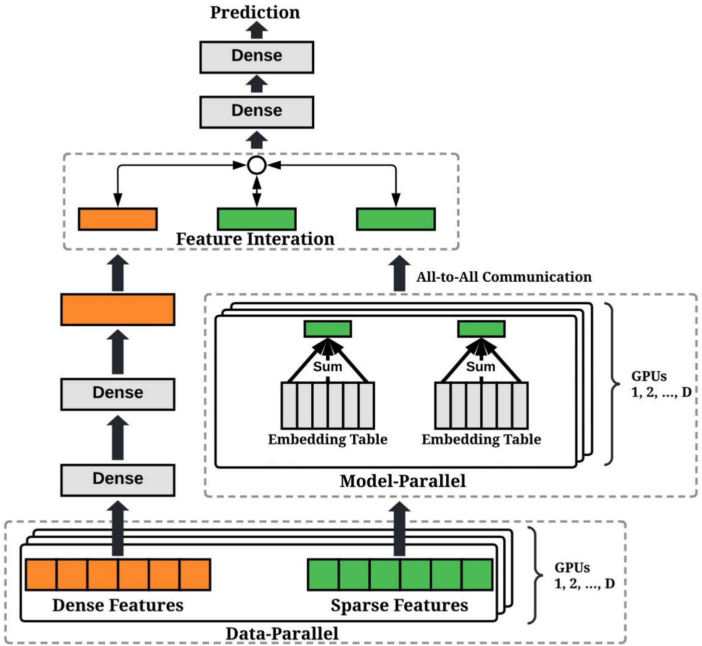
DreamShard: Generalizable Embedding Table Placement for Recommender Systems

Daochen Zha, Louis Feng, Qiaoyu Tan, Zirui Liu, Kwei-Herng Lai,
Bhargav Bhushanam, Yuandong Tian, Arun Kejariwal, Xia Hu

Rice University
Meta Platforms, Inc.
Texas A&M University

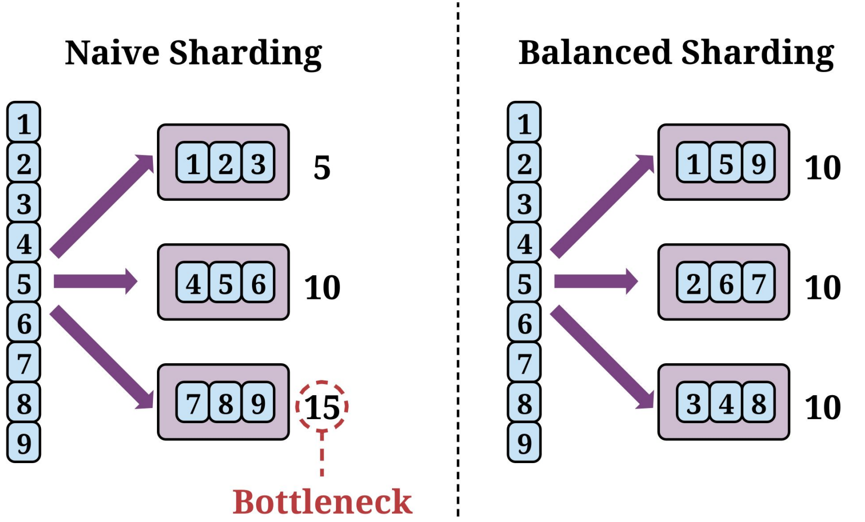
Distributed Recommender System

Combining data-parallelism and model-parallelism.

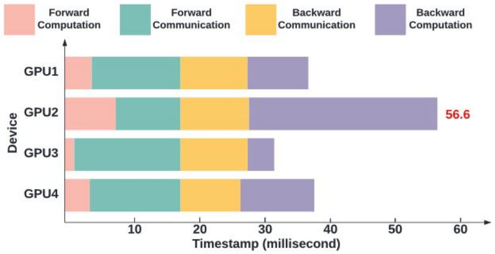


Embedding Table Placement Problem

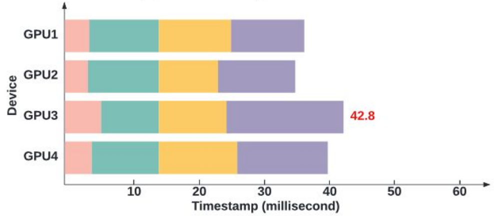
- **Problem Setting**
  - We consider embedding table placement on GPU devices.
  - Embedding accounts for 48% and 65% of the computation and communication costs in production model.



