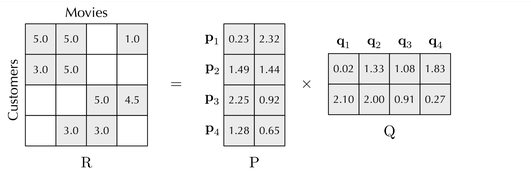
Efficient Matrix Factorization on Heterogeneous CPU-GPU Systems

Yuanhang Yu♮, Dong Wen♮, Ying Zhang♮, Xiaoyang Wang§, Wenjie Zhang†, and Xuemin Lin†

♮AAII, University of Technology Sydney, Australia § Zhejiang Gongshang University, China †The University of New South Wales, Australia

Matrix Factorization



A rating matrix R and a corresponding matrix factorization

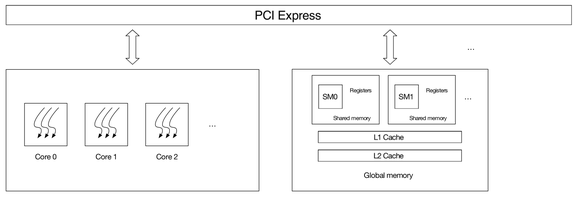
Matrix factorization aims to represent the matrix R as a dot product between two dense matrices

$$R \approx P \times Q, P \in \mathrm{R}^{m \times k}, Q \in R^{k \times n}$$

where k is the number of latent factors

Methods: alternating least squares (ALS)
coordinate descent (CD)
stochastic gradient descent (SGD) We use this !

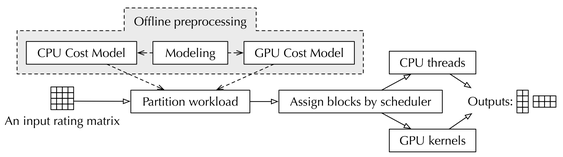
Heterogeneous CPU-GPU system



Heterogeneous CPU-GPU Systems

Heterogeneous CPU-GPU systems contains two types of hardware sources, which can transfer data via PCI Express.

*HSGD**



Overview of HSGD*

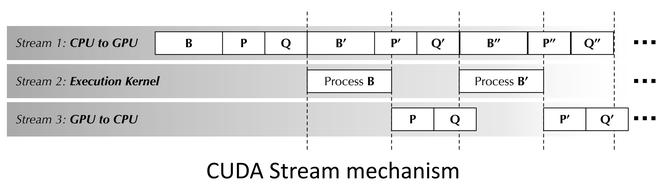
The offline phase derives a cost model which estimates the hardware performance.

In the online phase, HSGD* divides the matrix based on the cost model.

Then, the scheduler assigns blocks to worker threads.

Our cost model takes CUDA stream mechanism into consideration and establish models for data transfer and kernel execution respectively.

Non-linear model for GPU part

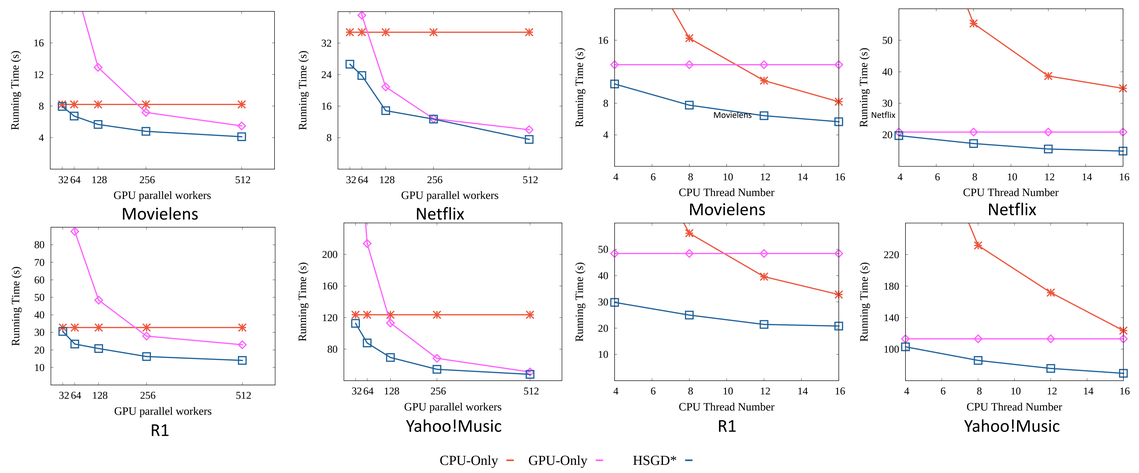


CUDA Stream mechanism

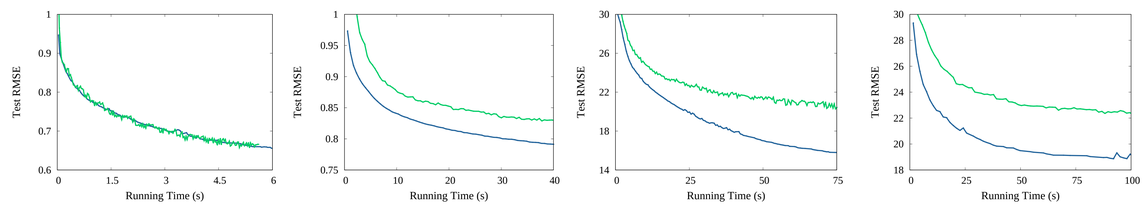
Contributions

- A parallel SGD algorithm on heterogeneous systems.
- A novel cost model to estimate the performance of GPUs.
- Extensive experiments on four benchmark datasets

Overall Efficiency



Training Quality



Workload Balance

TABLE II
COMPARISON OF COST MODELS

Datasets		MovieLens	Netflix	R1	Yahoo!Music
Workload proportion					
HSGD*-Q	C	49.56%	55.98%	56.07%	56.46%
	G	50.44%	44.02%	43.93%	43.54%
HSGD*-M	C	55.91%	49.02%	49.75%	53.61%
	G	44.09%	50.98%	50.25%	46.39%
Running time					
HSGD*-Q		0.92 s	15.87 s	13.07 s	40.88 s
HSGD*-M		**0.89** s	**13.02** s	**12.08** s	**35.41** s

Shorter running time

ICDE2021