Carnegie Mellon Univ.  
Dept. of Computer Science  
15-415/615 - DB Applications  

C. Faloutsos – A. Pavlo  
Lecture#24: Distributed Database Systems  
(R&G ch. 22)

Administrivia

- HW7 Phase 1: **Wed Nov 9th**
- HW7 Phase 2: **Mon Nov 28th**
- HW8: **Mon Dec 5th**
- Final Exam: **Tue Dec 13th @ 5:30pm**

Today’s Class

- High-level overview of distributed DBMSs.
- Not meant to be a detailed examination of all aspects of these systems.

Today’s Class

- Overview & Background
- Design Issues
- Distributed OLTP
- Distributed OLAP
Why Do We Need Parallel/Distributed DBMSs?

• PayPal in 2009…

• Single, monolithic Oracle installation.
• Had to manually move data every xmas.
• Legal restrictions.

Parallel/Distributed DBMS

• Database is spread out across multiple resources to improve parallelism.
• Appears as a single database instance to the application.
  – SQL query for a single-node DBMS should generate same result on a parallel or distributed DBMS.

Parallel vs. Distributed

• **Parallel DBMSs:**
  – Nodes are physically close to each other.
  – Nodes connected with high-speed LAN.
  – Communication cost is assumed to be small.

• **Distributed DBMSs:**
  – Nodes can be far from each other.
  – Nodes connected using public network.
  – Communication cost and problems cannot be ignored.
Database Architectures

• The goal is parallelize operations across multiple resources.
  – CPU
  – Memory
  – Network
  – Disk

Shared Memory

• CPUs and disks have access to common memory via a fast interconnect.
  – Very efficient to send messages between processors.
  – Sometimes called “shared everything”

• Examples: All single-node DBMSs.

Shared Disk

• All CPUs can access all disks directly via an interconnect but each have their own private memories.
  – Easy fault tolerance.
  – Easy consistency since there is a single copy of DB.

• Examples: Oracle Exadata, ScaleDB.
**Shared Nothing**

- Each DBMS instance has its own CPU, memory, and disk.
- Nodes only communicate with each other via network.
  - Easy to increase capacity.
  - Hard to ensure consistency.
- Examples: Vertica, Parallel DB2, MongoDB.

**Early Systems**

- **MUFFIN** – UC Berkeley (1979)
- **SDD-1** – CCA (1980)
- **System R*** – IBM Research (1984)
- **Gamma** – Univ. of Wisconsin (1986)
- **NonStop SQL** – Tandem (1987)

**Inter- vs. Intra-query Parallelism**

- **Inter-Query**: Different queries or txns are executed concurrently.
  - Increases throughput & reduces latency.
  - Already discussed for shared-memory DBMSs.
- **Intra-Query**: Execute the operations of a single query in parallel.
  - Decreases latency for long-running queries.

**Parallel/Distributed DBMSs**

- **Advantages**:
  - Data sharing.
  - Reliability and availability.
  - Speed up of query processing.
- **Disadvantages**:
  - May increase processing overhead.
  - Harder to ensure ACID guarantees.
  - More database design issues.
Today’s Class

• Overview & Background
• Design Issues
• Distributed OLTP
• Distributed OLAP

Design Issues

• How do we store data across nodes?
• How does the application find data?
• How to execute queries on distributed data?
  – Push query to data.
  – Pull data to query.
• How does the DBMS ensure correctness?

Database Partitioning

• Split database across multiple resources:
  – Disks, nodes, processors.
  – Sometimes called “sharding”
• The DBMS executes query fragments on each partition and then combines the results to produce a single answer.

Naïve Table Partitioning

• Each node stores one and only table.
• Assumes that each node has enough storage space for a table.
Naïve Table Partitioning

![Diagram of Naïve Table Partitioning]

**Ideal Query:**
```
SELECT * FROM table
```

Horizontal Partitioning

- Split a table’s tuples into disjoint subsets.
  - Choose column(s) that divides the database equally in terms of size, load, or usage.
  - Each tuple contains all of its columns.
- Three main approaches:
  - Round-robin Partitioning.
  - Hash Partitioning.
  - Range Partitioning.

