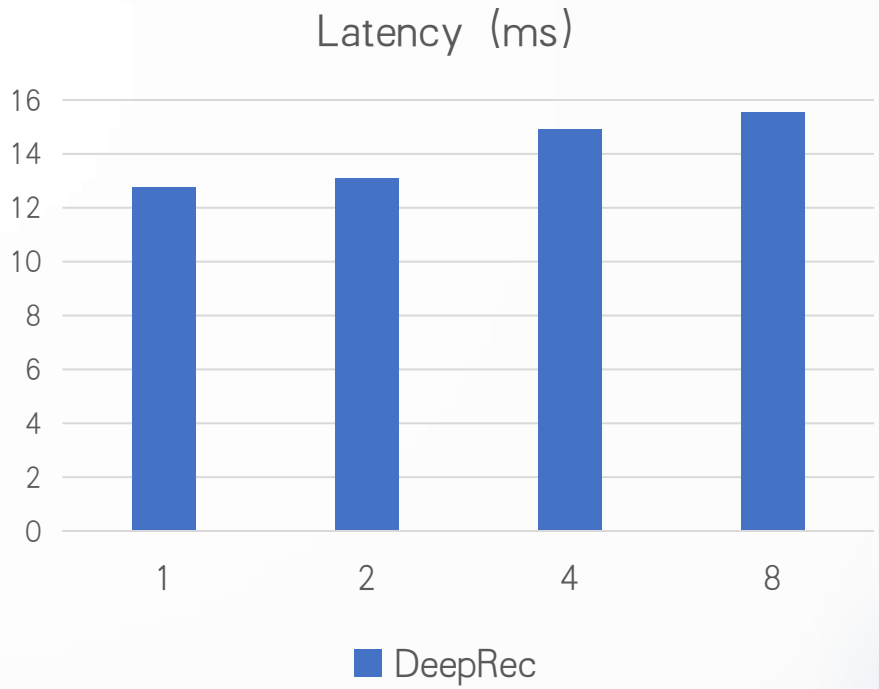
Share-nothing SessionGroup



DeepRec 概述及关键技术

Share-nothing SessionGroup

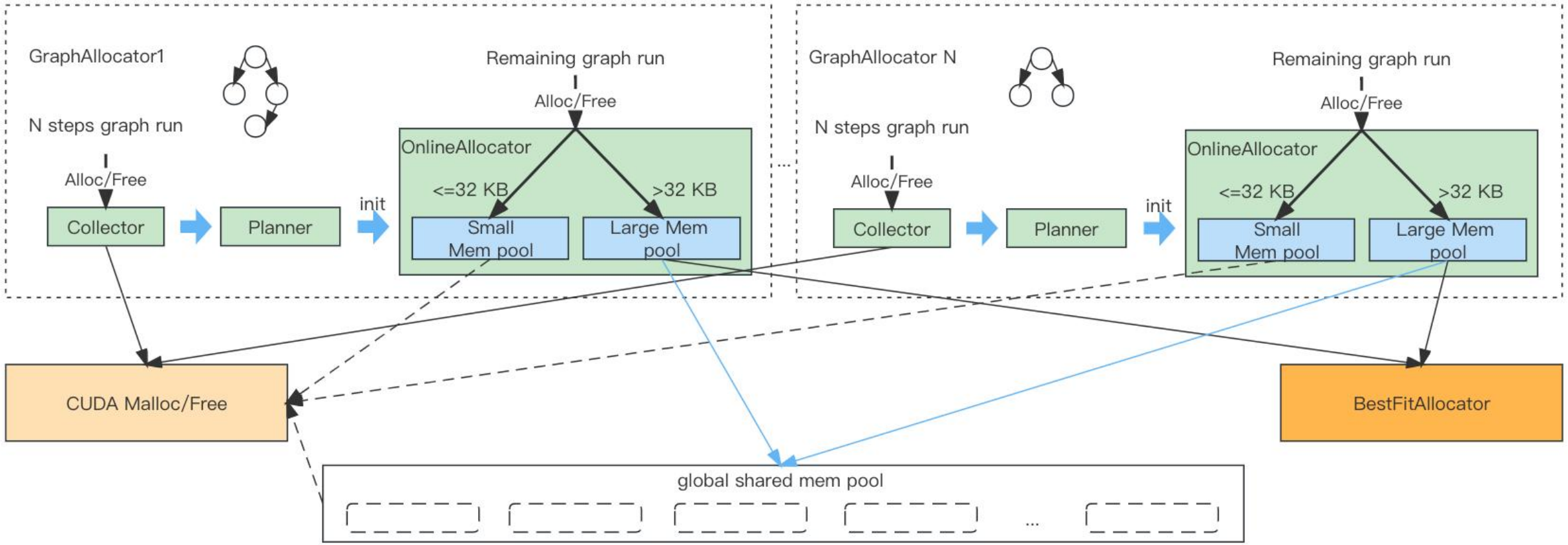




- 8 CUDA stream and share-nothing architecture could brings 6X QPS than TensorFlow

