CHEMNITZ UNIVERSITY OF TECHNOLOGY

A native InfiniBand Transporter for MySQL Cluster

Frank Mietke, Dirk Dunger, Torsten Mehlan, Torsten Höfler and Wolfgang Rehm

Computer Architecture Group
Department of Computer Science
Chemnitz University of Technology

February 8, 2007

CHEMNITZ UNIVERSITY OF TECHNOLOGY

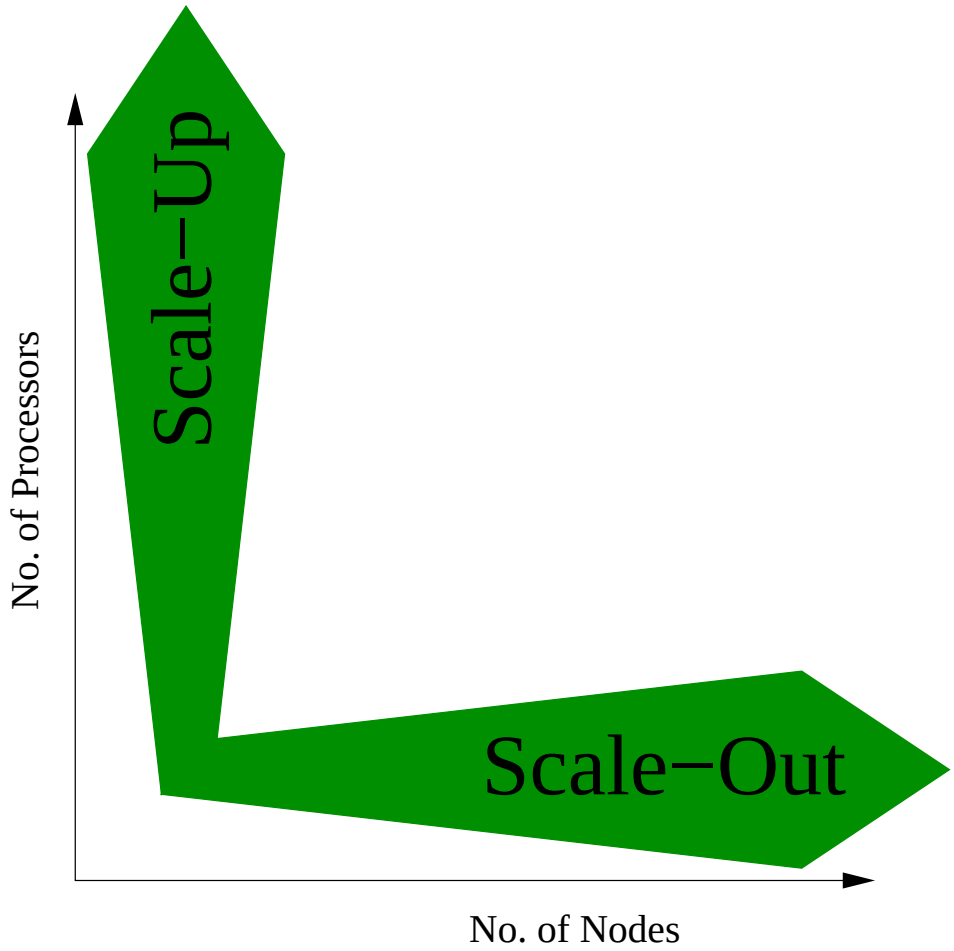
Outline

General Motivation

- Doubling of Storage Needs every 18 Months (IDC)
  (some apps grow faster)
- Faster Internet Connectivity (Consumer) up to 100 Mbit/s
- Ubiquitious for Business Processes
- Globalization – 24/7
- Driven by User Need (Tax Offices)

Scalability Issues

Where to store?



Clustered Database Systems

Examples:

- **Oracle RAC 10g**
  - Shared Storage Approach (SAN, NAS or DAS)
  - Using Local Caching
  - Fast Recovery but Cache Coherence
- **IBM DB2 UDB ESE**
  - Database Partitioning Feature (DPF)
  - Shared Nothing Approach
  - No Increase of Availability
- **MySQL Cluster**
  - Shared Nothing/In-Memory Approach
  - Increase of Availability
  - Based on NDB Cluster

