



Distributed Hierarchical GPU Parameter Server for Massive Scale Deep Learning Ads Systems

Presenter: **Weijie Zhao¹**

¹ **Cognitive Computing Lab, Baidu Research**

Joint work with

Deping Xie², Ronglai Jia², Yulei Qian², Ruiquan Ding³, Mingming Sun¹, Ping Li¹

² Baidu Search Ads (Phoenix Nest), Baidu Inc. ³ Sys. & Basic Infra., Baidu Inc.



Sponsored Online Advertising

The screenshot shows a Baidu search interface. The search bar contains the query "austin hotel". Below the search bar is a navigation menu with categories: 网页 (selected), 资讯, 视频, 图片, 知道, 文库, 贴吧, 采购, 地图, and 更多». A message indicates approximately 1,040 results found. The main content area is blurred, but a blue "Ads" label is visible in the bottom right corner of the blurred area.

Baidu

austin hotel

百度一下

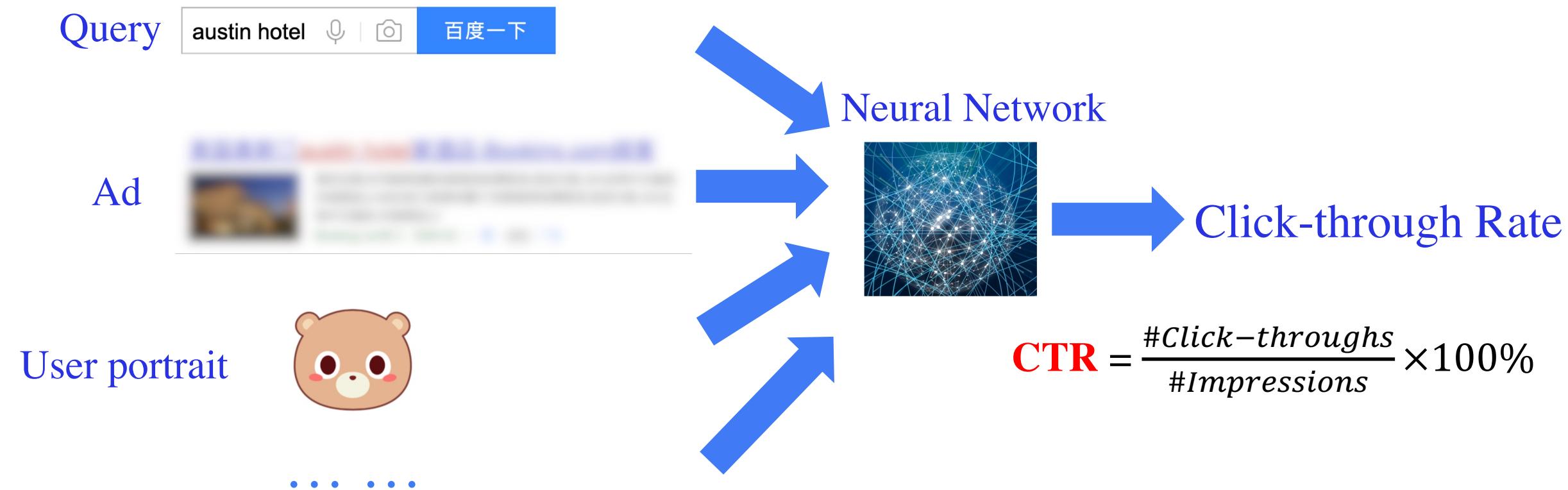
网页 资讯 视频 图片 知道 文库 贴吧 采购 地图 更多»

百度为您找到相关结果约1,040个

搜索工具

Ads

Sponsored Online Advertising



High-dimensional sparse vectors (10^{11} dimensions)

CTR Prediction Models at Baidu

2009 and earlier: Single machine CTR models

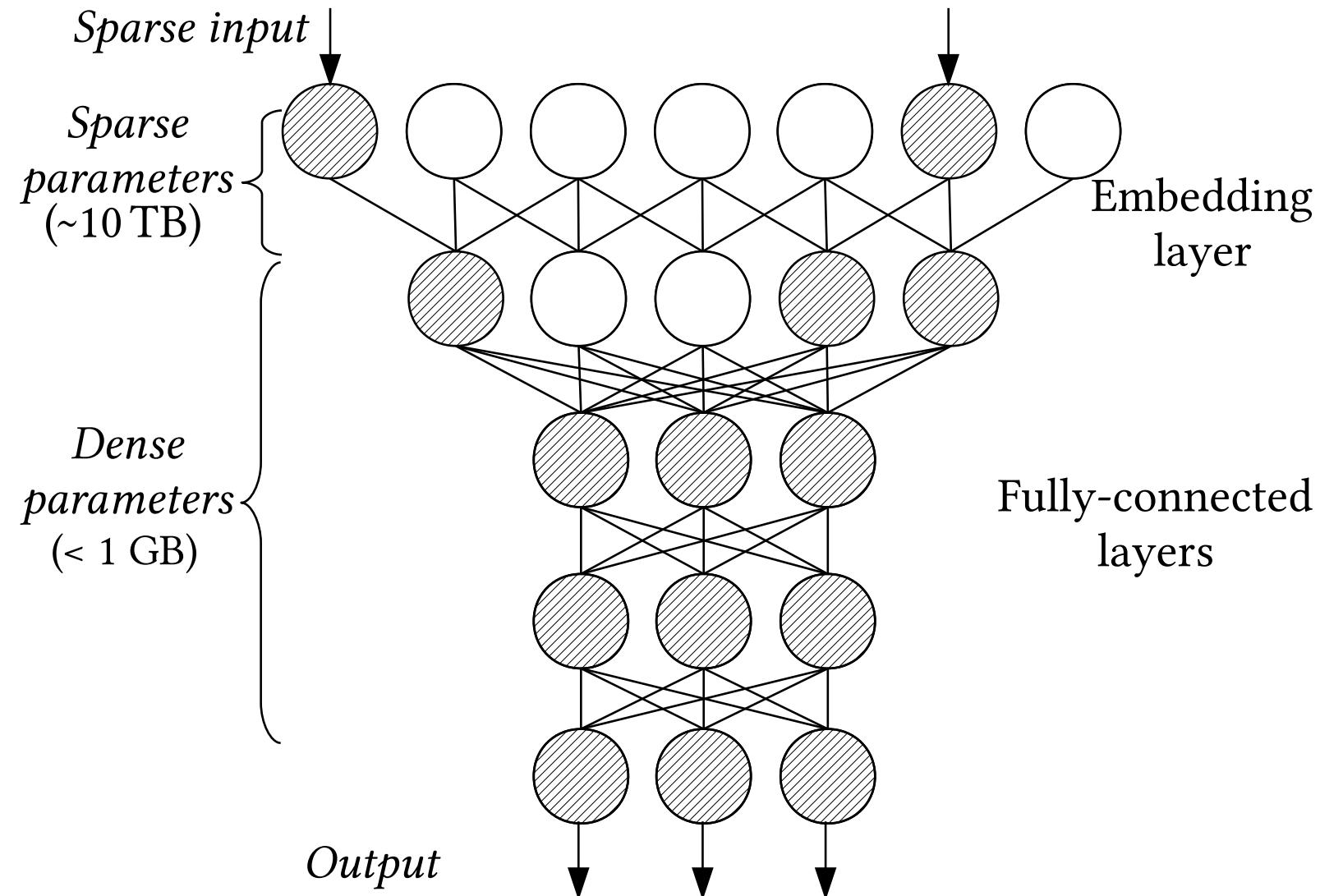
2010: Distributed logistic regression (LR) and distributed parameter server

