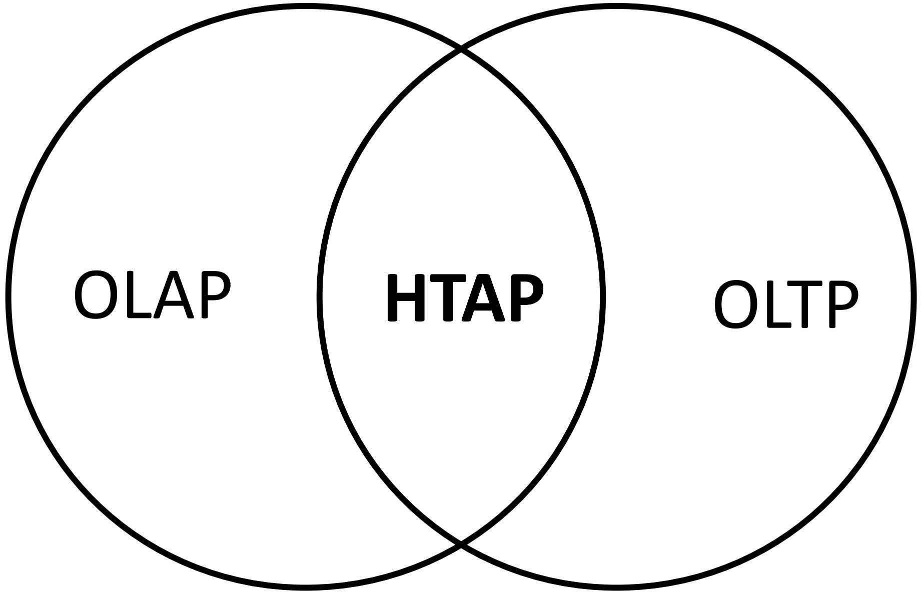
Hybrid Transactional/Analytical Processing Literature Review

CSE 5249 AU20

Team 3: Haiyang Qi, Yuting Fang

0. Introduction



- Online Transactional Processing (OLTP):

  atomic change of state, record insert/delete/update

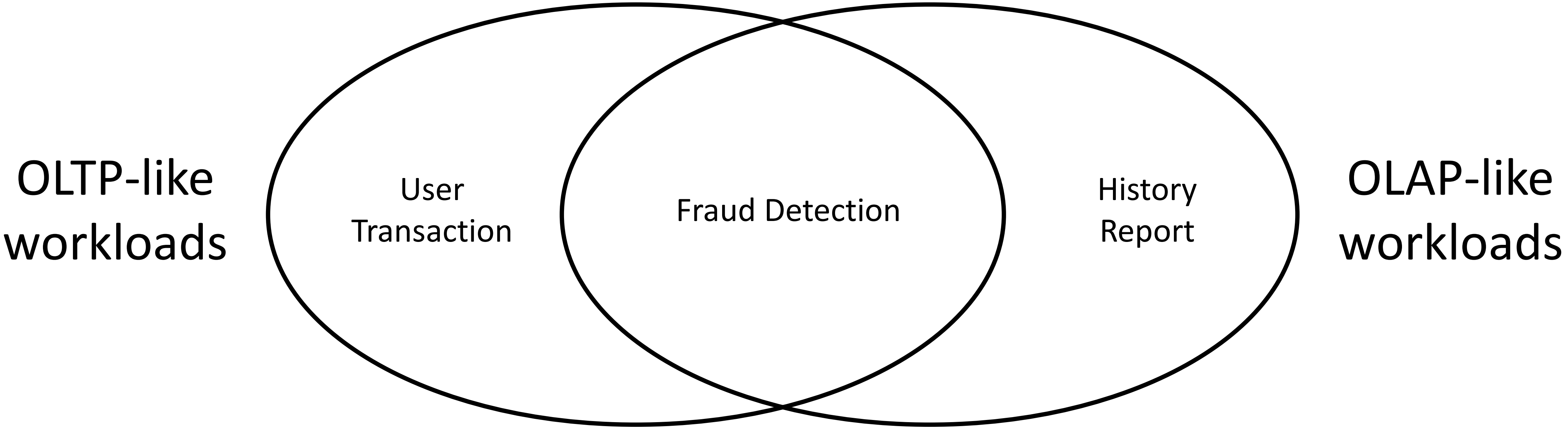
- Online Analytical Processing (OLAP):