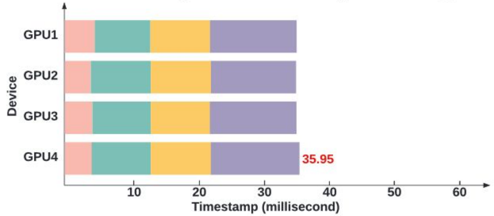
Embedding Table Placement Problem



(a) Random placement



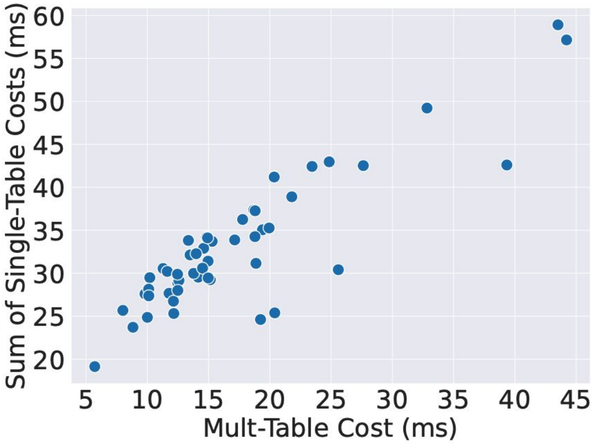
(b) The existing best human expert strategy



(c) DreamShard

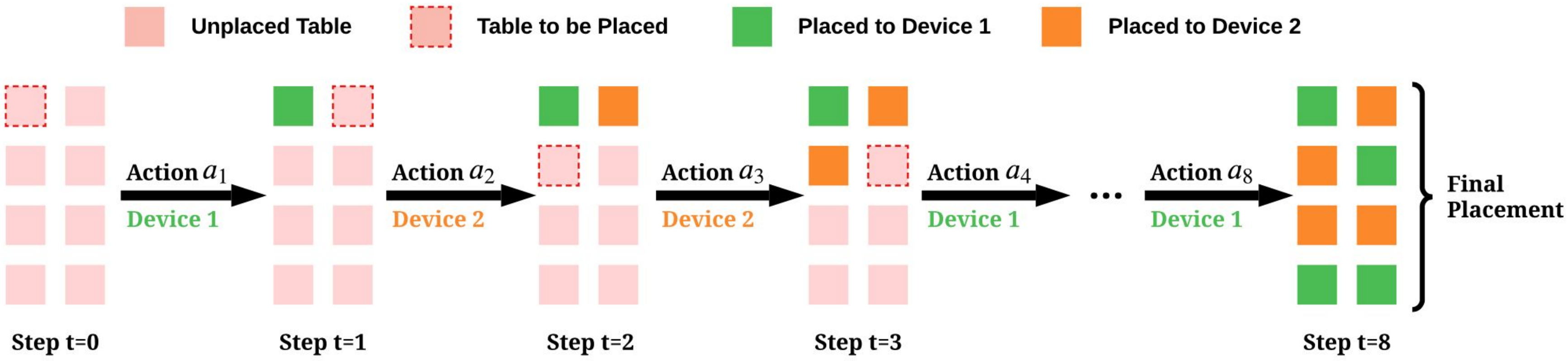
Key Challenges

- **Challenges**
  - Operation fusion, which uses a single operation to subsume multiple tables, makes it hard estimate cost.
  - The adopted embedding tables and the available devices can change frequently (e.g., machine learning engineers may conduct experiments with various table combinations and numbers of devices).