2013: Distributed deep neural networks (DNN), extremely large models

Since 2017: Single GPU AIBox, Multi-GPU Hierarchical Parameter Server, Approx. near neighbor (ANN) search, Maximum inner product search (MIPS)

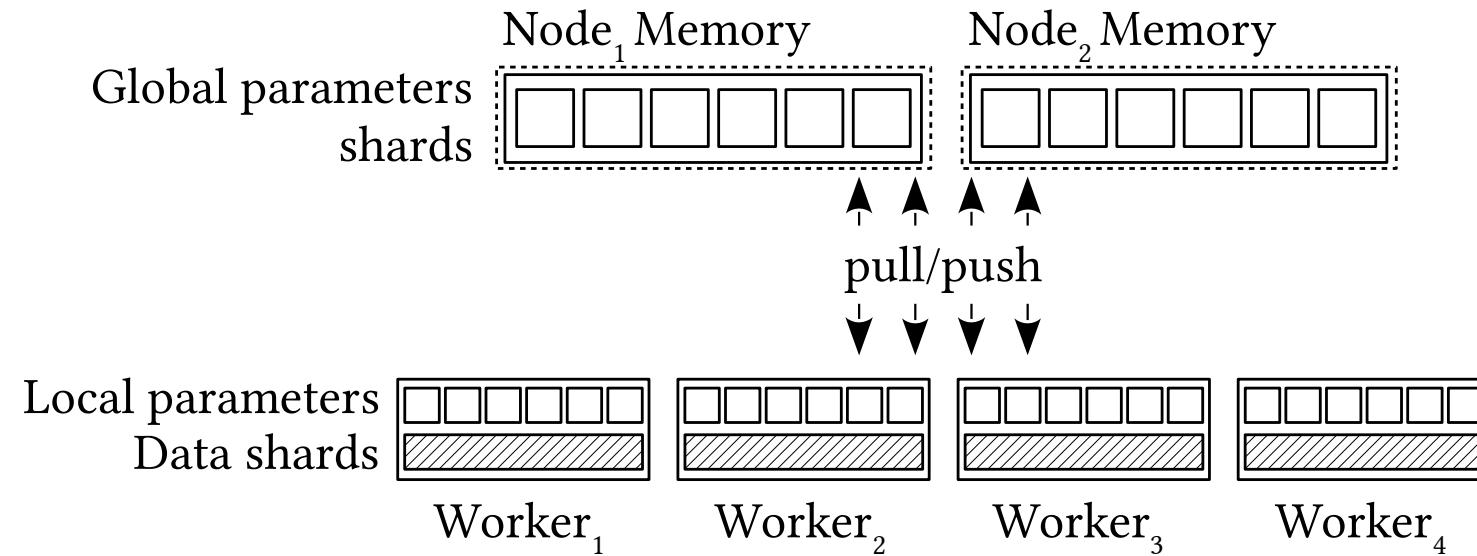
ANN and MIPS have become increasingly important in the whole pipeline of CTR prediction, due to the popularity & maturity of embedding learning and improved ANN/MIPS techniques

A Visual Illustration of CTR Models



MPI Cluster Solution

Distributed Parameter Server



Wait! Why do We Need Such a Massive Model?

Hashing For Reducing CTR Models

One permutation + one sign random projection (work done in 2015)

Table 1. OP+OSRP for Image Search Sponsored Ads Data

	# Nonzero Weights	Test AUC
Baseline LR	31,949,213,205	0.7112
Baseline DNN		0.7470
Hash+DNN ($k = 2^{34}$)	6,439,972,994	0.7407
Hash+DNN ($k = 2^{23}$)	3,903,844,565	0.7388
Hash+DNN ($k = 2^{22}$)	2,275,442,496	0.7370
Hash+DNN ($k = 2^{31}$)	1,284,025,453	0.7339
Hash+DNN ($k = 2^{30}$)	707,983,366	0.7310
Hash+DNN ($k = 2^{29}$)	383,499,175	0.7278
Hash+DNN ($k = 2^{28}$)	203,864,439	0.7245
Hash+DNN ($k = 2^{27}$)	106,824,123	0.7208
Hash+DNN ($k = 2^{26}$)	55,363,771	0.7175
Hash+DNN ($k = 2^{25}$)	28,479,330	0.7132
Hash+DNN ($k = 2^{24}$)	14,617,324	0.7113

Image search ads
is typically a small
source of revenue

1. Hashing + DNN significantly improves over LR (logistic regression)!
2. A fine solution if the goal is to use **single-machine** to achieve good accuracy!