  get insight of data, supports planning or forecasting

  usually require scans of the tables and process in batches

0. Introduction

- Application: **large-scale real-time analytics applications** like Internet of Things (IoT), risk analysis, mobile app personalized recommendation.



A Example of Bank Platform

1. Research Goal & Problems

	Latency	Volume	Concurrency
OLTP	Lower	Higher	Higher
OLAP	Higher	Lower	Lower

- Traditional Database Solutions:

	OLTP	OLAP
1. Indexing data for fast accessing	☑	☒
2. Using shared file systems for scans	☒	☑

→ Hard to balance

1. Research Goal & Problems

- Research Goal:

  Improve the overall efficiency of **simultaneously processing transactions and analyses streams**, specifically when the situation calls for **large amounts** of transactions and analyses to happen at the same time.

2. Sources selection and search

- Database
    - Google Scholar
- Keyword
    - “HTAP”, “OLTP”, “OLAP”
    - “CPU”, “GPU”
- Inclusion Criteria
    - Published in recent 10 years
    - Top Conferences: VLDB, SIGMON, ICDE
    - Papers that identify procedures or techniques of HTAP
    - Papers that present experiment on HTAP
    - Papers that discuss evaluation of HTAP
    - …

3. Selected Literatures

1. Scheduling Concurrent Applications on A Cluster of CPU-GPU Nodes, Vignesh T. Ravi (IEEE, 2012)
2. A Framework for Developing Real-Time OLAP algorithm using Multi-core processing and GPU: *Heterogeneous Computing, H I Alzeini* (ICOM, 2013)
3. The Case For Heterogeneous HTAP, Raja Appuswamy (CIDR, 2017)
4. Low-Latency Transaction Execution on Graphics Processors: Dream or Reality?, Iya Arefyeva (VLDB, 2018)
5. Memory Management Strategies in CPU/GPU Database Systems: A Survey, Iya Arefyeva (BDAS, 2018).

3.1 Literature Analysis – **Real–Time OLAP**

	Latency	Volume	Concurrency
OLAP	**High**	Low	Low

- **Traditional Methods**: Materialization
  - Pre-fetching data, pre-computing prospective queries
  - Example: PostgreSQL – `CREATE MATERIALIZED VIEW …`
  - ➢ answers do not include current updates

- **Problem**: cannot meet the Real-Time requirements

Literature: A Framework for Developing Real-Time OLAP algorithm using Multi-core processing and GPU: *Heterogeneous Computing*, H I Alzeini (ICOM, 2013)

3.1 Literature Analysis – Real–Time OLAP

- **Proposed Solution**:
  - Ignore Materialization
  - Compensate performance degradation with hardware development

1. **CPU+GPU Hybrid System**
   - Increase processing capability significantly
2. (Task/Processing) Distribution and Partition Algorithm
   - Assign different tasks to proper resource (CPU or GPU)
   - Utilize both CPU and GPU efficiently - research question

Literature: A Framework for Developing Real-Time OLAP algorithm using Multi-core processing and GPU: *Heterogeneous Computing*, H I Alzeini (ICOM, 2013)

