Parallel Databases

R&G Chapter 22

(slides adapted from content by J.Gehrke, J.Shanmugasundaram, and/or C.Koch)

Bloom Joins

- Based on Bloom Filters
  - A technique for "summarizing" sets.
    - Creates a "sketch" that can be used to speed up set membership tests.
- Summary: Use bloom filters to determine which tuples can participate in an equi-join.

Distributed Transactions

- Transactions can update multiple objects
  - … and data is replicated for performance/redundancy
- Isolation challenge: All sites participating in the transaction must be updated.
- Durability challenge: After a failure, some sites may not recover (e.g., Hurricane).

Durability

- New kinds of failure modes:
  - Network Partitions; Some nodes lose connectivity with other nodes.
    - Partial Failures; Some nodes fail permanently, others fail temporarily.
      - Important for replication.
- When have we safely committed?

2-Phase Commit

- Phase 1: Prepare
  - Ensure that all sites can safely commit.
  - Ensure that no site will need to abort.
- Phase 2: Notify
  - Communicate the commit to each site.
- After phase 1 completes successfully, the transaction will never abort.

2-Phase Commit

- One site selected as a coordinator.
  - Initiates the 2-phase commit process.
- Remaining sites are subordinates.

- Only one coordinator per xact.
  - Different xacts may have different coordinators.

2-Phase Commit

- Coordinator sends ‘prepare’ to each subordinate.
- When subordinate receives ‘prepare’, it makes a final decision: Commit or Abort.
- The transaction is treated as if it committed for conflict detection.
- The subordinate logs ‘prepare’, or ‘abort’
- The subordinate responds ‘yes’, or ‘no’

2-Phase Commit

- If coordinator receives ‘no’ from <u>any</u> subordinate, it tells subordinates to ‘abort’.
  - Can treat timeouts as ‘no’s
- If coordinator receives ‘yes’ from <u>all</u> subordinates, it tells subordinates to ‘commit’
- In both cases, the coordinator first logs the decision and forces the log to local storage.

2-Phase Commit

- Subordinates perform abort or commit as appropriate (logging as in single-site ARIES)
- Subordinates 'ack'nowledge the coordinator.
- The transaction is complete once the coordinator receives all 'acks'.

Recovery

- Network Partition (aka Net-Split)
  - What happens in Phase 1? Phase 2?
- Transient (or Permanent) Failure
  - Coordinator in Phase 1? Phase 2?
  - Subordinate in Phase 1? Phase 2?

Recovery

How do we recover from a Net Split?

Recovery

How do we recover from a (transient) coordinator crash in Phase 1?

Recovery

How do we recover from a (transient) coordinator crash in Phase 1?

What information/communication state is lost?

Recovery

How do we recover from a (transient) coordinator crash in Phase 1?

What information/communication state is lost?

Can it be recovered?

Recovery

How do we recover from a (transient) coordinator crash in Phase 1?

What information/communication state is lost?

Can it be recovered?

(Does it need to be?)

Recovery

How do we recover from a (transient) coordinator crash in Phase 2?

Recovery

How do we recover from a (transient) coordinator crash in Phase 2?

What information/communication state is lost?

Recovery

How do we recover from a (transient) coordinator crash in Phase 2?

What information/communication state is lost?

Can it be recovered?

Recovery

How do we recover from a (transient) subordinate crash in Phase 1?

Recovery

How do we recover from a (transient) subordinate crash in Phase 1?

What information/communication state is lost?

Recovery

How do we recover from a (transient) subordinate crash in Phase 1?

What information/communication state is lost?

Can it be recovered?

Recovery

How do we recover from a (transient) subordinate crash in Phase 2?

Recovery

How do we recover from a (transient) subordinate crash in Phase 2?

What information/communication state is lost?

Recovery

How do we recover from a (transient) subordinate crash in Phase 2?

What information/communication state is lost?

Can it be recovered?

3-Phase Commit

- The coordinator is a central point of failure
  - If the coordinator fails, everyone blocked.
- Solution: Allow other sites to take over.
  - Phase 1.5: Precommit; Signals to all nodes that the coordinator is ready to commit.
  - Coordinator waits for N acks before moving to the commit phase.

2PC for Replication

- Optimization: We don't need 100% responses from replicas.
  - Replicas can be reconstructed from others.
  - Asserting 'preparedness' can be difficult.
- How much failure tolerance do we want?
  - We can tolerate N failures by waiting for N+1 responses during the 'prepare' phase.