Hashing For Reducing CTR Models



One permutation + one sign random projection (work done in 2015)

Table 2. OP+OSRP for Web Search Sponsored Ads Data

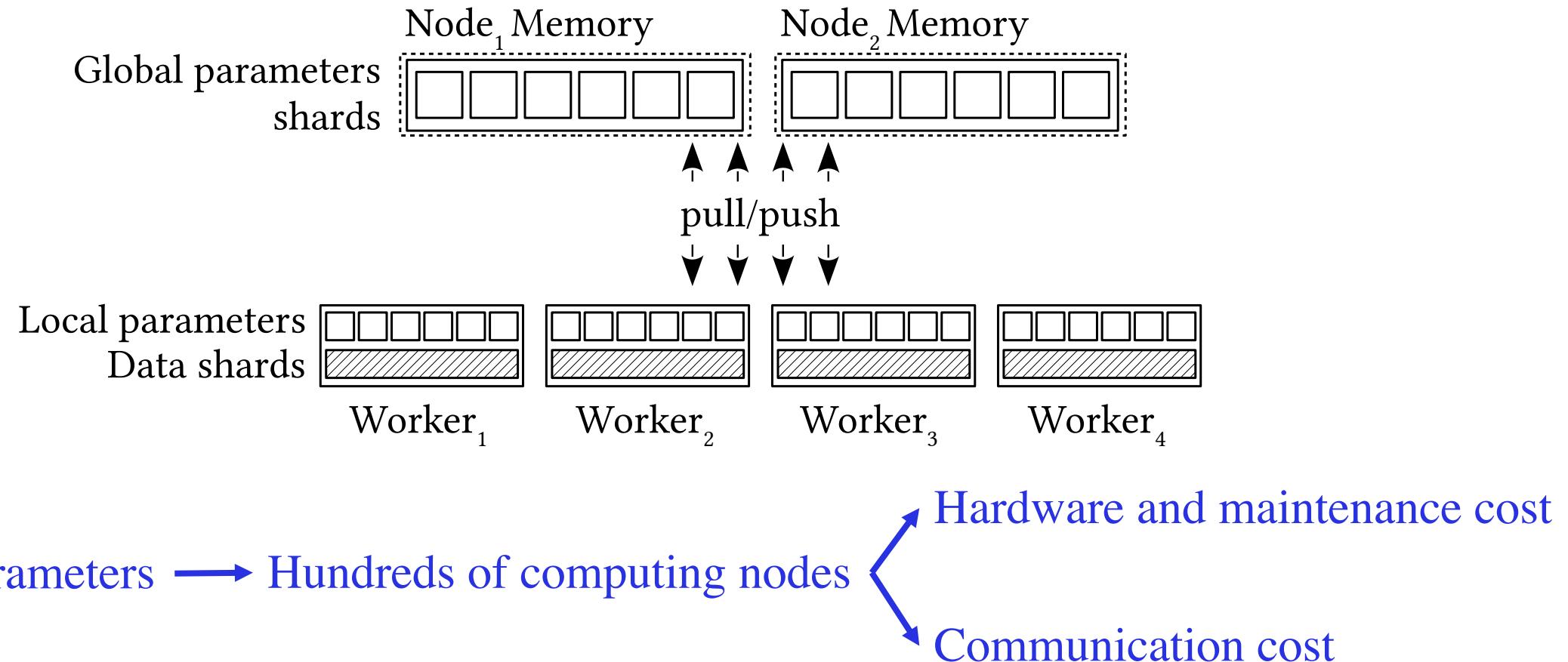
Web search ads
use more features
and larger models

	# Nonzero Weights	Test AUC
Baseline LR	199,359,034,971	0.7458
Baseline DNN		0.7670
Hash+DNN ($k = 2^{32}$)	3,005,012,154	0.7556
Hash+DNN ($k = 2^{31}$)	1,599,247,184	0.7547
Hash+DNN ($k = 2^{30}$)	838,120,432	0.7538
Hash+DNN ($k = 2^{29}$)	433,267,303	0.7528
Hash+DNN ($k = 2^{28}$)	222,780,993	0.7515
Hash+DNN ($k = 2^{27}$)	114,222,607	0.7501
Hash+DNN ($k = 2^{26}$)	58,517,936	0.7487
Hash+DNN ($k = 2^{24}$)	15,410,799	0.7453
Hash+DNN ($k = 2^{22}$)	4,125,016	0.7408

1. Even a 0.1% decrease in AUC would result in a noticeable decrease in revenue
2. Solution of using hashing + DNN + single machine is typically not acceptable