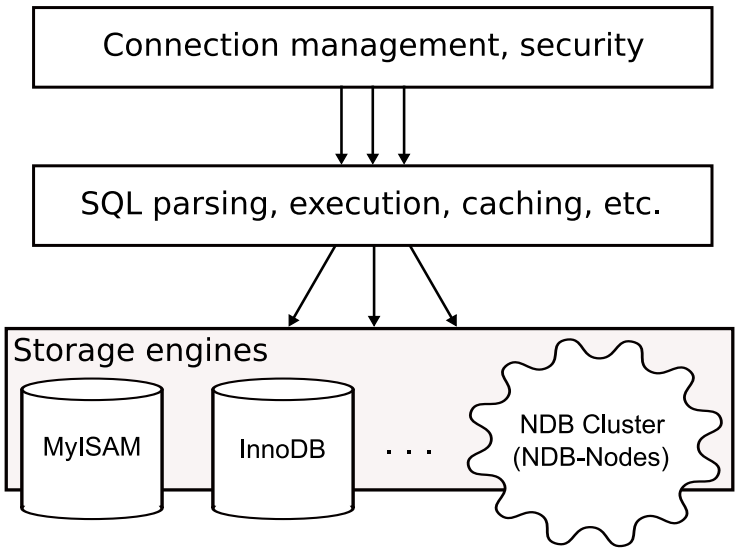
CHEMNITZ UNIVERSITY OF TECHNOLOGY

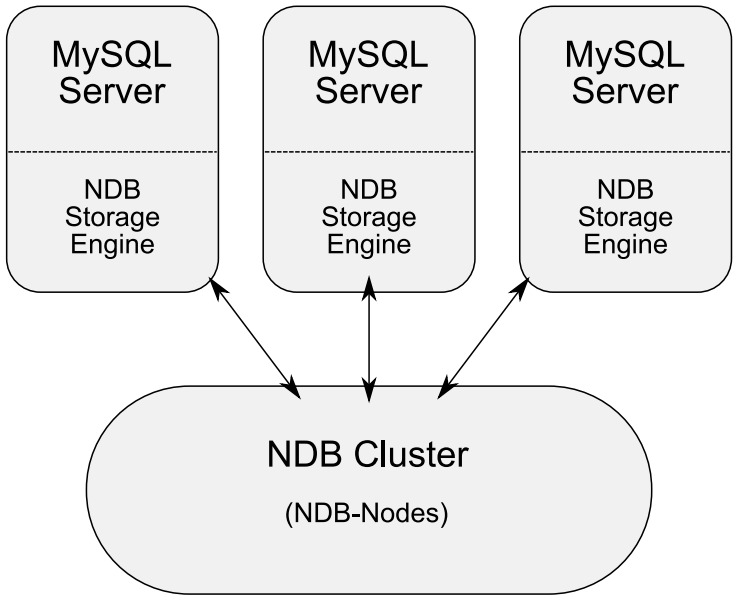
InfiniBand Interconnect

- Most affordable 10Gbit/s Interconnect
- Low Latency
- RDMA Capabilities
- HPC / Storage / Data Center
- Wide Vendor Acceptance (OFA)

