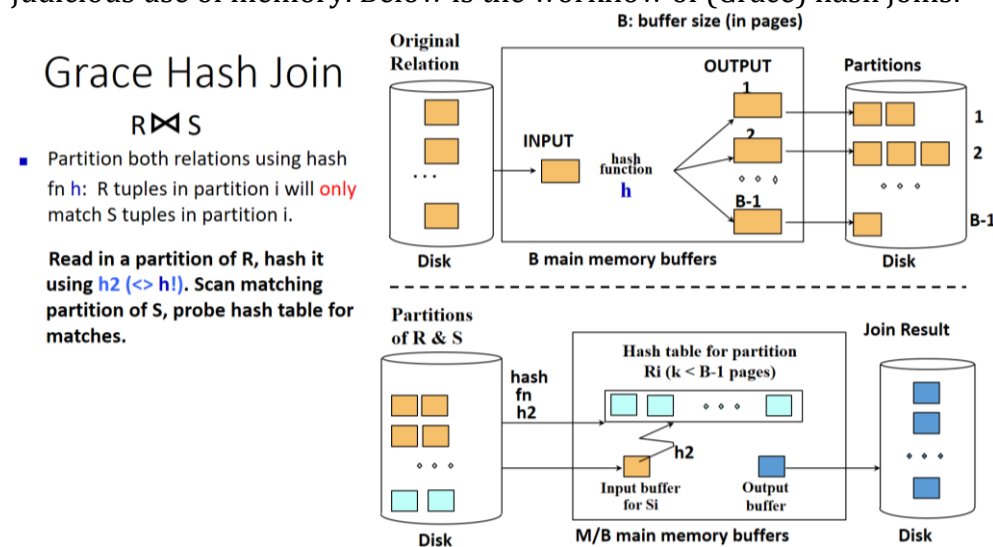


## CS561 Spring 2023 - Research Project

**Title:** *Boosting Join Implementation for Skew Correlation in Postgres*

**Background:** Join is a fundamental operator in relational database systems. Among different join implementations, hash join offers efficient performance and judicious use of memory. Below is the workflow of (Grace) hash joins:



**Problem:** Hash join uniformly distributes all the records regardless of the correlation of the join attributes. When there is a skew correlation during the join execution, a few partitions might be larger than the available memory which results in unnecessary I/Os. Existing skewness-based optimization in Postgres only caches the most frequent items in memory to avoid redundant I/Os. However, our analysis shows that only caching the most frequent items in memory may not be optimal for storage-based PK-FK (primary-key foreign-key) joins.

**Objective:** The objective of the project is to boost the hybrid hash join implementation for skew optimization in Postgres. The workflow is as follows:

- Understand the implementation of hybrid hash join in Postgres [1,2].
- Refine an existing partitioning strategy for partition-wise joins according to the existing skew optimization and integrate it into Postgres (contact **Zichen Zhu** to request material for the partitioning strategy).
- Benchmark the performance for skew join under memory pressure (a modified TPC-H generator is given to produce skew correlation between tables `orders` and `lineitem`).

**Responsible Mentor:** *Zichen Zhu (zczhu@bu.edu)*

**Postgres Repo:** <https://github.com/postgres/postgres>

**References:**

[1] [PostgreSQL Source Code: src/backend/executor/nodeHashjoin.c File Reference](#)

[2] [PostgreSQL Source Code: src/backend/executor/nodeHash.c File Reference](#)