MPI Cluster Solution

Distributed Parameter Server

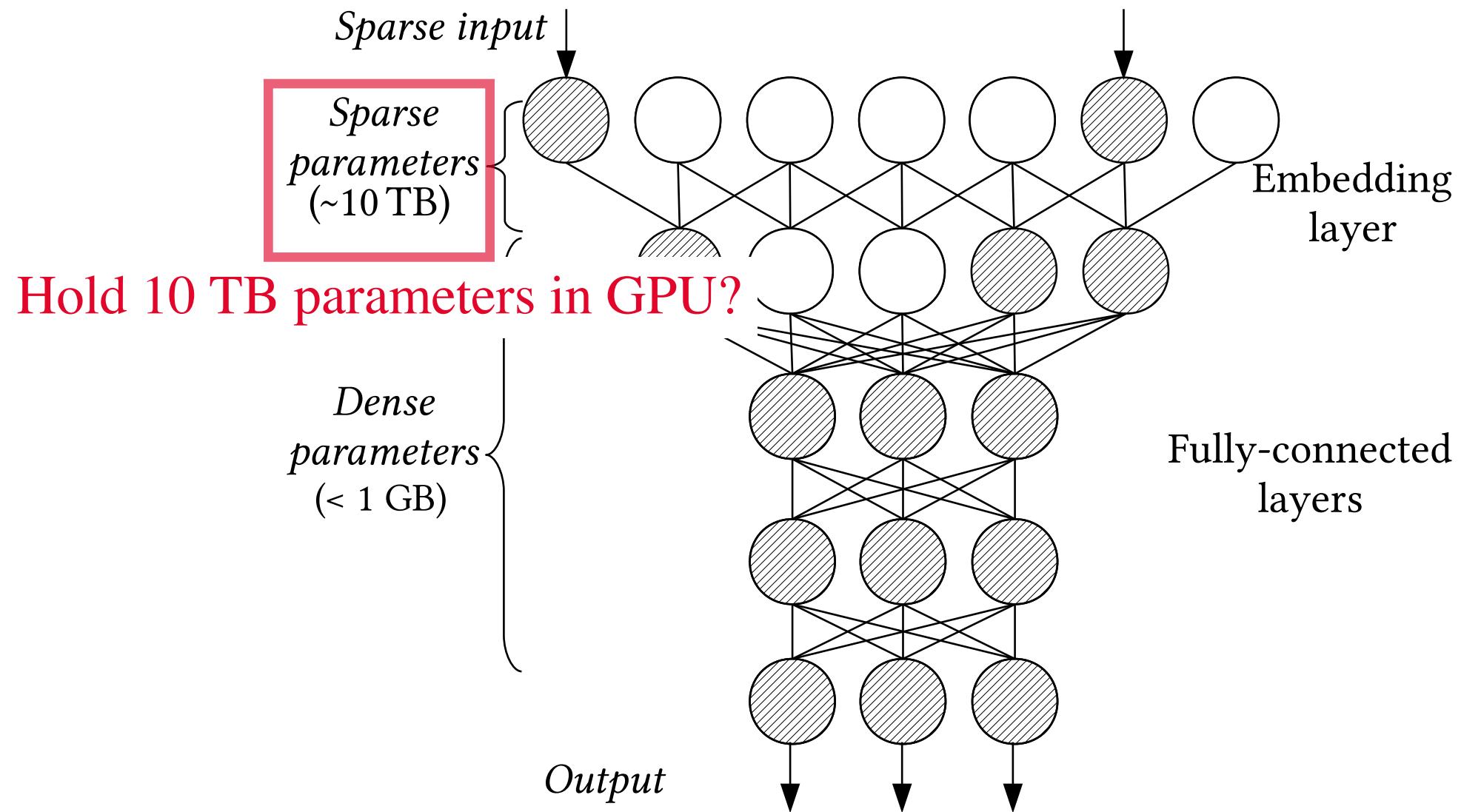


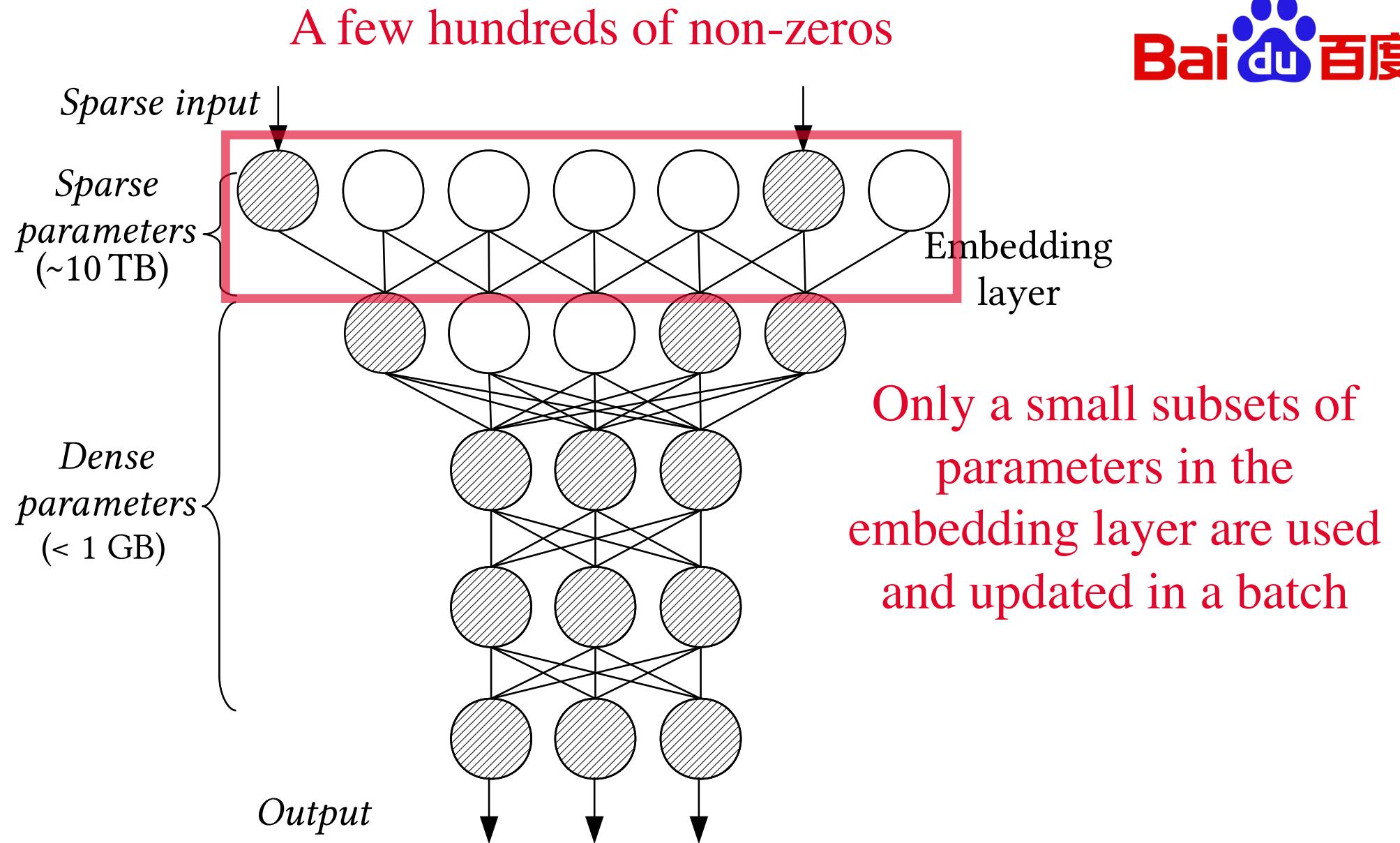