3.1 Literature Analysis – **OLTP on GPU**

	Latency	**Volume**	**Concurrency**
OLTP	Low	High	High

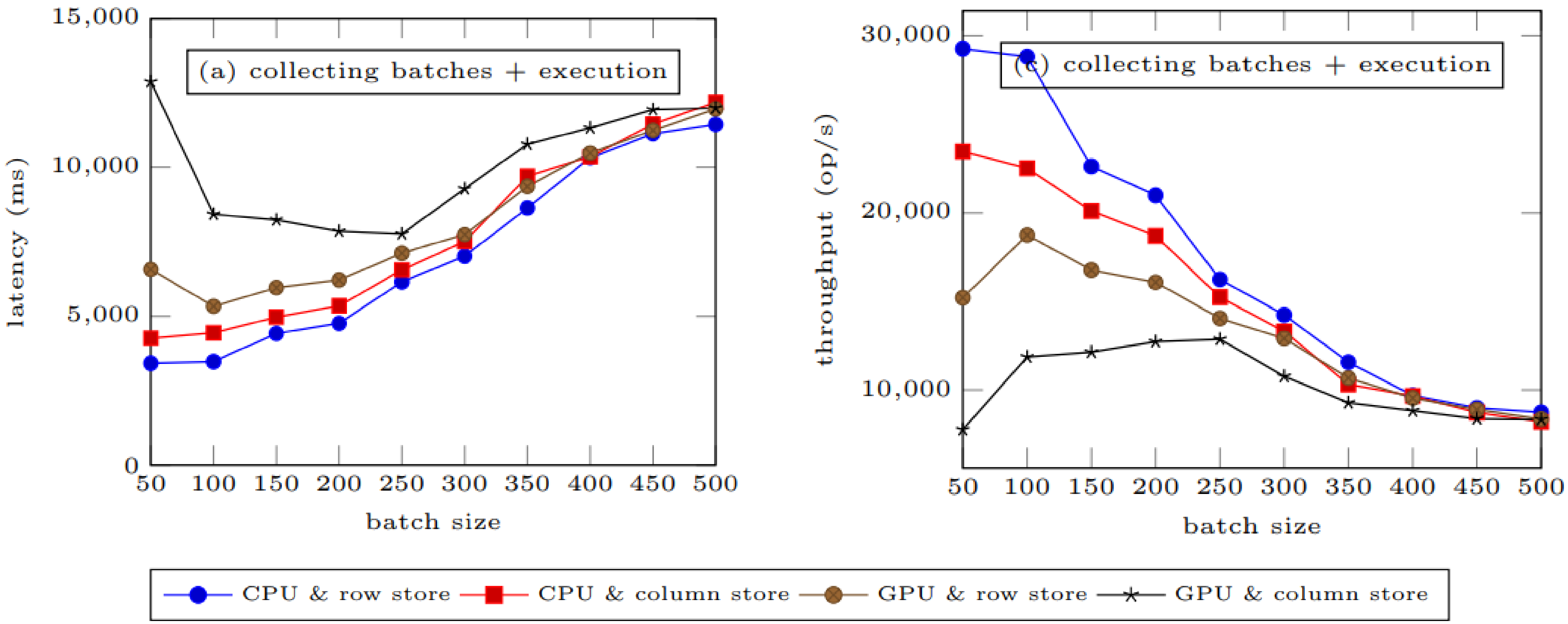
OLTP-like Workload: High volume of short transactions

- **Traditional Assumption:** GPUs cannot support OLTP efficiently.
  - If process in small batches: does not utilize GPU efficiently
  - If process in big batches: wait longer to collect and transfer

Literature: Low-Latency Transaction Execution on Graphics Processors: Dream or Reality?, Iya Arefyeva (VLDB, 2018).

3.1 Literature Analysis – **OLTP on GPU**

- **Traditional Assumption:** GPUs cannot support OLTP efficiently.



Literature: Low-Latency Transaction Execution on Graphics Processors: Dream or Reality?, Iya Arefyeva (VLDB, 2018).

3.1 Literature Analysis – OLTP on GPU

- **Problem**: In HTAP system, when the workload switches to OLTP mostly, GPUs will be underutilized significantly.

