UC SANTA CRUZ | Baskin Engineering

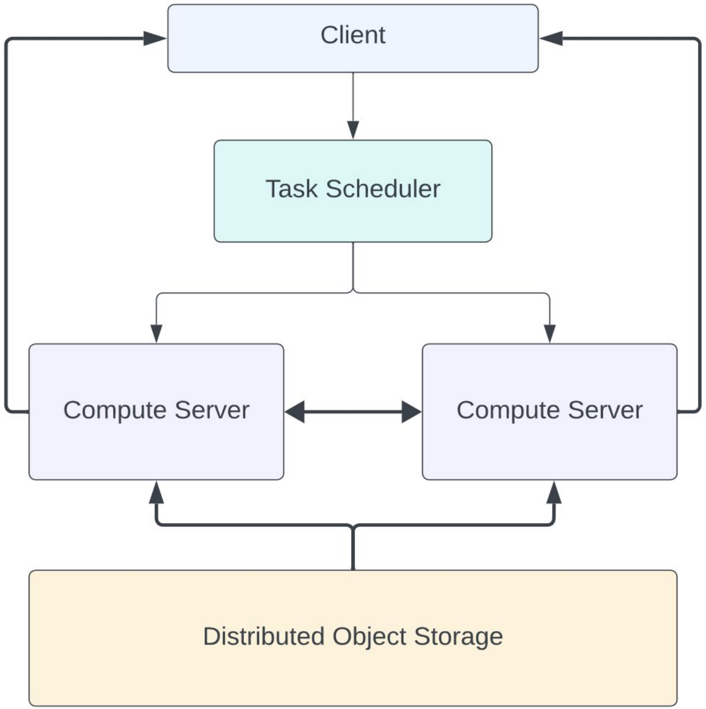
Thallus: An RDMA-based Columnar Data Transport Protocol

Jayjeet Chakraborty*, Matthieu Dorier[#], Philip Carns[#], Robert Ross[#], Carlos Maltzahn*, Heiner Litz*

UC Santa Cruz*, Argonne National Labs[#]

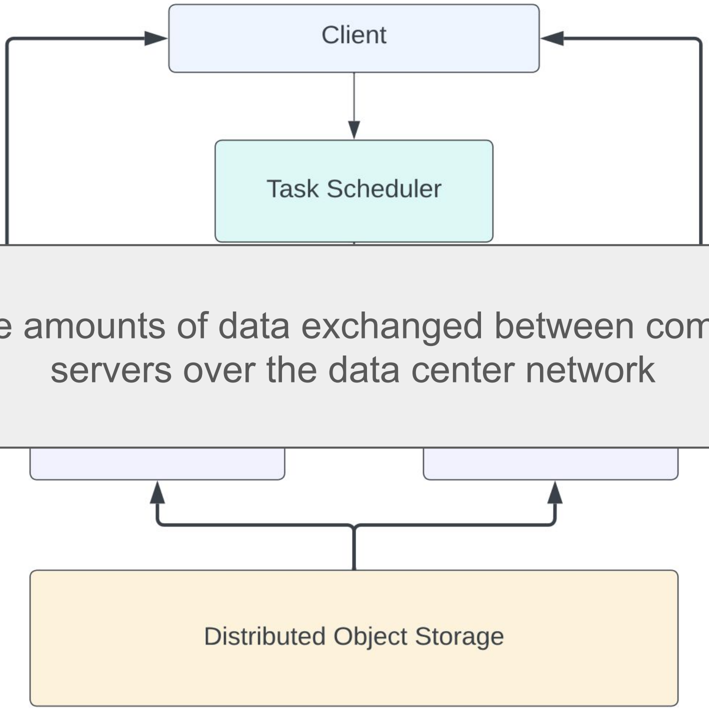
Distributed Data Processing Systems

DASK



APACHE Spark™

Distributed Data Processing Systems



Large amounts of data exchanged between compute servers over the data center network