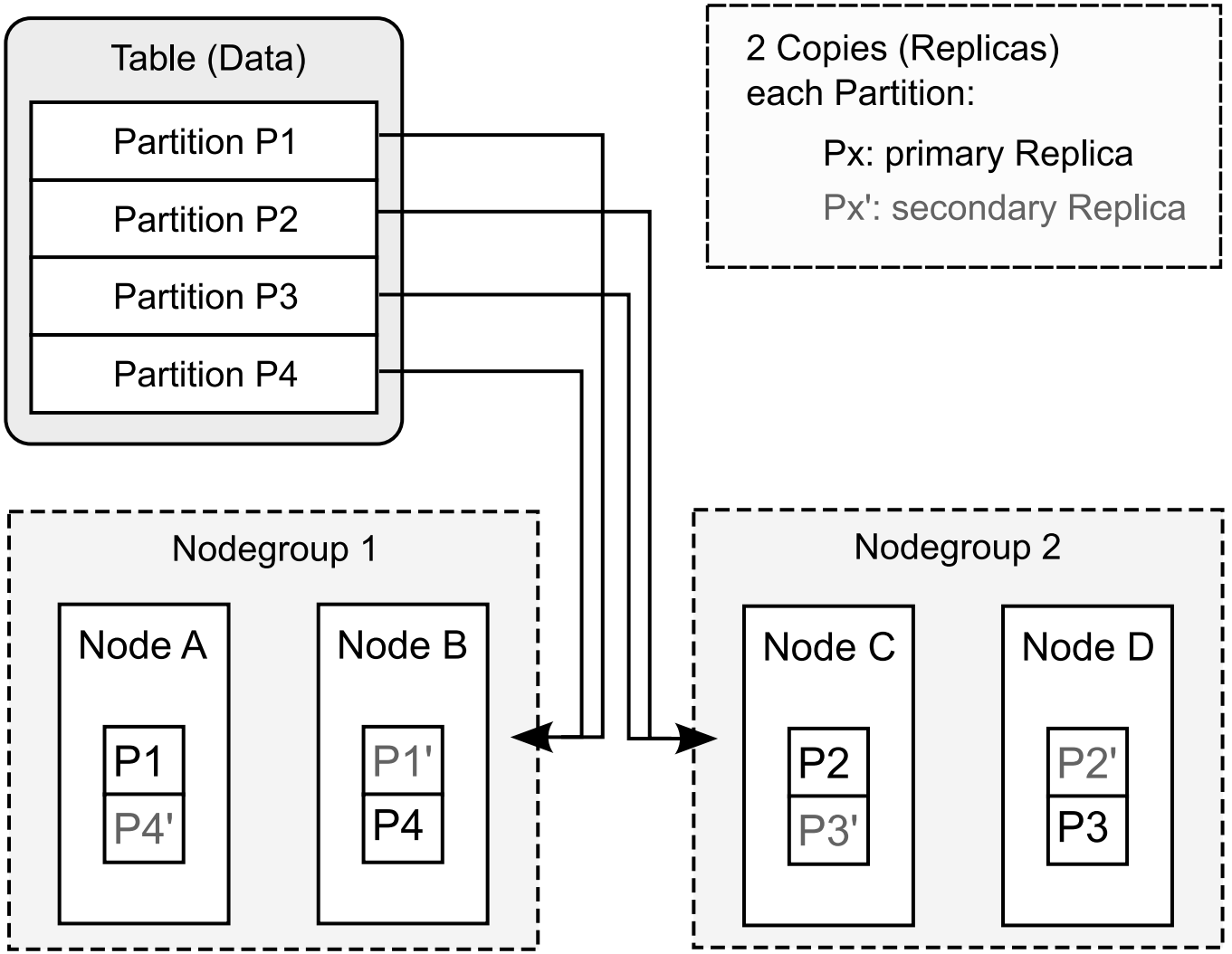
MySQL Cluster

- Transporter for SHM, TCP/IP and SCI
- Network Database (NDB)





Data Partitioning

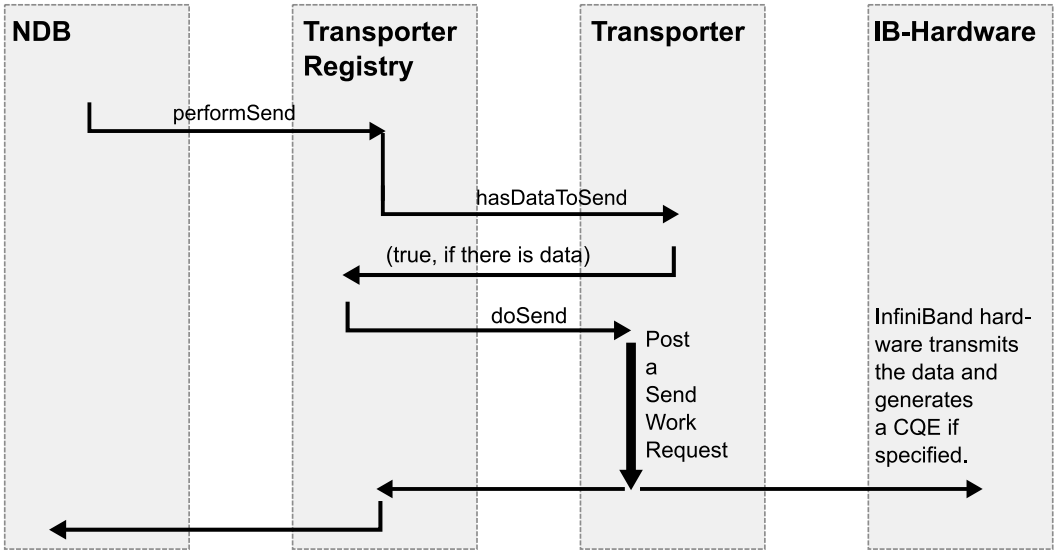


Access the Cluster Data

- **Primary Key:**
  - Using Hash Value
  - One Communication Step
- **Unique Key:**
  - Keyword UNIQUE (Hidden Table)
  - Two Communication Steps
- **Ordered Index:**
  - Using T-Tree (Range Search)
  - Communicate to all NDB Nodes
- **Complete Scan:**
  - SQL Node asks each NDB Node (MySQL 5.x)