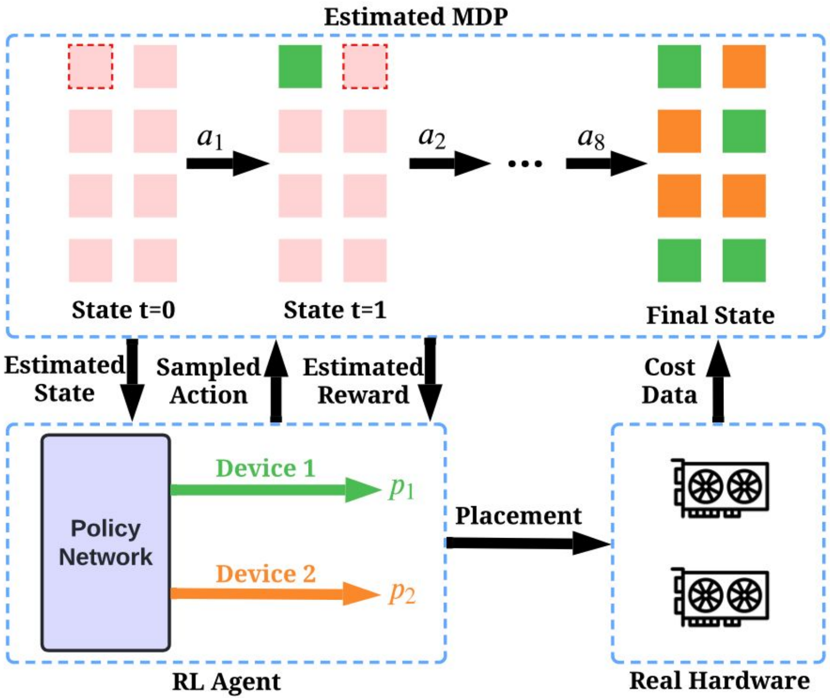


Formulation of MDP

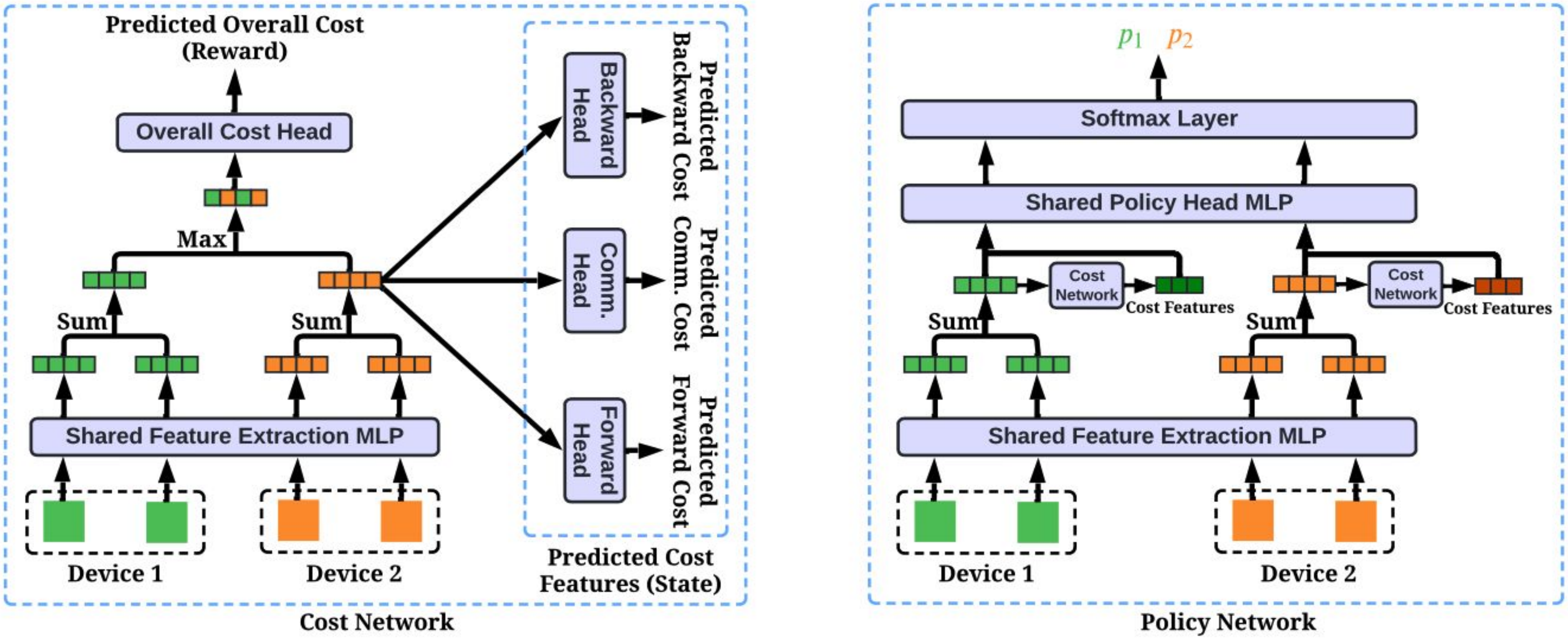
- **Markov Decision Process**



DreamShard Framework



DreamShard Framework



Main Results

- **Observations**
  - DreamShard outperforms baselines significantly.
  - DreamShard can generalize well (test performance is similar to train performance).