Problem

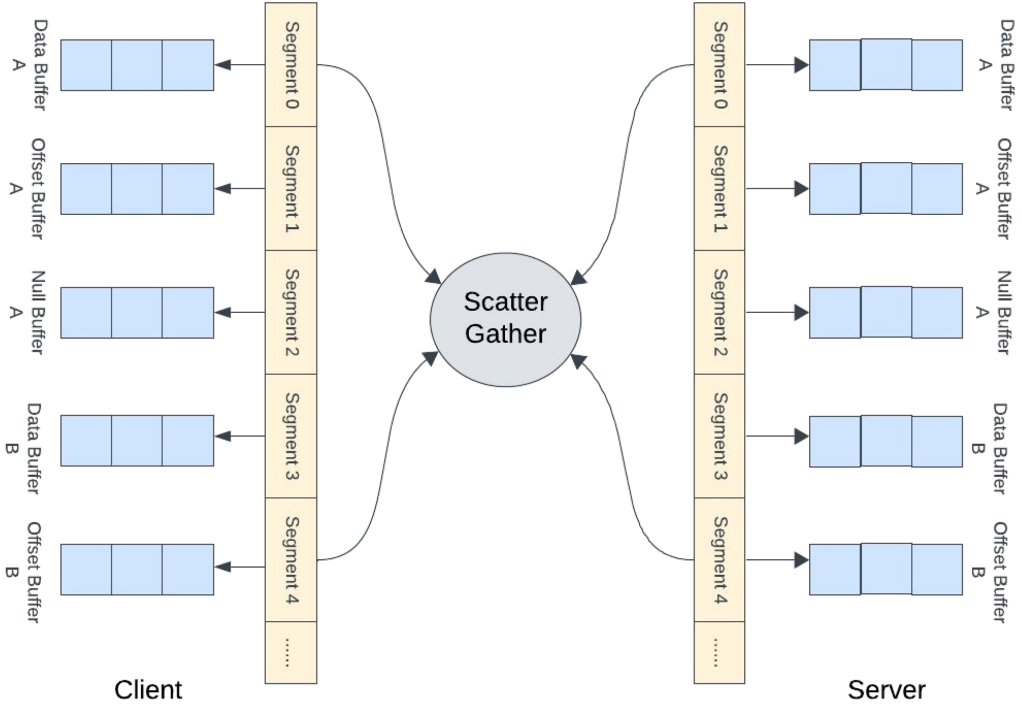
- Modern Fast Hardware
  - Fast multi-core CPUs with large caches, Fast NICs (ConnectX-7 @ 400 Gb/s), Fast SSDs (NVMe PCIe Gen 5 x16 @ 64 GB/s)
- Bottleneck is now in the software stack !
  - For distributed data processing systems, serializing data for transferring over the network has become a new bottleneck (“data center tax”)**
  - *Example*: When transporting Apache Arrow data over RPC, **~30%** of the CPU cycles is spent in serialization (memory copies to lay out buffers contiguously)
- Legacy Data Transport Protocols
  - JDBC, ODBC
  - All TCP/IP-based, RPC-over-Ethernet

