Spanner: Google's Globally-Distributed Database

Aranya Aryaman

April 4, 2025

1 Introduction

Spanner is a globally distributed database system developed by Google that provides strong consistency, high availability, and horizontal scalability. It is the first system to distribute data at global scale and support externally consistent distributed transactions.

2 Problems Faced

Spanner addresses several fundamental challenges in distributed systems:

- **Consistency vs. Availability**: Traditional distributed systems often had to trade off consistency for availability, as described by the CAP theorem.
- **Global Distribution**: Managing data across geographically distributed datacenters posed issues with latency, replication, and clock synchronization.
- **Scalability**: Handling millions of machines and databases required a highly scalable design.
- **Clock Uncertainty**: Synchronized clocks are necessary for strong consistency, but network latencies and clock skews create uncertainties.
- **Multi-Version Concurrency Control (MVCC)**: Supporting consistent snapshots across globally replicated data was non-trivial.

3 Importance

Spanner is a pioneering system due to several reasons:

- **Globally Consistent Database**: It combines the benefits of relational databases with the scalability of NoSQL systems.
- **External Consistency**: Spanner achieves serializability across wide-area networks, offering stronger guarantees than most distributed databases.
- **TrueTime API**: Introduction of TrueTime, a novel API that provides bounded clock uncertainty, enables external consistency.
- **Applications**: Spanner underpins critical Google services such as AdWords and Google Play, handling billions of dollars of revenue.

4 Architecture

Spanner's architecture includes several key components:

- **Directory-based Data Model**: Data is organized in directories that are the unit of data movement and placement.
- **Paxos State Machines**: Spanner uses Paxos to replicate each directory across multiple servers.
- **TrueTime**: A key innovation providing tight bounds on clock uncertainty using GPS and atomic clocks.
- **Two-phase Commit with TrueTime**: Transactions use two-phase commit enhanced by TrueTime to guarantee external consistency.
- **Automatic Sharding and Rebalancing**: Spanner automatically partitions data into tablets and rebalances them across servers based on load and size.
- **Replication and Failure Recovery**: Data is synchronously replicated across datacenters for fault tolerance.

5 Future Work

The Spanner paper and subsequent research indicate multiple avenues for future work:

- **Reducing Clock Uncertainty**: Further reducing the ϵ bounds in TrueTime would allow lower commit latencies.
- **Improved Multi-Tenancy**: Supporting more varied workloads while maintaining strong guarantees.
- **Relaxing Consistency for Specific Applications**: Allowing applications to choose consistency levels dynamically for better performance.
- **Integration with Machine Learning Workloads**: Scaling Spanner for hybrid transactional/analytical processing (HTAP) systems.
- **Increasing Geographic Coverage**: Expanding infrastructure to support even broader global distribution without increasing latencies.