**Ideal Query:**
```
SELECT * FROM table
WHERE partitionKey = ?
```
Horizontal Partitioning

![Horizontal Partitioning Diagram]

**Ideal Query:**
```
SELECT * FROM table
WHERE partitionKey = ?
```

Vertical Partitioning

- Split the columns of tuples into fragments:
  - Each fragment contains all of the tuples’ values for column(s).
- Use fixed length attribute values to ensure that the original tuple can be reconstructed.
- Column fragments can also be horizontally partitioned.

**Ideal Query:**
```
SELECT column FROM table
```
Replication

- **Partition Replication**: Store a copy of an entire partition in multiple locations.
  - Master – Slave Replication
- **Table Replication**: Store an entire copy of a table in each partition.
  - Usually small, read-only tables.
- The DBMS ensures that updates are propagated to all replicas in either case.

Data Transparency

- Users should not be required to know where data is physically located, how tables are partitioned or replicated.
- A SQL query that works on a single-node DBMS should work the same on a distributed DBMS.

Today’s Class

- Overview & Background
- Design Issues
- Distributed OLTP
- Distributed OLAP
OLTP vs. OLAP

- On-line Transaction Processing:
  - Short-lived txns.
  - Small footprint.
  - Repetitive operations.
- On-line Analytical Processing:
  - Long running queries.
  - Complex joins.
  - Exploratory queries.

Distributed OLTP

- Execute txns on a distributed DBMS.
- Used for user-facing applications:
  - Example: Credit card processing.
- Key Challenges:
  - Consistency
  - Availability

Single-Node vs. Distributed Transactions

- Single-node txns do not require the DBMS to coordinate behavior between nodes.
- Distributed txns are any txn that involves more than one node.
  - Requires expensive coordination.

Simple Example

- Application Server
- Begin
  - Node 1
  - Node 2
Transaction Coordination

- Assuming that our DBMS supports multi-operation txns, we need some way to coordinate their execution in the system.
- Two different approaches:
  - **Centralized**: Global “traffic cop”.
  - **Decentralized**: Nodes organize themselves.

TP Monitors

- Example of a centralized coordinator.
- Originally developed in the 1970-80s to provide txns between terminals + mainframe databases.
  - Examples: ATMs, Airline Reservations.
- Many DBMSs now support the same functionality internally.
Observation

• **Q:** How do we ensure that all nodes agree to commit a txn?
  – What happens if a node fails?
  – What happens if our messages show up late?

CAP Theorem

• Proposed by Eric Brewer that it is impossible for a distributed system to always be:
  – Consistent
  – Always Available
  – Network Partition Tolerant

• Proved in 2002.

CAP – Consistency

All up nodes can satisfy all requests.

Consistency
Availability
Partition Tolerant

Linearizability

No Man’s Land

Still operate correctly despite message loss.

Application Server

Application Server

Master

Replica

Node 1

Node 2

NETWORK

A=1
B=8

A=1
B=8
**CAP – Consistency**

Node 1: Set $A=2, B=9$

Node 2: Master

NETWORK

Application Server

Master

Replica

$A=2$

$B=9$

**Read A,B**

Application Server

Master

Replica

$A=2$

$B=9$

Node 1

Node 2

**CAP – Consistency**

Node 1: Set $A=2, B=9$

Node 2: Application Server

$A=1$

$B=8$

$A=2$

$B=9$

Must see both changes or no changes

**CAP – Availability**

Node 1: Application Server

Master

Replica

$A=2$

$B=9$

$A=1$

$B=8$

Network

Node 2

Node 2: Application Server

$A=1$

$B=8$
CAP – Partition Tolerance

Node 1

Node 2

NETWORK

Application Server

A=1
B=8

Application Server

A=1
B=8

X

Master

Master

Set A=2, B=9

Set A=3, B=6

A=2
B=9

A=3
B=6

Application Server

Application Server

Set A=2, B=9

Set A=3, B=6

A=2
B=9

A=3
B=6

X

Master

Master
**CAP Theorem**

- **Relational DBMSs:** CA/CP  
  – Examples: IBM DB2, MySQL Cluster, VoltDB
- **NoSQL DBMSs:** AP  
  – Examples: Cassandra, Riak, DynamoDB

> These are essentially the same!