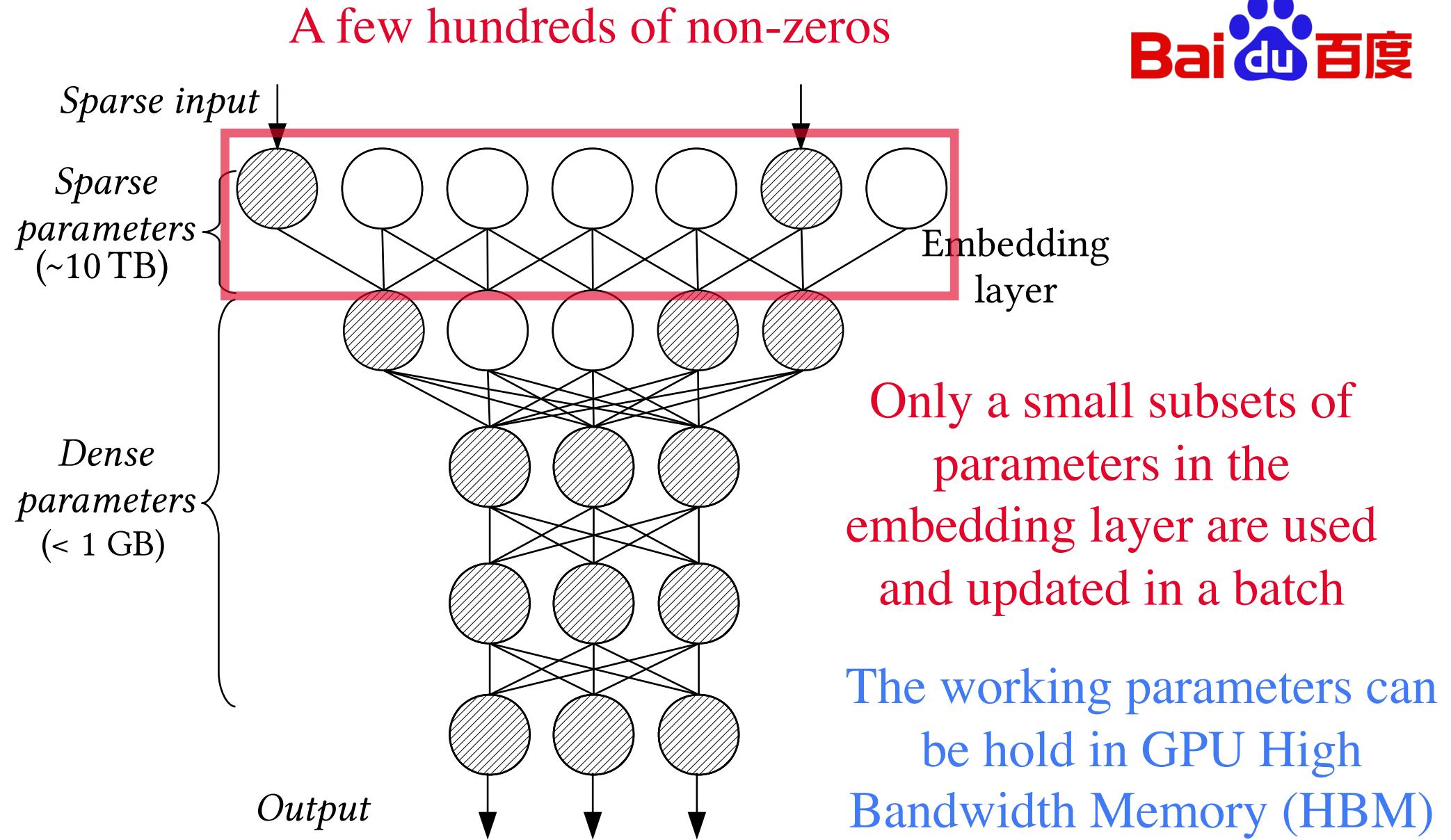
But all the cool kids use GPUs!

Let's train the 10-TB Model with GPUs!