** Raghavan et. al, Wolnikowski et. al

Our Approach

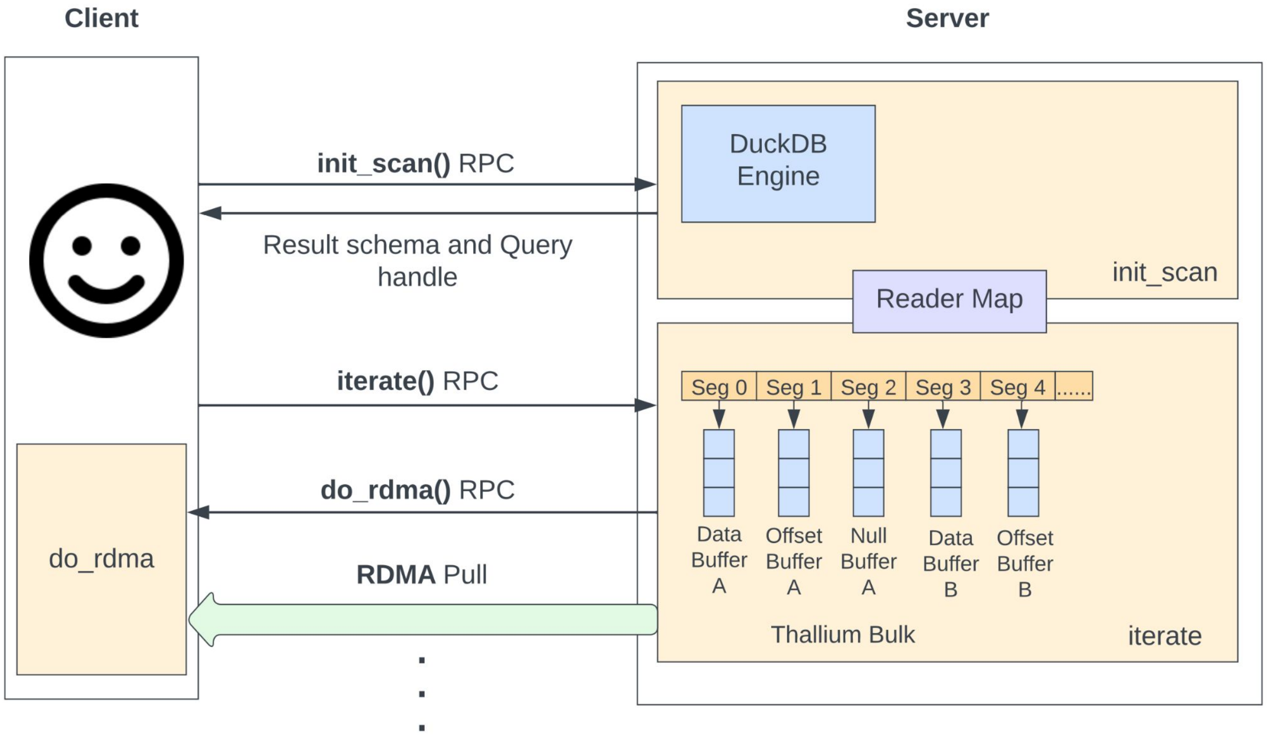
- Leverage hardware accelerated networking technologies such as RDMA-over-Infiniband
    - Squeeze out performance from modern NICs and free up CPUs for other processing tasks
- Exploit the knowledge of memory layout to co-design optimized transport protocols
    - Don’t just treat your data as just a byte blob when you know it’s memory layout
- Solution using Argonne’s Mochi ecosystem
    - **Thallium**: A C++-based RDMA / RPC framework (Memory managed wrapper around libfabric and libibverbs that can utilize Infiniband hardware)
    - Thallus is built using Thallium RDMA framework specialized to transport Apache Arrow record batches

Mapping Data Buffers to RDMA Segments



We map the data, offset, and null buffers to the **3i, 3i + 1, and 3i + 2 th** segment respectively

