Design and Evaluation of an RDMA-aware Data Shuffling Operator for Parallel Database Systems
Feilong Liu, Lingyan Yin and Spyros Blanas

Key contributions

- An RDMA-aware data shuffling operator for parallel database systems
  - The endpoint abstraction hides RDMA details
  - Multiple endpoints avoid thread contention

Key result

- Two-sided Send/Receive and unreliable delivery fully utilize the network bandwidth, accelerate TPC-H by 2×

Why is data shuffling important?

- Many queries are bottlenecked on the network bandwidth
- Shuffling needs to fully utilize the network bandwidth

Challenges

- Isolate the complexity of RDMA
  - Manage memory registration
  - Anticipate packets may arrive out of order
  - Support different implementations

- Identify promising design choices
  - Compare two-sided and one-sided primitives
  - Consider both UD and RC transport

Isolate the complexity of RDMA

- Propose the endpoint abstraction:
  - Hides the complexity of synchronization and memory management in RDMA communication
  - One shuffle operator can have one or multiple endpoints
- All functions are thread-safe

The endpoint abstraction hides the complexity of RDMA

Identify promising design choices

- Send/Receive and Read
  - Works with Unreliable Datagram
  - RDMA primitive
    - Single Endpoint
    - Multiple Endpoints
  - Avoids thread contention
  - Uses fewer Queue Pairs

No design choice is strictly better than the others

What is data shuffling?

- SQL Query
- Single node query plan
- Parallel query plan

Two communication patterns: Repartition & Broadcast

Evaluation

- FDR InfiniBand
- EDR InfiniBand
- RDMA is better than MPI

Conclusions

- Two-sided send/receive and unreliable delivery fully utilize the network bandwidth in a database system
- We propose the endpoint abstraction to hide RDMA details
- We design a shuffling operator with multiple endpoints to avoid thread contention, accelerate TPC-H queries by 2×

code.osu.edu/pythia

This research was partially supported by the National Science Foundation under grants IIS-1422077, IIS-1646483, CNS-1513120 and by a Google Research Faculty Award. The evaluation was conducted in part at the Ohio Supercomputer Center.