Transporter Registry

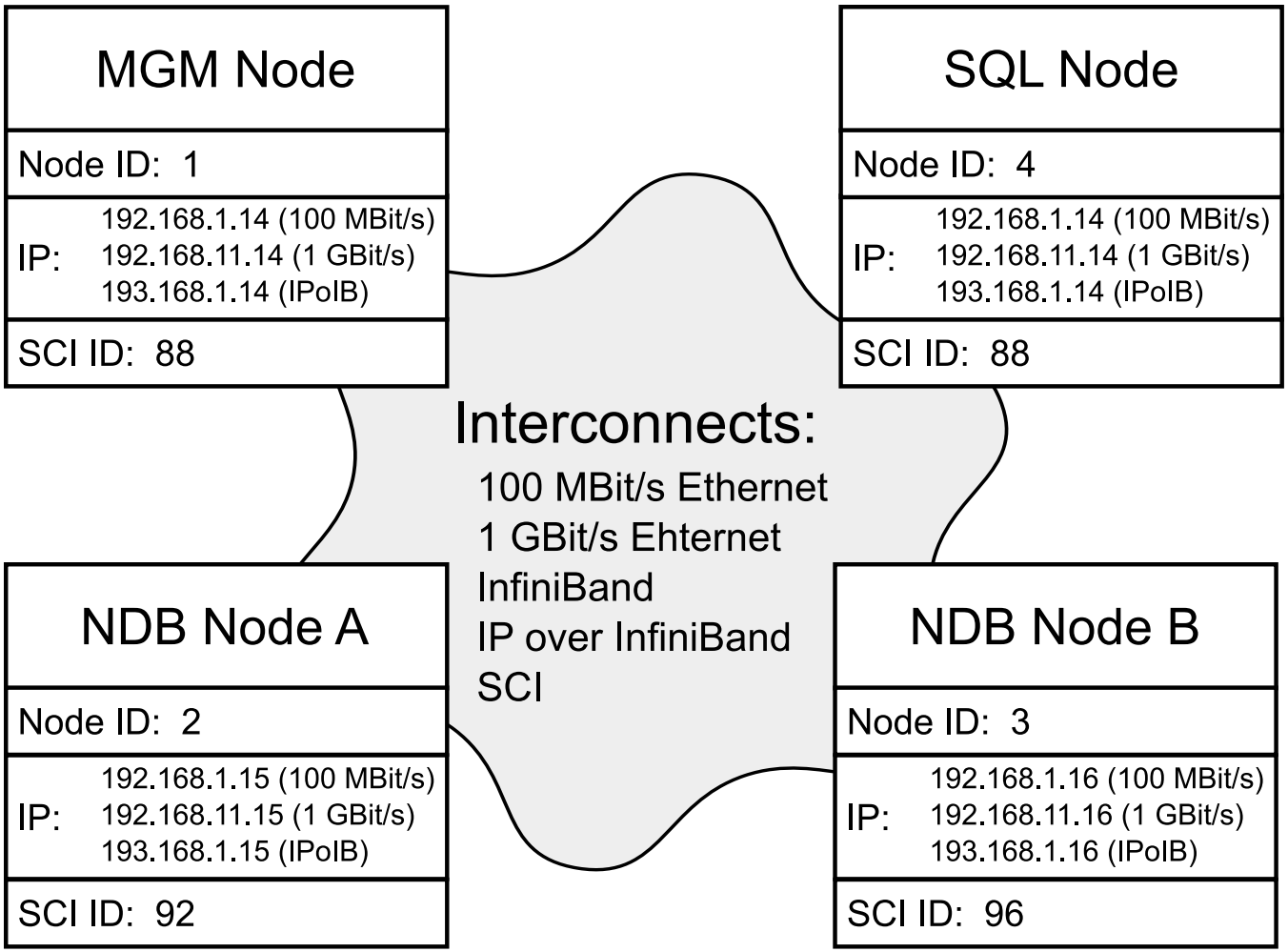
- Handles different Transporter Types (Creation and Management)
- Provides Methods to NDB Process
  - prepareSend, performSend, pollReceive and performReceive