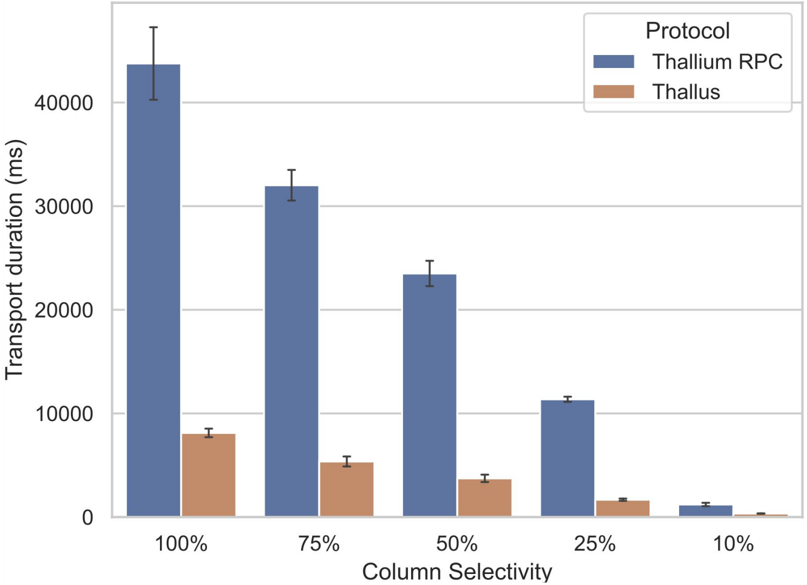
Design and Implementation of Thallus

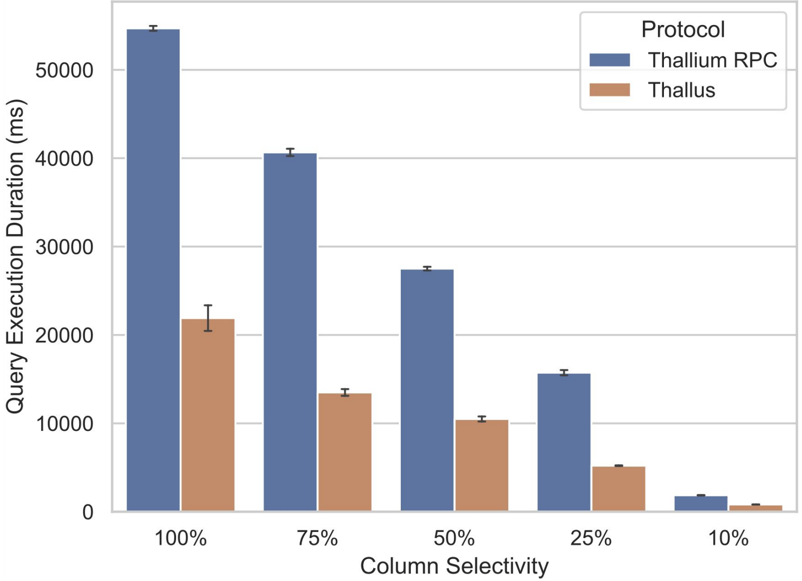


We design our protocol as a client/server but every compute node can act as both

Evaluations

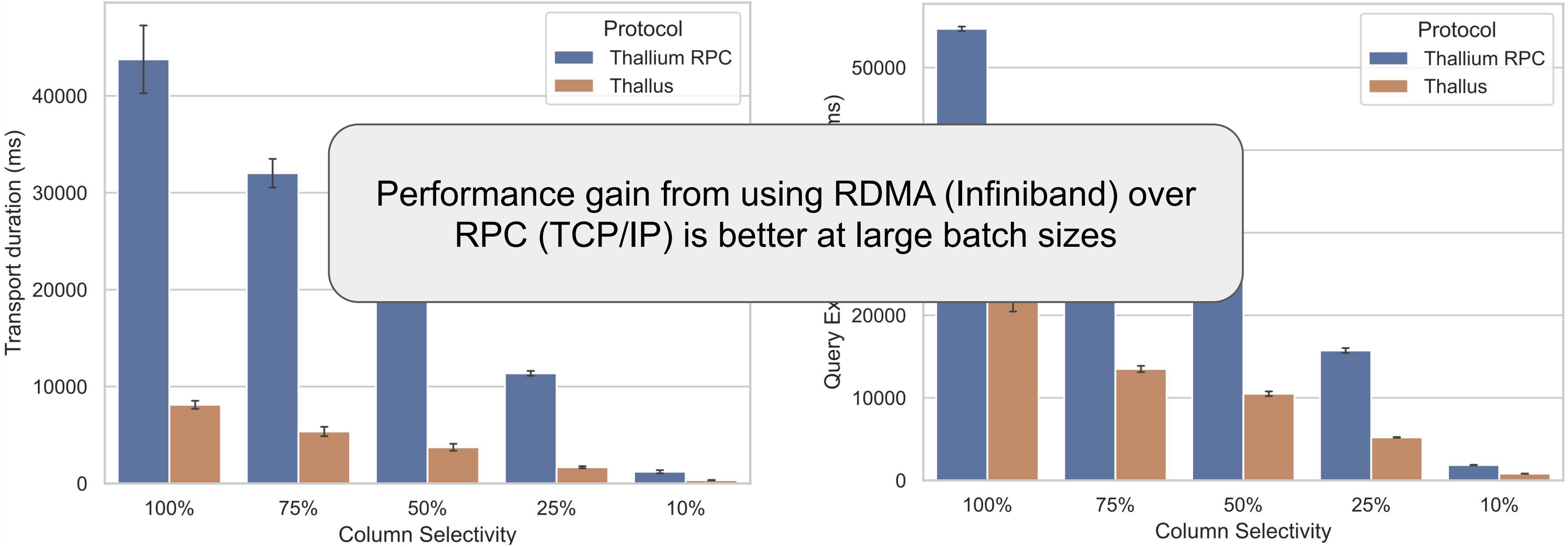
We compare **Thallus** with **Thallium RPC** on NYC Taxi Dataset





Evaluations

We compare **Thallus** with **Thallium RPC** on NYC Taxi Dataset



Performance gain from using RDMA (Infiniband) over RPC (TCP/IP) is better at large batch sizes

Thank You !

Questions?

jayjeetc@ucsc.edu