**Atomic Commit Protocol**

- When a multi-node txn finishes, the DBMS needs to ask all of the nodes involved whether it is safe to commit.  
  – All nodes must agree on the outcome
- **Examples:**  
  – Two-Phase Commit  
  – Three-Phase Commit *(not used)*  
  – Paxos

**Two-Phase Commit**

1. **Phase 1:** Prepare  
   - Coordinator requests all participants to prepare
   - Participants prepare and report back

2. **Phase 2:** Commit  
   - Coordinator collects prepare responses and commits if all participants agree
   - Application server receives commit request and commits

---

**Two-Phase Commit**

1. **Phase 1:** Prepare  
   - Coordinator sends prepare request to all participants
   - Participants prepare and report back

2. **Phase 2:** Commit  
   - Coordinator collects prepare responses and commits if all participants agree
   - Application server receives commit request and commits
Two-Phase Commit

Phase 1: Prepare

Phase 2: Commit
Two-Phase Commit

Phase 1: Prepare

Node 1
Node 2
Node 3

Phase 2: Abort

Node 1
Node 2
Node 3

OK
ABORT
OK
ABORT
Two-Phase Commit

• Each node has to record the outcome of each phase in a stable storage log.
• **Q:** What happens if coordinator crashes?
  – Participants have to decide what to do.
• **Q:** What happens if participant crashes?
  – Coordinator assumes that it responded with an abort if it hasn’t sent an acknowledgement yet.
• The nodes have to block until they can figure out the correct action to take.

Paxos

• Consensus protocol where a coordinator proposes an outcome (e.g., commit or abort) and then the participants vote on whether that outcome should succeed.
• Does not block if a majority of participants are available and has provably minimal message delays in the best case.

2PC vs. Paxos

• **Two-Phase Commit:** blocks if coordinator fails after the prepare message is sent, until coordinator recovers.
• **Paxos:** non-blocking as long as a majority participants are alive, provided there is a sufficiently long period without further failures.

Distributed Concurrency Control

• Need to allow multiple txns to execute simultaneously across multiple nodes.
  – Many of the same protocols from single-node DBMSs can be adapted.
• This is harder because of:
  – Replication.
  – Network Communication Overhead.
  – Node Failures.
**Distributed 2PL**

- **Set A=2, B=9**
- **Set A=0, B=7**

**Application Server**

**Node 1**

**Node 2**

**Network**

**Recovery**

- **Q:** What do we do if a node crashes in CA/CP DBMS?
- If node is replicated, use Paxos to elect a new primary.
  - If node is last replica, halt the DBMS.
- Node can recover from checkpoints + logs and then catch up with primary.
Today’s Class

• Overview & Background
• Design Issues
• Distributed OLTP
• Distributed OLAP

Distributed OLAP

• Execute analytical queries that examine large portions of the database.
• Used for back-end data warehouses:
  – Example: Data mining
• Key Challenges:
  – Data movement.
  – Query planning.

Distributed OLAP

Application Server

Single Complex Query

Partitions

P1
P2
P3
P4
P5

Distributed OLAP

Application Server

Single Complex Query

Partitions

P1
P2
P3
P4
P5
Distributed Joins Are Hard

```
SELECT * FROM table1, table2
WHERE table1.val = table2.val
```

- Assume tables are horizontally partitioned:
  - Table1 Partition Key → table1.key
  - Table2 Partition Key → table2.key
- **Q**: How to execute?
- Naïve solution is to send all partitions to a single node and compute join.

Semi-Joins

- Main Idea: First distribute the join attributes between nodes and then recreate the full tuples in the final output.
  - Send just enough data from each table to compute which rows to include in output.
- Lots of choices make this problem hard:
  - What to materialize?
  - Which table to send?

Summary

- Everything is harder in a distributed setting:
  - Concurrency Control
  - Query Execution
  - Recovery

Rest of the Semester

- **Mon Nov 28th** – Column Stores
- **Wed Nov 30th** – Data Warehousing + Mining
- **Mon Dec 5th** – SpliceMachine Guest Speaker
- **Wed Dec 7th** – Review + Systems Potpourri

http://cmudb.io/f16-systems