Task		No strategy	Human Experts				RL	
		Random	Size-based	Dim-based	Lookup-based	Size-lookup-based	RNN-based	DreamShard
DLRM-20 (4)	Train	24.0±0.6	22.7±0.0 (+5.7%)	21.3±0.0 (+12.7%)	19.1±0.0 (+25.7%)	19.1±0.0 (+25.7%)	22.4±0.5 (+7.1%)	**18.6±0.2 (+29.0%)**
	Test	23.0±0.5	21.7±0.0 (+6.0%)	19.9±0.0 (+15.6%)	18.3±0.0 (+25.7%)	18.4±0.0 (+25.0%)	20.9±0.3 (+10.0%)	**17.6±0.2 (+30.7%)**
DLRM-40 (4)	Train	41.3±0.2	39.6±0.0 (+4.3%)	37.4±0.1 (+10.4%)	33.6±0.0 (+22.9%)	33.6±0.1 (+22.9%)	39.2±0.7 (+5.4%)	**32.8±0.3 (+25.9%)**
	Test	41.1±0.5	40.3±0.0 (+2.0%)	37.3±0.0 (+10.2%)	33.0±0.1 (+24.5%)	33.2±0.0 (+23.8%)	39.2±1.1 (+4.8%)	**32.4±0.3 (+26.9%)**
DLRM-60 (4)	Train	57.7±0.8	56.6±0.1 (+1.9%)	52.9±0.0 (+9.1%)	49.2±0.1 (+17.3%)	49.3±0.0 (+17.0%)	55.5±0.9 (+4.0%)	**47.6±0.4 (+21.2%)**
	Test	58.1±0.6	59.6±0.1 (-2.5%)	53.7±0.0 (+8.2%)	48.7±0.2 (+19.3%)	49.1±0.1 (+18.3%)	56.0±0.7 (+3.8%)	**47.9±0.7 (+21.3%)**
DLRM-80 (4)	Train	75.7±1.0	76.0±0.0 (-0.4%)	70.0±0.3 (+8.1%)	64.8±0.0 (+16.8%)	65.3±0.1 (+15.9%)	73.2±2.7 (+3.4%)	**62.2±0.2 (+21.7%)**
	Test	74.5±0.8	77.7±0.2 (-4.1%)	69.9±0.4 (+6.6%)	64.1±0.2 (+16.2%)	65.1±0.0 (+14.4%)	72.9±2.4 (+2.2%)	**62.7±0.3 (+18.8%)**
DLRM-100 (4)	Train	91.8±1.7	94.1±0.3 (-2.4%)	86.7±0.3 (+5.9%)	81.2±0.4 (+13.1%)	82.2±0.2 (+11.7%)	94.5±10.7 (-2.9%)	**78.4±0.6 (+17.1%)**
	Test	94.5±6.5	95.4±0.0 (-0.9%)	84.7±0.4 (+11.6%)	79.5±0.3 (+18.9%)	80.8±0.3 (+17.0%)	94.8±13.0 (-0.3%)	**77.8±0.8 (+21.5%)**
DLRM-40 (8)	Train	15.6±0.4	14.1±0.0 (+10.6%)	13.4±0.1 (+16.4%)	**9.8±0.0 (+59.2%)**	9.9±0.0 (+57.6%)	16.2±0.8 (-3.7%)	**9.8±0.6 (+59.2%)**
	Test	15.2±0.2	14.5±0.0 (+4.8%)	13.2±0.0 (+15.2%)	9.5±0.0 (+60.0%)	9.5±0.0 (+60.0%)	16.0±1.1 (-5.0%)	**9.4±0.5 (+61.7%)**
DLRM-80 (8)	Train	25.0±0.2	24.0±0.0 (+4.2%)	21.7±0.0 (+15.2%)	17.1±0.0 (+46.2%)	17.5±0.0 (+42.9%)	51.4±3.9 (-51.4%)	**16.1±0.3 (+55.3%)**
	Test	25.2±1.3	25.6±0.5 (-1.6%)	20.8±0.0 (+21.2%)	16.7±0.2 (+50.9%)	16.9±0.1 (+49.1%)	53.4±4.6 (-52.8%)	**16.1±0.4 (+56.5%)**
