Hermes at Scale

Powering Distributed Queries with a Unified Memory Fabric

Tim Gubner, Matheus Nerone, Diego Tome, Vitor Oliveira, Rune Humborstad, Manyi Lu

Now: on sabbatical

Now: @Clickhouse

Huawei Europe

HUAWEI

Hermes – Single Node Accelerator for Analytics

- OLTP DBMS typically slow for OLAP but widely used
- Offload analytic queries to Hermes
  - Replicate data and in-flight transactions to Hermes
  - Currently only MySQL supported (but design is generic)
- Hermes:
  - Storage optimized for analytics (compressed columnar Main Store) and transactions (Delta Store)
  - Engine is optimized for analytics

Hermes – Single Node Accelerator for Analytics

- OLTP DBMS typically slow for OLAP but widely used
- Offload analytic queries to Hermes
  - Replicate data and in-flight transactions to Hermes
  - Currently only MySQL supported (but design is generic)
- Hermes:
  - Storage optimized for analytics (compressed columnar Main Store) and transactions (Delta Store)
  - Engine is optimized for analytics

But

- in-memory
- limited to single-node (fortunately many workloads are rather small*)

* Alexander Van Renen and Viktor Leis. Cloud Analytics Benchmark. VLDB 2023.

Scale Out

Add more nodes for memory & get compute power for free

But how?

Disaggregated Storage:

+ very flexible (easy to add/remove nodes cheaply)
– excessive data transfers

Shared Storage/Memory:

+ flexible
+ data can be local

Shared Nothing:

– rigid
+ data local
– does not really allow load-balancing within query

Memcom: Shared Memory with Affinities

Data stored via abstraction layer *Memcom*

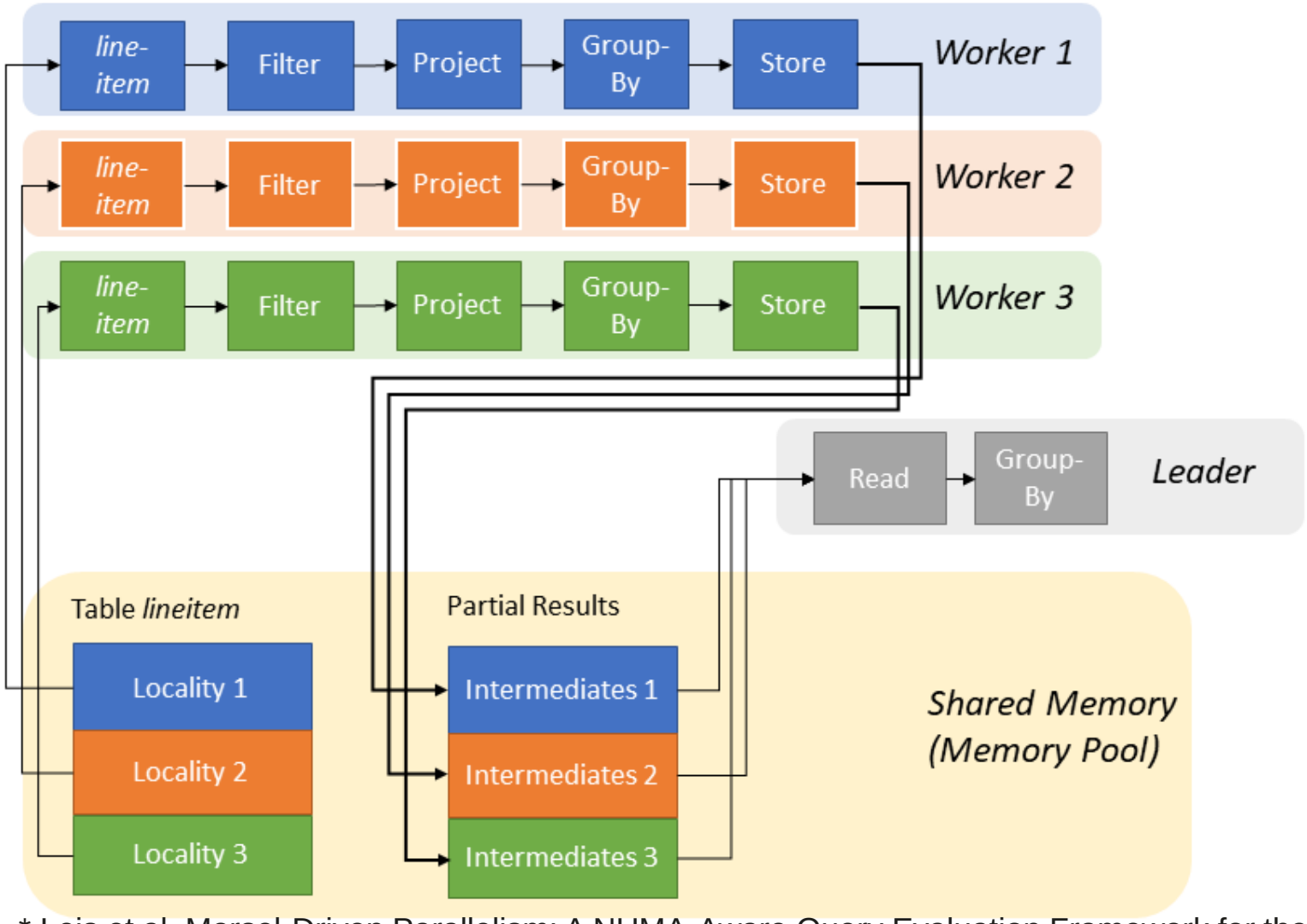
provides allocation, placement, access and latching