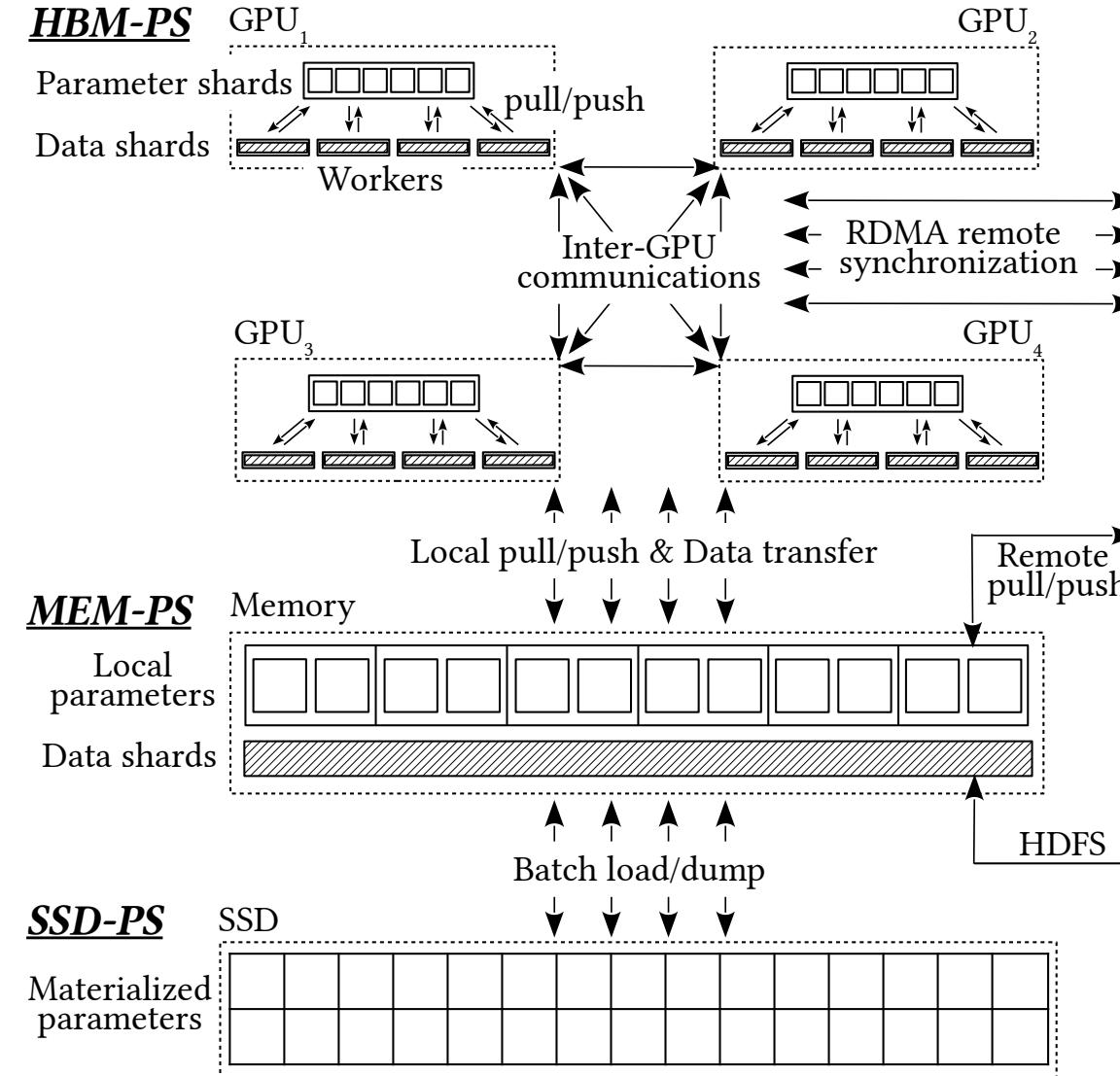


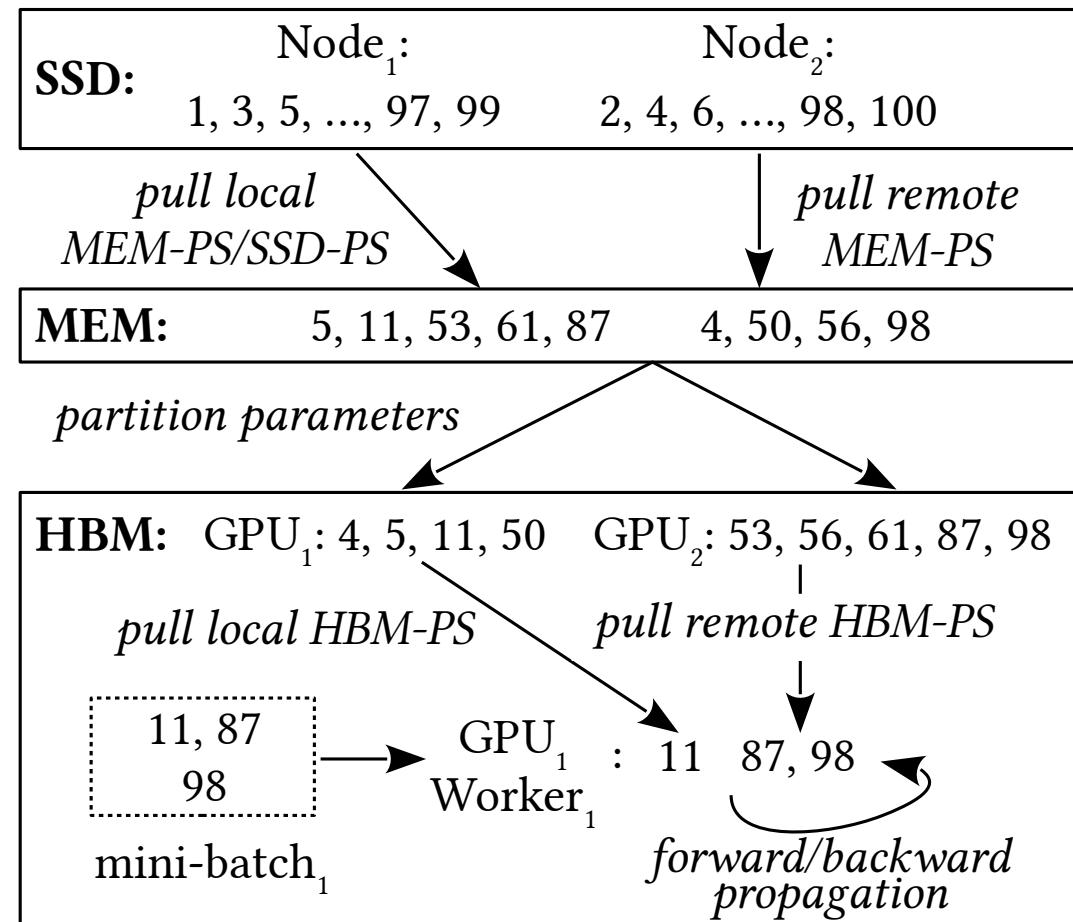
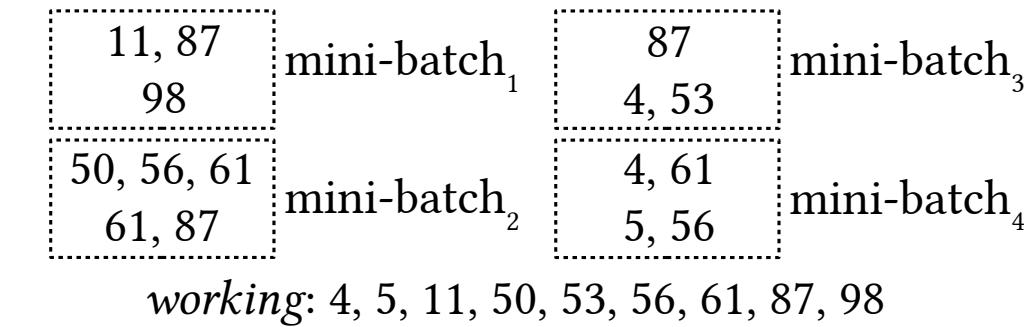






Solve the Machine Learning Problem in a System Way!