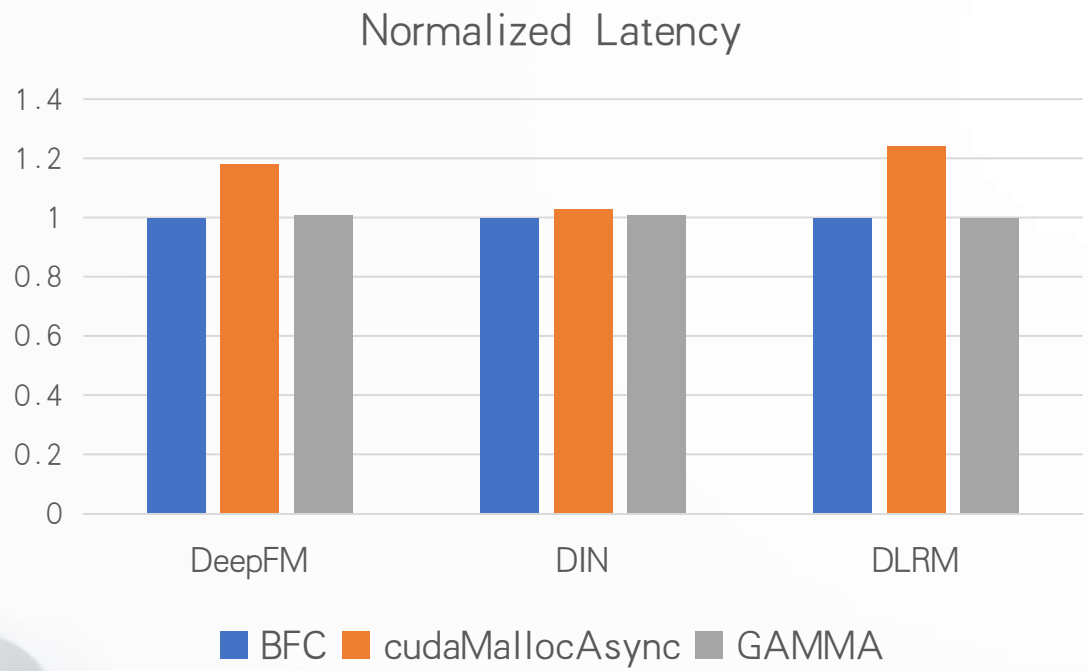
DeepRec 概述及关键技术

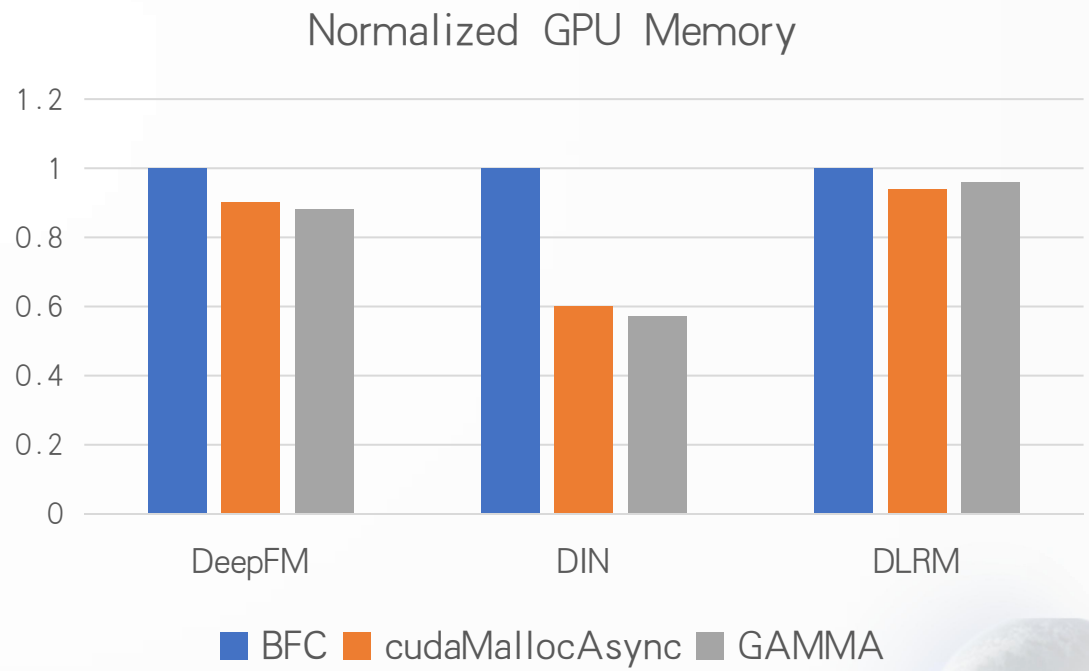
GAMMA - Graph Aware Allocator



DeepRec 概述及关键技术

GAMMA - Graph Aware Allocator

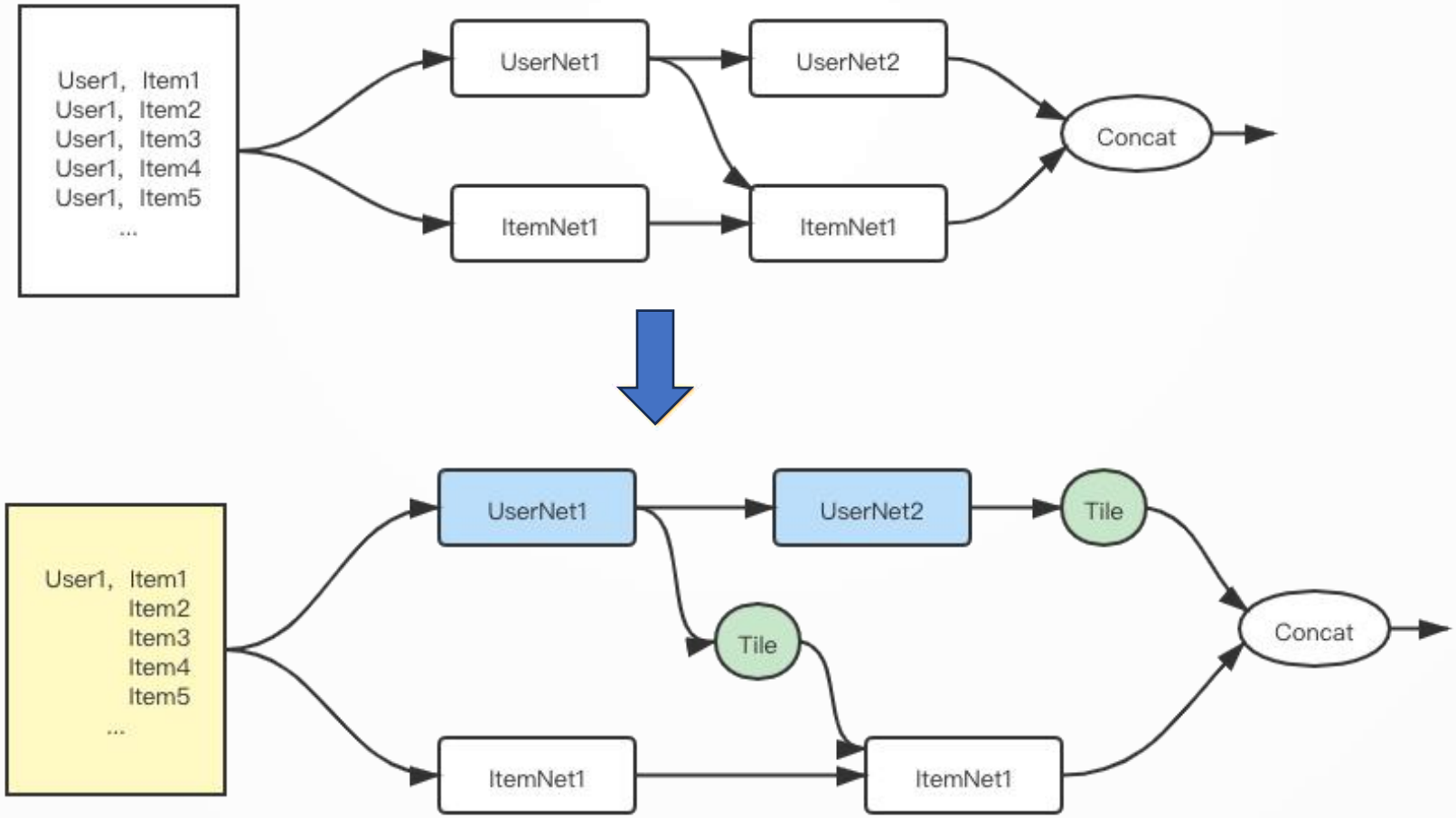