Specifically:

- Main Store + Meta Data, Delta Store, Catalog as well as Intermediates (to be shared between nodes)

- Main Store consists of row groups annotated with locality (home node where data is located)

Example



Data is partitioned (somehow)

Data access: *Mostly local*

- Workers read local data first (can read technically *all* data)
- Workers write partial results locally

Load-balancing:

- If worker is faster, it will consume more data -> as load-balancing as Morsel-driven Parallelism*

* Leis et al. Morsel-Driven Parallelism: A NUMA-Aware Query Evaluation Framework for the Many-Core Age. SIGMOD 2014

Intra-Query Parallelism

1. Exchange model parallelism* not easily possible, our query plans can be DAGs

2. Instead went for Portals:
   - *Store* operator materializes its input and stores it in Memcom
   - *Read* operator reads the data from Memcom
   - They can be in arbitrary spots of the plan

3. Introduced by Planner:
   - Runs after the regular optimizers (Filter pushdown, Join Order, Projection pushdown …)
   - Supports the frequent patterns:
   - Partitioned Join, Broadcast Join
   - Partitioned GroupBy, Local Pre-Aggregation + Final GroupBy
   - Rewrite Distributed Scan into regular Scan (for broadcasts, less synchronization)
   - Decided by basic cost function (i.e. cardinality)

* Graefe. Encapsulation of parallelism in the Volcano query processing system. SIGMOD Rec. 1990 (Vol 19, Issue 2)

Evaluation

TPC-H with SF100 or SF1000

Hardware (per node):

- 2x Kunpeng 920b
  - Huawei's ARM processor
  - 160 cores (total)
- HCCS (Huawei Cache-Coherent System)
  - Huawei's equivalent to CXL with cache coherency
- 512 GB main memory per node
  - 256 GB reserved for HCCS

Experiment: TPC-H

(4 nodes)

TPC-H SF100:

- Hermes outperforms:
  - DuckDB by 3x
  - TiFlash by 12x
  - SparkSQL by 53x

TPC-H SF1000:

- Hermes outperforms:
  - PolarDB IMCI by 1.5x (recently updated numbers)
  - PolarDB IMCI by 6x (before)

Experiment: Network Stack

#Nodes	Intel/TCP	Kunpeng/HCCS
	Total Time (sec)	
1	10.8	14.1
2	15.4	13.0
3	18.3	10.6
4	20.2	7.8

- Regular TCP overhead too high
- Good performance requires redesigned/optimized network stack

Summary

Hermes can now scale out, with

- Improved query performance
- Higher storage capacity

Modern hardware makes Shared Memory possible for Storage *and* intermediates

Overall, Hermes shows competitive performance.

End of presentation & Hermes project