Transporter Implementation

- C++ Class
- Exchange of Information through TCP Channel
- Using Mellanox Verbs API (Channel/Memory Semantic)
- Collecting short Messages
- Handling of CQE (many Send Operations)
- One CQ for all Instances of Transporter Class
- Impact of Buffer Structure (Small Packets)
- Usage of Inline Sends

Benchmark Setup

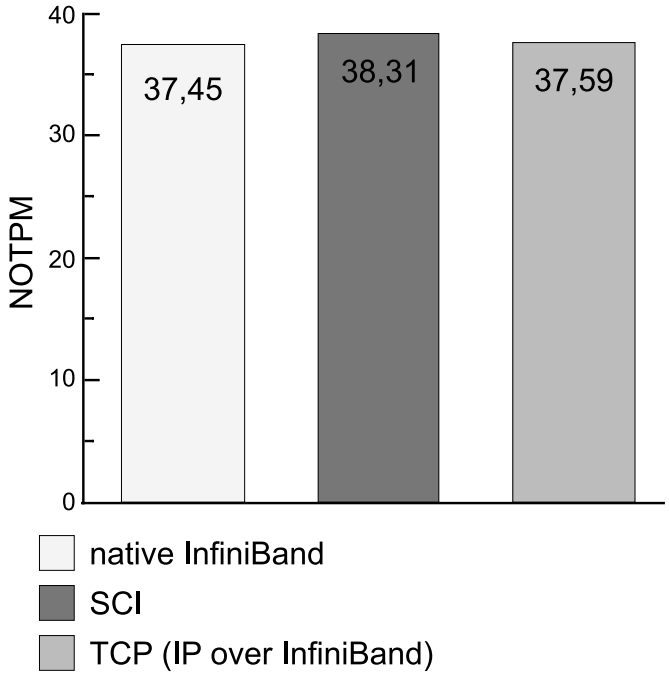


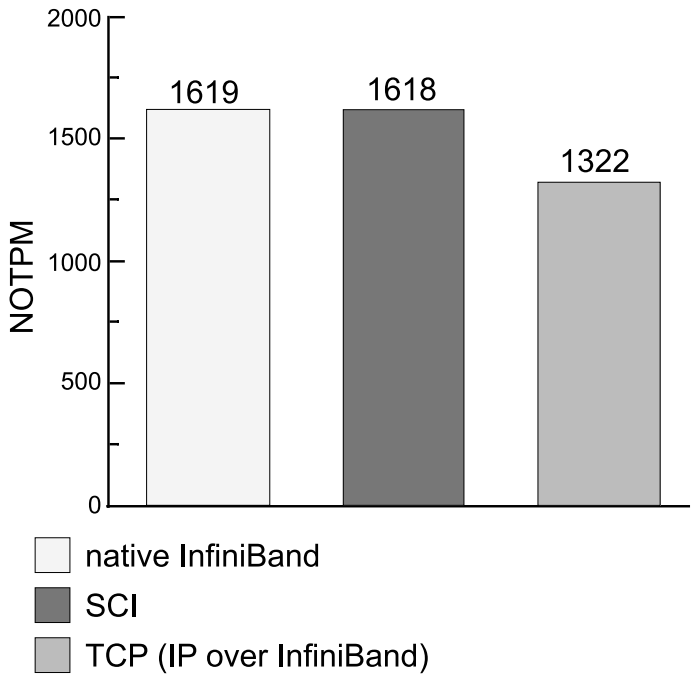
CHEMNITZ UNIVERSITY OF TECHNOLOGY

OSDL Database Test 2

- Based on TPC-C
- Simulates an Online Store
- New Order Transactions per Minute
- Background Transactions
  - ”Delivery”
  - ”Order Status”
  - ”Payment”
  - ”Stock Level”
- Two Modes of Benchmarking
  - Realistic (with Pauses)
  - Full Load