DLRM-120 (8)	Train	34.0±0.3	32.3±0.0 (+5.3%)	29.8±0.0 (+14.1%)	24.5±0.0 (+38.8%)	25.3±0.0 (+34.4%)	58.6±2.7 (-42.0%)	**23.3±0.2 (+45.9%)**
	Test	33.5±0.5	35.0±0.0 (-4.3%)	29.2±0.0 (+14.7%)	23.7±0.0 (+41.4%)	24.5±0.0 (+36.7%)	58.7±3.1 (-42.9%)	**22.8±0.2 (+46.9%)**
DLRM-160 (8)	Train	42.8±0.3	41.6±0.0 (+2.9%)	39.0±0.0 (+9.7%)	32.0±0.0 (+33.7%)	32.7±0.0 (+30.9%)	58.3±3.5 (-26.6%)	**30.3±0.2 (+41.3%)**
	Test	41.1±0.0	42.4±0.0 (-3.1%)	36.4±0.0 (+12.9%)	30.8±0.0 (+33.4%)	31.6±0.0 (+30.1%)	59.3±5.4 (-30.7%)	**29.6±0.2 (+38.9%)**
DLRM-200 (8)	Train	51.5±1.2	48.2±0.0 (+6.8%)	48.0±0.0 (+7.3%)	38.9±0.0 (+32.4%)	39.9±0.0 (+29.1%)	68.7±2.4 (-25.0%)	**37.2±0.2 (+38.4%)**
	Test	50.7±0.2	50.8±0.0 (-0.2%)	44.8±0.0 (+13.2%)	38.0±0.0 (+33.4%)	38.6±0.0 (+31.3%)	70.4±2.8 (-28.0%)	**36.4±0.3 (+39.3%)**
Prod-20 (2)	Train	41.3±0.7	43.4±0.0 (-4.8%)	37.0±0.0 (+11.6%)	44.2±0.0 (-6.6%)	45.8±0.0 (-9.8%)	38.0±0.3 (+8.7%)	**36.3±0.3 (+13.8%)**
	Test	42.8±0.4	46.1±0.0 (-7.2%)	39.5±0.0 (+8.4%)	45.9±0.0 (-6.8%)	45.7±0.0 (-6.3%)	39.3±0.6 (+8.9%)	**37.5±0.2 (+14.1%)**
Prod-40 (4)	Train	35.1±0.3	39.4±0.0 (-10.9%)	31.3±0.0 (+12.1%)	36.4±0.0 (-3.6%)	38.8±0.0 (-9.5%)	33.9±2.5 (+3.5%)	**28.3±0.3 (+24.0%)**
	Test	38.3±0.3	43.6±0.0 (-12.2%)	33.5±0.0 (+14.3%)	37.4±0.0 (+2.4%)	40.1±0.0 (-4.5%)	36.7±2.8 (+4.4%)	**30.4±0.7 (+26.0%)**
Prod-80 (8)	Train	43.2±0.2	44.3±0.0 (-2.5%)	39.0±0.0 (+10.8%)	43.7±0.0 (-1.1%)	49.3±0.0 (-12.4%)	56.6±6.8 (-23.7%)	**33.6±0.9 (+28.6%)**
	Test	47.7±0.4	53.9±0.0 (-11.5%)	41.9±0.0 (+13.8%)	46.1±0.0 (+3.5%)	49.6±0.0 (-3.8%)	62.5±4.2 (-23.7%)	**35.2±0.8 (+35.5%)**

Takeaways

- **Summary**
  - We explore embedding table placement/sharding, a direction that has been rarely explored.
  - We propose DreamShard, which learns estimated MDP and an RL agent.
  - DreamShard significantly outperforms heuristic baselines.

Paper

Code