- GAMMA save 12%-43% GPU memory compare to BFC
- GAMMA improve performance at most 2%-24% compare to cudaMallocAsync

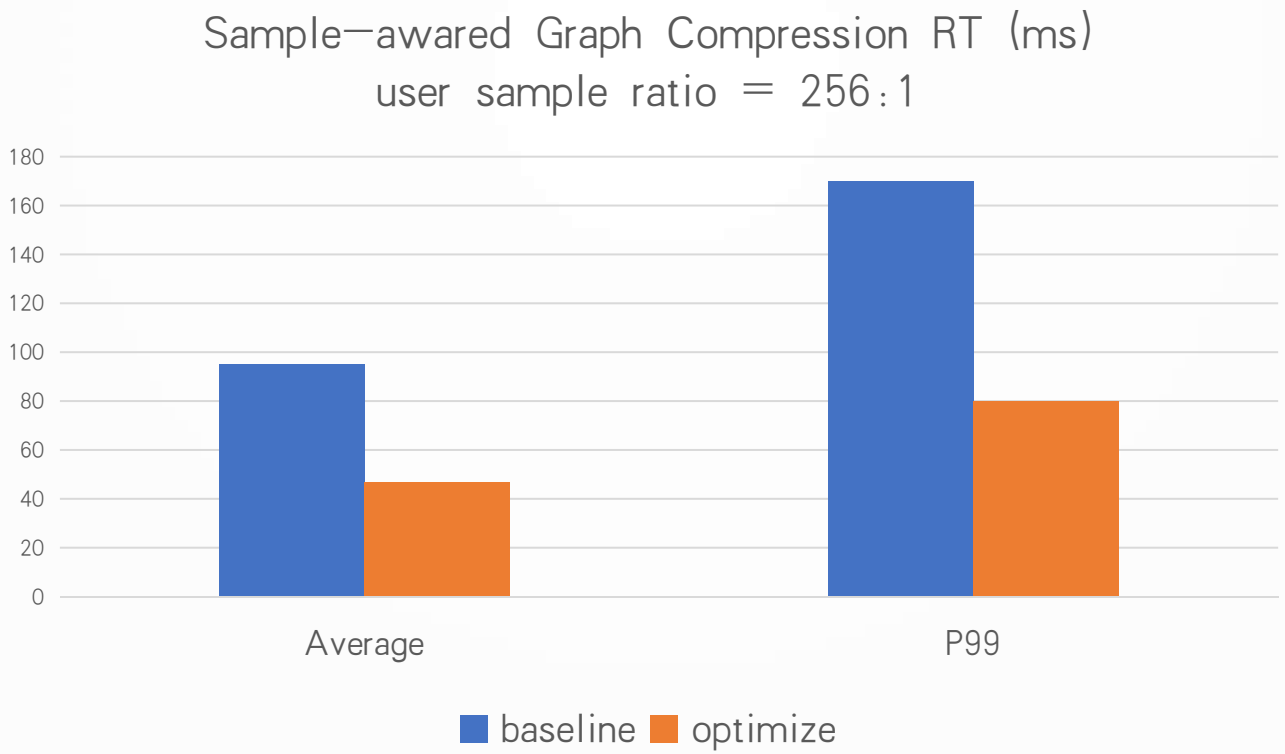
DeepRec 概述及关键技术

Sample-awared Graph Compression



DeepRec 概述及关键技术

Sample-awared Graph Compression



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

大纲

DeepRec GOTC

01 DeepRec 背景及开源历程

02 DeepRec 概述及关键技术

03 DeepRec 展望

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

GOTC

THANKS

全球开源技术峰会

THE GLOBAL OPENSOURCE TECHNOLOGY CONFERENCE