Results OSDL DB Test 2

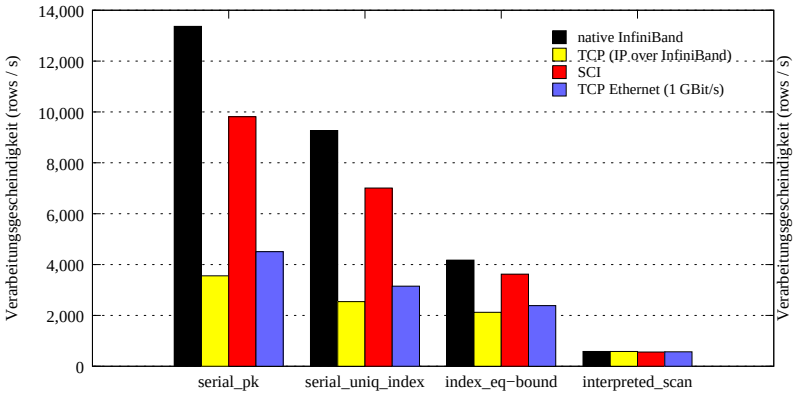


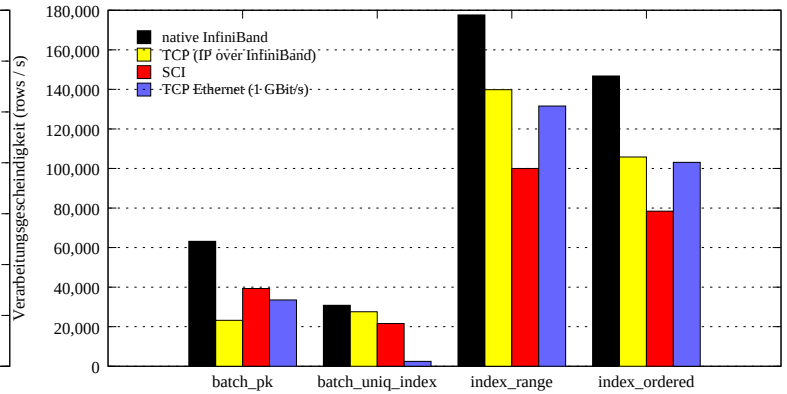


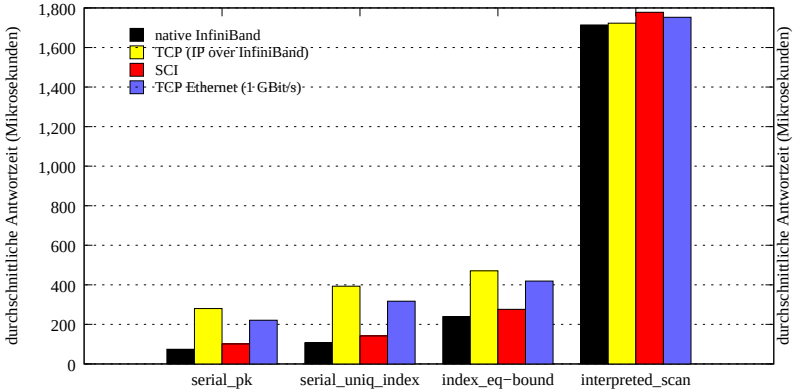
NDB testReadPerf

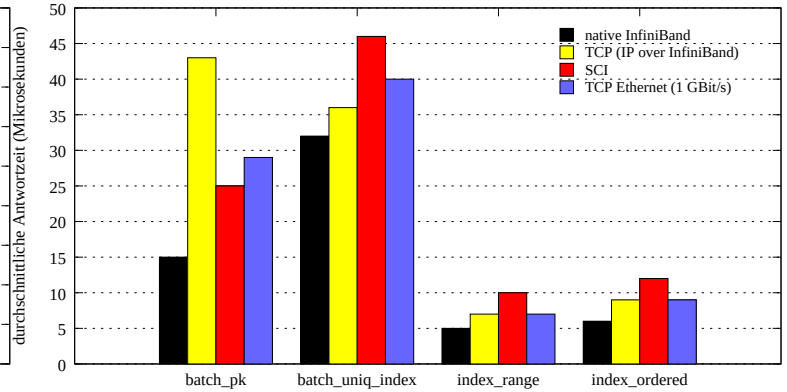
- Measure the Impact of Interconnects on NDB Cluster
- No SQL node necessary
- Benchmarks:
  - serial pk
  - batch pk
  - serial uniq index
  - batch uniq index
  - index eq-bound
  - index range
  - index ordered (batch)
  - interpreted scan

Results NDB testReadPerf









CHEMNITZ UNIVERSITY OF TECHNOLOGY

Summary

- Relatively easy Implementation Hurdles
- Good Replacement for SCI
- Up to 40% faster Response time (testReadPerf)
- Up to 70% more Processing Power (testReadPerf)

Outlook

- Finish RDMA Transporter
- Use OFED Verbs API