Experimental Evaluation

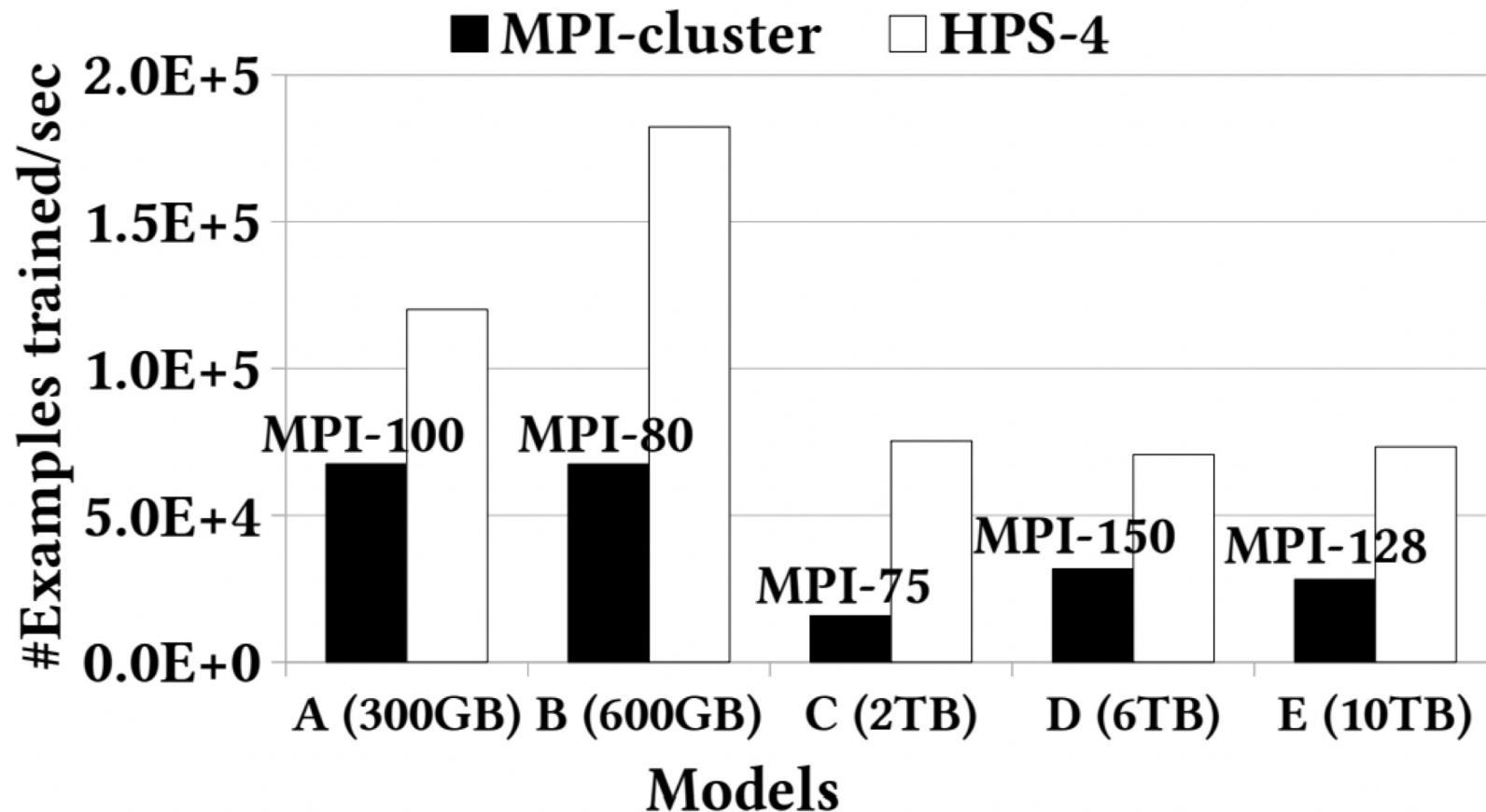
- 4 GPU computing nodes
 - 8 cutting-edge 32 GB HBM GPUs
 - Server-grade CPUs with 48 cores (96 threads)
 - ~1 TB of memory
 - ~20 TB RAID-0 NVMe SSDs
 - 100 Gb RDMA network adaptor
-

Experimental Evaluation

Table 3. Model specifications.

	#Non-zeros	#Sparse	#Dense	Size (GB)	MPI
A	100	8×10^9	7×10^5	300	100
B	100	2×10^{10}	2×10^4	600	80
C	500	6×10^{10}	2×10^6	2,000	75
D	500	1×10^{11}	4×10^6	6,000	150
E	500	2×10^{11}	7×10^6	10,000	128