- **Proposed Solutions**:
  - Precondition:
    1. High request arrival rate: little-to-no wait time for big batch
    2. Moderate request rate: break into sufficient parallel operations
  - **Concurrency control**
    - Example: GPU serves only large batches, scheduling
    - ➢ Utilize GPU efficiently in HTAP system - research question

Literature 3: Low-Latency Transaction Execution on Graphics Processors: Dream or Reality?, Iya Arefyeva (VLDB, 2018).

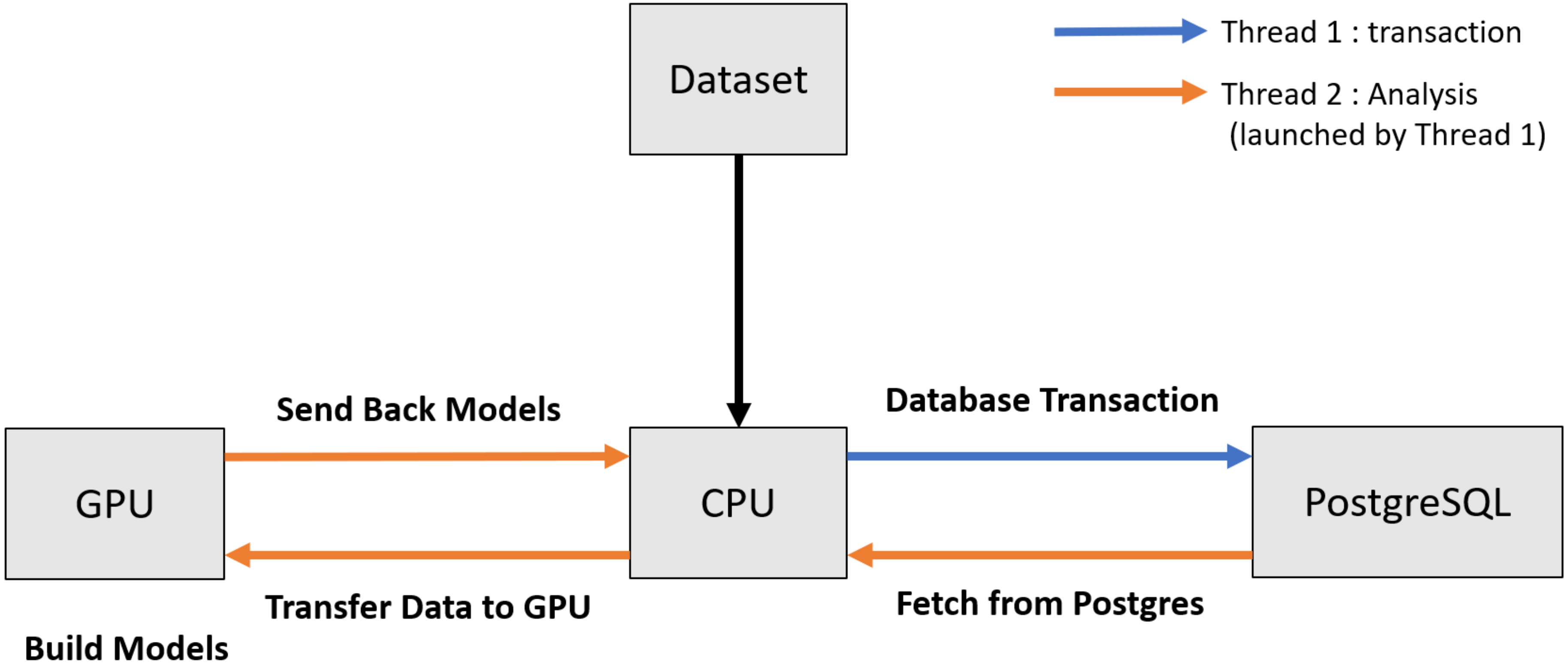
3.2 Literatures Analysis

- Problems: how to support efficient processing of transactional and analytical request simultaneously