Execution Time



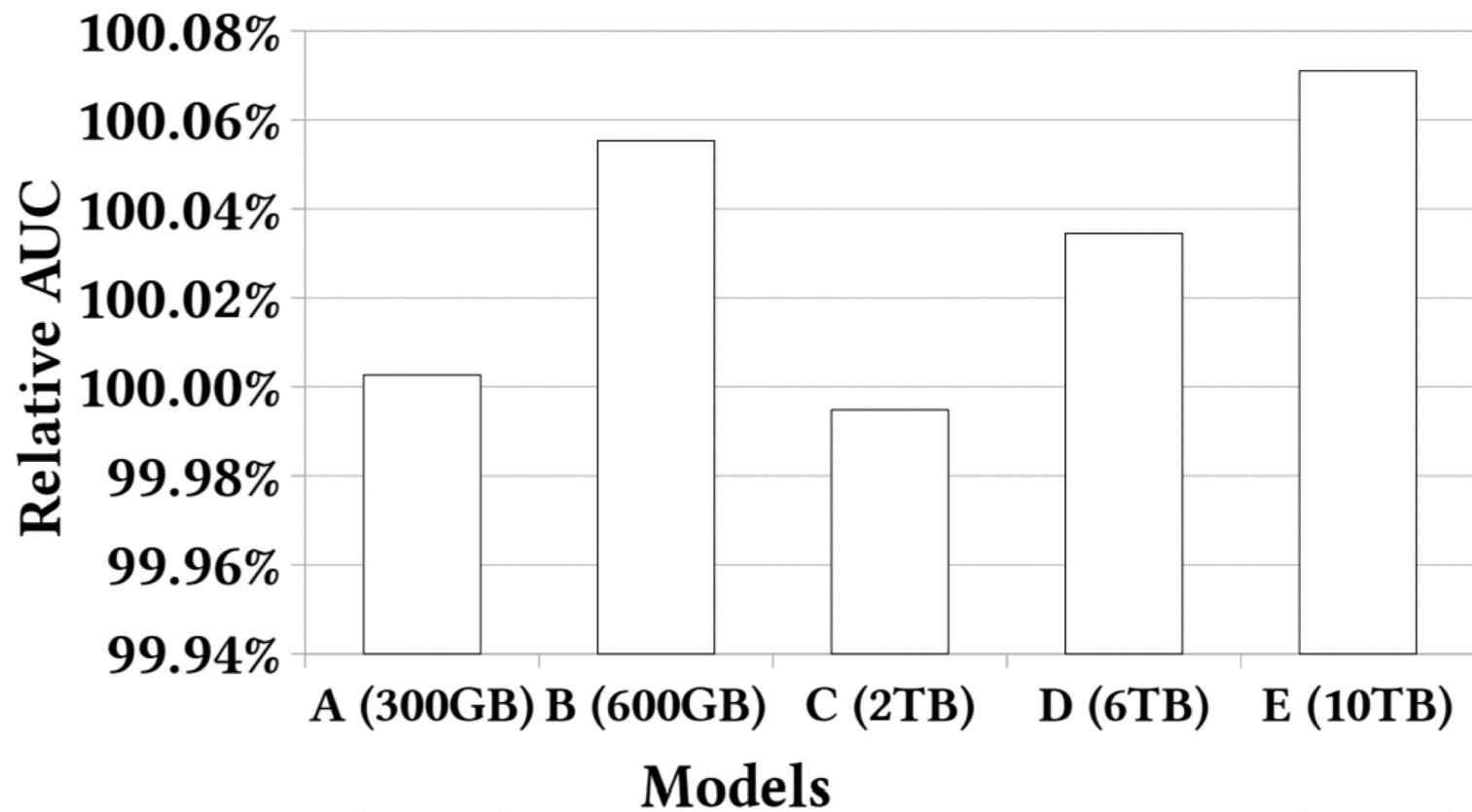
Price-Performance Ratio

- Hardware and maintenance cost: 1 GPU node ~ 10 CPU-only nodes
- 4 GPU node vs. 75-150 CPU nodes

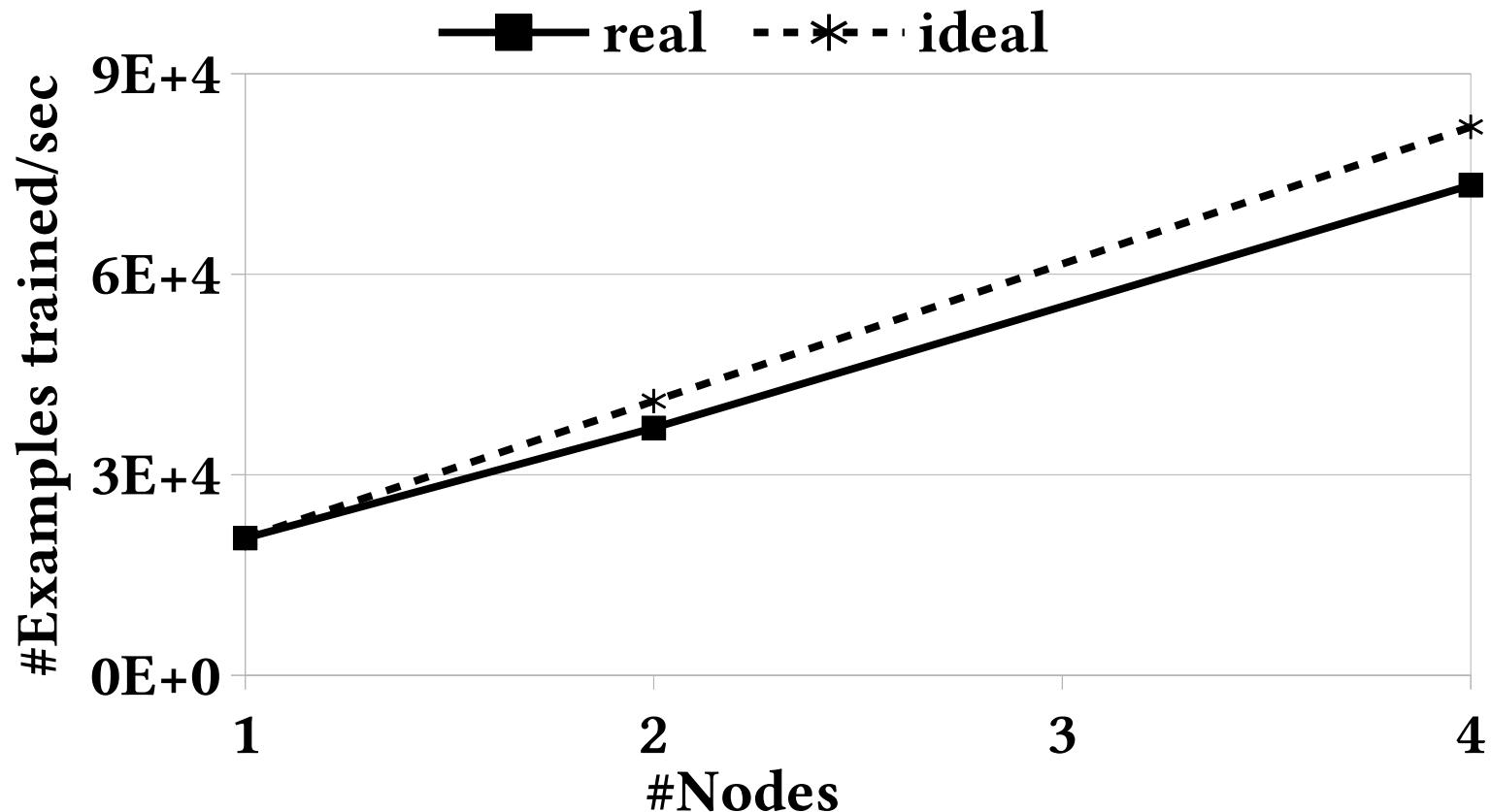
Table 4. The speedup over the MPI-cluster solution and the normalized speedup at the same hardware and maintenance cost.

	A	B	C	D	E
Speedup over MPI-cluster	1.8	2.7	4.8	2.2	2.6
Cost-normalized speedup	4.4	5.4	9.0	8.4	8.3

AUC



Scalability



Conclusions



- We introduce the architecture of a distributed hierarchical GPU parameter server for massive deep learning ads systems.
- We perform an extensive set of experiments on 5 CTR prediction models in real-world online sponsored advertising applications.
- A 4-node hierarchical GPU parameter server can train a model more than 2X faster than a 150-node in-memory distributed parameter server in an MPI cluster.
- The cost of 4 GPU nodes is much less than the cost of maintaining an MPI cluster of 75-150 CPU nodes.
- The price-performance ratio of this proposed system is 4.4-9.0X better than the previous MPI solution.
- This system is being integrated with the **PaddlePaddle** deep learning platform (<https://www.paddlepaddle.org.cn>) to become **PaddleBox**.