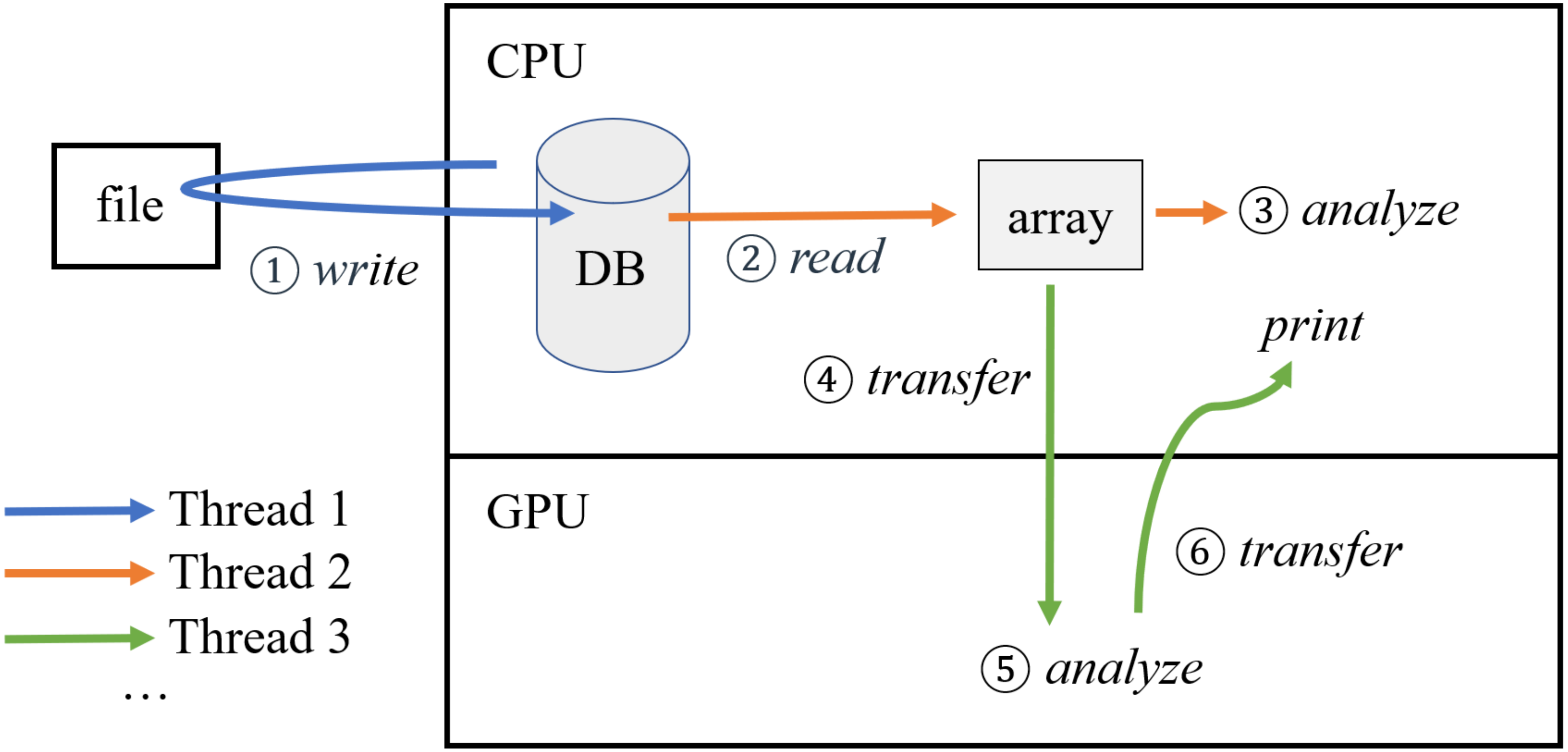
- Key Components:
  - CPU+GPU Hybrid System
  - Distribution
  - Scheduling
  - Memory Management
  - …

CPU – GPU Strategies

4. Proposed Guideline



4. Proposed Experiment



Thank you!

qi.359@osu.edu
fang.564@osu.edu